Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering

Chun Hong Yoon,[1] Peter Schwander,[1] Chantal Abergel,[2] Inger Andersson,[3] Jakob Andreasson,[4] Andrew Aquila,[5] Saša Bajt,[5] Miriam Barthelmess,[5] Anton Barty,[6] Michael J. Bogan,[7] Christoph Bostedt,[8] John Bozek,[8] Henry N. Chapman,[6,9] Jean-Michel Claverie,[2] Nicola Coppola,[10] Daniel P. DePonte,[6] Tomas Ekeberg,[3] Sascha W. Epp,[11,12] Benjamin Erk,[11,12] Holger Fleckenstein,[6] Lutz Foucar,[11,13] Heinz Graafsma,[5] Lars Gumprecht,[6] Janos Hajdu,[3] Christina Y. Hampton,[7] Andreas Hartmann,[14] Elisabeth Hartmann,[13] Robert Hartmann,[14] Gunter Hauser,[15,16] Helmut Hirsemann,[5] Peter Holl,[14] Stephan Kassemeyer,[11,13] Nils Kimmel,[15,16] Maya Kiskinova,[17] Mengning Liang,[6] Ne-Te Duane Loh,[7] Lukas Lomb,[11,13] Filipe R. N. C. Maia,[18] Andrew V. Martin,[6] Karol Nass,[6,9] Emanuele Pedersoli,[17] Christian Reich,[14] Daniel Rolles,[11,13] Benedikt Rudek,[11,12] Artem Rudenko,[11,12] Ilme Schlichting,[11,13] Joachim Schulz,[6] Marvin Seibert,[3] Virginie Seltzer,[2] Robert L. Shoeman,[11,13] Raymond G. Sierra,[7] Heike Soltau,[14] Dmitri Starodub,[7] Jan Steinbrener,[11,13] Gunter Stier,[13] Lothar Strüder,[15,16] Martin Svenda,[3] Joachim Ullrich,[11,12] Georg Weidenspointner,[15,16] Thomas A. White,[6] Cornelia Wunderer,[5] and Abbas Ourmazd[1,*]

[1]Department of Physics, University of Wisconsin-Milwaukee, 1900 East Kenwood Blvd, Milwaukee, Wisconsin 53211, USA

[2]Information Génomique et Structurale, CNRS-UPR2589, Aix-Marseille Université, Institut de Microbiologie de la Méditerranée, Parc Scientifique de Luminy, Case 934, 13288 Marseille Cedex 9, France

[3]Department of Molecular Biology, Swedish University of Agricultural Sciences, Uppsala Biomedical Centre, Box 590, S-751 24 Uppsala, Sweden

[4]Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden

[5]Photon Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany

[6]Center for Free Electron Laser Science DESY University of Hamburg, Notkestrasse 85, 22607, Hamburg, Germany

[7]PULSE Institute, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA

[8]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA

[9]University of Hamburg, Luruper Chaussee 149, Hamburg 22761, Germany

[10]European XFEL GmbH, Albert-Einstein-Ring 19, 22761 Hamburg, Germany

[11]Max Planck Advanced Study Group, Center for Free-Electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany

[12]Max-Planck-Institut für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany

[13]Max-Planck-Institut für Medizinische Forschung, Jahnstrasse 29, 69120 Heidelberg, Germany

[14]PNSensor GmbH, Römerstrasse 28, 80803 München, Germany

[15]Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse, 85741 Garching, Germany

[16]Max-Planck-Institut Halbleiterlabor, Otto-Hahn-Ring 6, 81739 München, Germany

[17]Fermi, Elettra Sincrotrone Trieste, SS 14 – km 163.5, 34149 Basovizza, Trieste, Italy

[18]NERSC, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 943-256, Berkeley, California 94720, USA

**ourmazd@uwm.edu*

Abstract: Single-particle experiments using X-ray Free Electron Lasers produce more than 10^5 snapshots per hour, consisting of an admixture of blank shots (no particle intercepted), and exposures of one or more particles. Experimental data sets also often contain unintentional contamination with different species. We present an unsupervised method able to sort experimental snapshots without recourse to templates, specific noise models, or user-directed learning. The results show 90% agreement with manual classification.

©2011 Optical Society of America

OCIS codes: (11.7440) X-ray imaging; (290.5840) Scattering, molecules; (320.7100) Ultrafast measurements.

