Coherent diffraction of single Rice Dwarf Virus particles using hard X-rays at the Linac Coherent Light Source

110 characters maximum, including spaces

Instructions from Scientific Data

Scope Guidelines

Data Descriptors submitted to *Scientific Data* should provide detailed descriptions of valuable research datasets, including the methods used to collect the data and technical analyses supporting the quality of the measurements. Data Descriptors focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses. Relevant datasets must be deposited in an appropriate public repository prior to Data Descriptor submission, and their completeness will be considered during editorial evaluation and peer review. The data must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (excepting reasonable controls related to human privacy issues or public safety).

Anna Munke[UU], Jakob Andreasson[ELI,UU], Andrew Aquila[SLAC], Salah Awel[CFEL], Kartik Ayyer[CFEL], Anton Barty[CFEL], Peter Berntsen[ARC], Johan Bielecki[UU], Sebastien Boutet[SLAC], Maximilian Bucher[SLAC,ANL,TUB], Benedikt J. Daurer[UU], Hasan DeMirci[PULSE,SLAC], Veit Elser[LASSP], Akifumi Higashiura[OSAKA], Brenda G. Hogue[CIDV, CASD, SOLS], Yoonhee Kim[GIST], Hemanth Kumar[UU], Ti-Yen Lan[LASSP], Daniel S. D. Larsson[UU], Haiguang Liu[CSRC], Max F. Hantke[UU], N. Duane Loh[CBIS], Filipe R. N. C. Maia[UU], Atsushi Nakagawa[OSAKA], Daewoong Nam[POSTECH], Garrett Nelson[ASU], Carl Nettelblad[UU3,UU], Max Rose[DESY], M. Marvin Seibert[UU], Jonas A. Sellberg[UU], Raymond G. Sierra[PULSE,SLAC], Changyong Song[POSTECH], Martin Svenda[UU], Kenta Okamoto[UU], Nicusor Timneanu[UU,UU2], Ivan A. Vartanyants[DESY,MEPhI], Daniel Westphal[UU], Garth J Williams[BNL], Paulraj Lourdu Xavier[CFEL,MPSD,UH], Chunhong Yoon[SLAC], James Zook[CASD]

Affiliations

ANL: Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439, USA

ARC: Australian Research Council Centre of Excellence in Advanced Molecular Imaging, La Trobe Institute for Molecular Science, La Trobe University, Melbourne 3086, Australia

ASU: Arizona State University, Department of Physics, Tempe, AZ 85287, USA

ARC: Australian Research Council Centre of Excellence in Advanced Molecular Imaging, La Trobe Institute for Molecular Science, La Trobe University, Melbourne 3086, Australia

BNL: Brookhaven National Laboratory, NSLS-II, Upton, NY 11973, USA

CASD: Center for Applied Structural Discovery, Biodesign Institute at Arizona State University, 85287, Tempe, USA

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under Contract No. DE-AC02-76SF00515.

CBIS: Centre for Bio-imaging Sciences, National University of Singapore, 14 Science Drive 4, BLK S1A, 117543, Singapore.

CFEL: Center for Free Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany

CIDV: Center for Infectious Diseases and Vaccinology, Biodesign Institute at Arizona State University, 85287, Tempe, USA

CSRC: Beijing Computational Science Research Center, 8 W Dongbeiwang Rd, Haidian, Beijing, China, 100193

DESY: Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22607 Hamburg, Germany

GIST: School of Materials Science and Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, Korea

LASSP: Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853, USA

MEPhI: National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, 115409 Moscow, Russia

MPSD: Max-Planck Institute for the Structure and Dynamics of Matter, CFEL, 22607, Hamburg, Germany

OSAKA: Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan;

POSTECH: Department of Physics, Pohang University of Science and Technology, Pohang 37673, Korea

PULSE: Stanford PULSE Institute, 2575 Sand Hill Road, Menlo Park, California 94025, USA

SLAC: SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA

SOLS: School of Life Sciences, Arizona State University, 85287, Tempe, USA

TUB: Institut für Optik und Atomare Physik, Technische Universität Berlin, Hardenbergstraße 36, 10623 Berlin, Germany

UU: Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-75124 Uppsala, Sweden

UU2: Department of Physics and Astronomy, Uppsala University, Lägerhyddsvägen 1 (Box 516), SE-75120 Uppsala, Sweden

UU3: Department of Information Technology, Science for Life Laboratory, Uppsala University, Lägerhyddsvägen 2 (Box 337), SE-75105 Uppsala, Sweden

UH: Department of Physics, University of Hamburg, Hamburg, Germany

ELI: Institute of Physics ASCR, v.v.i. (FZU), ELI-Beamlines Project, 182 21 Prague, Czech Republic

corresponding author: Andrew Aquila (aquila@slac.stanford.edu)

Abstract

170 words maximum

Single particle diffractive imaging data from Rice Dwarf Virus (RDV) was measured using the Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS). RDV was chosen as it is a well characterised model systems useful for proof-of-principle experiments, system optimization and algorithm development. RDV, an icosahedral virus of about 70 nm in diameter, was aerosolised and injected into the approximately 0.1 µm diameter focused hard X-ray beam at the CXI instrument of LCLS. Diffraction patterns from RDV were recorded to a resolution of 3.2 Ångström. The diffraction data is available through the Coherent X-ray Imaging Data Bank (CXIDB) as a resource for algorithm development, the contents of which are described here.

Background & Summary

700 words maximum

For several decades, X-ray crystallography has been the dominant technique to solve the three dimensional structure of biological macromolecules at atomic resolution. Structures of proteins, protein complexes and the machinery of entire biological reaction pathways have been elucidated, leading to numerous breakthroughs in our understanding of molecular architecture and function. However, not every protein complex crystallizes, a necessary condition for investigation using these methods. Radiation damage additionally limits the resolution for non-crystalline objects, which for high exposures leads to the determination of structures representative of a photodamaged state (Henderson 1995, Neutze, 2000, Howells 2009). The ultrashort and extremely bright pulses from X-ray free electron lasers (XFELs) were predicted to outrun the radiation damage processes and allow the recording of diffraction data from samples prior to any significant motion of the nuclei occurring (Neutze, 2000). This has been experimentally demonstrated at nanometer resolution in isolated objects (Chapman 2006, Seibert 2011) and Ångström resolution in micro/nanocrystals (Chapman 2011, Boutet 2012). Yet the goal of near atomic resolution single particle imaging remains elusive.

The single particle imaging (SPI) initiative is a large collaborative team of researchers from several institutions formed to identify and solve the challenges required for reaching high resolution imaging. The aim of the SPI initiative, as laid out in the roadmap, (Aquila, 2015) is to establish a community-wide approach to take up the scientific and technical challenges of single-molecule imaging with X-rays. In addition to developing solution to the technical challenges, a critical part of this project was selecting a well characterised model system

needed for demonstration experiments. After considering homogeneity, uniform size distribution, particle concentration, having a known structure, and the ability to be aerosolised for injection into the XFEL beam, Rice Dwarf Virus (RDV) was selected as the primary sample for the first SPI experiment (see methods).

RDV is an icosahedral virus of about 70 nm in diameter, and is the causative agent of rice dwarf disease. Which, creates severe economic damage in China, Japan and other Asian countries due to speck formation and its destructive effects on plant growth for rice, wheat, barley, and other gramineae plants. Leafhopper insects are the primary host in which the virus particle replicates and from which they are then transmitted to the leaves. The icosahedral virus particles consist of two shells, an inner and an outer capsid, enclosing a double stranded RNA genome. The genome encodes 12 products, seven of which are considered structural proteins. A thin layer of P3 capsid proteins (Kano 1990) make up the inner capsid and three proteins of mainly P8 (Omura 1989), but also P2 (Yan 1996) and P9 (Zhong 2003), form the outer capsid. Found in the core, together with the genome, are P1 (putative RNA polymerase (Suzuki 1992)), P5 (putative guanylyltransferase (Suzuki 1996)) and P7 (a nonspecific nucleic acid binding protein (Ueda 1997)). A three dimensional structure of the capsid was solved by X-ray crystallography at 3.5 Å resolution (PDB 1UF2) (Nagakawa 2003).

