



CHALMERS
UNIVERSITY OF TECHNOLOGY

A statistical approach to detect protein complexes at X-ray free electron laser facilities

Downloaded from: <https://research.chalmers.se>, 2026-04-18 01:32 UTC

Citation for the original published paper (version of record):

Pietrini, A., Bielecki, J., Timneanu, N. et al (2018). A statistical approach to detect protein complexes at X-ray free electron laser facilities. *Communications Physics*, 1.
<http://dx.doi.org/10.1038/s42005-018-0092-6>




N.B. When citing this work, cite the original published paper.

ARTICLE

DOI: 10.1038/s42005-018-0092-6

OPEN

A statistical approach to detect protein complexes at X-ray free electron laser facilities

Alberto Pietrini ¹, Johan Bielecki^{1,2}, Nicusor Timneanu^{1,3}, Max F. Hantke^{1,4}, Jakob Andreasson^{1,5,6}, N. Duane Loh^{7,8,9}, Daniel S.D. Larsson ¹, Sébastien Boutet¹⁰, Janos Hajdu^{1,5}, Filipe R.N.C. Maia ¹ & Carl Nettelblad ^{1,11}

The Flash X-ray Imaging (FXI) technique, under development at X-ray free electron lasers (XFEL), aims to achieve structure determination based on diffraction from individual macromolecular complexes. We report an FXI study on the first protein complex—RNA polymerase II—ever injected at an XFEL. A successful 3D reconstruction requires a high number of observations of the sample in various orientations. The measured diffraction signal for many shots can be comparable to background. Here we present a robust and highly sensitive hit-identification method based on automated modeling of beamline background through photon statistics. It can operate at controlled false positive hit-rate of 3×10^{-5} . We demonstrate its power in determining particle hits and validate our findings against an independent hit-identification approach based on ion time-of-flight spectra. We also validate the advantages of our method over simpler hit-identification schemes via tests on other samples and using computer simulations, showing a doubled hit-identification power.

¹Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3Box 596SE-751 24 Uppsala, Sweden.

²European XFEL GmbH, Holzkoppel 4, 22869 Schenefeld, Germany. ³Department of Physics and Astronomy, Uppsala University, Box 516SE-751 20, Uppsala, Sweden. ⁴Chemistry Research Laboratory, Department of Chemistry, Oxford University, 12 Mansfield Rd, Oxford OX1 3TA, UK. ⁵ELI Beamlines, Institute of Physics, Czech Academy of Science, Na Slovance 2, CZ-182 21, Prague, Czech Republic. ⁶Condensed Matter Physics, Department of Physics, Chalmers University of Technology, Gothenburg, Sweden. ⁷Centre for Bio-imaging Sciences, National University of Singapore, 117557 Singapore, Singapore.

⁸Department of Physics, National University of Singapore, 117551 Singapore, Singapore. ⁹Department of Biological Sciences, National University of Singapore, 117557 Singapore, Singapore. ¹⁰Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

¹¹Division of Scientific Computing, Department of Information Technology, Science for Life Laboratory, Uppsala University, Lägerhyddsvägen 2(Box 337)SE-751 05 Uppsala, Sweden. Correspondence and requests for materials should be addressed to C.N. (email: carl.nettelblad@it.uu.se)

To date, diffraction experiments have been one of the major techniques used in determining molecular structures up to high resolution, by shining X-rays on stable, static, and highly repetitive crystalline systems. Final results constitute an average over all occurrences of the crystal unit and also rely on proper phasing being achievable from the identified Bragg peaks¹. Unfortunately, many samples of biological interest do not lend themselves to crystallization^{2,3}. Thanks to the advent of X-ray free-electron lasers (XFELs), which provide very intense and ultrashort pulses with a brilliance 10^9 times greater than any other existing source and pulse lengths of ~ 50 fs, diffraction experiments on non-crystalline samples are becoming achievable. Over the last few years, successful reconstructions in two dimensions (2D) and three dimensions (3D) have been achieved for various larger biological samples (~ 50 – 500 nm in diameter), such as viruses and cell organelles^{4–8}.

In an XFEL imaging experiment, the sample can be delivered fixed on a solid substrate⁹ or in form of aerosols¹⁰. Aerosolisation has the advantage of removing background diffraction caused by the substrate; however, the position of the aerosolized sample in the X-ray beam is not well determined and the exact orientation and position of a particle when an X-ray pulse hits is generally purely stochastic. For a highly focused X-ray beam, intended to give the highest strength of the diffraction signal, the probability of hitting a particle thus also decreases. Thus, even though the typical problems associated with crystallization can be overcome, other issues can arise, such as the hit-finding problem that consists in discerning sample hits of any kind from background events. The background consists of signal occurring from beamline and optical components, scattering from remaining traces of sample buffer and injection or background gases, and detector electronic noise.

The main topic of this paper is a statistical hit-finder approach, based on the comparison of recorded individual patterns to a typical background signal, generally automatically extracted from the very same data set. Such a technique is relevant for any photon-counting pixel-based experimental modality with sporadic presence of sample signal on top of a background of comparable strength. For example, this would include imaging experiments using high-harmonic generation light sources, or attempts at short-pulsed operation at synchrotrons. Depending on the purity of the sample preparation and the reliability and tuning of the injection method, these identified hits then need to be classified to identify a subset of hits that actually correspond to single pristine particles.

In our experiments performed at the Coherent X-ray Imaging (CXI) end station⁹ at the Linac Coherent Light Source (LCLS), the sample is injected into a vacuum chamber, suspended in aerosol droplets. When these droplets interact with the X-ray beam, diffraction before destruction¹¹ principle holds and the scattered signal is collected by the photon detectors at CXI (Fig. 1). The principle states that femtosecond X-ray pulses can outrun sample damage, as single pulses are so short that they end before any damage to the sample occurs^{11,12}. This was demonstrated at a free-electron laser in Hamburg and at the LCLS on fixed targets^{12–14} and laid the foundations for serial X-ray femtocrystallography (SFX)¹², where sub-nanometer reconstructions are now successfully achieved routinely. SFX allows the study of several classes of samples that do not readily form large crystals, but where smaller crystals (at nanometer scales) can still be obtained. However, there is still a significant set of proteins and macromolecular complexes that cannot be studied even with this technique, but where FXI might prove successful.

The idealized view of our measurement assumes that the X-ray pulses hit only the sample macromolecules, as pure buffer droplets are expected to evaporate before reaching the X-ray

interaction region⁴. However, a variety of different cases can take place, which produce signal on the detector or even diffraction patterns that may or may not be representative of the sample of interest: buffer droplet or small impurities (quite common), impurities clusters, sample clusters, or single-sample molecules.

The two main factors affecting the detection capability are the nature of the sample, and the nature of the background. Smaller samples with weaker total scattering cross-section will be harder to detect, whereas a stable background signal will simplify detection.

There have been multiple techniques used for doing hit-finding in both SFX and FXI. These include methods based on ion time-of-flight spectra and plasma emission in serial femtosecond crystallography^{15,16} and photon-based flat-threshold hit-finders for virus particles, cells and organelles in FXI experiments^{4,17–19}. To date, no diffraction from a single-protein complex has been confirmed. We have developed a novel statistical method, based on the analysis of beamline background and relying on photon-counting statistics, which is able to discriminate sample hits from background in the recorded diffraction patterns, even when noise and sample signal are comparable. Such detection performance could allow definite validation of single-particle hits, if combined with suitable further analysis (e.g., size and orientation determination of the particle). Modeling the background signal is not only important for the hit-finding process, but also for all later reconstruction and analysis algorithms sensitive to noise, including methods for 3D reconstruction based on multiple diffraction patterns, such as the Expansion-Maximization-Compression (EMC) algorithm²⁰.

In this work, we discuss the analysis of the background and the method itself is introduced, followed by a validation of the method by applying it on FXI scattering data from a macromolecular machinery: RNA polymerase II²¹. It is at this point one of the smallest single particles ever studied at an XFEL. The experimental photon energy was ~ 6 keV, corresponding to ~ 2 nm photon wavelength, the beam pulse was of ~ 50 fs and the nominal focus beam size of 100 nm²². To further corroborate our results, we report a comparison with an independent time-of-flight ion detector (ToF) hit-finder, used in drift mode. Then, we show the efficiency of our statistical approach by testing it on data sets collected on other samples and by computer simulations.

