12-2012

Aspects of biomacromolecular dynamics at different scales

Dennis Christian Glass
dglass4@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Dennis Christian Glass entitled "Aspects of biomacromolecular dynamics at different scales." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Jeremy C. Smith, Major Professor

We have read this dissertation and recommend its acceptance:

Jerome Baudry, Xaolin Cheng, Hong Guo, Tongye Shen

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Aspects of biomacromolecular dynamics at different scales

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Dennis Christian Glass

December 2012

To my fiancée

Carolin Kern

and our son

Jonas Florian Glass

ACKNOWLEDGEMENT

I would like to express my gratitude to my advisor, Dr. Jeremy Smith, for giving me the opportunity to join his Computational Molecular Biophysics group at UT/ORNL. I specifically thank him and Dr. Jerome Baudry for their mentoring and also my other committee members, Drs. Xiaolin Cheng, Tongye Shen, and Hong Guo for their advice.

My thanks is extended to my collaborators, including, in alphabetical order of surnames, Drs. Liang Hong, Yinglong Miao, Kei Moritsugu, Marimuthu Krishnan, and Zheng Yi and to the entire CMB group for helpful discussions.

ABSTRACT

Biological functions of biomacromolecules are often indispensably linked to their internal dynamics. To investigate the dynamic nature of biomolecules, molecular dynamics (MD) simulation offers unique advantages by providing high spatial and temporal resolution over orders of magnitude in time- and length scales. Here, simulations at two different scales are used to investigate different aspects of biomolecular dynamics. At the atomistic scale, the first study investigates the relationship between the axial methyl group order parameter and the corresponding entropy in protein side chains. Three classes of methyl group are characterized based on the methyl group's "topological distance" from the backbone (that is the number of bonds between the methyl group axis and the closest backbone atom) even when direct effects of the topological distance are removed. This distinction implies that methyl groups at the same topological position share similar nonbonded environments. Furthermore, consideration of these classes of methyl group improves the accuracy of entropy-estimates based upon changes in order parameter. The second study investigates the deconstruction of crystalline cellulose, a problem relevant to bioenergy research. The large size of crystalline cellulose together with the associated long-time dynamics exceeds the capabilities of atomistic simulation. Thus, a residue-scale, coarse-grained model of cellulose is calculated using the REACH (Realistic Extension Algorithm via Covariance Hessian) method. The model is successfully validated against experiment using Young's moduli and the velocity of sound. The coarse-grained analysis of the cellulose fibril suggests that the intrinsic dynamics facilitates deconstruction of the crystalline cellulose fibril from the hydrophobic surface. Both applications share the same

concept of approach (that is, computational modeling and simulation at an appropriate scale), which reveals key insights into biomolecules by investigating their dynamic behavior.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	All-atom
AFM	Atomic force microscopy
BMRB	Biological Magnetic Resonance Data Bank
CG	Coarse-grained / coarse-graining
DP	Degree of polymerization
EM	Electron microscopy
ENM	elastic network model
FRET	Förster resonance energy transfer
HIV PR	Human immunodeficiency virus type 1 protease
ILE	Isoleucine
iRED	Isotropic Reorientational Eigenmode Dynamics
LEU	Leucine
MD	Molecular dynamics
MET	Methionine
MSF	Mean-square fluctuation
MUT	Mutant
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance spectroscopy
PDB	Protein data bank
PME	Particle Mesh Ewald
PMF	Potential of mean force
QM	Quantum mechanics
RDF	Radial distribution function
REACH	Realistic Extension Algorithm via Covariance Hessian
RMSF	Root mean-square fluctuation
THR	Threonine
VAL	Valine
WT	Wild-type
XRC	X-ray crystallography

INTRODUCTION

The present thesis investigates aspects of biomacromolecular dynamics at different scales.

A macromolecule consists of many identical or non-identical subunits and has a large mass. Important biological macromolecules (*i.e.* biomacromolecules) are nucleic acids (DNA/RNA), proteins, polysaccharides, and lipids. They have diverse biological functions, for example, to provide structural support, catalyze chemical reactions, store energy, or carry genetic information.

The traditional approach to understand the function of biomolecules is by studying their static structure [1]. More than 80,000 (August 2012) high-resolution structures of biological macromolecules such as proteins have been determined using X-ray crystallography (XRC) or nuclear magnetic resonance (NMR) studies [2], while much less is known about the internal dynamics of these molecules [3]. During the last decades, a wealth of evidence has been found for the internal dynamics of macromolecules [3-17], which is increasingly recognized as often important for function consistent with Richard P. Feynman's statement from 1963 that "...everything that living things do can be understood in terms of the jigglings and wigglings of atoms." This insight is even today underrepresented in biology textbooks, although motion is omnipresent in nature. At nonzero temperature, atoms are in motion due to *e.g.* thermal fluctuations around equilibrium positions [18, 19]. The first evidence for a more complex dynamics than these fluctuations was provided by NMR experiments, which found ring-flipping motions of aromatic protein side chains [8, 9]. Modern spectroscopy [10, 11], time-resolved XRC [12, 13], and computation [14-16] have provided evidence for protein backbone and side

chain motion on time scales ranging from picoseconds to seconds. Even very long time scales (*e.g.* milliseconds up to hours) have been probed by chemical methods such as hydrogen/deuterium exchange, revealing amplitudes of motion of up to 15 Å on the millisecond time scale [17]. These studies showed that many macromolecules are highly dynamic not only at their solvent-accessible surface, but also in their interior. In addition, dynamics has often been found as essential for function [20-23].

Various techniques are available to examine molecular motions on different time- and length scales (**Figure 1**; all figures are provided in the Appendix). While experimental techniques typically involve averaging over an ensemble of molecules within the sample, parts of each molecule, and/or time, molecular dynamics simulation has unmatched advantages in the investigation of molecular motion in that, in theory, it can provide a complete picture of the system, that is, the exact positions of the atoms in the system at any point in time. Moreover, simulations can reveal the underlying reason why a particular motion takes place because the forces and energies of each particle are calculated in the simulation [4].

Molecular dynamics simulation was introduced by Alder and Wainwright in 1957 and applied to the phase transition in a system consisting of hard spheres [24]. More realistic applications followed, for example, simulations of liquid argon [25] and water [26]. The lessons learned in these early days set the stage for the first macromolecular simulations of the protein bovine pancreatic trypsin inhibitor (BPTI) in 1977 [14] and nucleic acids in 1983 [27]. Today, thanks to numerous generally available MD programs, molecular dynamics simulations are carried out on a routine basis on time scales up to microseconds of simulated time. A few studies have already performed simulations up to the millisecond time scale using privileged access to specially

developed computational hardware [28]. Prominent success stories involve the investigation of the mechanism of folding of twelve fast-folding proteins [16, 28], explanation of the selectivity of the aquaporin water channel [29], the mechanism of voltage gating in potassium channels [30], the general microscopic interpretation of experimental data [31], and many more reviewed, for example, in reference [32].

The basic idea behind MD simulation is simple. Starting from an initial configuration usually obtained from experiment, the system's evolution in time is simulated by numerically integrating the Newtonian equations of motion of individual atoms that are initially assigned a velocity randomly pulled from the Maxwell-Boltzmann distribution at a given temperature.

Despite the high level of detail provided by simulation, experiment is indispensable. It is needed to validate simulation because the description of interactions within the system and the numerical algorithms used involve approximations – unfortunately at the expense of accuracy – to increase the computational efficiency. Still, a properly carried out, validated, and analyzed simulation can complement experiment or give insights in systems that are difficult to investigate using only experiment.

Intimately related to the capability of a simulation to reproduce experiment is the level of detail at which the system is modeled. Generally speaking, a higher level of detail promises higher accuracy and therefore is preferred if computational cost is disregarded. However, as biological processes usually involve relatively large length- and time scales of motions (**Figure 1**), a highly detailed simulation might not be practicable given the currently available computational technology. Also, simplistically waiting for increased computational power to become available

to tackle a problem is not an option, as, by that time, key problems might already have been answered due to a physically sound simplified model in accordance with Einstein's advice to "make everything as simple as possible, but not simpler". Therefore, the challenge is to find a model at a resolution that is just detailed enough to reproduce the investigated physics.

Typical levels of detail are at the electronic scale, atomistic scale, coarse-grained scale, mesoscale, or at a continuum representation (**Figure 2**). At the electronic scale that is only briefly mentioned here, quantum mechanical calculations are performed which involve electronic degrees of freedom. To speed up simulation, electronic degrees of freedom are often not treated explicitly. Instead, the system is modeled with classical physics using empirical energy functions that ideally are chosen to accurately reproduce quantum mechanical ground state energies or experimentally-determined condensed-phase properties. Further performance improvements can be obtained by condensing the explicit representation of several atoms into one coarse-grained bead. Such coarse-grained representations are possible at different scales. The united atom approach aims at incorporating the effect of hydrogen into the heavy atoms to which they are bonded and thus preserves the chemical detail of different monomers that make up the biopolymer. When less detail is required, the number of coarse-grained beads per monomer can be further reduced, for example, with one bead corresponding to one monomer. Additional simplifications may use one bead per biomolecular domain or per molecule, reaching the mesoscale. Finally, in continuum models, properties of the system become a function of a variable that depends on the coordinates. The different scales constitute a hierarchy. Ideally, a coarse-grained model is derived from a higher level of detail, so that relevant physical principles

present in highest resolution are translated along the hierarchy. Therefore, highly detailed simulation, *e.g.,* at the atomistic scale, is useful to derive a coarser model.

At any scale, the interactions within the system must be properly represented. In molecular dynamics simulation, these are described by the "force field", which is an empirical set of functions and parameters giving the potential energy of the system for a given configuration. Atomistic force fields are designed to represent accurately the interaction of atoms in typical bonded and nonbonded environments and thus preserve chemical detail. This makes atomistic force fields relatively transferrable within a class of biomacromolecules, for example, the same "protein" force field can be used for any protein but not for polysaccharides. Tremendous effort has been directed particularly toward developing atomistic force fields [33]. These have been successfully validated against many experiments. Still, refinement and validation of these force fields is an ongoing effort because, as technology advances, simulation becomes applicable to a wider range of problems, in which force fields could experience deficits unknown at present. The relatively large transferability of atomistic force fields is usually lost in coarse-grained force fields and a new coarse-grained force field must be developed for every system studied. This is because parameters become system-dependent[1], although exceptions, arguably at the expense of accuracy, exist [34].

The work reported herein has exploited the benefits of simulation for the investigation of biomolecular motion in applications on aspects of the dynamics of two systems using both

[1] Given the reduced number of parameters in coarse-grained approaches, the dependence of the parameters on the system arises from the need to include the effect of specific (*e.g*. structural) features into the parameters to model the system at sufficient accuracy.

atomistic and multiscale simulation to deliver in both cases insights that are inaccessible to experimental approaches.

The present thesis is organized as follows. The Methods section introduces molecular dynamics simulation of both atomistic and coarse-grained systems. It is followed by the two major scientific sections of this dissertation. These two sections each contain a specialized introduction, focusing on the relevance of the particular study in the respective field, specialized methods, introducing methods relevant only for the respective section, and conclusions. Finally, the last section concludes this thesis and also provides a brief outlook on the future of molecular dynamics simulation.

METHODS

The following sub-sections provide information on the methods used. The basic concepts of statistical mechanics which relate the simulations to experiments are presented. Then, the focus is set on atomistic simulation that is treated relatively general. Next, coarse-grained simulation is discussed with a bias towards its intended application in the present study.

STATISTICAL PHYSICS

Statistical physics provides the framework to relate data from the simulated trajectory quantities measurable in experiment for systems in equilibrium. The average of a quantity $A(p^{3N}, q^{3N})$ that depends on the $3N$ coordinates q and momenta p of N atoms is

$$\langle A \rangle = \iint dp^{3N} dq^{3N} \, A(p^{3N}, q^{3N}) \, \rho(p^{3N}, q^{3N}) \qquad \text{Eq. 1}$$

with the probability density $\rho(p^{3N}, q^{3N})$ being

$$\rho(p^{3N}, q^{3N}) = \frac{1}{Q}\exp\left(-\beta H(p^{3N}, q^{3N})\right). \quad \text{Eq. 2}$$

β is $(k_B T)^{-1}$, k_B the Boltzmann constant, T the temperature, H the Hamiltonian and Q the partition function which in turn is

$$Q = \iint dp^{3N} dq^{3N} \exp\left(-\beta H(p^{3N}, q^{3N})\right). \quad \text{Eq. 3}$$

The *ergodic* hypothesis implies that a system evolving in time explores the whole phase space accessible, *i.e.* it visits all possible states $\Gamma = (p^{3N}, q^{3N})$ in phase space. Thus results do not depend on the initial configuration. Therefore, the average in Eq. 1 can be calculated from a time average from an *ergodic* trajectory. A simulation trajectory that is sufficiently long can in principle visit all states Γ and thus be considered approximately *ergodic* and be used to approximate $\langle A \rangle$ (Eq. 1) as

$$\langle A \rangle = \lim_{t\to\infty} \frac{1}{\tau} \sum_{t=1}^{\tau} A\left(\Gamma(t)\right) \quad \text{Eq. 4}$$

with τ being the length of the simulation. The *ergodic* hypothesis is justified *a posteriori* due to agreement with experiment.

ATOMISTIC SIMULATION

FORCE FIELD (POTENTIAL ENERGY)

The potential energy function V is approximated as sum of pairwise interactions. Those interactions are designed to mimic typical interactions as they would arise from a more complex

treatment of the system. For example, a harmonic energy term mimics the covalent bonding between two atoms resulting from their electronic interaction. A typical functional form used in many empirical force fields defines bonded and nonbonded terms:

$$V = \underbrace{\sum V(b) + \sum V(\Theta) + \sum V(\phi) + \sum V(\omega)}_{bonded} + \underbrace{\sum V_{vdW}(r) + \sum V_{elec}(r)}_{nonbonded} \quad \text{Eq. 5}$$

BOND STRETCHING $V(b)$:

Covalent bonds between atoms i and j are described using a harmonic potential:

$$V_{ij}(b) = \frac{k_{ij}^{b}}{2}\left(b - b_{0,ij}\right)^2 \quad \text{Eq. 6}$$

where k_{ij}^{b} is the force constant, $b_{0,ij}$ is the equilibrium bond length, and b is the instantaneous bond length. The atoms i and j oscillate around $b_{0,ij}$ with frequency $\omega = \sqrt{k_{ij}/\mu_{ij}}$ (with the reduced mass $\mu_{ij} = m_i m_j/(m_i + m_j)$.

ANGLE BENDING $\mathrm{V}(\Theta)$:

The angle bending energy is also described with a harmonic potential:

$$V_{ij}(\Theta) = \frac{k_{ij}^{\Theta}}{2}\left(\Theta - \Theta_{0,ij}\right)^2 \quad \text{Eq. 7}$$

Θ is the angle between the vectors along the bonds connecting atom i to its neighbor and the neighbor to atom j (**Figure 3**). $\Theta_{0,ij}$ and k_{ij}^{Θ} are used in analogy to the above discussed bond

stretching. Forces for angle bending ($k_{ij}^{\Theta} \approx O(10^{-2})$ kcal/mol deg^2) can be four orders of magnitude smaller then for bond stretching ($k_{ij}^{\mathrm{b}} \approx O(10^{2})$ kcal/mol Å^2).

TORSIONAL TERMS ($\mathrm{V}(\Phi)$ AND $\mathrm{V}(\Omega)$):

Torsional terms involving four atoms are even softer than angle bending. See **Figure 4** for a definition of the proper dihedral angle ϕ. The proper torsional term is

$$V_{ij}(\phi) = \sum_{n} \frac{V_{n,ij}}{2}\left(1 + cos\left(n\phi - \gamma_{n,ij}\right)\right), n = 1,2,3, \ldots . \qquad \text{Eq. 8}$$

For each term in the sum, $V_{n,ij}$ is the height of the potential, n the multiplicity giving the number of minima as ϕ is rotated by 360°, and $\gamma_{n,ij}$ determines the angle of minimum energy. An improper dihedral may maintain the planarity or chirality about certain atoms as necessary [35] and is modeled as

$$V_{ij}(\omega) = \frac{k_{ij}^{\omega}}{2}\left(\omega - \omega_{0,ij}\right)^2 \qquad \text{Eq. 9}$$

with ω, $\omega_{\mathrm{o,ij}}$, and k_{ij}^{ω} in analogy to $V(b)$. See **Figure 3** for a definition of the improper dihedral angles ω.

VAN DER WAALS INTERACTION $\mathrm{V_{VDW}}(\mathrm{R})$:

Nonbonded energy terms are calculated between atoms separated by (usually) more than four covalent bonds. The van der Waals interaction is described using a Lenard-Jones potential

$$\mathrm{V_{vdW}}(\mathrm{r}) = 4\epsilon\left(\left(\frac{\sigma}{\mathrm{r}}\right)^{12} - \left(\frac{\sigma}{\mathrm{r}}\right)^{6}\right) \qquad \text{Eq. 10}$$

with the depth ϵ of the potential and the distance σ at which $V_{vdW}(\sigma) = 0$. The repulsive $+r^{-12}$ and attractive $-r^{-6}$ components mimic a hard repulsive core (Pauli's principle) and an attractive interaction due to the polarisability of the atoms, respectively.

ELECTROSTATIC POTENTIAL $V_{ELEC}(R)$:

Partially charged particles in the system interact electrostatically via Coulomb's law:

$$V_{elec}(r) = \frac{q_i q_j}{4\pi\epsilon_0 r} \qquad \text{Eq. 11}$$

where r is the distance between atoms i and j with partial charge $q_{i/j}$ and ϵ_0 is the dielectric constant.

ATOMISTIC FORCE FIELD PARAMETERS

Treating the interactions in the system not based on first principles but with a functional form derived from empirical knowledge introduces the need to wisely choose the parameters in the energy function. Selection of the parameters (e.g. bond length, partial charges) is crucial to obtain a realistic simulation for the system and a significant ongoing effort has been directed towards parameter development, resulting in diverse parameter sets, for example CHARMM [36], AMBER [37], or GROMOS [38] for proteins and CHARMM [39, 40], GLYCAM06 [41], GROMOS45a4 [42], or CSFF [43] for carbohydrates. Force field parameters are developed in an iterative fitting procedure to reproduce target data from experiment or higher levels of theory at selected target conditions (*e.g.* temperature, condensed phase). Therefore, parameters are strictly valid only for those properties and conditions they were developed for. However, transferability to other conditions (*e.g.* temperatures) is commonly assumed and needed to justify simulation in

the first place, as there is no sense in calculating only exactly those properties used in parameter development (because those are measured experimentally or calculated using a more accurate theory). Furthermore, transferability to other state points is also often assumed and may be justified *a posteriori* if agreement with experiment is obtained for relevant properties.

COARSE-GRAINED SIMULATION

INTRODUCTION

Coarse-grained (CG) simulation can only be an alternative to atomistic simulation if atomistic simulation is prohibited by resource requirements because the representation of the CG system and interactions therein are modeled in lesser detail to simplify the simulation.

Investigating a system using coarse-grained molecular dynamics (MD) simulation involves three steps. The first is the definition of the mapping from the atomisitc to the coarse-grained representation. The second step is to define the functional form of the CG potential energy function involving for example terms for pseudo-bonds, -angles, -dihedrals, van der Waals interaction, electrostatic terms, or a tabulated potential that does not follow a trivial functional form. The third step is to select the parameters in the CG potential energy function to best reproduce target properties from reference data that often originates from all-atom MD.

MAPPING FROM ATOMS TO CG-BEADS

In general, a CG-bead may correspond to any number of atoms and the intended application may further dictate the use of multiple types of CG-beads. Chemical intuition is often used to devise CG-beads in simple cases, *e.g.*, the mapping from the atomistic to the scale of residues, although

systematic approaches are available that aim at reproducing the essential dynamics as given by principle component analysis [44].

FUNCTIONAL FORM OF THE CG FORCE FIELD AND PARAMETERIZATION STRATEGIES

THE ELASTIC NETWORK MODEL (ENM)

Several functional forms may be used for the potential energy function. Of specific interest here is the elastic network model (ENM), which may be the simplest functional form used in coarse-graining. An ENM approximates the system as a number of classical masses connected by harmonic springs. A mass-point corresponds to a CG-bead. The energy function in the ENM is

$$\mathrm{E_{ENM}} = \sum_{i \neq j} E_{ij}(x) \;;\; E_{ij}(x) = \frac{1}{2} k_{ij} \left(x - x_{ij}^0\right)^2. \qquad \text{Eq. 12}$$

$\mathrm{E_{ENM}}$ is the sum over all pairwise interactions $E_{ij}(x)$ between CG-beads i and j with the force constant k_{ij} and equilibrium distance x_{ij}^0. As the $\mathrm{E_{ENM}}$ is harmonic, large anharmonic conformational changes are neglected and the model is correct only as long as fluctuations around the energy minimum are small.

The only parameters in Eq. 11 are the force constants k_{ij} and equilibrium distances x_{ij}^0. The equilibrium distances are defined by the structure that is investigated, for example, by the crystal structure. Several approaches exist for the selection of the force constants of elastic network models [45]. They can be defined to be constant, if the distance between two CG-beads is less than a cutoff (typically around 7 Å in protein applications using Gaussian network models) and zero otherwise, distance-dependent using empirical knowledge [46], dependent on the coordination number of the residue [47], different for covalently bonded than for nonbonded

interactions [48], dependent on amino acid type [49], or calculated from atomistic simulation to reproduce corresponding fluctuations [50].

In the coarse-graining work reported in the second application in this theses, the REACH ("Realistic Extension Algorithm via Covariance Hessian") method is used to develop force field parameters [50, 53-55]. The aim of REACH is to capture collective dynamics from all-atom models. REACH is a self-consistent multiscale approach that obtains elastic network model (ENM) force constants directly from the variance-covariance matrix calculated from all-atom MD as follows:

$$k_{ij} = -Tr(\boldsymbol{K_{ij}}) \qquad \text{Eq. 13}$$

where $\boldsymbol{K_{ij}}$ is the off-diagonal component of the Hessian associated with i and j. Making the harmonic approximation at constant temperature, T, allows the Hessian matrix to be calculated from the variance-covariance matrix $\boldsymbol{C} = (C_{nm}) = (\langle(r_n - \langle r_n\rangle)(r_m - \langle r_m\rangle)\rangle)$ as

$$\boldsymbol{K} = k_B T\, \boldsymbol{C^{-1}} \qquad \text{Eq. 14}$$

where k_B is the Boltzmann constant. k_{ij} is then derived by combining Eq. 13 and Eq. 14 as follows:

$$k_{ij} = k_B T\, Tr(\boldsymbol{C_{ij}^{-1}}) \qquad \text{Eq. 15}$$

OTHER FUNCTIONAL FORMS

If a non-ENM functional form is chosen for the force field, it is necessary to use different parameterization strategies. These strategies usually rely on reference data that is the target for

an iterative fitting procedure aiming at deriving a set of parameters that best reproduce the target data. Data from all-atom MD often serves as reference. Fitting this data can be done using one of multiple schemes, for example iterative Boltzmann inversion that aims at optimizing structural data from all-atom MD. Iterative Boltzmann inversion first requires a reference radial distribution function (RDF) [56, 57]. The RDF describes the probability to find two particles a and b (not necessarily of identical type) separated by the distance r relative to the distribution in an ideal gas [58]. The RDF, $g_{ab}^{ref}(r)$, from the reference simulation is inverted to yield the potential of mean force $F_0(r) = -k_B T ln[g_{ab}^{ref}(r)]$, where k_B is Boltzmann's constant and T the temperature. The potential of mean force $F_0(r)$ is not the exact pair-wise potential leading to $g_{ab}^{ref}(r)$, because many body effects from the interaction and packing of particles in the reference simulation are included in $g_{ab}^{ref}(r)$. However, $F_0(r)$ is a sufficient approximation to the target pair-wise potential to start an iterative optimization strategy. A simulation using $F_0(r)$ is carried out. The RDF, $g_{ab}^{0}(r)$, from this simulation is usually different from $g_{ab}^{ref}(r)$, and the information on this difference can be used to correct $F_0(r)$ [56, 57], which reads for the n^{th} iterative correction as follows:

$$F_{n+1}(r) = F_n(r) - k_B T ln\left(\frac{g_{ab}^{n}(r)}{g_{ab}^{ref}(r)}\right) \qquad \text{Eq. 16}$$

If more than one pair-wise potentials shall be used (e.g. van der Waals and electrostatic potentials), they, strictly speaking, cannot be obtained by performing independent iterations for each as in Eq. 16, because, in principle, the potentials depend on each other. However, the first potential can be optimized while others are kept constant and then the next is adjusted again with

all others constant. It has proven useful to optimize the potentials according to their relative strength, *i.e.*, in the order bond-, angle-, nonbonded-, and dihedral-potentials [57].

Other target properties than structure (as given by RDFs) involve fluctuations used only infrequently [59] and forces. The force matching approach performes least-squares optimization with respect to the parameters in the coarse-grained model in order to minimize the difference between the reference force data and predicted force data [60]. The initial force matching approach was has been refined to overcome problems arising due to increasing complexity when dealing with many parameters in *e.g.* systems with multiple types of atoms [61].

EQUATIONS OF MOTION

The system's evolution in in time is described by Newton's equations of motion that can be numerically integrated using finite difference methods. The integration is split into small steps δt during which forces are assumed constant. With an initial set of coordinates and momenta, $\Gamma(\mathrm{t}_0) = (p^{3N}, q^{3N})(t_0)$ at time t_0, the forces on all atoms are calculated and used to obtain $\Gamma(\mathrm{t}_0 + \delta t)$. This procedure is iterated until a trajectory of the desired length is obtained.