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011

References and links

1. 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. 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, C. 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. 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. Spence, "Femtosecond X-ray protein nanocrystallography," Nature **470**(7332), 73–77 (2011).
2. M. M. Seibert, T. Ekeberg, F. R. 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, C. 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. H. 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).
3. M. R. Howells, T. Beetz, H. N. Chapman, C. Cui, J. M. Holton, C. J. Jacobsen, J. Kirz, E. Lima, S. Marchesini, H. Miao, D. Sayre, D. A. Shapiro, J. C. H. Spence, and D. Starodub, "An assessment of the resolution limitation due to radiation-damage in x-ray diffraction microscopy," J. Electron Spectrosc. Relat. Phenom. **170**(1–3), 4–12 (2009).
4. 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).
5. M. J. Bogan, W. H. Benner, S. Boutet, U. Rohner, M. Frank, A. Barty, M. M. Seibert, F. 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).
6. J. H. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of International Conference on Machine Learning*, Banff, Canada (ACM, 2004), pp. 47–55.
7. V. L. Shneerson, A. Ourmazd, and D. K. Saldin, "Crystallography without crystals. I. The common-line method for assembling a three-dimensional diffraction volume from single-particle scattering," Acta Crystallogr. A **64**(2), 303–315 (2008).
8. 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).
9. J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000).
10. L. Hagen and A. B. Kahng, ""New spectral methods for ratio cut partitioning and clustering," IEEE Trans," Comput. Aided Des. **11**, 1074–1085 (1992).
11. L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: brining order to the web," Technical Report. Stanford InfoLab (1999).
12. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**(5500), 2319–2323 (2000).
13. R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph Laplacian tomography from unknown random projections," IEEE Trans. Image Process. **17**(10), 1891–1899 (2008).
14. B. Mohar, "The Laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications. Vol. 2*, Y. Alavi, G. Chartrand, O. Oellermann and A. Schwenk, eds. 871 – 898 (Wiley, 1991).
15. U. von Luxburg, "A tutorial on spectral clustering," in *Statistics and Computing,* **17**(4), 395–416 (Springer, 2007).
16. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," Adv. Neural Inf. Process. Syst. **14**, 849–856 (2001) (NIPS).
17. A. Basilevsky, *Statistical Factor Analysis and Related Methods* (John Wiley &Sons, Inc. 1994).
18. L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," Adv. Neural Inf. Process. Syst. **17**, 1601–1608 (2004) (NIPS).
19. L. Strüder, S. Epp, D. Rolles, R. Hartmann, P. Holl, G. Lutz, H. Soltau, R. Eckart, C. Reich, K. Heinzinger, C. Thamm, A. Rudenko, F. Krasniqi, K. Kühnel, C. Bauer, C. 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, C. 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

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

multipurpose chamber for experiments at 4th generation light sources," Nucl. Instrum. Methods Phys. Res. A **614**(3), 483–496 (2010).
20. J. D. Bozek, "AMO instrumentation for the LCLS X-ray FEL," Eur. Phys. J. Spec. Top. **169**(1), 129–132 (2009).
21. 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, Ph. 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).

1. Introduction

Recent results have established the potential of the so-called "diffract-and-destroy" approach for determining the structure of biological entities in a way which circumvents the damage limit faced by traditional techniques using ionizing radiation [1–3]. Ideally, these experiments intercept a succession of identical single-particles with a train of intense X-ray pulses to record diffraction snapshots before each particle is destroyed. In practice, several complications arise. (a) Particle injectors [4,5] are not synchronized with the X-ray pulses, resulting in a mixture of blank shots, where no injected object was intercepted, and successful exposures of injected objects. (b) Injected objects can contain none, one, or multiple copies of the particle of interest. (c) Unintentional contamination produces snapshots from a number of different species. (d) The incident beam intensity varies from shot-to-shot. (e) Only a part of the injected particle may be exposed to the incident beam. These factors give rise to unsorted data sets consisting of blank snapshots, and a variety of snapshots emanating from one or more particles belonging to a number of species. Some snapshots also exhibit artifacts due to the detection process. For example, an intense signal can exceed the dynamic range of the detector leading to saturation and charge bleeding into neighboring pixels. The fraction of snapshots from individual particles of one species is usually small (1–10% in our experiments), and must be extracted from large collections of snapshots. Experimental noise and stochastic processes, such as shot-to-shot variations in the beam intensity and position, make this task particularly challenging. The need for snapshots from identical single-particles is an essential pre-requisite for current diffract-and-destroy, three-dimensional structure recovery techniques, severely limiting the number of "useful" snapshots.

