Data Descriptor: Free-electron laser data for multiple-particle fluctuation scattering analysis

Kanupriya Pande1,2, Jeffrey J. Donatelli1,3, Erik Malmerberg1,2,4, Lutz Foucar5,6, Billy K. Poon2, Markus Sutter2, Sabine Botha5,7, Shibom Basu8,9, R. Bruce Doak5,8, Katerina Dörner8,10, Sascha W. Epp6,11,12, Lars Englert13,14, Raimund Fromme8, Elisabeth Hartmann12, Robert Hartmann15, Guenter Hauser13, Johan Hattne2,16, Ahmad Hosseinizadeh17, Stephan Kassemeyer13, Lukas Lomb2, Sebastian F. Carron Montero18,19, Andreas Menzel20, Daniel Rolles5,6,21, Artem Rudenko6,11,21, Marvin M. Seibert18, Raymond George Sierra18, Peter Schwander17, Abbas Ourmazd15, Petra Fromme11, Nicholas K. Sauer2, Michael Bogan18,22, John Bozek18,23, Christoph Bostedt18,24,25, Ilme Schlichting5,6, Cheryl A. Kerfeld2,26 & Petrus H. Zwart1,2

Fluctuation X-ray scattering (FXS) is an emerging experimental technique in which solution scattering data are collected using X-ray exposures below rotational diffusion times, resulting in angularly anisotropic X-ray snapshots that provide several orders of magnitude more information than traditional solution scattering data. Such experiments can be performed using the ultrashort X-ray pulses provided by a free-electron laser source, allowing one to collect a large number of diffraction patterns in a relatively short time. Here, we describe a test data set for FXS, obtained at the Linac Coherent Light Source, consisting of close to 100 000 multi-particle diffraction patterns originating from approximately 50 to 200 Paramecium Bursaria Chlorella virus particles per snapshot. In addition to the raw data, a selection of high-quality pre-processed diffraction patterns and a reference SAXS profile are provided.

1Center for Advanced Mathematics in Energy Research Applications, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 2Molecular Biophysics and Integrated Bio-imaging, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 3Computational Research Division, Dept. of Mathematics, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 4Hit Discovery, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden. 5Max-Planck-Institut für medizinische Forschung, Jahnstr. 29, 69120 Heidelberg, Germany. 6Max Planck Advanced Study Group, Center for Free Electron Laser Science (CFEL), Notkestrasse 85, 22607 Hamburg, Germany. 7University of Hamburg, Hamburg Germany. 8Arizona State University, Tempe, AZ, USA. 9Macromolecular Crystallography Group, Paul Scherrer Institute, 5232 Villigen – PSI, Switzerland. 10European XFEL GmbH, Schenefeld, Germany. 11Max-Planck-Institut für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany. 12Max Planck Institute for the Structure and Dynamics of Matter, Center for Free Electron Laser Science, Hamburg, Germany. 13Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse, 85741 Garching, Germany. 14Carl von Ossietzky Universität Oldenburg, Department of Physics, Oldenburg, Germany. 15PNSensor GmbH, Otto-Hahn-Ring 6, 81739 München, Germany. 16University of California, Los Angeles, Los Angeles, CA, USA. 17Department of Physics, University of Wisconsin-Milwaukee, 3135N. Maryland Ave, Milwaukee, WI 53211, USA. 18Linac Coherent Light Source, SLAC National Accelerator Laboratory, Stanford, CA, USA. 19Department of Physics, California Lutheran University, Thousand Oaks, CA, USA. 20Laboratory for Macromolecules and Bioimaging, Paul Scherrer Institute, 5232 Villigen – PSI, Switzerland. 21James R Macdonald Laboratory, Kansas State University, Manhattan, KS, USA. 22Traction on Demand, Burnaby, BC, Canada. 23Synchrotron SOLEIL, L’Orme des Merisiers, Saint-Aubin, BP 48, F-91922 GIF-sur-Yvette Cedex, France. 24Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA. 25Atomic, Molecular and Optical Physics, Advanced Photon Source, Argonne National Laboratory, Argonne, IL, USA. 26DOE Plant Research Laboratory, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. Correspondence and requests for materials should be addressed to P.H.Z. (email: phzwart@lbl.gov)
background & summary

