Supplementary Information

Common Processes Drive the Thermochemical Pretreatment of Lignocellulosic Biomass

Paul Langan, [a,b,c] Loukas Petridis, [d,e,f], Hugh M. O'Neill, [a,c] Sai Venkatesh □ingali,[a,c] Marcus Foston,[j] Yoshiharu Nishiyama, [h] Roland Schulz, [d,e,f] Benjamin Lindner, [d,e,f] B. Leif Hanson, [b] Shane Harton, [i] William T. Heller, [a] Volker Urban, [a,c] Barbara R. Evans, [j] S. Gnanakaran, [k] Arthur J. Ragauskas, [g] Jeremy C. Smith,[d,e,f] Brian Davison. [d*]

[a] Biology and Soft Matter Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.

[b] Department of Chemistry, University of Toledo, Toledo, OH 43606 USA.

[c]Center for Structural Molecular Biology, Oak Ridge National Laboratory, Oak Ridge TN 37831, USA

[d]Bioscience Division, Oak Ridge National Laboratory, Oak Ridge TN 37831, USA.

[e]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville TN 37996, U.S.A.

[f]UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge TN 37831, USA.

[g]Institute of Paper Science and Technology, School of Chemistry and Biochemistry and Georgia Institute of Technology, Atlanta, GA, 30332, USA.

[h]Centre de Recherches sur les Macromolécules Végétales (CERMAV-CNRS), BP 53, F-38041Grenoble Cedex 9, France.

[i]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA.

[j]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge TN 37831, USA.

[k]Theoretical Biology & Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

[1]Current address: Department of Energy, Environmental & Chemical Engineering, Washington University, St. Louis, MO 63130, USA.

[2]Current address: Pall Corporation, 25 Harbor Park Drive, Port Washington, NY 11050

Materials and Methods

Biomass Samples. Quaking aspen (*Populus tremuloides*) samples were obtained by Benchmark International in High Level, Alberta, Canada. Trees were destructively sampled to collect approximately 20-80 mm thick disks, or cross sections, from individual trees at 0.3 meters from the point of germination. All chips were taken from the outer portion of the section, or the mature wood, to minimize variability within a single chip. Chips had the following dimensions: 50.8 x 38.1 x 12.7 mm (2 x 1.5 x 0.5") in length, width, and thickness, respectively. The fiber axis ran along the length of the chip.

Steam Explosion Pretreatment (SEP). An aspen chip was placed into a woven metal mesh (Particle-sifting woven wire cloth type 316, 8 x 8 mesh, 0.025" wire diameter, McMaster Carr, CA, USA) basket, which was then suspended in a 4L Hastelloy steam reactor in the fashion of a tea bag, without the addition of any external catalysts. Steam for pretreatment was provided by a Fulton steam boiler (FB-075-L, Fulton Companies, Pulaski, NY, USA), which was controlled by setting the boiler pressure to the saturated steam pressure corresponding to the target temperature of 180°C. Pretreatments were performed at 180°C for 28 minutes. At the end of the reaction time, the temperature and pressure was suddenly dropped by opening a valve at the bottom of the vessel, during which all pretreatment liquid was discharged and not collected for analysis.

Dilute Acid Pretreatment (DAP). An aspen chip was presoaked in a 1% dilute sulfuric acid solution at 5% dry solids (w/w) for 4 hours at 25^{o}C in a glass beaker with continuous stirring on a temperature controlled stir plate. The chip was then washed with deionized water, and placed in a glass tube with fresh sulfuric acid solution, the tube was placed in a pressure vessel, and the sealed pressure vessel was placed in a sand bath furnace. The temperature of the furnace was raised to 160^{o}C over 30 minutes, held at this temperature for 10 minutes, and then quenched in a

water bath at 20°C. The sample was then placed in a vacuum desiccator and excess solvent removed at room temperature.

Ammonia Fiber Expansion (AFEX) Pretreatment. Conventional AFEX pretreatment conditions (*9*) were modified to maximize cellulose accessibility and simultaneously produce cellulose III_I (*5*). An Aspen chip was placed in a glass tube, within a pressure vessel, and then covered in liquified ammonia at a loading of about 1:10 (gram of biomass to milliliter of anhydrous ammonia). The sealed pressure vessel was placed in a sand bath furnace, the temperature of the furnace was raised to 140°C over 30 minutes, and then the temperature was held constant for 30 minutes. The ammonia gas was then released rapidly while the pressure vessel was still hot. The sample was allowed to cool to room temperature in the open reaction vessel.

Bleaching. Lignin was removed to leave holocellulose by treating aspen chips in 1% sodium chlorite solution at 70˚C for two days, at a pH below 5 by adding acetic acid.

X-ray Fiber Diffraction. Thin (~100μm in thickness) shavings were cut from the untreated and pretreated aspen chips and used to collect fiber diffraction data with an in-house Rigaku FR-E with R-Axis IV++ detector; λ was 1.5418Å; sample to detector distance was 15cm; 15 minute exposure time. The resulting images are shown in figure S4, with the meridional (fiber) direction vertical and the equatorial direction horizontal. For each image, first the beam center, sample tilt, and sample rotation were determined and then the image was transformed into polar reciprocal space coordinates d^* and ϕ, where ϕ is the azimuthal angle around the beam direction (ϕ has values of 0° and 180° in the equatorial direction and 90° and 270° in the meridional direction) and d^* is the reciprocal is the real space distance, d, where $\lambda = 2d\sin\theta$ and $q = 2\theta/d$. At each

value of d^* the intensity distribution in ϕ was fitted, using least-squares refinement, by Gaussian profile

$$G = I_A \exp\left(-4\ln 2\left(\frac{\phi - \phi_0}{\Delta\phi}\right)^2\right)$$

where the peak intensity I_A, peak position φ_0 and orientation fluctuation $\Delta\phi$ were refined together with an isotropic background, B(d^*) (which varies with d^* but is constant with ϕ). φ_0 and $\Delta\phi$ was constant for diffraction peaks, but fluctuated along the weak equatorial streak. Average values of φ_0 and $\Delta\phi$ were therefore determined for each diffraction pattern and applied to extract the final I(q) and B(q). The resulting fitted points for I(d^*) and B(d^*) are shown in red and green, respectively, in figure S2.