The Verlet algorithm [62] calculates the coordinates $\boldsymbol{r}$ at time $t + \delta t$ using the coordinates at time t and $t - \delta t$ and the acceleration $\boldsymbol{a}$ at time t as $\boldsymbol{r}(t + \delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t - \delta t) + \delta t^2 \boldsymbol{a}(t)$. Not before the calculation of the coordinates at the next point in time, $\boldsymbol{r}(t + \delta t)$, can the velocities at time t be obtained as $\boldsymbol{v}(t) = \big(\boldsymbol{r}(t + \delta t) - \boldsymbol{r}(t - \delta t)\big)/2\delta t$. This can be a technical problem in some MD algorithms. Furthermore, the algorithm is not self-starting, *i.e.*, the first set of velocities needs to be obtained differently. The Leap-Frog-Verlet algorithm [63] is used by many MD tools and provides higher accuracy. It calculates the velocities at time $t + 0.5\delta t$ from

those at $t - 0.5\delta t$ and from the acceleration: $\boldsymbol{v}(t + 0.5\delta t) = \boldsymbol{v}(t - 0.5\delta t) + \delta t\boldsymbol{a}(t)$. In a next step, coordinates are calculated as $\boldsymbol{r}(t + \delta t) = \boldsymbol{r}(t) + \delta t\boldsymbol{v}(t + 0.5\delta t)$. Apparently, velocities and coordinates are asynchronous, so that velocities at time t must be estimated as $\boldsymbol{v}(t) = 0.5\big(\boldsymbol{v}(t - 0.5\delta t) + \boldsymbol{v}(t + 0.5\delta t)\big)$. Alternative integrators are the Velocity Verlet [64] and the Beeman [65] algorithm synchronizing the calculation of $\boldsymbol{x}$, $\boldsymbol{v}$, and $\boldsymbol{a}$ or improving energy conservation, respectively [33].

WATER

Organisms evolved in and adapted to an aqueous environment for billions of years. Their machinery of life makes manifold use of the properties of water [66]. For example, water-protein interactions are important for protein folding, recognition and binding of a binding partner or catalysis [67-69]. Therefore, water must be adequately included in molecular dynamics simulation of biological molecules.

Models for water can be generally distinguished into implicit and explicit ones, the former represents water as a continuous medium while the latter treats water atoms explicitly, *i.e.*, similar to protein atoms. There are numerous explicit models of water (http://www.lsbu.ac.uk/water/models.html), differing in general aspects such as the polarizability, flexibility, usage of many-body interactions, or the number of interaction sites, as well as in detailed aspects such as the specific value of a parameter such as charge [70]. Today, most atomistic simulations are designed for use with an explicit, rigid, non-polarizable water model that uses only two-body interactions. These models are more accurate than implicit treatments and relatively good simulation performance can still be achieved. Many-body effects

are only included in a mean way in the parameterization of these models, *e.g.*, by using the condensed phase instead of the gas phase dipole moment. Furthermore, macromolecular force fields are developed together with one or few water models [36], so that, in principle, the parameters depend on those water models used during parameterization and, strictly speaking, no other water models can be used, although there is evidence for some flexibility in the choice of the water model [71, 72]. The present study uses the CHARMM protein [36] and carbohydrate [73] force fields together with the TIP3P water model [74] in atomistic simulation. The TIP3P model uses three interaction sites (one for every atom set at the exact position of the respective atom), adapts the experimental gas-phase geometry of water, uses values of the charges of $q_H = 0.417\ e$ and $q_O = -0.834\ e$, and uses a van der Waals interaction centered on the oxygen ($\sigma_{OO} = 3.536$ Å and $\epsilon_{OO} = 0.152$ kcal/mol). While inclusion of water in atomistic simulation can be considered standard, coarse-grained simulations have the tendency to not represent water in an explicit way, but include the effect of solvation in the parameters in a mean way or assume the effect as negligible at the coarse-grained resolution.

DYNAMICS OF METHYL GROUPS IN PROTEINS

This section includes modified parts of a manuscript entitled "Three Classes of Methyl Group Motion in a Globular Protein" by Dennis C. Glass, Marimuthu Krishnan, Jeremy C. Smith, and Jerome Baudry. The manuscript is to be submitted for publication to the *Journal of Physical Chemistry B*. The coauthors contributed the following to the submitted manuscript: discussions, correcting drafts of the manuscript.

INTRODUCTION

Understanding the principles of biomolecular interactions is of particular interest to biomedical applications. The overall strength of protein-ligand interactions can be quantified using the free energy, $\Delta A = \Delta H - T\Delta S$, consisting of an enthalpy, ΔH, and entropy, ΔS. The entropy can be decomposed into contributions from the ligand, the solvent, and the protein given that these contributions are independent. While much attention has been directed towards understanding the non-protein contributions to ΔS, the part of ΔS due to a change of protein entropy and in particular that due to protein conformational entropy, has only recently been found to be significant [75-80]. Protein conformational entropy can be indirectly measured using information on the abundance of different microscopic states, which can be obtained using, *e.g.*, methyl groups.

Methyl groups are abundant in proteins. They can be used as probes reporting on their microenvironment and as such inform on the global thermodynamics of biomacromolecules [80-82]. Significant changes in methyl group rotational dynamics are caused by environmental effects [83-85], specifically from changes to the methyl group's immediate microenvironment

[86-88]. In one series of applications on calmodulin using different ligands, changes in methyl group dynamics have been found useful to estimate changes in protein conformational entropy [78] and an empirically parameterized relation between changes in the dynamics of methyl groups and protein conformational entropy was developed [76, 77, 89].

Central to the characterization of methyl group dynamics is Nuclear Magnetic Resonance (NMR) spectroscopy. NMR spectroscopy describes methyl group dynamics using an order parameter, termed O^2 here. O^2 is a model free quantity and measures the angular restriction of an isotopically-labelled internuclear vector. O^2 varies between 1 indicating no motion and 0 indicating isotropic motion [90, 91]. The motion of methyl groups is decomposed into rotation of the C-H-bond, described by O^2_{rot}, and reorientation of the methyl group axis (X_1-C-bond in **Figure 6**), described by O^2_{axis}. If rotational and reorientational motions are uncoupled, the experimentally-measured order parameter is

$$O^2 = O^2_{rot} \cdot O^2_{axis}. \qquad \text{Eq. 17}$$

When assuming the methyl group geometry as constant in time and C_3-symmetric, O^2_{rot} depends only on the angle $\beta(X_1\text{-}C\text{-}H)$. Furthermore, if β is 109.47° (tetrahedral angle), O^2_{rot} is 0.111 and O^2_{axis} simplifies to $O^2/0.111$. The above approximations may in principle be not strictly true [92] and O^2_{axis} may also be influenced by motion from the part of the side chain preceding the methyl group axis [91]. Due to the influence of side chain dihedrals preceding the methyl group axis, a segregation of O^2_{axis} into distinct classes is expected depending on whether or not the preceding dihedrals undergo rotameric transitions (see **Figure 6**B) [79, 93, 94]. For example, the O^2_{axis}-distribution of ILEδ , which is influenced by rotamer transitions of χ_1 and χ_2) in principle

splits into four bands depending on whether neither χ_1 nor χ_2, only χ_1, only χ_2, or both χ_1 and χ_2 transition between rotamers [93].

Estimating entropy from order parameters has advanced our understanding of biomolecular interactions [76-79]. Typically, changes in order parameter are related to changes in entropy using a somewhat arbitrary motional model (for example the diffusion-in-a-cone model), which is a limitation because selection of a different model will in principle result in a different estimate of conformational entropy [95]. Moreover, the individual changes in entropy from methyl group probes may be correlated, but information on this correlation is not available via order parameters so that simple summation of all individual estimated changes in entropy gives only an upper limit for the cumulative change in methyl group entropy [96-98]. The arbitrary selection of a motional model is not needed when using molecular dynamics simulation to estimate O_{axis}^2 and entropy.

The present study has used atomistic molecular dynamics simulation to investigate the dynamics of methyl groups in proteins using HIV protease as model system. The exact relation between the methyl group axis entropy and order parameter and the potential influence of the nonbonded environment on this relation is investigated. Therefore, motion other than that form the methyl group has been subtracted from the simulated trajectories. Three classes of methyl group have been identified in the relation between the order parameter and entropy and in the potential of mean force for methyl group reorientation. Importantly, these classes differ from those expected from rotamer transitions [93]. Here, the classes depend on the topological position of the methyl group from the protein backbone and are suggested to arise mainly due to the average nonbonded

environment of the methyl group. This finding also suggests that methyl groups at different topological positions undergo somewhat different dynamics, which is relevant because, as a consequence, the change in entropy estimated from the change in order parameter varies with topological position.

SPECIALIZED METHODS

MODEL SYSTEM

Human immunodeficiency virus type 1 (HIV) protease (PR) is a model of choice because it is a methyl-rich, well-studied system with experimental methyl group order parameters available [99-102]. HIV PR is a homodimeric, aspartyl protease that cleaves newly synthezised viral proteins that are essential to viral maturation [103-106]. Thus, inhibition of HIV protease renders HIV uninfectious [107].

SYSTEM AND SIMULATION PROTOCOL

Atomistic molecular dynamics (MD) simulation of human immunodeficiency virus type 1 protease (HIV PR) were performed with NAMD 2.6 [108] using the CHARMM 27 all-atom force field for the protein [109] and the TIP3P water model [110]. The force field for the ligand was obtained as described later. Simulations were carried out for three different systems. Simulations of apo wild-type HIV PR used the 1.4 Å resolution X-ray structure, PDB ID 2PC0, as starting configuration [111], simulations of a mutant HIV PR bound to the inhibitor DMP323 used the 1.8 Å resolution X-Ray structure, PDB ID 1QBS [112], and the simulations of the wild-type structure bound to the inhibitor required *in-silicio* mutation of the mutant sequence (1QBS) back to the wild-type sequence (2PC0). Therefore, the modified residues, listed in **Table 1** (all

tables are placed in the Appendix), were energy-minimized to machine precision keeping the rest of the protein fixed.

The system was setup be placing the protein starting structure in a cubic box of water (76×76×76 $Å^3$) with at least 10 Å between the protein and the box edge. Water in hard contact with the protein, that is <2.4 Å, was removed. The total charge of the system was zero after adding 6 Cl^- ions to the MUT and 8 to the WT setup. For energy minimization, in a first step, the water and ions were energy minimized keeping the protein and ligand fixed, and then, in a second step, the energy of the whole system was minimized. MD simulations were performed at 310 K (37° C) and 1 atm using a Langevin thermostat and barostat with a damping coefficient of 5 ps^{-1}. Nonbonded interactions were smoothly truncated using a switching function between 10 and 12 Å. The pairlist distance was set to 14 Å. Electrostatic interactions were evaluated using the Particle Mesh Ewald algorithm switching from real-space to reciprocal space calculation at 12 Å. Bond distances and angles of the water molecules were kept rigid as defined in the TIP3P model; protein hydrogen atoms were not constrained. The equations of motion were integrated using a time step of 1 fs. Overall, 315 ns of simulation were created using the above protocol. These were distributed over 5 replicas of the apo-WT system, 5 replicas of the bound-WT system, and 5 replicas of the bound-MUT system. Each of the 15 replicas was simulated for 21 ns. Error estimates rely on the difference between the simulated replicas. Coordinates were stored every 100 fs, average velocities were calculated where needed using the difference in coordinates between consecutive frames.

To improve statistics, 1 μs additional simulation of the apo-WT system were performed, that is ten replicas to 100 ns each. GROMACS 4.5.3 [113] was used with the CHARMM 27 protein and TIP3P water force field. The PME method was used for electrostatic interactions using a real space cutoff of 12 Å, a Fourier grid spacing of 1.6 Å, and sixth-degree B-splines interpolation. Van der Waals interactions were truncated between 8 and 11 Å. The LINCS algorithm was used to constrain all bonds involving hydrogen atoms to their parameter values [114]. The integration time step was 2 fs and coordinates were written every 2 ps. Only apo-WT simulations were performed. These were equilibrated for 3 ns using the Berendsen pressure and temperature coupling algorithms with coupling time constants of τ=0.4 ps and 0.1 ps [115], followed by ten independent production simulations at 300 K, using a Parrinello-Rahman barostat [116] and a velocity-rescaling thermostat [117].

LIGAND PARAMETERIZATION

The standard CHARMM force field was missing part of the parameters needed to simulate the ligand, DMP323. These missing parameters were created as described in the publication "Three Classes of Methyl Group in a Globular Protein" by Glass *et al.* (to be submitted for publication to *JPCB*) following the parameterization protocol by MacKerell *et al.* [109, 118]. The procedure resulted in parameters reproducing the interaction energies within k_BT and the geometries within 0.4 Å heavy-atom RMSD between the structures obtained from quantum mechanics and molecular dynamics calculations.

LOCAL METHYL GROUP COORDINATE SYSTEM

The direct effect of the methyl group's position in the side chain (referenced to as "topological position", that is the number of bonds separating the methyl group axis from the backbone Cα

atom) is removed. Therefore, methyl group coordinates in the protein frame of reference are transformed into a local frame of reference by superposition of the side chain atoms preceding the methyl group as follows. The coordinate origin is defined as the X_1-atom (**Figure 7**), the x-axis is defined as being parallel to the X_1-X_2-bond, the x-y-plane is spanned by the X_1-X_2 and X_2-X_3-bonds, and the z-axis is defined to point along the positive direction of the cross-product between the vectors along the X_1-X_2 and X_2-X_3-bonds. The thus obtained local coordinate system, defined by basis vectors $\vec{e}_x$, $\vec{e}_y$, and $\vec{e}_z$, is illustrated in **Figure 7**.

METHYL GROUP ORDER PARAMETERS

METHYL GROUP ORDER PARAMETERS IN A PROTEIN FRAME OF REFERENCE

Methyl group order parameters calculated in a protein frame of reference contain, similar to experimental order parameters, contributions from all protein internal motions. Methyl group axis and backbone amide order parameters, O^2_{axis} and O^2_{NH}, are calculated using the isotropic Reorientational Eigenmode Dynamics (iRED) method [119, 120]. The elements of the isotropically averaged covariance matrix are calculated as $M_{ij} = 1/2\ \langle 3\left(\vec{\mu}_i\vec{\mu}_j\right) - 1\rangle$ with the vector μ_i ($\|\vec{\mu}_i\| = 1$) along bond i. The bracket $\langle\cdot\rangle$ denotes an average over all frames of the MD simulation. With eigenvalues λ_m and eigenvectors $|m\rangle$ from the eigenvalue problem $M|m\rangle = \lambda|m\rangle$, the order parameters can be calculated over the m' modes corresponding to internal motion: $O^2 = 1 - \sum_{m'} \lambda_m\ ||m'\rangle|^2$.

METHYL GROUP ORDER PARAMETERS IN A LOCAL FRAME OF REFERENCE

Methyl group axis order parameters, $O^2_{axis,loc}$, have been calculated in the local frame of reference using the x, y, and z coordinates of the tip of the methyl group axis (atom C in **Figure 6**) as follows:

$$O^2_{axis,loc} = \frac{3}{2}(\langle x^2\rangle^2 + \langle y^2\rangle^2 + \langle z^2\rangle^2 + 2\langle xy\rangle^2 + 2\langle xz\rangle^2 + 2\langle yz\rangle^2) - \frac{1}{2}. \qquad \text{Eq. 18}$$

RESULTS

COMPARISON BETWEEN EXPERIMENTAL AND SIMULATED AXIS ORDER PARAMETERS

Simulations are validated using the root mean-square fluctuation (RMSF) and the backbone amide and methyl group order parameters. **Figure 7**A shows the RMSF for the different simulations of HIV protease (apo/bound wild-type/mutant). As expected, the RMSF is small (~0.7 Å) in structurally ordered regions (for example residues 22-32), somewhat increased (up to 1.5 Å) at the termini and in solvent-exposed loops (for example around residue 17, 39, or 80), and large (~2.5 Å) in the flaps of apo-HIV protease [100, 101]. **Figure 7**B shows the backbone amide order parameter, O^2_{NH}. Small values of O^2_{NH} indicate angular flexibility of the corresponding backbone amide, and such values are found mostly for the regions that have large values of the RMSF confirming the above analysis. Unlike the RMSF, the order parameter can be measured by NMR. Values of O^2_{NH} compare favorably to experiment [121] (**Figure 7**C). Only the flexibility in the flaps of HIV protease appears overestimated in the simulations.

Figure 8 compares side chain order parameters, O^2_{axis}, from simulation with those from experiment [122]. Within error margins, near-quantitative agreement is obtained for many residues. Differences between simulated and experimental values of O^2_{axis} have been previously

attributed in part to current force field parameters that are not accurate enough to fully reproduce the detailed dynamics of protein side chains and in part to incomplete sampling of the motion of protein side chains [93, 123]. The order parameter $O^2_{axis,loc}$, calculated in a local methyl group frame of reference, shows a strong correlation with O^2_{axis}, calculated using iRED (**Figure 8**, right) [119, 120]. This indicates that, in the present simulation, most of the contributions to O^2_{axis} are from the motion of the methyl group axis itself and not from other parts of the protein.

ASSUMPTIONS WHEN OBTAINING METHYL GROUP AXIS ORDER PARAMETERS

Figure 9A shows $O^2 = O^2_{rot}O^2_{axis}$ against $O^{2\prime} = 0.111O^2_{axis}$ and thus probes the approximation $O^2_{rot} = 0.111$. If O^2_{rot} were 0.111, the data would lie on the diagonal ($f(x) = x$). However, this is not the case in **Figure 9**A. Instead, the function $g(x) = mx$ ($m = 0.91$) is the best linear fit to the data indicating that O^2_{rot} is not exactly 0.111. The probability distribution (not shown) of the calculated O^2_{rot} varies mainly between 0.090 and 0.114 with an average of $\overline{O^2_{rot}} = 0.099 \pm 0.005$ (~89% of 0.111). The order parameter O^2_{rot} can be also estimated from the X_1–C–H angle β if constant methyl group geometry and 3-fold symmetry is assumed. $O^{2\prime}_{rot}(\bar{\beta})$ is 0.093 with $\bar{\beta} = 111.1° \pm 4.3°$ [65]. Note that O^2_{rot} is sensitive to variations of β, with a 1° deviation from tetrahedral geometry ($\beta = 109.47°$) of the methyl rotator translating into a 9% variation of $O^{2\prime}_{rot}(\beta)$. The values obtained for O^2_{rot} are in agreement with various experiments (neutron crystallography [124, 125], gas electron diffraction [126], infrared absorption [127], NMR [128]) and simulations (quantum mechanical [129, 130] and MD calculations [131, 132]) on systems in the condensed or gas phase which have found β to be in the interval $110° - 113°$. **Figure 9**B plots the overall O^2 against $O^2_{rot}O^2_{axis}$. A linear fit $h(x) = mx$ to the data yields a slope of

approximately 1 indicating that the assumption $O^2 = O^2_{rot} O^2_{axis}$ is valid and that the relaxations described by O^2_{rot} and O^2_{axis} are uncoupled.

Finally, O^2_{rot} may be different before and after ligand binding while it is assumed as constant. This difference can cause potentially relevant changes in entropy impacting the ligand's binding affinity. **Figure 9** shows the rotational entropy for every methyl group from (C) apo and ligand-bound and (D) wild-type and mutant simulations of HIV PR. Both datasets scatter around the diagonals, except for two points (one per monomer of HIV PR) in D, which is explained by the mutation $V3I$ changing the type of a methyl group from $VAL\ \gamma 2$ to $ILE\ \delta$. The values of S_{rot} are directly related to those of O^2_{rot} [95], thus, **Figure 9**C and D show, admittedly in an indirect way (due to practical reasons), that the values of O^2_{rot} remain unchanged upon changes to the methyl group's nonbonded microenvironment arising from ligand binding or mutations because the data scatters around the diagonal in both cases C and D. Overall, these results suggest that the experimentally-determined order parameter O^2 may be best related to O^2_{axis} by assuming $O^2_{rot} = 0.099$ while no correction is necessary for a potential coupling between the relaxations described by O^2_{rot} and O^2_{axis} and also no correction is necessary for potentially different O^2_{rot} before and after ligand binding.

INFLUENCE OF ABOVE ASSUMPTIONS ON ESTIMATES OF METHYL GROUP AXIS ENTROPY

Order parameters can be related to entropy using a motional model, for example, the diffusion-in-a-cone model [95]

$$S_{axis} = k_B T ln\left(\pi\left(3 - \sqrt{1 + 8\sqrt{O_{axis}^2}}\right)\right) \quad \text{Eq. 19}$$

where k_B is Boltzmann's constant and T is the temperature. The relation between order parameter and entropy is non-linear [95], so that, two changes of order parameter that are of equal values $\Delta O_{axis}^2(1)$ and $\Delta O_{axis}^2(2)$ $[\Delta O_{axis}^2(1) = \Delta O_{axis}^2(2)]$, but take place around *e.g.* small and large values of order parameter, will lead to two different changes in entropy $\Delta S(1)$ and $\Delta S(2)$ $[\Delta S(1) \neq \Delta S(2)]$. This situation is apparent from the diffusion in a cone relation shown in **Figure 10** where $\Delta S(1)$ (corresponding to $\Delta O_{axis}^2(1)$ from $O_{axis}^2 = 0.1$ to 0.2) is different from $\Delta S(2)$ (corresponding to $\Delta O_{axis}^2(2)$ from $O_{axis}^2 = 0.8$ to 0.9). Using $O_{rot}^2 = 0.099$ instead of 0.111 to obtain a better estimate of the order parameter O_{axis}^2 has a two-fold impact, first, the value of ΔO_{axis}^2 is slightly changed, and second, O_{axis}^2 is scaled to a larger value (see numerical example in **Figure 10**). Both changes influence also the corresponding change in entropy. The numerical example in **Figure 10** depicts a ligand binding scenario with a change in order parameter, $\Delta O_{axis}^2(0.111)$, of 0.2 and a corresponding change in entropy, $\Delta S(0.111)$. However, if $O_{rot}^2 = 0.099$ is used to estimate O_{axis}^2, the updated change in order parameter, $\Delta O_{axis}^2(0.099)$, results in a different change in entropy, $\Delta S(0.099)$, and the two changes in entropy differ by a $\Delta\Delta S$ of approximately 0.1 kcal/mol (at 310 K).

To quantify $\Delta\Delta S$ for situations involving a real protein binding a ligand, three systems have been considered where both apo and ligand-bound order parameters have been reported: HIV PR with DMP323 [133], calcium saturated calmodulin with various ligands (Biological Magnetic Resonance Data Bank (BMRB) entries 4970, 15183, 15184, 15185, 15187, 15188, and 15191)

[78, 79], and barnase with barstar bound (BMRB entries 7139 and 7126) [134]. Only methyl groups are considered where both apo and ligand-bound values of O^2_{axis} have been reported and where the change of O^2_{axis} upon ligand-binding has been larger than 0.05 and significant (1σ) within the reported error, leaving 9, 19, and 146 methyl groups for HIV PR, barnase, and various ligand-bound calmodulin systems, respectively. Furthermore, only O^2_{axis} smaller than 0.95 are considered to avoid the potentially artificial (model-dependent) effect of the strong variation of entropy with order parameter for values of O^2_{axis} close to 1.0, leaving 145 methyl groups in total. Averages of the remaining order parameters are given in **Table 2**. The average O^2_{axis} increases from 0.43 to 0.53 upon ligand binding. Using the diffusion-in-a-cone model to estimate entropy from O^2_{axis} gives an average $\Delta\Delta S = -0.05$ kcal/mol per methyl group (**Table 2**). The cumulative $\Delta\Delta S_{tot}$ for HIV PR is 0.45 kcal/mol using only those 9 out of 41 reported methyl groups with significant changes in O^2_{axis} and assuming that their motion is independent from each other. Extrapolation of the HIV PR data from those methyl groups investigated in experiment to all of them leads to $T\Delta\Delta S_{all} \approx 1.6$ kcal/mol. This significant value of $T\Delta\Delta S_{all}$ indicates that it is important to assume an optimal value of O^2_{rot}, chosen either from simulation or experiment [124-132]. Furthermore, the average $\Delta\Delta S$ per methyl group depends on the system investigated. For example, $\Delta\Delta S = 0.13$ kcal/mol in the case of Barnease is a factor of 2.9 larger than for HIV PR (**Table 2**).

METHYL GROUP AXIS ENTROPY

The methyl group axis entropy is calculated as

$$S_X = -k_B T \int \rho(\vec{\mu}) \ln(\rho(\vec{\mu})) d\vec{\mu} \quad \text{Eq. 20}$$

where k_B is the Boltzmann constant, T is the temperature (310 K), and ρ is the probability distribution of the normalized methyl group axis vector $\vec{\mu}$ in the Cartesian coordinate system local to the methyl group. The probability distribution was obtained by binning the vectors $\vec{\mu}$ using a bin width of 0.1 Å.

