Automated identification and classification of single particle serial femtosecond X-ray diffraction data

Jakob Andreasson,[1,10] Andrew V. Martin,[2,10] Meng Liang,[2] Nicusor Timneanu,[1] Andrew Aquila,[3] Fenglin Wang,[2] Bianca Iwan,[1] Martin Svenda,[1] Tomas Ekeberg,[1] Max Hantke,[1] Johan Bielecki,[1] Daniel Rolles,[4,5] Artem Rudenko,[6] Lutz Foucar,[4,7] Robert Hartmann,[8] Benjamin Erk,[4,5] Benedikt Rudek,[4] Henry N. Chapman,[2,9] Janos Hajdu,[1,3] and Anton Barty[2,*]

[1]*Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-751 24 Uppsala, Sweden*
[2]*Center for Free-electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany*
[3]*European XFEL, Albert-Einstein-Ring 19, 22761 Hamburg, Germany*
[4]*Max Planck Advanced Study Group, Center for Free-Electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany*
[5]*DESY, Notkestrasse 85, 22607 Hamburg, Germany*
[6]*J.R. Macdonald Laboratory, Department of Physics, Kansas State University, 116 Cardwell Hall, Manhattan, Kansas 66506 USA*
[7]*Max-Planck Institute for Medical Research, Jahnstrasse 29, D-69120 Heidelberg, Germany*
[8]*PNSensor GmbH, Römerstrasse 28, 80803 München, Germany*
[9]*University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany*
[10]*These authors contributed equally to this work*
[*]*anton.barty@desy.de*

Abstract: The first hard X-ray laser, the Linac Coherent Light Source (LCLS), produces 120 shots per second. Particles injected into the X-ray beam are hit randomly and in unknown orientations by the extremely intense X-ray pulses, where the femtosecond-duration X-ray pulses diffract from the sample before the particle structure is significantly changed even though the sample is ultimately destroyed by the deposited X-ray energy. Single particle X-ray diffraction experiments generate data at the FEL repetition rate, resulting in more than 400,000 detector readouts in an hour, the data stream during an experiment contains blank frames mixed with hits on single particles, clusters and contaminants. The diffraction signal is generally weak and it is superimposed on a low but continually fluctuating background signal, originating from photon noise in the beam line and electronic noise from the detector. Meanwhile, explosion of the sample creates fragments with a characteristic signature. Here, we describe methods based on rapid image analysis combined with ion Time-of-Flight (ToF) spectroscopy of the fragments to achieve an efficient, automated and unsupervised sorting of diffraction data. The studies described here form a basis for the development of real-time frame rejection methods, e.g. for the European XFEL, which is expected to produce 100 million pulses per hour.

OCIS codes: (140.7240) UV, EUV, and X-ray lasers; (170.1650) Coherence imaging; (300.6350) Spectroscopy, ionization.

References and links

1. P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F.-J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, P. Hering, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H.-D. Nuhn, G. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu, G. Yocky, and J. Galayda, "First lasing and operation of an Ångström-wavelength free-electron laser," Nat. Photonics **4**(9), 641–647 (2010).

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

2. M. J. Bogan, W. H. Benner, S. Boutet, U. Rohner, M. Frank, A. Barty, M. M. Seibert, F. R. N. C. Maia, S. Marchesini, S. Bajt, B. Woods, V. Riot, S. P. Hau-Riege, M. Svenda, E. Marklund, E. Spiller, J. Hajdu, and H. N. Chapman, "Single particle X-ray diffractive imaging," Nano Lett. **8**(1), 310–316 (2008).
3. D. P. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. C. H. Spence, and R. B. Doak, "Gas dynamic virtual nozzle for generation of microscopic droplet streams," J. Phys. D Appl. Phys. **41**(19), 195505 (2008).
4. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu, "Potential for biomolecular imaging with femtosecond X-ray pulses," Nature **406**(6797), 752–757 (2000).
5. H. N. Chapman, A. Barty, M. J. Bogan, S. Boutet, M. Frank, S. P. Hau-Riege, S. Marchesini, B. W. Woods, S. Bajt, W. H. Benner, R. A. London, E. Plönjes, M. Kuhlmann, R. Treusch, S. Düsterer, T. Tschentscher, J. R. Schneider, E. Spiller, T. Möller, Ch. Bostedt, M. Hoener, D. A. Shapiro, K. O. Hodgson, D. van der Spoel, F. Burmeister, M. Bergh, C. Caleman, G. Huldt, M. M. Seibert, F. R. N. C. Maia, R. W. Lee, A. Szöke, N. Timneanu, and J. Hajdu, "Femtosecond diffractive imaging with a soft-X-ray free-electron laser," Nat. Phys. **2**(12), 839–843 (2006).
6. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, Ch. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Hömke, C. Reich, D. Pietschner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. Kühnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jönsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C.-D. Schröter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J. C. H. Spence, "Femtosecond X-ray protein nanocrystallography," Nature **470**(7332), 73–77 (2011).
7. M. M. Seibert, T. Ekeberg, F. R. N. C. Maia, M. Svenda, J. Andreasson, O. Jönsson, D. Odić, B. Iwan, A. Rocker, D. Westphal, M. Hantke, D. P. DePonte, A. Barty, J. Schulz, L. Gumprecht, N. Coppola, A. Aquila, M. Liang, T. A. White, A. Martin, C. Caleman, S. Stern, C. Abergel, V. Seltzer, J.-M. Claverie, Ch. Bostedt, J. D. Bozek, S. Boutet, A. A. Miahnahri, M. Messerschmidt, J. Krzywinski, G. Williams, K. O. Hodgson, M. J. Bogan, C. Y. Hampton, R. G. Sierra, D. Starodub, I. Andersson, S. Bajt, M. Barthelmess, J. C. Spence, P. Fromme, U. Weierstall, R. Kirian, M. Hunter, R. B. Doak, S. Marchesini, S. P. Hau-Riege, M. Frank, R. L. Shoeman, L. Lomb, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, C. Schmidt, L. Foucar, N. Kimmel, P. Holl, B. Rudek, B. Erk, A. Hömke, C. Reich, D. Pietschner, G. Weidenspointner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, I. Schlichting, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. U. Kühnel, R. Andritschke, C. D. Schröter, F. Krasniqi, M. Bott, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, H. N. Chapman, and J. Hajdu, "Single mimivirus particles intercepted and imaged with an X-ray laser," Nature **470**(7332), 78–81 (2011).
8. The European X-ray free-electron laser, Technical design report (2007).
9. V. Elser, "Noise limits on reconstructing diffraction signals from random tomographs," IEEE Trans. Inf. Theory **55**(10), 4715–4722 (2009).
10. R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, "Structure from fleeting illumination of faint spinning objects in flight," Nat. Phys. **5**(1), 64–67 (2009).
11. B. La Scola, S. Audic, C. Robert, L. Jungang, X. de Lamballerie, M. Drancourt, R. Birtles, J.-M. Claverie, and D. Raoult, "A giant virus in Amoebae," Science **299**(5615), 2033 (2003).
12. Ch. Bostedt, J. D. Bozek, P. H. Bucksbaum, R. N. Coffee, J. B. Hastings, Z. Huang, R. W. Lee, S. Schorb, J. N. Corlett, P. Denes, P. Emma, R. W. Falcone, R. W. Schoenlein, G. Doumy, E. P. Kanter, B. Kraessig, S. Southworth, L. Young, L. Fang, M. Hoener, N. Berrah, C. Roedig, and L. F. DiMauro, "Ultra-fast and ultra-intense x-ray sciences: first results from the Linac Coherent Light Source free-electron laser," J. Phys. At. Mol. Opt. Phys. **46**(16), 164003 (2013).
13. L. Strüder, S. Epp, D. Rolles, R. Hartmann, P. Holl, G. Lutz, H. Soltau, R. Eckart, Ch. Reich, K. Heinzinger, Ch. Thamm, A. Rudenko, F. Krasniqi, K.-U. Kühnel, Ch. Bauer, C.-D. Schröter, R. Moshammer, S. Techert, D. Miessner, M. Porro, O. Hälker, N. Meidinger, N. Kimmel, R. Andritschke, F. Schopper, G. Weidenspointner, A. Ziegler, D. Pietschner, S. Herrmann, U. Pietsch, A. Walenta, W. Leitenberger, Ch. Bostedt, T. Möller, D. Rupp, M. Adolph, H. Graafsma, H. Hirsemann, K. Gärtner, R. Richter, L. Foucar, R. L. Shoeman, I. Schlichting, and J. Ullrich, "Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources," Nucl. Instrum. Methods Phys. Res. A **614**, 483–496 (2010).
14. N. D. Loh, C. Y. Hampton, A. V. Martin, D. Starodub, R. G. Sierra, A. Barty, A. Aquila, J. Schulz, L. Lomb, J. Steinbrener, R. L. Shoeman, S. Kassemeyer, C. Bostedt, J. Bozek, S. W. Epp, B. Erk, R. Hartmann, D. Rolles, A. Rudenko, B. Rudek, L. Foucar, N. Kimmel, G. Weidenspointner, G. Hauser, P. Holl, E. Pedersoli, M. Liang, M. S. Hunter, L. Gumprecht, N. Coppola, C. Wunderer, H. Graafsma, F. R. N. C. Maia, T. Ekeberg, M. Hantke, H. Fleckenstein, H. Hirsemann, K. Nass, T. A. White, H. J. Tobias, G. R. Farquar, W. H. Benner, S. P. Hau-Riege, C. Reich, A. Hartmann, H. Soltau, S. Marchesini, S. Bajt, M. Barthelmess, P. Bucksbaum, K. O. Hodgson, L. Strüder, J. Ullrich, M. Frank, I. Schlichting, H. N. Chapman, and M. J. Bogan, "Fractal morphology, imaging

