APPLICATION OF CLASSICAL DYNAMICS SIMULATIONS TO INVESTIGATE CELLULOSE MICROFIBRIL TWISTING AND INFLUENZA H1 BINDING AFFINITY

by

JODI ANN HADDEN

(Under the Direction of Robert J. Woods)

ABSTRACT

As the most abundant biopolymer on Earth, cellulose serves as an important raw material for many industries, including paper, textiles, and more recently, biofuels. While high-resolution crystallographic data suggest that cellulose microfibrils occur as linearly oriented assemblies of cellulose chains, computational simulations predict a twisted structure. Through investigation of commonly employed theoretical approximations, this work establishes the physical origin of twisting behavior, indicating that it arises from a balance of competing forces. Overall, twisting appears to be driven by attractive van der Waals interactions, while mitigated by both the cellulose intrachain hydrogen bond network and solvent effects at the microfibril surface. As a result, modeling of simulated microfibrils is sensitive to monomeric charge distribution, solvent model, and the application of dummy atoms to mimic the influence of electron lone pairs. Further, analysis of back-calculated diffraction patterns for twisted and linear structures demonstrates that powder diffraction methodology cannot detect subtle twisting in cellulose samples, raising the possibility that crystals employed to resolve the original crystallographic coordinates could have incorporated twisting.

Adhesion of the influenza A virus is mediated by its primary surface antigen, hemagglutinin, which recognizes receptor glycans terminating in sialylated galactose. Host range is

determined by specificity for the glycosidic linkage displayed within this disaccharide motif, with avian viruses preferring α2-3 linkages and human viruses preferring α2-6 linkages. While experimental characterization of specificity is relatively straightforward, quantification of associated binding affinity represents a challenge due to the inherent multimeric nature of hemagglutinin structure. Through application of computational simulations, which allow investigation of a monomeric binding domain, this work computes highly accurate binding free energies associated with specificity determinants in the H1 hemagglutinin. Results include quantification of the effects of abrogating and specificity-altering point mutations, as well as avian and human receptor glycan contributions to binding affinity. Altogether, these data likely provide the most reasonable theoretical quantifications of binding currently available for the H1 system.

INDEX WORDS: classical dynamics, molecular dynamics simulation, thermodynamic integration, cellulose, microfibril twisting, powder diffraction, influenza hemagglutinin, binding affinity, free energy, GLYCAM

APPLICATION OF CLASSICAL DYNAMICS SIMULATIONS TO INVESTIGATE CELLULOSE MICROFIBRIL TWISTING AND INFLUENZA H1 BINDING AFFINITY

by

JODI ANN HADDEN

B.S., Armstrong Atlantic State University, 2007

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

Athens, Georgia

2014

APPLICATION OF CLASSICAL DYNAMICS SIMULATIONS TO INVESTIGATE CELLULOSE MICROFIBRIL TWISTING AND INFLUENZA H1 BINDING AFFINITY

by

JODI ANN HADDEN

Approved:

Major Professor: Robert J. Woods

Committee: Jeffrey L. Urbauer
I. Jonathan Amster
S. Mark Tompkins

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2014

DEDICATION

This work is dedicated to my loving parents, who always encouraged me to go as far as I could academically. I love them both more than words can say.

> Everything that living things do can be understood in terms of the jiggling and wiggling of atoms.
>
> — Richard P. Feynman

ACKNOWLEDGEMENTS

My sincerest gratitude goes out to the following people:

My late father, who long ago impressed upon me the value of hard work, and whose work ethic in the face of serious illness was a continual inspiration to me. He dedicated his life to ensuring I could have the brightest possible future, as well as the financial freedom to pursue my dreams. I like to think I have made him supremely proud by earning this Ph.D. (Thanks for everything, Daddy. I love you, I miss you, and I hope I grow up to be just like you.)

My mother, who endured countless scientific conversations during which I mulled over the challenges of my research. Only a phone call away, she always listened patiently, even when she had absolutely no idea what I was talking about. Clearly, I have inherited her remarkable passion, resolve, and strength of character, qualities that have proved indispensable in the course of my long academic journey. I am proud to be her daughter.

My late grandmother, whose adoration and encouragement I still feel all around me like the smile of warm, radiant sunshine. More than anyone else in my life, she convinced me that I was particuarly special and capable of anything. (I love and miss you, too, Nannie.)

My brother, as well as my extended family, whose unconditional love and support have ever been the wind beneath my wings. Even during the times when I struggled academically, they never had a doubt, and not one ever uttered a single negative or discouraging remark. Everything I've achieved has been made possible because they believed in me.

"Uncle" Al and Mary An, who have shown me kindess far beyond that warrented by a research collaboration, and have thus endeared themselves to me as treasured friends.

Matt, Lachele, Mark, Kausar, Xiao, Keigo, the *Desis*, and the rest of my Woods lab colleagues, with whom I have shared a unique relationship, far beyond friendship, forged in the crucible of my graduate school experience. Without them, I would never have made it this far, neither in my professional nor personal life. They each have my undying love and respect, and I wish them all the best of luck with their futures.

Professor Robert J. Woods, under whose advisement I have learned to be a rigorous, independent, and confident young scientist. No doubt the expertise I acquired and character I developed during my time in his lab will take me very far in life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:
INTRODUCTION

This dissertation comprises two separate research topics, applying classical dynamics simulations to investigate:

1. Cellulose microfibril twisting
2. Influenza H1 binding affinity

These topics, including review of their respective background information, relevant literature studies, and the computational theory/methods applied to them, are presented as follows:

CHAPTER 2: CARBOHYDRATES AND GLYCANS

Cellulose represents a crystalline assembly of carbohydrate polymers (polysaccharides), while influenza H1 binds host cell receptor glycans to accomplish viral adhesion. Chapter 2 describes the structural and conformational considerations relevant to dynamical study of these types of biomolecules.

CHAPTER 3: MOLECULAR DYNAMICS SIMULATIONS

Both cellulose and influenza H1 are investigated here through application of classical dynamics simulations, commonly referred to as molecular dynamics (MD) simulations. Chapter 3 covers the underlying theory of this computational methodology.

CHAPTER 4: THERMODYNAMIC INTEGRATION CALCULATIONS

Thermodynamic integration (TI) is an alchemical transformation/free energy technique further applied to influenza H1, in conjunction with MD simulation. Chapter 4 covers the theory pertaining to performing this type of calculation.

CHAPTER 5: SYSTEM UNDER STUDY: CELLULOSE

Chapter 5 provides relevant background and literature review of cellulose as a preface to the research studies presented in Chapters 6 and 7.

CHAPTER 6: UNRAVELING CELLULOSE MICROFIBRILS: A TWISTED TALE

Chapter 6 describes original research investigating the twisting behavior of cellulose microfibrils as observed during MD simulation. The results of this study were published in the scientific literature:

Hadden, J. A.; French, A. D.; Woods, R. J. *Biopolymers* **2013**, *99*, 746-756.

CHAPTER 7: EFFECT OF MICROFIBRIL TWISTING ON THEORETICAL POWDER DIFFRACTION PATTERNS OF CELLULOSE Iβ

Chapter 7 describes original research investigating the relationship between twisted cellulose microfibril samples and distortions observed in powder diffraction data. The results of this study were published in the scientific literature:

Hadden, J. A.; French, A. D.; Woods, R. J. *Cellulose* **2014**, *21*, 879-884.

CHAPTER 8: SYSTEM UNDER STUDY: INFLUENZA HEMAGGLUTININ

Chapter 8 provides relevant background and literature review of influenza H1 as a preface to the research study presented in Chapter 9.

CHAPTER 9: QUANTIFYING BINDING AFFINITY AND ITS RELATIONSHIP TO SPECIFICITY IN INFLUENZA H1

Chapter 9 describes original research employing TI calculations to compute highly accurate free energy changes relevant to the influenza H1 adhesion interaction. The results of this study will soon be published in the scientific literature.

CHAPTER 10: CONCLUSIONS AND RECOGNIZED CHALLENGES

In closing, Chapter 10 summarizes the major conclusions of the work, acknowledges challenges and limitations, and provides comment on possible future directions.

CHAPTER 2:
CARBOHYDRATES AND GLYCANS

Carbohydrates are a class of biomolecules that follow the general chemical formula $(C{\cdot}H_2O)_n$, and thus derive their name as hydrates of carbon. They may also be referred to as sugars, saccharides (stemming from the Latin word *saccharum* for sugar) [1], or glycans, specifically when they occur in conjugation with other biomolecules, such as proteins or lipids.

Linear Monosaccharides

The most basic carbohydrate unit is the monosaccharide, which can exist as a linear or cyclic form. In its linear (acyclic) form, a monosaccharide comprises an aldehyde (aldose) or ketone (ketose) whose carbon side chain is substituted with hydroxyl groups (Figure 2.1). With the exception of that on the chain terminal, each hydroxyl-substituted carbon represents a chiral center, resulting in 2^n possible stereoisomers per monosaccharide, where n is the number of chiral or stereocenters (Figure 2.3a). Any two isomers that differ in configuration at a single stereocenter are called epimers. For example, in Figure 2.3a, D-glucose and D-galactose differ in their configuration at C4, and are thus C4 epimers. When two isomers differ in configuration at every carbon position, that is, they form mirror images of each other, they are called enantiomers. The D- and L-series of each monosaccharide are examples of enantiomers (Figure 2.3a and 2.3b). Monosaccharides are classified as D or L depending on the orientation at the stereocenter furthest from the carbonyl group. By convention, D-series sugars have a chiral R configuration at this position, while L-series sugars have a chiral S configuration. Most naturally occurring monosaccharides belong to the D-series [1].

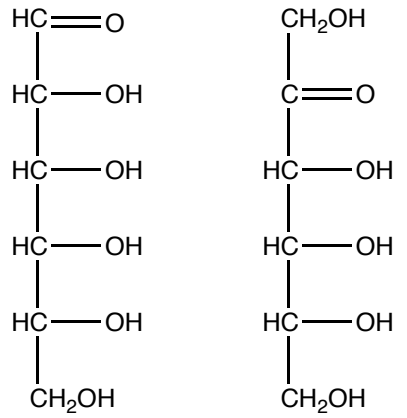


Figure 2.1: Example of a six-carbon aldose (aldohexose) and six-carbon ketose (ketohexose).

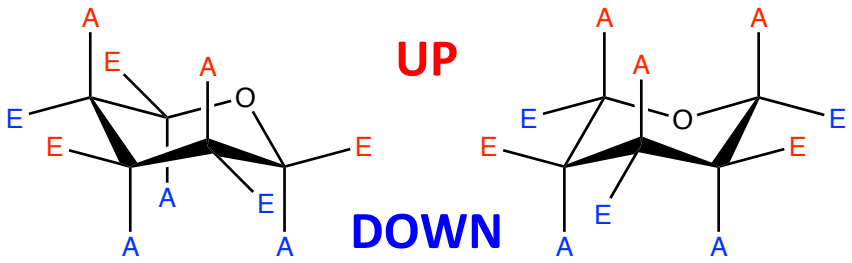
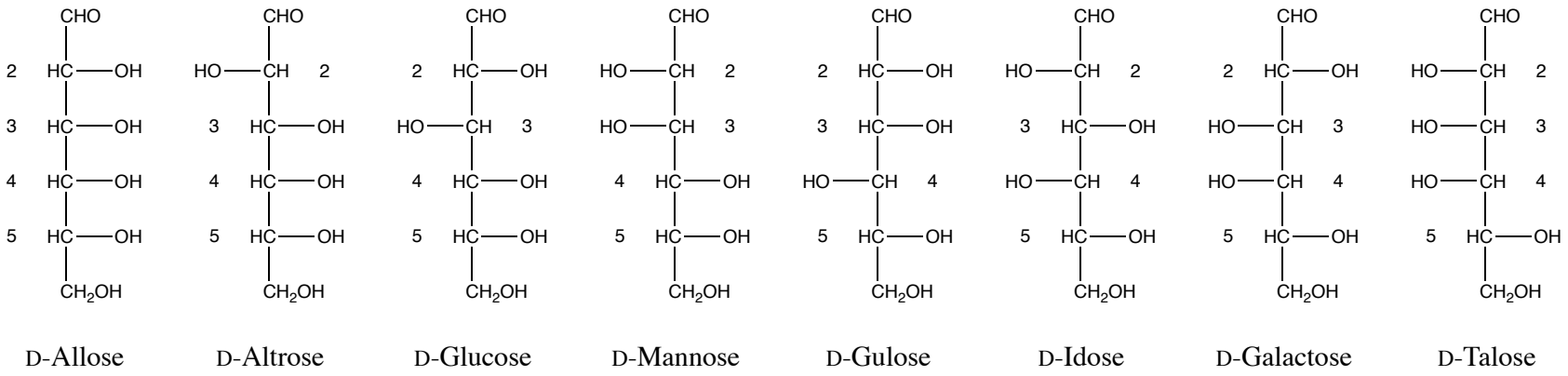
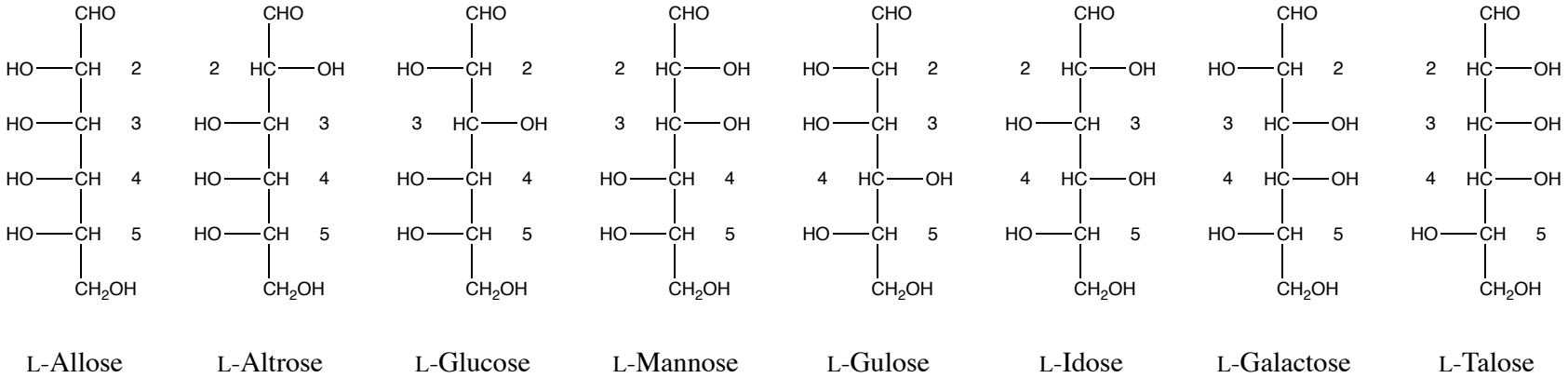


Figure 2.2: Groups falling above the ring following cyclization are defined as *up* (red), while groups falling below the ring are defined as *down* (blue).

Linear monosaccharides are generally represented with Fischer projections, as in Figure 2.3. Hydroxyl groups appearing on the left side of a projection will be in an *up* (above the ring) configuration once cyclized, and hydroxyl groups appearing on the right side will be in a *down* (below the ring) configuration. Figure 2.2 illustrates the concepts of *up* and *down* in relationship to axial and equatorial designations following cyclization.



a) Fischer projections for the D-aldohexoses



b) Fischer projections for the L-aldohexoses

Figure 2.3: All possible steroisomers of (a) D-aldohexose and (b) L-aldohexose. Each of these has four chiral centers, which gives $2^4 = 8$ possible isomers.

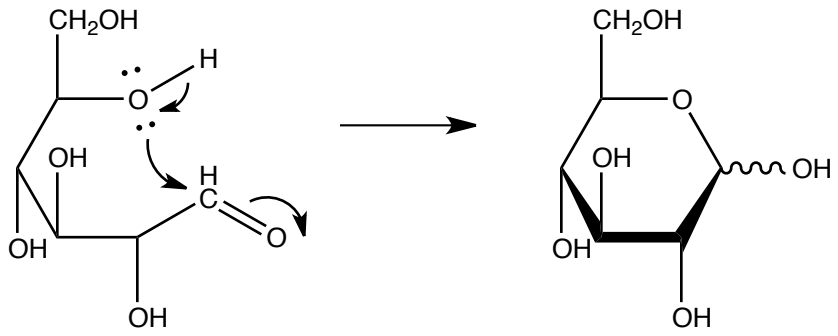


Figure 2.4: Mechanism for the cyclization reaction of D-glucose to form D-glucopyranose.

Cyclic Monosaccharides

The cyclic form of a monosaccharide results from an intramolecular interaction in which the carbonyl group of the aldose or ketose undergoes nucleophilic attack by a hydroxyl group from its own side chain (Figure 2.4). Depending on which hydroxyl group performs the attack, several ring forms are possible. For example, monosaccharides with four or more carbons in their side chain can form five-membered ring structures called furanoses, and monosaccharides with five or more carbons in their side chain can form six-membered ring structures called pyranoses; therefore, an aldohexose, which has has five carbons in its chain and six carbons total, could potentially form either a furanose or pyranose upon cyclization, although the pyranose form is preferred for most monosaccharides [1].

Further, depending on the planar face of the carbonyl group that undergoes attack, two different hydroxyl configurations are possible at the carbonyl carbon, which becomes a new chiral center following the reaction. This new chiral center is referred to as the anomeric carbon, and monosaccharides that differ in their configuration here are called anomers. The two possible configurations, designated α and β, are assigned relative to the configuration of the stereocenter furthest from the anomeric carbon. If these stereocenters exhibit configurations that place their hydroxyl groups on opposite sides of the sugar ring, then the anomeric configuration is denoted as α (Figure 2.5). Conversely, if these stereocenters ex-

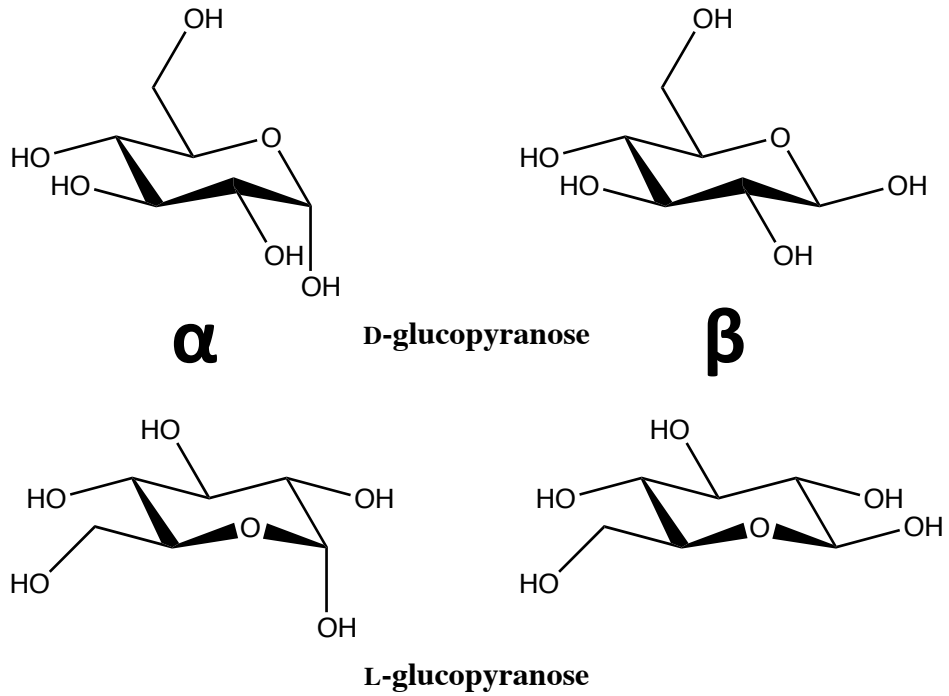


Figure 2.5: Anomeric configurations of glucopyranose.

hibit configurations that place their hydroxyl groups on the same side of the ring, then the anomeric configuration is denoted as β (Figure 2.5). The cyclic monosaccharide form exists in equilibrium with the acyclic form, though the cyclic form is predominant in solution [1].

Cyclic Monosaccharide Conformations

The cyclic ring structure itself gives rise to additional conformational and energetic considerations. Cyclic monosaccharide conformations are generally classified based on a plane designated through three (furanoses), four (furanoses and pyranoses), or occasionally five (pyranoses) atoms of the ring. The plane is oriented such that atom numbers increase in a clockwise direction, and out-of-plane atoms are described as being above or below the plane.

While the furanose structure is relatively uncommon, it plays an important role in RNA and DNA molecules, which contain the aldopentofuranoses ribose and deoxyribose respec-

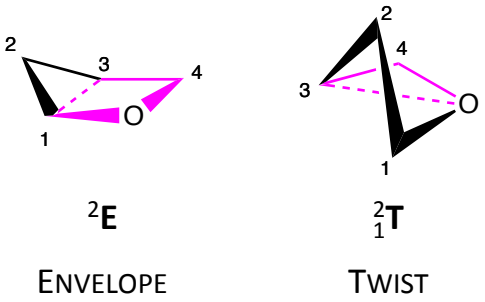


Figure 2.6: An example of the furanose envelope and furanose twist shapes.

tively. The furanose ring can adopt two major conformations, described as the envelope (E) and the twist (T) shapes (Figure 2.6). In the envelope shape, the plane comprises four atoms, and the remaining atom is out of the plane. The out-of-plane atom is denoted with a superscript, such as ^{2}E for the conformation where the plane is formed by C1-C3-C4-O4, and C2 falls above the plane. In the twist shape, the plane comprises only three atoms, and the remaining two are out of the plane, one above and one below. The out-of-plane atoms are denoted with a superscript/subscript pair, such as $^{2}_{1}$T for the conformation where the plane is formed by C3-C4-O4, and C2 falls above the plane, and C1 falls below. A total of ten conformations are possible for each of these two furanose ring shapes.

The major cyclic monosaccharide form, the pyranose ring, can also adopt several distinct conformations. These include the chair (C), boat (B), and skew/twist-boat (S) shapes (Figure 2.7). Transitions between these can also involve the half-chair (H) shape. The two chair shapes are the predominant and most energetically favorable of these conformations. They are denoted as $^{4}C_{1}$ and $^{1}C_{4}$, where superscripts indicate above-plane atoms, and subscripts indicate below-plane atoms. There are six possible boat shapes, each with two atoms either above or below the plane, such as ^{2,5}B or $B_{2,5}$, respectively. There are likewise six possible skew/twist-boat shapes, and as with the chairs, they have one above-plane and one below-plane atom, and are denoted, for example, as $^{1}S_{3}$ or $^{3}S_{1}$. Half-chairs can exists as either four- or five-atom planar forms, and some examples of these are given in Figure 2.7.

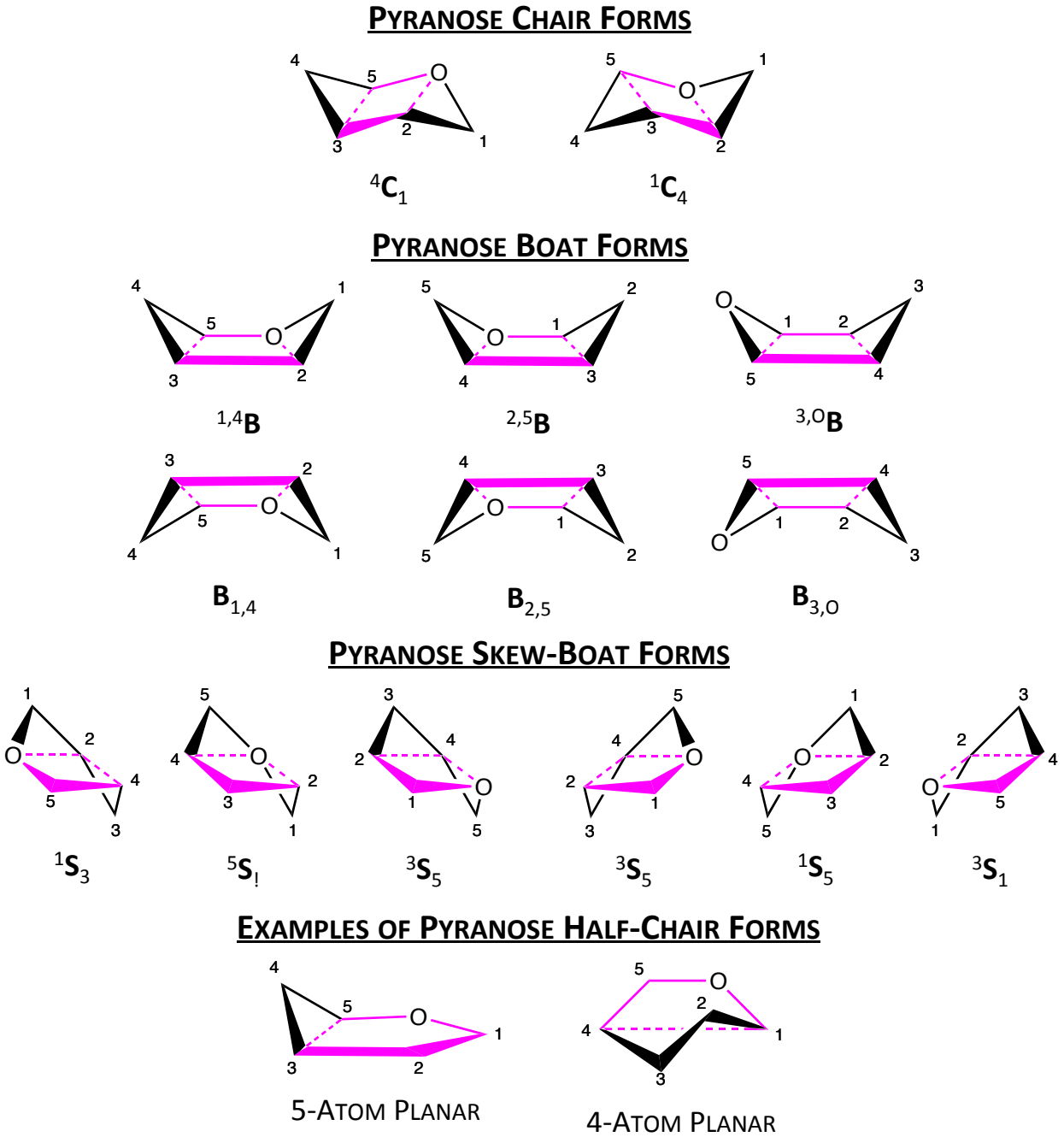


Figure 2.7: The cyclic conformations of pyranose.

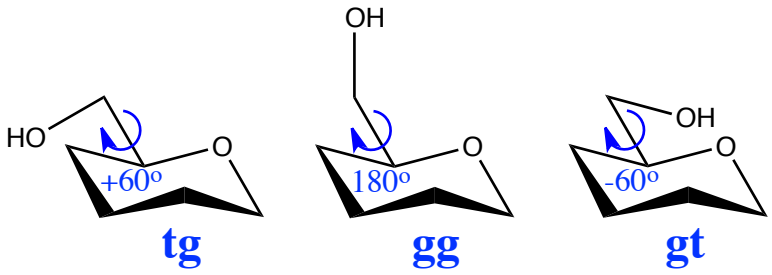


Figure 2.8: The preferred orientations of the D-pyranose hydroxymethyl substituent.

Exocyclic Substituent Conformations

The hydroxymethyl group of the hexopyranose ring, the exocyclic substituent at C5, can adopt several preferred orientations according to the torsional rotation around the C6-C5 bond. This dihedral angle, referred to as ω, is defined as O-C6-C5-H5. The preferred orientations are denoted *tg*, *gg*, and *gt*, where *t* indicates *trans* and *g* indicates *gauche* (Figure 2.8). The first letter describes the position of the hydroxymethyl substituent relative to O5, and the second relative to C4. For D-series sugars, the dihedral angle values characterizing these orientations are $\omega_{tg} = -60°$, $\omega_{gg} = 180°$, and $\omega_{gt} = 60°$. The preferences for these orientations are determined by steric interactions with the hydroxyl group at C4, thus $gt > tg > gg$ when the hydroxyl is axial and $gg > gt > tg$ when the hydroxyl is equatorial [1].

Hexopyranose ring substituents at C1, the anomeric carbon, are also subject to preferred orientations. First and foremost, the anomeric configuration itself is guided by a phenomenon called the endo-anomeric effect (often simply called the anomeric effect). Although the α and β configurations of monosaccharides interconvert as a result of the solution-phase equilibrium between the linear and cyclic forms, the anomeric effect induces a preference for the configuration that places the hydroxyl group into an axial orientation [1]. This is in contrast to steric considerations, which suggest the equatorial orientation would be more energetically favorable. Molecular orbital (MO) theory provides a possible explanation for

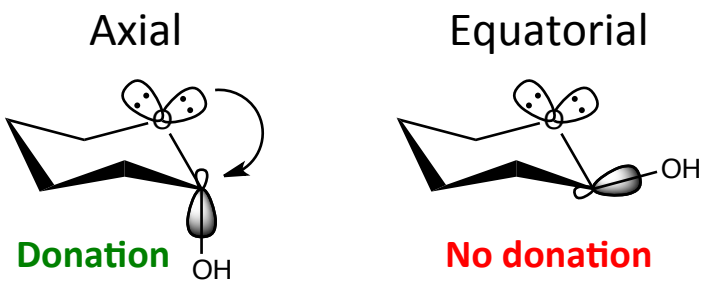


Figure 2.9: The endo-anomeric effect induces a preference for the hydroxyl substituent at C1 to be in an axial configuration. MO theory suggests this results from the alignment of the C1-O1 σ^* antibonding orbital with the O5 lone pair electrons achieved by an axial orientation, allowing donation of electron density to σ^*.

this phenomenon, illustrated in Figure 2.9. According to the MO model, the axial orientation is preferred because it aligns the C1-O1 σ^* antibonding orbital with the O5 lone pair electrons, facilitating donation of electron density and concomitant energetic stabilization [2]. Regardless of its origin, this phenomenon is thought to be responsible for the preference for the α configuration observed in D-series sugars [1].

Disaccharide and Higher Polymer Conformations

Disaccharides are formed via the condensation of the anomeric hydroxyl group of one monosaccharide with a hydroxyl group of a second monosaccharide, releasing a single water molecule. The C-O-C bridge formed between two such residues is called a glycosidic linkage (Figure 2.10). For 1-2, 1-3, and 1-4 connections, the dihedral angles around these bonds are referred to as ϕ (H1-C1-O-CX') and ψ (C1-O-CX'-HX'). For 1-6 connections, the dihedral angle around the additional linking bond of C6-C5 is referred to as ω (O-C6'-C5'-H5'), just as in monosaccharides. In the case of sialic acids, such as the common terminal glycan residue α-N-acetylneuraminic acid (αNeu5Ac), C2 is the anomeric carbon, and ϕ/ψ are alternatively defined as C1-C2-O-CX'/C2-O-CX'-HX'. Altogether, these torsional rotations around the

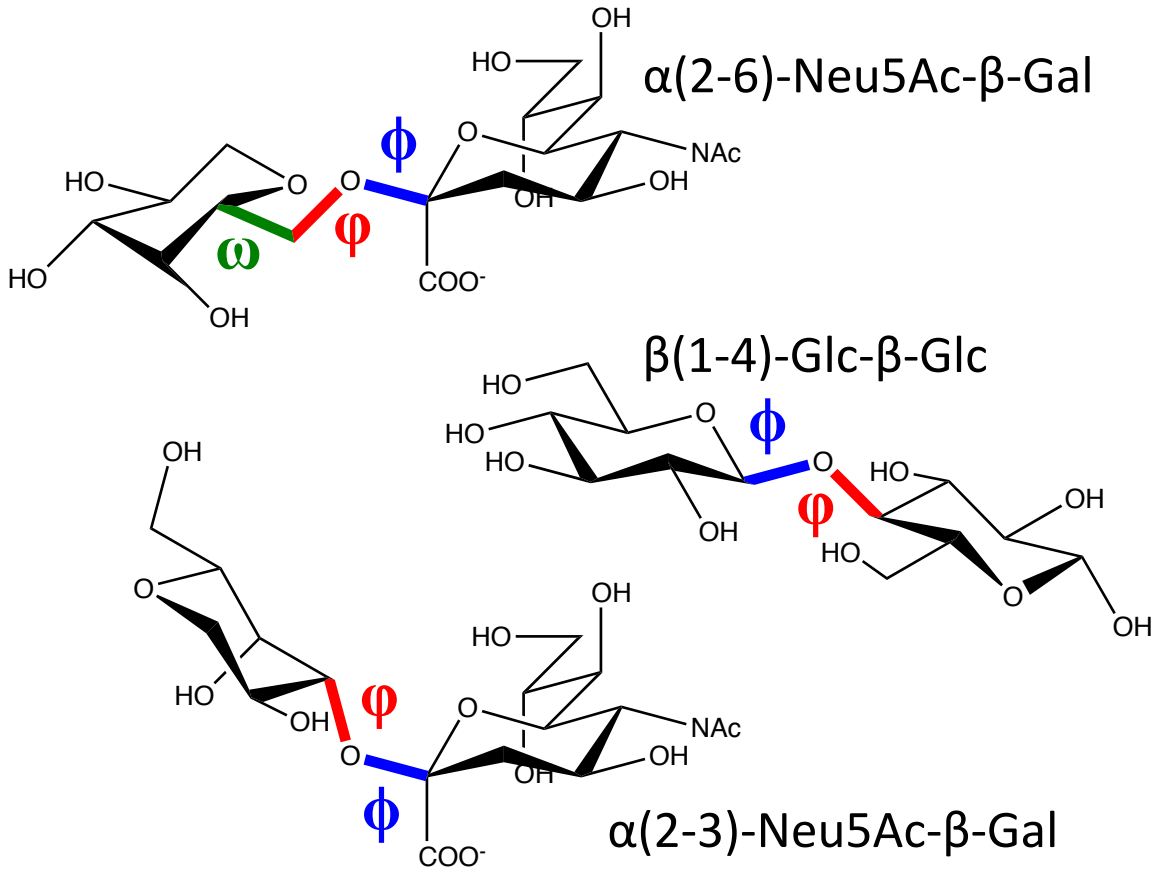


Figure 2.10: The dihedral angles ϕ, ψ, and ω characterize the torsional rotations of glycosidic linkages. Given examples represent, from top to bottom, the minimal motif for human-specific binding in influenza hemagglutinin (α2-6 linkage), the smallest polymer of cellulose, cellobiose (β1-4 linkage), and the minimal motif for avian-specific binding in influenza hemagglutinin (α2-3 linkage).

bonds of the glycosidic linkage determine global disaccharide conformation, as well as the conformations of higher carbohydrate polymers, including oligo- and polysaccharides.