Results

Implementation of the statistical hit-finding approach. Under stable experimental conditions (same injector, nozzle, pressure, buffer concentration, etc.) background has proved not to vary significantly during the same experimental shift (Fig. 2). Hence, we can establish an upper threshold on the amount of variation and denote as hits any event over this threshold. Whatever is below is then called a background event (or a miss).

Most previous hit-finding methods make use of arbitrary thresholds, basing the decision on, for example, the number of downsampled (binned) detector pixels (e.g., one pixel being constituted by 4×4 detector pixels) that are lit^{4,5}. Other thresholds can be defined by considering as hits patterns with a photon count above, e.g., four standard deviations from the mean value in the overall background photon count distribution after applying some arbitrary mask to ignore the least reliable sections of the image. Such a stringent threshold is too coarse to be effective in the case of weak scatterers. Therefore, we introduce a more sophisticated method, which has its foundations in the expected Poisson statistics of the number of photons hitting a pixel. A pixelwise comparison is made between each frame within a sample run, and the expected background. This approach also allows an automatic determination of pixels deviating from the

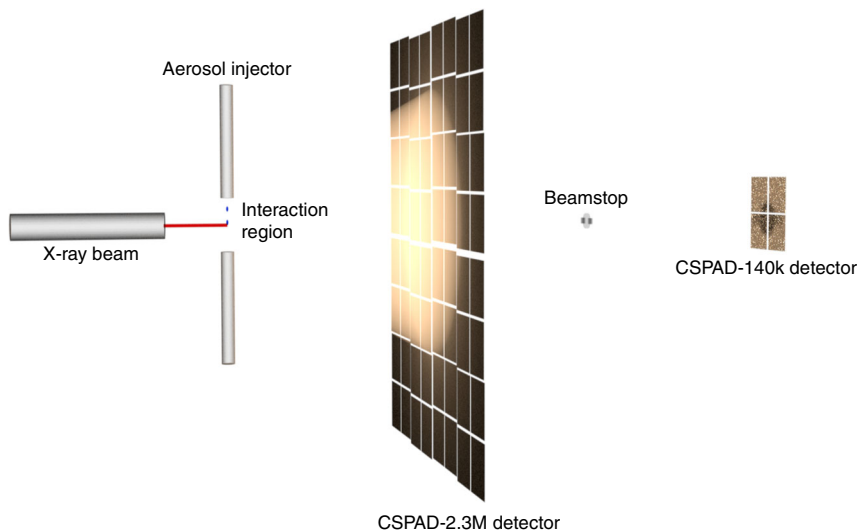


Fig. 1 Typical experiment setup for flash X-ray imaging (FXI) at the coherent X-ray Imaging (CXI) end station. The sample is injected in aerosol form by an injector. When the aerosol reaches the vacuum chamber, it is hit by the beam pulse in the interaction region. As the explosion time ($\sim 10^{-12}$ s) is believed to be longer than the beam pulse ($\sim 10^{-15}$ s), diffraction before destruction is possible: the scattered photons are recorded in the front (high resolution) detector and in the back (low resolution) detector. A beamstop is interposed between the two detectors to avoid the full intensity of the direct beam hitting and damaging the center of the 140,000 pixels Cornell-SLAC Pixel Array Detector (CSPAD-140k)

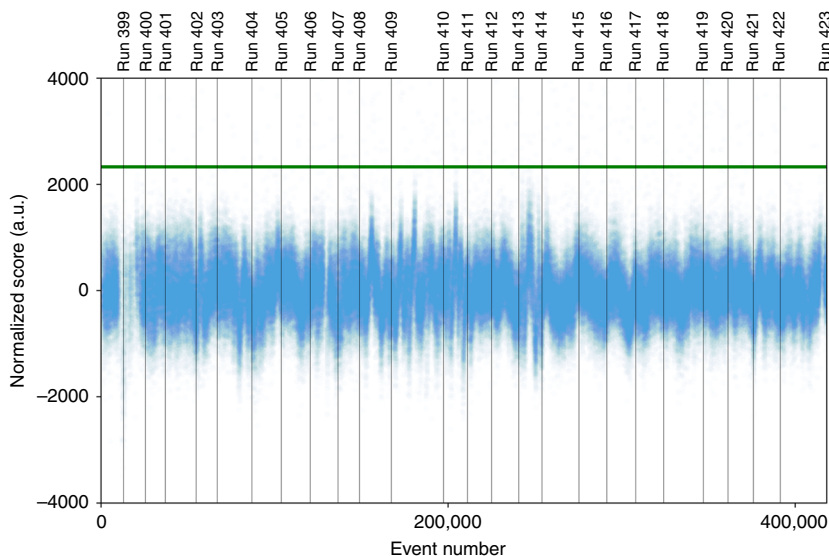


Fig. 2 Sample data as a time series. The vertical lines denote the different runs (from 399 to 423) used in our analysis. Those were collected in an experiment performed in May 2013, whose main object of study was RNA polymerase II. Although some background variations are present, the distribution is far more stable than, e.g., raw photon counts. The green line represents the threshold used to identify hits

model, which are then included in the mask (see Methods). This step guarantees the usage of a wider number of detector pixels—and thus better statistics—compared with more ad-hoc approaches of defining a suitable mask^{4,5,17}.

Our formal null hypothesis is the following: in the case of background patterns each detector pixel in each frame follows Poisson statistics (i.e., each pixel constitutes an observation n_{ki} with expected mean photon count λ_{ki} —see Fig. 3a, b). Furthermore, detector pixels are here considered to be independent from one another, with the exception that they are all a function of the total signal strength. In the low-photon emission regime, the first hypothesis constitutes a well known physical process,²³ the second one depends a lot on detector pixels design,

but it holds reasonably well for the Cornell-SLAC Pixel Array Detectors (CSPAD) used at the CXI end station²⁴.

Both n_{ki} and λ_{ki} are experimentally determined from the data (see Methods).

We calculate a score for each event—by means of multivariate analysis—in the form of a log-likelihood ratio²⁵:

$$s_i = \sum_k n_{ki} \log \frac{n_{ki}}{\lambda_{ki}} \quad (1)$$

Under the two hypotheses stated above, by taking advantage of the central limit theorem, we expect our s_i score distribution to be

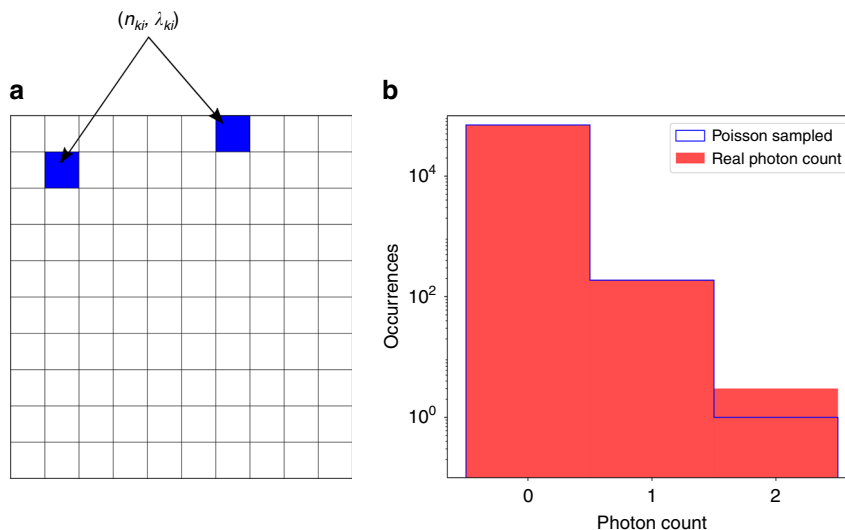


Fig. 3 Model structure based on independent pixels adhering to Poisson statistics. **a** A pixel-based detector and how each pixel in a background frame is provided with the observed photon count n_{ki} and its mean photon count λ_{ki} ; **b** histogram plot (semi-logscale) showing the distribution of a certain pixel in the background run collected during the June 2015 experiment, having an overall mean photon count equal to 0.022 (red), compared with a simulated Poisson distribution (blue), based on the same mean photon count and the same pulse energies per event as in the real case. The agreement is evident (in the case of the two-photon count, the histograms differ only for two occurrences), thus proving the validity of the assumption

normal. Owing to non-ideality in the data, we could not use the theoretical properties of the distribution directly to define reasonable test thresholds. Rather, we noted that the distribution of such scores for a stable background as a function of the expected photon count for each event gives a linear relationship. By fitting the log-likelihood scores, we obtain average expected scores for the same as it was pure background (Fig. 4a). We then subtract these values from our scores, (Fig. 4b), making the resulting distribution independent of the pulse energy. For this transformed distribution, we can express our hit threshold as $\mu + 4\sigma$ (μ being the mean of the new distribution and σ the corresponding standard deviation). The choice of a 4σ threshold ensures a theoretical false positive rate of 3.16×10^5 . When analyzing background data from a June 2015 experiment²⁶, the actual rate of false positives retrieved is consistent with this theoretical result: 2.83×10^5 (Supplementary Note 1 and Supplementary Fig. 1). A photon count threshold can also be similarly defined, by performing normalization in the photon count space (Fig. 4c, d).

