Software platform for European XFEL: towards online experimental data analysis

S. A. Bobkov,[1] A. B. Teslyuk,[1, 2, *] S. I. Zolotarev,[1] M. Rose,[3]
V. E. Velikhov,[1] I. A. Vartanyants,[3, 4] and V. A. Ilyin[1, 2, **]

(Submitted by A. A. Editor-name)

[1]*National Research Centre "Kurchatov Institute", Akademika Kurchatova pl. 1, Moscow, 123182 Russia*
[2]*Moscow Institute of Physics and Technology (State University), Dolgoprudny, Moscow Region, 141700 Russia*
[3]*Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, D-22607 Hamburg, Germany*
[4]*National Research Nuclear University MEPhI, Kashirskoe Shosse 31, Moscow, 115409 Russia*

Received March 12, 2018

Abstract—Large amount of data being generated at large scale facilities like European X-ray Free-Electron Laser (XFEL) requires new approaches for data processing and analysis. One of the most computationally challenging experiments at an XFEL is single-particle structure determination. In this paper we propose a new design for an integrated software platform which combines well-established techniques for XFEL data analysis with High Performance Data Analysis (HPDA) methods. In our software platform we use streaming data analysis algorithms with high performance computing solutions. This approach should allow analysis of the experimental dataflow in quasi-online regime.

2010 Mathematical Subject Classification: 68U01, 65Z05

Keywords and phrases: *Single particle imaging, Expectation-Maximization, X-ray Free-Electron Laser, High Performance Data Analysis, X-ray Cross-Correlation Analysis*

1. INTRODUCTION

A new approach of High Performance Data Analysis (HPDA) [1] aims to develop high performance computer efficient solutions for solving Big Data problems in different fields. One of the challenging HPDA applications is reconstruction of the three-dimensional (3D) structure of biomolecules for structural biology applications. One of the promising approaches for 3D structure recovery of biomolecules are Single Particle Imaging (SPI) experiments [2, 3] performed at the X-ray Free-Electron Laser (XFEL) facilities [4, 5].

In such experiments scattered radiation is measured from reproducible bioparticles in unknown orientations injected into the XFEL beam. The particles are destroyed during scattering process. However, due to ultra-short radiation pulses (about tens of femtoseconds) generated by the XFEL, the diffraction patterns are recorded on a detector before the radiation damage takes place. This approach is called *diffraction-before-destruction* [6]. No optical elements are used to produce an image of the biological particle under investigation. Instead, the SPI approach employs complex algorithms to reconstruct the structure from collected diffraction patterns. Diffraction pattern contains only limited information about the structure. Consequently, 3D structure reconstruction is possible from many diffraction patterns of the particle in random orientations. The reconstruction with high resolution requires large statistics of diffraction patterns to be measured. The SPI data analysis workflow includes three main steps: classification of images with diffraction patterns

[*] E-mail: anthony.teslyuk@grid.kiae.ru
[**] E-mail: ilyin0048@gmail.com

suitable for reconstruction, orientation recovery, and phase retrieval. Each step requires high performance data analysis.

Several successful reconstructions of 3D structures of biological particles were performed based on diffraction patterns collected during experiments at the XFELs [7–10]. Presently, there is no unified software solution available for the whole workflow of structure reconstruction procedure directly from the SPI experimental data, while effective software packages have been developed for individual steps of the workflow [11–13]. The reconstruction process requires detailed analysis of intermediate results after each step. Various software packages for individual steps of analysis use their own data formats and they are often not compatible with each other. Integration of heterogeneous programs into a unified data processing pipeline is an important task. Due to the complex nature of the applied algorithms and large data volumes, successful reconstruction and validation typically requires more than a year of research after data acquisition.

A complete reconstruction pipeline requires time-consuming computations. The spatial resolution of the reconstructed structures depends strongly on the number of single hit diffraction patterns. Computation time increases at least linearly with the number of diffraction patterns. Processing of the millions of diffraction patterns for sub-nanometer resolution can only become feasible with the HPDA approach using powerful computing resources.

