Development and application of an enhanced sampling molecular dynamics method to the conformational exploration of biologically relevant molecules

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2017

Irfan Alibay

School of Health Sciences / Division of Pharmacy and Optometry

Contents

Final word count: 39864

List of Figures

List of Tables

Abbreviations

Å	Angstrom
AMBER	Assisted Model Building with Energy Refinement
aMD	Accelerated molecular dynamics
amu	Atomic mass unit
CPU	Central processing unit
CUDA	Compute unified device architecture
ESP	Electrostatic potential
GAFF	Generalised AMBER force field
GAG	Glycosaminoglycan
GalNAc	N-acetylgalactosamine
GlcA	Glucuronic acid
GlcNAc	N-acetylglucosamine
GPU	Graphics processing unit
GROMACS	Groningen machine for chemical simulations
HMR	Hydrogen mass repartitioning
HS	Heparan sulfate
IdoA	iduronic acid
K	Kelvin
MAE	Mean absolute error
MD	Molecular dynamics
MM	Molecular mechanics
MPI	Message passing interface
ms	millisecond
msesMD	Multi-dimensional swarm-enhanced sampling molecular dynamics
NMR	Nuclear magnetic resonance
NPT	Isobaric-Isothermal ensemble
ns	nanosecond
NVT	Canonical ensemble
OpenMP	Open multi-processing
PMF	Potential of mean force

QM	Quantum mechanics
RESP	Restrained electrostatic potential
RMSD	Root-mean-squared deviation
sesMD	Swarm-enhanced sampling molecular dynamics
WHAM	Weighted histogram method
α-Glc	α-D-glucose
β-Glc	β-D-glucose
μs	microsecond

Abstract

This thesis describes the development a new swarm-enhanced sampling methodology and its application to the exploration of biologically relevant molecules. First, the development of a new multi-dimensional swarm-enhanced sampling molecular dynamics (msesMD) approach is detailed. Relative to the original swarm-enhanced sampling molecular dynamics (sesMD) methodology, the msesMD method demonstrates improved parameter transferability, resulting in more extensive sampling when scaling to larger systems such as alanine heptapeptide. The implementation and optimisation of the swarm-enhanced sampling algorithms in the AMBER software suite are also described. Through the use the newer *pmemd* molecular dynamics (MD) engine and asynchronous MPI routines, speedups of up to three times the original sesMD implementation were achieved.

The msesMD method is then applied to the investigation of carbohydrates, first looking at rare conformational changes in Lewis oligosaccharides. Validating against multi-microsecond unbiased MD trajectories and other enhanced sampling methods, the msesMD simulations identified rare conformational changes leading to the adoption of non-canonical unstacked core trisaccharide structures. Next, the use of msesMD as a tool to probe pyranose ring pucker events is explored. Evaluating against four benchmark monosaccharide systems, msesMD simulations accurately recover puckering details not easily obtained via multi-microsecond unbiased MD. This was followed by an exploration of the impact of ring substituents on conformation in glycosaminoglycan monosaccharides: through msesMD simulations, the influence of specific sulfation patterns were explored, finding that in some cases, such as 4-O-sulfation in N-acetyl-galactosamine, large changes in the relative stability of ring conformers can arise. Finally, the msesMD method was coupled with a thermodynamic integration scheme and used to evaluate solvation free energies for small molecule systems. Comparing against independent trajectory TI simulations, it was found that although the correct solvation free energies were obtained, the msesMD based method did not offer an advantage over unbiased MD for these small molecule systems. However, interesting discrepancies in free energy estimates arising from the use of hydrogen mass repartitioning were found.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning.

Copyright Statement

The Author

The author attained an MPharm in Pharmacy from the University of Manchester in 2012. Aside from a short Master's project as part of his MPharm, the author has no other research experience aside from that gained whilst undertaking the research presented in this thesis during the period of 2013 to 2017.

Acknowledgements

My supervisor:

Dr Richard Bryce for his generous continued guidance, support and patience.

Molecular modelling lab:

Dr Kepa Burusco for originally conceptualising the multi-dimensional swarm-enhanced sampling method, in addition to many useful discussions.

All other members of molecular modelling lab, both past and present for their support.

School Health Sciences, Division of Pharmacy & Optometry:

For funding this project and providing necessary training and support.

HPC services:

Manchester University IT services and the use of the Computational Shared Facility.

N8 consortium and EPSRC (Grant No. EP/K000225/1) for access to the N8 HPC.

HECBIOSIM and EPSRC (Grant No. EP/L000253/1) for access to ARCHER.

Family and friends:

My parents and sister; Nadine Soundarjee, Nazir Alibay and Farah Alibay for their infinite support and motivation.

My close friends; Jamie Holcroft and Chaan Iqbal, for the many useful discussions and endless hours of entertainment.

Daving Topping; for the many useful discussions and the constant supply of coffee.

Most importantly Jenny Zhou; for your continued love and support, for the millions of things which you have done for me, no words can describe how grateful I am.

Safety:

To all those whom I have unintentionally omitted, my apologies and thank you.

Dedication

Dedicated to my parent, grandparent, great-grandparents and all those before them both maternal and paternal, who sacrificed so much in order to ensure the happiness of their children.

Thank you for allowing me to pursue my dreams.

Chapter 1: Introduction

1.1 Thesis Introduction

Atomistic molecular dynamics (MD) simulations are a useful theoretical tool in understanding the time evolution of systems of interest. Through recent advances in hardware utilisation, access to long simulation timescales of up to microseconds for large systems is now possible using commodity hardware.[1-2] Unfortunately, even at such timescales, conventional MD simulations are frequently limited in their ability to effectively sample rare events.[3] Although it is possible to overcome this limitation by accessing even longer multi-microsecond to millisecond simulation lengths, achieving this within realistic time frames usually requires either highly specialised hardware[4] or distributed computing schemes.[5] To address this, enhanced sampling methods have been developed[6-10], allowing for easier traversal of energetic barrier, and consequently improved exploration of phase space. Of particular focus in this thesis is a recently introduced enhanced sampling scheme based on swarm intelligence, termed swarm-enhanced sampling molecular dynamics (sesMD).[11-12]

The aim of this research is to develop and apply an efficient high performance variant of the sesMD methodology for use in routinely exploring rare conformational events in biologically relevant molecules. This is achieved through the development of the multi-dimensional swarm-enhanced sampling (msesMD) scheme, which generalises the sesMD potential to improve sampling and reduce parameterisation costs. The msesMD methodology is then applied as a tool to explore the conformational behaviour of carbohydrates and the solvation free energies of small molecules.

Chapter 2 of this thesis details the development efforts in creating the msesMD method, including its implementation and optimisation within the AMBER molecular dynamics software package. Also discussed are some of the practical aspects of running swarm-enhanced sampling simulations. In chapter 3, the msesMD protocol is used to explore the

behaviour of Lewis oligosaccharides, detailing rare transitions away from canonically "closed" conformational states. Chapter 4 describes the validation and application of swarm-enhanced sampling to boost sampling of puckering in hexapyranose rings. This is then used to investigate the impact of post-translational ring modifications on the puckering behaviour of glycosaminoglycan monosaccharides. Moving away from carbohydrates, in chapter 5 the msesMD method is coupled with thermodynamic integration in order to explore the solvation free energies of small molecules. Overall conclusions are presented in chapter 6.

In the remainder of this chapter, some of the theoretical background surrounding this thesis work is briefly detailed, introducing key concepts and methods which will be used in later chapters.

1.2 Theoretical background

1.2.1 Modelling atomistic behaviour

In order to accurately describe an atomic system, a suitable set of potential energy functions detailing the behaviour of this system as a function of its coordinates must be chosen. Ideally one would use Schrodinger's equation[13] (equation 1.1), fully accounting for the complete electronic behaviour of the system at the quantum mechanical (QM) level of theory. However, doing so is intractable for systems of greater size than a few atoms molecular systems. Even using approximations to ab initio QM approaches, such as the semi-empirical neglect of diatomic differential overlap (NDDO) methods (e.g. AM1[14] and PM3[15]) is impractically time consuming when attempting to describe the time evolution of a large system.

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V\right\}\Psi(r) = E\Psi(r) \qquad [1.1]$$

Instead, empirical force field methods are typically used which disregard the explicit treatment of electrons in a system opting for a simpler model which purely treats the molecule(s) mechanistically based on its atomic coordinates. As shown in equation 1.2, a standard force field treats the potential energy $U(r)$ of molecular systems as a set of bonded (stretching, angle and dihedral) and non-bonded (electrostatic and van der Waals) interactions.[16]

$$U_{total} = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + \cos(n\phi - \gamma)\right] + \sum_{i<j}\left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} + \left(\frac{B_{ij}}{r_{ij}}\right)^{6} + \frac{q_i q_j}{\varepsilon r_{ij}}\right] \qquad [1.2]$$

where r is the bond length, θ is the angle formed between three bonded atoms, ϕ is the dihedral formed between four bonded atoms, r_{eq}, θ_{eq}, K_r, K_θ, V_n, γ, and n are constants for bonded terms, and q_i is the partial charge of atom i, A_{ij} and B_{ij} are Lennard-Jones parameters and r_{ij} the distance between atoms i and j; and ε is the dielectric constant. Assuming suitable parameter fitting (usually involving QM calculations), the resultant force field can typically describe accurately the behaviour of the class of system for which it was parameterised. Thus force fields have generally been developed for specific types of molecules, such as proteins, nucleic acids or carbohydrates. Although it is noted that force fields designed to be used with arbitrarily diverse sets of molecules have been successfully developed and used, such as the Generalised AMBER force field[17] and the CHARMM General Force Field.[18]

While usually effective, force fields do also suffer from some limitations. One such example is the treatment of electrostatics; in most cases, force fields use a static representation of the partial charges on the different atoms of the system. Therefore the charge distribution along a molecule modelled via a force field does not alter based on changes in the local environment. While this usually can be accounted for by using ensemble averaged charges to represent the most prominent conformations of a molecule, it can still be limiting in very polar systems. To address this issue, a newer generation of polarisable force fields have been developed such as the Drude[19] and AMOEBA[20-22] models. Whilst these have been proven quite effective, these methods tend to be limited in their computational efficiency.[23] Another limitation of force fields is an inability to handle the dynamic formation and breaking of bonds, although this again is being addressed by so-called third generation force fields such as ReaxFF.[24]

1.2.2 Molecular dynamics

Static representations of a molecule or assembly of molecules provide very little information about the propensity of the system to adopt different conformational arrangements. To do this, we must be able to describe its time evolution based on atomic coordinates and a given set of ensemble conditions. This can be achieved via the integration of the classical Newtonian equation of motion (equation 1.3).

$$F = m\mathbf{a} = \left(m\frac{\delta^2 r}{\delta t^2}\right) \quad [1.3]$$

Numerically, this can be achieved via integration schemes such as the velocity Verlet and the leapfrog methods (equations 1.4a-b and 1.5a-b respectively).[25] In the AMBER MD engines used in this thesis, use a special variant of the velocity Verlet integration.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t\boldsymbol{v}(t) + \frac{1}{2}\delta t^2\boldsymbol{a}(t) \quad [1.4a]$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \frac{1}{2}\delta t[\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)] \quad [1.4b]$$

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \delta t\boldsymbol{v}(t+\frac{1}{2}\delta t) \quad [1.5a]$$

$$\boldsymbol{v}\left(t+\frac{1}{2}\delta t\right) = \boldsymbol{v}\left(t-\frac{1}{2}\delta t\right) + \delta t\boldsymbol{a}(t) \quad [1.5b]$$

where **r**(*t*) represents a given set of coordinates at time *t*, **v**(*t*) their velocities, **a**(*t*) their acceleration, δt the time step over which the equations of motion are integrated. Generally the value of the time step, δt, is based on the fastest motion of the system, i.e. frequency of hydrogen bond vibration.[13, 26] Thus, if one explicitly takes into account the bonded motion

of the hydrogens in a system, a maximum time step of 1 fs can be used. However, if one uses constraints such as SHAKE[27] to limit the motion of these hydrogen bonds, then the time step can be increased up to 2 fs. This is unfortunately still quite small and requires millions of integration steps required to sample a few nanoseconds of simulation time. To overcome this, long time step methods have been introduced such as the multiple time step RESPA method.[28]One particularly popular method, hydrogen mass repartitioning (HMR), is used throughout this thesis. As originally described by Feenstra et al.[29], the idea is to increase the mass of hydrogens by transferring part of the mass of neighbouring heavy atoms to the hydrogen. This slows down the motion of the hydrogens and allows for the use of higher time steps, up to around 4 fs. Ideally, since the ensemble averages are mass independent[26], the use of HMR should not have much of an impact on the outcomes of a simulation. Whilst Feenstra et al.[29] did note some discrepancies in water box simulations, particularly surrounding the formation of hydrogen bonds between water molecules, these do not appear to be seen in practice.[26, 30-31] In this thesis, we use the default AMBER approach to HMR, as defined by Hopkins et al.[26, 32] where the solute hydrogen masses are increased by 3 amus, whilst the solvent masses are unaltered and handled via the SETTLE[33] algorithm.

As the time evolution of a system progresses, it will be able to access new configurational states in phase space. Eventually, given ergodic sampling and an appropriate ensemble, the probability of occupying any given microstate can be found. In the canonical ensemble, this probability density of all states, ρ, is given by[34]

$$\rho = Nexp\left[\frac{\psi - E}{\theta}\right] \qquad [1.6]$$

where N is the total number of microstates in the ensemble, E is the energy as a function of the coordinates and momenta, and ψ and θ are parameters of the distribution. This can be expressed in terms of the partition function Z[25]

$$Z = \sum_i \exp[-\beta E_i] \qquad [1.6]$$

where β is the reciprocal of the Boltzmann constant multiplied by the temperature and E_i is the total energy of microstate i. Given an appropriate approximation of the microstate density, correct estimations of the true ensemble averages of a given system can be obtained. For example, by knowing the microstate density of two states of interest, ρ_i and ρ_j, the relative Helmholtz free energy ΔA between them can be found[34]

$$\Delta A = -k_B T ln\left(\frac{\rho_i}{\rho_j}\right) \qquad [1.7]$$

Unfortunately, obtaining estimates of the microstate density is not always an easy task. Access to different probable states can often be hindered by the presence of low probability states (i.e. energy barriers). Thus it is sometimes possible that at the limit of the time scales which can be simulated, some microstates are not visited, thus resulting in non-ergodic sampling.[34] This is the point where enhanced sampling methods can be useful in overcoming these energetic barriers and achieving better sampling of phase space.

1.2.2 Enhanced sampling methods

As detailed above, conventional molecular dynamics methods are often limited in their ability to sample due to large energetic barriers preventing transitions between stable basins of interest. While this can be resolved by using extended simulation times, it often requires calculation lengths which are impractical. Additionally, without *a priori* knowledge of the existence of a specific state, one could easily assume that a system has been fully sampled even though important, albeit rarely accessed, microstates were not found. To seek to address this issue, several enhanced sampling methods have been proposed such as accelerated MD[6, 35-37], umbrella sampling[38], replica exchange schemes[8], metadynamics[7][and the swarm-enhanced sampling MD method[11-12]. The concept of enhanced sampling can be generalised as the idea of introducing a bias (perturbative or not), which allows the system under investigation to traverse energy barriers and explore larger portions of phase space. Whilst enhanced sampling approaches can offer a much more efficient approach to sampling phase space than convention MD, they can also come at the cost of increased complexity and some caveats in their application. Examples of such caveats include increased uncertainty due to reweighting[37]; system-dependent choices in boost coordinates and parameters[39]; and increased computational costs[40].

Below, we briefly detail selected enhanced sampling methods, which have been used in this thesis.

1.2.2.1 Accelerated molecular dynamics

The concept of accelerated molecular dynamics (aMD) is to apply additional potential energy non-specifically to a system, i.e. boosting all atoms in the system, in order to allow it to escape from potential energy minima. This is achieved by creating an artificial potential energy landscape, where the potential energy of a system is raised by a boost potential, $\Delta U(\boldsymbol{r})$, if the system potential energy is found to be below a given threshold energy, E[6, 36-37]

$$U^*(\boldsymbol{r}) = \begin{cases} U(\boldsymbol{r}) \; if \; U(\boldsymbol{r}) \geq E \\ U(\boldsymbol{r}) + \Delta U(\boldsymbol{r}) \; if \; U(\boldsymbol{r}) < E \end{cases} \quad [1.8]$$

where $U(\boldsymbol{r})$ is the potential energy of a system, $U_{total}(\boldsymbol{r})$, in a given state. The form of the boost potential, $\Delta U_{total}(\boldsymbol{r})$, is defined by,

$$\Delta U_{total}(\boldsymbol{r}) = \frac{(E_{total} - U_{total}(\boldsymbol{r}))^2}{\alpha_{total} + (E_{total} - U_{total}(\boldsymbol{r}))} \quad [1.9]$$

where α_{total} is a parameter that controls the shape of the boost potential. In addition to boosting along the total potential energy of a system, it is also possible to define $U(\boldsymbol{r})$ to be the total potential energy of the dihedrals of the system, $U_{dihedral}(\boldsymbol{r})$.

$$\Delta U_{dihedral}(\boldsymbol{r}) = \frac{(E_{dihedral} - U_{dihedral}(\boldsymbol{r}))^2}{\alpha_{dihedral} + (E_{dihedral} - U_{dihedral}(\boldsymbol{r}))} \quad [1.10]$$

Both equations 1.9 and 1.10 can be combined, leading to the so-called "dual-boost" aMD method[35]. Overall, if using the dual-boost aMD method, this means that four parameters must be provided for this method (α_{total}, $\alpha_{dihedral}$, E_{total}, $E_{dihedral}$). Thankfully empirical strategies for the derivation of these parameters have been previously proposed.[37]Reweighting of the biased distributions of observables, $X(\mathbf{r})$, can be achieved via the an exponential reweighting method:

$$\langle X \rangle = \frac{\langle X(\boldsymbol{r}) e^{\beta \Delta U(\boldsymbol{r})} \rangle}{\langle e^{\beta \Delta U(\boldsymbol{r})} \rangle} \quad [1.11]$$

where $\langle X \rangle$ is the ensemble average of observable X. One of the main advantages of aMD is that it does not require a specific boost coordinate to be defined, as it explicitly boosts all atoms in the system. However, this can lead to large fluctuations in the aMD biasing

potential energy, leading to inaccurate reweighting via the exponential reweighting method. To avoid this, new versions of aMD[36, 41] and improved reweighting methods have been proposed.[37]

1.2.2.2 Umbrella sampling

The main idea of the umbrella sampling method[10] is to use a biasing potential to sample different windows along a coordinate pathway of interest. This biasing potential usually takes the form of a harmonic restraint potential which keeps the system within the desired window. An example of such a harmonic potential is[42]

$$U^{bias}(\xi) = \frac{1}{2}k(\xi - \xi_0) \qquad [1.12]$$

where k is a force constant defining the restraint potential, ξ_0 is the target reaction coordinate for the window of interest, and ξ is the value of the reaction coordinate sampled by the system. Assuming good coverage of all windows along a path, it is possible to reweight the probability densities of occupying the different windows along a path and generate a potential of mean force for this path. Reweighting is usually achieved via the Weighted Histogram Analysis Method[43-44] (WHAM).

1.2.2.3 Swarm-enhanced sampling molecular dynamics

Swarm intelligence describes the behaviour exhibited by several decentralised individuals working cooperatively to achieve a collective goal. Originally characterised in natural systems e.g. bee swarms, bird flocks and ant colonies, it was eventually applied to artificial systems by G. Beni and J. Wang[45] and has since become a popular branch of artificial intelligence. Common swarm intelligence applications include molecular docking[46-48], protein folding prediction[49-50], crowd simulation[51] and financial forecasting.[52]

The idea of using swarm intelligence within the context of MD was first introduced by Huber and Van Gunsteren[53] via their SWARM-MD method. The aim of SWARM-MD is to encourage convergence between replicas by applying an attractive potential towards an average position in dihedral space (equation 1.11). Since its conception, the use of SWARM-MD in conjunction with simulated annealing has been shown to be effective in carrying out ab initio folding simulations of miniproteins and peptides.[53-54]

$$U_{tot}^{sw}(\{\phi^{\alpha}\}) = \sum_{\alpha=1}^{N} A\, exp\left[-Bd_{\alpha}^{RMS,av}(\phi^{\alpha})\right] \qquad [1.13]$$

Unlike SWARM-MD, the swarm-enhanced sampling molecular dynamics (sesMD) method was developed with an emphasis on enhancing the sampling of dihedral space by combining a mixture of attractive (A, B) and repulsive (C, D) terms with a pairwise evaluation of the root-mean-square dihedral distance ($d_{rms}^{\alpha\beta}(\phi^{\alpha}, \phi^{\beta})$).[11]

$$U_{tot}^{sw}(\{\phi^{\alpha}\}) = \sum_{\alpha=1}^{N} U_{\alpha}^{sw} = \sum_{\alpha=1}^{N}\left[\sum_{\beta=1,\beta\neq\alpha}^{N}\left\{\frac{A}{2}exp[-Bd_{rms}^{\alpha\beta}(\phi^{\alpha},\phi^{\beta})] + \frac{C}{2}exp[-Dd_{rms}^{\alpha\beta}(\phi^{\alpha},\phi^{\beta})]\right\}\right] \qquad [1.14]$$

As described later in this thesis (Chapter 2.4.2), the resultant biased distribution can be recovered using the exponential reweighting method of Torrie and Valleau.[10]

The effectiveness of sesMD in comparison to brute force MD has been demonstrated on several occasions, including studies of the p38α MAP kinase DFG loop motion and the free energies of butane. In the former case, the use of sesMD allowed the observations of a DFG loop transition from a "DFG-out" to a "DFG-in" state (referring to the loop's position relative to the kinase's ATP binding pocket), which was not obtainable via non-biased MD simulations.[11] In the latter case, sesMD was coupled with thermodynamic integration (termed sesTI), to improve the dihedral space sampling of the "common atoms" in a butane-to-butane alchemical transformation, resulting in a better estimation of the free energies when compared to independent trajectory thermodynamic integration.[12]

Chapter 2: Towards a fast and efficient swarm enhanced sampling methodology

2.1 Introduction

As previously described in Chapter 1, the swarm-enhanced sampling MD (sesMD) methodology has been successfully used in the past to investigate systems such as the p38α MAP kinase[11] and butane-to-butane alchemical transformations.[12] Unfortunately, the sesMD approach suffers from a few limitations which impacts its ease of application to new systems of interest. These limitations can be categorised into two areas: high compute costs and algorithmic issues leading to poor exploration of highly dimensional systems. One of the main aims of this project is to develop a swarm-enhanced sampling protocol that can be used to effectively and rapidly sample the configurational space of new systems of interest. Thus in order to achieve this, we must attempt to address these limitations.

This chapter provides an overview of the work which has been performed to improve the sesMD methodology, both in terms of its algorithmic and compute implementations. Firstly, we present a generalisation of the sesMD potential, termed the multi-dimensional swarm-enhanced sampling method (msesMD), which aims to both improve sampling and ease the parameter transferability limitations of the sesMD methodology. Secondly, optimisation efforts are detailed for both the sesMD framework within the AMBER14 *sander* MD engine[32], and the implementation of the msesMD methodology within the more efficient *pmemd* MD engine. Finally, some of the methodological aspects of conducting swarm simulations are discussed, including approaches to recover ensemble averages from swarm-biased results.

2.2 The multi-dimensional swarm-enhanced sampling method

2.2.1 Introducing the multi-dimensional swarm-enhanced sampling potential

As described in Chapter 1, the sesMD methodology enhances conformational sampling by coupling a series of *M* replicate trajectories via their relative proximity in dihedral space. This is achieved by application of a pair potential between replicates α and β:

$$U^{ses}(\{\varphi^{\alpha}\}) = \sum_{\alpha}^{M} U_{\alpha}^{ses} = \frac{1}{2}\sum_{\alpha}^{M}\sum_{\beta\neq\alpha}^{M}\left(Aexp\left[-Bd_{rms}^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})\right] + Cexp\left[-Dd_{rms}^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})\right]\right) \qquad [2.1]$$

where $d_{rms}^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})$ is the root-mean-square dihedral angle distance of *K* dihedrals *j* between swarm members α and β, namely $(M^{-1}\sum_{j}^{K}\left(\varphi_{j}^{\alpha}-\varphi_{j}^{\beta}\right)^{2})^{1/2}$; and A - D are parameters for attractive (A, B) and repulsive components (C, D). The swarm biasing potential seeks to balance the dispersing of replicas lying in similar areas of conformational space while also preventing them from becoming isolated from each other. In theory, assuming an adequate selection of parameters A – D, this should offer an improved rate of exploration in comparison to conventional molecular dynamics for any molecular system of choice. In practice however, this algorithmic approach is limited. As one investigates increasingly complex systems, it is found that the effectiveness of a given sesMD parameter set reduces.

A potential source of this issue is the use of the $d_{rms}^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})$ term to describe replica proximity. This term, which takes its origins from the original SWARM-MD methodology[53-54], can be effective in describing coordinated movement in systems, particularly for small numbers of dihedrals. However, as the number of dihedrals, *K*, increases, the use of a single averaging term to describe the conformational proximity between replicas can be limiting, particularly in cases where a single coupled motion along these dihedrals is not expected. There are two reasons for this: firstly, the use of a single

coordinate to represent a *K*-dimensional space means that as *K* increases, the nature and size of the available phase space changes accordingly. Thus, swarm parameters which are optimised for a specific *K*-dimensional space may not be effective when applied to a *K*+1 space, as the relative positions and number of energy minima in this new exploration space have changed. The second issue is that RMSD terms can lead to issues regarding outliers.[55] In terms of sesMD simulation, let us consider a scenario where a molecule is being boosted along several dihedrals and that a large motion along one of these dihedrals away from a highly stable state specifically relates to the access to a conformer of interest. It is possible that, in some cases, the distance between a pair of replicas, one of which exists in the stable state and the other of which occupies the conformer of interest, may be considered as very close to each other compared to another replica that may moderately deviate along several other dihedrals. Thus, sampling of the conformer of interest may be reduced relative to the other conformational states, as the boost potential actively attempts to steer replicas to new positions that are away from each other.

One could attempt to avoid these issues by reducing the number of dihedrals boosted by sesMD, or amending the number of replicas and/or swarm parameters accordingly. In both cases, this can be particularly difficult as it usually requires *a priori* knowledge of the system, which in a lot of cases defeats the purpose of using an enhanced sampling method like sesMD. Additionally, the reparameterisation of the swarm parameters can be a time consuming task, and unlike methods like aMD[37] the form of the potential is not so simple as to easily lend itself to a generalised empirical rule for parameter development.

In order to overcome the limitations of the RMSD term, we instead propose an extension of the swarm methodology, which we term the multi-dimensional swarm enhanced sampling method (msesMD). This approach, in part inspired by focused enhanced sampling methods such as partial replica exchange MD[56] and the generalised adaptive biasing force[57] decomposes the swarm potential into several locally evaluated sub-swarm potentials using distinct per-dihedral distance evaluations $d_j^{\alpha\beta}\left(\varphi^{\alpha}, \varphi^{\beta}\right)$ for *K* dihedrals *j* of interest, rather than the singular root-mean-square term. This results in the following potential:

$$U^{ses}(\{\varphi^{\alpha}\}) = \sum_{\alpha}^{M} U_{\alpha}^{ses} = \frac{1}{2}\sum_{\alpha}^{M}\sum_{\beta\neq\alpha}^{M}\sum_{j}^{K}\left(Aexp\left[-Bd_j^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})\right] + Cexp\left[-Dd_j^{\alpha\beta}(\varphi^{\alpha},\varphi^{\beta})\right]\right) \qquad [2.2]$$

The use of such a decomposed potential circumvents the above mentioned sesMD issues, as each dihedral is treated independently of the other. Thus the dimensionality of the search space for each sub-potential does not change with the number of dihedrals investigated. Since the swarm parameters now reflect the exploration of a single local dihedral, we expect that the choice of parameters will be less influenced by the nature and size of system under investigation. Therefore, we theorise that it may be possible to develop a set of parameters for "model" systems (e.g. alanine dipeptide) and apply them to larger systems without needing to explicitly re-parameterise. Admittedly, there are also some possible caveats to using this approach. The most obvious disadvantage is that the boost potential no longer explicitly accounts for coupled motion; instead one now has to rely on the possibility that a replica may concurrently explore the correct areas of dihedral space in tandem in order to achieve coupled motion. The second potential disadvantage is that since each locally applied sub-swarm is not explicitly aware of each other, the increased entropy may result in a less "well behaved" boost potential.

2.2.2 Comparing the effectiveness of the swarm methodologies

In order to investigate the effectiveness of msesMD in comparison to sesMD and unbiased MD, a set of benchmark simulations were carried out looking at how extensively each method explores the dihedral space of explicitly solvated alanine dipeptide and alanine heptapeptide. The reason for choosing these two systems is twofold: firstly alanine dipeptide is a commonly used small system to investigate sampling efficiency as the molecule can exhibit "rare" transitions across the φ angle to a stable well in α_L region. Thus using the alanine dipeptide model, we can judge as to how well each enhanced sampling method is performing. Secondly, alanine heptapeptide, being the slightly larger polypeptide version of the alanine dipeptide, was chosen in order to investigate the extent of transferability of the swarm parameters when increasing the number of torsions.

Methods. Both systems were built using the *leap* module in AMBER14[32], with the peptides parameterised using the AMBER14SB force field. Both peptides were solvated in a box of TIP3P waters with waters being placed up to 10 Å away from the solute. Solute and solvent hydrogen bonds where constrained using the SHAKE[27] and SETTLE[33] algorithms respectively. The simulations used a 1 fs time step, an 8 Å cutoff for short range non-bonded interactions and the particle mesh Ewald method to handle long range electrostatics. Thermal control was achieved using the Langevin thermostat[25] with a target temperature of 298 K and a collision frequency of 3 ps^{-1}. All production simulations were carried out in the isobaric-isothermal ensemble (NPT), using the Berendsen barostat[58] and a target pressure of 1 bar. The trajectories were sampled every 500 ps. It should be noted that two different AMBER MD engines were used, *pmemd* for the unbiased MD and msesMD simulations and *sander* for sesMD.

Both systems were first minimised using 1000 steps of steepest descents algorithm, followed by a further 1000 steps of conjugate gradients. The temperature was then slowly increased from 0 K to the 298 K target temperature over 500 ps under NVT conditions. This was then followed by 1 ns of NPT equilibration. From this equilibrated structure, six replicate trajectories were spawned and allowed to evolve independently for an additional 1 ns. For the unbiased MD simulation this was followed by a further 500 ps of

equilibration and then 1 ns of MD production for each independent replica. In both sesMD and msesMD, the six equilibrated trajectories were first subjected to 500 ps of swarm parameter annealing where the swarm coupling potential was slowly introduced. This was then followed by 1 ns of swarm production simulations. The swarm boost potential was applied to all backbone φψ dihedrals; thus two dihedrals were boosted in alanine dipeptide and twelve in the alanine heptapeptide. The results were analysed using the *cpptraj*[59] module in AMBER14, recovering 2D histograms of the φψ occupation density with a bin width of 2°. It is noted that the surfaces are not reweighted and therefore, in the swarm calculations, represent the biased distributions.

For this test, both the sesMD and msesMD were parameters were refined in order to ensure good exploration of the alanine dipeptide φψ surface. Firstly, a swarm size of six replicas was chosen as this was seen to be the lowest effective amount of replicas which could fully sample the alanine dipeptide using an initial parameter guess. Then the swarm parameters were iteratively refined in two steps: first coarsely testing different parameter combinations; then taking the best results from the first step, iteratively altering the parameters in order to ensure good sampling. The resultant swarm parameters for sesMD were A = -1.0 kcal mol^{-1}, B = 1.0 rad^{-1}, C = 8.0 kcal mol^{-1}, D = 2.0 rad^{-1}, whilst for msesMD they were A = -1.0 kcal mol^{-1}, B = 1.0 rad^{-1}, C = 4.0 kcal mol^{-1}, and D = 3.0 rad^{-1}. As the sesMD parameters are proportional to the number of dihedrals, in order to provide a fair comparison, the parameters were scaled by a factor of six when applied to the alanine heptapeptide.

Alanine dipeptide sampling. The aggregate 6 ns unbiased MD surface of alanine dipeptide is shown to sample only a few regions of conformational space during the simulation, mainly tracing the β (φ = -60° to -180°, ψ = 120° to 180°) and α_R (φ = -60°, ψ = -40°) regions. As expected, both sesMD and msesMD outperform unbiased MD in sampling the φψ surface, with both exploring the α_L and γ regions not visited during the unbiased MD trajectories (Figure 2.1**A**-**C**). Interestingly, the sesMD simulation appears to sample the major β and α_R states more frequently, resulting higher (darker) bin densities in those states, whilst the msesMD sampling appears to be more diffuse.

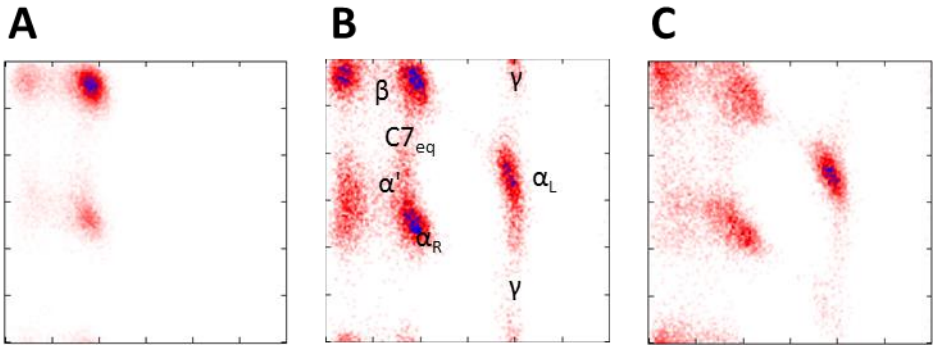

*Figure 2.1 φ (abscissa) vs ψ (ordinate) aggregate bin occupation sampling of alanine dipeptide using **A**) unbiased MD, **B**) sesMD and **C**) msesMD. Explored regions are labelled on the sesMD surface.*

Alanine heptapeptide sampling. We next look at the sampling of the backbone φψ dihedrals of the alanine heptapeptide. As seen from Figure 2.2, the unbiased MD simulations primarily occupy the β regions, with some slight sampling of the α_R, although the latter well is not found in $\varphi\psi_4$. The sesMD slightly improves sampling over the unbiased MD simulations, tracing more of the α regions, especially α′ (Figure 2.2). However, in all cases, the α_L regions seen in the alanine dipeptide are not sampled. The msesMD simulation samples a wider range of conformational states than both unbiased MD and sesMD, finding both α_L and γ regions for four of the six φψ surfaces (Figure 2.2).

The reduced sampling effectiveness of the sesMD method for the alanine heptapeptide is the expected result of using a low replica count and a set of parameters developed for a small system to explore a multi-dihedral search space. On the other hand, the msesMD methodology shows greater promise in terms of parameter transferability, sampling the alanine heptapeptide dihedral surfaces to nearly the same extent as the alanine dipeptide. Though not all φψ surfaces were fully explored using msesMD, it is likely that this would have been achieved with a slightly longer sampling time. Whilst further benchmarks will be required to define the full extent and scenarios under which msesMD outperforms the sesMD methodology, the suggested improved transferability of the swarm parameters would be a clear advantage for the msesMD method. Therefore our focus in this thesis will be the further development and application of the msesMD approach.

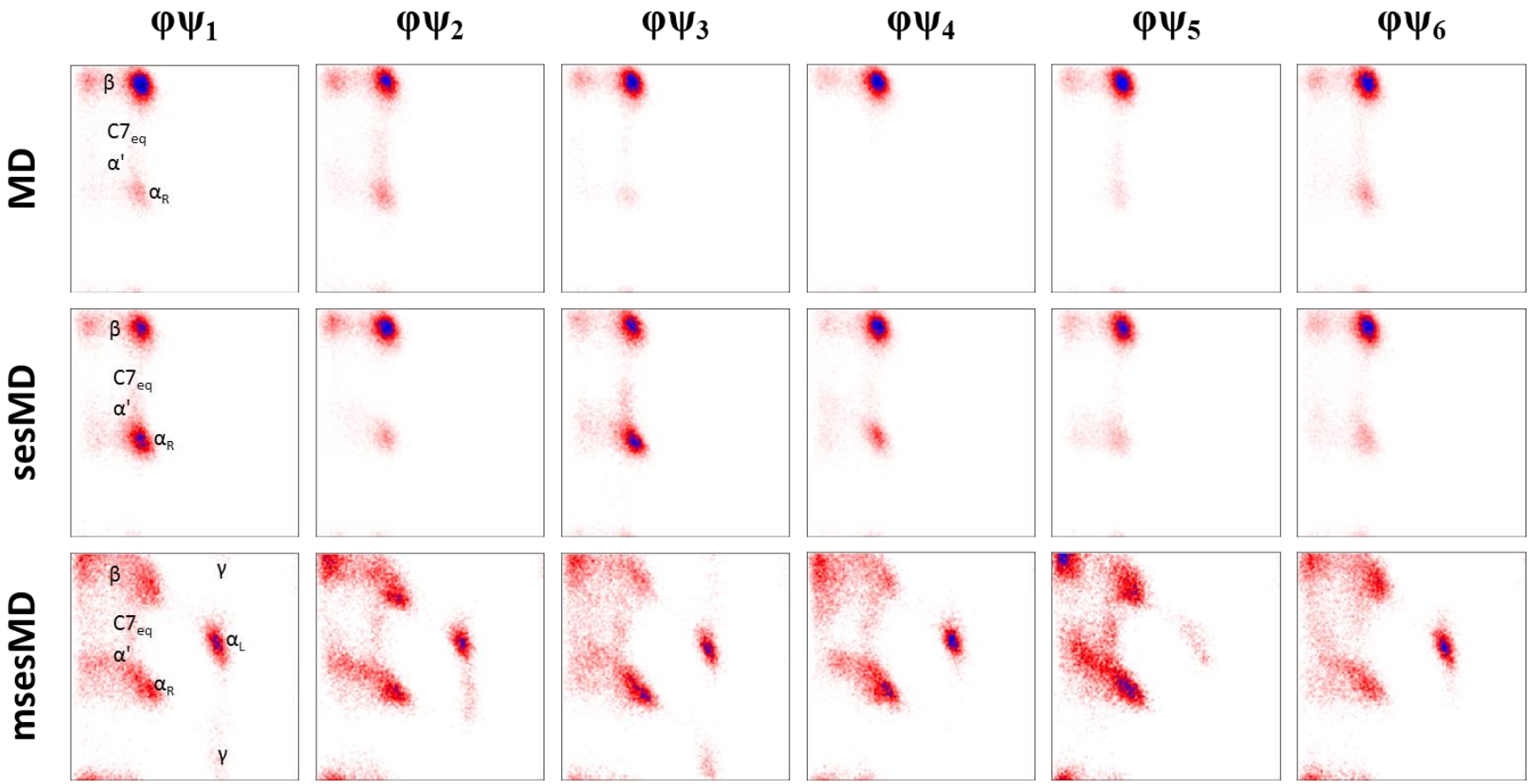

Figure 2.2 φ (abscissa) vs ψ (ordinate) aggregate bin occupation histograms of all alanine heptapeptide backbone residues as calculated by unbiased MD, sesMD and msesMD

2.3 Optimising the swarm compute performance

2.3.1 Developing an optimised implementation of sesMD in *sander*

Beyond algorithmic performance, the compute performance of the swarm-enhanced sampling models must also be addressed. Unfortunately, the original implementation of sesMD is associated with high compute costs such that for medium-sized molecules, several days of simulations are usually required to obtain tens of nanoseconds of simulation time per replica. In this section, the efforts to create a fast and efficient implementation of the sesMD algorithm within the AMBER14 *sander* MD routine are detailed.

The original implementation of the sesMD[11] was implemented in a custom variant of the *sander* MD engine of the AMBER8[60] suite and later ported to AMBER11.[61] The *sander* engine is one of the two main MD components of the AMBER molecular modelling suite, the other being the newer *pmemd* engine (and its GPU-accelerated variant). A small overview of the program structures of both *sander* and *pmemd* can be seen in Supplementary Figures A.1 and A.2. As of AMBER14, the multiprocessor frameworks for both sander and pmemd use the distributed memory Message Passing Interface (MPI)[62] parallel framework in order to share MD simulation costs over several CPU cores. This works by having a "*master*" thread decompose the coordinate space into a series of chunks which are then offloaded to "*slave*" MPI threads running on different cores. For multi-replica simulation methods which require communication between replicas, such as replica exchange[8] or sesMD, the multi-core hierarchical structure is set such that each individual replica runs its own MD engine instance and then communicates information to other replicas across a higher level MPI layer. Thus, a multi-replica simulation is one where each individual replica uses one or more cores to parallelise its local MD integration costs and also communicates data to and from the compute cores of other replicas in order to obtain the necessary information from them.

In the simplest of terms, MD engines have two main functions: first, at every time step, a set of forces are calculated based on the potential energy functions defining the system of interest (i.e. non-bonded and bonded terms as derived from the force fields used); secondly, these forces are then integrated using the equations of motion in order to obtain a new set of coordinates (e.g. using the velocity Verlet or leapfrog integrators). This is then repeated continually until the number of simulation steps requested has been achieved.

In terms of sesMD, the main idea is to introduce additional forces at each time step based on the swarm potential (equation 2.1). As detailed in Figure 2.3, this involves four unique steps. First, the dihedral angles over which the sesMD potential is applied must be calculated from the coordinates of each replica. Then the values of these dihedral angles must be shared explicitly with every other replica in the swarm. In terms of MPI, this step involves an operation named "AllGather". As evidenced by its name, the AllGather operation is a function that allows all MPI targets, in this case the individual replicas, to explicitly gather from each other the contents of an array (i.e. dihedral angle values). The next step in sesMD is to use the dihedral angles shared by all replicas to calculate the proximity between the local replica and all other replicas in the swarm (i.e. $d_{rms}^{\alpha\beta}(\varphi^{\alpha}, \varphi^{\beta})$). Once this is achieved, the energies and resultant forces can be calculated as per equation 2.1. For the sake of brevity, we henceforth will refer to this step as the "swarm energies calculation". Finally, once the swarm forces have been updated, they must then be reflected accordingly in the main MD force arrays prior to the next coordinate update. In the sesMD implementation, all four swarm calculation steps are carried out only on the primary *master* MPI thread, this means that the forces must be then shared across to all the *slave* MPI threads which handle the atoms over which the swarm forces are acting. In terms of the *sander* engine, all local MPI threads must explicitly have access to a copy of the final forces and coordinates at the end of each step. Therefore, the *master* must copy the swarm forces across to all other local threads, this is achieved using the MPI broadcast operation.

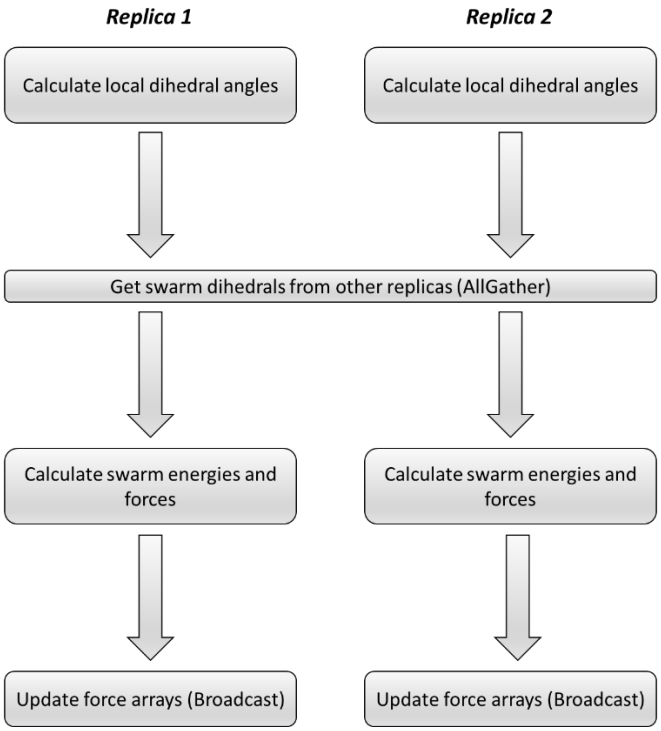

Figure 2.3 The key steps in a sesMD calculation undertaken at each time integration steps

In the original sesMD implementation, all four of these steps are executed in succession during each force evaluation after having calculated the other non-swarm MD forces (Figure 2.4). Unfortunately, the use of such an approach can lead to significant slowdowns in the simulation speed, as the communication costs between replicas, i.e. the AllGather call, can be quite high; this is especially true for cases where the replicas are being executed in different compute nodes. In those cases, data packets must be sent from the local memory of each replica across a network fabric to all other replicas. Even with high-end network fabrics, such as InfiniBand[63], this is usually a very expensive operation which can have costs on the order of up to a few milliseconds. Additionally, MPI AllGather operations are "blocking"; that is to say, they require all communicating processes, i.e. the *master* threads in each replica, to wait for the AllGather to complete fully, prior to doing any other work. Since the original sesMD implementation carries out the swarm routines after all the other forces have been calculated, this means that all the threads handling a replica are essentially idling, waiting for the *master* thread to complete the inter-replica communication.

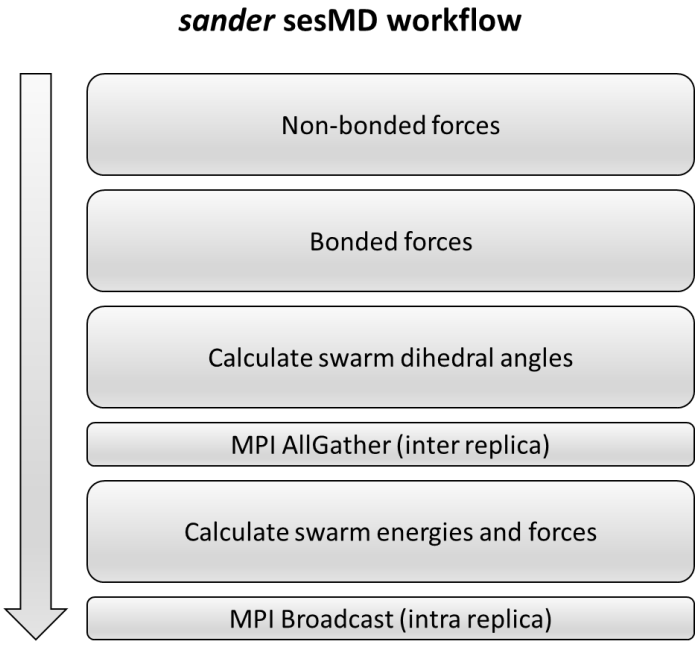

Figure 2.4 The original sesMD implementation within the sander *force evaluation routine*

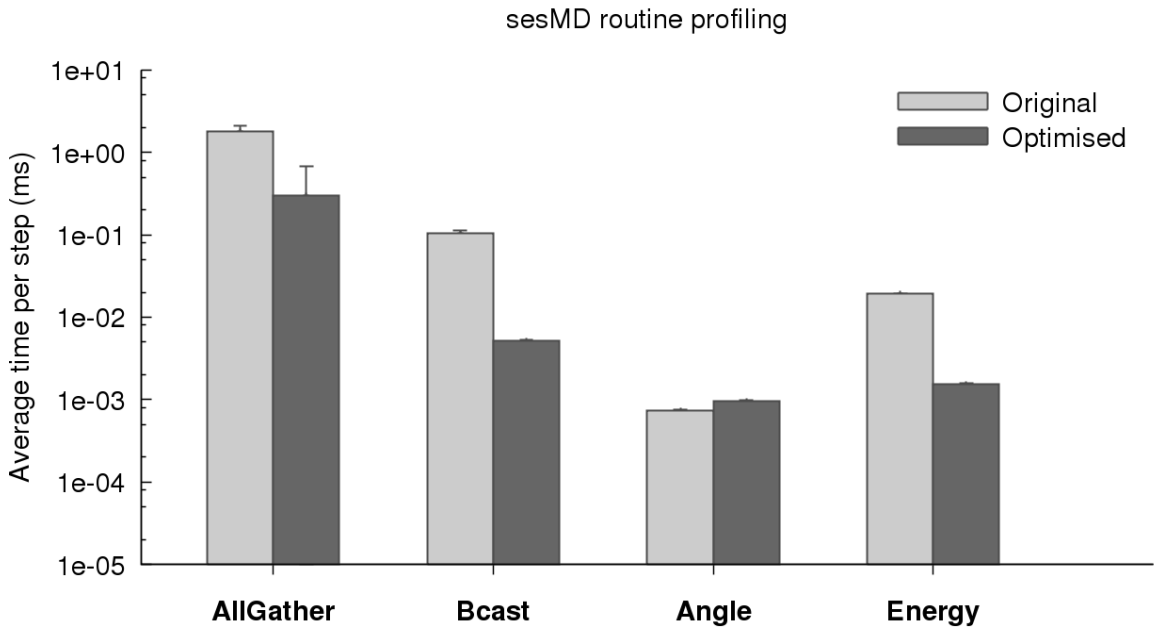

Figure 2.5 Log scaled bar chart detailing the timings (ms) of the different sesMD steps, calculated using a mannotriose benchmark system, comparing the original and optimised sesMD implementation

Being intra-replica, the MPI Broadcast usually has a lower cost, as all the MPI threads of sesMD replicas typically exist within the same compute node. This means that communication can be handled via direct memory copies, which are usually over an order of magnitude faster than communicating across a network fabric. The other steps, i.e. the calculation of the dihedral angles, energies and forces, are computationally cheaper, with negligible impacts on the simulation speed.

To demonstrate these relative costs, the sesMD routine was profiled using a solvated mannotriose benchmark system (Figure 2.5). The benchmark system consists of a pre-equilibrated, mannotriose, modelled by the Glycam06i force field and solvated in a TIP3P water box with 10 Å edges, under NVT conditions using the Langevin thermostat. An 8 replica swarm simulation was then calculated for 25 ps across two 24 core Haswell E5-2680 v4 nodes (i.e. 6 cores per replica), connected by a single switch QDR InfiniBand fabric. Timings were obtained using the high accuracy MPI time library for each of the four sesMD steps.

For this benchmark the AllGather communication usually incurs a cost of around 1.7 ms per step; this is significant considering that the average time taken to fully complete a step is 7.8 ms (Figure 2.5). By comparison, the Broadcast (Bcast, Figure 2.5) only takes 0.1 ms on average to complete. The swarm energy/force (Energy, Figure 2.5) and dihedral angle (Angle, Figure 2.5) calculations take orders of magnitude less time at 0.02 and 0.0007 ms per step respectively. This highlights the need to specifically optimise the MPI communication aspects of the sesMD algorithm.

Thankfully, since the initial msesMD implementation, there have been several updates to the MPI library. Of particular note is the inclusion of asynchronous versions of the collective operations such as AllGather and Broadcast as part of the MPI 3.0 standard.[64] Unlike the original “blocking” versions, asynchronous communication allows for MPI processes to immediately carry on with other compute tasks without having to wait for the communication task to have completed. MPI barriers are then used at a later point the code

in order to ensure that the requested MPI communication has completed. The advantage of this is that it can be used to overlap the communication phase with non-related calculations, effectively hiding most of the associated costs. Thus, by re-arranging the four msesMD steps, it is possible to overlap both the AllGather and Broadcast steps with time consuming MD force calculations. As shown in Figure 2.6, the code has been amended such that the swarm dihedral angles are first calculated and then shared between replicas while the very expensive non-bonded forces are evaluated. By the time that the first MPI barrier is reached, little to no waiting is required before moving onto the calculation of the swarm energies. This is then followed by an asynchronous Broadcast of the forces whilst the bonded forces are being calculated.

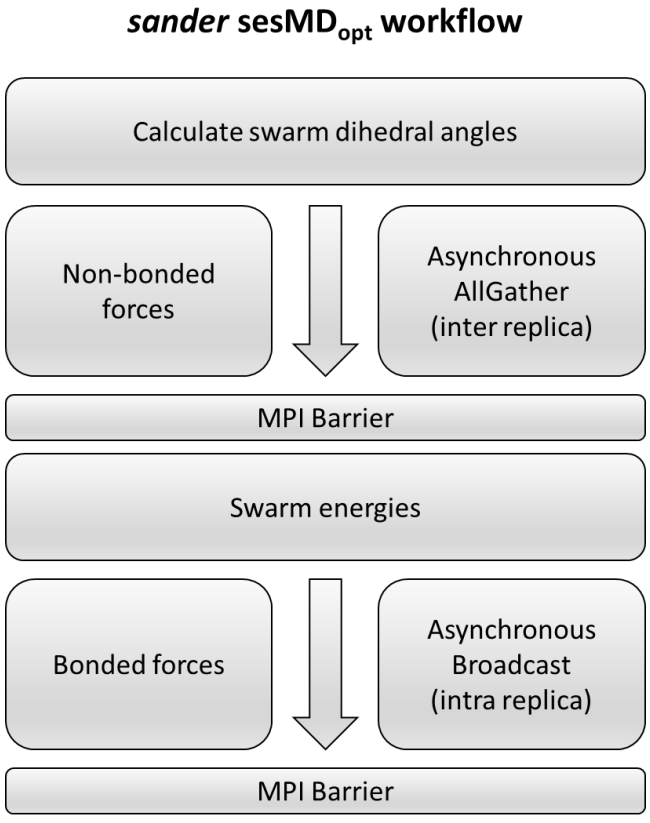

Figure 2.6 An optimised sesMD workflow taking advantage of MPI 3.0 asynchronous collectives

As seen from the profiling of this optimised sesMD code, the judicious use of asynchronous MPI routines results in an order of magnitude improvement in both the AllGather and Broadcast operations (Figure 2.5). It is noted that the standard deviation in the optimised AllGather timings is quite high; this is in part due to network traffic by other

jobs running concurrently on the InfiniBand fabric used for this benchmark, which at times causes large spikes in communication costs. Additionally, we note an order of magnitude decrease in the calculation cost of the swarm energy/force calculation routine; this is due to a mix of both a transfer of the code to AMBER14's *sander* engine and small code improvements not detailed here. A core scaling test for the 8 replica mannotriose benchmark was carried out to see how well this optimised algorithm performed relative to both the original sesMD code and non-swarm calculations (Figure 2.7). Unlike the routine profiling, this core scaling test was carried out using 500 ps of simulation time, a 2 fs time step and up to eight 24 core Cray XC30 nodes using the ARCHER national HPC resource. The optimised sesMD code ($sesMD_{opt}$) shows a speed up of up to 1.3 times the original sesMD implementation, reaching up to 53 ns per day of simulated time (Figure 2.7). Comparing against the *sander* unbiased MD simulation, the cost of the optimised sesMD simulation is near negligible, with only a 1 ns per day difference between the two. However, the faster *pmemd* engine is found to be around 2.3 times faster.

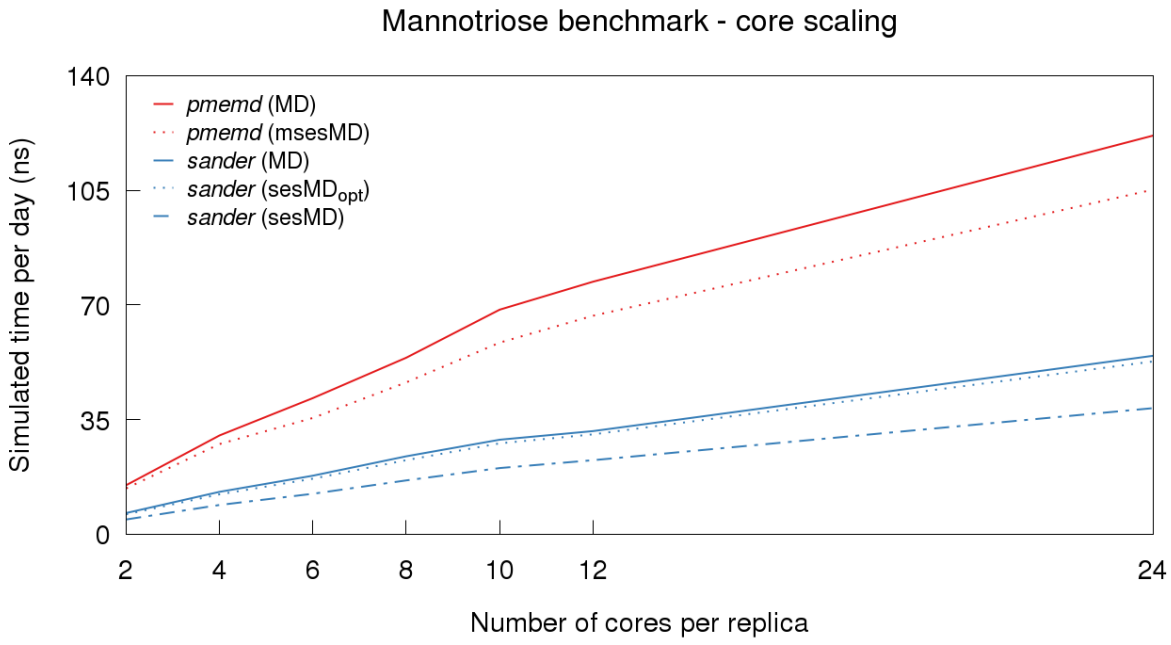

Figure 2.7 Core scaling test for the mannotriose 8 core benchmark test, comparing both unbiased MD simulations and m/sesMD simulations calculated via both sander *and pmemd*

2.3.2 Implementation of msesMD in *pmemd*

Having sufficiently optimised the sesMD routines in *sander*, it became apparent that further speedups would be limited by the MD engine itself. Therefore, when it came to implementing a fast version of the msesMD methodology, it was decided to instead explore the use of the faster *pmemd*. As shown in Figure 2.7, the *pmemd* MD engine is around twice as fast as *sander*, in part due to the use of a better parallel domain decomposition model which vastly reduces the rate of communication between MPI threads. The *pmemd* code is also more efficient in its code structures, allowing for better use of vector instructions at the individual thread level.

The swarm energy evaluations in msesMD, although requiring the evaluation of K times more energies (equation 2.2), can inherently be used with domain decomposition. Therefore, rather than carrying out all the swarm work on a single *master* MPI thread, it is possible to arrange the calculations such that individual *slave* threads calculate the energies for the dihedral atoms which they explicitly handle. It is noted that as shown in Figure 2.5, the time taken to complete the swarm energy evaluation is usually negligible for sesMD; however, the slightly higher cost of doing more energy evaluations in msesMD calculation can easily be offset by parallelising the swarm routines in this way. Furthermore, parallelising the swarm calculation also replaces the need to share the swarm force array from the local *master* thread to their *slave* threads; instead this requires the sharing of the usually smaller swarm dihedral angles array which is slightly more efficient.

An overview of the msesMD workflow in *pmemd* is shown in Figure 2.8. As can be seen, as per the optimised sesMD workflow (Figure 2.6), the dihedral angles are calculated and shared on the *master* threads at the start of each loop. Unlike sesMD, this is then followed by a local sharing of these angles so that the relevant *slave* processes can use them for their local swarm energy evaluations. Two additional changes relative to the sesMD workflow (Figure 2.6) can also be seen: first the AllGather operation is now a set of asynchronous send/receive calls; secondly the MPI barrier handling the inter-replica gathering of dihedral angles is now dynamic. These details of these two small optimisations are not discussed here, rather only mentioning that they can, in some cases, lower the inter-replica communication costs.

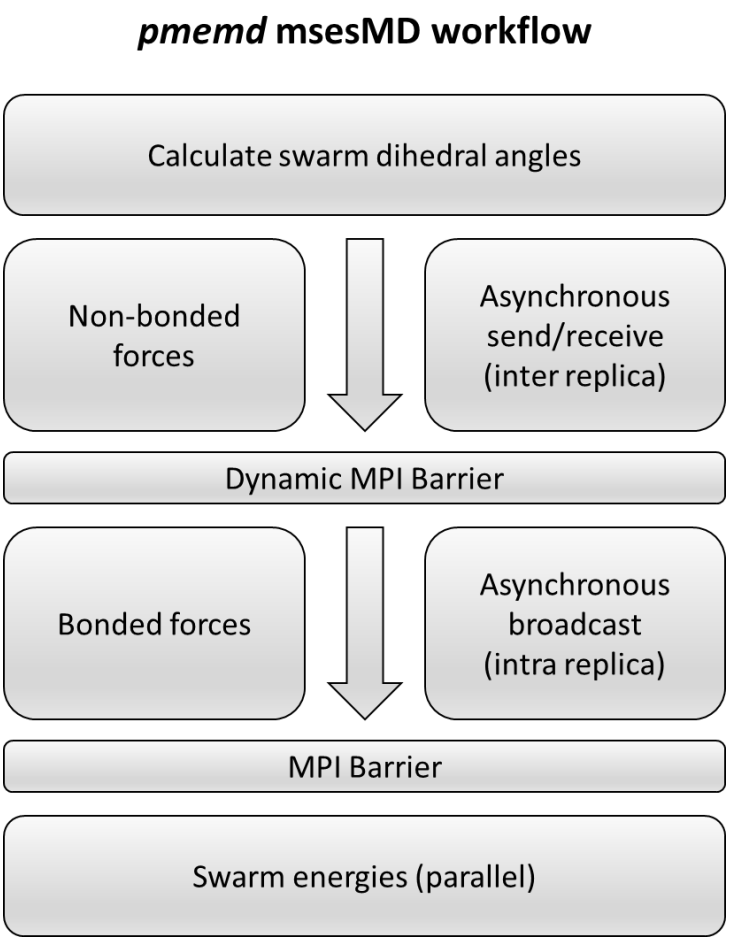

Figure 2.8 The msesMD workflow as implemented in pmemd

Assessing the efficiency of the msesMD implementation in *pmemd*, it is found that good scaling across core counts can be achieved (Figure 2.7). Comparing against unbiased MD *pmemd* simulations, the msesMD routines incur a reduction in the simulated time per day of 7 to 15%. This is larger than the difference between the optimised sesMD and unbiased MD in *sander*, which is expected considering that the *pmemd* code is more efficient, making it harder to hide the swarm communication costs. Finally, we find that the msesMD simulations in *pmemd* are around 3 and 2.3 times faster than the original and optimised sesMD implementations respectively.

2.4 Methodological aspects of msesMD calculations

2.4.1 The choice of replica count and swarm parameters

The appropriate choice of the number of replicas and associated swarm parameters can be a complex task. However, as demonstrated in Chapter 2.2, the msesMD method appears more suited to the use of a transferable set of parameters. Therefore, for this thesis, a single set of parameters for 8-replica swarms was developed and used as the basis for the work done in Chapters 3 to 5. However it is noted that in some cases (Chapter 3 and 5), the parameters were scaled down in order to reduce reweighting noise (indicated where this was applied). While 12 replicas has been the norm in previous sesMD simulations[11-12], the reason for choosing 8 replica is for practical reasons. First, even with modern high density HPC nodes, a 12 replica swarm can be quite resource intensive compared to 8 replicas, requiring the use of more nodes or less cores per replica. Secondly, the core count on modern HPC nodes has recently shown a tendency to be divisible by powers of 2; therefore, using 8 replicas, we usually find that we are more likely to be able to fit a whole swarm calculation within a single node.

Here we provide a summary of the parametrisation of the swarm potential for msesMD. A suitable set of parameters for the alanine dipeptide and Lewis^a oligosaccharide model systems was obtained by searching through the strongly repulsive and weakly attractive parameter space of the msesMD pair potential. Parameter suitability was assessed by looking at (i) the breadth of the dihedral space which was sampled during a short simulation, in addition to (ii) examining the distribution of the swarm boost potential energies. The latter criteria was chosen in order to avoid cases where the swarm potential would results in multi-modal energy distributions which would present issues upon reweighting of observables. Unfortunately, whilst a good set of parameters was found for alanine dipeptide, the strength of the boost potential was too high, leading to SHAKE failures in the Lewis^a system. Thus, a weaker set of pair potential parameters which sufficiently sampled Lewis^a was chosen, with A = -0.5 kcal mol^{-1}, B = 0.5 rad^{-1}, C = 2.13 kcal mol^{-1}, D = 2.625 rad^{-1}. The effectiveness of the potential was then checked on both

mannotriose and the alanine dipeptide, showing that effective sampling of dihedral space could still be achieved.

2.4.2 Reweighting swarm biased distributions

The use of a swarm boost potential biases the observed ensemble of configurations; therefore, in order to be able to recover properties of interest from such simulations, the original unbiased distribution must be recovered. As previously detailed[11-12, 65], the recovery of the Boltzmann-weighted ensemble averages of any observable *X* from an M-replica swarm can be achieved via the exponential reweighting method of Torrie and Valleau[38],

$$\langle X \rangle = M^{-1} \sum_{\alpha}^{M} \frac{\langle X(r_\alpha) \exp(\beta U_\alpha^{ses}) \rangle}{\langle \exp(\beta U_\alpha^{ses}) \rangle} \quad [2.3]$$

where $X(r_\alpha)$ is an an observable and U_α^{ses} is the total swarm potential acting on replica α at a given time point. Assuming that the individual replica averages converge given sufficient sampling, equation 2.3 can be further approximated to the following[12];

$$\langle X \rangle = \frac{\sum_{\alpha}^{M} \langle X(r_\alpha) \exp(\beta U_\alpha^{ses}) \rangle}{\langle \exp(\beta U_\alpha^{ses}) \rangle} \quad [2.4]$$

As explored in Chapter 5, this approximation usually holds except at very short simulation times. Therefore for simplicity of use, the latter form of the reweighting equations is primarily used in this thesis. Using this approach, it is possible to recover relative free energy (ΔA) estimates of surfaces such as dihedral rotations, by reweighting observed occupation density averages in conjunction with the canonical relationship,

$$\Delta A_n = -k_B T ln(p_n / p_m) \quad [2.5]$$

where ρ_n and ρ_m represent ensemble average estimates of the occupation density of different bins along a histogrammed surface, with ρ_m being the value of the highest observed bin density.

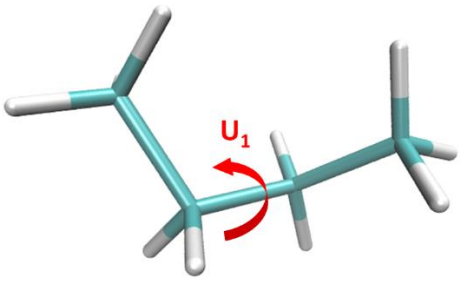

Figure 2.9 The butane U_1 torsion

To test the reweighting method, the recovery of a solvated butane U_1 torsional PMF (Figure 2.9) was compared using both msesMD and umbrella sampling (for simulation details, see Mehthods appended below). Comparing the 10 ns msesMD results with those of umbrella sampling, we see that there is good agreement with near identical profiles being generated. Some slight deviations at the higher energy states can be seen, particularly around U_1 = 135°; however this remains within 0.1 kcal mol^{-1}. These results demonstrate the effectiveness of the exponential reweighting method.