Fluctuation X-ray scattering (FXS) studies extend traditional small angle X-ray scattering (SAXS) methods by using X-ray snapshot with exposure times so short that the ensemble of illuminated particles can be well approximated as frozen in time and space. The resulting scattering patterns are no longer angularly isotropic, but instead exhibit small intensity fluctuations around the mean SAXS intensity. Angular correlations of these intensity fluctuations can be directly related to the underlying molecular structure of the sample, providing much more information than traditional 1D SAXS curves. The correlation function \( C_2(q, q', \Delta \phi) \) is defined as

\[
C_2(q, q', \Delta \phi) = \frac{1}{2\pi N} \sum_{l=1}^{N} f_l(q, q') \int_0^{2\pi} I_l(q, \phi) I_l(q', \phi + \Delta \phi) d\phi,
\]

where \( N \) is the total number of diffraction patterns, \( q \) and \( q' \) are the magnitudes of the scattering vectors (in inverse resolution), and \( \phi \) and \( \psi + \Delta \phi \) are the corresponding angular coordinates describing the intensity \( I_l(q, \phi) \) of the \( j \)th scattering pattern recorded on the detector. The correlation function \( C_2(q, q', \Delta \phi) \) can be written as a Legendre series:

\[
C_2(q, q', \Delta \phi) = \sum_{l=0,2,\ldots}^{\infty} B_l(q, q') P_l(\cos \theta_q \cos \theta_{q'} + \sin \theta_q \sin \theta_{q'} \cos \Delta \phi),
\]

where \( P_l \) is the Legendre polynomial of order \( l \) and \( \theta_q = \arccos(q\lambda/4\pi) \), \( \lambda \) is the wavelength of the incident X-rays, and \( k_l \) is a scale factor equal to the number of particles in the beam for \( l > 0 \), and equal to its square for \( l = 0 \). The expansion coefficients \( B_l(q, q') \) are in turn related to the spherical harmonic expansion coefficients \( I_{lm}(q) \) of the 3D intensity scattering volume \( I(q) \) of the scattering particle, where \( q = (q, \theta, \phi) \):

\[
B_l(q, q') = \sum_{m=-1}^{m=1} I_{lm}^*(q) I_{lm}(q'),
\]

where, in polar coordinates,

\[
I(q, \theta, \phi) = \int I_{lm}(q) Y_{lm}(\theta, \phi).
\]

The intensity function is equal to the square of the modulus of the Fourier transform of the real-space object \( \rho(r) \) under investigation:

\[
I(q) = |\mathcal{F}[\rho(r)]|^2,
\]

where \( \mathcal{F} \) denotes the Fourier transform.

Prior work in fluctuation scattering from biological samples has demonstrated that high-quality correlation data can be obtained from single-particle diffraction data and can be used for \textit{ab initio} structure determination using the multi-tiered iterative phasing algorithm. Although previous work has shown that the signal to noise ratio (SNR) of such data is independent of the number of particles per shot when the particles are in a vacuum, the relationship between number of particles and SNR in the presence of large buffer and detector backgrounds has not yet been studied. In this communication, we describe unprocessed, experimental multi-particle scattering data from which an FXS correlation data set can be derived. The data, obtained at the Atomic, Molecular and Optical (AMO) instrument at the Linac Coherent Light Source, consist of close to 60,000 high quality scattering images of the \textit{Paramecium bursaria} Chlorella virus 1 (PBCV-1, \(~190 \text{nm in diameter}) and 30,000 scattering images of the sample buffer. The images presented here provide the community with experimental data on which algorithms for processing fluctuation scattering data and structure solution can be tested. The data are deposited at the CXIDB in the form of hdf5 and xtc files.
Methods
Sample Preparation, Sample Delivery and Data collection
A batch of *Paramecium Bursaria Chlorella* virus 1 sample was prepared as described previously\(^{10}\). Here we used 1% triton instead of Nonidet and centrifuged the virus sample at 20,000 rpm in an ultracentrifuge. The pure virus sample was dialyzed against 50 mM 4-methylmorpholine\(^{11}\). The quality of the FXS data was gauged by comparing the derived SAXS data against a reference Small Angle Scattering curve obtained at the cSAXS beamline at the Swiss Light Source, at an energy of 11 keV. The sample used for FXS data collection was diluted with buffer to a concentration of approximately \(5 \times 10^{11}\) particles per ml.

