Noname manuscript No.
(will be inserted by the editor)

Libcel: A library of simulated cellulose microfibrils and scattering intensities

Yan Zhang · Michael Crowley · Jacob Hinkle · Lee Makowski

Received: date / Accepted: date

Abstract A library of cellulose microfibrils with different cross-sectional shapes and number of chains is generated. Dynamics of these fibrils is simulated by molecular dynamics using Charmm. Pair-distances and x-ray scattering intensities of untwisted fibrils as well as their twisted structures are simulated using the Debye formula in cylindrical coordinates. For untwisted fibrils of different shape and size, the intensity peak of the (2 0 0) reflection becomes narrower as the fibril size increases as expected. For twisted fibrils of same shape and size, intensities at (1 1 0)/(1 -1 0) reflections appear to flatten or form double peaks but the (2 0 0) reflections remains relatively insensitive to fiber twist. Plant tissue undergoing chemical pretreatment may consist an ensemble of different cellulose structures(shape, twist) which all contribute to a scattering intensity profile at x-ray data collection. A multi-component regression model is proposed to calculate the abundance of each microfibril structure by searching for the linear combination of simulated intensities in the library that best fits a given experimental intensity. Principal component analysis and spectral clustering methods are use to study the library redundancy and to reduce the biased of twisting profiles in the fitting. Hierarchical support vector machine and convolutional neural network are developed for microfibril shape classifi-

Yan Zhang
Material Science Division, Argonne National Laboratory, Lemont, IL 60561, USA
E-mail: yz@anl.gov

Michael F. Crowley
Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401, USA
E-mail: Michael.Crowley@nrel.gov

Jacob Hinkle
Computational Science Center, National Renewable Energy Laboratory, Golden, CO 80401, USA
E-mail: Jacob.Hinkle@nrel.gov

Lee Makowski
Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
E-mail: l.makowski@northeastern.edu

cation. The relation between scattering intensities and microfibril shapes and twist is systematically addressed by this study.

Keywords Cellulose Microfibril · Structure Library · X-ray Scattering · Intensity Simulation

1 Introduction

Cellulose plays a very important role in various industries such as paper, wood, renewable energy and others. The recalcitrance of cellulose microfibrils has been studied for years. This characteristic makes cellulose a useful building material but not an efficient renewable energy resources. One possible source of the recalcitrance of cellulose microfibrls is the twisting of fibrils which may render them resistant to pretreatment or enzymatic digestion. Many imaging techniques such as TEM and AFM have been used to provide experimental evidence to visulize the twist of cellulose fibrils [1][2][3][4][5]. These imaging results showed cellulose I microfibrils have right-handed twist in several plant tissues.

Theoretically, simulations of cellulose microfibril twist have been generated using molecular mechanics by CHARMM [6][7], GLYCAM [8] and GROMOS [9][10] carbohydrate force fields. Several studies showed that hydrogen bonding of cellulose fibrils has a critical energetic contribution to the twist [11][12]. These studies showed that the twisting was due to van der Waals forces rather than electrostatics [13].

Cellulose microfibrils in different plant tissues may exhibit different shapes and contain different numbers of chains, such as 36-chain hexagonal or diamond shape or 24-chain shape [18]. These may also impact the degree of twisting in equilibrium - less twisting as the the number of chains increases. In order to understand cellulose microfibrils in a systematic way, a library of cellulose structures with twist equilibrated was simulated (4-diamond, 9- diamond, 16-diamond, 25-diamond, 36-diamond, 49-diamond, 64-diamond, 81-diamond, 24-brick, 24-hexagonal and 36-hexagonal). The scattering intensities were calculated for each structure. In this study, we used multi-component regression analysis to find the coefficients that best fit synthetic mixture of intensities generated from the libray. Each shape later has 100 structures by selecting intensities every other 20 from 2000 structures to construct the library for fitting analysis.

2 Library of simulated cellulose microfibrils and scattering intensities

2.1 Twisted cellulose fibrils with different cross-sectional shapes and number of chains

Charmm (Chemistry at HARvard Molecular Mechanics) program [19] was used to simulate many hypothetical cellulose fibrous structures. The Charmm molecular dynamic software was developed by Martin Karplus group at Harvard University in late 1960s and the packages have been develped with wild range of features, force fields and analytical methods. The cellulose structure simulation was performed at the National Renewable Energy Laboratory (NREL) using DOMDEC fast parallel method [20]. Charmm was first used to generate an 81-chain (9x9) diamond shape cellulose microfibril which is 200 $\AA$ long and a cellulose Iβ crystalline structure [21]. The structure template was cut and reshaped into different cellulose microfibril shapes and lengths (each chain has 20 cellobiose units which is 200 $\AA$). These structures were subjected to a process of structural dynamics in water buffer until fibrils equilibrated (in $\sim$ 800 ns). The library of simulated cellulose microfibril structures includes diamond shape of 2x2-, 3x3-, 4x4-, 5x5-, 6x6-, 7x7-, 8x8- and 9x9-chains, and hexagonal shapes of 24- and 36-chains as well as a 24-chains brick shape. Cellulose microfibril structures were saved in dcd files after equilibration. Each dcd file contains 1 untwisted and 2000 twisted structures, the cross-section view of selected members from this library are shown in figure 1.

2.2 Pair-distribution and scattering intensity simulation

The equatorial scattering intensities, $I(R)$ of cellulose microfibrils were calculated using cylindrical Debye formula [14]:

$$I(R) = \sum_i \sum_j f_i(R) f_j(R) J_0(2\pi r_{ij} R) \tag{1}$$

where r_{ij} is the radial interatomic distance between atom i and j component of the interatomic vector perpendicular to the fibril axis. R is the equatorial distance in reciprocal space. f_i and f_j are atomic scattering factors for atom i and j, and J_0 is zero-order Bessel function of first kind. A fast Debye method was developed to accelerate the simulation process [17]. This method first calculates all interatomic distances and labels them by atom type into six groups, aka. C-C, H-H, O-O, C-H, H-O, C-O. Then, a histogram is used to compress the large number of interatomic distances into a small number of discrete distances. The size of the resulting interatomic distance arrays ($\hat{r_{ab}}$) contain population-distributions (w_{ab}) which are analogous to pair-distribution functions. Since the pair-distances are atom type indexed, the atomic scattering factors can be moved out of the summation in the Debye formula, which becomes:

$$I_{ab}(R) = f_a(R)f_b(R)\sum_i w_{ab,i}J_0(2\pi r_{\hat{ab},i}R) \tag{2}$$

The final scattering intensity is the sum of six sub-intensity profiles simulated by each atom type indexed pair-distance arrays.

$$I = \sum_{ab\in S} I_{ab} \tag{3}$$

where S=C-C, H-H, O-O, C-H, H-O, C-O. The fast Debye method not only accelerated the simulation process but also made it possible to visulize the pair-distribution functions for the large and complex cellulose structure dataset. The pair-distribution functions and scattering intensities are shown in figures 2 to 4.

