Data Driven Approaches for Characterization Techniques with Applications in Materials Science

A. Hexemer[1], S. Liu[1], C. N. Melton[1], R. Pandolfi[1], D. McReynolds[1], D. Kumar[1],

M. Noack[1], D. Ushizima[1] and J. Sethian[1]

[1]Lawrence Berkeley National Lab, 1 Cyclotron Rd, 94720 Berkeley, US

The materials discovery cycle contains many different components, including synthesis, characterization and data analysis and interpretation. In the past few decades, automatic synthesis pipelines have been established for many chemistry and materials systems. For characterization, many advanced techniques, such as X-ray scattering and NMR crystallography, have enabled the structure identification of various chemical, biological and materials systems, including polymers, inorganic materials and proteins. These techniques have been developed and improved substantially over the past few decades, which brings high-throughput experimental discovery into reach. Meanwhile, these breakthroughs produce very large data amounts. However, the process of understanding structural feature from data is still very labor-intensive. It requires many man hours of work by highly specialized and trained scientific staff to interpret the data and identify the structure correctly. Therefore, from the experimental side, the next generation of materials research requires a novel approach to address these challenges. From the computational side, high performance simulation methods have been developed to understand the structures of underlying materials systems. Even in the case that the simulation process is not the bottleneck, it is still not trivial to map the characterization result to the underlying structures. To accelerate materials research, a different and more flexible approach is needed to address these challenges from both experimental and computational sides and enable high-throughput materials discovery. Recently, machine learning, a branch of artificial intelligence, has demonstrated the capability to tackle many challenging chemistry and materials problems, including machine-learning-assisted materials discovery, drug design and crystal structure representations [1]. Previously, histogram of gradient (HOG) and Support Vector Machines (SVM) methods have been applied to predict X-ray scattering experiment configurations with more than 80 classes [2,3]. We propose a novel approach: integrate machine learning methods into characterization techniques to categorize and manage experimental data, identify the structures, understand the chemistry-nanostructure relationship, approximate state-of-art computational prediction results, and optimize characterization facility parameters.

X-ray scattering has been applied to characterize different materials chemistry systems. Nowadays, X-ray scattering experiments can be conducted in a high-throughput manner driven by high speed detectors and high brightness sources. It is important to develop scientific procedures to manage and analyse large-scale datasets. Herein, we propose a machine learning based hierarchical categorization approach to manage and classify the data, so that appropriate analysis pipelines can be applied autonomously. More importantly, this framework can be potentially integrated into an automatic materials chemistry discovery process, together with high-throughput synthesis and robotic X-ray scattering experiments. Scattering data can be categorized based on different criteria. For example, depending on the geometry of the experiment, the data can be categorized as transmission or grazing incidence small angle X-ray scattering data. In addition, the data can be categorized by characteristic features. Different features, such as rings, arcs, rods and Bragg peaks, usually require different analysis approached and data reduction methods. For example, ring patterns in transmission X-ray scattering data, usually require radial integration. To expand on this approach, we propose a framework for materials discovery using X-ray scattering by leveraging a large-scale experiment database and different machine learning methods. To apply machine

learning algorithm to experimental data we start by organizing the scattering data using a flexible database containing experiment information, labels from domain experts and predicted labels from trained machine learning models. We built a database containing more than 500,000 images in collaboration with users of the SAXS/WAXS beamline at the Advanced Light Source. A convenient web application for data labelling was developed, and about 11,000 experimental images were labelled. The labelled data is used to train machine learning models using our hierarchical approach shown in figure 1, which allows us to categorize each X-ray scattering data's features individually, starting from the coarse-grain information (such as geometry of X-ray scattering experiment), to the fine-grain information (such as ring or crystalline features). The trained network was then applied to in-situ experiment data to demonstrate its ability to recognize and automatically adjust to phase transitions of real-time data.