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

and mass spectrometry of single aerosol particles in flight," Nature **486**(7404), 513–517 (2012).
15. T. Gorkhover, M. Adolph, D. Rupp, S. Schorb, S. W. Epp, B. Erk, L. Foucar, R. Hartmann, N. Kimmel, K.-U. Kühnel, D. Rolles, B. Rudek, A. Rudenko, R. Andritschke, A. Aquila, J. D. Bozek, N. Coppola, T. Erke, F. Filsinger, H. Gorke, H. Graafsma, L. Gumprecht, G. Hauser, S. Herrmann, H. Hirsemann, A. Hömke, P. Holl, C. Kaiser, F. Krasniqi, J.-H. Meyer, M. Matysek, M. Messerschmidt, D. Miessner, B. Nilsson, D. Pietschner, G. Potdevin, C. Reich, G. Schaller, C. Schmidt, F. Schopper, C. D. Schröter, J. Schulz, H. Soltau, G. Weidenspointner, I. Schlichting, L. Strüder, J. Ullrich, T. Möller, and Ch. Bostedt, "Nanoplasma dynamics of single large Xenon clusters irradiated with superintense X-ray pulses from the Linac Coherent Light Source free-electron laser," Phys. Rev. Lett. **108**(24), 245005 (2012).
16. http://www.rmjordan.com/
17. W. C. Wiley and I. H. McLaren, "Time of flight mass spectrometer with improved resolution," Rev. Sci. Instrum. **26**(12), 1150–1157 (1955).
18. J. Andreasson, B. Iwan, A. Andrejczuk, E. Abreu, M. Bergh, C. Caleman, A. J. Nelson, S. Bajt, J. Chalupsky, H. N. Chapman, R. R. Fäustlin, V. Hajkova, P. A. Heimann, B. Hjörvarsson, L. Juha, D. Klinger, J. Krzywinski, B. Nagler, G. K. Pálsson, W. Singer, M. M. Seibert, R. Sobierajski, S. Toleikis, T. Tschentscher, S. M. Vinko, R. W. Lee, J. Hajdu, and N. Tîmneanu, "Saturated ablation in metal hydrides and acceleration of protons and deuterons to keV energies with a soft-x-ray laser," Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **83**(1), 016403 (2011).
19. B. Iwan, J. Andreasson, A. Andrejczuk, E. Abreu, M. Bergh, C. Caleman, A. J. Nelson, S. Bajt, J. Chalupsky, H. N. Chapman, R. R. Fäustlin, V. Hajkova, P. A. Heimann, B. Hjörvarsson, L. Juha, D. Klinger, J. Krzywinski, B. Nagler, G. K. Pálsson, W. Singer, M. M. Seibert, R. Sobierajski, S. Toleikis, T. Tschentscher, S. M. Vinko, R. W. Lee, J. Hajdu, and N. Tîmneanu, "TOF-OFF: A method for determining focal positions in tightly focused free-electron laser experiments by measurement of ejected ions," High Energy Density Phys. **7**(4), 336–342 (2011).
20. C. H. Yoon, P. Schwander, C. Abergel, I. Andersson, J. Andreasson, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, M. J. Bogan, Ch. Bostedt, J. Bozek, H. N. Chapman, J.-M. Claverie, N. Coppola, D. P. DePonte, T. Ekeberg, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, A. Hartmann, E. Hartmann, R. Hartmann, G. Hauser, H. Hirsemann, P. Holl, S. Kassemeyer, N. Kimmel, M. Kiskinova, M. Liang, N.-T. D. Loh, L. Lomb, F. R. N. C. Maia, A. V. Martin, K. Nass, E. Pedersoli, Ch. Reich, D. Rolles, B. Rudek, A. Rudenko, I. Schlichting, J. Schulz, M. M. Seibert, V. Seltzer, R. L. Shoeman, R. G. Sierra, H. Soltau, D. Starodub, J. Steinbrener, G. Stier, L. Strüder, M. Svenda, J. Ullrich, G. Weidenspointner, T. A. White, C. Wunderer, and A. Ourmazd, "Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering," Opt. Express **19**(17), 16542–16549 (2011).

