LA-UR-17-29420
Approved for public release; distribution is unlimited.

Title: Measurement and Interpretation of Diffuse Scattering in X-Ray Diffraction for Macromolecular Crystallography

Author(s): Wall, Michael E.
Sweet, Robert
Ando, Nozomi
Fraser, James S.
Phillips, George N. Jr.

Intended for: Report

Issued: 2017-11-15 (rev.1)

**Measurement and Interpretation of Diffuse Scattering in X-Ray Diffraction for Macromolecular Crystallography**

Workshop at the 2017 NSLS-II and CFN Users' Meeting, Brookhaven National Laboratory, Upton, NY, May 15, 2017

Organizers: Michael Wall, mewall@lanl.gov (LANL), Robert Sweet, rsweet@bnl.gov (NSLS-II, BNL), Nozomi Ando, nozomi.ando@princeton.edu (Princeton University), James S. Fraser, jfraser@fraserlab.com (University of California, San Francisco), George N. Phillips, Jr., georgep@rice.edu (Rice University)

X-ray diffraction from macromolecular crystals includes both sharply peaked Bragg reflections and diffuse intensity between the peaks. The information in Bragg scattering reflects the mean electron density in the unit cells of the crystal. The diffuse scattering arises from correlations in the variations of electron density that may occur from one unit cell to another, and therefore contains information about collective motions in proteins.

Leading researchers in diffuse scattering gathered May 15, 2017 for a one-day workshop at the NSLS-II Users' Meeting. A major focus of the workshop was to provide a roadmap to the acquisition of reliable data by surveying measurement methods and discussing the increase in measurement accuracy enabled by improved detectors, experimental methods, and data integration. Another major focus was to survey examples of information that can be extracted about the behavior of biomolecules that would guide the thinking of biochemists and biologists. A number of talks addressed the measurement of diffuse-scattering data and advances in the modeling of the data in terms of conformational variation. Below we give a short synopsis of each talk, and at the end an analysis of the results of the workshop in total.

Michael E. Wall (Los Alamos National Laboratory) focused on using three-dimensional diffuse datasets for model validation and refinement, including real-space validation using diffuse Patterson maps. In a case study of crystalline staphylococcal nuclease, the attenuation of the Patterson at long distances was captured by molecular dynamics simulations. Crystalline normal modes were highlighted as a possible means of obtaining refined models of conformational ensembles that can connect to biological interpretations.

George N. Phillips, Jr. (Rice University) spoke about some of the theory of diffuse scattering based on kinematical arguments, the main point being the value of describing the variance-covariance matrix as a key descriptor of the effect of displacements on the total diffraction patterns of protein crystals. He gave specific examples for conceptual context and for real world macromolecules.

Nozomi Ando and Steve Meisburger (Princeton University) focused on experimental strategies for accurately measuring and processing diffuse-scattering data, a

perspective recently published in a review article (Meisburger et al, 2017). Although diffuse scattering has been approached as an extension of crystallography, its measurement should be viewed as a scattering problem. This was demonstrated by revisiting lysozyme, a system with historical significance in the field, for which a variety of models have been applied. High quality data from lysozyme crystals in two different space groups were presented, and showed promising agreement with MD simulations performed by David Case. In addition, they discussed the importance for conveying biological significance.

Donald Caspar (Florida State University) noted that his phenomenological liquid-like motions model published in 1988 (Caspar et al, Nature) captured essential features of diffuse scattering from insulin crystals, indicating that much of the signal present is due to variations that can be described without mechanistic details. He emphasized the importance of identifying systems whose structure and dynamics can be reversibly controlled, e.g., using pH, to achieve the next level of understanding.

Peter Moore (Yale University) presented the work he and Yury Polikanov have published on the diffuse scattering produced by crystals of Thermus thermophilus 70S ribosomes (Polikanov and Moore, 2015) . Much of that scatter appears to be produced by acoustic vibrations of the crystal lattice, which are unlikely to interest most biochemists, rather than by motions that do not correlate between unit cells, which might interest them. Thus these observations provide evidence of just how hard it is going to be to develop a systematic way of extracting biologically relevant information about macromolecular dynamics from diffuse scattering patterns.

James Holton (Lawrence Berkeley National Laboratory) presented simulations of total scattering patterns considering a variety of models of crystal variation. He noted that diffuse scattering occurs not just between but also underneath the Bragg peaks, which might potentially lead to systematic errors in Bragg peak intensity measurements.