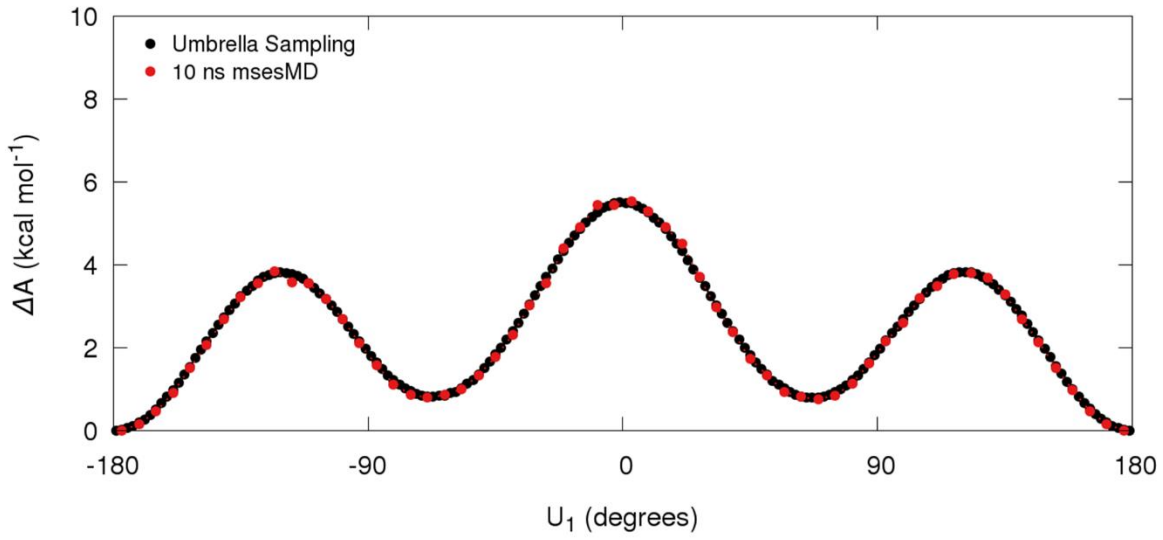

Figure 2.10 Free energy profiles of the butane U_1 rotation as calculated by both umbrella sampling and 10 ns msesMD

Methods. A butane was parameterised using the GAFF[17] force field in the *leap* and *antechamber* modules of AMBER14[32] and solvated in a TIP3P[66] water octahedral box with waters placed up to 12 Å away from the solute. A 2 fs time step was used for all simulations with the solute and solvent hydrogen bond motions constrained using the SHAKE[27] and SETTLE[33] algorithms respectively. Long-range electrostatics were treated using the particle mesh Ewald method[67], and short-range non-bonded interactions were truncated at 9 Å. Thermal control was achieved using the Langevin[25] thermostat with a target temperature of 298 K and a collision frequency of 3 ps^{-1}.

The system was first equilibrated using 50000 steps of minimisation (25000 steps of steepest descent and 25000 steps of conjugate gradient). The temperature was then slowly increased from 0 K to 298 K during a 500 ps NVT simulation. The box density was then equilibrated over a 1 ns period using the Monte Carlo[25] barostat and a target pressure of 1 bar. The system then went through a final 1 ns NVT equilibration round. All following production calculations were carried out under the NVT ensemble.

msesMD simulations. Eight independent replicas were spawned and allowed to evolve independently over a 500 ps simulation. The swarm coupling parameters were then slowly introduced to the target values defined in Section 2.4.1 over a 500 ps period. This was then followed by a 10 ns swarm production simulation.

Umbrella sampling simulations. A total of 120 evenly spread window simulations along the U_1 torsion were computed. Each was subject to a harmonic biasing force of 200 kcal mol^{-1} rad^{-2} ensuring good overlap between windows. At each window, the system was subject to a further equilibration, consisting of 2000 steps of minimisation and 200 ps of simulation under the influence of the biasing force. This was then followed by a 500 ps production simulation. The resultant PMF profile was computed using the WHAM code by the Grossfield lab.[68]

2.5 Conclusions

In this chapter we first discuss the algorithmic limitations of the sesMD methodology, especially with regards to poor parameter transferability and reduced sampling as the number of boosted dihedrals increases. The msesMD methodology which decomposes the swarm biasing potential into a series of per dihedral swarms is introduced as a solution to these limitations. Testing the effectiveness of the two swarm-enhanced sampling models against benchmark systems, both methods were found to be effective in sampling the φψ torsional space of alanine dipeptide, identifying the α_L and γ regions which are not visited in unbiased MD simulations (Figure 2.1). In the alanine heptapeptide benchmark simulations, the effectiveness of the sesMD method is reduced, only showing slight improvements in sampling over unbiased MD (Figure 2.2). However, the msesMD simulations are able to sample nearly all the relevant torsional states, including the α_L and γ regions which are not seen in any of the sesMD or unbiased MD simulations. It is expected that, if appropriately reparameterised, the sesMD approach may provide similar sampling of the alanine heptapeptide as msesMD. However this would involve high compute costs and therefore demonstrates the advantage of the msesMD, for which a single parameter set can be used semi-agnostically across different systems. Future work will be required to define the exact extent of the advantages of the msesMD over sesMD. Nevertheless, in this thesis we will be concentrating on the development and use of the msesMD methodology.

Next, the optimisation efforts were detailed for both the sesMD and msesMD methodology in the *sander* and *pmemd* MD engines respectively. We show that through the judicious use of asynchronous MPI collectives, the high costs associated with inter-replica communication can be efficiently hidden. This resulted in a speedup of 1.3x of the sesMD code in *sander* relative to its initial implementation, resulting in only a 5% cost in simulation time relative to unbiased MD using the same MD engine (Figure 2.7). One should note that in all cases, this assumes a high-end network fabric such as InfiniBand. It is expected that the results presented here would be drastically different if the swarm communications instead occurred across a low bandwidth Ethernet connection. In terms of implementing the msesMD routines within *pmemd*, this led to substantial speedups of up to 2.3x and 3x relative to the *sander* optimised and original sesMD simulations (Figure 2.7).

It is noted implementing msesMD within *pmemd* offers the opportunity for the integration of msesMD within some of the MD routines which are specific to *pmemd*, such as the single topology TI framework (see Chapter 5).

The significant speedups in the swarm routines shown here will enable the routine access to aggregate microsecond calculations in small to medium size systems such as mannotriose. This will be particularly useful in this project, allowing for rapid evaluations of different parameter sets and boost coordinates which, using the original sesMD implementation, would be impossible. Although a substantial amount of work has been done to optimise the speed of both the sesMD and msesMD routines, there is still scope for further work, including leveraging the use of hardware accelerators (e.g. GPUs) or using alternative parallelisation methods such as the recently introduced mixed-mode hybrid MPI+OpenMP compute model in AMBER16.[69]

Finally some of the methodological aspects of using swarm-enhanced sampling, specifically msesMD were outlined. Briefly covered are the development details of a unified set of swarm parameters which will be used throughout this thesis. Whilst the chosen parameters may not be ideal in most cases (see chapters 3 and 5), they serve a useful starting point for further adjustments if needed. The reweighting procedure for msesMD simulations was also covered, presenting two possible forms of the exponential reweighting equations that, as demonstrated in chapter 5, can be used somewhat interchangeably assuming the system is sufficiently sampled. As a test of the effectiveness of the reweighting procedure, the PMF profile of the U_1 torsion of butane as generated by both msesMD and umbrella sampling were compared. As can be seen in Figure 2.10, the reweighted msesMD PMF offers near identical results those of umbrella sampling.

Chapter 3: Identification of rare Lewis oligosaccharide conformations

3.1 Introduction

Glycoscience is an important and emerging area with potential applications ranging from personalised medicines and food security to biofuels and advanced biomaterials.[70] Key to the rational design of carbohydrate-based compounds, however, is the ability to accurately characterise their conformational manifolds[71]. Achieving this can be challenging in carbohydrates as they exhibit subtle conformational properties which may dictate their functional behaviour. The use of computational methods have proven successful in this endeavour; in particular, the use of molecular dynamics (MD) simulations is capable of providing conformational landscapes in atomistic detail which are not readily available via experimental approaches. For example, conformational analysis of the core pentasaccharide derived from the lipooligosaccharide of pathogenic bacterium *Moraxella catarrhalis* was recently performed by molecular simulation and NMR. It was shown that the addition of one glucosyl unit to each of the (1→4) and (1→2) branches of this core pentasaccharide led to a significant change in conformation, into a more folded compact structure[72], with significant potential implications for the minimum epitope to target in vaccine design.

In a second recent example, Blaum et al.[73] examined the pentasaccharide glycan component of ganglioside GM1, a membrane-bound ligand for a range of proteins, including those of viruses and bacteria. In their work, they identified a novel conformation of GM1, exhibited in its complex with a 155 kDa protein, Factor H. Subsequent long time scale (10 μs) MD simulations of GM1 in explicit solvent identified this conformation as a secondary low lying minimum on the free energy surface. Interestingly, this novel conformation of GM1, although consistent with the NMR data, does not correspond to a good set of NOE restraints and thus was not reported from previous combined NMR/modelling work.

The Lewis blood group oligosaccharides constitute key mediators of biomolecular recognition. This blood group mainly consists of six oligosaccharides (Figure 3.1,**1**-**6**), all of which share a common Gal-GlcNac-Fuc core. These six sugars are divided into two types (**I** and **II**) based on the linkage configuration of this core. For type **I** Lewises, the core trisaccharide is Galβ(1→3)(Fucα(1→4))GlcNAc depicted as Le^a (**1**) in Figure 3.1; this can be further sialylated (sLe^a, **3**) or fucosylated (Le^b, **5**). For type II sugars, the glycosidic linkage connectivity of the core trisaccharide is interchanged to give Galβ(1→4)(Fucα(1→3))GlcNAcβ or Le^x (**2**) in Figure 3.1; as for Type **I** Lewises, this can be further sialylated (sLe^x, **4**) or fucosylated (Le^y, **6**). While important contributors to normal physiological function, the Lewis oligosaccharides have also, due to their role in the immune system, been associated with several disease states including pathogenic inflammatory response and carcinomas.[74-76] Understanding their conformational behaviour is therefore advantageous in furthering our knowledge of their role in these disease states and potentially assisting in the development of novel therapeutics.

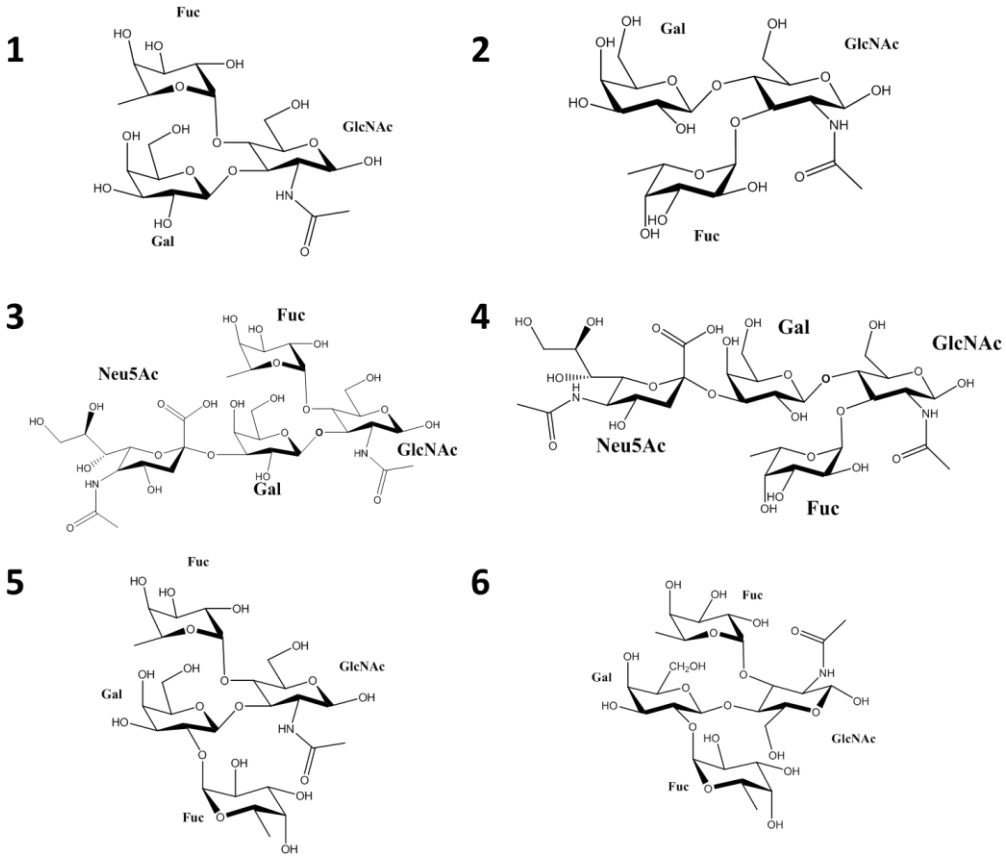

*Figure 3.1 The six Lewis oligosaccharides **1**) Le^a, **2**) Le^x, **3**) sLe^a, **4**) sLe^x, **5**) Le^b, and **6**) Le^y*

Historically, the Lewis oligosaccharides have been viewed as being relatively rigid structures, adopting a single "closed" conformation at equilibrium, where a stable stacking interaction is formed between the fucose (Fuc) and galactose (Gal) rings (Figure 3.2).[77] Although the presence of "open" conformers, i.e. non-stacked structures, has been previously suggested by both NMR and computational studies,[78-79] this could not be confirmed.[30] Recently, however, unusual conformations adopted by Lewis antigens Le^x and sLe^x were reported by Topin and co-workers.[80] In this case, crystal structures of Le^x and sLe^x bound to pathogenic *Ralstonia solanacearum* lectin (RSL) were found to adopt poses distinct from the usual closed shape. Subsequent long time scale MD simulation of Le^x in explicit solvent by Topin et al. detected transitions from closed to open conformations on the μs timescale. These transitions however were rather infrequent: of the 30 MD trajectories acquired, five showed transitions from the closed to open forms of Le^x; of the combined 75 μs of simulation, this comprised 1% of the total solution ensemble.

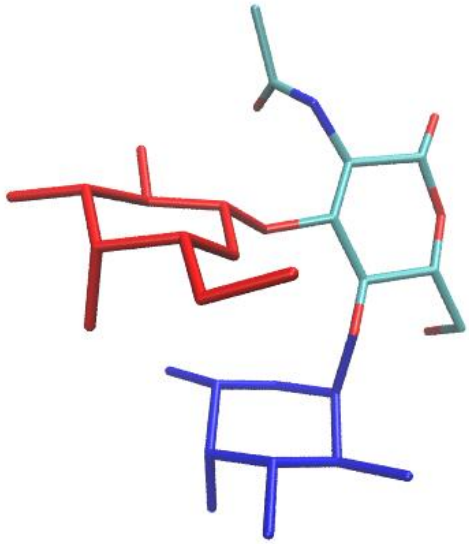

Figure 3.2 Closed conformation of Le^a with Fuc in blue, Gal in red, GlcNAc coloured by atom type

These results demonstrate the importance of characterising as fully as possible the ensemble of glycan structures and illustrate the power of hardware acceleration in probing longer time scales.[30, 81-82] However, despite recent advances in hardware development and utilisation,[1-2, 83-84] exhaustively characterising the multi-microsecond landscapes to identify rare conformational states, as achieved by Topin et al.,[80] remains a time consuming task. As discussed in Chapter 1.2.2, several advanced sampling techniques, such as accelerated molecular dynamics[6], umbrella sampling[38, 85], metadynamics[7] and replica-exchange schemes[8, 86-87], have been introduced to explore these time scales and examples have emerged of their application to the challenge of exploring carbohydrate conformation.[85, 88-94] Given the importance of glycosidic torsions in defining the shapes of oligosaccharides,[40] approaches which boost along torsional coordinates, such as the sesMD method, would appear well suited to the task of characterising carbohydrate dynamics. Having introduced the multi-dimensional swarm enhanced sampling method (msesMD) in Chapter 2, we now turn our attention to its use in probing the dynamics of oligosaccharides.

In this study, we investigate the conformational flexibility of the Lewis oligosaccharide cores in aqueous solution, via molecular dynamics simulations using msesMD in explicit solvent. Using blood sugar sLe^a (**3**), we compare the extent of sampling of msesMD with

multi-microsecond simulations (triplicates of 1 μs and of 10 μs trajectories), validating our outcomes using other biased sampling methods, namely umbrella sampling and the accelerated molecular dynamics (aMD) method. We then apply msesMD and unbiased MD simulations to investigate the ability of sLe^x (**4**) and unmodified Lewis cores, Le^a (**1**) and Le^x (**2**), to form open conformations in aqueous solution.

3.2 Methods

3.2.1 System details

Solvated oligosaccharides **1-4** (Figure 3.1) were modelled using the Glycam06j-1[95] and TIP3P[66] force fields for the sugars and water respectively. Each oligosaccharide was built and then solvated in an octahedral water box with solvent molecules placed up to a minimum of 12 Å away using the *leap* module in AmberTools14.[32] Sodium ions were added to neutralise the solute charge where appropriate.

Molecular dynamics simulations used modified version of the *pmemd* and *pmemd.cuda* modules of AMBER14[32]. The *pmemd* code was modified to include the msesMD methodology (Chapter 2), whilst the *pmemd.cuda* code was amended to fix a bug in the aMD routines which, in certain cases, prevented its use with the Glycam06 force field. A hydrogen mass repartitioning scheme[29] was used for all unbiased MD simulations, scaling solute hydrogen masses by 3 amu[26]; all msesMD, aMD and umbrella sampling simulations used standard masses. Unless otherwise indicated, a time step of 4 fs was used for mass repartitioned systems and 2 fs for standard mass systems. In all simulations, hydrogen atom motions were constrained using the SHAKE[27] and SETTLE[33] algorithms for the solutes and waters respectively. Long range electrostatic interactions were calculated using the particle mesh Ewald method and a 9 Å cutoff for short range nonbonded interactions. Thermal control was achieved using the Langevin thermostat[25] and a collision frequency of 3.0 ps^{-1}.

3.2.2 Simulation protocol

Equilibration. All systems were equilibrated using 500000 steps of energy minimisation (250000 steps of steepest descents followed by 250000 steps of conjugate gradients). The systems were then heated under NVT conditions, where the temperature was raised from 0 K to a target of 298 K over 500 ps. This was then followed by 1 ns of NPT simulation to equilibrate box densities at a pressure of 1 bar using the Monte Carlo barostat[25], with volume exchange attempts every 100 steps. This was then followed by a final 1 ns of NVT equilibration. At the end of the equilibration phase, all systems were observed to occupy the closed conformational state.

Unbiased simulations. For all oligosaccharide systems, three replicated unbiased NVT simulations at 298 K were carried out with a sampling frequency of 5 ps. Each replicate simulation was performed for a total of 10 μs.

Umbrella sampling calculations. Two one-dimensional potential of mean force (PMF) profiles were computed along the ψ angles of the Fucα(1→4)GlcNAc and Galβ(1→3)GlcNAc glycosidic linkages of sLe^a (denoted ψ_F and ψ_G respectively) using umbrella sampling (Figure 3.3). A total of 120 window simulations along each ψ torsion were computed, subjected to a harmonic biasing force of 200 kcal mol^{-1} rad^{-2}, ensuring good overlap between windows. Unlike other enhanced sampling simulations, a 1 fs time step was used to improve stability. The systems were equilibrated using the equilibration protocol above, and then each window, under the influence of its respective biasing potential, was subject to an additional 500000 steps of energy minimisation and 1 ns of NVT equilibration prior to 20 ns of NVT production sampling, extracting dihedral values every 50 steps.

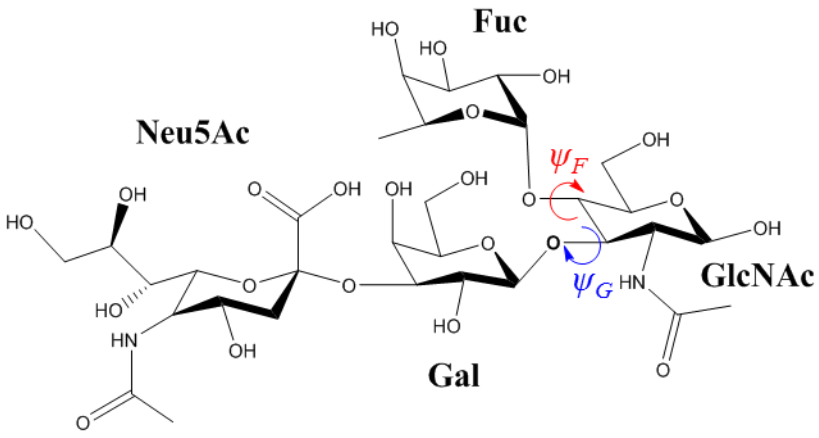

Figure 3.3 ψ_F and ψ_G angles in sLe^a indicated in red and blue respectively

Swarm-enhanced sampling simulations. Swarm-enhanced sampling simulations were conducted all systems using the msesMD potential, as described in Chapter 2. After system equilibration, eight independent NVT unbiased trajectories were propagated from the equilibrated structure at 298 K for 1 ns each. This was achieved in order to ensure that the structures diverge from each other in dihedral space, preventing the large boost values which would occur when replicas are very close to each other in dihedral space. The eight simulations were then coupled via msesMD in the NVT ensemble, first gradually introducing the swarm potential over a period of 600 ps, followed by a further 5.4 ns of equilibration under the full influence of the biasing potential. This was then followed by 250 ns per replica of production simulation, sampling every picosecond. Unlike the swarm parameters used in Chapter 2 and 4, due to excessive noise in the reweighted surfaces, a slightly downscaled set was used here, A = -0.375 kcal mol^{-1}, B = 0.5 rad^{-1}, C = 1.5975 kcal mol^{-1}, D = 2.625 rad^{-1}. The swarm boost potential was applied to every φ/ψ glycosidic torsion in the system of interest.

Accelerated molecular dynamics simulations. Similar to the msesMD simulations, eight independent trajectories were spawned from an equilibrated system and simulated for an extra 1 ns each in the in the NVT ensemble. An aMD dual boost potential was then applied using boost parameters E_{dih} = 17.7 kcal mol^{-1}, α_{dih} = 6.4 kcal mol^{-1} for the dihedral potential energies and E_{tot} = -15000.3 kcal mol^{-1}, α_{tot} = 822.7 kcal mol^{-1} for the total potential energies. It is noted that unlike msesMD, the aMD parameters were not specifically optimised for carbohydrate sampling, instead they were derived from previous

empirical parameter results for proteins[37], with the dihedral boost scaled by a factor of two to broaden sampling. Each independent aMD simulation was carried out for a total of 500 ns, with the first 6 ns being discarded as equilibration.

Analysis. Simulation analyses were primarily carried out using *cpptraj*[59] in AmberTools16[69] and in-house python scripts. Additionally, the WHAM code by the Grossfield group[68] and the PyReweighting scripts from the McCammon group[37] were used to analyse the umbrella sampling and aMD simulations respectively. Glycosidic torsions were measured using the IUPAC definition of φ = O_5-C_1-O-C_n, ψ = C_1-O-C_n-$C_{(n-1)}$ for α/β(1→n) linkages and φ = O_6-C_2-O-C_3, ψ = C_2-O-C_3-C_4 for the α(2→3) linkages of sLe^a and sLe^x. Free energy surfaces for the glycosidic linkages were generated using a bin size of 8° and the expression $\Delta A = -k_B T ln(\rho_x/\rho_{max})$, where ρ_x is the bin density, ρ_{max} is the highest occupied bin density, k_B *is* the Boltzmann constant and T is temperature. For the computation of the free energy profile associated with the Cremer and Pople θ pucker angle, the same approach was used but with a bin size of 6°. In unbiased simulations, the bin occupation densities were obtained via the "counting method", i.e. by directly summing observable time series across all trajectories. Unbiased bin densities were recovered from msesMD simulations using the approach of Torrie and Valleau[18] where replica contributions are reweighted according to the swarm energy term (see Chapter 2). Errors in the reweighted estimates of the $\varphi\psi$ free energy surfaces were obtained by bootstrap resampling the surfaces 25000 times and calculating the standard deviation in the bin estimates across all bootstrap samples. In order to further analyse the conformational diversity of the Lewis oligosaccharides, the unbiased and msesMD trajectories of systems **1-4** were clustered based on the RMS distance of core Gal-GlcNAc-Fuc ring atoms using the DBSCAN algorithm[96]. For msesMD, cluster densities were then reweighted according to the swarm contributions of the configurations in each cluster, treating the clusters as a one-dimensional generalised coordinate. For the umbrella sampling simulations, an implementation of the weighted histogram analysis method (WHAM)[68] by the Grossfield Lab was used to recover a one-dimensional PMF from the biased windows. The aMD simulations were reweighted according to the total boost energy in each frame, using a 10th order Maclaurin series expansion to reduce noise from the biasing potential.[37]

3.3 Results and discussion

3.3.1 Evaluation of sLe^a conformational sampling

3.3.1.1 Unbiased simulations

Starting from a closed conformation of sLe^a, unbiased triplicate 10 μs simulations in explicit aqueous solvent were obtained. We first start by noting that the Neu5Acα(1-4)Gal rotation is readily sampled at all simulation times due to a lack of intramolecular hindrance. The free energy surfaces of the Fucα(1→4)GlcNAc glycosidic linkages (termed **F**) and Galβ(1→3)GlcNAc glycosidic linkage (termed **G**) derived from the first 1 μs portion of the trajectories indicates that, at this timescale, sLe^a samples mainly around its initial closed conformational state (Figure 3.4). This conformer, occupying the F_1 and G_1 regions as labelled on the surfaces (Figure 3.4), will be denoted from this point as the C state. Although no crystallographic structure of sLe^a could be found in the PDB, it is noted that the C state corresponds to the majority of the Le^a protein crystal structures (Figure 3.17, Supplementary Table B.1). Less frequently occupied non-closed conformational wells are also predicted along the two torsions, which we denote at $F_{2\text{-}3}$ and $G_{2\text{-}3}$ respectively (Figure 3.4). These open conformers tend to maintain partial stacking interactions between the Fuc and Gal rings, forming T-shaped conformers as the fucose and galactose rings orient themselves in a perpendicular arrangement, unlike the parallel stacking exhibited by the closed state (Figure 3.5).

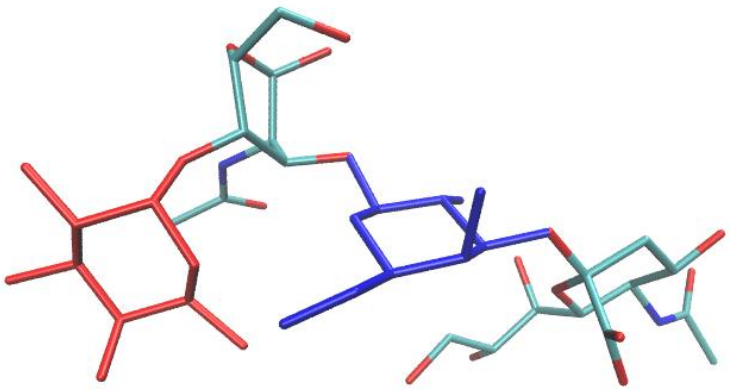

Figure 3.4 Example T-shaped open conformer of sLe^a, fucose in red, galactose in blue

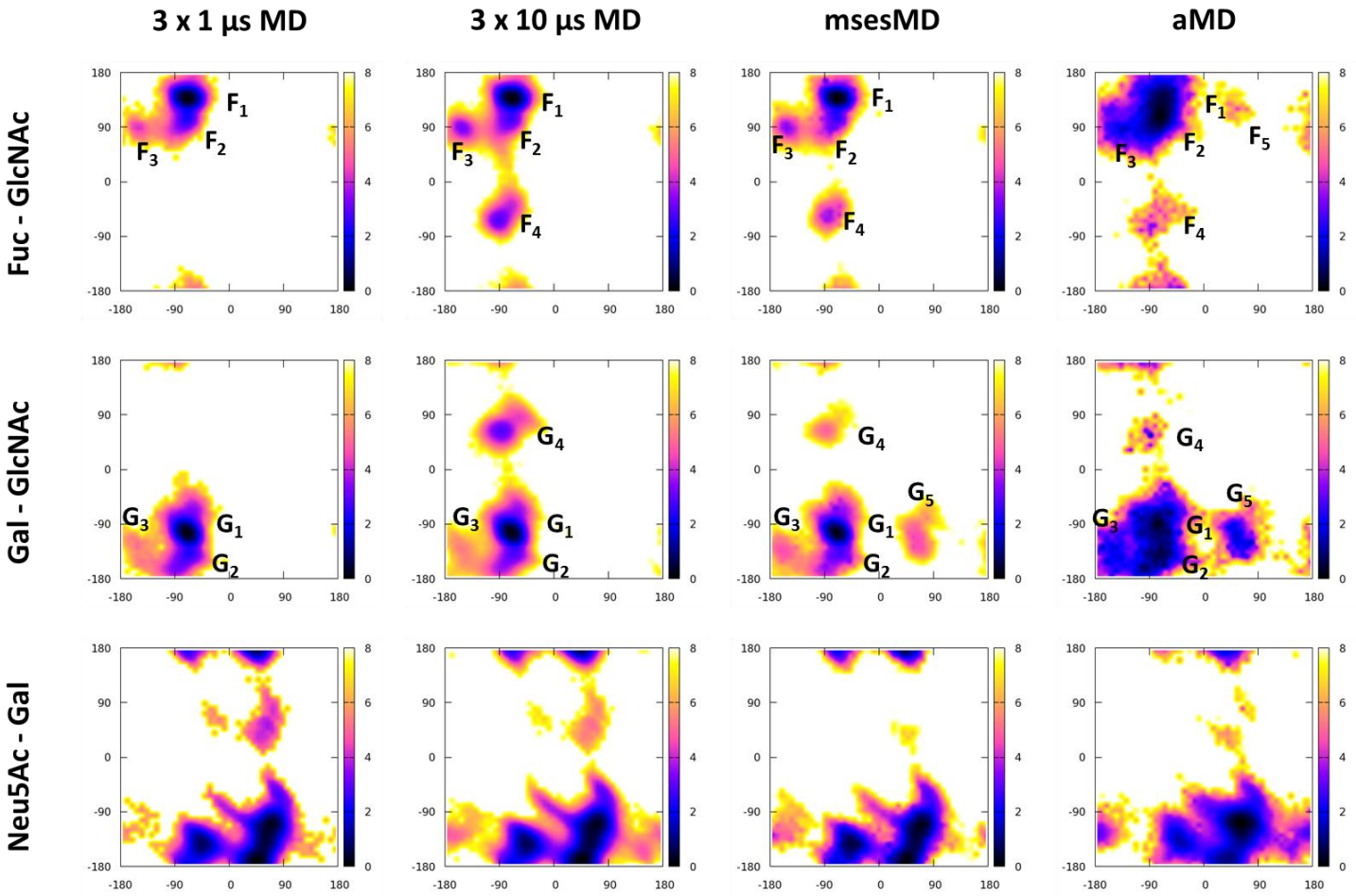

Figure 3.5 Free energy surfaces of $\phi\psi$ glycosidic torsions of sLe^a computed via MD, msesMD and aMD

More extensive sampling of sLe^a obtained by extending the three replicate simulations to 10 μs each leads to a greater population of F_3, in addition to the detection of new open conformers which occupy F_4 and G_4 regions; these have with ψ angle values which are over 180° away from the native C (F_1 and G_1) positions in their respective linkages (Figure 3.4). Inspection of the trajectories indicates that for both torsions, access to the F_4/G_4 regions occurs primarily via ψ transitions through the F_2/G_2 and F_3/G_3 regions, in a 1 → 2 → 3 → 4 sequence. Transitions to the G_4 and F_4 regions can occur simultaneously, resulting in a conformational state which is stabilised by the formation of a hydrogen bond between the hydroxyl groups at the C4 positions of both the fucose and galactose rings (Figure 3.6). Nevertheless, sampling of the F_4 and G_4 wells is highly infrequent, as can be seen from the time profiles for the ψ angle of both glycosidic torsions (Figure 3.7**A**-**B**). Aside from the first replicate trajectory (black), access to the F_4 and G_4 regions, with ψ values of 100° and -60° respectively, is infrequent with lifetimes of up to a few tens of nanoseconds. In fact, for one of the three 10 μs trajectories (red), this is only achieved once for a duration 10.8 ns and 270 ps for F_4 and G_4 respectively (Figure 3.7). We can therefore conclude that a 10 μs simulation could easily fail to identify either well during the lifetime of its time evolution. This observation accords with an earlier 10 μs study of sLe^a by Sattelle et al., where no open conformers were detected.[30]

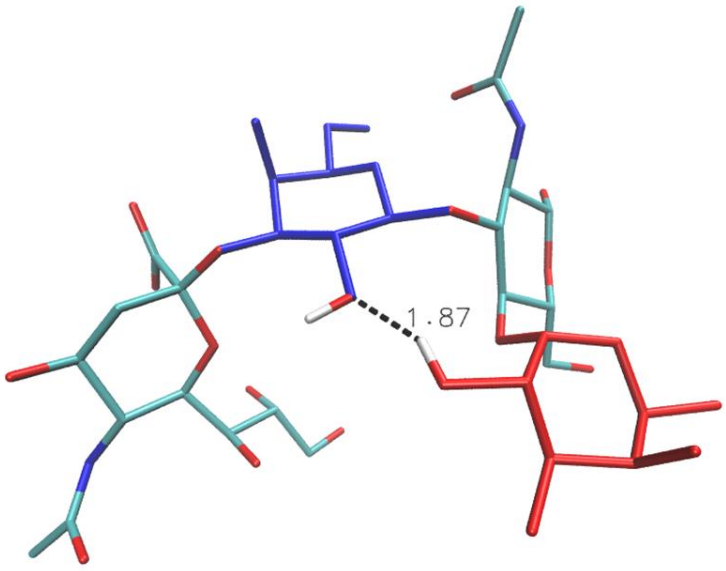

Figure 3.6 Hydrogen bond stabilised F_4/G_4 conformer of sLe^a

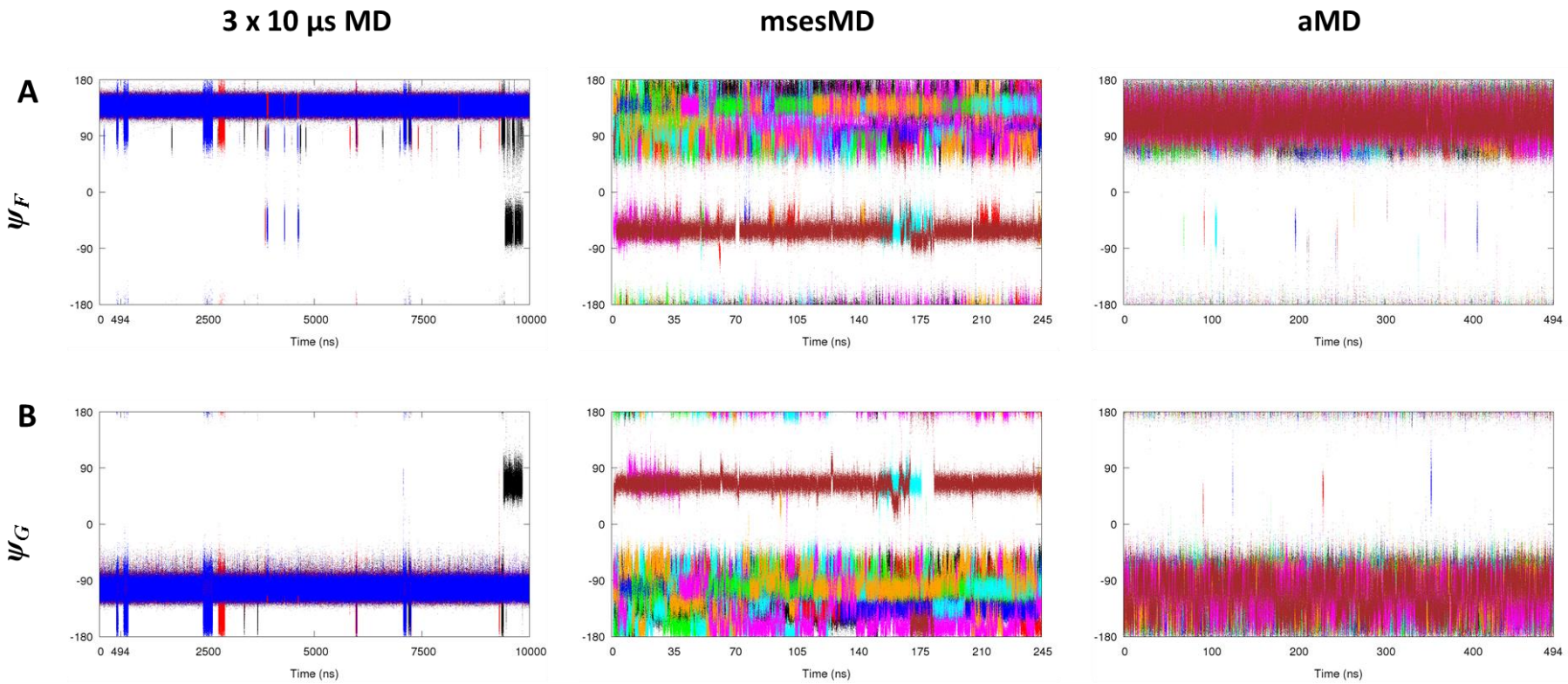

Figure 3.7 A) ψ_F and b) ψ_G time profiles of sLe^a for unbiased MD, msesMD and aMD

3.3.1.2 msesMD simulation

We now compare the unbiased 1 μs and 10 μs MD simulations with a msesMD simulation of 245 ns per replica, where eight coupled replicas of sLe^a in explicit solvent interact via their proximity in the values of the six glycosidic torsion angles. It is briefly noted that the flexible Neu5Acα(2→3)Gal linkage is readily sampled, with similar free energy landscapes to both the unbiased 1 μs and 10 μs simulations (Figure 3.4). However, the msesMD free energy surfaces of the core trisaccharide linkages of sLe^a correspond considerably more closely to the 10 μs simulations than 1 μs simulations, with broad sampling of the $F_{1\text{-}4}$ and $G_{1\text{-}4}$ low energy regions. The enhanced coverage of open forms by the swarm replicas is clearly shown by the sampling of ψ values for both glycosidic torsions (Figure 3.7**A**-**B**). Interestingly, on the Galβ(1→3)GlcNAc free energy surface obtained via msesMD, an additional low energy region is identified, which we denote G_5 (Figure 3.4).

Overall, the free energy surface computed via msesMD tends to predict similar stabilities to the profile obtained from 10 μs simulations; the exception is the G_4 well, where the free energy is predicted as 2.1 kcal mol^{-1} less stable than unbiased MD. Whilst inaccuracies in the msesMD methodology cannot be discounted, it is quite possible that this discrepancy is a lack of convergence in the unbiased simulations. As we can see from the time profiles (Figure 3.7**B**), the lifetime of the G_4 state occupancy is disproportionally longer in the first replicate trajectory (black) compared to the other two. As the G_4 state is sampled only once briefly in each unbiased trajectory, it is not possible to tell if this constitutes an outlier; however it does highlight the fact that the free energy estimates for the G_4 region are limited due to inadequate sampling.

The msesMD simulation also appears to reproduce more subtle features of the unbiased simulations, in particular sampling of the GlcNAc ring puckering. This can be seen from the free energy profiles as a function of Cremer-Pople pucker angle θ (Figure 3.8**A**). The msesMD simulations and unbiased aggregate 3 μs and 30 μs trajectories sample the full spectrum of puckers, from $^4C_1$ through skew-boat intermediates to $^1C_4$ conformations, tracing near identical profiles (Figure 3.8**A**). It is noted that the msesMD profile deviates

from the unbiased simulations at the less stable envelope/half-chair positions, which is likely due to poor sampling of these states in the msesMD simulation. For the other rings, the puckering behaviour explored to a lesser extent by msesMD (Figure 3.8**B**-**D**). In the case of the Neu5Ac residue, msesMD simulation is able to explore the complete chair inversion process of its pyranose ring (Figure 3.8**B**), finding similar wells to the 30 μs simulation for the $^{1}C_{4}$ (θ = 170°) and skew-boat/boat (θ = 90°) puckers. However the stability of the $^{4}C_{1}$ puckered conformer (θ = 20°) is overestimated relative to the unbiased 30 μs simulations by more than 2 kcal mol^{-1} due to poor sampling. For the Gal ring, the msesMD free energy profile matches the unbiased simulations by sampling mostly the 0-90° region of the θ profile (Figure 3.8**C**); however the envelope/half-chair positions are too high in energy, and the boat/skew-boat conformations too low in energy than estimates from the 30 μs unbiased simulations. For the Fuc ring, the msesMD simulation appears to be less effective than the 3 μs aggregate, with msesMD overestimating the boat/skew-boat stability by over 2 kcal mol^{-1} relative to the unbiased simulation (Figure 3.8**D**). The fact that the msesMD simulation does not effectively sample all ring puckering events is unsurprising as the biasing potential is not being applied specifically to these degrees of freedom. Further work will be required in order to efficiently apply msesMD to enhance sampling of ring puckering (Chapter 4).

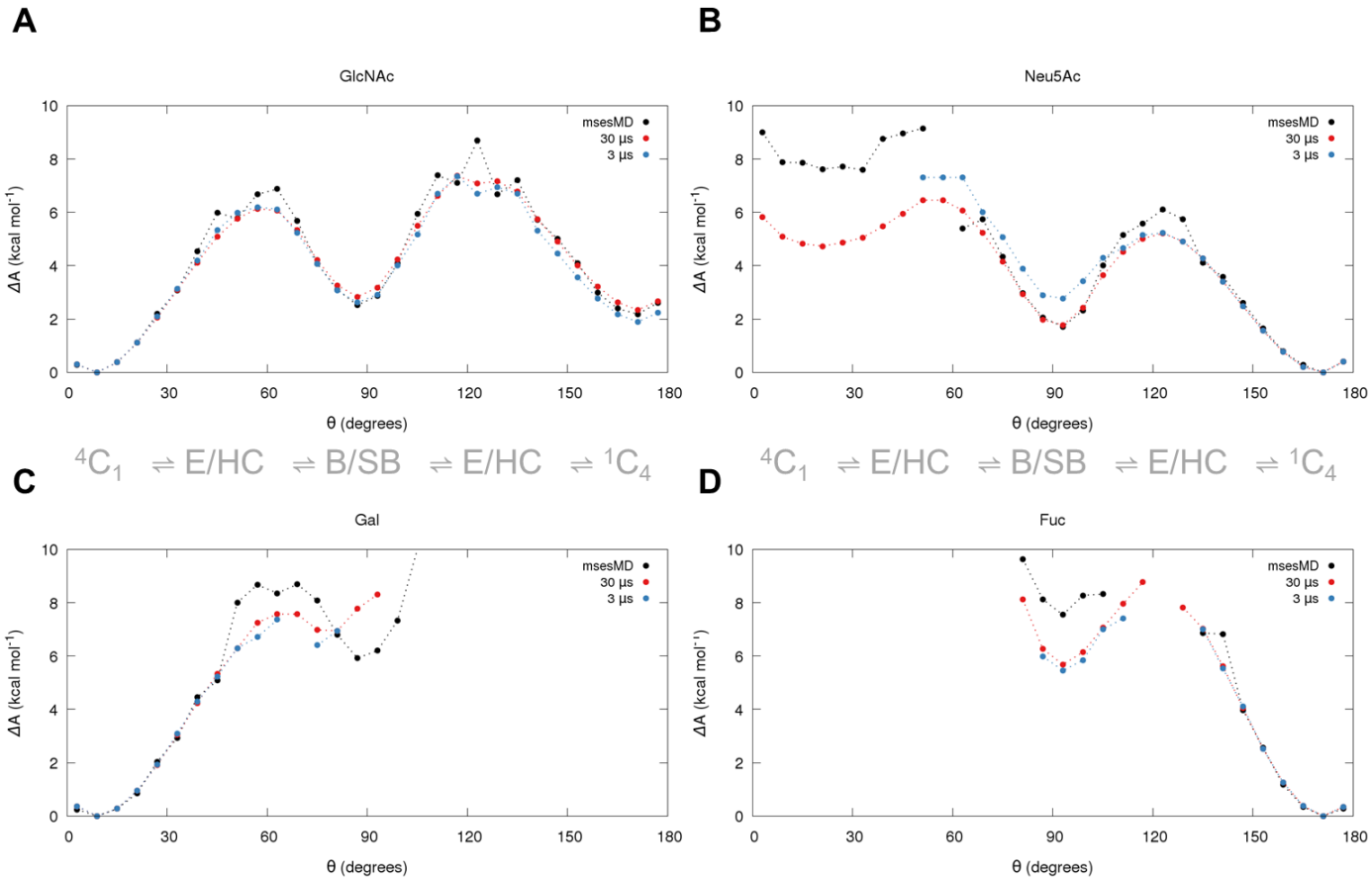

Figure 3.8 Cremer-Pople θ puckering profiles of sLe^a as calculated by aggregate 3 and 30 µs MD and msesMD for the four saccharide rings

3.3.1.3 Umbrella sampling

We also compare the conformational free energy profiles obtained from msesMD simulation of sLe^a with those of other biasing methods, including umbrella sampling. Firstly, in order to observe the rotational barrier along the $F_1 \rightarrow F_4$ transition, a 1-dimensional potential of mean force was computed using umbrella sampling around the ψ_F torsion angle of sLe^a, ie. ψ of the Fucα(1→4)GlcNAc . As expected, for this free energy profile, the closed form C is predicted as the lowest energy conformation, occupying the F_1 well (Figure 3.9). On this profile we also observe other local minima: $F_{2/3}$ (we group F_2 and F_3 as they occupy similar positions along ψ) and F_4. The $F_{2/3}$ and F_4 wells have calculated free energies relative to F_1 of 3.2 and 2.8 kcal mol^{-1} (Figure 3.9). The profile also indicates that direct transition of F_1 conformer to F_4 would encounter an energetic barrier of 9.8 kcal mol^{-1}; this compares with barriers of 4.6 and 7.9 kcal mol^{-1} for the indirect route via $F_{2/3}$ intermediates; this indeed corresponds to the $F_1 \rightarrow F_4$ path observed from the aggregate 30 μs unbiased MD simulations.

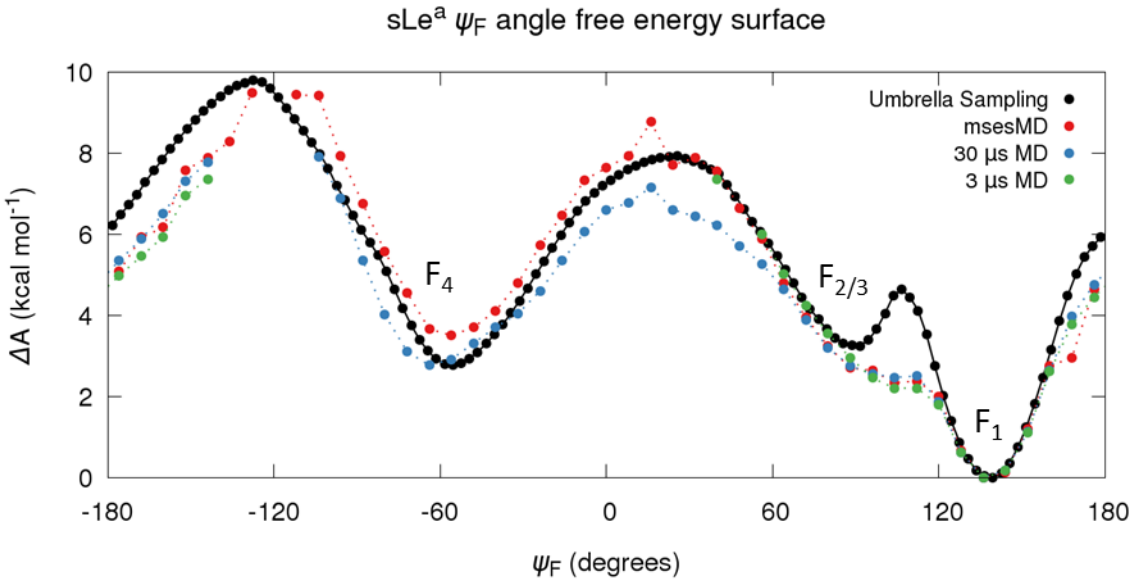

Figure 3.9 PMF of ψ_F rotation as calculated by umbrella sampling, msesMD and unbiased MD

The free energy profile constructed from the triplicate 10 μs simulations provides a reasonable approximation of the ψ_F torsional PMF (Figure 3.9), although the relative stability of the $F_{2/3}$ and F_4 wells is estimated as nearly degenerate. From the unbiased simulations, the $F_1 \rightarrow F_{2/3}$ transition also appears to have a significantly smaller barrier height, approximately 2 kcal mol^{-1} lower than the one predicted via umbrella sampling. The triplicate 1 μs simulations unfortunately fail to completely sample the ψ_F coordinate, but otherwise agree with the 10 μs simulations. An estimate of the free energy as a function of ψ_F, computed from the msesMD simulation, yields a similar profile to the unbiased simulation, although the msesMD barrier heights for the $F_{2/3} \rightarrow F_4$ and $F_4 \rightarrow F_1$ transitions are in closer agreement with the PMF from umbrella sampling. Nevertheless, it is noted that msesMD predicts the stability of the F_4 region to be 0.7 kcal mol^{-1} lower than the umbrella sampling and unbiased MD values. As for the unbiased simulations, the free energy barrier for $F_1 \rightarrow F_{2/3}$ is predicted by msesMD to be much lower than found via umbrella sampling. This discrepancy may be due to inadequate sampling of the φ_F rotation in the 1-D PMF, leading to an inaccurate description of the $F_1 \rightarrow F_3$ transition path. In fact, inspection of the φψ normalised average occupation density of visited φψ states reconstructed from all umbrella sampling windows shows that a lack of sampling of the F_2 well is evident, with the path seemingly hopping discontinuously from the F_1 to F_3 wells (Figure 3.10).

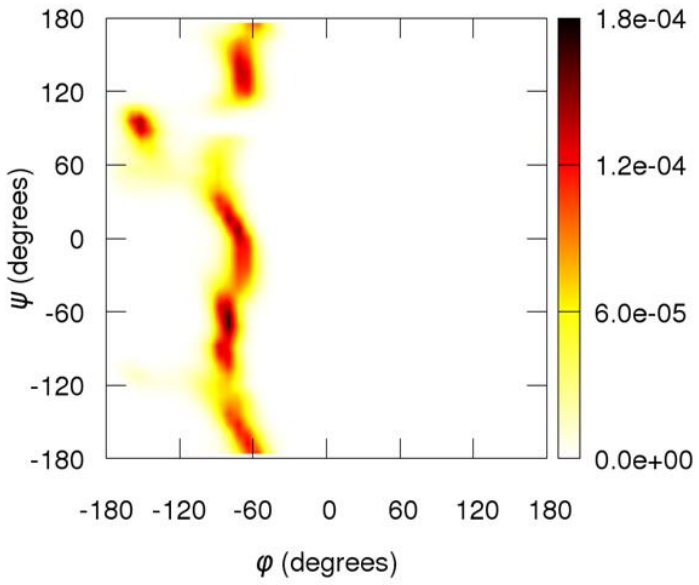

Figure 3.10 Normalised average occupation density of the φ_F/ψ_F rotational path generated from umbrella sampling

A second PMF profile was also calculated by umbrella sampling about the ψ rotation of the Galβ(1→3)GlcNAc glycosidic linkage (ψ_G), in order to examine transitions between G wells. The closed conformer C, occupying G_1, is predicted as the lowest energy state, as found for the ψ_F profile (Figure 3.11). The G_4 well is predicted as 5.7 kcal mol^{-1} above G_1; however, no clearly defined $G_{2/3}$ region was observed. Additionally, the profile indicates an energetic barrier of 10.3 kcal mol^{-1} going from the G_1 to G_4 wells directly, while the route which should involve the $G_{2/3}$ region shows a barrier of 12.6 kcal mol^{-1}. This would indicate that the most appropriate route is a direct transition from G_1 to G_4, which does not agree with the unbiased MD observed path of $G_1 \rightarrow G_2 \rightarrow G_3 \rightarrow G_4$.

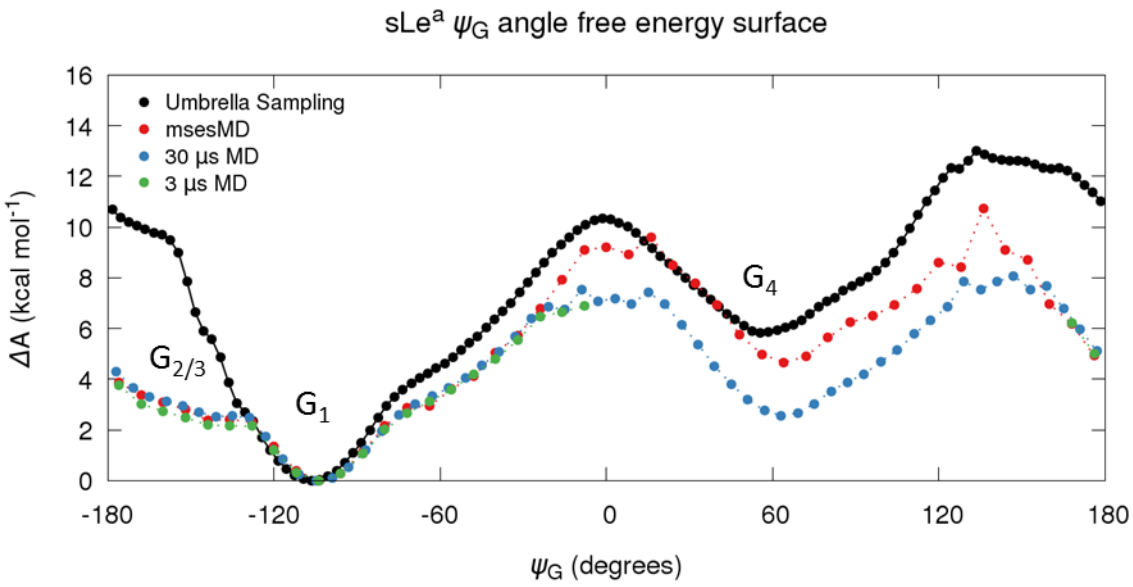

Figure 3.11 PMF of ψ_G rotation as calculated by umbrella sampling, msesMD and unbiased MD

The triplicate 10 µs MD profile is, as expected, more consistent with the previously observed behaviour with a defined $G_{2/3}$ well 2.5 kcal mol^{-1} above G_1. The G_4 well stability is also increased relative to the umbrella sampling surface, becoming equienergetic with $G_{2/3}$. The energetic barriers are also significantly lowered, with a value of 7.2 kcal mol^{-1} for the direct $G_1 \rightarrow G_4$ transition, compared to 2.5 and 5.3 kcal mol^{-1} via the indirect route; this supports the observed $G_1 \rightarrow G_{2/3} \rightarrow G_4$ path. The msesMD free energy profile for ψ_G shows agreement in overall shape with the unbiased MD prediction, but differs in finding a decreased stability of G_4 by ~2 kcal mol^{-1} and an increase of 2.0 and 2.7 kcal mol^{-1} in

energetic barrier of the direct and indirect $G_1 \rightarrow G_4$ routes respectively. As previously explained, this discrepancy between the msesMD and unbiased MD estimates is potentially due to a poor estimate of G_4 population in the unbiased MD simulation, arising from one of the replicate trajectories becoming kinetically trapped in the higher energy G_4 well for a disproportionate amount of time. The much larger differences, compared to msesMD, in the umbrella sampling profile appear to be, as for ψ_F, caused by a lack of sampling of the φ_G rotation. Although not as obvious as the ψ_F profile, there appears to be incomplete sampling of the G_3 region by the PMF, with the path consisting primarily of a transition from $G_2 \rightarrow G_4$ (Figure 3.12). The discrepancies seen in the umbrella sampling profiles of both the ψ_F and ψ_G highlights the limitation of using one-dimensional coordinates to describe complex systems such as carbohydrates. The use of two-dimensional PMF profiles for carbohydrates would be preferred, but are very computationally expensive, even for relatively small systems of such as sLe[a].

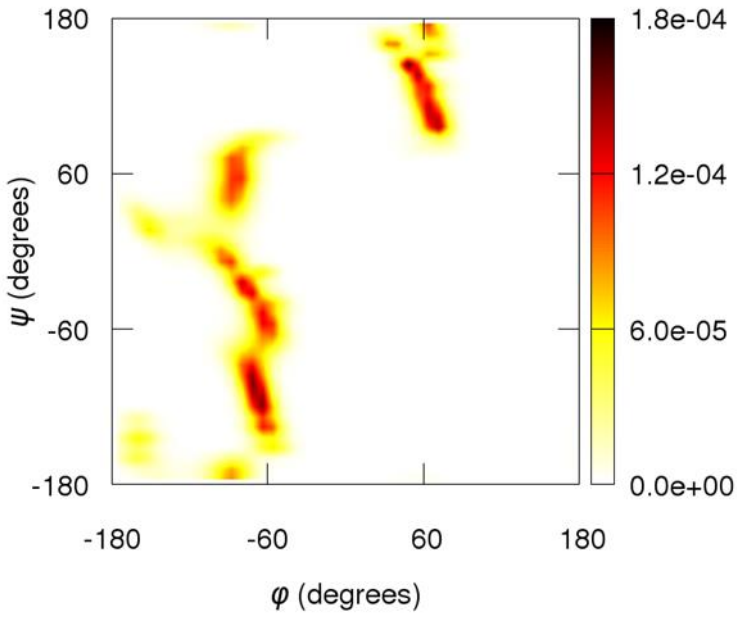

Figure 3.12 Normalised average occupation density of the φ_F/ψ_F rotational path generated from umbrella sampling

3.3.1.3 Accelerated molecular dynamics

A third biasing approach, accelerated molecular dynamics (aMD), was also employed for comparison. We report the free energy surface for sLe^a in explicit water obtained using dual boost aMD, such that a boost potential is applied based on the total and dihedral potential energies of the system. To compare with the eight msesMD replicas, we combine the results of eight independent 494 ns aMD simulations to obtain the final free energy surfaces (Figure 3.4). The aMD simulations provide good coverage of the Fucα(1→4)GlcNAc and Galβ(1-3)GlcNAc torsional surfaces, identifying all the main wells, $F_{1\text{-}4}$ and $G_{1\text{-}5,}$ in addition to an F_5 high energy well not seen in either unbiased MD or msesMD simulations. However, as demonstrated by the ψ_F and ψ_G time evolution profiles (Figure 3.7**A**-**B**), the F_4 and G_4 wells appear to be infrequently sampled resulting in noisy predictions of the well occupancies. Conversely, the aMD simulations offer greater sampling of the G_5 well (ψ_G = 90°) relative to msesMD as indicated by their respective ψ_G profiles (Figure 3.7**B**). We do note the rather flattened features of the reweighted aMD surfaces which appear to be masking finer details; this is particularly evident around the $F_{1\text{-}3}$ and $G_{1\text{-}3}$ regions which are much broader in aMD and appear to have merged into a single well (Figure 3.4). This coincides with an overestimation of the stability of F_3 and G_3 regions relative to both msesMD and unbiased MD, by around 2 kcal mol^{-1}. This may be a reflection of the magnitude of the applied aMD boost energy, which has a mean and range of 7.9 and 30.7 kcal mol^{-1} respectively; this compares to an applied swarm energy for msesMD of mean 4.1 kcal mol^{-1} and range 18.4 kcal mol^{-1} (Supplementary Figures B.1**A**-**B**). It is noted that this could be due to overboosting, arising from the choice of aMD parameters; however, softer boost parameters did not result in adequate transitions to the already undersampled F_5 and G_5 regions (data not shown). We predict that more advanced versions of the aMD method, such as the recently proposed Gaussian-aMD[41] may be more effective in sampling systems such as these.

3.3.2 Evaluating closed and open conformations of sLe^x in solution

Having established that the msesMD method provides a reasonable estimate of the conformational landscape of sLe^a, we apply msesMD alongside unbiased simulations to investigate the conformational flexibility of sLe^x. In particular, we examine the ability of the tetrasaccharide to adopt open conformational states in aqueous solution, such as the previously reported RSL-bound pose.[80]

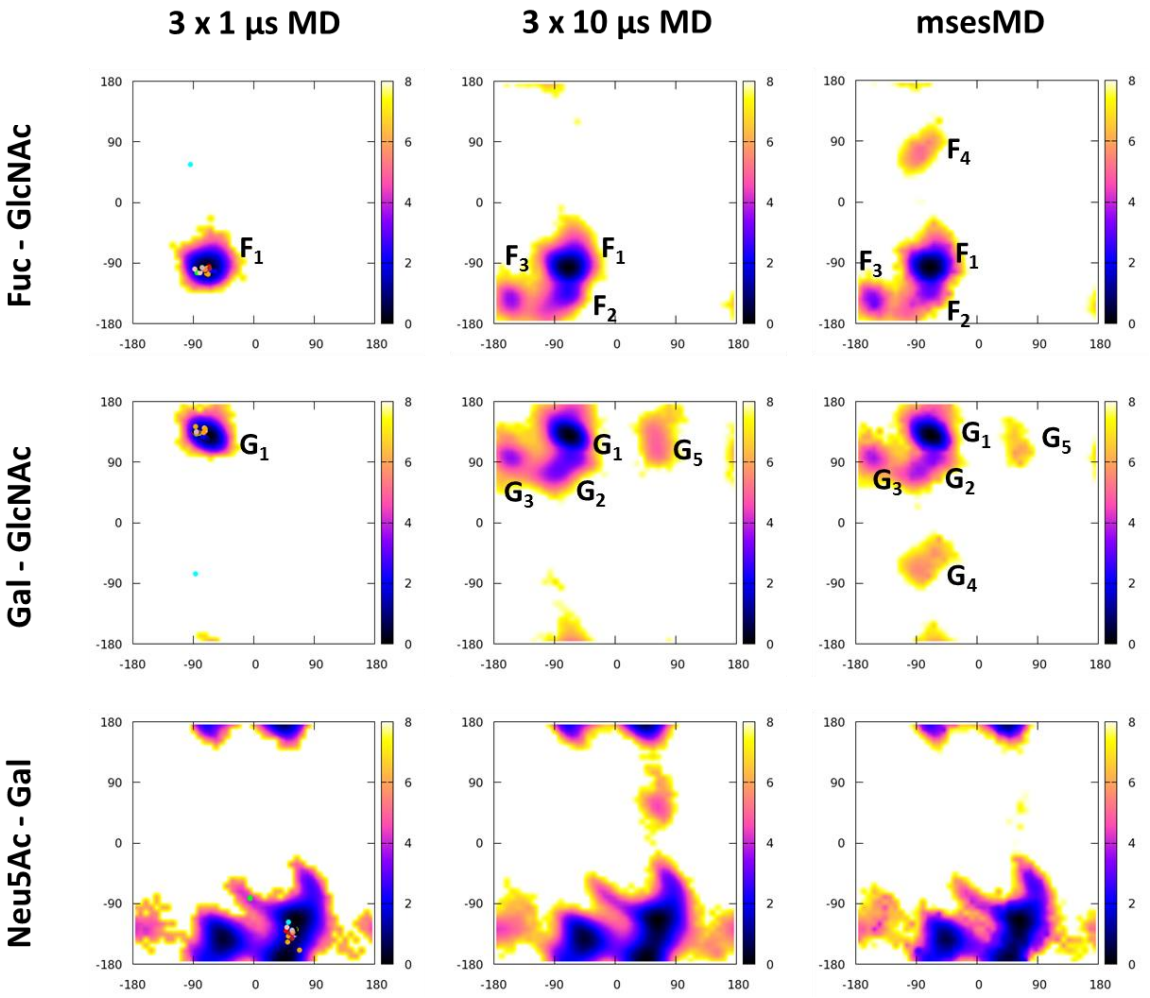

Figure 3.13 Free energy surfaces of the ϕψ glycosidic torsion of sLe^x computed via MD, msesMD and aMD. X-ray crystal positions are overlapped on the aggregate 3 µs profile.

3.3.2.1 Unbiased MD simulations

Firstly, we consider unbiased simulations of sLe^x in solution. Unlike sLe^a, triplicate 1 μs trajectories only sample the native closed conformations, with no sampling of neighbouring $F_{2\text{-}3}$ and $G_{2\text{-}3}$ regions (Figure 3.13). This suggests that sLe^x may be more rigid in solution than sLe^a. The longer 10 μs simulations explore non-closed minima of sLe^x, in the $F_{2\text{-}3}$ and $G_{2\text{-}3}$ regions, in addition to some infrequent sampling of the G_5 region (Figure 3.13). However, the F_4 and G_4 minima, which tend to be associated with the open protein-bound conformations of the tetrasaccharide[79-80], are not sampled. This is consistent with previous 25 μs simulations of sLe^x, which similarly did not appear to detect such conformations.[30] As for sLe^a, the Neu5Acα(2→3)Gal rotation is readily sampled (Figure 3.13).

Looking at the pucker free energy profiles, we find that for the GlcNAc ring, the triplicate 10 μs simulations sample the complete $^4C_1$-to-$^1C_4$ pucker transition (Figure 3.14 **A**); by contrast, the triplicate 1 μs simulations fail to sample the $^1C_4$ pucker. Given that the 1 μs simulations do not sample any non-native conformations, this difference in sampling of pucker would support the hypothesis of Topin et al.[80] that GlcNAc ring puckering is important to the exploration of open conformational states. This is further demonstrated in Supplementary Figure B.2, where a shift in the pucker states from is seen when moving along the path from F_1/G_1 to F_2/G_2. The GlcNAc ring primarily occupies a $^1C_4$ pucker when in the closed F_1/G_1 state; this changes to a mixture of skew-boat/boat and $^4C_1$ puckers in the F_2 and/or G_2 states. As for sLe^a simulations, access to the F_2 and G_2 wells is the primary route to open conformations, thus supporting the theory that ring inversion is necessary to achieve such events. For the remaining rings, in contrast to the 10 μs simulations, the triplicate 1 μs MD trajectories do not sample the boat/skew-boat puckers of the fucose and galactose rings, nor the $^4C_1$ pucker of Neu5Ac (Figure 3.14**B**-**D**).