The diffraction from the untreated, steam explosion pretreated, and dilute acid pretreated samples corresponds to cellulose I, the collective term for two crystal allomorphs, cellulose I_α and I_β, that occur in naturally in aspen. Cellulose I_α has space group P_1 with reduced unit cell of a = 6.717 Å, b = 5.962 Å, c = 10.40 Å, α =118.08°, β = 114.80 °, and γ = 80.37 ° ; cellulose I_β has space group $P2_1$ with a reduced unit cell of a = 7.784 Å, b = 8.201 Å, c =10.38 Å, and γ = 96.5 ° (*10,11*). The (1 0 0), (0 1 0) and (1 1 0) reflections from I_α overlap with the (1 -1 0), (1 1 0) and (2 0 0) reflections from I_β, respectively, to produce three equatorial diffraction peaks that were used in radial profile width analysis of the lateral fibril dimensions (LFD). These three equatorial peaks are referred to by their cellulose I_β Miller indices. The diffraction from the AFEX pretreated sample corresponded to cellulose III_I which has space group $P2_1$ with a reduced unit cell of a = 4.45 Å, b = 7.85 Å, c = 10.31 Å, γ = 122.4° (*12*). The (010) reflection

and the overlapping (100) and (1-10) reflections produce two diffraction peaks that were used in radial profile width analysis of LFD. In this analysis both the position and profile of each reflection were fitted along with a diffuse scattering background, as shown in figure S2.

The meridional (002) and (004) reflections from either cellulose III_I or cellulose I_β were also fitted in a profile width analysis to obtain the unit cell *c* axis parameter and the axial correlation length (ACL) along the fiber chain direction. No correction for instrumental broadening was applied and we therefore expect the values of LFD and ACL to systematically underestimated, especially for the ACL where the corresponding meridional peaks are very sharp. We treat the values of ACL and LFD as useful only for comparison amongst different pretreatments. The extracted parameters are shown in table S1.

X-ray fiber diffraction from native aspen contains diffraction features from crystalline elemental fibrils that have LFDs of 1.8 - 3.1 nm and ACLs of 28.9 nm. Fibrils are preferentially aligned with an angular distribution width of 32^o. A diffuse ring of background scattering peaks at ~4.5Å, which is close to the expected hydrophobic stacking distance (d_{200}) of glucose and may therefore be due to less well-ordered cellulose and/or hemicellulose. Diffraction features become sharper after pretreatment due to an increase in the LFD to 3.2 - 3.9 nm, 3.4 – 4.4 nm, and 5.8 – 6.7 nm after DAP, SEP and AFEX, respectively, Table S1. Interestingly, there are slight decreases in distances between hydrophobically stacked planes in the cellulose I form after DAP and SEP. The changes in crystal structure are more striking after AFEX and correspond to a conversion from cellulose I to cellulose III_I. The ACL is unchanged after AFEX, but decreases after DAP and SEP. The changes in ACL are inversely correlated to

changes in the length of the *c*-axis parameter along the fiber axis direction. The widths of the orientation distribution of fibrils decrease after pretreatment, most significantly after DAP and SEP, indicating that increased alignment. The diffuse rings are less pronounced after pretreatment, indicating decreases in the relative amounts of less-ordered cellulose and/or hemicellulose.

Small Angle Neutron Scattering (SANS). SANS measurements were carried out on thin (~1mm thick) circular wafers of 10mm diameter tangential section cut from the untreated and pretreated aspen chips. The wafers were sealed in quartz "Banjo" cuvettes (0.5mm and 14mm inner thickness and diameter) and with either D_2O or a mixture of 40% D_2O and 60% H_2O (representing the average neutron scattering contrast match point for biomass). For comparison some dried Aspen chips were also measured. All data were collected on the CG-2 instrument at the High Flux Isotope Reactor (HFIR) facility of Oak Ridge National Laboratory. A wavelength, λ, of 4.75 Å, a neutron spread $\Delta\lambda/\lambda$ of 0.15, and three sample-to-detector distances of 0.3m, 6m and 18.5m were used to cover a q range of 0.003 $Å^{-1}$ to 1 $Å^{-1}$, where $q = 2\theta/d$ as discussed above. The center of the area detector was offset to allow access to larger values of q. The instrument resolution was defined using a circular aperture diameter of 10mm diameter. The detector images were normalized to incident beam monitor counts, and corrected for detector dark current, pixel sensitivity and scattering from backgrounds. The images were then remapped into polar coordinates and then averaged in ϕ over sectors to produce intensity scattering curves, $I(q)$, in the meridional and equatorial directions, as shown in figure S3. For native aspen soaked in D_2O there is a distinct peak in the equatorial scattering at q ~0.17$Å^{-1}$, indicative of diffraction from elongated scattering objects oriented in the fiber axis direction

with a lateral side-by-side packing distance of ~3.7nm. Soaking biomass in D_2O causes exchange of labile hydrogen atoms by deuterium in the accessible hydrated lignin-hemicellulose matrix gel and disordered cellulose regions, but does not cause exchange in the crystalline cellulose regions (*15*). When the sample was soaked in 40% D_2O so that the scattering from water matched that from cellulose this diffraction peaks disappears. The origin of the diffraction peak must therefore be in the scattering contrast between side-by-side packing of aligned cellulose fibrils with water in between.

Significant changes in SANS equatorial scattering occur after pretreatment in the q range of $0.3Å^{-1}$ to $0.01Å^{-1}$, indicative of morphological changes over the range of 2nm to 60nm. The equatorial diffraction peak from cellulose broadens and moves to a lower q value of ~$0.07Å^{-1}$ for the SEP and AFEX samples and to ~$0.17Å^{-1}$ for the DAP samples. This corresponds to a large increase in the lateral side-by-side packing distance of fibrils to ~9nm after AFEX and SEP, but a smaller increase to 4nm after DAP. Computational studies discussed below provide a physical understanding of the processes leading to the increase of the side-by-side distance. We also collected SANS from the same SEP sample after bleaching to remove lignin and found that the equatorial scattering was similar, but with a stronger peak that had moved to slightly lower q, corresponding to 10nm. This confirms our interpretation of the equatorial diffraction peak as being due the cellulose. The cellulose peak is broader after pretreatment indicating a larger distribution of packing distances. However, this broadening may also reflect the emergence of a Guinier regime in which scattering from individual cellulose fibrils, or their bundles, becomes significant.