When two monosaccharides condense to form a disaccharide, the anomeric carbon involved in the reaction becomes locked into a given α/β configuration and cannot undergo further interconversion. This residue represents the nonreducing terminus of the disaccharide molecule, while the second residue represents the reducing terminus, which may still interconvert freely according to solution-phase equilibrium. The dihedral angle ϕ around the bond connecting the non-reducing anomeric carbon to the glycosidic oxygen (C1-O) exhibits

distinct rotational preferences depending on the designated α/β configuration, as well as the pyranose ring form. This preference results from a phenomenon called the exo-anomeric effect. MO theory suggests that, as with the endo-anomeric effect, dihedral angle orientation seeks to optimize the alignment of oxygen lone pair electrons with a σ^* antibonding orbital to facilitate donation of electron density [2]. In this case, the oxygen is the glycosidic oxygen, and the antibonding orbital belongs to the C1-O5 bond of the nonreducing sugar residue. Figure 2.11 illustrates the MO model for the exo-anomeric effect, based on examples of the $^4C_1$(D)-α and $^4C_1$(D)-β configurations, which tend to adopt rotations of $\phi = -60°$ and $\phi = 60°$ respectively.

Beyond this primary guiding phenomenon, the orientation of ϕ is determined by steric considerations. The orientation of ψ is determined relative to ϕ, and further by steric considerations. In the case of 1-6 linkages, the orientation of ω is determined relative to ϕ/ψ and sterics, as well as accounting for internal rotational preferences toward *tg*, *gg*, or *gt* configurations.

Challenges of Modeling Carbohydrate and Glycan Conformations

As a result of their innate structural and dynamic complexity, carbohydrates and glycans often prove exceptionally challenging to model. Unlike proteins, which are commonly comprised of only 20 unique residue types and form purely linear polymers, each monosaccharide unit exhibits 2^n possible stereoisomers that can assemble into both linear and highly branched polysaccharide chains. Each stereocenter represents a distinct linkage point for branching, meaning a single hexopyranose can secure up to five separate connections to ancillary residues. The stereocenters involved in a given connection determine the nature and properties of each glycosidic linkage, which may include two to three dihedral angles, each displaying structurally- and environmentally-influenced rotational preferences. Furthermore, the intrinsic flexibility of the glycosidic linkage, as well as the pliability of the carbohydrate

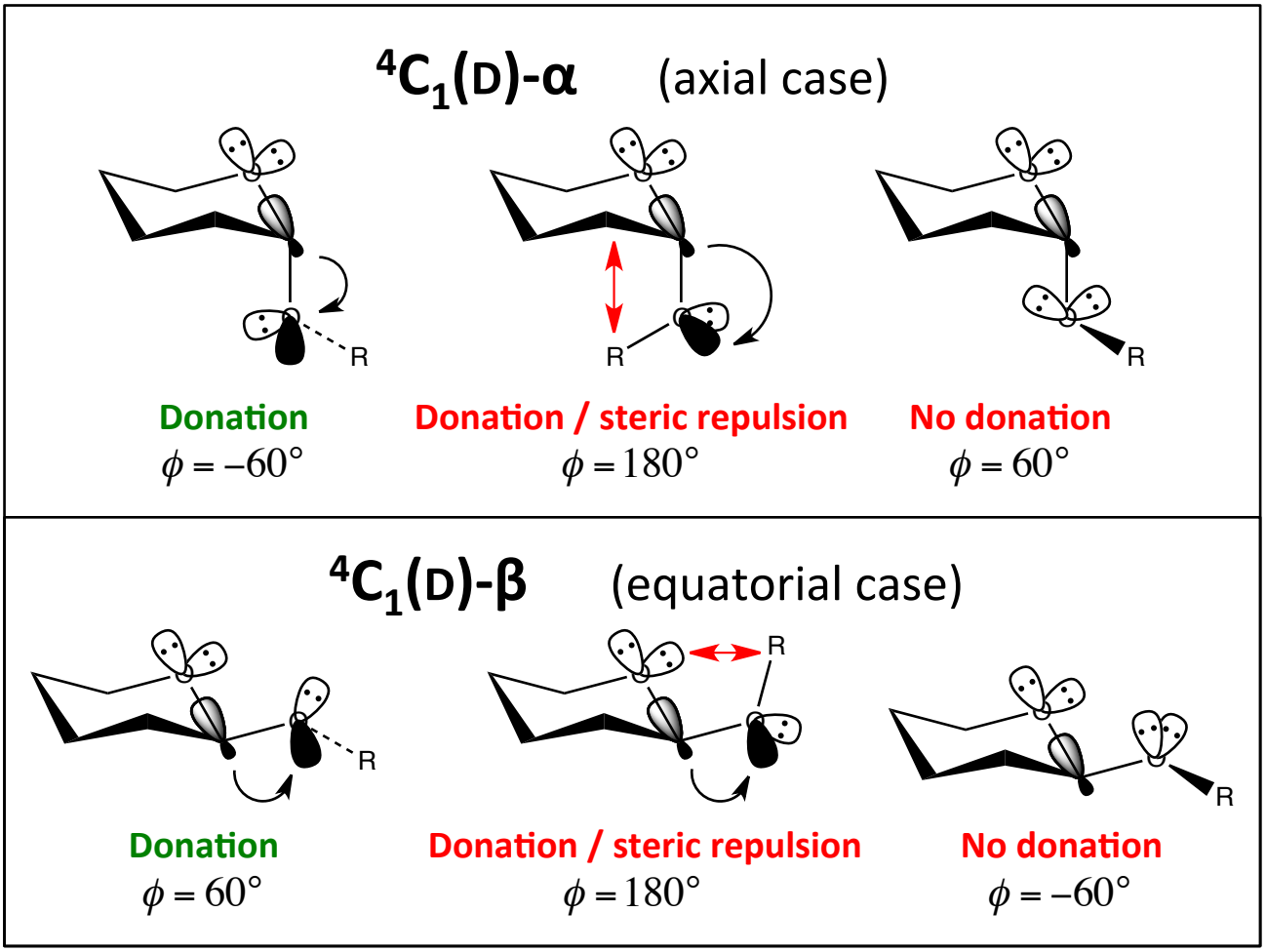


Figure 2.11: The exo-anomeric effect induces distinct conformational preferences of the glycosidic torsion angle ϕ, depending on anomeric configuration and pyranose ring form. MO theory suggests that this follows from optimization of the alignment of the lone pair electrons on the glycosidic oxygen with the σ^* antibonding orbital of the C1-O5 bond, allowing donation of electron density to σ^*. Configurations that facilitate donation in the absence of steric repulsion are preferred.

ring, together give rise to a vast number of configurational and conformational states that must not only be appropriately modeled, but sufficiently sampled.

Related to this, as carbohydrates and glycans are highly polar molecules, their complex structure and conformational behavior also lead to a series of complicated electrostatic considerations. For example, the fixed partial charges commonly applied in carbohydrate modeling (discussed in the following chapter) are quite sensitive to molecular conformation, and the spatial charge distribution within a given monosaccharide varies with stereoisomer and conformational form. Additionally, the endo- and exo-anomeric effects defy steric expectation, and must be purposefully accounted for by auxiliary means.

Thus, in order to accurately model the dynamic behavior of these highly complex molecules, many carbohydrate-specific force fields have been developed (recently reviewed by Foley et al. [3]) and are now widely employed in biomolecular simulations to explore the structure-function relationships of carbohydrates and glycans. Force fields, and their critical role in molecular modeling and simulation, are discussed in the following chapter.

CHAPTER 3:
MOLECULAR DYNAMICS SIMULATIONS

While quantum mechanical (QM) theory utilizes complex mathematical formulations to compute electronic structure for the characterization and prediction of atomic level interactions, the inherent computational expense of this methodology typically limits its application to problems involving only small numbers of atoms. As such, intensive QM study of the dynamical behavior of biomolecular systems is generally precluded, at least on meaningful timescales, as these systems, when fully solvated, may contain hundreds, thousands, even millions of atoms. Alternatively, by combining a simplified atomic model with basic Newtonian physics, classical mechanics (CM) provides a straightforward mechanism to predict molecular motion in a manner that can be applied to systems of biological relevance, on timescales that allow elucidation of conformational dynamics and structure-function relationships.

The use of high-performance computers to conduct molecular dynamics (MD) simulations for studying the time-dependent behavior of a system of many simultaneously interacting CM particles was proposed in 1959 by Alder and Wainwright [4]. Nearly two decades later, in 1977, McCammon et al. first extended this methodology to understand the dynamics of a macromolecule of biological interest at full atomistic detail [5]. Since this original study of the bovine pancreatic trypsin inhibitor (BPTI) protein, MD simulation has been applied to many other classes of biomolecules, including DNA and RNA, carbohydrates, lipids, glycoconjugates, and various covalent complexes of these [6–8].

General Theory

A fundamental law of physics underpinning all of CM theory is the concept of determinism/reversibility. That is, according to Isaac Newton's equations of motion, the explicit future of a particle moving through space is necessarily known, while all history of its past motion is necessarily retained. Thus, it is possible to apply Newton's equations to predict the trajectory of a moving particle, given a set of initial conditions. MD simulations comprise a computational technique that employs these principles to study the motion of model atomic particles within an environment or circumstances of interest.

While QM theory places emphasis on electronic structure when determining the behavior of molecular systems, the atomic model employed in CM MD simulations neglects electronic considerations and instead reduces the atom to a single particle, comprised of an electrostatic point charge surrounded by a shielding van der Waals (vdW) sphere. Each of these atomic particles inherently possess six degrees of motional freedom: three degrees describing location in configuration space (x_c, y_c, z_c), and three degrees describing direction of the associated momentum vector (x_p, y_p, z_p). This six-dimensional space encompassing the mathematical medium through which atomic motion is propagated is called phase space.

While motion through phase space is a time-dependent phenomenon, time in a CM sense is not considered continuous, but rather as occurring in discrete intervals or regularly spaced instances. The time step interval (Δt) represents a predefined value and is generally selected relative to the underlying timescales of motion under study. For example, MD simulations of biological systems often employ time steps of 1-2 fs, which require application of motional constraints to bonds involving hydrogen [9, 10].

A commonly employed method for integrating Newton's equations of motion over time is the Verlet algorithm, given in Equation 3.1.

$$x_{(t+\Delta t)} = 2x_{(t)} - x_{(t-\Delta t)} + a_{(t)}\Delta t^2 \tag{3.1}$$

As per the fundamental law of determinism/reversibility, this equation allows prediction of future $(t + \Delta t)$ atomic position, given current (t) and previous $(t - \Delta t)$ atomic positions and current (t) acceleration. While current and previous positions are known, current acceleration must be computed.

A common translation of Newton's First Law states that "an object in motion will stay in motion unless acted upon by an outside force." Newton's Second Law, shown in Equation 3.2, then provides the relationship characterizing how the motion of that object, in this case, an atomic particle, changes upon application of external force.

$$F = ma \tag{3.2}$$

That is, if the force on the particle is known, then the acceleration is also known. Furthermore, an auxiliary definition of force, given in Equation 3.3, indicates that it is also proportional to the slope of the atom's position on the potential energy surface of the system.

$$F = ma = -\frac{\partial V}{\partial x} \tag{3.3}$$

In CM MD simulations, the potential energy V, and subsequently the force F on an atom, is calculated through a mathematical expression commonly referred to as a force field (discussed in detail in the following section). Thus, during an MD simulation, motion is propagated via performing the following operation for every atom in the system upon every time step increment Δt:

1. The force field expression is evaluated to obtain current potential energy V (Equation 3.4, given in following section)

2. The position-derivative of the potential energy V is used to compute current acceleration a_t (Equation 3.3)

3. Current acceleration a_t is combined with current position x_t, previous position $x_{t-\Delta t}$, and time step Δt to predict future position $x_{t+\Delta t}$ (Equation 3.1)

4. The atom is moved to its new position, time is incremented by Δt, and this iterative process repeats (go back to step 1)

One caveat to this process is that, at the beginning of a simulation, the initial atomic coordinates are static, and as such, possess no current accelerations or previous positions. A commonly employed solution to this problem is to simply assign velocities to each atom, with these velocities typically being selected randomly based on a Maxwell-Boltzmann distribution (a probability distribution characterizing particle speeds) appropriate to the simulation temperature. The system is then allowed to evolve for one time step Δt, after which the Verlet algorithm (Equation 3.1) applies, and normal iteration may proceed.

Again as a consequence of the fundamental law of determinism/reversibility, there exists no analytical solution to Newton's equations of motion, and the integration process performed in MD simulations must be carried out numerically. That is, system configuration at time t_{Future} cannot be known until all previous time steps are computed to arrive at that time and its associated system configuration. The resulting computational expense for exploring long timescale dynamics of large, biologically relevant systems using MD simulations thus necessitates the use of high-performance supercomputers, and the state of MD methodology is thereby limited by the power of current computational technology and the availability of that technology to scientific researchers.

Force Fields

Empirical CM biomolecular force fields have two components:

1. A mathematical expression for describing the potential energy V of a system as a function of atomic configuration

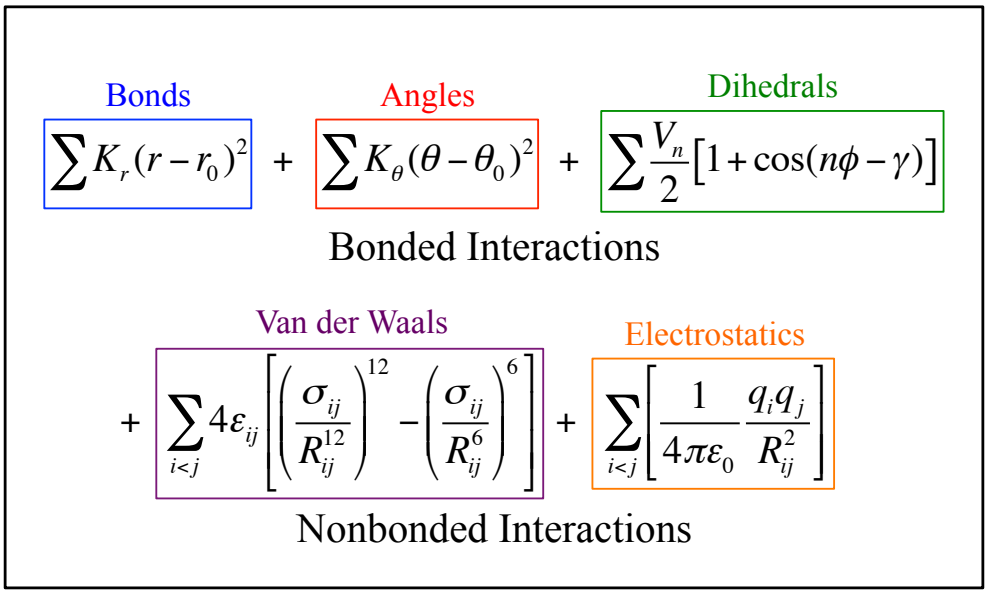


Figure 3.1: The force field expression, which is used to compute total potential energy of a system as a function of atomic configuration, includes consideration of bonds, angles, dihedral angles, and nonbonded vdW and electrostatic interactions.

2. A parameter set for use with this expression that contains all necessary predefined values for characterizing the dynamic behavior of a molecular system

The force field equation or potential energy function generally follows a form similar to that given in Equation 3.4, which is used by the AMBER-family force fields, such as the GLYCAM06 force field for carbohydrates [11].

$$V_{Total} = \sum^{Bonds} K_r(r-r_0)^2 + \sum^{Angles} K_\theta(\theta-\theta_0)^2 + \sum^{Dihedrals} \frac{V_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{i<j}^{van\ der\ Waals} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] + \sum_{i<j}^{Electrostatics}\left[\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{R_{ij}^2}\right] \tag{3.4}$$

While this equation appears complex, it merely represents a summation over all atomic interactions in the system, where bonded interactions are comprised of those between two

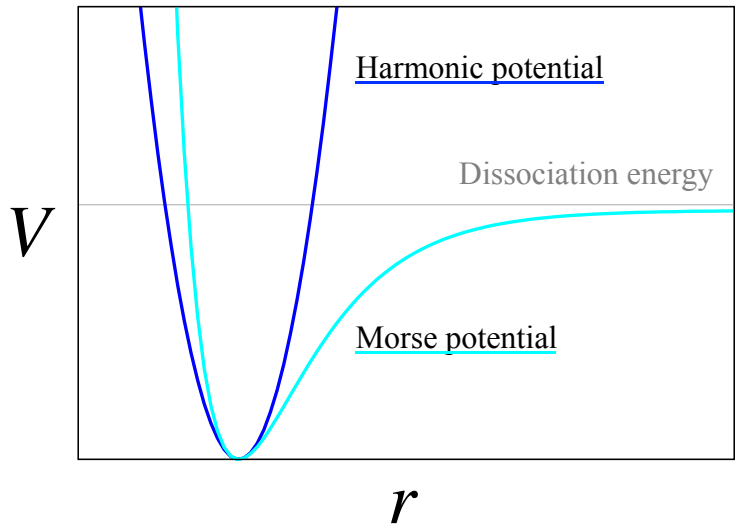


Figure 3.2: The harmonic and Morse potentials, two mathematical functions used to describe the behavior of atomic bonds. While the Morse potential more accurately describes bond behavior, with dissociation occurring at sufficient magnitude of r_0, the harmonic potential is generally employed to prevent bond breakage and improve system stability. The harmonic potential well-approximates the Morse potential for values of r close to equilibrium.

(bonds), three (angles), and four (dihedral angles) covalently attached atoms, and nonbonded interactions encompass through-space vdW and electrostatic effects (Figure 3.1).

Bond and Angle Terms

In the CM world, atoms and bonds are loosely treated as "balls" and "springs." As such, the stretching/compressing behavior of an atomic bond is generally modeled as a simple harmonic oscillation (Figure 3.2) using Hooke's Law, as showin in Equation 3.5.

$$V_{Bonds} = \sum^{Bonds} K_r(r - r_0)^2 \tag{3.5}$$

Technically, a true chemical bond would break if its length were to become sufficiently large, meaning that bond behavior could be more correctly modeled by employing a Morse potential (Figure 3.2), which becomes less steep and eventually plateaus beyond length

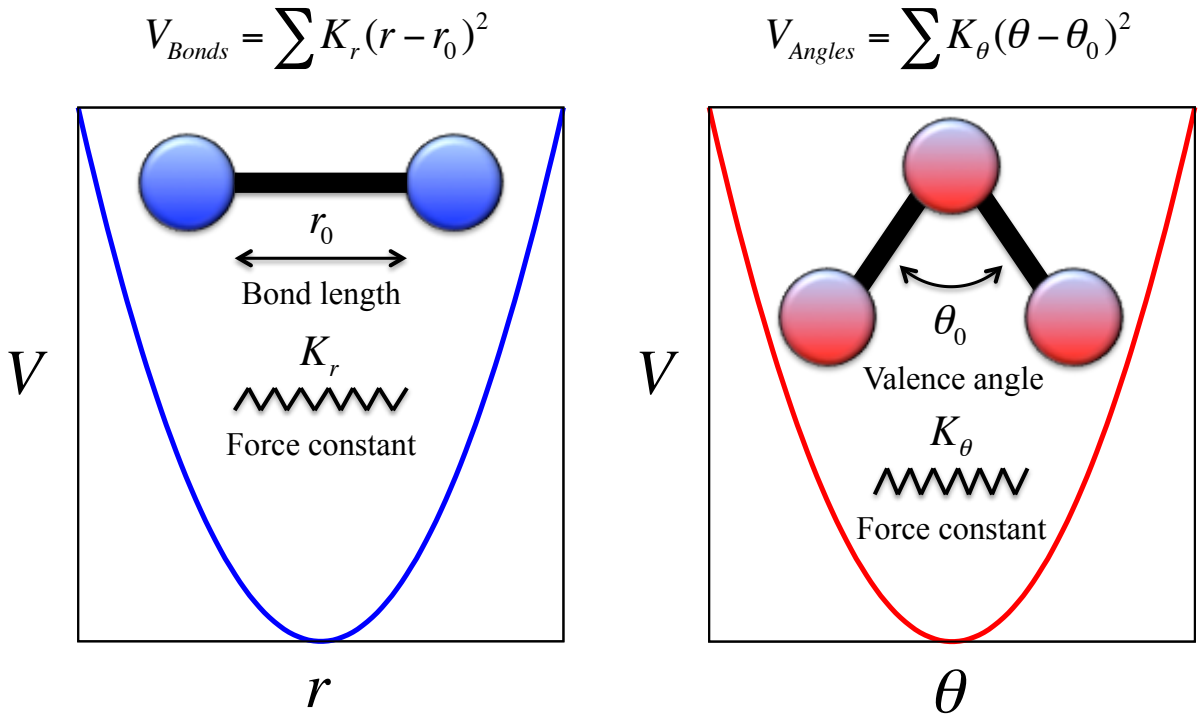


Figure 3.3: Harmonic potentials are used to model the dynamic behavior of atomic bonds and angles. The force constants K_r and K_θ determine the shape of the potential wells, while the equilibrium values r_0 and θ_0 determine the location of the well minimums.

magnitudes corresponding to bond dissociation. This consideration is typically neglected in CM MD simulations, however, both for the sake of simplicity and to enforce stability of the system. While the inability to break or form chemical bonds fundamentally precludes study of reaction mechanisms with purely CM models, this circumstance is often exploited to more effectively investigate dynamic behavior. For example, calculations may be performed at elevated temperatures to facilitate crossing of barriers on the potential energy surface without ripping apart the molecule under study due to increased kinetic motion.

The parameters for the bond portion of the force field expression include the equilibrium or reference bond length r_0 and the force constant for bond oscillation K_r. In terms of the harmonic potential described by Equation 3.5, K_r determines the overall shape and steepness of the potential well, and r_0 represents the location of the well minimum (Figure 3.3, left).

Deviations in current bond length r induced by stretching/compressing are ultimately limited by K_r, and r is thus restricted to remain close to the equilibrium value r_0. Initial parameters for r_0 are generally obtained from experimental diffraction data or QM geometry optimization of a representative molecule, while those for K_r are typically estimated based on vibrational spectroscopy, be it experimental (infrared, Raman) or calculated (QM).

Angles between atoms also exhibit stretching and compressing behavior in a strictly CM sense and are likewise modeled as harmonic oscillators, as per Equation 3.6.

$$V_{Angles} = \sum^{Angles} K_\theta(\theta - \theta_0)^2 \tag{3.6}$$

The parameters for the angle portion of the force field expression include the equilibrium or reference valence angle value θ_0 and the force constant for angle oscillation K_θ. As with the harmonic potential describing bond behavior, K_θ determines the overall shape and steepness of the potential well, θ_0 represents the location of the well minimum, and the value of θ is thus restricted to remain close to equilibrium (Figure 3.3, right). Initial parameters for angles are obtained through the same mechanisms as those for bond, selecting θ_0 based on crystallographic or QM optimized structures, and K_θ based on experimental or calculated vibrational spectroscopy.

vdW Term

Nonbonded interactions are comprised of both vdW and electrostatic effects. Vdw interactions encompass both attractive and repulsive forces between atoms, and in general, describe how atoms interact with each other sterically through space. The overall vdW interaction between two atoms is modeled as a Lennard-Jones potential, given in Equation 3.7, which is constructed from an attractive component representing the effect of London dispersion forces, and a repulsive component accounting for Pauli exclusion, or disallowed atomic overlap (Figure 3.4).

$$V_{van\ der\ Waals} = \sum_{i<j}^{van\ der\ Waals} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] \tag{3.7}$$

The parameters for the vdW portion of the force field expression include ϵ_{ij} and σ_{ij}. The value of ϵ_{ij} gives the depth of the Lennard-Jones well and indicates the strength of the interaction between atoms i and j. The value of σ_{ij} denotes the distance R_{ij} at which the potential between atoms i and j is zero, and represents the sum of the vdW radii for these two atoms. Initial parameters for ϵ and σ for individual atoms are generally determined based on reproducing pure liquid or crystal properties, such as enthalpies of vaporization/sublimation [12], with theoretical values being adjusted until calculations match experimental data.

At long range, vdW interactions become increasingly negligible and are often neglected beyond a minimum 8-9 Å cutoff distance to improve efficiency of MD simulations [13]. Some MD software implementations include a long-range correction factor to account for errors associated with this truncation [14].

Electrostatic Term

Electrostatic effects encompass the attractive and repulsive interactions that occur between positive and negative electric charges (Figure 3.5). In a QM sense, the charge on an atom or molecule is polarizable, with charge distribution shifting according to the surrounding electrostatic environment. While some CM force fields, such as AMOEBA [15], attempt to implement charge polarizability, this dramatically increases the computational expense of the potential calculation. Consequently, a far more common approach involves assigning fixed partial charges to atoms, and modeling their interaction using Coulomb's Law, shown in Equation 3.8.

$$V_{Electrostatics} = \sum_{i<j}^{Electrostatics}\left[\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{R_{ij}^2}\right] \tag{3.8}$$

Accordingly, the overall potential between two atomic charges is related to their signed magnitudes q and the distance R separating them. Partial charges for atoms are derived by

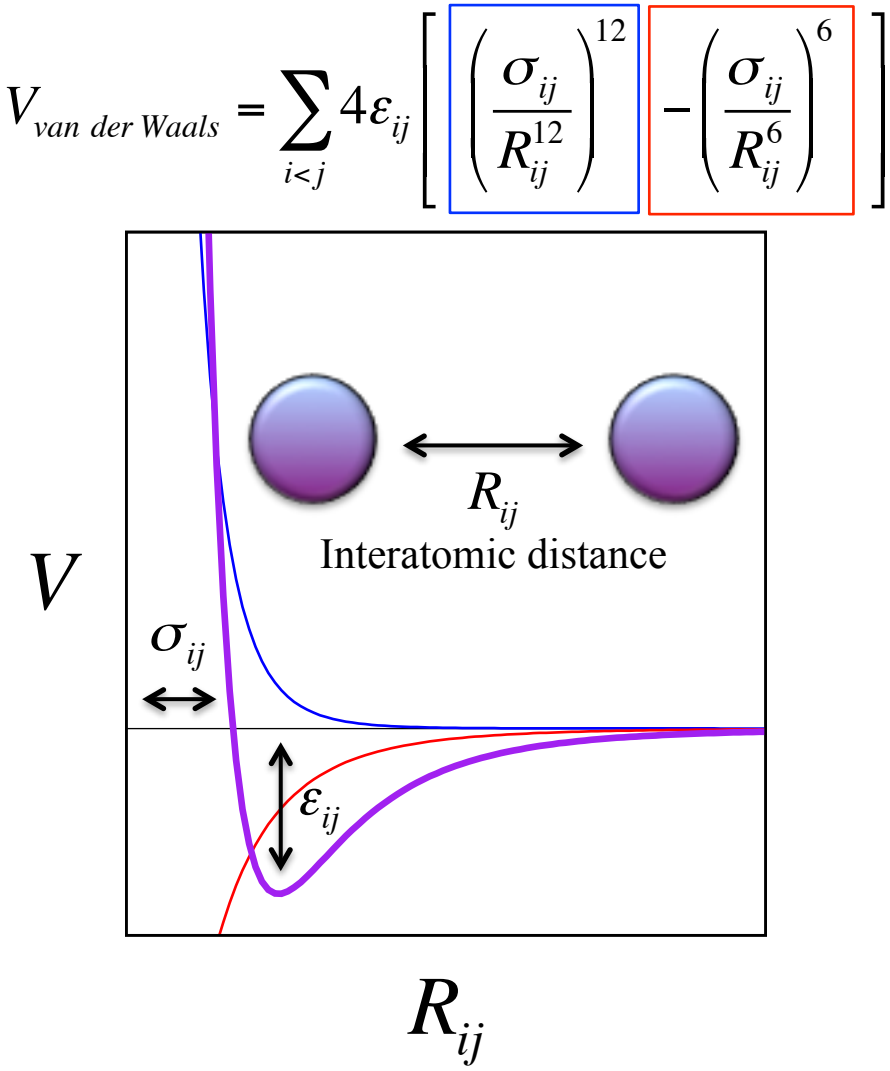


Figure 3.4: Lennard-Jones potentials are used to model the dynamic behavior of atoms resulting from vdW interactions. The potential is comprised of repulsive (blue) and attractive (red) components. Atomic repulsion is a consequence of disallowed atomic overlap, as per the Pauli exclusion principle, while weak atomic attraction follows from the effect of London dispersion forces. The value of σ_{ij} represents the sum of the atoms' vdW radii, and ϵ_{ij} characterizes the strength of the interaction between them, given by the depth of the Lennard-Jones well.

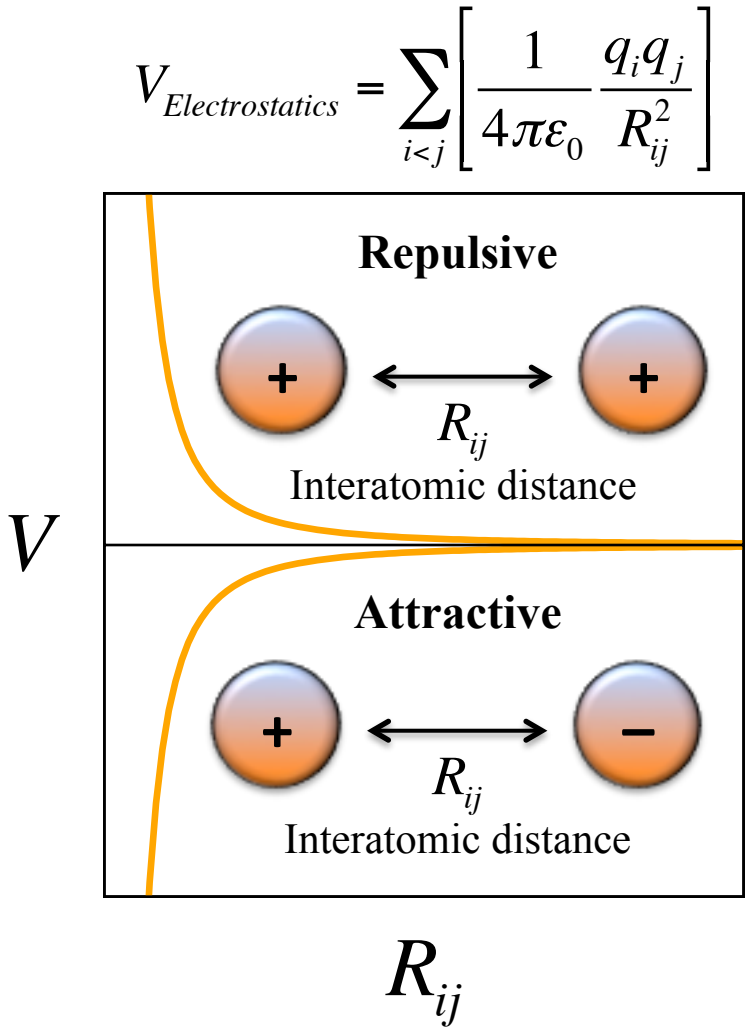


Figure 3.5: Electrostatic interactions are modeled with Coulomb's law. For atoms of similar charge (+/+ or −/−), the potential becomes increasingly repulsive as R_{ij} decreases. Conversely, for atoms of opposite charge (+/−), the potential becomes increasingly attractive as R_{ij} decreases.

fitting a CM electrostatic potential (ESP) to a QM ESP, calculated at grid points around a representative molecule or molecular fragment. In AMBER-family force fields, the 6-31G* basis set is the standard for QM calculation of ESPs because it represents a reasonable compromise between computional expense and suitability of the resulting charge sets for use in condensed phase simulations. However, the raw ESP charges generated with this basis set tend to overstate bond polarity. As such, a hyperbolic restraint function, described in Equations 3.9 and 3.10, is commonly applied during fitting in order to produce slightly attenuated resultant charges, referred to as restrained ESP (RESP) charges [16].

$$\chi^2_{RESP} = \chi^2_{ESP} + \chi^2_{Rstr} \tag{3.9}$$

where

$$\chi^2_{Rstr} = k_{Rstr} \sum_j \left((q_j^2 + b^2)^{1/2} - b \right) \tag{3.10}$$

Within Equation 3.10, the magnitude of b determines the tightness of the hyperbola and serves to constrain total charge to a given integer value. The constant k_{rstr} determines the strength of the restraint function, with weights of $k_{rstr} = 0.001$ and $k_{rstr} = 0.01$ typically employed for protein and carbohydrate charges, respectively [16, 17].

Long-range Nonbonded Interactions

The nonbonded interaction calculation, which must consider each atom interacting with all neighboring atoms in an N^2 manner, represents the computational bottleneck for predicting molecular motion with CM methods. As discussed above, vdW interactions are typically truncated beyond an 8 Å minimum cutoff, after which their contribution is considered negligible. Charge interactions have significant long-range effects, however, and truncation can thus produce serious calculation artifacts. In 1993, Darden et al. presented the particle-mesh Ewald (PME) method [18], which combines a pair-wise, direct-space calculation within the cutoff with an approximate, reciprocal-space calculation beyond the cutoff. This technique

depends on periodic boundary conditions (discussed later in this chapter), and involves interpolating charges onto a mesh or grid to evaluate long-range interactions in a manner that converges quickly. As truncation beyond a finite distance vastly reduces the accuracy of charge modeling, yet expansion of the cutoff distance to improve accuracy greatly increases computational expense, the implementation of PME serves to improve overall simulation fidelity while significantly enhancing computational efficiency.

Nonbonded Exclusions and Scalings

While covalently attached atoms also share nonbonded interactions, the methods employed for calculating the bond (1-2), angle (1-3), and dihedral angle (1-4) potentials partially account for nonbonded behaviors within these 1-2, 1-3, and 1-4 atom sets. Thus, to avoid double counting, CM force field implementations often exclude 1-2 and 1-3 calculations for vdW and electrostatic effects, and may also scale the 1-4 calculations by an empirically determined factor. For proteins, the AMBER-family force fields employ a scaling factor of 2.0 and 1.2 for vdW and electrostatics respectively [19]. For carbohydrates, 1-4 interactions are not scaled, because scaling was found to interfere with satisfactory modeling of hydrogen bond behavior [17].

Dihedral Angle Term

The remaining component of the force field expression, that describing covalent 1-4 interactions, is generally addressed during the final stages of force field parameterization. In an ideal sense, consideration of bonds, angles, and nonbonded interactions alone should be sufficient to reasonably compute the potential energy of a system as a function of atomic configuration. However, comparison of rotational profiles generated using this short-sighted approach with analogous profiles calculated via QM methods commonly demonstrates the failure of CM to adequately reproduce the energetics of torsional rotation. Thus, the dihe-

dral angle portion of the force field expression essentially represents a quantum correction factor designed to account for discrepancies with respect to calculated QM behavior.

Upon juxtaposing corresponding CM and QM rotational plots, parameters for V_n, n, and γ are determined such that Equation 3.11 represents a mathematical formulation for the energetic inconsistency observed between the two (Figure 3.6).

$$V_{Dihedrals} = \sum^{Dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] \tag{3.11}$$

The value of V_n controls the amplitude of the cosine function, n controls the frequency, and γ determines phase shift. For a simple curve, $n = 1$ is employed. However, if the difference function is sufficiently complex, $V_{Dihedrals}$ may require summation of several curves employing $n = 1, 2, 3...$ to achieve requisite fit. When Equation 3.11 is combined with the rest of the force field expression, the CM methodology for MD simulations becomes adequately robust to describe the dynamic behavior of biomolecular systems.