Here, we present an unbiased, accurate, and computationally efficient method for classifying experimental X-ray diffraction snapshots without recourse to operator supervision, specific noise models, or templates. The approach is based on spectral clustering, a kernel-based Principal Component Analysis (PCA) method [6], which uses the nonlinear correlations in the data set over a variety of length scales to classify snapshots. Our primary results, obtained without supervision, can be summarized as follows. (a) Experimental X-ray Free Electron Laser (XFEL) diffraction snapshots can be sorted with 90% accuracy, as judged by manual expert assignment. (b) Noting that in most cases, the number of pixels needed to sample the intensity diffracted from a single particle according to the Shannon-Nyquist theorem is less than or about 10^4 [7,8], we estimate a sorting capability of 10^6 snapshots in less than 10 hours with modest computing resources. Given the current experimental snapshot output of ~10^6 per day at 120Hz with 10% hit rate, this capability suffices for typical experimental runs until significantly higher source and injector repetition rates and/or source-injector synchronization become available.

This paper is organized as follows. Section 2 provides a brief outline of spectral clustering. Section 3 describes sorting results for experimental XFEL data sets, and compares these with those obtained by manual assignment, and standard (i.e., linear) PCA-based sorting. Section 4 concludes with a brief summary and directions for future work.

2. Spectral clustering

Spectral clustering is an unsupervised method for classifying data, with applications including image segmentation [9], circuit partitioning [10], data mining [11], and machine learning [12,13]. It exploits similarity relationships between data vectors to discover the key variables,

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

which nonlinearly determine the global structure of the data. Complex, high-dimensional data sets are then described in terms of a small number of dimensions stemming from these underlying variables.

Consider the problem of classifying high-dimensional data vectors stemming from two different objects. This comes about, for example, with snapshots from two sets of objects, when *n* pixel intensities are measured on each snapshot, and a snapshot is represented as a point in *n*-dimensional space. Under ideal circumstances, the local neighborhood arrangements would reveal two distinct groups of internally connected data points forming two disconnected classes, with class membership reflected in a single variable. Finding such a classification is an objective of spectral clustering. Below, we offer a brief technical outline of this approach, referring the reader to the extensive literature for further details [14–16].

Formally, given a set *X* of *s* snapshots, each represented by an *n*-dimensional data point, i.e., $X = \{x_1,...,x_s\} \in \Re^n$, spectral clustering provides a low-dimensional representation of the data $U = \{u_1,...,u_s\} \in \Re^m$, which captures the main features of the data set with $m << n$. The reduced coordinates are obtained from the eigenfunctions of the graph Laplacian (see, e.g., [14]). In contrast to standard PCA and factor analysis [17], which reflect the linear variance and covariance of the data, spectral clustering captures the nonlinear correlations in the data set.

2.1 Construction of the graph Laplacian

Let $G = (V, E)$ be an undirected similarity graph with vertices $V = \{v_1,...,v_s\}$, and edges $E = (e_{ij})$ for *i*, *j* = 1, …, *s*. Each vertex v_i represents a data point in *n*-dimensional space, $x_i \in \Re^n$, with the edge e_{ij} a binary representation of connectedness between v_i and v_j, where $e_{ij} = 1$ if connected and $e_{ij} = 0$ otherwise. Subgraphs not sharing any edges are called disconnected components. The unweighted affinity matrix of the graph is the matrix $W = E$. The node volume or the degree of a vertex $v_i \in V$ is defined to be $d_i = \sum_{j=1}^{s} e_{ij}$, with the sum running over edges connected to v_i only. A diagonal "degree matrix" D is defined as $D_{ii} = d_i$.

Several types of affinity matrix W can be constructed, each capturing a different aspect of the local neighborhood relationships in the data set [15]. Here, we exploit mutual κ-nearest neighborhood (*mknn*), where edges are formed between v_i and v_j if and only if v_j is among the κ-nearest Euclidean distance neighbors of v_i, and vice versa. The *mknn* construction is well suited to detecting clusters of different density [15]. Affinity matrix construction methods depend on the single parameter κ specifying the number of mutual nearest neighbors, which can be tuned to optimize the output. Self-tuning algorithms exist to automate this process [18].