As expected, the entropy decreases with order parameter (**Figure 11**). Three entropic classes of methyl group have been found differing in the number of bonds that separate them from the backbone. Class 1 consists of VAL$\gamma_{1,2}$, THRγ, and ILEγ, Class 2 of ILEδ and LEU$\delta_{1,2}$, and Class 3 of METϵ. The entropy corresponding to an order parameter increases with distance from the backbone from Class 1 to 2 to 3. This shows that a given value of the order parameter can correspond to different amounts of disorder or entropy. Furthermore, the data in the quasi-linear regime from 0.5 to 0.95 was fit with a linear function, $f_i(x) = m_i * x + c_i$, for classes i. The fitted slopes are -0.82 and -0.96 kcal/mol for classes 1 and 2, respectively. The ratio m_2/m_1 is approximately 1.17, suggesting that methyl groups in different classes yield different changes in entropy from a change in order parameter. As the order parameter calculated in the local coordinate system that included only methyl group axis motion ($O^2_{axis,loc}$) showed a strong correlation with the order parameter calculated in the protein frame of reference using iRED (O^2_{axis}), the above conclusion will also hold for the latter order parameter, which is the equivalent to the experimentally obtained one.

The existence of different classes is further investigated using the potential of mean force (PMF), $f(\chi) = -k_B T ln[\rho(\chi)]$, where k_B is the Boltzmann constant, T is the temperature, and $\rho(\chi)$ is

the ‘single-well’ probability distribution of the dihedral angle that influences the methyl group axis orientation, where ‘single-well’ means that the three dihedral angle energy minima are superimposed by overlapping the intervals $\chi = 0 - 120°$, $120 - 240°$, and $240 - 360°$. The inset in **Figure 12**A shows a probability distribution of MET that is well-sampled. For other types of methyl groups, the distribution is often narrower with insufficiently sampled wings prohibiting estimates of barriers to rotamer transitions. Therefore, only the width around the minimum of the PMF is characterized in terms of $\Delta\chi = \chi^+ - \chi^-$, where χ^+ and χ^- are the dihedral angles at which the PMF is 1 kcal/mol larger than at its minimum (compare **Figure 12**A). **Figure 12**B shows in good approximation that $\Delta\chi$ decreases with order parameter. Again, three classes are apparent differing in terms of their average values of $\Delta\chi$, *i.e.*, in the shape of the PMF around its energy-minimum.

CONCLUSIONS - DYNAMICS OF METHYL GROUPS IN PROTEINS

The above application of atomistic molecular dynamics simulation on methyl groups in proteins revealed three classes of methyl groups depending on the methyl group’s topological distance from the backbone. Class 1 consists of VAL$\gamma_{1,2}$, THRγ, and ILEγ, Class 2 of ILEδ and LEU$\delta_{1,2}$, and Class 3 of METϵ. The classes have been found in the relation between the methyl group axis order parameter and the corresponding entropy.

Interestingly, the classes are not only caused by the dihedral energy force field parameters controlling the orientation of the methyl group axis because the barrier to rotation in the force field is 3.6 kcal/mol for both Classes 1 and 2. Owing to the different chemical nature of MET (*i.e.* MET has a sulfur atom at the foot of the methyl group axis) the barrier for Class 3 (MET) is

1.86 kcal/mol, which explains its relatively large entropy (**Figure 11**). The similarity of the dihedral energy parameter and the simultaneous difference in axis entropy combined with the fact that the analysis was done in the local methyl group frame of reference, in which motion other than that of the methyl group axis is absent, suggests that the differences between the classes of methyl group at different topological distances from the backbone are caused by differences in the average nonbonded environment. Indeed characteristics of the potential of mean force close to the energy minimum differ between classes 1 and 2; the PMF for Class 2 appears more smeared out than for Class 1 motivating its increased entropy in **Figure 11**.

Hence, methyl groups on the same side chain are not necessarily similar. Instead, methyl groups share similar properties if they are at the same topological distance from the backbone. This also suggests that a change in entropy obtained from a change in order parameter will depend on topological distance, in agreement with the slopes fitted to the relationship between entropy and order parameter based on the data shown in **Figure 11**.

DYNAMICS OF CRYSTALLINE CELLULOSE

This section includes modified parts of a manuscript by the author published in *Biomacromolecules* (doi: 10.1021/bm300460f). The coauthors contributed the following to the afore-cited manuscript: discussions, correcting drafts of the manuscript. Furthermore, Dr. Kei Moritsugu made available source code for the calculation of the coarse-grained force field and dispersion relations.

INTRODUCTION

BIOMASS BASED BIOFUELS

The world was dependent on biomass to meet its energy demand until the first systematic exploration and drilling for crude oil began in the 19th century. Crude oil, as inexpensive energy source, powered the industrial revolution and promoted the growth of wealth. Consumption of oil increased subsequently because more and more economies became dependent on it. However, crude oil is a finite resource and the maximum level of global oil extraction may have already been passed (reference [135] and references therein). Together with the negative impact of fossil fuels on the environment, it became imperative to develop sustainable alternatives. Biomass may be such an alternative. Currently, organic carbon and liquid transportation fuels can only be sustainably derived from plant-biomass [136].

The production of first generation crop-based biofuels (e.g. sugarcane ethanol, rapeseed biodiesel) is commercialized today, but still accounts for only few percent of the consumption in transportation fuel [137]. The unsustainability of gasoline-fuel raises the need to scale up sustainable biofuel production. The social and environmental impact of any biofuel production

process must be carefully evaluated. Cultivation of biofuel-crops replaces food-crops so that food prices rise increasing the risk of food riots as happened in the developing world in December 2007. An expected change in land-use, *i.e*., the conversion of forest or grassland to farmland, was predicted to significantly increase overall greenhouse gas emissions [138] although first generation biofuels initially appeared to reduce greenhouse gas emission due to the uptake of carbon during plant growth [139].

Second generation biofuels have the potential to take center-stage. It was estimated that the U.S. could still meet its food and export demands and sustainably produce $1.3\ 10^9$ metric tons of dry biomass per year ($3.8\ 10^9$ barrels of oil energy equivalent) [140], which would account for $\sim 50\%$ of the current U.S. liquid fuel consumption ($7.3\ 10^9$ barrels of oil per year [141]).

Second generation biofuels are made from non-food components of biomass, which mainly consists of cheap lignocellulosic biomass that is available in abundance. Biomass can be obtained from specifically grown energy crops (*e.g.* on marginally fertile land), aquatic biomass (*e.g.* algae), or waste (*e.g.* agricultural, forest). The composition of the biomass varies by source (Table 3). The three main components are cellulose, hemicellulose, and lignin accounting for roughly 40%, 20%, and 20% of the dry weight of woody biomass, respectively [136, 142, 143]. The polysaccharide cellulose is a polymerized, unbranched glucose. Glucose can be easily fermented to biofuel, *e.g*. ethanol. Hemicelluloses consist of a variety of different sugar monomers that form a branched polymer and lignin is an aromatic, branched heteropolymer. All of those components can be converted into biofuels with varying efficiency [136]. However, the plant evolved to resist degradation and holds on tightly to its components. Cellulose exists to a

large fraction in crystalline forms, in which the sugar polymers strongly interact with each other to form a stable microfibril (Figure 13). These microfibrils are embedded in a matrix of, among others, hemicellulose that is further shielded with lignin (Figure 13) against cellulose degradation through *e.g.* bacteria and fungi.

To efficiently produce biofuel, cellulose and hemicellulose must be made accessible and broken into their monomeric units for further processing [144]. Therefore, breaking down the protection conferred by lignin and the crystallinity of cellulose is the primary goal of various pretreatment methods, which include mechanical and chemical treatment, steam explosion, ammonia fiber explosion and biological treatment [144]. Furthermore, an ideal pretreatment method minimizes inhibitory products that could adversely affect subsequent processing steps that involve the hydrolysis of the polymers into sugars using enzymes or acids.

Pretreatment is the economically most expensive processing step and a detailed molecular understanding of all biomass components, especially of cellulose, is desirable to further improve pretreatment efficiency. Herein, the focus is on cellulose.

CELLULOSE

Carbohydrates are a large group of organic compounds that consist only of carbon, oxygen and hydrogen. Carbohydrate is often used synonymous to saccharide. These are distinguished into mono-, di-, oligo-, and polysaccharides. Examples for these types are *e.g.* glucose ("grape sugar"), lactose ("milk sugar"), rafinose, and cellulose. Disaccharides are two linked monosaccharides. Oligo- and polysaccharides cannot be strictly distinguished, but a common

definition requires them to have three to ten and more than ten constituent sugar monosaccharides, respectively.

Cellulose is a polysaccharide of cellobiose, a disaccharide of D-glucopyranose. The chemical composition of cellulose is $C_6H_{11}O_5 - (C_6H_{10}O_5)_{n-2} - C_6H_{11}O_5$ with the degree of polymerization n. Glucose units are linked together by an β-1,4 glycosidic bond under the loss of one water molecule. The conformation of a glucose ring is non-planar and resembles a "chair" [145]. Successive glucose units are rotated by 180° around the approximately straight polymer axis relative to one another. The typical degree of polymerization for one molecule of cellulose ranges between 100 and 20,000 glucose monomers [146]. The top and bottom of each molecule is mainly hydrophobic while the sides are capable of hydrogen bonding [136].

Approximately 36 molecules of cellulose associate to form an elementary cellulose microfibril with linear cross-section dimensions ranging between 2 and 5 nm. The chains were proposed to interact via their hydrophobic and hydrophilic sides after being co-synthesized. In such a microfibril, a single molecule of cellulose is referred to as "chain" (atoms that are directly or indirectly connected by covalent bonds). Chains, whose sugar rings reside in the same plane, are referred to as "sheet" and sheets are stacked above one another to form the fibril.

POLYMORPHS

Cellulose exists in several polymorphs, I, II, III and IV, whose relative abundance depends on source and the history of treatments to the sample [147]. Cellulose I is predominant in nature. It can be further separated into cellulose $I\alpha$ that is prevalent in bacteria and algae, and cellulose $I\beta$ that is the main form in plants [148, 149]. Since cellulose $I\alpha$ can be irreversibly converted into

$I\beta$ by heat treatment, cellulose $I\beta$ was suggested to be the more stable form of Cellulose *I* [150, 151]. Cellulose *I* can be converted to cellulose *II* by alkali treatment and subsequent washing, or in some cases, may be directly biosynthesized depending on temperature [152] or mutations present in the bacteria studied [153]. Treatment with ammonia converts cellulose *I* and *II* into cellulose III_I and III_{II}, and treatment with temperatures larger than 240℃ can convert cellulose *I* and *II* into cellulose IV_I and IV_{II}, respectively [147]. Furthermore, a significant fraction of cellulose is amorphous.

Detailed crystallographic information is available for cellulose $I\alpha$, $I\beta$, II, and III_I, in part from combined neutron and X-ray diffraction experiments that solved both the hydrogen and heavy-atom positions and thus allowed conclusions on the hydrogen bonding pattern [154-157]. The polymorphs differ in their unit cells [154-157]. Cellulose $I\alpha$ and III_I have one chain in their triclinic and monoclinic unit cells, respectively. The monoclinic unit cells of cellulose $I\beta$ and II, with two chains per cell, in principle allow for antiparallel chain packing. Such chain packing involves alternating reducing and non-reducing chain ends at each tip of the fibril. As cellulose $I\beta$ and II can be interconverted by what is called mercerization, it was initially controversial that the two polymorphs could have different chain packing, until high resolution diffraction studies provided strong evidence for antiparallel packing in cellulose *II* [147].

HYDROGEN BONDS

The recalcitrance of cellulose towards hydrolysis is in part explained by a robust and strong hydrogen bonding network in cellulose. Hydrogen bonds can be classified terms of in-chain, interchain, and intersheet hydrogen bonds. In-chain hydrogen bonds connect two atoms within

the same chain and increase the stiffness of a single chain, interchain hydrogen bonds connect atoms that are within the same sheet, but not within the same chain, and intersheet hydrogen bonds connect atoms between sheets. The interchain and intersheet hydrogen bonds connect chains and thus stabilize the overall structure of the fibril. The chain arrangement in the polymorphs determines the possibilities for hydrogen bonding. In cellulose $I\alpha$ and $I\beta$, there are two kinds of in-chain and interchain hydrogen bonds and no intersheet hydrogen bonds of type $O \cdots H - O$ [154, 157]. Thus, the sheets are held together mainly by hydrophobic interactions, although some relatively weak intersheet $O \cdots H - C$ interactions may also be formed [154]. Besides in-chain hydrogen bonds, only intersheet and hardly any in-sheet hydrogen bonds have been found in cellulose II and III_I [156]. A more detailed comparison of the hydrogen bond networks in cellulose is available in reference [156]. Considering cellulose degradation, it is noteworthy that enzymatic sacchrification rates are increases 5-fold for cellulose III over $I\beta$ due to what was described as "amorphous-like" nature of the surface of cellulose III [158]. Destabilization of the hydrogen bond network suggests itself as straightforward route to facilitate cellulose deconstruction, however, the hydrogen bond network is particularly stable and can resist heat treatment until $\sim$220℃ in cellulose I [151, 159-161].

The strong hydrogen bonding interaction in cellulose is one of several factors explaining the recalcitrance of cellulose towards hydrolysis. Equally important might be the assembly of the plant cell wall with its other components that protect cellulose from being hydrolyzed.

PLANT CELL WALL

The primary plant cell wall is often described as having a fiberglass-like structure [162, 163] built from polysaccharides that are traditionally classified as cellulose and the matrix polysaccharides hemicellulose and pectin.

Cellulose is the only well-structured polysaccharide in the primary plant cell wall. Cellulose is synthesized, likely due to its insolubility, in the plasma membrane. Cellulose synthesizing proteins form hexameric "rosettes", as observed by electron microscopy [164]. Possibly six rosettes are suggested to assemble into arrays co-localizing 36 CesA proteins, each of which synthesizes a chain of cellulose [165] that then interact to form the cellulose microfibril.

Hemicelluloses are typically grouped into xyloglucan, xylans, mannans and glucomannans and have been recently reviewed in detail in reference [166]. Most hemicelluloses share a similar $\beta, 1-4$ linked backbone with cellulose, but, in contrast to cellulose, can be branched. Hemicellulose and other matrix polysaccharides are synthesized in the Golgi apparatus and secreted at the cell wall. Then, they can diffuse into the cell wall [167] driven by a pressure gradient [168]. Newly secreted fragments have been proposed to be integrated into the existing matrix by enzymes [169]. Several models were proposed to explain the cross-linking of polysaccharide chains. For example, hemicelluloses could interact with cellulose via hydrogen-bonds, hemicelluloses could be entrapped during the synthesis of the cellulose microfibril, or they may be covalently cross-linked to other wall-polysaccharides (reference [163] and references therein).

Pectins are a heterogeneous third class of primary plant cell wall polysaccharides also forming a cross-linked network. Due to their relatively low fraction of the dry-weight of biomass, pectins are of limited importance for bioenergy applications.

Another abundant component of woody biomass is lignin. Lignin is an integral part of the secondary cell wall, which is established in some cell types after the cell has stopped expanding. Lignin is a branched and cross-linked phenolic biopolymer formed by various monomers and linkages [170]. Lignin confers mechanical strength and is crucially involved in water transport. Furthermore, lignin's indigestibility confers resistance towards pathogens such as bacteria and fungi.

Unfortunately, lignin is also an important factor for biomass recalcitrance. After pretreatment of biomass, lignin precipitates back on cellulose posing a physical obstacle for the action of cellulases, the enzymes that hydrolyze cellulose [171-173]. Furthermore, lignin nonspecifically binds cellulases decreasing the concentration of active enzymes [174, 175].

PREVIOUS ATOMISTIC MD STUDIES OF CELLULOSE

After the hydrogen bond network was established for the cellulose polymorphs using high resolution neutron and X-ray crystallographic structures, simulation studies on the cellulose polymorphs have been undertaken. Initial studies often simulated a periodic cellulose crystal, where the crystal is quasi-infinitely extended due to periodic boundary conditions, often along all three base vectors. Such studies investigated, for example, the thermal response of cellulose lattice parameters, the density, or elastic properties [161, 176]. However, to avoid potentially artificial effects of the periodic boundary conditions, simulations of fully solvated cellulose

crystals are preferable as large-scale simulations of cellulose have become feasible [177]. Fibril twist is consistently observed in simulations of cellulose $I\alpha$ and $I\beta$ but not in cellulose III_I, consistent with the 2-dimensional and 3-dimensional hydrogen bond network of the polymorphs, respectively [178-180]. Conclusive experimental proof for fibril twist is not yet available. Also the hydrogen bond network received much attention. One study clarified an ambiguity considering the exact hydrogen bonding scheme in cellulose $I\beta$ [181], others investigated the hydrogen bonding in terms of in-chain, interchain, and intersheet hydrogen bonds in cellulose solvated in water [182] or in ammonia [183]. Also, aspects of surface solvation by *e.g.* water, benzene, or ionic liquids raised interest [182, 184, 185]. Furthermore, the interactions of the Carbohydrate-Binding Module and an entire cellulase with the cellulose surface were investigated [186-189]. Targeted towards bioenergy research have been studies on the free energy profile of chain removal from the fibril in water [190] and ionic liquids [191]. Other aspects covered are for example the structural conversion of cellulose III_I to cellulose I [192] or amorphous cellulose [193, 194].

PREVIOUS COARSE-GRAINED STUDIES OF CARBOHYDRATES

A first coarse-grained model of carbohydrates used structural and thermodynamic data to derive a CG model for gas phase and solvated $\alpha(1 \rightarrow 4)\, D-glucans$, representing each sugar monomer by three and each water by one CG-beads [195]. Bonded interactions have been derived using Boltzmann inversion and nonbonded interactions have been modeled with a Morse potential. The model successfully reproduces excluded volume interactions, the distribution of torsion angles, or the glass transition temperature of hydrated and dry samples. A subsequent approach systematically derived a three-site CG-model for α-D-glucopyranose and for a short

chain of cellulose (DP=14) demonstrating the feasibility and transferability of the multiscale force matching method used [196]. Later on, crystalline cellulose was coarse-grained to study the interactions of the carbohydrate-binding module of *Trichoderma reesei* cellobiohydrolase *I* with the fibril's hydrophobic face and predicted stable positions of the CBM approximately every 5 and 10 Å along the fibril, consistent with the repeat of sugar monomers and cellobiose, respectively [197]. Later on, a coarse-grained model for cellulose was derived compatible with the family of MARTINI coarse-grained force fields. In principle, this makes possible the simulation of combined protein-cellulose and other composite systems [198]. However, the initial parameterization based on partitioning free energies in water and cyclohexane did not yield the correct crystal structure of cellulose $I\beta$ and an additional interaction was added to overcome this problem at the expense of the strict compatibility with other MARTINE-style force fields. A restraint-free coarse-grained model of cellulose $I\beta$ has also been published using different parameters for the center and origin chains [199]. Furthermore, the model can be continuously switched to various degrees of amorphousness in terms of a coupling parameter λ.

THE PRESENT COARSE-GRAINED STUDY

The large size of cellulose together with the associated long-time dynamics pushes simulation of cellulose beyond the capabilities of atomistic MD and coarse-graining must be further followed. The above coarse-graining approaches have derived their force field parameters with an (often) iterative fitting procedure to reproduce target data. The somewhat arbitrary choices involved make these CG force fields less suitable for consistent multiscale approaches. Furthermore, part of the physical origin of atomistic force fields is arguably lost in the above CG strategies. Previous work introduced the REACH (Realistic Extension Algorithm via Covariance Hessian)

methodology that calculates elastic network model force constants from the variance-covariance matrix obtained using atomistic MD [50, 53-55]. REACH translates the information on collective dynamics contained in the atomistic variance-covariance matrix to the coarse-grained scale. Furthermore, REACH is a direct, one-step mapping with no iterative fitting and no need for experimental input data, which makes REACH particularly suitable for automated multiscale approaches. REACH was already successfully applied to coarse-graining various classes of proteins [50, 53-55].

In the present study, a REACH force field is developed for crystalline Iβ cellulose. This allows the characterization of elastic properties of large cellulosic fibrils and investigation of some aspects of cellulose deconstruction. The REACH model is calculated for a cellulose fibril in water at temperatures from 100 to 500 K in steps of 50 K and successfully validated against experiment. Analysis of the normal modes of motion at different surfaces of cellulose suggests that crystalline cellulose might be deconstructed from the hydrophobic surface.

SPECIALIZED METHODS

LATTICE DYNAMICS AND NORMAL MODE ANALYSIS

References [200, 201] describe the calculation of phonon dispersion relations in detail. A brief summary is given in what follows. The equation of motion for the displacement of residue n in unit cell α, $r_{\alpha n}$, is

$$m_n \ddot{r}_{\alpha n} + \sum_{\beta m} V_{\alpha n}^{\beta m} r_{\beta m} = 0 \qquad \text{Eq. 21}$$

where $V_{\alpha n}^{\beta m}$ is the force constant between residue n in unit cell α and residue m in unit cell β. $V_{\alpha n}^{\beta m}$ can be calculated from the REACH model function using the interatomic distance (Eq. 23). With the plane wave approach, $r_{\alpha n} = m_n^{-1/2}\, u_n(q) exp[i(qr_\alpha - \omega t)]$, this linear equation is derived:

$$-\omega^2(q)u_n(q) + \sum_m D_n^m(q)u_m(q) = 0 \qquad \text{Eq. 22}$$

where the dynamical matrix is $D_n^m(q) = \sum_\beta V_{\alpha n}^{\beta m} exp\left[iq\left(r_\beta - r_\alpha\right)\right] / \sqrt{m_n m_m}$. For every q, the dynamical matrix can be diagonalized leading to the dispersion relations $\omega(q)$.

Here, the dispersion relations were calculated for a 1-dimensional lattice (1-dimensional in the longitudinal fibril direction [0,0,1]) and at 41 points of the reduced wave number, q_{hkl} ($= q/(2\pi d_{hkl})$). This corresponds to sampling q in the range from 0 to 0.5 every $\Delta q = 0.0125$. The sound velocity is obtained as the gradient of $\omega(q)$ in the limit $q \to 0$, that is, in the range $q_{hkl} \leq 0.05$. Calculations were performed for a 36 chain cellulose fibril ($DP = 80$) using the X-ray crystal structure from reference [154].

ATOMISTIC SYSTEM AND SIMULATION PROTOCOL

The cellulose fibril starting structure consisted of 36 chains of β-1,4 linked D-glucopyranose monomers similar to that from Schultz and co-workers [202, 203], but shorter in length with a degree of polymerization (DP) of 40. The cellulose fibril was solvated in 28,367 water molecules, leaving a distance larger than 1 nm between the fibril and the edge of the simulation box. Hard contacts (<2.0 Å) between water molecules and cellulose were removed. The final

simulation system had dimensions of approximately 70×70×227 $Å^3$ and the number of atoms totaled to 115,509. After solvation, the water was energy minimized while cellulose atoms were fixed using harmonic restraints. Then, the full system was energy minimized.

Atomistic molecular dynamics simulations were carried out using the MD software GROMACS 4.5.3 [113] together with the CHARMM cellulose force field developed by Guvench et al. [73]. Water was represented by the TIP3P [74] model. The Particle Mesh Ewald (PME) electrostatics were evaluated using a real space cutoff of 12 Å. A Fourier grid spacing of 1.6 Å together with sixth-degree B-splines interpolation was used for better parallel performance. Van der Waals interactions were truncated between 8 and 11 Å using a switch function. All bonds involving hydrogen atoms were treated as rigid by constraining their bond lengths to equilibrium values using the LINCS method [114]. The equations of motion were integrated using a time step of 2 fs. The trajectory was recorded every 1 ps.

NPT simulations were performed at temparatures between 100 and 500 K in steps of 50 K and at atmospheric pressure. The Berendsen temperature and pressure coupling algorithms with coupling time constants of τ=0.1 ps and τ=0.4 ps were used during the initial equilibration of the system (1.5 ns) [115]. This was followed by additional 10 ns equilibration and 20 ns production dynamics using a Parrinello-Rahman barostat [116] and a velocity-rescaling thermostat [117] at each temperature.

COARSE-GRAINED SYSTEM AND SIMULATION PROTOCOL

In the coarse-grained simulations, one sugar monomer was represented by one CG-bead. The mass of one CG-bead was set accordingly to that of a sugar monomer. GROMACS was used to

perform coarse-grained MD simulations with the REACH force field derived in this study. The amount of force constants, proportional to the square of the number of CG-beads, rendered several tools, whcih are essential to GROMACS, useless, so that those force constants with values smaller than 10^{-3} kJ/mol nm^2 were set to zero, *i.e.*, they were not represented by a harmonic potential. The accompanying decrease in the number of interaction pairs circumvented previous issues with GROMACS tools and resulted in negligible changes in benchmark simulations. The velocity-rescale temperature coupling algorithm with a coupling time constant of τ=0.1 ps was used for simulations. These were performed using an integration time step of 2 fs and writing coordinates every 1 ps. A total of 100 ns simulation was carried out at each temperature.

RESULTS

REACH FORCE FIELD FROM ATOMISTIC SIMULATION OF CELLULOSE

Covariance matrices were calculated from all-atom MD simulation at various temperatures at the scale of single residues, *i.e.*, using coordinates of C_1 atoms, to then calculate the REACH force field using Eq. 15. This corresponds to coarse-graining sugar monomers to single CG-beads, so that the mass of one CG-bead was set to that of a sugar monomer.