Sometimes significant changes were made between experimental runs without recording new background data. Those can affect the reliability of the determination of μ and σ . Therefore, in order to characterize the parameters, we decided to select events in the sample runs that are unlikely to be hits (called “preliminary misses”) by using the photon count distribution as described in Methods. We are essentially using a crude identification method for non-hits to seed the method for identifying proper hits. This filtering ensures that the background statistics is not influenced by a small amount of very strong hits when calculating the rate parameters (λ_{ki}). When hit rates are low ($< 5\%$), some contamination of the background statistics with weak hits should not influence our detection power adversely. High pixel photon counts, on the other hand, can strongly bias the λ_{ki} outcomes and so alter the background model, especially reducing the detection power arising from those areas of the detector where the background signal is very clean. Again, one should also note that when using a narrow focus, like the 100 nm focus in our study, current injection techniques tend to only allow hit rates of single-digit percentages or less. This is in contrast to hit rates of

over 20% in ideal conditions at, e.g., the AMO end station²⁷, with a much wider focus. However, a small concentrated focus and the shorter wavelengths possible at CXI are critical for proper imaging of small samples. Although the filter could be considered somewhat coarse, it also allows an online mode²⁸ for our hit-finding methodology, where the background model would be adjusted in real time while data are being collected.

The plot in Fig. 5a illustrates how to perform hit-identification using this method: the density plot represents all the events collected, darker blue indicating higher density. A threshold (green line) is defined as described above, and the subset of events identified as hits by our statistical approach (blue circle outlines) are reported. The red circles represent the hits found via the independent time-of-flight ion hit-finder.

RNA polymerase II as an application. To give a thorough understanding of how the proposed method works, its application on a specific FXI experiment is reported. The data examined are diffraction patterns of RNA polymerase II collected on the CSPAD-140k detector^{29,30} at the LCLS CXI end station (Fig. 1), during an experiment in May 2013. The sample studied is a molecular machinery involved in DNA transcription²¹. An improved understanding of the in vivo structure and dynamics of this complex could improve the modeling of gene regulation.

The sample buffer consisted of water and ammonium acetate and the sample itself was labeled with gold spherical nanoparticles to increase its scattering power.

As shown in Fig. 5a, when a 4σ outlier threshold is applied, the bulk of the background is clearly separated from hits. We identified 1165 hits of varying strength over a total of 402,296 non-blank events considered over 25 runs (418,153 events in total). Furthermore, 828 hits were identified by another independent hit-finder: a ToF detector. Thus, we could prove that our hits are not spurious by noticing that most hits are shared between the ToF hit-finder (red circles) and our statistical model (blue circle outlines). The total fraction of ToF hits that are above our defined threshold is 94% (771 ToF hits); the remaining 57 belong to background (Fig. 5a). By looking at the average

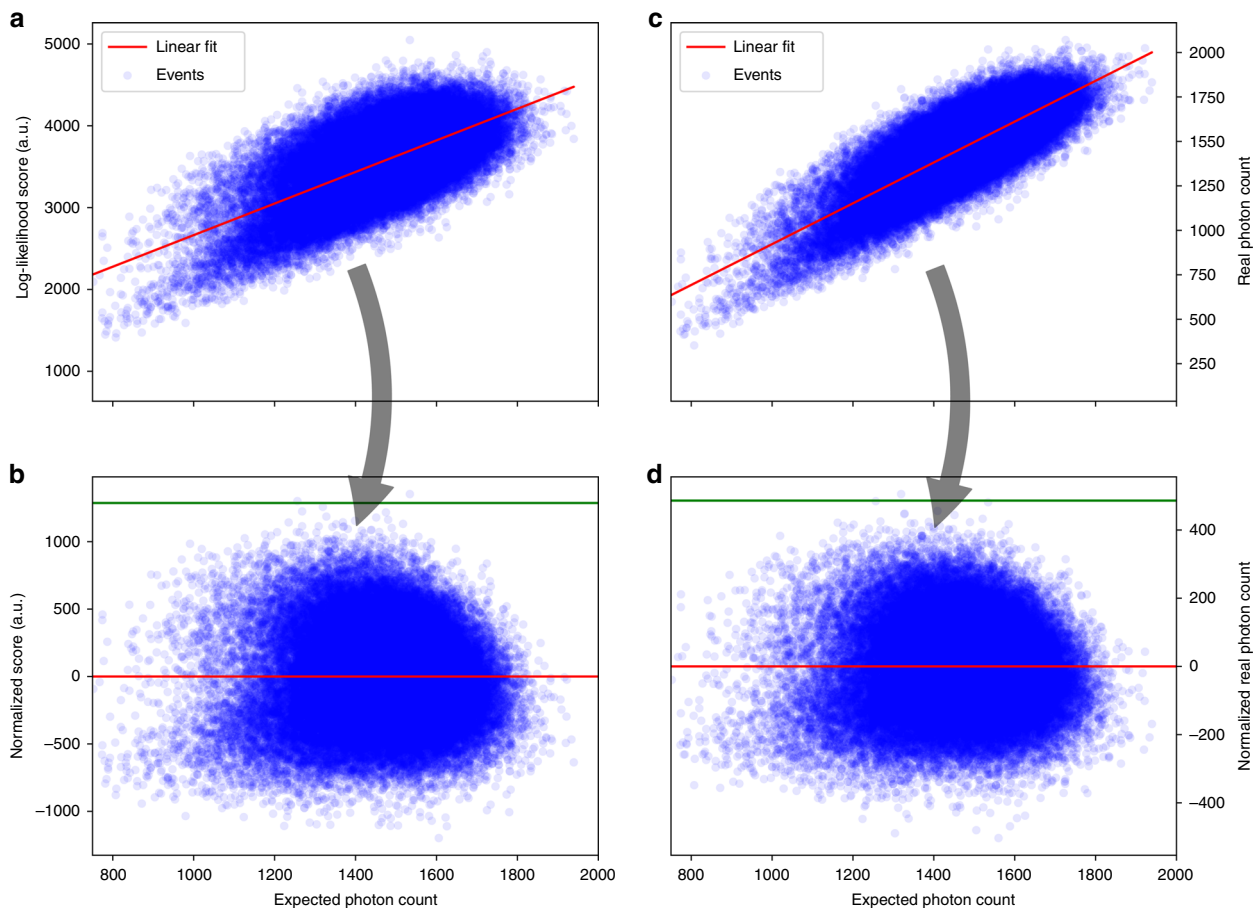


Fig. 4 Log-likelihood score and photon count normalization. **a** and **c** show, respectively, the original log-likelihood scores and real photon counts and their transformed counterparts for a background run (from the June 2015 experiment). Blue circles represent all the events in the data set, whereas the red line is a linear fit of the log-likelihood scores (**a**) and real photon count (**c**) to expected photon counts. **b** and **d** are obtained by subtracting the offset from the data. The resulting normalized distribution is roughly Gaussian and centered at $\mu = 0$, for both **a** and **c**, making a hit-finding threshold of $\mu + 4\sigma$ straightforward to define (green line). The thresholds here shown—1287 and 486—are the ones used to study the hit-finder’s efficiency through computer simulations in the present work. Whereas overall distributions look similar, the score-based method in fact provides tighter bounds

integrated ToF trace of those hits not confirmed using our method (Fig. 5b), as well as inspecting all of them separately (Supplementary Fig. 2), it appears none or few of these are actual hits, as the proton peak, which is the intended hit-finding criterion, is visually absent in the traces. Fifty-seven events out of 402,296 amounts to $\sim 0.013\%$, which is consistent with the expected false positive rate for the ToF method of $\sim 0.01\%$ previously reported¹⁵. One can also note that the distribution of these non-matching ToF hits follows the overall distribution of recorded events in terms of the likelihood scores (Supplementary Fig. 3). If they were instead very weak hits not picked up by our approach, one would expect them to cluster between the main background “cloud” and the threshold.