Validation

Once an empirical biomolecular force field has been fully parameterized, data collected from MD simulations employing it are compared against experimentally analogous data to assess merit. At this point, some parameters may be adjusted until experimental values are suitably reproduced. This validation step is essential to confirm the force field's capability to make accurate predictions regarding molecular motion.

Solvent Models

While a basic MD simulation of a biomolecule can be performed *in vaccuo*, that is, in the absence of any surrounding solvent, this provides little information about molecular behavior in a biologically relevant environment. These simulations are generally considered to represent gas phase conditions and comprise only a single, isolated copy of the molecule

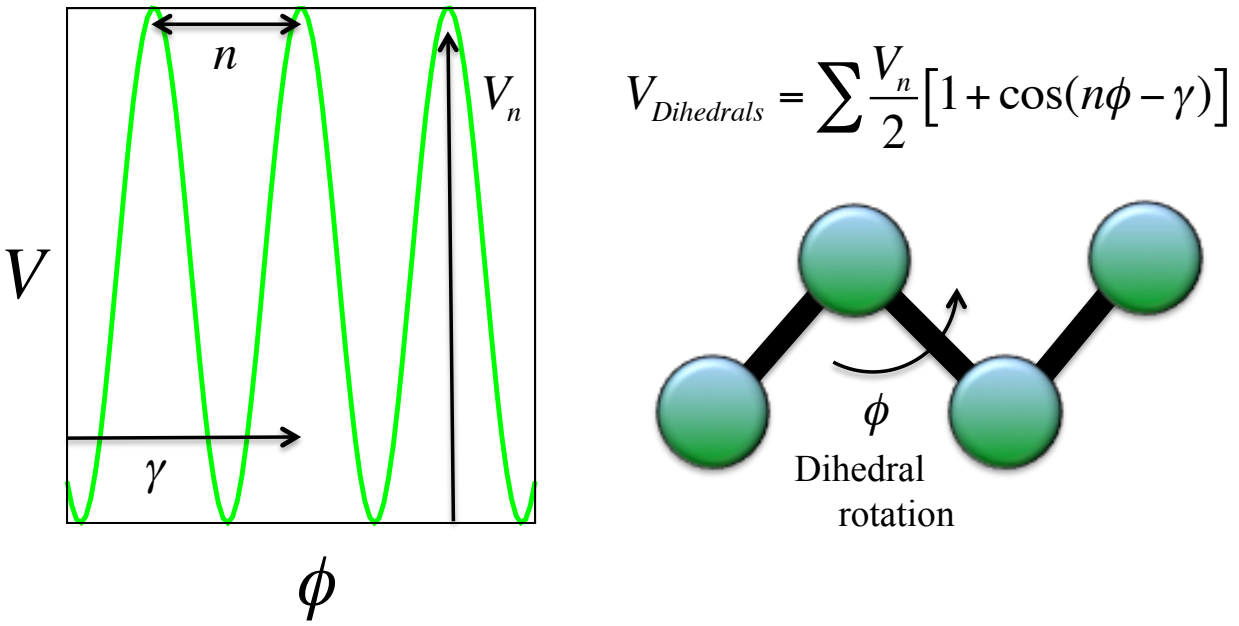


Figure 3.6: A force field expression that considers only bonds, angles, and nonbonded interactions typically fails to reproduce the rotational behavior predicted by QM methods. As such, a $V_{Dihedrals}$ term comprised of a modified cosine function is included to account for these energetic discrepancies, allowing CM methodology to adequately describe the rotational dynamics of biomolecular systems. For a simple molecule such as butane, the cosine function captures the repeating energy minimums associated with anti conformations, and energy maximums associated with eclipsed conformations. The parameter V_n controls the amplitude of the cosine function, n controls the frequency, and γ determines phase shift.

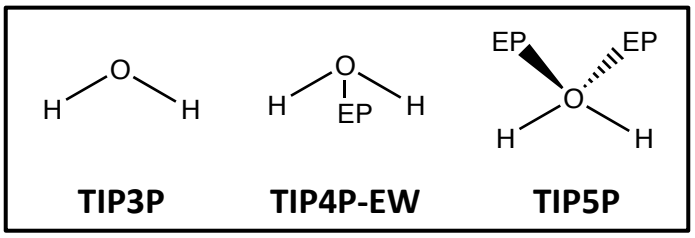


Figure 3.7: Diagrams of three commonly employed explicit solvent models, where EP represents an extra point or dummy atom.

of interest in otherwise empty space. Alternatively, the molecule can be modeled in the presence of solvent by means of two possible strategies.

The first of these is implicit solvent, which utilizes a mathematical formulation designed to effectively approximate the influence of bulk water as a continuum around the molecular surface. This method, while computationally inexpensive for small systems, suffers from a number of limitations, including the inability to form solute-solvent hydrogen bonds. In some cases, such hydrogen bonds mediate key interactions critical for ligand binding [20].

The second, more accurate, and more commonly employed method for modeling solvent is to describe it explicitly using discrete water molecules. Numerous explicit solvent models exist, each characterized by individual merits and limitations. Three conventional examples of models that focus on reproducing specific properties of water are TIP3P [21], TIP4P-EW [22], and TIP5P [23] (Figure 3.7).

The TIP3P framework comprises three atoms, simply representing an oxygen and two hydrogens. While this standard model is relatively computationally efficient and reasonably approximates the behavior of bulk solvent, its viscosity is too low to allow accurate calculation of diffusion rates [24].

The TIP4P-EW framework comprises four atoms, representing a typical water molecule with the addition of a dummy atom (also called an extra point) positioned along the bisection of the hydrogen-oxygen-hydrogen angle. The oxygen charge is alternatively placed on this

dummy atom, serving to refine electrostatic distribution around the water molecule and thus improve bulk water behavior [22].

The TIP5P framework comprises five atoms, augmenting the three atoms of water with two dummy atoms (extra points) designed to represent electron lone pairs on the oxygen. The oxygen charge is distributed between these dummy atoms, serving to improve both bulk water behavior and the tetrahedral geometry of contacting hydrogen bonds [23]. While TIP5P is the most computationally expensive of the solvent models discussed here, it has been shown to enhance specific solvation, produce more highly conserved and optimally coordinated water interactions, and generally affect the overall conformational dynamics of biomolecular systems as a consequence of increased surface hydration [25–27].

To reduce overall complexity and facilitate simulation time steps of up to 2 fs, all three of these water models are designed to be rigid and, therefore, do not undergo the internal motions of bond/angle stretching.

Periodic Boundary Conditions

When a biomolecule is surrounded by explicit solvent, a boundary or box must also be defined in order to contain the solvent molecules and prevent them from flying off into space. Furthermore, when such a boundary is introduced, the system becomes subject to calculation artifacts resulting from solvent molecules along the boundary contacting the vacuum of emptiness encompassing the system. This nonphysical circumstance, called an edge effect, can be addressed through the use of periodic boundary conditions (PBC).

When periodicity is in effect, copies of the primary simulation box, or unit cell, are tiled infinitely in all directions, such that each face of every cell is matched to its opposing face in the next image (Figure 3.8). On the occasion that an atom should cross the periodic boundary as if to move into an adjacent cell, it exits the primary box along one side and is wrapped to simply re-enter the primary box along the opposite side. In this way, only

a finite number of atoms are required to represent a fully solvated molecular system, while nonphysical boundaries and associated edge effects are eliminated.

There exists a single caveat that must be considered in order to avoid introducing additional calculation artifacts when utilizing PBC, referred to as the minimum image convention. This convention states that the shortest dimension describing the simulation box must be at least twice the value of the nonbonded interaction cutoff (Figure 3.9). In other words, the diameter of the cutoff for a given atom must not overlap across adjacent images. If the minimum image convention is not observed, neighboring atoms could be double-counted in the nonbonded interaction calculation, leading to erroneous results.

Thermodynamic Ensembles

Following appropriate design and parameterization of a biomolecular system, a further consideration of simulation setup is the choice of thermodynamic ensemble. While MD simulations can be used to predict time dependent motional behavior, in doing so, they also serve to generate a collection of structures that each describe a unique state of the system. This statistically relevant collection of structures is called an ensemble, and the constraints imposed on the system under study determine the thermodynamic properties of that ensemble.

The simplest thermodynamic ensemble is the canonical ensemble (nVE), where number of particles n, volume V (defined by the size of the simulation box), and total energy E of the system are held constant. Conservation of energy is another fundamental law of physics. That is, energy within a closed system is not created or destroyed, but simply converted between different forms. The total energy E of a closed CM system comprised of moving particles is the sum of the total kinetic U and potential V energies, as indicated in Equation 3.12.

$$E_{Total} = U_{Total} + V_{Total} \tag{3.12}$$

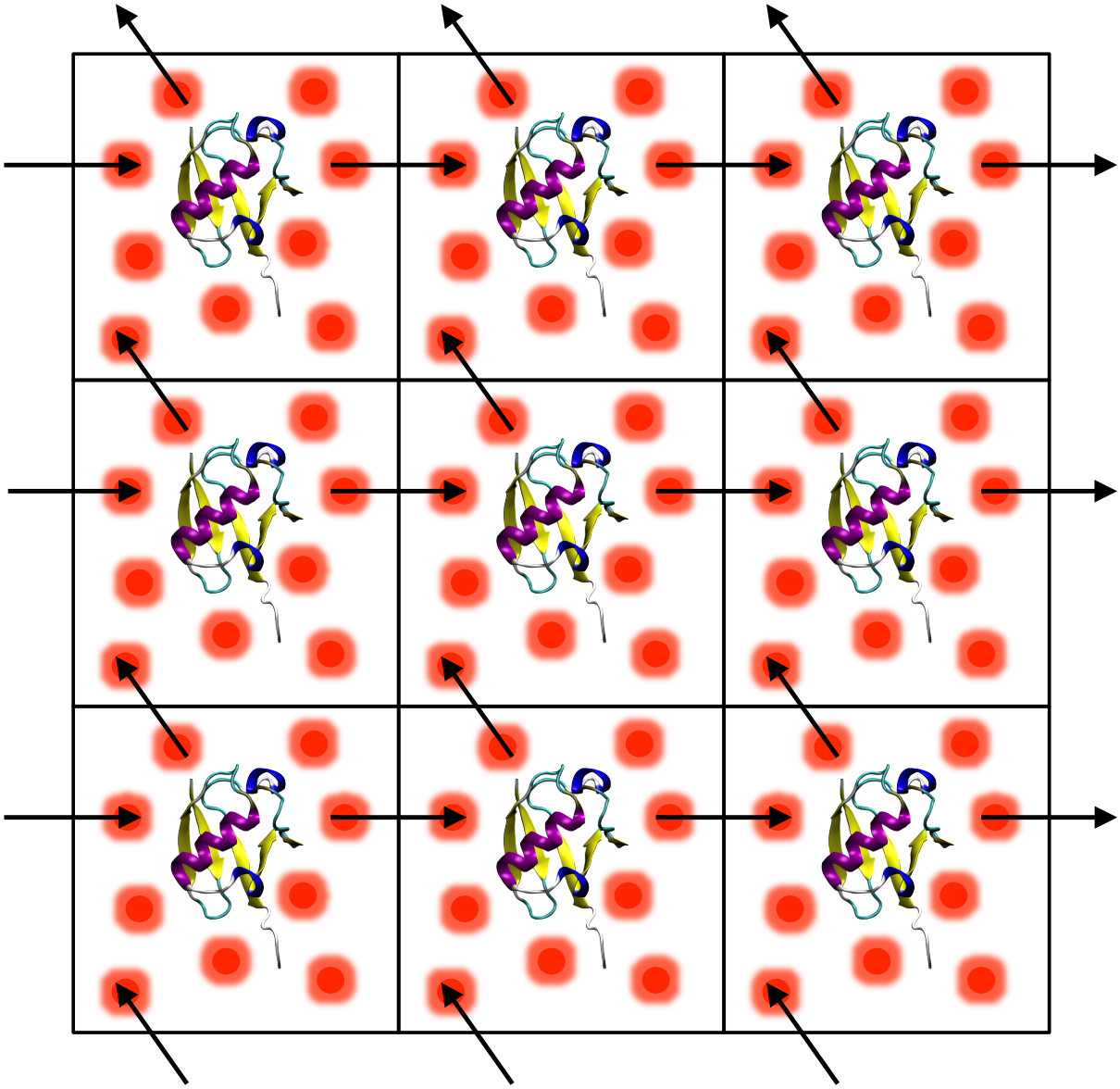


Figure 3.8: 2D representation of a system under periodic boundary conditions. As a molecule exits one side of the primary simulation box, it re-enters the box on the opposite side.

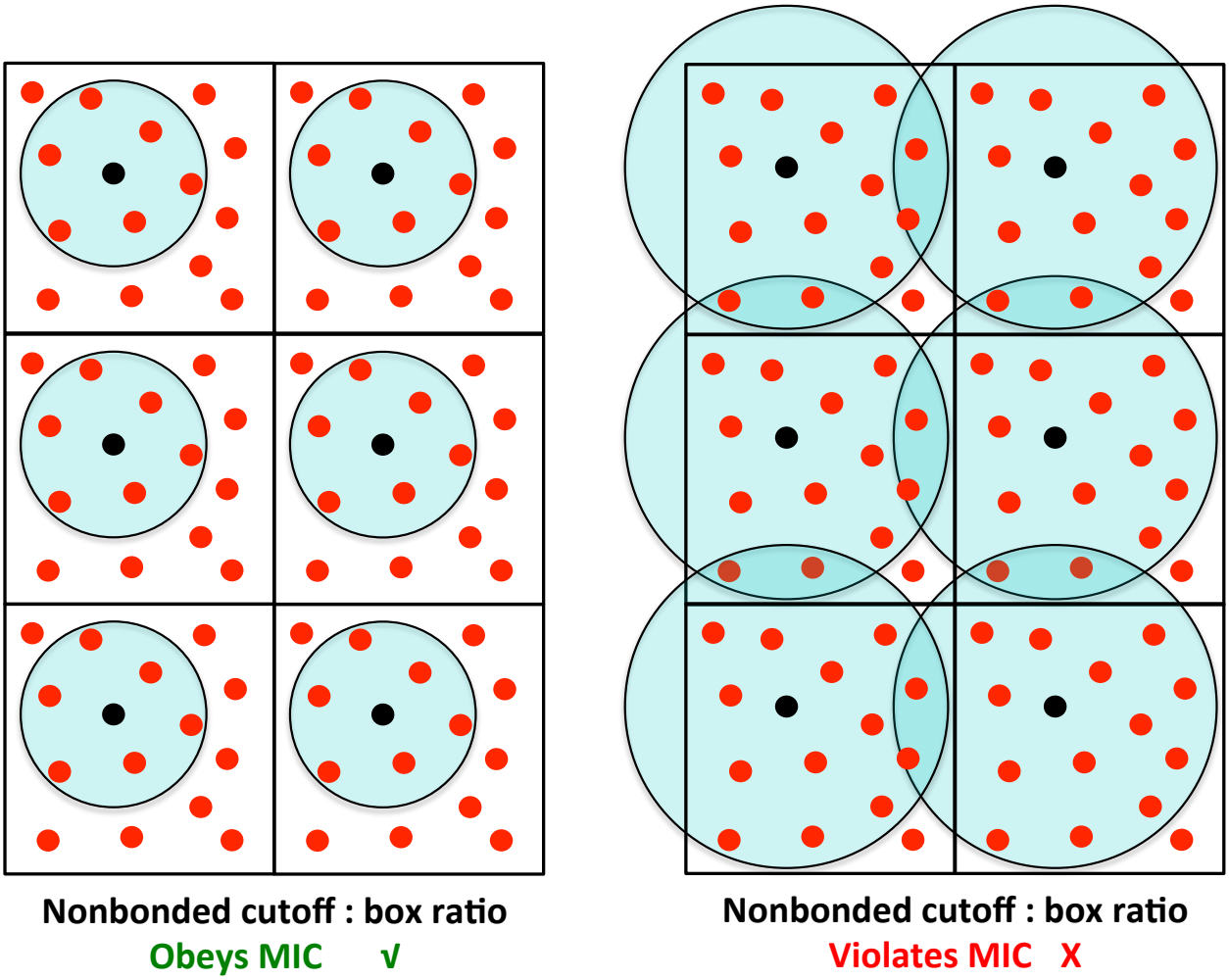


Figure 3.9: According to the minimum image convention, the shortest simulation box length must be at least twice the value of the nonbonded interaction cutoff. If this convention is not observed, the cutoff to box size ratio could allow multiple copies of atoms to be counted in the nonbonded interaction calculation.

While the overall kinetic and potential energies of the system may fluctuate, the sum of the two remains constant and conserved over time. In this way, energy represents a conserved quantity that defines the accessible phase space of the system. If left to its own devices, in the absence of any external controls, a physically accurate simulation of an isolated system will maintain a constant total energy.

An alternative and more commonly employed thermodynamic ensemble is the micro-canonical ensemble (nVT), where number of particles n, volume V, and system temperature T are held constant. This is achieved by coupling the system to a thermostat, or an algorithm designed to maintain a target temperature. Constant temperature conditions are more appropriate for biomolecular simulations, as these molecules exist in thermal equilibrium with their surroundings both *in vitro* and *in vivo*. Coupling to an external thermostat means the system is no longer isolated, and temperature manipulation is accomplished via energy exchange between the two. Thus, total energy of the system itself depends on current temperature. Likewise, the phase space available to the system is determined by current temperature, and this collection of accessible states will change if the value is adjusted. Numerous thermostat algorithms exist to facilitate temperature control, and some commonly employed examples implemented in the AMBER software package [28, 29] include the Berendsen [30], Andersen [31], and Langevin [32] thermostats.

A third, and yet more appropriate thermodynamic ensemble is the isothermal-isobaric ensemble (nPT), where number of particles n, system pressure P, and system temperature T are held constant. Again, a thermostat algorithm is employed to maintain temperature, and the system is further coupled to a barostat to maintain a target pressure. Two common barostat algorithms are the Berendsen-type [30] and Monte-Carlo [33] barostats.

According to the ideal gas law, given in Equation 3.13, pressure P and volume V are inversely proportional and interdependent. That is, if pressure is to remain constant, then the volume of the system may adjust accordingly, and vice versa.

$$PV = nRT \tag{3.13}$$

The relationship between pressure and volume in a biomolecular simulation ultimately determines whether the employed explicit solvent model reaches its target density, and it is generally recommended that a system be subjected to nPT conditions to facilitate density equilibration before an nVT ensemble is explored.

Energy Minimization

Since the future coordinate positions of a biomolecule under simulation are ultimately dependent on current coordinate positions, as per the fundamental law of determinism/reversibility, systems are typically subjected to energy minimization prior to dynamical study. This process endeavors to locate a minimum point on the potential energy surface of the system in order to eliminate steric clashes and resolve any other unfavorable aspects of conformation. The configuration corresponding to this low energy state is then employed as a starting coordinate structure for MD simulation, thereby enhancing calculation stability and allowing sampling to proceed from a reasonable area of phase space. Commonly encountered energy minimization methods include the steepest descent, conjugate gradient, and Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [34] algorithms. Due to limitations present in each of these mathematical methods, the energy minimization procedure may not successfully locate the absolute global minimum on the potential energy surface. However, even a local energy minimum generally corresponds to a structural configuration suitable to commense a stable MD simulation.

Sampling and Convergence

As previously discussed, MD simulations facilitate time dependent exploration of the phase space available to a system, subject to any external constraints. According to the concept

of determinism/reversibility, given these constraints and current system conditions, there is only one possible future coordinate position for each atom in the system upon each time step. Consequently, there is only one possible future position or state for the entire bimolecular system upon each time step, and thus only one possible trajectory through phase space the system can travel along. Also following from determinism/reversibility, this trajectory is continuous, essentially forming a complete circuit through phase space as long as system conditions and constraints (according to the thermodynamic ensemble) remain constant.

Any system of particles under motional study has $6N$ degrees of freedom (x_c, y_c, z_c of configuration and x_p, y_p, z_p of momentum), where N is the total number of particles. Given the size of typical bimolecular systems, and the fact that each atom contributes six degrees of freedom, sampling all available phase space of the system can prove computationally problematic. According to the ergodic hypothesis, all states comprising the ensemble for accessible phase space are equally probable, and as sampling time increases, the time average of state properties will increasingly approximate the overall ensemble average. Generally, the timescale necessary to sample all states of the system and thereby achieve true convergence of ensemble properties is computationally unreachable. Seemingly long sampling times for relatively small systems are proven inadequate to describe all possible motional behavior of the molecule [35], and the degree of convergence commonly achieved in such simulations may only characterize a small portion of phase space. Nevertheless, meaningful information may still be obtained, given simulation timescales sufficient to produce a reasonably representative and statistically relevant ensemble that, at the very least, demonstrates a measure of apparent convergence for the particular property of interest.

A commonly employed technique for expanding accessible simulation timescales is to constrain fast motions in the system, thus facilitating larger time steps. Examples of this include use of the SHAKE [9] and LINCS [10] algorithms, which constrain bonds to hydrogen and allow time steps of up to 2 fs. As further motional constraints are applied, greater time steps become possible, although at the increasing cost of atomistic detail.

Various other strategies also exist for achieving enhanced sampling, while maintaining computational feasibility. The simplest, safest, and most straightforward of these involves performing, in place of a single, long timescale simulation, multiple simulations on shorter timescales, employing independent, uncorrelated starting coordinates. This approach increases sampling by allowing exploration of different areas of phase space, beyond what might be visited along a single, relatively short trajectory. Other, more elaborate methodologies have likewise been developed, and are becoming increasingly popular for investigating extended timescale system behaviors. These include targeted MD (TMD) [36], umbrella sampling (US) [37], temperature- or Hamiltonian-based replica-exchange MD (REMD) [38], and accelerated MD (aMD) [39]. While such techniques require particular care to utilize with success, they have the capacity to produce scientifically compelling results that are generally unobtainable with brute-force MD methodology.

CHAPTER 4:
THERMODYNAMIC INTEGRATION CALCULATIONS[1]

Thermodynamic integration (TI), originally proposed by Kirkwood in 1935 [40], can be used to compute absolute binding free energies, or, more frequently, the difference in binding free energy between two closely-related states. In terms of protein-carbohydrate complexes, TI may, in principle, be employed to compute:

1. Absolute ligand binding energies, by employing TI with ligand annihilation.
2. The effects of protein side chain mutations, characterized by relative binding energies.
3. The effects of structural changes in a ligand, such as the loss or modification of a hydroxyl group, again characterized by relative binding energies.

In the first case, the ligand is annihilated over the course of the TI simulation, and the difference of the bound relative to the unbound state essentially gives rise to an absolute ligand binding energy. The latter two cases quantify the relative energies associated with structural differences, and can be useful for studying how mutations in a protein affect ligand binding, or how well different ligands bind to the same protein. Though much of the same information can be obtained at a lesser computational expense through endpoint methods, such as those involving molecular mechanics – Poisson-Boltzmann / generalized Born surface area (MM-PB/GBSA) [41] approximations, the TI approach offers the advantage of direct incorporation of desolvation and entropic effects. Over the years, TI has been applied in the

[1]The introductory paragraph, as well as the General Theory and Soft-core Potentials sections of this document were adapted from: Hadden, J. A.; Tessier, M. B.; Fadda, E.; Woods, R. J. In *Methods in Molecular Biology: Glycoinformatics*; Lutteke, T.; Frank, R., Eds.; Humana Press: Totowa, NJ; Chapter 8: Calculating binding free energies for protein-carbohydrate complexes; Submitted.

study of protein-binding interactions in complexes with carbohydrates [26, 42–47], as well as glycomimetic drugs, such as inhibitors of influenza neuraminidase [48, 49].

General Theory

TI is commonly referred to as computational alchemy. While MM-PB/GBSA is a post processing method that utilizes frames from a MD trajectory, data for TI calculations are collected numerically over the course of a MD simulation in which the initial state (state A) is alchemically mutated to the final state (state B). This mutation is accomplished through incorporation of a nonphysical mixing parameter λ, which is used to couple the two states and interpolate between them by mediating their contributions to the potential V of the mixed (mutating) system, given by Equation 4.1.

$$V(\lambda) = (1 - \lambda)V_A + \lambda V_B \tag{4.1}$$

Values of λ range from $\lambda = 0$, where the system is wholly state A with no coupling to B, to $\lambda = 1$, where the system is wholly state B with no coupling to A. For each step of the MD simulation, the potential is calculated as *what it would have been* for both state A and state B. These two potentials are combined via λ to generate the mixed potential, which is then applied to propagate the motion of the mixed system. It is therefore the mixed system, which lies somewhere between states A and B, as per the value of λ, that is effectively simulated, evolving according to its mixed potential and propagating as a single set of coordinates.

Generally, simulations are performed at a number of discrete λ windows between $\lambda = 0$ and $\lambda = 1$, during which ensemble averaged values of $< \partial V(\lambda)/\partial\lambda >$ are collected. The relative free energy change between states A and B can then be obtained via integration over the resulting function as $\lambda\ 0 \rightarrow 1$, according to Equation 4.2.

$$\Delta\Delta G = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{4.2}$$

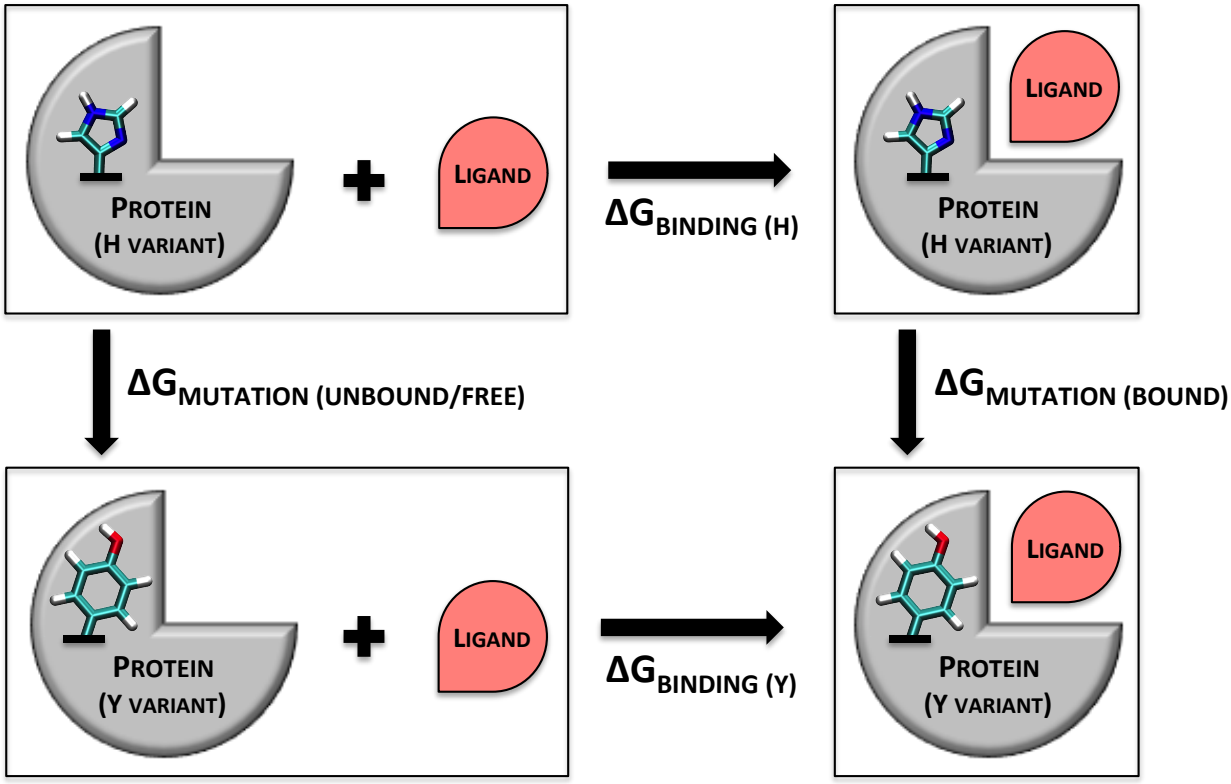


Figure 4.1: Representation of a full thermodynamic cycle for the mutation of a protein residue. The mutation must be completed in both the bound and unbound systems, then the free energy difference between these describes the free energy associated with binding.

In order to obtain a relative free energy of binding, a full thermodynamic cycle must be completed (Figure 4.1). That is, the mutation must be performed in both the bound and unbound systems, then the relative binding energy is given by the difference of these, as per Equation 4.3.

$$\Delta\Delta G_{Binding} = \Delta\Delta G_{Bound} - \Delta\Delta G_{Unbound} \tag{4.3}$$

Further, as TI is an equilibrium method, the value of $\Delta\Delta G$ can be calculated in either the forward or reverse directions, using state A or B as the initial state, as indicated by Equation 4.4.

$$\Delta\Delta G_{Forward} = -\Delta\Delta G_{Reverse} \tag{4.4}$$

Dummy Atoms

Several strategies exist for dealing with the circumstance of partially coupled/decoupled atoms necessitated during TI simulations. One of these is the dummy atom approach.

In CM methodology, the atomic model typically comprises a point charge in configuration space, shielded by a surrounding vdW sphere. Accordingly, dummy or ghost atoms are simply points with assigned charge and vdW parameters of zero, which may also maintain inconsequential covalent bonds to their neighbors. Dummy atoms therefore lack properties that would cause them to interfere with the behavior of other atoms in the system, allowing them to overlap with those atoms, while theoretically having no influence on them and feeling no influence from them. In the context of TI, they are simply a method used for bookkeeping when appearing (coupling) or disappearing (decoupling) atoms from a system under the conditions of standard (hard-core) Lennard-Jones and Coulomb potentials. For TI implementations requiring that the number of atoms and their coordinate positions remain constant between states A and B, a deficiency in either state is compensated for by the inclusion of dummy atoms, which essentially represent fully decoupled particles of the opposite state.

TI simulations employing dummy atoms are typically executed as two-step transformations, with separate steps for Lennard-Jones (vdW) and Coulomb (electrostatic) interactions (Figure 4.2). This is done so that an atomic charge never becomes unshielded by its corresponding vdW radius, leading to a catastrophic collapse in the system. If atoms are appearing or are becoming larger, the Lennard-Jones step must come first in order to generate appropriate vdW shielding to accommodate a forthcoming charge coupling. If atoms are disappearing or are becoming smaller, the Coulomb step must come first, with the vdW shielding curtailed behind the decoupled charge.

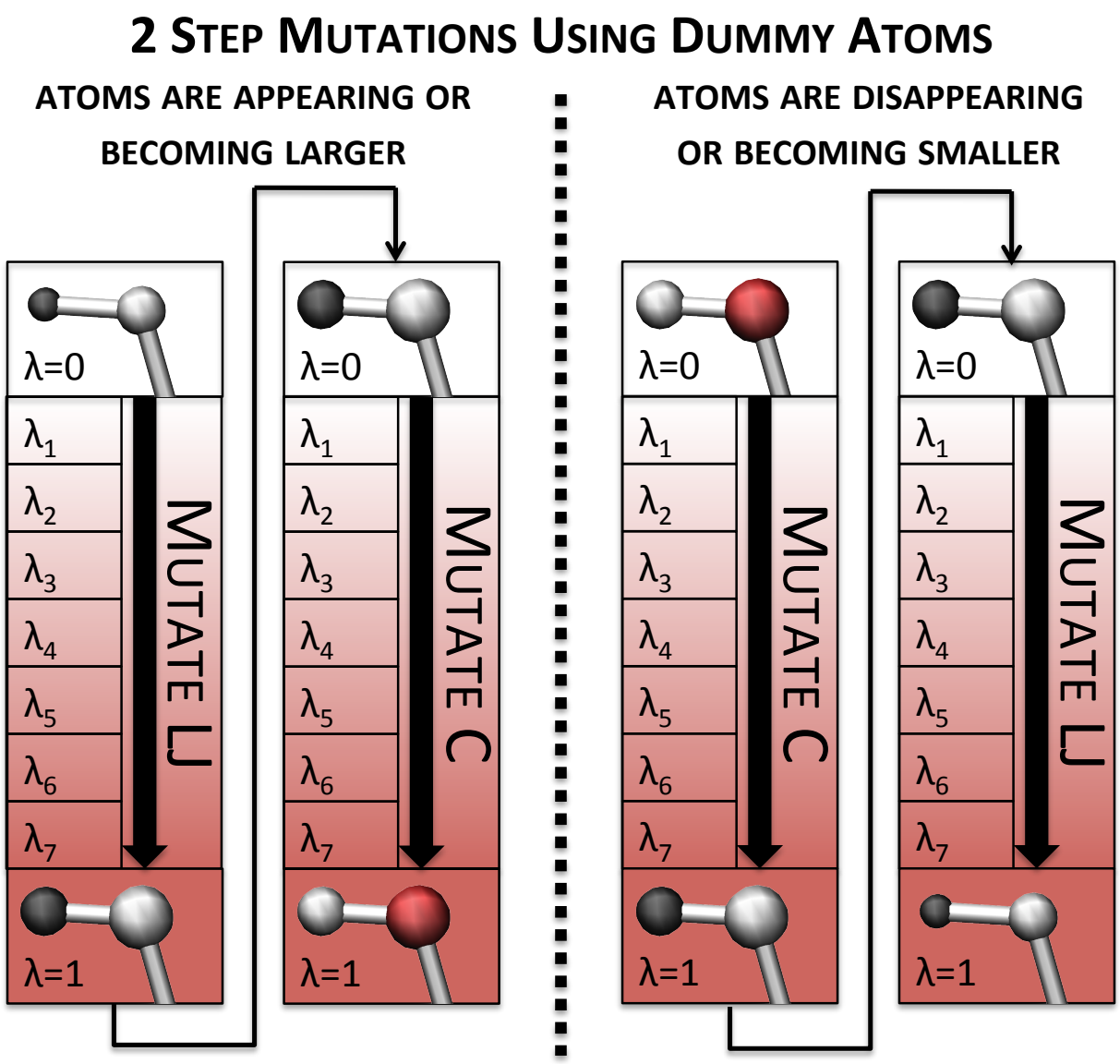


Figure 4.2: In these examples, a hydrogen (H) is mutated into a hydroxyl group (OH) as λ 0 $\rightarrow$ 1 [left panel], and a hydroxyl group (OH) is mutated into a hydrogen (H) as λ 0 $\rightarrow$ 1 [right panel]. The black spheres represent dummy atoms, or fully decoupled particles of the opposite state. If atoms are appearing, or becoming larger [left panel], then the Lennard-Jones interactions must be mutated $A \rightarrow B$ in the first step, followed by Coulomb interactions $A \rightarrow B$ in the second step. If atoms are disappearing or becoming smaller [right panel], then the Coulomb interactions must be mutated $A \rightarrow B$ in the first step, followed by Lennard-Jones interactions $A \rightarrow B$ in the second step. This protocol ensures that an atomic charge never becomes unshielded by its corresponding vdW radius.