SANS meridional scattering also changes after pretreatment in the q range of 0. 2Å^{-1} to 0.01Å^{-1}, indicative of morphological changes in the range of 3nm to 60nm. Features emerge that can be interpreted as Guinier regions with the most pronounced shoulder appearing after DAP. In SANS collected after SEP and then bleaching to remove lignin, the Guinier feature is considerably diminished and we therefore interpret it as being due mainly to the emergence of lignin aggregates. It is likely that this lignin scattering feature is isotropic but overwhelmed by the much stronger scattering contrast contributions from oriented cellulose on the equator. Accurately estimating the radius of gyration, R_g, for the lignin aggregates is difficult because of uncertainties associated with estimating the underlying scattering curve. For DAP R_g is in the range of 20 – 30nm, and it is slightly less for the SEP and AFEX samples. One question, addressed below by computational approaches, is whether lignin redistribution into aggregates requires cleavage of most lignin:hemicellulose crosslinks or if it is possible to have phase separation of lignin and hemicellulose while the two polymers remain crosslinked.

Compositional Analysis. HPLC based anion-exchange chromatography was performed on native and pretreated wood chips of aspen. Figure S1 summarizes the variation in normalized cell wall composition due to different pretreatments. The carbohydrate content includes the glucose distribution mainly representing the proportion of cell wall cellulose while the xylose, mannose, arabinose, and galactose distributions constitute the monomers forming major hemicellulose polysaccharides. The residual acid-insoluble material is referred to as the Klason lignin content.

Wood chip samples were size reduced by vibrational ball-milling at 20 Hz for 20 min. Samples for carbohydrate constituents and acid-insoluble lignin (Klason lignin) analysis were prepared using a two-stage acid hydrolysis protocol based on TAPPI methods T-222 om-88 with a slight modification. The first stage utilizes a severe pH and a low reaction temperature (72 vol. % H_2SO_4 at 30 °C for 1 h). The second stage is performed at much lower acid concentration and higher temperature (3 vol. % H_2SO_4 at 121 °C for 1 h) in an autoclave. The resulting solution was cooled to room temperature and filtered using G8 glass fiber filter (Fisher Scientific, USA). The remaining residue which is considered as Klason lignin was oven-dried and weighed to obtain the Klason lignin content. The filtered solution was analyzed for carbohydrate constituents of the hydrolyzed *Populus* samples determined by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD) using Dionex ICS-3000 (Dionex Corp., USA) (*16*).

DAP causes significant hemicellulose removal through acid-catalyzed hydrolysis, decreasing the relative xylose content found in native aspen from 24 to 12%. SEP also causes hemicellulose removal, through auto-catalyzed hydrolysis, although removing only approximately half the amount removed by DAP. AFEX leaves the cell wall composition unchanged. Pretreated samples, especially following DAP and SEP, display a slight increase in the relative glucose and Klason lignin content, which we attributed to the selective removal of hemicellulose. These trends are consistent with data reported in the literature, and indicate that different pretreatment strategies produce significantly different changes in biomass composition (*17*). However, the decrease in hemicellulose related monosaccharides is not as substantial as typically described even at short resonance times. This may be due to the

relatively large sample (particle) size used in this study. We believe that the wood chips used in this study, compared to the finer powdered or meshed samples used in others, are more representative of the biomass feed stock likely to be most used in a commercial biorefinery. It would appear that mass and heat transfer effects in these larger samples slows the kinetics of pretreatment with respect to finer ground samples.

Molecular Dynamics Simulations.

Cellulose Model. Seven hexagonal cellulose elementary fibrils of degree of polymerization 40 were constructed using the cellulose Iβ crystalline structure (*10*). The seven fibrils were arranged in a parallel manner with two hydration shells separating them initially. The system was then solvated in an orthorhombic simulation cell consisting of ~340,000 atoms. The simulated system consisted of 340k atoms.

Lignin and Hemicellulose Model. A structural model of a crosslinked lignin-hemicellulose copolymer was generated by using available experimental information on the average chemical composition of aspen lignin and hemicellulose (*3*). Aspen lignin is composed primarily of guaiacyl (G) and syringyl (S) monomeric units, with ratio S/G~ 1.86, connected by various linkages. Here, a linear lignin polymer was constructed that comprised 11 G and 20 S units connected *via* 22 β-O-4 (β-aryl ether), four 4-O-5 (bithenyl ether), two β-β (pinoresinol) and two β-1 linkages, similar to the experimentally determined average inter-linkage composition of aspen (3). Hemicelluloses are branched polymers, composed of sugar residues Here, the hemicellulose polymer consisted of a (1→4) linked backbone consisting of 28 β-xylose and four

β-manose monomers. Four 4-O-methyl-glucuronic acid (4-*O*-MeGlcA) side-chain monomers were bonded to the xylose backbone via a (1→2) link. The lignin and hemicellulose polymers were connected end-to-end *via* an ether bond between the γ carbon of the lignin and the O1 oxygen of the hemicelluose, resembling a ferulate lignin-hemicellulose crosslink (*40*). The lignin-hemicellulose polymer was then hydrated in a cubic box and four Na^+ ions were added to neutralize the system (4-*O*-MeGlcA, with pKa value of 3, was deprotonated). The system consisted of ~42,000 atoms when simulated at 27°C and 92,000 atoms at 160°C, the high temperature simulation requiring a larger simulation box to accommodate the extension of the polymer.

Molecular Dynamics Simulation Details. MD simulations were performed using the NAMD 2.9 software (*18*) and employed the CHARMM carbohydrate (*19,20*) and lignin (*21*) force fields and the TIP3P water model (*22*). Periodic boundary conditions were employed with the Particle Mesh Ewald method (*23,24*) with grid spacing of 1.1Å for the treatment of Coulomb interactions beyond 11Å and a force switching function to smoothly transition the Lennard-Jones forces to zero over the range of 10-11Å. Multiple timestep integration was used with timesteps of 2 fs for bonded and short-range nonbonded forces, and 4 fs for the long-range electrostatic forces. The neighbor list was updated every 10 steps with a pair-list distance of 12.5Å. The SHAKE algorithm (*25*) was used to constrain all covalent bonds involving hydrogen atoms to their equilibrium values. The simulations were performed in the NPT ensemble, and the temperature was kept constant using the Langevin dynamics algorithm implemented in X-PLOR with a damping coefficient of 5 ps^{-1}. The pressure was held constant at 1 atm using the Nosé-Hoover

Langevin piston algorithm (*26,27*) employing a piston oscillation period of 50 fs and a piston damping decay time of 100 fs.