The European XFEL facility started its operation in 2017. This facility has potential to acquire 27000 diffraction images per second. With this data rate, a million of diffraction images suitable for reconstruction can be recorded within one hour.

In this paper we formulate an approach to develop the software platform to implement the complete data analysis pipeline for SPI experiments at European XFEL. This pipeline will include all necessary steps to obtain the 3D structure of a biological particle from the raw experimental data. It will make use of existing software solutions for individual steps of analysis. Missing software components will be developed to complete the pipeline. As a result, the processing steps requiring manual intervention will be minimized or even excluded.

2. CURRENT IMPLEMENTATION OF THE DATA PROCESSING

Currently, structure reconstruction starts when the experiment is completed. Empty images are fitered out, diffraction patterns are classified by the original particle type. A dataset that include all selected diffraction patterns from the particle of interest is prepared. After that, orientations of diffraction patterns in the dataset are determined. Two-dimentional diffraction patterns in respective orientations are combined into 3D intensity map. Then, phases of the scattered radiation are retrieved. Phase information is combined with 3D intensity into 3D distribution of complex amplitude, which can be directly translated into 3D structure *via* Fourier transform.

The European XFEL facility will produce up to 27000 diffraction images per second. For such experiments a new AGIPD [14] detector has been developed, which has a resolution of 1024×1024 pixels, and the size of one image is 2 Megabytes. About 200 Terabytes of data will be acquired during one hour in a typical SPI experiment. As we mentioned above, there is no ready-to-use platform that would allow fast processing of the data flow of this scale.

Computing resources of about 0.7 teraflops is required for primary filtering and classification of such data flow [15]. Required performance can be achieved on a rather small cluster. For the next reconstruction steps, about $2 \cdot 10^5$ Teraflops of computing resources is necessary. These requirements were estimated on the basis of existing software in a numerical experiment. Processing of the data acquired within an hour of the experiment requires several days of computation on the laboratory clusters. To process experimental data at the generation rate, it is required to use a supercomputer with several thousands of cores and total performance of about hundreds of teraflops. Modern supercomputers and computing resources of the Kurchatov Institute, for example, provide a total performance of about thousand of teraflops, which is sufficient for the proposed objective.

Typically, collected data are stored in a specific format associated with a certain detector. These data have to be preprocessed and converted in more commonly used format for SPI datasets, such as CXI data format [16]. Preprocessing is a complex procedure that includes many steps [17]. Raw data has to be calibrated, background has to be determined and subtracted, response ratio of different detector panels has to be estimated, pixel coordinates have to be precisely computed, any detector artifacts need to be determined and corrected. Raw data has to be converted into a distribution of scattered photons.

Presently existing software solutions for separate steps of analysis are not integrated with each other and hardly use additional information that can be extracted from previous steps. Integration of these software packages along with new solutions that will appear in the future into a single computing platform will allow to automate the reconstruction process and to reduce the necessity of the manual control. The proposed approach should improve the process of evaluation of collected SPI data and open up the possibility to extract more valuable data in a single experiment. Therefore, continuous analysis and processing of dataflow can directly improve the resolution of the reconstructed structures.

3. IMPROVEMENTS IN DATAFLOW PROCESSING

The proposed improvements in data analysis are focused on two aspects. First, data analysis will start together with the experiment. Collected data will be directly sent to the analysis pipeline. Currently, only rough estimates of the number of images suitable for reconstruction can be obtained before the experiment is completed. All reconstruction stages will be optimized for analysis of dataflow instead of a complete measured dataset. Algorithms which were originally designed to process a complete dataset will be modified for batch processing. They will start processing of incomplete set of already collected data, and batches of new data will extend the processed dataset. Since the algorithms which process complete datasets are mostly iterative, it is convenient to update the dataset when a new iteration of an algorithm starts.

Second, various software packages will be integrated with each other. Compatible data formats will be used to reduce data conversion, improve performance and provide a uniform access to all available data and metadata for analysis. User intervention in data processing will be minimized as it is one of the main factors which slows down the analysis. Available software will be optimized for high performance and scalability in parallel computing architectures including hybrid architectures.