The fact that the statistical hit-finder recovered a higher number of hits in total is to be expected, as the ToF ion detector was used in drift mode with an experiment geometry that meant it only covered a small portion of the total solid angle. Therefore, only a limited fraction of the ions emanating from sample explosions could be picked up.

None of the hits found—both via our proposed method or ToF hit-finder—can be unambiguously attributed to a single RNA polymerase II complex, but they are hits indicating the presence of organic matter in the interaction region of the X-ray pulse,

based on the combination of the diffraction patterns and ToF ion traces (Supplementary Fig. 4a).

Below follows a more theoretical demonstration of the general validity and reliability of our hit-finding method, where a comparison with an idealized photon count threshold is shown and detection limits are explored.

Statistical hit-finder efficiency on larger samples. As further evidence of the effectiveness of the method, we tested our statistical approach on other datasets collected from larger particles collected at the CXI end station: the Omono River virus (OmRV)—same data set as presented in previous work⁵—and the bacteriophage PR772, the same sample used in an earlier experiment at the AMO end station³¹.

These represented two icosahedral viruses, respectively, of 40 nm and 70 nm in diameter, studied during experiments in April 2014 and April 2016 (Supplementary Note 1). The experimental setup was the same as described for the RNA polymerase experiment; the CSPAD back detectors shared the same revision (v. 1.6) as on May 2013. To reduce the total amount of incident photon flux (and so the scattered background), more aggressive aperturing of the beam was applied, reducing the background scattering significantly, but also decreasing the scattered signal

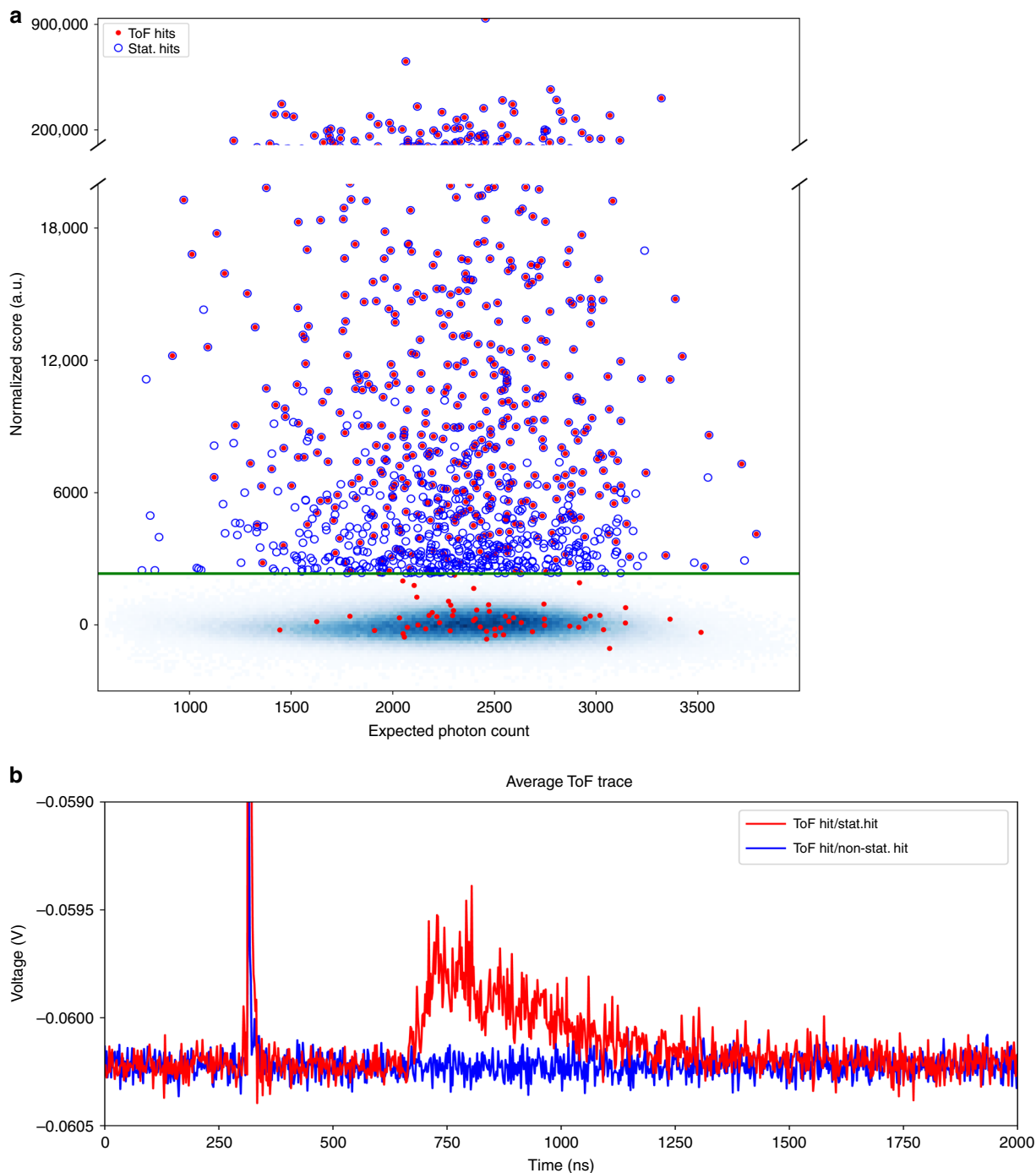


Fig. 5 Comparison of statistical hit-finder against ion based hit identification using a time-of-flight detector (ToF). **a** shows a density plot (darker blue implying higher density) of the normalized score distribution of all shots. The green line is the threshold defined by our statistical hit-finder, and whatever is above is considered a hit (blue circle outlines). ToF hits are also shown (red). **b** shows the average of 56 ToF non-blank traces that are not statistical hits (blue line) and a corresponding number of ToF hits that are statistical hits as well (red line). In the latter, the wide drift-mode proton peak is visible ~ 800 ns

from the samples. OmRV was injected as described in a recent study⁵. In the PR772 data set analyzed (Supplementary Note 1), the specific run used was collected while the injection system was being flushed with water, thus creating a slow elution of remaining sample particles towards the end.

We found 870 hits for OmRV and 460 hits for PR772 (see Supplementary Fig. 5a, b), meaning, respectively, 4.29% and 0.28% hit-rate. In Fig. 6a, b we show icosahedral patterns from

these hits: they are single hits, respectively, for OmRV virus and bacteriophage PR772. The snapshots show particles of different sizes, as the size distribution is quite broad for both viruses with the injection system used^{5,31}.

In the previously published analysis on the OmRV data set⁵, 421 hits were found in the same run we analyzed (Supplementary Note 1). Our set of 870 hits was a strict superset, thus including all the 421 hits previously identified (Supplementary Fig. 5a).

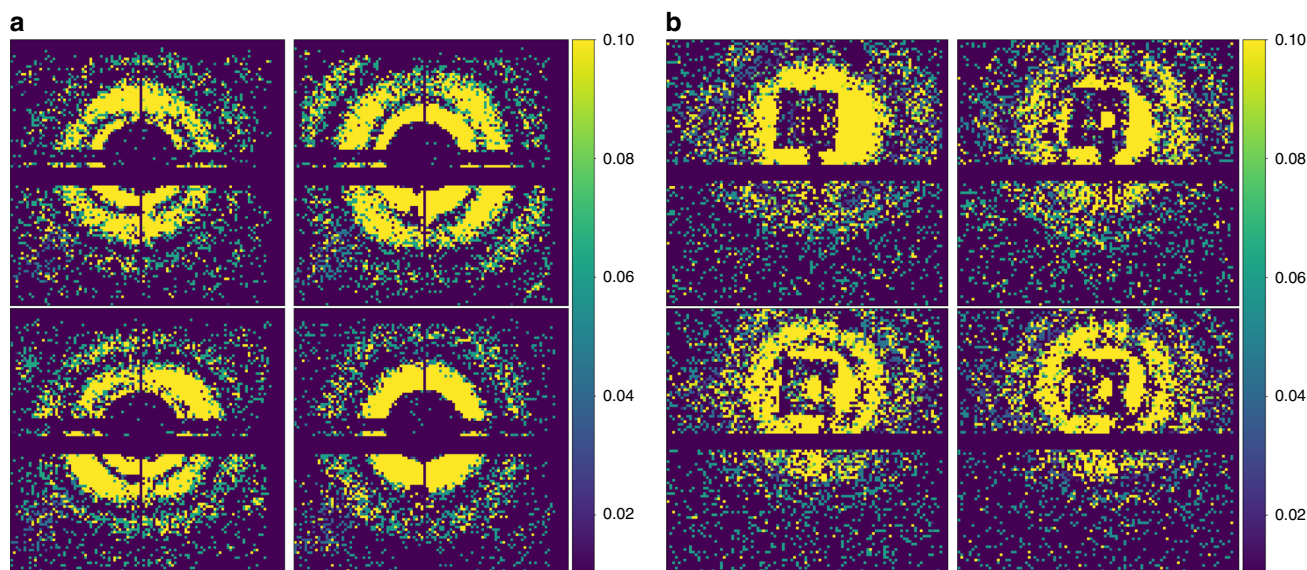


Fig. 6 Diffraction patterns representing two icosahedral viruses. **a** and **b** show, respectively, four single OmRV and PR772 hits (downsampled at 4×4 pixels). These patterns represent hits of different sizes, as the size distribution for those viruses is quite broad using the injection methods in place at the time

Moreover, we show that some of the additional hits are clearly representative of single Omono River virus particles, albeit weaker (Fig. 7).