3 Comparison of intensities in the library

3.1 Twisted microfibrils

The scattering intensities at (1 1 0)/(1 -1 0) reflection changed as these fibrils twisted, shown in figures 2 - 4. As the cellulose fibrils twisted, the periodicity of their crystal structure was lost which resulted in a progressively smoother pair-distribution function. The slope of the small-angle scattering region (1/d < 0.05 $\AA^{-1}$) became steeper for larger cellulose microfibrils but remained almost the same for each individual shape during twist. This confirmed that the size of structures remained very similar during the twist for each shape and has little or no influence on small-angle data. In the wide-angle scattering range, the larger the microfibril, the clearer the seperation of (1 1 0) and (1 -1 0) while twisting. The intensity curve at (1 1 0)/(1 -1 0) reflection of 36-chain diamond shape started to flatten after twist and the (1 1 0)/(1 -1 0) reflection of 49-, 64- and 81-chain formed two seperate peaks. The height of (2 0 0) reflection changed among crystal and twisted structures and in many cases the peak position was seen to shift slightly.

3.2 Untwisted microfibrils

By comparing scattering intensities from all untwisted structures, the relation between crystalline size and cross-sectional shape and (2 0 0) reflection width was understood. In figure 5 a, the scattering intensities of 36-chain diamond vs. hexagonal shape structures are shown and the differences between 24- vs. 36-chain hexagonal were small as seen in figure 5 c. Figure 5 b shows that 24-chain brick had much broader (2 0 0) corresponding to the shorter edge of the rectangular cross-section brick shape than in the 24-chain hexagonal shape. The 25-chain diamond shape gave an observable smaller peak width than 24-chain hexagonal. Figure 5 d is a systematic and complete intensity data comparison of structures with same shape (diamond) but increasing number of chains - as the crystal size gets larger, the (2 0 0) reflection becomes narrower.

4 Redundancy analysis and dimension reduction of the library

4.1 Visualization of library redundancy

The intensity library is a MxN matrix where M is the scattering q-range and N is the number of structures. In order to visualize the library redundancy in a 2-D plot, we computed a 2xN visulazing matrix by calculating the log-sum of intensities in small-angle and wide-angle respectively for each structures. In figure 6, x-axis is log-sum D_{SAXS} and y-axis is log-sum D_{WAXS}.

$$D_{SAXS} = log(\sum_{0}^{qmax/2-1} D) \tag{4}$$

$$D_{WAXS} = log(\sum_{qmax/2}^{qmax} D) \tag{5}$$

The scatter plot of the full library elements is shown in figure 6. If all intensites are projected in to x-axis, each shape can be roughly seperated but projection onto y-axis couldn't help on identification of either shape or twist of the library.

4.2 Principal Component Analysis

PCA [22] was used here to find significant component in the library and to reduce the dimension of the library. First, the library intensities were normalized by subtracting the mean of all intensities.

$$\mu = \frac{1}{n}(I_0 + I_1 + ... + I_{n-1}) \tag{6}$$

$$\bar{I} = I - \mu \tag{7}$$

$$(\bar{I}^T\bar{I})\overrightarrow{u} = \lambda\overrightarrow{u} \tag{8}$$

Singular Value Decomposition was used on $\bar{I}^T\bar{I}$ to calculate the principal components of the libraray. The singular values are shown in figure 7. There are 22 singular values that are small than ϵ = 1e-3, which equals to the number of intensities of 1 untwisted plus 1 twisted structures for the 11 shape library.

4.3 Spectral Clusteirng

In order to find a reasonable number of clusters needed to reduce the dimensionality while maintaining the features of the intensity library at same time, we adapted spectral clustering method which has been shown to work well on both spherical and elliptical data [23].

An affinity matrix A was formed:

$$A_{ij} = e^{-||I_i - I_j||^2}, \;\; if\, i \neq j \tag{9}$$

and $A_{ij} = 0$ for $i = j$. Define S as a diagonal matrix whose diagonal values were computed:

$$S_{ii} = \sum_{j=0}^{n-1} A_{ij} \tag{10}$$

Matrix L is constructed:

$$L = S^{\frac{1}{2}} A S^{\frac{1}{2}} \tag{11}$$

The eigenvalues and eigenvectors of L were calculated and matrix X was formed by selecting the first p-th largest eigenvectors, in another word, X was a q x p dimension reduced matrix. Here, k was chosen by identification of the knee point of eigenvalues of L. Normalized each row of matrix X gave a matrix Y whose row was in unit length:

$$Y_{ij} = \frac{X_{ij}}{(\sum_{j=0}^{k-1} X_{ij}^2)^{1/2}} \tag{12}$$

Each row in matrix Y was treated as a point in p dimensional space, $Y \in \mathbb{R}^p$ and p-means clustering method is used to partition the n points into p groups. Here the p centroids are $\bar{Y}_1, ...\bar{Y}_p$. The two-steps are following:

Step 1: assign p dimensional points into clusters by comparing the distance to centroids:

$$C_i^{(t)} = \{Y_j : \|Y_j - \bar{Y_i}^{(t)}\| \leq \|Y_j - \bar{Y}_m^{(t)}\|\} \tag{13}$$

and $m \in \{1, ..., p\}$.

Step 2: update new centroids for t+1 step:

$$\bar{Y_i}^{(t+1)} = \frac{\sum_{Y_j \in C_i^{(t)}} Y_j}{\sum_{Y_j \in C_i^{(t)}} 1} \tag{14}$$

In the last, the original intensities in the library were assigned into the p clusters. Affinity matrix and spectral clustering result are shown in figure 8 and figure 9 respectively. Most intensities in same shape were clustered into the same group except the 4-chain diamond which changed significantly due to the twist.

