Accelerated phase identification using deep learning for retrieving physical parameters from SAXS data

Smita Chakraborty smita.chakraborty@ri.se
Department of Computer Science
RISE Research Institutes of Sweden

Abstract

Small-angle X-ray Scattering (SAXS) provides crucial insights into the structure of cellulose fibres. However, extracting quantitative physical parameters such as fibril diameter, orientation distribution, and porosity from SAXS patterns is challenging due to signal complexity and model ambiguity. This ongoing work presents deep learning and physics-informed neural network (PINN) approaches that leverage both experimental SAXS data and the underlying physical laws governing X-ray scattering.

Keywords: causal inference, physics-informed neural network, SAXS, machine learning, deep learning, cellulose, lipid nanoparticles

1 Introduction

This is an early-stage research of the scientific application of machine learning in the interdisciplinary project AI-SAXS at RISE.

Small-angle X-ray scattering is a versatile scattering method for characterising materials in the order of nanometer scales ranging 1-100 nm. It is a non-invasive reciprocal space characterisation technique that provides statistically representative microstructural information about a material. By sending X-rays through a sample and recording how they scatter at small angles, it reveals details about the sample's internal architecture, such as pore size, distribution, and arrangement of particles or domains.

From Bragg's law, $\mathrm{n}\lambda = 2\mathrm{d}\sin\theta$ where λ is the X-ray wavelength, $d =$ distance between atomic planes in a crystal sample, and θ is the angle of incidence - it is evident that with decreasing scattering angles, larger structural features can be increasingly determined. The raw spatial arrangement of atoms or nanostructures in the material creates a unique scattering pattern when X-rays traverse it. In SAXS, the observed intensity profile is not a direct map of distances, but rather a result of mathematical transformations (specifically, the Fourier transform of the electron density distribution) which encodes information about different structural parameters in the resulting scattering curve. These include size distributions, shapes, and surface structures at the nanometer scale, all "overlaid" as features within the measured intensity profile. Additionally, q in a SAXS profile is directly proportional to frequency space ν, where q is given

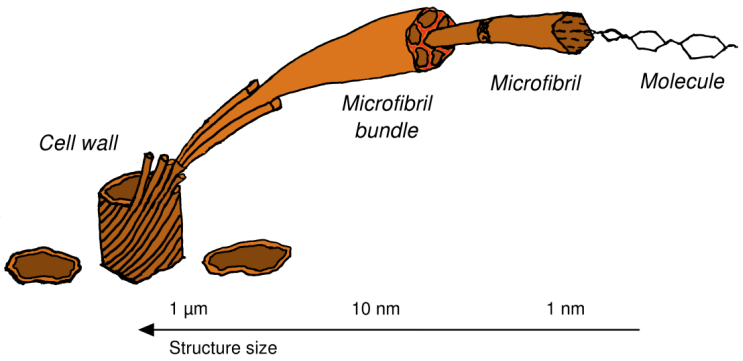

Figure 1: Schematic diagram of hierarchical structure of the wood cell wall from (Penttilä [2013]).

by $(4\pi \sin\theta)/\lambda$, so the scattering profile becomes independent of the incident wavelengths. So the SAXS profile of a material is a fingerprint of its structural complexity.

Purely theoretical inference of the inverse mapping of the SAXS profile to the parameter values of a material is a scientifically challenging task due to its nature, as mentioned before. Machine learning (ML) models have opened new avenues to this complex problem - the models can map within uncertainties which parameters are the most dominant ones for a given region in a SAXS curve (Röding et al. [2022], Anker et al. [2023]). However, one would also need the distribution ranges of parameters for such a SAXS profile to better understand the material behaviour, for example, in the manufacturing of paper straws.

2 Datasets: cellulose and lipid nanoparticles

The datasets in this project are cellulose fibres and lipid nanoparticles (LNPs). Cellulose is the most abundant biopolymer on Earth, which has an anisotropic structure with linear, unbranched fibres that form strong hydrogen-bonded microfibrils. This implies that there is almost no structural symmetry, which leaves us with a challenging simulated modelling of the SAXS profile.

LNPs represent a precise nanoscale assembly of lipids with programmable chemical and physical properties for biomedical applications. A lipid nanoparticle is a nanoscale assembly made up of multiple types of lipid molecules, forming a structured carrier rather than a single molecule. Its typical diameter ranges from 10 to 1000 nanometers, and it is engineered to encapsulate genetic material or drugs for delivery into cells. A lipid nanoparticle is simulated using a Gaussian random field and an electron density. LNPs have inherent structural symmetry.