RDV was aerosolized and delivered into the hard X-ray Coherent X-ray Imaging (CXI) nanofocus instrument of the LCLS (Boutet 2010, Liang 2015) using an aerodynamic lens injector (Bogan 2008). Diffraction patterns were recorded at a rate of 120 Hz using two Cornell-SLAC Pixel Array Detectors (CSPAD), a large 2Mpix detector located close to the sample for wide angle scattering and another smaller 0.25Mpix “2x2” detector located further downstream to detect small angle scattering (Hart 2012, Herrmann 2014) (pictorially shown in center panel of figure 1). The photon energy was 7 keV, the pulse duration was <50 fs, and the average pulse energy immediately after the undulator was 4 mJ. (see *Methods* for a detailed description)

In the data deposited in the Coherent X-ray Imaging Data Bank (CXIDB) we record clear diffraction from single RDV virus on the back detector, as well as elevated scattering on the front detector for single virus hits that are identified based on the data on the back detector, indicating that measurable photons are recorded from the sample up to 3.2 Ångström resolution.

Methods

Sample preparation

Nymphs of *Nephotettix nigropictus* were fed on diseased rice plants. The purification procedure of the RDV O strain (Kimura 1987) followed the procedure of Omura et al. (Omura 1982) with modified sucrose concentrations as follows. A meat chopper was used to grind the infected rice leaves and the resultant slurry was treated with CCl_4 and subjected to repeated precipitations and consecutive density gradient centrifugations in 40% to 60% and 40% to 70% sucrose. The pellet from the final centrifugation of the viral particle band was resuspended in a 0.1 M solution of histidine that contained 0.01 M $MgCl_2$ (pH 6.2). The

sample contained all viral components except the P2 protein, which was removed by the CCl_4 treatment. This removal prevents infection through oral intake by the insect and direct injection is instead required for vector infection.

Pre-characterization experiments

Reference samples need to have a known structure as well as be available in high concentration, have monodisperse size distribution, and need to be compatible with the available sample delivery techniques. Pre-characterisation was performed using Dynamic Light Scattering (DLS), Nanoparticle Tracking Analysis (NTA), Differential Mobility Analysis (DMA) and aerosol injection testing. RDV was determined to satisfy these requirements and was selected as the test sample for this data set.

Injection testing: Sample was injected using a setup identical to the subsequent LCLS experiment (see *Sample injection at the LCLS*) to investigate their ability to aerosolize and their resistance to the injection procedure (Hantke 2014). By placing a microscope glass slide covered by a gel piece (Gel-Pak) beneath the outlet of the aerodynamic lens (at the same position as the interaction region with the X-ray beam in the subsequent LCLS experiments), a particle dust could be observed through an objective lens mounted below. In a second set of experiments a formvar/carbon grid (#01754-F, F/C 400 mesh Cu, Ted Pella Inc.) was substituted for the glass slide, which captured RDV particles that had traversed the injector. These samples were examined without further fixation by an environmental scanning electron microscope (ESEM) (Quanta FEG 650, FEI). The pressure in the vacuum chamber was kept at c. 10^{-5} mBar.

Sample size and monodispersity in the liquid phase: The size and polydispersity of the RDV sample in solution (250mM Ammonium Acetate buffer) was measured using both the DLS method (w130i, AvidNano Ltd. and Spectrolight 600, Molecular Dimensions) and the NTA technique (NanoSight, model LM10, Malvern Instruments Ltd.). For DLS and NTA the sample was diluted to 10^9 particles mL^{-1} and 10^8 particles mL^{-1}, respectively. The size distribution is shown in figure 2a and 2b.

Sample size and monodispersity in the gas phase: The sample size and size distribution in the gas phase were measured by means of Electrophoretic DMA. RDV was aerosolized with a nano-Electrospray ionization (ESI) source (TSI model 3480) and passed through an electrostatic classifier (TSI model 3480) whose size selection window was continuously scanned. Transmitted particles were counted with a condensation particle counter (CPC, TSI model 3786). The size distribution is shown in figure 2c.

Sample injection at the LCLS

The experiment was carried out at the CXI instrument at the LCLS (Boutet 2010, Liang 2015). An aerosol injector (described in Hantke 2014) was used to introduce the particles into the X-ray beam. Purified RDV were transferred to a volatile buffer (250 mM ammonium acetate, pH 7.5) at a concentration of 10^{12} particles mL^{-1} and introduced to the injector via a gas

dynamic virtual nozzle (GDVN) (DePonte 2008) at a flow rate of 1-2 µL min^{-1}. The aerosol continued through a skimmer, relaxation chamber and lastly an aerodynamic lens, as described elsewhere. By regulating gas and liquid flow and the skimmer pressure the quality of the particle beam could be optimized. The particles intersected the X-ray beam in random orientation.

Experimental setup and Data collection

Data was collected at the CXI instrument at LCLS. The LCLS was tuned to a photon energy of 7 keV and produced pulses with ~4 mJ pulse energy and <50 fs duration. The selection of the photon energy, within the CXI operation range of 5 to 11 keV, was driven by the competition between the sample's scattering cross section, which tends to be high at lower energies, and the ability to discriminate single-photon events in the detector, which increases at the high end of the range. Additionally, there is a desire to remain below the Iron K-alpha excitation edge at 7.1 keV, as the isotropic fluorescence signal further complicates photon identification in the data.

X-rays were focused with a pair of Kirkpatrick-Baetz (KB) mirrors to a nominal size of 0.1 x 0.1 um. The focused beam passed through a set of beam-defining apertures to reduce the x-ray scattering imperfections in the optical system. An additional post sample aperture was used to limit background scatter. The post sample aperture also limited the collection angle to 3.2 Å resolution. Small angle diffraction patterns were recorded with a CSPAD 140 kPixel detector (also referred to as back detector), located 2.4 m downstream of the interaction and high-angle scattering was captured on a 1.5 MPixel CSPAD detector (also referred to as front detector) (Hart 2012), located 217.4 mm downstream of the interaction point, in a tandem arrangement as shown in the center panel of figure 1. All data events were recorded and synchronized with the repetition rate of the LCLS of 120 Hz. The back detector was offset with respect to the optical axis of the focusing optics, and extended to a maximal resolution of 15.2 nm and 11.6 nm on the edge and in the corner, respectively. A semitransparent beam-stop was utilized so that very low-q scattering could be collected, as well as provide a monitor for the direct beam. Data was analysed onsite using *Hummingbird*, a fast online analysis tool developed for single particle imaging (Daurer et al.) and *Cheetah*, a software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data (Barty et al. 2014).

Data processing steps used

In order to provide interpretable data in addition to the raw XTC files, we selected a small subset of diffraction events (175 frames) and converted them into a CXI file using *psana* (Damiani 2016). For both back and front detectors, data has been calibrated using psana's *ImgAlgos.NDArrCalib* module with pedestal subtraction (*do_peds*), common-mode correction (do_cmod), statistical correction (do_stat) and gain corrections (do_gain) turned on. Pixel gains were calculated by generating per-pixel histograms from a flat-field run and fitting a bimodal distribution with respect to the noise peak and the single photon peak. Using psana's *CSPadPixCoords.CSPadImageProducer* and *CSPadPixCoords.CSPad2x2ImageProducer*, the detector panels are assembled in order to form real images. For a list of provided files and data entries, see section Data Records. The

175 patterns have been selected manually by means of identifying strong diffraction signal showing similarities to simulations. See Technical Validation for more details.