5 Mixture of intensities and multi-component analysis using reduced library

Two synthetic mixture of intensity profiles were generated using the library. Candidate structures were drawn ramdomly from the library with weighting coefficients in random and Gaussian distribution respectively. The synthetic intensity profiles (measurements) are:

$$I_{s1} = \sum_{i=0}^{N} \alpha_i I_i + \epsilon, \quad \alpha \sim random() \tag{15}$$

and

$$I_{s2} = \sum_{i=0}^{N} \beta_i I_i + \epsilon, \quad \beta \sim \mathcal{N}(\frac{N}{2}, \sigma^2) \tag{16}$$

where α and β are in random and Gaussian distribution, $\epsilon \sim \mathcal{N}(0, 0.5)$ is additive Gaussian White noise. Since the full intensity library is heavily biased on the twisted structures, we used a reduced library with 1 untwisted and the 1 twisted intensity profiles for each shape. In this case, N = 22 and we set σ = 3 for Gaussian distribution measurement.

Having the two profiles and coefficients together in a vector notation: $\mathbf{I_s} = [I_{s1}, I_{s2}]$, $\alpha = [\alpha_0, ...\alpha_n]^T$, $\beta = [\beta_0, ...\beta_n]^T$ and $\mathbf{W} = [\alpha, \beta]$, the linear system becomes:

$$\mathbf{I_s} = \mathbf{IW} + \epsilon \tag{17}$$

Where $\mathbf{I_s} \in \mathrm{I\!R}^{\mathrm{mx2}}$, $\mathbf{I} \in \mathrm{I\!R}^{\mathrm{mxn}}$ and $\mathbf{W} \in \mathrm{I\!R}^{\mathrm{nx2}}$, and each individual coefficients ω_i in $\mathbf{W}$ are subjected to:

$$\sum_i \omega_{i,j} = 1, \quad 0 \leq \omega_{i,j} \leq 1 \tag{18}$$

The estimated coefficient vector $\hat{\mathbf{W}}$ can be obtained by minimizing

$$\hat{\mathbf{W}} = argmin_{\mathbf{W}} \left\{ \frac{1}{2} ||\mathbf{I_s} - \mathbf{IW}||^2) \right\} \tag{19}$$

or calculated by least-square method [24]:

$$\hat{\mathbf{W}} = (\mathbf{I^T I})^{-1} \mathbf{I^T I_s} \tag{20}$$

Here, we used the R-factor and sum of absolute error(SAE) as metric for intensity fitting and coefficient estimation:

$$R - factor = \frac{\sum |\sqrt{I_s} - \sqrt{\hat{I}_s}|}{\sum |\sqrt{I_s}|} \tag{21}$$

$$SAE_j = \sum_i |\omega_{i,j} - \hat{\omega}_{i,j}| \tag{22}$$

and the results in figure 10 and table 1 showed good accuracy between estimated and ground-truth synthetic data in both intensity fitting and coefficient estimation.

6 Microfibril shape classification using full library

Another important application is microfibril shape classification. Experimental collected scattering intensities of single-fibrils without information about the shape could be classified by pre-trained classifiers built based on this library.

6.1 Hierarchical Support Vector Machine

We first use support vector machine (SVM) as an example [25]. Since SVM is binary classifier, we need to built $N-1$ mini-classifiers for N classes (N=11 shapes) in a hierarchical architecture, shown in figure 11. We divide the 1100 intensity profile into 880 profiles for training and 220 for testing. The training in the first classifier can recognize the shapes smaller or larger than 30 chains. SVM performs binary classification in each layer until the bottom child was reached in this hierarchical decision tree model. This method has advantages over traditional two type of SVM classifiers: one vs. all and one vs. one method. One vs. all method has huge bias when the number of classes is increasing. The one vs. one method, on the other hand, increases the number of classifiers to be trained quandratically ($\frac{N(N-1)}{2}$).

6.2 Convolutional Nerural Network

Convolutional neural network (CNN) has been popular in recent years mainly because of the deep architecture which significantly improves the performance of traditional neural network [26]. CNN based classification now becomes the state-of-the-art method for natural images such as ImageNet challenge, MNIST and CIFAR-10 datasets. The CNN classification architecture used here is shown in figure 12, which has 2 convoluiton layers, 2 max-pooling layers and a fully connected layer. The first convolution layer extracts 5 features from each input intensity and second convolution layer further increases the total number of features to 10. Each max-pooling layer selects maximum value in a 1 x 5 kernel sliding across previous convolution layers. Both the convolution and max-pooling layers have stride of 3, which reduce the feature dimension by $\frac{1}{3}$ layer by layer. All features from the second max-pooling layers are stacked into a single vector as input for fully connected layer which has 100 neurons. The output has 11 shapes and cross-entropy is used for loss function. Stochastic gradient descent and back-propagation methods is used for training to update the system weights. After training, the system has a classification accuracy of 99 % and test result is in figure 13 showing the true and estimated shape labels with only one misclassification.

7 Discussion

In this paper, we built a library of cellulose fibrils with different shape, size and twist, and generated their scattering intensities. Both data sets will be open for download for cellulose research. These cellulose fibrils were generated using Charmm at NREL with three shapes: diamond, brick and hexagonal and number of chains varying from 4 to 81. All the structures were equilibrated in water solvent (water molecules removed afterwards) and each frame were extracted. The comparison of scattering intensities from all structures in our library showed that larger fibrils contribute to narrower and sharper (2 0 0) intensity peak. The library redundancy was further discussed using principal component analysis and spectral clustering method. A multi-component model was used to estimate the weighting coefficient of each structure in synthetic mixture of intensities (ensemble). More cellulose fibrils with different shapes will be added into this library in the future so that to address more scenarios - different plant tissue or chemical/physical deconstruction.

8 Library download

The library of cellulose microfibrils and intensity profiles in 11 shapes can be download at: https://github.com/yzhangresearch/libcel. The dataset will be later migrated to a website with larger space as library dimension grows.

9 Acknowledgement

This work was supported as part of the Center for Direct Catalytic Conversion of Biomass to Biofuels (C3Bio), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Science under award #DE-SC0000997.