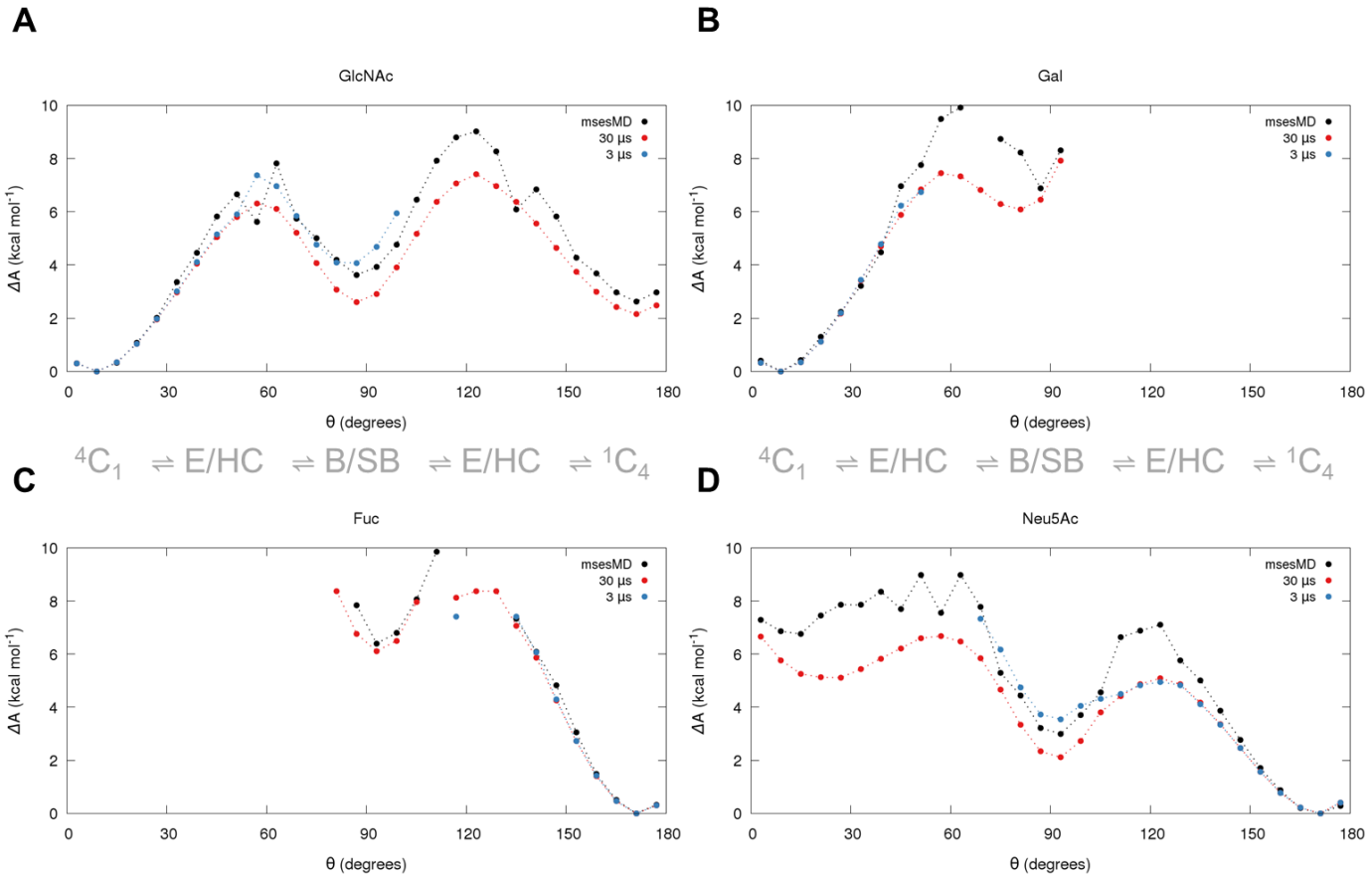

Figure 3.14 Cremer-Pople θ puckering profiles of sLe^x as calculated by aggregate 3 and 30 µs MD and msesMD for the four saccharide rings

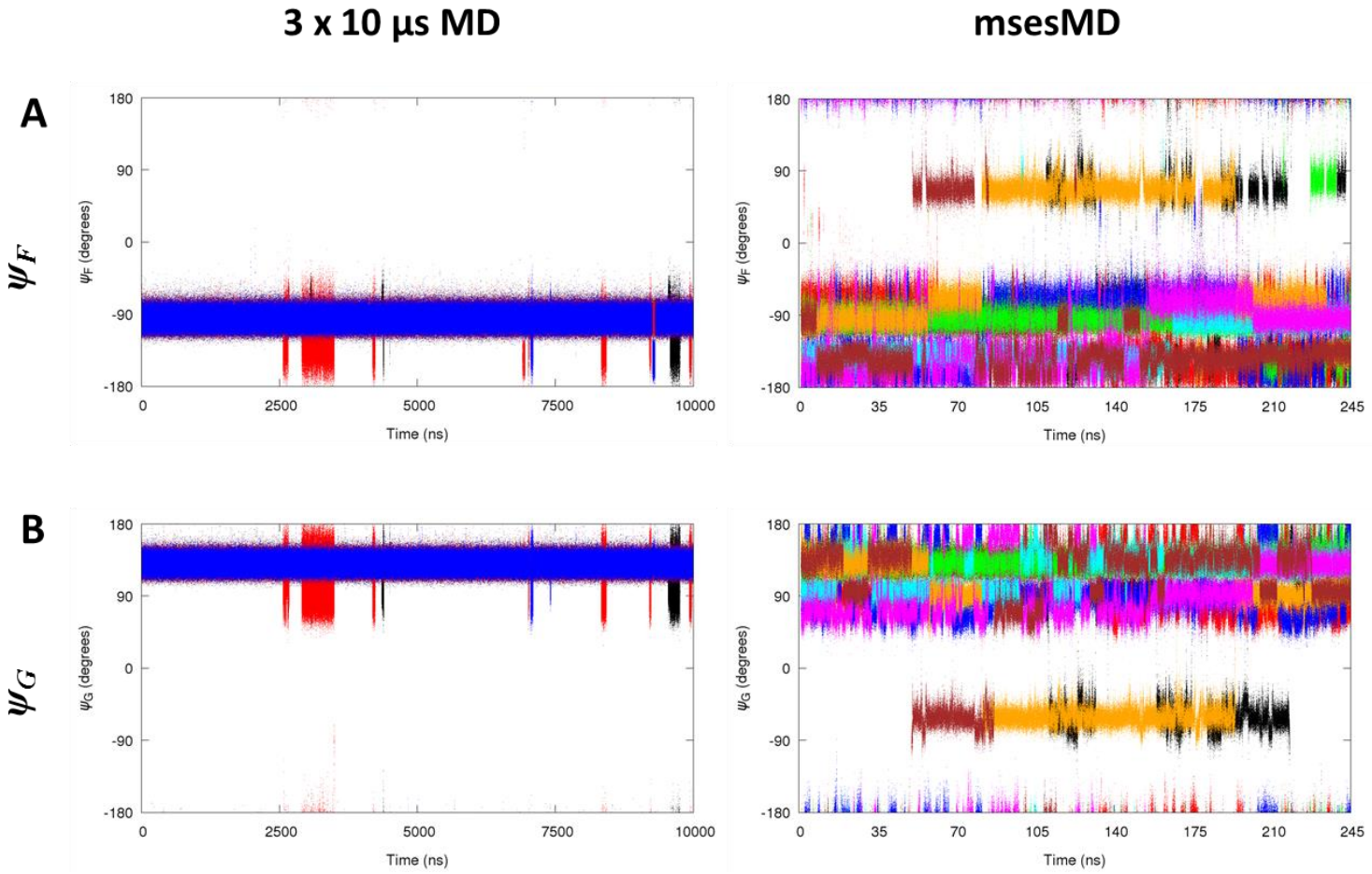

*Figure 3.15 (**A**) ψ_F and (**B**) ψ_G time series of sLe^x for unbiased MD, msesMD and aMD*

3.3.2.2 msesMD simulation

As for sLe^a, the 245 ns msesMD simulations of sLe^x in solution display greater similarity to the triplicate 10 μs rather than 1 μs unbiased simulations, with extensive sampling of the $F_{1\text{-}3}$ and $G_{1\text{-}3,5}$ regions (Figure 3.13). Additionally, F_4 and G_4 minima are identified by msesMD, although as seen from the ψ_F and ψ_G time evolution profiles (ψ_F = 60° and ψ_G = -60°), they are sampled to a lesser extent than sLe^a (Figures 3.7**A**-**B**, 3.15**A**-**B**). The F_4 and G_4 torsional values from msesMD match those of the RSL-bound sLe^x X-ray conformer, with φ,ψ values of (-70°,150°) and (-70°,-120°) for the Fucα(1→3)GlcNAc and Galβ(1→4)GlcNAc linkages respectively.[80] By clustering the msesMD trajectories, a structure closely matching the RSL bound conformer was found (Figure 3.16). It should be noted that whilst this RSL-like conformer was identified in the top 10 most occupied clusters (at position 9) upon reweighting, the reweighted normalised cluster density was only 9 x 10^{-5}, corresponding to an estimated relative free energy of around 5.5 kcal mol^{-1} above the closed state. This matches the msesMD predicted stability of the F_4 and G_4 wells as seen in the ϕψ torsional free energy maps (Figure 3.13), but does highlight the relatively low occupation of such conformations, which could easily be ignored if relying purely on clustering density to identify important conformers.

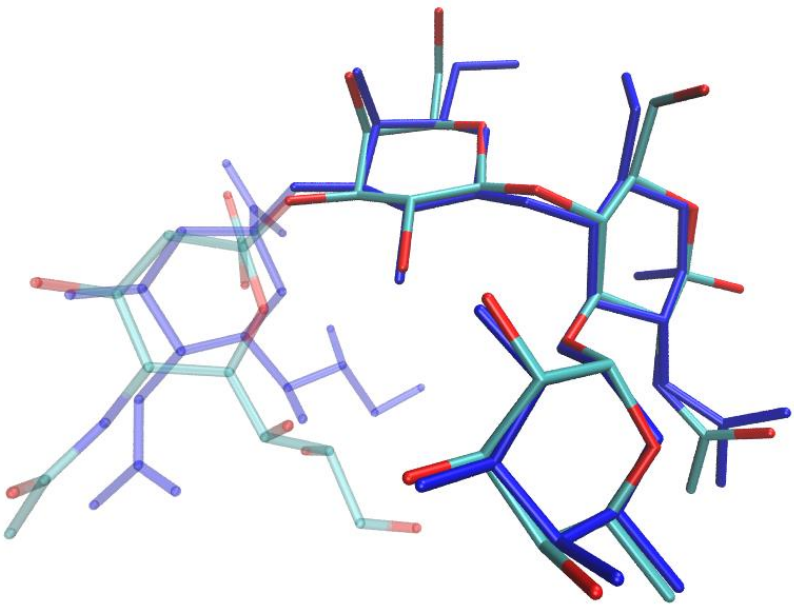

Figure 3.16 Overlap of 5AJC sLe^x RSL bound pose (blue) with clustered conformation from msesMD (coloured by atom type)

In regard to ring puckering in sLe^x, we find that for the GlcNAc residue, the complete $^4C_1$-to-$^1C_4$ landscape is sampled by msesMD simulation (Figure 3.14**A**), similar to the triplicate 10 μs unbiased MD simulations. However, somewhat large deviations of 1.0 and 1.7 kcal mol^{-1} from the unbiased MD profile are seen for the skew-boat/boat (θ = 90°) and the θ = 120° envelope/half-chair puckers respectively (Figure 3.14**A**). For the other pyranose rings, msesMD generally traces more complete profiles than the triplicate 1 μs MD simulations, but tends to deviate from the 10 μs estimates in the higher energy conformers (Figure 3.14**B**-**D**). As for sLe^a, this is particularly evident in the Neu5Ac ring, where msesMD estimate of the $^4C_1$ pucker sampling is very different from the triplicate 10 μs MD (Figure 3.14**D**).

3.3.3 Effect of sialylation on the conformational equilibrium

It is also interesting to assess the effect of sialylation of Le^a and Le^x on their closed/open equilibria in solution. Unbiased triplicate 10 μs molecular dynamics in aqueous solution of Le^a indicate a broadly similar pattern of sampling to the corresponding 30 μs trajectories of sLe^a (Figures 3.4 and 3.17). We note two small differences in the topologies of the free energy surface for glycosidic linkage Galβ(1-3)GlcNAc: first is the detection of an additional well, termed G_5, which is unseen in the unbiased sLe^a simulations (although seen in the enhanced sampling simulations of sLe^a); secondly, the G_3 region appears to be better sampled in simulations of Le^a. Furthermore, the G_4 and F_4 well stabilities differ slightly with the free energy estimates in Le^a being 0.8 and 0.7 kcal mol^{-1} less stable than for sLe^a (Figures 3.4 and 3.17). The Fucα(1→4)GlcNAc surfaces for both Le^a and sLe^a obtained via msesMD are near identical with similar well depths for all major regions (Figures 3.4 and 3.17). However, differences of up to 1.4 kcal mol^{-1} are seen in the reweighted Galβ(1-3)GlcNAc surfaces.

Conversely to the unbiased MD estimates, msesMD finds the G_3 well to be more stable in sLe^a by 0.5 kcal mol^{-1} and the G_4 well to be more stable in Le^a by 1.1 kcal mol^{-1} (Figures 3.4 and 3.17). Additionally, the G_5 well is estimated to be more stable in sLe^a by 1.4 kcal mol^{-1}. This discrepancy between unbiased MD and msesMD is in part explained by the poor sampling of these regions by unbiased MD, as can be seen by the ψ and φ time profiles (Figures 3.7, Supplementary Figures B.3, B.5 and B.7). Whilst the msesMD errors calculated via bootstrap sampling are generally within ±0.25 kcal mol^{-1} (Supplementary Figure B.9), one cannot discount the possibility that this may also be due to inaccuracies in the msesMD reweighted profiles.

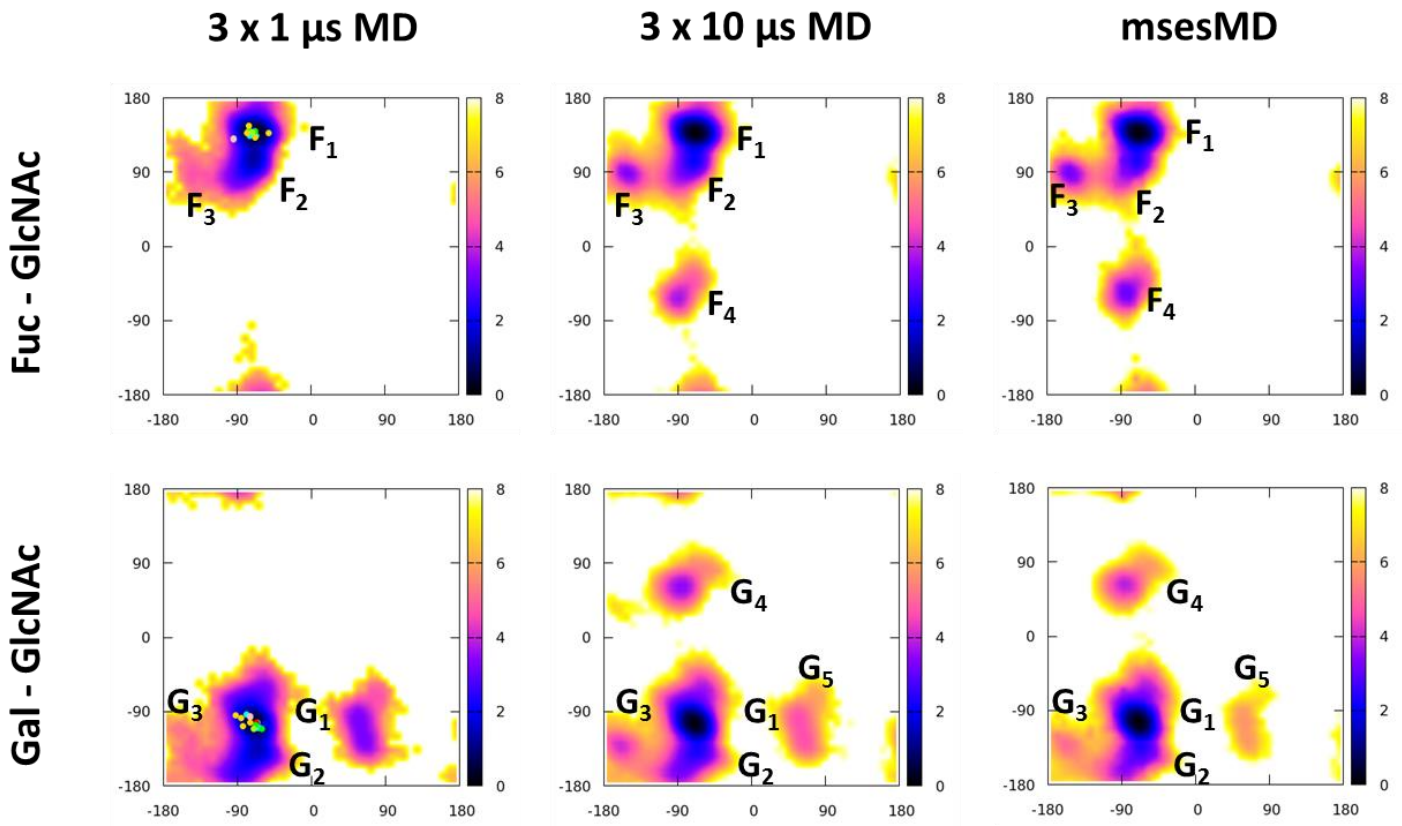

Figure 3.17 ϕ/ψ maps of the glycosidic torsion of Le^a generated via each method with different regions labelled accordingly. X-ray crystal positions are overlapped on the aggregate 3 µs profile.

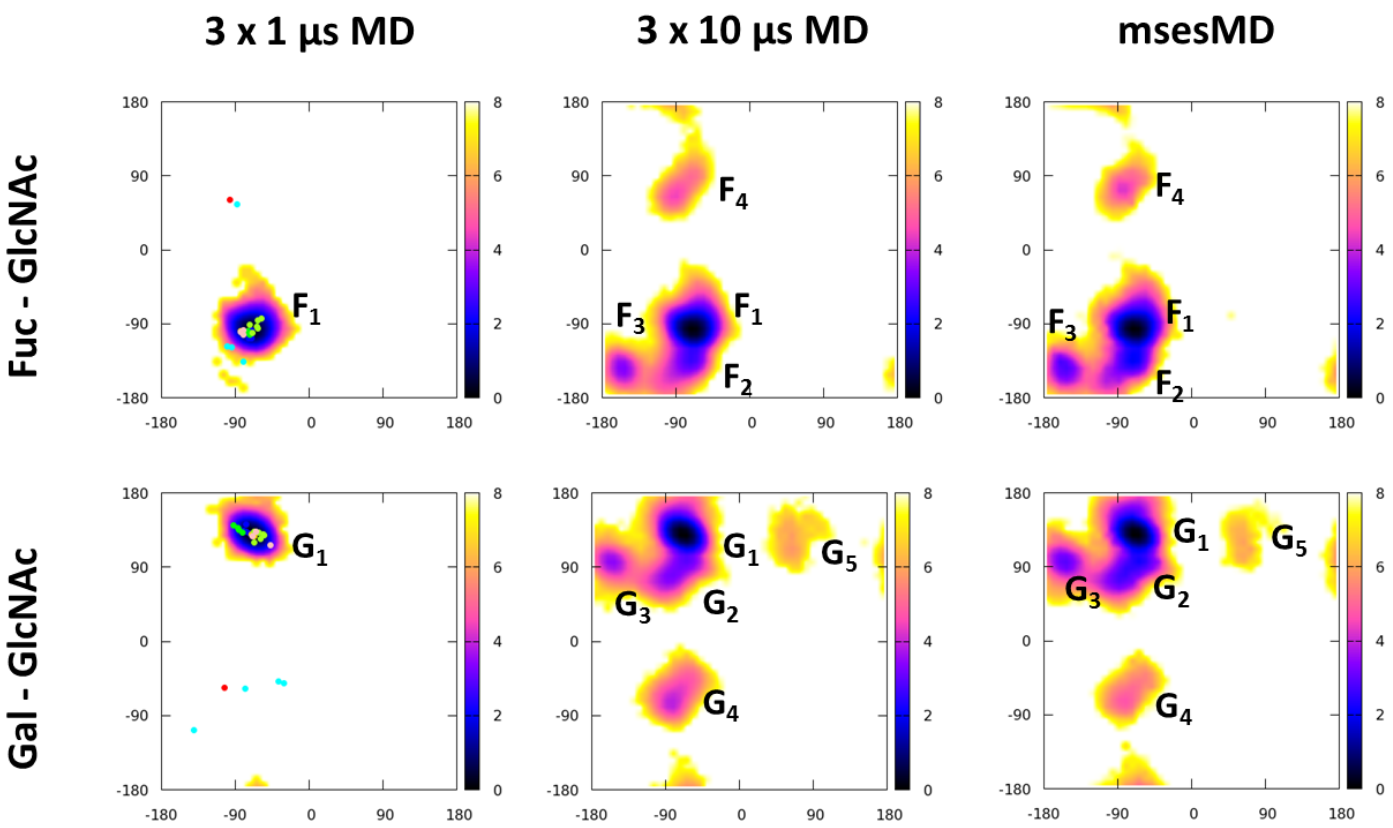

Figure 3.18 ϕ/ψ maps of the glycosidic torsion of Le^x generated via each method with different regions labelled accordingly. X-ray crystal positions are overlapped on the aggregate 3 µs profile.

Looking at the Le^x simulations, we find that sialylation appears to have a larger impact on conformer ensemble than for Le^a (Figures 3.13 and 3.18). The triplicate 10 μs MD simulations sample the F_4 and G_4 wells for Le^x, which were not identified in the sLe^x simulations. As previously discussed, these torsional wells are crucial to adopting the observed RSL-bound conformations of Le^x and sLe^x reported by Topin and co-workers.[80] Cluster analysis of both the unbiased and msesMD simulations found conformations which correspond to the bound crystal open poses (Figure 3.19**A**-**C**). The only exception is the X-ray conformer occupying the F_1/G_4 well, which was labelled the "Open III" conformer by Topin et al. (Figure 3.19**B**), for which a similar structure was only identified in the msesMD trajectories. In contrast to unbiased MD, the msesMD simulations were able to find the G_4 and F_4 wells in both the Le^x and sLe^x simulations (Figures 3.13 and 3.18). However, our premise that sialylation impacts on the stability of these conformers is supported, with reweighted msesMD free energy profiles estimating a 0.9 and 0.8 kcal mol^{-1} decrease in stability going from Le^x to sLe^x for G_4 and F_4 wells respectively. Interestingly, as discussed above, the msesMD simulations for Le^a and sLe^a also predict a decrease in stability for the G_4 well upon sialylation. This would indicate that sialylation may stiffen the Galβ(1-3)GlcNAc rotation, preventing the 180° flip in the ψ angle associated with the G_4 well.

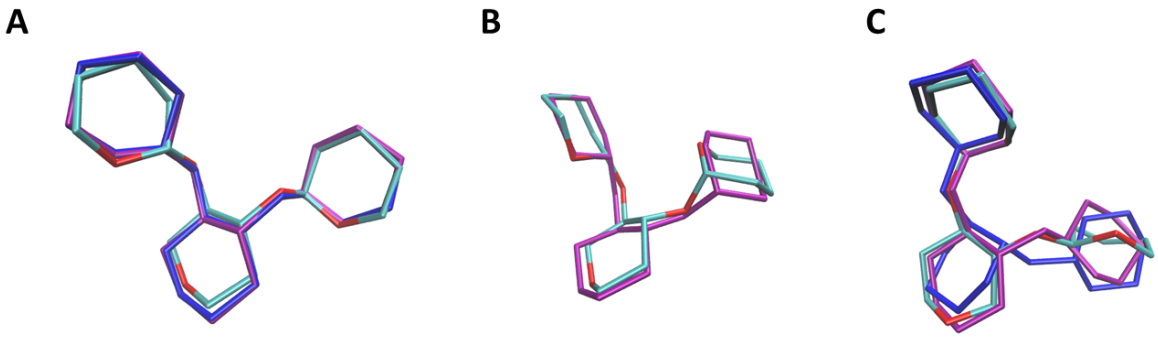

*Figure 3.19 Overlaps of the 5AJC and 5AJB RSL bound Le^x conformations, superimposing the core ring atoms of the crystal (false atomistic colours), with selected conformations of msesMD (purple) and unbiased MD (blue). These crystal structures correspond to the notation of Topin et al.[80] as **A**) Open I **B**) Open III and **C**) Open IV*

3.4 Conclusions

In this study, we have tested the use of the multidimensional variant of sesMD proposed in Chapter 2, exploring the propensity of the Lewis blood sugar antigens to form closed and open conformations. We find that 245 ns msesMD of sLe^a, using a swarm of eight simulation replicas, reproduces key features of triplicate 10 μs unbiased trajectories: the topological features of the glycosidic linkage free energy surfaces agree well; the extent of ring puckering of the key central GlcNAc ring is also in good agreement. It is noted that ring pucker sampling for other rings is limited, and should be a focus for future work (Chapter 4). The msesMD trajectories yield free energy profiles along ψ_F, the torsion angle most diagnostic of open and closed conformations, in closer accord with the profile from triplicate 10 μs simulations than calculated via umbrella sampling, avoiding the sampling issues associated with one-dimensional umbrella sampling. Furthermore, msesMD simulations identify a new G_5 well for sLe^a which is absent from the 10 μs simulations, but appears in aMD simulations and unbiased MD calculations of Le^a (Figures 3.4 and 3.17). Discrepancies in the well depths of sLe^a as calculated by unbiased MD and msesMD are found, particularly for the G_4 well. However, this is likely to arise from inadequate convergence in the unbiased MD simulations. Overall, these outcomes highlight the need for even longer unbiased MD simulation lengths in order to exhaustively sample the conformational space of sLe^a.

Applying msesMD simulations to four Lewises in solution suggests that the closed form is the lowest in energy conformer, comprising 93%, 89%, 92% and 97% of the reweighted cluster densities for Le^a, Le^x, sLe^a and sLe^x respectively. This agrees with the ubiquity of closed forms in both free and bound structures observed from X-ray spectroscopy (Figures 3.13, 3.17 and 3.18). Several clearly defined energy minima were identified along the torsional landscapes of the Gal-GlcNac and Fuc-GlcNac glycosidic linkages, corresponding to a large ensemble of non-closed conformational states. These minima have a range of stabilities, with those closer to the main closed (F_1 and G_1) wells, e.g. G_2 and F_2, lying within 1-2 kcal mol^{-1} of the closed form; and then more distant wells, e.g. G_5 and F_5, reaching over 5 kcal mol^{-1}. Access to the different open forms is achieved via transition paths visiting the different wells, most of which require initial access to the G_2 and F_2

states. As demonstrated for Le^x, this requires a shift in GlcNAc pucker, from $^4C_1$ to either skew-boat/boat or $^1C_4$. This agrees with previous proposals[80] that GlcNAc pucker acts as a mediator for closed-to-open transitions alongside the possibility of concerted torsional transitions in the core glycosidic linkages.

Access to the G_4 and F_4 regions, 180° in glycosidic torsion space from the closed regions, is crucial to formation of the RSL-bound geometries of Le^x and sLe^x; these open conformer regions appear to be lower in energy when sialylation is not present. Although there is a lack of agreement on this observation between msesMD and unbiased simulations, it is believed that a similar behaviour may also be true for Le^a/sLe^a, albeit possibly to a lesser extent. Interestingly, the K_d was measured for the binding of Le^x tetrasaccharide and sLe^x pentasaccharide to RSL by Topin et al.[80] These sugars have the primary structure Galβ(1→4)(Fucα(1→3))GlcNAcβ(1→3)Gal and Neu5Acα(2→3)Galβ(1→4)(Fucα(1→3)) GlcNAcβ(1→3)Gal respectively, differing from Le^x and sLe^x in possessing an extra galactose residue distal to the sialylation site. The K_d of Le^x tetrasaccharide was found to be half that of sLe^x pentasaccharide, at 26 and 58 μM respectively. This supports our theory that sialylation leads to an increased energetic cost of accessing the torsional wells of the bound pose. We also note that while conformers closely matching the RSL bound crystal poses were found (within 1 Å heavy atom RMSD), some slight deviations in the relative orientations of the rings were seen (Figure 3.19). This reflects the induced fit nature of protein-ligand binding, where conformational states which would normally be unnatural in solution may be adopted due to interactions with protein side chains. Nevertheless, our simulations were able to detail the subtle conformational flexibility of these four Lewis oligosaccharides, identifying rare conformers that would usually lie below the detection threshold for NMR.

The work presented in this chapter demonstrates that msesMD can be a useful tool in overcoming energy barriers to sample rare minima that are usually accessed at very long timescales. The msesMD method reasonably reproduces free energy surfaces obtained by multi-microsecond unbiased MD simulations and is approximately five times faster in achieving this, acquiring 245 ns in 5 days, relative to the 25 days required by HMR simulations (using four infiniband connected 24 core Haswell E5-2680 nodes and three

Nvidia K20 GPU nodes respectively). It is noted that without the use of HMR, the unbiased simulations would have required nearly two months of calculation. Indeed one could envisage the use of msesMD in tandem with HMR in order to lower further the time to solution (see Chapter 5). Furthermore, although not shown here, shorter msesMD simulation times on the order of 100 to 150 ns would have been suitable for these systems, albeit with some slight additional noise. Overall, application of the msesMD method is able to reduce the challenge of sampling these oligosaccharides from several weeks to a matter of days.

While *a priori* knowledge of possible boost coordinates is required when using msesMD, the use of glycosidic torsions, as previously noted by similar locally biased methods[40], seems to be effective. The form of the enhancing potential used here appears to be effective for these carbohydrates, only introducing a slight amount of noise as seen in sLe^a and sLe^x as the number of boosted dihedrals increases (Figures 3.5 and 3.13). Unlike aMD, smoothing-based reweighting methods such as the Maclaurin series are not required in order to recover reasonable estimates of ensemble averages. The GPU accelerated aMD simulations did only require 60% of the compute time of msesMD; however, the resultant surfaces (Figure 3.4) were too noisy to recover fine details about the dynamics of the Lewises. However, the advantage of aMD to quickly estimate potential modes of motion does suggest it could be used effectively alongside msesMD, guiding the appropriate choice of boost coordinates in complex systems.

In conclusion, the msesMD method appears well suited as a tool for the exploration of oligosaccharides. We therefore anticipate its use in future investigations of other small to medium sized carbohydrate systems, such as the fucosylated ABO blood group sugars. It is noted that sampling of slow modes of motion which were not explicitly boosted was at times limited. An important example of this is the ring pucker of the Gal, Fuc and Neu5Ac rings, which for the most part was only on a par with shorter unbiased simulations. Unfortunately, the choice of a suitable boost coordinate for carbohydrate rings is not intuitive. Nevertheless, in the next chapter, we will attempt to address this issue.

Chapter 4: Sampling carbohydrate ring dynamics

4.1 Chapter Introduction

Aside from the rotation of glycosidic torsions, there exists another important albeit somewhat more subtle slow degree of freedom in carbohydrates, ring motion. Events along this degree of freedom, also known as ring puckering, describe the adoption of different monosaccharide ring conformations, generally exchanging between stable chair (C) orientations, via boat (B), skew-boat (S), half-chair (H) and envelope (E) conformers. Being able to describe puckering in carbohydrates is important, as the adoption of specific ring conformers have frequently been identified as key steps in the biological action of carbohydrates. A well-known example of this is the adoption of the $^{2}S_{0}$ ring conformers in L-iduronic acid units of the therapeutically used glycosaminoglycan, heparin, which has been shown to be essential to the activation of antithrombin, heparin's main target for its action as an anticoagulant.[97] Another example is the role of puckering in the formation of transition states in post-translational carbohydrate processing by enzymes such as the glycoside hydrolases.[98] Beyond being directly involved as a component of carbohydrate structure-activity relationships, ring puckering is also thought to sometimes play an important role in subtlely modulating the shape and dynamics of polysaccharides. Such a case was seen in our investigations of the Lewis oligosaccharides (Chapter 3) whereby the puckering of the GlcNAc ring is thought to act as an accessory event to the "opening" of the trisaccharide core. A larger scale example of the impact of puckering on dynamics was shown in coarse grained simulations of long heparan sulfate chains, where the inclusion of ring puckering was demonstrated as essential to the accurate description of the macroscopic properties of the polymer.[82]

Unfortunately the free energy landscape of carbohydrate ring motion is difficult to traverse with high energy barriers on the order of several kcal mol^{-1} separating stable conformations. Due to this, $^{4}C_{1}$ to $^{1}C_{4}$ pucker transition events usually occur on the multi-nanosecond to microsecond timescales, with a minimum of at least 5 to 10 microseconds required to effectively sample the complete conformational landscapes of monosaccharide

rings.[31, 81-82] Such timescales place ring puckering dynamics in the inconvenient position of being too fast to be easily detailed by experimental approaches,[99] whilst also being too long to fully explore in systems consisting of more than a few monosaccharide units via conventional molecular dynamics. Investigations of puckering behaviour in simulations of carbohydrates therefore rely on very long simulation times (on the order of 5 to 10 μs) using long time-step methods (i.e. hydrogen mass repartitioning)[31, 81-82] or the use of conformational restraints to probe specific pucker conformers.[100-101] However, such approaches are limited: in the former case, simulating long time scales can be prohibitively time consuming for large system sizes. The use of restraints in the latter case not only unnaturally biases the system of interest, but also becomes too complex to use when more than a couple of ring conformers need to be investigated. To address this, enhanced sampling methods have been successfully adopted to probe ring dynamics. Examples of such methods include metadynamics schemes[102-104]; a combined adaptive biasing force/hamiltonian replica exchange protocol[105]; and the adaptive reaction coordinate force method[106]. However, the applications of such methods to different systems of interest have, unfortunately, been somewhat limited. This in some ways reflects the complexity of such schemes, particularly when scaling the number of degrees of freedom under investigation.

In Chapter 3, it was shown that the msesMD method can effectively be used to explore rare conformational changes in oligosaccharides. However, it was noted that due to the fact that only the glycosidic torsions were biased, the exploration of some of the ring conformations was limited. In this chapter, as a first attempt to address this issue, we investigate the applicability of the msesMD method in characterising the puckering of pyranose rings. We first validate the approach on four benchmark systems, namely the α and β anomers of D-glucose and the uronic acid epimers α-L-iduronic acid and β-D-glucuronic acid. Demonstrating good agreement with multi-microsecond unbiased MD, the msesMD method is then used to characterise the impact of different sulfation patterns on the puckering behaviour of the free monosaccharide units of the glycosaminoglycans heparan sulfate/heparin and chondroitin sulfate. Although previous attempts have been made at exploring the dynamics of these sulfated monosaccharides[31, 100, 107], their outcomes have usually been limited due to sampling issues. Thus, the results presented here describe the first exhaustive molecular mechanics characterisation of the influence of sulfation on the puckering behaviour of glycosaminoglycan monosaccharides.

4.2 Methods

4.2.1 Describing a set of boost coordinates for puckering

Unlike glycosidic linkages which can be easily decomposed into sets of torsions, the choice of boost coordinates to describe the puckering of rings is more complex. Ideally one would want to choose a set of coordinates that directly describes the transition between ring conformers, such as the Hill-Reilley[108] or Cremer-Pople[109] coordinates. Doing so is likely to ensure the most efficient use of the boost potential, which is why previous applications of enhanced sampling to the exploration of puckers have generally used them[106, 110]. Unfortunately, due to time constraints, implementing and validating such a coordinate systems within the framework of the msesMD method was not possible. Thus as a first approximation we instead decided to decompose the pyranose ring into two equidistant ring torsions, U_1 and U_2, defined by the O5-C1-C2-C3 and C3-C4-C5-O5 atoms respectively (Figure 4.1). Such a choice is obviously non-ideal as it relies on the perturbation of the ring atoms to implicitly result in pucker transitions. This could potentially lead to a somewhat "wasted" boost potential as the atoms are perturbed along non-relevant degrees of freedom and inaccuracies in describing the transition paths between conformational states. However, it is noted that such a choice of boost coordinates was successfully employed in a previous Hamiltonian replica exchange study of the puckering behaviour of the methylated iduronic and glucuronic acids.[105]

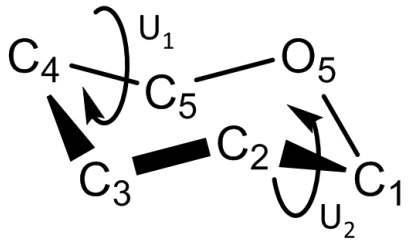

Figure 4.1 Puckering boost coordinates defined as two equidistant torsions, U_1 and U_2

4.2.2 Simulation details

System preparation: All systems were built using the *leap* module in AmberTools14[32]. Parameters for the carbohydrates were obtained from the Glycam06 (version j-1) force field[95], with additional parameters for the unsubstituted and N-sulfated glucosamine units obtained from a separate Glycam06 release for glycosaminoglycan monosaccharides[111]. Parameters for O-sulfation of the glycosaminoglycan monosaccharides were obtained using the transferable sulfation method for the Glycam06 force field as described by Singh et al.[111], where a fixed charge sulfate residue was added to position of interest followed by a +0.031 charge adjustment on the linking oxygen to achieve a net integer charge on the modified sugar. Such an additive approach to sulfation has previously been shown to successfully reproduce the behaviour of heparin/heparan sulfate chains[100, 111], although considering the relatively short simulation times used (~1 μs), the impact of excluding polarisation effects when probing rare conformers of such highly charged systems is unclear.

All systems (Figures 4.2 and 4.3) were built as free monosaccharides in the $^4C_1$ ring configuration with hydroxyl groups at the C1 position, except from α-L-iduronic acid and β-D-glucuronic acid for which O-methylated systems were also built. The systems were neutralised using sodium ions where appropriate and explicitly solvated in truncated octahedrons of TIP3P[66] waters with the waters placed up to distance of a minimum 12 Å away from the solute for the msesMD simulations or 13 Å for the HMR MD simulations. The reason for the discrepancy in the water box sizes is due to a bug in the CUDA implementation of *pmemd*[1-2] of AMBER14[32], which leads to random floating point errors in systems with less than ~3000 atoms. Thus, since the multi-microsecond HMR MD simulations used the CUDA accelerated *pmemd*, a larger water box was used to prevent this issue. Since the msesMD simulations use the CPU variant of *pmemd*, which is not prone to this bug, the smaller 12 Å distance was used to reduce simulation costs. This difference in box sizes is unlikely to have any measurable impact on conformational behaviour, as the systems built using the 12 Å solvation distance are adequately dilute.[112]

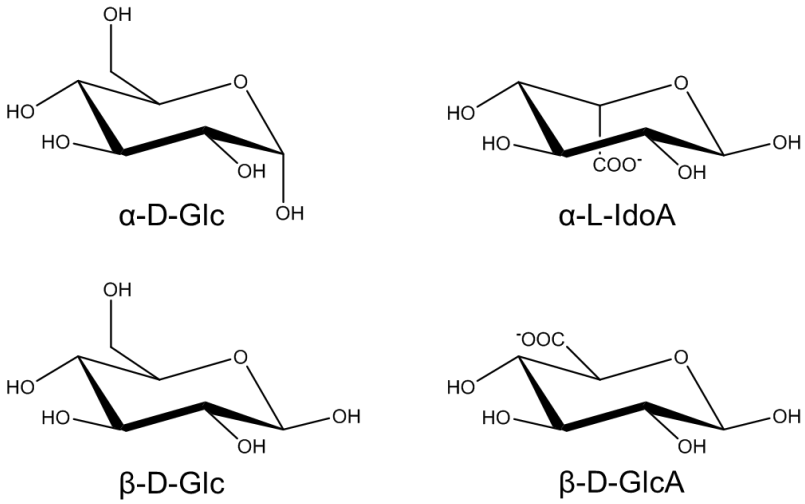

Figure 4.2 The four benchmark monosaccharide systems; α-D-glucose (α-D-Glc), β-D-glucose (β-D-Glc), α-L-iduronic Acid (α-D-Glc) and β-D-glucuronic Acid (β-D-GlcA)

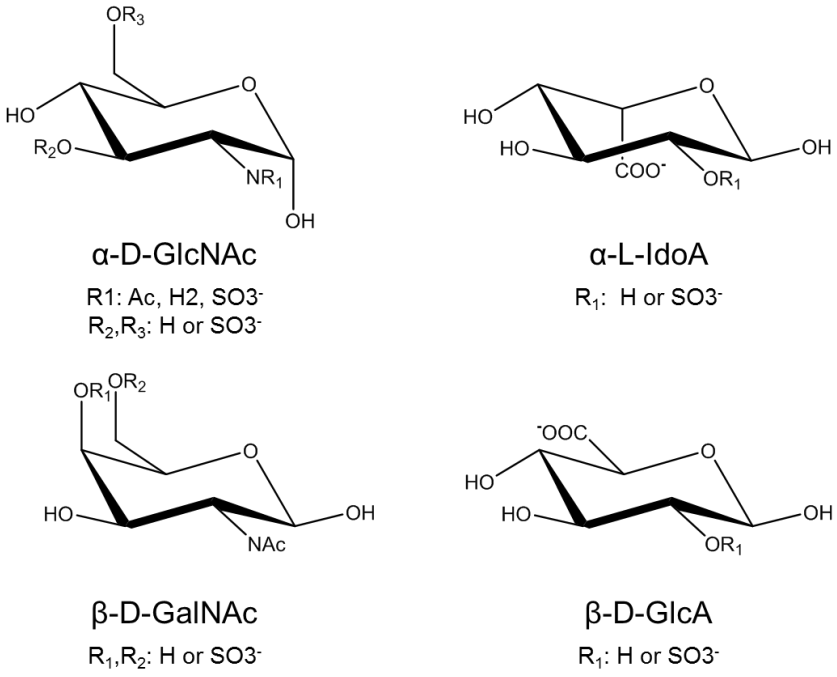

Figure 4.3 The four glycosaminoglycan monosaccharides and their sulfation patterns investigated in this study

Equilibration protocol and general simulation details: A generalised equilibration protocol was used for all simulations. In order to resolve any atom clashes, the solvated monosaccharide boxes were first energy minimised using 250000 steps of the steepest

descent algorithm followed by 250000 steps of conjugate gradients. The systems were then slowly heated over 500 ps under constant volume conditions (NVT) from 0 K to the target temperature of 298 K. The box density was then equilibrated to a target pressure of 1 bar via 1 ns of constant pressure (NPT) equilibration using the Monte Carlo barostat[25], with volume exchange attempts every 100 steps. The system was then further equilibrated under NVT conditions for a further 1 ns. A cut-off of 9 Å was used for short range non-bonded interactions, with long range electrostatics being handled via the particle mesh Ewald approach[67]. All production simulations were carried out under the canonical ensemble (NVT) and temperature control was achieved using the Langevin thermostat[25], with a collision frequency of 3 ps^{-1} and a target temperature of 298 K. To achieve long time-steps, solute hydrogen motion was constrained using both the SHAKE[27] and SETTLE[33] algorithms, applied to the solute and rigid TIP3P water model respectively.

msesMD simulation protocol: The msesMD simulations were carried out via the generalised simulation protocol discussed in Chapter 2.4.1, using 8 replicas and the pair potential parameters A = -0.5 kcal mol^{-1}, B = 0.5 rad^{-1}, C = 2.13 kcal mol^{-1}, D = 2.625 rad^{-1}. In all cases, the msesMD potential was applied to the two ring torsions defined in section 4.2.1. Upon equilibration, the system was replicated into 8 independent trajectories and allowed to diverge over a 1 ns period under NVT conditions. The msesMD potential was then slowly introduced across the replicas over a 600 ps period, and then allowed to equilibrate for a further 400 ps. This was then followed by 150 ns per replica (200 ns for the test systems) of msesMD production simulation, with the first 5 ns then discarded as a further swarm equilibration period. All msesMD simulations used a time integration step of 2 fs with system energies and trajectories saved every 1 ps.

HMR simulation protocol: Upon equilibration, as described above, individual production simulations using unbiased MD were propagated for a total of 10 μs. This was further extended by 5 μs for α-D-glucose and β-D-glucuronic due to poor sampling of chair interconversion events. In order to reduce simulations costs by increasing the time step to 4 fs, the default AMBER hydrogen mass repartitioning scheme was used, where solute hydrogen masses were scaled to 3 amu, reducing the adjacent heavy atom mass accordingly (see Chapter 1.2.2 for more details).[26] Unlike msesMD simulations,

trajectories were saved every 5 ps so as to reduce the data storage associated with such long simulations. It is noted that some puckering events are estimated to have transition state lifetimes of a few picoseconds[113]; thus it is possible that the chosen sampling frequency may be too low to accurately characterise such events. Nevertheless, considering that other multi-microsecond unbiased simulations have generally opted for longer sampling intervals (~10 ps)[30-31, 82], it is felt that an interval of 5 ps should be sufficient to validate the msesMD method.

Analysis methods: Analysis was achieved via a mix of in-house python scripts and the *cpptraj* program from AmberTools16.[59] In this study, we focus on use the Cremer-Pople[109] θ and ϕ puckering parameters as means of characterising the puckering behaviours of the different monosaccharides. Specifically, we compute the θ free energy surfaces to obtain information about the occupation density of the general ring conformational states ($^1C_4$, $^4C_1$, half chair/envelope and boat/skew-boat states). This allows us to readily identify deviations in the ring dynamics that may occur as the result of using different methods; simulation times or substitution of different functional groups. Furthermore, estimated free energy heat maps of the θϕ surfaces are calculated to gain more detailed insight into the occupation of skew-boat, boat, envelope and half-chair states. As per previous chapters, free energy estimates of the surfaces were calculated by histogramming the generalised coordinate surfaces and using the canonical relationship between the Helmholtz free energy and the microstate density (eq 4.1).

$$\Delta A_n = -k_B T ln(p_n/p_m) \quad [4.1]$$

For unbiased MD simulations, estimates of the microstate density was achieved by the "counting" approach[114], whilst for the msesMD simulations, the density was reweighted according to the biasing potential as described in Chapter 2.4.2. A histogram bin size of 6° was used, in addition to maximum energy cut-offs of 12 kcal mol^{-1} and 8 kcal mol^{-1} for the θ and θϕ surfaces respectively. In order to gain insights into the msesMD simulation errors, bootstrap analysis was carried out. This was achieved by randomly resampling the simulation dataset 100,000 times and calculating unbiased free energy surfaces for each resample. Error estimates for each bin was then calculated as the standard deviation in the bin energy estimate across all resamples.[12, 115]

4.3 Validation of the msesMD method for pucker exploration

4.3.1 Introduction

In this section, we evaluate the feasibility of using the msesMD approach to accurately probe carbohydrate ring dynamics. To achieve this, we compare the results of msesMD against those of the current "gold standard" in the investigation of puckering, namely long unbiased multi-microsecond MD simulations. We compare four benchmark systems; α-D-glucose (α-Glc), β-D-glucose (β-Glc), α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA), as seen in Figure 4.2. These were primarily chosen due to their biological relevance[116], and because they have been used in the validation of previous enhanced sampling methods[99, 105, 113, 117], the latter being particularly useful in cross-validating our results. Furthermore, the four systems can be seen as two pairs, one of anomers and the other of epimers, varying in the relative orientation of their ring substituents. Thus, being able to accurately describe the puckering differences which occur as a result of such changes should demonstrate the msesMD method's ability to investigate the impact of different ring substitutions.

4.3.2 Results and Discussion

4.3.2.1 Glucose anomers

We first consider thepuckering free energy profiles for both D-glucose anomers as a function of the Cremer-Pople angle θ(Figure 4.4). Analysis of the 10 μs unbiased MD simulations was performed for the first 1, 5 and full 10 μs in order to assess convergence. As seen from the α-D-glucose profile (Figure 4.4**A**), all but the 1 μs simulation reproduce a similar profile, exploring the full $^4C_1$ (θ = 10°) to $^1C_4$ (θ = 170°) chair interconversion pathway. As expected minima are found at the $^4C_1$, $^1C_4$ and boat/skew-boat (θ = 90°) positions. The $^4C_1$ is seen as the lowest energy state, with the $^1C_4$ and boat/skew-boat puckers found to be around 1.5 and 4.5 kcal mol^{-1} less stable respectively. For β-D-glucose, the full $^4C_1$ to $^1C_4$ profile is sampled by all unbiased MD simulation lengths, except for the envelope/half-chair pucker region around θ = 110-140° (Figure 4.4**B**). The β-D-glucose profile also shows the $^4C_1$ pucker to be more stable, however unlike α-D-glucose the $^1C_4$ and boat/skew-boat puckers are predicted as nearly equi-energetic, being around 3.5 kcal mol^{-1} less stable than $^4C_1$. It is however noted that the 1 μs simulation estimate of the $^1C_4$ stability differs from the other simulation lengths, estimating its stability to be 0.8 kcal mol^{-1} lower. This difference highlights the impact of inadequate sampling on the free energy estimates.

Comparing both msesMD and unbiased MD simulations we find that both approaches correctly identify the β anomer to have a less stable $^1C_4$ configuration compared to the α anomer (Figure 4.4). This behaviour is believed to be due to large steric interactions between the hydroxymethyl and the anomeric hydroxyl groups resulting from groups both taking axial positions in $^1C_4$ the arrangement. However, comparing the 10 μs MD results with those of msesMD, we find that whilst the β anomer profile is closely matched (aside from discontinuities over the 110-140° region of the unbiased MD profile) by both methods (0.2 kcal mol^{-1}), large discrepancies (1.6 kcal mol^{-1}) can be seen in the non-4C1 regions of the α anomer.

Considering at the variation in the value of θ over time for the α-D-glucose unbiased MD simulation (Figure 4.5**A**), we find that the cause of this inconsistency appears to be linked to relatively poor sampling of the $^4C_1$ to $^1C_4$ interconversion. As seen in Figure 4.5**A**, the $^1C_4$ conformer is only sampled twice during the first 10 μs. By comparison, the 10 μs β anomer trajectory demonstrates a higher frequency of chair interconversion events, albeit with much shorter $^1C_4$ lifetimes, with 8 such events observed over the 10 μs (Figure 4.5**B**). Additionally, boat/skew-boat conformers (θ = 90°) are frequently observed for β-D-glucose, whilst for the α anomer, this tends to only occur when starting from the less frequently occupied $^1C_4$ conformer (data not shown). Thus it is likely that, in the case of α-D-glucose, a 10 μs simulation may be insufficient to accurately sample ring dynamics, resulting in unconverged estimates of the microstate densities and, by association, an incorrect estimate of the free energy profile.

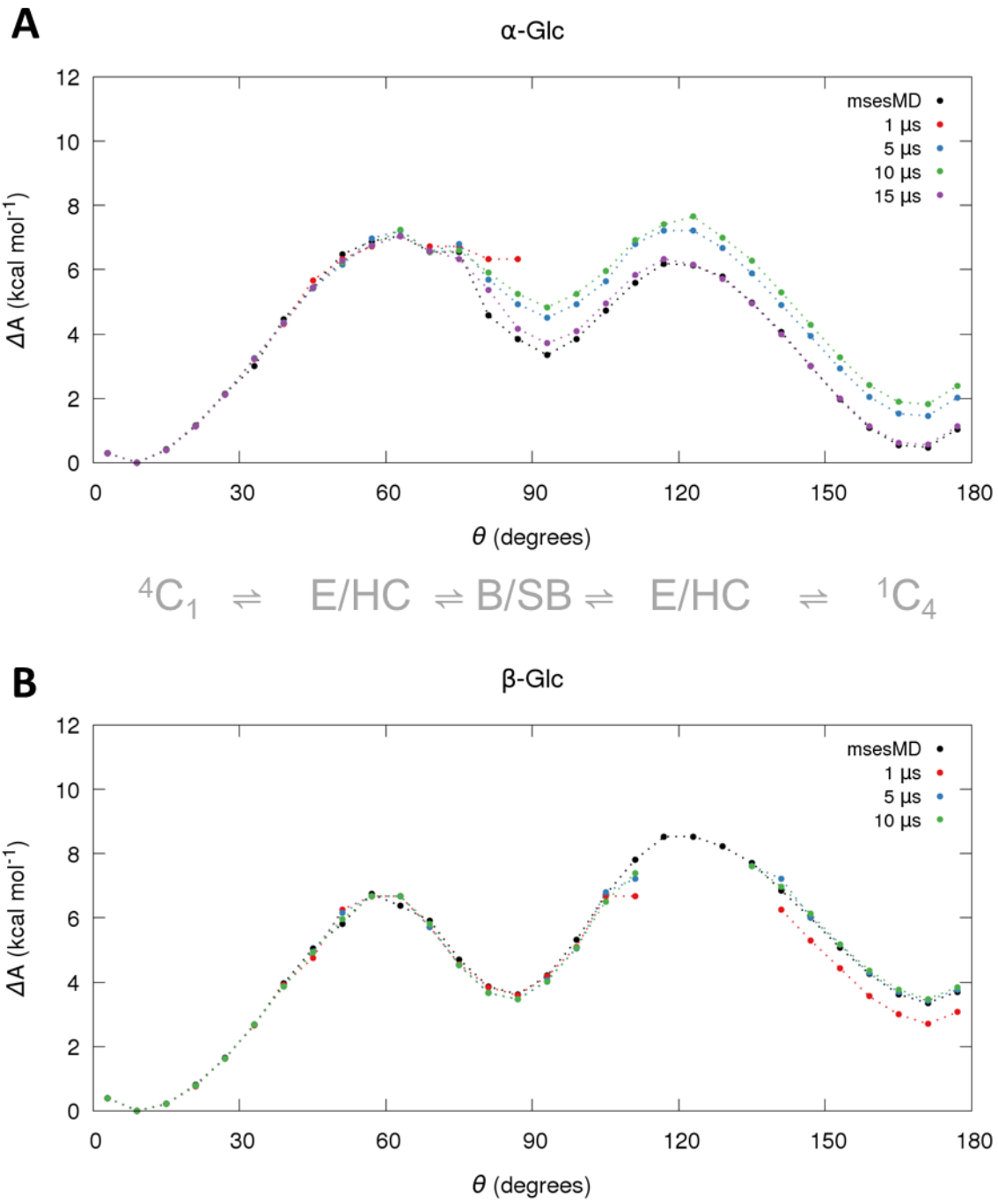

Figure 4.4 Cremer-Pople θ angle relative free energy profiles for both α-*D-glucose (α-Glc) and* β-*D-glucose (β-Gluc) calculated via both msesMD and unbiased MD*

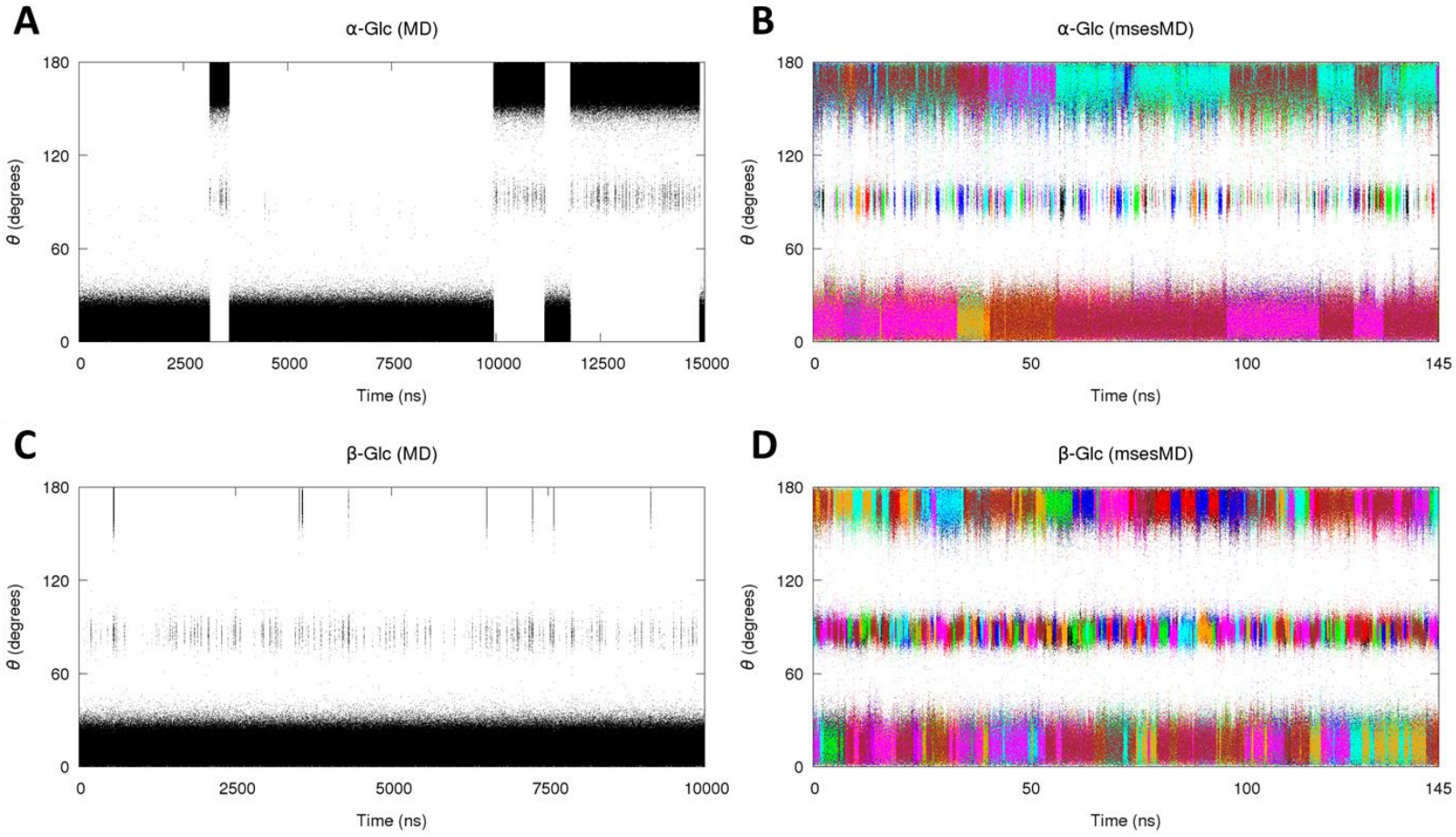

Figure 4.5 Profiles of the change in the Cremer-Pople θ angle over time for α-D-glucose (α-Glc) and β-D-glucose (β-Gluc)

In order to verify if the results are indeed limited by sampling, the unbiased MD simulation for α-D-glucose was extended for a further 5 μs. As we can see from the free energy profile (Figure 4.4**A**), the extended 15 μs simulation results are in better agreement with those of the msesMD simulation, with a maximum deviation of around 0.4 kcal mol^{-1}. Looking at the change in θ over time (Figure 4.5**A**), it seems that this shift in the computed free energy profile is the result of the monosaccharide occupying the $^{1}C_{4}$ conformer for most of the last 5 μs, which is in stark contrast to the first 10 μs. This tendency for the system to become kinetically trapped in a certain pucker for multiple microseconds at a time demonstrates that true convergence in the ring dynamics is probably only attained at timescales above 20 μs. This is far longer than the previously estimated 5-10 μs.[31, 81] One also notes that most of the differences between the 15 μs MD and 145 ns msesMD profiles are in the estimation of the boat/skew-boat conformational stability (θ = 75 to 105°) rather than that of the chairs. Thus, this warrants closer inspection of the distribution of the different pucker states visited.

From the pucker distribution along Cremer-Pople the θϕ coordinates(Figure 4.6), we find that both methods produce similar surfaces. In the case of α-D-glucose (Figure 4.5**A**-**B**), both the msesMD and MD simulations identify a relatively stable minimum in the $^{1}S_{5}$ and $B_{2,5}$ regions (θ = 90°, ϕ = 250-300°). However, whilst the path to this well from the $^{1}C_{4}$ state, going through the $^{1}H_{2}$ and E_{2} (θ = 120°, ϕ = 250-300°) conformers, is well defined by both methods, the path from $^{4}C_{1}$ shows some differences. In particular, the msesMD method does not properly identify the presence of a high energy region (around 5 kcal mol^{-1} above $^{4}C_{1}$) corresponding to the $^{5}S_{1}$ conformer (Figure 4.6**B**) which is seen in the unbiased MD simulation (Figure 4.6**A**). Instead, the msesMD replicas predominantly sample a puckering path via the ^{2}E pucker (θ = 60°, ϕ = 120°). Unfortunately, due to poor sampling of the chair interconversion process in the unbiased MD simulation (Figure 4.6**A**), it is unclear as to whether or not the $^{5}S_{1}$ conformer plays an important role as a transition state in this. In fact, out of 6 such events during the 15 μs MD simulation, only one was recorded as accessing the $^{5}S_{1}$ conformer (Supplementary Figure C.1). Nonetheless, this may indicate that the msesMD methodology applied here, is susceptible to poor sampling of rare high energy transition states.

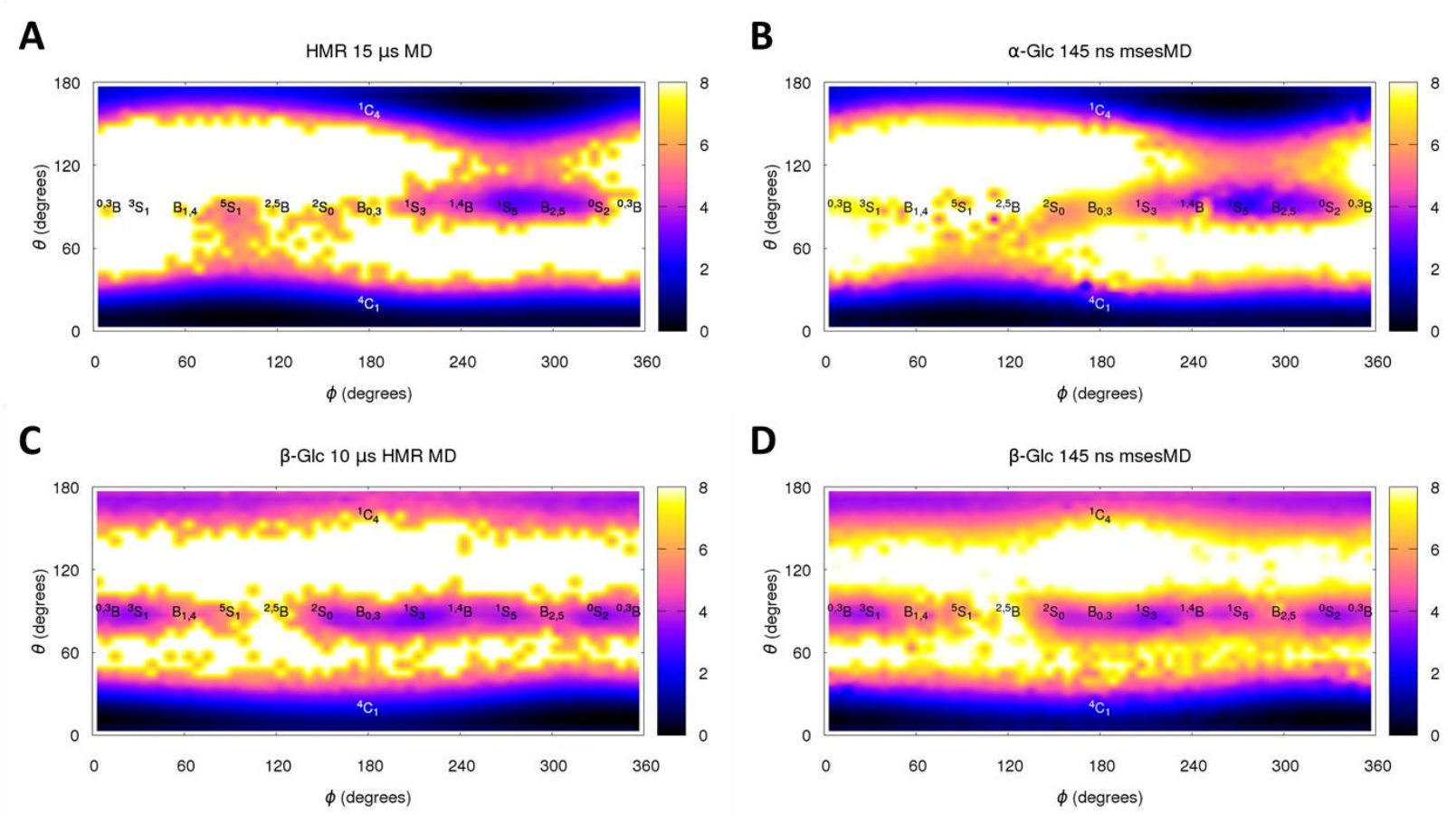

Figure 4.6 Cremer-Pople θ vs ϕ puckering free energy profiles for α-D-glucose (α-Glc) and β-D-glucose (β-Gluc)

In this particular scenario, a possible cause is that a combination of the relatively strong swarm biasing potential and non-optimal boost coordinate is pushing the ring away from the $^{4}C_{1}$ state along the "shortest" path to the stable $^{1}S_{5}$ / $B_{2,5}$ (θ = 90°, ϕ = 250-300°) region, ignoring alternate (and potentially lower energy) transition states. Evidence for this can be observed from the results of an msesMD simulation of α-D-glucose where the biasing potential is effectively reduced by a factor of 4 (parameters A = -0.125 kcal mol^{-1}, C = 0.5325 kcal mol^{-1}). As we can see from the $\theta\phi$ profile (Supplementary Figure C.2), this results in much better sampling of the $^{5}S_{1}$ conformer, although it reduces the number of chair interconversion events observed by each replica (Supplementary Figure C.3), whilst still reasonable in this case, one believes that in systems which higher energetic barriers, such as in the case of β-D-glucuronic acid seen below, this could lead to sampling issues.

For the β-D-glucose, the $\theta\phi$ free energy profiles show good agreement between both msesMD (Figure 4.6**D**) and unbiased MD (Figure 4.6**C**). Both methods demonstrate a generally flexible ring which adopts a variety of different stable non-chair conformers, covering nearly all skew-boat and boat orientations except from $^{2,5}B$ (θ = 90°, ϕ = 120°). This "floppy" ring behaviour echoes the findings of previous semi-empirical QM investigations, particularly using AM1 and PM3, of β-D-glucose puckering.[118] Unlike the α anomer, the chair interconversion path cannot be easily defined from this profile. This is mainly due to the fact that there appears to be several possible transition paths, resulting in occupation of many different envelope and half-chair conformers. This agrees with previous studies of β-D-glucose[113, 118].

Finally, we compare our results against those of previously published studies. Overall, the general trends which have been observed (e.g. the $^{1}C_{4}$ conformer being more stable in the α anomer relative to β) agree with previously reported simulations.[99, 102, 113, 118] However, due to force field differences, the magnitude of the relative free energy for the varying conformational states is quite different from our results (on the order of several kcal mol^{-1}). In fact, both the unbiased MD and msesMD results presented here demonstrates one of the lowest reported free energy differences between the $^{4}C_{1}$ and $^{1}C_{4}$ states for α-D-glucose. [99, 102, 113, 118] Interestingly, our results do not match those of the two previous studies of glucose ring puckering using the same Glycam06 fore field. In a study by Spiwok et al.[117],

which looked solely at β-D-glucose in water, the estimated relative free energy difference between the two chair states of the solvated monosaccharide is around 1 kcal mol^{-1} lower than our estimates. Curiously, the author's reported vacuum results are in better agreement with our solvated results with a $\Delta A(^4C_1\text{-}^1C_4)$ value of 3.4 kcal mol^{-1}. In a study by Plazinski et al.[99], the relative free energy estimates for both the α and the β anomers are very different, with deviations of greater than 4 kcal mol^{-1} from our $\Delta A(^4C_1\text{-}^1C_4)$ values.

It is unclear as to why such differences are seen. Whilst it could be partly linked to methodological differences, as both studies used metadynamics-based enhanced sampling schemes, the large differences between the two studies appear to indicate force field implementation issues. In fact, looking at the Plazinski et al. results we find that the relative free energies are closer to the "AMBER scaled" results of the Spiwok et al. study, whereby the 1-4 terms are scaled to the values of the AMBER protein force field (1.2 and 2.0 for the electrostatics and van der Waals interactions respectively) rather than using unscaled 1-4 terms, as per the standard Glycam06 implementation.[95] This hints at the possibility that the Plazinski et al. study may have used incorrect scaling parameters when adapting Glycam06 parameters for use within the Gromacs MD engine. With regards to the differences between our results and the Spiwok et al. study, it is likely a mixture of both small differences in the force field implementation and methodology. Considering that such differences can have significant implications on the dynamics of carbohydrates, this outlines the importance of being able to quickly and accurately profile the puckering properties of monosaccharides, using methods such as msesMD, when developing and validating new force fields.

4.3.2.2 Uronic acids

We now look at our two uronic acid benchmark systems; α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA). As seen from the Cremer-Pople θ coordinate puckering free energy profiles (Figure 4.7**A**), the unbiased MD simulation results for IdoA show relatively good agreement between the different simulation lengths. Unlike the glucose simulations, both chair conformers ($^4C_1$ and $^1C_4$) show similar stability, being around 2.3 kcal mol^{-1} more stable than the boat/skew-boat puckers. One does note that there are some slight

fluctuations in the estimates of the chair stabilities, of up to 0.5 kcal mol^{-1}, across the different simulation times. For example, the 1 μs simulation results shows the $^{1}C_{4}$ to be 0.4 kcal mol^{-1} higher than the $^{4}C_{1}$, whilst at 5 μs, the inverse is seen with the $^{4}C_{1}$ being 0.5 kcal mol^{-1} less stable than the $^{1}C_{4}$. This fluctuation eventually converges at the 10 μs timescale, with both pucker states seen as around 0.1 kcal mol^{-1} of each other. Conversely, the GlcA profiles calculated via unbiased MD shows large differences in the relative stability of the two chair puckers, with the $^{1}C_{4}$ conformer being 2.7 kcal mol^{-1} less stable (Figure 4.7**B**). Whilst there is good agreement on the stability of the $^{4}C_{1}$ and boat/skew-boat puckers between the three unbiased MD simulation lengths, only the 10 μs simulation results identify the presence of the $^{1}C_{4}$ conformer.