Integrated platform will provide a full processing pipeline, from the experimental dataset to reconstructed structure. Software components will be optimized for performance, scalability and resource management effectiveness. Data formats for input and output on every pipeline stage will be standardized and all software components will be updated to support these formats. The proposed scheme of data processing workflow is presented in Fig. 1. The processing procedure includes the following steps:

- Empty image filtering. After this step empty patterns are being filtered out based on the intensity threshold.

- Diffraction patterns classification. Stream of diffraction patterns is being sorted using machine learning algorithms into three categories: single hit patterns, multiple hit patterns and the rest which are considered as non informative patterns.

- Single hit patterns are used as an input for x-ray cross correlation analysis (XCCA) to determine preliminary information about the structure. The latter can be used to estimate the initial guess for a Fourier density map FDM_1. Current approximation of density map FDM_i and a stream of single hit patterns reshaped into a set of batches with a number of patterns in a single batch are input data for Expectation-Maximisation (EM) orientation determination algorithm. The output from the orientation determination algorithm is a refined Fourier density map FDM_{i+1}. The refined Fourier density map is used for further iterations of the EM algorithm together with new batches of diffraction patterns from experiment and already processed diffraction patterns.

- When an obtained FDM meets the established criterion, it is used as input for Phase retrieval iterative algorithms. These algorithms are used to reconstruct corresponding Phase map (PM), which when combined with FDM can be transformed into 3D electron density map (3D-DM) of a particle.

- The resolution of a 3D density map is estimated and validated. When the resolution exceeds the required level, the obtained 3D density map is used as the final 3D structure of the single particle under study.

The data processing workflow is grounded on ideas of High Performance Data Analysis (HPDA) approach. The most computationally intensive steps of analysis: *Orientation Determination* and *Phase Retrieval* are designed to operate in asynchronous batch mode. One can vary batch size, frequency of *Phase Retrieval* calculations, the number of different initial map initializations for FDMs and PMs. Calculations of individual batches as well as reconstructions for different initial parameters sets can be parallelized across high performance computing infrastructure. With asynchronous operations and parallel computing one can balance the load of individual analysis operations thus avoiding possible bottlenecks and improving total performance.

