Tissue-specific variation of softwood nanostructure revealed by scanning wide-angle X-ray scattering and clustering analysis

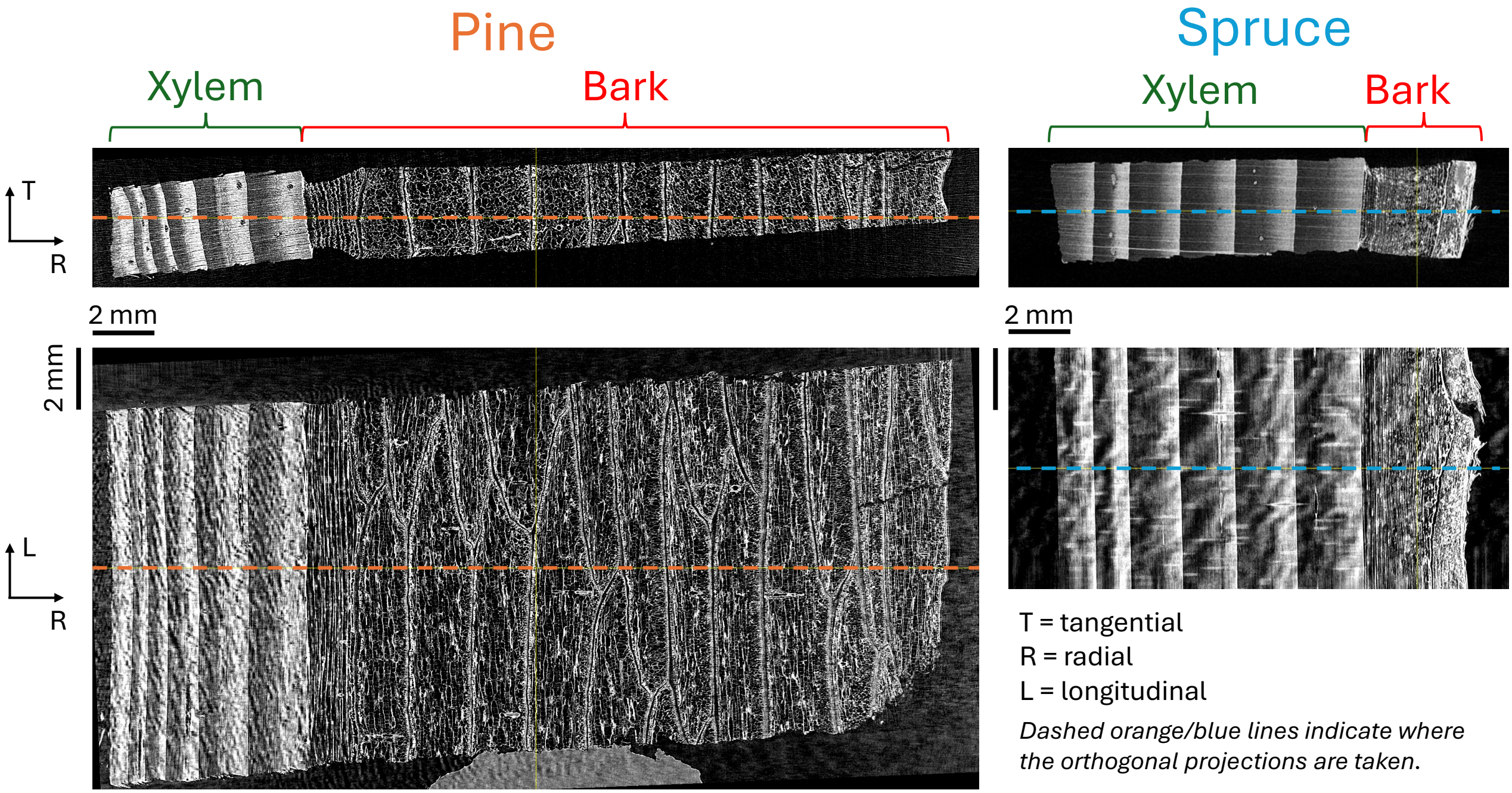
Patrik Ahvenainen[1*,2], Enriqueta Noriega Benitez[1], Tuula Jyske[3,4*], Mira Viljanen[2,5*], Heikki Suhonen[2], Paavo Penttilä[1]

*[1]Department of Bioproducts and Biosystems, Aalto University, Finland; [2]Department of Physics, University of Helsinki, Finland; [3]Natural Resources Institute Finland; [4]Department of Forest Sciences, University of Helsinki, Finland; [5]MAX IV Laboratory, Lund University, Sweden; *=current affiliation*

Corresponding authors: Patrik Ahvenainen (patrik.ahvenainen@aalto.fi) and Paavo Penttilä (paavo.a.penttila@jyu.fi)