Data Records

The data is deposited in the (CXIDB) (Maia 2012) and stored in the CXIDB data format which is based on the HDF5 format. HDF5 files are readable in many computing environments, including Python using the h5py module and MATLAB using e.g. the h5read function. Convenient functions for accessing the CXIDB data file exist in the libspimage package for C and Python (Maia 2010). For visualizing data in the CXIDB format, the Owl software is convenient (https://github.com/FilipeMaia/owl/). In addition to the CXI file, we are providing the conversion script (create_dataset.py) and additional metadata files (selection.h5, psana.cfg) along with usage instructions. Detector panel calibration files mapping data to real space are also provided. Configuration files for psana and Cheetah are provided for completeness of describing processing performed on the deposited data.

Technical Validation

Background scattering and direct beam scatter.

A background scattering pattern was derived by averaging 1000 frames, which did not include any hits or dark frames. This background, as well as suggested masks for non-responsive pixels and beamstops are shown in Figure 6.

Additionally, for the back detector data, manifold-embedding methods were used to detect and identify the nature and origin of such stochastic changes, and quantify the necessary corrections. The manifold of raw RDV single-particle snapshots from shift-5 is shown in figure 7, where each point represents a diffraction pattern(Giannakis 2012)(Schwander 2012)(Hosseinizadeh 2014). The parabolic nature of this manifold reveals that a single parameter dominates the changes from snapshot to snapshot. Namely fluctuations in pulse intensity consistent with the self-amplified spontaneous emission process of the FEL, and can be corrected by appropriate normalization procedures (Hosseinizadeh 2015). In addition to a monotonic intensity change along the parabola, a prominent deviation is evident. This is from a shifts of about one pixel along the lateral direction of the beam. The cause of this shift in the shift is a drift in pointing of the offset mirrors on the beam-defining aperture of the KB mirrors.

Simulated diffraction data of expected size

In Figure 3, two diffraction patterns from different single particle hits are shown in comparison to simulated diffraction from homogeneous spheres of size 78nm. In the simulation we use a photon energy of 7000 eV and assume the mass density of RDV to be 1.381 g/cm^3. The back detector was simulated using a detector distance of 2.4 m, a pixel size of 110 microns and a signal conversion rate of 33 ADUs per photon.

Signal above background on front detector

The front detector was located 214.7mm downstream of the sample interaction region and collects diffraction at resolutions up to 3.2 Å resolution. Although hits are immediately apparent on the back detector, and this signal is used used for hit finding, determining whether there is useful signal from the sample on the front detector above background levels is not immediately apparent from any individual image. A radial average of the sum of frames determined by Cheetah to be hits shows that there is indeed consistently elevated signal above background when sample is detected to be in the beam based on the downstream detector (Figure 5). This intensity distribution falls off with the expected q-dependence, and stops at a resolution of 3.2 Å, this being the resolution limit set by clipping from the post-sample aperture. Beyond this resolution both radial sums are identical, further supporting the notion that signal up to 3.2 Å resolution comes from individual particles. This validates the potential usefulness of signal on the front detector for image analysis. Cheetah processing scripts are included in the archive.

Validation that scattering comes from RDV

Although the selected 175 frames are not enough for an intensity reconstruction, it is interesting to examine how well they agree with the known structure of RDV (PDB 1UF2) (Nagakawa 2003). The method of least surprise was used to determine whether signal on the front detector corresponded to the expected signal from RDV.

Data was converted from Analog-to-Digital Units (ADU) measured by the detector into photon counts using the relation

$$k_i \;=\; ceil[\,(A_i \;-\; 0.5\,\gamma)\,/\,\gamma\,]$$

where k_i is the photon count at pixel i, A_i is the dark, common-mode, gain-corrected ADU measured at pixel i, and γ is the average ADUs per photon for the detector which was calculated from a flat-field run. In this analysis, we use front detector data up to 6.67 Å resolution, which corresponds to a radius of 265 pixels.

Assuming Poisson statistics, we define the surprise function as the negative log-likelihood

$$S(K;\Phi,\Omega_j) \;=\; -\sum_{i=1}^{N} \quad log(\frac{n_i^{k_i} e^{-n_i}}{k_i!}) \;\equiv\; -\sum_{i=1}^{N} \quad log\; P(n_i,k_i),$$

where K denotes the dependence on data, with k_i being the measured photon count at pixel i, n_i is the average photon number at pixel i when the fluence is Φ and the RDV particle has orientation Ω_j, and the summation runs over all the pixels. Minimizing the surprise function, or maximizing the log-likelihood, across different orientations and fluence values, we assign each data frame with the orientation and fluence at which it was most likely recorded. To help us assess the quality of these assignments, we further "normalize" the surprise function. Given the RDV model with an estimate of the particle orientation and fluence, we calculate the mean

$$\langle\, S(\Phi,\Omega_j)\,\rangle \;=\; -\sum_{i=1}^{N} \quad \langle\, log\; P(n_i,k)\,\rangle$$

and the standard deviation

$$\sigma_S(\Phi,\Omega_j) \;=\; [\sum_{i=1}^{N} \quad \langle\,(log\;P(n_i,k))^2\,\rangle - \langle\,log\;P(n_i,k)\,\rangle^2\,]^{1/2}$$

of the surprise function, where $\langle\cdots\rangle$ denotes the expectation value under the Poisson distribution. Note that $\langle\,S(\Phi,\Omega_j)\,\rangle$ and $\sigma_S(\Phi,\Omega_j)$ are independent of the data K. The normalized surprise function, or its z-score,

$$z(K;\Phi,\Omega_j) \;\equiv\; \frac{S(K;\Phi,\Omega_j) \;-\; \langle\,S(\Phi,\Omega_j)\,\rangle}{\sigma_S(\Phi,\Omega_j)}$$

measures the agreement of the data with a known model: The data is inconsistent with the model when the absolute value of the z-score is much greater than unity: a z-score much greater than unity is consistent with the data being 'surprising' given the assumed model.

The z-scores of all the selected frames versus particle size are shown in Figure 8. The particle sizes were determined by fitting back detector data to a homogeneous sphere model with adjustable size. Frames with particle size close to the diameter (70.8 nm) of the RDV model generally have smaller z-scores, though some still manifest inconsistency with the model. The source of this could be the presence of a water layer on the particle surface. This model-based surprise function calculation may potentially be useful for hit-finding, especially when the signal is as weak as the front detector data. Unsurprisingly, we find that presence of RDV scattering signal on the front detector is correlated with minimum values of the surprise function.

Usage Notes (optional)

The dataset (CXIDB ID 36) contains the full data stream recorded during the experiment in .xtc format. The dataset also contains a set of pre-selected hits (as described above) from both CSPAD detectors plus instrument metadata. XTC files are the native format of LCLS can be read using analysis frameworks provided by the LCLS (see https://confluence.slac.stanford.edu/display/PSDM/LCLS+Data+Analysis).

Acknowledgements

"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."

ANL: This work is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical, Geological, and Biological Sciences, under Contract No. DE-AC02-06CH11357.

LASSP: U.S. Department of Energy (DOE) Grant No. DE-SC0005827

OSAKA: Japan Society for the Promotion of Science, The Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT); Grants-in-Aid for Scientific Research and

Bilateral Joint Research Projects by Japan Society for the Promotion of Science; X-ray Free Electron Laser Priority Strategy Program (The Ministry of Education, Culture, Sports, Science and Technology of Japan, MEXT)

GIST & POSTECH: NRF through the SRC (Grant No. NRF- 2015R1A5A1009962), MSIP and PAL, Korea

UU: The Swedish Research Council (VR), the Röngen-Ångström Cluster, the Swedish Foundation for Strategic Research (SSF), the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Knut and Alice Wallenberg Foundation, European Research Council (ERC)

ELI: The Ministry of Education, Youth and Sports of the Czech Republic (ELI-Beamlines Registered No. CZ.1.05/1.1.00/02.0061) and the Academy of Sciences of the Czech Republic (M100101210).