The mixed potential for a TI simulation employing dummy atoms is typically calculated according to Equation 4.5.

$$V(\lambda) = (1-\lambda)^k V_A + \left[1 - (1-\lambda)^k\right] V_B \tag{4.5}$$

When $k = 1$, Equation 4.5 reduces to the standard form given in Equation 4.1, and mixing is performed linearly. However, non-linear mixing, invoked when $k > 1$, is necessary when fully decoupling atoms under the conditions of standard Lennard-Jones and Coulomb potentials in order to avoid singularities (further discussed in the following section). A value of $k \geq 4$ generally ensures the integrand remains finite at the endpoints when dummy atoms are employed [50].

Soft-core Potentials

An alternative and widely used strategy for dealing with partial atomic couplings in TI simulations is the application of soft-core potentials for the mutation of nonbonded interactions.

For a system where state A has N atoms and state B has $N + 1$ atoms, the mutation corresponds to appearing the additional atom as λ 0 $\rightarrow$ 1. At $\lambda = 0$, the atom is fully decoupled from the overall system and should have no effect on it, while at $\lambda = 1$, the atom is fully coupled and should interact normally. The standard Lennard-Jones potential between two atoms i and j, separated by distance R_{ij} is given by Equation 4.6.

$$V(\lambda)_{ij}^{LJ} = \lambda 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] \tag{4.6}$$

However, when dealing with decoupled atoms, Equation 4.6 has the limitation that as $\lambda \rightarrow 0$ a singularity can occur. This is known as the origin singularity or vdW endpoint problem [50]. As the atom is decoupled, overlap with coupled atoms becomes possible, allowing $R_{ij} \rightarrow 0$, where the Lennard-Jones potential becomes undefined. Concomitantly, as $R_{ij} \rightarrow 0$, the Lennard-Jones interactions become increasingly repulsive, causing instabilities in the

potential calculation. This affects the numerical integration of the MD simulation, and furthermore, the value of $< \partial V(\lambda)/\partial\lambda >$ when performing TI.

Various methods have been developed to address these issues, including analytical fitting schemes [51, 52], slow growth methods [53], and nonlinear mixing functions for $V(\lambda)$ [50, 54–59]. Nevertheless, most modern TI implementations employ soft-core potentials [59, 60] to allow overlap of decoupled particles, while avoiding endpoint singularities. The functional form of the soft-core Lennard-Jones potential, as given by Beutler et al. [59] and implemented in the AMBER software package [28, 29], is shown in Equation 4.7.

$$V(\lambda)_{ij}^{LJ} = \lambda 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{(\alpha_{LJ}(1-\lambda)\sigma_{ij}^{m} + R_{ij}^{m})^{1/m}}\right)^{12} - \left(\frac{\sigma_{ij}}{(\alpha_{LJ}(1-\lambda)\sigma_{ij}^{m} + R_{ij}^{m})^{1/m}}\right)^{6}\right] \quad (4.7)$$

The supplement of α_{LJ}, a positive constant, prevents the denominator from approaching zero and becoming undefined as $R_{ij} \to 0$. Furthermore, due to the factor of $(1-\lambda)$, the contribution of α_{LJ} will only be invoked as $\lambda \to 0$, when decoupling of the atom necessitates soft-core behavior (Figure 4.3).

Coulomb interactions have the standard form given in Equation 4.8.

$$V(\lambda)_{ij}^{C} = \lambda\left[\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{R_{ij}^{2}}\right] \quad (4.8)$$

These can also be modeled using soft-core potentials, as in the form given by Beutler et al. [59] and implemented in the AMBER software package [28, 29], shown in Equation 4.9.

$$V(\lambda)_{ij}^{C} = \lambda\left[\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\left(\beta_C(1-\lambda) + R_{ij}^{m}\right)^{1/m}}\right] \quad (4.9)$$

Again, the supplement of β_C prevents endpoint singularity and only applies as $\lambda \to 0$.

Note: If the atom is being disappeared as $\lambda\ 0 \to 1$, and is instead fully decoupled at $\lambda = 1$, the value of λ must be replaced with $(1-\lambda)$ and vice versa for the above equations.

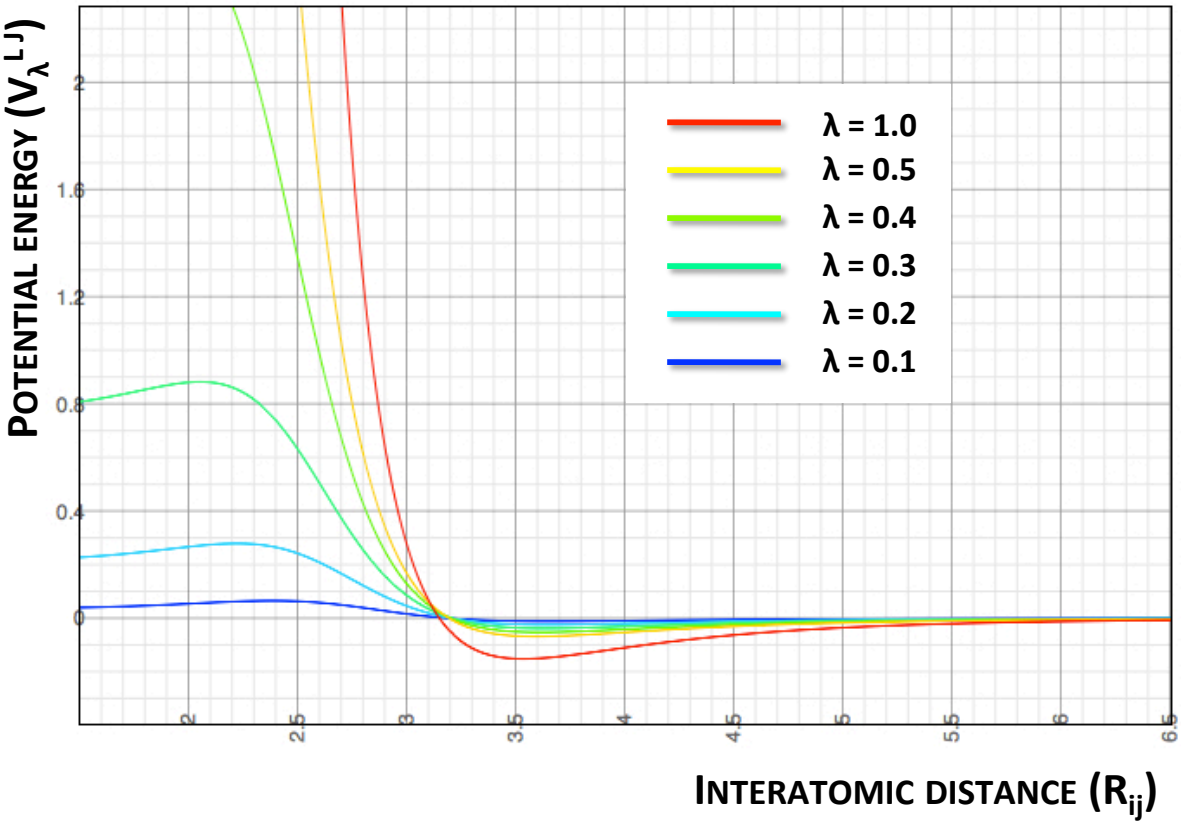


Figure 4.3: This example shows plots for the Lennard-Jones interaction between two equivalent theoretical particles, evaluated at a range of λ values. As $\lambda \to 0$, the potential begins to exhibit soft-core behavior and remains defined despite the overlap of partially decoupled atoms with others in the system.

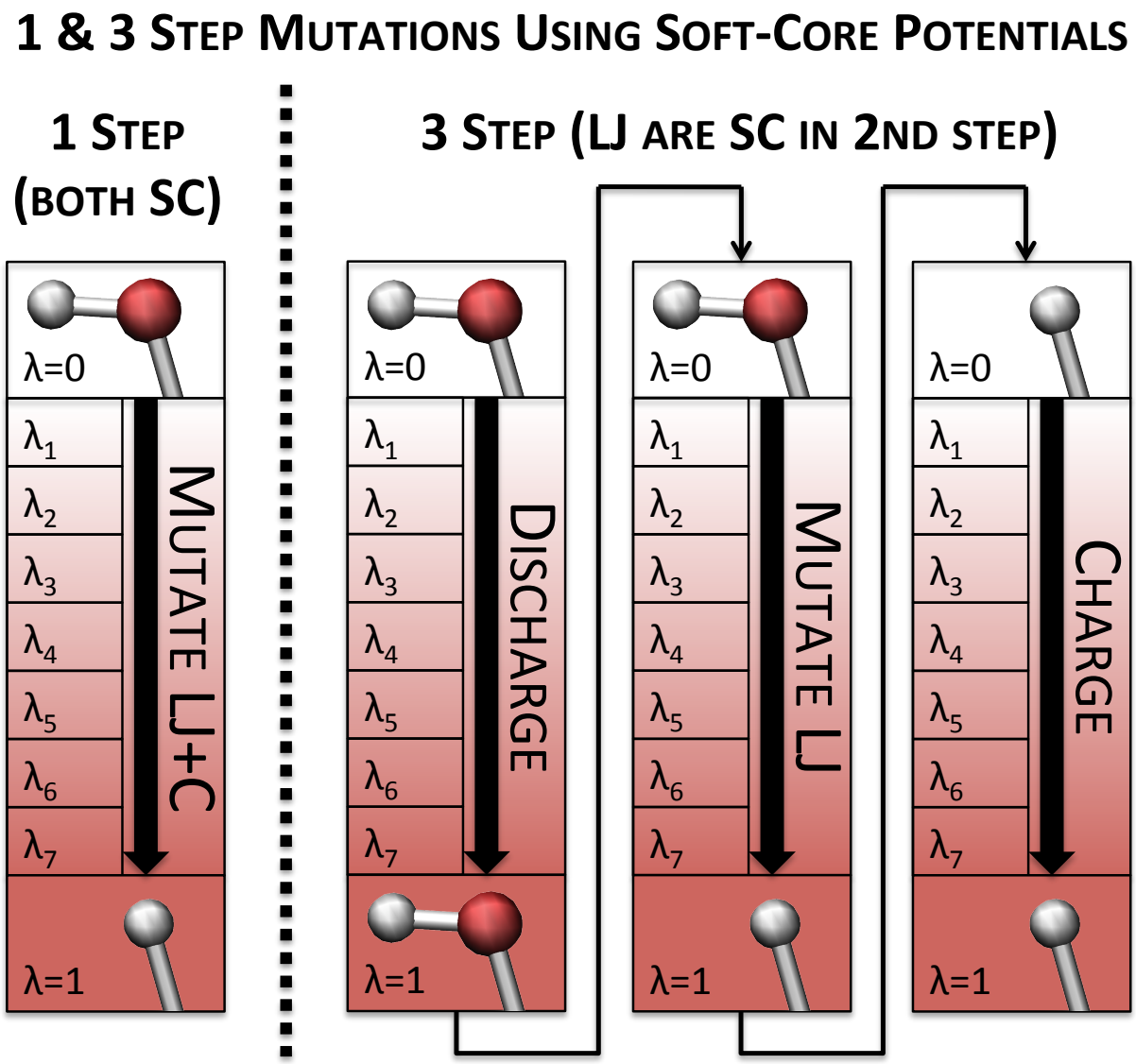


Figure 4.4: In this example, a hydroxyl group (OH) is mutated into a hydrogen (H) as λ $0 \rightarrow 1$. When soft-core potentials are used for both Lennard-Jones and Coulomb interactions, TI simulations can be run as single-step mutations, with no transition states. However, if soft-core potentials are only employed for Lennard-Jones interactions, Coulomb interactions must be mutated as a separate step in order to avoid the unshielding of atomic charges. The most innocuous way to do this is to perform the vdW mutation in the absence of charge, which requires three mutation steps: (1) Discharge the mutating region of state A, (2) Mutate the Lennard-Jones interactions $A \rightarrow B$, and (3) Charge the mutating region as state B.

When scaling nonbonded properties, care must always be taken that an atomic charge never becomes unshielded by its corresponding vdW radius, as this would lead to a catastrophic collapse in the system. For some TI implementations, this requires dividing the mutation of Lennard-Jones and Coulomb interactions into separate transformation steps, which increases the computational expense. However, if soft-core potentials are implemented for both Lennard-Jones and Coulomb interactions, mutations can be performed in a single step, given suitable optimization of α_{LJ} and β_C [61] (Figure 4.4).

Integration Methods

During a TI simulation, ensemble averaged values of $< \partial V(\lambda)/\partial\lambda >$ are collected over discrete λ windows in order to construct a function of $< \partial V(\lambda)/\partial\lambda >$ vs. λ for the mutation. Integration over this function as $\lambda\ 0 \rightarrow 1$ produces a value for the corresponding free energy change, as previously shown in Equation 4.2. Ideally, the calculation is performed over many λ windows, such that sufficient points are generated to well-describe the $< \partial V(\lambda)/\partial\lambda >$ vs. λ function, thus increasing accuracy of the free energy estimate. However, due to the inherent computational expense of TI simulations, it is generally only feasible to compute a relatively small subset of λ windows. This is particularly true for systems as large as protein-carbohydrate complexes. Within the number of computationally accessible λ windows for a given system, specific λ values are generally selected based on constraints or recommendations relevant to the integration method that will be employed to calculate the final free energy difference.

A simple and widely used integration technique is the trapezoidal rule, which performs linear interpolation between consecutive λ values to evaluate the integral, as shown in Equation 4.10. This method can be used with any number of λ values, separated by any spacing increment, equal or otherwise.

$$\int_0^1 f(\lambda)\ d\lambda = \sum_{i=1}^{N-1} (\lambda_{i+1} - \lambda_i) \frac{f(\lambda_{i+1}) + f(\lambda_i)}{2} \tag{4.10}$$

However, most TI simulations produce functions that are far from linear, with the consequence that the trapezoidal rule fails to well-approximate the function and may give rise to free energy differences that contain significant inherent error.

A more accurate integration method commonly employed in the literature is Simpson's rule, which interpolates between three consecutive λ values using a quadratic polynomial, as in Equation 4.11. This method requires an odd number of λ values, with the constraint that the spacing between any three consecutive points be equal.

$$\int_0^1 f(\lambda)\, d\lambda = \sum_{i=1}^{(N-1)/2} (\lambda_{2i+1} - \lambda_{2i-1}) \frac{f(\lambda_{2i-1}) + 4f(\lambda_{2i}) + f(\lambda_{2i+1})}{3} \tag{4.11}$$

Several studies have previously compared the effect of the trapezoidal rule and Simpson's rule on the accuracy and computational efficiency of TI calculations, and have found Simpson's rule to be the superior of the two [62, 63].

Alternatively, higher accuracy may be achieved by fitting the $< \partial V(\lambda)/\partial\lambda >$ vs. λ dataset to a functional form that well-describes its behavior and performing the integration analytically. Several independent studies have proposed techniques for executing this strategy and demonstrated its merit over more simplistic integration methods [64–66].

A further strategy for obtaining accuracy in TI calculations with a limited number of λ windows is to employ a quadrature scheme, such as Gaussian quadrature, which utilizes a recommended set of λ values especially selected to well-approximate the integral over a polynomial of up to degree $2n - 1$. The ensemble averaged values of $< \partial V(\lambda)/\partial\lambda >$ collected for each λ window are combined as a weighted sum, according to a corresponding set of weights, to estimate the overall free energy change, as shown in Equation 4.12.

$$\Delta\Delta G = \sum_i w_i \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_i \tag{4.12}$$

Table 4.1 provides a listing of λ/weight sets for up to 12-point Gaussian quadrature. This method for estimating the value of an integral is particularly useful in the case of large

Table 4.1: λ values and weights for integral estimation with Gaussian quadrature.

n	λ_i	$1-\lambda_i$	w_i
1	0.5		1.0
2	0.21132	0.78867	0.5
3	0.1127	0.88729	0.27777
	0.5		0.44444
5	0.04691	0.95308	0.11846
	0.23076	0.76923	0.23931
	0.5		0.28444
7	0.02544	0.97455	0.06474
	0.12923	0.87076	0.13985
	0.29707	0.70292	0.19091
	0.5		0.20897
9	0.01592	0.98408	0.04064
	0.08198	0.91802	0.09032
	0.19331	0.80669	0.13031
	0.33787	0.66213	0.15617
	0.5		0.16512
12	0.00922	0.99078	0.02359
	0.04794	0.95206	0.05347
	0.11505	0.88495	0.08004
	0.20634	0.79366	0.10158
	0.31608	0.68392	0.11675
	0.43738	0.56262	0.12457

systems, such as protein-carbohydrate complexes, because it significantly reduces the number of λ windows necessary for an accurate free energy estimate, thereby reducing computational expense and increasing computational feasibility.

Finally, a practical tactic for increasing the accuracy of any of these integration methods is to assess the smoothness of the resulting $< \partial V(\lambda)/\partial\lambda >$ vs. λ function and add in additional λ windows (subject to the spacing or quadrature requirements of the method) as necessary to improve the characterization of the curve.

Error and Uncertainty Estimates

As with any experimentally measured or computationally predicted quantity, the free energy values calculated with TI must be qualified with a corresponding estimate of statistical precision or uncertainty.

It should first be noted that statistical errors are a separate consideration from systematic errors that arise directly from TI methodology as applied in a given calculation. For example, a commonly employed technique for estimating methodologically inherent error is to perform a TI simulation in both the forward (λ 0 $\rightarrow$ 1) and reverse directions (λ 1 $\rightarrow$ 0) in order to evaluate any hysteresis in the resulting free energies. Because TI is an equilibrium method that represents a thermodynamically reversible process (Equation 4.4), an accurate calculation should be free of hysteresis. While such an error assessment is useful, it is often erroneously applied to estimate uncertainty values in TI data [67], although it provides no statistically relevant quantification.

A more suitable technique to gauge statistical uncertainty is to calculate the standard deviation of the mean (SDM) for the free energy, accomplished by calculating SDM for $< \partial V(\lambda)/\partial\lambda >$ within each λ window and combining them as a weighted sum over all λ windows, as shown in Equations 4.13 and 4.14 [68].

$$\sigma_{\Delta G_i} = \sqrt{\sum_i w_i^2 \sigma_{SDMi}^2} \tag{4.13}$$

where

$$\sigma_{SDMi} = \sigma_{\partial V(\lambda)/\partial\lambda} \; / \sqrt{t_{sim}/2\tau} \tag{4.14}$$

Within Equation 4.14, $\sigma_{\partial V(\lambda)/\partial\lambda}$ is the standard deviation, τ is the autocorrelation time of $\partial V(\lambda)/\partial\lambda$, and t_{sim} is the total length of the simulation. This metric assumes $\partial V(\lambda)/\partial\lambda$ values are statistically uncorrelated over subsequent λ windows, however Lawrenz et al. [48] point out that this cannot be strictly true given that $\sigma_{\Delta G_i}$ includes the physically based fluc-

tuations of $\partial V(\lambda)/\partial\lambda$, and that overlap of phase space at subsequent λ windows is necessary to generate smooth $\partial V(\lambda)/\partial\lambda$ vs. λ curves. While $\sigma_{\Delta G_i}$ is thus deemed a questionable measure of uncertainty, it nevertheless represents the lowest possible uncertainty for free energy differences calculated from a single standard TI simulation [48], and as a result, is widely used in the literature.

In general, the primary factor that contributes to both systematic and statistical error in an otherwise methodologically sound TI calculation is inadequate sampling of thermally relevant phase space, or the resulting lack of convergence in the simulation. This is most often an unavoidable consequence of the inherent computational expense of TI calculations, especially when applied to large biological systems such as protein-carbohydrate complexes. Moreover, simulations performed on modest timescales can falsely appear to have converged, especially if the system under study is subject to long timescale configurational fluctuations.

An alternative computational approach that both improves phase space sampling and produces an uncertainty estimate with clear statistical validity is independent trajectory TI (IT-TI) [48]. As the name implies, this method involves running multiple TI simulations from independent, uncorrelated starting structures to generate a more comprehensive and informative dataset. The resulting free energies are averaged over the N runs to obtain a final ΔG, and the uncertainty is calculated as SDM, as per Equation 4.15.

$$\sigma_{\Delta\overline{G}} = \frac{\sigma_{\Delta G}}{\sqrt{N}} \tag{4.15}$$

According to a comparative study by Lawrenz et al., free energies computed with standard single-trajectory TI simulations ranged from a 19 % underestimation to a 29 % overestimation of an experimental reference value, while IT-TI reproduced the reference value to a 2 % relative difference [48]. This enhanced capacity to achieve accuracy, combined with more statistically relevant error estimation, thus supports IT-TI as a preferred strategy for performing free energy calculations.

Finally, once $\Delta G_{Binding}$ is obtained through completion of a full thermodynamic cycle (Equation 4.3), its associated uncertainty is computed based on standard propagation of statistical errors, as given in Equation 4.16, regardless of how the respective errors were determined.

$$\sigma_{\Delta G_{Binding}} = \sqrt{\sigma^2_{\Delta G_{Bound}} + \sigma^2_{\Delta G_{Unbound}}} \tag{4.16}$$

Although the above measures of $\sigma_{\Delta G}$ (Equations 4.13 and 4.15) provide reasonable estimates of statistical uncertainty resulting from configurational fluctuations and sampling limitations in TI simulations, they do not account for errors arising from the integration step of TI calculations. As such, an additional metric that may be employed to gauge error related to integration is the finite difference curvature, shown in Equations 4.17 (even λ spacing) and 4.18 (uneven λ spacing), which measures the smoothness of a free energy curve [61].

$$C = \frac{1}{N-2}\sum_{i=2}^{N-1} = | < \partial V(\lambda)/\partial\lambda >_{i-1} -2 < \partial V(\lambda)/\partial\lambda >_i + < \partial V(\lambda)/\partial\lambda >_{i+1} | \tag{4.17}$$

for evenly spaced λ windows, or

$$C = \frac{1}{N}\sum_{i=1}^{N} | \ \partial^2 V(\lambda)/\partial\lambda^2 \ | \tag{4.18}$$

for unevenly spaced λ windows.

If the free energy function is not smooth, or is otherwise described by an insufficient number of points to well-approximate its true shape, then integration according to any rule or polynomial fit will likely lead to discrepancies in the final free energy value. Smaller magnitudes of C indicate a smoother curve, and a corresponding smaller error from integration.

As a closing point, it should be noted that while TI methodology is capable of determining highly precise free energy values (with small associated statistical uncertainties) that compare reasonably well with experiment for simple systems containing relatively few

numbers of atoms (for which simulation convergence is readily attainable), for example, in the well-known work of Shirts et al. [67], it is frequently not the most practical technique for quantitatively computing free energy differences. In particular, recent studies comparing popular free energy methods generally recommend Bennet acceptance ratio (BAR) as a more robust and computationally efficient approach than TI [63, 69–73].

CHAPTER 5:
SYSTEM UNDER STUDY: CELLULOSE

Cellulose Structure

Cellulose is the common name describing the polymer of β(1-4)-D-glucose. A primary structural component of plant cell walls, it is the most abundant biopolymer on Earth. The smallest unit of cellulose is the disaccharide cellobiose, which can be obtained via hydrolysis. Otherwise, cellulose does not occur as a single polysaccharide chain, but exists, from synthesis, as an ordered bundle of chains referred to as a microfibril. Internally, this primary structural unit is characterized by a well-defined hydrogen bond network that associates arrays of parallel cellulose chains into layers, which are then stacked hydrophobically. The details of this hydrogen bond network determine the specific cellulose crystalline phase. Cellulose I represents the natural form, comprised by two distinct, and often coexistent phases, Iα and Iβ. These differ only in relative alignment of hydrophobic layers. Bacterial and algae cellulose exists primarily as Iα, while plant cellulose exists primarily as Iβ. Through high-resolution X-ray and neutron diffraction studies, Nishiyama et al. resolved comprehensive crystallographic coordinates describing both variants [74, 75], and structural models based on these provide a solid experimental foundation for computational work employing biomolecular simulation.

Diffraction Methodology as Applied to Cellulose

Experimental diffraction studies of cellulose involve bombarding samples with X-ray or neutron beams in order to observe the resultant scattering effect produced by the crystalline

lattice. While incident neutrons are scattered by atomic nuclei, incident X-rays are scattered only through contact with electron density. For this reason, the sensitivity of X-ray diffraction is limited to the detection of heavy atoms, such as carbon and oxygen. Neutron diffraction may be employed further as a complementary technique to additionally detect the positions of hydrogen atoms. Upon striking the sample, a portion of the beam intensity is re-radiated by each scatterer as a spherical wave. If the scatterers are arranged in a regular, repeating pattern, such as in a crystal lattice, these waves can add constructively in specific directions. The angle for coherent scattering is determined according to Bragg's law, given in Equation 5.1.

$$n\lambda = 2d\ sin\ \theta \tag{5.1}$$

Here, n is any integer, λ is the wavelength of the incident beam, d is the spacing between the diffracting planes comprised of regularly spaced scatterers, and θ is the angle between the incident beam and the diffracting planes. This means that waves will add constructively only in directions such that the interplanar spacing d is proportional to an integer multiple of the beam wavelength $n\lambda$. In this case, the difference in path length between the waves is also equivalent to $n\lambda$, thus the waves remain in phase. The coherent scattering produced by a crystalline sample is detected as a series of regularly spaced reflections referred to as a diffraction pattern. From these data, a crystallographer can infer information that describes the spatial and structural properties of the sample, including specific relative atomic positions and unit cell dimensions for the crystal lattice. Various diffraction techniques are available, each involving special sample preparation and producing distinct types of diffraction patterns. Two techniques relevant to cellulose study are fiber and power diffraction.

Molecular Dynamics Simulations of Cellulose

The starting coordinates for computational simulations of complex biological molecules are often taken from experimentally-derived structures in order to ensure a solid and realistic

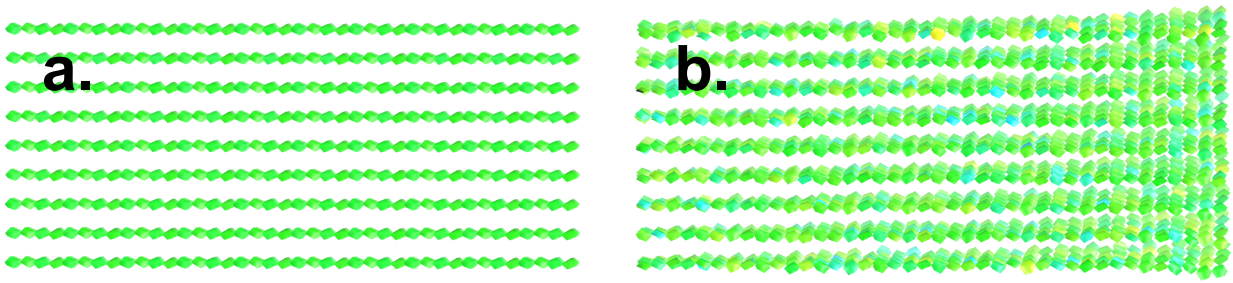


Figure 5.1: Side view of a model microfibril based on the cellulose Iβ crystallographic coordinates (a) before and (b) after MD simulation.

foundation for predictions. In the case of cellulose, theoretical studies routinely employ models based on the benchmark crystallographic data produced by Nishiyama et al. [74, 75]. While these high-resolution structures suggest that the polysaccharide chains comprising a given microfibril are perfectly parallel, and that microfibrils should display an overall linear conformation, MD simulations of model microfibrils based on these data tend to undergo a series of distortions, including the adoption of an overall right-handed twist (Figure 5.1). This phenomenon was first reported by Matthews et al. employing CHARMM CSFF [76], but was later observed by numerous other researchers employing a range of alternative empirical carbohydrate force fields [77–84]. Such dramatic discrepancy between experimentally-determined and computationally-predicted structures has incited significant controversy regarding the application of biomolecular simulation methodology to cellulose.

Significance of this Study

Cellulose comprises the major chemical component of cotton and rayon fiber, as well as serves as a primary raw material for the manufacture of paper. MD simulations of cellulose microfibrils in interaction with molecules of water or other solvents could therefore serve to guide the design and refinement of novel fabric and paper treatment processes. Additionally, cellulose is currently under earnest investigation by the biofuels industry as a potential source

for the environmentally sustainable production of ethanol. Computational studies involving MD simulation are directed at understanding how cellulase enzymes interact with the microfibril surface to facilitate natural cellulose degradation. Ideally, the results of such studies may suggest strategies for engineering enzymes to perform degradation at highly accelerated rates, thus enabling commercial-scale bioethanol production. While MD simulations have the capability to further our knowledge of cellulose structure, dynamic behavior, and intermolecular interaction toward improvement of the many industrial processes that employ it as a raw material, the success of all future computational endeavors to model cellulose will ultimately depend on the robustness of available biomolecular simulation methodologies and the manner in which they are applied to the unique case of cellulose microfibrils. The original research studies described in this document aim to enhance general understanding of microfibril twisting, assessing the current state of cellulose modeling capability and further suggesting how it might be improved by:

1. Elucidating the physical forces that drive twisting behavior.

2. Probing the sensitivity of twisting behavior to commonly employed approximations.

3. Examining how structural twisting manifests in theoretical diffraction data.

CHAPTER 6:

UNRAVELING CELLULOSE MICROFIBRILS: A TWISTED TALE[1]

[1]Hadden, J. A.; French, A. D.; Woods, R. J. *Biopolymers* **2013**, *99*, 746-756.
Reprinted here with permission of publisher.

Abstract

Molecular dynamics (MD) simulations of cellulose microfibrils are pertinent to the paper, textile, and biofuels industries for their unique capacity to characterize dynamic behavior and atomic-level interactions with solvent molecules and cellulase enzymes. While high-resolution crystallographic data have established a solid basis for computational analysis of cellulose, previous work has demonstrated a tendency for modeled microfibrils to diverge from the linear experimental structure and adopt a twisted conformation. Here, we investigate the dependence of this twisting behavior on computational approximations and establish the theoretical basis for its occurrence. We examine the role of solvent, the effect of nonbonded force field parameters [partial charges and van der Waals (vdW) contributions], and the use of explicitly modeled oxygen lone pairs in both the solute and solvent. Findings suggest that microfibril twisting is favored by vdW interactions, and counteracted by both intrachain hydrogen bonds and solvent effects at the microfibril surface.

Keywords: cellulose, microfibril twist, molecular dynamics, GLYCAM

Introduction

Naturally occurring cellulose, termed cellulose I, is the most abundant of all biomolecular polymers. Composed of repeating β(1-4)-D-glucosyl residues, it manifests as a crystalline array of parallel chains associated into layers via hydrogen-bonding between equatorial hydroxyl and hydroxymethyl groups. This primary structural unit, known as a microfibril, can be thousands of residues in length, whereas the cross-sectional thickness of a given microfibril, defined by the number of constituent polysaccharide chains, is entirely dependent upon the cellulosic source. The specific cellulose synthase complex employed for biosynthesis is thought to determine not only the number of chains but also microfibril shape and packing arrangement [85]. While cellulose I occurs as two distinct, yet coexistent crystal phases, Iα

and Iβ, that differ only in the relative alignment of the polysaccharide layers, plant-based cellulose microfibrils are believed to exist primarily as the Iβ phase.

Of longstanding importance to the paper and textile industries, cellulose has recently garnered significant interest from the biofuels sector as a potential source for the environmentally sustainable production of ethanol. Computational analysis of microfibril interactions with solvent molecules and cellulase enzymes could guide the design of novel paper manufacturing processes and fabric treatments, as well as shed light on efficient mechanisms for cellulose degradation relevant to ethanol production. High-resolution crystal structures of both the Iα and Iβ polymorphs have established a solid experimental basis for theoretical studies [74, 75], yet molecular dynamics (MD) simulations generally produce results that disagree to varying extents with the crystallographic data [76–84].

Matthews et al. first reported the tendency for cellulose microfibrils to rapidly (<200 ps) diverge from the crystallographic coordinates and adopt a right-handed twist during MD simulation using the CHARMM CSFF force field [76]. Shortly thereafter, Yui et al. reported similar behavior of microfibrils in simulations performed with the GLYCAM06 force field [77, 78]. Since then, further studies employing GLYCAM06, a range of other atomistic carbohydrate force fields including OPLS, CHARMM C35, and GROMOS 45a4, as well as several specifically adapted coarse-grain force fields, have likewise produced twisted structures [79–83]. In addition, simulations with the MARTINI coarse-grain force field have been shown to produce microfibrils with either a right- or left-handed twist depending on the number of constituent cellulose chains, or have otherwise been intentionally parameterized to prevent twisting in finite models [84, 86].

In defense of the tendency for simulated microfibrils to diverge from the crystallographic coordinates and adopt a twisted conformation, it should be noted that such twisting has also been observed under some experimental conditions. Hanley et al. reported visual evidence for twisting in individual microfibrils with cross-sectional thicknesses of 20–50 nm based on over 100 measurements with transmission electron microscopy (TEM), atomic force microscopy

(AFM), and tapping mode AFM (TM-AFM) [87]. These analyses revealed the presence of a periodic right-handed twist occurring in short segments along the microfibrils. The authors proposed that the segmented nature of twisting was likely an artifact of sample preparation and that when the microfibrils were suspended in aqueous solution, prior to being dried down onto substrates for study, they would have exhibited smooth, uniform twisting over their lengths.

Given the experimental evidence to support a twisted microfibril conformation in water, it might be anticipated that these systems would adopt twisted structures under typical biomolecular simulation conditions. Even so, recent work reported by Matthews et al. has served to further complicate interpretation of simulation data. The authors demonstrated that cellulose Iβ microfibrils studied at elevated temperature (500 K) using the CHARMM C35 and GLYCAM06 force fields develop an inter-layer hydrogen bond network, resulting from widespread reorientation of hydroxymethyl groups, and subsequently untwist to a linear structure representing the high-temperature intermediate (I-HT) for phase transformation between cellulose Iα and Iβ [79]. While the I-HT structure appears to agree with high-temperature experimental data and thus suggests appropriate behavior under these conditions, the results of a later study showed that microfibrils simulated at room temperature (300 K) with the same force fields also untwist to the I-HT form on a near-microsecond timescale [81]. Because of insufficient conformational sampling, it remains unclear whether this behavior represents structural convergence in the simulation, or whether the twisted and linear states would occur in some equilibrium.

While the results of MD simulations may reasonably reproduce some experimental data, both the initial twisted and eventual I-HT structures they predict at ambient temperature exhibit significant deviations from the crystallographic coordinates for cellulose Iβ. These structural changes raise questions regarding the driving forces responsible for deviations, as well as draw into question the suitability of classical force fields for application to cellulose. To shed light on the structural and dynamic complexities observed in these simulations,

the present work probes the initial onset of microfibril twisting by evaluating the effect of microfibril model dimensions, the role of solvent and effect of solvent model, the effect of charge set and application of explicitly modeled oxygen lone pairs, and the overall role of nonbonded interactions. Altogether, the results of this study should serve to enhance the general understanding of cellulose microfibril behavior in water, as well as suggest how best to apply MD simulations for the study of cellulose in broader contexts.

Computational Methods

Initial Structures

All initial structures were generated in Mercury 2.0 [88] based on the coordinates for cellulose Iβ reported by Nishiyama et al. [75], with exposed crystallographic faces corresponding to the $1\bar{1}0$ and 110 planes (Figure 6.1a). A microfibril of 81 total chains (9 per face), each consisting of 20 glucosyl residues [degree of polymerization (DP) 20] was taken as the representative model for this study, as discussed further in the text.