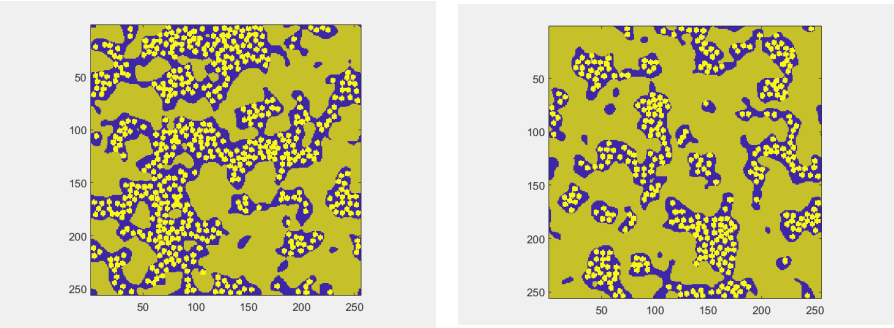

Figure 2: Fibril aggregates in cellulose model with two different area fractions. The length scale of the fibres is obtained from the model fit. Image credit: Sandra Barman.

2.1 Cellulose dataset

The cellulose dataset has two counterparts: experimental SAXS data collected from the Max IV laboratory and its simulated dataset. The simulated SAXS dataset contains physical parameters and generating properties of the spatial model, including the mean half-size of crystals (`alpha_mean`), the associated standard deviation, the spectral filter for the Gaussian Random Field (GRF) function that generates the amorphous regions (`pow_exp`) and the smoothing of the same at low q values, called μ. There are SLD values that are the relative electron densities compared with the crystals ($\text{SLD}_{\text{crystal}} = 1$ is the default value). The other parameters are the volume fractions of the three different phases: crystals, GRF phase 1, and GRF phase 2. The last parameter is the surface area, which is the total length of the boundaries between all phases. An example figure of such a cellulose model is shown in figure 2. In total, we have 10 parameters in the simulated dataset.

2.1.1 Generating the simulated SAXS dataset

Each simulated scattering profile (q vs $I(q)$) has a setting for each of the generating parameters set of values. Each column consists of 500 sampled values of the SAXS curve. There are 10000 different parameter settings and corresponding 10000 SAXS scattering profiles. The ML model is trained on a subset of this simulated dataset.

3 Preliminary results

For a pilot test to see how well a simple regressor ML model can learn about the physical parameters from the SAXS data, we trained a model on a subset of the parameter dataset and the corresponding SAXS data. We then test and predict the physical parameters using this model.

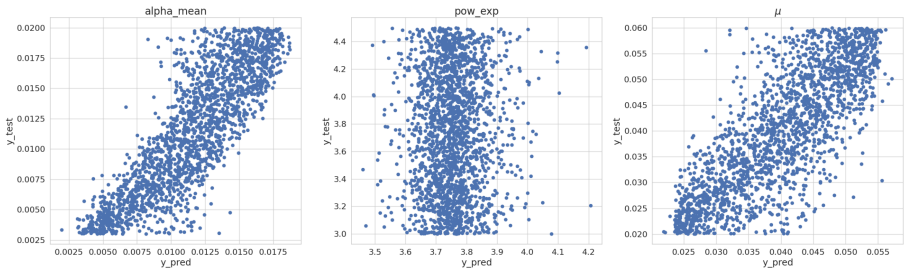

Figure 3: Scatter plots showing the `y_test` vs `y_pred` for the parameters `alpha_mean` (left), `pow_exp` (center), and μ (right).

We used randomly selected indices from the parameter dataset and the same indices to select SAXS datasets for training our model. For regression tasks, we utilised the XGBoost algorithm, implemented via the XGBRegressor class. The model was configured with a squared error objective function, 100 boosting rounds (`n_estimators` = 100), a learning rate of 0.1, and a maximum tree depth of 5. These hyperparameters were selected based on preliminary tests to balance model complexity and generalisation. To ensure reproducibility, we fixed the random seed to 42. Hyperparameter tuning was performed using grid search with cross-validation, optimising for mean squared error on the validation folds.