CFEL: Helmholtz Association through project oriented funds

BNL: Brookhaven National Laboratory is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-SC0012704.

ARC: This work was supported by the Australian Research Council Centre of Excellence in Advanced Molecular Imaging (CE140100011) www.imagingcoe.org.

Author contributions

NOTE: If you have not uploaded them directly to the google drive please do so ASAP.

Competing interests

IF YOU HAVE COMPETING INTERESTS, DECLARE THEM HERE, OTHERWISE:

The authors declare no competing interests.

A competing financial interests statement is required for all papers accepted by and published in *Scientific Data*. If there is no conflict of interest, a statement declaring this must still be included in the manuscript.

Figures

Figure 1

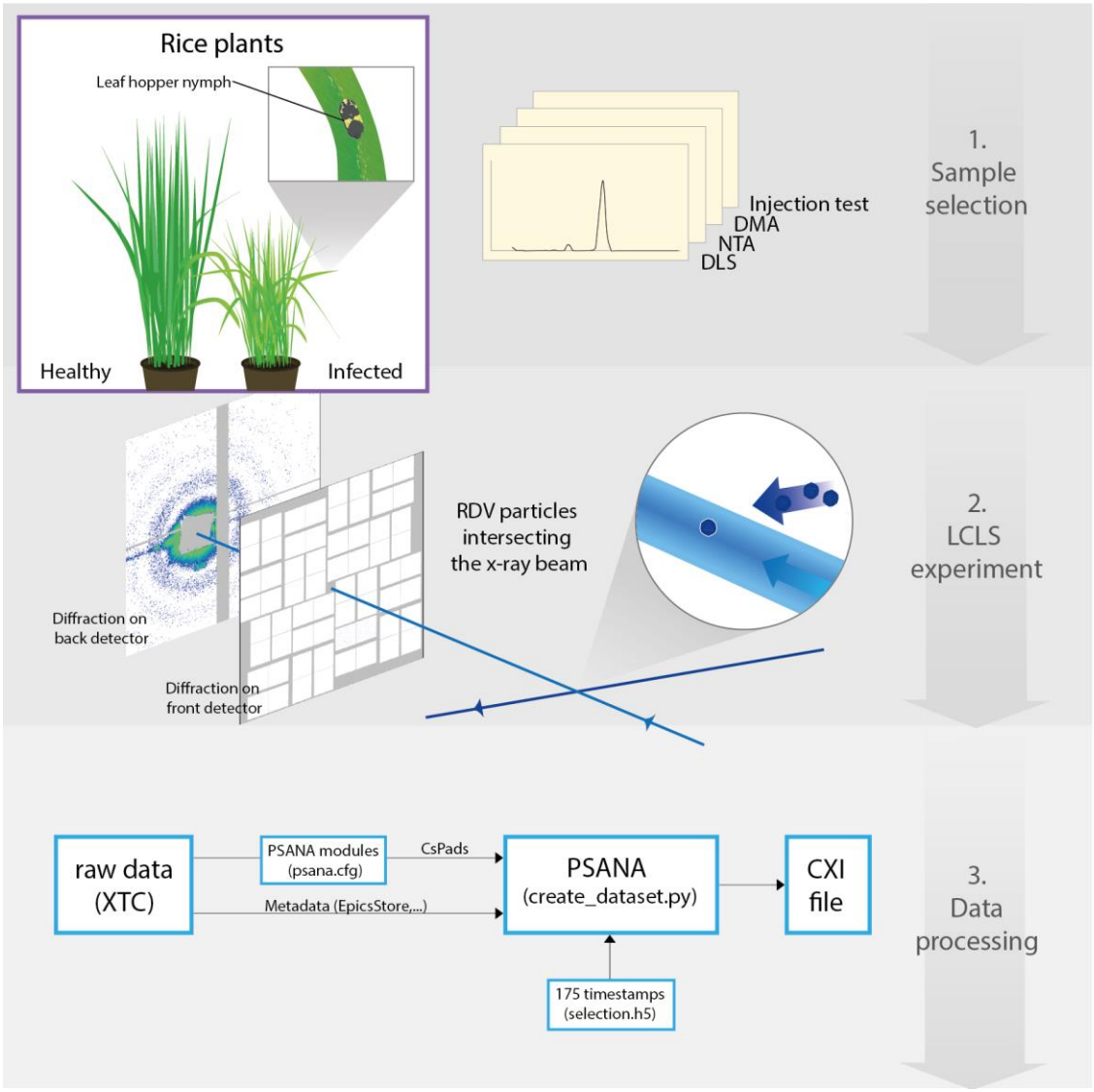

Figure 2a)

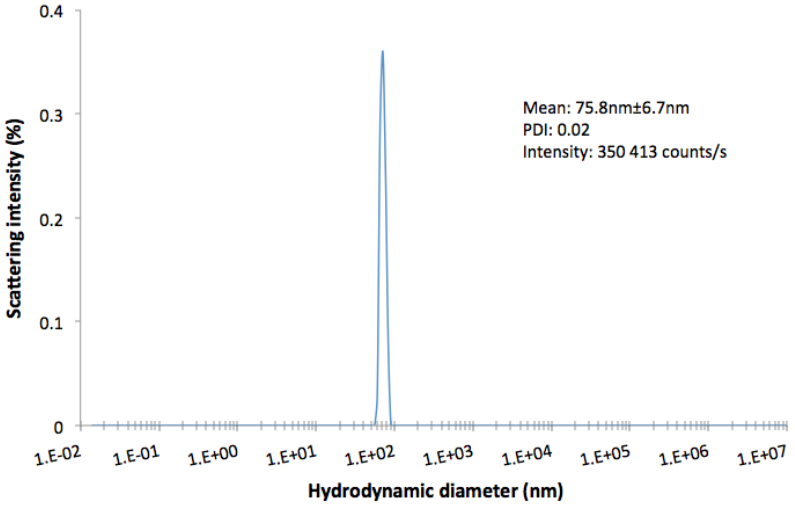

Figure 2b)

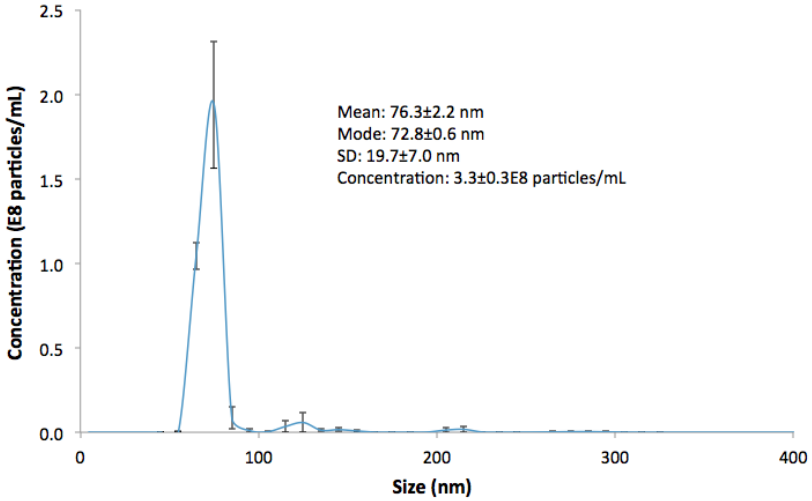

Figure 2c)