The cellulose system was simulated for 50 ns at temperature T=27°C, then for an additional 60 ns at T=160 °C and finally for an additional 80 ns at T=27°C. The lignin-hemicellulose system was simulated for 300ns at 27°C. The structure of the copolymer at 211ns was then solvated in a larger water box and simulated for an additional 100ns at 160 °C. After 393 ns the system was then gradually cooled back to 27 °C in at a rate of 21.7 °C/ns, and thus the length of the cooling simulation was 6 ns. From 398 ns to 535 ns the system was subsequently simulated at a constant temperature of 27 °C. Unless otherwise stated, atomic coordinates were saved every 1 ps. All calculations were performed on the Hopper Cray XE6 supercomputer at the National Energy Research Scientific Computing Center. Analysis of all MD trajectories was performed with the VMD software (28) using local scripts. As found by Matthews et al. (38) the *tg* conformation of the primary alcohol, representative of the cellulose I_β cellulose allomorph, was not maintained in simulations of the CHARMM C37 force field employed here.

Cellulose X-ray diffraction spectra. Wide-angle X-ray diffraction spectra were calculated from the MD trajectory of the cellulose macrofibril using the DEBYER software (http://code.google.com/p/debyer/). WAXS spectra were first calculated from single frames saved every 1 ns, and were subsequently averaged over the time periods 5-50 ns and 110-160 ns, see Fig. 5A. During the first 50 ns of the simulation, when the temperature is kept constant at 27 °C, the (200) Bragg peak is broad and centered at q ~1.53 $Å^{-1}$ (i.e. d ~ 4.1 Å). For t=110-160ns, where the temperature was first increased to 160 °C and then cooled down to 27 °C, the (200)

peak becomes sharper and moves to q ~1.56 $Å^{-1}$ (i.e. 4.0 Å). The (200) peak narrowing and shift to higher q resulting from heating the system and then cooling it down to RT is qualitatively the same as the changes resulting from SEP pretreatment in the experimental diffraction data of Fig. S2.

Cellulose core water molecules. Core water molecules were defined as those with their oxygen atoms within 7 Å of the central elementary fibril and at a distance of at least 6 Å from the end of the fibril. For the calculation of the mean square displacement (MSD) of the core water molecules, the trajectory was first divided into forty-five 1 ns parts, from t=5-50ns. For each trajectory part, the MSD of a subselection of core water molecules, identified as being within 100 Å of the central fibril midpoint at the beginning of the trajectory part, was calculated. The same procedure was repeated for a 25 ns simulation of 4,692 water molecules in a cubic simulation box. The confined core water molecules display considerably slower dynamics compared to bulk water, manifested by the smaller mean-square displacements in Fig. S5B.

Hydrogen bonds. A geometric criterion was employed to determine the presence of hydrogen bonds (HB), in which two molecules are considered bonded if their hydrogen-acceptor separation is less than 3.5 Å and the donor-hydrogen-acceptor angle is greater than 150°. For the calculations of the hydrogen bonds of the core water molecules, the trajectory was first divided into forty-five 1 ns parts from t=5ns-50ns. For each trajectory part, the list of core water molecules was determined at the first frame of the part and was not updated subsequently. The same procedure was repeated for a 25 ns simulation of 4,692 water molecules in a cubic

simulation box. The dynamic behavior of the HB was examined by defining a binary function $h(t)$ equal to 1 when a HB is present and 0 otherwise. The autocorrelation function $C(t)$ of $h(t)$:

$$C(t) = \frac{\langle h(0)h(t)\rangle}{\langle h^2\rangle} \quad \text{(S1)}$$

was computed and averaged over all possible time origins and similar HB types: between core water and cellulose, between two core water molecules and between two bulk water molecules. Before SEP, the confined core water molecules display significantly slower breaking of hydrogen bonds than bulk, manifested by the slow dynamics compared to bulk water, and the slow decay of the HB autocorrelation function, C(t) (Eq. S1) in Fig. 2E. Consequently, the standard Gibbs energy of activation, ΔG‡, is higher for core water molecules than bulk. For $t \lesssim 0.1$ ps, $C(t)$ exhibits a decay due to librational motion. The initial decay is followed by a plateau where $C(t)$ remains approximately constant. A second decay region due to bond breaking by diffusion follows the plateau at $t \gtrsim 0.1$ ps for bulk water and $t \gtrsim 10$ ps for core water molecules. A relaxation time τ can be obtained by fitting $C(t)$ at long-time relaxation (t>10 ps) to a stretched exponential:

$$C(\tau) = exp\left(-(t/\tau)^{\beta}\right). \quad \text{(S2)}$$

For bulk-bulk HB β_{bb}=0.29 and τ_{bb}~1.4 ps, for core-core β_{ww}=0.30 and τ_{ww}~160 ps, and for core-cellulose β_{ww}=0.20 and τ_{bb}~780 ps.

Entropy of water molecules. The entropy of water molecules was calculated using a two-phase thermodynamic (2PT) model (30). For translational entropy, the translational velocity autocorrelation function (VACF)

$$C_{trn} = \sum_{i=1}^{W} m\langle \vec{v_i}(t) \cdot \vec{v_i}(0)\rangle \quad \text{(S4)}$$

was first derived, where W is the number of water molecules in the system, m is the mass and $\overrightarrow{v_i}(t)$ the velocity of the i^{th} water molecule at time t. The translational density of states was then obtained as the Fourier transform of $C_{trn}(t)$

$$g_{trn}(\omega) = \frac{2}{k_B T}\int_{-\infty}^{+\infty} C_{trn}(t)e^{-i\omega t}. \quad \text{(S5)}$$