The REACH analysis (Eq. 13 to Eq. 15) provides one elementary force constant for every pair of CG-beads and thus in principle fully defines the CG force field for the exact system under study. To obtain a CG force field that is also applicable to different system sizes and to improve the understanding of various interactions in the system, the CG force field was further simplified by defining a small number of representative classes of interactions (see **Figure 16**A). For the

present system, these are two-fold: First, there are classes of local in-chain interactions between CG-beads separated by 1, 2, or 3 pseudo-bonds modeled by force constants k_{12}, k_{13}, and k_{14}, respectively. These k_{ij} are obtained as the average of the elementary force constants (Eq. 15) belonging to that class. Second, there are classes of interchain interactions. One of these, k_{HB}, is meant to include the effect of interchain hydrogen bonding in the atomistic model. The force constant k_{HB} is also the average of the respective elementary force constants (Eq. 15). The remaining interchain interactions are modeled with a distance-dependent function

$$k_{nb}^{fit}(d) = A\, e^{-(d/d_0)} \qquad \text{Eq. 23}$$

with the force constant $k_{nb}^{fit}(x_{ij}^0)$ between two beads i and j separated by the equilibrium distance x_{ij}^0.

Figure 16B shows a scatter plot of the elementary force constants belonging to the class k_{12}, that is used for example between the black and the red CG-bead illustrated in **Figure 16**A. Also shown are contributions from the hydrophobic (k_{12}^{hrphob}) and hydrophilic (k_{12}^{hrphil}) surface and from the internal ($k_{12}^{internal}$) of the fibril. The distributions feature a single cluster within a narrow range in distance. The few outlying data points, *e.g.* those smaller than 400 kJ/mol Å^2, are attributed to the ends of the chains. The difference between the average values from the hydrophobic ($k_{12}^{hrphob} = 501$ kJ/mol Å^2) and hydrophilic ($k_{12}^{hrphil} = 528$ kJ/mol Å^2) surface and from the internal ($k_{12}^{internal} = 577$ kJ/mol Å^2) of the fibril indicate that the internal is stiffer with lesser fluctuations than the surface of the fibril. In the subsequent study, only the overall average of $k_{12} = 548$ kJ/mol Å^2 is used because of the proximity and overlap of the distributions

belonging to k_{12}^{hrphob}, k_{12}^{hrphil}, and $k_{12}^{internal}$. The value $k_{12} = 548$ kJ/mol Å^2 is consistent with the values 459 and 770 kJ/mol Å^2 obtained in a previous study using Boltzmann inversion as coarse-graining scheme [199]. The averages for k_{13}, k_{14}, and k_{HB} are 34, 2, and 32 kJ/mol Å^2, respectively. Instances of negative elementary force constants (in *e.g.* k_{13}, k_{14} shown in reference [204]) are unphysical and caused by numerical errors in the matrix diagonalization required by REACH and by the anharmonicity in the AA MD, as described in reference [50]. Usage of the average of k_{12} eliminates the unphysical effect of negative force constants. **Figure 16**C shows the distance-dependence of the nonbonded force constant model function, $k_{nb}^{fit}(d)$ (Eq. 23). The function $k_{nb}^{fit}(d)$ was fitted to the nonbonded force constant data at 300 K resulting in fitting parameters $A = 531$ kJ/mol Å^2 and $d_0 = 2.0$ Å. The data is well-reproduced. The function k_{nb}^{fit} decreases to zero within 12 Å. Thus, the nonbonded interactions modeled by $k_{nb}^{fit}(d)$ are more short-ranged than the bonded in-chain interactions that decrease less strong, for example, k_{13} and k_{14} corresponding to CG-bead separations of 10.38 and 15.57 Å still have values of 34 and 2 kJ/mol Å^2, which are larger than k_{nb}^{fit} at these distances. At the distance $d_{HB} = 8.4$ Å corresponding to typical distances between interchain hydrogen bonded sugar monomers, $k_{nb}^{fit}(8.4\text{ Å})$ is 9 kJ/mol Å^2. This is significantly smaller than k_{HB} (33 kJ/mol Å^2) and thus justifies the use of k_{HB} as separate class of force constants for hydrogen bonded interchain CG-beads.

The temperature-dependence of the classes of force constant introduced above is discussed next. **Figure 17**A,B and **Table 4** show that the force constants k_{12}, k_{13}, k_{14}, and k_{HB} decrease with temperature. For k_{12}, the temperature variation of the components belonging to the hydrophobic

($k_{12}^{h\prime phob}$) and hydrophilic ($k_{12}^{h\prime phil}$) surface and to the internal ($k_{12}^{internal}$) of the fibril are also shown. These follow $k_{12}^{internal} > k_{12} > k_{12}^{h\prime phil} > k_{12}^{h\prime phob}$ at all temperatures indicating that motion in the inside of the fibril is more restricted than at the surface, and also that motion at the hydrophilic surface is more restricted than that at the hydrophobic surface. Also the rate of decrease is larger at the surfaces than in the inside of the fibril indicating that the surface softens more than the inside as temperature increases. While the decrease with temperature of k_{13} and k_{14} does not have noteworthy specifics, k_{HB} shows a marked drop around 300 K, coinciding with a decrease in the average number of hydrogen-bonds in the simulation of the atomistic cellulose fibril at this temperature (see **Figure 17**C). **Figure 17**D shows that the nonbonded force constant model function, $k_{nb}^{fit}(d)$, also decreases with temperature (illustrated by the arrow in **Figure 17**D) indicating that also the nonbonded interactions in the crystalline cellulose fibril soften with increasing temperature. This is further highlighted by the inset of **Figure 17**D showing K_{nb}^{T}, the integral over $k_{nb}^{fit}(d)$ from $d = 5$ to 12 Å, which is the range from approximately the next-neighbor distance until $k_{nb}^{fit}(d)$ is approximately zero. The values of K_{nb}^{T} decrease monotonically with temperature.

The above section has fully defined the REACH force field for crystalline cellulose at temperatures between 100 and 500 K. The coarse-graining significantly reduced the number of particles involved in the simulation relative to the initial atomistic model and significantly reduced the complexity of the potential energy function from an anharmonic function to one with only harmonic terms. This enabling rapid simulation and analytical calculations on crystalline cellulose using the REACH force field.

DEPENDENCE OF REACH FORCE FIELD ON FIBRIL MODEL

The structure of the elementary cellulose micro-fibril depends on the synthesizing organism. Therefore, a potential dependence of the REACH force field on the structure of the cellulose model must be clarified. This is done by calculating and comparing the REACH force constants from simulations of two different cellulose models "A" and "B" whose cross-section is shown in **Table 5**. Please note that only the calculations in this section are based on the older CHARMM cellulose force field by Kuttel et al. [43] and a stretched-exponential nonbonded force constant model function $\kappa_{nb}^{fit}(d) = A' \exp[d/d_0']^{\beta}$. **Table 5** compares the obtained REACH force fields for models A and B. The force constants and parameters obtained for both models are in close agreement indicating that the REACH method is robust towards relatively realistic variations of the cellulose model.

SPEEDUP OF REACH CG MD SIMULATIONS

The performance of atomistic and coarse-grained simulation is briefly compared. The general protocol for AA and CG MD simulation is given in the methods section. The AA simulations use the advantageous domain decomposition algorithm for electrostatic interactions [113]. Similar time steps (2 fs) were chosen for the AA and CG MD, although a larger time step could have been chosen for the CG calculations.

Benchmark simulations were carried out on a the local cluster "moldyn" at ORNL, featuring 12-core/2.4 GHz Infiniband interconnected nodes. Simulations of a cellulose fibril (DP=40) yielded 7.2 ns/day and 172.5 for the AA and CG systems, respectively, using 48 cores (**Table 6**). This is

a speedup-factor of 24, that remained approximately similar, *i.e.* >20, with increasing system size (DP=80,160).

COMPARISON OF MEAN-SQUARE FLUCTUATION FROM CG AND AA MD CALCULATIONS

The root mean-square fluctuations (RMSF) from all-atom (AA) molecular dynamics simulation are compared to those obtained using the coarse-grained REACH force field to examine the quality of the force field derived. The RMSF from AA MD is calculated as:

$$RMSF_{MD,m} = \sqrt{\frac{1}{L}\sum_{t=1}^{L}(r_m(t) - \overline{r_m})^2} \qquad \text{Eq. 24}$$

where $r_m(t)$ is the position vector of atom m at time t, $\overline{r_m}$ is the average position vector of this atom, and L is the trajectory length. The RMSF for only the C_1 atom in the backbone of a sugar monomer is calculated and taken as representative for the fluctuation of this monomer.

The RMSF from CG MD is calculated twofold, first using Eq. 24 and the CG MD trajectory, and second using normal mode analysis (NMA) from the normal mode eigenvalues λ_n and eigenvectors $\vec{c}_{m,n}$ as

$$RMSF_{NMA,m} = \sqrt{\frac{k_B T}{M}\sum_{n}\frac{\vec{c}_{m,n}}{\lambda_n}} \qquad \text{Eq. 25}$$

where k_B is Boltzmann's constant, T is the temperature, M is the mass of one sugar monomer m, and n is the normal mode number. Normal mode analysis solves the equations of motion of the same harmonic system that is used in REACH CG MD simulation. Thus, this elegant analytical calculation is expected to give the same results as CG MD. However, NMA becomes difficult for

large systems as it requires to hold in memory and diagonalize a matrix of size $3N \times 3N$ with the number of CG-beads N.

Figure 18 compares the RMSF from AA MD simulation with that from REACH CG MD and REACH NMA at representative temperatures [for the fibril (DP=40) described in methods]. At $T \leq 150\,K$, the atomistic data agrees with the coarse-grained data as there is mostly harmonic dynamics in the system at these temperatures. At larger temperatures, the fluctuation from the atomistic simulation is larger than that from the coarse-grained calculations. This indicates the presence of anharmonic motion, which is reproduced only imperfectly using linear normal modes.

A part of the difference between the AA and CG MSF might also be due to usage of the average of the elementary force constants in REACH calculations while the MSF is non-linear in the force constant k, *i.e.* proportional to $1/k$. Therefore, more fluctuations would be removed by replacing small force constants with the average than added by replacing large force constants with the average. However, averaging is necessary to obtain a force field transferrable to other system sizes. Averaging using non-constant weights, *i.e.*, proportional to k^{-1}, is in principle possible but unfeasible in practice because numerical errors in the small nonbonded force constants lead to diverging weights. The effect of relatively small REACH-RMSF at large temperatures is indicated to be of limited relevance.[2]

[2]To investigate a possible effect of the relatively small RMSF in REACH calculations, the force constants were iteratively rescaled by a constant value to reproduce the RMSF from AA MD, sacrificing the elegance of REACH, *i.e.*, the non-iterative direct calculation of force constants, for only this test-case and the calculations for validation were also performed with the rescaled force constants. Qualitative features are similar to the original REACH

MEAN-SQUARE FLUCTUATION FROM NORMAL MODE ANALYSIS

Normal mode calculations allow detailed analysis of the mean-square fluctuations in terms of a decomposition into contributions along different directions or from different parts of the system. **Figure 19**A shows the NMA-derived mean-square fluctuations along the longitudinal axis of a cellulose fibril (DP=80). The MSF is the largest at the fibril ends (positions #1 and #80), somewhat large in the center of the fibril (position #40), and the smallest at "knot" positions between the end and the center of the fibril (approximately positions #20 and #60). This pattern exists in the relatively large MSF along the hydrophilic axis (**Figure 19**B) and in the even larger MSF along the hydrophobic axis (**Figure 19**C) while the fluctuations along the longitudinal axis (**Figure 19**D) are approximately an order of magnitude smaller in amplitude. The shape of the MSF transversal to the fibril axis resembles the fluctuations of for example an elastic rod and is due to the shape of the lowest frequency normal modes, shown in **Figure 19**E. The motion described by the lowest frequency (largest amplitude) normal modes is i) bending along hydrophobic axis (mode 1), ii) bending along hydrophilic axis (mode 2), and iii) fibril twist (mode 3). This three-membered pattern is in principle repeated with increasing mode number, however, one knot is added per repetition of the pattern resulting in overtones. The repetition is soon interrupted at mode 12 when different normal modes become relevant, *e.g.* due to fibril stretching. **Figure 19**F illustrates the motion from typical normal modes, which describe a basic bending motion (mode 1), an overtone bending motion (mode 31), a torsion motion (mode 3), a

---

analysis, particularly pertaining to differences between the hydrophobic and hydrophilic face of cellulose. Quantitatively, the reweighted force constants lead to a softer fibril with a shift of the density of states to lower frequencies, a decrease in the transverse and a small decrease in the longitudinal Young's moduli, a somewhat small decrease in the persistence length, and an increase in the chain-specific fluctuations. These results are not shown here.

stretching motion (mode 12), and also a breathing motion (mode 60). The breathing motion changes the distance between neighboring chains on the surface of cellulose and is capable of displacing chains towards the solvent environment, which suggests that such motion could be relevant for the deconstruction of the fibril. **Figure 19**G shows the fraction of the overall MSF that is cumulatively contributed by the N lowest normal modes. The ten lowest modes capture on average >60% of the total MSF, a fraction which is reached already by the two lowest modes at some positions along the fibril. Overall, **Figure 19** illustrates some representative types of motion that the fibril undergoes and shows that this motion contributes most of total fluctuations.

In addition to the overall MSF (its square-root, the RMSF, is defined in Eq. 25), we define the effective MSF along a direction $\vec{d}_m$ perpendicular to the fibril surface

$$x^2_{\perp,\mathrm{m}} = \frac{k_B T}{M} \sum_n \frac{\vec{p}_{m,n}}{\lambda_n} \qquad \text{Eq. 26}$$

where k_B is Boltzmann's constant, T is the temperature, M is the mass of one sugar monomer m, n is the normal mode number, and $\vec{p}_{m,n} = (\vec{c}_{m,n} \cdot \vec{d}_m)\vec{d}_m$ is the projection of the normal mode vector $\vec{c}_{m,n}$ along the direction $\vec{d}_m$ $(\|\vec{d}_m\| = 1)$ perpendicular to the fibril surface. **Figure 20**A,B show the temperature-dependence of the overall and effective MSF averaged over the CG-beads at the hydrophobic and hydrophilic surface. The MSF increases superlinearly due to the softening of the force constants with temperature (compare **Figure 17**). At all T, the MSF at the hydrophilic is larger than that at the hydrophobic surface, as reported by Beckham et al. [205], while the effective MSF from chains at the hydrophobic surface is significantly larger than from those on the hydrophilic surface (**Figure 20**B). Hence, the fraction between the effective

and total MSF $\langle f_m \rangle_{side} = \langle x_{\perp,\mathrm{m}}^2 / x_m^2 \rangle_{side}$ is much larger for CG-beads at the hydrophobic side. Also, the values of $\langle f_m \rangle_{side}$ are approximately constant with temperature. The data shown above demonstrates that the MSF perpendicular to the fibril surface is the largest at the hydrophobic surface. This is expected, because fluctuations in the direction perpendicular to the hydrophilic surface are suppressed by the presence of interchain hydrogen bonds along that direction, while such strong $O-H\cdots O$ -hydrogen bonds are absent in the direction perpendicular to the hydrophobic surface.

DENSITY OF STATES

Next, the density of states is derived from the normalized velocity autocorrelation function as

$$g(\omega) = \int_{-\infty}^{\infty} \frac{\langle \vec{v}(0) \cdot \vec{v}(t) \rangle}{\langle \vec{v}^2 \rangle} e^{-i\omega t} \qquad \text{Eq. 27}$$

where $\vec{v}(t)$ is the velocity at time t and the brackets denote the average over all C_1 atoms (or CG-beads) and the trajectory. To obtain error bars, ten AA and CG MD simulations were performed for 200 ps and velocities from every 5 fs were used to calculate the density of states.

Figure 21A shows the density of states for the internal of the fibril, the hydrophilic surface and the hydrophobic surface. The density of states calculated from the internal of the fibril is shifted to larger frequencies than that from the surfaces indicating larger effective force constants for beads in the interior of the fibril and thus smaller-amplitude motion than on the surface. Moreover, a difference appears in the density of states between the hydrophobic and hydrophilic surfaces. The density of states from the hydrophobic surface has larger intensity at small (approximately 10-30 cm^{-1}) and lesser intensity at large (larger than 80-100 cm^{-1}) frequencies

than the density of states from the hydrophilic surface. This suggests that the hydrophobic surface is softer than the hydrophilic one. The CG MD densities of states are in qualitative agreement with the data from AA MD in that the blue-shift of the density of states from the interior is reproduced and in that the density of states from the hydrophobic surface is larger than that from the hydrophilic surface at small frequencies (**Figure 21**B). The CG densities of states shift to smaller frequencies with increasing temperature (**Figure 21**C) indicating the expected temperature-dependent softening of the motion of the fibril.

YOUNG'S MODULUS AND VELOCITY OF SOUND

Young's modulus E_Y is a characteristic elastic property of materials describing the "stiffness" of a material, that is, its relative increase in length $\Delta L/L$ under the influence of a force F:

$$F = E_Y \cdot A \frac{\Delta L}{L} \qquad \text{Eq. 28}$$

where A is the material's cross section. The Young's modulus was calculated in the longitudinal and transversal fibril directions and compared to experiment. To obtain the relative increase in length for the longitudinal fibril direction (**Figure 22**A), a force F=50,000 kcal/mol nm was exerted on the fibril until the fibril length converged (within 10 ns). Such pulling simulations were performed at temperatures between 100 and 500 K in steps of 50 K. The maximal relative change in length was 5.1%. The cross-section area A was calculated using the cellulose lattice parameters to 11.4 nm^2 and considered temperature-independent. Only for the transversal fibril direction, a similar protocol was followed using a force F=20,000 kcal/mol nm and different fibril models to reduce noise. These models consisted of 6, 30, and 5 (30, 6, and 5) cellulose unit

cells along the transversal base vectors *a, b*, and the longitudinal *c* to obtain the Young's modulus along b (a) (**Figure 22**B).

Figure 22C shows the temperature dependence of the longitudinal Young's modulus that decreases monotonically between 100 and 500 K. At 300 K, E_Y is 162 GPa in agreement with experimental and theoretical estimates that range from 93 to 220 GPa [206-216]. The transversal Young's modulus was calculated at only 300 K to 25 and 41 GPa along the base vectors *a* and *b*, respectively, which is consistent with previous modeling studies (Young's modulus along *a* and *b* estimated as 51 and 57 GPa [217] or 15 and 55 GPa [215], respectively). This is also consistent with experiments that found the transversal Young's modulus to be 15 GPa (using inelastic X-ray scattering) [218] or to be between 18 and 50 GPa using atomic force microscopy [219]. However, experiments were not able to distinguish a specific transversal direction. Altogether, the calculated values show that the modulus is significantly larger along the longitudinal than along the transversal direction and thus underline the anisotropy of this property in cellulose.

The Young's modulus is related to the velocity of sound, v_s, via Christoffel's equation

$$E_Y = \rho v_s^2 \qquad \text{Eq. 29}$$

where ρ is the density of cellulose (taken as 1.67 g/cm^3 [220]). Thus obtained v_s is 9831 m/s in satisfactory agreement with experiment ($v_{s,exp} = 11{,}450 \pm 1{,}290\ m/s$ corresponding to $E_Y = 220\ GPa$ [218]). The velocity of sound can also be estimated from the phonon dispersion relation $\omega(q)$ from lattice dynamics calculations as $v_s = \partial\omega/\partial q$ for $q \to 0$. The dispersion relations have been calculated from the harmonic REACH CG force field and v_s was calculated to 9,785 m/s.

PERSISTENCE LENGTH

After previous validation of the REACH force field, it is now applied to estimate the persistence length of a cellulose microcrystal. The persistence length, L_P, of a polymer describes its stiffness. The value of L_P of a cellulose microcrystal is estimated from the elastic bending energy using Hook's law

$$V(\Theta) = \frac{1}{2} k_B T L_P \frac{\Theta^2}{s} \qquad \text{Eq. 30}$$

where k_B is the Boltzmann's constant, T is the temperature, s is the arc length of the bended polymer, and Θ is the bending angle. Eighteen conformations of cellulose microcrystals were created and relaxed at 300 K, annealed to 30 K, and relaxed again for 50, 100, and 50 ps, respectively. The relaxed structures were then energy minimized using steepest descent and conjugate gradient minimization algorithms yielding values of the bending energy at different bending angles. Eq. 30 was fit to this data to obtain the persistence length L_P.

The average value of L_P is 378 μm, values varied between 332 and 423 μm depending on the thickness of the fibril in the given bending direction. Bending the fibril in a direction along which it is thick consumes more energy because the further away a chain is from the center of the fibril the more will it be stretched. It results intuitively that the resulting persistence length depends on the model of the fibril and a different model will in principle result in a different value of L_P. Experimental values for the persistence length of single chains of cellulose were previously reported to 160 [221], 252 [222], 110 [223], or 130 Å [224]; a reported theoretical value is 145 Å [225]. However, values of L_P, not for single chains, but for a cellulose microcrystal are expected to be considerably larger. Accordingly, atomic force microscopy

images and electron micrographs show cellulose microcrystals that are elongated on the length-scale of several hundreds of nanometers [226-231] and start to show some bending on the length-scale of tens of micrometer. This is consistent with the current estimate of approximately 380 μm. The persistence length for other materials is for example within 4-17 μm for F-actin [232], within 1-8 mm for microtubules [232], or from ~20 μm to several mm for carbon nanotubes [233-236], depending on *e.g.* tube diameter. To the best of the author's knowledge, no other value of the persistence length of a cellulose microcrystal is known.

CONCLUSIONS - DYNAMICS OF CRYSTALLINE CELLULOSE

In the present work on coarse-graining, a harmonic REACH coarse-grained force field was calculated for a 36 chain fibril of crystalline $I\beta$ cellulose. The mean-square fluctuation, vibrational density of states, longitudinal and transversal Young's moduli, and velocity of sound were calculated for a coarse-grained cellulose fibril and these properties compare favorably to atomistic simulation and experiment validating the force field derived. The persistence length of a crystalline cellulose fibril has not yet been reported to the best knowledge of the author. It was calculated to $\sim 380\ \mu m$.

Several lines of experimental evidence suggest that the cellulose fibril may be deconstructed from the hydrophobic surface. The carbohydrate-binding module of family I Cel7A has a flat surface with hydrophobic groups that are spaced apart similar to the sugar rings in cellulose suggesting that the CBM might preferentially bind the hydrophobic surface of cellulose [229, 237]. Furthermore, single molecule fluorescence experiments of CBM linked to green fluorescence protein indicate the binding of cellulase to the hydrophobic surface [238]. Studies

performing atomic force microscopic imaging experiments suggest said binding and, in addition, motion along the binding-surface [228, 230]. The present study adds to this evidence by finding that the MSF perpendicular to the hydrophobic surface of cellulose is larger than that perpendicular to the hydrophilic surface, by finding that the transversal Young's modulus along the base vector almost perpendicular to the hydrophobic surface is smaller than that perpendicular to the hydrophilic surface, and by finding that the vibrational densities of states in both atomistic and coarse-grained simulation have increased intensity at small frequencies at the hydrophobic surface. With these findings, the present study also suggests that the crystalline $I\beta$ cellulose may be more easily deconstructed from the hydrophobic surface.

The present REACH force field is harmonic and thus useful for the investigation of elastic properties, as done here. But larger-scale motions involve a considerable anharmonic component that can be introduced using a Morse-potential for inter-chain interactions. The associated parameters are related to the REACH force constants around the energy minimum. This anharmonic extension to the present model will be done in future work.

CONCLUSIONS

The dynamics of biological macromolecules is encoded in their three-dimensional structures and is often relevant for their function [15]. In the last decades, structural methods have provided high-resolution (atomistic) models for many biomacromolecules while less information has been found on the corresponding dynamics using experiment. At the same time, molecular dynamics simulation has established itself as a key method for investigating macromolecular motion. As a "computational microscope", simulation provides fine details of the time-evolution of single atoms so that properties of the system can be analyzed with a precision and at time scales that are otherwise inaccessible.

The study of methyl group entropy illustrates the capabilities of analyzing simulation data in a way that is inaccessible to experiment. The simulated motion of the protein has been decomposed with great detail into various contributions of interest not only to calculate the value of a quantity but also to understand what influences that quantity. More specifically, simulations allows not only the calculation of the methyl group order parameter but also facilitates investigations of how certain kinds of motion (*e.g.* side chain rotation about a dihedral angle) affect the order parameter, or to study what influences the relation between the order parameter and the corresponding entropy. Such analyses were not possible with only experiment; however, experiment still was required for validation of the calculated order parameters to then perform the detailed analyses. Furthermore, future experiments will benefit from this work because the present findings enable more accurate estimates of methyl group entropy.

The study on crystalline cellulose illustrates how simulation can help to understand a system that cannot be easily investigated using experiment. Furthermore, the study shows how a large-scale simulation can be obtained in the first place using coarse graining. In this study, a coarse-grained model of cellulose Iβ was systematically calculated in a one-step process using the REACH method. Calculated properties agree surprisingly well with experiment, indicating that physical principles present in atomistic simulation have been successfully translated to a coarser scale. This success and the fact that, unlike many coarse-graining approaches REACH requires only a minimum of decisions to be made by the researcher (*e.g.* selection of parameters), suggests the integration of the REACH method into a future, fully-automated, multiscale, coarse-graining toolset that will allow the development of a coarse-grained force field for a given system at the push of a button. Such a toolset would be an efficient workaround to the hurdle of limited transferability of common coarse-grained force fields as it would significantly reduce the effort required to derive a coarse-grained model. Thus, future use of coarse-graining would be greatly facilitated and the corresponding cost would be greatly reduced, which increases for many researchers the time scale they can access with simulation.