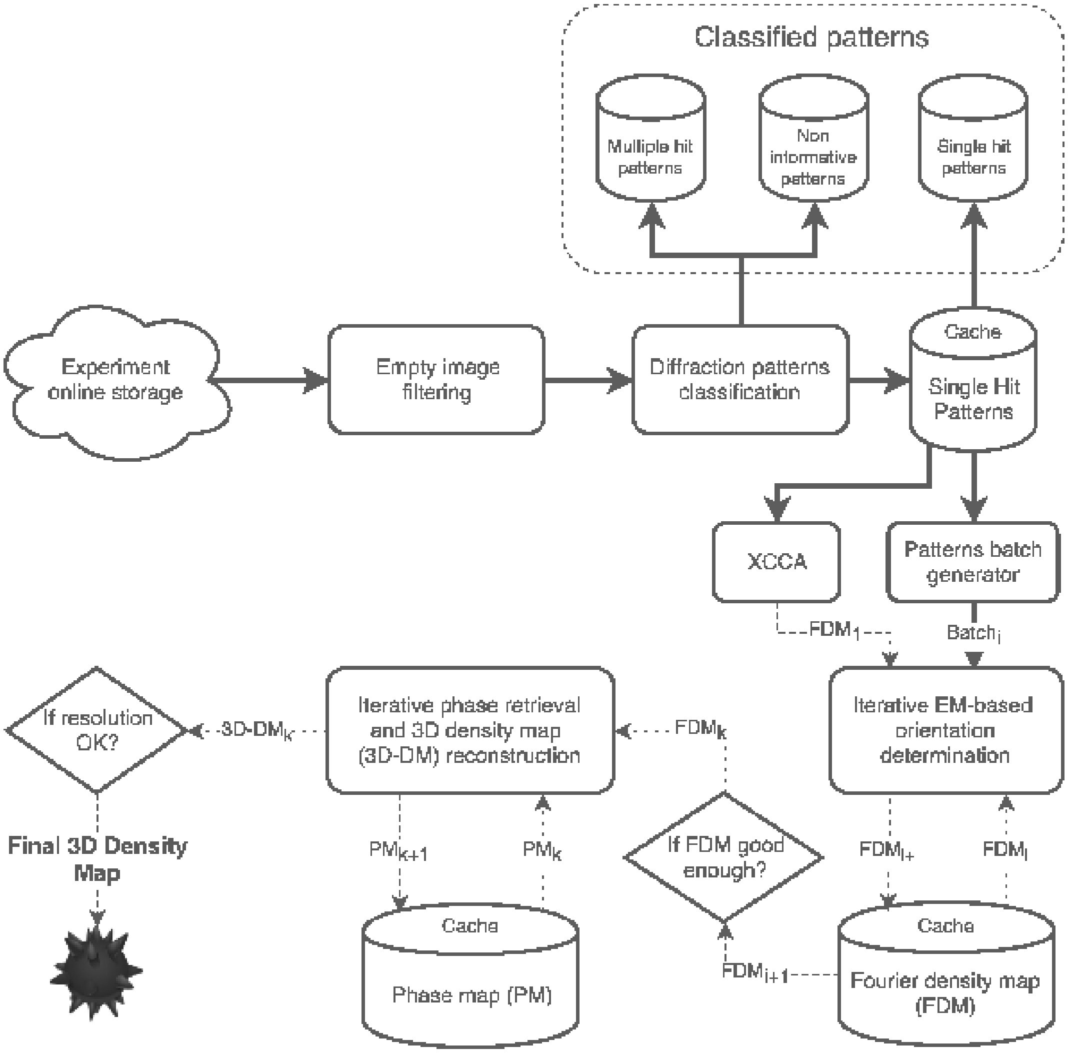

Figure 1. Asynchronous data processing workflow.

Below we describe each step in more details.

3.1. Preprocessing

The preprocessing stage involves transformation of detector output into diffraction images suitable for further analysis. This process is relatively fast and every image is processed separately. Therefore, preprocessing can be performed online on the basis of available software. Current workflow for data preprocessing with AGIPD detector includes a calibration stage when calibration constants that are required to convert the detector output into photons are determined. We plan to integrate the calibration software into the platform, as we can estimate and improve calibration quality on the basis of subsequent results.

3.2. Classification and filtering of diffraction patterns

The first step of data analysis after preprocessing is selection of diffraction patterns that are suitable for reconstruction. Typically in the SPI experiments less than one percent of collected data is suitable for reconstruction. More than 95% of images are empty frames, they can be filtered out with well-established methods [18]. Such images can be used to estimate background scattering level for calibration purposes. Collected data also include empty images without scattered signal, as well as images containing scattering from different particles or multiple particles. Only diffraction patterns from the single particles of interest are suitable for reconstruction.

Selecting suitable single-particle images among other non-empty images is a more complex problem. Filtering of such images can be performed *via* classification by structure type. Recently, a new approach for classification of such images has been proposed [19]. Comparative study of different classification approaches has been presented in Ref. [15], where classification precision, sensitivity and performance were investigated, and optimal sizes of training sets for machine learning were determined. Considered approaches and their implementations allow to perform classification of data measured at the acquisition rate of European XFEL.

Due to necessity for algorithm training before classification, analysis of the dataflow is possible with only a small delay for training [15]. First non-empty images will be classified manually to create a relatively small dataset to train classification software. After that, all experimental data will be processed at acquisition rate. Diffraction images of single particles of interest will form a reduced dataflow for reconstruction and for angular x-ray cross-correlation analysis (XCCA) [10]. A feedback from subsequent analysis stages will be used to improve classification. Images of other structure types will be transferred to separate datasets which can also be used in the analysis.