The multi-particle FXS scattering data were collected at the Atomic, Molecular and Optical (AMO) instrument at the LCLS\(^{6,7}\). The experiment was performed in the CFEL-ASG Multi-Purpose chamber (CAMP)\(^{12}\). The PBCV-1 solution described above was injected into the XFEL interaction region as a microjet of approximately 5 \(\mu\)m diameter, using a gas dynamic virtual nozzle (GDVN) injector\(^{13,14}\) at a flow rate of \(~20\ \mu\text{l/min}\) (Fig. 1). The diffraction data were collected in the water window, using a photon energy of 514 eV, an electron bunch length (pulse length) of 100 fs and a repetition rate of 120 Hz. The average number of photon per pulse was \(10^{12}\). The focus size was approximately 25 \(\mu\text{m}^2\) (FWHM). Diffraction patterns were collected on two pairs of p-n junction charge-couple device (pnCCD) detectors\(^{15}\) read out at 120 Hz. The front and back panels of the detector consisted of two pairs of 1024 \(\times\) 512 arrays of square 75 \(\mu\text{m} \times 75 \mu\text{m}\) pixels. The front panels were placed 224 mm from the interaction region, separated by a horizontal gap of 23 mm. The back detector was placed 741 mm from the sample/XFEL interaction zone, with a horizontal gap of 1.73 mm (Table 1). The maximum resolution achievable under these conditions is 14.3 nm (8.9 nm) at the edge (corner) of the front and 46.8 nm (32.7 nm) at the edge (corner) of the back detector. The X-ray scattering patterns and associated metadata were stored as xtc files. These diffraction images were pre-processed using the CFEL-ASG Software Suite (CASS)\(^{16}\). The images were corrected for dark current; pixels systematically producing outlying intensity values were flagged. The resulting data was cast in larger arrays to include the detector gaps. The detector halves were placed roughly symmetrically around the X-ray beam. The back detector, with gaps included, is contained in an array of 1024 \(\times\) 1047, with the mean beam center located at (506, 526). The front detector (with gaps included) is contained in an 1331 by 1031 array, with a mean beam center of (657, 552). The back detector was at gain mode 1, corresponding to 1250 ADU per 1keV photon. The gain mode of the front pnCCD was 4, corresponding to 78 ADU per 1keV photon. The resulting arrays are stored in an hdf5 file and are deposited in the CXIDB with accession number 79 (Data Citation 1). A Globus end-point is available for high-speed data-transfer.

From the concentration, the jet diameter and the focus size the particle count per exposure is estimated to 60 particles per shot. Given that the focussed X-ray beam has extended tails beyond the focal limit, there is an uncertainty that likely places the true particle count somewhat higher. A conservative estimate of the bounds of the particle count of 50 to 200 is proposed.

Pattern Selection
Due to liquid jet and X-ray beam instabilities, not all scattering patterns collected are of sufficient quality for correlation analyses. The set of diffractions patterns of the sample collected contain, besides multi-

---

**Figure 1.** A schematic overview of a Fluctuation X-ray Scattering. (a) An FXS experiment at an XFEL can be performed by intersecting a liquid jet containing the sample particles with the XFEL pulse and recording the diffraction pattern on a multi-panel detector. The data generated by the setup described in this manuscript resulted in a maximum resolution of 14.3 nm (8.9 nm) at the edge (corner) of the front and 46.8 nm (32.7 nm) at the edge (corner) of the back detector. (b) Calculation of the intensity correlations is performed according to equation 1. The scattering patterns shown above are experimental data discussed in this report.
particle hits, a set of blanks, where no virus was intersected by the XFEL beam, as well as images characterized by a very high total scattered intensity and extensive intensity streaks where the X-rays hit the edge of the jet or part of the liquid-jet nozzle, Fig. 2. Similar observations were made for the buffer run. A selection of patterns was made on the basis of the total integrated intensity on the back panel. A histogram analysis of the integrated intensity reveals a bimodal distribution, with high quality patterns occurring around the most-populated mode of the distribution.

### Code Availability
CASS is publicly available on github (https://gitlab.gwdg.de/p.lfoucar/cass). The reading of xtc files is supported by the psana libraries distributed by the LCLS (https://stanford.io/2lhTEwT).

