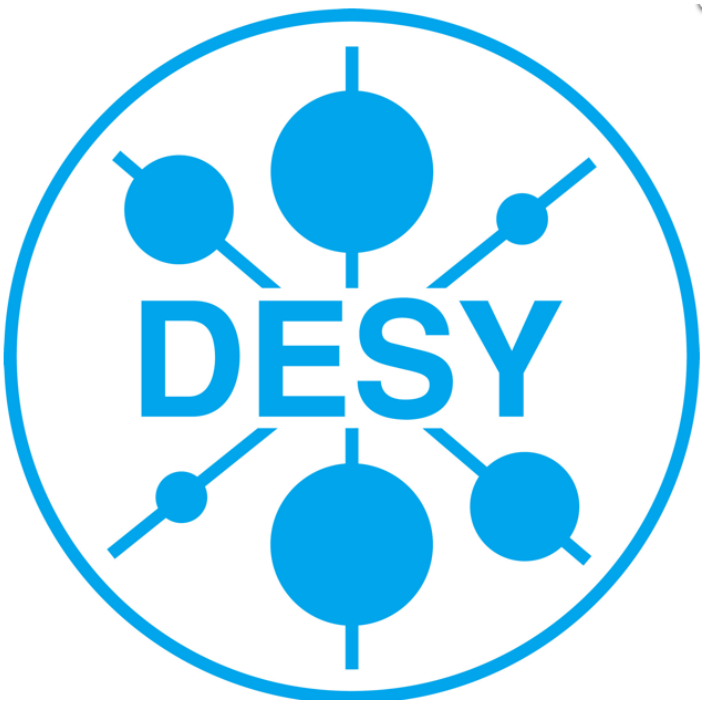

Single particle experiments feasibility at LCLS

Dameli Assalauova, dameli.assalauova@phystech.edu
Moscow Institute of Physics and Technology, Moscow, Russian Federation
CFEL, DESY, Hamburg, Germany

Abstract

The most common way to determine the structure of a biological specimen is X-ray crystallography. Unfortunately this method has several disadvantages, among which the most important one is the fact that not all structures can be crystalized. Therefore it is very important to find a way to determine the structure of single biological specimens. One of the very promising techniques is called single particle imaging (SPI). This method became possible after the invention of Free Electron Lasers (FEL) in 2004, but until now application of this method to real structure still struggles from many technical issues. Therefore it is very important to understand how the SPI experiments have to be performed to be successful. In this work we take experimental parameters (realistic flux distribution in x-ray focus and photon noise at the detector) from hard X-ray FEL build in Stanford (LCLS). We apply these parameters to simulations of different biological specimens to determine the range of particles which could be investigated at current generation of FELs. The diffraction patterns made in this way are oriented in 3D reciprocal space for future electron density reconstruction.

September, 2015

Contents

1 Introduction

Earlier the only way to reconstruct the structure of the molecules was crystallography. This is due to the fact that it is very difficult to get enough scattered signal from one protein, therefore having N unit cells the intensity in Bragg peaks grows as N^2. However crystallography has its disadvantages. It works only if protein can be crystalized. Plus often some proteins have to be modified to form a crystal. Another problem is that measurements are often performed at cryo-temperature, therefore it's impossible to measure biological specimen in native environment. Therefore we need to measure single specimens.

So why do people actually use crystallography? X-rays are weakly interacting with matter – that's why x-rays are used to 'look' through things. Therefore, as it was said earlier, one bio specimen scatters not enough x-ray photons. Of course one can increase the time of illumination, but the radiation damage will be too high and the sample just 'dies'. Free electron laser (FEL) opens a way to overcome this problem – using very intense and very shot pulses[1]. Structure survives during the pulse and then dies but we don't really care because scattered light is already measured.

One of the promising techniques for structure determination is called single particle imaging (SPI). This method requires many identical particles in random orientation, each illuminated by very intense femtosecond FEL pulse. In this way many diffraction patterns could be measured, each corresponding to different sections of the 3D reciprocal space of the object. Therefore we can assemble all patterns into a single 3D intensity distribution. And then reconstruct electron density (in real space) using standard phase retrieval technique.

The main problem is that each pattern (measurement) corresponds to unknown orientation of the sample. So we need to use orientation determination program, for example, EMC algorithm. Many others algorithms exist but EMC proved to be working. Getting the three-dimensional model of the particles or molecules from the big amount of diffraction two-dimensional 'snapshots' is what we are now trying to do.