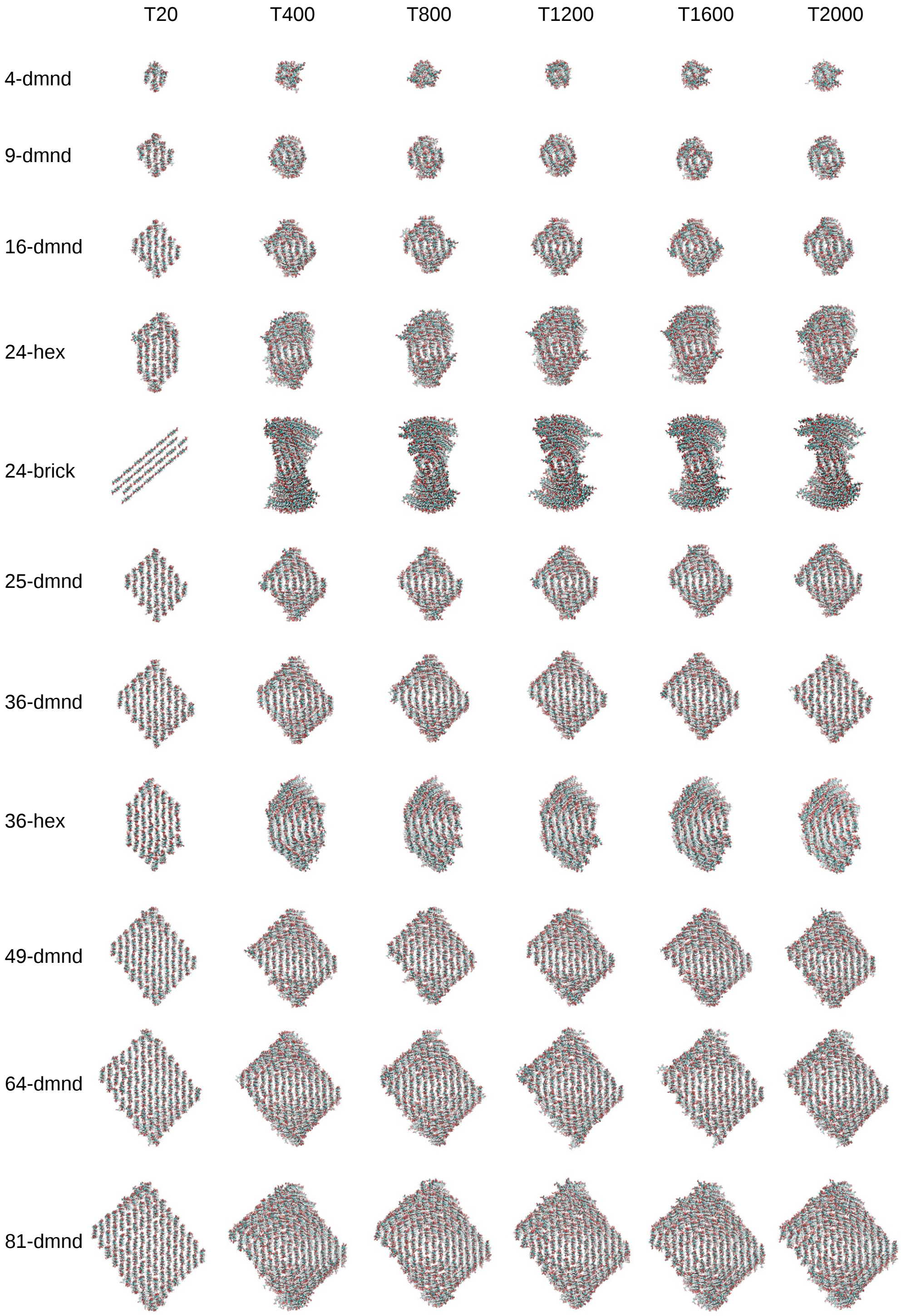
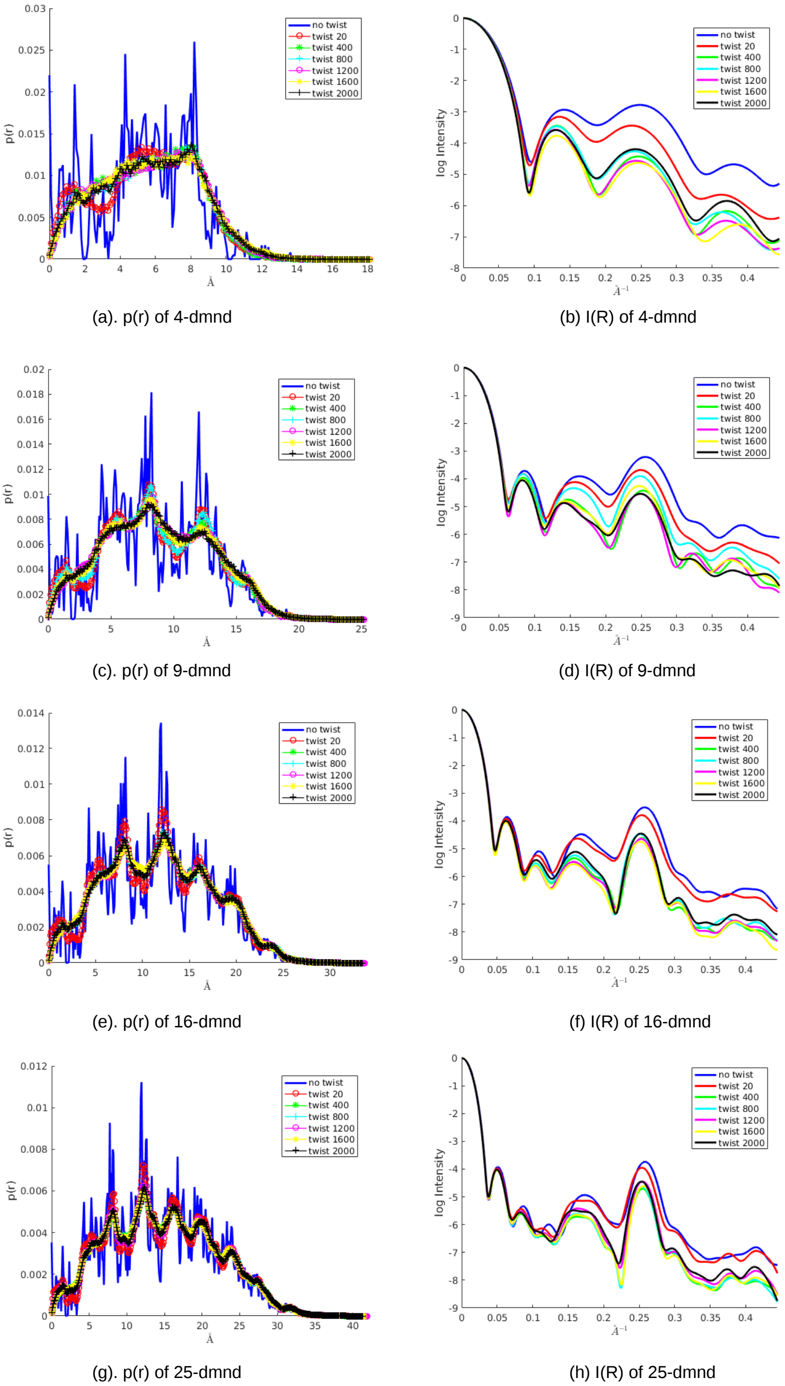


Fig. 1 Cross-section view of library of cellulose fibirls (11 shapes and 6 sampled twisted structure for each shape).