Charge Calculations

Charges were either employed as developed for the monosaccharide β-D-glucose (Glcβ) in GLYCAM06 [11], or recomputed for the methyl glycoside of the trisaccharide Glcβ(1-4)Glcβ(1-4)Glcβ generated from the coordinates for cellulose Iβ [75] (Figures 6.2 and 6.3). Any water molecules included in the charge derivation were constrained to have charges corresponding to the TIP3P explicit solvent model [21]. Quantum mechanical (QM) molecular electrostatic potentials (MEPs) were computed with Gaussian03 [89] at the HF/6-31G* level of theory to maintain consistency with the GLYCAM06 charge development protocol [11, 17]. The MEPs were sampled at grid points according to the CHELPG scheme [90]. Partial atomic charges were obtained via fitting a classical electrostatic potential (ESP) to the QM MEPs using the restrained ESP (RESP) method [16, 28]. Alternate charges for the nonreducing terminal residue in each cellulose chain (GLYCAM residue 0GB) were

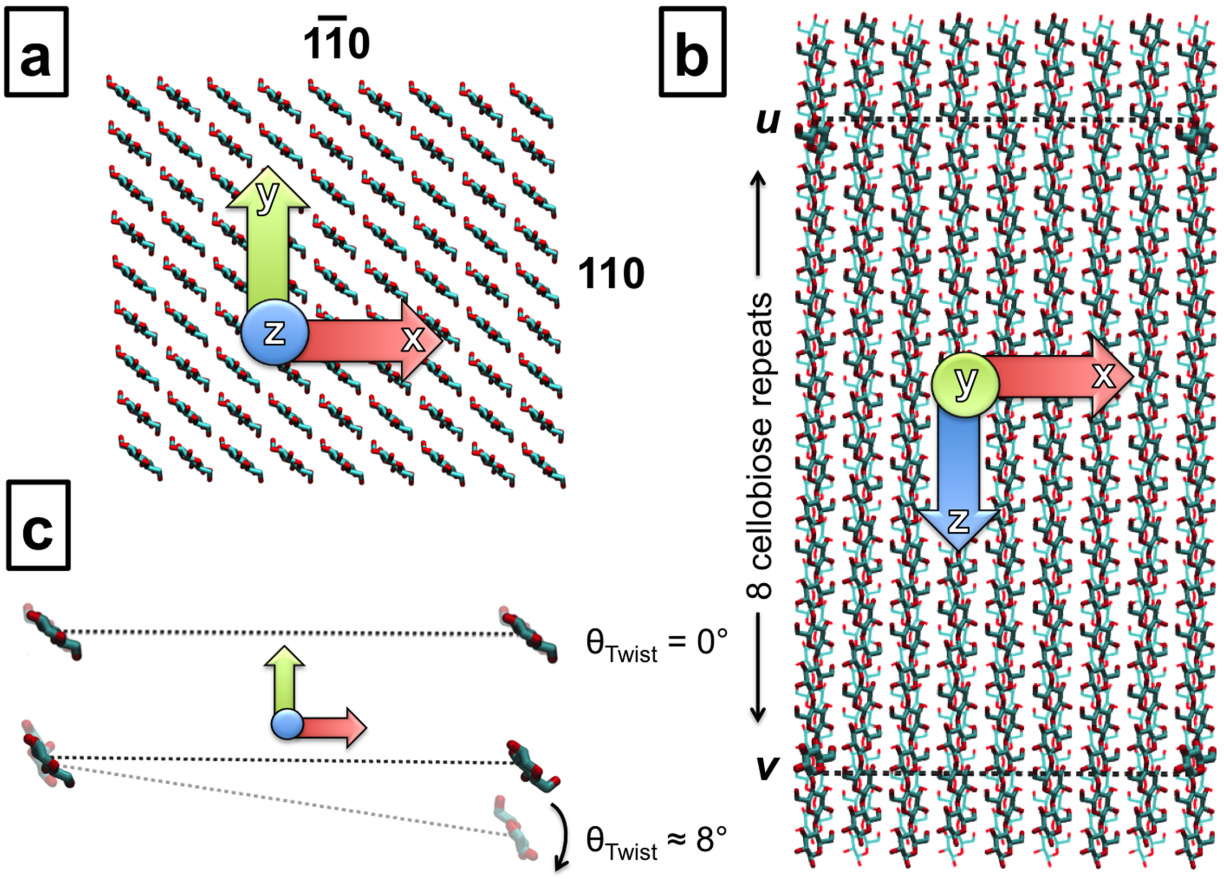


Figure 6.1: (a) Model microfibrils are oriented along the z-axis with the exposed crystallographic faces corresponding to the $1\bar{1}0$ and 110 planes on the y- and x-axes, respectively. (b) Vectors v and u are designated across the $1\bar{1}0$ plane, perpendicular to the z-axis. (c) Prior to twisting, vectors v and u are parallel and the angle between them, θ_{Twist}, is zero. Upon twisting, the two vectors diverge, resulting in a θ_{Twist} of $\sim 8^{\circ}$ per cellobiose unit for the 81-chain, DP 20 model.

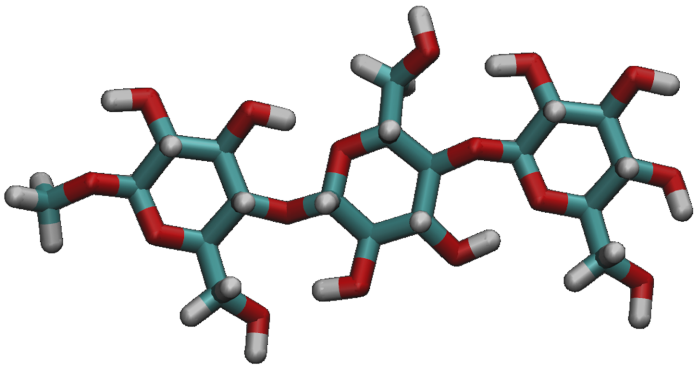


Figure 6.2: A trisaccharide fragment of cellulose Iβ was employed for derivation of charges to account for internal polarization of the cellulose polymer. This is referred to as the chain-polarized charge model.

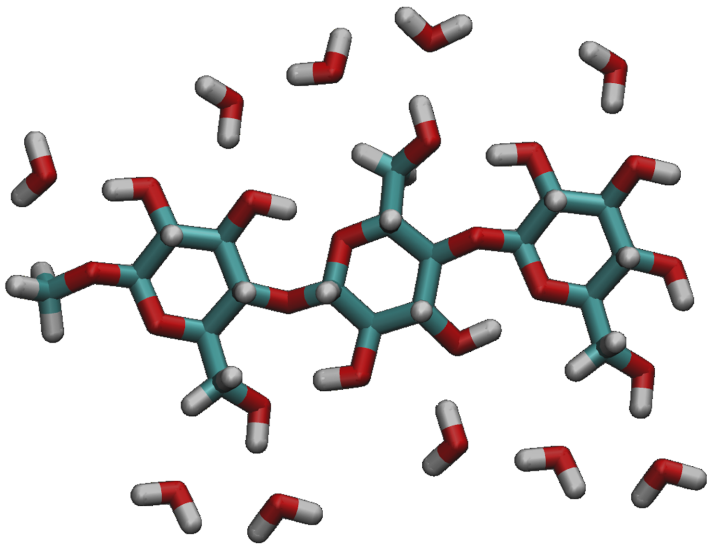


Figure 6.3: The chain-polarized charge model was augmented with 10 water molecules to represent contacting hydroxyl groups from neighboring polysaccharide chains according to the crystallographic coordinates. This accounts for polarization arising from both the polymeric and crystalline aspects of cellulose structure, and is referred to as the crystal-polarized charge model.

taken from the nonreducing glucosyl unit of the trisaccharide model, while charges for the repeating internal residue (4GB) were taken from the center glucosyl unit.

MD Simulations

MD simulations were performed with the GPU implementation of pmemd, pmemd.cuda_SPDP [91, 92], from AMBER12 [28], using the GLYCAM06 [11] (version h) force field for carbohydrates and the TIP3P [21] water model, unless otherwise indicated.

Model microfibrils were solvated with a 1.2 nm water buffer, which was subjected to energy minimization (12,500 steps steepest descent, 12,500 steps conjugate gradient). Full systems were then subjected to further energy minimization (12,500 steps steepest descent, 12,500 steps conjugate gradient), followed by heating from 0–300 K over 25 ps. Production simulations were performed at constant pressure (NPT) with a pressure relaxation time of 1 ps. A Berendsen-type thermostat with a time coupling constant of 1 ps was invoked for temperature regulation. All covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm [9], allowing a simulation time step of 2 fs. Scaling factors for 1-4 nonbonded interactions were set to unity [93], and a nonbonded interaction cutoff of 0.8 nm was employed. Long-range electrostatics were computed with the particle mesh Ewald (PME) method [18]. Systems were equilibrated for 1 ns prior to data collection, with the exception of an extended-length (DP 106) model, which was equilibrated for 2 ns. The timescale for production simulations was 10 ns unless otherwise noted.

Quantifying Microfibril Twist

To quantitatively assess twisting behavior, a metric to characterize the angle of twist (θ_{Twist}) along the microfibril axis was defined. Two vectors were designated on the $1\bar{1}0$ face of the microfibril, perpendicular to the axis, between the C1 (v) and O4 (u) atoms of the $n-2$ glucosyl residues of the outermost cellulose chains of the face (Figure 6.1b). Antepenultimate ($n-2$) residues were chosen to avoid artifacts from any disorder in the terminal residues that

might arise during simulation. These vectors are parallel in the crystallographic structure ($\theta_{Twist} = 0°$), but diverge by θ_{Twist} as the microfibril twists (Figure 6.1c). The value of θ_{Twist} is readily calculated from the dot product of the two vectors, v and u (Equation 6.1).

$$\theta_{Twist} = \frac{180}{\pi} cos^{-1} \frac{v \cdot u}{|v||u|} \tag{6.1}$$

To allow for comparisons between microfibrils of varying dimension, all calculated values of θ_{Twist} were normalized by the number of cellobiose repeats encompassed by the vectors. Normalization to cellobiose repeats instead of DP facilitated comparison to values reported per cellobiose unit in previous studies [76]. Values of θ_{Twist} were time-averaged over simulation trajectories to give $< \theta_{Twist} >$. Error estimates reported for values of $< \theta_{Twist} >$ represent standard deviation of the mean (SDM) and were calculated by averaging results from the two statistical inefficiency methods detailed by Foley et al. [94].

Results and Discussion

Model Microfibril

Previous simulation studies of finite microfibril models report bulk twisting along the axis in which outer chains circumscribe inner chains in a concentric fashion [76–84], suggesting a direct relationship between cross-sectional thickness and the magnitude of $< \theta_{Twist} >$ observed. Larger models composed of greater numbers of chains should be expected to exhibit smaller values of $< \theta_{Twist} >$, and simulation results for a series of microfibrils ranging from 9 to 289 constituent chains follow this trend (Figure 6.4). These values represent the characteristic magnitudes of $< \theta_{Twist} >$ observed for microfibrils of given thicknesses according to the GLYCAM06 force field under typical simulation conditions.

The largest model studied here (289 chains) measures ∼11.1 nm across the $1\bar{1}0$ face, ∼9.6 nm across the 110 face, and displays a $< \theta_{Twist} >$ of 0.35° per cellobiose (Figure 6.4). According to TEM, AFM, and TM-AFM studies by Hanley et al. [87], a smooth, uniform

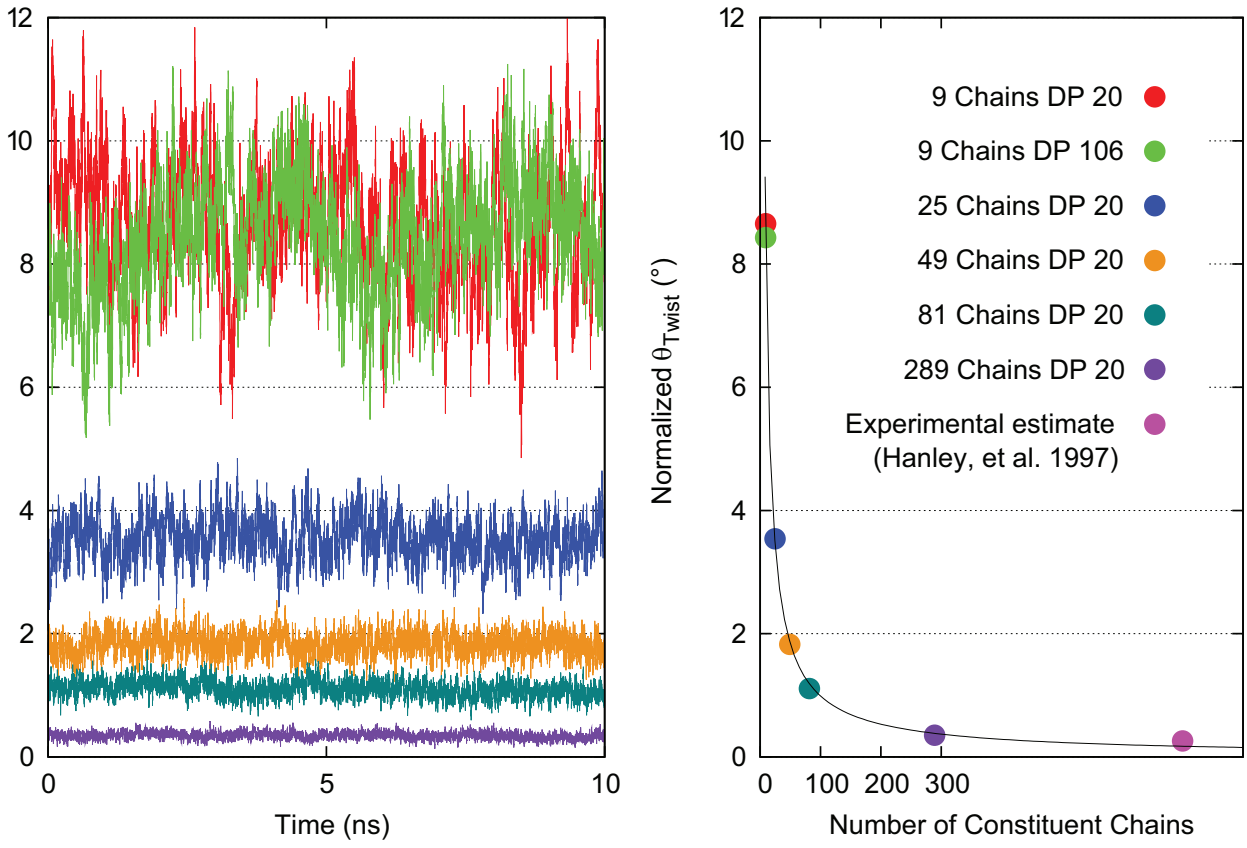


Figure 6.4: Calculated values of θ_{Twist} plotted over time and averaged over 10 ns simulation trajectories for models ranging from 9 to 289 constituent chains. Values are per cellobiose unit.

twisting could be extrapolated to have a 1400 nm repeat length. This corresponds to 360° of twist per 2800 glucosyl units, or 0.26° of twist per cellobiose unit for microfibrils in the 20–50 nm thickness regime, implying the degree of twisting predicted by simulation may not be unreasonable. Furthermore, recent work by Nishiyama et al. [95] on theoretical fiber diffraction patterns calculated from model crystals suggests that a subtle degree of twist, such as that observed by Hanley et al., would not interfere with the development of a diffraction pattern and is not necessarily inconsistent with the crystallographic data.

Although microfibril thickness is known to depend on cellulosic source, experimental evidence to support exact sizes for microfibrils of a given origin remains controversial. An elementary fibril of 36 chains has become widely accepted as the representative structure for plant-based cellulose; however, Nishiyama et al. [95] have recently presented results suggesting a structure containing 64–100 chains (8–10 per face) for cotton cellulose. As the present work ultimately seeks to lay a foundation for understanding the interactions of water molecules with cotton fiber, microfibrils corresponding to this cross-sectional thickness regime were of primary interest. A model of 81 chains (9 per face) was therefore selected as the representative microfibril structure for further use in this study.

While microfibrils can range naturally to thousands of residues in length with intermittent amorphous regions, previous simulation work has explored lengths corresponding to only DP 10–40 [76–84]. Evaluation of a 9-chain, DP 106 model showed the magnitude of $< \theta_{Twist} >$ to be highly comparable to the value obtained for a 9-chain, DP 20 model (Figure 6.4), suggesting the characteristic magnitude of $< \theta_{Twist} >$ observed for a microfibril of given thickness is independent of microfibril length. Since no notable length dependence was observed for values of $< \theta_{Twist} >$, the representative structure of 81 chains was assigned DP 20.

As the focus of this study was the initial tendency of microfibrils to adopt a twisted conformation, preliminary simulations explored timescales of only 10 ns to collect statistics on the twisted structure after transition from the crystallographic coordinates. Once dimen-

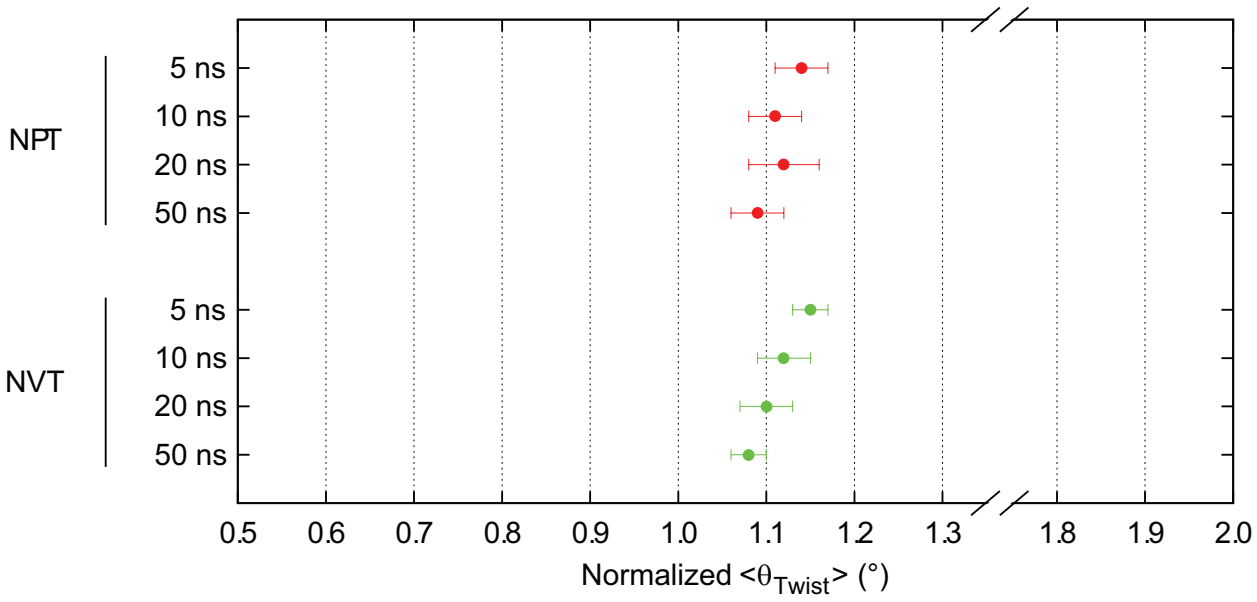


Figure 6.5: Values of $< \theta_{Twist} >$ collected for the representative microfibril of 81 chains, DP 20 over timescales of up to 50 ns with NPT and NVT ensembles. Values are per cellobiose unit. Error bars are SDM. Root-mean-squared fluctuations were within 13°–14°.

sions for the representative structure were established, simulations were extended to the 50 ns timescale under both constant pressure (NPT) and constant volume (NVT) conditions (Figure 6.5). While the use of different thermodynamic ensembles produced no significant variation in twisting behavior, the magnitude of $< \theta_{Twist} >$ was observed to trend downward when averaged over progressively longer timescale increments. This cannot necessarily be interpreted as suggesting convergence of the twisted structure over time, as similar systems have been shown to gradually converge to a structure that no longer corresponds to cellulose Iβ [81]. Nevertheless, beyond 5 ns of simulation, the rate of decrease of $< \theta_{Twist} >$ was relatively slow. Given this, and in order to avoid the possibility that structures would begin to deviate markedly from the cellulose Iβ form, a 10 ns timescale was deemed acceptable for drawing relative comparisons between models. A summary of data collected in this study for the representative 81-chain, DP 20 microfibril averaged over 10 ns simulation trajectories is presented in Figure 6.6.

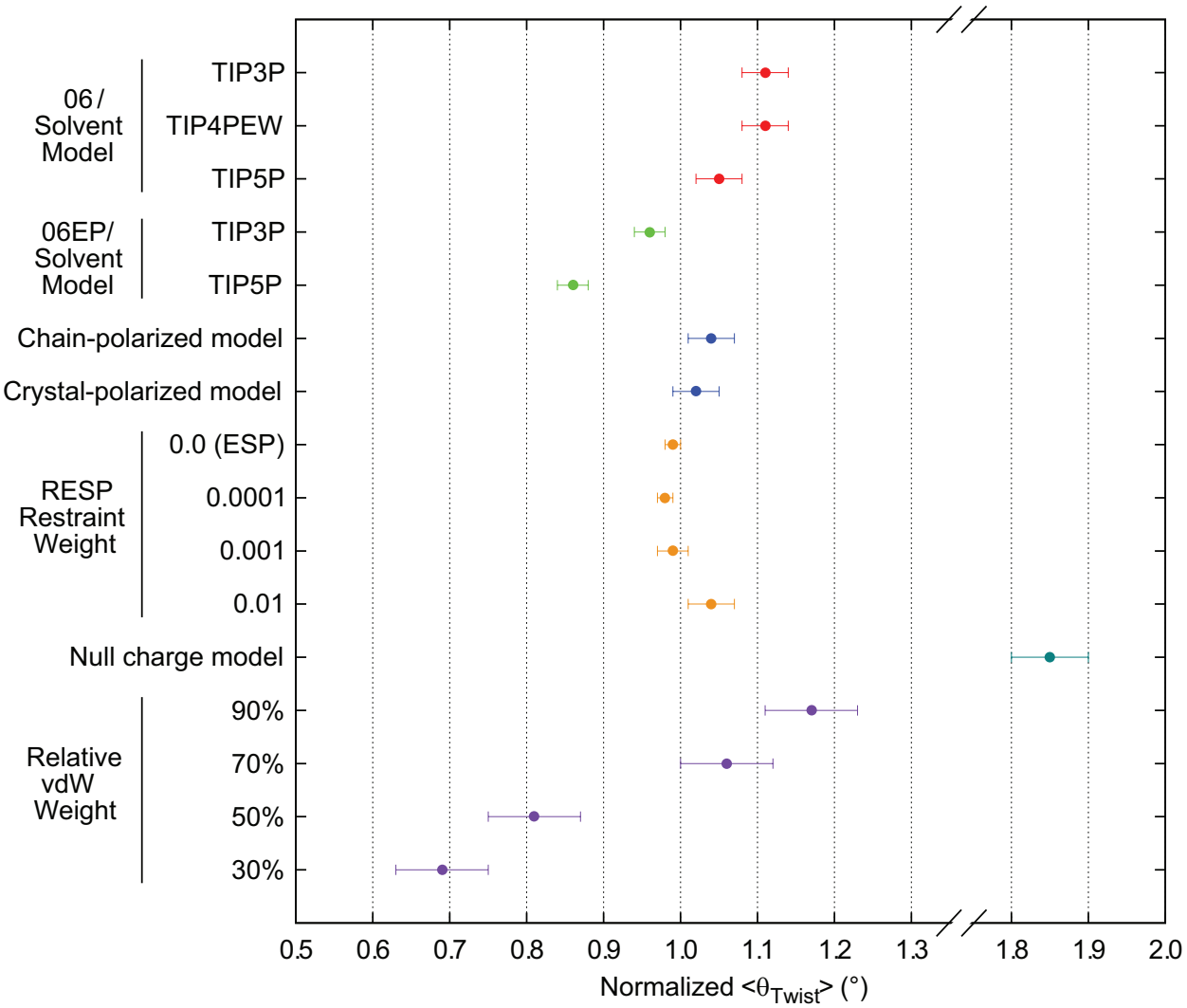


Figure 6.6: Summary of $< \theta_{Twist} >$ data collected in this study, plotted as a function of force field, solvent model, and varied nonbonded force field parameters. Values are per cellobiose unit, averaged over 10 ns simulation trajectories. Error bars are SDM.

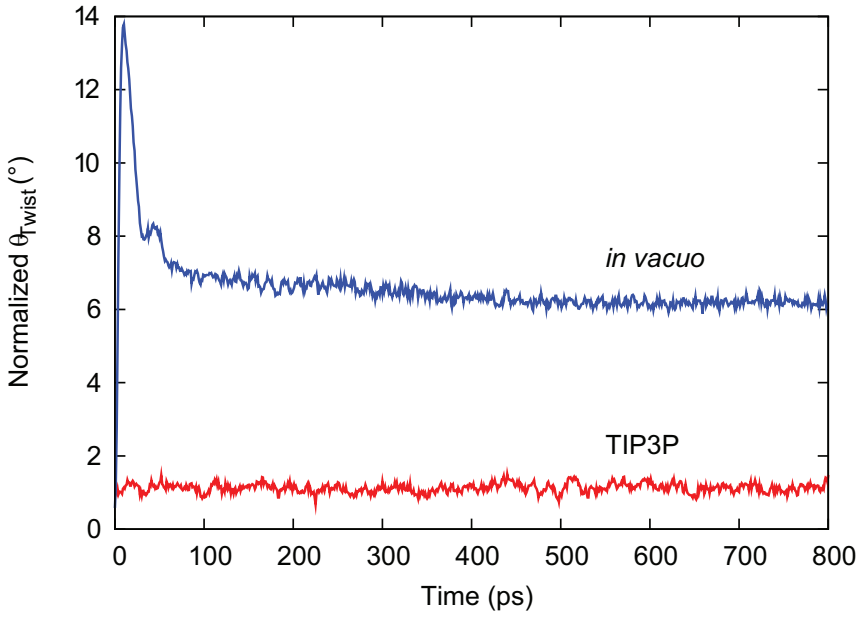


Figure 6.7: Calculated values of θ_{Twist} plotted over time for a short (800 ps) in vacuo simulation, and a simulation employing TIP3P explicit water. Values are per cellobiose unit.

Solvent Effects

While the crystallographic coordinates for cellulose represent the ordered interior of a large, solid phase structure, cellulose microfibrils based on these coordinates constitute comparatively small, isolated crystalline assembles. Furthermore, simulation studies of microfibrils generally aim to understand the dynamic behavior and intermolecular interactions of these assemblies in the context of an aqueous environment, necessitating the addition of solvent. A short (800 ps) in vacuo simulation displayed extreme deformation, indicating that the physical presence of water plays a critical role in mitigating the extent of twisting that would otherwise occur (Figure 6.7). Although the representative 81-chain, DP 20 model displayed normalized $< \theta_{Twist} >$ values of $\sim 1^{\circ}$ when solvated with TIP3P (Figures 6.4–6.6), the value from the in vacuo simulation rose to nearly 14° before equilibrating to $\sim 6^{\circ}$. Implicit solvent simulations employing a series of generalized Born models as implemented in AMBER12

(igb = 1 [96–98], 2 [99, 100], 5 [100], 8 [101, 102]) all produced microfibrils that fragmented or peeled into constituent cellulose chains, further demonstrating the importance of explicit solvent in this system.

Solvent Model

Having established the critical role of water in microfibril simulations, two additional commonly employed explicit solvent models were evaluated. Employing TIP4P-EW [22] produced results that were remarkably comparable to that of TIP3P (Figure 6.6), indicating that improved modeling of bulk water properties has no effect on microfibril behavior. Alternatively, employing TIP5P [23], which includes lone pairs on oxygen atoms to better reproduce both bulk water properties and the tetrahedral geometry of hydrogen bonds, resulted in a noticeable reduction of $< \theta_{Twist} >$ (Figure 6.6). This implicates solvent hydrogen bonding as a determining factor in twisting behavior.

Analysis of the radial distribution function (RDF) for water oxygen atoms around a representative solvent-exposed hydroxymethyl hydrogen (H6O) on the $1\bar{1}0$ face of the microfibril (Figure 6.8) indicates that TIP5P binds more tightly, and displays a significantly higher occupancy in the first solvation shell as compared to TIP3P. Decomposition of the molecular mechanical (MM) contributions to interaction energy between microfibril and solvent also shows that TIP5P is preferred over TIP3P by more than 2000 kcal/mol due to more favorable electrostatic interactions with the microfibril surface (Figure 6.9). Previous computational studies have demonstrated that TIP5P enhances specific solvation and results in more highly conserved and optimally coordinated water interactions in biomolecules, which can impact the dynamics and conformational preferences of flexible systems [25, 26]. While the mechanism by which tighter interaction with solvent leads to a decrease in $< \theta_{Twist} >$ is unclear, significant solvent structuring is known to occur around cellulose surfaces [76, 103, 104], and it may be that increased order in this structure serves to restrict twisting motion. This is

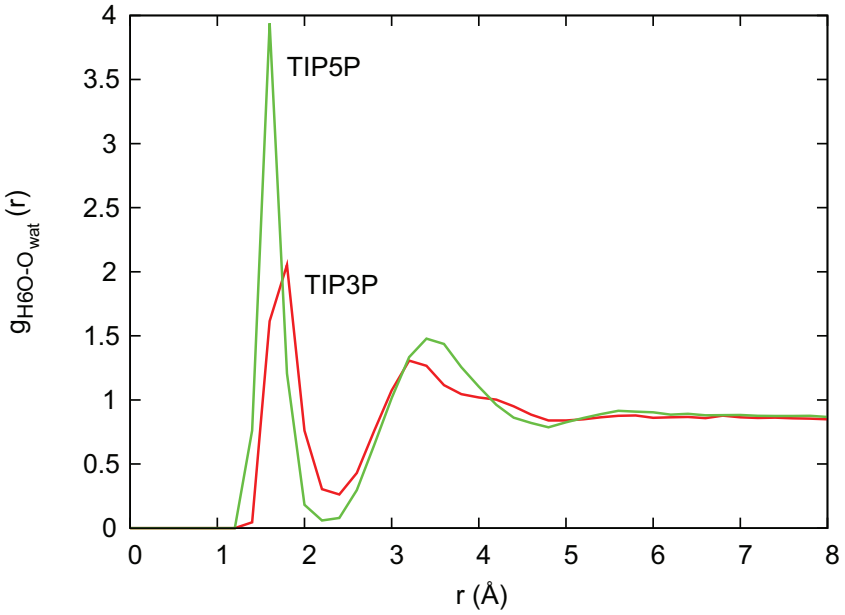


Figure 6.8: Radial distribution functions for TIP3P (red) and TIP5P (green) water oxygen atoms around a representative solvent-exposed hydroxymethyl hydrogen atom (H6O) on the $1\bar{1}0$ microfibril surface.

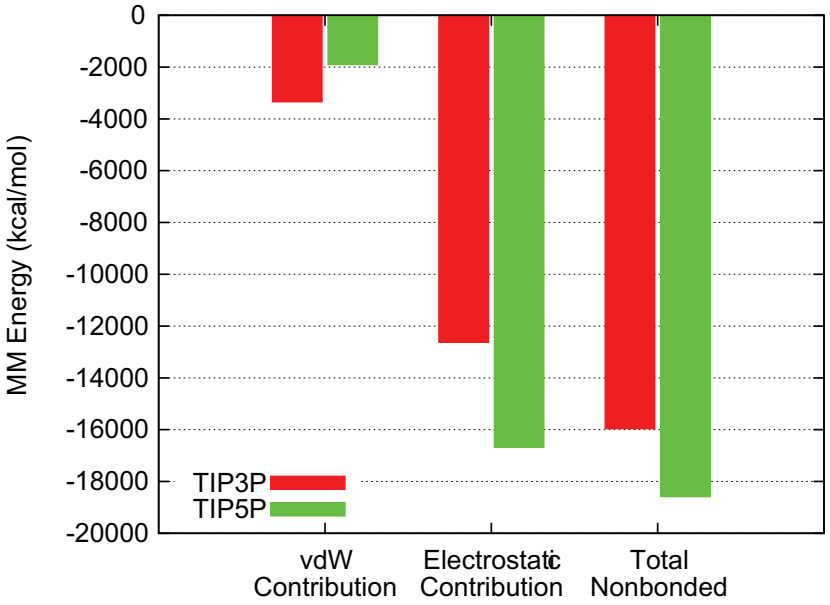


Figure 6.9: Molecular mechanics interaction energies between microfibril and water, decomposed into vdW and electrostatic contributions for TIP3P and TIP5P solvent models.

consistent with the observation that the physical presence of water restricts twisting relative to vacuum conditions.

Solute Lone Pairs

As the use of an explicit solvent model employing lone pairs on oxygen atoms produced a noticeable effect on the magnitude of $< \theta_{Twist} >$, additional simulations were performed with the GLYCAM06EP [105] carbohydrate force field, which includes lone pairs on carbohydrate oxygen atoms. While the magnitude of $< \theta_{Twist} >$ was markedly reduced with GLYCAM06EP relative to GLYCAM06 in TIP3P solvent, this value was substantially reduced when GLYCAM06EP was combined with TIP5P solvent (Figure 6.6). Comparing the standard model (GLYCAM06 with TIP3P) to the full lone pair model (GLYCAM06EP with TIP5P), the overall values for $< \theta_{Twist} >$ differ by 0.25° per cellobiose. As with TIP5P, the addition of lone pairs in GLYCAM06EP improves hydrogen bond directionality and has been shown to better preserve unit cell dimensions when simulating crystalline carbohydrate assemblies [105]. Furthermore, use of lone pairs on the solute also appears to enhance the surface solvent effects observed with TIP5P, resulting in increased order in the surrounding solvent structure, which apparently serves to mitigate twisting. These data underscore the role of hydrogen bonds as critical determinants of cellulose structure, both in terms of the internal network, as well as at the microfibril surface. In particular, simulations with GLYCAM06EP suggest that the internal hydrogen bond network resists the tendency to twist and that this effect is sensitive to the manner in which these charge interactions are modeled. Tighter error bars for the calculated values of $< \theta_{Twist} >$ (Figure 6.6) indicate that refinement of the internal hydrogen bond network enhances overall structural stability.

Charge Model

Given the importance of internal hydrogen bonds to cellulose microfibril structure and the sensitivity of modeled hydrogen bonds to charge parameterization protocols, the ef-

fect of charge model was evaluated. To reduce computational expense, most biomolecular force fields do not account for charge polarization, but instead employ invariant partial atomic charges whose molecular distributions are dependent on the conformation of the model used for their development. While the charges in GLYCAM06 were derived based on isolated monosaccharides in solution to facilitate modularity and broad applicability, such fixed charges are unable to adjust to changes in local environment, including assembly of monomeric units into polymers or a crystalline lattice, as found in cellulose.