SUPPLEMENTARY

Supplementary figure S1. Orthogonal projections of X-ray microtomography reconstructions



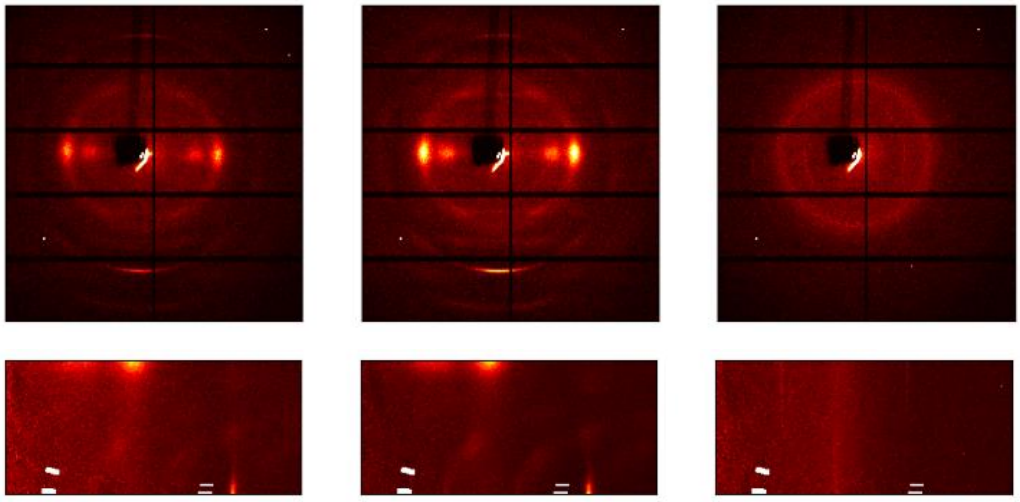
T = tangential
R = radial
L = longitudinal

Dashed orange/blue lines indicate where the orthogonal projections are taken.

X-ray scattering data pre-processing

The X-ray scattering data was transformed into a standardized format of azimuthal angle versus scattering vector length q. For the benefit of data reduction, the scattering patterns were limited to an azimuthal region of 0 to 90 degrees and q-values of 0.5 to 3.0 $Å^{-1}$. Using symmetry, data points in the scattering pattern outside that azimuthal range were mirrored into that range and the final values are an average of all obtained symmetry points. Before mirroring, the 0 azimuthal angle of the scattering pattern was selected to be the position of the cellulose Iβ 200-peak, to correct for any remaining tilt in the sample in the longitudinal-radial plane. Microtomography taken before the scanning-WAXS measurement was used to adjust the angle of the specimen along the longitudinal axis to align the WAXS beam to be fully perpendicular to the annual growth rings. For the pine specimen, alignment was done with respect to the bark. Information from all symmetry points were not available due to detector geometry (gaps between detector elements, **Supplementary Figure S2**).

Clustering methods

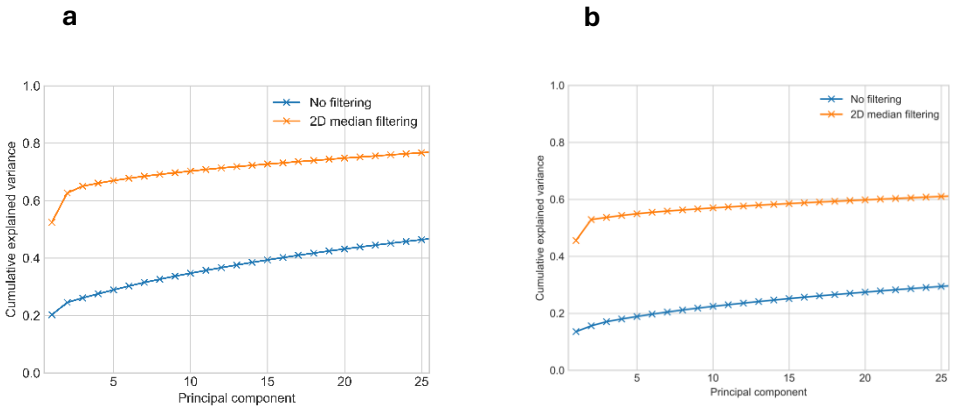


Supplementary Figure S2: The top row shows the raw images from the WAXS detector, with gaps between the elements (black horizontal and vertical stripes). The bottom row shows the mirrored and averaged scattering patterns with 0-to-90-degree azimuthal angles (vertical axis, pointing downwards) and 0 to 3.0 $Å^{-1}$ scattering vector length q (horizontal axis, from left to right). White spots indicate lack of complete information from all four symmetry points, and they were masked out in the analysis.