Another important factor for data selection is particle size analysis. Particles measured in SPI experiments can have different sizes either due to intrinsic heterogeneity, specific sample preparation protocol, or sample delivery scheme. Images corresponding to a limited range of particle sizes should be selected for successful structure reconstruction. Size determination can be performed by autocorrelation [18] or by fitting of power spectral density function with a form factor of a spherical particle [10] or spheroid [8]. Size determination can be done separately for individual images, its performance is sufficient for online analysis in SPI experiments at the European XFEL.

3.3. X-Ray Cross-Correlation Analysis

Angular x-ray cross-correlation (XCCA) approach allows to estimate dataset quality and to retrieve valuable structural information about the particle of interest [10, 20]. This information can be used for early detection of experimental problems; experiment can be aborted if collected data will significantly differ from the expected parameters estimated by means of XCCA. XCCA can also be used to create additional constraints for reconstruction methods which could improve convergence and robustness of the reconstruction step.

3.4. Orientations recovery

When informative diffraction patterns are selected from the data stream they need to be aligned according to particle spatial orientation. Each diffraction image represents a 2D section of 3D distribution of the scattered intensity in reciprocal space. Orientation recovery is one of the most time consuming step of structure recovery. Nowadays, it requires several days of computation on a modern server to reconstruct orientations of the experimental dataset consisting of 10000 diffraction images.

One of the most popular methods to find orientations of diffraction patterns and thus the 3D intensity distribution is the Expansion-Maximization-Compression (EMC) algorithm [12, 21]. It has been successfully applied in a number of structure reconstructions [7, 9]. The EMC method is based on a powerful machine learning Expectation Maximization (EM) algorithm [22] which has been used in a large number of applications in various fields of research.

The EM algorithm in its general formulation needs all the data to be available prior to the analysis. This is not the case when we have a stream of data and want to do online learning while data are collected from the stream. For this case several modifications of the EM method were proposed, namely batch EM [23], incremental EM [24] and stepwise EM [25, 26]. All three methods were shown to produce correct results and improve the likelihood function after each iteration.

Other promising modifications of the EM method which can accelerate calculations are Stochastic EM [27, 28] and Monte Carlo EM [29] methods. The idea of these methods is to replace the probability density function, which is used to calculate the likelihood, with a random sample of n (Monte Carlo) or even one (Stochastic) example. This technique significantly simplifies calculations of the likelihood function, that is the most computationally intensive part of the algorithm.

Implementation of EM algorithm modifications will allow solving the problem of orientation determination in the interactive regime, and getting a rough approximation of the 3D intensity distribution just after first data have been measured. However, calculation of the final result might need to reuse the whole dataset for refinement. In this case a draft structure computed during data collection will be used as initial approximation. This will significantly improve further reconstruction since EM-based methods strongly depend on the choice of initial parameters. If bad initial values were chosen the algorithm may converge to a wrong structure or may not converge at all. The choice of the most efficient algorithm for the orientation determination problem is a subject of our further research.

The EMC reconstruction will start once sufficient dataset is collected. A criterion for a sufficient dataset can be based on the XCCA analysis. Another option is to estimate the statistical significance of image density for given q-values in reciprocal space.

Many EMC reconstructions will run in parallel with different initial states (guesses). These states can be chosen randomly, they can be based on XCCA and size analysis, or be a combination of both options. Reconstructions which do not converge will be detected and discarded automatically. Successful reconstructions will be used in phase retrieval, and will provide feedback for previous stages of analysis. When more suitable images will be collected in the experiment, a new set of EMC reconstructions will be launched with initial states which are based on previous successful reconstructions. Successful results will be compared by the likelihood metrics.

3.5. *Phase retrieval*

When orientations are recovered, diffraction images can be combined into a 3D intensity distribution in reciprocal space. To determine the structure of a particle, it is necessary to retrieve phases of the scattered radiation. The 3D intensity distribution complemented with phase information can be converted into 3D electron density of the particle through a single 3D Fourier transform.