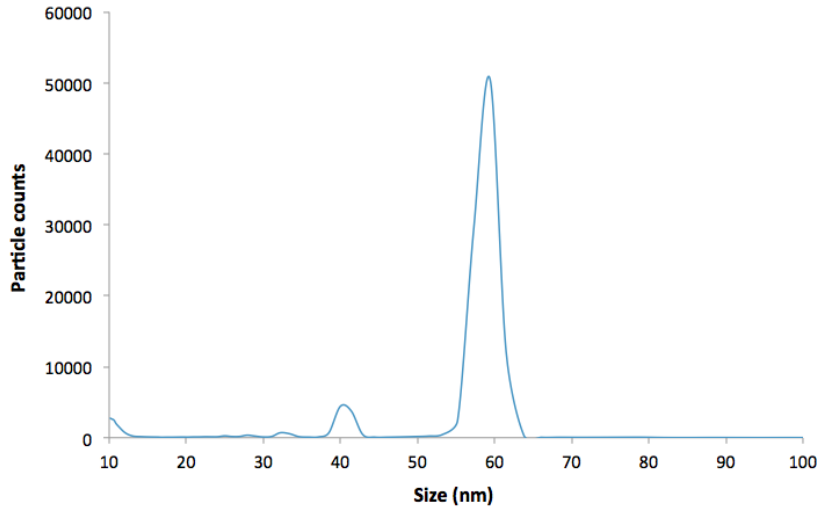
60000
50000
40000
30000
20000
10000
0
Particle counts
10
20
30
40
50
60
70
80
90
100
Size (nm)

Figure 3a)

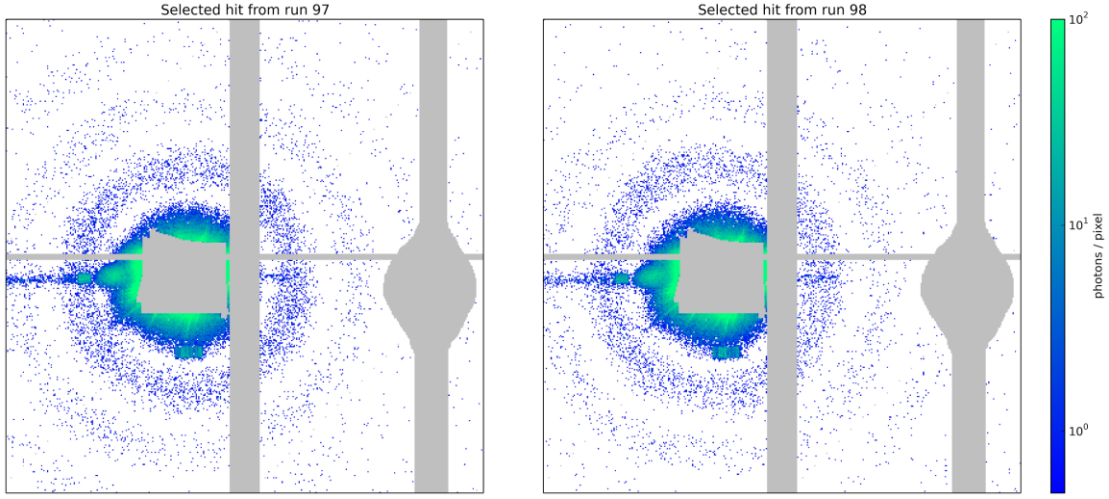
Selected hit from run 97
Selected hit from run 98
10^2
10^1
10^0
photons / pixel

Figure 3b)

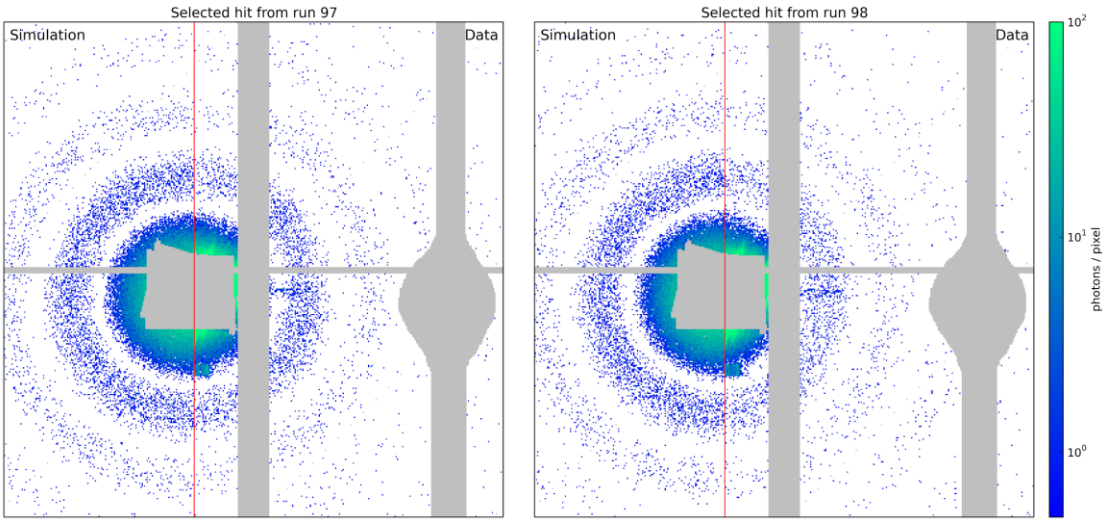

Figure 4a)[Note the front detector images display poorly with google's rescaling.]