1. Introduction

The advent of free-electron lasers in the X-ray regime [1] has opened new avenues for the structural determination of cells, viruses and aerosol particles in free flight. X-ray pulses of only a few femtoseconds duration intersect a stream of particles contained in either an aerosol jet [2] or liquid stream [3], capturing structural information in the form of X-ray diffraction before the sample explodes [2,4–7]. A fast area detector reads out the scattered signals after each pulse (Fig. 1) and ideally every frame read out from the detector will consist of X-ray diffraction data from a single particle. However, in practice the particle density in the interaction region is relatively low, and the FEL beam randomly intersects the particle beam such that useful particle hits are interspersed with blank data frames. Furthermore, diffraction occurs from anything in the FEL beam including clusters, water droplets, and contaminant material in addition to the desired single particles. This mixture of data is randomly sampled in time and must be separated according to particle type during data analysis. The comparatively large data volumes generated by such experiments motivates the development of automated and unsupervised data reduction techniques designed to work efficiently on data sets containing millions of data frames.

Current X-ray free-electron lasers operate at repetition rates of 10 to 120 Hz, future facilities plan for bursts of X-ray pulses at megahertz spacing. Each X-ray pulse represents a separate measurement requiring specialized X-ray pixel detectors capable of reading out full frame images on each X-ray pulse. At the LCLS, this produces a data stream of over 400,000 images per hour and data rates in excess of 1.6 TB per hour. Individual experiments can easily generate 200 TB of raw data today. The ability to automatically reduce this deluge of data into a compact form by selecting good hits from other frames is a critical first step in data analysis. This challenge will become more critical as new facilities with higher pulse repetition rates

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

come online in the near future: the European XFEL [8] for example, promises X-ray pulse rates of up to 27,000 pulses per second, motivating the development of efficient and unsupervised real-time frame rejection strategies. An ability to reject frames before the pixel detector is read out would be highly advantageous, and could be based on secondary diagnostics, including a real-time analysis of the ion spectrum from the sample explosion. Here, we present two strategies for automated selection of single particle diffraction events and evaluate their effectiveness on data sets collected at LCLS.

2. Finding particle diffraction events in a sea of blank frames

During aerosol injection particles enter the interaction zone at velocities of 50-100 m/s and are intercepted by X-ray pulses of femtosecond duration (Fig. 1). Since it is currently not possible to trigger the FEL on demand when a particle is known to be in the interaction region, interaction between particles and X-ray pulses occurs at random. With the particle densities achievable using current sample handling technology not every X-ray pulse hits a particle [2,7]. Laser-based particle tracking may be able to predict when a particle is near the interaction region, however when the particle beam diameter is much larger than the FEL focal spot, determining whether or not a particle was actually in the X-ray beam at the time of measurement can only be achieved from data measured during the FEL pulse.

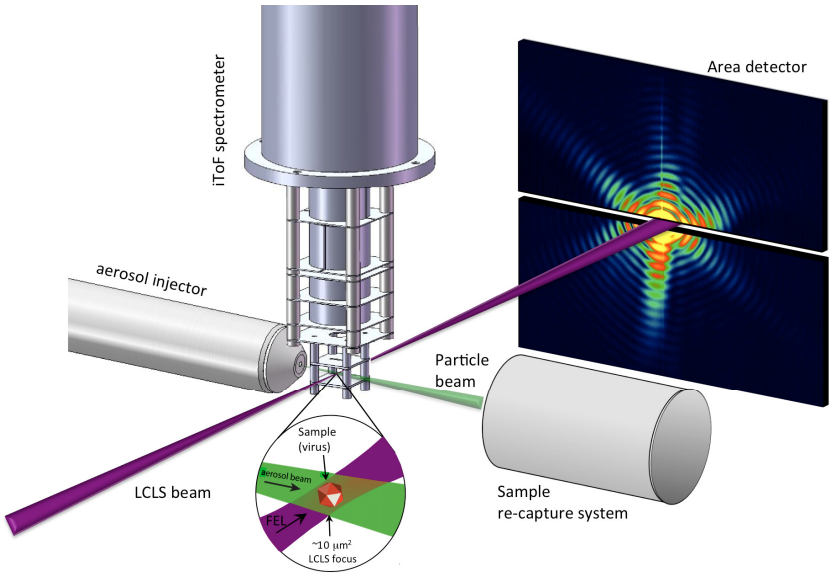

Fig. 1. Femtosecond coherent imaging using aerosol sample delivery technology. Sample is delivered into the vacuum environment of the experiment using an aerodynamic lens aerosol injector. A fast detector reads out after each pulse and the strength of X-ray scattering recorded on the detector depends primarily on incident X-ray pulse energy density, particle size (scattering strength), detector efficiency, and the particle location within the FEL beam intensity profile. Due to the typically Gaussian distribution of intensity in the focal spot, collection of weak hits is more likely than strong hits. For part of the experiments an ion Time-of-Flight (iToF) spectrometer was used together with the injector and imaging detector. The iToF is aligned such that injected samples are hit between two metal plates (repeller and extractor) and an electric field between the plates sends the positively charged ions towards a multi-channel plate (MCP) detector. The iToF is mounted on a motorized translation stage and can be moved out of the interaction region when not used. The instrument is also equipped with x and y deflection plates and an Einzel lens, which were kept at 0 V in the present experiments.