Comparing the θ coordinate free energy profiles of IdoA for msesMD and unbiased MD, we see that there is good agreement between the two methods (Figure 4.8**A**). Both methods describe the near equivalent occupancy of both chair states, with a 2.3 kcal mol^{-1} less stable boat/skew-boat state. Small deviations are seen in the estimates of the chair stabilities, however they are within simulation error (less than 0.2 kcal mol^{-1}) and therefore likely irrelevant. The Cremer-Pople θϕ free energy surfaces are also similar (Figure 4.8**A-B**) for both methods; these identify the $^{3}S_{1}$ and $^{2}S_{0}$ skew-boat conformers (θ = 90°, ϕ = 30° and θ = 90°, ϕ = 160° respectively) as the predominant non-chair low energy wells. Higher energy boat/skew-boat conformers are also occupied along the profile, including a fairly populated path between the $^{3}S_{1}$ and $^{2}S_{0}$ wells via the $^{5}S_{1}$ pucker (θ = 90°, ϕ = 100°).

Considering the dynamics of θ over time (Figure 4.9**A**), we can see that chair interconversion events occur frequently for iduronic acid, with around 24 such events occurring over the 10 μs simulation. This explains the relatively good estimate of the free energy surface obtained using the first microsecond of the simulation. Nevertheless, as previously mentioned, there are some deviations in estimated PMF across the 1, 5 and 10 μs simulation lengths (Figure 4.6). This error (e.g. ±0.5 kcal mol^{-1} in the chair stabilities), demonstrates that even for flexible monosaccharides, simulation of up to around 10 μs are required, rather than the 3 μs described in previous studies of iduronic acid.[81]

In terms of β-D-glucuronic acid, most of the free energy profile shows good agreement between unbiased MD and msesMD simulations (Figure 4.7**B**). However whilst both methods identify a relatively high energy $^{1}C_{4}$ conformer, caused by an unfavourable equatorial to axial reorientation of all the ring substituents, there is an up to 0.6 kcal mol^{-1} difference in the estimate. From the change in θ over time (Figure 4.9**C**) it appears that, as for α-D-glucose, the source of the discrepancy is limited sampling by the unbiased MD simulation. In fact, the $^{1}C_{4}$ chair is only briefly observed once throughout the 10 μs trajectory. Unfortunately extending the simulation by an additional 5 μs, whilst decreasing the $^{1}C_{4}$ stability estimate (by around 0.2 kcal mol^{-1}), did not yield more chair transitions. Such a low rate of chair interconversion reflects the high barrier (~ 6.5 kcal mol^{-1}) associated with the skew-boat/boat to $^{1}C_{4}$ transition; likely, much longer simulation lengths are require possibly ~ 50 μs, to obtain sufficient sampling of the $^{1}C_{4}$ conformer. Unlike $^{1}C_{4}$, skew-boat/boat conformations are observed frequently, and are well defined by both methods. A closer look at the conformer distribution (Figure 4.9**C**-**D**) shows that the $^{1}S_{3}$ and $B_{0,3}$ forms (θ = 90°, ϕ = 180-230°) are preferentially adopted by the pyranose ring, with less frequently visited states along the $^{1,4}B$ to $^{0,3}B$ range (θ = 90°, ϕ = 240-360°).

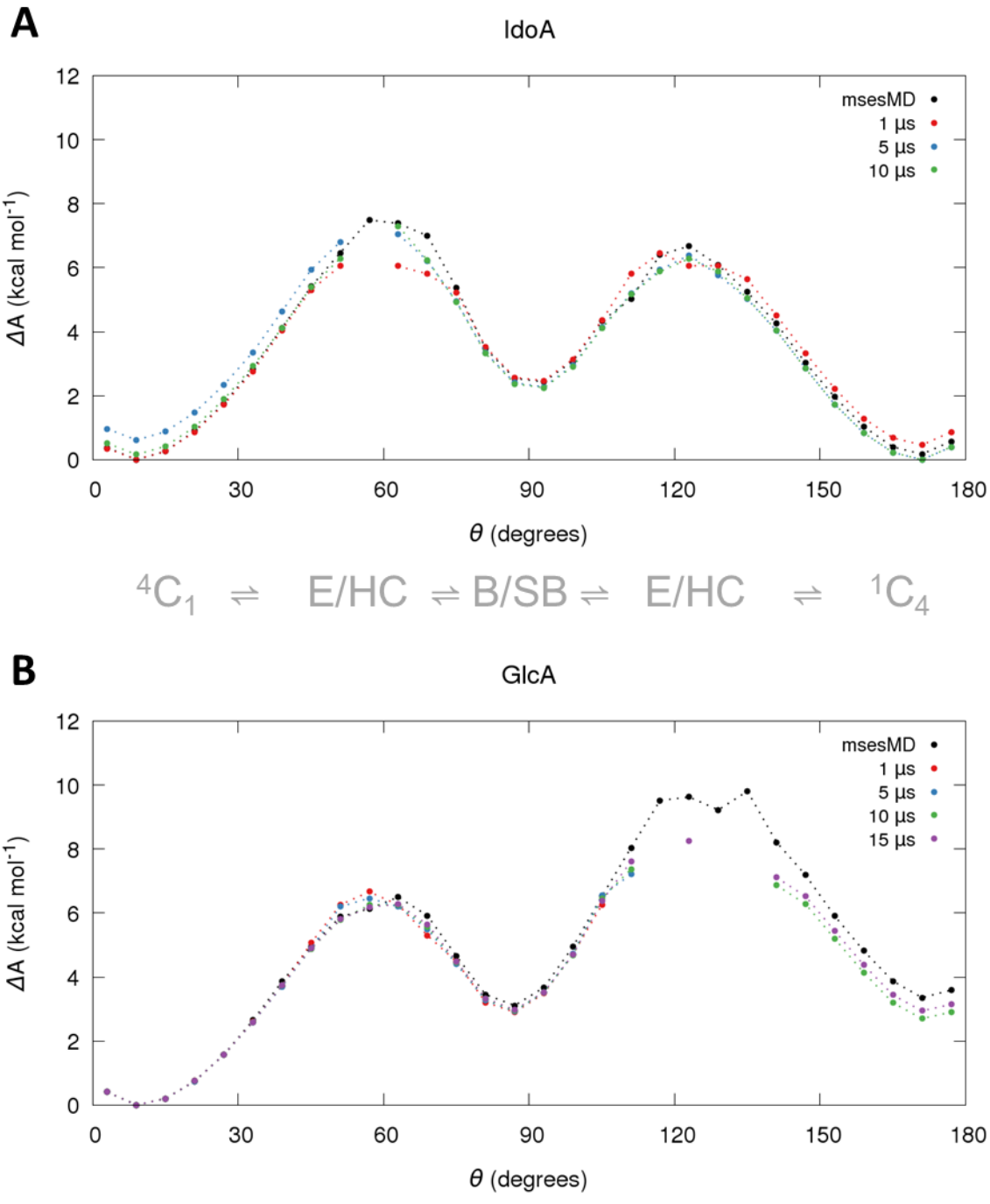

Figure 4.7 Cremer-Pople θ angle relative free energy profiles for both α-*L-iduronic acid (IdoA) and* β-*D-glucuronic acid (GlcA) calculated via both msesMD and unbiased MD*

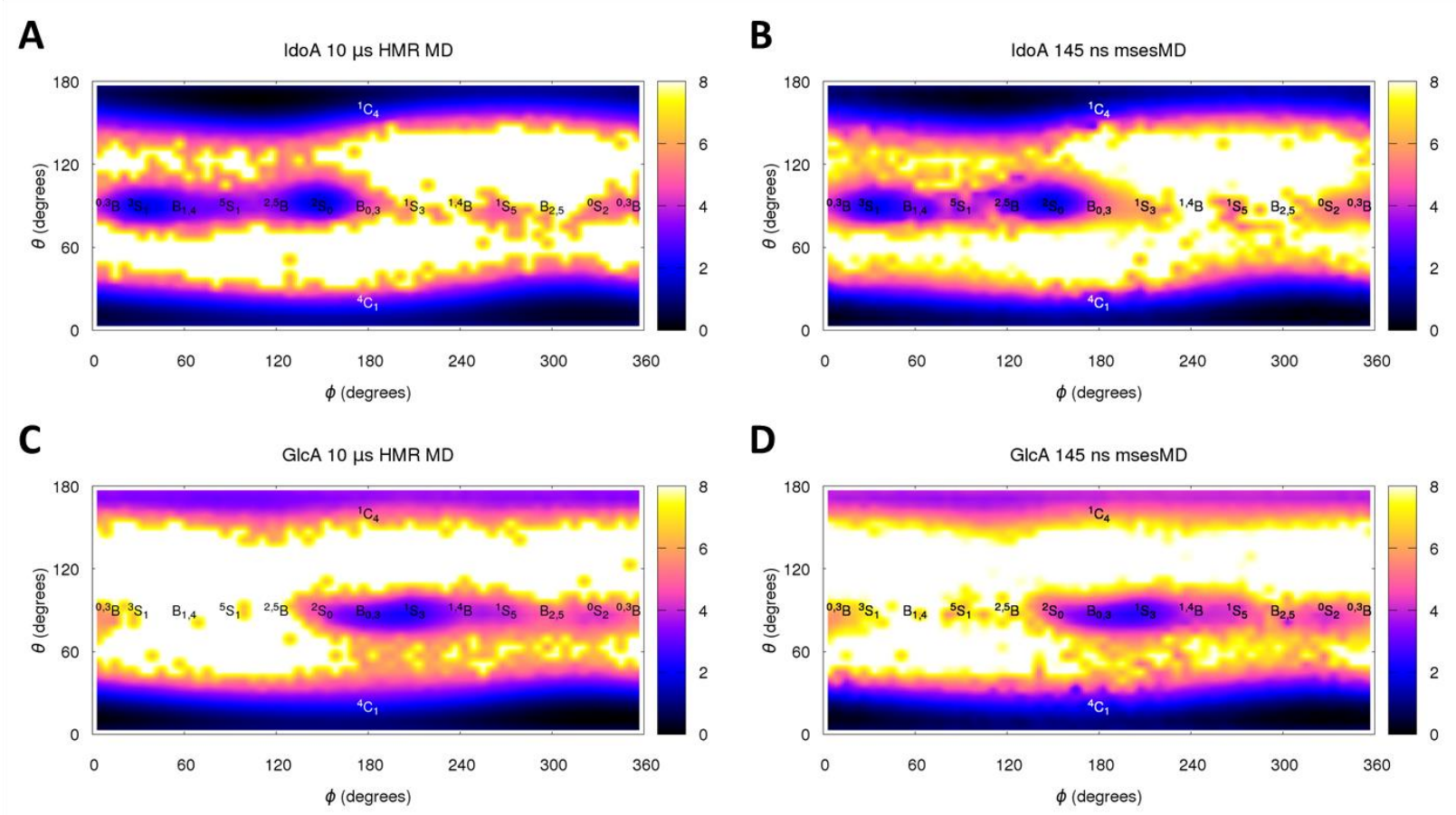

Figure 4.8 Cremer-Pople θ vs ϕ puckering free energy profiles for α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA)

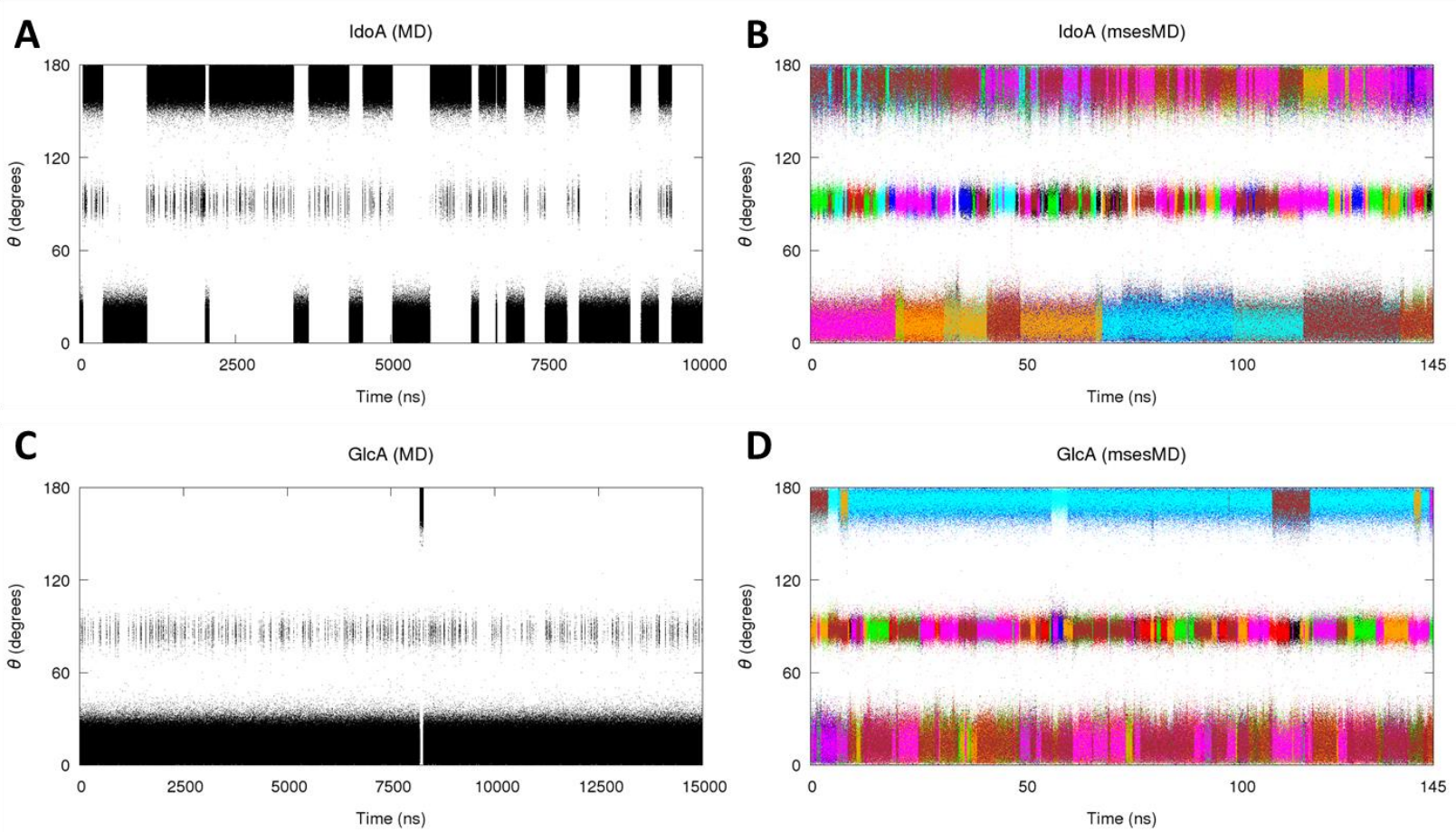

Figure 4.9 Profiles of the change in the Cremer-Pople θ angle over time for α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA)

Unfortunately comparing the results presented here against published data is difficult as the only two prior studies which investigated the ring dynamics of the uronic acids using the Glycam06 force field, opted for monosaccharides which are O-methylated at the anomeric position. As this substitution on the anomeric carbon can have a significant impact on the ring dynamics[31, 116], due to changes in interactions between ring substituents that stems from the presence of the methyl, a direct comparison would yield differences. Thus, in order to effectively cross-check our results against the literature, O-methylated variants of both uronic acids were also calculated using the msesMD method (Figure 4.10**A**-**B**).

For α-L-iduronic acid, introduction of an O-methyl at the anomeric position destabilises the $^{4}C_{1}$ conformer, resulting in an estimated stability 0.6 kcal mol^{-1} higher than that of the $^{1}C_{4}$ conformer (Figure 4.10**A**). The reason for this change in the relative free energy of the two conformers may be linked to a stabilisation of the axial position due to the anomeric effect. The estimated value is in good accord with previous Hamiltonian replica exchange[105] and multi-microsecond unbiased MD simulations[81] which estimate $^{4}C_{1}$ to be less stable than $^{1}C_{4}$ by 0.7 kcal mol^{-1} and 0.9 kcal mol^{-1} respectively. It is noted that the higher-than-error deviation for the latter study is likely due to unconverged sampling: the study only sampled from two independent 5 μs simulations; and differences in the MD engine, using ACEMD rather than AMBER. Interestingly, in their analysis, Sattelle et al. state that, compared to experiment, their calculated results appear to underestimate the iduronic acid $^{4}C_{1}$ stability by ~ 0.5 kcal mol^{-1}.[81] This would place the msesMD results closer to experiment. O-methylation had very little impact on the boat/skew-boat conformer occupation (Figure 4.11**A**,**C**), with the only differences stemming from a decrease in the stability of the $^{2,5}B$ conformer and an increase in the stability of the $^{1}S_{3}$ and $^{1}S_{5}$ states. The $^{3}S_{1}$ and $^{2}S_{0}$ forms are still the dominant non-chair puckers which matches the results from the multi-microsecond MD study.[81]

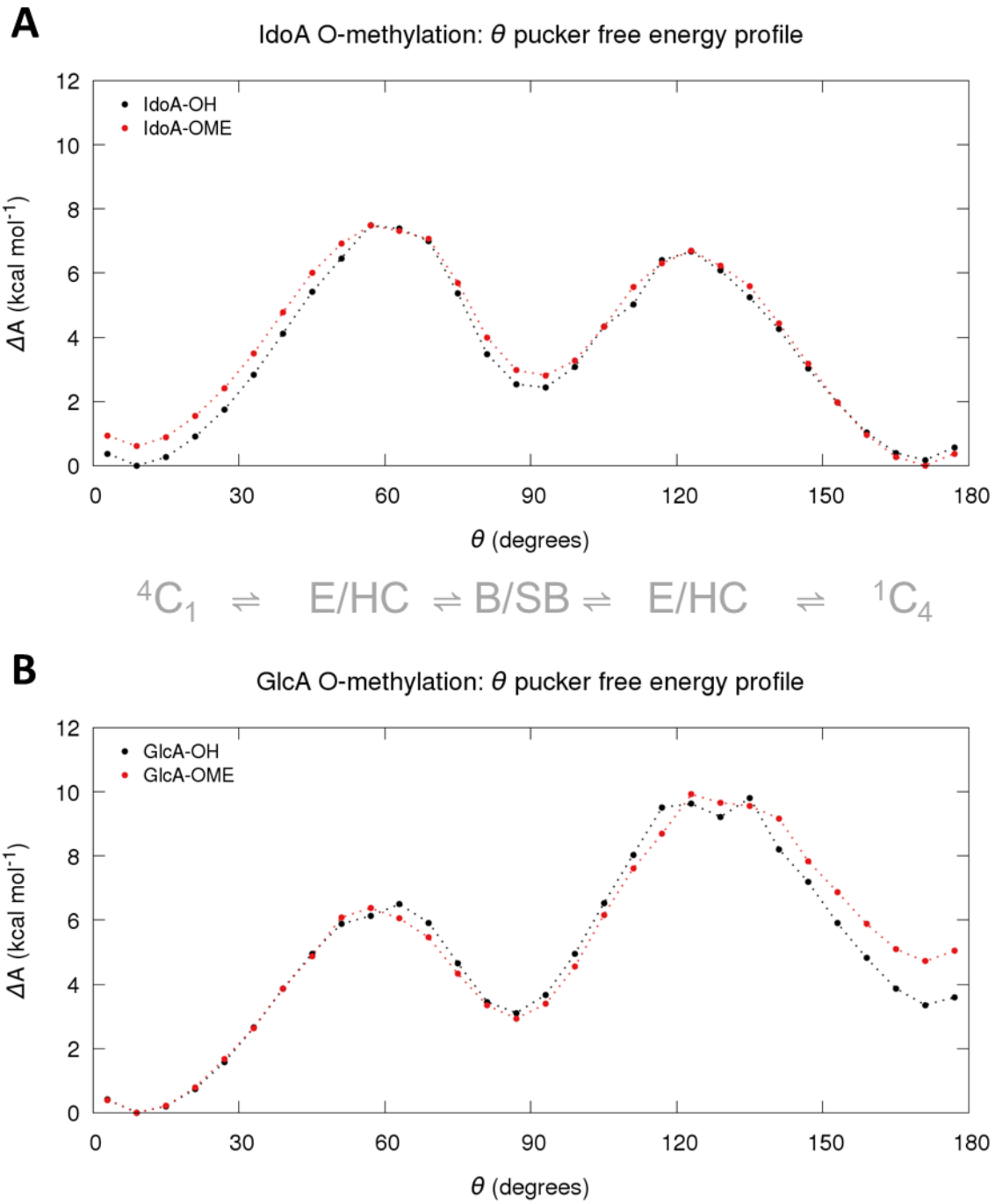

Figure 4.10 Cremer-Pople θ angle relative free energy profiles calculated via msesMD evaluating the O-Methylation of α-L-iduronic acid (IdoA) and β-D-Glucuronic acid (GlcA)

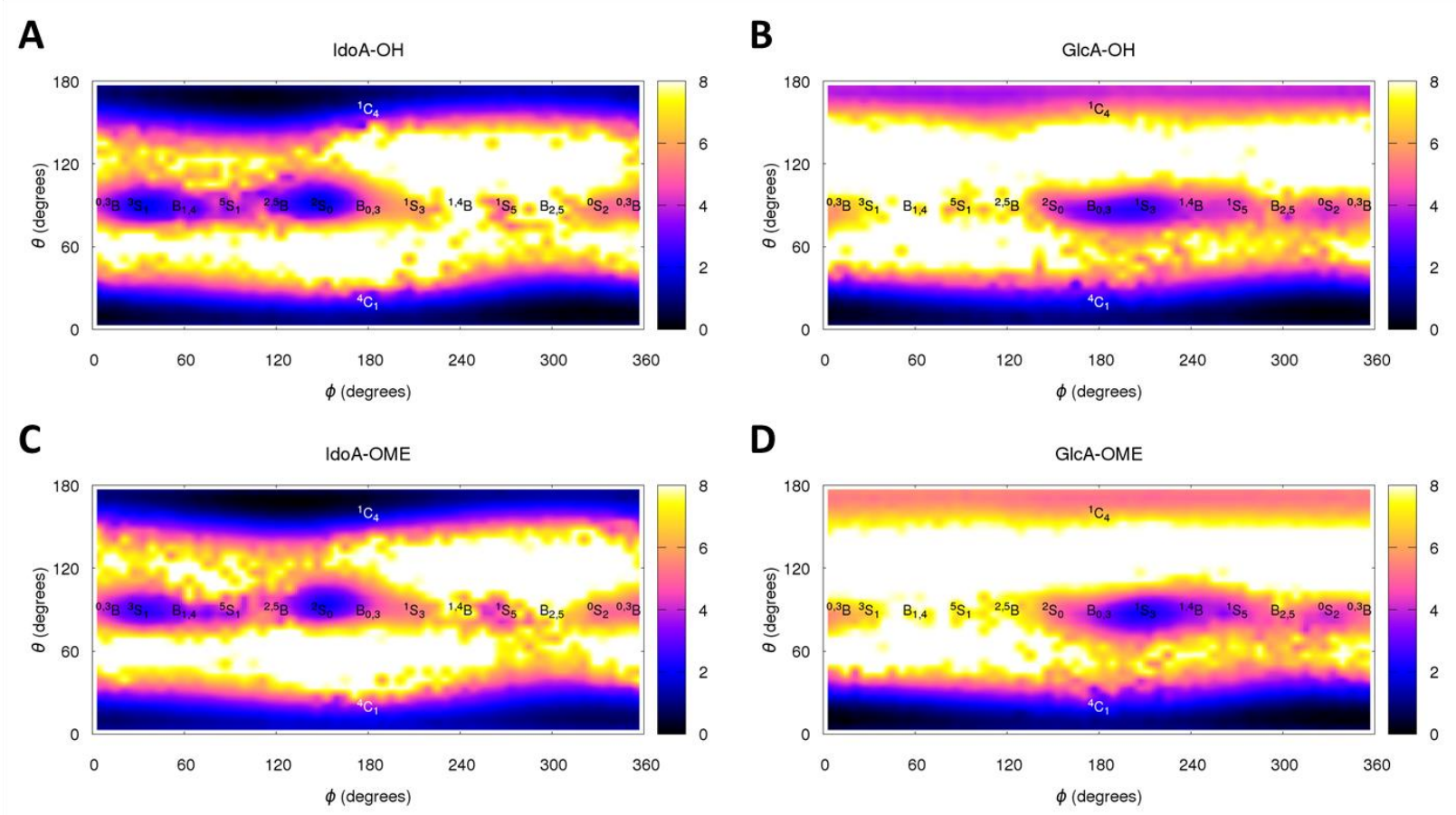

Figure 4.11 Cremer-Pople θ vs ϕ puckering free energy profiles calculated via msesMD comparing the impact of O-Methylation for α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA)

For β-D-glucuronic acid, O-methylation has a significant impact on the chair distribution by destabilising the $^{1}C_{4}$ state, resulting in a free energy state of 4.7 kcal mol^{-1} relative to the $^{4}C_{1}$ chair (Figure 4.10**B**). The reason for this is likely due to a loss of hydrogen bonding coupled with diaxial steric interactions between the axial carboxylate and O-methyl group in the $^{1}C_{4}$ state. This estimated difference in the chair stabilities agrees closely with the above mentioned Hamiltonian replica exchange study, which predicts a value of 4.5 kcal mol^{-1}.[105] For the multi-microsecond MD study[81], the $^{1}C_{4}$ chair was not found during the 2 x 5μs simulations. This is understandable considering that our own 15 μs MD simulations did not sufficiently sample 1C4 conformers in the more stable hydroxylated monosaccharide. Given its high free energy, one would expect that the $^{1}C_{4}$ chair of the O-methylated β-D-glucuronic acid is not accessible at room temperature. As for iduronic acid, the distribution of boat/skew-boat conformers is negligibly affected the O-methylation of the monosaccharide, with the only noticeable change being a reduction in stability of the $^{2}S_{0}$ conformer, $\theta = 90°$, $\phi = 160°$ (Figure 4.11**B**,**D**). The $^{1}S_{3}$ at $\theta = 90°$, $\phi = 200°$ state remains the dominant non-chair pucker which agrees with the unbiased MD results of Sattelle et al.[81]

4.3.2.3 Evaluating msesMD simulation convergence

As part of evaluating the effectiveness of the msesMD protocol in probing ring dynamics, it is important to look at the magnitude of the uncertainties in predicted free energies. One of the approach is to evaluate how well the simulation results converge over time. To do this, each msesMD simulation for the four benchmark sugar systems was extended to a total of 200 ns per replica. The estimated free energy profile for the θ coordinate was then plotted at various simulation lengths, allowing for a simple evaluation of the impact of extending the simulation length on the free energy profile (Figure 4.12).

As seen in Figure 4.12, a simulation length of 45 ns per replica is usually sufficient to adequately describe the puckering free energy profile. However, we can see that for the 45 ns simulations, the profiles can be noisy, particularly for poorly sampled conformers such as the envelope and half-chair puckers (e.g. Figures 4.12**B**-**D**). Furthermore, when comparing against longer simulation times, the 45 ns surface underestimates the stability of

the glucuronic acid $^{1}C_{4}$ pucker by around 0.5 kcal mol^{-1} relative to the 4C1 (Figure 4.12**D**). In this case convergence appears to be only achieved at simulation times of 145 ns per replica. The reason for this appears to be due to a low rate of replica transitions to and from the $^{1}C_{4}$ pucker over time (Figure 4.9**D**), thus leading to a slow convergence in the swarm behaviour. This reflects the very large energetic barrier of around 6 kcal mol^{-1} for $^{1}C_{4}$ to boat/skew-boat conformers. Contrastingly, as seen by their θ time series (Figures 4.9**A**-**C**), the other monosaccharide systems exhibit a much higher rate of exchange between conformational states. Taking this into account, although shorter simulation lengths could potentially be used, the choice of 145 ns per replica appears to be sufficient to ensure converged results.

In order to gain further insights into sampling errors associated with the 145 ns msesMD simulations, a bootstrap sampling analysis of the θ free energy profiles is carried out for all systems (Supplementary Figures C.4 to C.8). Overall it can be seen that the error is usually very low, with values within ± 0.2 kcal mol^{-1}. Although, it is noted that in a few cases, such as glucuronic acid, the error can be much higher, up to ±0.7 kcal mol^{-1} for poorly sampled ring conformations such as half-chair and envelope puckers. However, as these are very high energy conformers, the increased error would likely have no impact on analyses of ring behaviour. Overall, we can be confident that the free energy profiles recovered from the msesMD simulations are well converged.

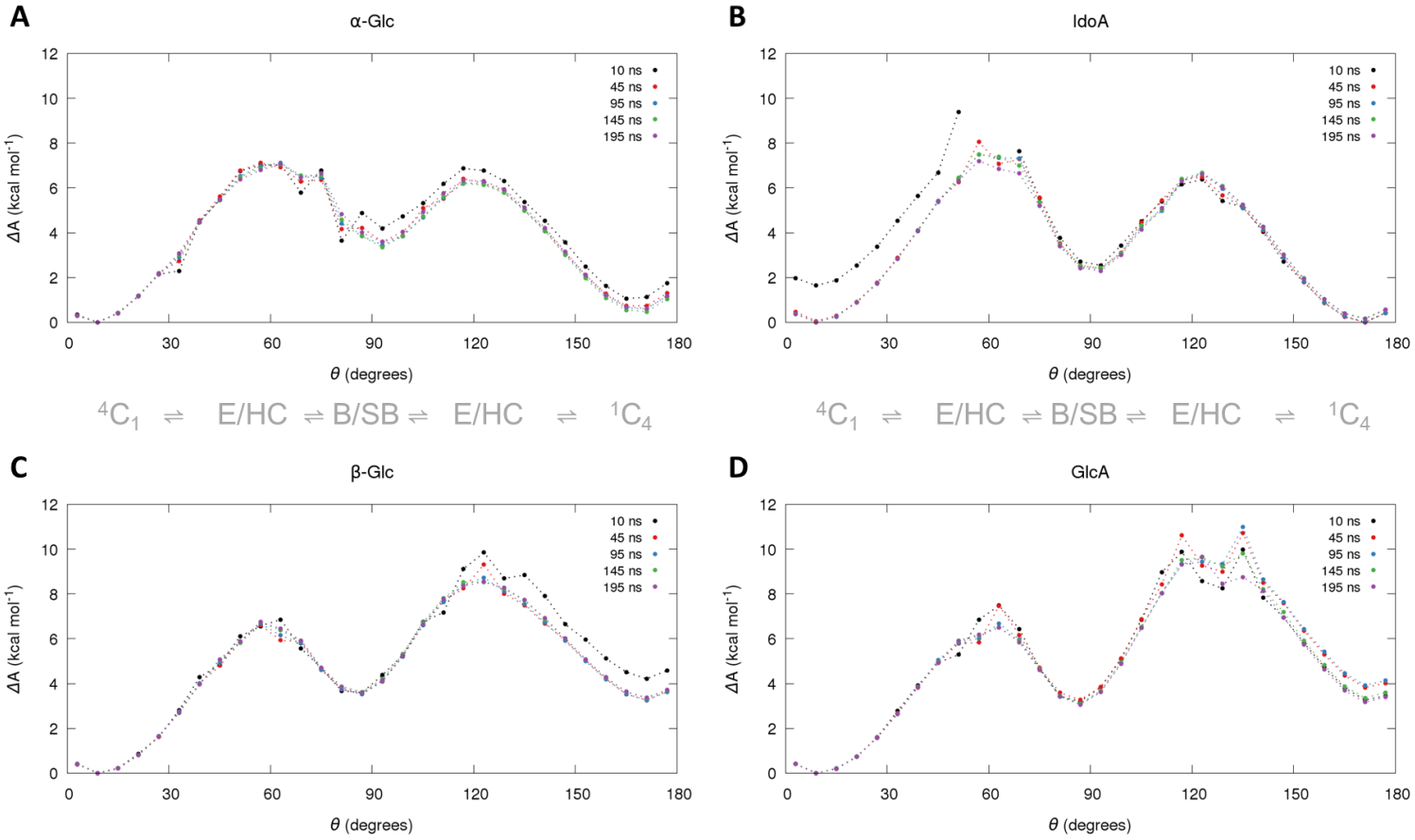

Figure 4.12 Evaluation of the msesMD convergence of the Cremer-Pople θ free energy profile over time for all four benchmark systems; α-D-glucose (α-Glc), β-D-glucose (β-Glc), α-L-iduronic acid (IdoA) and β-D-glucuronic acid (GlcA)

4.3.3 Conclusion

In this section, we have demonstrated that the msesMD method can be used to rapidly and accurately probe the ring dynamics of four benchmark monosaccharide systems. Our results show that using the msesMD method, puckering details which would usually require weeks of unbiased MD simulations, can be recovered in the time frame of a few days. Additionally, the results also highlight the fact that unbiased MD simulations require much longer simulation times than previously thought[31, 81] to ensure truly converged results. This shows that only relying on conventional MD methods in order to describe puckering is unfeasible, for some sugars, using current hardware technology. It is noted that evidence of potential issues relating to the use of a sub-optimal set of boost coordinates and relatively strong swarm potential are seen in the calculated puckering profile of α-D-glucose. Considering that this was only seen to affect a single high energy state in one of the four benchmark systems, this issue is relatively minor in impact on the computed free energy profiles. Nevertheless, it is encouraging to see that lowering the swarm potential can reduce this error whilst still sampling changes in ring conformation, albeit at a slower rate than when using the stronger potential. This means that when applying this method to the puckering of larger systems, as per our investigation of the Lewis oligosaccharides (Chapter 3) one could use a smaller boost potential to reduce reweighting noise. However this would come at the cost of slower sampling and therefore would require longer simulation times to ensure convergence. Ultimately, the msesMD protocol used in this benchmark study remains sufficiently adequate to be used agnostically to investigate the ring puckering of monosaccharides of interest.

4.4 Investigating the sulfation patterns of glycosaminoglycan monomers

4.4.1 Introduction

Having validated the use of the msesMD method in exploring the conformational behaviour of selected pyranose rings, we now turn our attention to elucidating the puckering behaviour of glycosaminoglycans (GAG). The GAGs are a family of endogenous complex linear polysaccharides which are ubiquitous in mammalian cells. They are involved in a variety of biological functions such as regulating cell growth[119], coagulation[120-122] and ensuring the structural integrity of tissues.[123-124] The most notable of the GAGs is heparin which has been used as a therapeutic anticoagulant for several decades.[125-126] Structurally, the GAG family is characterised by a repeating [hexosamine - uronic acid (galactose in keratan)] disaccharide unit which can undergo a variety of post-translational modifications such as epimerisation, sulfation, and deacetylation. In terms of the monosaccharide constituents, the hexosamine can either be glucosamine or galactosamine, whilst the uronic acid can either be glucuronic or iduronic acid. The nature of the monomer composition, stereochemistry, glycosidic linkages and post-translational modifications are the major characteristics that differ amongst members of the GAG family. It is this variability, existing even within the same GAG family member, e.g. heparan sulfate chains exhibiting varying levels of sulfation, which imparts selectivity in the function of the GAGs.[127-128] An example of this can be seen in the regulation of the angiogenic activity of FGF2 and $VEGF_{165}$ by the O6 sulfation of heparan sulfate as demonstrated by Ferreras et al.[129]

The focus of this study is to look at the impact of post-translational modification on ring puckering of GAG monosaccharides. Puckering has a significant role in the behaviour of GAGs, controlling both biological activity[100] and macroscopic chain behaviour[82]. The presence of chain decoration has been shown to alter ring flexibility[31, 82, 100], although the extent to which this is achieved by specific substitutions is unclear. While the presence of neighbouring sugar units is known to influence ring dynamics in GAG chains[100], accurately describing how on-ring substitutions affect free monosaccharides is an

important first step to deciphering the complex behaviour of polysaccharides. This has been attempted in previous MD studies of simulation lengths varying from 20 ns to 20 μs[31, 107]. However, in all cases, the accuracy of the results was restricted due to poor sampling. This is understandable considering that our benchmark results (Chapter 4.3) indicate that several tens of microseconds are likely required to accurately describe puckering profiles. Through the use of the msesMD method, we hope to overcome such limitations and provide a complete characterisation of the puckering profiles of these monosaccharides.

Table 4.1 Composition of the glycosaminoglycans

Glycosaminoglycan	*Disaccharide Unit*	*Decoration Sites*
Hyaluronic Acid	β-D-GlcA(1→3)β-D-GlcNAc(1→4)	None
Keratan Sulfate	β-D-Gal(1→4)β-D-GlcNAc(1→3)	Sulfation: O6 of Gal and GlcNAc
Heparan Sulfate / Heparin	β-D-GlcA(1→4)α-D-GlcNAc(1→4)	Sulfation: O2 of GlcA and IdoA; N2, O3 and O6 of GlcNAc
	α-L-IdoA(1→4)α-D-GlcNAc(1→4)	Deacetylation: N2 of GlcNAc
Chondroitin Sulfate	β-D-GlcA(1→3)β-D-GalNAc(1→4)	Sulfation: O2 and O3 of GlcA; O4 and O6 of GalNAc
Dermatan Sulfate	β-D-GlcA(1→3)β-D-GalNAc(1→4)	Sulfation: O2 of IdoA; O4 and O6 of GalNAc
	α-L-IdoA(1→3)β-D-GalNAc(1→4)	

Four of the most commonly occurring GAG monosaccharides are investigated here; α-L-iduronic acid (IdoA), β-D-glucuronic acid (GlcA), α-D-Glucosamine (GlcNH), and β-D-N-Acetylgalactosamine (GalNAc) (Figure 4.3). As shown in Table 4.1, these four monosaccharides are the key constituents of three out of the five major GAG families. The reason for choosing these four monosaccharides is two-fold. Firstly, the three aforementioned GAG families; heparin/heparan sulfate, dermatan sulfate, and chondroitin sulfate, have been specifically linked to a large variety of medicinally relevant functions.[125, 130-132] For example, all three have been shown to be implicated in tumour growth due to their natural roles as promoters of cell proliferation and angiogenesis[130, 133-134]. Thus, improving our understanding their conformational behaviour will help develop

advanced structure-activity relationships for their endogenous targets (e.g. pro-oncogenic proteins such as FGF2 and CXCL12) and may allow us to find novel inhibitors. The second reason for choosing the monosaccharides of these three GAGs is that they exhibit a more varied range of post-translational decoration, and resulting interactions. It is therefore expected that these modifications would play a larger role in dictating their conformational behaviour. Ultimately it is hoped that, at a later date, this investigation will be extended to all monosaccharide constituents of the GAGs.

4.4.2 Results and discussion

4.4.2.1 Glucuronic Acid

Post-translational modification of glucuronic acid in GAGs usually occurs as sulfation at the O2 position in GAGs, which we denote as GlcA(2S) (Figure 4.3). It is noted that there is also potential for sulfation at the O3 position in chondroitin sulfate, leading to the formation of so-called "oversulfated chondroitin sulfate"[135]. However, due to the rarity of this sulfation pattern, it was not investigated here.

We begin by computing the free energy profile as a function of the Cremer-Pople puckering angle θ for both the unsulfated and sulfated GlcA (Figure 4.13**A**). As seen in the earlier investigation of glucuronic acid puckering (Chapter 4.3.2.2), the puckering profile of the unsulfated GlcA details the global energy minimum to be the $^4C_1$ conformer, with the $^1C_4$ and boat/skew-boat conformers occupying energy wells that are around 3 kcal mol^{-1} less stable than the $^4C_1$. The energetic barriers between these minima is 6.5 and 9.6 kcal mol^{-1} for the $^4C_1$ → boat/skew-boat and $^1C_4$ → boat/skew-boat transitions respectively. The O2-sulfated monomer, GlcA(2S), traces a near identical profile, indicating that sulfation has little impact on the ring dynamics. The only difference seen is a decrease of around 0.5 kcal mol^{-1} in the barrier between the boat/skew-boat and $^1C_4$ minima (θ = 100-150°). However, this deviation is mostly accounted for by the relatively large estimate error for these states (Supplementary Figures C.4 and C.5). Interestingly, one would usually expect that the substitution of the O2 hydroxyl with a bulkier O-sulfate to lead to a decrease of the $^1C_4$ population, where the O2 group is in the axial orientation. This lack of change in the $^1C_4$ stability appears to be due to the presence of hydrogen bonding between the C4 hydroxyl and the O-sulfate group (Figure 4.15). The θϕ free energy profiles (Figure 4.14**C**,**D**) confirms the similarity between the two systems, with both GlcA and GlcA(2S) occupying a near equivalent distribution of puckers.

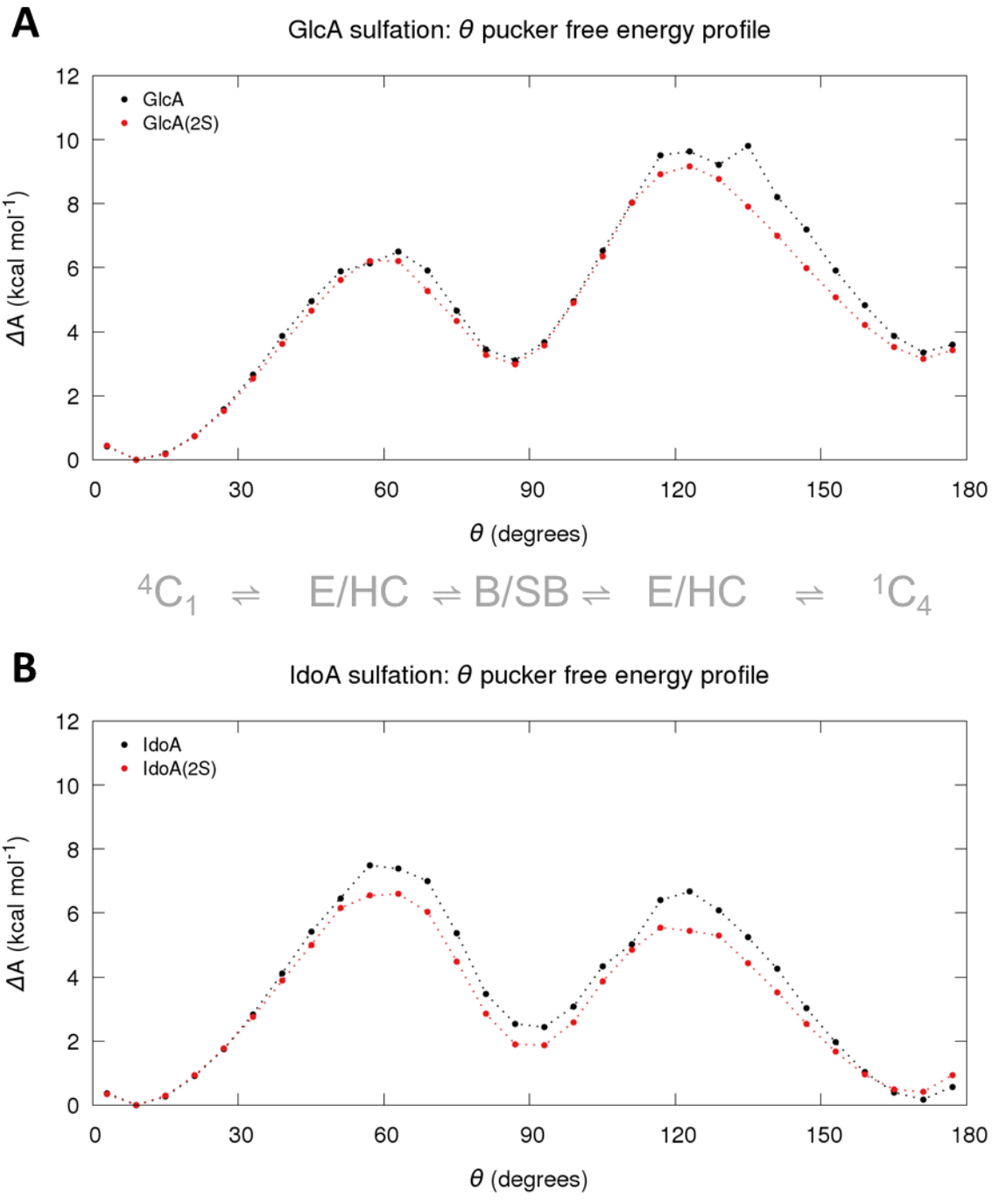

Figure 4.13 Cremer-Pople θ angle free energy profiles evaluating the impact of 2-O-sulfation in iduronic acid (IdoA) and glucuronic acid (GlcA)

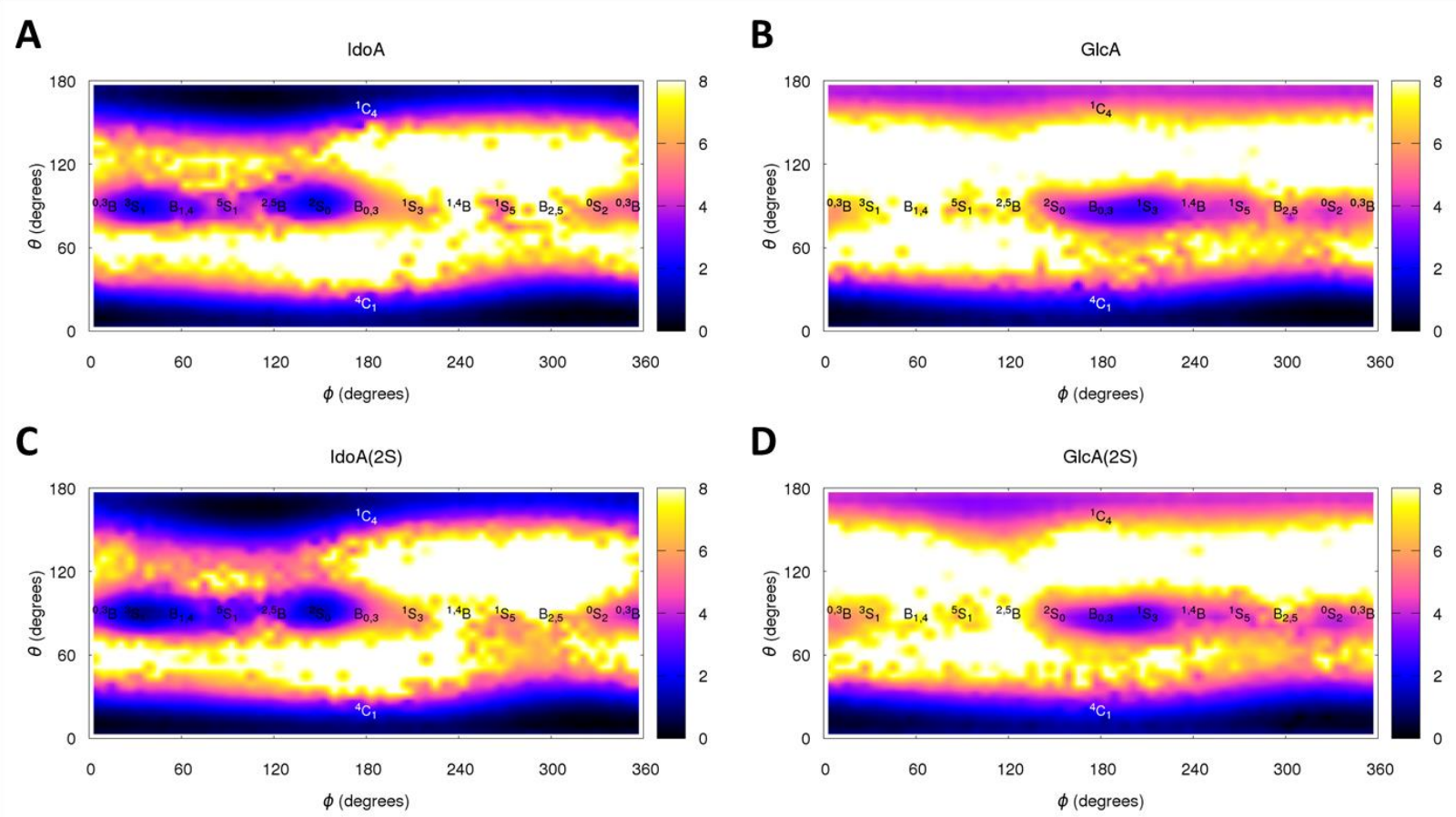

Figure 4.14 Cremer-Pople θ vs ϕ free energy profiles evaluating the impact of 2-O-sulfation in iduronic acid (IdoA) and glucuronic acid (GlcA)

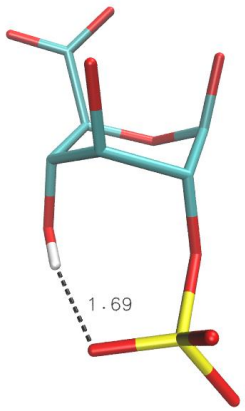

Figure 4.15 Intra-molecular hydrogen bond formation (with representative distance in Å) in the $^{1}C_{4}$ conformer of GlcA(2S)

4.4.2.2 Iduronic Acid

The presence of iduronic acid (IdoA) in GAG polysaccharides is a result of the post-translational epimerisation of GlcA via the action of uronosyl 5-epimerase[136-138]. Similarly to its precursor, only one type of ring substitution occurs for IdoA, which is sulfation at the O2 position, denoted here as IdoA(2S) (Figure 4.3).

As detailed in the benchmark simulations of IdoA (Chapter 4.3.2.2), the Cremer-Pople θ free energy profile predicted by 145 ns msesMD simulations details equi-energetic $^{4}C_{1}$ and $^{1}C_{4}$ chair conformers which are around 2.5 kcal mol^{-1} more stable than the boat/skew-boat puckers (Figure 4.13**B**). The energetic barrier to access this boat/skew-boat conformers is 0.9 kcal mol^{-1} higher when starting from the $^{4}C_{1}$, with barriers of 7.5 and 6.6 kcal mol^{-1} for the $^{4}C_{1}$ → boat/skew-boat and $^{1}C_{4}$ → boat/skew-boat transitions respectively. Unlike GlcA, the introduction of sulfation alters the pucker distribution, resulting in a 1.0 and 0.5 kcal mol^{-1} increase in stability for the envelope/half-chair and boat/skew-boat puckers respectively (Figure 4.13**B**). The presence of the sulfate does not affect the range of puckers accessed; instead it increases the stability of the two main skew-boats, $^{3}S_{1}$ and $^{2}S_{0}$ (θ = 90°, ϕ = 30° and 150° respectively), in addition to the E_{3}, ^{5}E and E_{5} (θ = 120°, ϕ = 30°-150°) envelope puckers (Figure 4.14**A**,**C**). As previously reported, this increase in non-

chair stability appears to be due to the formation of intramolecular hydrogen bonding between ring hydroxyls and the O-sulfate group.[100] Analysis of the msesMD trajectories demonstrates hydrogen bonding occurs with the C1 and C3 hydroxyls when in the $^{3}S_{1}$ and $^{2}S_{0}$ states respectively (Figure 4.16).

This increased preference for the skew-boat puckers upon sulfation concurs with previous analyses of the IdoA ring dynamics in heparan sulfate chains, where the $^{2}S_{0}$ was predominantly occupied when sulfated and the $^{1}C_{4}$ chair when not[100, 139]. The major driving forces for this effect have previously been attributed to electrostatic interactions with sulfates on neighbouring glucosamines. It is therefore interesting that a similar trend can also be seen in the msesMD simulations of the free monosaccharide. This indicates that the formation of intra-ring hydrogen bonds may play an important role in defining polysaccharide dynamics.

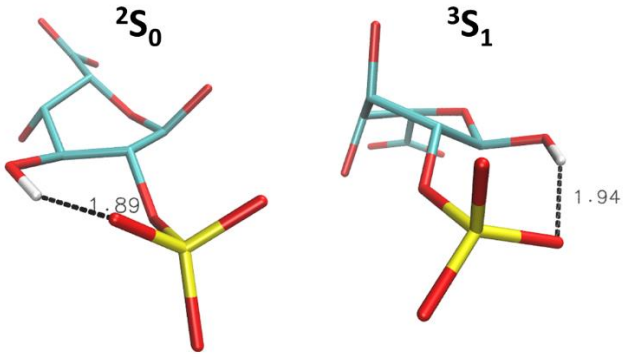

Figure 4.16 Intro-molecular hydrogen bond formation (with representative distance in Å) in the $^{2}S_{0}$ and $^{3}S_{1}$ conformers of IdoA(2S)

4.4.2.3 N-Acetyl Galactosamine

The N-acetyl galactosamine (GalNAc) hexosamine can be modified by sulfation at either the O4 (GalNAc(4S)), O6 (GalNAc(6S)) or both (GalNAc(4S,6S)) positions (Figure 4.3). In accordance with previous findings[101], the Cremer-Pople θ profile calculated via 145 ns msesMD details the GalNAc ring as rigid, occupying primarily the $^{4}C_{1}$ chair, with a high energy boat/skew-boat minimum of 5.4 kcal mol^{-1} (Figures 4.17**A** and 4.18**A**). Only one of

the replicas in the msesMD simulation sampled the $^{1}C_{4}$ chair (Supplementary Figure C.10), thus resulting in the 11.2 kcal mol^{-1} estimate seen in the free energy profile (Figure 4.17**A**). This large difference in stability between the chair conformers is due to the high steric strain involved in moving four of the ring substituents from equatorial to axial positions when transitioning from the $^{4}C_{1}$ to $^{1}C_{4}$ chair. Three of these substituents, the C1-hydroxyl, C3-hydroxyl and hydroxymethyl groups, lead to unfavourable 1,3-diaxial interactions when axial. The introduction of sulfation at the O6 position only reinforces this effect, resulting in little to no impact on the ring dynamics, as seen from the free energy profiles (Figure 4.17**A**).

Conversely, O4 sulfation significantly stabilises the $^{1}C_{4}$ pucker to ~ 5 kcal mol^{-1}. This in part due to the bulky O4 sulfate group occupying a strained axial position in the $^{4}C_{1}$ chair, thus shifting the conformational equilibrium. The $^{1}C_{4}$ is also further stabilised by the formation of intra-ring hydrogen bonding between both the ring hydroxyls and both the sulfate and hydroxymethyl groups (Figure 4.19). The boat/skew-boat conformations also because marginally (~ 0.6 kcal mol^{-1}) more stable, with some additional sampling of high energy conformers in the $^{0,3}B$ and $^{3}S_{1}$ puckers (Figure 4.18).

When sulfation occurs at both positions, the presence of the O6 sulfate slightly counters the increased flexibility imparted from the O4 sulfate resulting in a ~ 1.5 kcal mol^{-1} decrease in the $^{1}C_{4}$ stability. The fact that the O4 sulfation appears to have a greater influence on the ring dynamics may be because the O6 sulfate can adopt a hydrogen bond stabilised *gauche-trans* rotamer when in the $^{1}C_{4}$ chair, thus reducing the steric impact of having to adopt an axial position (Figure 4.19). It is important to note that for both the GalNAc(4S) and GalNAc(4S,6S), the rate of pucker transition in the msesMD simulations was very low, in fact the number of replicas observing the $^{1}C_{4}$ conformer were 3 and 2 respectively (Supplementary Figure C.10). Thus, although the bootstrap errors were relatively low (Supplementary Figure C.6), there is a possibility that the swarm has not converged and therefore as shown in the glucuronic acid benchmark study (Section 4.3.2.2), the free energy estimates are potentially prone to change if run for a longer period of time.

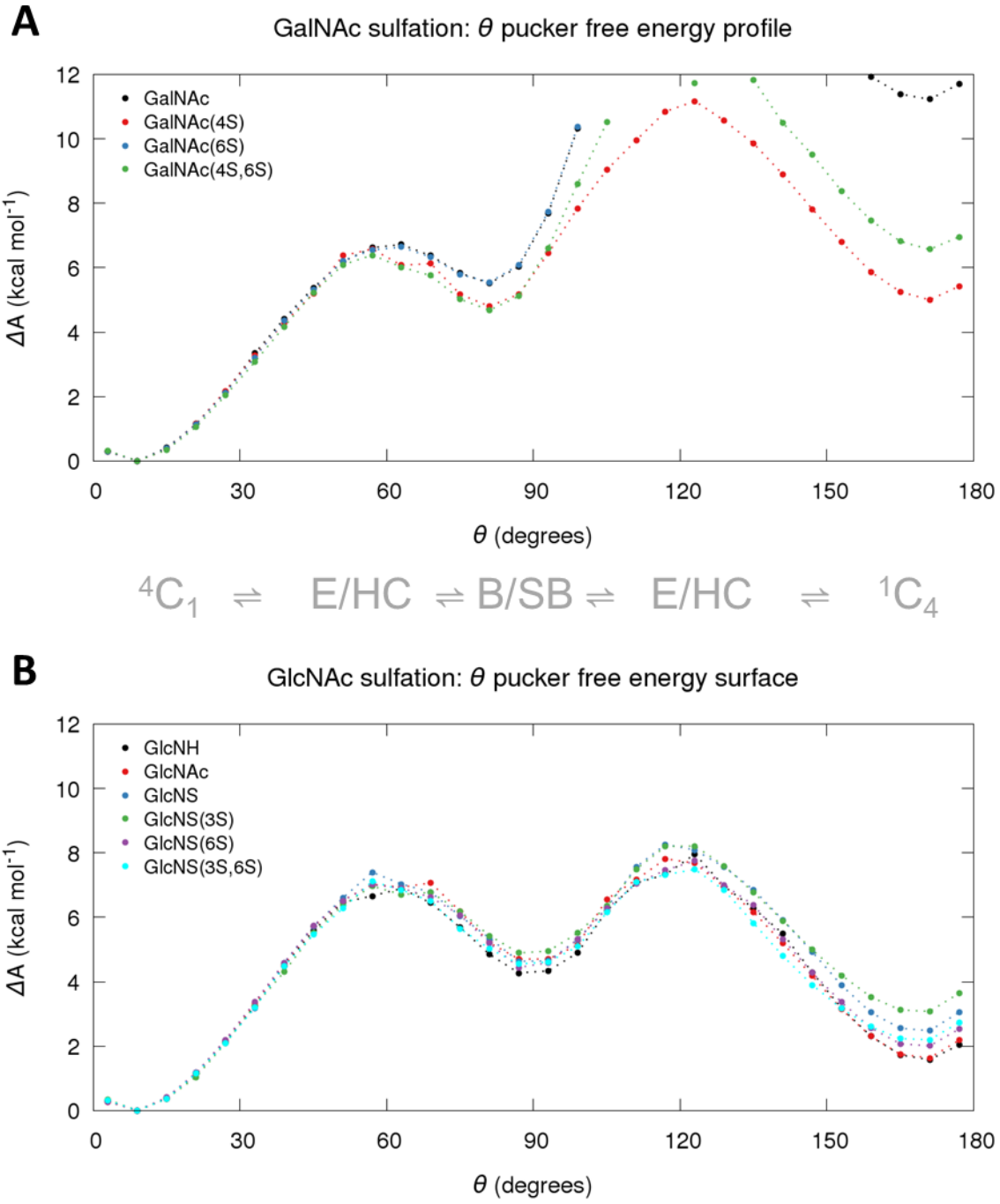

Figure 4.17 Cremer-Pople θ angle free energy profiles evaluating the impact of ring modification on N-Acetyl-galactosamine (GalNAc) and N-Acetyl-glucosamine (GlcNAc)

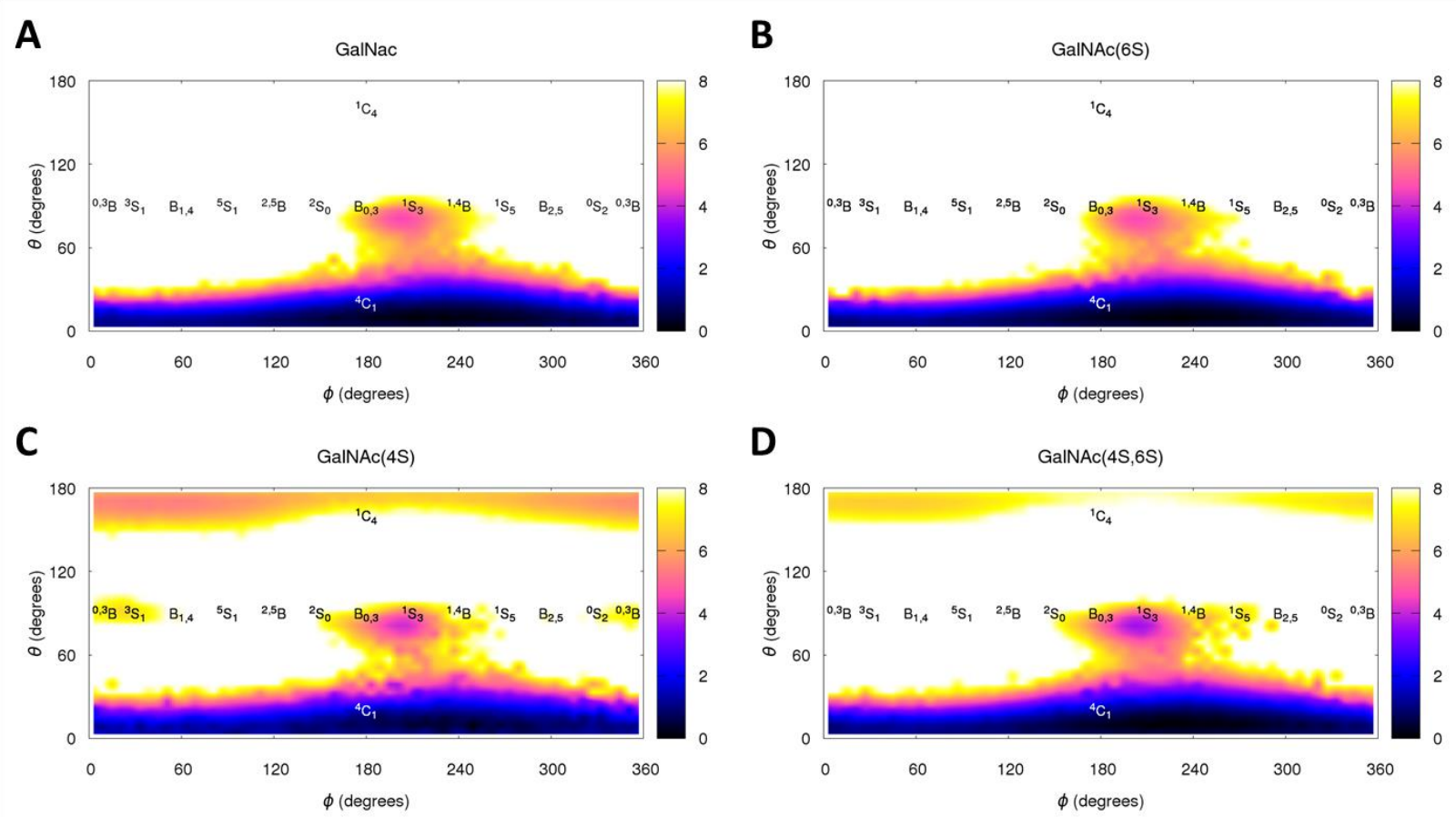

Figure 4.18 Cremer-Pople θ vs ϕ puckering free energy profiles of the different O-sulfation patterns of N-Acetyl-galactosamine (GalNAc)

These results demonstrate that the current idea that galactosamine ring is rigid may be incorrect, with selective sulfation potentially playing a part in increasing chain flexibility. This is particularly interesting when considering that dermatan sulfate, unlike the heavily sulfated chondroitin sulfate, usually only undergoes O4 sulfation.[116, 140] One may therefore predict that the inclusion of such a flexibility-inducing modification, and the presence of a more flexible uronic acid (IdoA) would yield more flexible chains relative to chondroitin sulfate. This may in part account for the functional difference between the two otherwise similar GAG families.

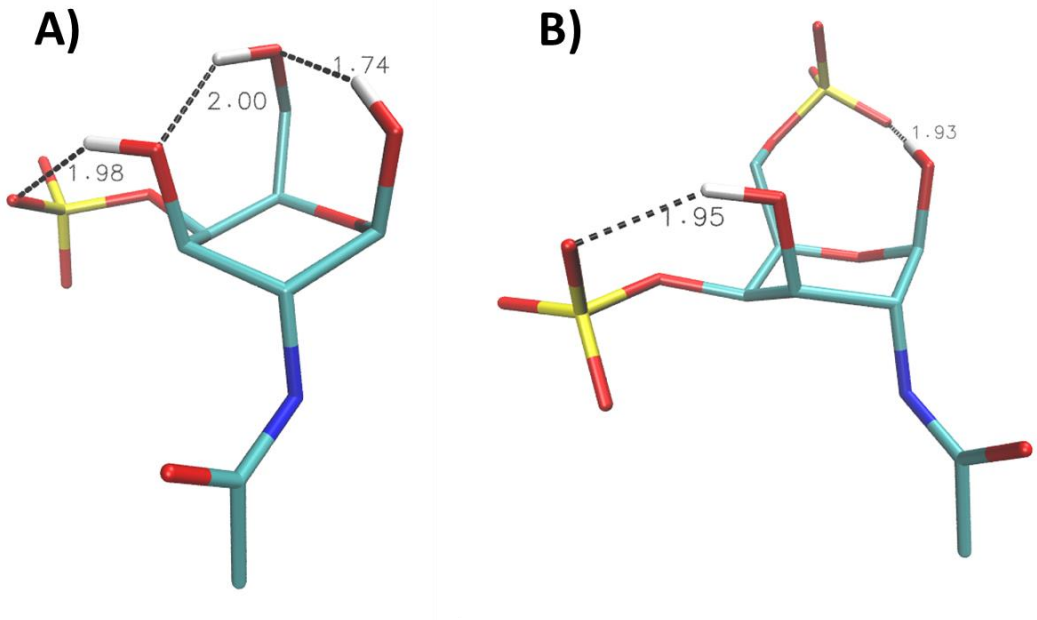

Figure 4.19 Intra-molecular hydrogen bond formation (with representative distances in Å) in the $^{1}C_{4}$ conformers of A) GalNAc(4S) and B) GalNAc(4S,6S)

4.4.2.4 Glucosamine

The N-acetyl glucosamine (GlcNAc) is the most widely modified GAG monomer, resulting in a total of six major structural states (Figure 4.3). It can first undergo N-deacetylation forming the unsubstituted glucosamine (GlcNH), and then N-sulfation to form N-sulfoglucosamine (GlcNS). The GlcNS monomer can then undergo further O-sulfation via the action of heparan sulfate sulfotransferases at either the O3 (GlcNS(3S)) or O6 (GlcNS(6S)) positions or both (GlcNS(3S,6S)). The msesMD calculated θ coordinate free energy profile predicts the $^{4}C_{1}$ chair as the most stable conformer (Figure 4.17**B**). However, the GlcNAc ring offers more flexibility than GalNAc, predicting minima for the $^{1}C_{4}$ and boat/skew-boat puckers which are 2.5 and 4.6 kcal mol^{-1} less stable than the $^{4}C_{1}$ (Figure 4.17**B**). This flexibility is also reflected in the range of ring conformations adopted, with several near equivalently stable states along the $^{2}S_{0}$ to $^{0}S_{2}$ puckers and higher energy occupation of the remaining skew-boat puckers, except from $^{2,5}B$ (Figure 4.21**A**).

The removal of the N-acetyl group (GlcNH) appears to have only a minor impact on the ring dynamics, slightly increasing the stability of the skew-boat puckers by an average of 0.5 kcal mol^{-1} (Figures 4.17**B** and 4.21**B**). However, N-sulfation (GlcNS), decreases the $^{1}C_{4}$ stability by 0.9 kcal mol^{-1} relative to the $^{4}C_{1}$ chair. This seems to be caused by the formation of an intra-ring hydrogen bond between the N-sulfate and the C3 hydroxyl when in the $^{4}C_{1}$ chair, thus increasing its stability (Figure 4.20).