Protein hits simulated on top of true background. We have also evaluated the efficiency of our approach using computer simulations. We compared the statistical hit-finder with one that makes use of pure photon count statistics, by normalizing to the expected photon count (Fig. 4c, d) and setting a threshold as described for the log-likelihood scores. The ratio of correctly identified hits given various photon beam intensities is reported in Fig. 8, in the special cases of diameter size 8, 13, and 40 nm spherical particles with protein-like scattering power. Focus-centered spherical hits of protein material were simulated (using Condor online³²) at different intensities (4.46×10^9 – 4.46×10^{18} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$) for the three sizes. These were superimposed on a background run for the CSPAD-140k detector (from a June 2015 experiment).

In an actual FXI experiment, the hits are rarely perfectly focused relative to the X-ray pulse. However, owing to broad tails of the beam profile²², we still get detectable scattered signal. Therefore, we also simulated particles hit by a Lorentzian and a Gaussian beam (by multiplying the simulated patterns with scalar intensities sampled from a Lorentzian and a Gaussian 2D function—see Methods).

In the case of a particle hit perfectly in focus by a tophat beam (Fig. 8a), perfect efficiency for larger 40 nm particles is reached already around 10^{10} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$; for smaller 13 nm particles, 50% and 100% correct hit-identification are reached at 5.00×10^{11} and 10^{12} , respectively.

On the other hand, when looking at the case of non-centered hits (Fig. 8b, c), we can see that at 10^{12} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$, the efficiency of the implemented algorithm is not yet at 50% in any case.

Provided greater beam intensity—e.g., $>10^{13}$ photons \times pulse $^{-1} \times \mu\text{m}^{-2}$ —evaluating hits of diameter size ~ 13 nm (or even smaller, ~ 8 nm) become feasible ($>50\%$ of correct identifications with the proposed statistical approach).

A quick comparison of the two methods for the realistic cases (Lorentzian and Gaussian beam profiles) shows that for hits of 13

nm our method can obtain a 50% recovery rate at less than a half of the intensity. As a matter of fact, the statistical hit-finder (straight lines) reaches this point, respectively, at 2.10×10^{13} photons/pulse/ μm^2 and 9.73×10^{13} photons/pulse/ μm^2 , whereas the photon count hit-finder (dash-dotted lines) achieves the same result at 4.61×10^{13} photons/pulse/ μm^2 and 2.12×10^{14} photons/pulse/ μm^2 . As we kept the background signal constant in our experiments, doubling the intensity corresponds to increasing the signal-to-noise ratio by ~ 3 dB. At the intensity where the statistical hit-finder recovers 50% of hits, the median shot contains 244 sample photons on top of a background of 1486 photons, giving a $\text{SNR} = -7.85$ dB. The corresponding median sample photon count for the photon threshold hit-finder was 442 photons. The average hit, on the other hand, contains a far higher number of photons, owing to the presence of a smaller number of well-focused hits. The average hit at the 50% recovery level for the statistical hit-finder contains ~ 4500 photons, equivalent to $\text{SNR} = 4.89$ dB, still at a level where the background component is non-negligible.

In all, our statistical approach needs less photon flux. When considering the 8 nm particle, we can conclude that reliable detection would require intensities that are unachievable at CXI²² so far.

Relying on a sound way to reveal sample hits is a first step toward single-molecule imaging. Having a reliable set of hits and models of the background component in those hits, further analysis can be performed, such as size determination, classification, and reconstruction^{4,5,17,33}.

Discussion

In previous works, Poisson log-likelihood models have been used for unmeasured backgrounds, as a concordant step during 3D reconstruction of the sample. This is a highly computationally intensive process, which would not easily lend itself to the processing of huge data sets with single-digit hit rates³⁴. Similar schemes are still relevant as a later step, to filter out correct single-particle hits.

Without need of making assumption regarding the nature of the particles giving rise to diffraction, our statistical approach addresses the hit-finding problem in a more sophisticated way

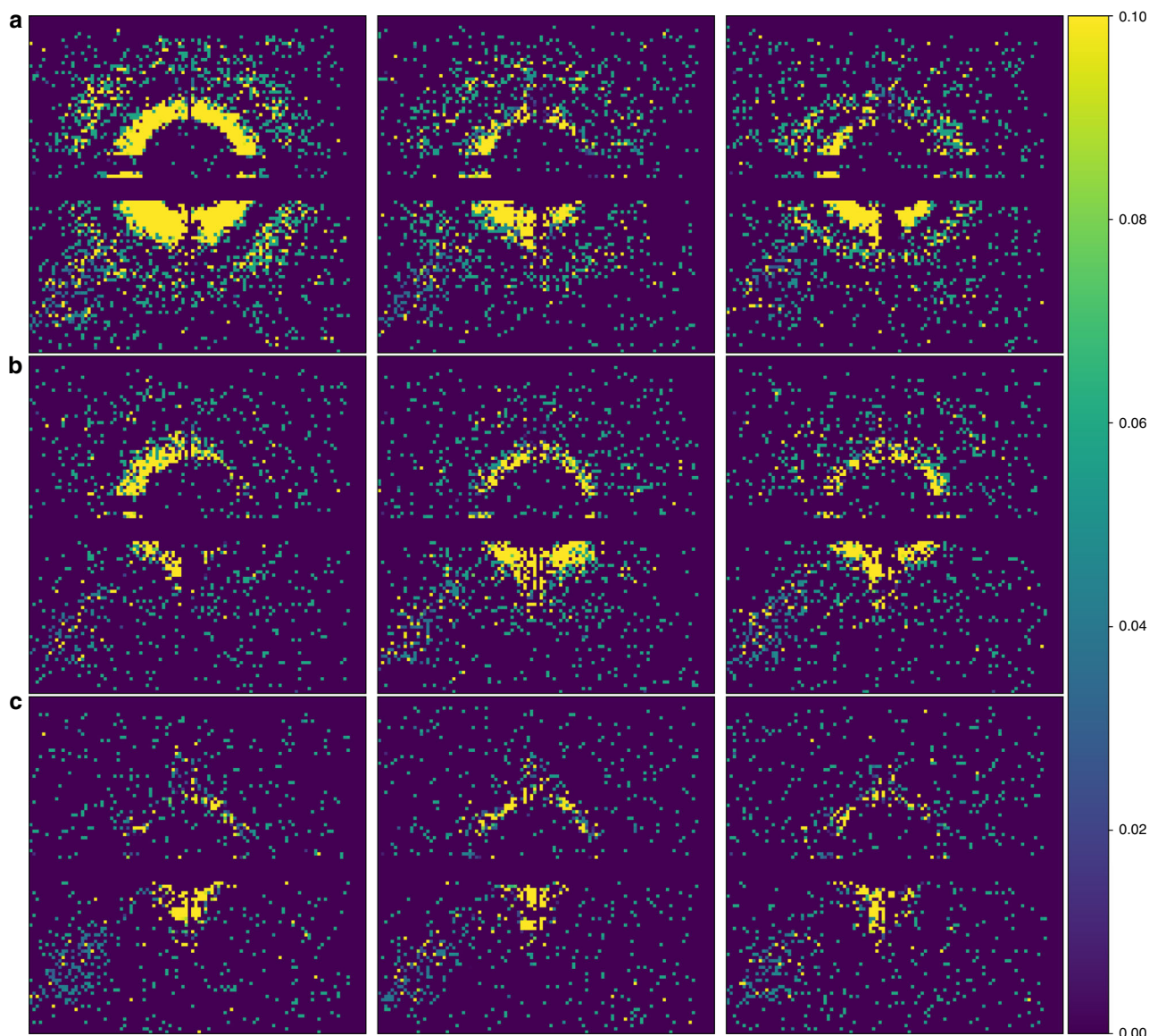


Fig. 7 Downsampled patterns belonging to the OmRV data set. Row **a** and **b** show events identified as hits by our hit-finder—in the score range 4000–6000 (**a**) and in the range 2000–4000 (**b**)—that are not identified by a simpler hit-identification scheme (such as the one used by Cheetah software); **c** shows three blank events (belonging to background), to be used as a reference to the eye for discerning hits/misses. In particular, **a** shows that our hit-finder can find particle hits from a single OmRV virus that were excluded by a standard hit-finder