The most obvious method for finding frames in which the FEL beam happened to intercept a particle is to look for the presence of X-ray scattering on the detectors. In principle

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

this should be an easy task, and finding the strongest hits based on scattered photons is indeed comparatively trivial. However, in practice particles of interest are very small, thus the X-ray scattering from individual particles can be very weak, with theoretical calculations suggesting that frames with as few as 100 scattered photons may be useful for structural determination [9,10]. Furthermore, the focal spot intensity distribution typically has long low-intensity tails, and particles are much more likely to be intercepted by weak parts of the FEL beam rather than the most intense portion of the beam. As a result weak hits that may still be useful for structure determination [9] will make up the bulk of the acquired data. At the other end of the spectrum, diffraction patterns for the strongest hits are frequently affected by saturation due to limitations in dynamic range of available detectors, causing large regions of missing data. Finding *useful hits* thus becomes an exercise in identifying relatively weak particle scattering signals above experimental background noise, especially when averaging of many weak hits can be used to build up signal from many very weak data frames.

Careful subtraction of experimental background signals is essential for photon-based hit finding, and is complicated by factors such as the presence of X-ray scattering from apertures and beamline optics, X-ray scattering from water jets or particle carrier gas, shot-to-shot variability in the X-ray beam, and detector properties that slowly drift over time. We exploit the blank frames interleaved between hits, and the fact that hits can be sparsely distributed through data frames, to provide a running estimate of background signal in the data. A simple procedure for obtaining current background estimate is to calculate the pixel-wise median through a ring buffer of depth *n* populated only with non-hit frames. A schematic is shown in Fig. 2. Although somewhat computationally intensive, a pixel-wise median is used to reduce the effect of outlier bright pixels and cumulative effects of very weak hits. This running background serves as an up-to-date estimate of X-ray scattering from beamline optics and residual background gas, as well as any slow drifts in detector offsets during the *n* non-hits immediately prior to the current frame. Hot pixels are identified as any pixels with abnormally high signal in more than 80% of frames in the buffer and are excluded from subsequent analysis.

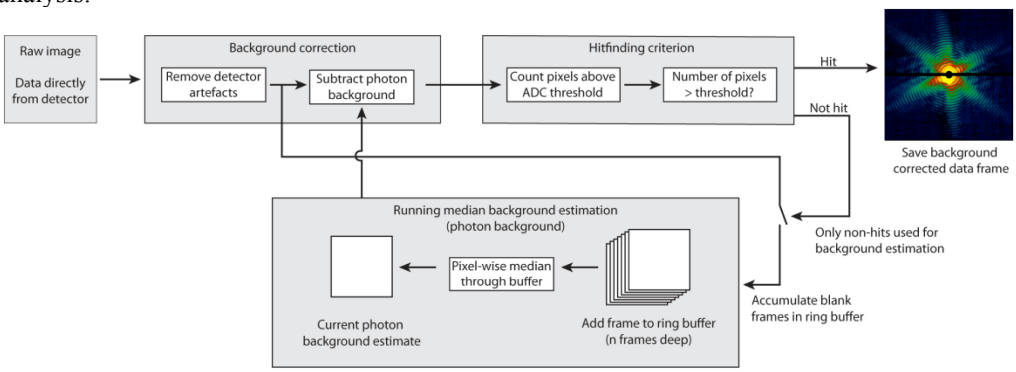

Fig. 2. Moving median background correction. Frames identified as non-hits are added to a ring buffer of n images depth. A pixel-wise median through this rung buffer provides an estimate of the current photon background signal, where a median filter is used to reduce bias due to outlier frames.

After background subtraction, identification of hits is performed based on scattered signal above threshold. Counting pixels containing more than *n* photons (i.e.: above a constant threshold after background subtraction) is observed to be a more reliable discriminator of weak hits than total integrated image intensity, and has been applied in several single particle imaging experiments at LCLS.

In recent experiments, Mimivirus (Acanthamoeba polyphaga mimivirus) particles [11] were injected into the FEL interaction region at the Atomic, Molecular Optical Science (AMO) beamline [12] at LCLS using an aerosol injector [7]. At a photon energy of 2 keV and

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

the pulse duration about 50 fs, diffraction patterns were collected using a pair of pnCCD detectors mounted in the CFEL-ASG Multi-purpose (CAMP) instrument [13]. The in-focus beam diameter was approximately 10 μm^2 giving a peak intensity on the sample of up to 10^{17} W/cm^2 for a perfect hit (virus particle in the center of the LCLS pulse). Of 8,474,596 data frames collected during the beamtime, 840,241 frames are identified as having hits. The hit rate varies significantly through the experiment (Fig. 3), being close to zero during initial setup and alignment, peaking at over 40% in best cases, and averaging just under 10% over the entire beamtime. Over 90% of data collected during the beamtime is identified as blanks, enabling over 54 TB of raw data to be rapidly reduced to a more manageable 5TB of processed data. False positives passing through this initial filter are weeded out in subsequent analysis (section 3 below). Importantly for processing large volumes of data, this classification step is very efficient and can be executed at over 60 frames per second today.

Sorting the observed hits according to scattering strength shows that weak hits are much more prevalent than strong hits: 99% of images contain less than 10% of the integrated intensity contained in the strongest hits (Fig. 4). This is to be expected when the aerosol beam distribution is significantly larger than the focal spot distribution: particles are intersected at random locations in space, so the distribution of intensities should fall off according to the focal spot intensity distribution.

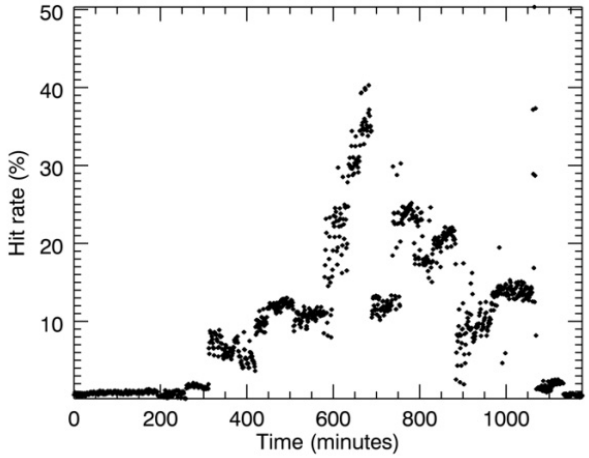

Fig. 3. Hit rate as a function of time in a typical aerosol injection experiment using the Uppsala injector in the CAMP instrument on the AMO beamline at LCLS. After initial alignment, hit rates peak at ~40% of data frames with an average of 10% hits over the course of the entire experiment, which includes initial alignment. Fluctuations in hit rate are expected as experimental conditions are changed.