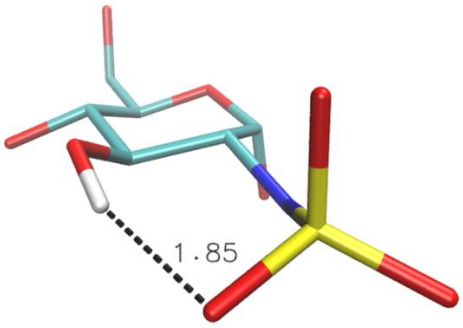

Figure 4.20 Intra-molecular hydrogen bond formation (with representative distance in Å) in the $^{1}C_{4}$ conformer of GlcNS

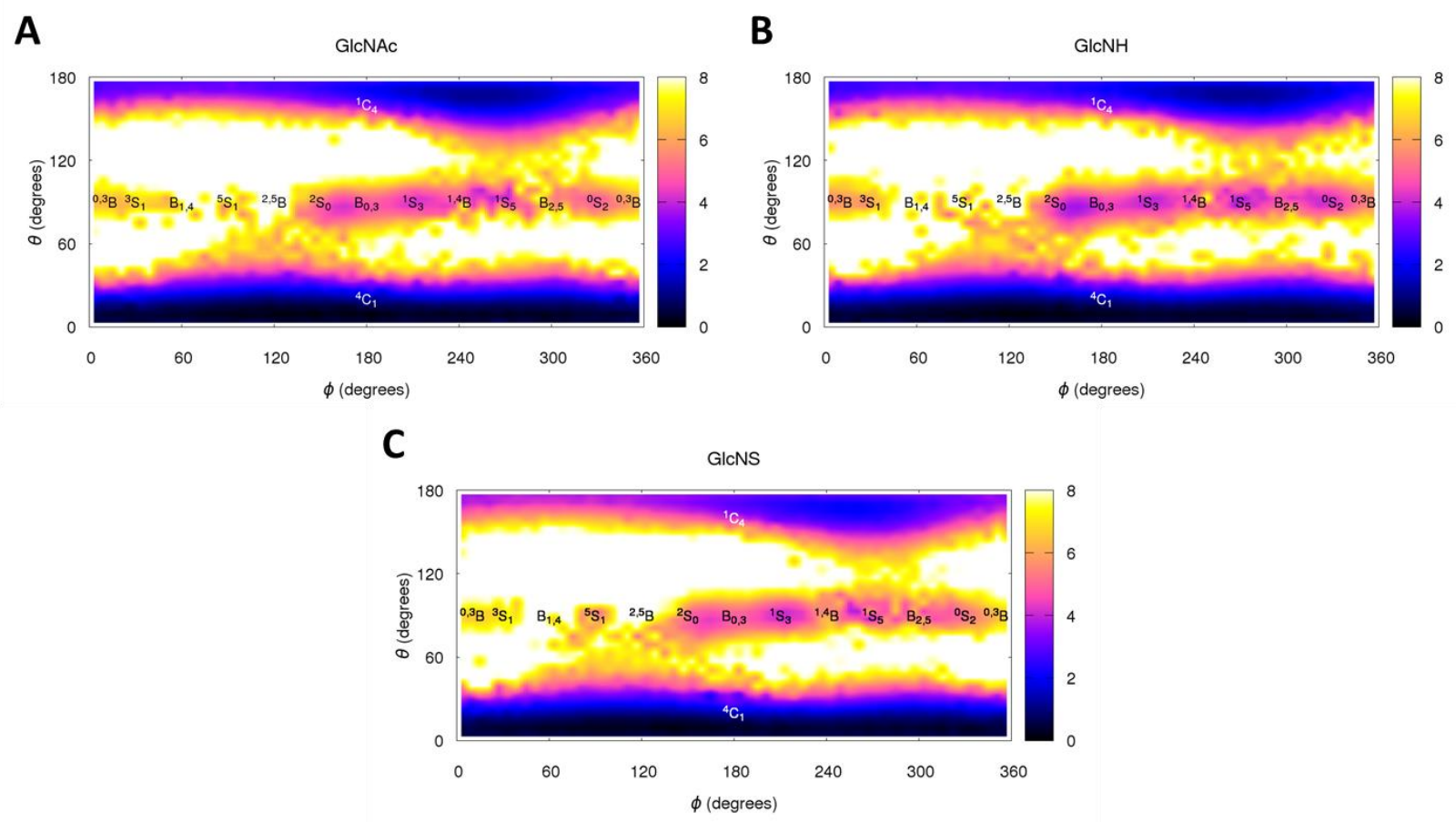

Figure 4.21 Cremer-Pople θ vs ϕ puckering free energy profiles for the different N modifications of GlcNAc

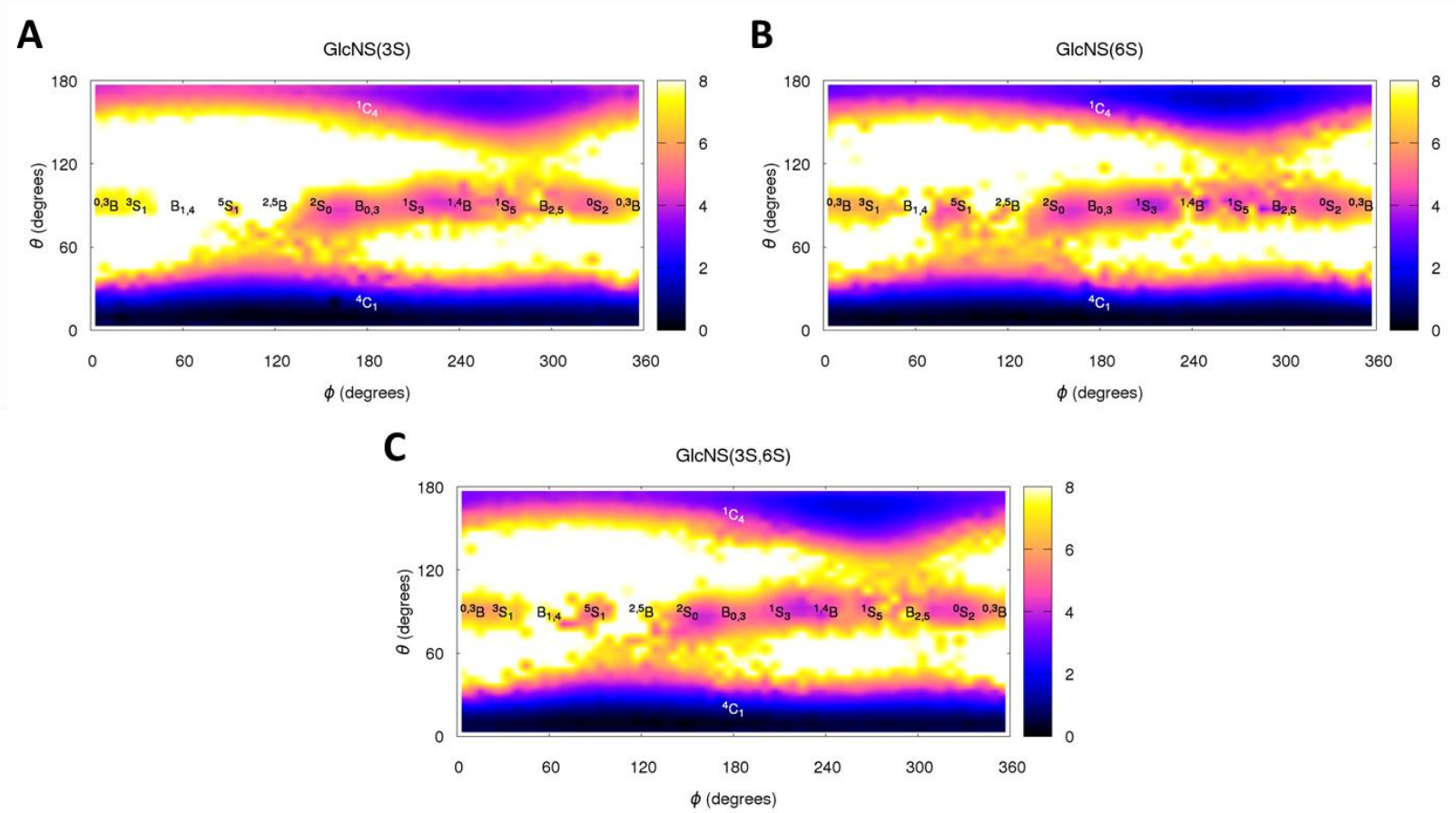

Figure 4.22 Cremer-Pople θ vs ϕ puckering free energy profiles for the different O-sulfations of GlcNAc

The addition of 3-O-sulfation (GlcNS(3S)) further destabilises the $^{1}C_{4}$ chair, by 0.5 kcal mol^{-1} relative to GlcNS. This cause of this outcome is due to both the formation of a strong hydrogen bond between the O3-sulfate and the C4-hydroxyl (Figure 4.23**A**), in addition to increased steric strain caused by the bulkier sulfate group when going to an axial orientation in the $^{1}C_{4}$ chair. It is also noted that the presence of the O3-sulfate also reduces the stability of the skew-boat puckers, particularly the higher energy states ($^{0,3}B$, $^{3}S_{1}$, $B_{1,4}$ and $^{5}S_{1}$) which are far less sampled (Figure 4.22**A**). On the other hand, O6-sulfation (GlcNS(6S)) increases $^{1}C_{4}$ stability by 0.5 kcal mol^{-1} relative to GlcNS (Figure 4.17**B**). This appears to be due to the formation of a hydrogen bond between the O6-sulfate and the C3-hydroxyl (Figure 4.23**B**) when in the $^{1}C_{4}$ conformation.

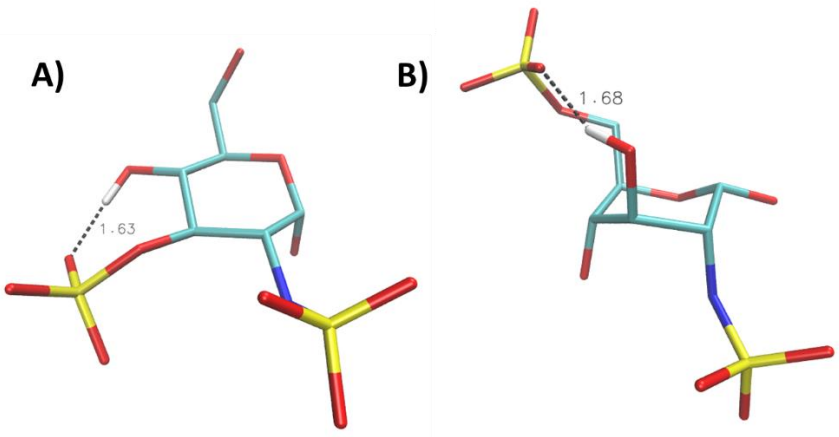

Figure 4.23 Intra-molecular hydrogen bond formation (with representative distances in Å) in A) the $^{4}C_{1}$ pucker of GlcNS(3S) and B) the $^{1}C_{4}$ pucker of GlcNS(6S)

However, as seen from the free energy profile (Figure 4.17**B**), this O6-sulfation stabilisation effect is still seen when both the O3 and O6 positions are sulfated. In fact both GlcNS(6S) and GlcNS(3S,6S) have, within error (Supplementary Figures C.7 and C.8), nearly identical puckering profiles, indicating that the O6-sulfate has a dominant contribution towards determining ring dynamics than the O3-sulfate. This is particularly interesting considering that when this occurs, the replacement of the C3-hydroxyl prevents the $^{1}C_{4}$ intra-ring hydrogen bond seen in GlcNS(6S) from existing (Figure 4.23**B**). Moreover, analysis of the O3-sulfate to C4-hydroxyl hydrogen bond (i.e. the bond displayed in Figure 4.23**A**) formation in GlcNS(3S,6S) reveals an occurrence probability of 76%, which is only a 3.8 % reduction compared to GlcNS(3S). It is therefore likely that

the source of this increased access to the $^{1}C_4$ state may be the electrostatic repulsion between the negatively charged sulfate groups. The exact mechanism as to how this is happening is unfortunately unclear, as conventional wisdom would tell us that the close proximity of the O3 and O6 sulfates as seen in the $^{1}C_4$ chair should only serve to further destabilise the system. It is possible that this behaviour is due to the limitation of describing a highly charged system using an additive fixed charged sulfation method.

Overall we can conclude that non-sulfated glucosamines exhibit higher ring flexibility relative to the sulfated variants. This concurs with previous simulation findings showing that undecorated portions of HS chains are more flexible than decorated ones.[141] Such behaviour is thought to play an important role in correctly placing the sulfated domains within their active sites. Furthermore, we can also predict that the presence of sulfation at the O6 position plays a role in increasing ring flexibility within decorated HS domains. This partially agrees with previous findings by Sattelle and Almond[31] who found the GlcNS(6S) as the only sulfated variant of GlcNAc to undergo chair interconversion on the multi-microsecond timescale. Interestingly, whilst the GlcNAc results presented here generally agree with the findings of Sattelle and Almond, deviations in terms of the breadth and stability of puckering states are seen for the remaining substitution patterns. Considering the high energetic barriers of up to 7.4 and 6.5 kcal mol^{-1} for the $^{4}C_1 \rightarrow$ boat/skew-boat and $^{1}C_4 \rightarrow$ boat/skew-boat transitions respectively (Figure 4.17**B**), it is likely that their 2 x 10 μs unbiased MD simulations were sampling limited, thus resulting in a lack of exploration of the boat/skew-boat and $^{1}C_4$ forms. Unlike the GalNAc simulations, the msesMD replicas readily crossed energetic barriers, ensuring adequate swarm convergence (Supplementary Figures C.11 and C.12). We can therefore be relatively confident in the energy estimate, within the remit of the force field model employed.

4.4.3 Conclusion

Although recognised as important, the role of flexibility in GAGs and how it is encoded by selective chain decoration has yet to be elucidated. *In vitro* work has generally been challenged by the tasks of achieving site specific modifications of GAG polysaccharides[142-143] and capturing meaningful data from such flexible systems. *In silico* work has been hampered by the need for long sampling timescales and a difficulty in properly representing the highly charged nature of these systems.[31, 82, 144] Through the use of the msesMD method, we overcome limitations in sampling. Thus, the results presented here are, to date, the most accurate theoretical depictions of ring flexibility for the GAG monosaccharides, as described by the Glycam06 force field. This information provides a good starting point to both improve our understanding of larger chain effects and further refine currently available models of GAG polysaccharides.

Beyond providing a rationale for polysaccharide behaviour, our study has outline two specific points which merit further attention. The first is the role of O4 sulfation in increasing the flexibility of N-acetyl galactosamine. As far as we aware, this is the first time this effect has been reported. Not only does this challenge previous assumptions of galactosamines as rigid ring systems, but it also may potentially offer some explanation for the impact of 4-O-sulfation on the activity, such as the findings by Pavao et al. who demonstrated that presence of the 4-O-sulfate is necessary for the anticoagulant effect of dermatan sulfate.[145] The second outcome of particular interest is the influence of 6-O-sulfation on increasing the flexibility of glucosamine. Whilst the importance of a sulfate at the O6 position has been widely reported experimentally[100, 146-147], the fact that it appears to increase flexibility within the glucosamine ring may influence polysaccharide conformation. Though this effect is likely to be subtle it may provide some insights into the relative flexibility of different decorations of the sulfated HS domains.

Finally we also note that our findings have also outlined potential limitations of the Glycam06 force field, particularly with regards to the use of a transferable sulfation model.[111] The seemingly unnatural behaviour of GlcNS(3S,6S), demonstrates that the model is prone to error as the system charge increases. Further work using models with

more complex representations of charge such as via semi-empirical QM/MM methods will likely be required to evaluate the nature and extent of any such errors. This issue also highlights the need to ensure that full evaluations of ring puckering should ideally be used as part of the development and validation of carbohydrate force fields.

Chapter 5: Calculating solvation free energies using msesMD

5.1 Introduction

Calculating the free energy contributions to a system is a particularly useful task, allowing one to understand the favourability of certain changes in a system of interest. For example, in the context of pharmaceutical research, using such approaches could allow for a quantitative look into the propensity of different ligands to bind to a particular protein target. This could then be used to both identify and refine leads prior to synthesis. Other uses include the estimation of key physiochemical properties of ligands such as calculating the solvation free energy to obtain solubility.[148] Several molecular dynamics based methods (e.g. free energy perturbation[149] and thermodynamic integration[150]) have been developed in order to measure free energy changes along alchemical paths of interest. Whilst the use of such methods has become increasingly popular[151], particularly due to significant improvements in simulation costs, they are frequently limited in terms of their accuracy relative to experiment. Two major contributors to these inaccuracies can be identified: force field[151-152] and sampling[153-154] limitations.

In the former case, this usually occurs due to a poor description of electrostatics by fixed-charge force field models. As fixed-charge models do not allow for dynamic changes in partial point charge values, they are unable to accurately describe the impact of conformational change on the electrostatic distribution. This can sometimes lead to large inaccuracies when systems of interest adopt conformational states that are not accounted for in the parameterisation procedure. This effect is particularly evident in generalised force fields (e.g. GAFF; the AMBER generalised force field[17]), where partial charge assignment not only uses a single conformational state, but also relies on approximate methods for the electrostatics (e.g. AM1/BCC[155]). The use of polarisable force fields, such as the AMOEBA[20-22] or the CHARMM Drude[19] models, offer a potentially promising solution to this issue[156-157]. However their lack of generalisability and increased computational costs may hinder their real world applications. In any case, if one can

overcome sampling limitations and exhaustively explore free energy paths, it is possible to identify the impact of force field limitations by comparing against experimentally-derived free energies, as attempted by collaborative projects such as the SAMPL challenges[158-160], which can be particularly useful in directing future force field development efforts.

With regards to sampling limitations, these occur as a result of slow motions in the system of interest which are not captured by the usually short simulation windows employed in most free energy calculations. This is particularly evident in simulations describing complex systems such as ligands binding to proteins, whereby large scale conformational events can occur as a function of the presence of the ligand.[154] However, it can also be an issue in much smaller systems, as demonstrated in the solvation free energy calculations of carboxylic acid containing ligands, where poor sampling of the O-C-O-H torsion leads to errors in the free energy estimate.[153, 161] To address such issues, several combined enhanced sampling free energy methods have been proposed including; λ dynamics[162], Hamiltonian replica exchange schemes[163-164], accelerated molecular dynamics coupled TI[165], and λ-metadynamics[166]. Unfortunately, the use of such approaches not only increases computational costs but also the complexity of setting up and running the simulation.

In this chapter, we investigate the applicability of using the msesMD methodology to enhance sampling in alchemical free energy calculations. Having demonstrated in previous chapters that the method can be intuitively used to boost conformational transitions along large energetic barriers, one predicts that it could be readily incorporated as part of a free energy protocol to ensure exhaustive sampling of conformational space. It is noted that the swarm-enhanced sampling method (sesMD) has previously been used within a thermodynamic integration (TI) framework, termed swarm-enhanced sampling TI (sesTI), to accurately recover the relative free energy of butane-to-butane alchemical transformations.[12] Whilst this approach was shown to be very effective relative to the unbiased independent trajectory TI (IT-TI) approach[167], it nevertheless faces several limitations preventing it from being easily applied to other systems of interest. These limitations include high compute costs and poor swarm parameter transferability.

These limitations are inherent to the sesMD boost function and its implementation within the *sander* MD engine and are for the most part discussed in Chapter 2.3. However, their complexity is increased in the context of free energy simulations. In terms of high compute costs, as previously discussed, the initial sesMD implementation has a poor compute performance due to high swarm communication costs and the non-optimised nature of the *sander* MD routines. In alchemical free energy simulations, performance is further hindered due to the replicated dual-topology scheme employed in *sander*. This scheme requires two concurrent processes to simulate both the initial and final states of the system in order to recover the combined potential energy representing a certain λ state. This is exemplified in equation 5.1 where $U_{initial}$ describes the interactions involving atoms only present in the initial state, U_{final} involving interactions only present in the final state, and U_{common} describing the interactions that exist in both states. As discussed by Kaus et al.[168] such an approach has an algorithmic efficiency of only ~50% as U_{common} describes the majority of all interactions in both states; this means that excess compute resources are wasted by duplicating the same potential energy evaluations. In terms of the combined sesTI methodology, this implies that in the best case scenario (i.e. disregarding additional communication overheads) an N replica swarm would require a minimum of $2N$ cores to achieve a similar simulation speed as a 1 core per replica sesMD simulation. This unfortunately renders the simulation of any but the smallest of systems (such as a butane) prohibitively expensive, even on modern high core density compute architectures.

$$U(r,\lambda) = (1-\lambda)[U_{common}(r) + U_{initial}(r,\lambda)] + \lambda[U_{common}(r) + U_{final}(r,\lambda)] \quad [5.1]$$

Fortunately, this issue can be avoided in the current msesMD implementation, as the *pmemd* engine in which it is incorporated uses a single topology scheme instead (equation 5.2).[168]

$$U(r,\lambda) = U_{common}(r) + (1-\lambda)[U_{initial}(r,\lambda)] + \lambda[U_{final}(r,\lambda)] \quad [5.2]$$

This approach "merges" both end states into a single system where the perturbed atoms are

treated as separate λ-scaled regions. This allows for the U_{common} term to be evaluated once per simulation cycle, whilst also treating the perturbed regions ($U_{initial}$, U_{final}) independently of each other within a single simulation process (Figure 5.1). Whilst this technically means that a single process must evaluate the potential energy of more atoms per cycle (i.e. by having to account for the atoms in both end states), this small overhead is insignificant relative to the communication costs of the dual topology approach. This is demonstrated by Kaus et al. who show an algorithmic efficiency speedup of 3 to 3.5x relative to the *sander* implementation.[168] Considering that msesMD shows a 10-15% performance cost relative to unbiased *pmemd* simulations (Chapter 2.3.2), one expects a similar speedup, which would significantly reduce compute costs relative to sesTI.

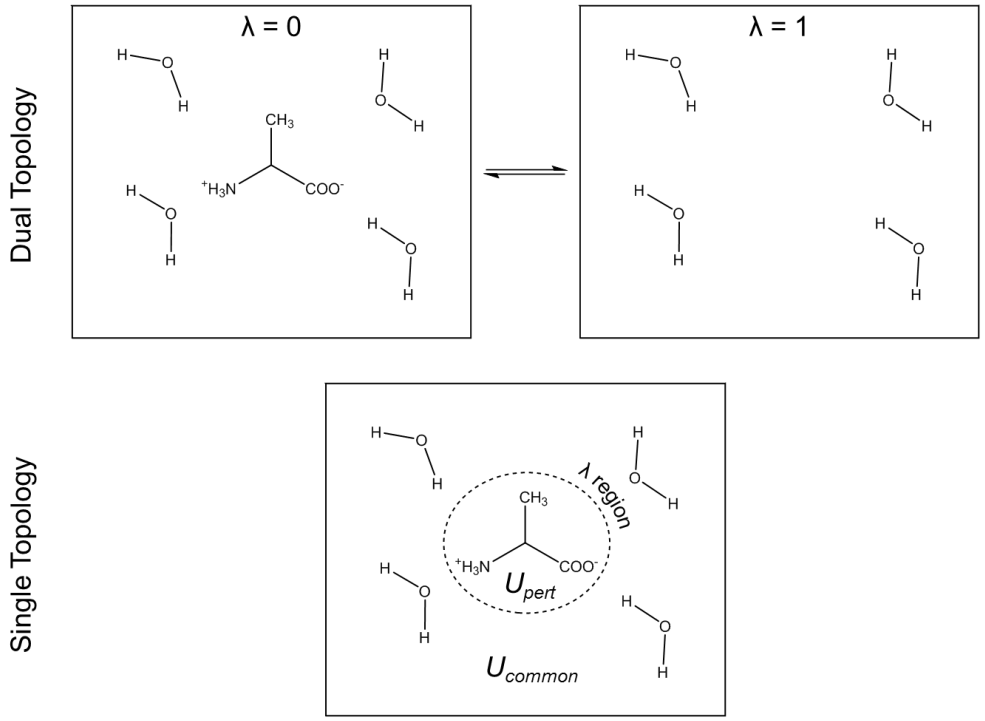

Figure 5.1 Comparison of dual and single topology schemes for solvation free energy calculations

With regard to the latter limitation, namely poor parameter transferability, this issue is directly related to the functional form of the sesMD potential. As discussed in Chapter 2.2.1, the issue of parameter transferability from one system to another mainly stems from the use of a root-mean-square dihedral distance term (d_{rms}) in sesMD (equation 2.1). This is particularly problematic in free energy calculations as the environment in which a system exists, and by extension the energy barriers to conformational change, can change

significantly from one end state to another. This can lead to cases where the effectiveness of a certain set of parameters can alter as a function of the change in λ. One could re-parameterise the swarm parameters to fit the new system of interest if required, though this would incur significant costs in terms of user time and thus would not be suitable in high throughput scenarios. As demonstrated in previous chapters, this parameterisation issue is alleviated through the use of the generalised form of the msesMD potential. Although scaling of the parameters is sometimes necessary to ensure that the swarm potential does not lead to large uncertainties in the estimates (see Chapter 3), such amendments are intuitive and can be usually avoided by opting for a softer parameter set.

As stated above, the primary aim of this chapter is to assess whether or not the msesMD protocol can be used to accurately recover alchemical free energies. To do so, we combine the msesMD method with a soft-core thermodynamic integration scheme, which in the spirit of the previous sesTI implementation, we shall refer to as msesTI. To assess the msesTI results, they are compared against unbiased independent-trajectory TI (IT-TI)[169] simulations. For this evaluation, we focus on the calculation of absolute solvation free energies for small ligand sized molecules, i.e. the free energy cost of taking a ligand out of aqueous phase and putting it into gas phase. Within the context of a soft-core simulation, this essentially amounts to the decoupling of a solute from its solvent environment over λ space, resulting in a pure box of solvent (Figure 5.2).

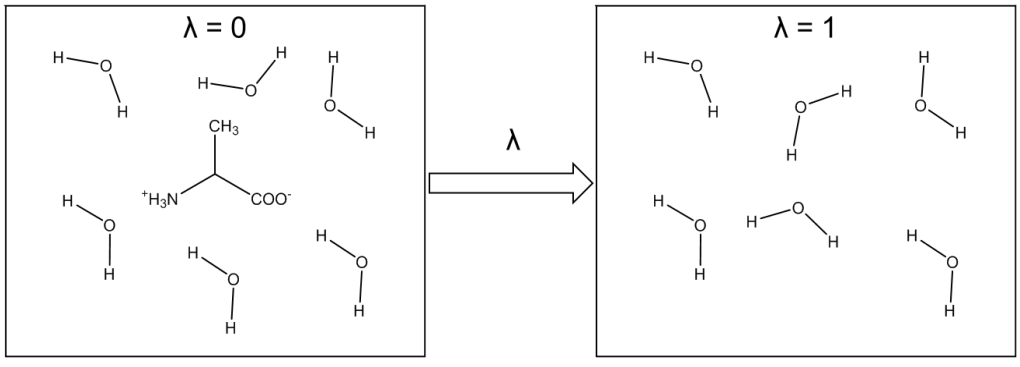

Figure 5.2 Solvation free energy transformation decoupling a solute from its solvent box

The reason for focusing on such transformations is twofold. First, solvation free energies are easier to obtain via both experimental and theoretical methods. The results of which have been extensively reported on by community resources such as the SAMPL challenges[158-160] and FreeSolv database[170-171]. It will therefore be simpler to compare the outcomes of msesTI calculations against experimental free energies of solvation and calculated standard TI values. Secondly, whilst the use of the butane-to-butane alchemical mutation was effective in demonstrating the effectiveness of the sesTI method[12], such a validation introduces some ambiguity in terms of the validating the reweighting procedure. This is because the transformation amounts to going from two equivalent biased endpoints, which means that even without reweighting one would expect the free energy difference to be zero. Obviously, one could opt for an alchemical mutation with different end points; however very little experimental data is available on such test cases. Additionally, there are limitations in the current *pmemd* thermodynamic integration code implementation which can at times introduce some abnormal behaviour during alchemical mutations. For the sake of this evaluation, it is best to avoid such transformations in this particular evaluation. We hope to expand our validation set in the future to include a wider range of alchemical transformation, possibly once the above mentioned code limitations have been fully accounted for.

This solvation free energy study primarily looks at seven flexible small molecule systems of varying sizes: butan-1-ol, prop-2-en-1-ol, glycerol, 2-propoxyethanol, 1-butoxy-2-propanol, mannitol and malathion (molecules **1**-**7**, Figure 5.3). These systems are all derived from the FreeSolv database[170-171], and were chosen based on: the presence of one or more rotatable torsions; the simplicity of the structures; and deviations of the calculated FreeSolv free energies relative to experiment by more than 1 kcal mol^{-1}. The reason for the latter discriminant was to ensure that there is scope for improvement in the free energy calculation, either through improved sampling or a better force field representation of the model. As an unintended consequence, the majority of the systems chosen are hydroxyl rich, although it is hoped that the presence of the organophosphate malathion should provide some insights on the impact of other types of functional group. As detailed in the Method section, the solutes are described by the GAFF force field and the AM1-BCC charge method so as to match the FreeSolv database[170-171]. Nonetheless, in order to test the impact of using more complex partial charges, particularly in terms of describing the

higher energy states frequently visited by msesMD simulations, we also test a RESP[16] charge assignment protocol.

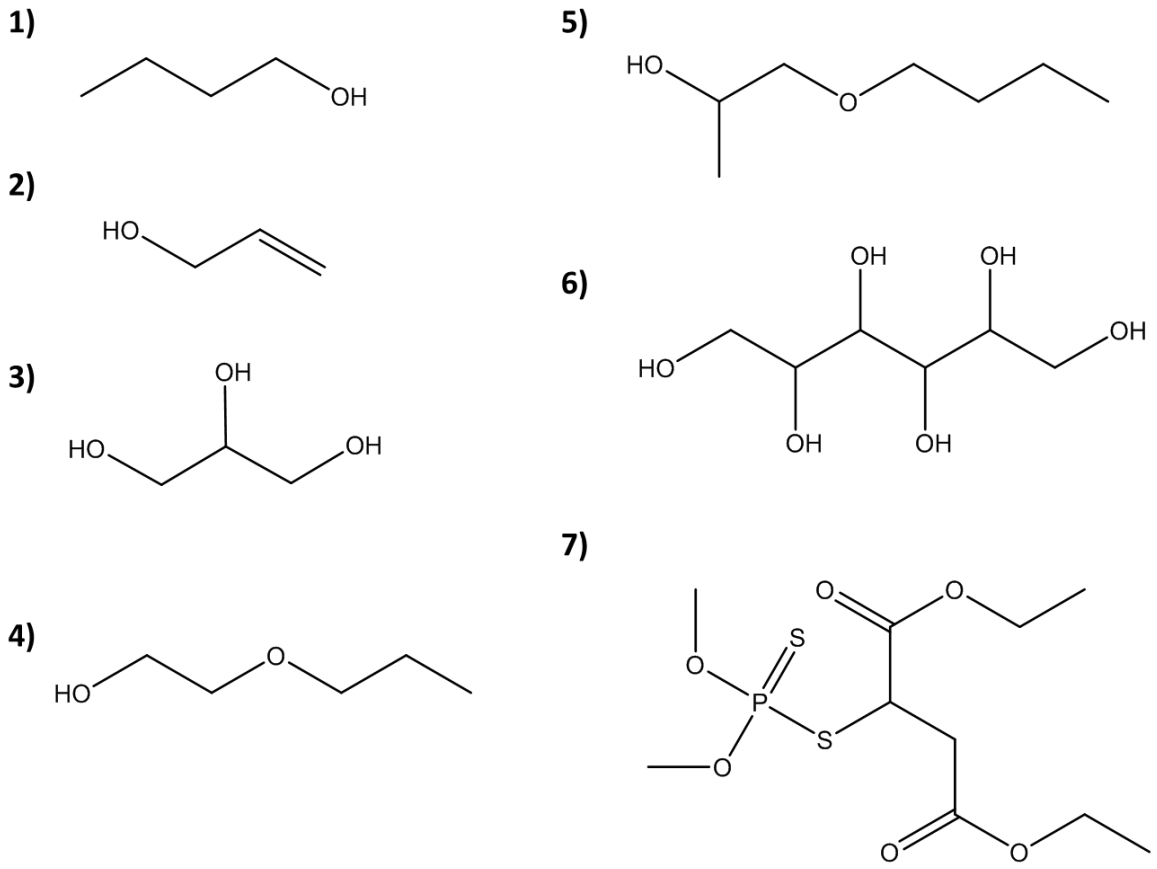

*Figure 5.3 Small molecule systems investigated. **1**) butan-1-ol, **2**) prop-2-en-1-ol, **3**) glycerol, **4**) 2-propoxyethanol, **5**) 1-butoxy-2-propanol, **6**) mannitol and **7**) malathion*

Beyond evaluating the msesTI methodology, we also take the opportunity to look at the impact of using hydrogen mass repartitioning (HMR) in free energy calculations. Although the use of HMR in the context of conformational sampling has been extensively investigated[26], very little is known about the true impact of such a mass altering scheme on the recovery of thermodynamic averages along alchemical paths. In fact, whilst HMR is frequently used in restraint-based free energy methods[26, 29], only one example of using a 4 fs HMR scheme during an alchemical transformation is known to have been previously reported.[172] This study, which looks at host-guest binding free energies as part of the SAMPL5 challenge[173], did demonstrate relatively good agreement with the results of standard mass attach-pull-release simulations reported by Yin et al.[174] However, deviations

did exceed 1 kcal mol^{-1} in some cases, and without an accurate standard mass comparison using the same alchemical method, it is difficult to tell if the use of HMR inadvertently biases free energy estimates. As previously described (Chapter 1.2.2), ensemble averages are not mass dependent, thus one may expect the recovery of $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle$ to be accurate, particularly in the context of soft-core simulations where the bonded terms have no contribution towards the λ derivative of the potential. Nevertheless, the use of HMR is known to have unintended influences on the dynamics, particularly with regard to hydrogen bond formation.[29] Proper evaluation of the HMR method, especially in cases where the system takes on unusual chemical properties, is therefore necessary.

5.2 Methods

5.2.1 Implementation of msesMD within a soft-core TI framework

The free energy change, ΔA, along an alchemical path, defined by the scaling parameter λ, can be obtained via[150]

$$\Delta A = \int_0^1 \frac{\delta A}{\delta \lambda} \delta\lambda = \int_0^1 \langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle \, \delta\lambda \qquad [5.3]$$

where $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle$ describes the ensemble average of the partial derivative with respect to λ of the potential energy of the system at a given position along the λ path, U(r,λ). As detailed in this chapter's introduction, the potential energy of the system can be described by a linear switch over λ between two alchemical states, where the initial and final states are fully described at λ = 0 and λ = 1 respectively. This leads to the following description of the potential energy (note: this is the same as equation 5.2)

$$U(r,\lambda) = U_{common}(r) + (1-\lambda)[U_{initial}(r,\lambda)] + \lambda[U_{final}(r,\lambda)] \qquad [5.4]$$

where U_{common} describes the sum of the potential energy contributions only involving atoms present in both end states, whilst $U_{initial}$ and U_{final} are the sum of any potential energy contributions involving one or more atoms in the initial and final states respectively. Unfortunately, using such a linear scaling approach is problematic as it tends to lead to numerical instabilities in the evaluation of the non-bonded terms at the end states. This is because the effective size of perturbed atoms with small contributions to the system near the end states is small, allowing them to get close to other atoms. This causes the reciprocal of the atomic distance terms, r_{ij}, in the nonbonded contributions, to yield extremely large values. This is particularly true for the r_{ij}^{-12} term in the Lennard-Jones (LJ) potential but also to an extent the r_{ij}^{-1} in the Coulomb potential. As a solution to this, so-called soft-core

versions of the nonbonded potentials have been introduced which smoothly switches off interatomic interactions as a function of λ, avoiding the formation of singularities.[175] The LJ soft-core potential for the disappearing state takes the form of[69, 175]

$$U_{VdW}^{\lambda=0} = 4\varepsilon(1-\lambda)\left[\frac{1}{\left[\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6}\right] \quad [5.5]$$

with σ describing the collision diameter, ε the well depth, and α being a scaling parameter that defines the smoothness of the soft-core potential. Similarly, the soft-core Coulomb potential for the disappearing state takes the form,

$$U_{Eel}^{\lambda=0} = (1-\lambda)\left[\frac{q_i q_j}{4\pi\varepsilon_0\sqrt{\beta\lambda + r_{ij}^2}}\right] \quad [5.6]$$

with q_i and q_j representing the charges of atoms i and j respectively, ε_0 the vacuum permittivity, and β the Coulombic analogue of the soft-core scaling parameter α.

In soft-core simulations, the potential mixing function is altered in order to allow for the soft-core regions to become decoupled from the system. This is arranged in such a way that the nonbonded interactions, $U_{initial}^{nbsc}$ and U_{final}^{nbsc}, are scaled as a function of λ, whilst the bonded contributions, $U_{initial}^{bsc}$and U_{final}^{bsc} remain intact.[176] The potential energy of the system therefore becomes

$$U(r,\lambda) = U_{common}(r) + U_{initial}^{bsc}(r,\lambda) + U_{final}^{bsc}(r,\lambda) + (1-\lambda)U_{initial}^{nbsc}(r,\lambda) + \lambda U_{final}^{nbsc}(r,\lambda) \quad [5.7]$$

By doing this, the only contributions to $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle$ stem from the soft-core nonbonded terms. Unlike the linear scaling approach (eq 5.4), this has the advantage of ensuring that the perturbed atoms experience the full force arising from their bonded contributions no matter the λ value, hence preventing the sampling issues that would otherwise occur as the potential energy tends to zero.

As previously described[12], when applying a boost potential such as msesMD to a simulation, the resultant biased total potential, $U^*(r_n)$, on replica *n* can be described as

$$U^*(r_n) = U^{MM}(r_n) + U^{ses}(r_n) \quad [5.8]$$

where $U^{MM}(r_n)$ represents the potential energy contributions arising from the force field based on the atomic coordinates **r**, and $U^{ses}(r_n)$ those of the swarm biasing potential acting on replica *n* (as defined in Chapter 2.2.1). In the context of a soft-core simulation, if we treat the swarm biasing potential as an additional bonded term which is applied to dihedrals atoms, then we can define the resultant potential on replica *n* to be

$$\begin{aligned} U^*(r_n,\lambda) = U_{common}(r_n) + U^{ses}_{common}(r_n) + U^{bsc}_{initial}(r_n,\lambda) + U^{ses}_{initial}(r_n,\lambda) \\ + U^{bsc}_{final}(r_n,\lambda) + U^{ses}_{final}(r_n,\lambda) + \\ (1-\lambda)U^{nbsc}_{initial}(r_n,\lambda) + \lambda U^{nbsc}_{final}(r_n,\lambda) \end{aligned} \quad [5.9]$$

where $U^{ses}_{common}(r_n)$, $U^{ses}_{initial}(r_n,\lambda)$ and $U^{ses}_{final}(r_n,\lambda)$ represent the swarm potential contributions to the common, initial and final regions to replica *n* respectively. Through this approach, the resultant swarm potential is independent of λ, which means that the forces can be applied to the coordinates in the same manner as a normal msesMD calculation. Thus, the total swarm potential acting on replica *n*, $U^{ses}_{tot}(r_n,\lambda)$, can be defined as a linear addition of the swarm potentials added to each portion of the system:

$$U_{tot}^{ses}(r_n, \lambda) = U_{initial}^{ses}(r_n, \lambda) + U_{final}^{ses}(r_n, \lambda) + U_{common}^{ses}(r_n) \quad [5.10]$$

This definition holds for all λ states except from the end states. At the end states, since one of the perturbed regions has no influence on the system, any swarm boost applied to that region has no impact on the ensemble properties. Therefore, in those specific cases, the total swarm potential can be expressed as,

$$U_{tot}^{ses}(r_n, \lambda) = (1-\lambda)U_{initial}^{ses}(r_n, \lambda) + \lambda U_{final}^{ses}(r_n, \lambda) + U_{common}^{ses}(r_n) \quad [5.11]$$

As previously described (see Chapter 2.4.2), the expectation value of ensemble averages can be recovered from biased simulations through the exponential reweighting method.[12, 38] Within the context of a TI calculation, the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_\lambda$ from an M replica swarm can be recovered through the "independent replica" reweighting scheme in the following manner[12]

$$\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_\lambda = M^{-1} \sum_n^M \frac{\langle \frac{\delta U^*(r_n,\lambda)}{\delta \lambda} \exp(\beta U_{tot}^{ses}(r_n,\lambda)) \rangle_\lambda}{\langle \exp(\beta U_{tot}^{ses}(r_n,\lambda)) \rangle_\lambda} \quad [5.12]$$

where β represents the Boltzmann constant. If one assumes that the averages of each swarm replica converges over the simulation, then the "group reweighting" scheme can also be used:

$$\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_\lambda = \frac{\sum_n^M \langle \frac{\delta U^*(r_n,\lambda)}{\delta \lambda} \exp(\beta U_{tot}^{ses}(r_n,\lambda)) \rangle_\lambda}{\sum_n^M \langle \exp(\beta U_{tot}^{ses}(r_n,\lambda)) \rangle_\lambda} \quad [5.13]$$

Given sufficient sampling, these two approaches should, within error, lead to identical results. As demonstrated in previous chapters, the latter form of the exponential reweighting scheme (eq. 5.13) has been shown to accurately recover free energy surfaces. However, in this study, we instead primarily use the "independent replica" reweighting

scheme (eq 5.12). There are two reasons for this; first the sesTI paper demonstrated that at very short simulation times, such as the sub-nanosecond simulations used here, the use of the "independent replica" scheme can lead to slightly improved estimates of the mean.[12] Secondly, the calculation of the uncertainty via the "independent replica" reweighting scheme follows the same approach as the unbiased IT-TI scheme used in this study. For both the "independent replica" reweighting scheme and IT-TI, the uncertainty is calculated as the standard error, SE, of the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates of the *n* replica simulations[169]

$$SE = \frac{\sigma_{\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda}}{\sqrt{n}} \qquad [5.14]$$

where $\sigma_{\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda}$ is the standard deviation of the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates. However, for the "group reweighting" scheme, the uncertainty is instead calculated as the standard deviation of the bootstrap sampling mean estimates:

$$SD = \sqrt{\sum_{i}^{bootstraps} \left(\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_{\lambda,i} - \mu_{\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda}\right)^2} \qquad [5.15]$$

where $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_{i,\lambda}$ represents the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimate of a single bootstrap and $\mu_{\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda}$ the mean of all $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates across all bootstraps. Statistically, both of these uncertainty estimates are valid, but their values can differ significantly depending on the distribution of the underlying data. Therefore, using the "independent replica" reweighting scheme (eq 5.12) to compare against the IT-TI results would be more useful. Nevertheless, in order to demonstrate the validity of using either schemes, our analysis will also compare the difference in the results via both the "independent replica" and "group" reweighting schemes.

5.2.2 Simulation details

System preparation: All systems were parameterised via the *leap* module in AmberTools14[32]. For all small molecules (Figure 5.3), in order to match the FreeSolv parameters[171], solute force field parameters were obtained from the general AMBER force field (GAFF)[17] version 1.7. Initial conformations and AM1-BCC partial charges were directly obtained from the FreeSolv database version 0.32.[170] A separate set of simulations were also carried out for these systems using partial charges derived via RESP[16, 177] as detailed below. All systems were solvated in a TIP3P[66] water octahedral box with waters placed up to a minimum of 12 Å away from the solute. Since all systems are neutral, no counterions were added. For the HMR simulations, the default AMBER repartitioning scheme was used, redistributing 3 amu from heavy atoms to their bonded hydrogens for the solutes only.

RESP charge assignment procedure: An msesMD-based RESP charge derivation procedure is used here in order to scope its potential use as a force field development tool. The idea of this protocol is to use msesMD to fully explore conformational space and recover suitable representative conformations for charge derivation. To achieve this, the AM1-BCC msesTI production simulations were first generated using the simulation protocols defined below. The 8 msesTI replica trajectories from the $\lambda = 0.0$ window, i.e. the solute fully present in the solvent, were then clustered in order to identify representative conformations. The clustering was achieved using the DBSCAN algorithm[96] in AmberTools16's *cpptraj* module.[59] The clustering criteria was adjusted such that the standard deviation of the frames within each cluster was below 0.6 Å, ensuring that the representative conformer for each cluster is similar to all other conformers in that cluster. The cluster densities were then reweighted accordingly, based on the swarm potential energy contribution at each frame, and the most occupied cluster conformation was used to derive the RESP charges. The chosen conformation was first optimised at the HF/6-31G* level of theory, followed by a single point electrostatic potential (ESP) calculation at the same level of theory using Gaussian09d[178]. Finally, the RESP partial charges were then obtained from the ESP grid using the *antechamber* and *resp* modules in AmberTools14.

Simulation details: Simulations were carried out under periodic boundary conditions, using a 9 Å cut-off for short-range nonbonded interactions and the particle mesh Ewald (PME)[67] method to handle long-range electrostatics. Thermal control was achieved using the Langevin thermostat[25], with a target temperature of 298 K and a collision frequency of 3 ps^{-1}. Hydrogen bond motion in the solute and solvent molecules were constrained using the SHAKE[27] and SETTLE[33] algorithms respectively. Standard mass simulations used an integration time step of 2 fs whilst the HMR simulations used 4 fs. Pressure equilibration was achieved in the NPT ensemble using the Monte Carlo barostat with exchange attempts every 200 fs. All production free energy simulations were carried out under the canonical ensemble (NVT). It is noted that this means that resultant free energies are Helmholtz free energies rather than the experimentally reported Gibbs free energies in the FreeSolv database.[170] However for small molecules of the size investigated here, the contribution to the pressure volume term, $P\Delta V$, is expected to be minimal. Therefore, in this case, $\Delta A \approx \Delta G$ can be approximated within simulation uncertainty. The alchemical free energy simulation was carried out using a one-step soft-core transformation along 21 equidistant windows, ranging from λ values of 0 to 1, with a spacing of 0.05. The AMBER default soft-core scaling parameters were used, ie. $\alpha = 0.5$ Å^2 and $\beta = 12$ Å^2, as validated by Steinbrecher et al.[175]. During production simulations, the energies, including the $\frac{\delta U(r,\lambda)}{\delta \lambda}$ values, were sampled every 400 fs whilst the trajectories were written every 5 ps.

Equilibration protocol: Both IT-TI and msesTI simulations follow the same initial equilibration protocol. Under the complete presence of the solute, i.e. $\lambda = 0.0$, the systems were first energy minimised to remove atomic clashes using 25000 steps of steepest descents optimisation, followed by 25000 steps of conjugate gradients. The systems were then slowly heated under NVT conditions over 500 ps from 0 K to the 298 K target temperature. The box density was then equilibrated over 1 ns of NPT simulation to a target pressure of 1 bar. The system was then further equilibrated under NVT conditions for a further 1 ns. This NVT equilibrated state with the complete presence of the ligand was then used as a starting point for all λ states. For each λ value, the system was replicated into 8 independent trajectories which were further equilibrated for 500 ps under the λ scaling influence.

IT-TI simulations: Upon equilibration, the 8 replicated trajectories at each λ window were further equilibrated for an additional 500 ps. This was then followed by the production simulation. For the small molecule systems (Figure 5.3, **1**-**7**), the production simulation was 5 ns per replicate trajectory. For the monosaccharides (Figure 5.3, **8**-**11**), the production simulation length was increased to 10 ns per replica.

msesTI simulations: Upon equilibration, the 8 replicas were coupled using the msesMD pair potential which was slowly introduced over a 300 ps period and then allowed to equilibrate for a further 200 ps. This was then followed by the production simulation, which as per the IT-TI simulation, was calculated for 5 ns per replica. In this investigation, the swarm potential used "half-scaled" swarm parameters (relative to those used in Chapter 2 and 4): A = -0.25 kcal mol^{-1}, B = 0.5 rad^{-1}, C = 1.065 kcal mol^{-1} and D = 2.625 rad^{-1}. Such a parameter choice leads to a weaker boost potential. The reason for this choice is so as to reduce the magnitude of the statistical errors in the reweighted ensemble averages that stem from the noise introduced by the use of a boost potential.[37] Considering that these small molecules do not have excessively large barriers to conformational change, the reduction in the boost potential is believed to have little reduction in sampling efficiency. To evaluate the impact of this parameter choice, msesTI calculations using the "full" swarm parameters, A = -0.5 kcal mol^{-1}, B = 0.5 rad^{-1}, C = 2.13 kcal mol^{-1} and D = 2.625 rad^{-1}, were also carried out. No obvious low frequency motion could be identified in the seven small molecules, so the swarm boost potential was instead applied to all unique dihedrals consisting of four heavy atoms.

Analysis: The $\frac{\delta U(r,\lambda)}{\delta\lambda}$ profiles collected during the course of the free energy simulations were decorrelated according to the statistical inefficiency, g_A, which is defined as[179-180]

$$g_A = 1 + 2\tau_A \qquad [5.16]$$

where τ_A is the autocorrelation time. This autocorrelation time is the integral of the autocorrelation function, C_A, over a sufficiently large lag time limit[179]

$$\tau_A = \sum_{t=1}^{lim} C_A(t)\left(\frac{1-t}{N}\right) \quad [5.17]$$

where *t* is the lag time, *N* is the total number of samples, and *lim* is the maximum lag time limit. Ideally, the lag time limit should be the entire simulation length, but due to statistical noise, using longer lag times results in erroneous correlation time estimates. Thus, a limit is usually chosen that is less than half the length of the simulation time. The normalised autocorrelation function, C_A, of the observable *A*(x) time series is calculated as[179],

$$C_A = \frac{\langle (A(x_0) - \mu_A)(A(x_t) - \mu_A) \rangle}{\sigma_A} \quad [5.18]$$

where μ_A and σ_A represent the average and standard deviation of *A*(x) respectively. For these simulations, the autocorrelation function was calculated over the first 500 ps of each replica simulation and then integrated over the first 200 ps. The maximum statistical inefficiency across all replicas of all λ windows was then used to decorrelate all $\frac{\delta U(r,\lambda)}{\delta \lambda}$ time series. Mean and uncertainty estimates were obtained via bootstrap sampling and the reweighting methods outlined above (section 5.2.1) using a total of 250000 bootstrap resamples per time series. The $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_\lambda$ estimates were then integrated across all windows using the trapezoidal rule and the uncertainty was calculated using error propagation. The statistical inefficiencies and bootstrap resampling calculations were carried out using a custom OpenMP parallelised C++ code using the Intel MKL libraries[181].

All other analyses were carried out using the AmberTools16 *cpptraj* module and custom python scripts. Hydrogen bond analyses were carried out by looking at the formation of hydrogen bonds between the solute and solvent molecules, using the default *cpptraj* settings. Radial distribution functions were calculated from the geometric centres of the solute atoms of interest to all water oxygens up to a distance of 4 Å using a grid spacing of 0.1 Å. As in previous chapters, the dihedral surfaces were binned with a spacing of 8° and a maximum free energy of 8 kcal mol^{-1}.

5.3 Results and discussion

5.3.1 Estimating solvation free energy using IT-TI

We first start by evaluating the effectiveness of the IT-TI protocol in estimating the free energies of solvation for the seven small molecules (Figure 5.3) using AM1-BCC charges. Considering the absolute deviations in the solvation free energy estimates for the 5 ns per replica trajectories (Figure 5.4, **A**), we can see that all systems exhibit a greater than 1 kcal mol^{-1} error. This not unexpected considering that these systems were specifically chosen from the FreeSolv database due to their large deviation between the calculated and experimental values. Looking at the free energy estimates we find that for all molecules except from malathion, the free energies are overestimated relative to experiment (Supplementary Table D.1). Furthermore, we find that this error increases as a function of the number of hydroxyl groups present, with glycerol (**3**) and mannitol (**6**) showing much larger errors than systems **1**, **2**, **4** and **5**, which only contain one hydroxyl each. This trend is consistent with previous findings by Mobley et al.[182], who found systematic solvation free energy overestimation errors for alcohol-containing compounds using the GAFF/AM1-BCC parameterisation method.

Overall, we find that the IT-TI results are well converged, with uncertainties in the estimate within ±0.1 kcal mol^{-1}. In fact, using shorter simulations times, as demonstrated by the 500 ps and 1 ns per replica estimates, does not significantly affect the results with deviations remaining within error. It is however noted that using smaller simulation times does, as expected, lead to increased uncertainties in the estimates; this is especially the case for the larger systems such as malathion, which sees the uncertainty triple from ±0.05 kcal mol^{-1} to ±0.15 kcal mol^{-1} (Supplementary Tables D.1-D.3). Nevertheless, even at small simulation times, the calculated uncertainties remain smaller than the experimental uncertainties, indicating such small timescales could be used in conjunction with IT-TI to rapidly obtain free energy estimates.

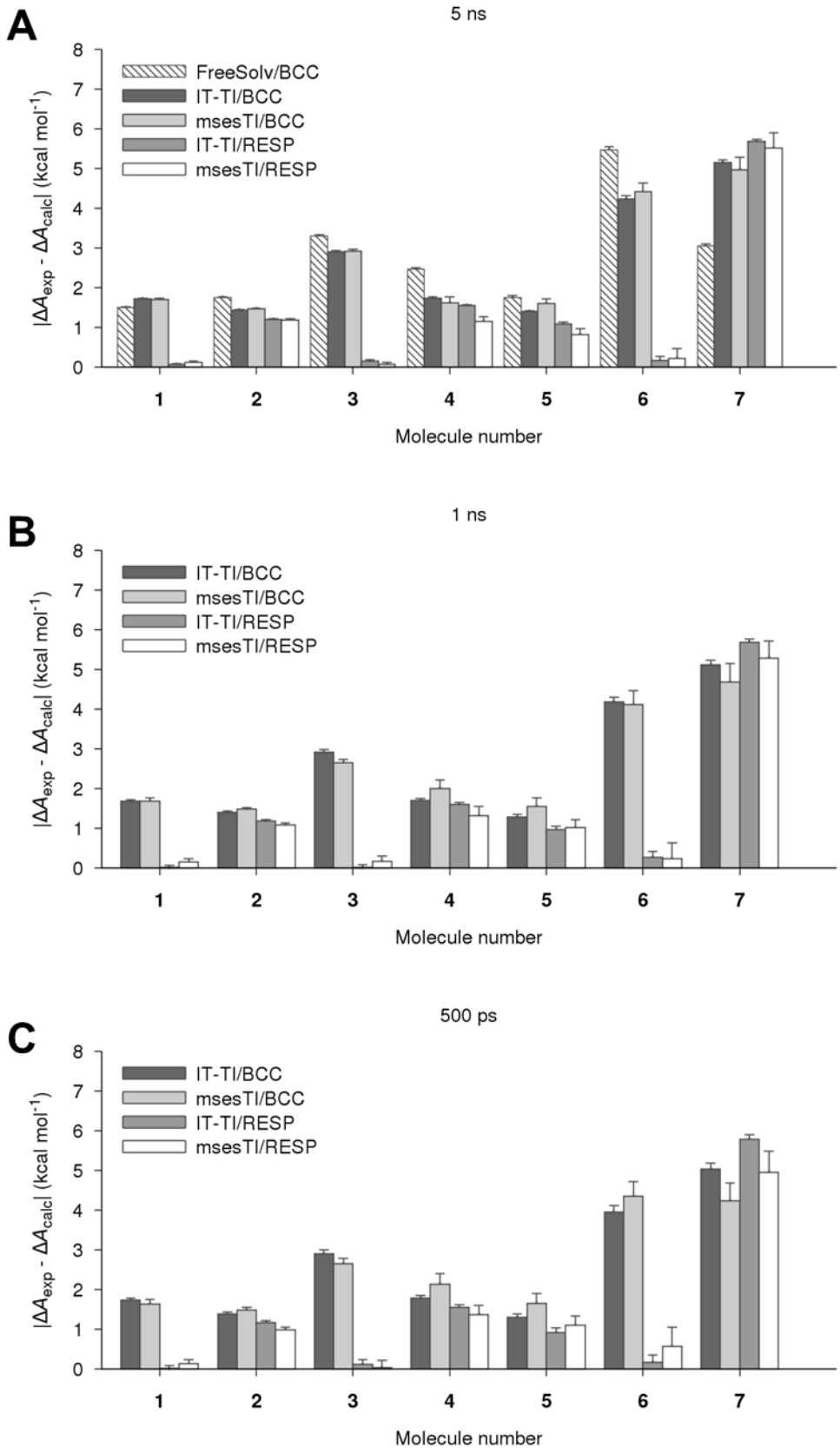

Figure 5.4 Absolute deviations from experiment for FreeSolv, IT-TI and msesTI free energy estimates

Comparing the IT-TI/AM1-BCC results against the FreeSolv calculated estimates, we find that there is good agreement for systems **1** to **5**. However, greater than 1 kcal mol^{-1} deviations are seen for the two larger systems, mannitol (**6**) and malathion (**7**) (Figure 5.4, **A**). In both cases, the deviation stems from IT-TI estimating lower free energy values than FreeSolv (Supplementary Table D.1). In fact, this general trend of underestimating free energies relative to FreeSolv appears to be consistent across all systems except for butan-1-ol (**1**). A paired two-tailed student t-test on the signed errors reveals a probability of 0.051, which although slightly above significance ($p < 0.05$) indicates the possibility for this difference to be systematic. Considering that the same force field and partial charges were used in both the FreeSolv and IT-TI simulations, the reason for such a difference in the estimates, particularly the several kcal mol^{-1} deviation seen in the larger systems, is unclear. The fact that the estimate does not change much based on simulation time means that this is unlikely to be due to sampling limitations. Part of the difference for the larger systems could potentially be due to the fact that the FreeSolv simulations are done under NPT conditions whilst ours are under NVT. However, the change in density due to solute decoupling is only 1.4×10^{-2} g cm^{-3} for malathion (the largest solute), which should not have a major influence on $P\Delta V$. The most likely reason is the difference in the way in which the electrostatics are treated in both protocols. Unlike the one-step soft-core decoupling scheme used in our simulations, the FreeSolv protocol uses a more rigorous two-step transformation, which does not rely on the use of a soft-core Coulomb potential. Whilst one-step soft-core has previously been shown to be equally as effective as a two-step approach[175], deviations between the methods are expected as the electronic perturbation increases. This, coupled with inconsistencies in the internal Coulombic constants between MD engines, as previously reported by Shirts et al.[183] could result in the differences seen here.

5.3.2 Impact of RESP charges

Having detailed the outcomes of using IT-TI with the approximate AM1-BCC partial charge assignment method, we now look at the use of RESP charges. Overall, the inclusion of RESP charges generally leads to a decrease in the free energy estimate error relative to AM1-BCC calculations, except from malathion (**7**) where we instead see an increase of around 0.5 kcal mol^{-1} (Figure 5.4 **A**, Supplementary Table D.1). The fact that the use of RESP charges improves solvation free energy estimates in alcohols (systems **1**-**6**) concurs with previous studies[184], and appears to stem from an increased polarisation of the C-O-H partial charges. We find that marked improvements, of up to 4.1 kcal mol^{-1}, are mainly seen in the simpler polyhydroxylated systems, butan-1-ol (**1**), glycerol (**3**) and mannitol (**6**), with free energy estimates falling within error of the experimental values. However, for the remaining systems, prop-2-en-1-ol (**2**), 2-propoxyethanol (**4**) and 1-butoxy-2-propanol (**5**), changes from the AM1-BCC estimates are minimal at around 0.2 kcal mol^{-1}. The reason for this difference is unclear; it may be due to incorrect polarisation of hydrocarbon moieties, particularly near the tail ends of the molecule. An example of this can be seen in Figure 5.5 where the partial charge of terminal carbons of 1-butoxy-2-propanol become more negative. In comparison, the relative distribution of partial charges in mannitol does not change much between AM1-BCC and RESP. In the cases of 2-propoxyethanol (**4**) and 1-butoxy-2-propanol (**5**), this may in part be due to the re-orientation of the hydroxyl group towards the ether oxygen during the HF/6-31G* gas phase optimisation, leading to charge distributions that are less representative of the condensed phase (Figure 5.6).

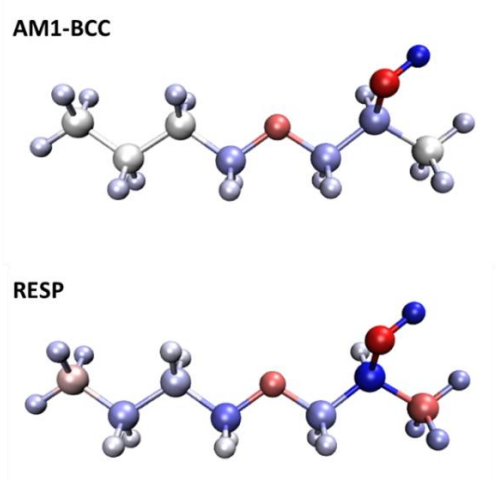

Figure 5.5 Comparison of partial charge distribution for 1-butoxy-2-propanol with both BCC and RESP partial charge assignment. Positive and negative charges are coloured as blue and red respectively.

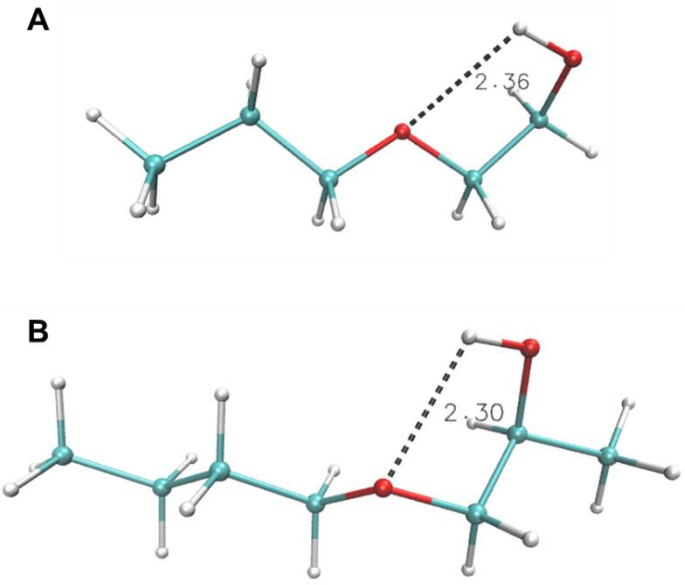

Figure 5.6 Orientation of the hydroxyls post gas phase HF/6-31G optimisation for A) 2-propoxyethanol, B) 1-butoxy-2-propanol. The hydrogens, initially pointing away from the rest of the molecule re-orient to form intramolecular hydrogen bonds.*

In terms of malathion (**7**), the increase in error seen here is contrary to a previous study by Mobley et al.[185]. The latter work found that, for **7**, using a single conformation RESP charge assignment protocol reduced error relative to experiment. It is likely that this difference stems from the molecule being too flexible for its charge to be adequately represented in a single conformation charge fit. Evidence for this can be seen from the

clustering densities, which show that the conformer used to derive the malathion charges (**7** in Figure 5.7) only represented around 13% of its total conformational states. This, in addition to the charge assignment issues for the hydroxyl-rich systems, indicates that due to the conformational sensitivity of RESP charge assignment, a more complex multi-conformational fitting approach would be preferable.

As discussed in the Methods section, part of the aim of using this RESP charge assignment protocol was to test the potential for the msesMD method to be used as a tool for force field development. Looking at the conformations obtained via msesMD clustering (Figure 5.7), we find that, except from malathion, they do match the original FreeSolv OmegaTK[186] generated conformers quite well. Comparing the two sets of conformers, we find that most deviations stem from symmetric torsional rotations and therefore would not have much of an impact on the charge distribution. Whilst this does not provide a justification for the use of the relatively more complex msesMD protocol, it does show that the methodology may have some promise. It therefore warrants further study in the future, particularly in the context of multi-conformational charge fitting strategies.

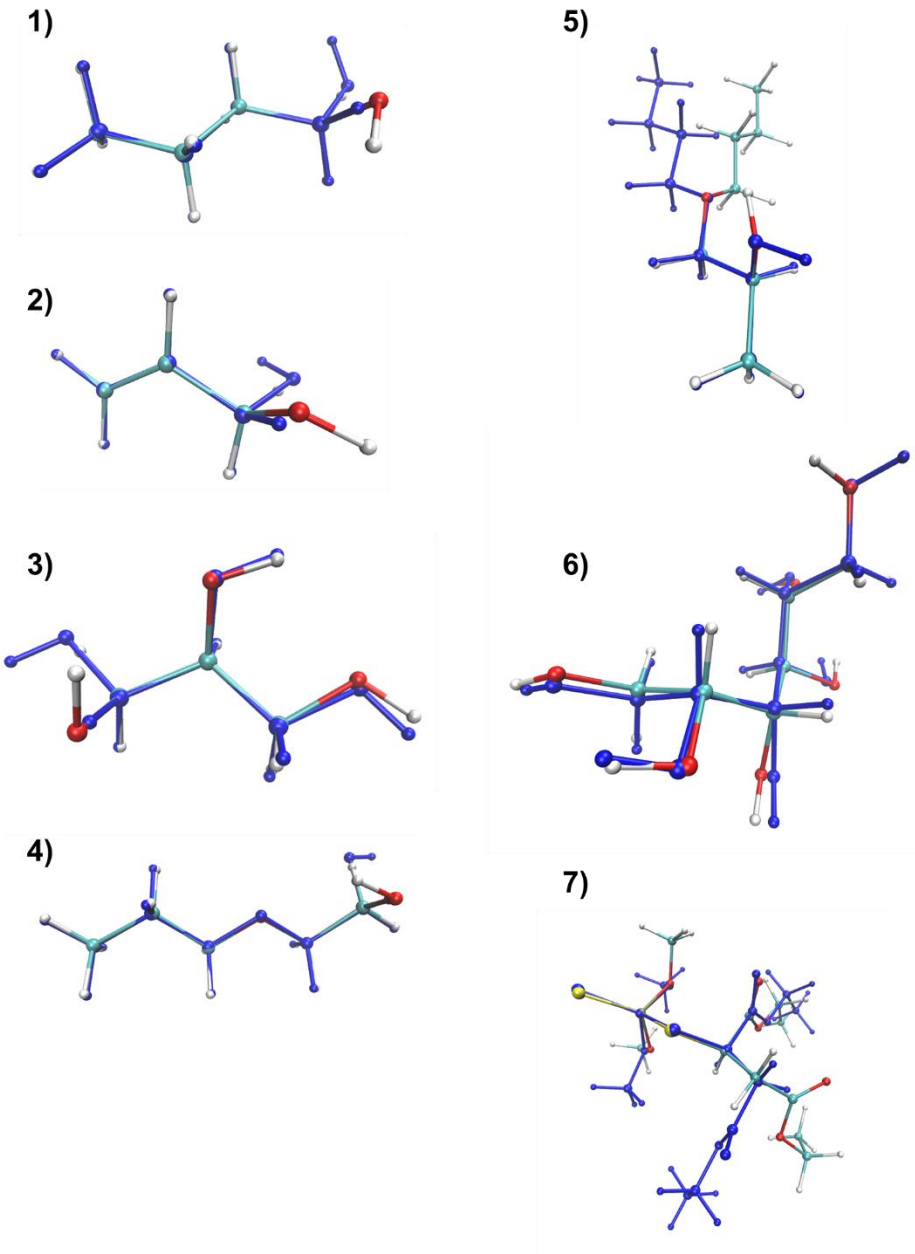

Figure 5.7 Comparison of the FreeSolv OmegaTK (blue) conformers with the gas phase optimised msesMD clustered conformations (elemental coloured) for 1) butan-1-ol, 2) prop-2-en-1-ol, 3) glycerol, 4) 2-propoxy-ethanol, 5) 1-butoxy-2-propanol, 6) mannitol, 7) malathion

5.3.3 Validation of msesTI

5.3.3.1 Comparison with IT-TI

We now turn our attention to the evaluation of the msesTI methodology in estimating the free energy of the small molecule set. As seen from the 5 ns per replica free energy estimates (Figure 5.4**A**, Supplementary Table D.1), the msesTI results are, within error, equivalent to those obtained via IT-TI. Some greater than error improvements are shown in the RESP charge estimates of solvation free energy for molecules **4** and **5**; however, the difference is marginal, with a value of 0.4 and 0.2 kcal mol^{-1} respectively. Comparing the difference in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ profiles (Supplementary Figures D.1-D.2), we find that the msesTI $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ values are close to those of IT-TI, with deviations mainly occurring at small λ values. The likely cause of these low λ deviations is that at low λ values, the solute, and by extension its conformation, have a larger influence on $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$, thus leading to larger changes in the estimate as msesTI samples more conformational states.