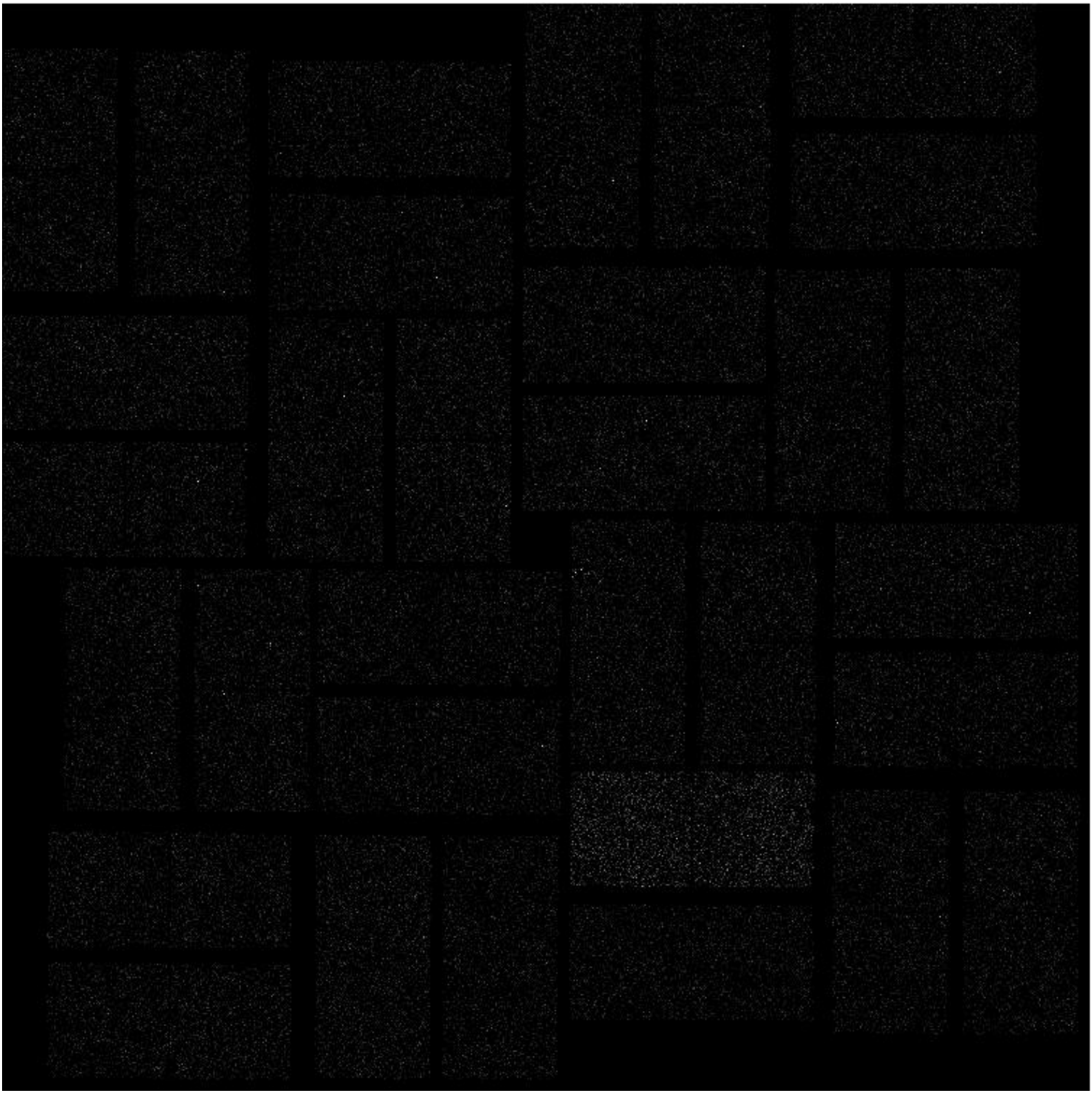

Figure 4b) [Note the front detector images display poorly with google's rescaling.]