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

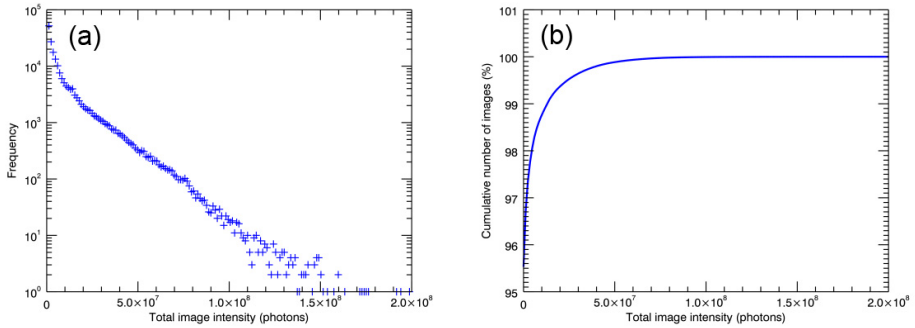

Fig. 4. Frequency of hits and cumulative image count as a function of total scattered intensity. Weak hits are significantly more likely to occur than strong hits.

3. Identifying hits using an ion Time-of-Flight spectrometer

Analyzing scattered photon signals is a very direct method for finding the strongest hits, but requires reading out the detector in order to perform analysis. At high frame rates even reading data off the detector becomes a frame rate-limiting factor, motivating the use of secondary diagnostics to veto frames before they are read out. One possible approach is to study the ion signal produced by disintegration of the sample [2,14,15] for locating and identifying particle hits.

When particles are struck by the intense XFEL beam, they not only scatter photons but also fragment in a coulomb or hydrodynamic explosion [4]. During the sample explosion, ions and electrons are ejected from the interaction region together with an emission of plasma radiation for larger samples. In principle, these processes could be detected and used to identify hits independently of scattered photon signal, providing an alternative veto signal for deletion of data frames not yet read out from the detector or not yet saved to disk. To investigate this possibility, we installed an ion Time-of-Flight (iToF) mass spectrometer (MS) in the sample chamber to measure the fragments from exploding single particles and the corresponding diffraction patterns simultaneously (Fig. 1).

This instrument is a customization of a linear ToF MS supplied by Jordan TOF Products, Inc [16]. Our customization is in the ion extraction region where the plate configuration has been altered to comply with the requirements of the imaging application. Due to a divergence in the particle beam exiting the injector the tip of the sample injector should approach close to the FEL to maximize the hit rate in the single particle imaging experiment. Furthermore, detection of useful diffraction patterns requires no obstacles in the path of the scattered photons towards the detectors. This means that the ion extraction plates must be relatively small and with a large spacing. The iToF accelerates the positive ions towards the detector by fields applied between three parallel plates, the repeller, extractor and ground [17]. All plates are squares with the repeller and extractor having 20 mm sides while the ground plate has 38 mm sides. The distance between the repeller and extractor plates (between which the interaction takes place) was 8 mm and the distance between the extractor and ground plates was 12.5 mm. The applied voltages on the repeller and extractor plates were 2500 V and 500 V respectively. The injector was kept 2 mm from the sides of the repeller and extractor plates to avoid significant effects from the grounded injector tip on the accelerating fields. This results in a distance between the injector tip and the FEL beam of about 12 mm compared to the 3 mm distance normally used. We suggest that this increase is the main cause of the drop in hit-rate by a factor of four observed when introducing the iToF. The extractor and ground plates have 1 mm x 10 mm slits aligned along the particle beam. This reduces the amount of background signal caused by ionization of the residual gas across the long Rayleigh length of the LCLS beam. Behind the ground plate is an Einzel lens and two sets of

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

deflector plates. In the present experiments these were all kept at 0 V and make up the beginning of a 700 mm long grounded linear flight tube that terminates on a multi channel plate (MCP) detector (triple plate, Z-gap). In the experiments, each spectrum was 20 μs in duration, corresponding to a maximum detectable mass to charge ration (m/q) of about 150. Despite this we observed no significant amount of complex fragments above 40 m/q. Finally, the iToF is adapted to a motorized translation stage for alignment to the interaction region. This stage also allows the iToF to be retracted when not in use to allow the injector to be repositioned close to the FEL beam.

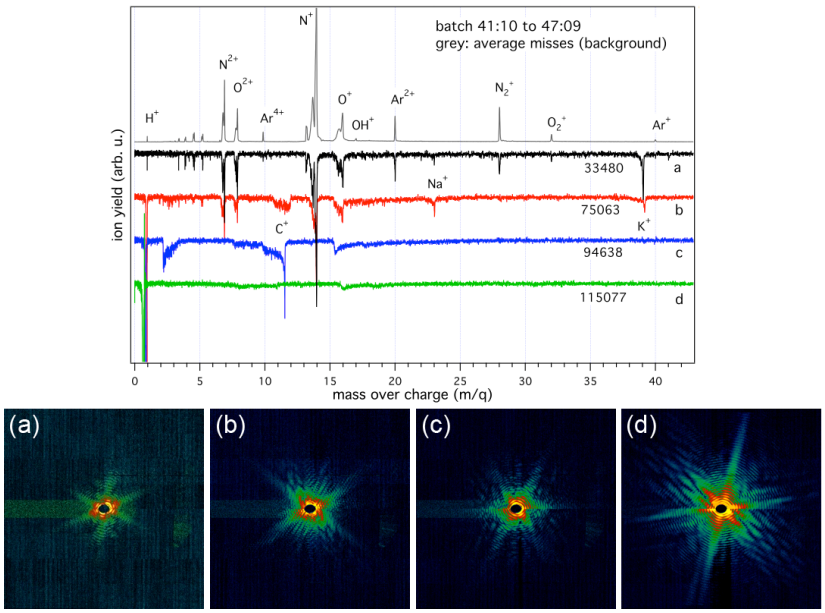

Fig. 5. Simultaneous diffraction and iToF measurements. (Top) Averaged background spectrum (gray) and single-shot iToF spectra with increasing incident intensity on the sample (a-d). The single shot spectra are inverted and individually shifted for clarity. For a low intensity hit (a) a potassium (K^+) peak can be observed, followed by a significant increase in the proton (H^+) peak (b-d). (Bottom) Simultaneous photon scattering as measured on the CCD detector for each of the spectra (a)-(d) respectively.

Figure 5 shows four single shot iToF spectra along with the simultaneously obtained diffraction patterns and an averaged background spectrum from the aerosol carrier gas (a mix of helium and air). The background spectrum is obtained from ionization of carrier gas by the focused LCLS pulses in frames where no photon scattering was detected above background levels. The average background spectrum shows characteristic spectral peaks from the ionized carrier gas. It is dominated by ionic and molecular nitrogen and oxygen signals (including high charge states) with an additional contribution from water (OH and protons) and trace elements (most prominently argon that is easily ionized at this wavelength).