The present studies and those briefly mentioned in the introduction [16, 28-32, 239] give examples of the current capabilities of biomolecular simulation. While the findings in some of the studies are already impressive, an even brighter future for the method is to come. With microsecond-long atomistic simulations being routinely possible and the apparent success of today's force fields even on the millisecond time scale [16, 28, 240], molecular dynamics simulations are beginning to access time scales at which many critical biological processes inaccessible to experiments are known to occur [241]. These processes can involve, for example,

protein folding, *i.e.*, the folding of a 1-dimensional amino acid chain into its native three-dimensional structure. Protein folding is one of the oldest unsolved problems addressed by biological research and there are many diseases associated with erroneous protein folding, *e.g.,* Alzheimer's or Parkinson's. It is possible that a fully atomistic molecular dynamics simulation will mechanistically explain the detailed origins of misfolding of certain proteins and thus ultimately enable development of cures for the diseases they cause.

LIST OF REFERENCES

1. Huang, Y.J. and G.T. Montelione, *Structural biology: Proteins flex to function.* Nature, 2005. **438**(7064): p. 36-37.
2. Bernstein, F.C., et al., *The protein data bank: A computer-based archival file for macromolecular structures.* Archives of Biochemistry and Biophysics, 1978. **185**(2): p. 584-591.
3. Falke, J.J., *A Moving Story.* Science, 2002. **295**(5559): p. 1480-1481.
4. Henzler-Wildman, K. and D. Kern, *Dynamic personalities of proteins.* Nature, 2007. **450**(7172): p. 964-972.
5. Austin, R.H., et al., *Dynamics of ligand binding to myoglobin.* Biochemistry, 1975. **14**(24): p. 5355-5373.
6. Wagner, G. and K. Wuthrich, *Dynamic model of globular protein conformations based on NMR studies in solution.* Nature, 1978. **275**(5677): p. 247-248.
7. *NMR Characterization of the Dynamics of Biomacromolecules.* Chemical Reviews, 2004. **104**(8): p. 3623-3640.
8. Wüthrich, K. and G. Wagner, *Internal dynamics of proteins.* Trends in Biochemical Sciences, 1984. **9**(4): p. 152-154.
9. Wüthrich, K. and G. Wagner, *NMR investigations of the dynamics of the aromatic amino acid residues in the basic pancreatic trypsin inhibitor.* FEBS Letters, 1975. **50**(2): p. 265-268.
10. Palmer III, A.G., *NMR PROBES OF MOLECULAR DYNAMICS: Overview and Comparison with Other Techniques.* Annual Review of Biophysics and Biomolecular Structure, 2001. **30**(1): p. 129-155.
11. Hubbell, W.L., D.S. Cafiso, and C. Altenbach, *Identifying conformational changes with site-directed spin labeling.* Nat Struct Mol Biol, 2000. **7**(9): p. 735-739.
12. Schotte, F., et al., *Watching a Protein as it Functions with 150-ps Time-Resolved X-ray Crystallography.* Science, 2003. **300**(5627): p. 1944-1947.
13. Moffat, K., *Time-Resolved Biochemical Crystallography: A Mechanistic Perspective.* Chemical Reviews, 2001. **101**(6): p. 1569-1582.
14. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-590.
15. Karplus, M. and J. Kuriyan, *Molecular dynamics and protein function.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(19): p. 6679-6685.
16. Shaw, D.E., et al., *Atomic-Level Characterization of the Structural Dynamics of Proteins.* Science, 2010. **330**(6002): p. 341-346.
17. Careaga, C.L. and J.J. Falke, *Thermal motions of surface α-helices in the d-galactose chemosensory receptor: Detection by disulfide trapping.* Journal of Molecular Biology, 1992. **226**(4): p. 1219-1235.
18. Debye, P., *Interferenz von Röntgenstrahlen und Wärmebewegung.* Annalen der Physik, 1913. **348**(1): p. 49-92.
19. Waller, I., *Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen.* Zeitschrift für Physik A Hadrons and Nuclei, 1923. **17**(1): p. 398-408.
20. Ostermann, A., et al., *Ligand binding and conformational motions in myoglobin.* Nature, 2000. **404**(6774): p. 205-208.
21. Ferrand, M., et al., *Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering.* Proceedings of the National Academy of Sciences, 1993. **90**(20): p. 9668-9672.
22. Eisenmesser, E.Z., et al., *Enzyme Dynamics During Catalysis.* Science, 2002. **295**(5559): p. 1520-1523.
23. Parak, F., et al., *Evidence for a correlation between the photoinduced electron transfer and dynamic properties of the chromatophore membranes from Rhodospirillum rubrum.* FEBS Letters, 1980. **117**(1–2): p. 368-372.
24. Alder, B.J. and T.E. Wainwright, *Phase Transition for a Hard Sphere System.* The Journal of Chemical Physics, 1957. **27**(5): p. 1208-1209.
25. Rahman, A., *Correlations in the Motion of Atoms in Liquid Argon.* Physical Review, 1964. **136**(2A): p. A405-A411.
26. Rahman, A. and F.H. Stillinger, *Molecular Dynamics Study of Liquid Water.* The Journal of Chemical Physics, 1971. **55**(7): p. 3336-3359.

27. Levitt, M., *Computer Simulation of DNA Double-helix Dynamics.* Cold Spring Harbor Symposia on Quantitative Biology, 1983. **47**: p. 251-262.
28. Lindorff-Larsen, K., et al., *How Fast-Folding Proteins Fold.* Science, 2011. **334**(6055): p. 517-520.
29. Tajkhorshid, E., et al., *Control of the Selectivity of the Aquaporin Water Channel Family by Global Orientational Tuning.* Science, 2002. **296**(5567): p. 525-530.
30. Jensen, M.Ø., et al., *Mechanism of Voltage Gating in Potassium Channels.* Science, 2012. **336**(6078): p. 229-233.
31. Hong, L., et al., *Three Classes of Motion in the Dynamic Neutron-Scattering Susceptibility of a Globular Protein.* Phys. Rev. Lett., 2011. **107**(14): p. 148102.
32. Schlick, T., et al., *Biomolecular modeling and simulation: a field coming of age.* Quarterly Reviews of Biophysics, 2011. **44**(02): p. 191-228.
33. Adcock, S.A. and J.A. McCammon, *Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins.* Chemical Reviews, 2006. **106**(5): p. 1589-1615.
34. Monticelli, L., et al., *The MARTINI Coarse-Grained Force Field: Extension to Proteins.* Journal of Chemical Theory and Computation, 2008. **4**(5): p. 819-834.
35. Brooks, B.R., et al., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.
36. MacKerell, A.D., et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins.* The Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
37. Pearlman, D.A., et al., *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules.* Computer Physics Communications, 1995. **91**(1-3): p. 1-41.
38. Scott, W.R.P., et al., *The GROMOS Biomolecular Simulation Program Package.* The Journal of Physical Chemistry A, 1999. **103**(19): p. 3596-3607.
39. Guvench, O., et al., *Additive empirical force field for hexopyranose monosaccharides.* Journal of Computational Chemistry, 2008. **29**(15): p. 2543-2564.
40. Hatcher, E.R., O. Guvench, and A.D. MacKerell, *CHARMM Additive All-Atom Force Field for Acyclic Polyalcohols, Acyclic Carbohydrates, and Inositol.* Journal of Chemical Theory and Computation, 2009. **5**(5): p. 1315-1327.
41. Kirschner, K.N., et al., *GLYCAM06: A generalizable biomolecular force field. Carbohydrates.* Journal of Computational Chemistry, 2008. **29**(4): p. 622-655.
42. Lins, R.D. and P.H. Hünenberger, *A new GROMOS force field for hexopyranose-based carbohydrates.* Journal of Computational Chemistry, 2005. **26**(13): p. 1400-1412.
43. Kuttel, M., J.W. Brady, and K.J. Naidoo, *Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations.* Journal of Computational Chemistry, 2002. **23**(13): p. 1236-1243.
44. Zhang, Z., et al., *A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules.* Biophysical journal, 2008. **95**(11): p. 5073-5083.
45. Bahar, I., et al., *Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins.* Chemical Reviews, 2009. **110**(3): p. 1463-1497.
46. Hinsen, K., et al., *Harmonicity in slow protein dynamics.* Chemical Physics, 2000. **261**(1–2): p. 25-37.
47. Jernigan, R. and T. Sen, *Optimizing the Parameters of the Gaussian Network Model for ATP-Binding Proteins*, in *Normal Mode Analysis*. 2005, Chapman and Hall/CRC. p. 171-186.
48. Kondrashov, D.A., Q. Cui, and G.N. Phillips, *Optimization and Evaluation of a Coarse-Grained Model of Protein Motion Using X-Ray Crystal Data.* Biophysical Journal, 2006. **91**(8): p. 2760-2767.
49. Hamacher, K. and J.A. McCammon, *Computing the Amino Acid Specificity of Fluctuations in Biomolecular Systems.* Journal of Chemical Theory and Computation, 2006. **2**(3): p. 873-878.
50. Moritsugu, K. and J.C. Smith, *Coarse-Grained Biomolecular Simulation with REACH: Realistic Extension Algorithm via Covariance Hessian.* Biophysical Journal, 2007. **93**(10): p. 3460-3469.

51. Doruker, P., R.L. Jernigan, and I. Bahar, *Dynamics of large proteins through hierarchical levels of coarse-grained structures.* Journal of Computational Chemistry, 2002. **23**(1): p. 119-127.
52. Lu, M. and J. Ma, *The Role of Shape in Determining Molecular Motions.* Biophysical Journal, 2005. **89**(4): p. 2395-2401.
53. Moritsugu, K. and J.C. Smith, *REACH Coarse-Grained Biomolecular Simulation: Transferability between Different Protein Structural Classes.* Biophysical Journal, 2008. **95**(4): p. 1639-1648.
54. Moritsugu, K., V. Kurkal-Siebert, and J.C. Smith, *REACH Coarse-Grained Normal Mode Analysis of Protein Dimer Interaction Dynamics.* Biophysical Journal, 2009. **97**(4): p. 1158-1167.
55. Moritsugu, K. and J.C. Smith, *REACH: A program for coarse-grained biomolecular simulation.* Computer Physics Communications, 2009. **180**(7): p. 1188-1195.
56. Soper, A.K., *Empirical potential Monte Carlo simulation of fluid structure.* Chemical Physics, 1996. **202**(2–3): p. 295-306.
57. Reith, D., M. Pütz, and F. Müller-Plathe, *Deriving effective mesoscale potentials from atomistic simulations.* Journal of Computational Chemistry, 2003. **24**(13): p. 1624-1636.
58. Leach, A., *Molecular modelling : principles and applications*. 2001: Pearson Prentice Hall.
59. Chu, J.-W. and G.A. Voth, *Coarse-Grained Modeling of the Actin Filament Derived from Atomistic-Scale Simulations.* Biophysical Journal, 2006. **90**(5): p. 1572-1582.
60. Izvekov, S., et al., *Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching.* The Journal of Chemical Physics, 2004. **120**(23): p. 10896-10913.
61. Izvekov, S. and G.A. Voth, *A Multiscale Coarse-Graining Method for Biomolecular Systems.* The Journal of Physical Chemistry B, 2005. **109**(7): p. 2469-2473.
62. Verlet, L., *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules.* Physical Review, 1967. **159**(1): p. 98-103.
63. Hockney, R.W., *Potential calculation and some applications.* Methods Comput. Phys., 1970.
64. Swope, W.C., et al., *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters.* The Journal of Chemical Physics, 1982. **76**(1): p. 637-649.
65. Beeman, D., *Some multistep methods for use in molecular dynamics calculations.* Journal of Computational Physics, 1976. **20**(2): p. 130-139.
66. Finney, J.L., *Water? What's so special about it?* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2004. **359**(1448): p. 1145-1165.
67. Halle, B., *Protein hydration dynamics in solution: a critical survey.* Philos Trans R Soc Lond B Biol Sci, 2004. **359**(1448): p. 1207-23; discussion 1223-4, 1323-8.
68. Chaplin, M., *Do we underestimate the importance of water in cell biology?* Nat Rev Mol Cell Biol, 2006. **7**(11): p. 861-866.
69. Chaplin, M.F., *Water: its importance to life.* Biochemistry and Molecular Biology Education, 2001. **29**(2): p. 54-59.
70. Guillot, B., *A reappraisal of what we have learnt during three decades of computer simulations on water.* Journal of Molecular Liquids, 2002. **101**(1–3): p. 219-260.
71. Glass, D.C., et al., *Temperature Dependence of Protein Dynamics Simulated with Three Different Water Models.* Journal of Chemical Theory and Computation, 2010. **6**(4): p. 1390-1400.
72. Nutt, D.R. and J.C. Smith, *Molecular Dynamics Simulations of Proteins: Can the Explicit Water Model Be Varied?* Journal of Chemical Theory and Computation, 2007. **3**(4): p. 1550-1560.
73. Guvench, O., et al., *Additive empirical force field for hexopyranose monosaccharides.* J. Comput. Chem., 2008. **29**(15): p. 2543-2564.
74. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**(2): p. 926-935.
75. Karplus, M., T. Ichiye, and B.M. Pettitt, *Configurational entropy of native proteins.* Biophys J, 1987. **52**(6): p. 1083-5.

76. Moorman, V.R., K.G. Valentine, and A.J. Wand, *The dynamical response of hen egg white lysozyme to the binding of a carbohydrate ligand.* Protein Science, 2012. **21**(7): p. 1066-1073.
77. Marlow, M.S., et al., *The role of conformational entropy in molecular recognition by calmodulin.* Nat. Chem. Biol., 2010. **6**(5): p. 352-358.
78. Frederick, K.K., et al., *Conformational entropy in molecular recognition by proteins.* Nature, 2007. **448**(7151): p. 325-3.
79. Lee, A.L. and A.J. Wand, *Microscopic origins of entropy, heat capacity and the glass transition in proteins.* Nature, 2001. **411**(6836): p. 501-504.
80. Lee, A.L., S.A. Kinnear, and A.J. Wand, *Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex.* Nat. Struct. Biol., 2000. **7**(1): p. 72-77.
81. Krishnan, M. and J.C. Smith, *Response of Small-Scale, Methyl Rotors to Protein-Ligand Association: A Simulation Analysis of Calmodulin-Peptide Binding.* J. Am. Chem. Soc., 2009. **131**(29): p. 10083-10091.
82. Hajduk, P.J., et al., *NMR-Based Screening of Proteins Containing 13C-Labeled Methyl Groups.* J. Am. Chem. Soc., 2000. **122**(33): p. 7898-7904.
83. Hayward, R.L., et al., *Dynamics of crystalline acetanilide: Analysis using neutron scattering and computer simulation.* J. Chem. Phys., 1995. **102**(13): p. 5525-5541.
84. Alvarez, F., et al., *Origin of the Distribution of Potential Barriers for Methyl Group Dynamics in Glassy Polymers: A Molecular Dynamics Simulation in Polyisoprene.* Macromolecules, 2000. **33**(21): p. 8077-8084.
85. Nair, S., et al., *Methyl rotational tunneling dynamics of p-xylene confined in a crystalline zeolite host.* J. Chem. Phys., 2004. **121**(10): p. 4810.
86. Baudry, J., *van der Waals Interactions and Decrease of the Rotational Barrier of Methyl-Sized Rotators: A Theoretical Study.* J. Am. Chem. Soc., 2006. **128**(34): p. 11088-11093.
87. Baudry, J. and J.C. Smith, *Can Proteins and Crystals Self-Catalyze Methyl Rotations?* J. Phys. Chem. B, 2005. **109**(43): p. 20572-20578.
88. Hembree, W.I. and J.Y. Baudry, *Three-Dimensional Mapping of Micro-Environmental Control of Methyl Rotational Barriers.* J. Phys. Chem. B, 2011: p. 8575-8580.
89. Schwalbe, H. and J. Rinnenthal, *Thermodynamics: The world is flat.* Nat. Chem. Biol., 2010. **6**(5): p. 312-313.
90. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity.* J. Am. Chem. Soc., 1982. **104**(17): p. 4546-4559.
91. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results.* J. Am. Chem. Soc., 1982. **104**(17): p. 4559-4570.
92. Chatfield, D.C., A. Szabo, and B.R. Brooks, *Molecular Dynamics of Staphylococcal Nuclease: Comparison of Simulation with 15N and 13C NMR Relaxation Data.* J. Am. Chem. Soc., 1998. **120**(21): p. 5301-5311.
93. Best, R.B., J. Clarke, and M. Karplus, *The origin of protein sidechain order parameter distributions.* J. Am. Chem. Soc., 2004. **126**(25): p. 7734-7735.
94. Hu, H., J. Hermans, and A.L. Lee, *Relating side-chain mobility in proteins to rotameric transitions: Insights from molecular dynamics simulations and NMR.* J. Biomol. NMR, 2005. **32**(2): p. 151-162.
95. Yang, D.W. and L.E. Kay, *Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: Application to protein folding.* J. Mol. Biol., 1996. **263**(2): p. 369-382.
96. Killian, B.J., et al., *Configurational Entropy in Protein-Peptide Binding:: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an HIV-Derived PTAP Nonapeptide.* J. Mol. Biol., 2009. **389**(2): p. 315-335.
97. Li, D.-W., S.A. Showalter, and R. Brüschweiler, *Entropy Localization in Proteins.* J. Phys. Chem. B, 2010. **114**(48): p. 16036-16044.

98. Zhou, H.-X. and M.K. Gilson, *Theory of Free Energy and Entropy in Noncovalent Binding.* Chem. Rev., 2009. **109**(9): p. 4092-4107.
99. Collins, J.R., S.K. Burt, and J.W. Erickson, *Flap opening in HIV-1 protease simulated by /`activated/' molecular dynamics.* Nat. Struct. Mol. Biol., 1995. **2**(4): p. 334-338.
100. Ishima, R., et al., *Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function.* Structure, 1999. **7**(9): p. 1047-1055.
101. Hornak, V., et al., *HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.* Proc. Natl. Acad. Sci., 2006. **103**(4): p. 915-920.
102. Trylska, J., et al., *HIV-1 Protease Substrate Binding and Product Release Pathways Explored with Coarse-Grained Molecular Dynamics.* Biophys. J., 2007. **92**(12): p. 4179-4187.
103. Elder, J.H., et al., *Identification of proteolytic processing sites within the Gag and Pol polyproteins of feline immunodeficiency virus.* J. Virol., 1993. **67**(4): p. 1869-1876.
104. Shehu-Xhilaga, M., et al., *Proteolytic processing of the p2/nucleocapsid cleavage site is critical for human immunodeficiency virus type 1 RNA dimer maturation.* J. Virol., 2001. **75**(19): p. 9156-64.
105. Lin, Y.-C., et al., *Altered Gag Polyprotein Cleavage Specificity of Feline Immunodeficiency Virus/Human Immunodeficiency Virus Mutant Proteases as Demonstrated in a Cell-Based Expression System.* J. Virol., 2006. **80**(16): p. 7832-7843.
106. Navia, M.A., et al., *Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1.* Nature, 1989. **337**(6208): p. 615-620.
107. Kohl, N.E., et al., *Active human immunodeficiency virus protease is required for viral infectivity.* Proc. Natl. Acad. Sci., 1988. **85**(13): p. 4686-90.
108. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J. Comput. Chem., 2005. **26**(16): p. 1781-1802.
109. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins.* J. Phys. Chem. B, 1998. **102**(18): p. 3586-3616.
110. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**: p. 926-935.
111. Heaslet, H., et al., *Conformational flexibility in the flap domains of ligand-free HIV protease.* Acta Cryst., 2007. **D63**(8): p. 866-875.
112. Lam, P.Y.S., et al., *Cyclic HIV protease Inhibitors: synthesis, conformational analysis, P2/P2? structure activity relationship, and molecular recognition of cyclic ureas.* J. Med. Chem., 1996. **39**(18): p. 3514-3525.
113. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* J. Chem. Theory Comput., 2008. **4**(3): p. 435-447.
114. Hess, B., et al., *LINCS: A linear constraint solver for molecular simulations.* J. Comput. Chem., 1997. **18**(12): p. 1463-1472.
115. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* J. Chem. Phys., 1984. **81**(8): p. 3684-3690.
116. Parrinello, M. and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method.* J. Appl. Phys., 1981. **52**(12): p. 7182-7190.
117. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling.* J. Chem. Phys., 2007. **126**(1): p. 014101.
118. MacKerell Jr, A.D., *http://mackerell.umaryland.edu/CHARMM_ff_params.html*. http://mackerell.umaryland.edu/CHARMM_ff_params.html.
119. Prompers, J.J. and R. Bruschweiler, *General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation.* J. Am. Chem. Soc., 2002. **124**(16): p. 4522-4534.
120. Prompers, J.J. and R. Bruschweiler, *Reorientational Eigenmode Dynamics: A Combined MD/NMR Relaxation Analysis Method for Flexible Parts in Globular Proteins.* J. Am. Chem. Soc., 2001. **123**(30): p. 7305-7313.
121. Nicholson, L.K., et al., *Flexibility and function in HIV-1 protease.* Nat. Struct. Biol., 1995. **2**(4): p. 274-280.

122. Ishima, R., J.M. Louis, and D.A. Torchia, *Characterization of two hydrophobic methyl clusters in HIV-1 protease by NMR spin relaxation in solution.* J. Mol. Biol., 2001. **305**(3): p. 515-521.
123. Krishnan, M. and J.C. Smith, *Reconstruction of Protein Side-Chain Conformational Free Energy Surfaces From NMR-Derived Methyl Axis Order Parameters.* J. Phys. Chem. B, 2012.
124. Lehmann, M.S., T.F. Koetzle, and W.C. Hamilton, *Precision neutron diffraction structure determination of protein and nucleic acid components. I. Crystal and molecular structure of the amino acid L-alanine.* JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 1972. **94**(8): p. 2657-2660.
125. Koetzle, T.F., et al., *Precision neutron-diffraction structure determination of protein and nucleic acid components. 15. Crystal and molecular structure of the amino acid L-valine hydrochloride.* Journal of Chemical Physics, 1974. **60**(12): p. 4690-4696.
126. Briant, P., et al., *The methyl group geometry in trichloromethyltitanium: a reinvestigation by gas electron diffraction.* JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 1989. **111**(9): p. 3434-3436.
127. McKean, D.C., et al., *Infrared spectra of MeTiCl3 species, methyl group geometry and a force field.* Journal of Molecular Structure, 1991. **247**: p. 73-87.
128. Ottiger, M. and A. Bax, *How Tetrahedral Are Methyl Groups in Proteins?: A Liquid Crystal NMR Study.* JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 1999. **121**(19): p. 4690-4695.
129. Nandini, G. and D.N. Sathyanarayana, *Ab initio studies on geometry and vibrational spectra of N-methyl formamide and N-methylacetamide.* Journal of Molecular Structure-Theochem, 2002. **579**: p. 1-9.
130. Williamson, R.L. and M.B. Hall, *Calculations of the geometric and electronic structure of trichloromethyltitanium: is there an agostic hydrogen interaction?* JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 1988. **110**(13): p. 4428-4429.
131. Krishnan, M. and J.C. Smith, *Response of Small-Scale, Methyl Rotors to Protein-Ligand Association: A Simulation Analysis of Calmodulin-Peptide Binding.* Journal of the American Chemical Society, 2009. **131**(29): p. 10083-10091.
132. Chatfield, D.C., A. Szabo, and B.R. Brooks, *Molecular Dynamics of Staphylococcal Nuclease: Comparison of Simulation with 15N and 13C NMR Relaxation Data.* Journal of the American Chemical Society, 1998. **120**(21): p. 5301-5311.
133. Ishima, R., J.M. Louis, and D.A. Torchia, *Characterization of two hydrophobic methyl clusters in HIV-1 protease by NMR spin relaxation in solution.* Journal of Molecular Biology, 2001. **305**(3): p. 515-521.
134. Zhuravleva, A., et al., *Propagation of Dynamic Changes in Barnase Upon Binding of Barstar: An NMR and Computational Study.* Journal of Molecular Biology, 2007. **367**(4): p. 1079-1092.
135. de Almeida, P. and P.D. Silva, *The peak of oil production—Timings and market recognition.* Energy Policy, 2009. **37**(4): p. 1267-1276.
136. Huber, G.W., S. Iborra, and A. Corma, *Synthesis of Transportation Fuels from Biomass: Chemistry, Catalysts, and Engineering.* Chemical Reviews, 2006. **106**(9): p. 4044-4098.
137. Naik, S.N., et al., *Production of first and second generation biofuels: A comprehensive review.* Renewable and Sustainable Energy Reviews, 2010. **14**(2): p. 578-597.
138. Searchinger, T., et al., *Use of U.S. Croplands for Biofuels Increases Greenhouse Gases Through Emissions from Land-Use Change.* Science, 2008. **319**(5867): p. 1238-1240.
139. Farrell, A.E., et al., *Ethanol Can Contribute to Energy and Environmental Goals.* Science, 2006. **311**(5760): p. 506-508.
140. Perlack, R.D., et al., *Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply*. 2005, Oak Ridge National Laboratory: Oak Risge.
141. *Annual energy outlook 2012 early release.* 2012.
142. Towler, G.P., A.R. Oroskar, and S.E. Smith, *Development of a sustainable liquid fuels infrastructure based on biomass.* Environmental Progress, 2004. **23**(4): p. 334-341.
143. Mosier, N., et al., *Features of promising technologies for pretreatment of lignocellulosic biomass.* Bioresource Technology, 2005. **96**(6): p. 673-686.
144. Kumar, P., et al., *Methods for Pretreatment of Lignocellulosic Biomass for Efficient Hydrolysis and Biofuel Production.* Industrial & Engineering Chemistry Research, 2009. **48**(8): p. 3713-3729.