than methods previously used, resulting in several improvements. First, it gives us an unprecedented accuracy in hit-identification, providing a large number of hits with a verified low false-positive rate. Second, the direct comparison of each pattern with its expected background is able to disentangle sample hits even in a very low signal-to-noise regime: a necessary development to be able to reveal sample hits in the ~ 13 nm size range. Thus, the method has general validity, working in any signal-to-noise range and allowing the identification of hits with diameters from 13 nm and up. Moreover, it is resilient to different backgrounds, as shown by testing it on different data sets. For all those reasons the statistical hit-finding approach constitutes a step towards realistic FXI experiments for protein imaging, providing new insights in the hit-finding problem and background signal treatment in general. Improvements in detectors, injection, intensity and background would make detection possible even for smaller particles (diameter < 8 nm). For larger particles, it is clear that our method is able to recover a higher number of relevant hits,

including the subset of single-particle hits of actual interest in downstream analysis.

Furthermore, the background model developed for the hit-finding algorithm can be used for other crucial steps: size determination of the hits retrieved, classification and EMC reconstruction: all essential for an eventual recovering of the 3D electronic density^{5,17,33}. We believe that renewed experiments at CXI, with current advances in detector performance as well as hit-finding technology, could succeed in single-protein complex imaging. The use of a background model is essential, since the majority of hits will be weak: most of the photons identified (even after some masking) arise from the background noise, not the sample itself.

For our main sample, the RNA polymerase II data, we identified 1165 hits from 402,296 events, meaning a $\sim 0.3\%$ hit-rate. Low-resolution 3D reconstruction is possible already with 200 single-particle diffraction patterns³³. Considering a five shifts (12 h each) experiment—a quite standard allocation time at the

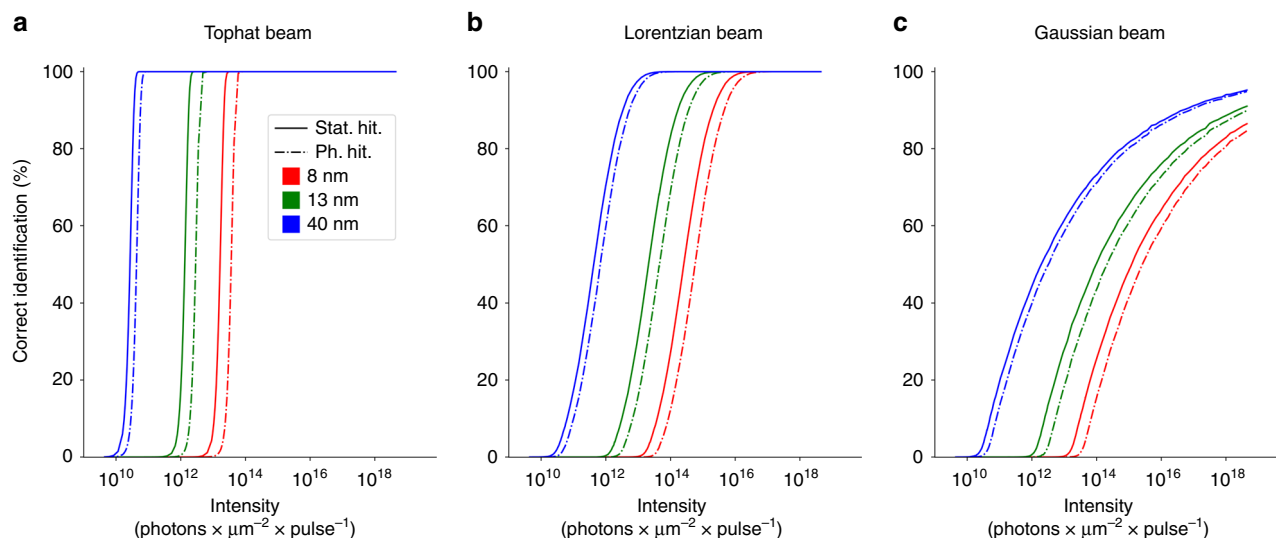


Fig. 8 Hit-finding performance of simulated hits superimposed on real background. Three representative spherical particle sizes (8, 13, 40 nm) were simulated for varying pulse intensities, particle sizes, focal properties, and hit-finding methods. We present our hit-finding method and a simpler scheme using our derived pixel mask and the total expected photon count given the pulse energy. Hit-finding was also performed in three distinct beam settings: particles hit perfectly on focus by a top-hat beam (**a**) and the more realistic cases of a Lorentzian or a Gaussian beam hitting the particle (**b** and **c**). In all the plots, normalized scores are shown as a solid line, whereas photon count based detection is dash-dotted. For the most interesting cases of a 13 nm particle hit by a Lorentzian and Gaussian beam, we found a 50% recovery, respectively, at intensities of 2.10×10^{13} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$ and 9.73×10^{13} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$ with our statistical hit-finder; at 4.61×10^{13} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$ and 2.12×10^{14} photons \times pulse $^{-1} \times \mu\text{m}^{-2}$ with the pure photon-based one

LCLS—and the 120 Hz acquisition rate at the CXI end station, and given even our low 0.3% hit-rate achieved in May 2013, that would mean collecting >72,000 particle hits. If we manage to improve sample injection and tune it so to increase the chance of having single-particle hits, we could have enough single RNA polymerase II hits to reconstruct its 3D structure at 5 Å resolution or better. The most critical improvement of injection would be to reduce the initial size of the aerosol droplets by using an electrospray injector³⁵. At that point, the ongoing research efforts in extending existing 3D reconstruction methods to model the full conformational landscape of the sample becomes even more pressing, a concern that so far has remained largely theoretical in FXI applications.

The need for robust hit-finding solutions is of increasing relevance given the current developments at XEFL sources, which focus on higher repetition rate, rather than stronger individual pulses. For example, the AGIPD detector at the European XFEL allows the readout of ~ 3500 patterns/second³⁶ and similar kHz readout rates of diffraction data are also expected for LCLS-II³⁷. Being able to identify shots containing weak diffraction patterns will be crucial in order to be able to fully leverage the properties of these sources for single-particle imaging. Filtering out >90% of all shots rapidly will allow more sophisticated classification and reconstruction schemes to be applied to the remaining subset.

Methods

Experimental setup. The experiment was performed at the CXI Instrument at the LCLS XEFL facility. Here, two X-ray detectors were available: CSPAD-2.3 M (to study high-resolution features) and CSPAD-140k (to study low-resolution features). The sample was injected in an aerosol form into the vacuum chamber, where it is hit by the XFEL beam pulse. The resulting scattering is then recorded as a diffraction pattern on the pixelated detector.

Sample preparation. The sample studied was the RNA polymerase II, an enzyme involved in DNA transcription. The sample was labeled with gold nanoparticles, using clusters of Au102(p-MBA)₄₄ covalently attached to specific sites on the molecule. The sample buffer was exchanged to 25 mM ammonium acetate by

dialysis. The other samples (OmRV and PR772) included in the analysis have been treated as described in^{5,31}.