An alternative charge set designed to account for changes in charge distribution induced by the polymeric nature of cellulose structure was developed based on a trisaccharide fragment of cellulose Iβ (Figure 6.2). Simulations employing these chain-polarized charges resulted in a reduction in the magnitude of $< \theta_{Twist} >$, with error bars comparable to that of standard GLYCAM06 in TIP3P under equivalent simulation conditions (Figure 6.6). This refined charge distribution serves to polarize and thus strengthen the hydrogen bond network that extends down the length of each cellulose chain. Two key hydrogen bonds are those that span each of the glycosidic linkages (O3–O5' and O6–O2'), and variations in their relative strength may be expected to directly impact the torsional properties of these linkages.

A second alternative charge set designed to account for changes in charge distribution induced by both the polymeric and crystalline aspects of cellulose structure was also developed. This was accomplished by augmenting the chain-polarized charge model with 10 water molecules representing contacting hydroxyl groups of neighboring polysaccharide chains according to the crystallographic coordinates (Figure 6.3). Simulations applying these crystal-polarized charges displayed only a slight reduction in the value of $< \theta_{Twist} >$ beyond that already imparted by the chain-polarized charge model (Figure 6.6). This result indicates that, while interchain hydrogen bonds are clearly associated with organization of cellulose chains into layers, electrostatic polarization from such interactions contributes only a modest stabilizing force. In contrast, polarization of the intrachain hydrogen bonds plays a signifi-

cant role in resisting the tendency to twist by constraining the individual torsional properties of the glycosidic linkages.

Charge Restraint Weight

An additional factor of charge development protocol that can influence the strength of modeled hydrogen bonds is the choice of restraint weight (k_{Rstr}). In GLYCAM06 and other AMBER-family force fields, the 6-31G* basis set is generally employed for the QM calculation of MEPs for charge derivations, as it suitably reproduces biomolecular properties for use in condensed phase simulations. However, the ESP charges produced with this basis set tend to overstate bond polarity, such that a hyperbolic restraint function is commonly applied during fitting in order to compensate (Equations 6.2 and 6.3) [16, 17].

$$\chi^2_{RESP} = \chi^2_{ESP} + \chi^2_{Rstr} \tag{6.2}$$

where

$$\chi^2_{Rstr} = k_{Rstr} \sum_j [(q_j^2 + b^2)^{1/2} - b)] \tag{6.3}$$

The standard GLYCAM06 charge development protocol uses a restraint weight of 0.01 for calculation of these RESP charges [17, 106]. To probe the sensitivity of the cellulose internal hydrogen bond network to attenuation of bond polarity, as determined by the choice of this value, a series of RESP charges were developed based on the chain-polarized trisaccharide model described above (Figure 6.2), employing restraint weights ranging from 0–0.01. Simulations with these charge sets all resulted in a reduction of $< \theta_{Twist} >$ (Figure 6.6). The majority of this effect arises from use of the chain-polarized charge model, which strengthens the intrachain hydrogen bonds. Use of restraint weights less than 0.01 led to a further reduction of $< \theta_{Twist} >$ that may be directly attributed to enhanced bond polarity imparting additional strength to this network. The error ranges for values of $< \theta_{Twist} >$ decrease

with increasing bond polarity (Figure 6.6), indicating greater overall microfibril stability. As noted in previous sections, the cellulose internal hydrogen bond network, particularly the intrachain network, is a critical determinant of the extent of twisting, and accurate modeling of microfibril behavior will likely depend on the force fields ability to capture the characteristics of charge interactions in the context of this crystalline lattice.

Internal Nonbonded Interactions

To probe the overall role of electrostatics in microfibril behavior, a simulation was performed in which all atoms in the microfibril were assigned a charge of zero to create a null charge model. In the absence of all internal electrostatic interactions, the magnitude of $< \theta_{Twist} >$ was considerably enhanced (Figure 6.6), indicating that twisting behavior is not fundamentally driven by electrostatics. Although this increase might stem partially from a lack of electrostatic repulsion between layers, it is likely also related to the absence of the internal hydrogen bond network, particularly the intrachain network, which serves as an essential stabilizing framework that resists the tendency to twist.

The observation that significant twisting occurs in the absence of any solute electrostatics implicates van der Waals (vdW) interactions as a key contributing factor. Notably, recent ab initio QM studies of cellulose structure suggest that dispersion interactions are largely responsible for the stability of stacked layers in the crystalline assembly [107].

Classical force fields, such as those from the AMBER family, often employ a 12-6 Lennard-Jones potential to model vdW interactions between atoms (Equation 6.4).

$$V_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{6.4}$$

where ϵ defines the well depth, or strength of the pairwise association.

To investigate the overall function of vdW interactions with regard to twisting behavior, a series of simulations were performed in which ϵ was scaled to percentages ranging from

90–30 % of the default. While this alteration induced some structural instability, as was to be expected, the results nevertheless showed the magnitude of $< \theta_{Twist} >$ decreasing dramatically as ϵ was reduced (Figure 6.6). That is, as the vdW interactions within the microfibril were artificially diminished, the structure became less twisted. Combined with the results from the null charge model simulation, this analysis suggests that attractive vdW interactions within the microfibril contribute a driving force responsible for twisting behavior, likely seeking to maximize crystal packing efficiency in the twisted structure. In contrast, electrostatic interactions provide a balancing resistance to twisting.

Conclusions

Previous simulation studies of cellulose Iβ microfibrils have demonstrated the tendency for these structures to diverge from the linearly oriented crystallographic coordinates and adopt a twisted conformation [76–84]. The present work sought to understand the driving forces behind this behavior, as well as to determine how computational methodology might play into it, through evaluation of model microfibril dimensions, the role of solvent and effect of solvent model, the effect of charge set and application of explicitly modeled oxygen lone pairs, and the overall role of nonbonded interactions. The results indicate that a balance of competing forces ultimately determines the extent of microfibril twisting observed in a given simulation. While twisting appears to be driven by attractive vdW (dispersion) interactions seeking to maximize crystal packing efficiency between layers, it is mitigated by both the intrachain hydrogen bond network, which influences the torsional properties of the glycosidic linkages, and solvent effects at the microfibril surface. As the strength and geometry of modeled hydrogen bonds are sensitive to charge development protocols and to the relative positioning of oxygen partial charges, the most accurate atomistic modeling of cellulose structure likely necessitates use of prepolarized charge distributions that describe β-D-glucose in the context of the cellulose crystalline lattice. Further, to optimally reproduce the internal hydrogen bond network, unit cell dimensions, and the resulting magnitude of

microfibril twist, improved electrostatic directionality, such as imparted by the oxygen lone pairs employed in GLYCAM06EP, is required. The physical presence of solvent also serves to significantly mitigate the extent of twisting, and this effect is enhanced by the ability of the explicit solvent model to induce well-ordered, tightly coordinated water structure at the cellulose surface.

The magnitude of the microfibril twist resulting from the balance of these inter- and intramolecular forces depends on the cross-sectional thickness of the model employed for simulation, with models composed of greater numbers of constituent chains exhibiting increasingly subtle twisting behavior. Comparison of $< \theta_{Twist} >$ values collected in this study with experimental estimates based on data from Hanley et al. [87] suggest that the degree of twist predicted by MD simulation is not unreasonable and that current computational methodology is thus adequate to provide a suitable representation of microfibril behavior in the presence of aqueous solution, at least on relatively short simulation timescales. Further work is needed to assess the affect of system parameterization on long timescale behavior, given the report by Matthews et al. [81] describing eventual transition of the twisted cellulose Iβ structure to the I-HT form. As this transition results from a gradual disruption of the internal hydrogen bond network due to widespread reorientation of hydroxymethyl groups, improved modeling of partial atomic charge distributions and hydrogen bond geometries in both solute and solvent, as discussed in this study, may be required to enhance microfibril stability and preserve the cellulose Iβ structure over extended simulation timescales.

CHAPTER 7:

EFFECT OF MICROFIBRIL TWISTING ON THEORETICAL POWDER DIFFRACTION PATTERNS OF CELLULOSE Iβ[1]

[1]Hadden, J. A.; French, A. D.; Woods, R. J. *Cellulose* **2014**, *21*, 879-884.
Reprinted here with permission of publisher.

Abstract

Previous studies of calculated diffraction patterns for cellulose crystallites suggest that distortions that arise once models have been subjected to molecular dynamics (MD) simulation are the result of both microfibril twisting and changes in unit cell dimensions induced by the empirical force field; to date, it has not been possible to separate the individual contributions of these effects. To provide a better understanding of how twisting manifests in diffraction data, the present study demonstrates a method for generating twisted and linear cellulose structures that can be compared without the bias of dimensional changes, allowing assessment of the impact of twisting alone. Analysis of unit cell dimensions, microfibril volume, hydrogen bond patterns, glycosidic torsion angles, and hydroxymethyl group orientations confirmed that the twisted and linear structures collected with this method were internally consistent, and theoretical powder diffraction patterns for the two were shown to be effectively indistinguishable. These results indicate that differences between calculated patterns for the crystal coordinates and twisted structures from MD simulation can result entirely from changes in unit cell dimensions, and not from microfibril twisting. Although powder diffraction patterns for models in the 81-chain size regime were shown to be unaffected by twisting, suggesting that a modest degree of twist is not inconsistent with available crystallographic data, it may be that other diffraction techniques are capable of detecting this structural difference. Until such time as definitive experimental evidence comes to light, the results of this study suggest that both twisted and linear microfibrils may represent an appropriate model for cellulose Iβ.

Keywords: Cellulose, Microfibril twist, Molecular dynamics, X-ray diffraction

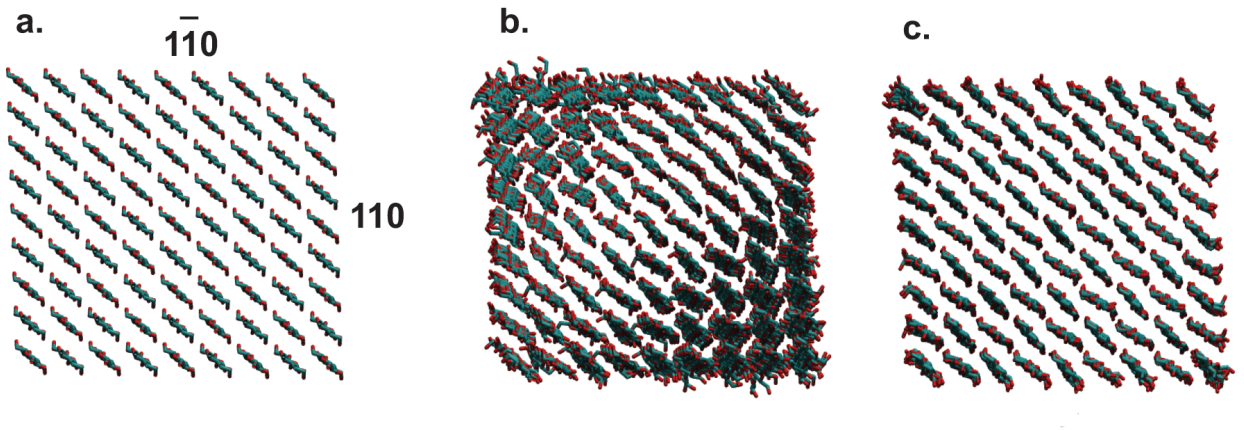


Figure 7.1: View down the microfibril axis for (a) the crystallographic structure of cellulose Iβ, (b) a finite (twisted) structure, and (c) an infinite (linearly constrained) structure with chain termini bonded across the simulation box periodic boundary.

Introduction

Just over a decade ago, Nishiyama et al. [75] combined X-ray and neutron diffraction data to develop a high-resolution crystallographic structure of cellulose Iβ, describing it in terms of both heavy and hydrogen atom positions. This structure, based on samples of tunicin, is characterized by an ordered array of perfectly parallel glucosyl chains associated into layers by a well-defined hydrogen bond network (Figure 7.1a). In recent years, molecular dynamics (MD) simulation studies of finite or mini-crystal microfibril models based on these coordinates have demonstrated a preference to distort from the linearly oriented crystal structure and adopt a twisted conformation (Figure 7.1b) [27, 76, 77, 79–83]. While some experimental data supports the existence of twisted microfibrils [87, 108], the apparent contradiction with high-resolution crystallographic data has been the source of considerable controversy in the cellulose structure community.

More recently, Nishiyama et al. [95] published work comparing theoretical diffraction patterns for the crystallographic coordinates with those calculated from twisted structures produced by MD simulation. They observed significant changes in peak positions and in-

tensities and attributed these distortions both to microfibril twisting and to changes in unit cell dimensions induced by the empirical force field. To separate the individual contributions of these two effects and provide a better understanding of how twisting in crystalline cellulose manifests in diffraction data, the present study demonstrates a method for generating twisted and linear cellulose structures that can be compared without the bias of dimensional changes. While the twisted structure represents a typical finitely modeled microfibril (Figure 7.1b), the linear structure is constrained from twisting by bonding chain termini across the simulation box periodic boundary, making it essentially infinite (Figure 7.1c). As the infinite model is nevertheless permitted to adapt to dimensional changes induced by the force field, it represents a linearly oriented form of cellulose Iβ, similar to the crystallographic structure, with unit cell dimensions that are comparable to those of the twisted form. Comparison of powder diffraction patterns calculated for these two structures thus allows an unbiased assessment of the impact of twisting alone.

Computational Methods

Microfibril starting structures for MD simulations were prepared with Mercury 2.0 [88], based on the cellulose Iβ crystallographic coordinates reported by Nishiyama et al. [75]. Models consisted of 81 total chains (9 per face), and were constructed such that the $1\bar{1}0$ and 110 planes formed the exposed surfaces (Figure 7.1a). Finite microfibrils were generated with 20 glucosyl units per chain [degree of polymerization (DP) 20], while infinite microfibrils were generated with DP 18, with terminal residues bonded across the simulation box periodic boundary. Models were surrounded with a 12 Å solvent buffer.

Microfibrils were parameterized with the GLYCAM06 [11] (version h) force field for carbohydrates, and solvent was modeled as TIP3P water [21]. Simulation files were generated using the tleap module of AmberTools12 [28], and subsequently converted to GROMACS format with the glycam2gmx.pl script from Wehle et al. [109]. The double precision parallel

implementation of the mdrun module from GROMACS 4.5.5 [110–112] was employed for all simulation work.

An energy minimization protocol consisting of 12,500 cycles of steepest descent, followed by 12,500 cycles of conjugate gradient, was applied to the solvent and subsequently to the entire system prior to the start of dynamics. MD simulations were performed under isothermal, isobaric (NPT) conditions utilizing Berendsen thermo- and barostats (time constants of 1 ps) to maintain a reference temperature of 300 K and pressure of 1 bar. A nonbonded cutoff of 8 Å was applied, beyond which van der Waals interactions were truncated and long-range electrostatics were handled with particle mesh Ewald [18]. Covalent bonds to hydrogen were constrained using the LINCS algorithm [10] to allow a simulation time step of 2 fs. Systems were equilibrated for 1 ns, followed by 10 ns of production dynamics. The periodic-molecules option was set for the infinite microfibril model to allow coupling of chain termini across the periodic boundary.

GROMACS trajectories were converted to AMBER format for analysis using VMD [113]. Water molecules were removed, and microfibril models were cropped to contain only the center 16 glucose repeats (DP 16). This ensured comparison between models of equivalent dimension, as well eliminated noise resulting from terminal fraying during simulation of the finite model. Unit cell dimensions, hydrogen bond percentages, glycosidic torsion angles, and hydroxymethyl group orientations were determined using the ptraj module of AmberTools12 [28]. Average microfibril volume was calculated with Mol_Volume [114], based on 100 evenly spaced frames extracted from simulation trajectories. Theoretical powder diffraction patterns were calculated with Debyer [115] for this ensemble of 100 trajectory frames, employing a value of $k = 1.5418$ Å to denote CuKa radiation. Reference patterns for the crystallographic coordinates were calculated with Mercury 2.0 [88], which assumes an infinite crystallite size and eliminates artifacts associated with low-angle scattering seen in patterns calculated with Debyer.

Twisting was quantified using the metric for calculating θ_{Twist} defined by Hadden et al. [27]. This method designates two vectors (v, u) across the microfibril $1\bar{1}0$ face, perpendicular to the axis. While the vectors are parallel in the linear starting structure, their dot product describes the angle by which they diverge as the microfibril twists (Equation 7.1).

$$\theta_{Twist} = \frac{180}{\pi} cos^{-1} \frac{v \cdot u}{|v||u|} \tag{7.1}$$

Values of θ_{Twist} were normalized by the number of cellobiose repeats between the vectors and averaged over time to give $< \theta_{Twist} >$.

Results and Discussion

MD simulations were performed on the 10 ns timescale for finite (twisted, Figure 7.1b) and infinite (linearly constrained, Figure 7.1c) cellulose microfibrils. The twisted structure displayed a $< \theta_{Twist} >$ value of 1.17° per cellobiose, with root-mean-squared fluctuation of 0.14° and standard deviation of the mean of 0.02°. Although the unit cell showed some deviation from the crystallographic coordinates in all three dimensions (-1.3 % in **a**, +0.5 % in **b**, and +4.0 % in **c**), as has previously been observed during simulations of cellulose microfibrils [76], there were no significant differences in dimensions between the twisted and linear structures. Values for microfibril volume showed an increase of approximately 5 % from the crystallographic coordinates, yet corresponded exactly for the two MD models. Additionally, the twisted and linear structures contained equivalent, as well as experimentally consistent, internal hydrogen bond patterns, glycosidic torsion angles, and hydroxymethyl group orientations. These data are summarized in Tables 7.1 and 7.2, and together indicate that with the exception of twisting, the twisted and linear structures produced with this method are essentially identical, allowing for straightforward, unbiased comparison of theoretical diffraction data.

Table 7.1: Values used for quantitative comparison of twisted and linear microfibril models.

Model	$< \theta_{Twist} >$ (°)	Unit cell dimensions (Å)						Volume ($\times 10^5$, Å^3)	200 2-θ (°)	PWHM (°)
		a	RMSF[a]	b	RMSF	c	RMSF			
Twisted	1.17	7.69	0.18	8.24	0.15	10.80	0.10	1.97	23.24	1.75
Linear	–	7.68	0.17	8.24	0.14	10.79	0.09	1.97	23.27	1.75
Iβ crystal	–	7.78	–	8.20	–	10.38	–	1.87	22.98	1.64

[a]Root-mean-squared fluctuation

Table 7.2: Values used for quantitative comparison of twisted and linear microfibril models.

Model	Hydrogen bonds (%)			ϕ (%)	RMSF[a]	ψ (%)	RMSF	Hydroxymethyl group (%)		
	H_2O–O_6	H_3O–O_5	H_6O–O_3					tg	gt	gg
Twisted	99.6	99.7	98.5	27.2	6.7	-24.3	6.8	98.4	1.3	0.3
Linear	99.4	99.5	98.4	26.8	6.7	-24.3	6.8	99.3	0.6	0.1
Iβ crystal	100	100	100	23.9	–	-27.4	–	100	0	0

[a]Root-mean-squared fluctuation

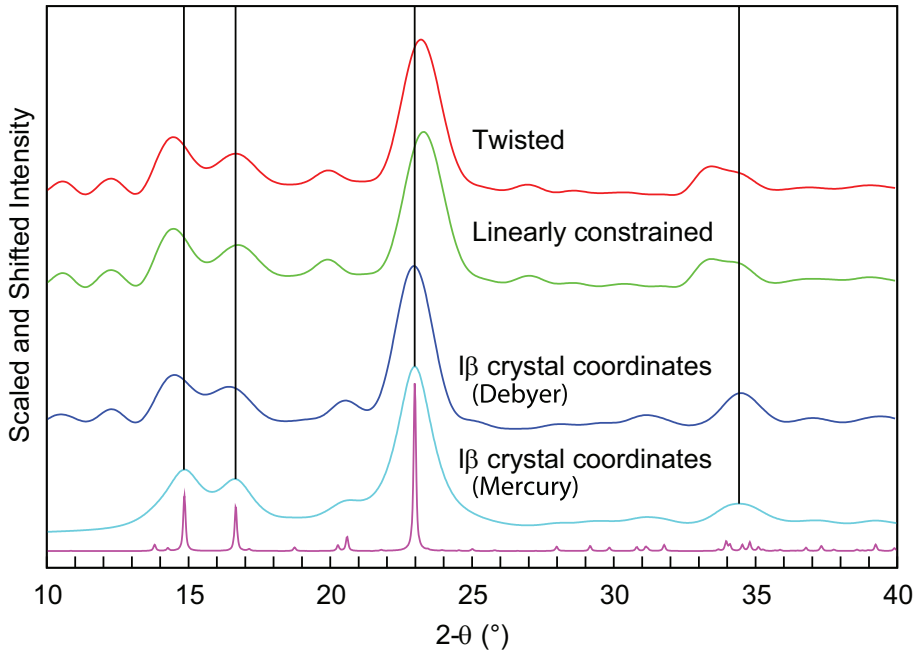


Figure 7.2: Theoretical powder diffraction patterns for twisted and linear microfibril models and three reference patterns corresponding to the original crystallographic coordinates for cellulose Iβ. The upper three patterns were calculated with Debyer as non-periodic 81-chain, DP 16 mini-crystals, while the lower two patterns were calculated with Mercury 2.0 with the assumption of infinite crystallite size (PWHM of 0.1° and 1.5°) to demonstrate the absence of artifacts associated with low-angle scattering present in the Debyer patterns. Peak heights were each normalized to 100.

Powder diffraction patterns calculated based on the final frames of simulation trajectories for both the twisted and linear microfibrils are presented in Figure 7.2, along with three reference patterns corresponding to the original crystallographic coordinates. The lower two reference patterns assume infinite crystallite size, and demonstrate the absence of artifacts associated with low-angle scattering present in the remaining patterns, which were calculated based on non-periodic 81-chain, DP 16 mini-crystals. This is especially obvious in the 10°–13° region of 2-θ. The upper reference pattern was calculated in an identical manner to that of the twisted and linear models and provides a direct comparison to assess distortions arising from MD simulation.

While the patterns for the twisted and linear structures indeed display distortions relative to that of the crystallographic structure, as previously discussed by Nishiyama et al. [95],

they are effectively indistinguishable from each other. The PWHM for the 200 peak (averaged over patterns for 100 simulation frames) was equivalent for the two MD models, but was broader (by less than 10 %) than that of the crystal coordinates, with the value for 2-θ shifted by less than 2 % (Table 7.1). Because the twisted and linear structures give rise to equivalent theoretical diffraction patterns, the distortions relative to the pattern for the crystallographic structure must arise from dimensional changes induced by the force field during simulation, as suggested by Nishiyama et al. [95], and not from microfibril twisting. This is expected to hold for all microfibrils, provided that the model is large enough that twisting does not excessively disorder the unit cell of the crystalline assembly.

As discussed by Nishiyama et al. [95], low-angle scattering can influence both peak position and intensity, in particular for the $1\bar{1}0$ reflection at 14.5° 2-θ. While some low-angle scattering can manifest in experimental patterns as a result of discontinuity between crystallites or inhomogeneity in the sample, the extent of it observed in the present theoretical patterns represents the worst-case scenario for non-periodic models of this size. Along with issues of preferred orientation of crystals in experimental patterns, low-angle scattering may present a greater concern for making comparisons between theory and experiment than modest twisting.

While this study has demonstrated that powder patterns are unaffected by a limited degree of microfibril twisting, it may be that other experimental techniques are capable of detecting this subtle structural difference. For example, Nishiyama et al. [95] presented theoretical fiber diffraction patterns that displayed wedge-shaped layer lines for twisted models that were not subjected to convolution to account for crystallite tilt distribution. These wedges were shown to have larger angles for models with fewer numbers of chains, which display a greater degree of twisting. It should be noted, however, that the models used in that study were relatively small compared to the tunicate samples (10–20 nm) used to solve the cellulose Iβ crystal structure. Hadden et al. [27] estimated that twisted microfibrils observed experimentally by Hanley et al. [87], which ranged in size from 20 to 50 nm, likely would

have had a uniform twist of only 0.26° per cellobiose when suspended in aqueous solution. This observation suggests that the degree of twisting present in large samples, such as those used for X-ray diffraction, is sufficiently subtle that it may be difficult to detect experimentally. Even with the use of comparatively small microfibril models, both the present work and that of Nishiyama et al. [95] imply that a modest degree of twisting is not necessarily inconsistent with crystallographic data.

Conclusions

The method presented here for generating internally consistent twisted and linear structures provides a previously unexploited strategy for probing the effect of microfibril twisting on experimental cellulose diffraction data. While this study has demonstrated that theoretical powder patterns for models in the 81-chain size regime are unaffected by twisting, suggesting that a modest degree of twist is not inconsistent with available crystallographic data, it may be that other diffraction techniques are capable of detecting this subtle structural difference. Further analysis is thus required to assess the effect of twisting on other back- calculated diffraction properties, as well as to confirm these observations experimentally. Until such time as definitive experimental evidence comes to light, the results of this study suggest that both twisted and linear microfibrils may represent an appropriate model for cellulose Iβ.

CHAPTER 8:
SYSTEM UNDER STUDY: INFLUENZA HEMAGGLUTININ

Influenza A

Influenza A is a negative-sense RNA virus of the family *Orthomyxoviridae* that causes acute respiratory infection in humans. Over the course of the last century, this pathogen has given rise to four distinct global pandemics. The first and most severe of these was the 1918 Spanish Flu (H1N1) outbreak, estimated to have killed up to 50 million people worldwide [116]. The subsequent 1957 Asian Flu (H2N2) and 1968 Hong Kong Flu (H3N2) outbreaks were significantly less virulent, resulting in around one million and 700,000 deaths, respectively [117]. While the fourth and most recent pandemic, that of the 2009 Swine Flu (novel H1N1) was relatively mild, it spread alarmingly fast, and estimates cite up to 5.7 million cases of infection within the first four months in the United States alone [118].

Pandemics occur when a strain of influenza previously unseen in humans, and toward which humans have no preexisting immunity, emerges and becomes easily transmissible between hosts. Once a pandemic has run its course and general immunity is established in the population, the strain is relegated to the pool of seasonal influenza strains, which cause the well-known annual epidemics seen during the autumn/winter months in temperate regions. These annual outbreaks result in an estimated 500,000 deaths worldwide each year [119].

Hemagglutinin and Neuraminidase

The influenza A viral envelope displays two major antigens, the glycoproteins hemagglutinin (HA) and neuraminidase (NA), which extend from the viral surface as spike-like structures

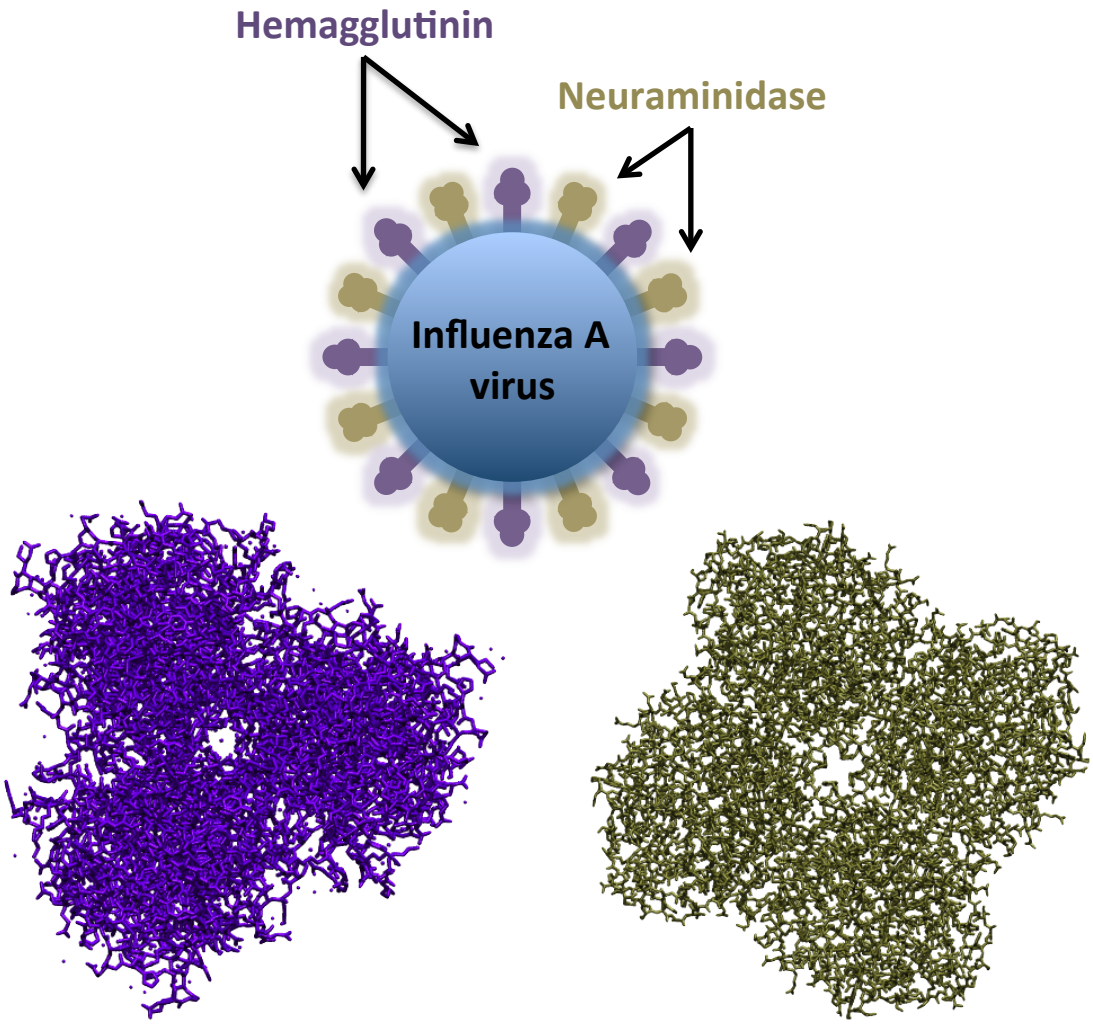


Figure 8.1: Hemagglutinin (HA) and neuraminidase (NA) are the primary surface antigens of influenza A. Shown are top-down representations of these proteins, as well as a schematic illustrating their placement as spike-like structures extending from the viral surface.

(Figure 8.1). HA, the principal antigen, serves to initiate influenza infection by binding glycan receptors on the host cell, leading to viral entry by endocytosis. Following replication, the secondary antigen, NA, cleaves the host cell glycans to allow escape of progeny virions. There are currently sixteen known types of HA (H1-H16) [120, 121] and nine known types of NA (N1-N9) [120], all of which are found in wild birds. Recently, two novel types of HA (H17-H18) and NA (N10-N11) were also identified as components of highly divergent influenza viruses found in Central/South American bat species [122, 123]. Influenza subtypes are classified based on their respective HA/NA combinations, such as H1N1, H2N2, and H3N2. Only H1-H3 and N1-N2 have been seen regularly circulating in the human population, although H2 subtypes have been extinct in humans since 1968 [124].

Aquatic birds, in particular, are regarded as the natural reservoir of influenza viruses [125], which manifest in avians as an asymptomatic intestinal condition. Infected avians excrete virus particles into their surrounding water environment, and these are subsequently ingested by other avians, who then also become infected. An ongoing infection cycle involving many HA/NA influenza subtypes is thus propagated, representing an ever-present public health risk. It is now widely accepted that all strains of influenza originated in avian sources and crossed over to humans and other species [126], predominantly through specificity-altering mutation of their surface antigens, HA and NA.

Hemagglutinin Structure

Wilson et al. resolved the first crystal structure of an HA in 1981 [127]. The particular HA used for crystallization, of the H3 subtype, was the largest biological molecule for which a structure had been determined at that time. Today, crystallographic coordinates are available for HAs of multiple subtypes, including H1, H2, H3, H5, H7, H9, H17, and H18 [123, 128–133].

Structurally, the HA glycoprotein comprises a homotrimeric assembly (Figure 8.2a) in which each monomeric subunit consists of two domains (Figure 8.2b). The globular head do-

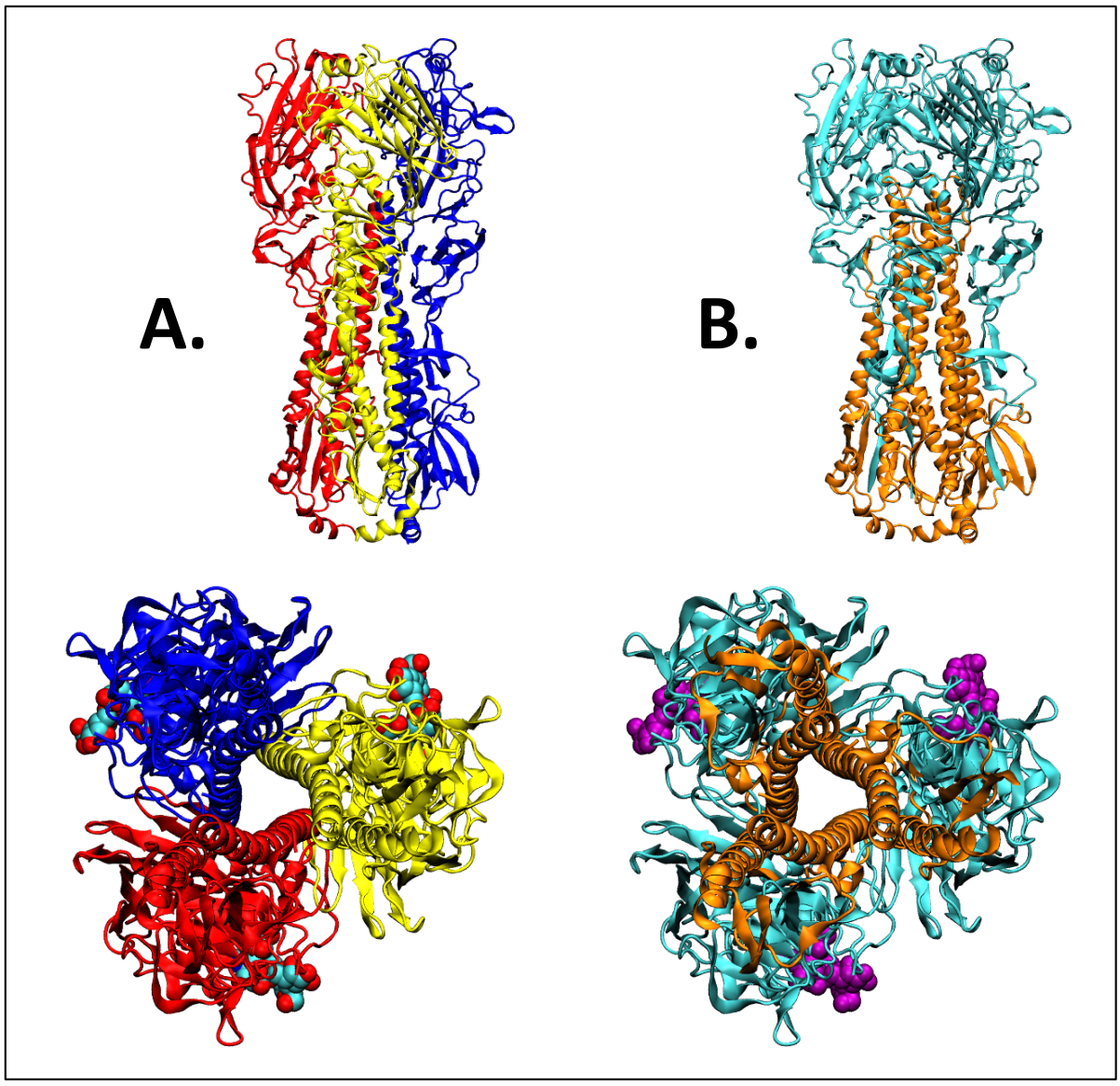


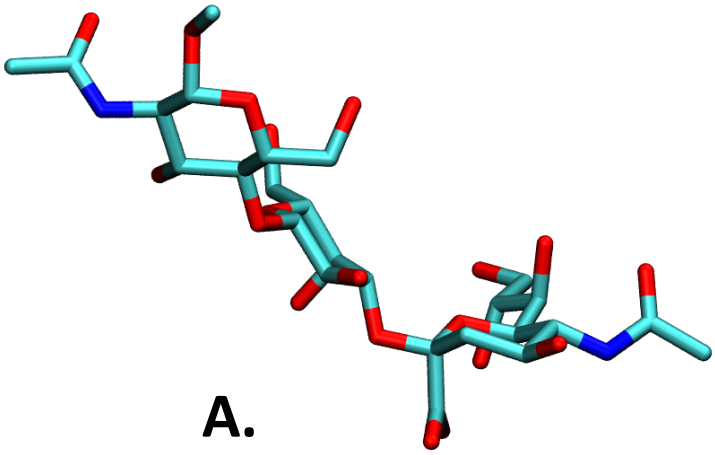
Figure 8.2: Cartoon representations of HA illustrating (a) homotrimeric subunits, and (b) globular head (cyan) and stalk (orange) domains. Bound glycan analogs are included in the lower panel to indicate the positions of receptor binding sites.

main (Figure 8.2b, cyan) encompasses the HA binding site, distinguished by a characterisitic jelly-roll motif formed by eight antiparallel β-strands. As each monomer exhibits a single binding site, a given HA exhibits three binding sites total (Figure 8.2a and 8.2b), facilitating notable avidity effects. The remainder of the HA structure represents the stalk domain (Figure 8.2b, orange), which is defined by a central α-helical coiled-coil anchored into the influenza viral envelope. HA is thus classified as a type I integral membrane glycoprotein, which presents with its amino-terminus projecting out and away from the viral surface as a hydrophilic spike.

