Development of Automatic Phase Change Recognition for Accelerated Materials Discovery

Sam Dawley, David Elbert[2], Tyrel McQueen[1], Apurva Mehta[3]
[1]Department of Chemistry, Johns Hopkins University, Baltimore, MD, 21218
[2]Hopkins Extreme Materials Institute, Johns Hopkins University, Baltimore, MD, 21218
[3]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, San Mateo County, CA, 94025

Abstract

How do we know if a spectral peak is meaningful? Answering this question is central to extraction of information such as the onset of phase changes. Recent advances in experimental methods and detectors allows collection of more data more quickly and efficiently than ever before, providing the opportunity to leverage data-intensive methods, such as artificial intelligence and machine learning, to more rigorously evaluate spectra in real, or near real-time. Such on-the-fly data analysis provides the opportunity to drive the decision-making process during experimentation with live-streamed data, as it is being collected from an instrument. On-the-fly analysis is central to creating autonomous experimental control and characterization of fundamental phenomena such as phase changes. Firstly, however, an understanding of noise and signal structure must be established to allow rigorous, repeatable analysis of the spectral data stream. Herein a statistical algorithm and programmatic implementation for signal structure and phase change detection is introduced as a means of advancing methods for the on-the-fly data analysis.

Methods

Diffraction patterns were collected with a Bruker D8 X-ray powder diffractometer. All represent the diffraction of a lanthanum hexaboride sample with minor impurities. Patterns were collected over multiple time intervals, as in Figure 1. All analysis was performed in Python using SciPy.

Peak detection was performed by making pairwise comparisons of local variation. The primary metric of variation used is the coefficient of variation, defined as $\gamma = \sigma/\mu$ for the mean and standard deviation of a set of intensities, μ and σ, respectively. Variation was measured by first partitioning a diffraction pattern into discrete intervals. Then, the coefficient of variation in adjacent regions can be measured and compared. The distribution of sample coefficient of variations follows a chi-squared distribution and therefore the ratio of any two sample coefficients of variation follow an *F*-distribution. The statistic used to test the significance of any pairwise comparison is given by Forkman:

$$F = \frac{c_1^2/(1 + c_1^2(n_1 - 1)/n_1)}{c_2^2/(1 + c_2^2(n_2 - 1)/n_2)}$$

Results

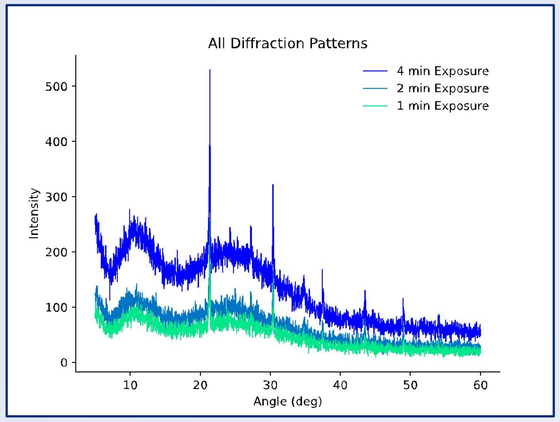
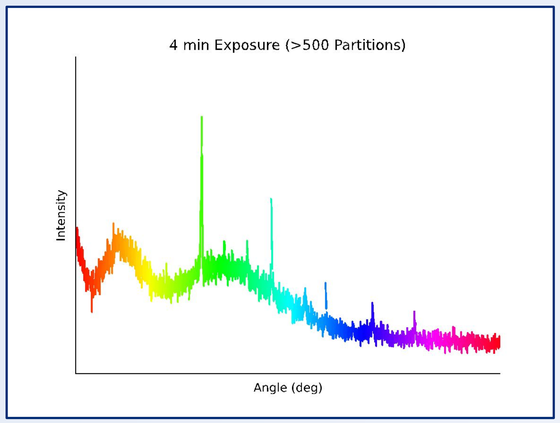


Figure 1. Illustration of X-ray diffraction data over different exposure times. The noise levels appear similar for each series. Most notably, the intensity of any one point is larger and the signal to noise ratio is increased for larger exposure times.

Figure 3. Example of partitioning data set into discrete regions before pairwise comparisons of sample coefficient of variation can be made. The partitions shown are smaller than would be used in practice. The partition size used here is the full-width half-max of the most intense signal



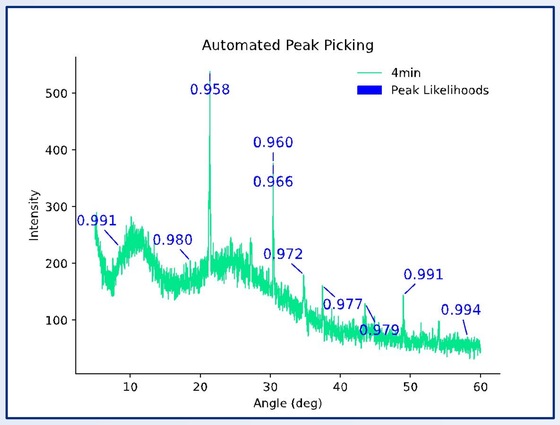


Figure 4. Assignment of peak probabilities from peak picking algorithm on four-minute exposure of lanthanum hexaboride sample. In practice, peak likelihoods for all partitions would be returned in an output file.

Discussion

Averaging intensity over many short exposure-time experiments affords more discernable diffraction patterns than does an equivalent exposure time within a single experiment. Though, the SNR remains consistently lower for short exposure-time experiments no matter the correction for noise reduction.

Future Directions

Current research seeks to quantify the information content contained within a noisy spectrum using a combination of statistical analysis and machine learning. Ultimately, building a model to differentiate significant deviations from baseline noise, i.e., peaks, for on-the-fly data analysis will allow for data-driven experimentation.

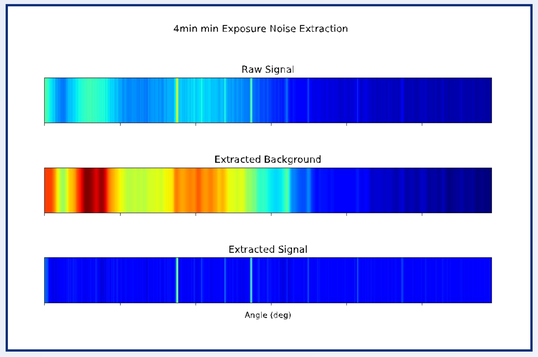


Figure 4. Extraction of noise from four-minute exposure using neural network.

Sources

[1] The Materials Genome Initiative (https://www.mgi.gov/)
[2] Johannes Forkman (2009) Estimator and Tests for Common Coefficients of Variation in Normal Distributions, Communications in Statistics - Theory and Methods, 38:2, 233-251, DOI: 10.1080/03610920802187448.
[3] Ratner, D., Sumpter, B., Alexander, F., et al. *[Office of Basic Energy Sciences (BES)] Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning*. United States. https://doi.org/10.2172/1630823

Acknowledgements

The author would like to thank Dr. David Elbert for his continued guidance and support within and outside of the entire project, and Dr. Apurva Mehta for his contributions to the research and statistical methods. Additionally, a large thanks goes to Prof. Tyrel McQueen for his unwavering support for the project as well as the graduate students within his lab who offer their time to effort to guide visiting students in using the instruments within the lab.

This material is based upon work partially supported by PARADIM and VariMat, organizations funded through the NSF under respective Award Numbers DMR-2150446 and OAC-2129051