Clustering algorithms are essential tools for grouping similar data points based on shared characteristics, often used in combination with PCA to analyze complex datasets. There are

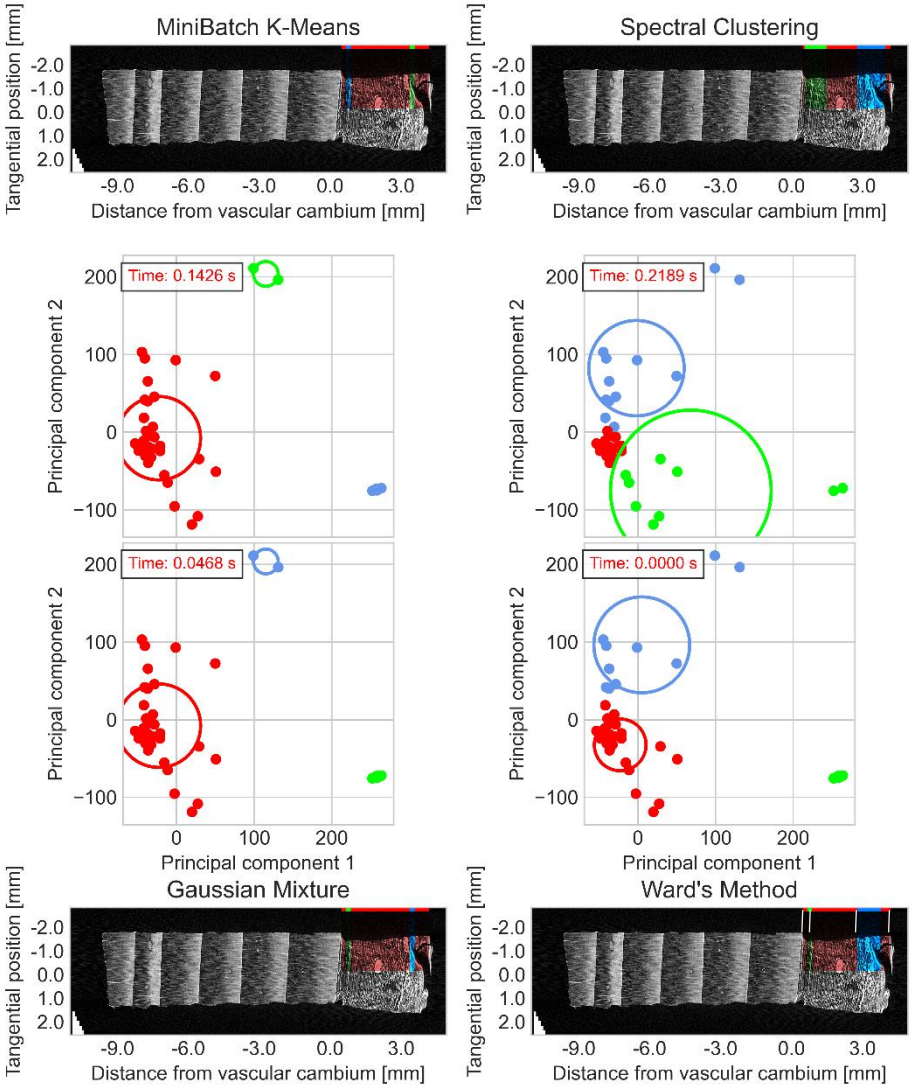
various types of clustering algorithms, such as k-means, spectral clustering, Gaussian mixture models (GMM), and Ward's method, each with unique approaches.

K-means assigns data points to a fixed number of clusters by minimizing the distance to centroids (Ding & He, 2004), which is useful for partitioning distinct groups like earlywood and latewood in wood tissues. Spectral clustering leverages eigenvalues of similarity matrices to group points, making it effective for complex structures where clusters may not be spherical (Aggarwal and Reddy, 2018), like separating bark from annual rings. Gaussian mixture models assume that data is generated from a mixture of several Gaussian distributions, enabling soft clustering where points can belong to multiple clusters with varying probabilities, which could help identifying transitional tissue regions (Reynolds, 2009). Ward's method minimizes variance within clusters at each step and builds a tree of clusters, offering detailed insights into nested groupings such as differentiating between large-scale structures and finer features within the wood sample (Ward, 1963). Some algorithms assume equal-sized spherical clusters, while others have more flexible structures. In practice, the selection is mainly guided by the interpretation of the results.



Supplementary Figure S3 Scree plots from PCA of WAXS patterns for a) spruce and b) pine showing the cumulative explained variance as function of number of principal components with and without 2D median filtering (5x5 kernel). Median filtering was shown here only to highlight how much of the variance in the data is noise that could be removed by filtering. Noise was not an issue for data analysis, so filtering was not used to avoid the filtering effects on peak height determination.