(a). p(r) of 4-dmnd

(b) I(R) of 4-dmnd

(c). p(r) of 9-dmnd

(d) I(R) of 9-dmnd

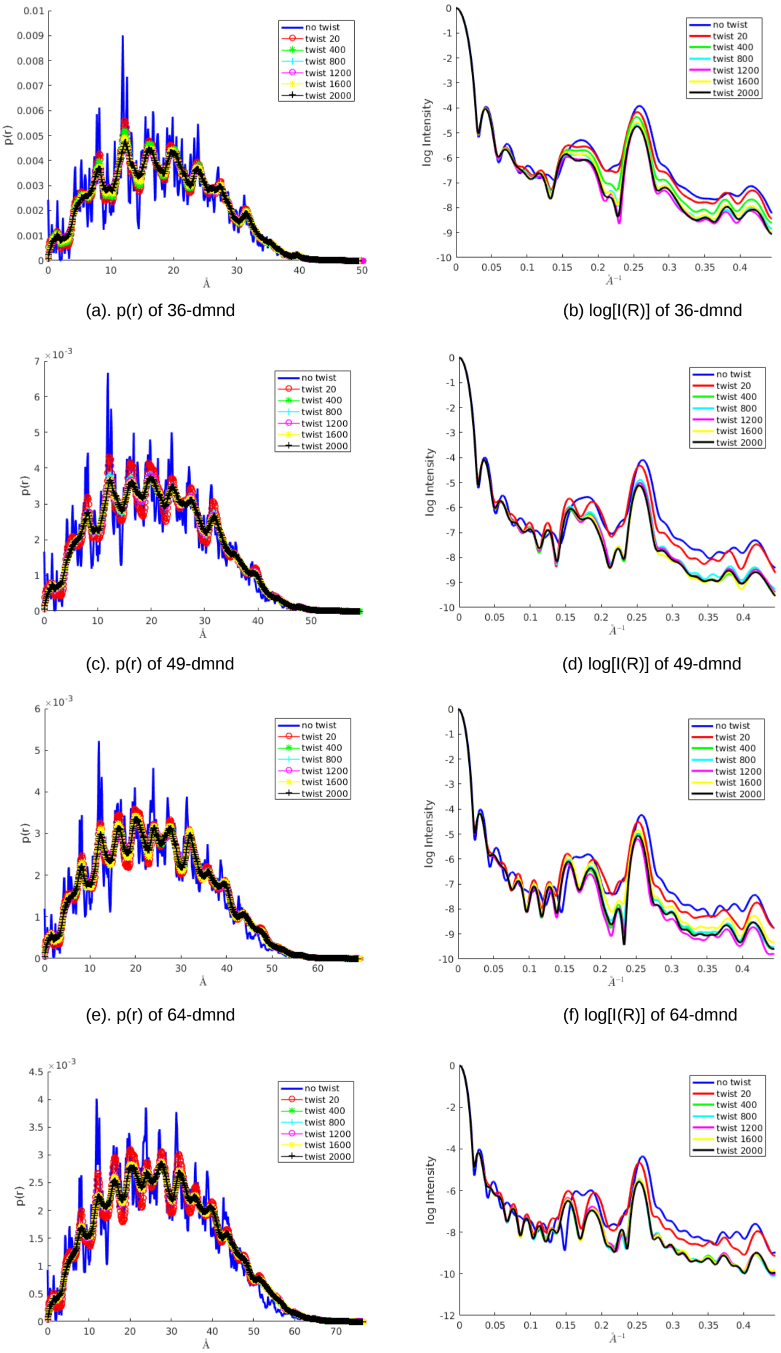
(e). p(r) of 16-dmnd

(f) I(R) of 16-dmnd

(g). p(r) of 25-dmnd

(h) I(R) of 25-dmnd

Fig. 2 Pair-distribution and scattering intensity of cross-section in r and R of 4-, 9-, 16- and 25-chain diamond shape.



(a). p(r) of 36-dmnd

(b) log[I(R)] of 36-dmnd

(c). p(r) of 49-dmnd

(d) log[I(R)] of 49-dmnd

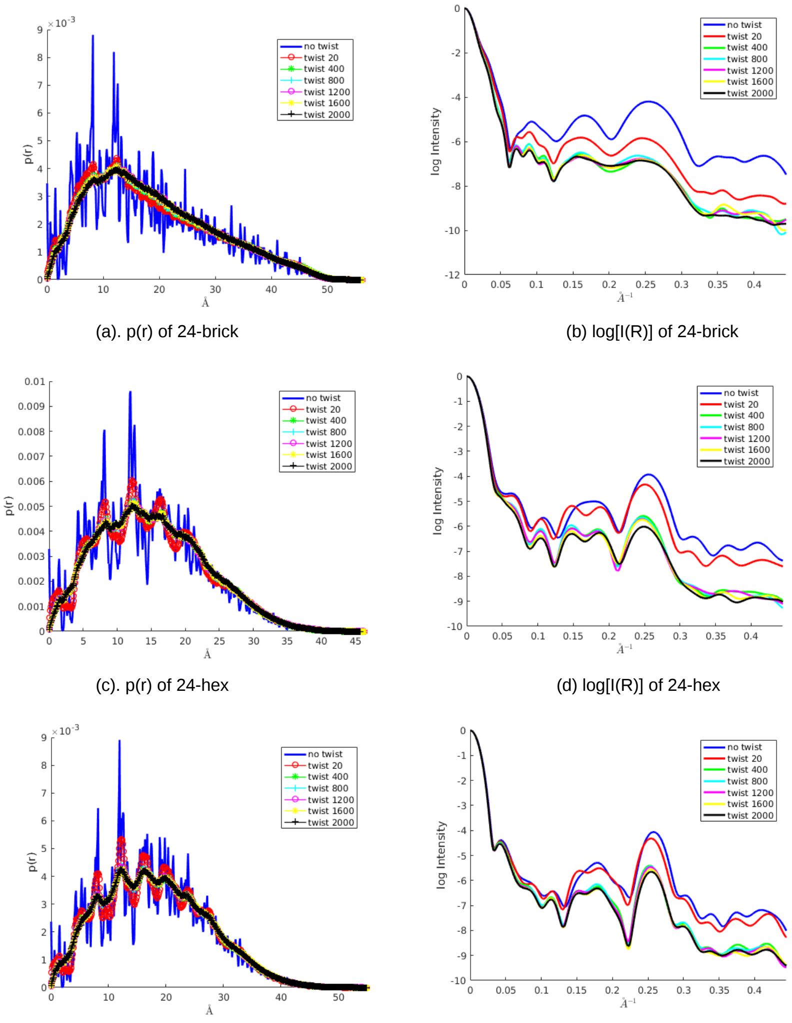
(e). p(r) of 64-dmnd

(f) log[I(R)] of 64-dmnd

(g). p(r) of 81-dmnd

(h) log[I(R)] of 81-dmnd

Fig. 3 Pair-distribution and scattering intensity of cross-section in r and R of 36-, 49-, 64- and 81-chain diamond shape.