The main difference between mesTI and IT-TI is in terms of the magnitude of the uncertainties. Whilst the uncertainties are similar in the smaller molecules (**1**-**2**), we find that as system size increases, the msesTI uncertainties become larger than those of IT-TI, ranging from between three to six times the values for IT-TI (Supplementary Table D.1). For systems like malathion, where the uncertainty reaches up to ±0.39 kcal mol^{-1}, this can be problematic, as it makes quantitative comparisons between free energy estimates difficult to make. Unfortunately, large uncertainties are a known consequence of using enhanced sampling methods, as the simulation samples a wider range of states relative to unbiased simulations, leading to a less defined estimate of the averages. Although reweighted, the width of the distributions are large, thus leading to large fluctuations in the mean estimates. Solutions to this could involve either using longer simulation times, or a more complex free energy scheme such as the multistate Bennett acceptance ratio (MBAR) approach. In the former case, this would be counterintuitive to the idea of using an enhanced sampling method, as it would only lead to increased simulation costs. In the latter case, methods such as MBAR and the non-Boltzmann Bennett acceptance ratio approach have previously been used successfully in biased simulations.[187] It could

therefore be suitable for use with msesMD; unfortunately due to time restrictions, this could not be achieved within the time frame of this project. A further point to make is that due to the large perturbations involved, absolute free energies are generally prone to higher uncertainties compared to other thermodynamic cycles. It is therefore possible that large uncertainties may not be as much of an issue in other types of simulations such as alchemical mutations.

Looking at the free energy estimates from shorter simulations times (Figure 5.4**B**-**C**), we see more fluctuations in the estimate than in the case of IT-TI, but the deviations usually remain within error. The exception to this are 2-propoxyethanol (**4**) and malathion (**7**), where the AM1-BCC estimates for the 5 ns and 500 ps per replica simulations deviate by 0.51 kcal mol^{-1} and 0.74 kcal mol^{-1} respectively. In the case of 2-propoxyethanol, the deviation is around five times larger than the uncertainty in the 5 ns simulations. As expected, the use of shorter simulation times also leads to increases in the uncertainty, with the 500 ps simulations reaching a maximum of ±0.53 kcal mol^{-1} for the RESP charged malathion.

The overall lack of change in the free energy estimates when using either enhanced sampling or longer simulation times does appear to indicate that any sampling limitations seen in these small molecule systems do not have a large influence on the estimates. This is primarily due to the fact that when sampling limitations occur, it only tends to affect a few λ windows. An example of this can be seen when looking at the conformational changes during simulations of mannitol (with AM1-BCC charges). As detailed in Figure 5.8, we represent the conformation of mannitol in terms of its five backbone dihedrals (U_1 to U_5), with two symmetric two dimensional plots, U_1/U_2 and U_4/U_5, and a one dimensional plot of the central torsion U_3.

Looking at the fully solvated simulations (λ = 0), the msesTI simulations, as expected, readily explore the conformational space of mannitol, although some of the higher energy regions are less well defined at shorter timescales (Figure 5.9, **A**). For IT-TI this is not the case; even at 5 ns per replica timescales, certain conformational wells are not sampled

(Figure 5.9, **B**). This is particularly evident in the U_4/U_5 rotational map where the -120° to 0° region of U_4 is not explored by any of the independent replicas. Additional sampling issues include missing high energy wells, such as the -60°, -180° region of U_1/U_2, and a non-equivalent estimation of the stability of the U_3 60° and -60° regions, which due to symmetry should be equienergetic. This sampling issue is only accentuated at lower simulation times, with just two of the four main wells of U_1/U_2 and U_4/U_5 being explored by the IT-TI replicas during the first 500 ps. This difference in the range of conformers explored explains the relatively large 10 kcal mol^{-1} deviation in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates at $\lambda = 0$ (Supplementary Figure D.2).

However, as we move to higher λ values sampling no longer becomes limited. For example at $\lambda = 0.5$, both IT-TI and msesTI are able to sample the main conformational wells to a reasonable extent, even at short timescales (Figure 5.10, **A**-**B**). The reason for this is that the reduced solvent influence allows for easier sampling of conformational changes. This, in addition to a reduced solute contribution to the λ derivative, leads to better convergence in $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ as λ progresses. Therefore with sampling primarily affecting low λ values, the net contribution of poor sampling to the free energy estimate is small. This explains why only small differences in the free energy estimates are seen, with 0.28 kcal mol^{-1} between the 5 ns and 500 ps IT-TI simulations, and 0.2 kcal mol^{-1} between msesTI and IT-TI. One should note that in larger solutes, and when doing alchemical mutations, it would be expected that sampling would have a larger influence throughout a λ path, which may benefit more from the msesTI methodology.

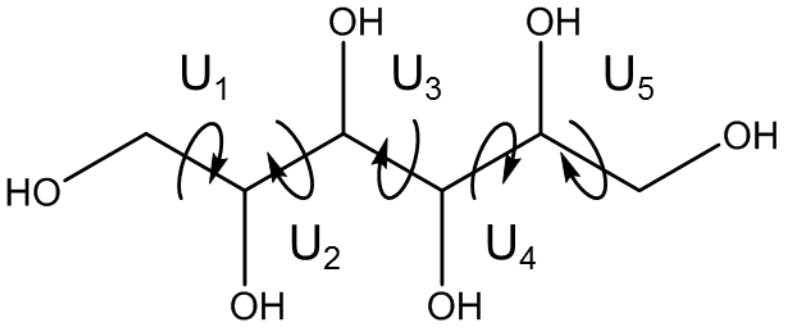

Figure 5.8 Dihedral angles of mannitol

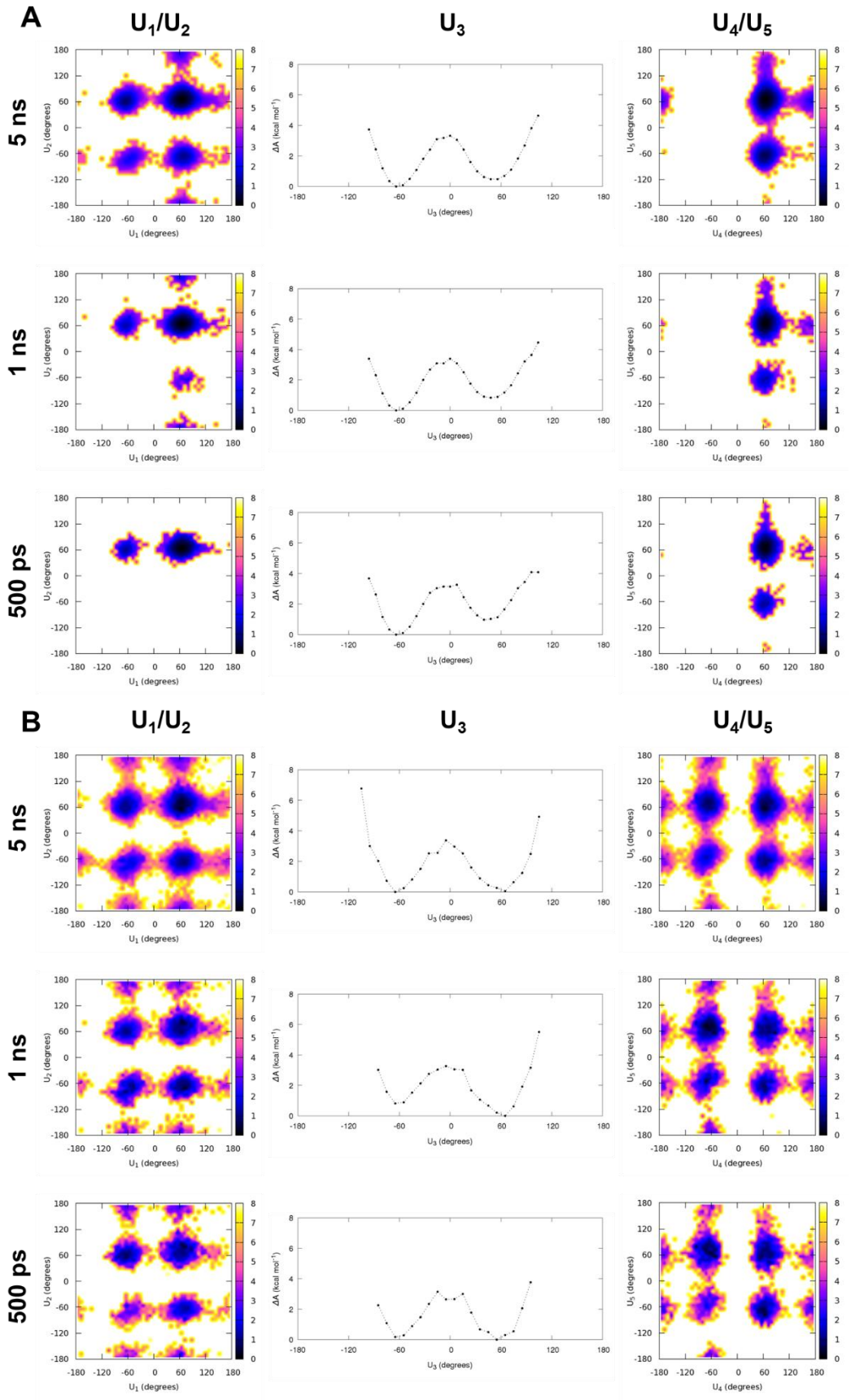

*Figure 5.9 **A**) IT-TI and **B**) msesTI approximate free energy maps of the mannitol dihedrals at varying sampling times for λ= 0.0*

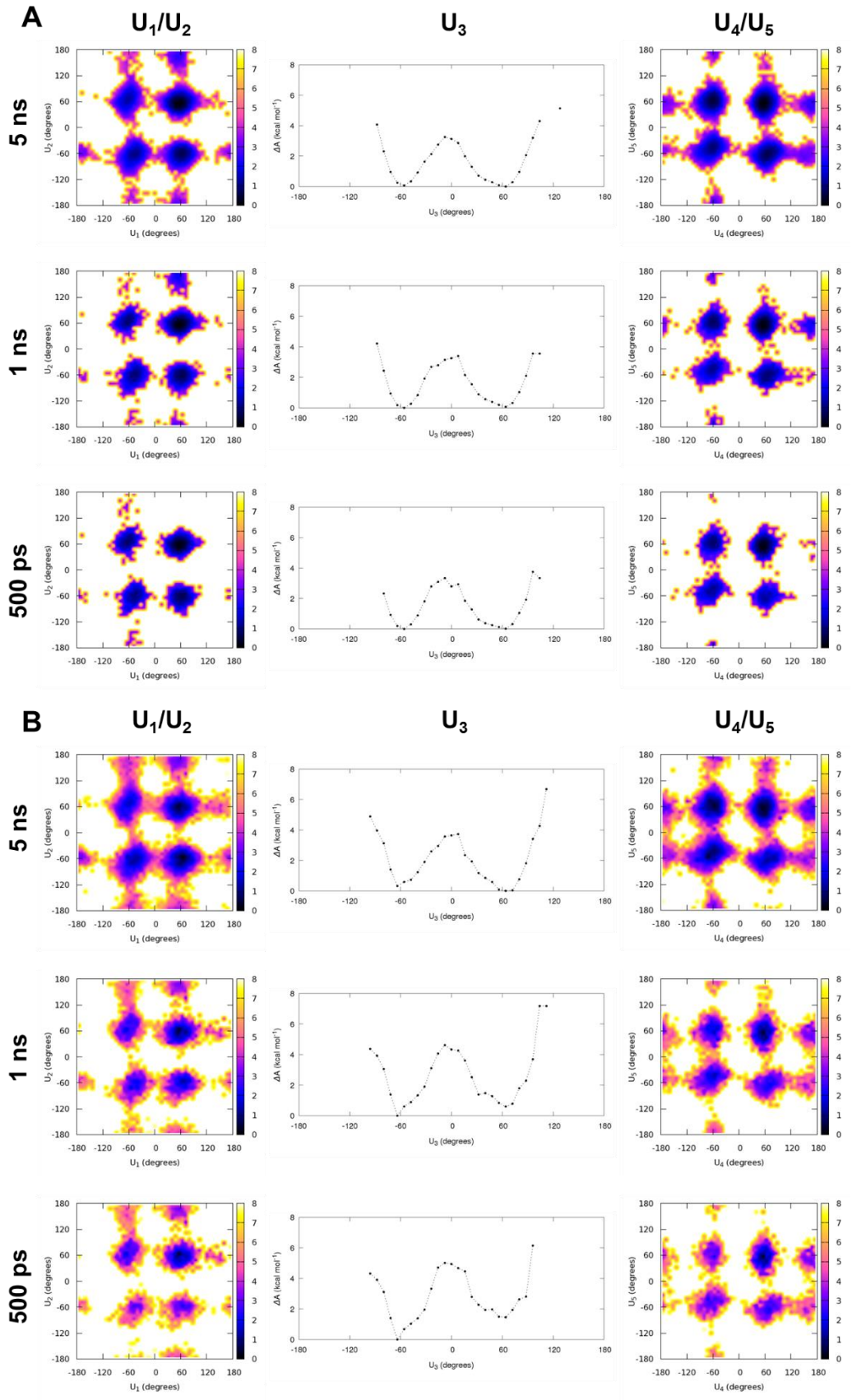

*Figure 5.10 **A**) IT-TI and **B**) msesTI (λ = 0.5) approximate free energy maps of the mannitol dihedrals at varying sampling times*

5.3.3.2 Impact of the choice of reweighting method

As discussed in the methods section, two possible approaches to the reweighting of msesMD can be used, either the "individual replica" (IndRw) or the "group" (GroupRw) methods. Assuming sufficient sampling, the two approaches should converge within uncertainty. However it has previously been shown that deviations between the two occur at small timescales[12], such as the ones used in this solvation free energy study. Thus the IndRw approach, which is less approximate and more consistent with IT-TI with regards to the calculation of uncertainties, has been used in this chapter. Here, we test the impact of choosing either reweighting strategy on the free energy estimates.

Comparing the free energy estimates calculated from the full 5 ns per replica simulations, both methods are seen to give near-equivalent answers, particularly for the smaller solutes (**1**-**3**) (Figure 5.11**A**, Supplementary Table D.4). There are some slight differences of ~0.1 kcal mol^{-1} in the estimates of the larger solutes (**4**-**7**), but they remain within uncertainty. As expected, there are differences in the calculated uncertainties, with the GroupRw method reporting uncertainties which are on average 1.3 times larger than those of the IndRw method. As explained in the Methods section, this is due to the fact that different measures of uncertainty are used for each method, with the IndRw method reporting the standard error of the means from each replica, whilst GroupRw reports the standard deviation of the bootstrap sampling means. The difference between the reweighting methods becomes more apparent at shorter timescales, where larger deviations in the free energy estimates can be seen in molecules **5** to **7** (Figure 5.11**B**-**C**). In fact in some cases, the GroupRw free energy estimate of malathion (**7**) varies from the IndRw values by over 1 kcal mol^{-1}, at both 1 ns and 500 ps timescales. This coincides with large uncertainties in the estimate of up to ±1.16 kcal mol^{-1} (Supplementary Tables D.5 and D.6).

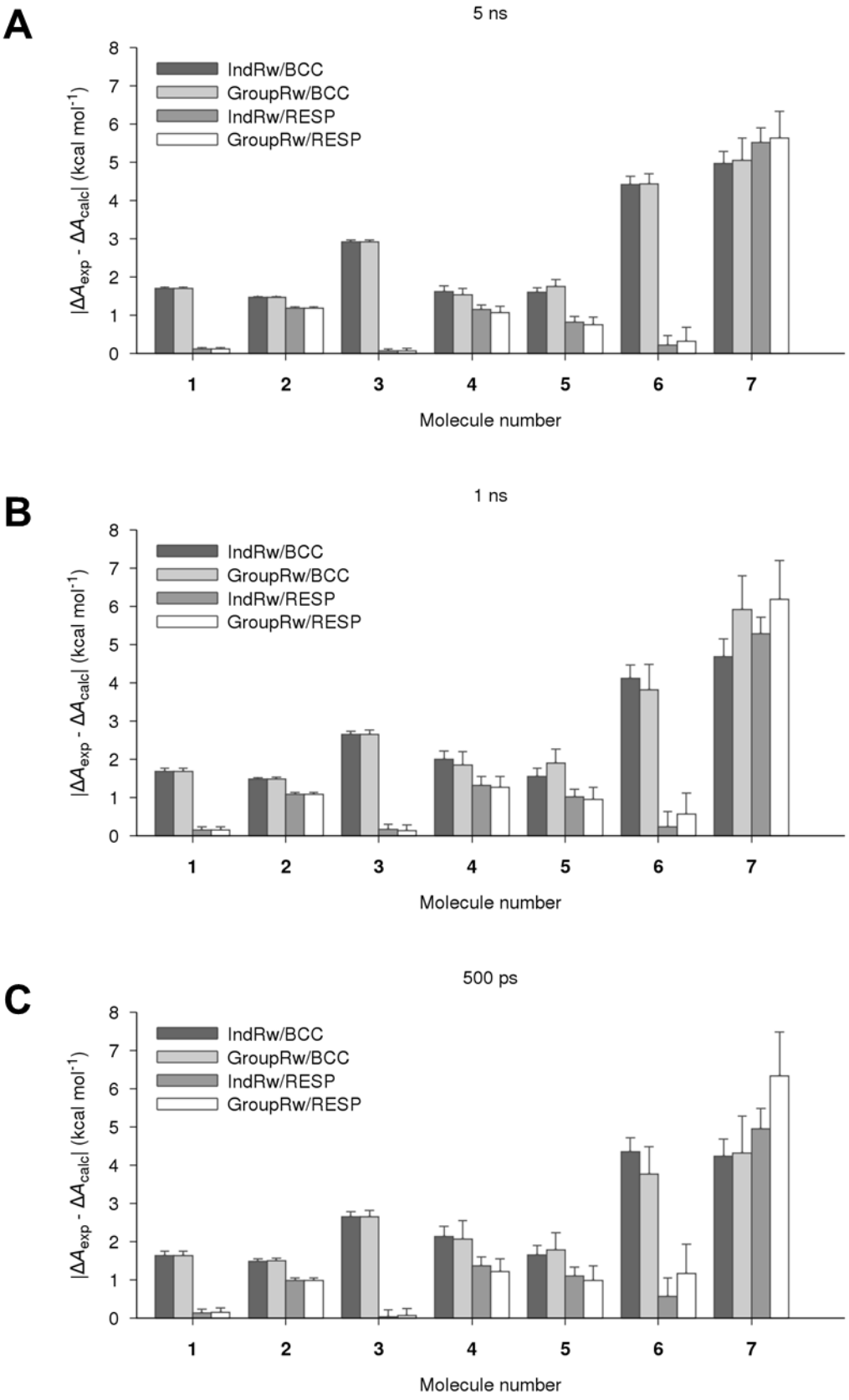

Figure 5.11 Absolute deviation from experiment for both the "independent replica" (IndRw) and "group" (GroupRw) msesTI reweighting schemes

Considering that the free energy estimates recovered via the IndRw have been shown in the previous section (5.3.3.1) to accurately match those of IT-TI, we can conclude that the GroupRw approach inadequately recovers ensemble averages at very short sampling times for complex systems such as malathion. This agrees with previous findings[12] and is likely caused by a lack of convergence across the swarm replicas. It would therefore be preferable to opt for the IndRw method when looking at short simulation times, as attempted here. For longer sampling times, such as the ones used in previous chapters, both approaches eventually converge and as a consequence either could be used. In fact, one could envision the use of both approaches in tandem so as to check if the system has been sufficiently sampled to ensure that the time averages across the swarm replicas have converged.

5.3.3.3 Impact of swarm parameter scaling

As previously detailed in the Methods section, the swarm boost parameters employed in this study were scaled down by a factor of two in order to reduce the noise, and by consequence, the uncertainty associated with high boost potentials. Whilst this was necessary when attempting to recover fine details of the glycosidic surfaces of carbohydrates (Chapter 3), the impact of such a parameter choice when attempting to recover free energy averages which are coarser in detail relative to the torsional surfaces detailed in previous Chapters is unknown. Here we compare the stronger swarm boost parameters ("Full"), with the half-scaled parameters ("Half") which were chosen for this msesTI study.

At the longer simulation lengths, 5 ns and 1 ns per replica, the choice of boost potential does not have much impact on the estimates (Figure 5.12**A**-**B**, Supplementary Table D.7-D.8). Some small changes are seen in the larger two molecules, mannitol (**6**) and malathion (**7**), with the most notable difference being a 0.65 kcal mol^{-1} decrease in the 5 ns per replica free energy estimate of the RESP charged malathion when using the "Full" boost potential (Figure 5.12**A**). As expected, we see an increase in uncertainty when using the higher boost potential, with the standard error being on average 1.9 and 1.5 times larger in the 5 ns and 1 ns per replica estimates respectively.

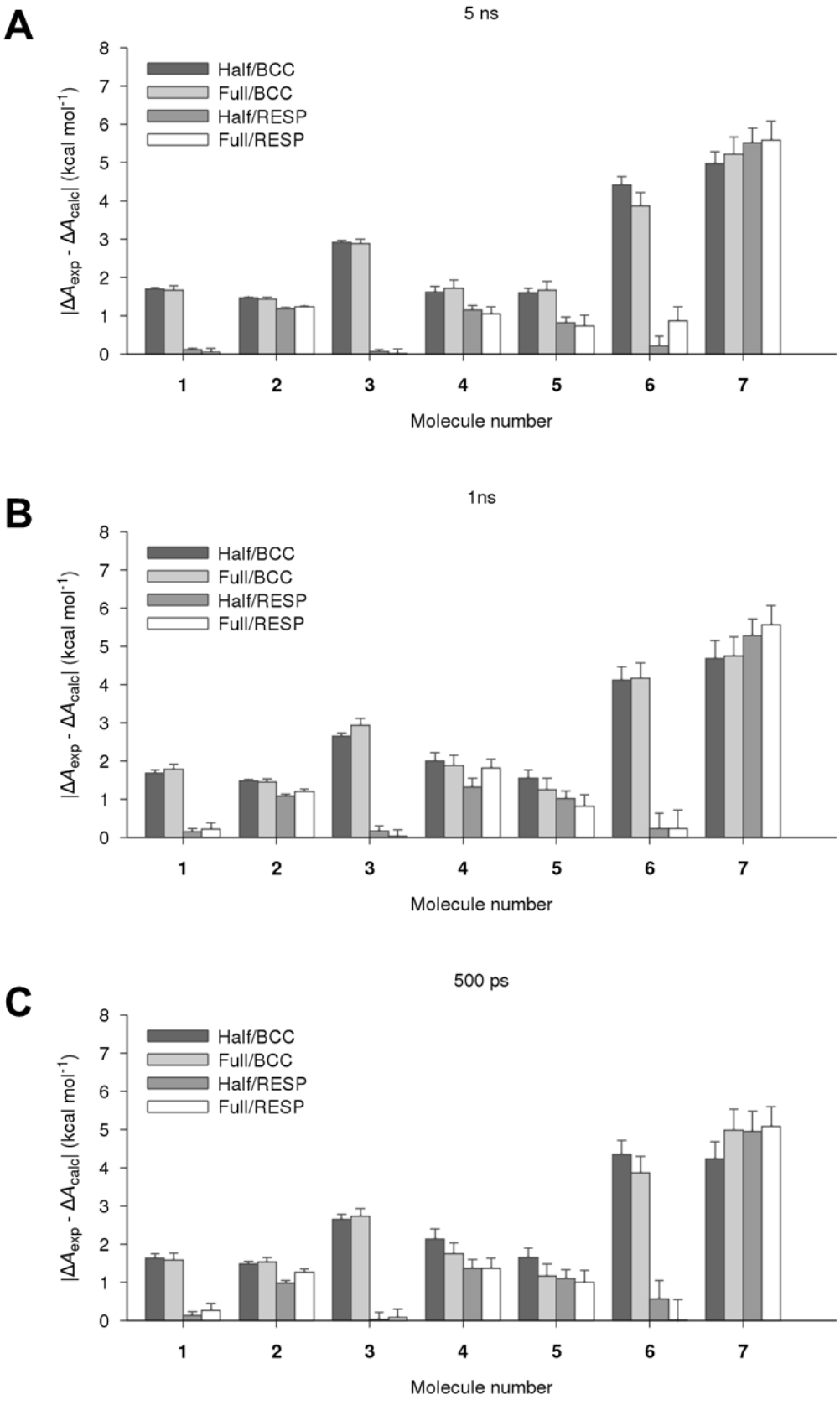

Figure 5.12 Absolute deviation in calculated free energy estimates for the "Half" and "Full" msesTI boost potentials

At the shorter 500 ps per replica timescale, larger deviations between the two boost potentials are seen (Figure 5.12**C**, Supplementary Table D.9). The free energy estimates of the AM1-BCC charged 2-propoxyethanol (**4**), 1-butoxy-2-propanol (**5**) and both charge assignments of mannitol (**6**), show improvements of between 0.4 and 0.5 kcal mol^{-1} relative to experiment when using the higher boost potential. Conversely, the free energy of solvation estimate for the AM1-BCC charged malathion (**7**) increases relative to experiment by 0.76 kcal mol^{-1}. However considering the large uncertainties exhibited by both parameter sets at such low sampling times, significant deviations are expected. Interestingly, the average statistical error from the "Full" parameter set free energy estimates is only 1.3 times larger than the average statistical errors of the "Half" parameter set estimates, which is a smaller difference than what is seen at higher sampling times. In fact, whilst the magnitude of the uncertainties decreases with sampling time, the difference in the values between the two parameter sets increases. The reason for this is unclear, but it may be due to the "Full" parameter set offering improved sampling for each replica which is more appreciable at lower simulation times.

In conclusion, for the two swarm parameter sets investigated here, the strength of the boost potential does not appear to have a large impact in the msesTI results. The increase in uncertainty, though not excessive, should be of more concern when choosing an appropriate set of boost parameters. Ultimately, unless investigating a system with a conformational landscape particularly difficult to explore, it is advisable, as done in this study, to use a weaker potential to reduce the impact of uncertainties.

5.3.4 Evaluating the use of HMR in calculating solvation free energies

Having detailed the use of both IT-TI and msesTI in recovering absolute free energies of solvation, we now turn our attention to the inclusion of hydrogen mass repartitioning (HMR) in order to accelerate sampling. Comparing the influence of HMR on the IT-TI results using 5 ns per replica, both approaches show similar trends in free energy estimates (Figure 13, Supplementary Tables D.1 and D.10). However, the difference in the estimates between the two approaches are in all cases greater than error, ranging from 0.15 kcal mol^{-1} in butan-1-ol (AM1-BCC) to 2.47 kcal mol^{-1} in malathion (RESP). The average difference in the estimate between the normal mass (NORM) and HMR schemes is 0.54 kcal mol^{-1} with a standard deviation of 0.61 kcal mol^{-1} which is much larger than the average 0.15 kcal mol^{-1} difference between the IT-TI and msesTI results using normal masses. Nonetheless, a paired student t-test comparison of the deviations from experiment, both signed and unsigned, does not show the two datasets to be different with probabilities of p = 0.47 and 0.41 respectively. However it should be noted that, due to the relatively small sample size, the robustness of such a statistical test is quite low. It is therefore possible that using a larger dataset may yield different results.

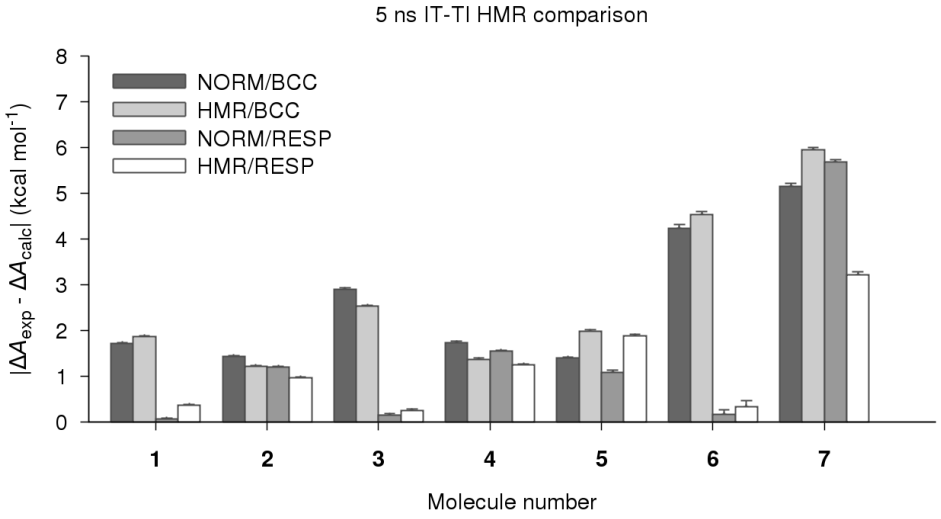

Figure 5.13 Absolute deviation in solvation free energies for standard mass (NORM) and hydrogen mass repartitioning (HMR) IT-TI simulations

Considering the magnitude of the deviations seen in 1-butoxy-2-propanol (**5**) and malathion (**7**), it is worth a deeper look into any possible causes for this. Looking at the difference in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ profiles for the RESP charged simulations (which exhibit the larger differences), a tendency for the values to deviate by a substantial amount at λ values between 0.5 and 0.8 can be seen (Figure 5.14). Although not as pronounced, this trend also occurs in the other systems (Supplementary Figures D.3-D.4). Considering the non-uniformity of the deviations, this indicates that the inclusion of HMR is somehow having an effect on the dynamics at such λ values. Looking at the solute $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ contributions at λ = 0.7, this appears to stem from changes in the van der Waals (VdW) contributions with the use of HMR, leading to an average decrease relative to the normal mass simulations of 6.3 kcal mol^{-1} and 1.5 kcal mol^{-1} for malathion and 1-butoxy-2-propanol respectively. In comparison, the electrostatic contributions increased by 0.47 kcal mol^{-1} and 0.07 kcal mol^{-1} respectively.

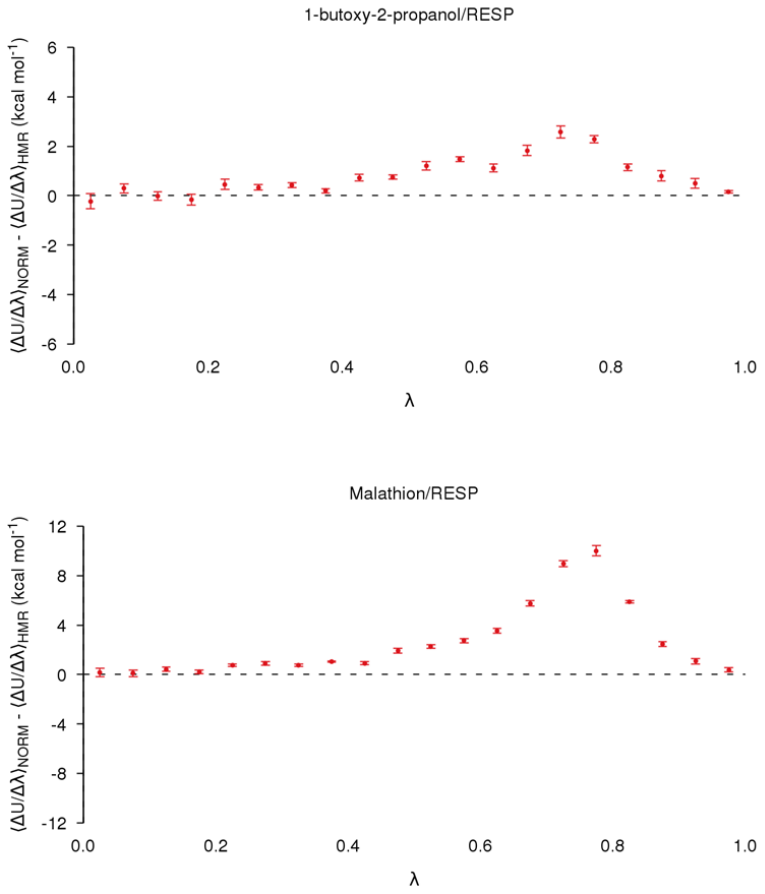

Figure 5.14 Differences in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ profiles between the standard mass (NORM) and hydrogen mass repartitioned (HMR) IT-TI simulations for RESP charged 1-butoxy-2-propanol and malathion

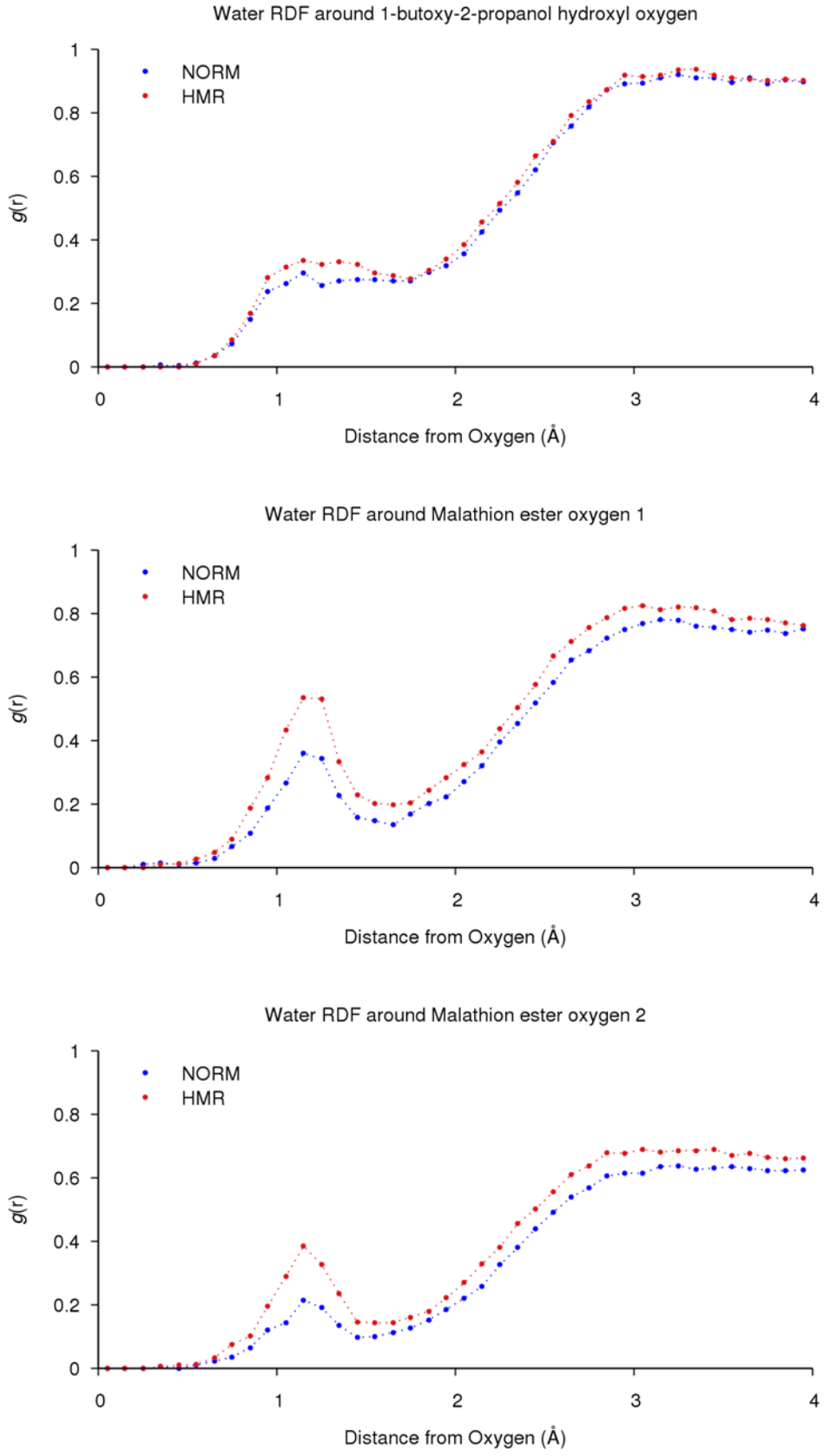

Figure 5.15 Normalised water RDF around the hydrogen bonding oxygens of 1-butoxy-2-propanol and malathion

A possible explanation for this observation is that the increased time step of 4 fs in HMR is allowing soft-core atoms to get closer to the solvent than would normally occur at 2 fs. This then leads to abnormal increases in interatomic interactions. Evidence of this can be seen from the RDF profiles of the water distribution around the hydrogen bond acceptor oxygens in malathion and 1-butoxy-2-propanol (Figure 5.15). The use of HMR leads to an increase in the first peak density, demonstrating that waters more frequently approach the solutes to form hydrogen bonds. Interestingly, in their original description of the HMR method, Feenstra et al.[29] detailed that altering hydrogen masses in SPC water boxes led to slight increases in hydrogen bond lifetimes. Although not appearing to have much influence on the dynamics of normal simulations simulations,[26, 29] it seems that the use of the soft-core potentials employed in this study leads to specific states where such effects are accentuated. It is possible that alternative soft-core scaling parameters may modulate this effect and therefore warrants further study.

The use of HMR in conjunction with msesTI does not appear to have a significant impact on the free energy estimates compared to the HMR IT-TI values (Figure 5.16, Supplementary Table D.11). As per previous msesTI simulations, there is an increase in uncertainty relative to IT-TI, but any deviation in the free energy estimates usually remains within this uncertainty limit. The only exceptions to this are the RESP charged calculations of mannitol (**6**) and malathion (**7**), where the deviation from experiment improves by 0.39 and 0.54 kcal mol^{-1} respectively, albeit with significant increases in the uncertainties. The cause of this appears to be linked to slightly increased fluctuations in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates around the aforementioned λ = 0.5 to 0.8 range, with the uncertainty in the $\langle\frac{\delta U(r,\lambda)}{\delta\lambda}\rangle_\lambda$ estimates for both systems peaking at λ = 0.7 (Figure 5.17). This reinforces the thought that the use of HMR leads to abnormal states at low solute coupling states. The otherwise good agreement indicates that the msesMD method can be used in conjunction with HMR without altering the value of ensemble averages. This could be very useful when attempting to rapidly probe the dynamics of a system, as it would effectively half the simulation time.

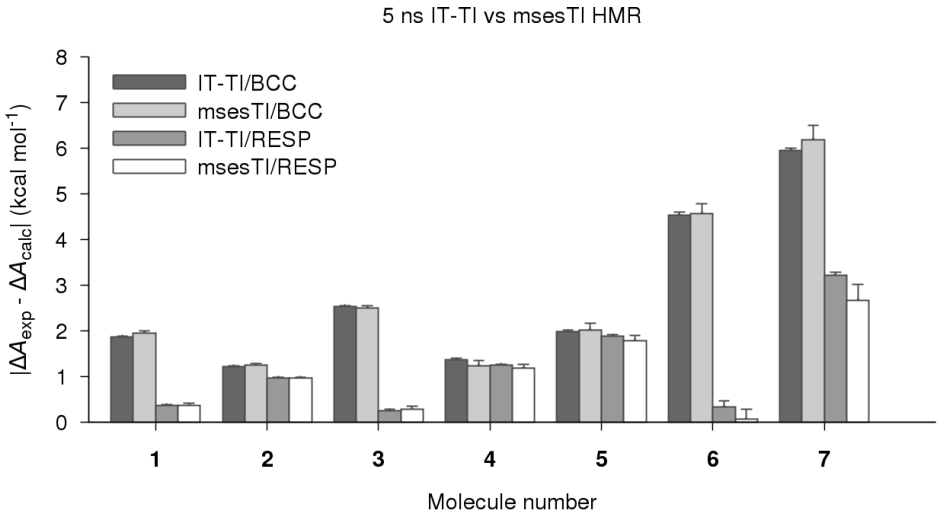

Figure 5.16 Absolute deviations from experiment for the IT-TI and msesTI hydrogen mass repartitioning (HMR) free energy estimates

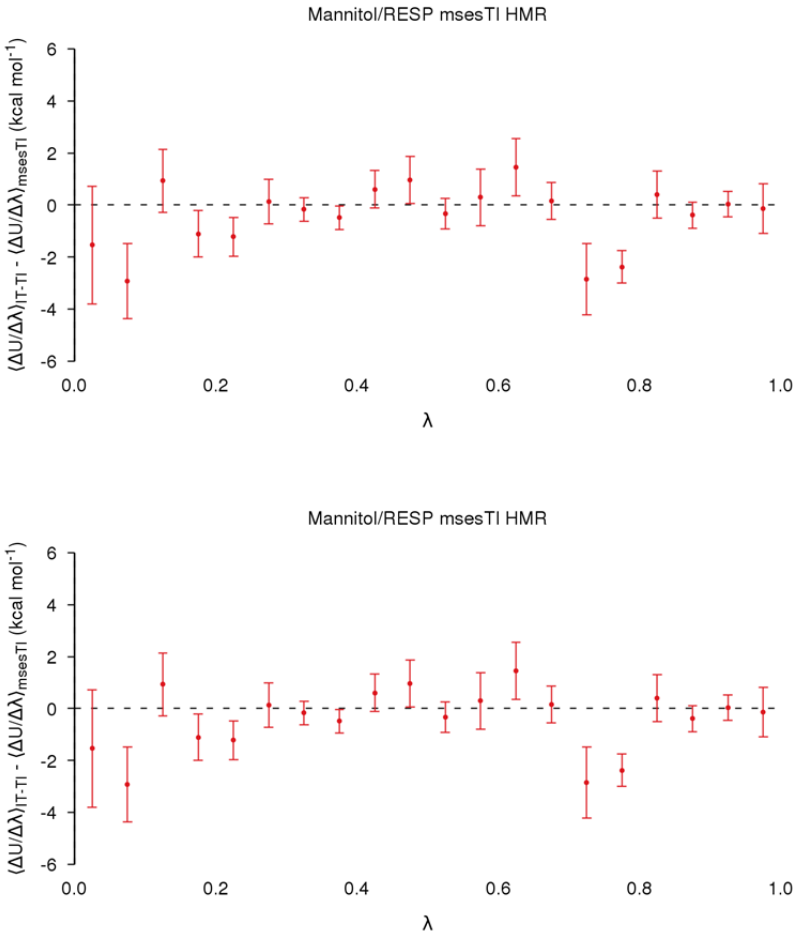

Figure 5.17 Differences in the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_{\lambda}$ profiles between the hydrogen mass repartitioned (HMR) IT-TI and msesTI simulations of RESP charged mannitol and malathion

5.4 Conclusions

In this chapter, the use of IT-TI and msesTI in calculating solvation free energies for small solutes was detailed. The impact of different partial charge assignment schemes and the hydrogen mass repartitioning method was also assessed. The IT-TI method was found to be quite effective in recovering free energy estimates, even at short simulation times. Somewhat large deviations in the estimates were seen relative to the FreeSolv[171] database results, which are likely due differences in the treatment of electrostatics, both in terms of the chosen soft-core potentials and internal MD engine differences. Future tests using two-step perturbation protocols will hopefully allow us to narrow down the exact nature of this issue.

The RESP partial charge assignment protocol employed in this study resulted in inconsistent improvements in the free energies. Whilst some of the simpler polyhydroxylated systems saw large improvements in the estimates, we noted some abnormal polarisation in some of the system. This appears to limit improvements seen by using a higher quality charge assignment method. These abnormal polarisation effects appear to be at least in part due to use of a single conformation charge assignment protocol. Since an msesMD/clustering protocol was able to recover reasonable solvated structures relative to the OmegaTK[186] FreeSolv structures[170-171], further tests of its use as a parameterisation tool are warranted. In the future, we hope to test its use in conjunction with a more complex multi-conformational charge fitting protocol, such as the recently introduced IpolQ framework in AMBER17's *mdgx* module[69, 188].

Whilst the enhanced conformational sampling offered by the msesTI method had little impact on the free energy estimates, it was able to adequately match the IT-TI results. This demonstrates that the methodology can potentially be used to recover free energies in systems, although at the cost of increased uncertainty. Its use may not be justified for the solvation free energy of small solutes, but one would expect potential uses in systems where sampling does have a significant impact on the free energy along an alchemical path, such as for protein-ligand binding scenarios. The choice of reweighting protocol was

also tested and it was demonstrated that the use of the "independent replica" approach, as previously reported[12], was more effective at shorter timescales. Thus, when calculating free energies, it would be more prudent to use this reweighting scheme. As for the choice of parameters, the use of increased boost potential was not found to have a large impact on the estimates, although it did result in an increase in the uncertainties. We therefore conclude that unless sampling limited, the use of a smaller boost potential is more advisable in order to minimise error bars associated with the enhanced sampling method. In the future, we hope to test even smaller boost potentials to see if the uncertainties can be further reduced whilst still benefiting from enhanced sampling.

Finally, the use of HMR in the context of free energy simulations was evaluated. It was found that whilst the free energy estimates were not significantly different from the standard mass calculations, in some cases large deviations in the estimates could be seen. Closer inspection at those cases indicate that the inclusion of HMR within a soft-core transformation leads to cases where interatomic distances are reduced relative to standard mass simulations. Ultimately leading to increased nonbonded interactions, such as the formation of hydrogen bonds. It is possible that this may be a result of the specific alchemical soft-core protocol employed here, therefore further work will be required to test as to how much of an impact this has on alternate protocols e.g. two-step transformations and different soft-core scaling parameters. Nonetheless, this does serve as a warning against the black box use of HMR within the context of free energy calculations.

Chapter 6: Concluding remarks

The main objective of this research was the development and application of an efficient high performance swarm-enhanced sampling MD (sesMD) methodology for use in investigating the dynamics of biomolecular systems of interest.

Firstly, targetting key limitations in the original sesMD methodology, particularly with regards to parameter transferability, the development of a new generalised swarm-enhanced sampling potential, termed multi-dimensional swarm-enhanced sampling molecular dynamics (msesMD) is detailed. Comparing against both the original sesMD method and unbiased MD simulations, we find that, using parameters developed for alanine dipeptide, the msesMD methodology was able to sample rare conformational regions in the backbone φψ maps of alanine heptapeptide, which neither of the other two methods were able to find. This indicates that the msesMD method is more amenable than sesMD to the transferability of swarm parameters between systems of interest. As demonstrated in later chapters, msesMD parameters developed for the Lewis^a trisaccharide were effective in exploring other systems of interest, with only intuitive scaling in the A and C terms necessary to reduce the reweighting noise associated with larger swarm boost potential energies. Due to this, the msesMD method was chosen as the enhanced sampling method of choice throughout the rest of this thesis. However, without exhaustive testing, we note that the exact extent of the particular cases in which msesMD is preferable to use over sesMD is still unknown; we expect that in some cases, particularly when observing highly concerted motion across a set of dihedrals, that a well parameterised sesMD simulation may prove more useful.

Attempts were then made at optimising the compute efficiency of sesMD and msesMD algorithms within both the *sander* and *pmemd* MD engines of the AMBER software suite. Through the appropriate use of new asynchronous MPI routines, the relatively high cost of inter-replica communication was efficiently hidden by overlapping it with equally time consuming non-bonded force calculations. For the sesMD routines in *sander*, this led to a 1.3x improvement over the original sesMD implementation using blocking MPI routines.

Additionally, the msesMD code was implemented in *pmemd* using similar MPI optimisations as those used in the *sander* sesMD code. This change in MD engine resulted in up to a 2.3x improvement in simulation speed over the optimised *sander* sesMD code. Through these optimisations, simulations of several hundreds of nanoseconds per replica can now be readily computed for oligosaccharide-sized systems, something which was not possible using the original sesMD simulations. As shown in later chapters, this proved particularly useful as simulations of over a hundred nanoseconds per replica were required to adequately profile rare conformational changes within a reasonable level of accuracy.

Having developed a fast and efficient msesMD method, it was then applied to the investigation of rare conformational changes in Lewis oligosaccharides. Comparing against triplicate 10 μs unbiased MD simulations, 245 ns per replica msesMD simulations were able to accurately reproduce the key topological features of the free energy surfaces associated with rotation around the glycosidic torsions. Both sets of simulations were able to identify rare transitions away from canonically closed conformations to a variety of different open forms. Using trajectory clustering, it was possible to extract conformational states from the msesMD simulations which closely matched previously observed open crystallographic protein-bound poses of Le^x and sLe^x. The effectiveness of the msesMD method was also compared against other enhanced sampling methods, namely umbrella sampling and accelerated molecular dynamics. In both cases msesMD was found to be either on a par or more effective than the other enhanced sampling approach, not suffering from the sampling limitations of one-dimensional umbrella sampling nor noisy reweighted surfaces due to the large boost energies as found for the aMD approach.

From the Lewis oligosaccharide msesMD simulations, one of the limitations which was identified was that purely boosting glycosidic torsions did not always lead to efficient sampling of other slow coordinates, specifically ring puckering. Therefore, we then attempted to use the msesMD potential to directly boost the internal torsions of pyranose rings in order to sample different ring puckers. Evaluating this on four benchmark monosaccharides and comparing against 10 μs unbiased MD simulations, it was found that msesMD can readily profile puckering surfaces, something not as easily obtained via unbiased MD simulations. In fact, for two of the four benchmark cases, α-D-glucose and β-

D-glucuronic acid, the unbiased MD simulations needed to be lengthened by an additional 5 μs due to inadequate sampling. This challenges previous estimates that ring equilibrium is obtained at timescales of around 3 μs in unbiased MD simulations[81] and demonstrates the advantage of using enhanced sampling methods to observe puckering events. After validating the use of msesMD in observing ring deformation events, the method was used to explore the impact of post-translational ring substitutions on the dynamics of glycosaminoglycan monosaccharides, namely iduronic acid, glucuronic acid, N-acetyl-glucosamine and N-acetyl-galactosamine. It was found that in some cases, the introduction of sulfation at specific positions in the rings can lead to significant changes in ring behaviour. For example the introduction of 4-O-sulfation in N-acetyl-galactosamine leads to significant increases in the stability of the $^{1}C_{4}$ pucker.

Finally, the msesMD method was combined with a thermodynamic integration scheme, termed msesTI, to evaluate the solvation free energies of seven small molecule systems. Comparing against independent trajectory TI simulations (IT-TI), it was found in all cases that although the correct solvation free energies were obtained, there was no advantage in using the msesTI method over IT-TI. In fact, the use of msesTI generally led to increases in the uncertainty of the free energy estimates. This lack of improvement in estimates reflects the fact that for small ligands, conformational change can usually be sampled using short unbiased MD simulations. We hope the msesTI method may perform better in more complex systems where conformational behaviour has a larger impact on the free energies of the system, for example in protein-ligand binding scenarios.

Overall, the msesMD method shows potential as a intuitive tool to probe the conformational flexibility of small-to-medium-sized biomolecules. Although at times requiring slight alterations in the swarm potential parameters to reduce the noise associated with large swarm boost energies, the application of msesMD to new systems of interest where the choice of dihedral coordinates to enhance are obvious, has been relatively straightforward. However, the method has yet to be tested in large complex systems such as proteins, where selecting suitable boost coordinates is a more difficult prospect. One would expect that this would present a far more challenging task. Furthermore, whilst msesMD was shown in chapter 2 to sometimes be more effective than other enhanced

sampling methods, the question of why should one use msesMD over the myriad of other enhanced sampling methods still remains. It is likely that in most cases, the appropriate choice of enhanced sampling method is case specific. Nevertheless, future work would seek to address this point through the use of comprehensive benchmarks to categorise the specific use cases of different enhanced sampling methods.

References

1. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012,** *8* (5), 1542-1555.

2. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **2013,** *9* (9), 3878-3888.

3. Pan, A. C.; Weinreich, T. M.; Shan, Y.; Scarpazza, D. P.; Shaw, D. E., Assessing the Accuracy of Two Enhanced Sampling Methods Using EGFR Kinase Transition Pathways: The Influence of Collective Variable Choice. *J Chem Theory Comput* **2014,** *10* (7), 2860-2865.

4. Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. B.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C., Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE Press: New Orleans, Louisana, 2014; pp 41-53.

5. Larson, S. M.; Snow, C. D.; Shirts, M., Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology. **2002**.

6. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* **2004,** *120* (24), 11919-11929.

7. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002,** *99* (20), 12562-12566.

8. Swendsen, R. H.; Wang, J.-S., Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters* **1986,** *57* (21), 2607-2609.

9. Jang, S.; Shin, S.; Pak, Y., Replica-Exchange Method Using the Generalized Effective Potential. *Physical Review Letters* **2003,** *91* (5), 058305.

10. Torrie, G. M.; Valleau, J. P., Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem Phys Lett* **1974,** *28* (4), 578-581.

11. Atzori, A.; Bruce, N. J.; Burusco, K. K.; Wroblowski, B.; Bonnet, P.; Bryce, R. A., Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics. *J Chem Inf Model* **2014,** *54* (10), 2764-2775.

12. Burusco, K. K.; Bruce, N. J.; Alibay, I.; Bryce, R. A., Free Energy Calculations using a Swarm-Enhanced Sampling Molecular Dynamics Approach. *ChemPhysChem* **2015,** *16* (15), 3233-3241.

13. Leach, A. R., *Molecular modelling : principles and applications*. 2nd ed.; Prentice Hall: Harlow, England ; New York, 2001; p xxiv, 744 p., 16 p. of plates.

14. Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J., Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985,** *107* (13), 3902-3909.

15. Stewart, J. J. P., Optimization of parameters for semiempirical methods I. Method. *J Comput Chem* **1989,** *10* (2), 209-220.

16. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.

17. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-1174.

18. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **2010,** *31* (4), 671-690.

19. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J Chem Theory Comput* **2013,** *9* (12), 5430-5449.

20. Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B* **2010,** *114* (8), 2549-2564.

21. Ren, P.; Ponder, J. W., Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J Comput Chem* **2002,** *23* (16), 1497-1506.

22. Ren, P.; Ponder, J. W., Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **2003,** *107* (24), 5933-5947.

23. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of physics. Condensed matter : an Institute of Physics journal* **2009,** *21* (33), 333102-333102.

24. Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T., The ReaxFF reactive force-field: development, applications and future directions. **2016,** *2*, 15011.

25. Allen, M. P.; Tildesley, D. J., *Computer simulation of liquids*. Clarendon Press ;

Oxford University Press: Oxford England

New York, 1987; p xix, 385 p.

26. Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E., Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2015,** *11* (4), 1864-1874.

27. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **1977,** *23* (3), 327-341.

28. Tuckerman, M.; Berne, B. J.; Martyna, G. J., Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics* **1992,** *97* (3), 1990-2001.

29. Feenstra, K. A.; Hess, B.; Berendsen, H. J. C., Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* **1999,** *20* (8), 786-798.

30. Sattelle, B. M.; Almond, A., Shaping up for structural glycomics: a predictive protocol for oligosaccharide conformational analysis applied to N-linked glycans. *Carbohydrate Research* **2014,** *383*, 34-42.

31. Sattelle, B. M.; Almond, A., Is N-acetyl-d-glucosamine a rigid 4C1 chair? *Glycobiology* **2011,** *21* (12), 1651-1662.

32. Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham Iii, T.; Darden, T.; Duke, R.; Gohlke, H., Amber 14. **2014**.

33. Miyamoto, S.; Kollman, P. A., Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* **1992,** *13* (8), 952-962.

34. Tolman, R. C., The principles of statistical mechanics. **1979**.

35. Hamelberg, D.; Oliveira, C. A. F. d.; McCammon, J. A., Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *The Journal of Chemical Physics* **2007,** *127* (15), 155102.

36. Wereszczynski, J.; McCammon, J. A., Using Selectively Applied Accelerated Molecular Dynamics to Enhance Free Energy Calculations. *J Chem Theory Comput* **2010,** *6* (11), 3285-3292.

37. Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A., Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J Chem Theory Comput* **2014,** *10* (7), 2677-2689.

38. Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977,** *23* (2), 187-199.

39. Shaffer, P.; Valsson, O.; Parrinello, M., Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proceedings of the National Academy of Sciences of the United States of America* **2016,** *113* (5), 1150-1155.

40. Yang, M.; Huang, J.; MacKerell, A. D., Enhanced Conformational Sampling Using Replica Exchange with Concurrent Solute Scaling and Hamiltonian Biasing Realized in One Dimension. *J Chem Theory Comput* **2015,** *11* (6), 2855-2867.

41. Miao, Y.; Feher, V. A.; McCammon, J. A., Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J Chem Theory Comput* **2015,** *11* (8), 3584-3595.

42. Kästner, J., Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011,** *1* (6), 932-942.

43. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* **1992,** *13* (8), 1011-1021.

44. Souaille, M.; Roux, B. t., Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications* **2001,** *135* (1), 40-57.

45. Beni, G.; Wang, J., Swarm Intelligence in Cellular Robotic Systems. In *Robots and Biological Systems: Towards a New Bionics?*, Dario, P.; Sandini, G.; Aebischer, P., Eds. Springer Berlin Heidelberg: 1993; Vol. 102, pp 703-712.

46. Namasivayam, V.; Günther, R., PSO@ AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. *Chemical biology & drug design* **2007,** *70* (6), 475-484.

47. Chen, K.; Li, T.; Cao, T., Tribe-PSO: A novel global optimization algorithm and its application in molecular docking. *Chemometrics and intelligent laboratory systems* **2006,** *82* (1), 248-259.

48. Liu, Y.; Zhao, L.; Li, W.; Zhao, D.; Song, M.; Yang, Y., FIPSDock: A new molecular docking technique driven by fully informed swarm optimization algorithm. *J Comput Chem* **2013,** *34* (1), 67-75.

49. Liu, J.; Wang, L.; He, L.; Shi, F., Analysis of Toy Model for Protein Folding Based on Particle Swarm Optimization Algorithm. In *Advances in Natural Computation*, Wang, L.; Chen, K.; Ong, Y., Eds. Springer Berlin Heidelberg: 2005; Vol. 3612, pp 636-645.

50. Xiaolong, Z.; Tingting, L. In *Improved Particle Swarm Optimization Algorithm for 2D Protein Folding Prediction*, Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, 6-8 July 2007; 2007; pp 53-56.

51. Ying-yin, L.; Ying-ping, C. In *Crowd control with swarm intelligence*, Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 25-28 Sept. 2007; 2007; pp 3321-3328.

52. An-Pin, C.; Chien-Hsun, H.; Yu-Chia, H. In *A novel modified particle swarm optimization for forecasting financial time series*, Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on, 20-22 Nov. 2009; 2009; pp 683-687.

53. Huber, T.; van Gunsteren, W. F., SWARM-MD: Searching conformational space by cooperative molecular dynamics. *J Phys Chem A* **1998,** *102* (29), 5937-5943.

54. Bruce, N. J.; Bryce, R. A., Ab Initio Protein Folding Using a Cooperative Swarm of Molecular Dynamics Trajectories. *J Chem Theory Comput* **2010,** *6* (7), 1925-1930.

55. Pontius, R. G.; Thontteh, O.; Chen, H., Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* **2008,** *15* (2), 111-142.

56. Cheng, X.; Cui, G.; Hornak, V.; Simmerling, C., Modified Replica Exchange Simulation Methods for Local Structure Refinement. *The Journal of Physical Chemistry B* **2005,** *109* (16), 8220-8230.

57. Chipot, C.; Lelièvre, T., Enhanced Sampling of Multidimensional Free-Energy Landscapes Using Adaptive Biasing Forces. *SIAM Journal on Applied Mathematics* **2011,** *71* (5), 1673-1695.

58. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984,** *81* (8), 3684-3690.

59. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-3095.

60. Case, D.; Darden, T.; Cheatham III, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Merz, K.; Wang, B.; Pearlman, D., AMBER 8. *University of California, San Francisco* **2004,** *5*, 39.

61. Case, D. A.; Darden, T.; Cheatham, T.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R.; Zhang, W.; Merz, K. *Amber 11*; University of California: 2010.

62. Walker, D. W.; Dongarra, J. J., MPI: a standard message passing interface. *Supercomputer* **1996,** *12*, 56-68.

63. Liu, J.; Chandrasekaran, B.; Wu, J.; Jiang, W.; Kini, S.; Yu, W.; Buntinas, D.; Wyckoff, P.; Panda, D. K. In *Performance comparison of MPI implementations over InfiniBand, Myrinet and Quadrics*, Proceedings of the 2003 ACM/IEEE conference on Supercomputing, ACM: 2003; p 58.

64. Forum, M., A Message-Passing Interface Standard. Version 3.0. **2012**.

65. Malevanets, A.; Wodak, Shoshana J., Multiple Replica Repulsion Technique for Efficient Conformational Sampling of Biological Systems. *Biophys J* **2011,** *101* (4), 951-960.

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

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

68. Grossfield, A., WHAM: the weighted histogram analysis method. *Disponıvel em:< http://membrane. urmc. rochester. edu/content/wham* **2012**.

69. Case, D.; Cerutti, D.; Cheateham, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Greene, D.; Homeyer, N., AMBER16 Package. **2016**.

70. de la Fuente, J. M.; Penadés, S., Glyconanoparticles: Types, synthesis and applications in glycoscience, biomedicine and material science. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2006,** *1760* (4), 636-651.

71. Woods, R. J.; Tessier, M. B., Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan–protein complexes. *Current Opinion in Structural Biology* **2010,** *20* (5), 575-583.

72. Frank, M.; Collins, P.; Peak, I.; Grice, I.; Wilson, J., An Unusual Carbohydrate Conformation is Evident in Moraxella catarrhalis Oligosaccharides. *Molecules* **2015,** *20* (8), 14234.

73. Blaum, B. S.; Frank, M.; Walker, R. C.; Neu, U.; Stehle, T., Complement Factor H and Simian Virus 40 bind the GM1 ganglioside in distinct conformations. *Glycobiology* **2016,** *26* (5), 532-539.

74. Yuriev, E.; Farrugia, W.; Scott, A. M.; Ramsland, P. A., Three-dimensional structures of carbohydrate determinants of Lewis system antigens: Implications for effective antibody targeting of cancer. *Immunol Cell Biol* **2005,** *83* (6), 709-717.

75. Lorant, D. E.; Topham, M. K.; Whatley, R. E.; McEver, R. P.; McIntyre, T. M.; Prescott, S. M.; Zimmerman, G. A., Inflammatory roles of P-selectin. *Journal of Clinical Investigation* **1993,** *92* (2), 559-570.

76. Sugasaki, A.; Sugiyasu, K.; Ikeda, M.; Takeuchi, M.; Shinkai, S., First Successful Molecular Design of an Artificial Lewis Oligosaccharide Binding System Utilizing Positive Homotropic Allosterism. *Journal of the American Chemical Society* **2001,** *123* (42), 10239-10244.

77. Lemieux, R. U.; Bock, K.; Delbaere, L. T. J.; Koto, S.; Rao, V. S., The conformations of oligosaccharides related to the ABH and Lewis human blood group determinants. *Canadian Journal of Chemistry* **1980,** *58* (6), 631-653.

78. Mishra, S. K.; Kara, M.; Zacharias, M.; Koča, J., Enhanced conformational sampling of carbohydrates by Hamiltonian replica-exchange simulation. *Glycobiology* **2014,** *24* (1), 70-84.

79. Haselhorst, T.; Weimar, T.; Peters, T., Molecular Recognition of Sialyl Lewisx and Related Saccharides by Two Lectins. *Journal of the American Chemical Society* **2001,** *123* (43), 10705-10714.

80. Topin, J.; Lelimousin, M.; Arnaud, J.; Audfray, A.; Pérez, S.; Varrot, A.; Imberty, A., The Hidden Conformation of Lewis x, a Human Histo-Blood Group Antigen, Is a Determinant for Recognition by Pathogen Lectins. *ACS Chemical Biology* **2016,** *11* (7), 2011-2020.

81. Sattelle, B. M.; Hansen, S. U.; Gardiner, J.; Almond, A., Free Energy Landscapes of Iduronic Acid and Related Monosaccharides. *Journal of the American Chemical Society* **2010,** *132* (38), 13132-13134.

82. Sattelle, B. M.; Shakeri, J.; Almond, A., Does Microsecond Sugar Ring Flexing Encode 3D-Shape and Bioactivity in the Heparanome? *Biomacromolecules* **2013,** *14* (4), 1149-1159.

83. Harvey, M. J.; Giupponi, G.; Fabritiis, G. D., ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J Chem Theory Comput* **2009,** *5* (6), 1632-1639.

84. Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S., Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J Comput Chem* **2009,** *30* (6), 864-872.

85. Hansen, H. S.; Hünenberger, P. H., Using the local elevation method to construct optimized umbrella sampling potentials: Calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water. *J Comput Chem* **2010,** *31* (1), 1-23.

86. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **1999,** *314* (1–2), 141-151.

87. Hansmann, U. H. E., Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* **1997,** *281* (1–3), 140-150.

88. Islam, S. M.; Richards, M. R.; Taha, H. A.; Byrns, S. C.; Lowary, T. L.; Roy, P.-N., Conformational Analysis of Oligoarabinofuranosides: Overcoming Torsional Barriers with Umbrella Sampling. *J Chem Theory Comput* **2011,** *7* (9), 2989-3000.

89. Perić-Hassler, L.; Hansen, H. S.; Baron, R.; Hünenberger, P. H., Conformational properties of glucose-based disaccharides investigated using molecular dynamics

simulations with local elevation umbrella sampling. *Carbohydrate Research* **2010,** *345* (12), 1781-1801.

90. Spiwok, V.; Tvaroška, I., Conformational Free Energy Surface of α-N-Acetylneuraminic Acid: An Interplay Between Hydrogen Bonding and Solvation. *The Journal of Physical Chemistry B* **2009,** *113* (28), 9589-9594.

91. Yang, M.; MacKerell, A. D., Conformational Sampling of Oligosaccharides Using Hamiltonian Replica Exchange with Two-Dimensional Dihedral Biasing Potentials and the Weighted Histogram Analysis Method (WHAM). *J Chem Theory Comput* **2015,** *11* (2), 788-799.

92. Mallajosyula, S. S.; MacKerell, A. D., Influence of Solvent and Intramolecular Hydrogen Bonding on the Conformational Properties of O-Linked Glycopeptides. *The Journal of Physical Chemistry B* **2011,** *115* (38), 11215-11229.

93. Plazinski, W.; Drach, M., The influence of the hexopyranose ring geometry on the conformation of glycosidic linkages investigated using molecular dynamics simulations. *Carbohydrate Research* **2015,** *415*, 17-27.

94. Biarnés, X.; Ardèvol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M., The Conformational Free Energy Landscape of β-d-Glucopyranose. Implications for Substrate Preactivation in β-Glucoside Hydrolases. *Journal of the American Chemical Society* **2007,** *129* (35), 10686-10693.

95. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J Comput Chem* **2008,** *29* (4), 622-655.

96. Ester, M.; Kriegel, H.-P.; #246; Sander, r.; Xu, X., A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press: Portland, Oregon, 1996; pp 226-231.

97. Das, S. K.; Mallet, J. M.; Esnault, J.; Driguez, P. A.; Duchaussoy, P.; Sizun, P.; Herault, J. P.; Herbert, J. M.; Petitou, M.; Sinay, P., Synthesis of conformationally locked L-iduronic acid derivatives: direct evidence for a critical role of the skew-boat 2S0 conformer in the activation of antithrombin by heparin. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2001,** *7* (22), 4821-34.

98. Speciale, G.; Thompson, A. J.; Davies, G. J.; Williams, S. J., Dissecting conformational contributions to glycosidase catalysis and inhibition. *Current Opinion in Structural Biology* **2014,** *28*, 1-13.