Finally the water entropy was computed using the 2PT model that partitions the translational density of ate into gas-like and solid-like components. A similar procedure was performed for the rotational entropy. The rotational autocorrelation function is given by

$$C_{rot} = \sum_{j=1}^{3}\sum_{i=1}^{W}\langle I_j \omega_{ij}(t)\omega_{ij}(0)\rangle \quad \text{(S6)}$$

I_j and ω_{ij} are the j^{th} principal moment of inertia and angular velocity of water molecule i. The rotational density of states was then obtained as the Fourier transform of $C_{rot}(t)$

$$g_{rot}(\omega) = \frac{2}{k_B T}\int_{-\infty}^{+\infty} C_{rot}(t)e^{-i\omega t}. \quad \text{(S7)}$$

Differences in water dynamics between the bulk and the core water molecules in the evident in the translational and rotational VACFs in Figure S5. Negative values of the VACF are due to water molecules rebounding from collisions with their neighbors. The deep negative minimum in the core water molecule function thus arises from water confinement by the cellulose surface. A peak in the density of states represents the population of a mode of a given frequency. Therefore, the shift of the main peak of the core water molecules toward higher frequencies, more pronounced in the translational spectra, arises from water molecules on the surface of the

cellulose that are more tightly bound, and thus librate at a higher frequencies than does water in the bulk.

The translational and rotational diffusion constant

$$D = \frac{k_B T g(\omega = 0)}{12\, W\, m}$$

of the core water is significantly lower than the bulk, see Table S2. This 30x decrease in the translational *D* is consistent with the small MSD of core water molecules in Fig. S5B.

In order to employ the 2PT theory to calculate the translational and rotational entropy, the density of core and bulk water must be first obtained. While this is straightforward for bulk water, the definition of the density of core water is challenging because the volume core water occupies is ill-defined. However, the exact value chosen does not affect the qualitative results found. The translational and rotational entropy per water molecule is obtained in Table S3 assuming the density of core water to be the same as bulk (=33.73 molecules per nm^3).

Lignin and hemicellulose conformations. To probe whether phase separation of lignin and hemicellulose can occur while the two polymers remain crosslinked, simulations were also performed of a model copolymer in aqueous solution made of a lignin segment covalently bonded at its one end to a hemicellulose segment. Atomic contacts were defined as two atoms separated by less than 3 Å. The fraction, *f* of lignin monomeric units in contact with hemicellulose was computed and is shown in Fig. 3B.

To assess whether lignin and hemicellulose are in collapsed or extended coil conformations r_g of component segments was calculated. For hemicellulose at both 27 °C and 160 °C, r_g is found to follow a power-law behavior with $r_r \propto N^{\nu}$, in which the segment consists of (N+1) monomers, and the scaling exponent ν=0.68. This behavior is indicative of coil polymers in good solvent, for which the theoretical value of the exponent is ν=2/3.

For lignin at 27°C, r_g obeys a power law only up to N~20, with an exponent ν=0.36, and flattens for N>20. The presence of the plateau at large N and a value of the exponent close to ν=1/3 is consistent with theoretical predictions for collapsed "equilibrium" globules and similar to what has been found in previous simulations of isolated lignin in aqueous solution, *i.e.* in the absence of hemicellulose (*7*). However, at 160 °C the plateau disappears indicating the lignin acquires a "crumpled" globule conformation (*7*). The major difference between a "crumpled" and "equilibrium" globule is that whereas in the former lignin monomeric units that are distant along the chain are distant in space, for the latter monomeric units distant along the chain have a relatively high probability of being proximal in space. As a result "equilibrium" globules have kinks and "crumpled" globules do not (*8*).

Supplementary References and Notes:

S1. Himmel, ME, Ding, SY, Johnson, DK, Adney, WS, Nimlos, MR et al., *Science*, 2007, 315: 804.

S2. Alvira, P, Tomas-Pejo, E, Ballesteros, M, Negro, MJ, *Bioresource Technology,* 2010, 101**:** 4851.

S3. Sannigrahi, P, Ragauskas, AJ, Tuskan, GA, *Biofuels, Bioproducts and Biorefining.* 2010, 4: 209.

S4. Saito, T, Kimura, S, Nishiyama, Y, Isogai, A, *Biomacromolecules* 2007, 8:2485.

S5. Chundawat, SPS, Bellesia, G, Uppugundla, N, da Costa Sousa, L, Gao, D et al., *Journal of the American Chemical Society, 2011,* 133**:** 11163.

S6. Abe, K, Iwamoto, S, Yano, H, *Biomacromolecules,* 2007, 8: 3276.

S7. Petridis, L, Schulz, R, Smith, JC, *Journal of the American Chemical Society*, 2011*,* **133:** 20277.

S8. Grosberg, A, Rabin, Y, Havlin S, Neer, A, *Europhysics Letters.* 1993*,* **23**:373.

S9. Chundawat, SPS, Donohoe, B, Sousa, L, Elder, T, Agarwal, U, Lu, F, Ralph, J, Himmel, M, Balan V, Dale, B, *Energy and Environmental Sciences.* 2011, **4**: 973.

S10. Nishiyama, Y, Langan, P, Chanzy, H, *Journal of the American Chemical Society***,** 2002, **124**: 9074.

S11. Nishiyama, Y, Sugiyama, J, Chanzy, H, Langan, P *Journal of the American Chemical Society*, 2003, **125**:14300.

S12. Wada, M, Chanzy, H, Nishiyama, Y, Langan, P, *Macromolecules* 2004, **37**:8548 .

S13. Igarashi, K, Wada, M, Samejima, M, *FEBS J* 2007, **274 :**1785.

S14. Wada, M, Nishiyama, Y, Bellesia, G, Forsyth, T, Gnanakaran, S, Langan, P, *Cellulose* 2001, **18:**191.

S15. Nishiyama, Y, Kim, UJ, Kim, DY, Katsumata, KS, May, RP, Langan, P, *Biomacromolecules* 2003, **4**: 1013.

S16. Davis, M, *Wood Chemistry Technology* 1998, **18**:235.

S17. Kumar, R, Mago, G, Balan, V, Wyman, C. *Bioresource Technology*, 2009,**100**:3948..

S18. Phillips, JC, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, RD, Kale, L, Schulten, K, *J. Comput. Chem.* 2005, **26**:1781.

S19. Guvench, O, Greene, SN, Kamath, G, Brady, JW, Venable, RM, Pastor, RW, Mackerell, AD, *Journal of Computational Chemistry*. 2008, **29**:2543.