### Data Records
Four individual datasets have been deposited on the CXIDB website (Data Citation 1). The deposited data consists of the raw xtc file of the experimental PBCV-1 and buffer scattering data. Selected patterns, pre-processed with CASS, involving dark-current subtraction and common mode corrections, of PBCV-1 and buffer have been deposited also in separate hdf5 files for the back and front pnCCD detectors. The xtc files contain close to 100 000 scattering patterns, whereas the selected pre-processed files contain close to 60 000 patterns for PBCV-1 and 30 000 patterns for the buffer. A reference buffer subtracted SAXS data set from PBCV-1 at the same concentration, collected at 11 keV using the CSAXS beamline at the Swiss Light Source has been deposited as well. The data records are summarized in Table 2.

| Parameter                        | Value                        |
|----------------------------------|------------------------------|
| X-ray energy                     | 514 eV / 2.4 nm              |
| X-ray pulse length               | 100 fs                       |
| X-ray repetition rate            | 120 Hz                       |
| Front panel gap                  | 23 mm                        |
| Front panel distance             | 224 mm                       |
| Back panel gap                   | 1.73 mm                      |
| Back panel distance              | 741 mm                       |

Table 1. Data collection parameters.

Figure 2. Pattern selection. (a). The total integrated intensities from the back panel form a bimodal distribution, for both the buffer data and the sample. In the buffer run, decent buffer-only shots (b1) dominate the low end, whereas streak dominated images (b2) and shots containing residual virus particles (b3) are found with higher integrated intensities. For the sample run, blank shots (c1) dominate the low end, and streak-dominated images (c3) are found in shots with integrated intensities residing in the extended tail (>200 AU) of the distribution. Diffraction patterns of the sample falling in the major peak (c2), with integrated intensities between 50 and 200 AU can be used to obtain experimental intensity correlations. All diffraction patterns are shown with a logarithmic colormap.
Technical Validation
The quality of the data can be assessed by the mean intensity as a function of resolution, Fig. 3. The SAXS curve was obtained by angular integration of the images after masking out the strong jet scattering streaks seen in the diffraction patterns, Fig. 2c. A mask covering the jet streak included in the deposited hdf5 files.

The experimental SAXS data were fitted in the low-q region (up to 0.015 Å\(^{-1}\)) with the theoretical scattering curve of a hard sphere with a diameter of 174 nm. Given that the diameter of an icosahedron is 17% larger than the sphere that touches the midpoint of each vertex\(^1\), the hard sphere model derived here would correspond to a maximum particle dimension of little over 200 nm, consistent with the available model\(^8\). The analyses of the reference data collected at the Swiss Light Source can be modelled (at low q) with a hard sphere with a radius of 168 nm, corresponding to an icosahedron with diameter of 197 nm. The difference in estimated size between the reference data and the curve obtained from AMO can be ascribed to changes in relative contrast at lower X-ray energies, the effects of radiation damage on the sample at synchrotron sources, variations in sample preparations or concentration effects on the shape of the low q data.

Table 2. Data Records.

| Data Identifier | Contents | Comments |
|-----------------|----------|----------|
| CXIDB-79: Raw XTC Files | PBCV-1 and buffer scattering patterns in xtc format. | |
| CXIDB-79: PBCV Diffraction Patterns | Selected scattering images of PBCV-1 in solution, common mode and geometry corrected by CASS. Contains ~60000 patterns. | The hdf5 path '/data/data' contains the data. The '/data/mask' field contains a mask covering the overloaded jet streak. '/data/timestamp_str' contains timestamps of the images. |
| CXIDB-79: Buffer Diffraction Patterns | Selected scattering images of sample buffer, common mode and geometry corrected by CASS. Contains 20000 patterns. | The hdf5 path '/data/data' contains the data. The '/data/mask' field contains a mask covering the overloaded jet streak. '/data/timestamp_str' contains timestamps of the images. |
| CXIDB-79: SAXS Data | A reference SAXS curve of PBCV-1 collected at the cSAXS beamline at the Swiss Light Source, collected at an energy of 11keV. A SAXS curve in plain ASCII. Each line contains q (in inverse Ångström) and the corresponding mean intensity. | |

