Blogpost No.2: Deep Learning Pipeline for Surface Scattering Data

L. Pithan, V. Starostin, A. Gerlach, A. Hinderhofer, F. Schreiber
University of Tübingen

Paper: End-to-End Deep Learning Pipeline for Real-Time Processing of Surface Scattering Data at Synchrotron Facilities

Challenge

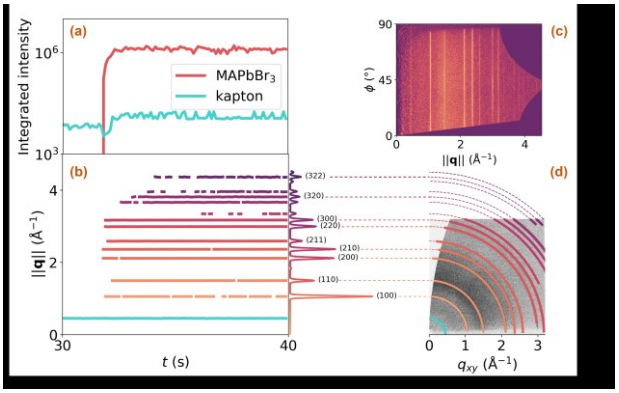
Data analysis is a major bottleneck in many fields of experimental science. With increasing data rates of beamlines at modern synchrotron facilities, this bottleneck becomes even critical for the success of surface diffraction experiments, in order to keep pace with the high-speed data streams.

Traditionally, data analysis of experiments performed at synchrotron or neutron facilities takes place after the beamline visit in the researcher's home institution. Due to the ever-increasing size and acquisition rates of modern area detectors, for many synchrotron users, transferring terabytes of scattering data to the home institutes has become a challenge in itself, not to mention analyzing those data on local computing resources. Even beyond the data transfer and storage challenge there is the demand to make data-driven decisions during the experiment, e.g., for real-time scattering studies of kinetic processes. Therefore, online data processing and analysis have become key factors for many experiments.

Today, large-scale research facilities such as synchrotrons provide a powerful computing infrastructure along with their experimental instruments. While computer clusters at synchrotrons advance rapidly and give users access to a wide range of computing resources, real-time data analysis based on user-developed software, usable at different beamlines and facilities remains a challenge.

Solution

In our article (Synchrotron Radiation News, 35:4, 21-27, *npj Comput Mater* 8, 101, 2022 https://doi.org/10.1038/s41524-022-00778-8), we discuss challenges and possible approaches for building a Grazing Incidence X-ray Diffraction (GIXD) data processing pipeline with emphasis on machine learning applied to data analysis. We demonstrate the implementation of such a software framework using gixi (Grazing Incidence X-ray diffraction Intelligent pipeline), an open-source package based on a



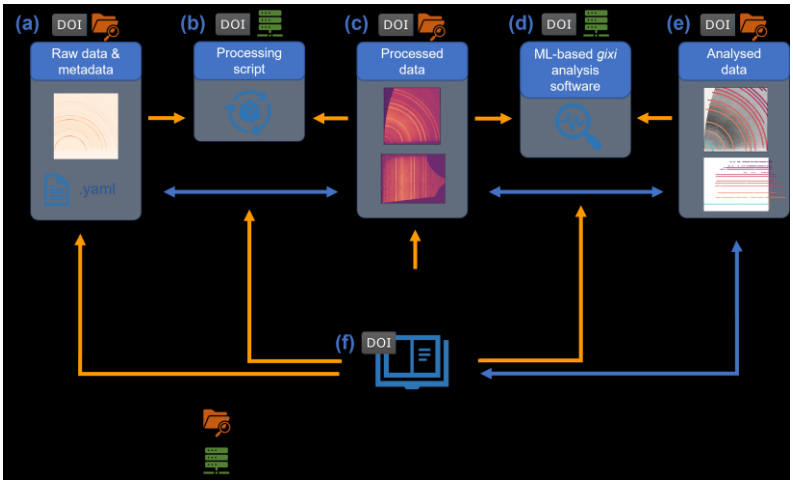
Detected diffraction peaks are identified and matched against the diffraction profile of $MAPbBr_3$

deep learning approach. It provides an end-to-end solution for automated GIXD analysis, including image processing, detection of the diffraction peaks, peak intensity extraction, and crystal structure identification. Our user-designed implementation allows a straightforward integration into any cluster infrastructure based on the commonly used Slurm Workload Manager.

Our Approach

We have developed a pipeline-based approach of GXID data processing. In general, GIXD images contain rich information about the sample. Based on the specific scientific problem, different types of analysis are required to extract the relevant properties of the studied system. These may include the lattice parameters, the texture, and fractions of co-existing mixtures in case of powder diffraction, time-dependent properties of the studied process in case of *in situ* measurements. Due to the large number of related quantities, it can may be inefficient to develop separate software solutions for each type of analysis. This is where customizable pipelines gain in importance.

In the core of our pipeline, we embed a tailor-made neural network (npj Comput Mater 8, 101, 2022, https://doi.org/10.1038/s41524-022-00778-8). Deep learning is a promising choice for analyzing complex 2D scattering data with various experimental artifacts and diffraction features. In order to be able to use the neural network for peak detection, the raw diffraction images obtained from the detector require certain pre-processing steps in the beginning of the pipeline before they can be fed into the CNN (convolution neural network). Following the peak detection, in a further step towards the end of the pipeline, crystal structures present in the sample are identified based on the obtained peak positions and comparison to known crystal structures (provided as CIF files). Due to the modular approach of the software architecture, we emphasize that this model can be easily combined with other indexing algorithms for structure identification.



Scheme of the bidirectionally linked open-data publication chain

Practicing Open Science

With the DAPHNE vision in mind we followed the FAIR principles when publishing our work. DAPHNE4NFDI works towards a standardized, transparent and traceable chain of all steps from the raw data to the final peer-reviewed scientific publication.

Applying the guidelines provided in the DAPHNE proposal (https://www.daphne4nfdi.de/downloads/Daphne_proposal.pdf) we provide interlinked

example datasets from raw- to analyzed data as well as the source code and parameters for all data processing steps involved in the data analysis chain using Zenodo infrastructure.

In the future DAPHNE is addressing this challenge by encouraging and supporting the use of SciCat (https://scicatproject.github.io), a data catalog developed within the PaN community to facilitate traceability from data to publication and vice versa.

Additional links

Project DAPHNE4NFDI: https://www.daphne4nfdi.de/english/index.php

Further resources on this topic can be found on the Schreiber Research Group website:

http://www.soft-matter.uni-tuebingen.de/machine_learning_GIWAXS.html

If you are interested in a postdoc, PhD or master position in Prof. Schreiber's group, please send an e-mail to softmatter [at] ifap.uni-tuebingen.de.