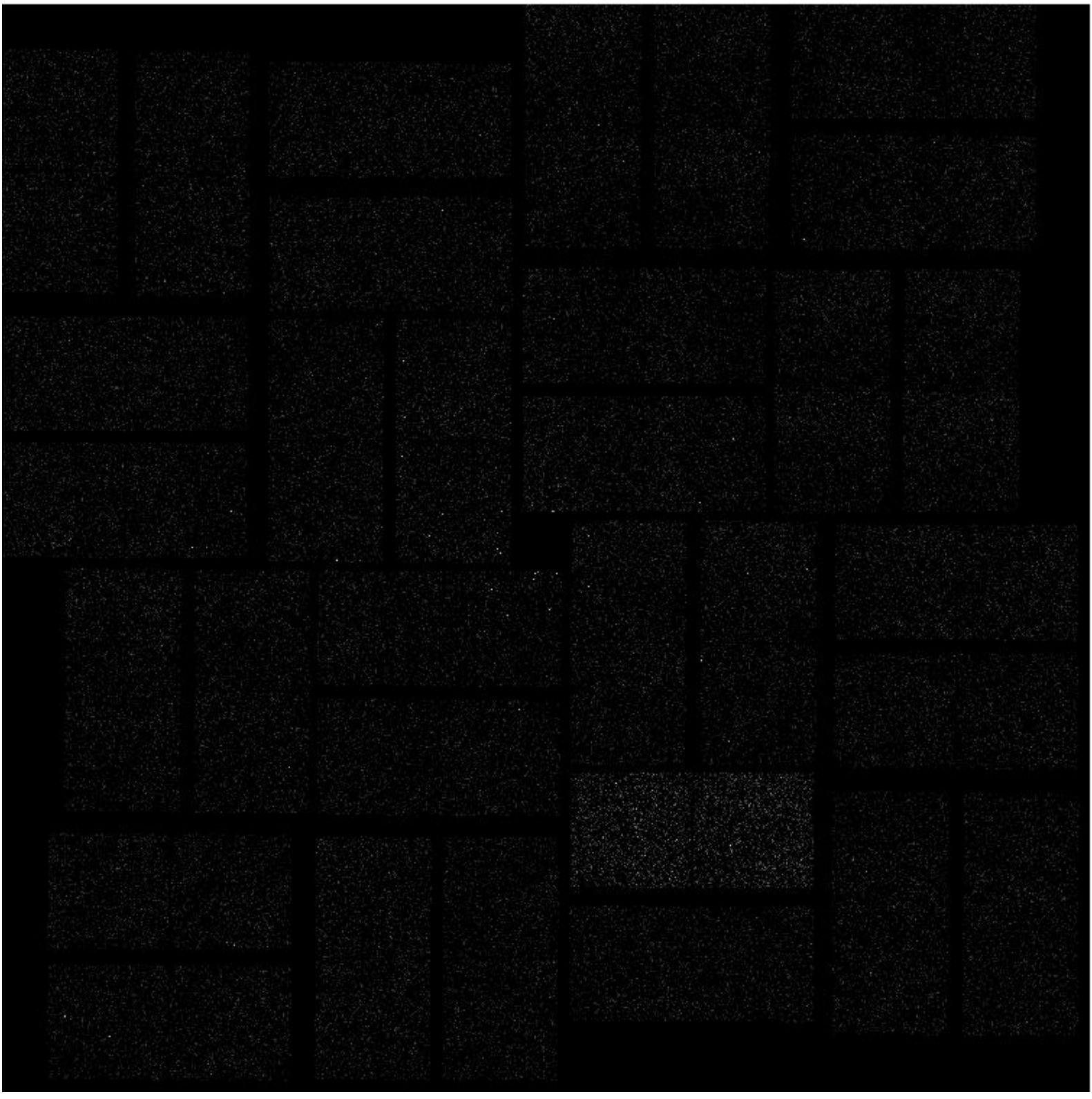

Figure 5

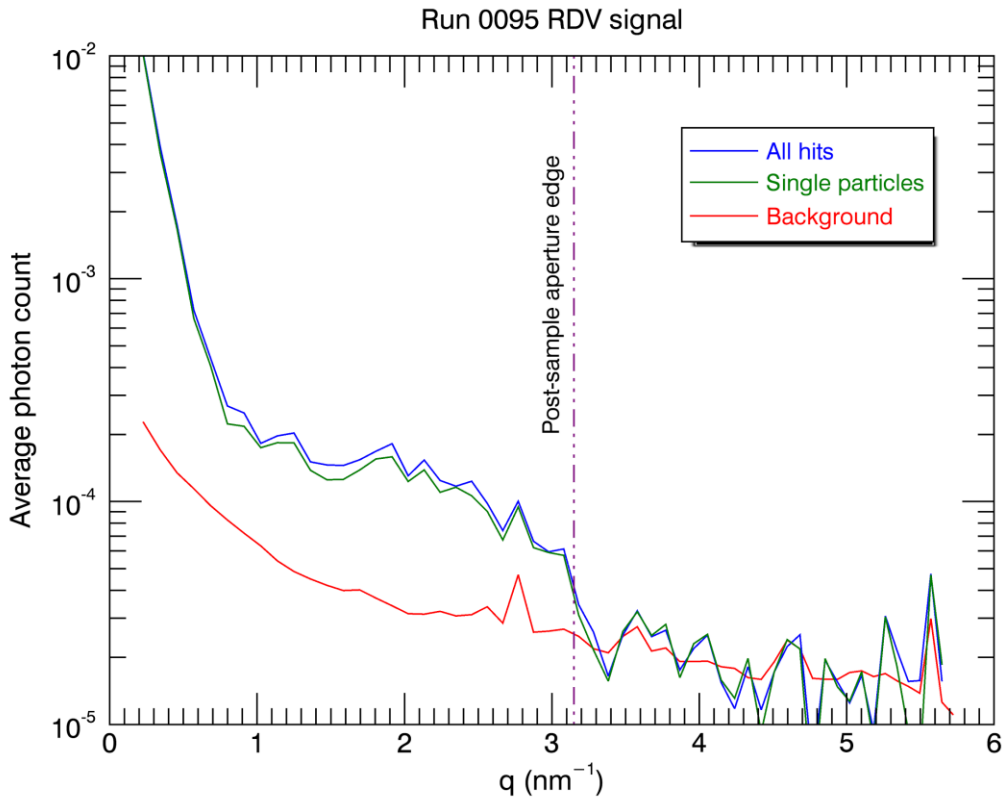

Figure 6

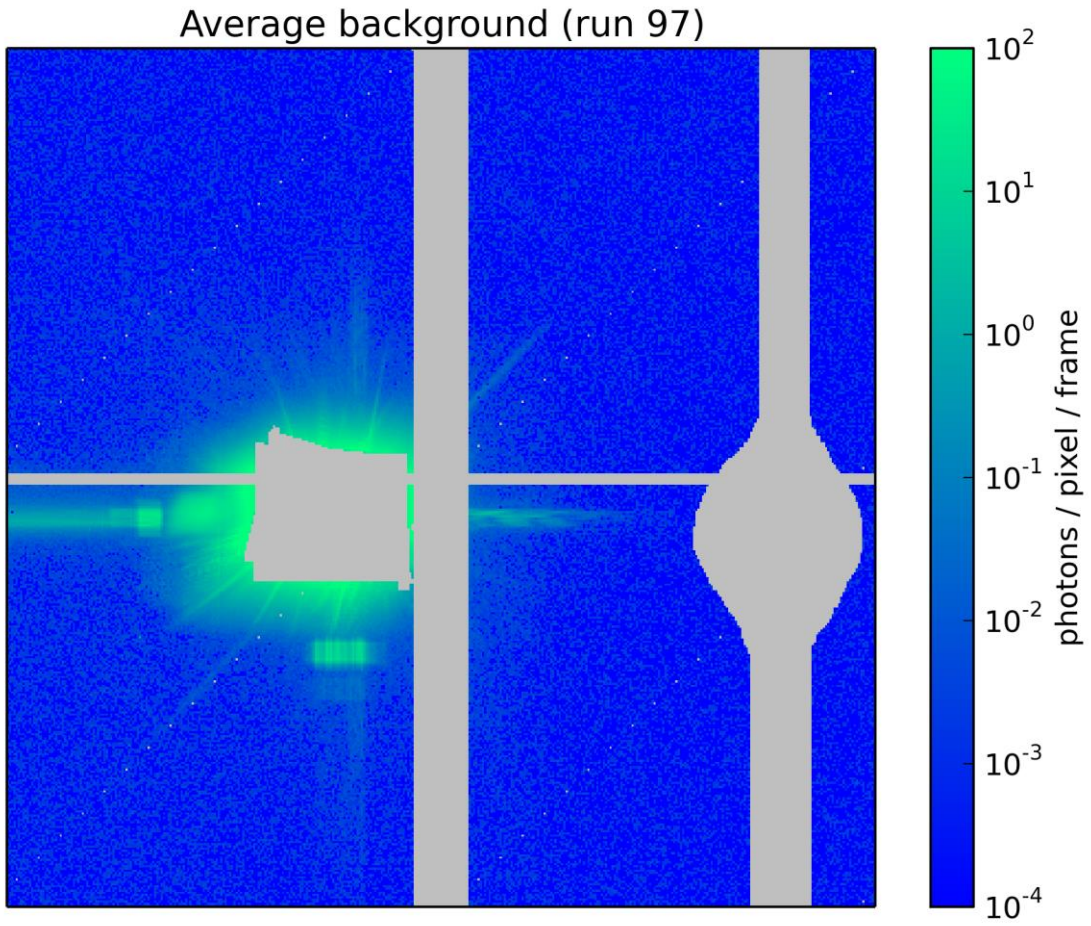
Average background (run 97)
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
photons / pixel / frame

Figure 7

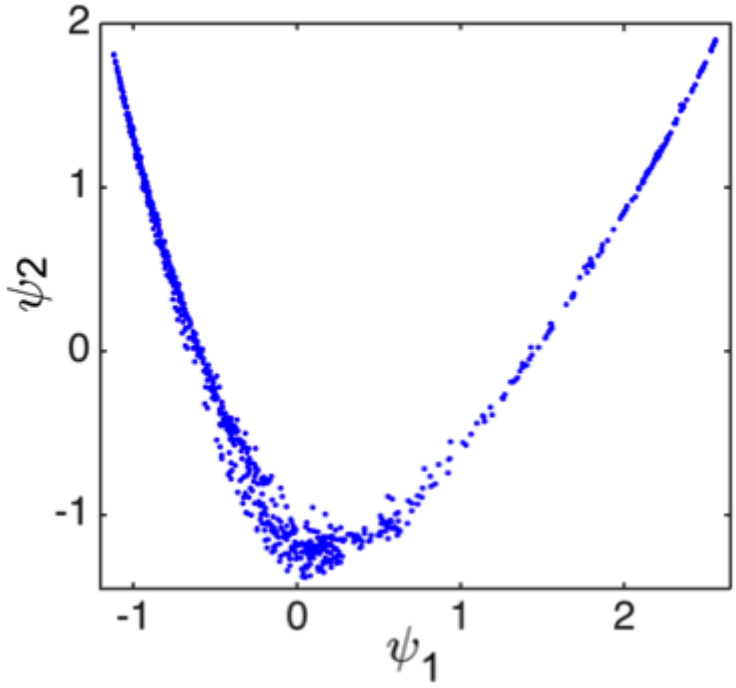

Figure 8

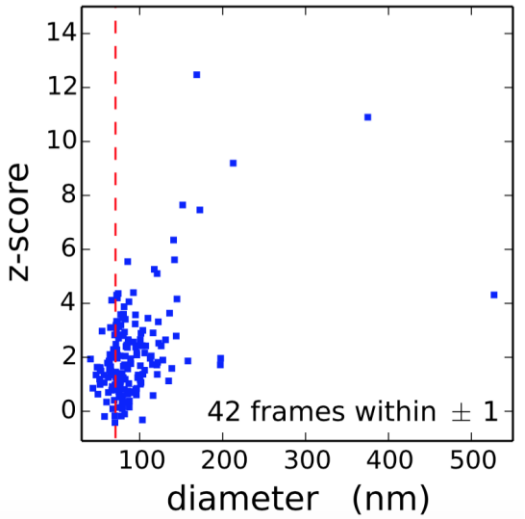

Figure Legends

Figure 1. Experimental design. The generated dataset presented is a result from three steps. In the first step, pre-characterization analysis of candidate samples were carried out and a primary sample, the RDV, was selected. The RDV was prepared by feeding grasshopper nymphs with infected rice plants. Secondly, the virus particles were injected into the x-ray beam of the LCLS and diffraction patterns on the front and back detector were recorded. The data was in the final step pre-processed using psana. In addition to the XTC files, 175 frames were selected and converted into a CXI file.

Figure 2. The purity of RDV analysed with DLS, NTA and DMA. a) Size distribution determined with DLS. The average size of the RDV particles was approximately 76 nm; b) Size distribution determined with NTA. The mean and mode of the RDV particles were approximately 76 and 73 nm, respectively; and c) the diameter peak from the DMA measurement was approximately 60 nm. The actual gas flow during the DMA measurement was lower than the set value, which resulted in a diameter low by about 10%. The gas flow was disregarded since the purpose of the DMA measurement was to assess sample purity and not size.

Figure 3. Two selected diffraction patterns of single particle hits recorded on the CSPAD 140k back detector showing the expected size for RDV and highest diffraction intensity. a) shows two of the brightest diffraction patterns found, while b) shows the same diffraction patterns displayed next to simulated diffraction from a homogenous sphere of size 78 nm and with a mass density of 1.381 g/cm^3. The simulation assumes a photon energy of 7000 eV, a detector distance of 2.4 m, a pixel size of 110 microns and a conversion of 33 ADUs per photon. Regions of beam-stops and gaps between the detector panels are masked in grey.