99. Plazinski, W.; Drach, M., Kinetic characteristics of conformational changes in the hexopyranose rings. *Carbohydrate Research* **2015,** *416*, 41-50.

100. Hsieh, P.-H.; Thieker, D. F.; Guerrini, M.; Woods, R. J.; Liu, J., Uncovering the Relationship between Sulphation Patterns and Conformation of Iduronic Acid in Heparan Sulphate. **2016,** *6*, 29602.

101. Sattelle, B. M.; Shakeri, J.; Roberts, I. S.; Almond, A., A 3D-structural model of unsulfated chondroitin from high-field NMR: 4-sulfation has little effect on backbone conformation. *Carbohydrate Research* **2010,** *345* (2), 291-302.

102. Autieri, E.; Sega, M.; Pederiva, F.; Guella, G., Puckering free energy of pyranoses: A NMR and metadynamics-umbrella sampling investigation. *The Journal of Chemical Physics* **2010,** *133* (9), 095104.

103. Qian, X.; Liu, D., Free energy landscape for glucose condensation and dehydration reactions in dimethyl sulfoxide and the effects of solvent. *Carbohydrate Research* **2014,** *388*, 50-60.

104. Petersen, L.; Ardèvol, A.; Rovira, C.; Reilly, P. J., Mechanism of Cellulose Hydrolysis by Inverting GH8 Endoglucanases: A QM/MM Metadynamics Study. *The Journal of Physical Chemistry B* **2009,** *113* (20), 7331-7339.

105. Babin, V.; Sagui, C., Conformational free energies of methyl-α-L-iduronic and methyl-β-D-glucuronic acids in water. *The Journal of Chemical Physics* **2010,** *132* (10), 104108.

106. Naidoo, K. J., FEARCF a multidimensional free energy method for investigating conformational landscapes and chemical reaction mechanisms. *Science China Chemistry* **2011,** *54* (12), 1962-1973.

107. Samsonov, S. A.; Theisgen, S.; Riemer, T.; Huster, D.; Pisabarro, M. T., Glycosaminoglycan Monosaccharide Blocks Analysis by Quantum Mechanics, Molecular Dynamics, and Nuclear Magnetic Resonance. *BioMed Research International* **2014,** *2014*, 11.

108. Hill, A. D.; Reilly, P. J., Puckering Coordinates of Monocyclic Rings by Triangular Decomposition. *J Chem Inf Model* **2007,** *47* (3), 1031-1035.

109. Cremer, D.; Pople, J. A., General definition of ring puckering coordinates. *Journal of the American Chemical Society* **1975,** *97* (6), 1354-1358.

110. Sega, M.; Autieri, E.; Pederiva, F., On the calculation of puckering free energy surfaces. *The Journal of Chemical Physics* **2009,** *130* (22), 225102.

111. Singh, A.; Tessier, M. B.; Pederson, K.; Wang, X.; Venot, A. P.; Boons, G.-J.; Prestegard, J. H.; Woods, R. J., Extension and validation of the GLYCAM force field parameters for modeling glycosaminoglycans. *Canadian Journal of Chemistry* **2016,** *94* (11), 927-935.

112. de Souza, O. N.; Ornstein, R. L., Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh Ewald method. *Biophys J* **1997,** *72* (6), 2395-2397.

113. Plazinski, W.; Drach, M., The dynamics of the conformational changes in the hexopyranose ring: a transition path sampling approach. *RSC Advances* **2014,** *4* (48), 25028-25039.

114. Meirovitch, H., Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Current Opinion in Structural Biology* **2007,** *17* (2), 181-186.

115. Grossfield, A.; Zuckerman, D. M., Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual reports in computational chemistry* **2009,** *5*, 23-48.

116. Rao, V. R., *Conformation of carbohydrates*. CRC Press: 1998.

117. Spiwok, V.; Králová, B.; Tvaroška, I., Modelling of β-d-glucopyranose ring distortion in different force fields: a metadynamics study. *Carbohydrate Research* **2010,** *345* (4), 530-537.

118. Barnett, C. B.; Naidoo, K. J., Ring puckering: A metric for evaluating the accuracy of AM1, PM3, PM3CARB-1, and SCC-DFTB carbohydrate QM/MM simulations. *The Journal of Physical Chemistry B* **2010,** *114* (51), 17142-17154.

119. Asimakopoulou, A. P.; Theocharis, A. D.; Tzanakakis, G. N.; Karamanos, N. K., The biological role of chondroitin sulfate in cancer and chondroitin-based anticancer agents. *In vivo (Athens, Greece)* **2008,** *22* (3), 385-9.

120. Senzolo, M.; Coppell, J.; Cholongitas, E.; Riddell, A.; Triantos, C. K.; Perry, D.; Burroughs, A. K., The effects of glycosaminoglycans on coagulation: a thromboelastographic study. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis* **2007,** *18* (3), 227-36.

121. Liu, J.; Pedersen, L. C., Anticoagulant Heparan Sulfate: Structural Specificity and Biosynthesis. *Applied microbiology and biotechnology* **2007,** *74* (2), 263-272.

122. Hamad, O. A.; Ekdahl, K. N.; Nilsson, P. H.; Andersson, J.; Magotti, P.; Lambris, J. D.; Nilsson, B., Complement activation triggered by chondroitin sulfate released by

thrombin receptor-activated platelets. *Journal of thrombosis and haemostasis : JTH* **2008,** *6* (8), 1413-1421.

123. Goulas, A.; Papakonstantinou, E.; Karakiulakis, G.; Mirtsou-Fidani, V.; Kalinderis, A.; Hatzichristou, D. G., Tissue structure-specific distribution of glycosaminoglycans in the human penis. *The International Journal of Biochemistry & Cell Biology* **2000,** *32* (9), 975-982.

124. Caterson, B.; Mahmoodian, F.; Sorrell, J. M.; Hardingham, T. E.; Bayliss, M. T.; Carney, S. L.; Ratcliffe, A.; Muir, H., Modulation of native chondroitin sulphate structure in tissue development and in disease. *Journal of Cell Science* **1990,** *97* (3), 411-417.

125. Jaques, L. B.; McDuffie, N. M., The chemical and anticoagulant nature of heparin. *Seminars in thrombosis and hemostasis* **1978,** *4* (4), 277-97.

126. Lever, R.; Page, C. P., Novel drug development opportunities for heparin. *Nat Rev Drug Discov* **2002,** *1* (2), 140-148.

127. Gama, C. I.; Tully, S. E.; Sotogaku, N.; Clark, P. M.; Rawat, M.; Vaidehi, N.; Goddard, W. A.; Nishi, A.; Hsieh-Wilson, L. C., Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. *Nat Chem Biol* **2006,** *2* (9), 467-473.

128. Esko, J. D.; Lindahl, U., Molecular diversity of heparan sulfate. *Journal of Clinical Investigation* **2001,** *108* (2), 169-173.

129. Ferreras, C.; Rushton, G.; Cole, C. L.; Babur, M.; Telfer, B. A.; van Kuppevelt, T. H.; Gardiner, J. M.; Williams, K. J.; Jayson, G. C.; Avizienyte, E., Endothelial heparan sulfate 6-O-sulfation levels regulate angiogenic responses of endothelial cells to fibroblast growth factor 2 and vascular endothelial growth factor. *The Journal of biological chemistry* **2012,** *287* (43), 36132-46.

130. Knelson, E. H.; Nee, J. C.; Blobe, G. C., Heparan sulfate signaling in cancer. *Trends in biochemical sciences* **2014,** *39* (6), 277-288.

131. Trowbridge, J. M.; Gallo, R. L., Dermatan sulfate: new functions from an old glycosaminoglycan. *Glycobiology* **2002,** *12* (9), 117R-125R.

132. Malavaki, C.; Mizumoto, S.; Karamanos, N.; Sugahara, K., Recent Advances in the Structural Study of Functional Chondroitin Sulfate and Dermatan Sulfate in Health and Disease. *Connective Tissue Research* **2008,** *49* (3-4), 133-139.

133. Afratis, N.; Gialeli, C.; Nikitovic, D.; Tsegenidis, T.; Karousou, E.; Theocharis, A. D.; Pavão, M. S.; Tzanakakis, G. N.; Karamanos, N. K., Glycosaminoglycans: key players in cancer cell biology and treatment. *FEBS Journal* **2012,** *279* (7), 1177-1197.

134. Denholm, E. M.; Lin, Y.-Q.; Silver, P. J., Anti-tumor activities of chondroitinase AC and chondroitinase B: inhibition of angiogenesis, proliferation and invasion. *European Journal of Pharmacology* **2001,** *416* (3), 213-221.

135. Guerrini, M.; Beccati, D.; Shriver, Z.; Naggi, A. M.; Bisio, A.; Capila, I.; Lansing, J.; Guglieri, S.; Fraser, B.; Al-Hakim, A.; Gunay, S.; Viswanathan, K.; Zhang, Z.; Robinson, L.; Venkataraman, G.; Buhse, L.; Nasr, M.; Woodcock, J.; Langer, R.; Linhardt, R.; Casu, B.; Torri, G.; Sasisekharan, R., Oversulfated Chondroitin Sulfate is a major contaminant in Heparin associated with Adverse Clinical Events. *Nature biotechnology* **2008,** *26* (6), 669-675.

136. Hallak, L. K.; Collins, P. L.; Knudson, W.; Peeples, M. E., Iduronic Acid-Containing Glycosaminoglycans on Target Cells Are Required for Efficient Respiratory Syncytial Virus Infection. *Virology* **2000,** *271* (2), 264-275.

137. Lindahl, U., Approaches to the Synthesis of Heparin. *Pathophysiology of Haemostasis and Thrombosis* **1990,** *20(suppl 1)* (Suppl. 1), 146-153.

138. Jacobsson, I.; Lindahl, U.; Jensen, J. W.; Roden, L.; Prihar, H.; Feingold, D. S., Biosynthesis of heparin. Substrate specificity of heparosan N-sulfate D-glucuronosyl 5-epimerase. *The Journal of biological chemistry* **1984,** *259* (2), 1056-63.

139. Muñoz-García, J. C.; Corzana, F.; de Paz, J. L.; Angulo, J.; Nieto, P. M., Conformations of the iduronate ring in short heparin fragments described by time-averaged distance restrained molecular dynamics. *Glycobiology* **2013,** *23* (11), 1220-1229.

140. Silbert, J. E.; Sugumaran, G., Biosynthesis of Chondroitin/Dermatan Sulfate. *IUBMB Life* **2002,** *54* (4), 177-186.

141. Mobli, M.; Nilsson, M.; Almond, A., The structural plasticity of heparan sulfate NA-domains and hence their role in mediating multivalent interactions is confirmed by high-accuracy 15N-NMR relaxation studies. *Glycoconj J* **2008,** *25* (5), 401-414.

142. Jayson, Gordon C.; Miller, Gavin J.; Hansen, Steen U.; Barath, M.; Gardiner, John M.; Avizienyte, E., The development of anti-angiogenic heparan sulfate oligosaccharides. *Biochem Soc T* **2014,** *42* (6), 1596-1600.

143. Avizienyte, E.; Cole, C. L.; Rushton, G.; Miller, G. J.; Bugatti, A.; Presta, M.; Gardiner, J. M.; Jayson, G. C., Synthetic Site-Selectively Mono-6-O-Sulfated Heparan Sulfate Dodecasaccharide Shows Anti-Angiogenic Properties In Vitro and Sensitizes Tumors to Cisplatin In Vivo. *Plos One* **2016,** *11* (8), e0159739.

144. Mulloy, B.; Forster, M. J., Conformation and dynamics of heparin and heparan sulfate. *Glycobiology* **2000,** *10* (11), 1147-1156.

145. Pavao, M. S.; Mourao, P. A.; Mulloy, B.; Tollefsen, D. M., A unique dermatan sulfate-like glycosaminoglycan from ascidian. Its structure and the effect of its unusual sulfation pattern on anticoagulant activity. *The Journal of biological chemistry* **1995,** *270* (52), 31027-36.

146. Pye, D. A.; Vives, R. R.; Turnbull, J. E.; Hyde, P.; Gallagher, J. T., Heparan sulfate oligosaccharides require 6-O-sulfation for promotion of basic fibroblast growth factor mitogenic activity. *Journal of Biological Chemistry* **1998,** *273* (36), 22936-22942.

147. Sugaya, N.; Habuchi, H.; Nagai, N.; Ashikari-Hada, S.; Kimata, K., 6-O-sulfation of heparan sulfate differentially regulates various fibroblast growth factor-dependent signalings in culture. *The Journal of biological chemistry* **2008,** *283* (16), 10366-76.

148. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001,** *46* (1), 3-26.

149. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.

150. Kirkwood, J. G., Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics* **1935,** *3* (5), 300-313.

151. Mobley, D. L.; Gilson, M. K., Predicting binding free energies: Frontiers and benchmarks. *bioRxiv* **2016**.

152. Rocklin, G. J.; Boyce, S. E.; Fischer, M.; Fish, I.; Mobley, D. L.; Shoichet, B. K.; Dill, K. A., Blind prediction of charged ligand binding affinities in a model binding site. *J Mol Biol* **2013,** *425* (22), 4569-83.

153. Paluch, A. S.; Mobley, D. L.; Maginn, E. J., Small Molecule Solvation Free Energy: Enhanced Conformational Sampling Using Expanded Ensemble Molecular Dynamics Simulation. *J Chem Theory Comput* **2011,** *7* (9), 2910-2918.

154. Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K., Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. *Journal of Molecular Biology* **2009,** *394* (4), 747-763.

155. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem* **2002,** *23* (16), 1623-1641.

156. Baker, C. M.; Lopes, P. E. M.; Zhu, X.; Roux, B.; MacKerell, A. D., Accurate Calculation of Hydration Free Energies using Pair-Specific Lennard-Jones Parameters in the CHARMM Drude Polarizable Force Field. *J Chem Theory Comput* **2010,** *6* (4), 1181-1198.

157. Bradshaw, R. T.; Essex, J. W., Evaluating Parametrization Protocols for Hydration Free Energy Calculations with the AMOEBA Polarizable Force Field. *J Chem Theory Comput* **2016,** *12* (8), 3871-3883.

158. Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J., The SAMPL2 blind prediction challenge: introduction and overview. *J Comput Aid Mol Des* **2010,** *24* (4), 259-279.

159. Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K., The SAMPL4 host–guest blind prediction challenge: an overview. *J Comput Aid Mol Des* **2014,** *28* (4), 305-317.

160. Skillman, A. G., SAMPL3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. *J Comput Aid Mol Des* **2012,** *26* (5), 473-474.

161. Klimovich, P. V.; Mobley, D. L., Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J Comput Aid Mol Des* **2010,** *24* (4), 307-316.

162. Knight, J. L.; Brooks, C. L., λ-Dynamics free energy simulation methods. *J Comput Chem* **2009,** *30* (11), 1692-1700.

163. Woods, C. J.; Essex, J. W.; King, M. A., The Development of Replica-Exchange-Based Free-Energy Methods. *The Journal of Physical Chemistry B* **2003,** *107* (49), 13703-13710.

164. Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J., Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (39), 13749-13754.

165. de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A., Coupling Accelerated Molecular Dynamics Methods with Thermodynamic Integration Simulations. *J Chem Theory Comput* **2008,** *4* (9), 1516-1525.

166. Wu, P.; Hu, X.; Yang, W., λ-Metadynamics Approach To Compute Absolute Solvation Free Energy. *The Journal of Physical Chemistry Letters* **2011,** *2* (17), 2099-2103.

167. Lawrenz, M.; Baron, R.; Wang, Y.; McCammon, J. A., Independent-Trajectory Thermodynamic Integration: A Practical Guide to Protein-Drug Binding Free Energy Calculations Using Distributed Computing. In *Computational Drug Discovery and Design*, Baron, R., Ed. Springer New York: New York, NY, 2012; pp 469-486.

168. Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammont, J. A., Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *J Chem Theory Comput* **2013,** *9* (9), 10.1021/ct400340s.

169. Lawrenz, M.; Baron, R.; McCammon, J. A., Independent-Trajectories Thermodynamic-Integration Free-Energy Changes for Biomolecular Systems: Determinants of H5N1 Avian Influenza Virus Neuraminidase Inhibition by Peramivir. *J Chem Theory Comput* **2009,** *5* (4), 1106-1116.

170. Mobley, D. L.; Guthrie, J. P., FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J Comput Aid Mol Des* **2014,** *28* (7), 711-720.

171. Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L., Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *Journal of Chemical & Engineering Data* **2017,** *62* (5), 1559-1569.

172. Bosisio, S.; Mey, A. S. J. S.; Michel, J., Blinded predictions of host-guest standard free energies of binding in the SAMPL5 challenge. *J Comput Aid Mol Des* **2017,** *31* (1), 61-70.

173. Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K., Overview of the SAMPL5 host–guest challenge: Are we doing better? *J Comput Aid Mol Des* **2017,** *31* (1), 1-19.

174. Yin, J.; Henriksen, N. M.; Slochower, D. R.; Gilson, M. K., The SAMPL5 host–guest challenge: computing binding free energies and enthalpies from explicit solvent simulations by the attach-pull-release (APR) method. *J Comput Aid Mol Des* **2017,** *31* (1), 133-145.

175. Steinbrecher, T.; Joung, I.; Case, D. A., Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations. *J Comput Chem* **2011,** *32* (15), 3253-3263.

176. Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammon, J. A., Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *J Chem Theory Comput* **2013,** *9* (9), 4131-4139.

177. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.

178. FrischmJ, T., Gaussian09, revisionD. 01 [CP]. *GaussianInc.: Wallingford, CT* **2009**.

179. Straatsma, T. P.; Berendsen, H. J. C.; Stam, A. J., Estimation of statistical errors in molecular simulation calculations. *Molecular Physics* **1986,** *57* (1), 89-95.

180. Shirts, M. R.; Chodera, J. D., Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics* **2008,** *129* (12), 124105.

181. Wang, E.; Zhang, Q.; Shen, B.; Zhang, G.; Lu, X.; Wu, Q.; Wang, Y., Intel Math Kernel Library. In *High-Performance Computing on the Intel® Xeon Phi™: How to Fully Exploit MIC Architectures*, Springer International Publishing: Cham, 2014; pp 167-188.

182. Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A., Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J Chem Theory Comput* **2009,** *5* (2), 350-358.

183. Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D., Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J Comput Aid Mol Des* **2017,** *31* (1), 147-161.

184. Mobley, D. L.; Dumont, É.; Chodera, J. D.; Dill, K. A., Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. *The Journal of Physical Chemistry B* **2007,** *111* (9), 2242-2254.

185. Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A., Predictions of hydration free energies from all-atom molecular dynamics simulations. *The journal of physical chemistry. B* **2009,** *113* (14), 4533-4537.

186. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **2010,** *50* (4), 572-584.

187. König, G.; Boresch, S., Non-Boltzmann sampling and Bennett's acceptance ratio method: How to profit from bending the rules. *J Comput Chem* **2011,** *32* (6), 1082-1090.

188. Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T., Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J Chem Theory Comput* **2016,** *12* (8), 3926-3947.

Appendix A

A.1 The Structure of *sander* in the view of sesMD

Shown below is the generalised flow of the AMBER *sander* engine as it relates to sesMD. At the very start of the simulation the multisander.F90 and sander.F90 routines set up all variables used during the simulation. Once done, the runmd.F90 subroutine is called which handles the integration of the equations of motion, subsequent calls to the force.F90 subroutine are made in order to calculate the forces from a given set of coordinates. In *sander* multi-core parallelism is only achieved in the force calculations, with the forces and coordinates being explicitly shared amongst all MPI processes before and after any calls to force.F90. Several subroutines are not detailed in Figure S2.1, such as those handling temperature and pressure control.

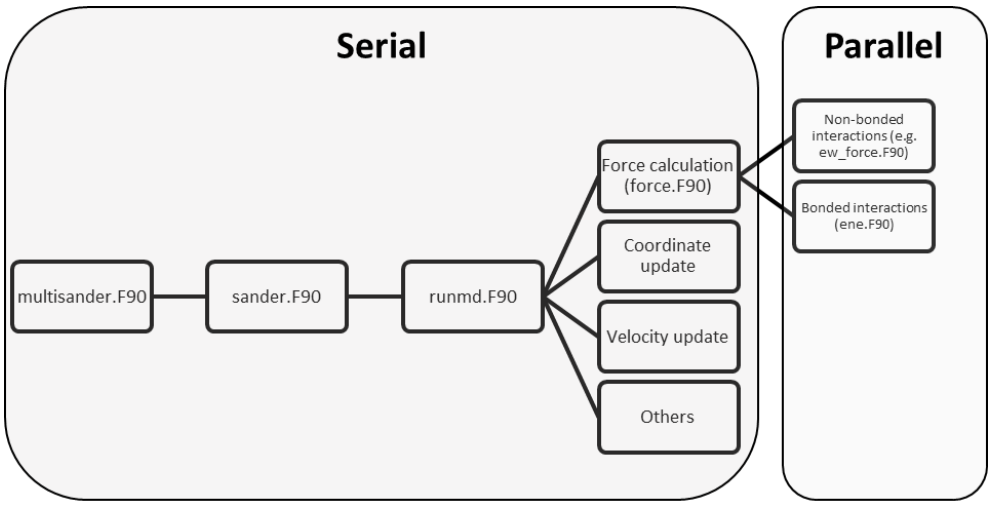

Figure A.1 Hierarchical structure of the main sander MD engine compute routines

A.2 The Structure of *pmemd* in the view of msesMD

The *pmemd* MD engine, shares a very similar code structure to *sander*, which is expected considering that it was initially developed from an earlier version of *sander*. However, unlike *sander*, the parallelism is achieved at a much higher level, with each MPI process handling the calculation and integration of the equations of motion for their own portion of the coordinates independently.

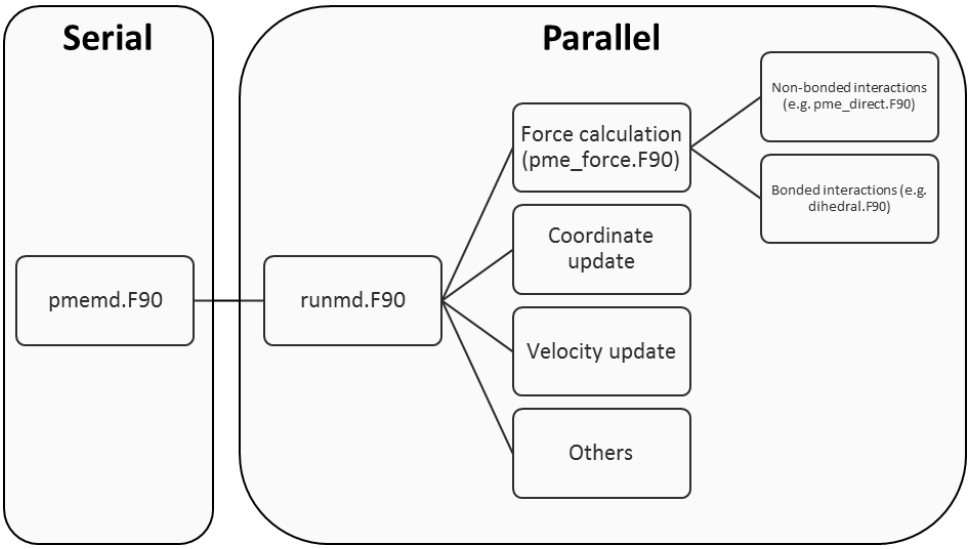

Figure A.2 Hierarchical structure of the main pmemd MD engine compute subroutines

Appendix B

Table B.1 Glycosidic linkage angle values of Le[a] from crystallographic PDB structures

Le[a]	**Fuc α(1→4) GlcNAc**		**Gal β(1→3) GlcNAc**	
PDB code	***φ***	***ψ***	***φ***	***ψ***
1W8H	-75.0	133.6	-62.3	-110.3
	-72.8	136.5	-66.1	-110.2
	-66.9	133.8	-65.7	-103.5
	-77.1	138.8	78.3	-93.9
3ASR	-70.8	138.0	-69.5	-111.5
4P3I	-77.6	137.9	-83.0	-108.1
	-75.4	146.0	-74.8	-103.4
	-51.3	137.6	-85.2	-98.2
	-67.8	131.4	-91.5	-94.4
4RM0	-68.0	139.7	-60.1	-110.6
	-76.0	141.2	-67.5	-107.0
4UT5	-69.9	140.6	-56.8	-104.9
4WZL	-93.9	130.4	-73.9	-96.3
	-93.7	129.7	-74.4	-96.1
5A6Z	-70.4	140.9	-66.4	-103.1

Table B.2 Glycosidic linkage angle values of Le^x from crystallographic PDB structures

Le^x	**Fuc α(1→3) GlcNAc**		**Gal β(1→4) GlcNAc**	
PDB code	***φ***	***ψ***	***φ***	***ψ***
1SL6	-69.2	-100.8	-61.6	131.4
	-63.7	-94.0	-66.9	132.8
	-72.7	-91.7	-66.6	119.9
	-61.9	-95.2	-54.6	128.4
	-62.0	-86.3	-58.8	124.6
	-58.1	-83.6	-56.0	129.7
1UZ8	-81.4	-98.6	-69.9	129.9
	-82.7	-99.4	-66.7	128.0
2OX9	-73.0	-107.6	-76.2	133.2
	-69.7	-97.4	-79.4	133.8
	-69.7	-100.6	-72.1	130.7
	-70.0	-107.4	-76.3	141.9
4P2N	-70.8	-102.4	-85.6	135.3
	-72.3	-102.9	-92.4	140.3
	-73.9	-99.4	-81.4	132.1
	-76.4	-100.1	-86.8	137.4
4X0C	-79.4	-103.1	-64.4	133.4
	-82.9	-100.2	-68.8	127.2
	-80.2	-98.3	-46.7	116.1
5AJB	-79.5	-135.9	-140.3	-108.1
	-88.0	55.9	-78.1	-57.8
	-99.6	-117.2	-31.1	-50.8
	-93.9	-118.9	-37.7	-49.0
5AJC	-96.5	61.3	-103.2	-56.7

Table B.3 Glycosidic linkage angle values of sLe^x from crystallographic PDB structures

Sle^x	**Fuc α(1→3) GlcNAc**		**Gal β(1→4) GlcNAc**		**Neu5Ac α(2→3) Gal**	
PDB code	***φ***	***ψ***	***φ***	***ψ***	***φ***	***ψ***
1G1T	-76.9	-97.0	-85.8	135.6	56.5	-129.7
2KMB	-70.5	-105.9	-72.1	139.1	67.9	-159.5
	-69.9	-98.4	-74.9	134.3	51.5	-138.3
	-68.2	-107.1	-86.6	143.1	50.3	-147.4
2R61	-78.4	-96.9	-69.1	125.4	55.3	-140.9
2RDG	-70.4	-101.4	-86.2	135.4	55.5	-137.8
2Z8L	-71.9	-100.6	-84.8	133.0	58.4	-133.1
3PVD	-87.6	-99.1	-72.8	135.9	49.3	-125.4
	-80.4	-104.7	-73.6	140.9	60.3	-136.4
4CSY	-83.4	-104.3	-85.9	133.1	-5.1	-81.8
	-84.4	-104.5	-85.1	132.3	-5.7	-82.6
4DXG	-75.5	-100.4	-81.4	133.4	62.6	-129.1
4RCO	-79.8	-97.5	-85.5	134.6	61.5	-129.0
	-72.6	-101.4	-88.1	135.6	62.4	-133.8
4RFB	-58.3	-102.5	-74.8	127.3	54.2	-135.0
	-71.5	-100.5	-74.5	129.5	61.5	-136.8
	-69.4	-99.0	-78.2	133.9	58.7	-136.9
	-64.3	-102.5	-75.8	129.1	60.0	-135.4
4USO	-66.5	-96.2	-73.8	128.8	48.0	-130.9
5AJC	-94.0	56.7	-86.8	-76.3	51.9	-117.9
5I4D	-78.6	-97.7	-82.5	134.1	57.2	-141.7
	-65.1	-97.7	-89.0	134.5	53.6	-134.0

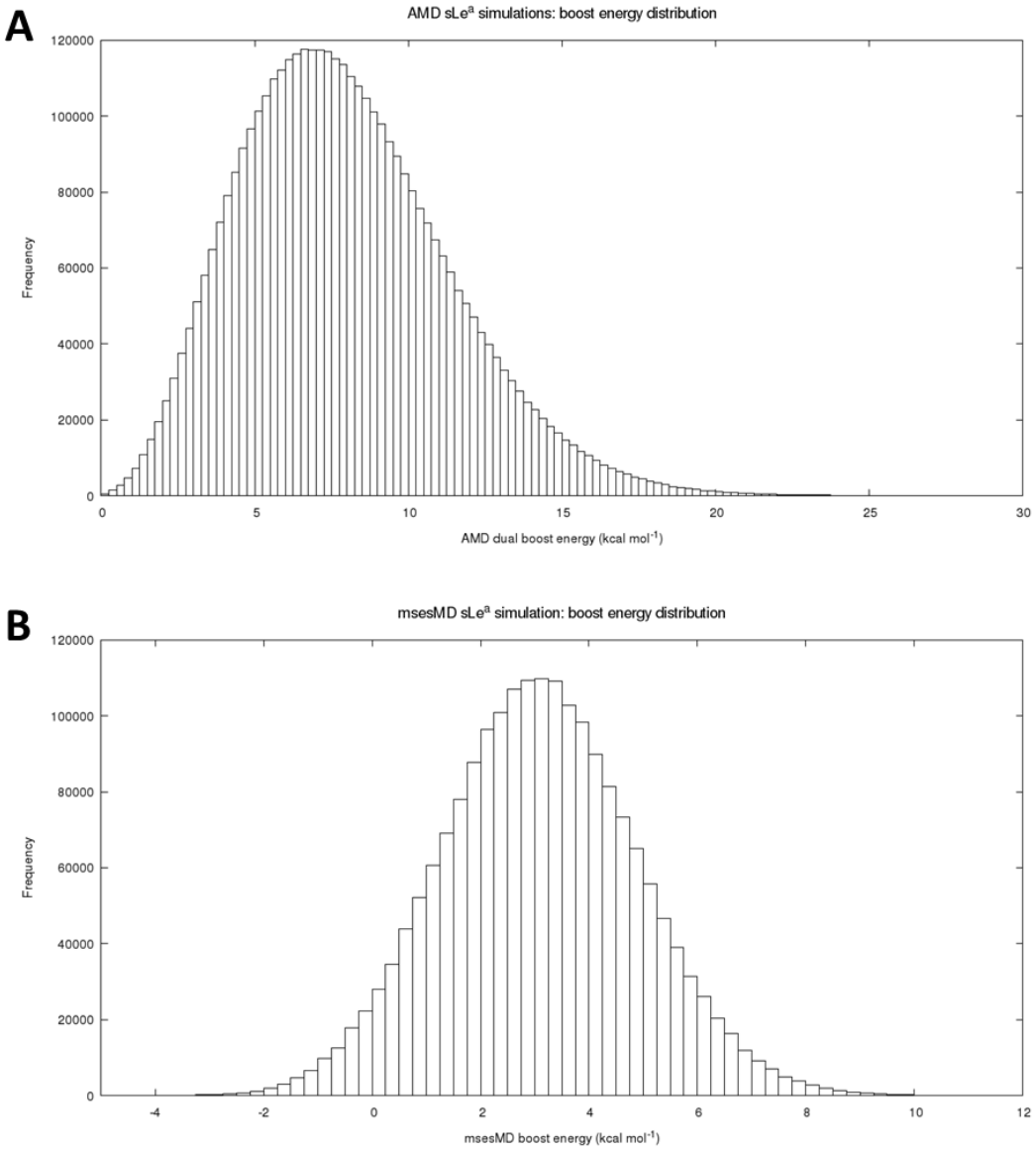

*Figure B.1 Histograms of boost potential energies from sLe^a calculations (kcal mol^{-1}) for **A)** AMD dual boost, **B**) msesMD*

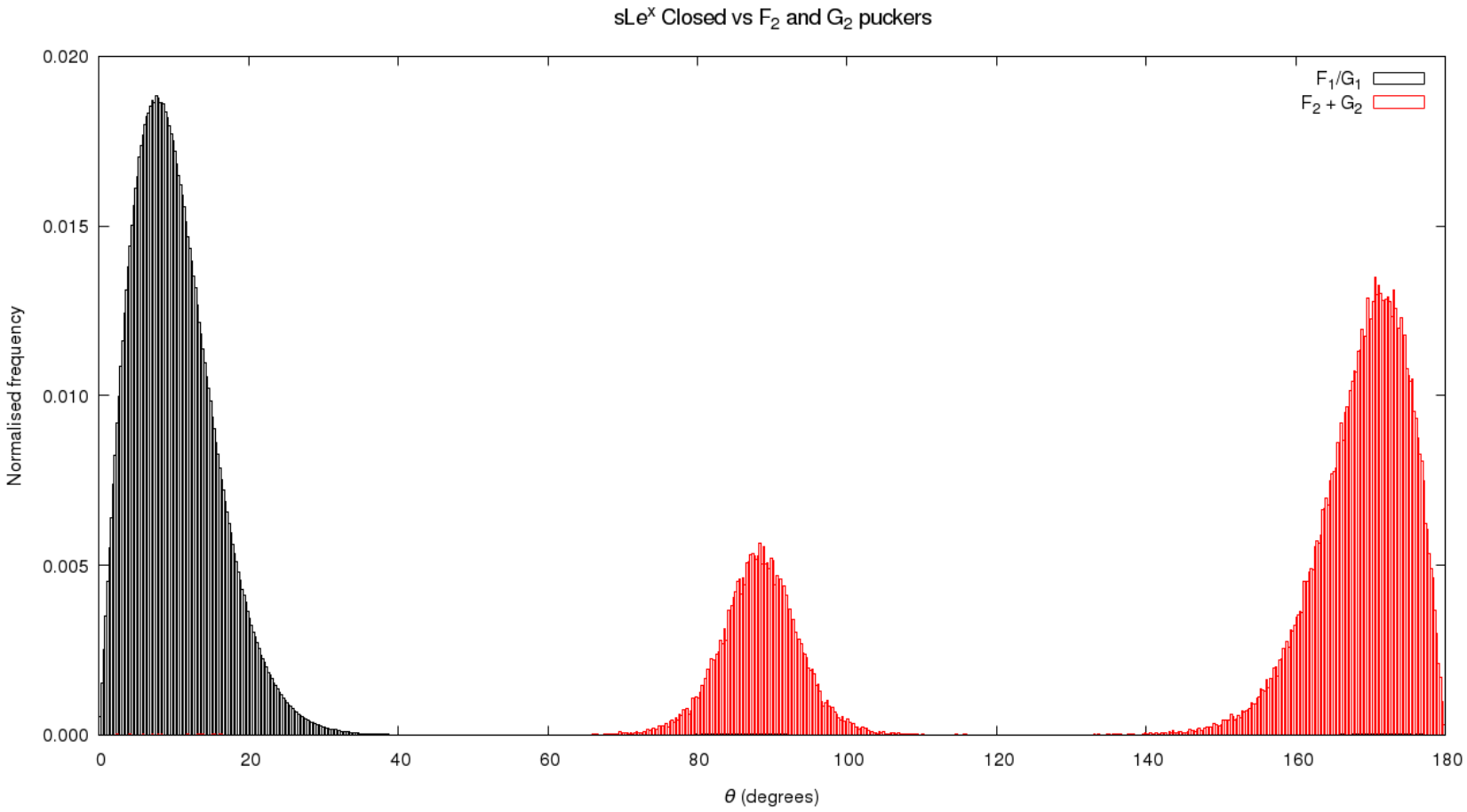

Figure B.2 GlcNAc Cremer-Pople θ coordinate normalised frequency distribution histogram for HMR MD frames that exist in the F_1/G_1 wells (black) and either the F_2 or G_2 wells (red)

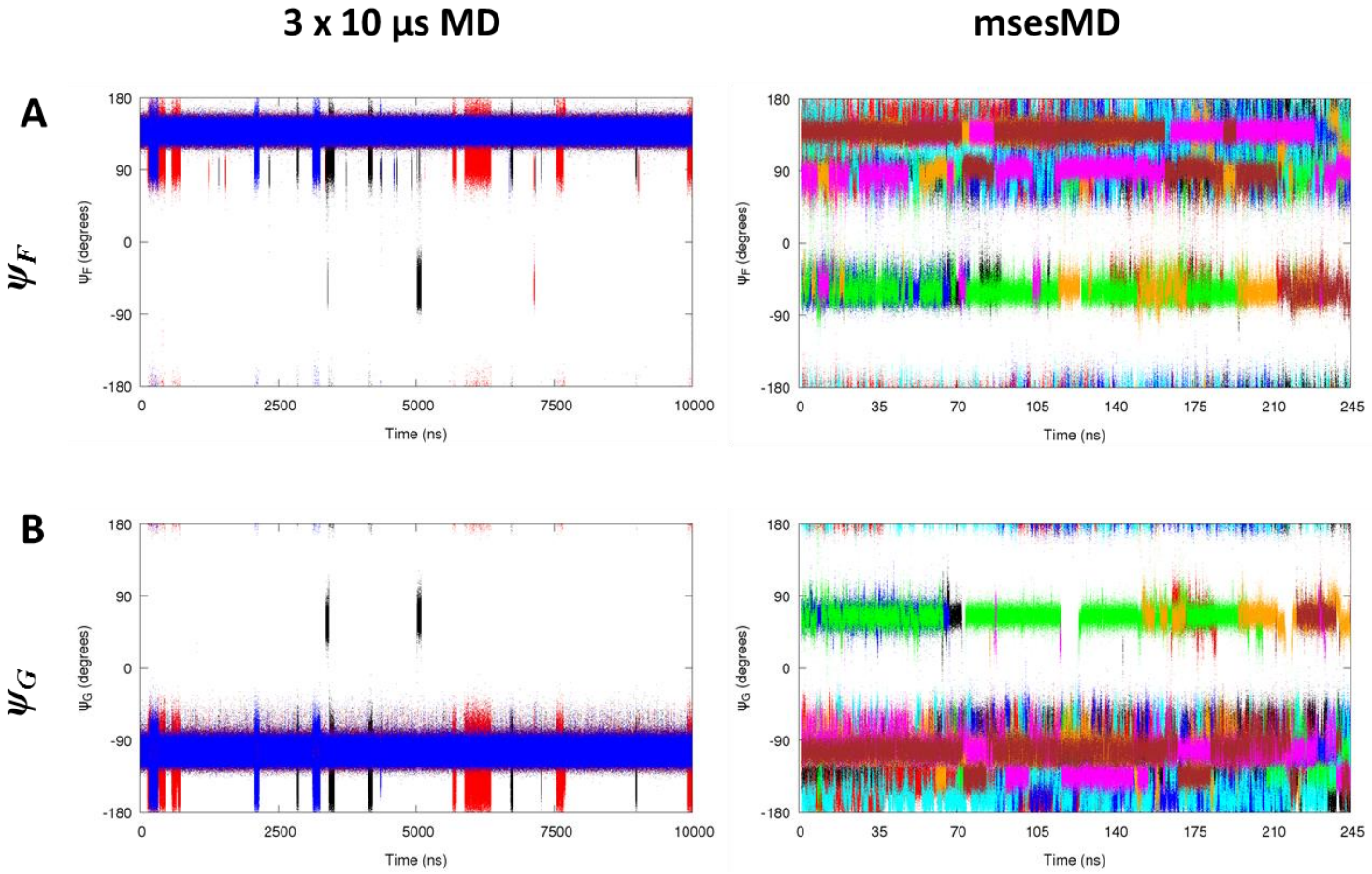

Figure B.3 A) ψ_F and b) ψ_G time profiles of Le^a for unbiased MD and msesMD

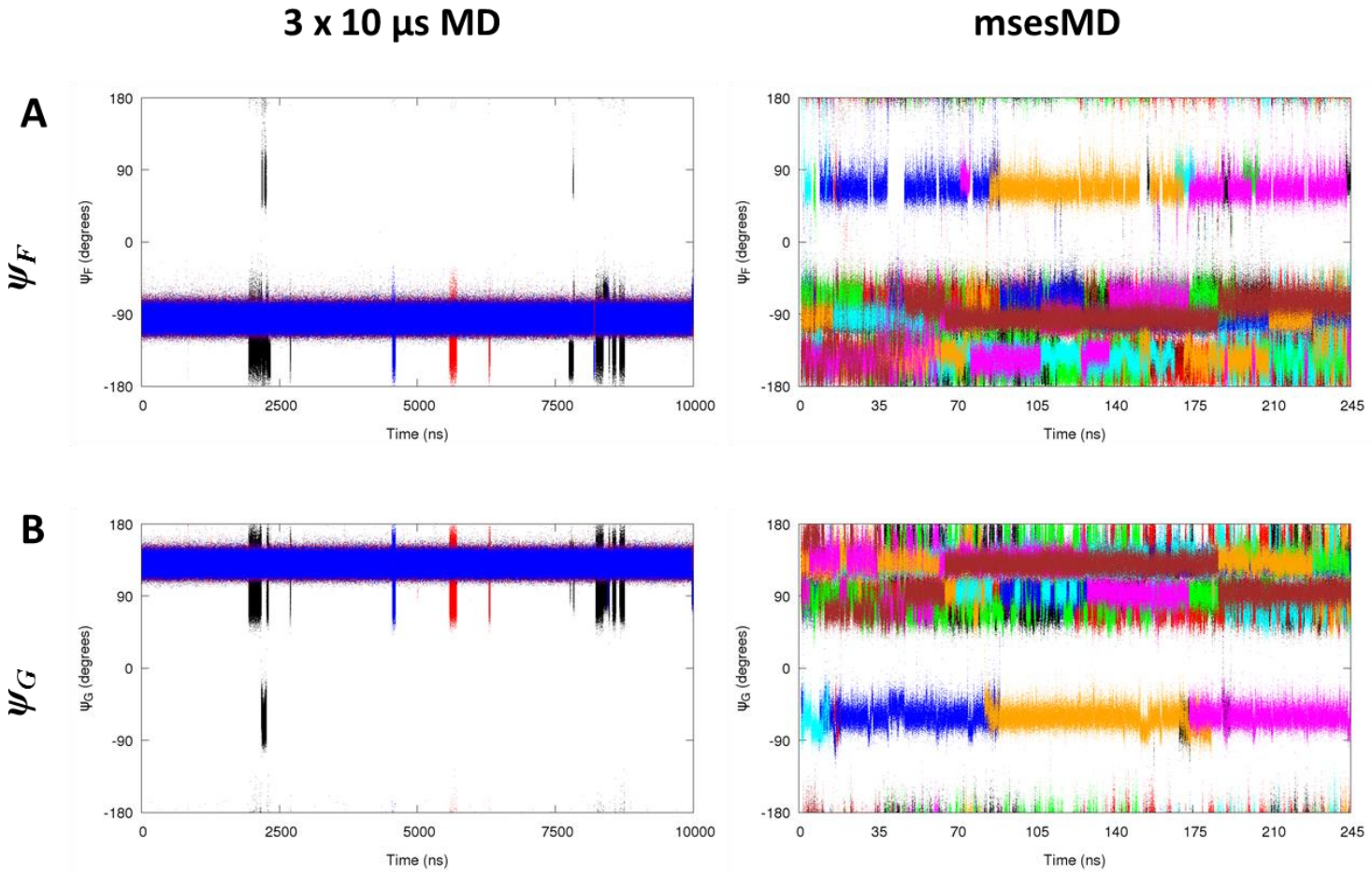

Figure B.4 A) ψ_F and b) ψ_G time profiles of Le^x for unbiased MD and msesMD

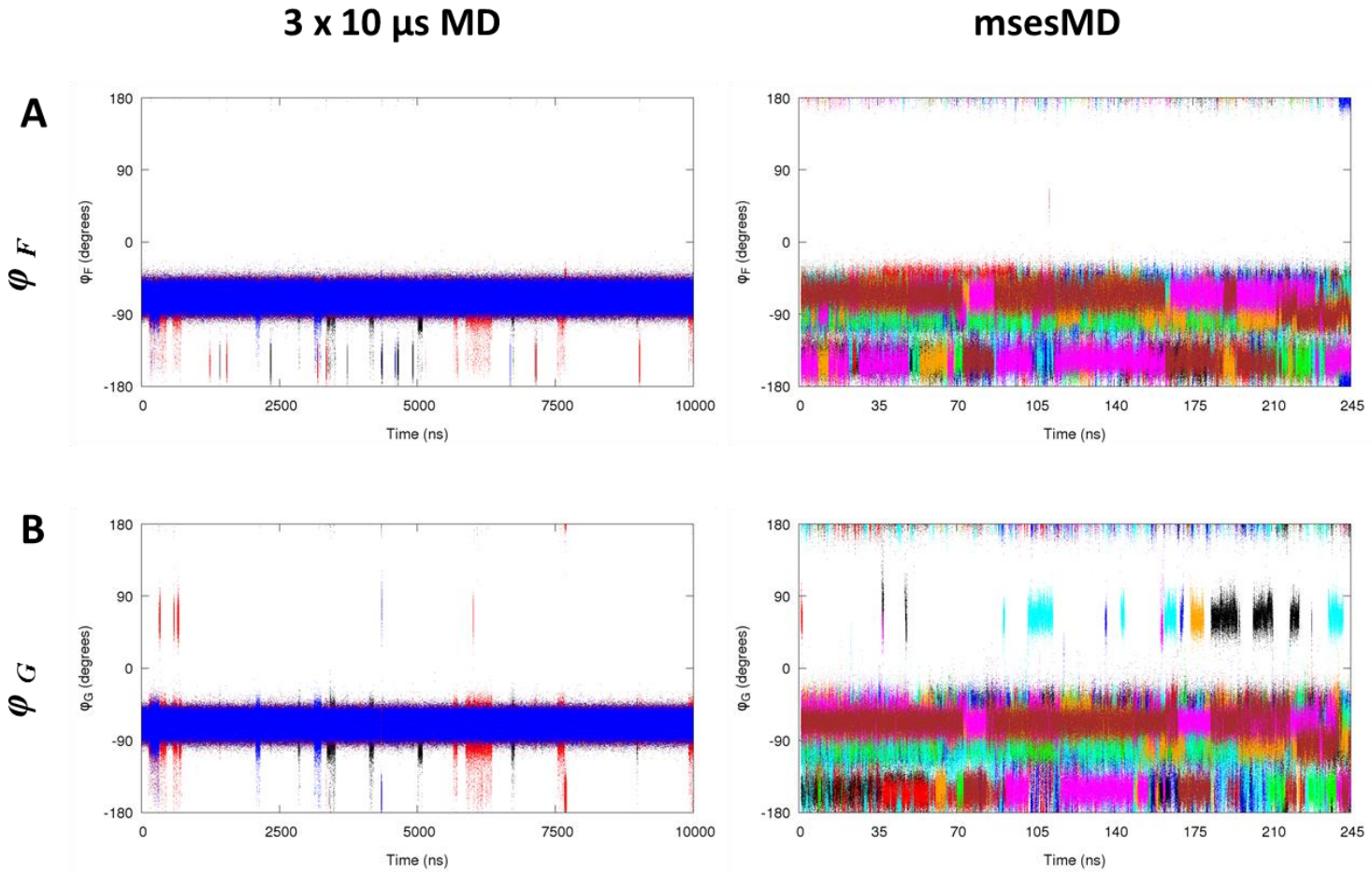

Figure B.5 A) φ_F and b) φ_G time profiles of Le^a for unbiased MD and msesMD

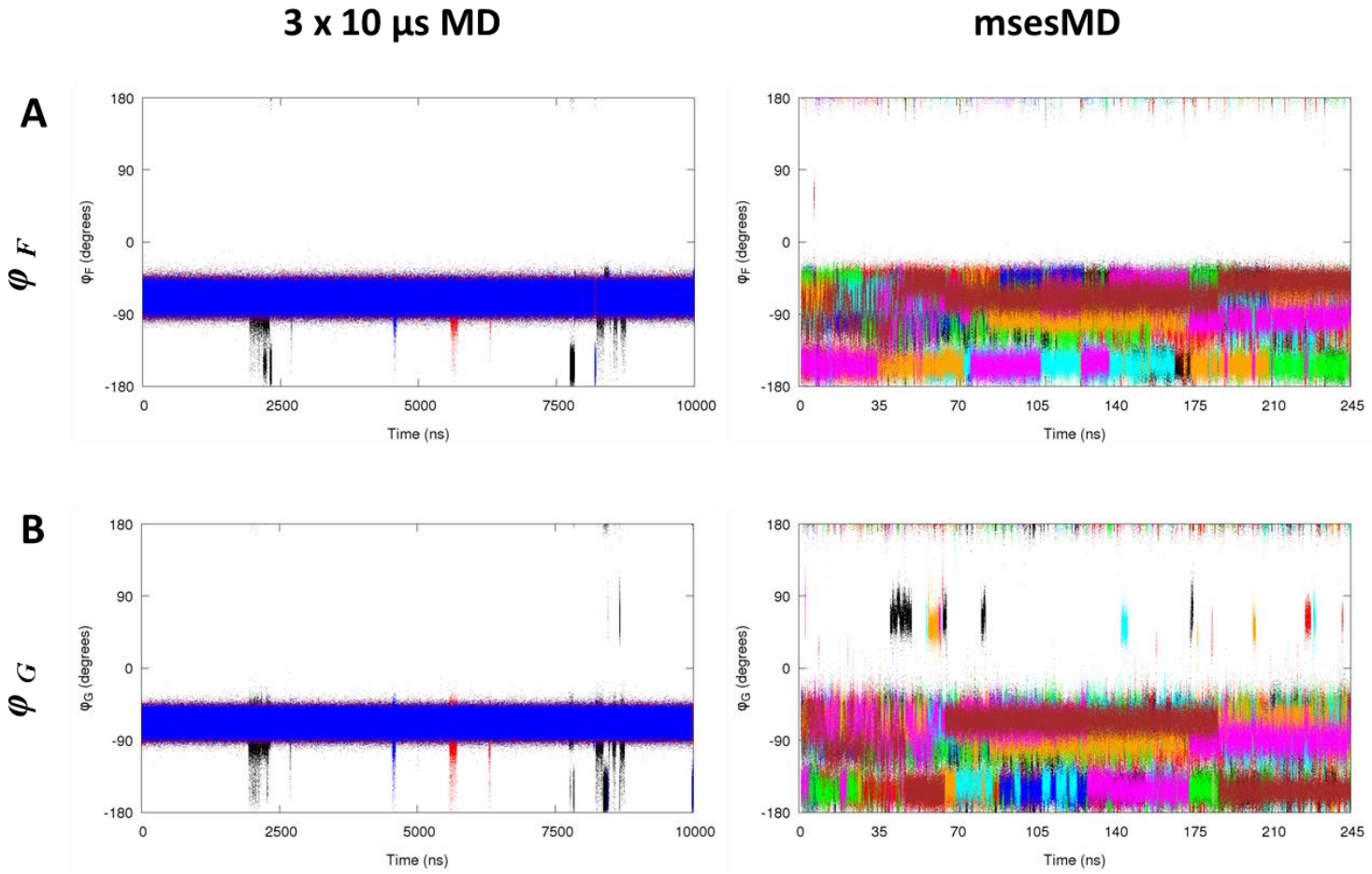

Figure B.6 A) φ_F and b) φ_G time profiles of Le^x for unbiased MD and msesMD

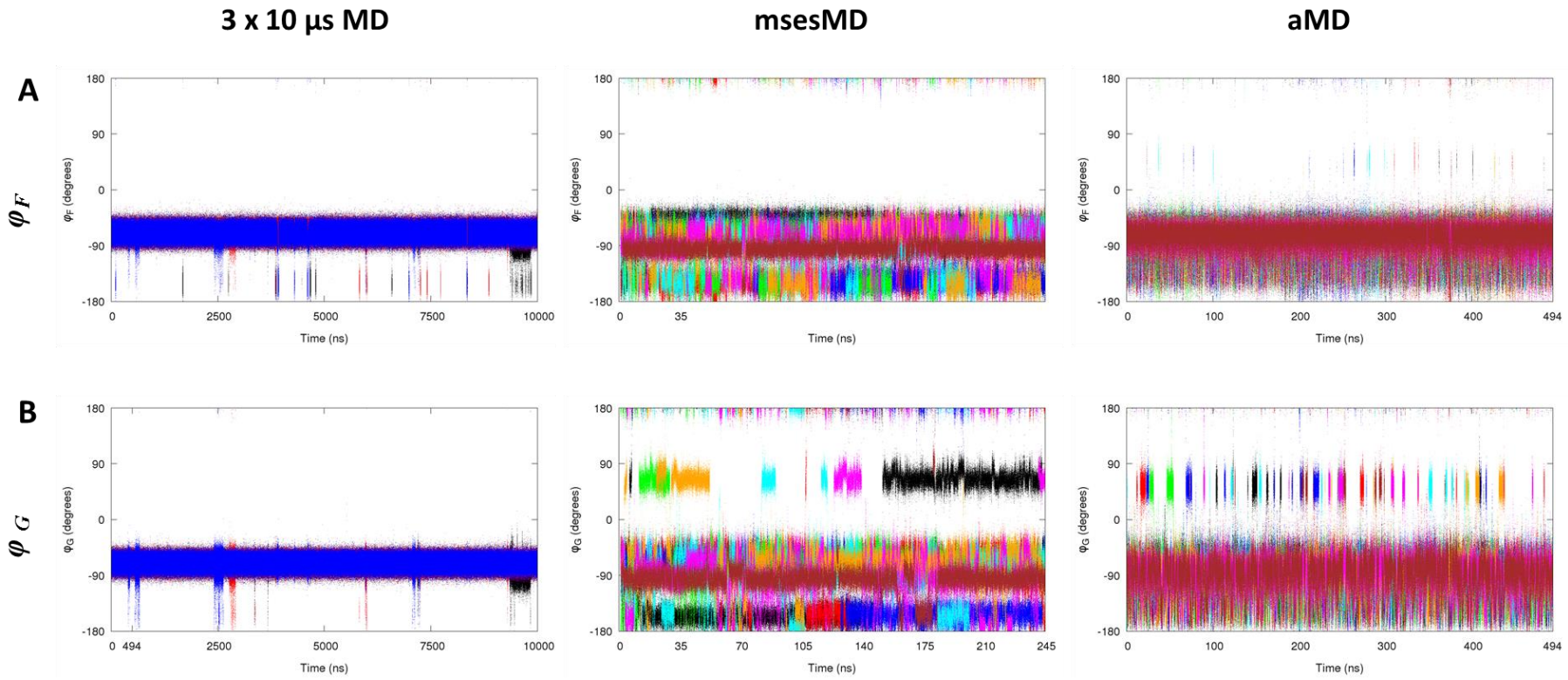

Figure B.7 A) φ_F and b) φ_G time profiles of sLe^a for unbiased MD, msesMD and aMD

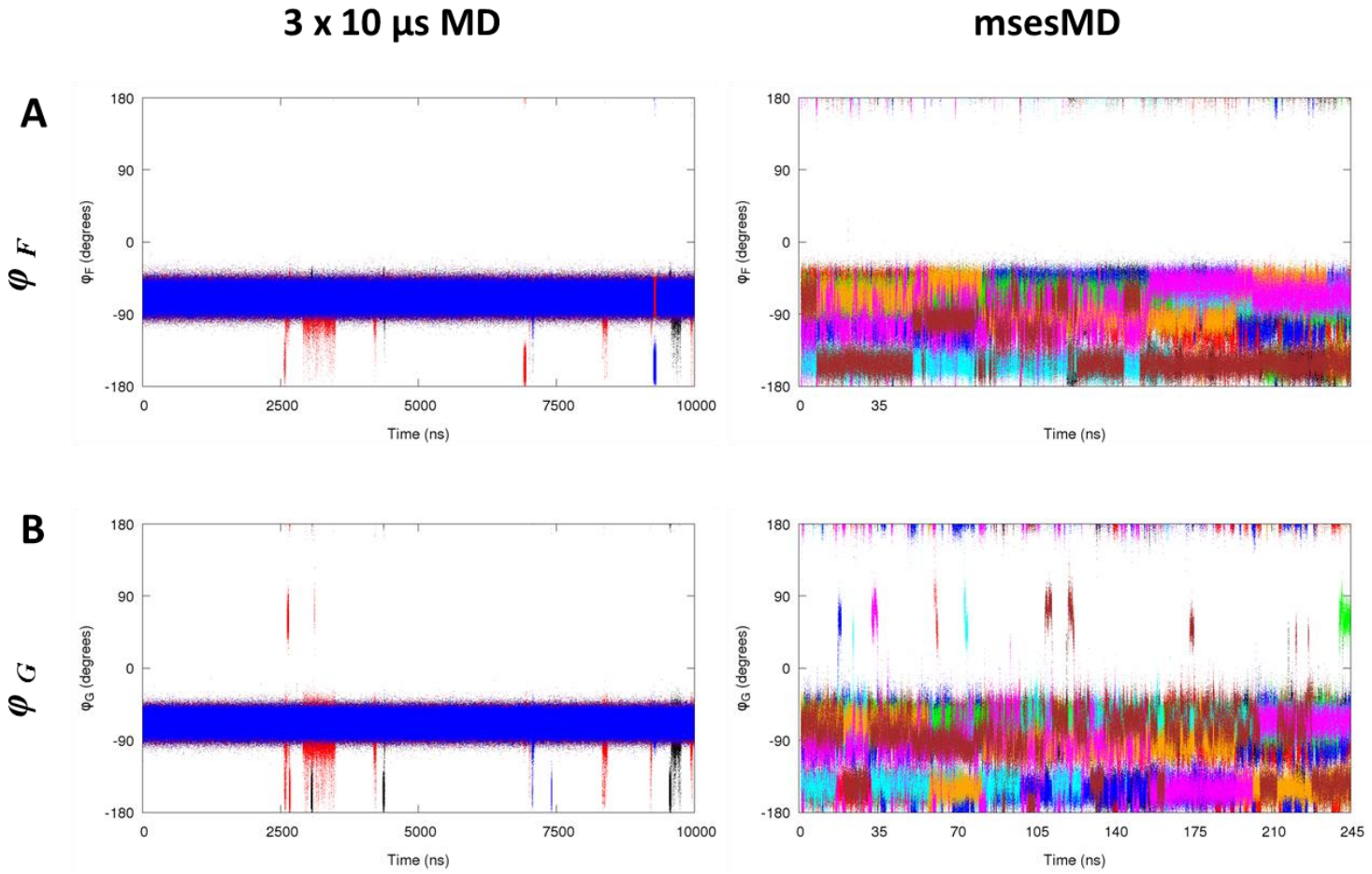

Figure B.8 A) φ_F and b) φ_G time profiles of sLe^x for unbiased MD and msesMD

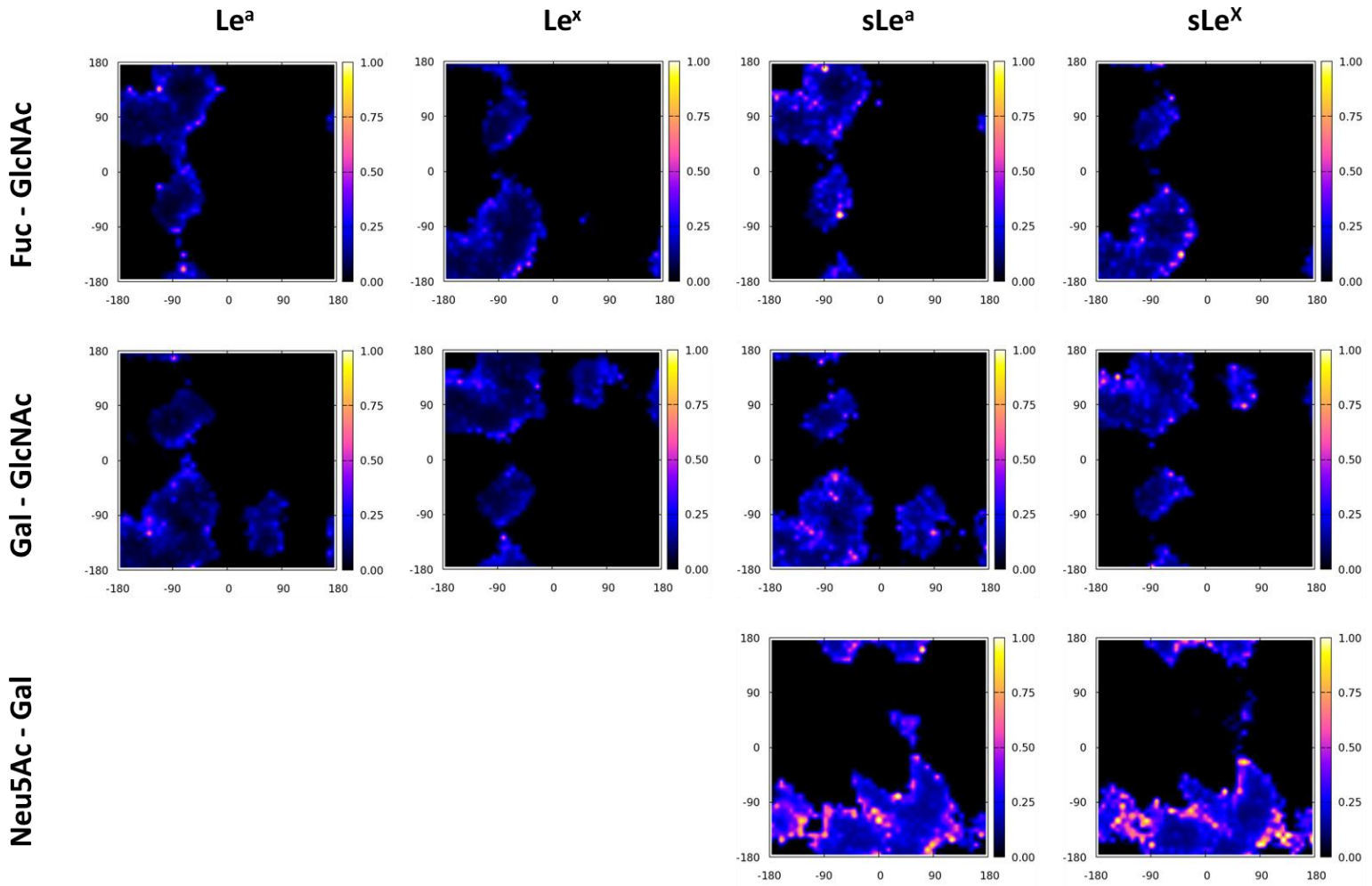

Figure B.9 Bootstrap sampling calculated errors in the $\phi\psi$ glycosidic torsions of Le^a, Le^x, sLe^a, and sLe^x (kcal mol^{-1})

Appendix C

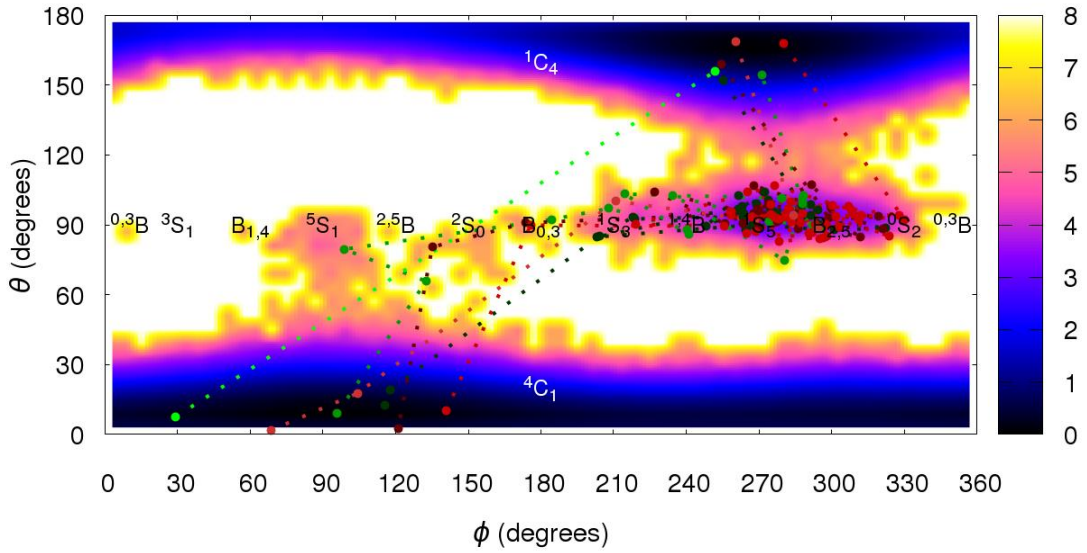

Figure C.1 Cremer-Pople θ vs ϕ free energy plot of the unbiased 15 microsecond simulation of α-D-glucose, overlaid with the six $^{4}C_1$ to $^{1}C_4$ transition paths observed during the simulation, forward and backward transitions are coloured in green and red respectively

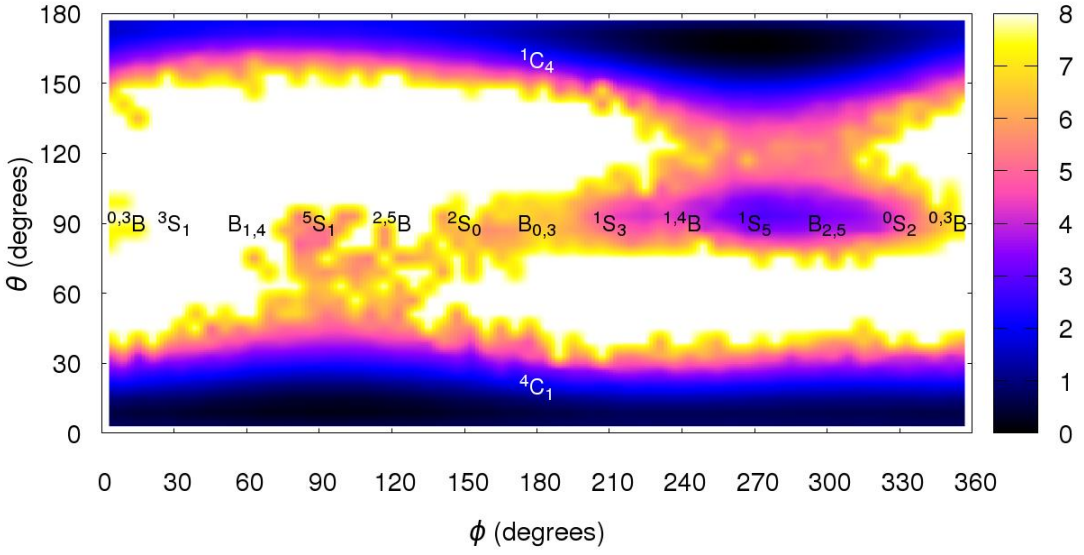

Figure C.2 Cremer-Pople θ vs ϕ free energy plot of α-D-glucose calculated via msesMD using a factor of 4 reduced boost potential