A variety of spectral clustering algorithms exist. We adopt the normalized spectral clustering algorithm developed by Shi and Malik [16], which proceeds as follows. First, choose the number of clusters to compute *m*. The unnormalized graph Laplacian is computed by subtracting the degree matrix from the affinity matrix:

$$L = D - W. \tag{1}$$

The normalized graph Laplacian is computed by normalizing each vertex by the node volume:

$$L_{norm} = I - D^{-1}W. \tag{2}$$

The eigenvalues λ and eigenvectors u are computed by solving the generalized eigenvalue problem:

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

$$L_{norm}u = \lambda u. \quad (3)$$

Let U be the s x m matrix with the eigenvectors $u_1,...,u_m$ as columns associated with the m largest eigenvalues. The i-th row of U is the reduced coordinate u_i of data point x_i. The reduced coordinates are partitioned using K-means clustering into $C_1,...,C_m$.

3. Application to experimental XFEL data

3.1. Data set

The experiments were carried out with the "CFEL ASG multi purpose" (CAMP) instrument [19] on the Atomic, Molecular and Optical Science beamline [20] at the Linac Coherent Light Source [21] in Stanford, California, essentially as described previously [2]. Nanorice, 250 nm long ellipsoidal particles of iron oxide coated with silicon dioxide (Corpuscular Inc., Cold Spring, NY) suspended in ethanol, was concentrated to ~10^{13} particles/ml and sonicated prior to aerosolization with a Burgener Mira Mist CE nebulizer (AHF Analysentechnik, Tübingen, Germany). *Acanthamoeba polyphaga* mimivirus was purified as described previously [2], transferred into 250 mM ammonium acetate, pH 7.5, and the suspension aerosolized with helium in a gas dynamic nebuliser [4]. The aerosols were injected into the CAMP instrument using an aerodynamic lens stack [5].

Fraunhofer diffraction patterns were recorded on two rectangular pnCCD detectors consisting of 512 × 1,024 pixels, each 75 × 75 μm^2 [19]. The detectors were placed 738 mm from the interaction point and separated by a horizontal gap of 21 pixels to minimize detector damage by the XFEL beam. The X-ray energy was 1.2 keV, and the measured electron bunch length was 150fs. The scattering angle to the edge of the detector corresponded to 20 nm resolution. Although each experimental run involved a single species, cross-contamination from the preceding runs produced snapshots from several species, including, but not limited to *Enterobacteria phage* T4, nanorice, and mimivirus. A selection of diffraction snapshots is shown in Fig. 1. The data set analyzed consisted of 7,214 diffraction snapshots collected by intentionally injecting nanorice and mimivirus, loosely pre-filtered based on total diffracted intensity to reduce the number of blank shots but retain as many hits as possible. As expected, the remaining data set used here for analysis still included some blank shots, and exposures of injected objects ("droplets") containing none, one, or multiple particles. Our aim is to extract useful snapshots of the intentionally injected particles, i.e. to identify clusters each containing all orientations of single nanorice, or mimivirus.

Most of the signal is concentrated in the low spatial frequencies, i.e., the central region of the detector. Prior to analysis, the high spatial frequencies were truncated by a circular mask with a radius of 150 pixels, i.e. at a maximum frequency of 61 nm (Fig. 1(a)). Due to the large dynamic range of ~50× in diffracted intensities recorded by the detector, square-roots of the absolute intensities were used in the analysis. No other intensity scaling was applied, and no attempt was made to compensate for shot-to-shot fluence fluctuations.