Henry Chapman (Deutsches Elektronen-Synchrotron) presented methods for processing and analysis of serial crystallography diffraction images for diffuse scattering using an X-ray free-electron laser. He argued for the interpretation that diffuse scattering (which also goes by the name “continuous diffraction” in the coherent diffractive imaging community) is proportional to the Fourier transform of rigidly moving molecular units, and the use of this interpretation for resolution extension and phasing of charge density maps.

Sarah Perry (University of Massachusetts, Amherst) described graphene-based microfluidics as a potential fixed-target mounting strategy capable of providing both sample stability and the ultra-low background necessary for diffuse scattering experiments.

Mitchell Miller (Rice University) described ongoing efforts in George Phillips's lab to collect and process diffuse scattering. He stressed especially the benefit of limiting background scattering from air, crystal mounts, and beamline components over trying to remove these sources computationally. Current pixel-array detectors, with their very low "dark-current" noise and near perfect photon-counting, facilitate simultaneous Bragg and diffuse-scattering measurements, and multi-pass collection strategies can fill in the gaps between detector modules. Currently, the group is using the program *XCAVATE* (Esterman et al, 1998) to map the scattering intensity from detector images into a Cartesian grid in reciprocal space, with further analysis and symmetry averaging in *MATLAB*.

Alexander Wolff (University of California, San Francisco) described his efforts in James Fraser's lab, in collaboration with Michael Wall, to develop an automated pipeline for the integration and analysis of diffuse scattering data. The goal of this work is to simplify data analysis, while retaining a modular workflow so that users can incorporate custom processing steps. An open-source, modular library for diffuse-data reduction will enhance transparency between labs and accelerate our ability to test disorder models across a variety of macromolecules.

Nicholas Sauter (Lawrence Berkeley National Laboratory) considered the tradeoffs made in serial crystallography when choosing between still-shot diffraction (used at XFEL sources) and traditional, even finely sliced, rotation shots (normally used at synchrotron sources). With XFEL stills, one can avoid radiation damage, perform time-domain work, and observe the system under room-temperature physiological conditions. However, it is more difficult to refine the experimental geometry and integrate the Bragg spot intensities. Moreover, the spots are always partially integrated measurements, and the conversion to the equivalent structure factor relies on variables that are imperfectly known. Computational approaches have not been settled yet, and it requires many more stills than rotation shots to achieve the same map accuracy.

Ariana Peck (Stanford University) presented analysis performed with Frédéric Poitevin and TJ Lane that surveyed models of disorder previously used to interpret diffuse scattering, and compared their ability to reproduce three experimental maps. Models of intramolecular liquid-like motions and rigid-body rotations showed modest correlation with the experimental maps but were unable to reproduce experimental speckles indicating a correlated disorder spanning multiple unit cells. These results suggest a need for models of disorder that account for correlations coupled across a range of length scales.

David Case (Rutgers University) discussed the use of molecular dynamics simulations of crystals, and their applications to the analysis of diffuse scattering. It is now feasible to carry out simulations of multiple unit cells of small globular proteins on time scales of ca. 5 microseconds in a few weeks of computer time using GPU acceleration. The average structures are typically 0.3 to 0.5 Å away from the refined x-ray structures, and atomic fluctuations are close to those extracted from

refinement of atomic displacement parameters using experimental data. Diffuse scattering requires averaging over about 10,000 snapshots for good convergence, but the general behavior can be obtained with less sampling. Comparison of calculated and observed diffuse scattering for tetragonal lysozyme (using experimental data from Ando and Meisburger described above) showed agreement to a level better than seen in previous comparisons. Solvent molecules contribute in an important way to what is observed.

Henry van den Bedem's (SLAC National Accelerator Laboratory) contribution focused on methods for modeling the protein conformational ensemble from Bragg spots. Traditionally, a crystal structure is presented as a single, unique conformer with isotropic or anisotropic atomic-displacement parameters, or B-factors. By contrast, the multi-conformer modeling algorithm qFit introduces up to four main-/side-chain conformations for each residue as needed to collectively, locally explain the experimental data. In a qFit crystal structure, B factors represent harmonic deviations, whereas conformers represent anharmonic deviations. In combination with room temperature crystallography, qFit has uncovered 'hidden' conformers, revealed molecular mechanisms, and established a relation between fast dynamics in crystals and in solution. Henry presented recent insights into catalytic motions of isocyanide hydratase (ICH) obtained from serial crystallography and analysis of qFit models, in collaboration with Mike Wall and Mark Wilson (University of Nebraska, Lincoln).