Inspecting the correlation between the iToF and diffraction data enables us to identify two complementary signs of a hit: (i) the presence of new peaks (e.g. C^+, Na^+, K^+) and (ii) an increase and broadening of the proton signal. When the FEL beam intercepts a mimivirus particle, the resulting single shot iToF spectrum contains peaks from ions present in either the virus particle or the buffer solution. This is exemplified in Fig. 5 by the appearance of C, Na and K signal at m/q = 12, 23 and 39 respectively and an accumulation between m/q = 2 and 3 of what likely is highly charged ions accelerated by plasma effects. Significant X-ray driven acceleration of light ions has been observed from solid density samples [18] earlier and was used for an accurate determination of the beam focused beyond the Rayleigh length [19].

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

When sorted according to the strength of the X-ray signal as measured on the CCD detector, peaks can be seen to evolve according to increasing intensity on the particle (Fig. 5). We observe that a small proton signal and a significant K^+ peak characterize weak hits whereas the proton signal dominates for strong hits. This suggests that a combined analysis of the proton and K^+ peaks may be used for hit detection. Total proton signal is successful at identifying frames with high photon-count hits, but misses the low photon count data. Meanwhile, the integrated K^+ signal successfully identifies most low and medium photon-count hits but misses the high photon-count hits.

Figure 6 shows the magnitude of the H^+ and K^+ peaks against hit finding performed based on the scattered photon signal. A clear cluster of iToF signal associated with blank frames (blue crosses) is located in a region of parameter space distinct from frames with photon scattering (red circles). Of 84,190 events collected with the iToF, photon diagnostics identified 439 hits and 83751 blank frames, for an overall hit rate of 0.5% in this particular run. Blank frames (blue crosses) are seen to cluster clearly into a portion of the iToF signal with low H^+ and K^+ signal, delineated by the box drawn in Fig. 6, whilst hits (red circles) generally fall outside this region. In the present experiments, a hit finder based on both proton and K^+ signal strength captures 85% of scattering events identified by photon diagnostics. Only 10 events lying outside of this region were identified as blanks using photon diagnostics – in other words the iToF is able to correctly reject 99.99% of blank frames (a false positive rate of roughly 0.01%).

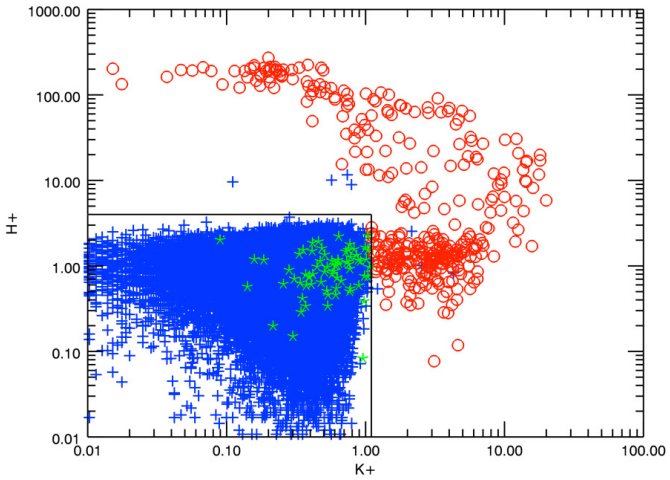

Fig. 6. Correlation of iToF signal against hits found using photon scattering. Red circles agree with hits found by elevated photon scattering, whilst blue crosses are blank frames. A clear cluster of signal with low H^+ and K^+ signal can be seen corresponding to blank frames. The iToF is able to correctly reject 99.99% of blank frames using a dual threshold based on H^+ and K^+ signal. Green stars are false negatives: events that produced photon scattering but had an iToF signal consistent with a non-hit.

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

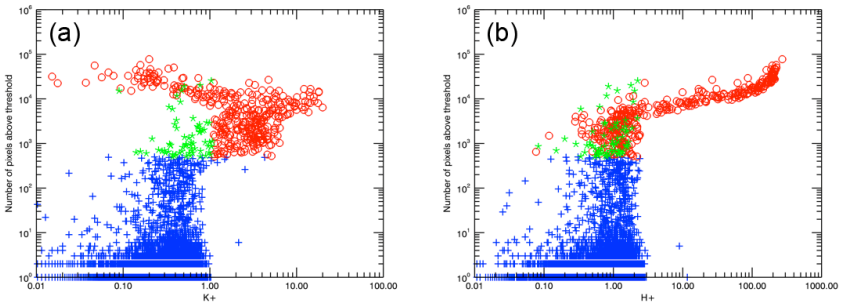

Fig. 7. Scattered photon signal plotted as a function of H + and K + signal separately. False negatives (green stars) do not necessarily correspond to weak hits, and in some cases contain appreciable photon scattering. We hypothesize that such frames correspond to particles intercepted by the X-ray beam outside the narrow observation region of the iToF.

The cases where scattered photons were observed on the detector even though the iToF did not produce signal corresponding to a hit are identified by green stars in Figs. 6 and 7. Specifically, 70 frames out of a total of 439 hits identified as hits using photon diagnostics fell within the region of H^+ and K^+ parameter space associated with blank shots, corresponding to a false negative rate of ~15%. Although it may be tempting to presume these false negatives correspond to very weak hits, this turns out not to be the case. Plotting both the H^+ and K^+ signal separately against the scattered photon signal (Fig. 7), we see that some of these hits identified as blanks by the iToF indeed contain significant photon scattering (Fig. 8). To limit the back-ground signal from residual gas ionized within the Rayleigh length of the FEL the field of view of the iToF is restricted by a 1x10 mm entrance slit in the iToF extractor plate. We hypothesize that hits missed by the iToF (the false negatives) correspond to particles intersected by the FEL beam outside this field of view and observe that increasing the acceptance angle could decrease the number of false negatives at the price of an increasing background.

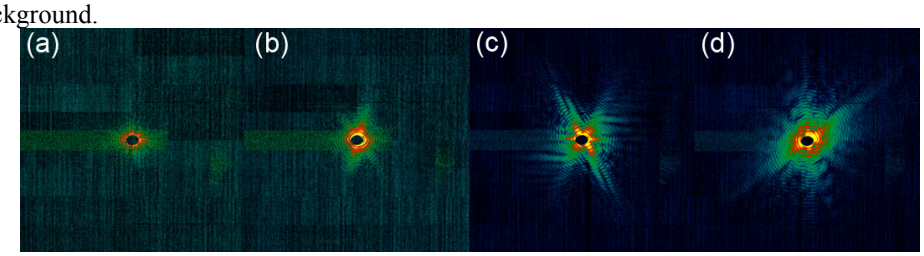

Fig. 8. False negatives identified by the iToF hit finder containing appreciable X-ray scattering (random selection of 4 frames from the green stars in Figs. 6 and 7).