Data processing. Raw data as recorded by the detector pixels consist of electronic signals, which are acquired by means of a 14-bit clock counter for the number of ticks until pixel voltage matches a reference voltage ramp²⁵, and so the output is expressed as arbitrary digital units (ADU) values ranging from 0 to $2^{14} = 16384$. They are stored in the XTC format, which is then converted into the CXIDB format³⁸ (a specific HDF5-based schema for X-ray Coherent Diffraction Imaging data) using the Cheetah software³⁹ and Hummingbird software²⁸. In our analysis, most preprocessing steps in Cheetah (for RNA polymerase II data) and in Hummingbird (for Omono River virus and bacteriophage PR772 data) were disabled and we carried out our own. Going from raw data to photon count and log-likelihood score data, requires the following: (i) per-column common mode subtraction (as described in previous work⁴⁰); (ii) gainmap: the ADU count corresponding to the 1-photon peak is estimated by merging the values over all the frames in the sample runs for each pixel separately. Gaussian distributions are fitted to the resulting data: the 0-photon peak is identified, followed by the 1-photon peak. The 0-photon peak fitting is constrained to be in the region where the ADU count x satisfies $x < 12$, and 1-photon peak is searched in the range $u_0 + 4s_0 < x < 50$ (u_0 being the mean fitted value of the 0-photon Gaussian—usually ranging between -3 and 3 —and s_0 its standard deviation—typically < 6). These constraints stabilized our fits, especially for pixels with relatively low, or relatively high, photon counts in the background distribution; (iii) photon count: the photon count is obtained by dividing the frame ADU data corrected through step i, by the gainmap and then rounding this value to the closest integer, truncating negative numbers to 0; (iv) rate parameters: see later section—mean photon count per pixel per frame; (v) pixel mask: we used a mask to exclude the most aberrant pixels from our calculations. Such problems can arise from detector damage or fabrication issues, signal levels where the detector linearity is compromised, or simply from regions where the beamline background is spatially unstable. We consider a “good” pixel to be one meeting our null hypothesis (as described in the first paragraph of Results section). The set of rate parameters (λ_{ki}) as calculated above are sampled according to Poisson statistics. The scalar product of the normalized vectors obtained from the histogram counts for the two different sets (binned in the same range) was taken:

$$S = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (2)$$

\mathbf{a} represents the photon counts, obtained dividing the ADU vector (after applying step i) by the gain for that pixel (obtained in step ii)); \mathbf{b} represents instead the expected photon counts (calculated as for **a**) of the same pixel.

The resulting scalar S ranges between 0 and 1: the closer S is to one the more equal the two histograms are and the better they fulfill our null hypothesis. The threshold we recommend for this purpose is 0.9999 for those data sets; (vi) log-likelihood ratio: observed photon counts in the events are compared with the expected photon count for that specific event under an assumption of pure background, as in the formula shown in the main body (Eq. 1). To some extent, the estimates in Eq. 1 will also account for deviations or errors owing to rounding effects or mismatches in the gainmap, as long as those effects influence background and sample shots identically.

The execution order of steps (i–v) is shown graphically in Supplementary Fig. 6.

Pulse energy detection. Located after the exit slit of the monochromator, the Gas Monitoring Detector (GMD) uses a rarified gas ionization system to record the energy of each beam pulse: the pulse intensity ionizes the gas, thus producing an ion drift current proportional to the energy of the shot. The proportionality factor is known and so pulse energy values are retrieved. These are associated to each frame collected by the data acquisition system, and are expressed in mJ. Pulse energy is expected to be linearly proportional to the number of photons N in each recorded pattern, as beam pulse energy is equal NE_{photon} . Instead, the trend in our data is roughly polynomial with a higher degree than 1 (third order has been tried out—see Supplementary Fig. 7). Inquiries with LCLS staff indicates that this effect can be owing to a lack of GMD calibration during the experiment in question. To account for this, re-fitted pulse energy values are used in the computation of the mean photon count per pixel (see Mean photon count per pixel). If we denote the pulse energy of the i th shot x_i , and Φ_i the expected photon count, we have:

$$\Phi_i = p_0 x_i^3 + p_1 x_i^2 + p_2 x_i + p_3 \quad (3)$$

where $p_{0,1,2,3}$ are the fitted parameters of a third order polynomial fit.

Usually, a quick look to the “photon count vs pulse energy” plot is necessary in order to remove extreme outliers (when present) in the background distribution, otherwise those will strongly affect the fit. This can include shots where the beam path was actually obstructed between the beam energy detector and the experiment chamber. Furthermore, also events with very low pulse energy values are removed for the same reason. All these events have in common that they do not represent stable operation of pulses being transmitted all the way to the end station.

Preliminary misses. In order to construct a list of preliminary misses, the photon count distribution is chronologically ordered and binned. Each bin contains at least 100 events, in order to provide reasonably stable estimates of the mean (μ_{bin}) and standard deviation (σ_{bin}) of that specific bin. The expected photon counts in each bin are fitted as explained in the section above and are then subtracted from the original photon counts of the events. μ_{bin} and σ_{bin} are then calculated and the elements of the bin are selected in the interval $(\mu_{\text{bin}} - 4\sigma_{\text{bin}}, \mu_{\text{bin}} + 4\sigma_{\text{bin}})$, calculated with the method of moments estimation⁴¹, which works well as long as the true background distribution is Gaussian. This operation is performed iteratively, reworking the fit of all parameters based on the current set of shots within the range, as long as the bin contains >100 elements or until σ_{bin} does not change anymore.

The preliminary misses based on this photon count criteria, as well as the constraints on pulse energy and photon count that exclude blank outliers per bin, are then combined to form the total set of preliminary misses. These are then used as the background events in our approach, used to compute all parameters (including the expected photon count polynomial) used for the remaining processing steps.

Mean photon count per pixel per frame. The mean photon count of each pixel in each frame can be affected both by systematic and statistical errors. We correct for those, by considering the expected photon count given a specific pulse energy (and thus a specific expected photon count Φ_i), as follows:

$$\lambda_{ki} = \bar{N}_k \frac{\Phi_i}{\Phi} \quad (4)$$

where \bar{N}_k is the actual mean photon count per pixel calculated as $\frac{\sum_j n_{jk}}{N}$, with N and n_{jk} being, respectively, the total number of frames and the observed photon count in the background, whereas the subscripts jk indicate the k th pixel in the j th frame.

Φ is the mean expected photon count per frame, calculated as $\frac{\sum_i \Phi_i}{N}$.

In total, we have:

$$\lambda_{ki} = \frac{\Phi_i \sum_j n_{jk}}{\sum_j \Phi_j} \quad (5)$$

ToF (time-of-flight) hit-finder. To validate our proof of concept method, another independent hit-finder was used: a ToF detector, which consists of a Multi-Channel Plate (MCP) detector at a distance of ~ 50 cm from the interaction region. The

MCP detector is a Z-gap triple plate detector with an active area of 40 mm. It was used in “drift mode” with no potential field across the interaction point to accelerate ions in the direction of the detector. Thus, the recorded flight times reflects directly the kinetic energy gained by the ions from the explosion of the sample particle. The detector used can be equipped with a high-pass electrostatic filter on the detector that would allow discrimination between specific ion species and charge states⁴². In the present analysis, a ToF event is considered to come from sample if more than 1 proton (~ 2 mV) signal is detected.

(For examples of ToF signals, see Fig. 5b and Supplementary Fig. 2 and 4).

Simulations—Condor online. We simulate three spherical hits (8, 13, and 40 nm in diameter) of protein material at $\sim 10^{11}$ photons \times pulse⁻¹ \times μm^{-2} of intensity and of 7 KeV in photon energy (E_{ph}), by using Condor online. Then, for each size, scaled versions are tested at 90 different intensities in the range 10^{10} – 10^{18} photons \times pulse⁻¹ \times μm^{-2} . Poisson samplings of the patterns are created and summed on top-specific background photon count frame. The total number of hits simulated for each size/intensity combination is identical to the total number of background events. We can then apply our hit-finder to this data, and determine the ratio between identified hits and the total number generated. As the background data are known, we can also estimate the false positive rate (Supplementary Fig. 1), by applying our hit-finder to unmodified background data.

Lorentzian and Gaussian beam. We simulated sampling of small hits within a Lorentzian beam, by using a 2D Lorentzian function $\frac{1}{(1+\frac{x^2}{a^2})(1+\frac{y^2}{b^2})}$; and a Gaussian beam of 400 nm, using a 2D Gaussian function $e^{-\frac{x^2}{2a^2} - \frac{y^2}{2b^2}}$. For both functions $a=b=0.2$ μm , and x, y sampled uniformly in $[-0.5, 0.5]$ μm .

Computational time and computational environment. The data analysis reported was implemented in Python and run on a cluster private to the Uppsala LMB group. Experiments used a variable number of nodes (1–8) and worker processes (10–80) in parallel, taking advantage of the *mpi4py* and *h5py* modules. Obtaining the log-likelihood scores (step vi) in Data processing section) for the RNA polymerase II sample runs (418,153 frames in total— 370×388 pixels per frame) takes $\sim 20,000$ seconds on ~ 80 CPUs. This level of performance is adequate for online as well as off-shift operation during a beamtime, given proper adaptations. For consistent online operation, however, a reasonable pixel mask and a detector gainmap for the specific beam energy will be required. The mean photon counts per pixel per frame can be estimated online from incoming samples, as already discussed.