Supplementary Figure S4 Clustering results for the PCA representation of the WAXS data from spruce bark only, projected on principal components 1 and 2. A total of 2 principal components were used for the spruce bark clustering, performed on PCA of only the bark portion of the scattering data. Tomographic reconstructions shown for visual reference. The thresholds used for spruce bark classification with Ward's method are shown with light grey vertical lines in the tomographic reconstructions.



Principal component analysis noise reduction

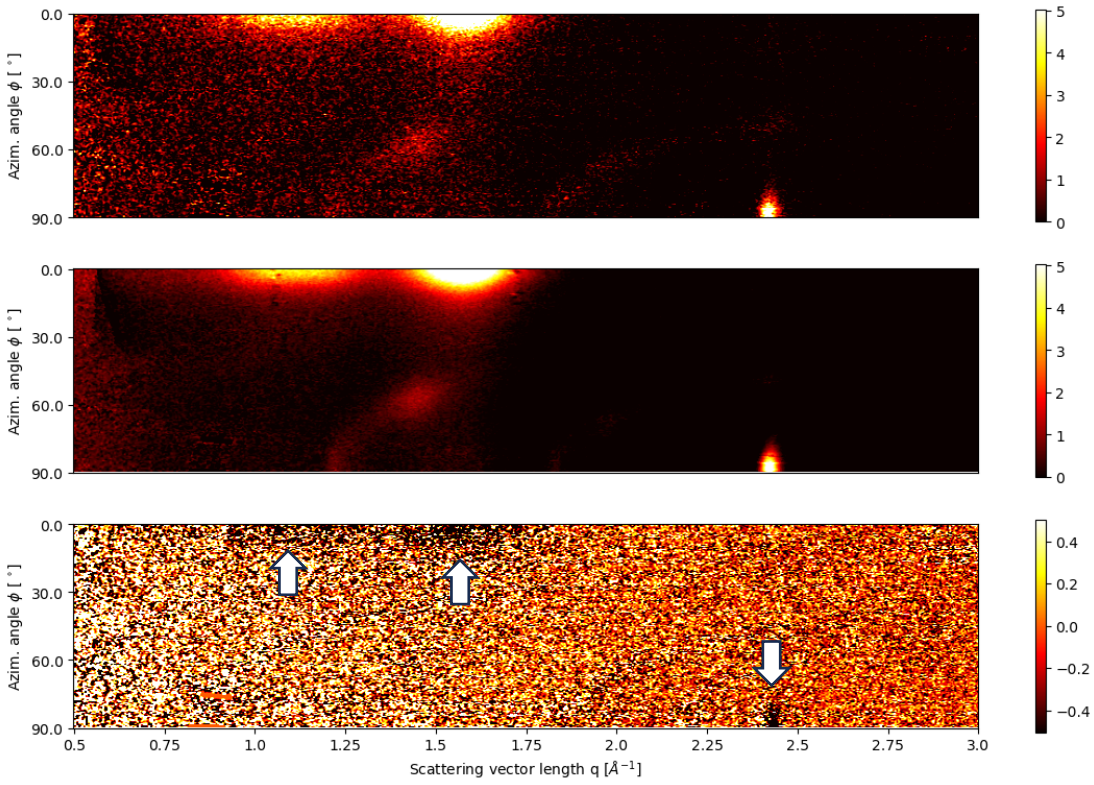
Supplementary Table S1: The PCA noise reduction affects the averages of the mean microfibril angle values for first and middle annual growth, but not for the last annual growth.

	Spruce			Pine		
	First annual growth†	Middle annual growth†	Last annual growth	First annual growth†	Middle annual growth†	Last annual growth
MMFA°*	9.3±0.5	7.9±0.7	6.7±0.1	11.3±1.8	9.6±1.0	8.8±0.6
MMFA°*, PCA noise-reduced	7.46±0.09	7.08±0.18	6.85±0.04	8.60±0.02	8.48±0.07	8.44±0.04
Data points	10	43	10	8	11	12

MMFA = mean microfibril angle, determined by the Cave T-method

PCA performed with 10 principal components

† = The average values of noise-reduced and non-noise reduced parameters are statistically significantly different (two-sided t-test, significance value 0.01).



Supplementary Figure S5: A pine scattering pattern in the azimuthal angle vs. scattering vector length, q, format. Top image is the original, noisy data and the middle is reconstructed from the principal components as part of the whole data set, using the 10 first principal components. The bottom image is the difference image that shows that mainly noise is reduced between the images (except around the strong cellulose diffraction

maxima, white arrows). Noise reduction is significant but microfibril orientation information is easily lost. Colorbars indicate intensity (in arbitrary units), with a different color scaling used for the difference image.

References

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. Proceedings of the 21st International Conference on Machine Learning, 1–9. https://doi.org/10.1145/1015330.1015408

Reynolds, D. A. (2009). Gaussian mixture models. Encyclopedia of Biometrics, 741–745. https://doi.org/10.1007/978-0-387-73003-5_196

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845