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

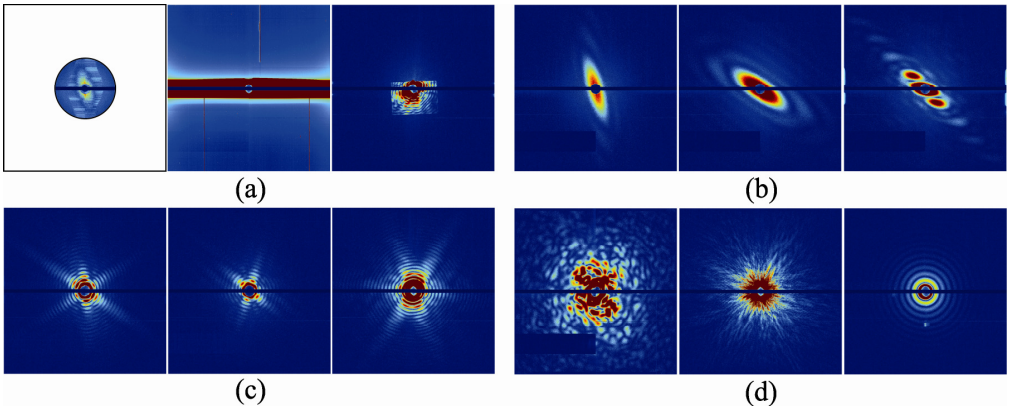

Fig. 1. Selection of diffraction snapshots from the data set. (a) Blank or unusual snapshots. (b) Nanorice in different orientations and incident beam intensities. (c) Mimivirus in different orientations and incident beam intensities. (d) Other viruses and particles. The mask shown in the upper left corner (with radius = 150 pixels) was applied to the data and analysis performed with the remaining pixels.

3.2. Results

The unweighted edges e_{ij} were used to construct the affinity matrix, W with nearest neighbor parameter $\kappa = 30$. This parameter is not critical; a range of 10-40 was found to be suitable by trial and error. The *mknn* construction then creates 2,451 disconnected clusters. The primary cluster contains 4,568 snapshots comprising the majority of the injected particles. The remaining 2,450 clusters, each consisting mainly of a single snapshot, were discarded from further analysis. Manual inspection of the 2,646 discarded snapshots revealed only about 50 snapshots of nanorice and mimivirus.

The eigenvectors $U = \{u_1,...,u_m\}$ of the graph Laplacian of the primary cluster were computed with the number of principal components $m = 24$. The parameter m is the number of clusters expected from the data set, and was set to be higher than the number of injected species. To partition the data into discrete classes, K-means clustering was applied to the reduced coordinates U, and the returned clusters ordered in decreasing size. Although not used here, one possible procedure for automatically choosing the clustering parameters is given in [18].

Inspection of the clusters revealed the following. The intentionally injected species of interest were assigned primarily to clusters 2, 5, 7 and 8. Clusters 2 and 5 (Fig. 2(a)) correspond to nanorice in various orientations. Clusters 7 and 8 (Fig. 2(b)) correspond to mimivirus in different orientations. The discarded clusters, for example, 10, 9, and 14, correspond to miscellaneous, blank, and saturated snapshots (Fig. 2(c-e)). A few other clusters contained snapshots with very high intensity levels, which may or may not have reached saturation.

In order to assess the reliability of the results, manual sorting was performed independently by two experts, each assigning snapshots to one of three classes regardless of the level of saturation: (1) single nanorice; (2) single mimivirus; and (3) other/unknown. The agreement between two manual classifications was 95%. The agreement between expert manual assignment and the outcome of unsupervised algorithmic sorting was 90%. We note that the algorithm differentiates between saturated and unsaturated shots, while the manual assignment did not. As such, the comparison may represent an overly pessimistic assessment of the performance of the algorithm.

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

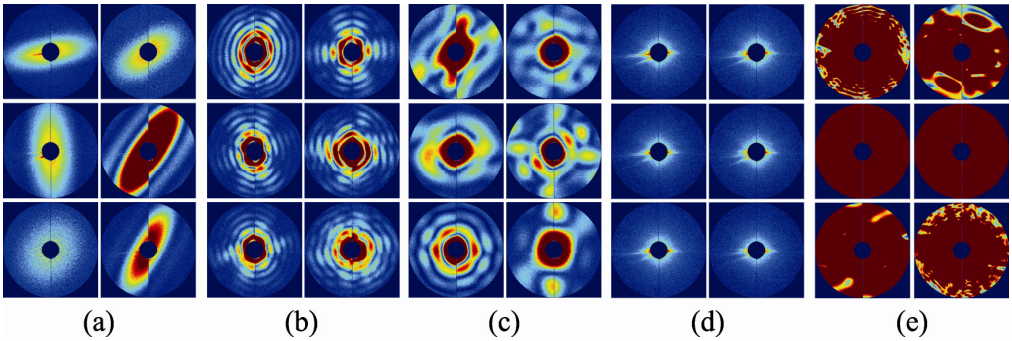

Fig. 2. Randomly selected representatives from (a) Cluster 2: nanorice snapshots; (b) Cluster 7: mimivirus snapshots; (c) Cluster 10: miscellaneous snapshots; (d) Class 9: blank snapshots; and (e) Cluster 14: saturated snapshots.