(a). p(r) of 24-brick

(b) log[I(R)] of 24-brick

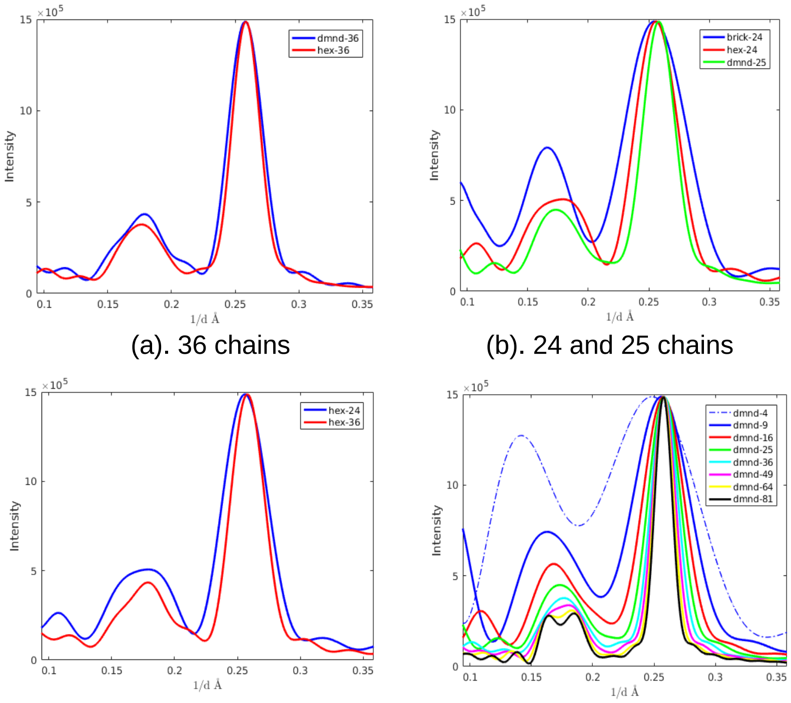
(c). p(r) of 24-hex

(d) log[I(R)] of 24-hex

(e). p(r) of 36-hex

(f) log[I(R)] of 36-hex

Fig. 4 Pair-distribution and scattering intensity of cross-section in r and R of 24-chain brick, 24- and 36-chain hexagonal shape.



(a). 36 chains

(b). 24 and 25 chains

(c). hexagonal shape

(d). diamond shape

Fig. 5 Simulated scattering intensities all all untwisted cellulose fibrils

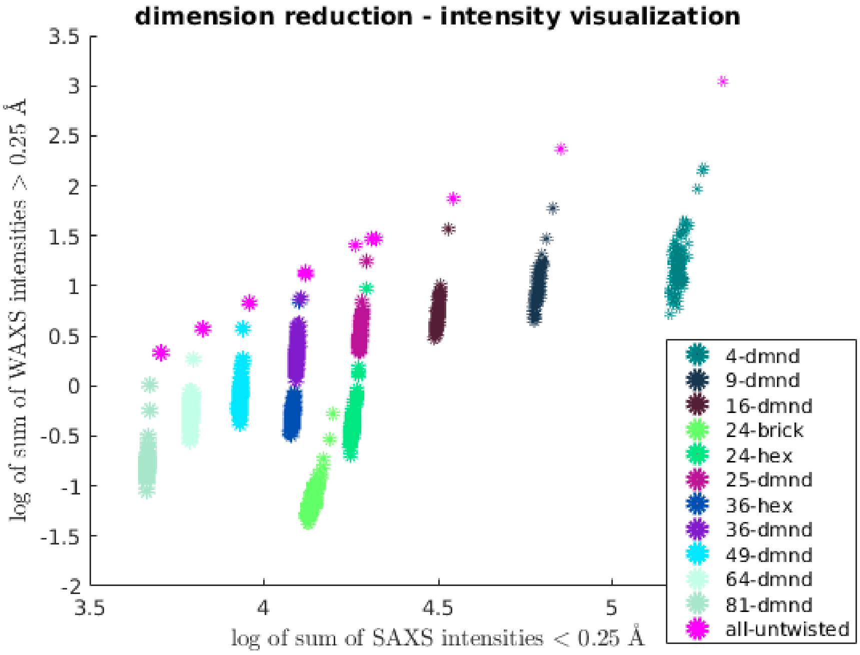
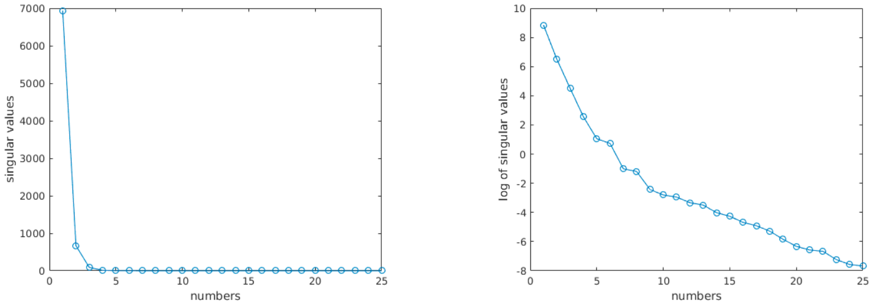


Fig. 6 Visualization of intensities library. Horizontal axis is the log-sum of small-angle scattering intensities and vertical axis is the log-sum of wide-angle scattering intensities.



(a) Singular values

(b) Log of singular values

Fig. 7 Principal component analysis and sigular values.