However, phase information is not recorded by a detector and has to be retrieved. Phase retrieval can be performed based, for example, on well-known Error-Reduction (ER) and Hybrid Input-Output (HIO) algorithms [30, 31]. These approaches are based on iterations between real and reciprocal spaces. At each iteration, various types of constraints are imposed, e.g., amplitude constraint in reciprocal space, and support constraint for electron density distribution in real space. Eventually, the algorithm converges to the phase distribution, which satisfies all imposed constraints, and produces the 3D structural image of a particle in real space.

Currently, phase retrieval is performed after successful determination of orientations of all images in a complete dataset. We propose to start phase retrieval on intermediate results that were computed for an incomplete dataset. Iteration process starts with phase distribution that is chosen randomly and can significantly differ from true distribution. Algorithm convergence is slow and typically requires few thousands of iterations. As a result, a lot of iterations have to be performed until an intermediate result gets close to the true values. These iterations can be computed on intermediate data. When more images are collected and better amplitude distribution is determined, phase retrieval can start from the intermediate result for faster convergence.

Phase retrieval will be organized similarly to the process of orientations determination. Many reconstructions will start in parallel with different random initial phases. Each computation will use one of the successful EMC results as an amplitude constraint. EMC results with better likelihood value will be used in more phase retrieval computations. The reconstructed electron density distribution will serve as an intermediate reconstruction result. These results will be analyzed to provide feedback for previous stages of analysis and for experiment. When the updated EMC result will be available, a new set of parallel computations will start. A new set of initial phases will be based on converged phase retrieval results obtained at a previous incomplete dataset collection step.

The presented data processing scheme makes it possible to isolate conformational changes of the structure with small differences when enough data is collected. If EMC results tend to converge to different intensity distributions, this tendency can be considered in the classification to separate dataflows for different conformations. Orientations and phases for these dataflows will be computed separately according to the presented scheme.

4. RESOURCES ESTIMATION

4.1. Network and storage

During the experiment diffraction patterns are saved at XFEL online cluster which can be thought of as an intermediate data cache and then are transferred to offline storage deployed at IBM General Parallel File System (GPFS). GPFS is a high-performance clustered file system widely used in supercomputing environments. Typical delay between data generation and data availability on GPFS is few minutes. To be able to process the incoming data in near real-time regime we need to have data transfer rate grater than data generation rate. First experiment at XFEL SPB station took place in December 2017. During 5 days of experiments about 115 Terabytes of data were generated. One can estimate data generation rate to be about 4.6 Gbps, taking for account that experiment time was about 12 hours per day. This rate is easilly achievable for big data centres such as Kurchatov Institute where several external 10 Gpbs network connections are available. However, for smaller centres network bandwidth tests and reservations should be performed. In the first SPB experiment in December 2017 we have used FTP service with multiple concurent connections to transfer the data. For automated data synchronization and replication one can consider distributed storage systems like dCache [32], which is widely used in LHC Computing Grid collaboration.

4.2. Computing resources

Data analysis includes three major steps: diffraction patterns classification, orientation determination and phase retrieval. For support vector machine (SVM) based classification method [19] we have measured the following performance results: 353 images per second when use of Intel Xeon E5-2650 v3 CPU facilities over 24 parallel threads and 354 images per second with 2xNVIDIA Tesla K80 GPU. Note that during latest experiment at XFEL, data generation rate was 270 diffraction patterns per second. In the future this rate is expected to be two orders of magnitude greater, but the fraction of informative images with diffraction patterns is estimated to be around 1%. Our computing experiments show that a single server with two GPU cards or with 24 CPU cores will be sufficient to classify patterns stream on the fly.

The most time consuming step is orientation determination. For model data we have the following measurements: one iteration of EMC algorithm on 1000 diffractions patterns using one Intel Xeon E5-2650 v3 core takes about 4200 seconds. During 4200 seconds a dataset of $4200 * 270 = 1.13 * 10^6$ diffraction patterns will be generated at the experiment. If use batch EM-mode we can distribute data across batches of diffraction patterns. Each batch will be processed separately. For one iteration per batch we will need about 1100 CPU cores. The precise estimation of optimal batch size and the number of iterations per batch is a subject of our further research. However, very rough estimations show that computing cluster with several thousands CPU cores will be sufficient to process the flow of diffraction patters from the experiment on the fly. The cluster does not require to be tightly connected (e.g. with InfiniBand fabric) because of independent processing of each batch. Up to the moment of writing no GPU-based EMC implementation is available.

