Robust Nanostructure from High Throughput Powder Diffraction Data

Simon J. L. Billinge[1,2], Christopher J. Wright[1], Chia-Hao Liu[1], Michael Waddell[1], Pavol Juhas[3], Eric Dooryhee[4], Sanjit Ghose[4], Milinda Abeykoon[4], Arman Arkilic[4], Daniel Allan[4] and Thomas Caswell[4]

[1.] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA
[2.] Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY, USA.
[3.] Computational Science Initiative, Brookhaven National Laboratory, Upton, NY, USA.
[4.] National Synchrotron Light Source-II, Brookhaven National Laboratory, Upton, NY, USA.

Modern materials under study for next generation technologies in energy conversion and storage, environmental remediation, and health are highly complex, often heterogeneous and nano-structured. Here we refer to these as *real materials*. A full understanding of the structure requires us to *go beyond crystallography* and to study the local structure, which is a major experimental and theoretical challenge [1].

Although we have been able to solve the structure of crystals for more than 100 years, when the structure of the material is not periodically long-range ordered, as is the case in real-materials in general, we need characterization methods that extend crystallography, and for nano-materials we are finding that we even need to *replace* crystallography. Microscopy and imaging is playing a critical role in this endeavor, of course. However, in this talk we focus on the nanostructure problem: obtaining the 3D coordinates of atoms in space with high precision, which goes beyond obtaining images.

A recent issue that we have been encountering is that as the structure becomes more complex at the nanoscale, the amount of information in the scattering tends to decrease due to the overlap of broad scattering peaks and the inability to obtain information from single particles. At the same time, the complexity of the structural model needed to describe the structure increases. At some point the structure inverse problem becomes ill-posed and there is not a unique structural model consistent with the data but, in principle a large number of, degenerate solutions. This is illustrated in Fig. 1 where simulations of nanoparticle clusters have been generated consistent with atomic pair distribution function (PDF) data. Each dot gives the fitting figure of merit for a cluster generated during the reconstruction. On the x-axis is the agreement factor between the PDF calculated for the nanoparticle and measured data. On the y-axis is another figure of merit related to surface energy that measures the compactness of the candidate simulated nanoparticle. There is a large number of dots that have as good or better agreement with the data than the correct solution indicating that there is a large number of degenerate models when we just use the PDF data to measure against. We can also use some kind of physical plausibility of the model to differentiate between good and bad models, and one such measure is the mean number of bonds per atom. This measure discriminates against models that have low-coordinated atoms extending out from the surface. Even taking these two measures together we do not find a unique solution. We believe that this problem of model degeneracy is quite widespread, though maybe not widely appreciated.

To address this problem of the posedness of the inverse problem we have been developing an approach called "Complex Modeling" that allows multiple datasets to be complexed or considered together in a multi-modal fit [2]. I will describe this approach in the talk. In this mode we want to give multiple

different experimental and theoretical inputs as constraints to the inverse-problem regression algorithm, as illustrated in Fig. 2.

Extending this idea even further, we can imagine carrying out more experiments on more samples in a high-throughput fashion, then analyzing all the data in the aggregate to obtain a robust solution. We have been developing high-throughput measurement capabilities at the XPD beamline (28-ID-2) at the National Synchrotron Light Source-II (NSLS-II) at Brookhaven National Laboratory and developing highly flexible, but highly automatable software for quasi-real-time data acquisition, reduction and analysis. The software is implemented mostly in Python and utilizes powerful linked databases to capture data and metadata, and a highly configurable workflow that keeps track of provenance and allows rich investigation of the data during the discovery phase, but may be automated for rapid analysis of large amounts of data when the analysis workflow is more mature. In the future, this will feed directly into our Complex Modeling software, taking us towards an end-to-end workflow for high-throughput experiments, built on databases that contain annotated data that is ready for novel data analytic discovery such as machine learning approaches [3].

References:

[1] S. J. L. Billinge and I. Levin. Science **316** (2007), p. 561.
[2] P. Juhás *et al,* Acta Crystallogr. A **71** (2015), p. 562.
[3] We acknowledge useful discussions and help from many members of the NSLS-II software and hardware teams. Funding in the Billinge group was provided by U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences (DOE-BES) under contract No. DE-SC00112704. National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704.
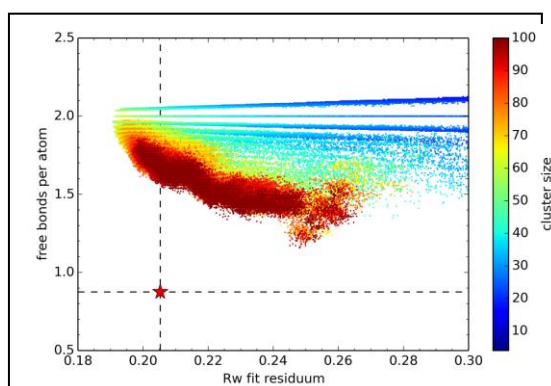
Figure 1. Plot of cluster model degeneracy. Each dot is from one cluster generated in a PDF data-fitting regression process. Dots towards the left axis give good fits to the PDF, dots lower down are from clusters that are more compact. The star shows the correct solution.
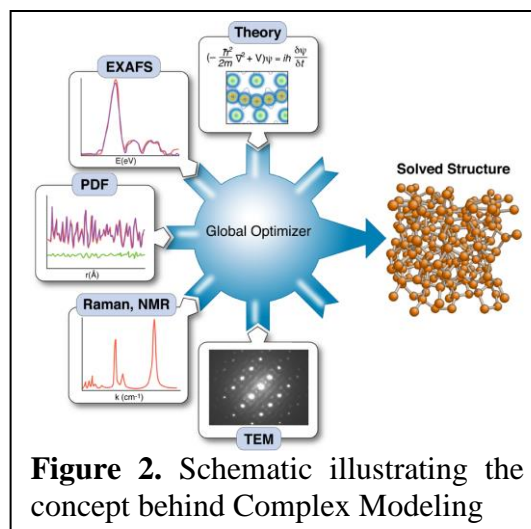

Figure 2. Schematic illustrating the concept behind Complex Modeling