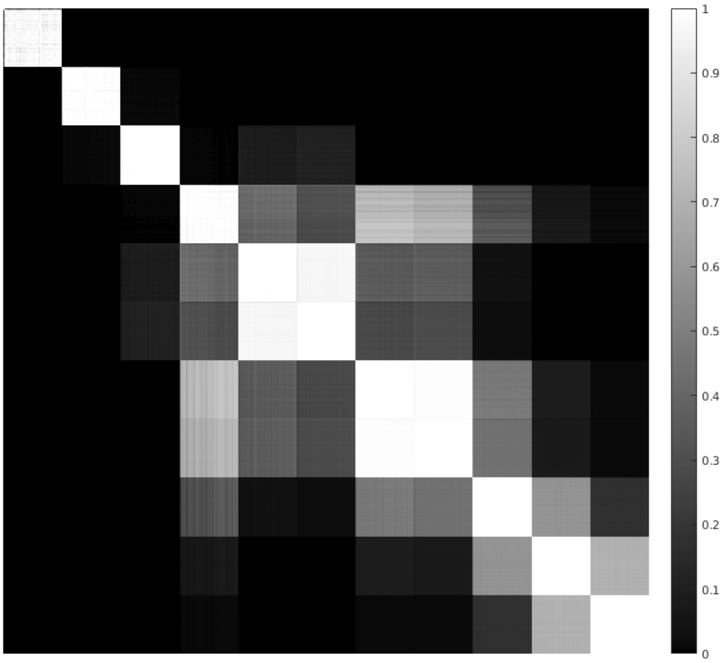


Fig. 8 Affinity matrix of the intensity library.

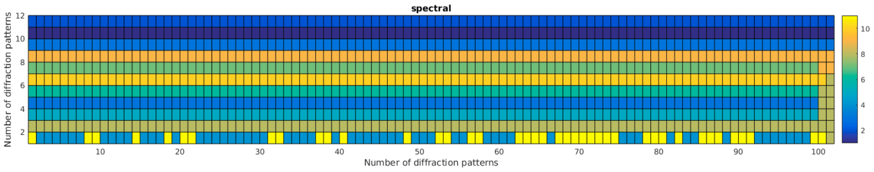
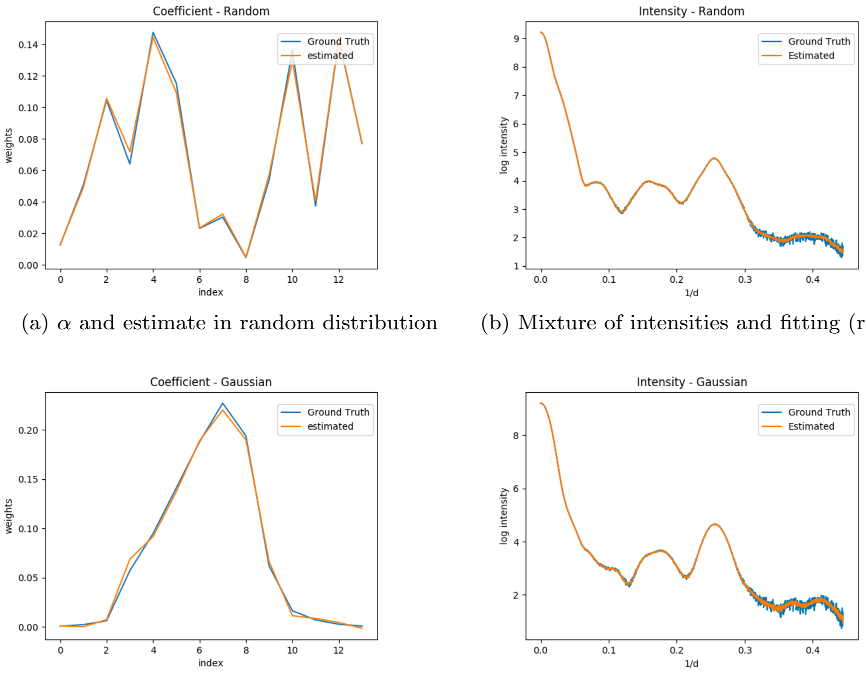


Fig. 9 Spectral clustering result of the intensity library.



(a) α and estimate in random distribution (b) Mixture of intensities and fitting (random)

(c) β and estimate in Gaussian distribution (d) Mixture of intensities and fitting (Gaussian)

Fig. 10 Multi-component analysis of synthetic mixture of intensities in random and Gaussian distribution.

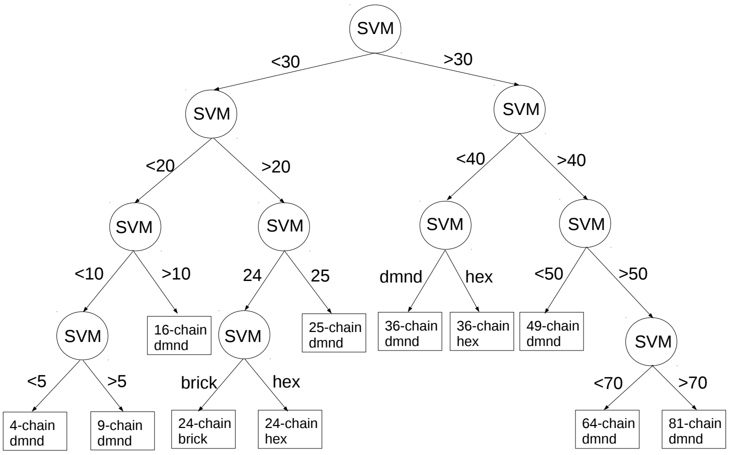


Fig. 11 Hierarchical support vector machine (H-SVM) classification.

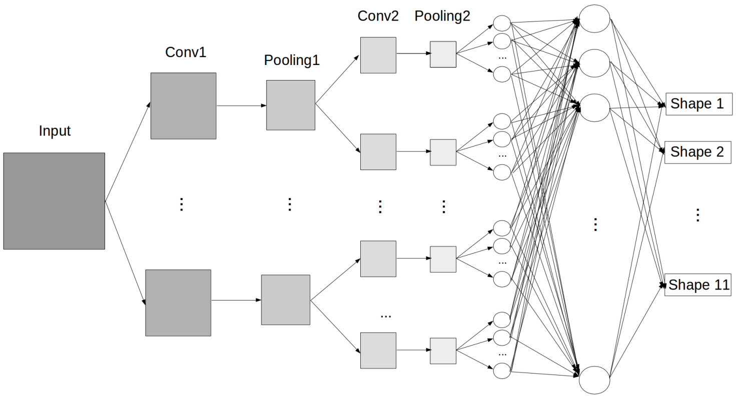


Fig. 12 Convolutional Neural Network architecture for cellulose microfibril shape classification.

