Machine Learning Analysis on X-ray Scattering Data of Cellulose Microfibrils

Yan Zhang[1], Michael Crowley[2], Jacob Hinkle[3], Charudatta Phatak[1], Lee Makowski[4,5]

[1]Material Science Division, Argonne National Laboratory, Lemont, IL 60439
[2]Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401
[3]Computational Science Center, National Renewable Energy Laboratory, Golden, CO 80401
[4]Department of Bioengineering, Northeastern University, Boston, MA 02115
[5]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115

X-ray scattering at the APS is an important tool for observing molecular processes during the deconstruction of cellulose microfibrils due to chemical pretreatment. Nevertheless, application of this tool to biomass degradation is challenging because the heterogeneous nature of the fibrils undergoing deconstruction results in scattering patterns that are difficult to interpret in terms of the molecular processes involved. Here, we apply machine learning approaches on simulated and experimental x-ray scattering intensities to characterize the ensemble of microfibril structures in native and treated samples of cellulose. A set of atomic coordinates for cellulose fibrils of different cross-sectional shapes, size and twist are generated by molecular dynamics using Charmm. Pair-distance distributions of these structures are calculated and scattering intensities are simulated using the Debye formula in cylindrical coordinates. The simulated intensities (organized into a dictionary) are used to fit the experimentally collected intensities (measurements) by determining the relative abundances of the conformations in the dictionary by regression. Principal component analysis as well as spectral clustering analysis demonstrated the redundancy of the dictionary of cellulose structures. Hierarchical support vector machines and neural network methods are applied for cellulose shape classification. A codebook mapping the pair-distance distribution and scattering intensities is trained and generated as the inverse Fourier Transform operator for scattering intensities. The pair-distance distribution function can be calculated fast and accurately by the multiplication of codebook and measured intensity. These methods provide the means to track alterations in the structural ensemble of cellulose fibrils in biomass undergoing chemical deconstruction processes.

MACHINE LEARNING ANALYSIS OF X-RAY SCATTERING DATA FROM CELLULOSE MICROFIBRILS

Yan Zhang, Charudatta Phatak, Material Science Division, Argonne National Laboratory, Lemont, IL
Michael Crowley, Biosciences Center, National Renewable Energy Laboratory, Golden, CO
Jacob Hinkle, Computational Science Center, National Renewable Energy Laboratory, Golden, CO
Lee Makowski, Department of Bioengineering, Northeastern University, Boston, MA

X-ray scattering of cellulose microfibrils in maize stover

- **X-ray scattering at the APS:** Important tool for observing molecular processes during the deconstruction of cellulose microfibrils due to chemical pretreatment.
- **The heterogeneous nature of cellulose:** Cellulose fibrils undergoing deconstruction results in scattering patterns that are difficult to interpret in terms of the molecular processes involved.
- **Molecular and scattering simulation:** Molecular dynamics and Debye formula for scattering intensity simulation of large number of structures.
- **Machine learning approaches:** Apply on simulated and experimental x-ray scattering intensities to characterize the ensemble of microfibril structures in native and treated samples of cellulose.

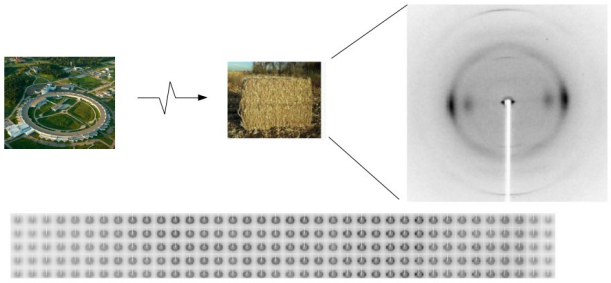


Fig 1. X-ray diffraction pattern of cellulose in maize stover and data collection at Advanced Photon Source.

Dictionary of simulated scattering intensities

- **Dictionary of simulated cellulose microfibril structures:** 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.
- **Dictionary of simulated scattering intensities:** Fast Debye method is used to simulated the x-ray scattering intensities of cellulose microfibrils.
- **Pair-distance and Intensity mapping:** A codebook is trained to map the pair-distance distributions with the scattering intensities.