The `y_test` vs `y_pred` scatter plot from an XGBoost regressor visually examines how well the model's predictions align with the ground truth values. This is shown in figure 3 for the three parameters, `alpha_mean`, `pow_exp` and μ. Each point in the plot represents one observation, where the x-axis is the predicted value (`y_pred`) and the y-axis is the true value (`y_test`). As is visible, the model is not able to learn the `pow_exp` parameter well, but μ and `alpha_mean` are better predicted, with better results in predicting the latter. Also, the model does not inherently constrain predictions to the observed range, sometimes resulting in predictions beyond realistic bounds, therefore, constraints on the prediction parameter values can improve the performance of the model.

These observations emphasise that small-angle X-ray scattering (SAXS) intensity profiles represent convoluted functions of the underlying cellulose electron density. Consequently, reconstructing physical parameters such as `alpha_mean` and μ from reciprocal-space scattering data I(q) constitutes an ill-posed inverse problem, wherein the mapping from Fourier to real-space is inherently unstable and non-unique.

We are working on a pilot study of using a physics-informed machine learning model, where we do not use a parameter dataset for a given SAXS profile to train our ML model, instead, we encode the structural guiding equation in the loss function of the model. Physics-informed machine learning models have been used in the wide-angle X-ray scattering region, benefiting from Bragg's law. The law is still valid in the small-angle regime; however, due to the small-angle approximation $\sin\theta \approx \theta$, the resolution in the parameter space becomes quite

narrow, hence hard to distinguish between changes in the parameter values. We are working on solving this issue in the SAXS regime.

4 Future work

By explicitly encoding SAXS scattering equations and structural constraints into the model's loss function, the PINN learns to predict relevant physical parameters while maintaining consistency with the structural constraints of the material sample and the physical theory of SAXS. Also, other neural network architectures are being explored to estimate parameters directly from experimental datasets, which reduces the need for simulated datasets and lowers the computation and time required for this task (Molodenskiy et al. [2022], Wong et al. [2024]). These approaches enable robust inverse modelling across diverse sample types, reduce reliance on labelled datasets (such as parameters from simulated models), and improve the interpretability of machine learning outcomes for cellulose fibre analysis.

In upcoming publications, we intend to show how causal inference models can predict the parameter distribution of the sample and predict other environmental properties, such as relative humidity in a cellulose sample. Currently, this information is available only from experimental SAXS data and is determined only via fitting the SAXS curve with simulated datasets. We aim to determine properties such as the relative humidity using a trained ML model on a simulated dataset. This would be beneficial for better modelling cellulose and reducing reliance on experimental measurements, in this case.

5 Acknowledgements

The work is done in collaboration with Sandra Barman, Shun Yu, Sepideh Pashami, Maria Bånkestad, and Jerk Rönnols from RISE and with Ann Terry in Max IV laboratory, Eskil Andreasson in Tetra Pak, Magnus Röding and Erik Kaunisto in AstraZeneca.

AI-SAXS project is funded through Vinnova's Advanced Digitalization program. The project is an example of how new technologies and interdisciplinary collaboration can pave the way for innovations that strengthen Swedish industry and research in materials science. The project is a collaboration among RISE, Tetra Pak, AstraZeneca, and the Max IV laboratory.

References

A. Anker et al. Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry. *Chemical Science*, 14:14003–14019, 2023. doi: 10.1039/D3SC05081E. URL `https://pubs.rsc.org/en/content/articlelanding/2023/sc/d3sc05081e`.

D. S. Molodenskiy, D. I. Svergun, and A. G. Kikhney. Artificial neural networks for solution scattering data analysis. *Structure*, 30(6):900–908.e2, 2022. ISSN 0969-2126. doi: 10.1016/j.str.2022.03.011. URL https://doi.org/10.1016/j.str.2022.03.011.

P. Penttilä. *Structural characterization of cellulosic materials using x-ray and neutron scattering*. PhD thesis, University of Helsinki, 2013. URL http://hdl.handle.net/10138/40782.

M. Röding, P. Tomaszewski, S. Yu, M. Borg, and J. Rönnols. Machine learning-accelerated small-angle x-ray scattering analysis of disordered two- and three-phase materials. *Frontiers in Materials*, Volume 9 - 2022, 2022. ISSN 2296-8016. doi: 10.3389/fmats.2022.956839.

K. Wong, R. Qi, Y. Yang, Z. Luo, S. Guldin, and K. T. Butler. Predicting colloidal interaction parameters from small-angle x-ray scattering curves using artificial neural networks and markov chain monte carlo sampling. *JACS Au*, 4(9):3492–3500, 2024. doi: 10.1021/jacsau.4c00368. URL https://doi.org/10.1021/jacsau.4c00368.