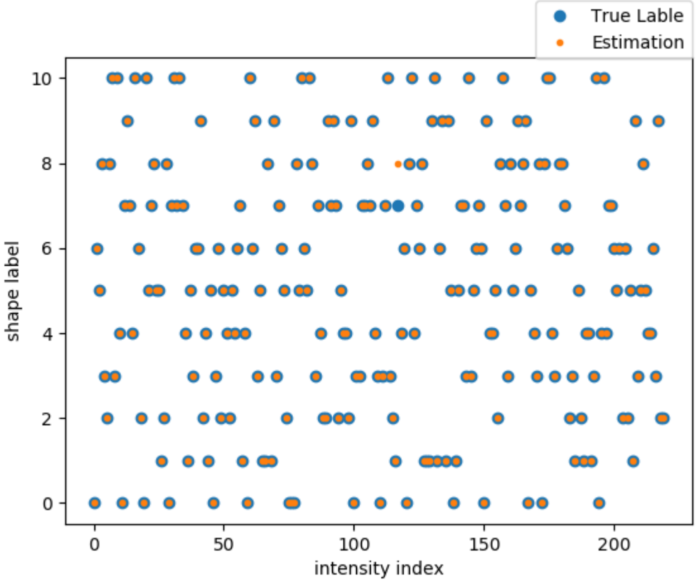


Fig. 13 True and estimated labels using Convolutional Neural Network.

	Random	Gaussian
R-fac, intensity	0.005	0.008
SAE, coefficient	0.036	0.049

Table 1: R-factor and sum of absolute error in mixture of intensities.

References

1. Bowling AJ, Amano Y, Lindstrom R, and Brown Jr. RM (2001) Rotation of cellulose ribbons during degradation with fungal cellulase. Cellulose 8(1): 91-97.
2. Brown Jr RM, Haigler CH, Suttie JA, White AR, Roberts ER, Smith CR, Itoh TA, Cooper K (1983) The biosynthesis and degradation of cellulose. J. Appl. Polym. Sci. 37:33-78.
3. Elazzouzi-Hafraoui S, Nishiyama Y, Putaux JL, Heux L, Dubreuil F, Rochas C (2007) The shape and size distribution of crystalline nanoparticles prepared by acid hydrolysis of native cellulose. Biomacromolecules. 9(1):57-65.
4. Hanley SJ, Revol JF, Godbout L, Gray DG (1997) Atomic force microscopy and transmission electron microscopy of cellulose from Micrasterias denticulata; evidence for a chiral helical microfibril twist. Cellulose 4(3):209-20.
5. Khandelwal M, Windle A. Origin of chiral interactions in cellulose supra-molecular microfibrils (2014) Carbohydr. Polym. 106:128-31.
6. Guvench O, Greene SN, Kamath G, Brady JW, Venable RM, Pastor RW, Mackerell AD. Additive empirical force field for hexopyranose monosaccharides (2008) J. Comput. Chem. 29(15):2543-64.
7. Hatcher ER, Guvench O, MacKerell Jr AD (2009) CHARMM additive all-atom force field for acyclic polyalcohols, acyclic carbohydrates, and inositol. Journal of chemical theory and computation. 5(5):1315-27.
8. Kirschner KN, Yongye AB, Tschampel SM, GonzlezOuteirio J, Daniels CR, Foley BL, Woods RJ (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J. Comput. Chem. 29(4):622-55.
9. Hansen HS, Hnenberger PH (2011) A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers, and glycosidic linkage conformers. J. Comput. Chem. 32(6):998-1032.
10. Lins RD, Hnenberger PH (2005) A new GROMOS force field for hexopyranosebased carbohydrates. J. Comput. Chem. 26(13):1400-12.
11. French AD, Concha M, Dowd MK, Stevens ED (2014) Electron (charge) density studies of cellulose models. Cellulose 21(2):1051-63.
12. Altaner CM, Thomas LH, Fernandes AN, Jarvis MC (2014) How cellulose stretches: synergism between covalent and hydrogen bonding. Biomacromolecules 15(3):791-8.
13. Hadden JA, French AD, Woods RJ (2013) Unraveling cellulose microfibrils: a twisted tale. Biopolymers 99(10):746-56.
14. Inouye H, Fraser PE, Kirschner DA (1993) Structure of beta-crystallite assemblies formed by Alzheimer beta-amyloid protein analogues: analysis by x-ray diffraction. Biophys. J. 64(2):502.
15. Inouye H, Zhang Y, Yang L, Venugopalan N, Fischetti RF, Gleber SC, Vogt S, Fowle W, Makowski B, Tucker M, Ciesielski P (2014) Multiscale deconstruction of molecular architecture in corn stover. Sci. Rep. 4, 3756.
16. Zhang Y, Inouye H, Yang L, Himmel ME, Tucker M, Makowski L (2015) Breakdown of hierarchical architecture in cellulose during dilute acid pretreatments. Cellulose 22(3):1495-504.
17. Zhang Y, Inouye H, Crowley M, Yu L, Kaeli D, Makowski L (2016) Diffraction pattern simulation of cellulose fibrils using distributed and quantized pair distances. J. Appl. Crystallogr. 49(6): 2244-2248.

18. Fernandes AN, Thomas LH, Altaner CM, Callow P, Forsyth VT, Apperley DC, Kennedy CJ, Jarvis MC (2011) Nanostructure of cellulose microfibrils in spruce wood. Proc. Natl. Acad. Sci. 108(47):E1195-203.
19. Brooks BR, Brooks CL, MacKerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A (2009) CHARMM: the biomolecular simulation program. J. Comput. Chem. 30(10):1545-614.
20. Hynninen AP, Crowley MF (2014) New faster CHARMM molecular dynamics engine. J. Comput. Chem. 35(5):406-13.
21. Langan P, Sukumar N, Nishiyama Y, Chanzy H (2005) Synchrotron X-ray structures of cellulose I and regenerated cellulose II at ambient temperature and 100 K. Cellulose 12(6):551-62.
22. Jolliffe I (2002) Principal component analysis. John Wiley & Sons, Ltd.
23. Ng A, Jordan M, and Weiss J (2001) On spectral clustering: Analysis and an algorithm. NIPS 14(2): 849-856.
24. Kay S (2013) Fundamentals of statistical signal processing: Practical algorithm development. Pearson Education.
25. Vapnik V (2013) The nature of statistical learning theory. Springer science & business media.
26. Krizhevsky A (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097–1105.