The latest part of data analysis is phase retrieval. Using one Nvidia Tesla K80 and libspimage software [33] it takes about 600 seconds for phase reconstruction for 3D volume of size 183x183x183 pixels. In our software platform we plan to use phase retrieval in asynchronous manner in parallel with orientation determination. In this way one node with NVIDIA K80 GPU will be sufficient for this step of analysis. CPU-only computations for phase retrieval are also possible but take one order of magnitude more time.

5. SUMMARY

We presented a possible approach for analysis and processing of large data volumes generated during SPI experiments at a large scale facilities as the European XFEL. A combination of the HPDA approach and knowledge of diffraction physics allows to create an integrated platform which comprises a complete reconstruction pipeline from detector output to the 3D structure of a particle under investigation. The proposed platform may allow producing first reconstruction results within a day after the start of data collection in the SPI experiment. Application of the platform may dramatically reduce time for obtaining the first images of the reconstructed biological particles and will allow better planning of the whole SPI experiment.

Acknowledgments. We acknowledge fruitful discussions and careful reading of the manuscript by R.P. Kurta. The research was supported by Helmholtz Association's Initiative and Networking Fund and Russian Science Foundation (project No. 18-41-06001). This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute" (ministry subvention under agreement RFMEFI62117X0016), http://ckp.nrcki.ru/.

REFERENCES

1. A. Osseyran and M. Giles, *Industrial Applications of High-Performance Computing: Best Global Practices* (2015).
2. K. J. Gaffney and H. N. Chapman, *Imaging atomic structure and dynamics with ultrafast X-ray scattering*, Science **316** (5830), 1444–1448 (2007).
3. 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 (2000).
4. M. Altarelli, R. Brinkmann, M. Chergui, W. Decking, B. Dobson, S. Düsterer, G. Grübel, W. Graeff, H. Graafsma, J. Hajdu et al, *The European x-ray free-electron laser*, Technical design report, DESY **97**, 1–26 (2006).
5. A. Aquila, A. Barty, C. Bostedt, S. Boutet, G. Carini, D. DePonte, P. Drell et al. *The linac coherent light source single particle imaging road map.* Structural Dynamics **2** (4), 041701 (2015).
6. 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 et al, *Femtosecond diffractive imaging with a soft-X-ray free-electron laser*, Nature Physics **2** (12), 839 (2006).
7. M. M. Seibert, T. Ekeberg, F. Maia, M. Svenda, J. Andreasson, O. Jönsson, D. Odić, B. Iwan, A. Rocker, D. Westphal et al, *Single mimivirus particles intercepted and imaged with an X-ray laser*, Nature **470** (7332), 78 (2011).
8. M. F. Hantke, D. Hasse, F. R. N. C. Maia, T. Ekeberg, K. John, M. Svenda, N. Duane Loh, A. V. Martin, N. Timneanu, D. S. D. Larsson et al, *High-throughput imaging of heterogeneous cell organelles with an X-ray laser*, Nature Photonics **8** (12), 943 (2014).
9. T. Ekeberg, M. Svenda, C. Abergel, F. Maia, V. Seltzer, J. Claverie, M. Hantke, O. Jönsson, C. Nettelblad, G. Van Der Schot et al, *Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser*, Phys. Rev. Lett. **114** (9), 098102 (2015).
10. R. P. Kurta, J. J. Donatelli, C. H. Yoon, P. Berntsen, J. Bielecki, B. J. Daurer, H. DeMirci, P. Fromme, M. Hantke, F. Maia et al, *Correlations in scattered x-ray laser pulses reveal nanoscale structural features of viruses*, Phys. Rev. Lett. **119** (15), 158102 (2017).
11. A. Barty,R. A. Kirian, F. R. N. C. Maia, M. Hantke, C. H. Yoon, T. A. White and H. Chapman, *Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data*, J. Appl. Cryst. **37** (3), 1118 (2014).
12. K. Ayyer, T. Lan, V. Elser and N. D. Loh, *Dragonfly: an implementation of the expand–maximize–compress algorithm for single-particle imaging*, Int. J. Appl. Cryptogr. **49** (4), 1320–1335 (2016).
13. F. R. N. C. Maia, T. Ekeberg, D. van der Spoel and J. Hajdu, *Hawk: the image reconstruction package for coherent x-ray diffractive imaging*, J. Appl. Cryst. **43**, 1535 (2010).
14. B. Henrich, J. Becker, R. Dinapoli, P. Goettlicher, H. Graafsma, H. Hirsemann, R. Klanner, H. Krueger, R. Mazzocco, A. Mozzanica et al, *The adaptive gain integrating pixel detector AGIPD a detector for the European XFEL*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **633**, S11–S14 (2011).
15. S. A. Bobkov, *Comparison Study of Different Approaches to Classification of Diffraction Images of Biological Particles Obtained in Coherent X-Ray Diffractive Imaging Experiments*, Mathematical Biology and Bioinformatics **12** (2), 411–434 (2017).
16. F. Maia, *The Coherent X-ray Imaging Data Bank.* Nat. Methods **9** (9), 854–855 (2012).
17. B. J. Daurer, K. Okamoto, J. Bielecki, F. Maia, K. Mühlig, M. Seibert, M. Hantke, C. Nettelblad, W. H. Benner, M. Svenda et al, *Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses* IUCrJ **4** (3), 854–855 (2017).
18. J. Andreasson, A. V. Martin, M. Liang, N. Timneanu, A. Aquila, F. Wang, B. Iwan, M. Svenda, T. Ekeberg, M. Hantke et al, *Automated identification and classification of single particle serial femtosecond X-ray diffraction data*, Optics express **22** (3), 2497–2510 (2014).