145. Hardy, B.J. and A. Sarko, *Conformational analysis and molecular dynamics simulation of cellobiose and larger cellooligomers.* Journal of Computational Chemistry, 1993. **14**(7): p. 831-847.
146. Zhang, Y.H.P. and L.R. Lynd, *Determination of the Number-Average Degree of Polymerization of Cellodextrins and Cellulose with Application to Enzymatic Hydrolysis.* Biomacromolecules, 2005. **6**(3): p. 1510-1515.
147. French, A.D. and G.P. Johnson, *Cellulose Shapes, Cellulose: Molecular and Structural Biology*, R.M. Brown and I.M. Saxena, Editors. 2007, Springer Netherlands. p. 257-284.
148. ATALLA, R.H. and D.L. VANDERHART, *Native Cellulose: A Composite of Two Distinct Crystalline Forms.* Science, 1984. **223**(4633): p. 283-285.
149. Sugiyama, J., R. Vuong, and H. Chanzy, *Electron diffraction study on the two crystalline phases occurring in native cellulose from an algal cell wall.* Macromolecules, 1991. **24**(14): p. 4168-4175.
150. Sugiyama, J., et al., *Transformation of Valonia cellulose crystals by an alkaline hydrothermal treatment.* Macromolecules, 1990. **23**(12): p. 3196-3198.
151. Masahisa, W., K. Tetsuo, and O. Takeshi, *Thermally induced crystal transformation from cellulose I[a] to I[ß]*. Vol. 35. 2003, Avenel, NJ, ETATS-UNIS: Nature Publishing Group. 155-159.
152. Hirai, A., M. Tsuji, and F. Horii, *TEM study of band-like cellulose assemblies produced by Acetobacter xylinum at 4 °C.* Cellulose, 2002. **9**(2): p. 105-113.
153. Kuga, S., S. Takagi, and R.M. Brown Jr, *Native folded-chain cellulose II.* Polymer, 1993. **34**(15): p. 3293-3297.
154. Nishiyama, Y., P. Langan, and H. Chanzy, *Crystal Structure and Hydrogen-Bonding System in Cellulose I β from Synchrotron X-ray and Neutron Fiber Diffraction.* Journal of the American Chemical Society, 2002. **124**(31): p. 9074-9082.
155. Langan, P., Y. Nishiyama, and H. Chanzy, *X-ray Structure of Mercerized Cellulose II at 1 Å Resolution.* Biomacromolecules, 2001. **2**(2): p. 410-416.
156. Wada, M., et al., *Cellulose IIII Crystal Structure and Hydrogen Bonding by Synchrotron X-ray and Neutron Fiber Diffraction.* Macromolecules, 2004. **37**(23): p. 8548-8555.
157. Nishiyama, Y., et al., *Crystal Structure and Hydrogen Bonding System in Cellulose I α from Synchrotron X-ray and Neutron Fiber Diffraction.* Journal of the American Chemical Society, 2003. **125**(47): p. 14300-14306.
158. Chundawat, S.P.S., et al., *Restructuring the Crystalline Cellulose Hydrogen Bond Network Enhances Its Depolymerization Rate.* Journal of the American Chemical Society, 2011. **133**(29): p. 11163-11174.
159. Watanabe, A., S. Morita, and Y. Ozaki, *Study on Temperature-Dependent Changes in Hydrogen Bonds in Cellulose I β by Infrared Spectroscopy with Perturbation-Correlation Moving-Window Two-Dimensional Correlation Spectroscopy.* Biomacromolecules, 2006. **7**(11): p. 3164-3170.
160. Watanabe, A., S. Morita, and Y. Ozaki, *Temperature-Dependent Changes in Hydrogen Bonds in Cellulose I α Studied by Infrared Spectroscopy in Combination with Perturbation-Correlation Moving-Window Two-Dimensional Correlation Spectroscopy: Comparison with Cellulose I β.* Biomacromolecules, 2007. **8**(9): p. 2969-2975.
161. Bergenstråhle, M., L.A. Berglund, and K. Mazeau, *Thermal Response in Crystalline I β Cellulose: A Molecular Dynamics Study.* The Journal of Physical Chemistry B, 2007. **111**(30): p. 9138-9145.
162. Carpita, N.C. and D.M. Gibeaut, *Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth.* The Plant Journal, 1993. **3**(1): p. 1-30.
163. Cosgrove, D.J., *Growth of the plant cell wall.* Nat Rev Mol Cell Biol, 2005. **6**(11): p. 850-861.
164. MUELLER, S.C., R.M. BROWN, and T.K. SCOTT, *Cellulosic Microfibrils: Nascent Stages of Synthesis in a Higher Plant Cell.* Science, 1976. **194**(4268): p. 949-951.
165. Ding, S.-Y. and M.E. Himmel, *The Maize Primary Cell Wall Microfibril: A New Model Derived from Direct Visualization.* Journal of Agricultural and Food Chemistry, 2006. **54**(3): p. 597-606.
166. Scheller, H.V. and P. Ulvskov, *Hemicelluloses.* Annual Review of Plant Biology, 2010. **61**(1): p. 263-289.

167. Ray, P.M., *Radioautographic Study of Cell Wall Deposition in Growing Plant Cells.* The Journal of Cell Biology, 1967. **35**(3): p. 659-674.
168. PROSEUS, T.E. and J.S. BOYER, *Turgor Pressure Moves Polysaccharides into Growing Cell Walls of Chara corallina.* Annals of Botany, 2005. **95**(6): p. 967-979.
169. Nishitani, K. and R. Tominaga, *Endo-xyloglucan transferase, a novel class of glycosyltransferase that catalyzes transfer of a segment of xyloglucan molecule to another xyloglucan molecule.* Journal of Biological Chemistry, 1992. **267**(29): p. 21058-21064.
170. Lu, F. and J. Ralph, *Chapter 6 - Lignin*, in *Cereal Straw as a Resource for Sustainable Biomaterials and Biofuels*. 2010, Elsevier: Amsterdam. p. 169-207.
171. Selig, M.J., et al., *Deposition of Lignin Droplets Produced During Dilute Acid Pretreatment of Maize Stems Retards Enzymatic Hydrolysis of Cellulose.* Biotechnology Progress, 2007. **23**(6): p. 1333-1339.
172. Donohoe, B.S., et al., *Visualizing lignin coalescence and migration through maize cell walls following thermochemical pretreatment.* Biotechnology and Bioengineering, 2008. **101**(5): p. 913-925.
173. Pingali, S.V., et al., *Breakdown of Cell Wall Nanostructure in Dilute Acid Pretreated Biomass.* Biomacromolecules, 2010. **11**(9): p. 2329-2335.
174. Berlin, A., et al., *Weak lignin-binding enzymes.* Applied Biochemistry and Biotechnology, 2005. **121**(1): p. 163-170.
175. Palonen, H., et al., *Adsorption of Trichoderma reesei CBH I and EG II and their catalytic domains on steam pretreated softwood and isolated lignin.* Journal of Biotechnology, 2004. **107**(1): p. 65-72.
176. Zhang, Q., et al., *A molecular dynamics study of the thermal response of crystalline cellulose I β.* Cellulose, 2011. **18**(2): p. 207-221.
177. Schulz, R., et al., *Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer.* Journal of Chemical Theory and Computation, 2009. **5**(10): p. 2798-2808.
178. Yui, T., et al., *Swelling behavior of the cellulose I β crystal models by molecular dynamics.* Carbohydrate Research, 2006. **341**(15): p. 2521-2530.
179. Yui, T. and S. Hayashi, *Molecular Dynamics Simulations of Solvated Crystal Models of Cellulose I α and IIII.* Biomacromolecules, 2007. **8**(3): p. 817-824.
180. Matthews, J.F., et al., *High-Temperature Behavior of Cellulose I.* The Journal of Physical Chemistry B, 2011. **115**(10): p. 2155-2166.
181. Nishiyama, Y., et al., *Neutron Crystallography, Molecular Dynamics, and Quantum Mechanics Studies of the Nature of Hydrogen Bonding in Cellulose I β.* Biomacromolecules, 2008. **9**(11): p. 3133-3140.
182. Gross, A.S. and J.W. Chu, *On the Molecular Origins of Biomass Recalcitrance: The Interaction Network and Solvation Structures of Cellulose Microfibrils.* Journal of Physical Chemistry B, 2010. **114**(42): p. 13333-13341.
183. Wada, M., et al., *Neutron crystallographic and molecular dynamics studies of the structure of ammonia-cellulose I: rearrangement of hydrogen bonding during the treatment of cellulose with ammonia.* Cellulose, 2011. **18**(2): p. 191-206.
184. Heiner, A.P., L. Kuutti, and O. Teleman, *Comparison of the interface between water and four surfaces of native crystalline cellulose by molecular dynamics simulations.* Carbohydrate Research, 1998. **306**(1–2): p. 205-220.
185. Liu, H., et al., *Understanding the Interactions of Cellulose with Ionic Liquids: A Molecular Dynamics Study.* The Journal of Physical Chemistry B, 2010. **114**(12): p. 4293-4301.
186. Nimlos, M.R., et al., *Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface.* Protein Engineering Design and Selection, 2007.
187. Yui, T., et al., *Systematic Docking Study of the Carbohydrate Binding Module Protein of Cel7A with the Cellulose I α Crystal Model.* The Journal of Physical Chemistry B, 2009. **114**(1): p. 49-58.
188. Zhong, L., et al., *Interactions of the complete cellobiohydrolase I from Trichodera reesei with microcrystalline cellulose I β.* Cellulose, 2008. **15**(2): p. 261-273.

189. Zhong, L., et al., *Computational simulations of the Trichoderma reesei cellobiohydrolase I acting on microcrystalline cellulose I[beta]: the enzyme-substrate complex.* Carbohydrate Research, 2009. **344**(15): p. 1984-1992.
190. Beckham, G.T., et al., *Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs.* The Journal of Physical Chemistry B, 2011. **115**(14): p. 4118-4127.
191. Cho, H.M., A.S. Gross, and J.-W. Chu, *Dissecting Force Interactions in Cellulose Deconstruction Reveals the Required Solvent Versatility for Overcoming Biomass Recalcitrance.* Journal of the American Chemical Society, 2011. **133**(35): p. 14033-14041.
192. Yui, T., N. Okayama, and S. Hayashi, *Structure conversions of cellulose III_I crystal models in solution state: a molecular dynamics study.* Cellulose, 2010. **17**(4): p. 679-691.
193. Chen, W., G.C. Lickfield, and C.Q. Yang, *Molecular modeling of cellulose in amorphous state. Part I: model building and plastic deformation study.* Polymer, 2004. **45**(3): p. 1063-1071.
194. Mazeau, K. and L. Heux, *Molecular Dynamics Simulations of Bulk Native Crystalline and Amorphous Structures of Cellulose.* The Journal of Physical Chemistry B, 2003. **107**(10): p. 2394-2403.
195. Molinero, V. and W.A. Goddard, *M3B: A Coarse Grain Force Field for Molecular Simulations of Malto-Oligosaccharides and Their Water Mixtures.* The Journal of Physical Chemistry B, 2004. **108**(4): p. 1414-1427.
196. Liu, P., S. Izvekov, and G.A. Voth, *Multiscale Coarse-Graining of Monosaccharides.* The Journal of Physical Chemistry B, 2007. **111**(39): p. 11566-11575.
197. Bu, L., et al., *The Energy Landscape for the Interaction of the Family 1 Carbohydrate-Binding Module and the Cellulose Surface is Altered by Hydrolyzed Glycosidic Bonds.* The Journal of Physical Chemistry B, 2009. **113**(31): p. 10994-11002.
198. Wohlert, J. and L.A. Berglund, *A Coarse-Grained Model for Molecular Dynamics Simulations of Native Cellulose.* Journal of Chemical Theory and Computation, 2011. **7**(3): p. 753-760.
199. Srinivas, G., X. Cheng, and J.C. Smith, *A solvent-free coarse-grain model for crystalline and amorphous cellulose fibrils.* Journal of Chemical Theory and Computation, 2011. **7**(8): p. 2539–2548.
200. Meinhold, L., F. Merzel, and J.C. Smith, *Lattice Dynamics of a Protein Crystal.* PHYSICAL REVIEW LETTERS, 2007. **99**(13): p. 138101.
201. Micu, A.M., et al., *Collective Vibrations in Crystalline L-Alanine.* The Journal of Physical Chemistry, 1995. **99**(15): p. 5645-5657.
202. Schulz, R., et al., *Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer.* J. Chem. Theory Comput., 2009. **5**(10): p. 2798-2808.
203. Lindner, B. and J.C. Smith, *Sassena — X-ray and neutron scattering calculated from molecular dynamics trajectories using massively parallel computers.* Comput. Phys. Commun., 2012. **183**(7): p. 1491-1501.
204. Glass, D.C., et al., *REACH Coarse-Grained Simulation of a Cellulose Fiber.* Biomacromolecules, 2012. **accepted**.
205. Beckham, G.T., et al., *Molecular-Level Origins of Biomass Recalcitrance: Decrystallization Free Energies for Four Common Cellulose Polymorphs.* J. Phys. Chem. B, 2011. **115**(14): p. 4118-4127.
206. Sakurada, I. and T. Ito, *Experimental Determination of Elastic Moduli of the Crystalline Regions in Oriented Polymers.* Kobunshi Kagaku, 1962. **19**: p. 300.
207. Šturcová, A., G.R. Davies, and S.J. Eichhorn, *Elastic Modulus and Stress-Transfer Properties of Tunicate Cellulose Whiskers.* Biomacromolecules, 2005. **6**(2): p. 1055-1061.
208. Nishino, T., K. Takano, and K. Nakamae, *Elastic modulus of the crystalline regions of cellulose polymorphs.* J. Polym. Sci. B, 1995. **33**(11): p. 1647-1651.
209. Matsuo, M., et al., *Effect of orientation distribution and crystallinity on the measurement by x-ray diffraction of the crystal lattice moduli of cellulose I and II.* Macromolecules, 1990. **23**(13): p. 3266-3275.
210. Sakurada, I., Y. Nukushina, and T. Ito, *Experimental determination of the elastic modulus of crystalline regions in oriented polymers.* J. Polym. Sci., 1962. **57**(165): p. 651-660.
211. Tanaka, F. and T. Iwata, *Estimation of the Elastic Modulus of Cellulose Crystal by Molecular Mechanics Simulation.* Cellulose, 2006. **13**(5): p. 509-517.

212. Tashiro, K. and M. Kobayashi, *Calculation of crystallite modulus of native cellulose.* Polym. Bull., 1985. **14**(3): p. 213-218.
213. Bergenstråhle, M., L.A. Berglund, and K. Mazeau, *Thermal Response in Crystalline Iβ Cellulose: A Molecular Dynamics Study.* J. Phys. Chem. B, 2007. **111**(30): p. 9138-9145.
214. Santiago Cintrón, M., G. Johnson, and A. French, *Young's modulus calculations for cellulose Iβ by MM3 and quantum mechanics.* Cellulose, 2011. **18**(3): p. 505-516.
215. Tashiro, K. and M. Kobayashi, *Theoretical evaluation of three-dimensional elastic constants of native and regenerated celluloses: role of hydrogen bonds.* Polymer, 1991. **32**(8): p. 1516-1526.
216. Iwamoto, S., et al., *Elastic Modulus of Single Cellulose Microfibrils from Tunicate Measured by Atomic Force Microscopy.* Biomacromolecules, 2009. **10**(9): p. 2571-2576.
217. Jaswon, M.A., P.P. Gillis, and R.E. Mark, *The Elastic Constants of Crystalline Native Cellulose.* Proc. R. Soc. A, 1968. **306**(1486): p. 389-412.
218. Diddens, I., et al., *Anisotropic Elastic Properties of Cellulose Measured Using Inelastic X-ray Scattering.* Macromolecules, 2008. **41**(24): p. 9755-9759.
219. Lahiji, R.R., et al., *Atomic Force Microscopy Characterization of Cellulose Nanocrystals.* Langmuir, 2010. **26**(6): p. 4480-4488.
220. Nishiyama, Y., P. Langan, and H. Chanzy, *Crystal Structure and Hydrogen-Bonding System in Cellulose Iβ from Synchrotron X-ray and Neutron Fiber Diffraction.* J. Am. Chem. Soc., 2002. **124**(31): p. 9074-9082.
221. McCormick, C.L., P.A. Callais, and B.H. Hutchinson, *Solution studies of cellulose in lithium chloride and N,N-dimethylacetamide.* Macromolecules, 1985. **18**(12): p. 2394-2401.
222. Bianchi, E., et al., *Mesophase formation and chain rigidity in cellulose and derivatives. 4. Cellulose in N,N-dimethylacetamide-lithium chloride.* Macromolecules, 1985. **18**(4): p. 646-650.
223. Kamide, K., M. Saito, and K. Kowsaka, *Temperature Dependence of Limiting Viscosity Number and Radius of Gyration for Cellulose Dissolved in Aqueous 8% Sodium Hydroxide Solution.* Polym. J., 1987. **19**(10): p. 1173-1181.
224. Burchard, W., et al., *Cellulose in Schweizer's Reagent: A Stable, Polymeric Metal Complex with High Chain Stiffness.* Angew. Chem., Int. Ed., 1994. **33**(8): p. 884-887.
225. Braccini, I., A. Heyraud, and S. Pérez, *Three-dimensional features of the bacterial polysaccharide (1 → 4)-β-D-glucuronan: A molecular modeling study.* Biopolymers, 1998. **45**(2): p. 165-175.
226. Ding, S.-Y. and M.E. Himmel, *The Maize Primary Cell Wall Microfibril: A New Model Derived from Direct Visualization.* J. Agric. Food Chem., 2006. **54**(3): p. 597-606.
227. Imai, T., et al., *Unidirectional processive action of cellobiohydrolase Cel7A on Valonia cellulose microcrystals.* FEBS Lett., 1998. **432**(3): p. 113-116.
228. Igarashi, K., et al., *Traffic Jams Reduce Hydrolytic Efficiency of Cellulase on Cellulose Surface.* Science, 2011. **333**(6047): p. 1279-1282.
229. Lehtiö, J., et al., *The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules.* Proc. Natl. Acad. Sci., 2003. **100**(2): p. 484-489.
230. Liu, Y.-S., et al., *Cellobiohydrolase hydrolyzes crystalline cellulose on hydrophobic faces.* J. Biol. Chem., 2011. **286**: p. 11195.
231. Wagner, R., A. Raman, and R. Moon, *Transverse elasticity of cellulose nanocrystals via atomic force microscopy.* 10th International Conference on Wood & Biofiber Plastic Composites and Cellulose Nanocomposites Symposium, 2010: p. 309-316.
232. van Mameren, J., et al., *Leveraging Single Protein Polymers To Measure Flexural Rigidity†.* J. Phys. Chem. B, 2009. **113**(12): p. 3837-3844.
233. Duggal, R. and M. Pasquali, *Dynamics of Individual Single-Walled Carbon Nanotubes in Water by Real-Time Visualization.* Phys. Rev. Lett., 2006. **96**(24): p. 246104.
234. Fakhri, N., et al., *Diameter-dependent bending dynamics of single-walled carbon nanotubes in liquids.* Proc. Natl. Acad. Sci., 2009. **106**(34): p. 14219-14223.
235. Zhou, W., et al., *Small angle neutron scattering from single-wall carbon nanotube suspensions: evidence for isolated rigid rods and rod networks.* Chem. Phys. Lett., 2004. **384**(1–3): p. 185-189.

236. Yakobson, B. and L. Couchman, *Persistence Length and Nanomechanics of Random Bundles of Nanotubes.* J. Nanopart. Res., 2006. **8**(1): p. 105-110.
237. Tormo, J., et al., *Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose.* EMBO J., 1996. **15**(21): p. 5739-5751.
238. Dagel, D.J., et al., *In Situ Imaging of Single Carbohydrate-Binding Modules on Cellulose Microfibrils.* J. Phys. Chem. B, 2010. **115**(4): p. 635-641.
239. Piana, S., K. Lindorff-Larsen, and D.E. Shaw, *Protein folding kinetics and thermodynamics from atomistic simulation.* Proceedings of the National Academy of Sciences, 2012.
240. Shaw, D.E., et al., *Millisecond-scale molecular dynamics simulations on Anton*, in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. 2009, ACM: Portland, Oregon. p. 1-11.
241. Dror, R.O., et al., *Biomolecular simulation: a computational microscope for molecular biology.* Annu Rev Biophys, 2012. **41**: p. 429-52.
242. Dror, R.O., et al., *Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations.* J Gen Physiol, 2010. **135**(6): p. 555-62.

APPENDIX

TABLES

Table 1. Differences in primary sequence between PDB IDs 2PC0 and 1QBS.

2PC0	residue	1QBS
ILE	3	VAL
LYS	7	GLU
ASP	37	SER
CYS	95	ALA

Table 2. Average of methyl group axis order parameters for apo and ligand bound systems.[a]

System	$O^2_{rot} = 0.111$		$O^2_{rot} = 0.099$		$\Delta\Delta S$ $[kcal/mol]$
	$O^2_{axis,apo}$	$O^2_{axis,bound}$	$O^2_{axis,apo}$	$O^2_{axis,bound}$	
BARNASE	0.57	0.67	0.64	0.75	-0.133
HIV PR	0.42	0.52	0.48	0.58	-0.046
CALMODULIN	0.44	0.51	0.49	0.57	-0.052
ALL	0.43	0.53	0.49	0.59	-0.051

[a] Systems investigated: Barnase [134], HIV PR [133], Calmodulin [78]. Methyl group axis order parameters are shown using $O^2_{rot} = 0.111$ or 0.099 (compare Figure 10). Also reported is the average difference in the change of entropy when using $O^2_{rot} = 0.099$ instead of 0.111 to obtain O^2_{axis}.

Table 3. Composition of feedstock.[a]

Feedstock	Cellulose	Hemicellulose	Lignin
Corn stover	37.5	22.4	17.6
Corn fiber	14.28	16.8	8.4
Pine wood	46.4	8.8	29.4
Popular	49.9	17.4	18.1
Wheat straw	38.2	21.2	23.4
Switchgrass	31	20.4	17.6

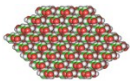
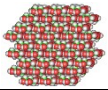
[a] Units: % of dry weight. Not listed are minor components. Data from reference [143] and references therein.

Table 4. REACH force constants at various temperatures.[a]

	Temperature [K]								
	100	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
k_{12}	584	580	56	563	548	535	527	517	507
k_{13}	50	48	45	41	34	29	27	24	21
k_{14}	9	7	3	3	2	2	2	1	1
k_{HB}	58	55	52	48	33	17	16	15	14
A	652	627	568	521	531	671	721	814	924
d_0	2.0	2.0	2.0	2.1	2.0	1.9	1.8	1.7	1.6

[a] Force constants in units of kJ/mol Å^2, d_0 in Å.

Table 5. REACH force constants for two different models of cellulose.[a]

	Model	
	A	**B**
		
κ_{12}	834	831
κ_{13}	36	36
κ_{14}	4	4
κ_{HB}	33	32
A'	48	44
d_0'	8	8
β	10.9	10.5

[a] The older CHARMM cellulose force field by Kuttel et al. [43] and a stretched-exponential nonbonded model function $\kappa_{nb}^{fit}(d)$ was used.