Since the iToF provides an independent hit-finder and can be evaluated relatively quickly it can be utilized for vetoing data read out from a photon detector array. Optimizing the detector for proton detection, decreasing the length of the flight tube, and increasing the accelerating voltage should make it possible to perform hit detection based on the proton signal within the 200 ns pulse separation of the European XFEL. The negative impact on the hit-rate from the introduction of the iToF can be decreased by further reducing the size of the iToF plates to allow the injector to be brought closer to the FEL beam while the iToF is used. For cases where detection times considerably shorter than 200 ns are needed, hit detection could be performed using electron ToF spectroscopy or even plasma emission detection.

4. Automated sorting of diffraction events by particle size

We now turn our attention to the task of identifying diffraction patterns from inhomogeneous samples. In addition to hits on single particles, there will be scattering from multiple particle hits and contaminant material. In less than optimal circumstances, the injection system may be

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

contaminated with sample from the previous run. All of these samples can give rise to photon scattering events. For three-dimensional imaging of reproducible samples, it is crucial to sort the ensemble of diffraction patterns to identify single particle hits of the desired species. We have found that the simple task of sizing particles based on their autocorrelation function can be used to reject most of the outliers in a data set.

Particle size can be measured from the autocorrelation function of the object, obtained by taking the inverse Fourier transform of the measured diffraction pattern. The domain covered by the autocorrelation function is twice as wide as that covered by the object and a measurement of autocorrelation size can thus be used to determine the particle size.

In practice, the direct XFEL beam is not measured as it would harm the detector. A gap in the detector allows the direct beam to pass through, so there is unmeasured data at low scattering angles. The missing data region from the detector acts as a high pass filter, which introduces fringes and artifacts on the autocorrelation function. Using the detector gap, a mask $M(\boldsymbol{q})$ can be defined for the regions where the data was measured. The term $\boldsymbol{q}$ denotes a vector in the plane of the detector and $\boldsymbol{r}$ denotes a vector in the plane of the object. The Fourier transform of this mask defines a point spread function $P(\boldsymbol{r}) = F[M(\boldsymbol{q})]$, where F denotes the Fourier transform. The autocorrelation calculated from the data $A_{calc}(r)$ is related to the autocorrelation function of the object $A_{obj}(r)$ by

$$A_{calc}(\boldsymbol{r}) = A_{obj}(\boldsymbol{r}) \otimes P(\boldsymbol{r}). \quad (1)$$

An example of the autocorrelation function, calculated from directly Fourier transforming the diffraction pattern taken on the pnCCD detectors, is given in Fig. 9. Since there is missing data, we cannot deconvolve $P(\boldsymbol{r})$ in Eq. (1) to obtain $A_{obj}(\boldsymbol{r})$. However, we can apply further operations to the diffraction data in order to obtain a function with sharp signal at the edge of the object autocorrelation function as follows:

i) Apply a broader high pass filter, such that the point spread function becomes narrower. The spatial extent of the point spread function can be further reduced by defining a smooth edge to filter with high-order exponential function, i.e. the filter can be defined as $1 - \exp(-(r/\sigma)^{16})$, where σ is the width of the filter in pixels. The width can be dynamically set according to the extent of scattering in the pattern, which we set to 40% of the highest measured q value. For weak scattering patterns, the minimum width can be specified, which in our case is 60 pixels.

ii) Scale the intensity by $I^{0.1}$ to accentuate the contribution to high frequencies.

The resulting high-pass-filtered autocorrelation functions have a strong signal at the edges of the object autocorrelation function, as shown by the examples in Fig. 10. Figures 10(a) and 10(b) show single mimivirus hits, while Fig. 10(c) shows a larger particle.

To identify the object size, a mask is created which specifies the domain of the object autocorrelation function. The point in the mask that is furthest from the centre is used to measure the size. A vertical and horizontal stripe was excluded from the analysis because of strong features from the detector gaps. The mask is defined by an iterative statistical separation of signal and noise. On the first iteration, the mean (μ) and standard deviation (σ) of the autocorrelation are calculated, and all points above 2.5σ are treated as signal. Then μ and σ are recalculated from all the pixels identified as noise, and the threshold is applied again. This can be iterated multiple times, however, it was found that two passes were sufficient. Outlying noisy pixels that had values above the threshold were removed from the mask by binary morphological operations. A binary-closing operation was applied, using a 4 pixel structure in the shape of a 'T', followed by a binary opening operation with a 5-pixel structure in shape of a ' + '. These structures were chosen to remove single isolated pixels above the statistical threshold. Some examples of the resulting masks are shown in Fig. 10. There is good coverage of the key intense features of the autocorrelation function.

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

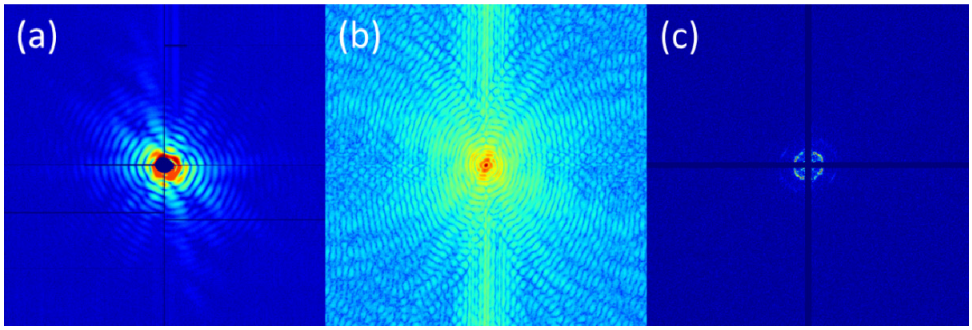

Fig. 9. (a) A single-shot diffraction pattern of a mimivirus. (b) The autocorrelation function formed by directly Fourier transforming (a). (c) The autocorrelation after high-pass filtering and scaling to the power of 0.1.

In the case of strong hits, the noise of the autocorrelation function was correlated by the filters and missing data region. This led to an overestimation of the size for strong single hits. To counter this effect, the threshold was raised to 4σ for $I > 10^7$ adu (analog to digital units) and 6σ for $I>10^8$ adu. We expect that this threshold variation is dependent on the properties of the detector, and not on the sample.