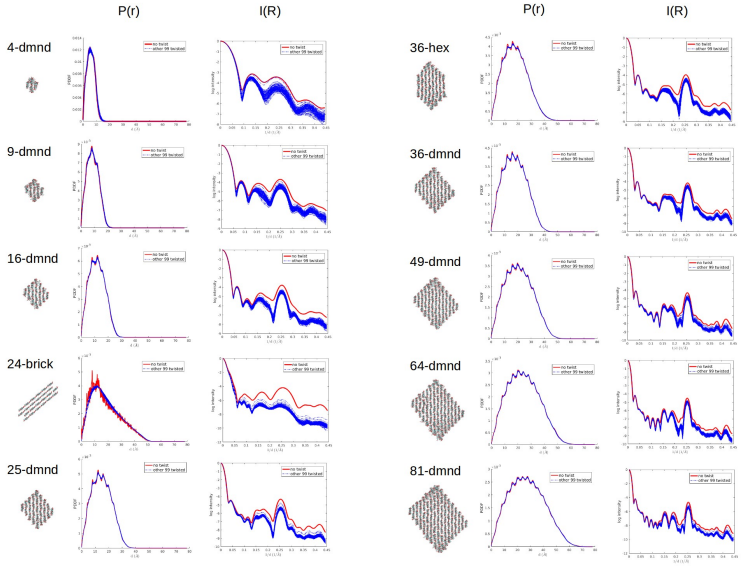


Fig 2. Pair-distance distribution and scattering intensities of cellulose microfibrils.

Classification and clustering

Hierarchical support vector machines is applied for cellulose shape classification. Since SVM is binary classifier, we need to train N−1 mini-classifiers for N classes in the hierarchical SVM architecture.

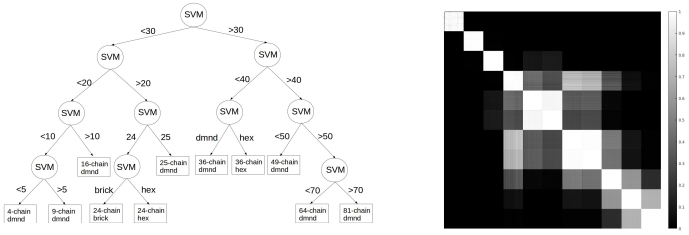
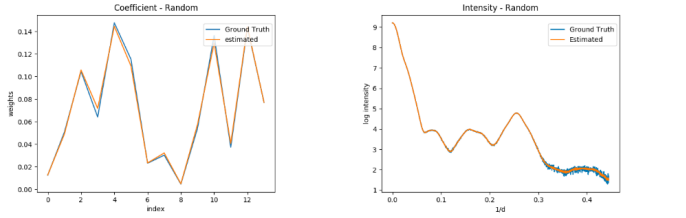


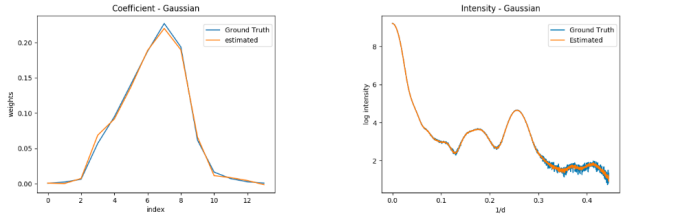
Fig 5. Hierarchical SVM (left) and Affinity matrix for spectral clustering (right).

Mixture of intensities and regression:

The simulated intensities are used to fit the measurements by determining the relative abundances of the conformations in the dictionary.



(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 3. Synthetic mixture of intensities and the estimation of coefficients (right).

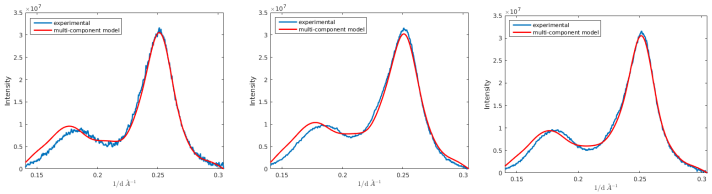


Fig 4. Wide-angle scattering intensities and fitting of untreated, dilute acid and acidic iron treated maize stover (from left to right).

Acknowledgement

APS: GM/CA, beamline 23ID-B. The authors would like to thank Dr. Robert Fischetti and Dr. Nagarajan Venugopalan on data collection.
C3Bio: Center for Direct Catalytic Conversion of Biomass to Biofuels.
LDRD: Integrated imaging at Argonne National Laboratory.

Reference

[1] Zhang, Yan, et al. "Breakdown of hierarchical architecture in cellulose during dilute acid pretreatments." Cellulose 22.3 (2015): 1495-1504.
[2] Zhang, Yan, et al. "Diffraction pattern simulation of cellulose fibrils using distributed and quantized pair distances." Journal of Applied Crystallography 49.6 (2016).