Hemagglutinin Specificity

In general, HA recognizes glycan sequences that terminate in sialylated galactose, however, host range depends on the nature of the glycosidic linkage contained within this disaccharide motif, as it controls the overall 3D shape of the receptor (Figure 8.3). In birds, HA exhibits specificity for receptors containing α2-3 linkages (Figure 8.3a), which are found in the avian intestinal epithelium [134]. Alternatively, in humans, HA exhibits specificity for receptors containing α2-6 linkages (Figure 8.3b), which are found on the epithelial cells of the human upper respiratory tract [134]. An avian influenza virus may only pose a significant threat to humans following a mutation that allows HA to recognize, and preferentially bind, human-type receptor glycans.

There are two possible mechanisms by which this can occur. The first, called antigenic shift, results from the reassortment of genetic material from multiple virus strains within a shared host, often referred to as a mixing vessel [125]. Swine species often serve as mixing vessels, as they exhibit both avian- and human-type receptor glycans within their tracheal epithelium [135]. The 1957 Asian Flu (H2N2), 1968 Hong Kong Flu (H3N2), and the more recent 2009 Swine Flu (novel H1N1) are all thought to have crossed over via antigenic shift [136, 137].



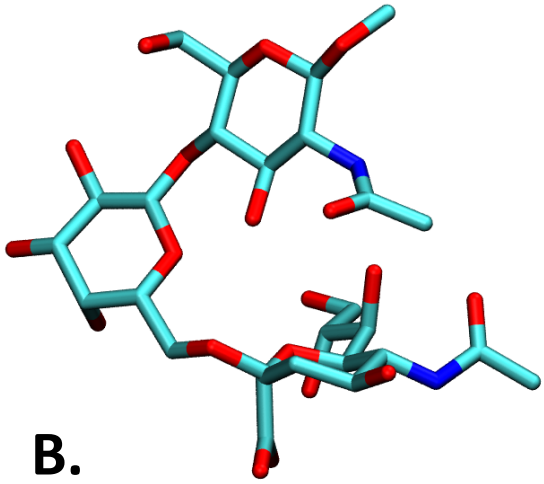


Figure 8.3: Trisaccharide analogs of (a) the avian influenza receptor glycan, which is characterized by sialic acid linked to the 3-position of galactose, and (b) the human influenza receptor glycan, which is characterized by sialic acid linked to the 6-position of galactose.

The second mechanism by which an avian influenza may acquire the ability to infect humans is called antigenic drift. Because viral RNA polymerases lack basic proofreading functionality, frequent mutations occur during genome transcription [125]. Over time, accumulation of favorable mutations can lead to adaptation of an avian HA to recognize, and preferentially bind, human-type receptor glycans. The 1918 Spanish Flu (H1N1) is thought to be the only pandemic influenza that crossed over directly from birds through antigenic drift [138, 139].

Experimental Characterization of Specificity and Affinity

The structural changes required to alter HA specificity are often extremely subtle. Experimental studies, most notably glycan array screenings, have demonstrated that two amino acid substitutions are generally sufficient to accomplish a specificity switch [140, 141]. In particular, the E190D/G225D mutation pair is associated with the avian-to-human switch in H1, while Q226L/G228S is implicated in H2 and H3 [140–142].

Glycan array screening involves washing a fluorescent-tagged protein of interest over a plate array that displays a library of affixed glycan samples. Binding is detected via fluorescence measurement, and commonalities between recognized glycans are used to infer specificity. While this technique is straightforward and provides valuable qualitative information about the binding preferences of a given protein, it is not well-suited to quantify the strength of binding interactions.

Kinetic assays, such as those employing nuclear magnetic resonance (NMR) or surface plasmon resonance (SPR) detection techniques, can be used to obtain binding affinity measurements, but their application to the HA system has been limited [143–147]. Furthermore, the inherent multimeric nature of HA structure and concomitant avidity effects complicate interpretation of these data and introduce a notable measure of uncertainty in the final affinity values.

NMR-based studies aim to determine equilibrium kinetics, however it is unclear to what degree avidity contributes to the resulting affinity measurements. Alternatively, SPR-based studies utilize chip-immobilized glycans, which practically manifest as an artificially high local glycan concentration, unlike the distribution found in a solution environment. While this experimental setup more reasonably mimics the presentation of glycans extending from a cellular surface, it neither facilitates determination of affinity values representative of a 1:1 binding mode. High glycan concentration translates to a high potential for rebinding events to occur before there is an opportunity for the protein to diffuse away from the chip surface. Thus, the experimental setup may bias measurements toward slower off rates (K_{off}), with faster on rates (K_{on}), producing artificially high final affinity (K_D) values. This situation is further exacerbated in the case of a multivalent protein, such as HA [148]. Indeed, K_D data describing HA binding determined by NMR is reported in the millimolar (mM) range [144, 145], while that determined by SPR is reported in the micromolar (μM) range [143].

Significance of this Study

Unlike experimental studies, whose outcomes are complicated by multivalency and biased kinetic events, computational approaches provide a mechanism to infer HA binding affinity for a simplified system representing a single binding site, without the ambiguity introduced by avidity effects. Techniques such as molecular mechanics – Poisson-Boltzmann / generalized Born surface area (MM-PB/GBSA) and thermodynamic integration (TI) may be applied to compute free energy changes (ΔG) associated with structural alterations affecting the HA binding interaction, which can be related to binding affinities (K_D) through Equation 8.1.

$$\Delta G = RT\ ln\ K_D \tag{8.1}$$

The original research study described in this document applies TI methodology with the aim of quantifying free energy changes, and thus relative affinities, associated with the H1 subtype binding interaction, including:

1. The effects of abrogating mutations Y98F and L194A.

2. The effects of specificity-altering mutations D190E and D225G.

3. The unique contributions of α2-3 and α2-6 receptor glycans to affinity and specificity.

CHAPTER 9:
QUANTIFYING BINDING AFFINITY AND ITS RELATIONSHIP TO SPECIFICITY IN INFLUENZA H1[1]

[1]Hadden, J. A.; Foley, B. F.; Woods, R. J. To be submitted.

Abstract

The hemagglutinin protein mediates adhesion of the influenza virus through binding of host cell receptor glycans. Despite the critical nature of this interaction, limited experimental data have been reported to quantify it. Here, independent trajectory thermodynamic integration (IT-TI) is employed to compute binding free energies associated with adhesion of 2009 Swine Flu H1 and 1918 Spanish Flu H1. The data produced in this study serve to quantify the effects of known specificity-altering mutations D190E and D225G (09H1), as well as quantify the contributions of individual glycan residues to H1 binding (09H1 and 18H1). Due to the weak affinity of these interactions and the multimeric nature of hemagglutinin structure, equivalent data is difficult to obtain experimentally. Altogether, these results likely represent the most reliable theoretical quantifications related to affinity and specificity currently available for the H1 system.

Keywords: swine flu, Spanish flu, influenza hemagglutinin, binding affinity, thermodynamic integration, free energy, GLYCAM

Introduction

In the spring of 2009, reports of a novel H1N1 influenza A subtype displaying widespread infection in humans heralded the first flu pandemic in more than 40 years [125]. Popularly referred to as Swine Flu, this strain was antigenically distinct from the earlier H1N1 Spanish Flu that killed up to 50 million people globally following the conclusion of the First World War [116]. While significantly less virulent than its predecessor, Swine Flu proved efficiently transmissible and resulted in up to 5.7 million estimated cases in the U.S. within the first four months of its emergence [118]. While there were fewer than 20,000 laboratory-confirmed mortalities [149], it has been estimated that between 151,700 and 575,400 perished worldwide, approximately 70 % from respiratory complications and the remainder from car-

diovascular issues [150]. As its byname suggests, the subtype was found to be of swine origin [151], yet achieved a successful foothold in the human population through the ability of its hemagglutinin (HA) adhesion protein to bind human-type receptor glycans.

While HAs generally recognize glycan sequences terminating in sialylated galactose, host range has long been correlated with specificity for the nature of the glycosidic linkage displayed within this disaccharide motif. According to the paradigm, avian influenza viruses exhibit specificity for sialic acid (Neu5Ac) linked to the 3-position of galactose (Neu5Ac-α2-3-Gal), as found in the intestinal epithelium of birds, while human influenza viruses exhibit specificity for sialic acid linked to the 6-position of galactose (Neu5Ac-α2-6-Gal), as found on the epithelial cells of the human upper respiratory tract [134]. Both of these receptor types are found in the tracheal epithelium of swine [135], and as a result, swine may serve as mixing vessels engendering reassortment or antigenic shift of influenza strains. As with Swine Flu, the 1957 Asian Flu (H2N2) and 1968 Hong Kong Flu (H3N2), each also responsible for a global pandemic, are thought to have entered the human population via this mechanism [136, 137]. In contrast, Spanish Flu is thought to have crossed over directly from avian species via antigenic drift [138, 139], a consequence of only two point mutations in the HA sequence [140, 141].

Complicating the established paradigm that human infection is dependent on α2-6 specificity, a recently published glycomic analysis of human respiratory tract tissue revealed that a wide range of α2-3-linked glycans are present throughout the lung and bronchus, not solely in the alveolar junctures and linings of the deeper airways, as previously thought [152]. While flu strains displaying α2-3 specificity are sporadically identified in humans, including highly pathogenic avian influenza (HPAI) viruses, isolates displaying distinct α2-6 specificity are observed far more commonly and characterize the vast majority of human-infective influenzas. Nevertheless, as α2-3-binding variants are generally associated with cases of particular severity, HA specificity remains an important consideration.

As with the 1918 Spanish Flu H1 (18H1), the species specificity of the 2009 Swine Flu H1 (09H1) may be altered by a single amino acid substitution [140], demonstrating that a subtle change in HA structure can significantly effect the affinity of glycan binding. Glycan array screening of H1 indicates that the mutation pair E190D/G225D controls the switch from avian (α2-3) to human (α2-6) specificity [140, 141]. Both 18H1 and 09H1 most frequently display the D190/D225 dyad, consistent with human specificity, and experimental analyses confirm a preference for α2-6-linked receptors [140, 141, 143, 153, 154]. Single mutant variants, E190/D225 or D190/G225, are associated with dual specificity and introduce the ability to bind both α2-3 and α2-6 glycan types [128, 140, 143, 154]. Notably, the D190/G225 variant was observed in 1-2 % of 09H1 sequences [154] and correlated with increased infection severity [155–158].

While HA specificity is relatively straightforward to characterize by glycan array screening, quantification of binding affinity represents a challenge. Only limited K_D data have been reported [143–147], and the inherent trimeric nature of HA structure, along with concomitant avidity effects, introduces a notable measure of uncertainty in interpreting kinetic data. In addition, glycan array screening is insensitive to low affinity interactions [159] and can indicate little or no binding to glycans that, based on related HA–glycan crystallographic structures, display satisfactory shape complementarity.

Computational methods offer an alternative approach, providing a mechanism to calculate affinities for monomeric HA–glycan interactions. Structural ensembles collected from molecular dynamics (MD) simulations can be evaluated to compute free energy changes (ΔG) relevant to binding, which can be directly related to affinity (K_D). Here, a free energy method referred to as independent trajectory thermodynamic integration (IT-TI) [48] is employed to quantify the effects of specificity-altering mutations D190E and D225G in 09H1, as well as glycan receptor contributions to binding in 09H1 and 18H1. Where possible, TI data have been compared with affinity data determined by experimental or alternative computational techniques. While all of these methodologies include sources of error, among the theoretical

Table 9.1: System information for simulated HAs.

HA	Glycan	PDB ID	Resolution[a]	Chain ID	Glycan B-Factor[b]
09H1	α2-6 analog	3UBE [164]	2.15	A	47.99
	α2-3 analog	3UBQ [164]	2.00	C	62.03
	apo (unbound)	3LZG [165]	2.60	A	–
18H1	α2-6 analog	2WRG [129]	3.00	J	118.32

[a]In Å, [b]Averaged over all atoms

methods, TI simulations may be expected to provide the most accurate quantification of affinity values for 1:1 binding modes currently available for the H1 system.

Computational Methods

Initial Structures

Starting coordinates for all H1 systems were extracted from crystallographic structures obtained through the RCSB Protein Data Bank (http://www.rcsb.org), as described in Table 9.1. In order to enhance computational efficiency, protein chains were cropped to contain a consensus of residues encompassing the binding site, extending 16-17 Å out from the bound glycans (residues 94-102, 127-162, 180-202, 213-233, and 246-254). Previous studies have shown that only residues within 12-16 Å need be considered when studying binding site activity [160, 161], and simulated systems may be truncated down to 15 Å without significantly affecting energies calculated via alchemical transformation [162]. Chain termini were capped with neutral peptide residues ACE (acetyl group) and NME (N-methyl amide). Proteins were parameterized with the AMBER ff99SB force field [163]. Starting coordinates for protein mutations were taken from ff99SB default residue templates so as to avoid any bias that might be introduced by selecting seemingly likely side chain orientations. H1 residues are referred to according to H3 numbering.

Trisaccharide analogs representing the native glycan receptors were employed, as specificity is thought to be determined by HA discrimination of these three terminal residues (Neu5Ac-Gal-GlcNAc, referred to as Sia-1, Gal-2, and GlcNAc-3) [164]. Glycans were parameterized with the GLYCAM06 (version h) force field [11] and capped with methoxy groups (OME) to prevent the possibility of hydrogen bonding at the reducing terminal position, which would not occur in complex with a larger receptor structure. Any crystallographic waters found within 6 Å of the glycans were retained. Starting coordinates for unbound glycan simulations were taken from bound glycan conformations. Each system was surrounded by a truncated octahedron of TIP3P explicit solvent [21], with a minimum 12-Å buffer between the solute and box edge. Simulation files were generated with the tleap module of AmberTools14 [29].

Simulation Details

To address the geometries of HA–glycan hydrogen bonds and ensure their initial presence prior to simulation, energy minimization was first applied to all hydrogens in the system. Subsequently, energy minimization was applied to the solvent, followed by all atoms in the system. Minimizations were performed with the AMBER14 [29] XMIN implementation (convergence criterion of $1.0e^{-4}$ kcal mol^{-1} Å^{-2}) in the sander.MPI module under constant volume (nVT) conditions.

MD simulations were performed with the pmemd.cuda [91, 92, 166] module of AMBER14 [29] under constant pressure (nPT) conditions. A Berendsen-type barostat with time constant of 1 ps was employed for pressure regulation, while a Langevin thermostat with collision frequency of 2 ps^{-1} was employed for temperature regulation. The random number seed was updated to a new value based on current wall-clock time upon each simulation restart to avoid correlation artifacts [167]. A nonbonded interaction cutoff of 8 Å was applied, beyond which long-range electrostatics were treated with the particle-mesh Ewald (PME) method [18]. Scaling factors for 1-4 nonbonded interactions were set to unity for glycans [17], and

1.2 and 2.0 for proteins to handle electrostatics and van der Waals (vdW), respectively [163]. Covalent bonds to hydrogen were constrained with the SHAKE algorithm [9] to allow a simulation time step of 2 fs. Because of the truncated nature of the HA model, Cartesian restraints were applied to the protein backbone (Cα, 10 kcal mol^{-1} Å^{-2}), as well as the terminal peptide caps (ACE and NME, 100 kcal mol^{-1} Å^{-2}) to maintain the binding site fold. The HA 220-loop was left unrestrained during investigation of the D225G mutation. Systems were heated to 300 K over 50 ps, followed by 5 ns of production dynamics, from which five structures were extracted at 1-ns intervals to provide independent starting coordinates for IT-TI simulations.

TI simulations were performed with a recently updated alchemical transformation implementation [168], available in the pmemd.MPI module of AMBER14 [29]. The simulation protocol included a short steepest-descent energy minimization (5000 steps), reheating to 300 K over 50 ps, and a further 50 ps of equilibration preceding 25 ns of production dynamics. Soft-core potentials were employed for both electrostatic and vdW interactions to facilitate a 1-step mutation process. Parameters of $\alpha_{LJ} = 0.4$, $\beta_C = 7.84$, and m = 2 were applied for protein mutations, while $\alpha_{LJ} = 0.5$, $\beta_C = 244$, and m = 6 were applied for glycan mutations [61]. Otherwise, simulation inputs were identical to those described above. Each independent TI simulation was performed over seven λ windows, and integration was estimated according to seven-point Gaussian quadrature. Values were averaged over five IT-TI replicates to produce final free energy quantifications. Uncertainties are reported as standard errors of the mean (SEM), or standard deviations over five IT-TI replicates divided by the square root of the number of replicates.

Results and Discussion

Abrogating Mutations as Positive Controls

Two mutations known to significantly abrogate H1 binding, Y98F and L194A, were employed as positive controls to validate TI simulation protocols. Results proved both inter-

Table 9.2: Calculated changes in the relative binding energy[a] for 09H1 protein mutations.

09H1 Mutation	α2-3 analog	α2-6 analog
Y98F	5.1 ± 0.4[b]	4.9 ± 0.8
L194A	4.3 ± 0.2	4.4 ± 0.4
D190E	-2.8 ± 0.5	-3.3 ± 0.7
D225G	-1.4 ± 0.1	-1.0 ± 0.1

[a]kcal mol^{-1}, [b]Uncertainties represent SEM

nally consistent and consistent with experiment, providing quantification of the effects of these mutations, as well as supporting the applicability of TI methodology for quantitative study of further binding site transformations in an HA system. Uncertainty estimates were systematically larger for simulations involving the α2-6 receptor analog, likely owing to the greater flexibility of the glycosidic linkages contained in this glycan.

The Y98F mutation involves loss of a critical hydrogen bond to Sia-1 and has been shown experimentally to reduce erythrocyte binding up to 95 % [169]. As residue Y98 contacts the conserved portion of the receptor, any reduction in affinity should affect binding of both α2-3 and α2-6 glycans to a similar extent. TI data are consistent with this expectation, indicating that Y98F impacts binding unfavorably by approximately 5 kcal mol^{-1}, with the values for α2-3 and α2-6 glycans falling within 0.2 kcal mol^{-1} of each other (Table 9.2).

The L194A mutation involves loss of a critical vdW contact for Sia-1 that provides important shape complementarity in the binding pocket. Erythrocyte studies indicate abrogation of binding by up to 96 % [169]. Residue L194 likewise contacts the conserved portion of the receptor, such that both α2-3 and α2-6 glycans should again experience a similar reduction of affinity. TI data are consistent with this expected outcome, indicating that L194A impacts binding unfavorably by over 4 kcal mol^{-1}, with the values for α2-3 and α2-6 glycans falling within 0.1 kcal mol^{-1} of each other (Table 9.2).

Specificity-altering Mutation D190E

The D190E/D225G mutation pair is associated with a specificity switch in H1 from α2-6 to α2-3 [140, 141]. As a human HA, 09H1 exhibits the D190/D225 dyad, consistent with specificity for α2-6-linked receptors. TI simulations were applied to crystallographic structures of H1 complexes containing both α2-3 and α2-6 glycan analogs in order to quantify the relative effects of these individual mutations on binding affinity.

For the α2-3 glycan, the D190E mutation should have a favorable effect on binding, corresponding to an expansion of 09H1 specificity to include this receptor type. TI data indeed support the favorability of this mutation, quantifying it at value of -2.8 kcal mol^{-1} (Table 9.2). An increase in affinity corresponding to a free energy change by this magnitude is thus sufficient to facilitate binding of α2-3-linked glycans, where binding was unlikely before. MD simulations of the λ_0 (D190/D225) and λ_1 (E190/D225) endpoint complexes suggest the structural origin of this result, indicating that the carboxylate moiety of D190 contacts the glycan through the 6-hydroxyl of Gal-2, while the E190 carboxylate can form additional contacts with the 7- and 9-hydroxyls of the Sia-1 glyceryl chain. This is in accordance with crystallographic data describing a 1934 H1 variant, which contains the mixed-specificity E190/D225 dyad and demonstrates contact of the E190 carboxylate with the Sia-1 9-hydroxyl [128].

As crystallographic data indicate that an equivalent interaction is possible for the α2-6 glycan upon substitution of E190 [128], it is reasonable to assume that the D190E mutation should also have a favorable effect on binding for this receptor type. TI data indeed predict favorability, quantifying the mutation at a value of -3.3 kcal mol^{-1} (Table 9.2). MD simulations of the λ_0 (D190/D225) and λ_1 (E190/D225) endpoint complexes indicate that the structural basis for increased binding affinity is, in fact, similar to that observed for the α2-3 case. When D190 is present, the carboxylate moiety contacts the glycan primarily through the amine of GlcNAc-3. Fleeting, weak contacts to the 7- and 9-hydroxyls of Sia-1 occur

(<3 % occupancy), but a stable, preferred contact to the 9-hydroxyl develops when E190 is substituted, owing to the extended length of the carboxylate chain.

Specificity-altering Mutation D225G (α2-3 case)

For the α2-3 glycan, the D225G mutation should have a favorable effect on binding, serving to impart dual specificity to 09H1. TI data indeed support the favorability of this mutation, quantifying it at a value of -1.4 kcal mol^{-1} (Table 9.2). An increase in affinity correlating to this relatively small free energy change – less than half the magnitude imparted to the free energy by D190E – is thus sufficient to facilitate binding of α2-3-linked glycans, where binding was unlikely before.

A recent SPR study detected no binding to wild type 09H1, yet reported an absolute binding affinity corresponding to a free energy of -7.8 kcal mol^{-1} for the G225 variant (09H1-D225G) [143]. As a significant portion of affinity no doubt derives from interactions conserved between 09H1 and 09H1-D225G, this value does not represent the affinity conferred by the D225G mutation alone, as computed here by TI. While a D225 substitution is sufficiently unfavorable to restrict binding to α2-3 receptors, the SPR experiment fails to quantify this associated magnitude.

Alternatively, a computational study applying the more approximate MM-GBSA method reported calculated affinity data for the D225G mutation corresponding to a free energy of -4.3 $\pm$ 0.66 kcal mol^{-1} [170]. While the limitations and associated inaccuracies of MM-GBSA relative to TI are widely known [171], further explanation for the discrepancy between these results lies in the structural interactions sampled by the respective simulations. For example, the MM-GBSA study indicates that Q226 plays an important role in the specificity-altering mechanism: While stable hydrogen bonds were observed between Q226 and the Sia-1 carboxylate and glyceryl hydroxyl groups in the 09H1–α23 complex, these contacts were dramatically reduced in the 09H1-D225G mutant, leading to an increase in the calculated nonpolar contribution of Q226 and ultimately resulting in an increase in the overall favorable

contribution of the residue. This is in contrast to a proposal based on crystallography [143], which suggests that Q226 only interacts with the α2-3 glycan in the 09H1-D225G mutant, where it is positioned close enough to form hydrogen bonds with Sia-1 and Gal-2. In the present study, MD simulations of the λ_0 (D190/D225) and λ_1 (D190/G225) endpoint complexes demonstrated roughly equivalent hydrogen bond behaviors of Q226 with the α2-3 analog, regardless of the substitution at position 225.

Discrepancies between simulations from the MM-GBSA study and those reported here can likely be attributed to subtle structural differences in the modeled complexes. While the present work employed crystallographic structures of 09H1 with bound receptor analogs, the MM-GBSA study utilized apo crystal structures with glycans superimposed into the binding sites. The hydrogen bond networks that developed during simulation of the constructed systems were apparently different from those seen in the crystallographic complexes. As such, the use of different starting coordinates describing the 09H1–glycan interaction likely led to exploration of alternative binding modes during the respective simulations. Ultimately, this may suggest insufficient sampling of the binding interaction by both theoretical studies.

Sampling inadequacies comprise a well-known source of uncertainty affecting TI simulations, however IT-TI methodology dramatically improves sampling relative to a single trajectory approach. The IT-TI protocol employed here facilitated exploration of five separate areas of phase space over 25 ns timescales and resulted in relatively low SEM error estimates. Nevertheless, HA–glycan complexes may be expected to display a multitude of valid binding modalities, and as a result, convergence is likely limited to structures closely resembling the crystallographic complex. Even so, the free energy values presented in this study successfully capture the effective trends associated with specificity-altering mutations in the 09H1 system within reasonable magnitudes.

Specificity-altering Mutation D225G (α2-6 case)

For the α2-6 glycan, experimental studies suggest that the D225G mutation should have a favorable effect on binding, and recent SPR analysis reports binding affinities corresponding to a free energy change of -1.2 kcal mol^{-1} [143]. TI data for this mutation agree well with experiment, quantifying it at a value of -1.0 kcal mol^{-1} (Table 9.2) and supporting the favorability of a D225G substitution.

It should be noted that experimental uncertainty is introduced when K_D values are determined based on curve-fitting to SPR data that display very fast kinetic on-rates, as exemplified in the cited study [143]. In addition, because the SPR experiment utilized glycans immobilized on a sensor surface, mimicking an artificially high local glycan concentration, the measured affinities may have been significantly influenced by avidity effects, potentially leading to overestimation of K_D [148]. However, because wild type 09H1 and 09H1-D225G likely demonstrate similar capacities for avidity, overestimation in the absolute affinity values may be expected to cancel upon taking a difference. That is, the relative affinity change between 09H1 and 09H1-D225G should correspond to the effect of the mutation on monomeric binding, as computed by TI. Thus, the theoretical result presented here, which reproduces the experimentally determined value to within 0.2 kcal mol^{-1}, demonstrates the ability of TI methodology to accurately quantify the relative effects of point mutations on H1 binding.

Glycan Receptor Contributions to Affinity and Specificity

TI simulations were further employed to determine the contributions to binding free energy imparted by the unique portions of α2-3 and α2-6 trisaccharide receptor analogs (Table 9.3). In both of these glycans, the position of Sia-1 is essentially equivalent, while the remainder of the analog, Gal-2-GlcNAc-3, is rendered unique by the difference in linkage. The origin of specificity, or discrimination between receptor types, thus lies primarily in the spatial positions and resulting binding site contacts adopted by the unique portions of the glycans. Due to the low affinity of monomeric HA–glycan interactions, no experimental data is available

Table 9.3: Calculated contributions to binding energy[a] for carbohydrate residues in the α2-3 (09H1) and α2-6 (09H1, 18H1) glycan analogs.

Ligand Component	09H1		18H1
	α2-3 analog	α2-6 analog	α2-6 analog
Gal-2-GlcNAc-3	-1.3 ± 1.3[b]	-4.9 ±1.6	-4.3 ± 1.7
GlcNAc-3	1.5 ± 0.5	2.0 ± 0.9	0.3 ± 0.6
Gal-2[c]	-2.8 ± 1.4	-6.9 ± 1.8	-4.6 ± 1.8

[a]kcal mol^{-1}, [b]Uncertainties represent SEM, [c]Obtained by difference

to indicate the relative importance of individual glycan residues within the binding motifs. To obtain such values computationally, two TI simulations were performed. In the first, the Gal-2-GlcNAc-3 disaccharide fragment of the trisaccharide receptor analog was decoupled. In the second, only the GlcNAc-3 monosaccharide fragment was decoupled. This facilitated characterization of the individual contributions of Gal-2 and GlcNAc-3 by difference, i.e., Gal-2-GlcNAc-3 contribution minus GlcNAc-3 contribution gives the Gal-2 contribution. It was necessary to obtain Gal-2 values by difference rather than by direct calculation, that is, by utilizing a disaccharide receptor analog and decoupling a terminal Gal-2, because Gal-2 was observed to display slightly different positioning and contact preferences within the binding site according to simulations of disaccharide versus trisaccharide complexes.

As a human HA, 09H1 exhibits α2-6 specificity, corresponding to a preference for receptors containing α2-6 linkages over α2-3. TI data support this, indicating that the unique portion of an α2-6-linked trisaccharide contributes -4.9 kcal mol^{-1} to binding, as opposed to only -1.3 kcal mol^{-1} for α2-3 (Table 9.3). Decomposing these values into per-residue contributions, Gal-2 is shown to impart -6.9 kcal mol^{-1} to binding for the α2-6 analog, compared to only -2.8 kcal mol^{-1} for the α2-3 analog. In both cases, the GlcNAc-3 contribution is shown to be weakly unfavorable to binding. MD simulations of disaccharide versus trisaccharide analogs suggest that this unfavorable effect results from destabilization of HA–Gal-2 contacts as GlcNAc-3 competes to form contacts of its own within the binding

site. This manifests essentially as a tug-of-war between Gal-2 and GlcNAc-3, as each vies to obtain optimal spatial positioning to establish strong hydrogen bonds with neighboring protein residues. Such behavior may be related to the use of protein backbone restraints, which could possibly hinder structural relaxations of the binding domain around the bound glycans, thus preventing Gal-2 and GlcNAc-3 from forming ideal binding site contacts simultaneously. Be that as it may, such restraints proved essential in this study to obtaining reasonably converged free energy values within accessible timescales.

For the 18H1 system, which is likewise characterized by α2-6 specificity, the Gal-2-GlcNAc-3 glycan component was shown to contribute a comparable amount (within 0.6 kcal mol^{-1}) to binding of an α2-6-linked trisaccharide to that seen in 09H1 (Table 9.3). As before, the majority of this contribution stemmed from Gal-2. In contrast to the 09H1 system, however, GlcNAc-3 appeared to provide no significant contribution to binding.

Taken together, TI data for 09H1 and 18H1 clearly indicate that differences in receptor affinity and specificity are attributable to Gal-2, while GlcNAc-3 provides no stabilization to the complex. These conclusions are supported by results from a recent MM-GBSA study of 09H1 [170], which also indicate that Gal-2 contributes most significantly to binding of an α2-3-linked glycan (-1.74 kcal mol^{-1}), whereas the GlcNAc-3 makes a negligible contribution (-0.25 kcal mol^{-1}). Discrepancies between the absolute free energy values reported by these two theoretical studies are likely the result of approximations inherent in MM-GBSA analyses.

Conclusions

While HA affinity is difficult to quantify experimentally, computational methods provide a mechanism to compute free energy changes that can be directly related to theoretical affinities. Here, IT-TI was employed to determine free energy changes associated with binding specificity in the H1 system. Results of this study confirm the expected relationships between specificity-altering mutations and experimentally observed receptor specificities. TI data provide quantification of the effects of these mutations and suggest that a loss of ex-

perimentally measurable binding likely equates to a free energy change of around 1-2 kcal mol^{-1} (Table 9.2). Further, data characterizing point mutations, as well as per-residue contributions of glycan binding, indicate that a similar shift in binding energy is associated with specificity switching. Analysis of per-residue contributions of glycans also confirms that the differences in affinity for α2-3 versus α2-6 receptors primarily result from interactions of Gal-2, as expected.

Altogether, through examination of the binding properties of the terminal trisaccharide typical of biologically relevant glycans, this study suggests that binding affinity in the H1 system stems from interactions of Sia-1-Gal-2, with negligible or unfavorable contribution from residues beyond this disaccharide motif. Nevertheless, not all glycans containing this terminal disaccharide demonstrate binding of similar affinity in glycan array screenings [141, 172–174]. This discrepancy likely relates to the differences in size and composition of the non-terminal portions of the screened glycans. It may be anticipated that larger glycan structures potentially introduce steric clashes, as indicated in computational carbohydrate grafting (CCG) studies [175, 176]. Otherwise, larger structures may serve to constrain the conformational flexibility of terminal glycans, either directly or through additional HA–glycan interactions, thus hindering their ability to sufficiently satisfy shape complementarity in the binding pocket.

Future work aimed at probing the binding properties of larger glycans, combined with experimental studies utilizing HA structures representing monomeric binding domains, may serve to enhance characterization of the HA adhesion interaction in a more biologically relevant context, including elucidation of the extent of avidity effects. Meanwhile, the data presented her provide quantification of the monomeric binding interaction between H1 and the terminal glycan residues responsible for affinity, as well as the effects of well-known abrogating and specificity-altering mutations. Despite the potential for error associated with finite sampling common to all large-scale TI analyses, these data serve to capture the effective trends associated with mutations within reasonable magnitudes and with relatively

low SEM error estimates. While these values may most accurately characterize HA–glycan binding modes that are not far deviated from that of the crystallographic complexes, they likely represent the most reliable theoretical quantifications of affinity for 1:1 binding modes currently available for the H1 system.

CHAPTER 10:
CONCLUSIONS AND RECOGNIZED CHALLENGES

Applying MD Simulations to Cellulose Microfibrils

The original research study presented in Chapter 6 (Unraveling Cellulose Microfibrils: A Twisted Tale) sought to explain why microfibrils modeled on linear crystallographic coordinates tend to adopt a right-handed twist during MD simulation. Further, through evaluation of commonly employed computational approximations, it aimed to determine whether simulation methodology was inherently responsible for any artifacts related to twisting behavior.