Measuring particle size is a potentially effective technique for determining whether a hit contains diffraction from only a single particle or diffraction from multiple particles. Multiple particles will either be aggregated or separated. Aerosol injection can produce a significant number of aggregated groups if the particle concentration is high relative to the droplet size. In both cases, a size measurement using the method described above should return a larger value than the expected size. Particles separated in a direction transverse to the beam axis are easiest to identify, because the "size" measurement will actually be the sum of the single particle size and the inter-particle separation distance. For aggregated particles, the size should indicate the size of the aggregated cluster. This will be larger than the single particle size in almost all cases except, for example, when two particles are aligned along the beam axis. However, it was found that the size of multiple particle hits (not aligned with the beam axis) was often estimated close to the single-particle size. The problem arose because multiple particle hits contain a strong feature in the autocorrelation function at the single-particle size. If the statistical threshold is too high, then the size of aggregates is often incorrect. However, lowering the threshold tends to overestimate each size measurement. This effect can be seen in Fig. 11, which was generated with a threshold designed to accurately identify single particle hits. The maximum of the peak is located at 520 nm, which is larger than the actual mimivirus size of 450 nm (without hair).

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

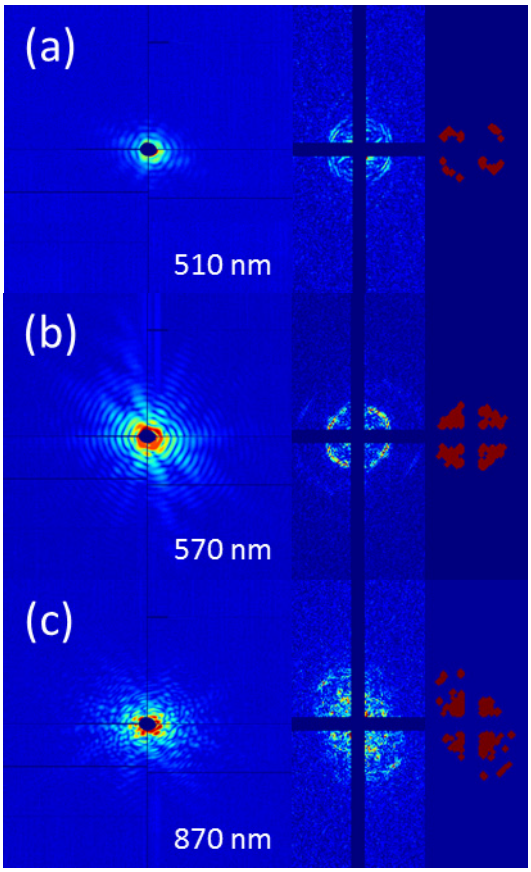

Fig. 10. Single-shot diffraction patterns alongside a high pass filtered autocorrelation function, and the mask (red) generated from the autocorrelation function to determine size: (a) a single mimivirus with a weak hit, (b) a strong single mimivirus hit, (c) a large cluster.

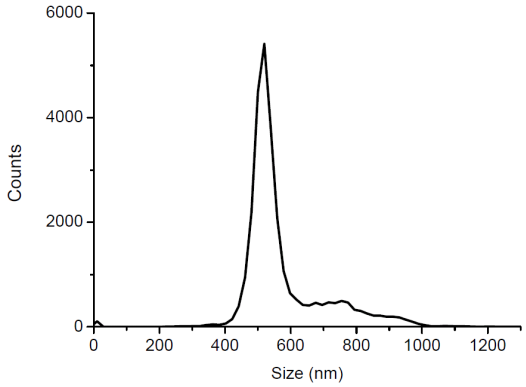

Fig. 11. Calculated size distribution from ~23,000 detected hits in a run of mimivirus. Note the strong peak corresponding to the approximate diameter of single mimivirus particles. Although the sizing procedure inherently leads to a systematic overestimation of the particle size, the size distribution is still useful for identifying single particle hits together with additional components present in the particle beam.

#200920 - $15.00 USD Received 11 Nov 2013; revised 23 Dec 2013; accepted 7 Jan 2014; published 29 Jan 2014

5. Conclusions

We have described a set of criteria for rapidly and efficiently sorting through millions of diffraction patterns collected in single-particle coherent diffractive imaging experiments. The criteria described here enable us to identify frames containing useful diffraction data, and to perform a first pass at sorting the data into different classes. We conclude that an iToF-based hit finder can effectively identify particle hits. Hit finding based on analysis of the proton and potassium peaks from mimivirus particles finds over 85% of the hits that can be identified by analyzing scattered photons hitting the CCD detectors. Further, this method is extremely good in weeding out false positive hits (only 0.01% of all hits identified by the iToF turned out to be false positives). With increased efficiency in proton detection the iToF diagnostics could be particularly useful in identifying weak hits not readily determined using photon diagnostics alone. Optimizing the iToF geometry and fields should enable a response time within the 200 ns pulse separation of the European XFEL.

Subsequent frame analysis by particle efficiently rejects outliers from the homogenous particle class desired for single particle imaging, thus reducing data volumes by another order of magnitude. Data vetoed by these hit finding criteria can be used as the input to more computationally intensive sorting algorithms [20] that are not yet fast enough for processing data in near-real time.

The initial screening steps are simple, fast to execute, can run at over 60 frames per second on current off-the-shelf hardware, and could in principle be readily scaled to pulse rates well in excess of 120 Hz, providing a viable path to real-time frame sorting. The speed and accuracy of the screening algorithms is critical in dealing with the large volumes of data generated by single particle imaging experiments. Extension of the methods described here to real-time frame rejection will be highly advantageous at future sources such as the European XFEL that promise in excess of 10^8 pulses per hour.

Acknowledgments

This work was supported by the following agencies: The Swedish Research Council, the Knut and Alice Wallenberg Foundation, the European Research Council, the Rontgen-Angstrom Cluster, the Helmholtz Association, the DFG Cluster of Excellence at the Munich Centre for Advanced Photonics (MAP), the Max Planck Society in the development and operation of the CAMP instrument within the ASG at CFEL; the Joachim Herz Stiftung; the German Federal Ministry of Education and Research project 05K2012. The ion ToF spectrometer was built and implemented through a grant from Stiftelsen Olle Engkvist Byggmästare. Portions of this research were carried out at the Linac Coherent Light Source, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. We are grateful to the scientific and technical staff of the LCLS for their support and acknowledge Daniel Westphal for his contribution to the development of the instruments used in this research. The mimivirus sample was provided by C. Abergel, V. Seltzer and J.-M. Claverie at the Structural and Genomic Information Laboratory, Aix-Marseille Université, France.