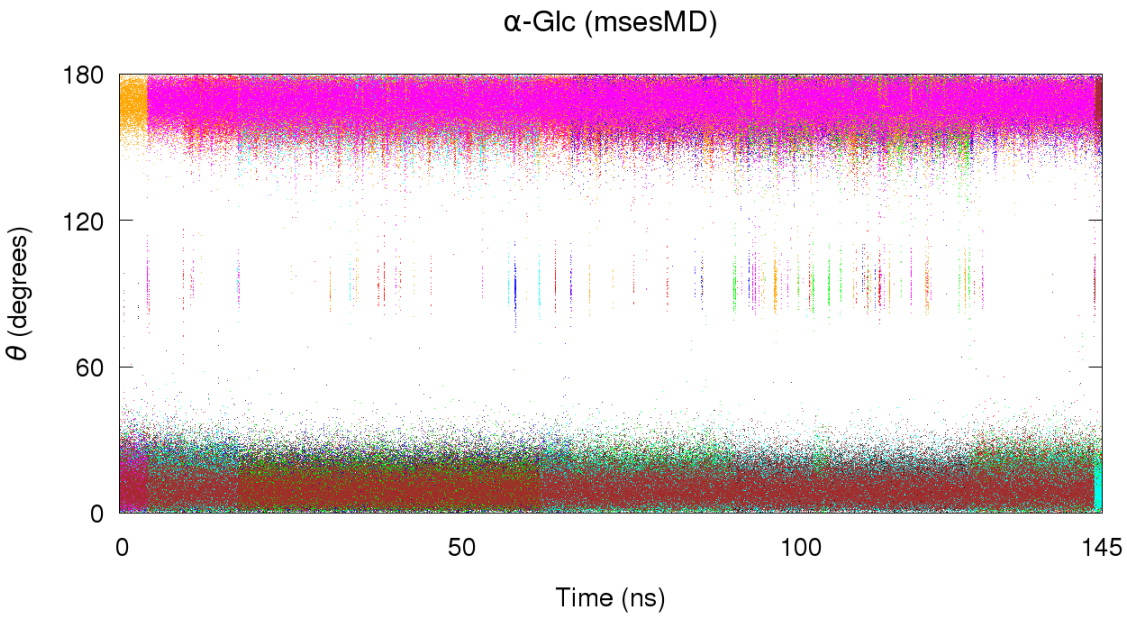

Figure C.3 Change in the Cremer-Pople θ angle over simulation time for all 8 replica (individually coloured) during the α-D-glucose msesMD simulation using a factor of 4 reduced boost potential

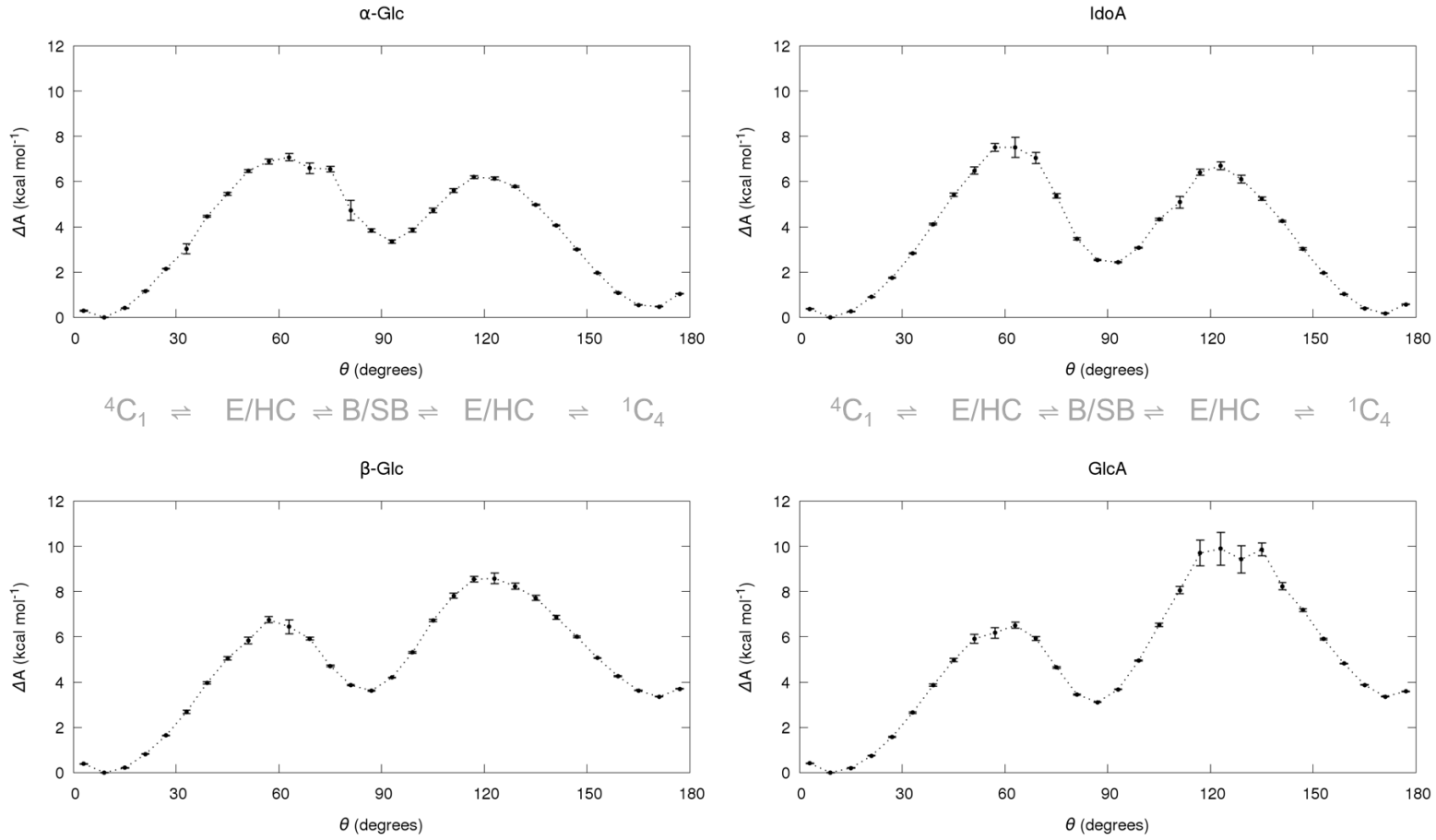

Figure C.4 Bootstrap error analysis of the Cremer-Pople θ free energy profiles of the benchmark systems calculated via msesMD

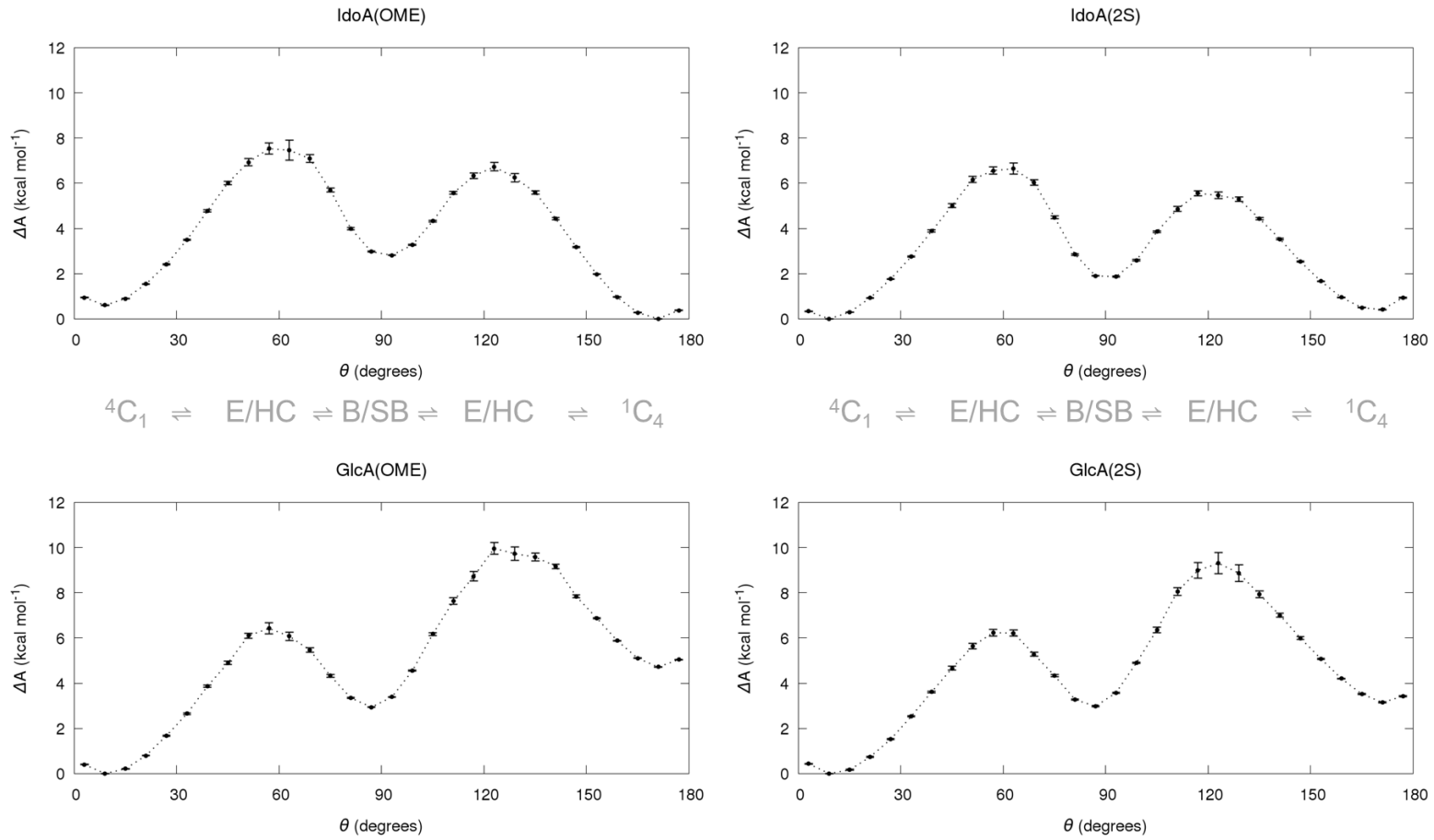

Figure C.5 Bootstrap error analysis of the Cremer-Pople θ free energy profiles of the O-methylated (OME) and 2-O-sulfated (2S) uronic acids systems calculated via msesMD

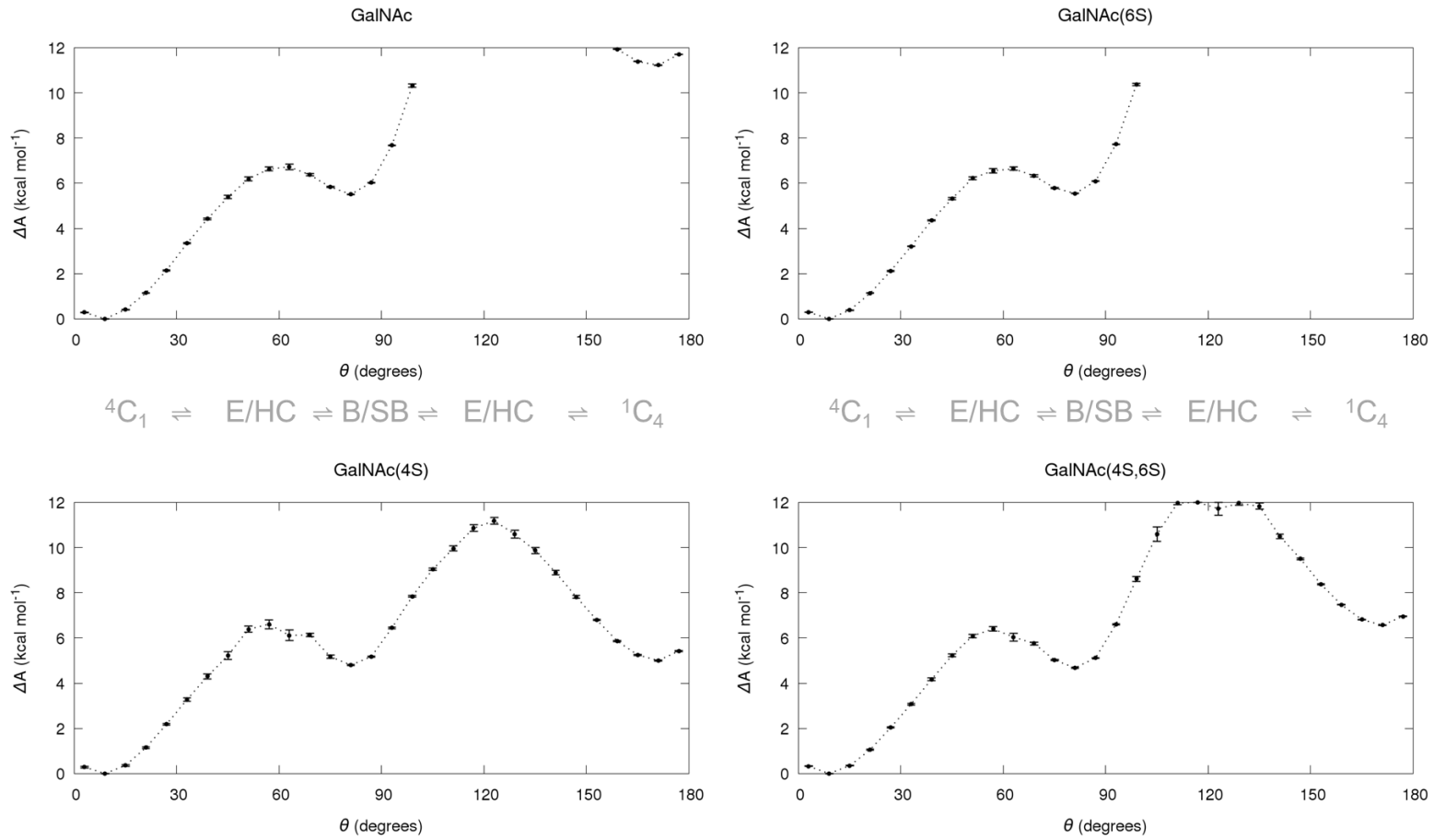

Figure C.6 Bootstrap error analysis of the Cremer-Pople θ free energy profiles of the varying decoration patterns of N-Acety- galactosamine (GalNAc) calculated via msesMD

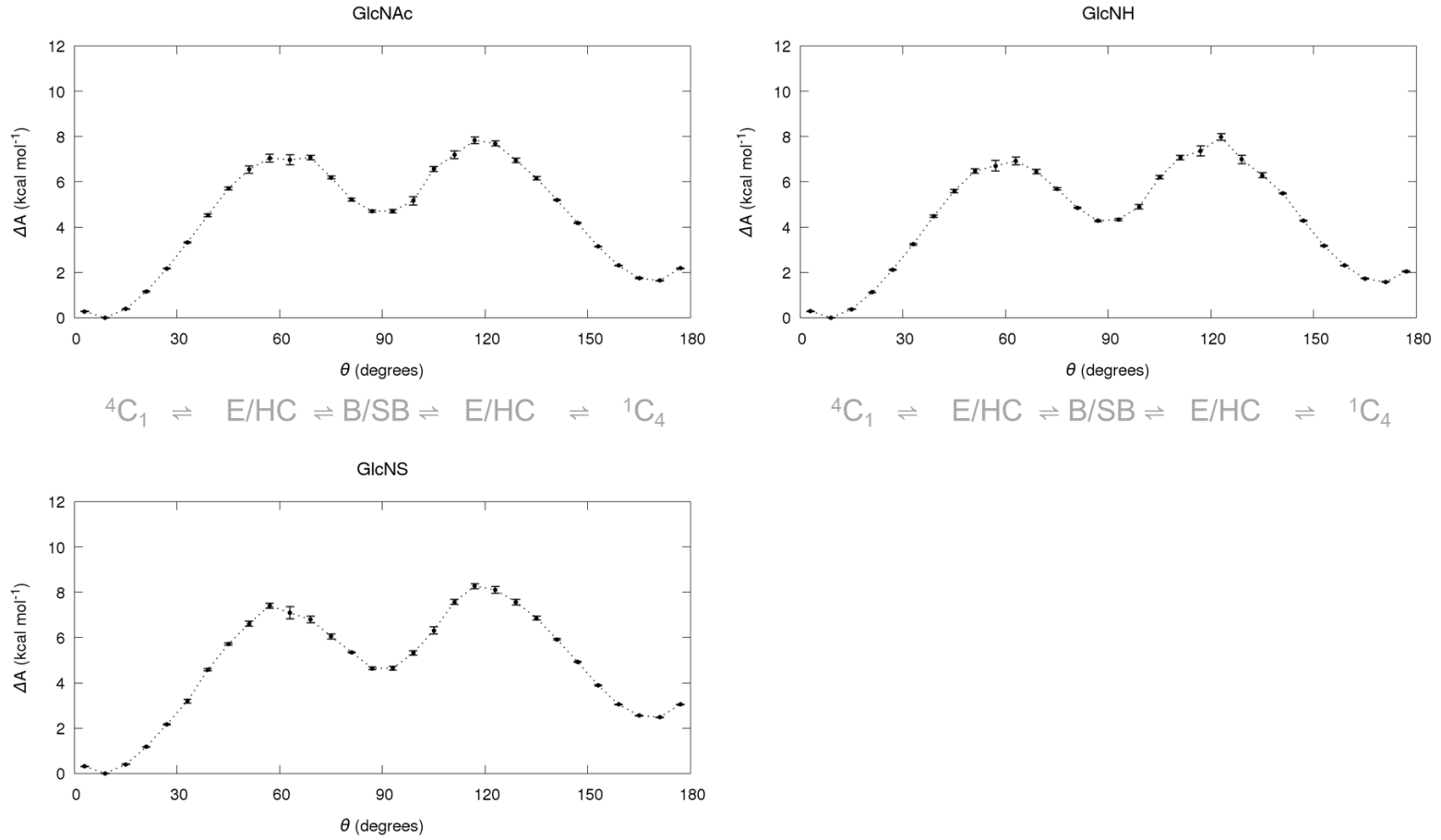

Figure C.7 Bootstrap error analysis of the Cremer-Pople θ free energy profiles of the varying N substitutions of glucosamine (GlcN) calculated via msesMD

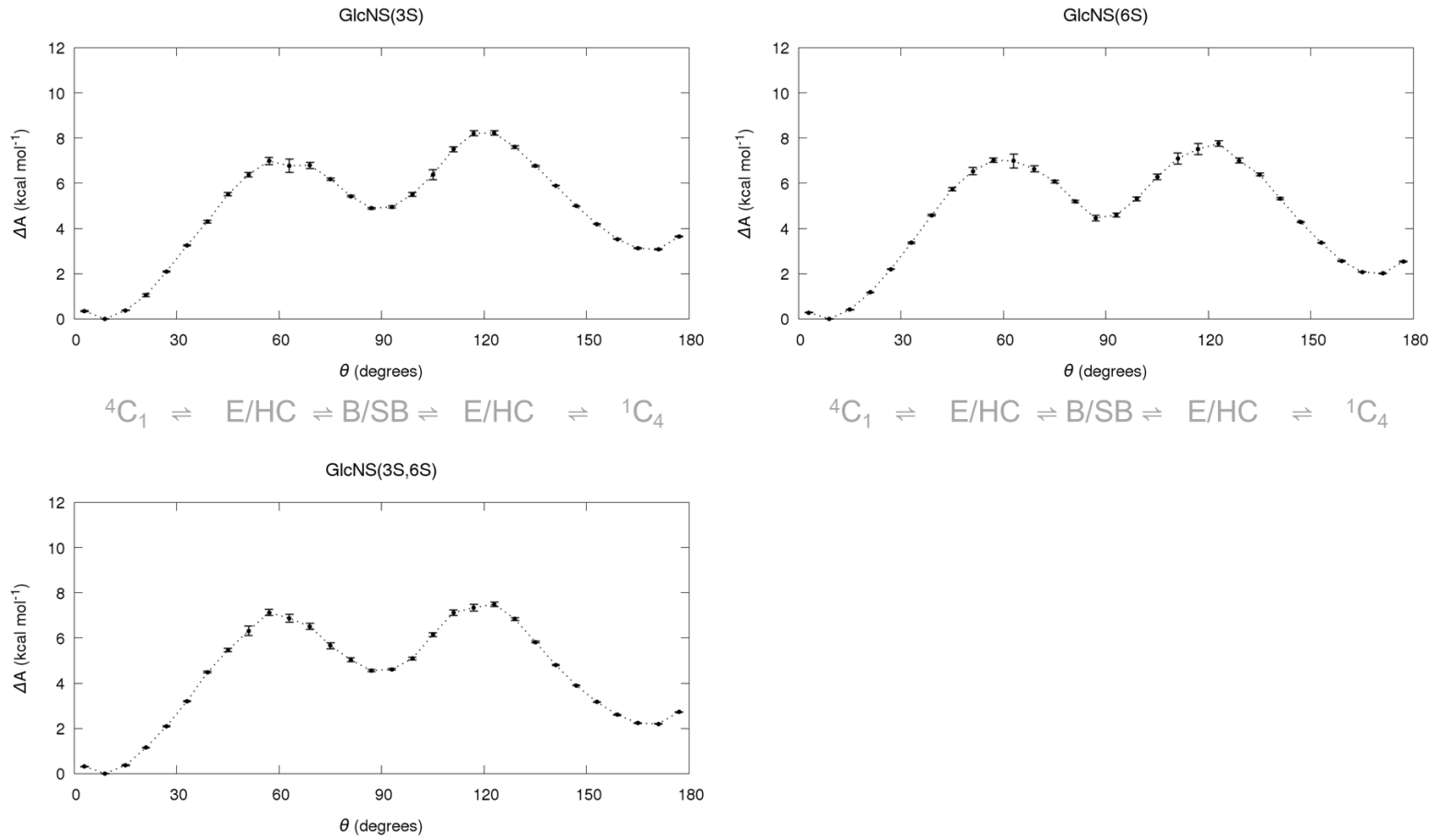

Figure C.8 Bootstrap error analysis of the Cremer-Pople θ free energy profiles of the varying O-sulfation patterns of N-sulfo-glucosamine (GlcNS) calculated via msesMD

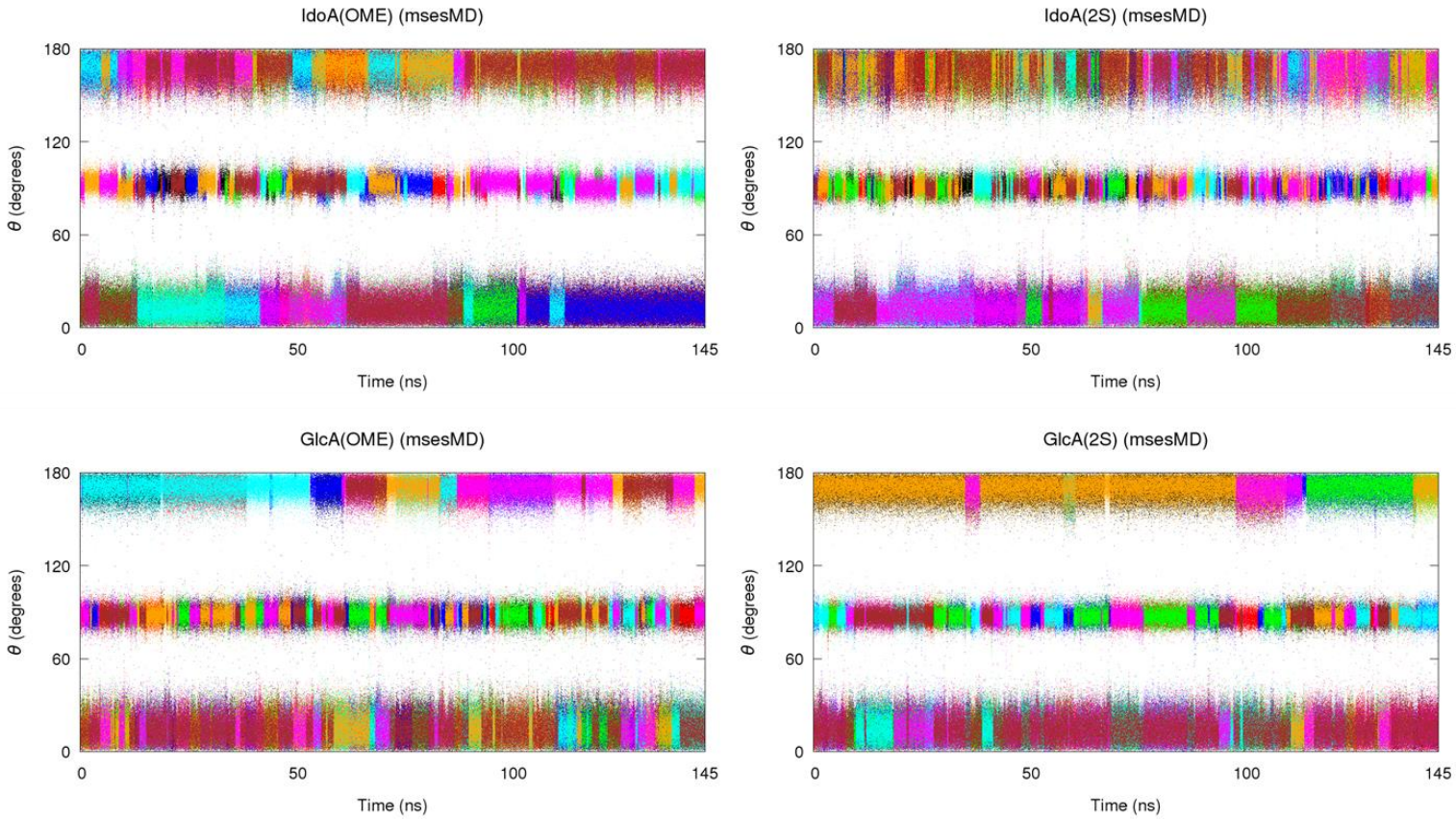

Figure C.9 Change in the Cremer-Pople θ angle over msesMD simulation time for all 8 replica (individually coloured) for the O-methylated (OME) and 2-O-sulfated (2S) uronic acid

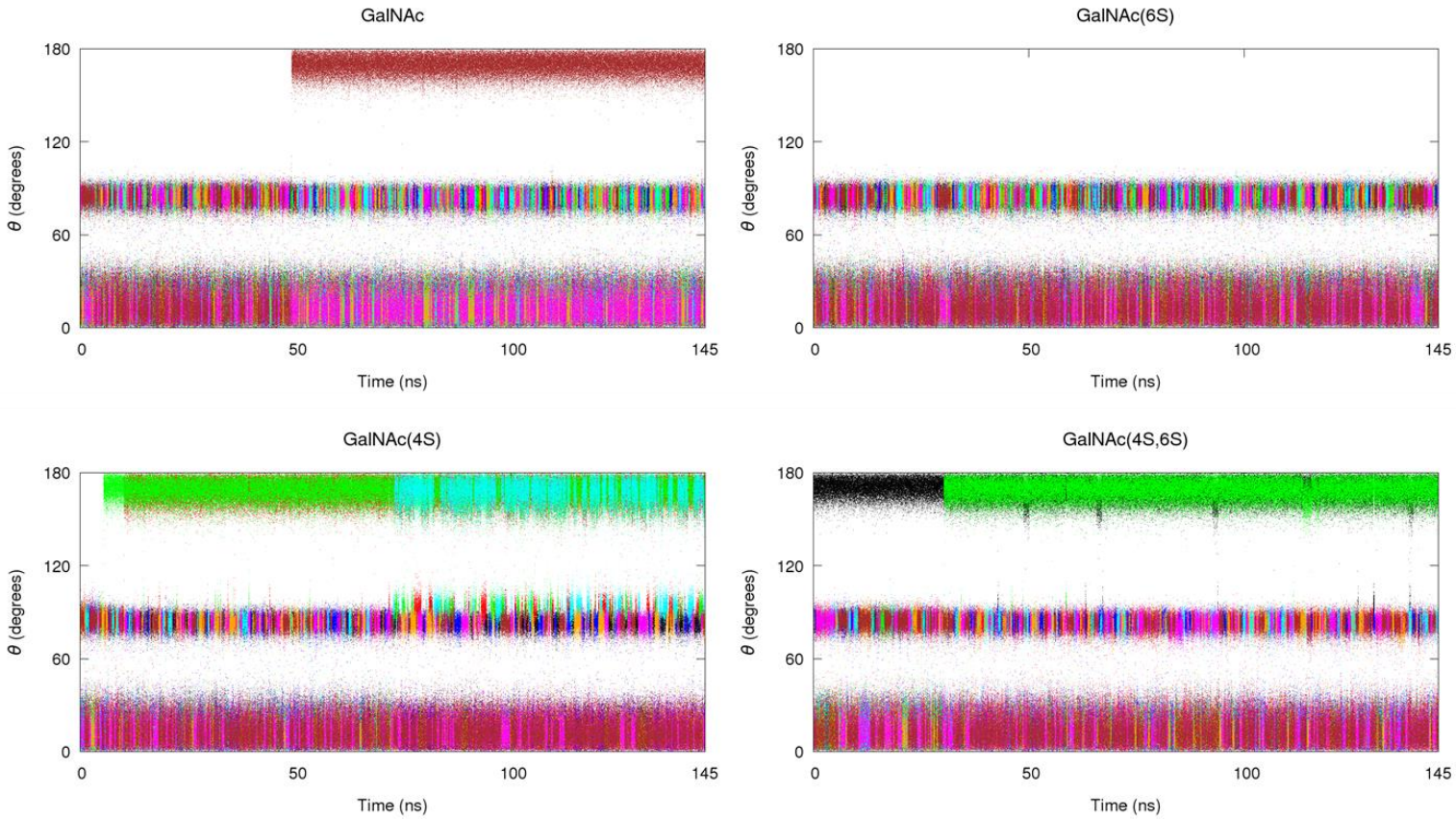

Figure C.10 Change in the Cremer-Pople θ angle over msesMD simulation time for all 8 replica (individually coloured) for the different decoration patterns of N-Acetyl-galactosamine (GalNAc)

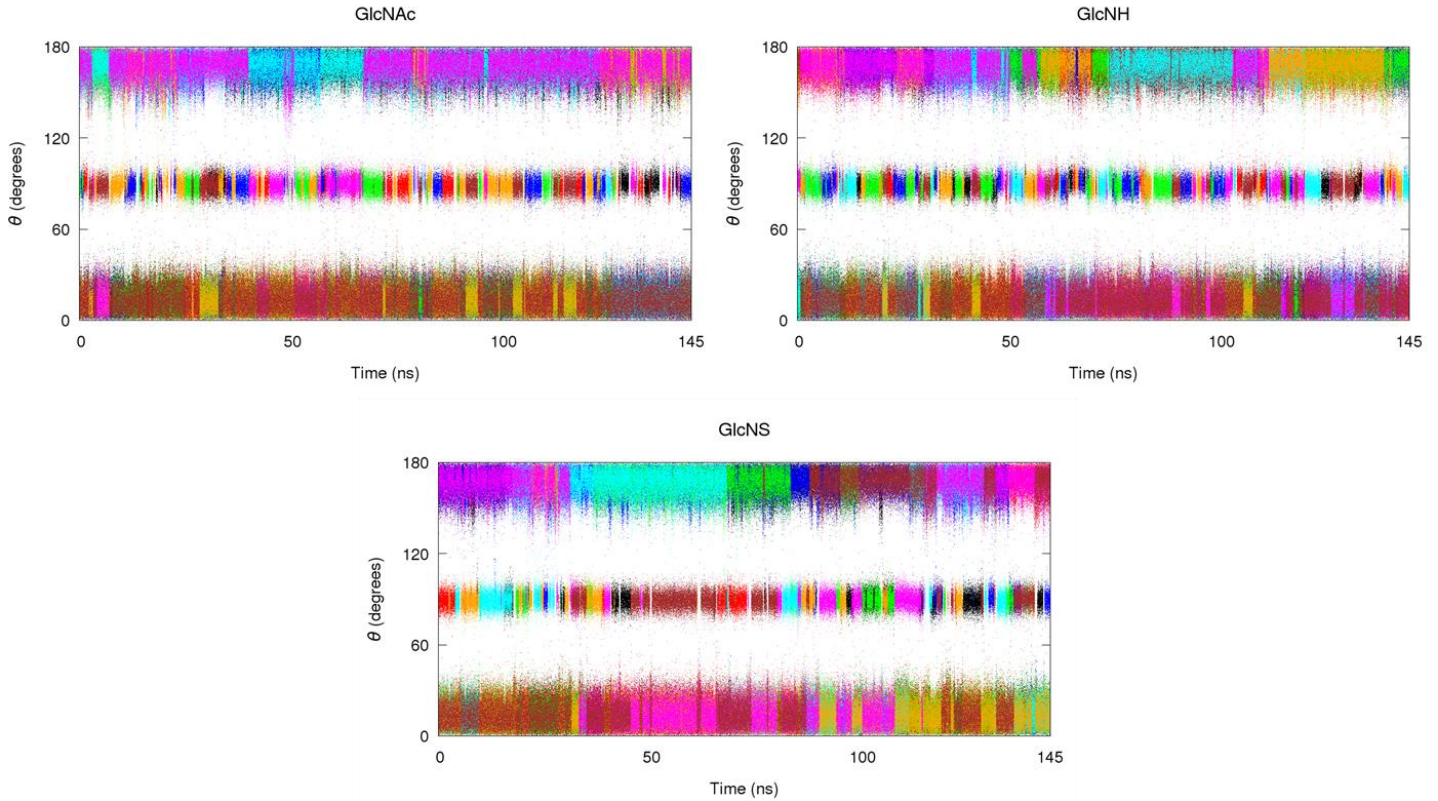

Figure C.11 Change in the Cremer-Pople θ angle over msesMD simulation time for all 8 replica (individually coloured) for the different N substitutions of glucosamine (GlcN)

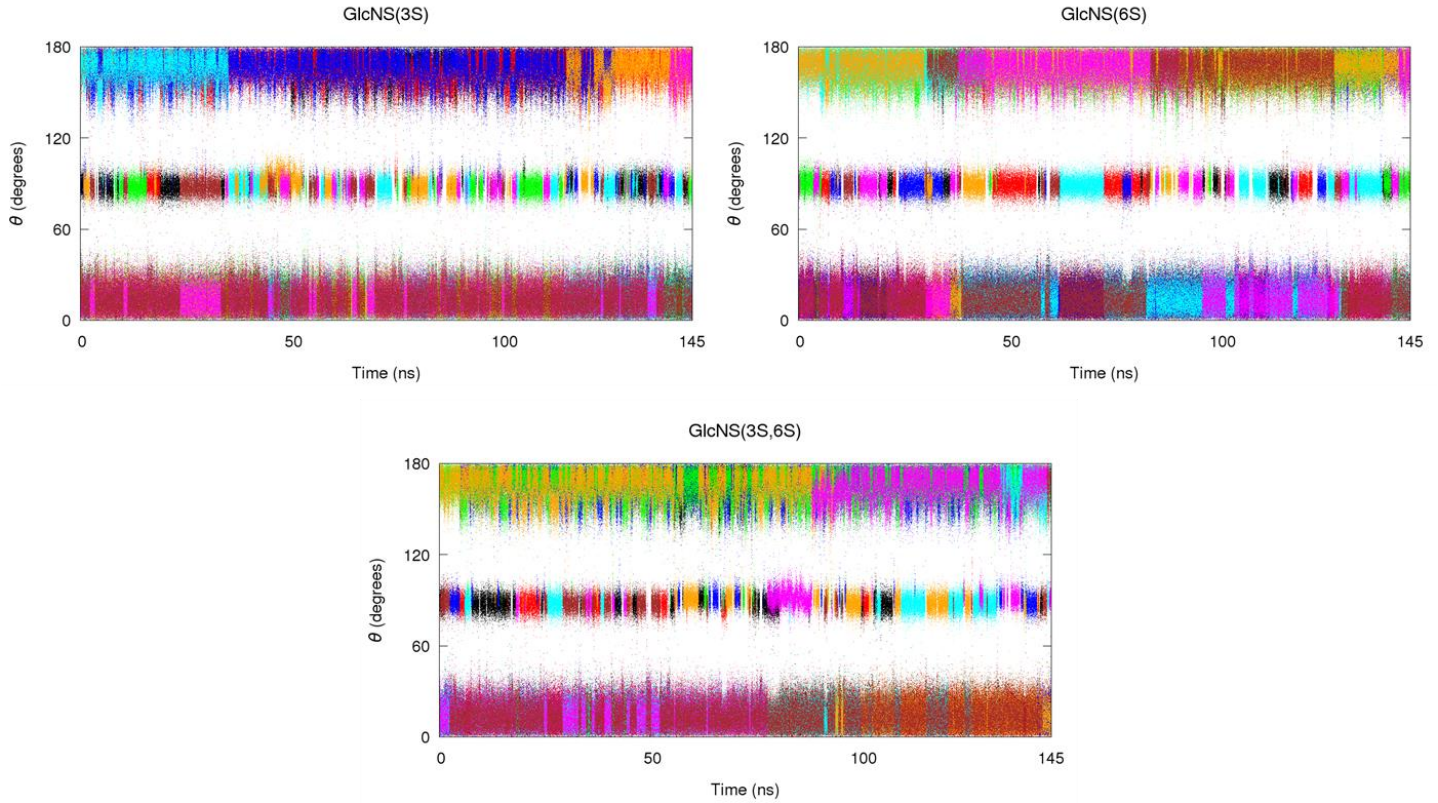

Figure C.12 Change in the Cremer-Pople θ angle over msesMD simulation time for all 8 replica (individually coloured) for the different O-sulfation patterns of N-sulfo-glucosamine (GlcNS)

Appendix D

Table D.1 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing 5 ns per replica IT-TI and msesTI with both calculated (calc) and experimental (exp) FreeSolv database results

Molecule	Charges	IT-TI		msesTI		FreeSolv$_{calc}$		FreeSolv$_{exp}$	
1	*BCC*	-3.01	±0.02	-3.03	±0.04	-3.23	±0.03	-4.72	±0.60
	RESP	-4.78	±0.02	-4.83	±0.03	-	-	-	-
2	*BCC*	-3.61	±0.02	-3.57	±0.02	-3.29	±0.02	-5.03	±0.60
	RESP	-3.84	±0.02	-3.85	±0.02	-	-	-	-
3	*BCC*	-10.53	±0.03	-10.52	±0.06	-10.14	±0.04	-13.43	±1.00
	RESP	-13.29	±0.04	-13.37	±0.06	-	-	-	-
4	*BCC*	-4.67	±0.02	-4.78	±0.14	-3.94	±0.03	-6.4	±0.60
	RESP	-4.86	±0.02	-5.25	±0.11	-	-	-	-
5	*BCC*	-4.34	±0.03	-4.14	±0.12	-3.98	±0.04	-5.73	±0.15
	RESP	-4.65	±0.04	-4.91	±0.14	-	-	-	-
6	*BCC*	-19.39	±0.08	-19.21	±0.22	-18.16	±0.09	-23.62	±0.32
	RESP	-23.47	±0.11	-23.83	±0.26	-	-	-	-
7	*BCC*	-13.31	±0.05	-13.12	±0.32	-11.19	±0.06	-8.15	±0.21
	RESP	-13.83	±0.04	-13.66	±0.39	-	-	-	-

Table D.2 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing 1 ns per replica IT-TI and msesTI with both calculated (calc) and experimental (exp) FreeSolv database results

Molecule	Charges	IT-TI		msesTI		FreeSolv$_{calc}$		FreeSolv$_{exp}$	
1	*BCC*	-3.05	±0.04	-3.04	±0.07	-3.23	±0.03	-4.72	±0.60
	RESP	-4.74	±0.04	-4.87	±0.08	-	-	-	-
2	*BCC*	-3.64	±0.04	-3.56	±0.05	-3.29	±0.02	-5.03	±0.60
	RESP	-3.85	±0.03	-3.94	±0.04	-	-	-	-
3	*BCC*	-10.52	±0.07	-10.79	±0.10	-10.14	±0.04	-13.43	±1.00
	RESP	-13.42	±0.08	-13.27	±0.13	-	-	-	-
4	*BCC*	-4.70	±0.05	-4.41	±0.22	-3.94	±0.03	-6.4	±0.60
	RESP	-4.80	±0.05	-5.08	±0.22	-	-	-	-
5	*BCC*	-4.44	±0.07	-4.18	±0.21	-3.98	±0.04	-5.73	±0.15
	RESP	-4.77	±0.08	-4.72	±0.21	-	-	-	-
6	*BCC*	-19.44	±0.11	-19.51	±0.36	-18.16	±0.09	-23.62	±0.32
	RESP	-23.36	±0.16	-23.85	±0.39	-	-	-	-
7	*BCC*	-13.27	±0.11	-12.84	±0.46	-11.19	±0.06	-8.15	±0.21
	RESP	-13.83	±0.09	-13.42	±0.45	-	-	-	-

Table D.3 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing 500 ps per replica IT-TI and msesTI with both calculated (calc) and experimental (exp) FreeSolv database results

Molecule	Charges	IT-TI		msesTI		FreeSolv$_{calc}$		FreeSolv$_{exp}$	
1	*BCC*	-3.00	±0.06	-3.08	±0.10	-3.23	±0.03	-4.72	±0.60
	RESP	-4.70	±0.06	-4.85	±0.10	-	-	-	-
2	*BCC*	-3.65	±0.05	-3.54	±0.07	-3.29	±0.02	-5.03	±0.60
	RESP	-3.86	±0.05	-4.05	±0.06	-	-	-	-
3	*BCC*	-10.53	±0.10	-10.79	±0.13	-10.14	±0.04	-13.43	±1.00
	RESP	-13.55	±0.11	-13.46	±0.18	-	-	-	-
4	*BCC*	-4.62	±0.06	-4.27	±0.27	-3.94	±0.03	-6.4	±0.60
	RESP	-4.86	±0.07	-5.04	±0.24	-	-	-	-
5	*BCC*	-4.43	±0.09	-4.08	±0.25	-3.98	±0.04	-5.73	±0.15
	RESP	-4.81	±0.12	-4.64	±0.25	-	-	-	-
6	*BCC*	-19.67	±0.17	-19.27	±0.36	-18.16	±0.09	-23.62	±0.32
	RESP	-23.46	±0.20	-24.18	±0.49	-	-	-	-
7	*BCC*	-13.18	±0.15	-12.38	±0.45	-11.19	±0.06	-8.15	±0.21
	RESP	-13.93	±0.11	-13.11	±0.53	-	-	-	-

Table D.4 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 5 ns per replica "independent replica" (IndRw) and "group" (GroupRw) msesTI reweighting results

Molecule	Charges	msesTI (IndRw)		msesTI (GroupRw)		FreeSolv$_{exp}$	
1	*BCC*	-3.03	±0.04	-3.03	±0.04	-4.72	±0.60
	RESP	-4.83	±0.03	-4.83	±0.04	-	-
2	*BCC*	-3.57	±0.00	-3.57	±0.02	-5.03	±0.60
	RESP	-3.85	±0.02	-3.85	±0.02	-	-
3	*BCC*	-10.52	±0.06	-10.52	±0.06	-13.43	±1.00
	RESP	-13.37	±0.06	-13.37	±0.07	-	-
4	*BCC*	-4.78	±0.14	-4.87	±0.16	-6.4	±0.60
	RESP	-5.25	±0.11	-5.34	±0.17	-	-
5	*BCC*	-4.14	±0.12	-3.98	±0.18	-5.73	±0.15
	RESP	-4.91	±0.14	-4.98	±0.21	-	-
6	*BCC*	-19.21	±0.22	-19.19	±0.27	-18.16	±0.09
	RESP	-23.83	±0.26	-23.94	±0.36	-	-
7	*BCC*	-13.12	±0.32	-13.20	±0.58	-8.15	±0.21
	RESP	-13.66	±0.39	-13.78	±0.71	-	-

Table D.5 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 1 ns per replica "independent replica" (IndRw) and "group" (GroupRw) msesTI reweighting results

Molecule	Charges	msesTI (IndRw)		msesTI (GroupRw)		FreeSolv$_{exp}$	
1	*BCC*	-3.04	±0.07	-3.04	±0.08	-4.72	±0.60
	RESP	-4.87	±0.08	-4.86	±0.08	-	-
2	*BCC*	-3.56	±0.05	-3.55	±0.05	-5.03	±0.60
	RESP	-3.94	±0.04	-3.95	±0.05	-	-
3	*BCC*	-10.79	±0.10	-10.78	±0.12	-13.43	±1.00
	RESP	-13.27	±0.13	-13.30	±0.14	-	-
4	*BCC*	-4.41	±0.22	-4.55	±0.34	-6.4	±0.60
	RESP	-5.08	±0.22	-5.14	±0.28	-	-
5	*BCC*	-4.18	±0.21	-3.84	±0.37	-5.73	±0.15
	RESP	-4.72	±0.21	-4.78	±0.32	-	-
6	*BCC*	-19.51	±0.36	-19.80	±0.66	-18.16	±0.09
	RESP	-23.85	±0.39	-24.18	±0.55	-	-
7	*BCC*	-12.84	±0.46	-14.07	±0.87	-8.15	±0.21
	RESP	-13.42	±0.45	-14.34	±1.01	-	-

Table D.6 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 500 ps per replica "independent replica" (IndRw) and "group" (GroupRw) msesTI reweighting results

Molecule	Charges	msesTI (IndRw)		msesTI (GroupRw)		FreeSolv$_{exp}$	
1	*BCC*	-3.08	±0.10	-3.09	±0.11	-4.72	±0.60
	RESP	-4.85	±0.10	-4.87	±0.11	-	-
2	*BCC*	-3.54	±0.07	-3.53	±0.07	-5.03	±0.60
	RESP	-4.05	±0.06	-4.05	±0.07	-	-
3	*BCC*	-10.79	±0.13	-10.78	±0.16	-13.43	±1.00
	RESP	-13.46	±0.18	-13.49	±0.19	-	-
4	*BCC*	-4.27	±0.27	-4.34	±0.48	-6.4	±0.60
	RESP	-5.04	±0.24	-5.19	±0.34	-	-
5	*BCC*	-4.08	±0.25	-3.94	±0.43	-5.73	±0.15
	RESP	-4.64	±0.25	-4.75	±0.38	-	-
6	*BCC*	-19.27	±0.36	-19.86	±0.72	-18.16	±0.09
	RESP	-24.18	±0.49	-24.78	±0.76	-	-
7	*BCC*	-12.38	±0.45	-12.46	±0.96	-8.15	±0.21
	RESP	-13.11	±0.53	-14.48	±1.16	-	-

Table D.7 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 5 ns per replica msesTI results using the "Half" and "Full" boost parameter sets

Molecule	Charges	msesTI (Half)		msesTI (Full)		FreeSolv$_{exp}$	
1	*BCC*	-3.03	±0.04	-3.05	±0.10	-4.72	±0.60
	RESP	-4.83	±0.03	-4.76	±0.10	-	-
2	*BCC*	-3.57	±0.02	-3.59	±0.04	-5.03	±0.60
	RESP	-3.85	±0.02	-3.81	±0.03	-	-
3	*BCC*	-10.52	±0.06	-10.55	±0.12	-13.43	±1.00
	RESP	-13.37	±0.06	-13.45	±0.11	-	-
4	*BCC*	-4.78	±0.14	-4.69	±0.22	-6.4	±0.60
	RESP	-5.25	±0.11	-5.35	±0.18	-	-
5	*BCC*	-4.14	±0.12	-4.07	±0.23	-5.73	±0.15
	RESP	-4.91	±0.14	-5.00	±0.28	-	-
6	*BCC*	-19.21	±0.22	-19.76	±0.36	-18.16	±0.09
	RESP	-23.83	±0.26	-24.48	±0.37	-	-
7	*BCC*	-13.12	±0.32	-13.36	±0.46	-8.15	±0.21
	RESP	-13.66	±0.39	-13.73	±0.50	-	-

Table D.8 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 1 ns per replica msesTI results using the "Half" and "Full" boost parameter sets

Molecule	Charges	msesTI (Half)		msesTI (Full)		FreeSolv$_{exp}$	
1	*BCC*	-3.04	±0.07	-2.94	±0.14	-4.72	±0.60
	RESP	-4.87	±0.08	-4.94	±0.17	-	-
2	*BCC*	-3.56	±0.05	-3.58	±0.08	-5.03	±0.60
	RESP	-3.94	±0.04	-3.83	±0.06	-	-
3	*BCC*	-10.79	±0.10	-10.50	±0.19	-13.43	±1.00
	RESP	-13.27	±0.13	-13.46	±0.18	-	-
4	*BCC*	-4.41	±0.22	-4.52	±0.27	-6.4	±0.60
	RESP	-5.08	±0.22	-4.59	±0.24	-	-
5	*BCC*	-4.18	±0.21	-4.49	±0.30	-5.73	±0.15
	RESP	-4.72	±0.21	-4.91	±0.29	-	-
6	*BCC*	-19.51	±0.36	-19.46	±0.41	-18.16	±0.09
	RESP	-23.85	±0.39	-23.85	±0.48	-	-
7	*BCC*	-12.84	±0.46	-12.90	±0.50	-8.15	±0.21
	RESP	-13.42	±0.45	-13.71	±0.51	-	-

Table D.9 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 500 ps per replica msesTI results using the "Half" and "Full" boost parameter sets

Molecule	Charges	msesTI (Half)		msesTI (Full)		FreeSolv$_{exp}$	
1	*BCC*	-3.08	±0.10	-3.13	±0.18	-4.72	±0.60
	RESP	-4.85	±0.10	-4.98	±0.19	-	-
2	*BCC*	-3.54	±0.07	-3.50	±0.12	-5.03	±0.60
	RESP	-4.05	±0.06	-3.77	±0.09	-	-
3	*BCC*	-10.79	±0.13	-10.70	±0.20	-13.43	±1.00
	RESP	-13.46	±0.18	-13.51	±0.22	-	-
4	*BCC*	-4.27	±0.27	-4.65	±0.29	-6.4	±0.60
	RESP	-5.04	±0.24	-5.04	±0.27	-	-
5	*BCC*	-4.08	±0.25	-4.57	±0.31	-5.73	±0.15
	RESP	-4.64	±0.25	-4.74	±0.32	-	-
6	*BCC*	-19.27	±0.36	-19.75	±0.43	-18.16	±0.09
	RESP	-24.18	±0.49	-23.62	±0.54	-	-
7	*BCC*	-12.38	±0.45	-13.14	±0.54	-8.15	±0.21
	RESP	-13.11	±0.53	-13.24	±0.52	-	-

Table D.10 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 5 ns per replica IT-TI results using normal atomic masses (NORM) and hydrogen mass repartitioning (HMR)

Molecule	Charges	IT-TI (NORM)		IT-TI (HMR)		FreeSolv$_{exp}$	
1	*BCC*	-3.01	±0.02	-2.86	±0.02	-4.72	±0.60
	RESP	-4.78	±0.02	-4.33	±0.02	-	-
2	*BCC*	-3.61	±0.02	-3.82	±0.01	-5.03	±0.60
	RESP	-3.84	±0.02	-4.06	±0.02	-	-
3	*BCC*	-10.53	±0.03	-10.90	±0.03	-13.43	±1.00
	RESP	-13.29	±0.04	-13.68	±0.03	-	-
4	*BCC*	-4.67	±0.02	-5.03	±0.03	-6.4	±0.60
	RESP	-4.86	±0.02	-5.16	±0.02	-	-
5	*BCC*	-4.34	±0.03	-3.76	±0.04	-5.73	±0.15
	RESP	-4.65	±0.04	-3.85	±0.04	-	-
6	*BCC*	-19.39	±0.08	-19.09	±0.08	-18.16	±0.09
	RESP	-23.47	±0.11	-23.29	±0.13	-	-
7	*BCC*	-13.31	±0.05	-14.09	±0.06	-8.15	±0.21
	RESP	-13.83	±0.04	-11.36	±0.06	-	-

Table D.11 Solvation free energies (kcal mol^{-1}) for small molecules 1-7, comparing the 5 ns per replica hydrogen mass repartitioning (HMR) IT-TI and msesTI results

Molecule	Charges	IT-TI (HMR)		msesTI (HMR)		FreeSolv$_{exp}$	
1	*BCC*	-2.86	±0.02	-2.77	±0.04	-4.72	±0.60
	RESP	-4.33	±0.02	-4.35	±0.04	-	-
2	*BCC*	-3.82	±0.01	-3.78	±0.02	-5.03	±0.60
	RESP	-4.06	±0.02	-4.07	±0.02	-	-
3	*BCC*	-10.90	±0.03	-10.93	±0.05	-13.43	±1.00
	RESP	-13.68	±0.03	-13.72	±0.06	-	-
4	*BCC*	-5.03	±0.03	-5.18	±0.12	-6.4	±0.60
	RESP	-5.16	±0.02	-5.22	±0.08	-	-
5	*BCC*	-3.76	±0.04	-3.71	±0.15	-5.73	±0.15
	RESP	-3.85	±0.04	-3.94	±0.10	-	-
6	*BCC*	-19.09	±0.08	-19.06	±0.21	-18.16	±0.09
	RESP	-23.29	±0.13	-23.68	±0.21	-	-
7	*BCC*	-14.09	±0.06	-14.34	±0.32	-8.15	±0.21
	RESP	-11.36	±0.06	-10.82	±0.34	-	-

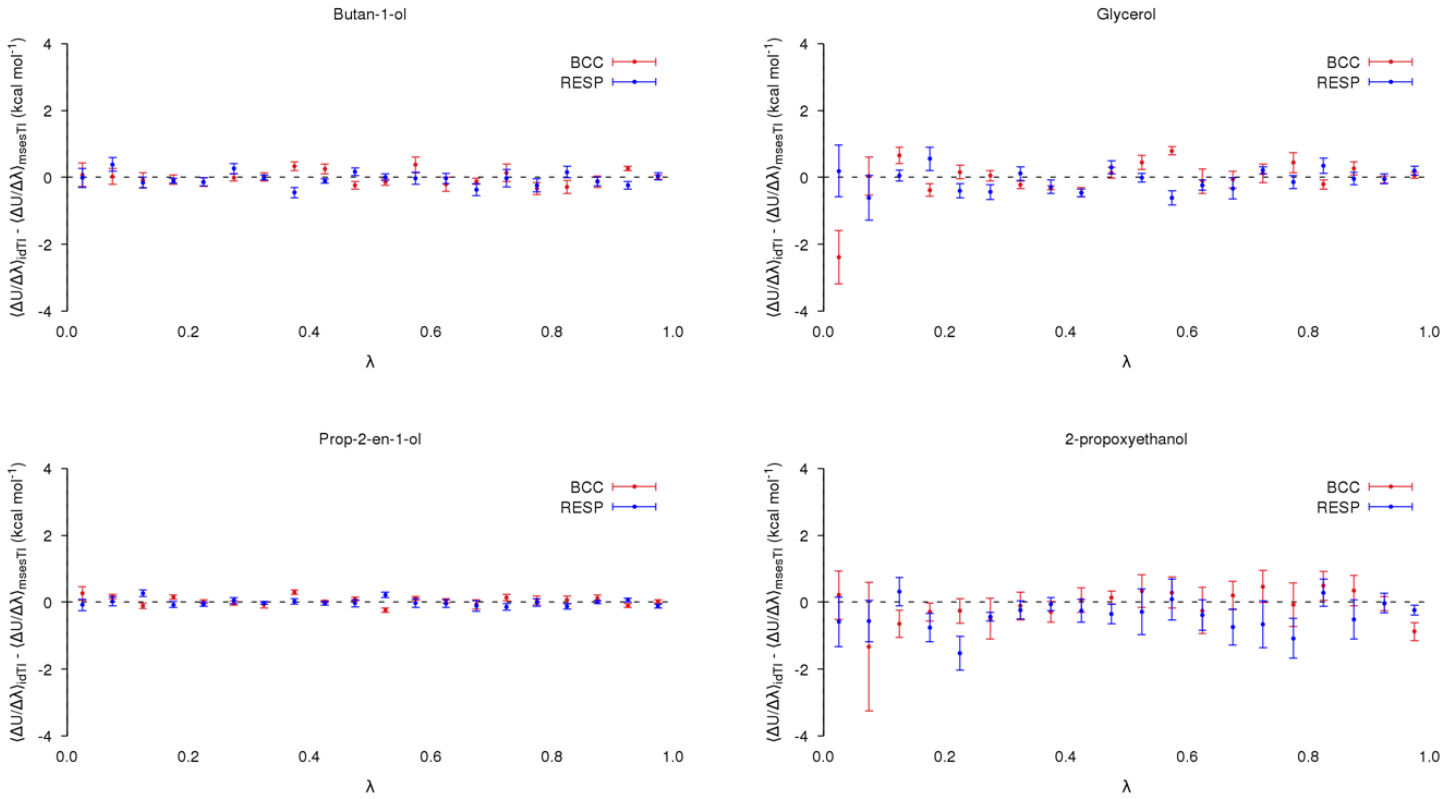

Figure D.1 Differences in the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_{\lambda}$ profiles between the 5 ns per replica IT-TI and msesTI simulations using both AM1-BCC and RESP partial charge assignments for; butan-1-ol, prop-2-en-1-ol, glycerol and 2-propoxyethanol

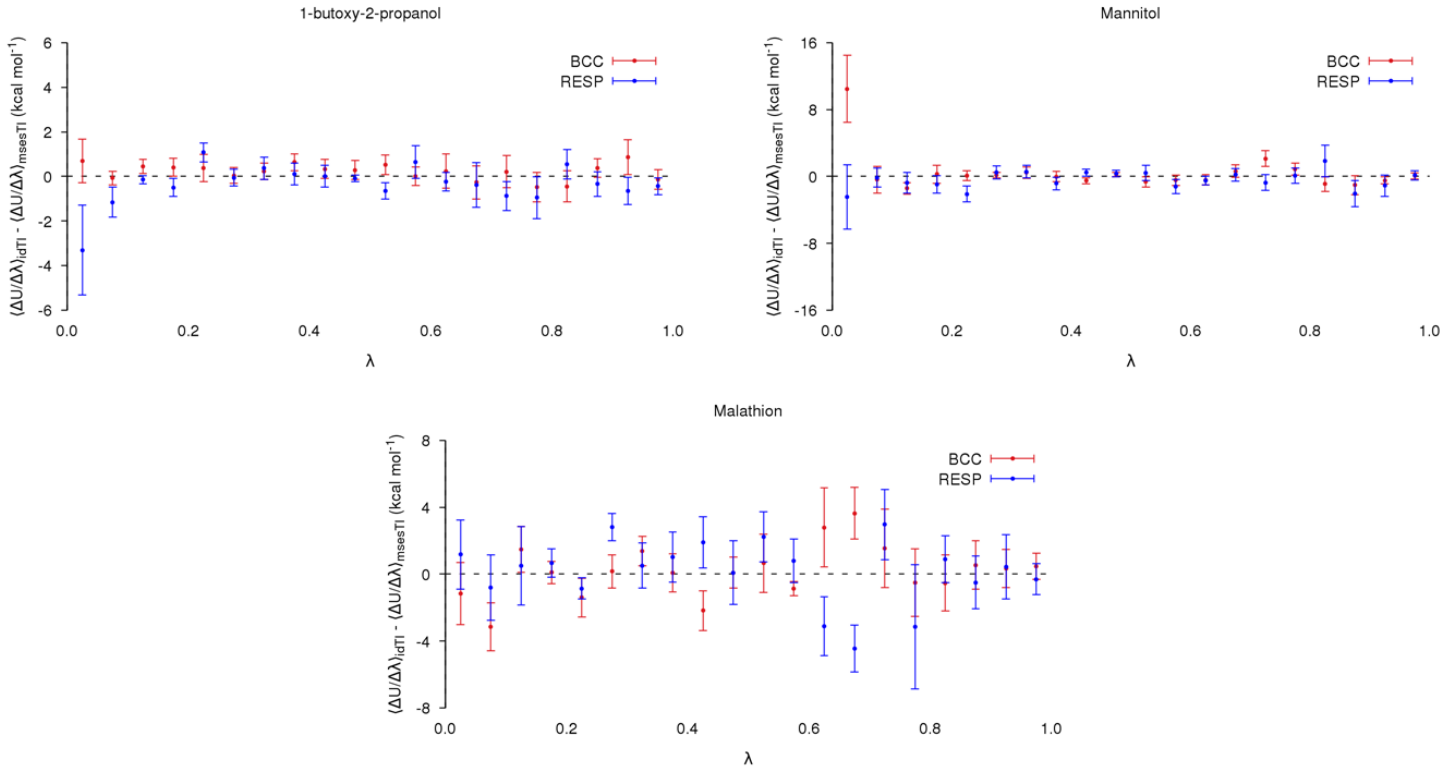

Figure D.2 Differences in the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_{\lambda}$ profiles between the 5 ns per replica IT-TI and msesTI simulations using both AM1-BCC and RESP partial charge assignments for; 1-butoxy-2-propanol, mannitol and malathion

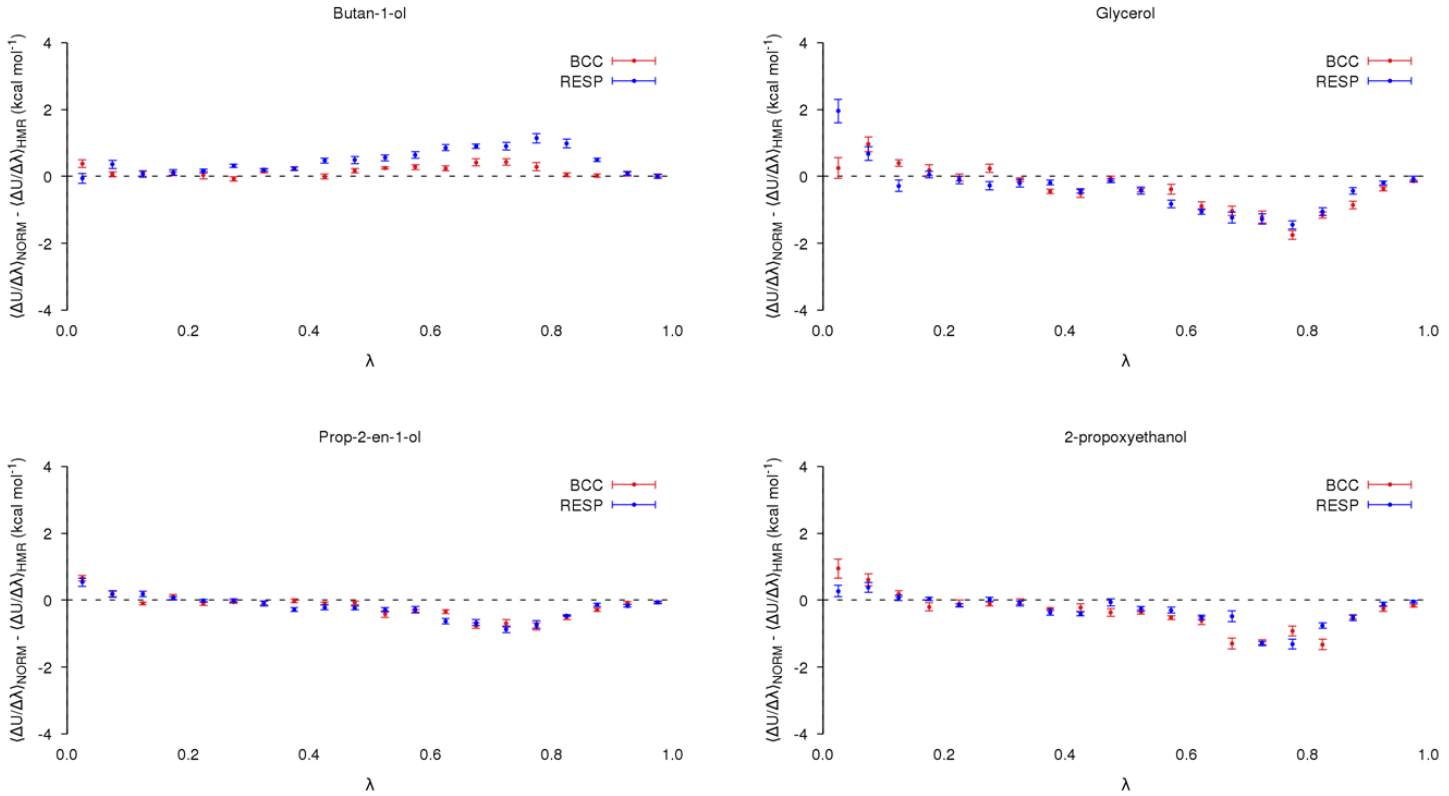

Figure D.3 Difference in the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_{\lambda}$ profiles between the 5 ns per replica IT-TI normal mass (NORM) and hydrogen mass repartitioning (HMR) simulations using both AM1-BCC and RESP partial charge assignments for; butan-1-ol, prop-2-en-1-ol, glycerol, and 2-propoxyethanol

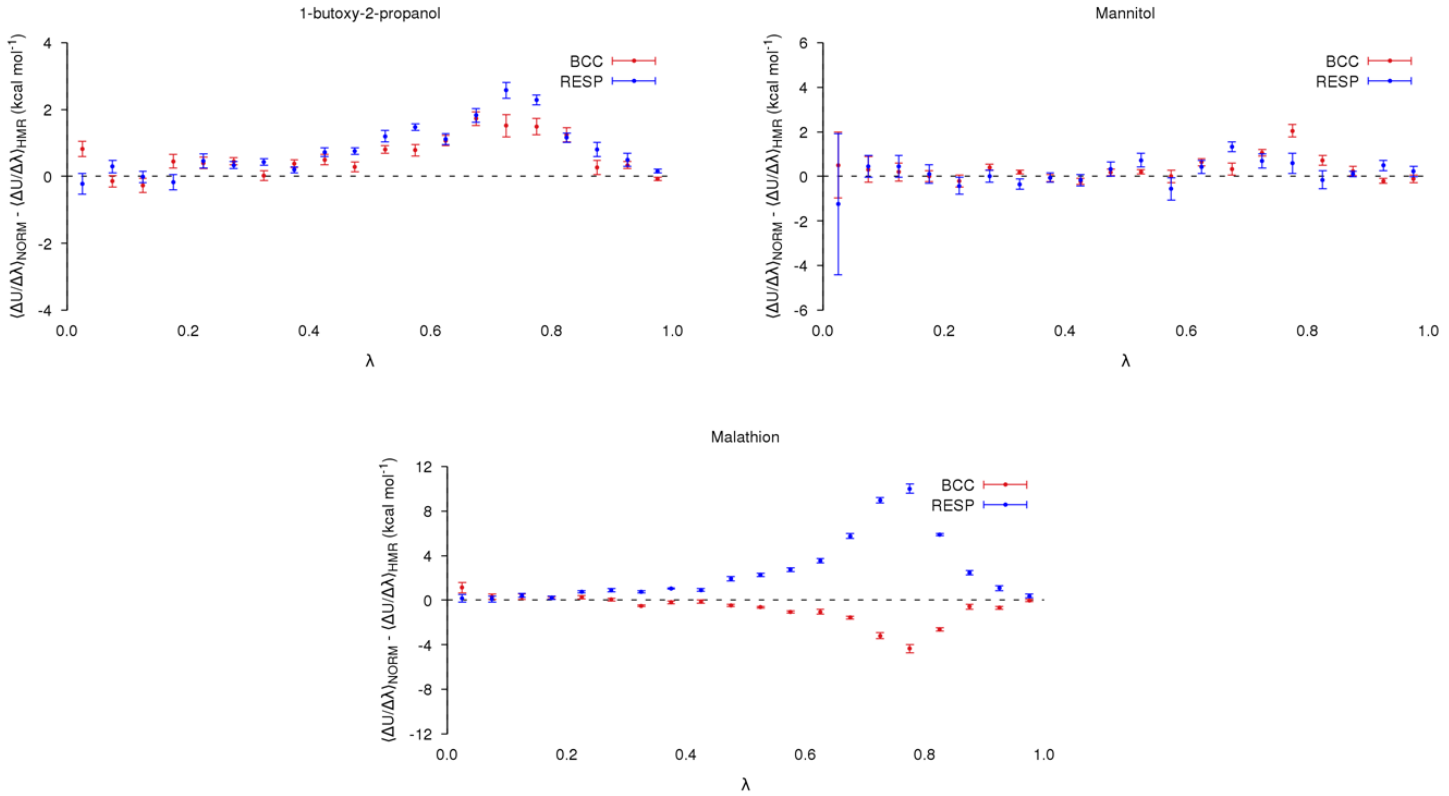

Figure D.4 Difference in the $\langle \frac{\delta U(r,\lambda)}{\delta \lambda} \rangle_{\lambda}$ profiles between the 5 ns per replica IT-TI normal mass (NORM) and hydrogen mass repartitioning (HMR) simulations using both AM1-BCC and RESP partial charge assignment for; 1-butoxy-2-propanol, mannitol and malathion