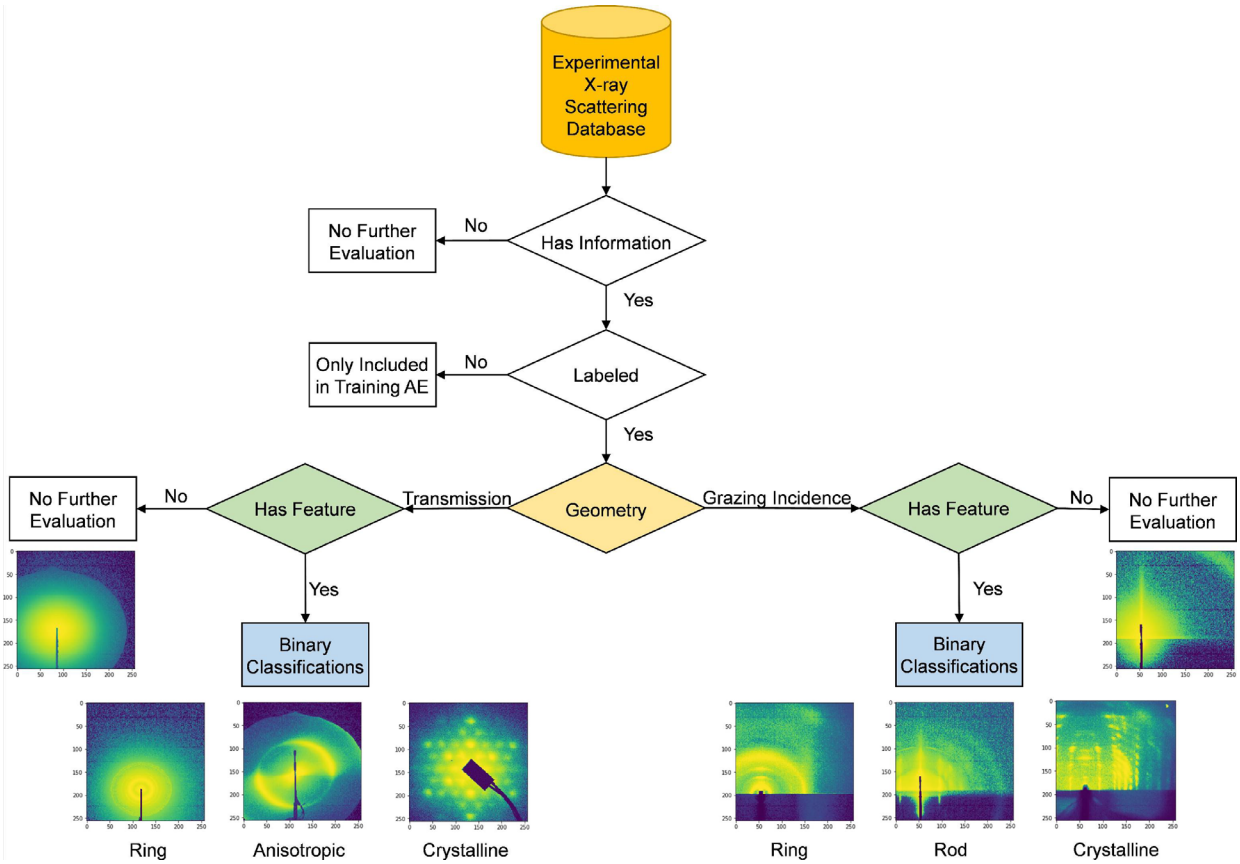

Figure 1: The hierarchical categorization method for X-ray scattering data.

References

[1] Shuai Liu et al. Materials Research Society - Special Issue on Artificial Intelligence, pp.1-7, 2019.

[2] Wang, B.; Yager, K.; Yu, D.; Hoai, M. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp 697–704

[3] Kiapour, M. H.; Yager, K.; Berg, A. C.; Berg, T. L. In IEEE Winter Conference on Appli- cations of Computer Vision, 2014, pp 933–940.