S120. Guvench, O, Hatcher, E, Venable, RM, Pastor, RW, Jr. MacKerell, AD *Journal of Chemical Theory and Computing* 2009, **5**:2353.

S21. Petridis, L, Smith, JC, *Journal of Computational Chemistry,* 2009, **30**:457.

S22. Jorgensen, WL, Chandrasekhar, J, Madura, JD, Impey, RW, Klein, ML, *Chinese Journal of Chemical Physics,* 1983, **79**:926.

S23. Darden, T, York, D, Pedersen, L, *Journal of Chemical Physics,* 1993, **98**:10089.

S24. Essmann, U, Perera, L, Berkowitz, ML, Darden, T, Lee, H, Pedersen, LG, *Journal of Chemical Physics,* 1995, **103**:8577.

S25. Ryckaert, JP, Ciccotti, G, Berendsen, HJC, *Journal of Computational Physics.* 1977, **23**:327.

S26. Martyna, GJ, Tobias, DJ, Klein, ML, *Journal of Chemical Physics,* 1994, **101**:4177.

S27. Feller, SE, Zhang, YH, Pastor, RW, Brooks, BR, *Journal of Chemical Physics,*1995, **103**:4613.

S28. Humphrey, W, Dalke, A, Schulten, K, *Journal of Molecular Graphics,* 1996, **14**:33.

S29. Matthews et al., Journal of Chemical Theory and Computing, 2012 **8 (2),** 735-748.

S30. Lin, ST, Blanco, M, Goddard, WA, *Journal of Chemical Physics,* 2003, **119**:11792.

Supplementary Figures

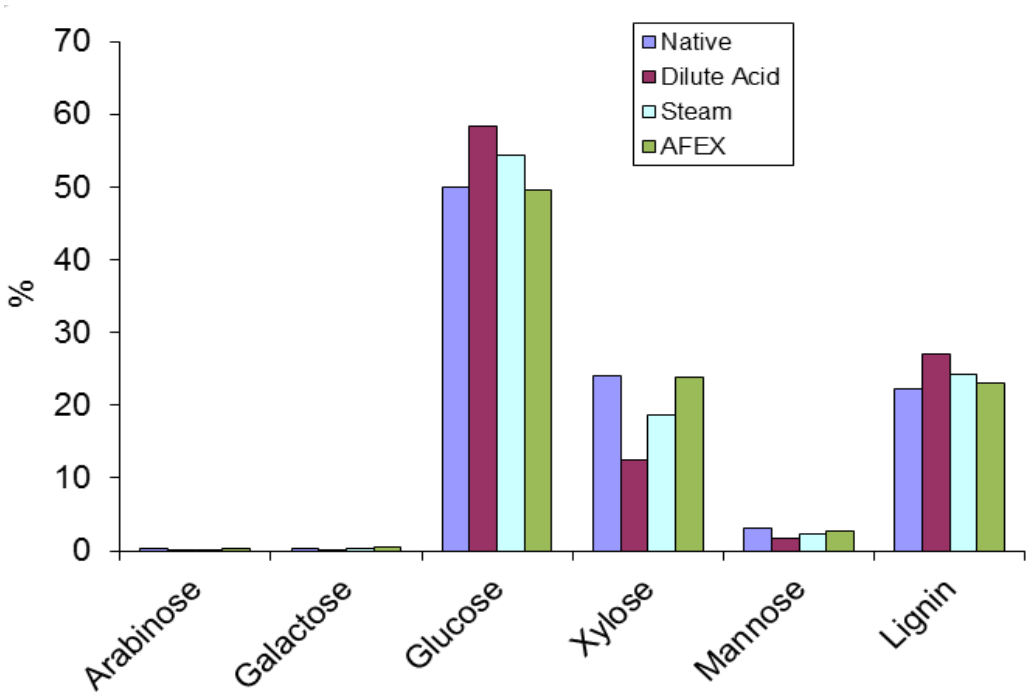

Fig.S1 Carbohydrate and Klason lignin distribution of aspen samples as determined by HPLC and normalized by the sum of all the measured components. Samples were native or pretreated by the listed method.

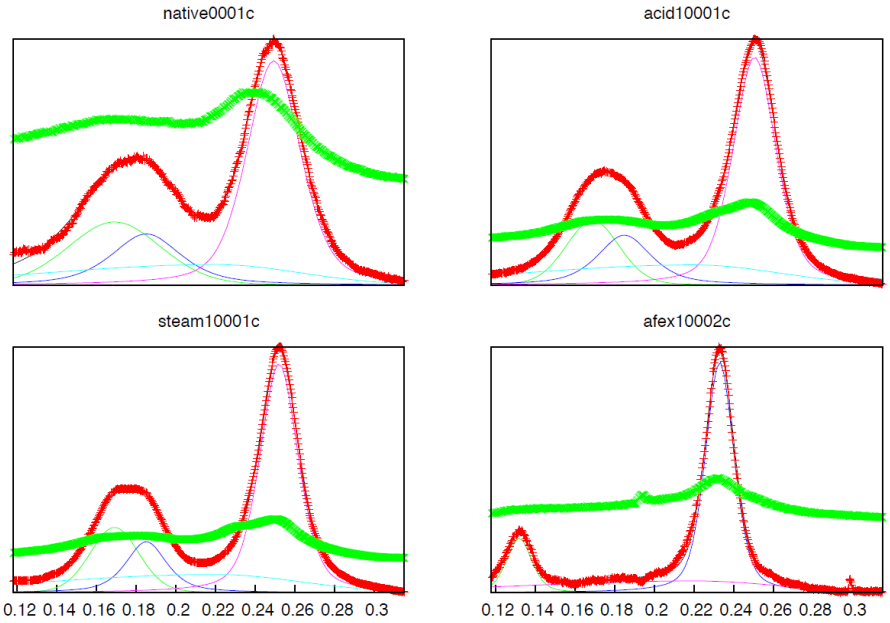

Fig.S2 The resulting fitted points for I(q) (red marks) and B(q) (green marks) for the X-ray fiber diffraction patterns collected from a) untreated Aspen b) dilute acid pretreated c) steam explosion pretreated d) AFEX-III pretreated. The three fitted profiles for the equatorial peaks (1-10), (110) and (200) for cellulose I are represented as green, dark blue and purple solid lines. The two fitted equatorial peaks (010) and (100) for cellulose III_I are represented in green and dark blue solid lines. In each case the fitted diffuse scattering background is represented as a light blue solid line and the sum of peaks and background is represented as a solid red line. The wave-vector q along the horizontal axes is the reciprocal of the distance in Angstroms.