It is important to note that the algorithmic classification is not simply based on total diffracted intensity, since the distributions of total diffracted intensity in the nanorice and mimivirus clusters overlap almost completely. (The mean and standard deviation of total diffracted intensity distributions for clusters 2 (nanorice) and 7 (mimivirus) are 0.575 ± 0.187 (a.u.) and 0.643 ± 0.163 (a.u.), respectively.).

The final sorting yield of the algorithm is summarized in Table 1. Of the 611 missed snapshots, 486 from single nanorice and mimivirus belong to discarded clusters 1 and 4, which consisted mainly of strongly diffracting snapshots of various species (The mean and standard deviation of the total diffracted intensity distributions are 1.186 ± 0.319 (a.u.) and 2.729 ± 0.741 (a.u.), respectively).

Table 1. Snapshot Sorting Yield for Nanorice and Mimivirus

		After Clustering		
Total	*After Prefiltering*	*Correctly extracted*	*Missed*	*Other snapshots*
1,569,406	7,214	1,270	611	5,332
100%	0.46%	0.08%	0.04%	0.34%

The computational cost of the algorithm is determined primarily by the Euclidean distance calculations. The *mknn* approach produces sparse matrices, whose dominant eigenvectors can be efficiently computed. For example, 7,214 snapshots can be sorted in 150 seconds on a desktop computer with a 2.66 GHz, 32 GB RAM, Quad-Core Intel Xeon CPU. The Euclidean distance calculation scales as $O(s^2)$, where s is number of snapshots. Assuming 10^4 Shannon pixels are needed per snapshot, we estimate a computation time of less than 10 hours to sort 10^6 snapshots using a cluster of 30 nodes, each consisting of two 2.5 GHz Quad-Core Intel Xeon CPUs with 16 GB RAM at half load with an existing parallel implementation. This estimate is an extrapolation based on the dominant scaling behavior only, and must be verified in practice.

We now benchmark the performance of our spectral clustering approach against one based on standard PCA analysis followed by K-means clustering (see Table 2). Spectral clustering is superior in two respects. In order to obtain similar results, (1) five times more clusters must be considered in PCA; and (2) the species of interest are spread over nine times more clusters after PCA analysis. These differences persist even when intensity and variance normalization of the data are employed. The superior performance of spectral clustering points to the nonlinear structure of XFEL data sets, which is not well-suited to analysis by (linear) PCA-based methods.

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA

Table 2. Comparison between Standard PCA and our Spectral Clustering Algorithm (SC)

	Total number of Clusters	*Clusters of interest*	*Yield (%)*	*Accuracy (%)*
PCA	97	36	66	86
Spectral Clustering	24	4	68	88

4. Conclusions

We have presented an unbiased, accurate, and algorithmically efficient method for sorting experimental snapshots produced by XFEL-based diffract-and-destroy techniques, without recourse to supervisor-directed learning, specific noise models, or templates. Our current implementation is capable of dealing with data sets consisting of ~10^6 snapshots, the number of hits typically collected in an XFEL run extending over a day. This is an essential step in realizing the full potential of the possibilities demonstrated by recent results on biological and nanocrystalline particles. Future algorithms must aim to improve the sorting yield and accuracy, and offer online capability to enable informed termination of experimental runs.

Acknowledgments

We are grateful to D. Giannakis, R. Fung and F. L. Wang for discussions, and acknowledge support from: the U.S. Department of Energy Office of Science (SC-22, BES) awards #DE-SC0002164 and #DE-SC0002164, and through the PULSE Institute at the SLAC National Accelerator Laboratory; the Max Planck Society for funding the development and operation of the CAMP instrument within the ASG at CFEL; the Hamburg Ministry of Science and Research and Joachim Herz Stiftung as part of the Hamburg Initiative for Excellence in Research (LEXI); and the Hamburg School for Structure and Dynamics in Infection. This work was also supported by the following agencies: the Swedish Research Councils; Stiftelsen Olle Engkvist Byggmastare; the Swedish University of Agricultural Sciences; the Helmholtz Association (VH-VI-302); the DFG Cluster of Excellence at the Munich Centre for Advanced Photonics; the Centre National de la Recherche Scientifique; Agence Nationale de la Recherche (ANR-BLAN08-0089). 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 US Department of Energy, Office of Basic Energy Sciences.

#148527 - $15.00 USD Received 2 Jun 2011; revised 5 Aug 2011; accepted 6 Aug 2011; published 12 Aug 2011
(C) 2011 OSA