19. S. A. Bobkov, A. B. Teslyuk, R. P. Kurta, O. Yu. Gorobtsov, O. M. Yefanov, V. A. Ilyin, R. A. Senin and I. A. Vartanyants, *Sorting algorithms for single-particle imaging experiments at X-ray free-electron lasers*, Journal of synchrotron radiation **22** (6), 1345–1352 (2015).
20. R. P. Kurta, M. Altarelli and I. A. Vartanyants, *Structural analysis by x-ray intensity angular cross correlations*, Advances in Chemical Physics **161** , 1–39 (2016).
21. N. D. Loh and V. Elser, *Reconstruction algorithm for single-particle diffraction imaging experiments*, Phys. Rev. E **80** (2), 026705 (2009).
22. A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 1–38 (1977).
23. M. Kevin, *Machine learning: a probabilistic perspective* (2012).
24. R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, 355–368 (1998).
25. M. A. Sato and S. Ishii, *On-line EM algorithm for the normalized Gaussian network*, Neural Comput. **12** (2), 407–432 (2000).
26. O. Cappé and E. Moulines, *On-line expectation–maximization algorithm for latent data models*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **71** (3), 593–613 (2009).
27. G. Celeux, D. Chauveau and J. Diebolt, *On stochastic versions of the EM algorithm* (1995).
28. S. F. Nielsen et al, *The stochastic EM algorithm: estimation and asymptotic results*, Bernoulli **6** (3), 457–489 (2000).
29. G. Wei and M. A. Tanner, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*, J. Amer. Statist. Assoc. **85** (411), 699–704 (1990).
30. J. R. Fienup, *Phase retrieval algorithms: a comparison*, Applied Optics **21** (15), 2758 (1982).
31. S. Marchesini, *A unified evaluation of iterative projection algorithms for phase retrieval*, Rev. Sci. Instrum. **78** (4), 011301 (2007).
32. G. Behrmann et al, *A distributed storage system with dCache*, Journal of Physics: Conference Series. Vol. **119**. No. 6. IOP Publishing, 2008.
33. F. Maia et al, *Hawk: the image reconstruction package for coherent X-ray*, Journal of applied crystallography, **43**, 1535-1539.