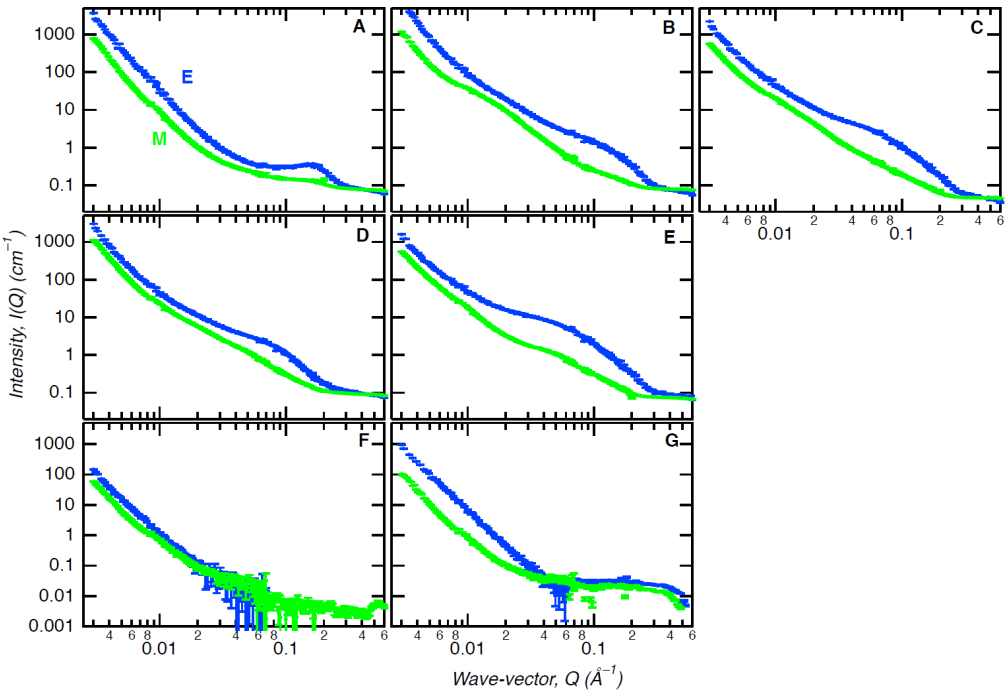

Fig.S3 Small Angle Neutron Scattering intensity scattering curves, I(q), in the meridional (green) and equatorial (blue) directions collected from D_2O soaked Aspen chips that were A) untreated B) dilute acid pretreated C) steam explosion pretreated D) AFEX-III pretreated. Scattering curves shown E) are from a sample that had been steam explosion pretreated and then bleached to remove lignin, in F) are from sample that had been dried, and in G) are from a sample that had been soaked in 40% D_2O and 60% H_2O.

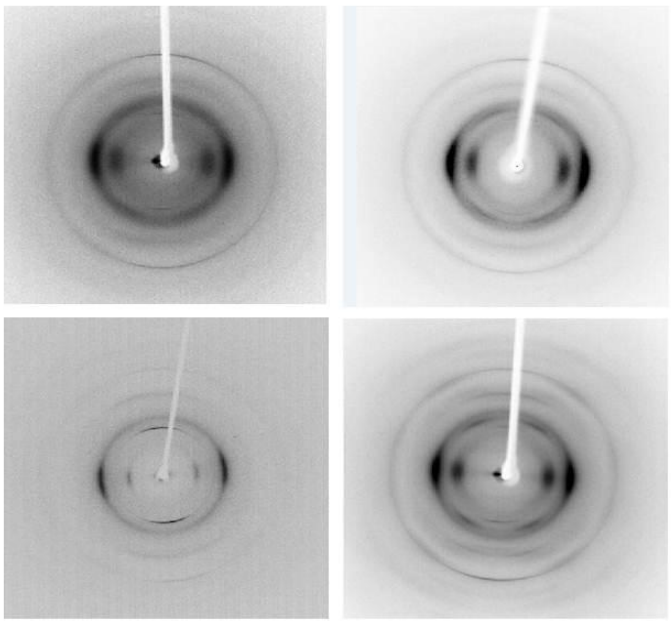

Fig.S4 X-ray fiber diffraction patterns collected from Aspen chips that were (top left) untreated, (top right) dilute acid pretreated, (bottom right) steam explosion pretreated, and (bottom left) AFEX-III pretreated. The fiber axis (meridional) direction is vertical and the equatorial direction is horizontal.