Figure 3. The experimental SAXS data derived from the selected snapshots on the back detector display a characteristic oscillatory behaviour consistent with spherical-like particles. This SAXS curve is close to a reference curve obtained from PBCV-1 at the Swiss Light Source’s CSAXS beamline at 11000 eV. Minor discrepancies between the soft X-ray and hard X-ray curve are likely due to contrast differences in the water window. A hard sphere model of the AMO data and the SLS data suggest a diameter of 174 nm and 168 nm, which would correspond to an icosahedron with diameter of 204 and 197 nm respectively.
References
1. Kam, Z. Determination of Macromolecular Structure in Solution by Spatial Correlation of Scattering Fluctuations. Macromolecules 10, 927–934 (1977).
2. Saldin, D. K. et al. New light on disordered ensembles: ab initio structure determination of one particle from scattering fluctuations of many copies. Phys Rev Lett 106, 115501 (2011).
3. Malmerberg, E., Kerfeld, C. A. & Zwart, P. H. Operational properties of fluctuation X-ray scattering data. IUCrJ 2, 309–316 (2015).
4. Kurtz, R. P. et al. Correlations in Scattered X-Ray Laser Pulses Reveal Nanoscale Structural Features of Viruses. Phys Rev Lett 119, 158102 (2017).
5. Kirian, R. A., Schmidt, K. E., Wang, X., Doak, R. B. & Spence, J. C. Signal, noise, and resolution in correlated fluctuations from snapshot small-angle x-ray scattering. Phys Rev E Stat Nonlin Soft Matter Phys 84, 011921 (2011).
6. Bozek, J. D. AMO instrumentation for the LCLS X-ray FEL. The European Physical Journal Special Topics 169, 129–132 (2009).
7. Ferguson, K. R. et al. The Atomic, Molecular and Optical Science instrument at the Linac Coherent Light Source. Journal of synchrotron radiation 22, 492–497 (2015).
8. Zhang, X. et al. Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. Proc Natl Acad Sci USA 108, 14837–14842 (2011).
9. Maia, F. R. The Coherent X-ray Imaging Data Bank. Nat Methods 9, 854–855 (2012).
10. Van Etten, J. L., Burbank, D. E., Xia, Y. & Meints, R. H. Growth cycle of a virus, PBCV-1, that infects Chlorella-like algae. Virology 126, 117–125 (1983).
11. Kassevemeyrey, S. et al. Femtosecond free-electron laser x-ray diffraction data sets for algorithm development. Opt Express 20, 4149–4158 (2012).
12. Strüder, L. et al. Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multi-purpose chamber for experiments at 4th generation light sources. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 614, 483–496 (2010).
13. DePonte, D. P. et al. Gas dynamic virtual nozzle for generation of microscopic droplet streams. Journal of Physics D - Applied Physics 41 (2008).
14. Weierstall, U., Spence, J. C. H. & Doak, R. B. Injector for scattering measurements on fully solvated biospecies. The Review of scientific instruments 83 (2012).
15. Strüder, L. High-resolution imaging X-ray spectrometers. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 454, 73–113 (2000).
16. Foucar, L. et al. CASS—CFEL-ASG software suite. Computer Physics Communications 183, 2207–2213 (2012).
17. Sloane N. J. A. editor The On-Line Encyclopedia of Integer Sequences, entries A019881 & A019863 https://oeis.org (2018).

Data Citations
1. Pande, K. Coherent X-ray Imaging Data Bank http://dx.doi.org/10.11577/1437269 (2018).

Acknowledgements
This research was supported by the Max Planck society and in part, by the Advanced Scientific Computing Research and the Basic Energy Sciences programs, which are supported by the Office of Science of the US Department of Energy under Contract DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract DE-AC02-05CH11231. Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. This material is also based upon work partly supported by the National Science Foundation under Grant Nos 1240590 and 1733552. The research conducted at UWM was supported by the US Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0002164 (algorithm design and development), and by the US National Science Foundation under awards STC 1231306 (numerical trial models and data analysis) and 1551489 (underlying analytical models). Further support originates from the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Awards R01GM109019 and GM117126. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Author Contributions
C.A.K., P.F., A.O., N.K.S., P.H.Z., I.S., C.B. arranged beamtime; P.H.Z., E.M., I.S. conceived the experiment. K.P., J.J.D., E.L.M., L.F., B.K.P., I.S. and P.H.Z. analysed the data. The manuscript was written with input from K.P., J.J.D., E.L.M., L.F., I.S., A.O., C.A.K. and P.H.Z. The experiment was performed with input from E.L.M., L.F., B.K.P., M.S., S.Bo., S.Ba., R.B.D., K.D., S.W.E., L.E., R.F., E.H., R.H., G.H., J.H., A.H., S.K., L.L., S.F.C.M., D.R., A.R., M.M.S., R.G.S., P.S., A.O., P.F., N.K.S., M.B., J.B., C.B., I.S., C.A.K. and P.H.Z. The reference SAXS data was collected by A.M.

Additional Information
Competing interests: The authors declare no competing interests.

How to cite this article: Pande, K. et al. Free-electron laser data for multiple-particle fluctuation scattering analysis. Sci. Data. 5:180201 doi: 10.1038/sdata.2018.201 (2018).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