Code availability. The code is available at <https://github.com/albpi/Statisticalhitfinder/>.

Data availability

Sample data (including raw frames, photon counts, and hit-identification scores) have been deposited into the CXIDB repository (entry 78).

Received: 2 May 2018 Accepted: 5 November 2018

Published online: 07 December 2018

References

1. Taylor, G. L. Introduction to phasing. *Acta Crystallogr.* **D66**, 325–338 (2010).
2. Johansson, L. C., Wöhri, A. B., Katona, G., Engström, S. & Neutze, R. Membrane protein crystallization from lipidic phases. *Curr. Opin. Struct. Biol.* **19**, 372–378 (2009).
3. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18**, 581–586 (2008).
4. Hantke, M. F. et al. High-throughput imaging of heterogeneous cell organelles with an X-ray laser. *Nat. Photonics* **8**, 943–949 (2014).
5. Daurer, B. J. et al. Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCr* **4**, 251–262 (2017).
6. Siebert, M. M. et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature* **470**, 78–81 (2011).
7. Hosseinzadeh, A. et al. Conformational landscape of a virus by single-particle X-ray scattering. *Nat. Methods* **14**, 877–881 (2017).
8. Kurtan, P. R. et al. Correlations in scattered X-ray laser pulses reveal nanoscale structural features of viruses. *Phys. Rev. Lett.* **119**, 158102 (2017).
9. Mengning, L. et al. The coherent X-ray imaging instrument at the Linac Coherent Light Source. *J. Synchrotron Rad.* **22**, 514–519 (2015).
10. Siebert, M. M. et al. Femtosecond diffractive imaging of biological cells. *J. Phys. B At. Mol. Opt.* **43**, 19 (2010).

11. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).
12. Chapman, H. N. et al. Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73–77 (2011).
13. Chapman, H. N. et al. Femtosecond time-delay X-ray holography. *Nature* **448**, 676–679 (2007).
14. Barty, A. et al. Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nat. Photonics* **6**, 35–40 (2012).
15. Andreasson, J. et al. Automated identification and classification of single particle serial femtosecond X-ray diffraction data. *Opt. Express* **22**, 2497–2510 (2014).
16. Jönsson, H. O., Caleman, C., Andreasson, J. & Timneanu, N. Hit detection in serial femtosecond crystallography using X-ray spectroscopy of plasma emission. *IUCrJ* **4**, 778–784 (2017).
17. van der Schot, G. et al. Imaging single cells in a beam of live cyanobacteria with an X-ray laser. *Nat. Commun.* **6**, 5704 (2015).
18. Loh, N. D. et al. Fractal morphology, imaging and mass spectrometry of single aerosol particles in flight. *Nature* **486**, 513–517 (2012).
19. Barke, I. et al. The 3D-architecture of individual free silver nanoparticles captured by X-ray scattering. *Nat. Commun.* **6**, 6187 (2015).
20. Loh, N. D. & Elser, V. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E* **80**, 026705 (2009).
21. Hahn, S. Structure and mechanism of the RNA Polymerase II transcription machinery. *Nat. Struct. Mol. Biol.* **11**, 394–403 (2004).
22. Nagler, B. et al. Focal spot and aversefront sensing of an X-ray free electron laser using Ronchi shearing interferometry. *Sci. Rep.* **7**, 13698 (2017).
23. Kirkpatrick, J. M. & Young, B. M. Poisson statistical methods for the analysis of low-count gamma spectra. *IEEE T. Nucl. Sci.* **56**, 1278–1282 (2009).
24. Hart, P. et al. *The Cornell-Pixel Array Detector at LCLS*, presented at Nuclear Science Symposium, Medical Imaging Conference, Anaheim, CA (2012).
25. Baglivo, J. & Olivier, D. Methods for exact goodness-of-fit tests. *J. Am. Stat. Assoc.* **87**, 464–469 (1992).
26. Munke, A. et al. Coherent diffraction of single rice dwarf virus particles using hard X-rays at the Linac Coherent Light Source. *Sci. Data* **3**, 160064 (2016).
27. Ferguson, K. R. et al. The atomic, molecular and optical science instrument at the Linac Coherent Light Source. *J. Synchrotron Rad.* **22**, 492–497 (2015).
28. Daurer, B. J., Hantke, M. F., Nettelblad, C. & Maia, F. R. N. C. Hummingbird: monitoring and analyzing flash X-ray imaging experiments in real time. *J. Appl. Crystallogr.* **49**, 1042–1047 (2016).
29. Hermann, S. et al. Cspad-140k: a versatile detector for lcls experiments. *Nucl. Instr. Meth. A* **718**, 550–553 (2013).
30. Hermann, S. et al. *CSPAD upgrades and CSPAD V1.5 at LCLS*, presented at Journal of Physics: Conference Series (2014).
31. Reddy, H. K. N. et al. Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac Coherent Light Source. *Sci. Data* **4**, 170079 (2017).
32. Hantke, M. F., Ekeberg, T. & Maia, F. R. N. C. Condor: a simulation tool for flash X-ray imaging. *J. Appl. Crystallogr.* **49**, 1356–1362 (2016).
33. Ekeberg, T. et al. Three-dimensional reconstruction of the giant mimivirus particle with an X-ray free-electron laser. *Phys. Rev. Lett.* **114**, 098102 (2015).
34. Loh, N. D. A minimal view of single-particle imaging with X-ray lasers. *Philos. Trans. R. Soc. B Lond. B Biol. Sci.* **369** (2014).
35. Bielecki, J. et al. Electrospray sample injection for single-particle imaging with X-ray lasers. Preprint at *bioRxiv* <https://doi.org/10.1101/453456> (2018).
36. Allahgholi, A. et al. The adaptive gain integrating pixel detector. *J. Instrum.* **11**, C02066 (2016).
37. Blaj G. et al. *Future of ePix detectors for high repetition rate FELs*, presented at AIP Conference Proceedings (2016).
38. Maia, F. R. N. C. The coherent X-ray imaging data bank. *Nat. Methods* **9**, 854–855 (2012).
39. Barty, A. et al. Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *J. Appl. Crystallogr.* **47**, 1118–1131 (2014).
40. Pietrini, A. & Nettelblad, C. Artifact reduction in the CSPAD detectors used for LCLS experiments. *J. Synchrotron Rad.* **24**, 1092–1097 (2017).
41. Hazelton, M. L. Methods of Moments Estimation. In *International Encyclopedia of Statistical Science* (ed. Lovric, M.) (Springer, Berlin, Heidelberg, 2011).
42. Andreasson, J. et al. Saturated ablation in metal hydrides and acceleration of protons and deuterons to keV energies with a soft-X-ray laser. *Phys. Rev. E* **83**, 016403 (2011).

Acknowledgements

This work was supported by the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the European Research Council, the Röntgen-Ångström Cluster, the projects Advanced research using high intensity laser produced photons and particles (ADONIS) (CZ.02.1.01/0.0/0.0/16_019/0000789) and Structural dynamics of biomolecular systems (ELIBIO) (CZ.02.1.01/0.0/0.0/15_003/ 0000447) from the European Regional Development Fund, the Swedish Foundation for Strategic Research, the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Wellcome Trust (204732/Z/16/Z), the Ministry of Education, Youth and Sports as part of targeted support from the National Programme of Sustainability II and the Chalmers Area of Advance Materials Science. Use of the 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.

Author contributions

C.N. conceived the study, in discussions with J.B., J.H., and F.R.N.C. J.B. and C.N. performed preliminary analysis. A.P. and C.N. developed the statistical hit-finding methodology. A.P. implemented the source code, and performed all data analysis and simulations. J.B. implemented the ToF hit-finding. J.B., N.T., and J.A. contributed specifically to the theory of ToF hit-identification and methodology comparison. M.F. is the original author of the simulation software Condor, and provided assistance in its use together with F.R.N.C. J.B., N.T., M.F.H., J.A., D.N.L., D.S.D.L., S.B., and C.N. were all participating in the original May 2013 experiment, for which F.R.N.C. was the principal investigator and experiment contact-person. A.P. and C.N. drafted the manuscript, which was then edited jointly by all co-authors.


Additional information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s42005-018-0092-6>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018