The results of this study indicate that a balance of competing forces ultimately controls cellulose microfibril twisting, which appears to be driven predominantly by attractive vdW interactions, while mitigated by both the cellulose intrachain hydrogen bond network and solvent effects at the microfibril surface. The magnitude of twisting observed for a given microfibril model is additionally dependent on its cross-sectional thickness, or number of constituent cellulose chains.

Computational approximations found to affect twisting behavior include choice of charge development protocol, choice of solvent model, and the use of dummy atoms to mimic the influence of electron lone pairs. As such, the results of this study suggest that the most accurate modeling of cellulose microfibrils may require application of partial charges that capture the electrostatic distribution across β-D-glucose as it occurs within the cellulose crystalline lattice, as well as explicitly modeled electron lone pairs to optimally reproduce, and sufficiently strengthen, the cellulose internal hydrogen bond network. Further, a solvent model

capable of inducing well-ordered, tightly coordinated water structure around the microfibril surface may also be required.

While comparison against experimentally-derived estimates suggest that the magnitude of twisting predicted by MD simulation might be reasonable, no definitive evidence currently exists to describe exactly the degree of twist that should be exhibited by a microfibril of given dimensions. Future experimental work is necessary to provide this characterization, after which theoretical microfibril behavior can be fine-tuned to correspond with experimental expectation.

Microfibrils of the cotton cellulose size regime, which served as the primary models in this study, display relatively subtle structural twisting. As such, twisting behavior in these and larger cellulose assemblies may prove to have a negligible effect on the results of MD simulations investigating interactions with cellulase enzymes or other molecules of interest. However, the widely-accepted model characterizing plant-based cellulose contains only 36-chains and, according to estimates obtained herein, should demonstrate a twist of around 2°-3° per cellobiose unit. Twisting of this magnitude could potentially alter the nature of the cellulose surface structure to a sufficient extent to effect the interaction profiles of ancillary molecules. As such, future studies are advised not to purposefully neglect twisting under the false assumption that it is unimportant or experimentally invalid, as has been done previously [84, 86].

A major issue in cellular modeling that remains unresolved is the degradation of the microfibril internal structure over the course of long simulation timescales. According to Matthews et al., hydroxymethyl groups gradually reorganize, breaking their equatorial contacts within the intralayer hydrogen bond network to form alternative interlayer contacts [81]. Once these local alterations to the crystalline lattice have become sufficiently widespread, the microfibril transitions to an I-HT form, no longer representative of natural cellulose under ambient conditions. Further work is necessary to address this concern. However, incorporation of the suggestions derived from the present study, including improved monomeric

charge distributions and explicitly modeled electron lone pairs, could serve to enhance stability of the internal hydrogen bond network and possibly preserve the microfibril structure over extended timescales.

In years to come, MD simulations of cellulose will no doubt play an essential role in guiding the technological advancement of many industrial processes, including biomass degradation relevant to ethanol production. The results of the original research presented herein – which provide explanation for a previously misunderstood dynamic behavior of cellulose, support the validity of that behavior through comparison with available experimental data, and suggest strategies for adjusting computational methodology to improve the accuracy of modeled structures – will thus serve to empower future simulation studies, allowing them to proceed with confidence in the underlying cellulose microfibril model.

Determining the Effect of Microfibril Twisting on Cellulose Diffraction Data

The original research study presented in Chapter 7 (Effect of Microfibril Twisting on Theoretical Powder Diffraction Patterns of Cellulose Iβ) sought to understand how microfibril twisting manifests in diffraction data collected for cellulose samples.

While previous work suggested that distortions in theoretical diffraction data observed for models produced by MD simulation could be attributed to both microfibril twisting and changes in cellulose unit cell dimensions [95], the results of the present study demonstrate that twisting is not a factor. Given the cross-sectional thickness of microfibrils employed for analysis, this leads to the conclusion that twisting on the order of $<1^\circ$ per cellobiose cannot be detected by powder diffraction methods. Further, if diffraction methodology is not sensitive to the subtle twisting displayed by large cellulose assemblies, such as the tunicin crystals used to infer the high-resolution crystallographic coordinates, then it stands to reason that these original structures could have been twisted in nature. Thus, the twisted models produced by MD simulation might be more representative of the reality of cellulose microfibril structure than the linear model suggested by crystallography.

The present study employed powder diffraction methodology to evaluate the effect of microfibril twisting on theoretical diffraction patterns. While powder diffraction was shown to be insensitive to subtle twisting, future work is necessary to determine whether other diffraction techniques, such as fiber diffraction, are capable of detecting this structural variation. Available softwares for predicting theoretical fiber patterns such as Calcdiff from Nishiyama et al. [95] and Sassena from Lindner et al. [177] might be employed toward such an investigation. Until definitive experimental evidence comes to light, the results of the original research presented herein indicate that there are currently no grounds on which to assume that twisted microfibrils do not represent valid exemplifications of cellulose Iβ structure.

Employing TI Calculations to Quantify Binding Free Energies in Influenza H1

The original research study presented in Chapter 9 (Quantifying Binding Affinity and Its Relationship to Specificity in Influenza H1) sought to compute highly accurate free energy changes associated with the H1 adhesion interaction. While avidity effects in HA systems render 1:1 binding affinities difficult to measure experimentally, computational methods may be applied to monomeric models to calculate free energies of binding, which can be directly related to theoretical affinities.

The results of this study provide quantification of the effects of abrogating and specificity-altering mutations in H1, as well as glycan contributions to affinity and specificity. Data characterizing the impact of point-mutations represent relative values that may have been impossible to obtain experimentally due to the insensitivity of binding assays to weak-affinity mutants. Data quantifying per-residue glycan contributions to binding for the case of α2-3 versus α2-6 receptors also represent experimentally inaccessible values for which calculation is made possible through free energy simulations.

While experimentally measured affinities may contain error introduced by multivalent interactions, biased kinetic events (owing to sample presentation strategy), or curve-fitting to binding data that display very fast kinetic on-rates, free energies calculated with TI also

contain sources of error. In general, the primary limitation of TI methodology is finite sampling. TI is, by nature, a computationally expensive technique, and convergence may not be feasibly attainable for large, complex systems that display a multitude of dynamic binding modalities, such as HA. As a result, free energy predictions may be highly dependent on the initial molecular coordinates employed for simulation, and final data may only well-characterize binding states that are not far deviated from this original structure. This is particularly true for cases requiring application of protein backbone restraints to achieve satisfactory convergence within the accessible timescale, as in the present study. Ultimately, given the inherent expense of TI calculations, alternative free energy methods may be better suited for investigation of HA, and other systems whose binding interactions display high degrees of motional freedom. For example, Bennet acceptance ratio (BAR) has been extensively compared against TI, and is reported to be both more robust and computationally efficient [63, 69–73].

Despite limitations of finite sampling common to all large-scale TI studies, the data presented herein likely provide the most reasonable theoretical quantifications characterizing affinity and specificity currently available for the H1 system. As pertinent experimental data contains uncertainty associated with avidity effects, explicit experimental validation of these results is not possible at the present time. In this case, future work is required to develop and screen structures that represent monomeric HA binding domains to provide experimental affinity measurements that are uncomplicated by multivalency. Additionally, experiments comparing data for monomers versus trimers could allow estimation of the proportion of overall binding affinity that derives from avidity effects in the HA system.

REFERENCES

[1] Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*; Harwood Academic: Amsterdam, 1998.

[2] Wong, C.-H. *Carbohydrate-based Drug Discovery*; Wiley-VHC: Weinheim, 2003.

[3] Foley, B. L.; Tessier, M. B.; Woods, R. J. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 652–697.

[4] Alder, B. J.; Wainwright, T. E. *Journal of Chemical Physics* **1959**, *31*, 459–466.

[5] McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

[6] Cheatham, T. E., III. *Current Opinion in Structural Biology* **2004**, *14*, 360–367.

[7] Fadda, E.; Woods, R. J. *Drug Discovery Today* **2010**, *15*, 596–609.

[8] Feller, S. E. *Current Opinion in Colloid & Interface Science* **2000**, *5*, 217–223.

[9] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327–341.

[10] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472.

[11] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *Journal of Computational Chemistry* **2008**, *29*, 622–655.

[12] Cox, J. D.; Pilcher, R. *Thermochemistry of Organic and Organometallic Compounds*; Academic: London, 1970.

[13] Piana, S.; Lindorff-Larsen, K.; Dirks, R. M.; K., S. J.; Dror, R. O.; Shaw, D. E. *Public Library of Science ONE* **2012**, *7*, e39918.

[14] Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. J. *Journal of Physical Chemistry B* **2007**, *111*, 13052–13063.

[15] Ponder, J. W.; Wu, C. J.; Ren, P. Y.; 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. *Journal of Physical Chemistry B* **2010**, *114*, 2549–2564.

[16] Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *Journal of Physical Chemistry* **1993**, *97*, 10269–10280.

[17] Woods, R. J.; Chappelle, R. *Journal of Molecular Structure-Theochem* **2000**, *527*, 149–156.

[18] Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

[19] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguso, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[20] Rini, J. M. *Annual Review of Biophysics and Biomolecular Structure* **1995**, *24*, 551–577.

[21] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926–935.

[22] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *Journal of Chemical Physics* **2004**, *120*, 9665–9678.

[23] Mahoney, M. W.; Jorgensen, W. L. *Journal of Chemical Physics* **2000**, *112*, 8910–8922.

[24] Gonzlez, M. A.; Abascal, J. L. *Journal of Chemical Physics* **2010**, *132*, 096101.

[25] Florova, P.; Sklenovsky, P.; Banas, P.; Otyepka, M. *Journal of Chemical Theory and Computation* **2010**, *6*, 3569–3579.

[26] Fadda, E.; Woods, R. J. *Journal of Chemical Theory and Computation* **2011**, *7*, 3391–3398.

[27] Hadden, J. A.; French, A. D.; Woods, R. J. *Biopolymers* **2013**, *99*, 746–756.

[28] Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M., Jr.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M. J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California: San Francisco, 2012.

[29] Case, D. A.; Babin, V.; Berryman, J.; Betz, R. M.; Cai, Q.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossvary, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M., Jr.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *AMBER 14*; University of California: San Francisco, 2014; Development version: Accessed January 9, 2013.

[30] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684–3690.

[31] Andersen, H. C. *Journal of Chemical Physics* **1980**, *72*, 2384–2393.

[32] Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523–535.

[33] Faller, R.; de Pablo, J. J. *Journal of Chemical Physics* **2002**, *116*, 55–59.

[34] Liu, D. C.; Nocedal, J. *Mathematical Programming B* **1989**, *45*, 503–528.

[35] Pierce, L. C. T.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.; Walker, R. C. *Journal of Chemical Theory and Computation* **2012**, *8*, 2997–3002.

[36] Schlitter, J.; Engels, M.; Krger, P.; Jacoby, E.; Wollmer, A. *Molecular Simulation* **1993**, *10*, 291–308.

[37] Torrie, G. M.; Valleau, J. P. *Journal of Computational Physics* **1977**, *23*, 187–199.

[38] Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141–151.

[39] Hamelberg, D.; Mongan, J.; McCammon, J. A. *The Journal of Chemical Physics* **2004**, *120*, 11919–11929.

[40] Kirkwood, J. G. *Journal of Chemical Physics* **1935**, *3*, 300–313.

[41] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; W., W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III. *Accounts of Chemical Research* **2000**, *33*, 889–897.

[42] Kadirvelraj, R.; Foley, B. L.; Dyekjaer, J. D.; Woods, R. J. *Journal of the American Chemical Society* **2008**, *130*, 16933–16942.

[43] Zacharias, M.; Straatsma, T. P.; McCammon, J. A.; Quiocho, F. A. *Biochemistry* **1993**, *32*, 7428–7434.

[44] Liang, G.; Schmidt, R. K.; Yu, H. A.; Cumming, D. A.; Brady, J. W. *Journal of Physical Chemistry* **1996**, *100*, 2528–2534.

[45] Pathiaseril, A.; Woods, R. J. *Journal of the American Chemical Society* **2000**, *122*, 331–338.

[46] Ganguly, D.; Mukhopadhyay, C. *Biopolymers* **2006**, *83*, 83–94.

[47] Bucher, D.; Grant, B. J.; McCammon, J. A. *Biochemistry* **2011**, *50*, 10530–10539.

[48] Lawrenz, M.; Baron, R.; McCammon, J. A. *Journal of Chemical Theory and Computation* **2009**, *5*, 1106–1116.

[49] Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A. *Proteins-Structure Function and Bioinformatics* **2010**, *78*, 2523–2532.

[50] Simonson, T. *Molecular Physics* **1993**, *80*, 441–447.

[51] Lin, C. L.; Wood, R. H. *Journal of Computational Chemistry* **1994**, *15*, 149–154.

[52] Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. *Faraday Symposia of the Chemical Society* **1982**, 55–67.

[53] Pearlman, D. A.; Kollman, P. A. *Journal of Chemical Physics* **1989**, *90*, 2460–2470.

[54] Maye, P. V.; Mezei, M. *Theochem-Journal of Molecular Structure* **1996**, *362*, 317–324.

[55] Mezei, M. *Journal of Computational Chemistry* **1992**, *13*, 651–656.

[56] Resat, H.; Mezei, M. *Journal of Chemical Physics* **1993**, *99*, 6052–6061.

[57] Steinbrecher, T.; Mobley, D. L.; Case, D. A. *Journal of Chemical Physics* **2007**, *127*, 25–33.

[58] Pitera, J. W.; Van Gunsteren, W. F. *Molecular Simulation* **2002**, *28*, 45–65.

[59] Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; Vangunsteren, W. F. *Chemical Physics Letters* **1994**, *222*, 529–539.

[60] Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *Journal of Chemical Physics* **1994**, *100*, 9025–9031.

[61] Steinbrecher, T.; Joung, I.; Case, D. A. *Journal of Computational Chemistry* **2011**, *32*, 3253–3263.

[62] Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. *Journal of Chemical Theory and Computation* **2010**, *6*, 1018–1027.

[63] Bruckner, S.; Boresch, S. *Journal of Computational Chemistry* **2011**, *32*, 1320–1333.

[64] Swope, W.; Andersen, H. *Journal of Physical Chemistry* **1984**, *88*, 6548–6556.

[65] Hummer, G.; Pratt, L. R.; Garca, A. E. *Journal of Physical Chemistry* **1996**, *100*, 1206–1215.

[66] Shyu, C.; Ytreberg, F. M. *Journal of Computational Chemistry* **2009**, *30*, 2297–2304.

[67] Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *Journal of Chemical Physics* **2003**, *119*, 5740–5761.

[68] Straatsma, T.; Berendsen, H.; Stamp, A. *Molecular Physics* **1986**, *57*, 89–95.

[69] Shirts, M. R.; Pande, V. S. *Journal of Chemical Physics* **2005**, *122*, 144107–144116.

[70] Konig, G.; Brooks, B. R. *Journal of Computer-Aided Molecular Design* **2012**, *26*, 543–550.

[71] Paliwal, H.; Shirts, M. R. *Journal of Chemical Theory and Computation* **2011**, *7*, 4115–4134.

[72] Bruckner, S.; Boresch, S. *Journal of Computational Chemistry* **2011**, *32*, 1303–1319.

[73] de Ruiter, A.; Boresch, S.; Oostenbrink, C. *Journal of Computational Chemistry* **2013**, *34*, 1024–1034.

[74] Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. *Journal of the American Chemical Society* **2003**, *125*, 14300–14306.

[75] Nishiyama, Y.; Langan, P.; Chanzy, H. *Journal of the American Chemical Society* **2002**, *124*, 9074–9082.

[76] Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydrate Research* **2006**, *341*, 138–152.

[77] Yui, T.; Nishimura, S.; Akiba, S.; Hayashi, S. *Carbohydrate Research* **2006**, *341*, 2521–2530.

[78] Yui, T.; Hayashi, S. *Biomacromolecules* **2007**, *8*, 817–824.

[79] Matthews, J. F.; Bergenstrahle, M.; Beckham, G. T.; Himmel, M. E.; Nimlos, M. R.; Brady, J. W.; Crowley, M. F. *Journal of Physical Chemistry B* **2011**, *115*, 2155–2166.

[80] Paavilainen, S.; Rog, T.; Vattulainen, I. *Journal of Physical Chemistry B* **2011**, *115*, 3747–3755.

[81] Matthews, J. F.; Beckham, G. T.; Bergenstrahle-Wohlert, M.; Brady, J. W.; Himmel, M. E.; Crowley, M. F. *Journal of Chemical Theory and Computation* **2012**, *8*, 735–748.

[82] Bu, L. T.; Beckham, G. T.; Crowley, M. F.; Chang, C. H.; Matthews, J. F.; Bomble, Y. J.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R. *Journal of Physical Chemistry B* **2009**, *113*, 10994–11002.

[83] Glass, D. C.; Moritsugu, K.; Cheng, X. L.; Smith, J. C. *Biomacromolecules* **2012**, *13*, 2634–2644.

[84] Wu, S.; Zhan, H. Y.; Wang, H. M.; Ju, Y. *Chinese Journal of Chemical Physics* **2012**, *25*, 191–198.

[85] Doblin, M. S.; Kurek, I.; Jacob-Wilk, D.; Delmer, D. P. *Plant and Cell Physiology* **2002**, *43*, 1407–1420.

[86] Wohlert, J.; Berglund, L. A. *Journal of Chemical Theory and Computation* **2011**, *7*, 753–760.

[87] Hanley, S. J.; Revol, J. F.; Godbout, L.; Gray, D. G. *Cellulose* **1997**, *4*, 209–220.

[88] Macrae, C. F.; Bruno, I. J.; Chisholm, J. A.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Rodriguez-Monge, L.; Taylor, R.; van de Streek, J.; Wood, P. A. *Journal of Applied Crystallography* **2008**, *41*, 466–470.

[89] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, revision D.01*; Gaussian, Incorporated: Wallingford, CT, 2004.

[90] Breneman, C. M.; Wiberg, K. B. *Journal of Computational Chemistry* **1990**, *11*, 361–373.

[91] Goetz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *Journal of Chemical Theory and Computation* **2012**, *8*, 1542–1555.

[92] Salomon-Ferrer, R.; Gtz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. *Journal of Chemical Theory and Computation* **2013**, *9*, 3878–3888.

[93] Kirschner, K. N.; Woods, R. J. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 10541–10545.

[94] Foley, B. L.; Woods, R. J. *International Journal of Quality, Statistics, and Reliability* **2014**; In preparation.

[95] Nishiyama, Y.; Johnson, G. P.; French, A. D. *Cellulose* **2012**, *19*, 319–336.

[96] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chemical Physics Letters* **1995**, *246*, 122–129.

[97] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry* **1996**, *100*, 19824–19839.

[98] Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275–291.

[99] Onufriev, A.; Bashford, D.; Case, D. A. *Journal of Physical Chemistry B* **2000**, *104*, 3712–3720.

[100] Onufriev, A.; Bashford, D.; Case, D. A. *Proteins-Structure Function and Bioinformatics* **2004**, *55*, 383–394.

[101] Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *Journal of Chemical Theory and Computation* **2007**, *3*, 156–169.

[102] Shang, Y.; Nguyen, H.; Wickstrom, L.; Okur, A.; Simmerling, C. *Journal of Molecular Graphics & Modelling* **2011**, *29*, 676–684.

[103] Heiner, A. P.; Kuutti, L.; Teleman, O. *Carbohydrate Research* **1998**, *306*, 205–220.

[104] Heiner, A. P.; Teleman, O. *Langmuir* **1997**, *13*, 511–518.

[105] Tschampel, S. M.; Kennerty, M. R.; Woods, R. J. *Journal of Chemical Theory and Computation* **2007**, *3*, 1721–1733.

[106] Basma, M.; Sundara, S.; Calgan, D.; Vernali, T.; Woods, R. J. *Journal of Computational Chemistry* **2001**, *22*, 1125–1137.

[107] Li, Y.; Lin, M. L.; Davenport, J. W. *Journal of Physical Chemistry C* **2011**, *115*, 11533–11539.

[108] Hirai, A.; Tsuji, M.; Horii, F. *Sen'i Gakkaishi* **1998**, *54*, 506–510.

[109] Wehle, M.; Vilotijevic, I.; Lipowsky, R.; Seeberger, P. H.; Varon Silva, D.; Santer, M. *Journal of the American Chemical Society* **2012**, *134*, 18964–18972.

[110] Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. *Computer Physics Communications* **1995**, *91*, 43–56.

[111] Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *Journal of Chemical Theory and Computation* **2008**, *4*, 435–447.

[112] van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

[113] Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

[114] Balaeff, A. *Mol_Volume 1.0*; Theoretical Biophysics Group, Beckman Institute, and Board of Trustees of the University of Illinois: Urbana-Champaign, 2001; Accessed August 2013. `http://www.ks.uiuc.edu/Development/MDTools/molvolume`.

[115] Wojdyr, M. *Debyer r77*; Marcin Wojdyr, 2009; Accessed June 2012. `http://code.google.com/p/debyer`.

[116] Johnson, N. P.; Mueller, J. *Bulletin of the History of Medicine* **2002**, *76*, 105–115.

[117] Rajagopal, S.; Treanor, J. *Seminars in Respiratory and Critical Care Medicine* **2007**, *28*, 159–170.

[118] Reed, C.; Angulo, F. J.; Swerdlow, D. L.; Lipsitch, M.; Meltzer, M. I.; Jernigan, D.; Finelli, L. *Emerging Infectious Diseases* **2009**, *15*, 2004–2007.

[119] Stohr, K. *Lancet Infectious Diseases* **2002**, *2*, 517.

[120] Gamblin, S. J.; Skehel, J. J. *Journal of Biological Chemistry* **2010**, *285*, 28403–28409.

[121] Fouchier, R. A. M.; Munster, V.; Wallensten, A.; Bestebroer, T. M.; Herfst, S.; Smith, D.; Rimmelzwaan, G. F.; Olsen, B.; Osterhaus, A. D. M. E. *Journal of Virology* **2005**, *79*, 2814–2822.

[122] Tong, S. X.; Li, Y.; Rivailler, P.; Conrardy, C.; Castillo, D. A. A.; Chen, L. M.; Recuenco, S.; Ellison, J. A.; Davis, C. T.; York, I.; Turmelle, A. S.; Moran, D.; Rogers, S.; Shi, M.; Tao, Y.; Weil, M. R.; Tang, K.; Rowe, L. A.; Sammons, S.; Xu, X. Y.; Frace, M.; Lindblade, K. A.; Cox, L. J., N. J.and Anderson; Rupprecht, C. E.; Donis, R. O. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 4269–4274.

[123] Tong, S.; Zhu, X.; Li, Y.; Shi, M.; Zhang, J.; Bourgeois, M.; Yang, H.; Chen, X.; Recuenco, S.; Gomez, J.; Chen, L.; Johnson, A.; Tao, Y.; Dreyfus, C.; Yu, W.; McBride, R.; Carney, P. J.; Gilbert, A. T.; Chang, J.; Guo, Z.; Davis, C. T.; Paulson, J. C.;

Stevens, J.; Rupprecht, C. E.; Holmes, E. C.; Wilson, I. A.; Donis, R. O. *Public Library of Science Pathogens* **2013**, *9*, e1003657.

[124] Ma, W.; Vincent, A. L.; Gramer, M. R.; Brockwell, C. B.; Lager, K. M.; Janke, B. H.; Gauger, P. C.; Patnayak, D. P.; Webby, R. J.; Richt, J. A. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 20949–20954.

[125] Cox, N. J.; Subbarao, K. *Annual Review of Medicine* **2000**, *51*, 407–421.

[126] Webster, R. G.; Bean, W. J.; Gorman, O. T.; Chambers, T. M.; Kawaoka, Y. *Microbiological Reviews* **1992**, *56*, 152–179.

[127] Wilson, I. A.; Skehel, J. J.; Wiley, D. C. *Nature* **1981**, *289*, 366–373.

[128] Gamblin, S. J.; Haire, L. F.; Russell, R. J.; Stevens, D. J.; Xiao, B.; Ha, Y.; Vasisht, N.; Steinhauer, D. A.; Daniels, R. S.; Elliot, D. C., A. Wiley; Skehel, J. J. *Science* **2004**, *303*, 1838–1842.

[129] Liu, J.; Stevens, D. J.; Haire, L. F.; Walker, P. A.; Coombs, P. J.; Russell, R. J.; Gamblin, S. J.; Skehel, J. J. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 17175–17180.

[130] Ha, Y.; Stevens, D. J.; Skehel, J. J.; Wiley, D. C. *Virology* **2003**, *309*, 209–218.

[131] Ha, Y.; Stevens, D. J.; Skehel, J. J.; Wiley, D. C. *European Molecular Biology Organization Journal* **2002**, *21*, 865–875.

[132] Yang, H.; Chen, L. M.; Carney, P. J.; Donis, R. O.; Stevens, J. *Public Library of Science Pathogens* **2010**, *6*, e1001081.

[133] Zhu, X.; Yu, W.; McBride, R.; Li, Y.; Chen, L. M.; Donis, R. O.; Tong, S.; Paulson, J. C.; Wilson, I. A. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 1458–1463.

[134] Rogers, G. N.; Paulson, J. C. *Virology* **1983**, *127*, 361–373.

[135] Ito, T.; Couceiro, J.; Kelm, S.; Baum, L. G.; Krauss, S.; Castrucci, M. R.; Donatelli, I.; Kida, H.; Paulson, J. C.; Webster, R. G.; Kawaoka, Y. *Journal of Virology* **1998**, *72*, 7367–7373.

[136] Scholtissek, C.; Rohde, W.; Vonhoyningen, V.; Rott, R. *Virology* **1978**, *87*, 13–20.

[137] Kawaoka, Y.; Krauss, S.; Webster, R. G. *Journal of Virology* **1989**, *63*, 4603–4608.

[138] Reid, A. H.; Taubenberger, J. K.; Fanning, T. G. *Nature Reviews Microbiology* **2004**, *2*, 909–914.

[139] Taubenberger, J. K.; Reid, A. H.; Lourens, R. M.; Wang, R. X.; Jin, G. Z.; Fanning, T. G. *Nature* **2005**, *437*, 889–893.

[140] Glaser, L.; Stevens, J.; Zamarin, D.; Wilson, I. A.; Garcia-Sastre, A.; Tumpey, T. M.; Basler, C. F.; Taubenberger, J. K.; Palese, P. *Journal of Virology* **2005**, *79*, 11533–11536.

[141] Stevens, J.; Blixt, O.; Glaser, L.; Taubenberger, J. K.; Palese, P.; Paulson, J. C.; Wilson, I. A. *Journal of Molecular Biology* **2006**, *355*, 1143–1155.

[142] Connor, R. J.; Kawaoka, Y.; Webster, R. G.; Paulson, J. C. *Virology* **1994**, *205*, 17–23.

[143] Zhang, W.; Shi, Y.; Qi, J.; Gao, F.; Li, Q.; Fan, Z.; Yan, J.; Gao, G. F. *Journal of Virology* **2013**, *87*, 5949–5958.

[144] Sauter, N. K.; Bednarski, M. D.; Wurzburg, B. A.; Hanson, J. E.; Whitesides, G. M.; Skehel, J. J.; Wiley, D. C. *Biochemistry* **1989**, *28*, 8388–8396.

[145] Sauter, N. K.; Hanson, J. E.; Glick, G. D.; Brown, J. H.; Crowther, R. L.; Park, S. J.; Skehel, J. J.; Wiley, D. C. *Biochemistry* **1992**, *31*, 9609–9621.

[146] Takemoto, D. K.; Skehel, J. J.; Wiley, D. C. *Virology* **1996**, *217*, 452–458.

[147] Hidari, K. P. J.; Shimada, S.; Suzuki, Y.; Suzuki, T. *Glycoconjugate Journal* **2007**, *24*, 583–590.

[148] Lortat-Jacob, H.; Chouin, E.; Cusack, S.; van Raaij, M. J. *Journal of Biological Chemistry* **2001**, *276*, 9009–9015.

[149] World Health Organization; *Pandemic (H1N1) 2009 – update 112*; 2010; Accessed March 2014. `http://www.who.int/csr/don/2010_08_06/en`.

[150] Dawood, F. S.; Iuliano, A. D.; Reed, C.; Meltzer, M. I.; Shay, D. K.; Cheng, P. Y.; Bandaranayake, D.; Breiman, R. F.; Brooks, W. A.; Buchy, P.; Feikin, D. R.; Fowler, K. B.; Gordon, A.; Hien, N. T.; Horby, P.; Huang, Q. S.; Katz, M. A.; Krishnan, A.; Lal, R.; Montgomery, J. M.; Molbak, K.; Pebody, R.; Presanis, A. M.; Razuri, H.; Steens, A.; Tinoco, Y. O.; Wallinga, J.; Yu, H. J.; Vong, S.; Bresee, J.; Widdowson, M. A. *Lancet Infectious Diseases* **2012**, *12*, 687–695.

[151] Dawood, F. S.; Jain, S.; Finelli, L.; Shaw, M. W.; Lindstrom, S.; Garten, R. J.; Gubareva, L. V.; Xu, X. Y.; Bridges, C. B.; Uyeki, T. M. *New England Journal of Medicine* **2009**, *360*, 2605–2615.

[152] Walther, T.; Karamanska, R.; Chan, R. W.; Chan, M. C.; Jia, N.; Air, G.; Hopton, C.; Wong, M. P.; Dell, A.; Malik Peiris, J. S.; Haslam, S. M.; Nicholls, J. M. *Public Library of Science Pathogens*.

[153] Liu, Y.; Childs, R. A.; Matrosovich, T.; Wharton, S.; Palma, A. S.; Chai, W.; Daniels, R.; Gregory, V.; Uhlendorff, J.; Kiso, M.; Klenk, H. D.; Hay, A.; Feizi, T.; Matrosovich, M. *Journal of Virology* **2010**, *84*, 12069–12074.

[154] Chutinimitkul, S.; Herfst, S.; Steel, J.; Lowen, A. C.; Ye, J.; van Riel, D.; Schrauwen, E. J.; Bestebroer, T. M.; Koel, B.; Burke, D. F.; Sutherland-Cash, K. H.; Whittleston, C. S.; Russell, C. A.; Wales, D. J.; Smith, D. J.; Jonges, M.; Meijer, A.; Koopmans, M.;

Rimmelzwaan, G. F.; Kuiken, T.; Osterhaus, A. D.; Garcia-Sastre, A.; Perez, D. R.; Fouchier, R. A. *Journal of Virology* **2010**, *84*, 11802–11813.

[155] Chen, H.; Wen, X.; To, K. K.; Wang, P.; Tse, H.; Chan, J. F.; Tsoi, H. W.; Fung, K. S.; Tse, C. W.; Lee, R. A.; Chan, K. H.; Yuen, K. Y. *Journal of Infectious Diseases* **2010**, *201*, 1517–1521.

[156] Kilander, A.; Rykkvin, R.; Dudman, S. G.; Hungnes, O. *Euro Surveillance* **2010**, *15*, pii=19498.

[157] Mak, G. C.; Au, K. W.; Tai, L. S.; Chuang, K. C.; Cheng, K. C.; Shiu, T. C.; Lim, W. *Euro Surveillance* **2010**, *15*, pii=19534.

[158] Anton, A.; Marcos, M. .; Martnez, M. J.; Ramn, S.; Martinez, A.; Cardeosa, N.; Godoy, P.; Torner, N.; De Molina, P.; Isanta, R.; Jimnez de Anta, M. T.; Pumarola, T. *Diagnostic Microbiology and Infectious Disease* **2010**, *67*, 207–208.

[159] Taylor, M. E.; Drickamer, K. *Glycobiology* **2009**, *19*, 1155–1162.

[160] Kaukonen, M.; Soderhjelm, P.; Heimdal, J.; Ryde, U. *Journal of Chemical Theory and Computation* **2008**, *4*, 985–1001.

[161] Hu, L. H.; Eliasson, J.; Heimdal, J.; Ryde, U. *Journal of Physical Chemistry A* **2009**, *113*, 11793–11800.

[162] Genheden, S.; Ryde, U. *Journal of Chemical Theory and Computation* **2012**, *8*, 1449–1458.

[163] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins-Structure Function and Bioinformatics* **2006**, *65*, 712–725.

[164] Xu, R.; McBride, R.; Nycholat, C. M.; Paulson, J. C.; Wilson, I. A. *Journal of Virology* **2012**, *86*, 982–990.

[165] Xu, R.; Ekiert, D. C.; Krause, J. C.; Hai, R.; Crowe, J. E.; Wilson, I. A. *Science* **2010**, *328*, 357–360.

[166] Le Grand, S.; Gtz, A. W.; Walker, R. C. *Computer Physics Communications* **2013**, *184*, 374–380.

[167] Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. *Journal of Chemical Theory and Computation* **2008**, *4*, 1669–1680.

[168] Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammont, J. A. *Journal of Chemical Theory and Computation* **2013**, *9*.

[169] Martin, J.; Wharton, S. A.; Lin, Y. P.; Takemoto, D. K.; Skehel, J. J.; Wiley, D. C.; Steinhauer, D. A. *Virology* **1998**, *241*, 101–111.

[170] Pan, D.; Xue, W.; Wang, X.; Guo, J.; Liu, H.; Yao, X. *Journal of Molecular Modeling* **2012**, *18*, 4355–4366.

[171] Hadden, J. A.; Tessier, M. B.; Fadda, E.; Woods, R. J. In *Methods in Molecular Biology: Glycoinformatics*; Lutteke, T.; Frank, R., Eds.; Humana Press: Totowa, NJ; Chapter Chapter 8: Calculating binding free energies for protein-carbohydrate complexes; In press.

[172] Childs, R. A.; Palma, A. S.; Wharton, S.; Matrosovich, T.; Liu, Y.; Chai, W. G.; Campanero-Rhodes, M. A.; Zhang, Y. B.; Eickmann, M.; Kiso, M.; Hay, A.; Matrosovich, M.; Feizi, T. *Nature Biotechnology* **2009**, *27*, 797–799.

[173] Wang, Z.; Chinoy, Z. S.; Ambre, S. G.; Peng, W. J.; McBride, R.; de Vries, R. P.; Glushka, J.; Paulson, J. C.; Boons, G. J. *Science* **2013**, *341*, 379–383.

[174] Stevens, J.; Blixt, O.; Paulson, J. C.; Wilson, I. A. *Nature Reviews Microbiology* **2006**, *4*, 857–864.

[175] Tessier, M. B.; Grant, O. C.; Heimburg-Molinaro, J.; Smith, D.; Jadey, S.; Gulick, A. M.; Glushka, J.; Deutscher, S. L.; Rittenhouse-Olson, K.; Woods, R. J. *Public Library of Science ONE* **2013**, *8*, e54874.

[176] Grant, O. C.; Smith, H. M. K.; Firsova, D.; Fadda, E.; Woods, R. J. *Glycobiology* **2014**, *24*, 17–25.

[177] Lindner, B.; Smith, J. C. *Computer Physics Communications* **2012**, *183*, 1491–1501.