Figure 4. show the corresponding high-angle scattering data collected with the front CSPAD.

Figure 5. Radial average of signal on the front detector from blank frames compared to radial average from frames determined to be hits. Elevated photon counts from the sample are visible up to 3.2 Å resolution, this being the resolution limit set due to clipping by the post-sample aperture.

Figure 6. Background image of the back detector (CPSAD 140k), averaged from 1000 non-hits and non-dark frames. The grey areas, corresponding to the beamstop, gap between sensors, and a shadow mask of an additional beamstop form the beamstop holder, are masked out. Unbonded pixels, that do not read out signal appear in a diagonal line and are likewise masked out.

Figure 7. Manifold of RDV single-particle raw data (820 snapshots) in two dimensions. The coordinates ψ_1 and ψ_2 are the first two eigenfunctions of Laplace-Beltrami operator.

Figure 8. Front detector normalized surprise (z-score) versus back detector particle size fits. The dashed red line indicates the diameter (70.8 nm) of the RDV model.

References

Aquila, A., et al. The linac coherent light source single particle imaging road map. Struct. Dyn. 2, 041701 (2015).

Barty, Anton et al. " *Cheetah*: Software for High-Throughput Reduction and Analysis of Serial Femtosecond X-Ray Diffraction Data." *Journal of Applied Crystallography* 47.Pt 3 (2014): 1118–1131.

Boutet, S. and Williams GJ. "The Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS)," New Journal of Physics 12 (2010) 035024

Boutet, S. et al. High-resolution protein structure determination by serial femtosecond crystallography Science 337 (2012) DOI: 10.1126/science.1217737

Chapman, H.N., et al. High-resolution ab initio three-dimensional x-ray diffraction microscopy. JOSA A, 23 5 (2006).

Chapman, H.N., et al. Femtosecond X-ray protein nanocrystallography Nature 470 (2011) doi:10.1038/nature09750

Damiani, D. et al., LCLS data analysis using PSANA. *manuscript submitted*

Daurer et al., *manuscript in preparation*

DePonte, D. P., et al. Gas dynamic virtual nozzle for generation of microscopic droplet streams. J. Phys. D 41, 195505 (2008).

Ekeberg, T., et al. Three-dimensional reconstruction of the giant mimivirus particle with an X-ray free electron laser. Physical Review Letters 112 (2015) 098102

Giannakis D., Schwander P. and Ourmazd A., "The symmetries of image formation by scattering. I. Theoretical framework", Opt. Express 20, 12799-12826 (2012).

Hantke, M., et al. High-throughput imaging of heterogeneous cell organelles with an X-ray laser Nature Photonics 8 (2014) doi:10.1038/nphoton.2014.270

Hart, P., et al. "The CSPAD megapixel x-ray camera at LCLS" Proc. SPIE 8504, X-Ray Free-Electron Lasers: Beam Diagnostics, Beamline Instrumentation, and Applications, 85040C (October 15, 2012); doi:10.1117/12.930924

Henderson, R. "The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules." Q. Rev. Biophys. 28, 171-193 (1995).

Herrmann, S., et al. "CPSAD upgrades and CSPAD V1.5 at LCLS" Journal of Physics: Conference Series 493 (2014) 012013 doi:10.1088/1742-6596/493/1/012013

Hosseinizadeh A., et al., "High-resolution structure of viruses from random diffraction snapshots", Phil. Trans. R. Soc. B 369, 20130326 (2014).

Hosseinizadeh A., et al., "Single-particle structure determination by X-ray free electron lasers: Possibilities and challenges", Struct. Dyn. 2, 041601 (2015).

Howells, M.R. et al, "An assessment of the resolution limitation due to radiation-damage in X-ray diffraction microscopy" Journal of Electron Spectroscopy and Related Phenomena 170 (2009) 4–12

Kano, H., et al. Nucleotide sequence of rice dwarf virus (RDV) genome segment S3 coding for 114 K major core protein. Nucleic Acids Res. 18, 6700 (1990).

Kimura, I., Minobe, Y. and Omura, T. Changes in a Nucleic Acid and a Protein Component of Rice Dwarf Virus Particles Associated with an Increase in Symptom Severity. J. gen. Virol. 68, 3211-25 (1987).

Liang M., et al, "The Coherent X-ray Imaging instrument at the Linac Coherent Light Source," J. Synchrotron Rad. (2015). 22, 514-519
doi: 10.1107/S160057751500449X

Maia, F.R.N.C. The Coherent X-ray Imaging Data Bank. Nat. Methods 9, 854–5 (2012).
Chapman, H.N., et al. Femtosecond diffractive imaging with a soft X-ray free electron laser Nature Physics 2 (2006) doi:10.1038/nphys461

Maia, F.R.N.C., Ekeberg, T., van der Spoel, D. & Hajdu, J. Hawk: the image reconstruction package for coherent X-ray diffractive imaging. J. Appl. Crystallogr. 43, 1535–1539 (2010).

Nakagawa, A. et al. The Atomic Structure of Rice dwarf Virus Reveals the Self-Assembly Mechanism of Component Proteins. Structure. 11, 1227-38 (2003).

Neutze, R., Wouts R., van der Spoel, D., Weckert, E. and Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. Nature. 406, 752-7.

Omura, T., Morinaka, T., Inoue, H. and Saito, Y. Purification and some properties of rice gall dwarf virus, a new Phytoreovirus. Phytopath. 72, 1246-9 (1982).

Omura, T., et al. The outer capsid protein of rice dwarf virus is encoded by genome segment S8. J Gen Virol. 70, 2759–2764 (1989). doi: 10.1099/0022-1317-70-10-2759

Schwander P., Giannakis D., Yoon C.H., Ourmazd A., "The symmetries of image formation by scattering. II. Applications", Opt. Express 20, 12827–12849 (2012).

Seibert, M.M., et al. Single mimivirus particles intercepted and imaged with an X-ray laser. Nature 470 (2011) doi:10.1038/nature09748

Suzuki, N., et al. Molecular analysis of rice dwarf phytoreovirus segment S1: interviral homology of the putative RNA-dependent RNA polymerase between plant- and animal-infecting reoviruses. Virology. (1992) 190:240-7. doi:10.1016/0042-6822(92)91210-L

Suzuki, N., Kusano, T., Matsuura, Y., and Omura, T. Novel NTP binding property of rice dwarf phytoreovirus minor core protein P5. Virology. (1996) 219:471-4. doi:10.1006/viro.1996.0273

Ueda, S. and Uyeda, I. The rice dwarf phytoreovirus structural protein P7 possesses non-specific nucleic acids binding activity in vitro. Molecular Plant Pathology On-Line (1997) (http://www.bspp.org.uk/mppol/1997/0123ueda/)

van der Schot, G., et al. Imaging single cells in a beam of live cyanobacteria with an X-ray laser Nature Communications 6 (2015) doi:10.1038/ncomms6704

Xu, R., et al. "Single-shot three-dimensional structure determination of nanocrystals with femtosecond X-ray free-electron laser pulses", Nature Communications (2014) 5(4061) DOI: 10.1038/ncomms5061

Yan, J., et al. P2 protein encoded by genome segment S2 of rice dwarf phytoreovirus is essential for virus infection. Virology. 224, 539-41 (1996) doi:10.1006/viro.1996.0560

Zhong, B, et al. A minor outer capsid protein, P9, of Rice dwarf virus. Arch Virol. (2003) 148:2275-80. 10.1007/s00705-003-0160-3