Table 6. Performance of atomistic (AA) and coarse-grained (CG) simulation using REACH.[a]

Length (DP)	CG	AA	Speed increase
40	172.5	7.2	24
80	88.1	3.7	24
160	35.2	1.6	22

[a] Units in ns/day. Also given is the factor increase in speed

FIGURES

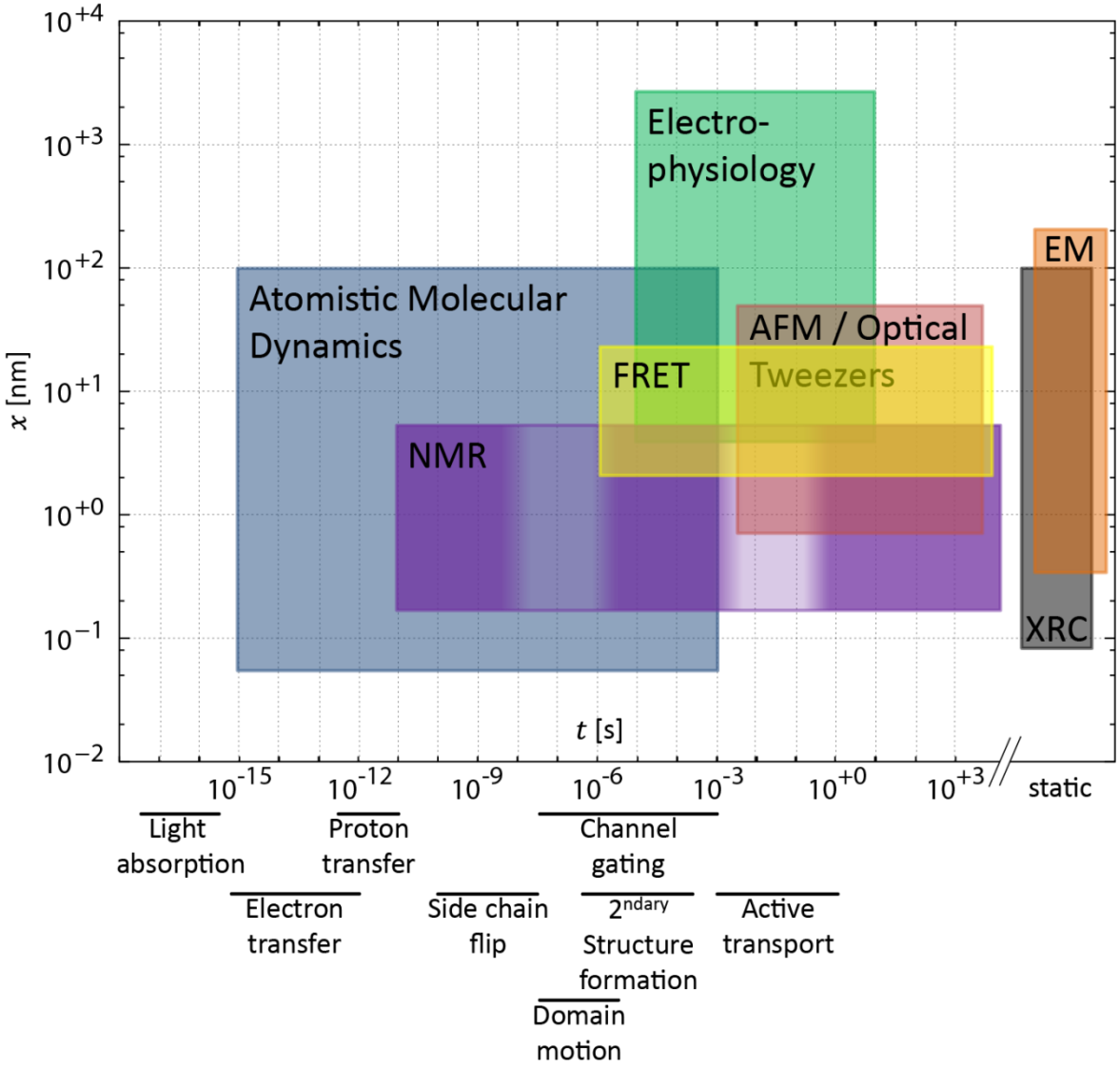


Figure 1. Time- and length scales (t and x, respectively) of various experimental techniques. Also shown is the time scale of dynamic processes in macromolecules, specifically proteins (bottom). Non-standard abbreviations: X-ray crystallography (XRC), electron microscopy (EM). Figure adapted from references [241, 242].

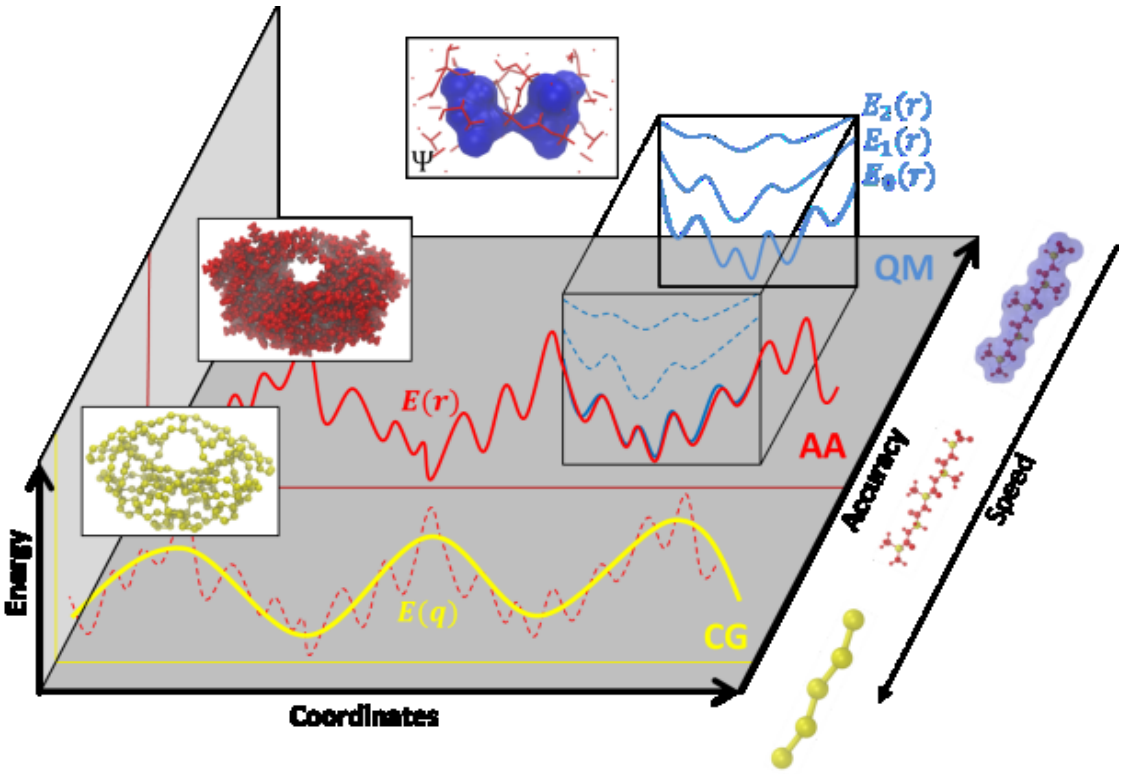


Figure 2. Illustartion of modeling at different scales. (Blue, QM) Quantum mechanical modeling also considers electronic degrees of freedom and is limited to simulation of tens of atoms for picoseconds. Ground or excited states can be addressed. (Red, AA) All-atom modeling approximates the QM-ground state energy and only implicitly treats the effect of electronic degrees of freedom in terms of an empirical potential energy function and parameters. (Yellow, CG) Coarse-grained modeling further reduces the degrees of freedom and simulates beads representative of a group of atoms using an effective potential.

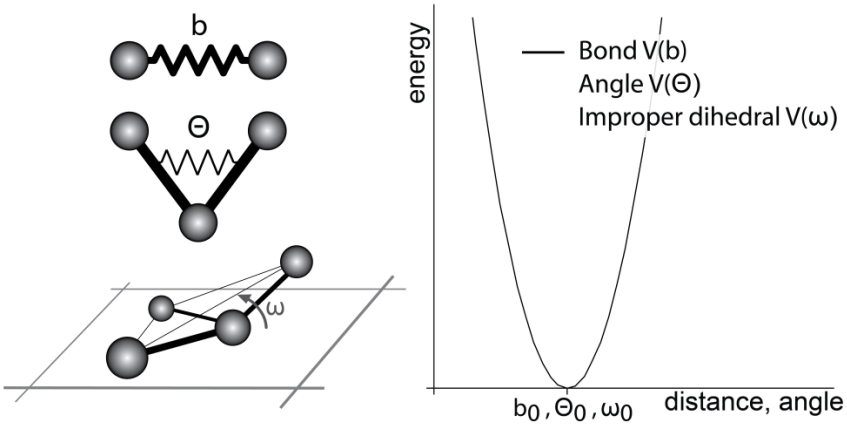


Figure 3. Harmonic potential for bond stretching (V(b)), angle bending (V(Θ)), and improper dihedrals (V(ω)).

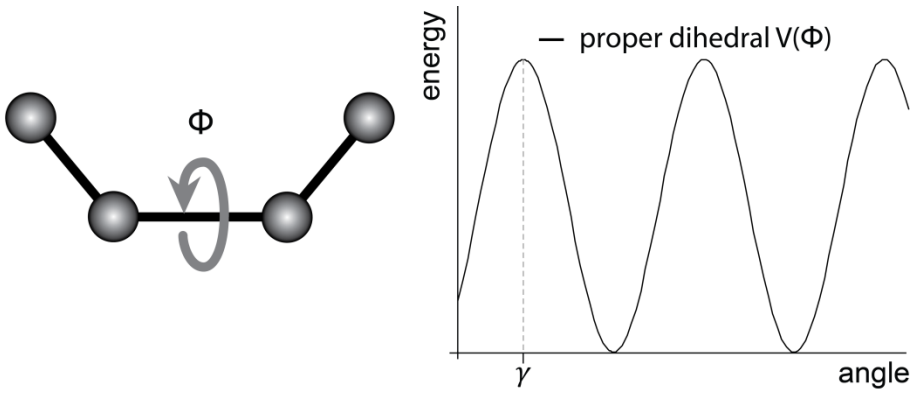


Figure 4. Proper torsional potential ($V(\Phi)$) when expanding the cosine series until $n = 1$.

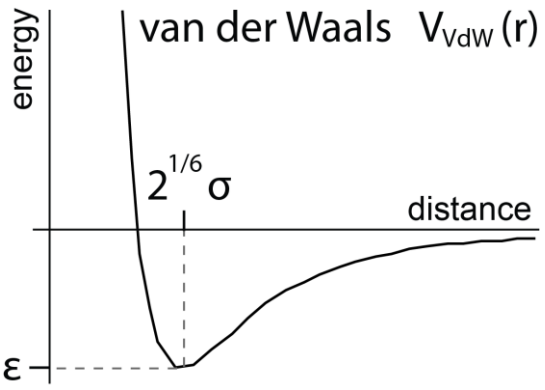


Figure 5. Van der Waals potential. Energy minimum of depth ε at distance $2^{1/6}\sigma$.

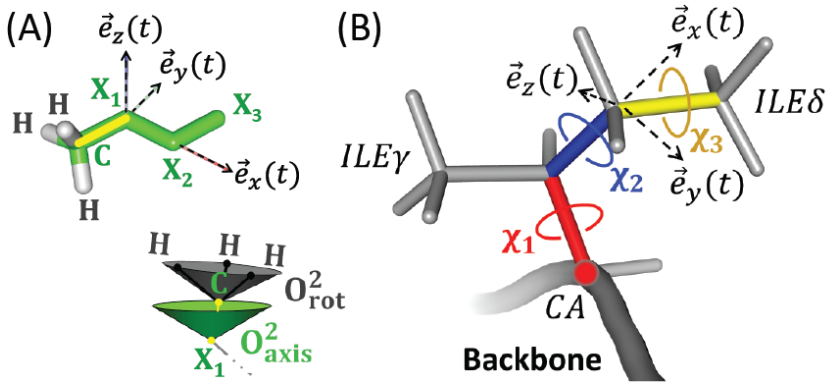


Figure 6. (A) Illustration of a methyl group on a side chain. The methyl group axis is shown as yellow line between atoms C and X_1. X_i represents any eligible atom type. A local coordinate system is shown spanned by the vectors $\vec{e}_i$. Methyl group order parameters O^2_{axis} and O^2_{rot} are also illustrated. (B) Illustration of an ILE residue and of side chain dihedral angles χ_i influencing the ILEδ methyl group. ILEδ methyl group axis is shown in yellow.

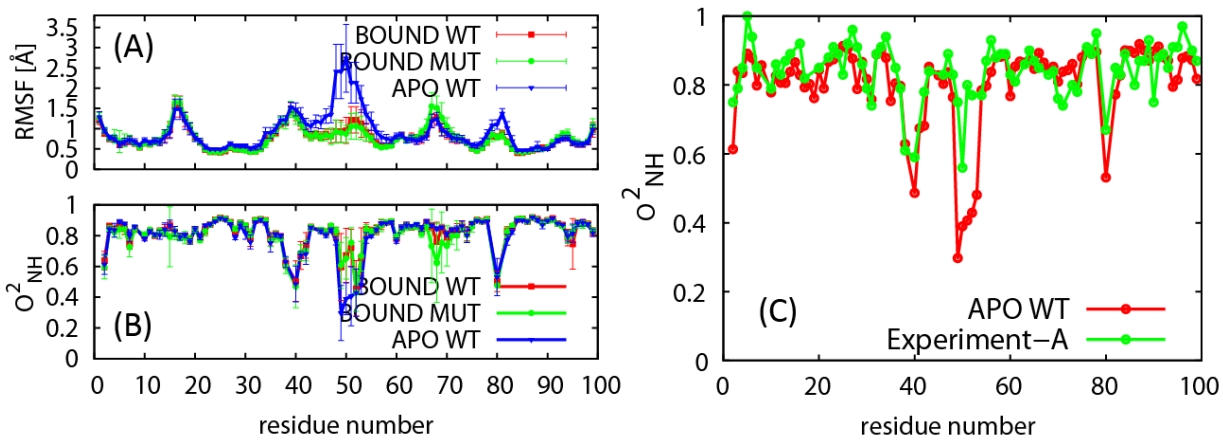


Figure 7. Simulated (A) root mean-square fluctuation and (B) backbone amide order parameters as function of residue number. (C) Comparison between backbone amide order parameters from simulation and experiment for apo wild-type HIV protease.

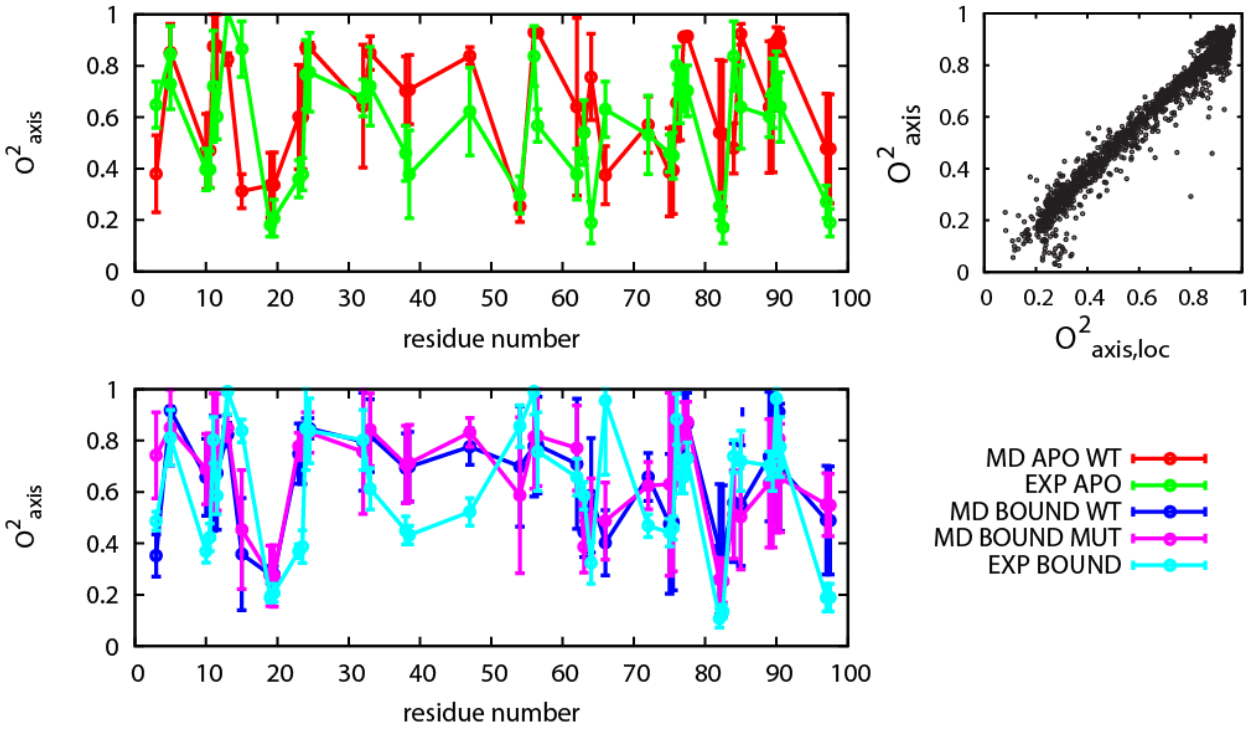


Figure 8. Side chain order parameter from simulation and experiment as function of residue number. Only residues with experimental data available are shown. Error bars from simulation estimated based on differences among simulated replicas. Also shown is the correlation between the axis order parameter calculated in a local and protein frame of reference (right).

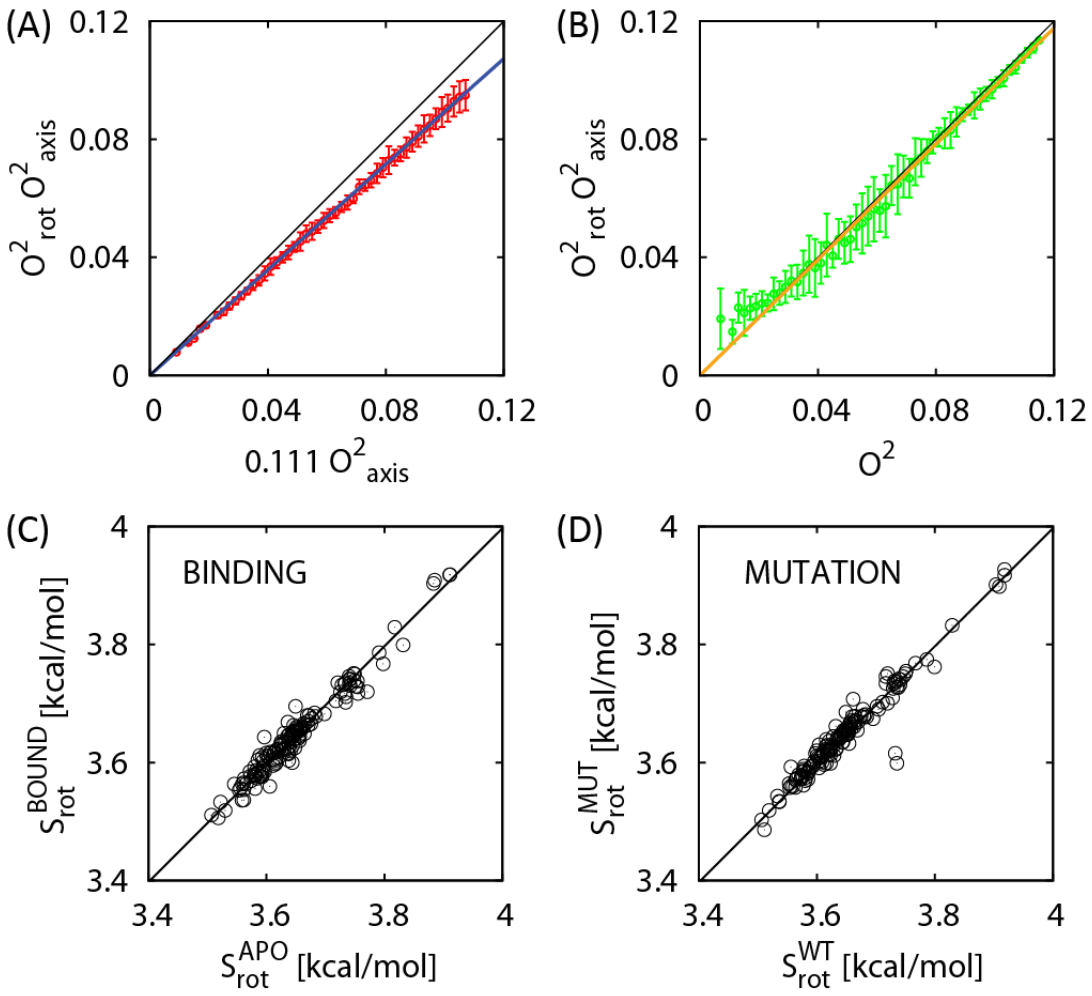


Figure 9. (A) The order parameter 0.111 O^2_{axis} *i*s shown against $O^2_{rot}\,O^2_{axis}$ (red). A linear fit (blue) and the diagonal (black) are shown as full lines. The inset shows the probability distribution of O^2_{rot}. (B) The order parameter O^2 is shown against $O^2_{rot}\,O^2_{axis}$ (green). A linear fit (orange) and the diagonal (black) are shown as full lines. (C) The methyl group rotational entropy from apo and ligand-bound wild-type simulations. (D) Same as C but from apo wild-type and mutant simulations. One point represents one methyl group in C and D.

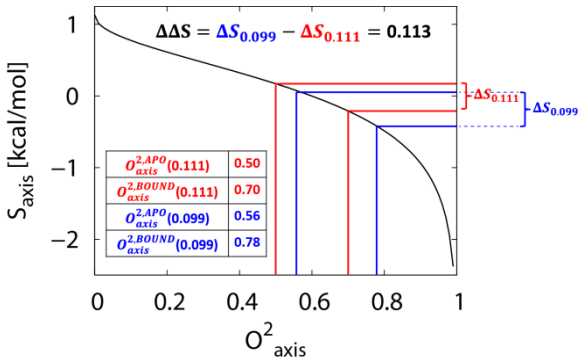


Figure 10. Diffusion-in-a-cone relation to obtain entropy from order parameter (T=310 K) [95]. Also shown is a numerical example of the ΔΔS when using 0.099 instead of 0.111 as value of O^2_{rot}.

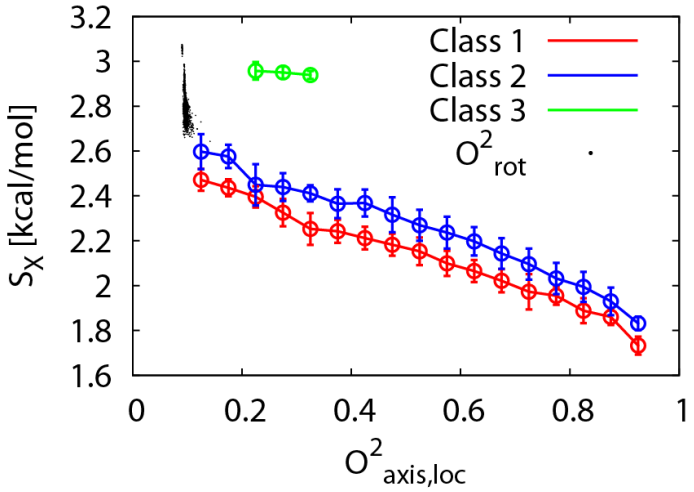


Figure 11. Methyl group axis conformational entropy as function of order parameter for Classes 1 to 3. Values of the axis order parameter are the mean in a bin of width 0.05; error bars are the corresponding standard deviation. Also shown are values of the rotational order parameter around 0.111.

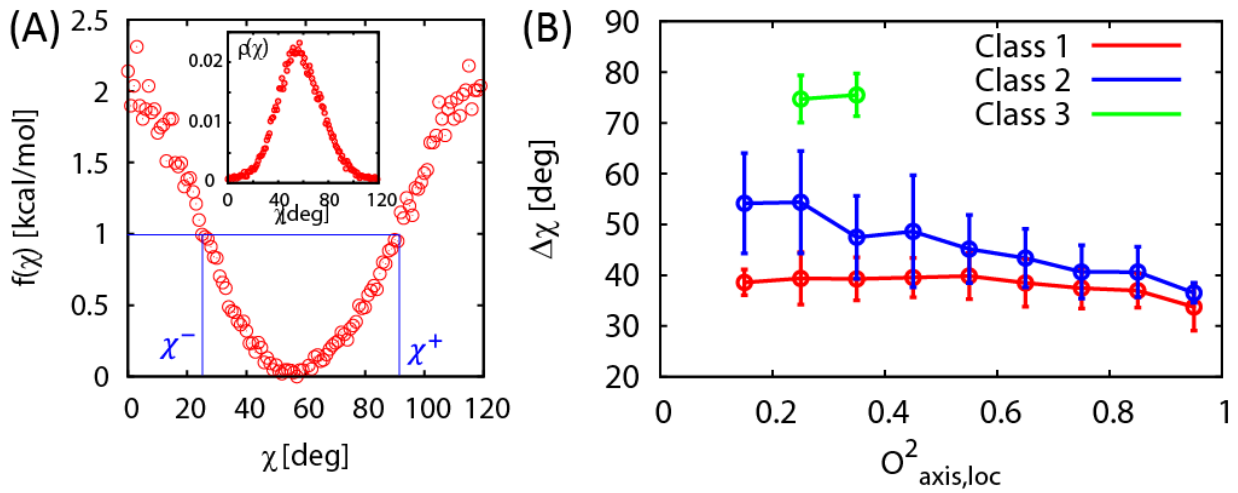


Figure 12. (A) Potential of mean force, f(χ), around one dihedral energy minimum and corresponding probability distribution, ρ(χ) (inset). The values of χ^+ and χ^- are defined by the angles at which the PMF is 1 kcal/mol larger than at its minimum. Shown data is from a MET residue. (B) Average width $\Delta\chi=\chi^+-\chi^-$ as function of order parameter. Averages taken for bins of width 0.1 of the order parameter. The three classes are shown separately.

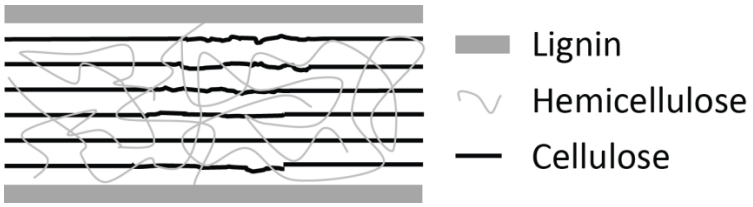


Figure 13. Crystalline and amorphous cellulose, hemicellulose, and lignin form complex structure.

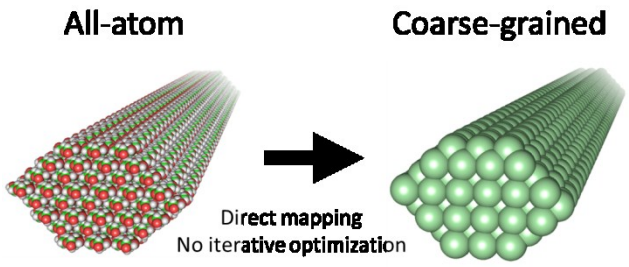


Figure 14. Illustration of the mapping of the atomistic force field onto the REACH coarse-grained model.