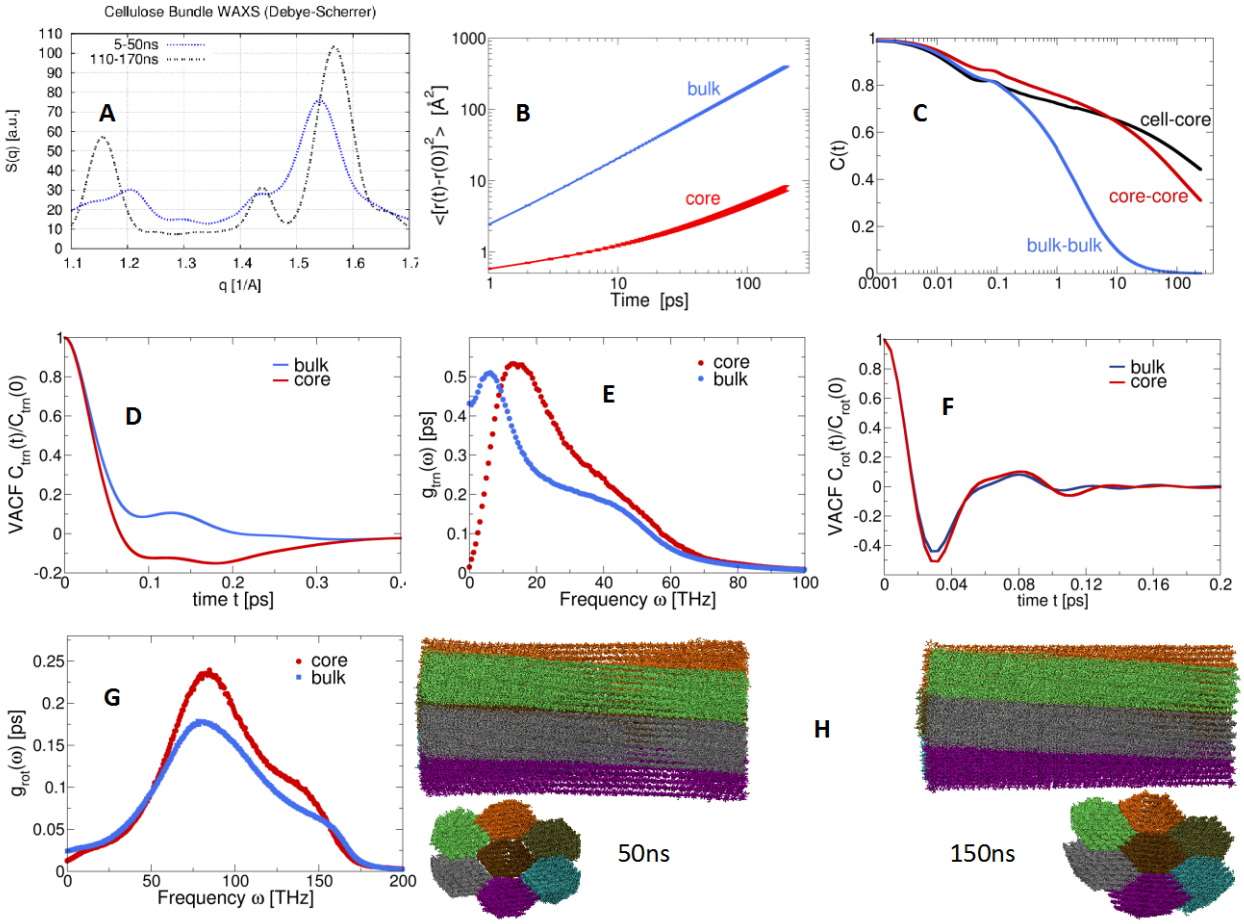

Fig.S5 (A) Wide-angle X-ray scattering computed from the simulation model shown in Fig. 2. Data are averaged between 5-50 ns and 110-170ns. (B) Mean square displacement of core and bulk water molecules. (C) Hydrogen bond autocorrelation function $C(t)$ for cellulose:core-water, core-water:core-water bulk-water:bulk-water pairs of molecules. (D) Translational velocity autocorrelation function for core and bulk water molecules. (E) Translational density of states (F) Rotational velocity autocorrelation function. (G) Rotational density of states. (H) Views of the

cellulose macrofibrillar bundle with water removed and individual cellulose elementary fibrils represented in different colors.

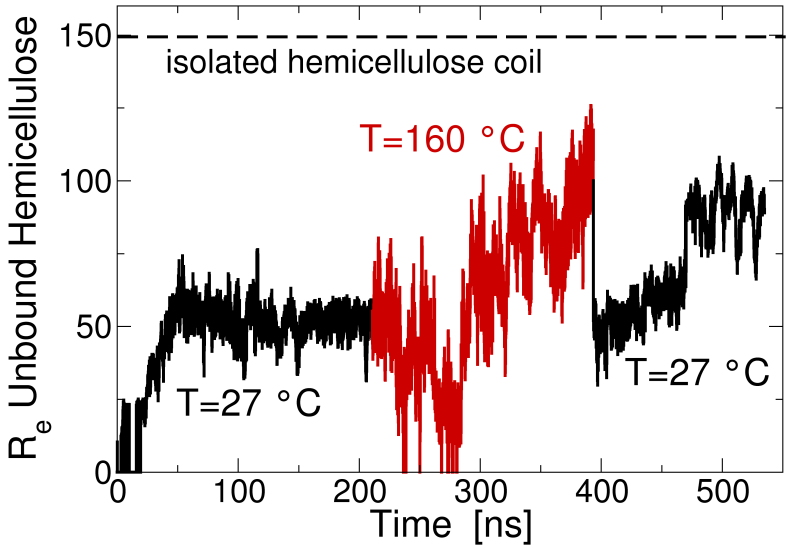

Fig.S6 End-to-end distance of the segment of hemicellulose not in contact with lignin as a function of simulation time as the temperature is cycled between 27 C and 160 C. The dotted line shows the theoretically maximum end-to-end distance for the isolated hemicellulos coil.

Table S1 A: We have added to the caption for the table "Cellulose elementary fibril parameters; LFD is the lateral fibril dimension, and ACL is the axial correlation length, as discussed in the text. The parameters were obtained using the Scherrer formula with a value of 1 for the shape constant."

	Native	Acid	Steam	AFEX III
Δφ (˚)	32	25	23	27
$LFD_{1\text{-}10}$ (nm)	1.8	3.2	3.4	NA
LFD_{110} (nm)	2.4	3.1	4.1	NA
LFD_{200} (nm)	3.1	3.9	4.4	NA
d_{200} (Å)	4.01	3.99	3.97	NA
LFD_{010} (nm)	NA	NA	NA	6.7
ACL (nm)	28.9	24.6	27,4	28.6
c (Å)	10.36	10.40	10.39	10.34

Table S2

Comparison of translational (D_T) and rotational diffusion (D_R) constants, translational (f_T) and rotational diffusion (f_R) fluidicity, translational (S_T), rotational (S_R) and total (S) entropy of bulk and core water molecules at 27°.

	Bulk	Core
D_T [$Å^2 ps^{-1}$]	0.580±0.020	0.017±0.002
D_R [ps^{-1}]	0.500±0.005	0.22±0.009
f_T	0.355±0.006	0.056±0.004
f_R	0.0800±0.0004	0.0510±0.0010
$T \cdot S_T$ [J/mol/K]	53.63±0.07	38.60±0.17
$T \cdot S_R$ [J/mol/K]	13.52±0.01	12.36±0.03
$T \cdot S$ [J/mol/K]	67.13±0.08	50.97±0.19