James Fraser (University of California, San Francisco) discussed the tantalizing possibilities of exploiting diffuse scattering for improved modeling of biological macromolecules. Significant progress has been made on the three main challenges (measurement, modeling, and validation) identified in the 2013 Diffuse Scattering Workshop (Wall et al, 2014), but there are several obstacles, discussed at this meeting, that remain to be overcome. Although measurement of diffuse-scattering data is now easier because of more sensitive detectors, it remains unclear whether high quality diffuse data can be measured simultaneously with Bragg data or if specialized protocols are required. Other remaining needs include standardized ways of processing, storing (e.g. mtz or hkl formats), and representing data to facilitate comparisons. Diffuse data continue to be well modeled by simple liquid-like-motions models that do not provide the sort of atomistic detail that would be useful to contribute to our understanding of biochemical mechanism. Although more sophisticated ensemble and TLS models are relatively commonly applied to increase the fit with Bragg data, they agree poorly with the diffuse data. Normal-modes-type models may represent a path forward and can be validated against careful MD simulations of crystalline proteins. Finally, he discussed the validation issue, with particular attention paid to metrics for agreement between model and data (including "exact" data calculated from MD) and new features present in the resolution-extension continuous-diffraction approaches developed by Chapman and colleagues.

The presentations were followed by a lively discussion about roadblocks in measurement, modeling, and biological interpretation of the data. Here is a summary of that discussion.

Several common themes emerged in the talks and the discussions. One theme is the importance of moving beyond analysis of individual aspects of the data and developing more comprehensive models that can simultaneously explain the large-scale diffuse features, related to variations correlated within the unit cell, and small-scale features, related to variations correlated on longer length scales. As with the Bragg data, one will need to develop a complete picture of the diffuse-scattering data to extract more detailed mechanistic insights.

A second theme was the importance of using controls and reversible perturbations, with the possible use of anomalous scatterers, to extract biological meaning from the diffuse data. A third theme is the potential of diffuse scattering to enable validation of the types of variations that are actually present in the crystal, but which cannot be distinguished using Bragg analysis alone. Another theme is the increasing success of MD simulations in capturing the diffuse scattering, and the potential for using diffuse scattering as a routine means of validating crystalline MD. A final theme is the potential for using normal modes models as a possible path forward for developing mechanistic insights that can be validated against diffuse scattering data and compared to MD simulations.

Going forward, it is critical to continue to validate models not only using subjective comparison of the results (diffraction, Patterson function) between model and data, as was common in earlier studies, but also with direct numerical refinement that depends on maximum likelihood or some other minimization of differences, as is increasingly done in modern studies. It also will be important when publishing to release all the raw data, along with the workflow: source code, compiled code, and command scripts that are needed to reproduce the published analysis. Data archives exist to help us: Sbgrid.org, proteindiffraction.org, and cxidb.org.

Overall, much as been accomplished since a workshop on diffuse scattering at Lawrence Berkeley National Laboratory in 2013 (Wall et al, 2014), with notable advances in data collection, data processing, and molecular dynamics simulations. The number of active researchers in the field has grown substantially. More progress is needed, in both accurate data collection and modeling, to increase the overall correlations between models and experiment, and to improve the ability to discriminate among alternatives. In addition, it is important for the field to clarify what new information we can learn about proteins with diffuse scattering. More applications to biomedically important systems are needed to deepen the connection between the experiments and modeling, thereby creating actionable information for biochemistry and biology.

D.L. Caspar, J. Clarage, D.M. Salunke, M. Clarage. Liquid-like movements in crystalline insulin. Nature, 332 (1988), pp. 659-662

Estermann, M. A. and W. Steurer (1998). “Diffuse scattering data acquisition techniques.” *Phase Transitions* **67**(1): 165-195.

Steve P. Meisburger, William C. Thomas, Maxwell B. Watkins, and Nozomi Ando. X-ray Scattering Studies of Protein Structural Dynamics. *Chemical Reviews* (2017) In Press (DOI: 10.1021/acs.chemrev.6b00790)

Y. Polikanov, P. Moore. Acta Cryst. (2015). D71, 2021–2031

Wall ME, Adams PD, Fraser JS, Sauter NK. Structure. 2014 Feb 4;22(2):182-4.