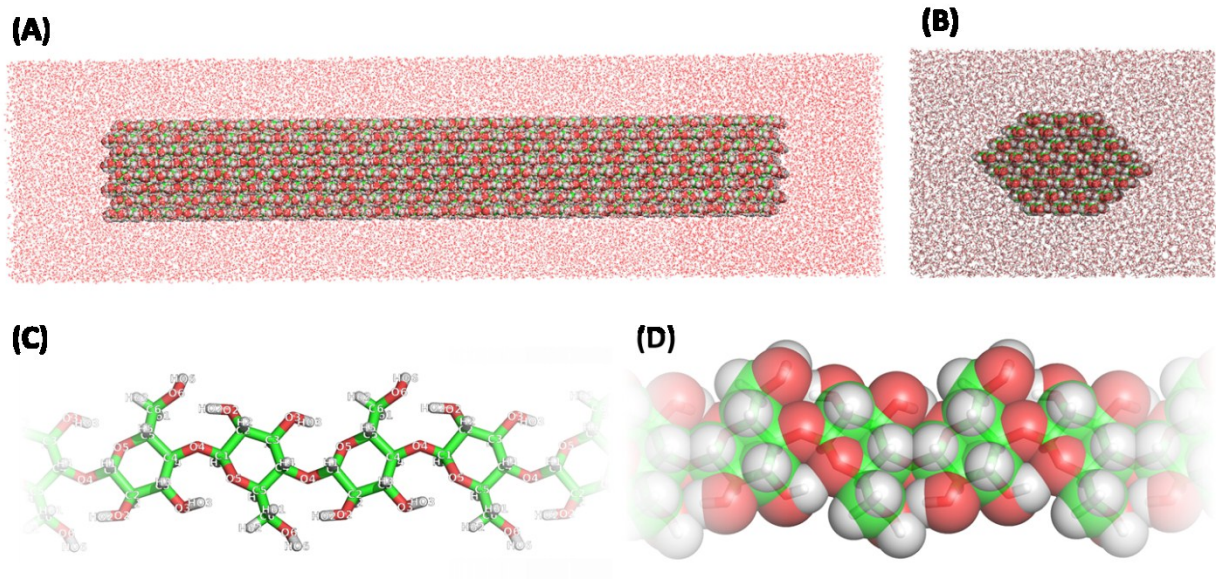


Figure 15. Illustrations of the atomistic system. (A) Side view of the cellulose fibril in a box of water. (B) Cross-section view of the cellulose fibril. (C) Top view of part of a chain shown as sticks. Atom labels are also shown. (D) Same as (C), but in van der Waals representation.

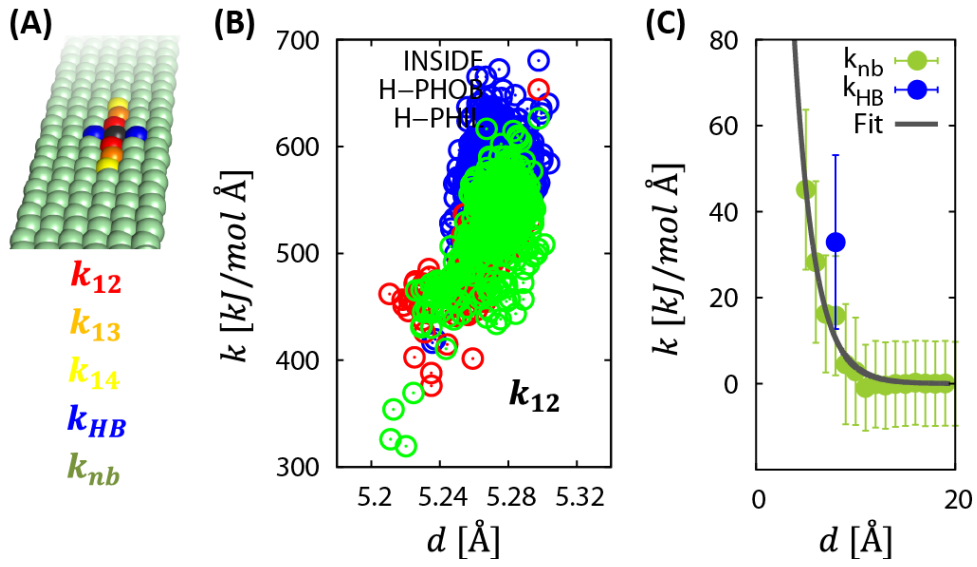


Figure 16. (A) Classes of force constants k that are considered: k_{12} (black-red), k_{13} (black-orange), k_{14} (black-yellow), k_{HB} (black-blue), and the remaining k_{nb} (black-green). (B) Elementary force constants k_{12} as function of distance. (C) The data points for k_{nb} obtained as the average of elementary force constants within a bin of width 1 Å. The full line shows a fit of Eq. 23 to the data.

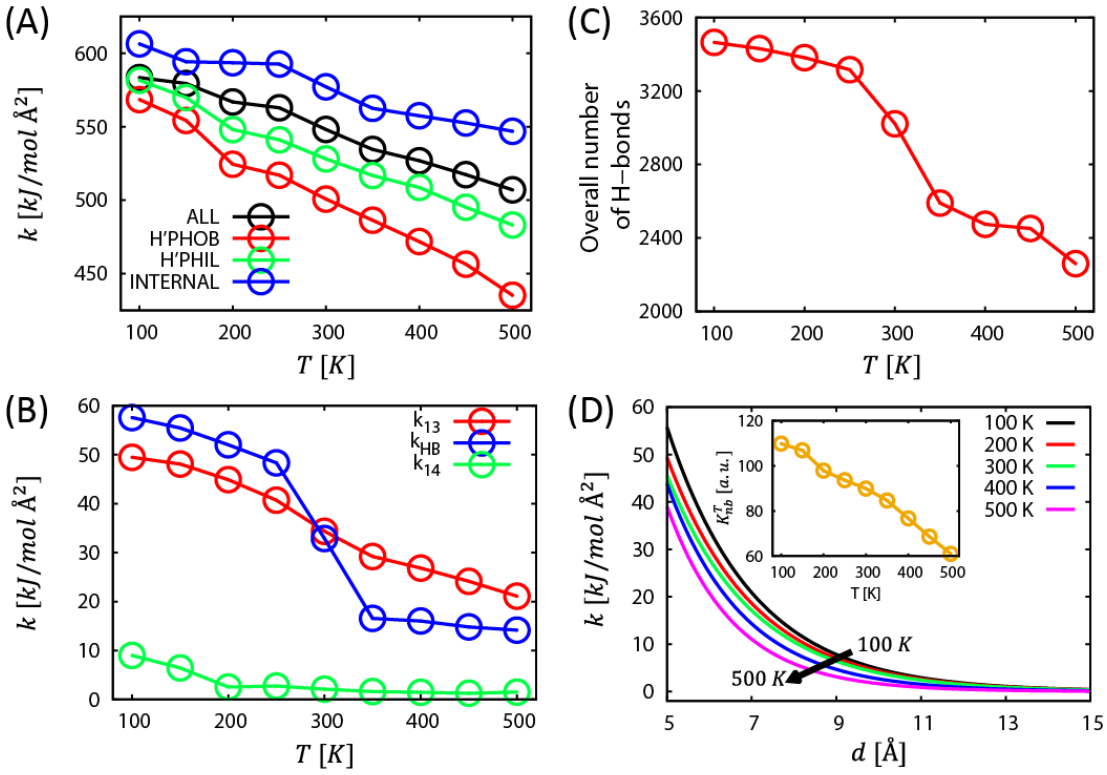


Figure 17. Temperature dependence of force constants for (A) k_{12} and the components of k_{12} for the internal and the hydrophobic and hydrophilic surfaces of the fibril and for (B) k_{13}, k_{14}, and k_{HB}. (C) Total number of hydrogen bonds in the fibril. (D) The nonbonded force constant model functions and the integral of those between 5 and 12 Å (see inset) for various temperatures

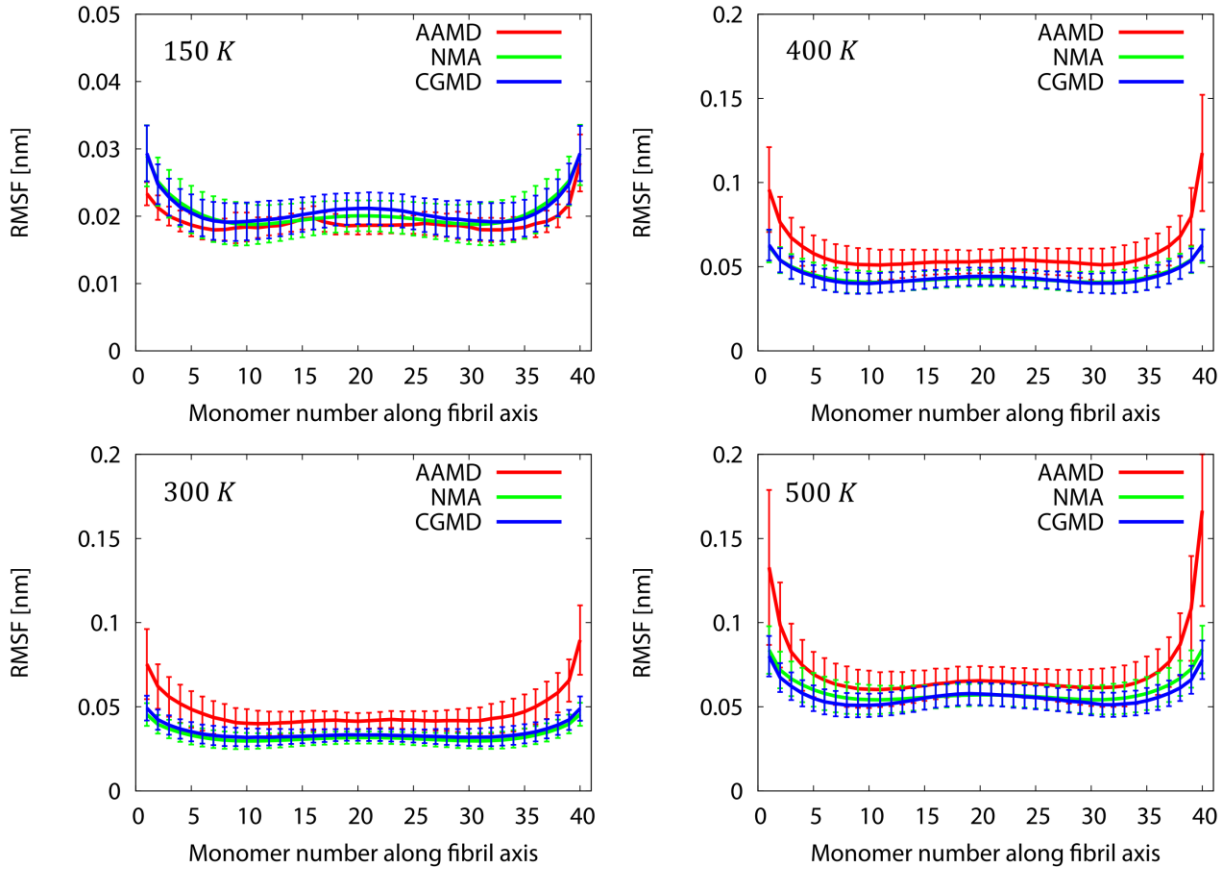


Figure 18. Comparison of the root mean-square fluctuation from AA MD with that from REACH CG MD and REACH NMA at representative temperatures. Values of the RMSF are the average over all CG-beads at a given position along the fibril and the error bars represent the corresponding standard deviation.

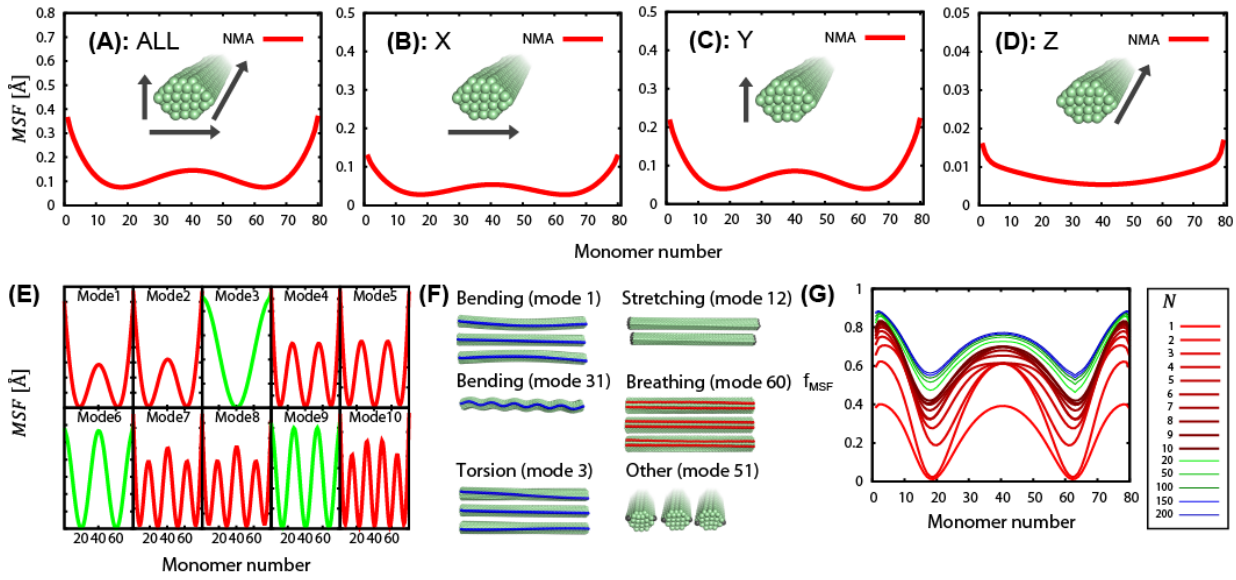


Figure 19. MSF from a different REACH NMA, *i.e.*, of a fibril 80 monomers in length. Total MSF (A) and MSF in X (B), Y (C), and Z (D) direction. (E) Shape of the contribution to the MSF of the lowest 10 normal modes. Red and green color illustrates bending and twisting modes, respectively. (F) Illustration of shape of normal modes, listing is non-exhaustive. (G) Cumulative fraction of MSF from the N lowest modes.

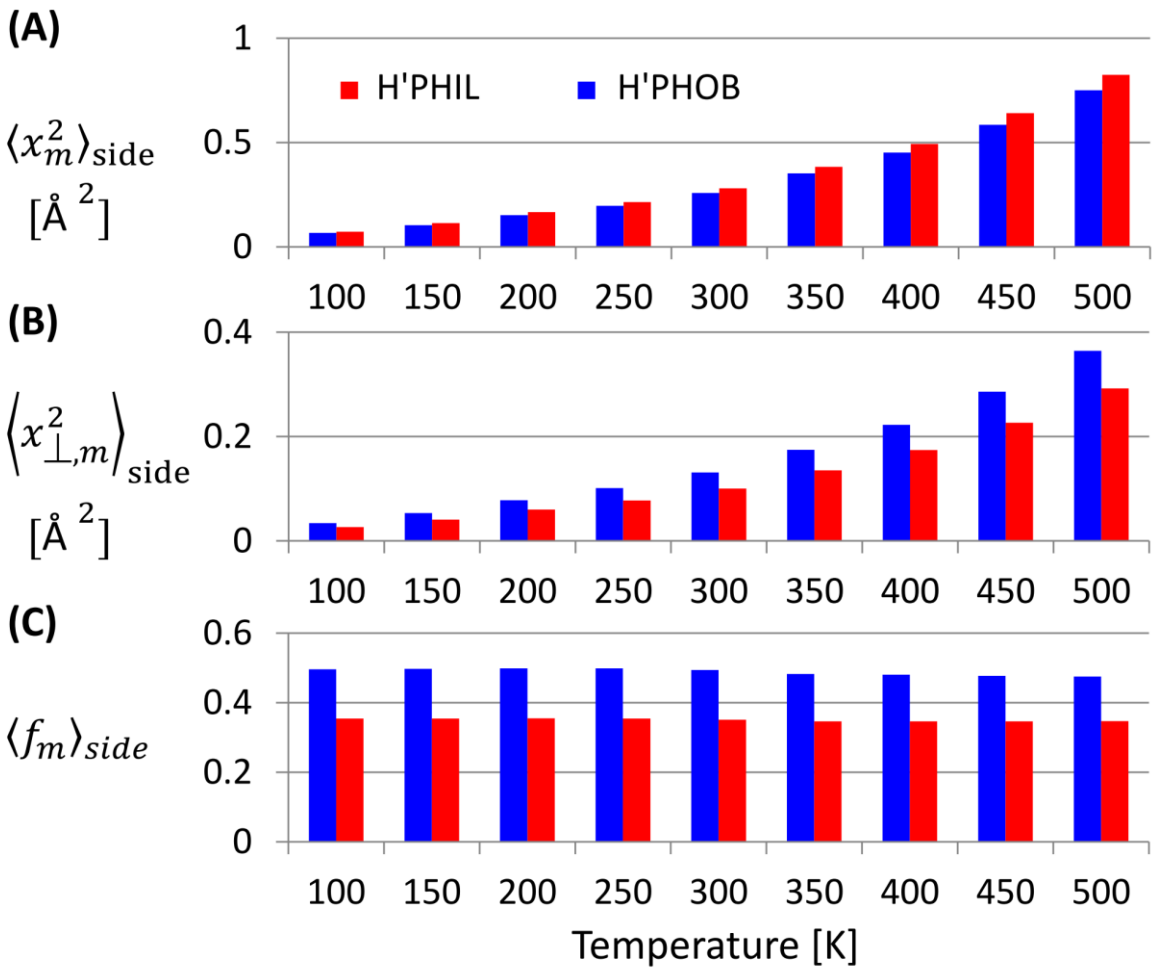


Figure 20. (A) Average mean-square fluctuation (MSF), (B) effective MSF, and (C) fraction between effective and average MSF. All data is shown as function of temperature for CG-beads from the hydrophilic (red) and hydrophobic (blue) surface.

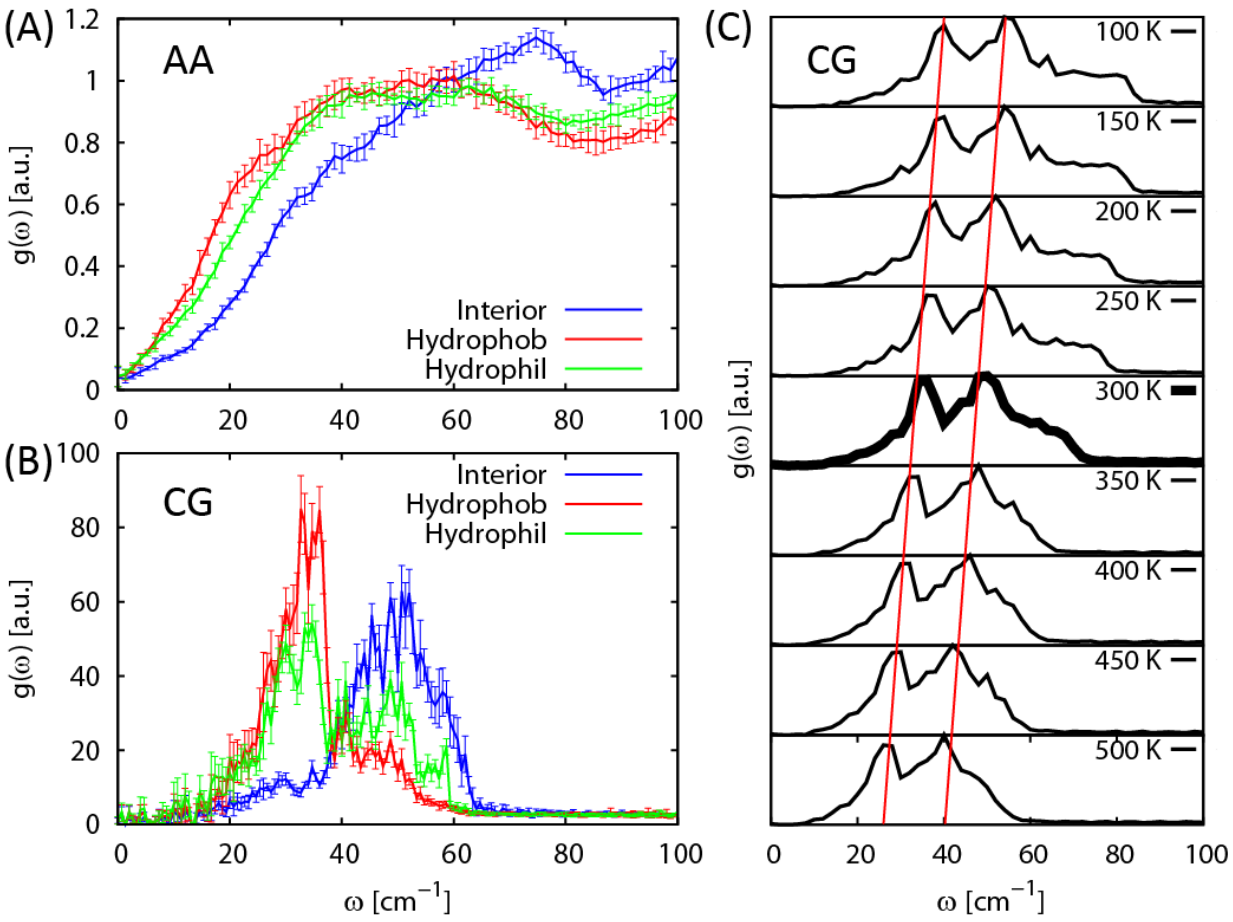


Figure 21. (A) Density of states for the interior of the fibril (blue) and for the hydrophobic and hydrophilic surface (red and green) from AA MD. (B) Density of states as in panel A but from CG MD. (C) Estimate of density of states for the entire cellulose fibril for various temperatures, obtained by counting the frequencies of CG normal modes in bins of width 2 cm^{-1}. Red lines as guide to the eye to illustrate red-shift of the density of states with increasing temperature.

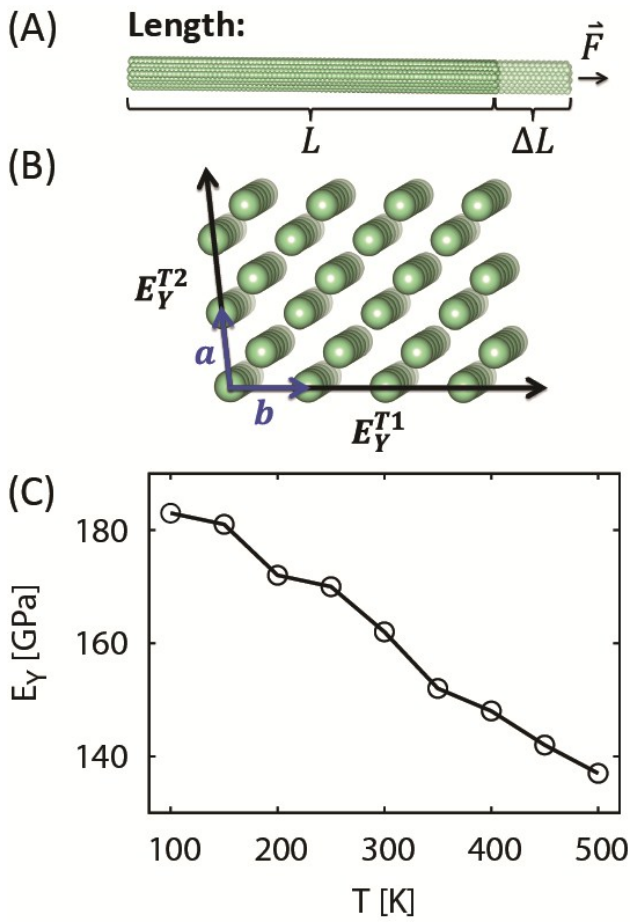


Figure 22. (A) Illustration for the relative change in length ΔL/L obtained by performing separate pulling simulations. (B) Illustration of transversal base vectors. (C) Temperature dependence of Young's modulus.

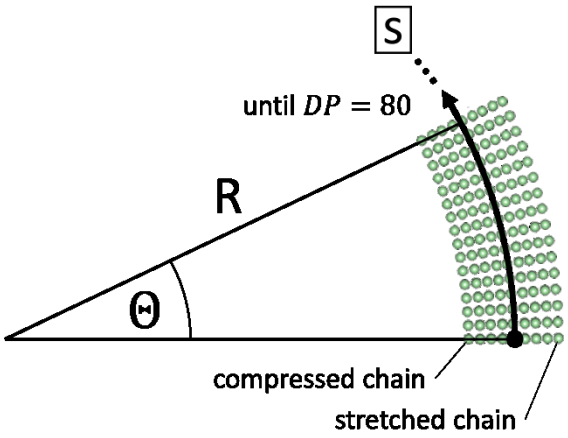


Figure 23. (A) Illustration of bended configurations. Values of R between 450 and 10^6 Å were used at constant fibril length (DP=80).

VITA

Dennis Christian Glass was born in Temeschburg, Romania, in 1984, but lived in Germany since age 5. Dennis received his Diploma in Physics at the Ruprecht-Karls University of Heidelberg, Germany, in 2008. In 2009, he joined the Genome, Science & Technology program at the University of Tennessee and Oak Ridge National Laboratory to obtain his doctoral degree in Life Sciences in 2012.