Parameters of the SPI experiments have to be decided before the experiment. This is our task – to understand what has to be done for the SPI to be successful. Here we are trying to add simulated diffraction patterns (for some known object) with the realistic flux and real flux distribution in focus and real noise. After analysing this data we will have the opportunity to find the minimum size of the object which can be measured using most recent experimental conditions available at LCLS (CXI beamline).

2 Estimation of X-ray flux incident on samples in a liquid jet

A. LCLS experiments with Granulovirus (GV)

Typical experimental set-up is shown on the Figure 1. Single X-ray pulses hit every structure (crystal or molecule) and then diffraction patterns are read out by the detectors. To avoid the radiation damage pulses briefer than the timescale of most damage processes are used.

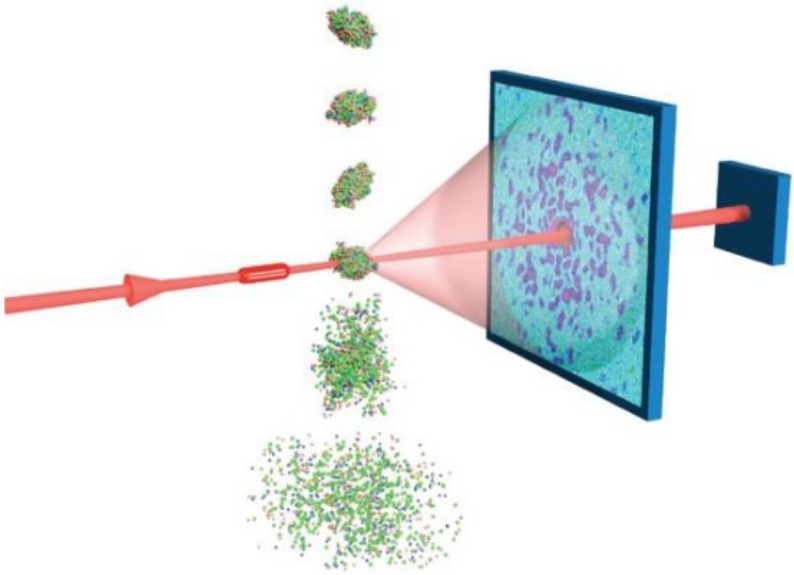

Figure 1. Experimental set-up.

One of the critical parts of an SPI experiment is sample delivery system. Currently gas-jet injectors are used. These injectors are spraying particles into the chamber ideally into the x-ray beam focus. Now this system is not very well controllable; therefore most of the x-ray pulses do not hit the injected structures. Moreover due to the fact that the particle 'beam' size is about 100 um in diameter and the x-ray focus is of the order of 1 um the event of illumination of a particle by the centre of the beam is very rare. So most of the patterns with measurable signal correspond to the case of illumination of a particle by the 'wings' of focused beam. Our first goal is to understand the distribution of the flux which illuminates different particles.

For such estimation x-ray serial femtosecond crystallography experiments were provided. Such experiments are very similar to the SPI experiments; the difference is just that small crystals are used instead of single particles. However the amount of data measured during such experiment is very high and the process of hit-finding (sorting out which measured patterns are 'good') is much easier. Here comes one issue: crystals used for such experiments are usually rather wide size distribution. Therefore scattered signal can depend not only on the

incident intensity but also on the crystal size and its quality. To overcome this problem we need data measured for crystals with very reproducible size and quality[2].

Luckily such crystal exists – Granulovirus[3]. These crystals look like 'sausage' with the diameter of approx. 200 nm and the length of 400 nm (Figure 2). The data from three different experiments in 2012, 2013 and 2015 were used. Parameters like number of patterns, energy, focus and numbers for converting counts to photons are presented in the Table 1.

Figure 2. Granulovirus structure.

Year		**Number of patterns**	**Energy**	**Focus**	**Adu_per_eV**
October 2012	Chapman	31190	6 keV	100 nm	0.00105
February 2013	Seeman	61133 (181 'best')	7.9 keV	1um	0.00105
June 2015	Obertuer	69162 (170 'best')	8 keV	100 nm	0.00338

Table 1. Parameters of the experiments ('best' are the most intensive patters).

The program that was used to analyse the measured patterns is called CrystFEL. It consists of several programs that deal with viewing, indexing, integrating and other steps of data analysis. Indexing process allows getting intensities for each h, k, l (Miller indices) of the structure. On the graph (Figure 3) integrated intensity of the peaks predicted by CrystFEL for each pattern can be seen[4]. Histograms for each three experiments were plotted. The intensity of all (indexed and real) Bragg peaks is the horizontal axis and the number of the patterns is on the vertical axis. It can be noticed that the number of the most intensive patterns is dropping.

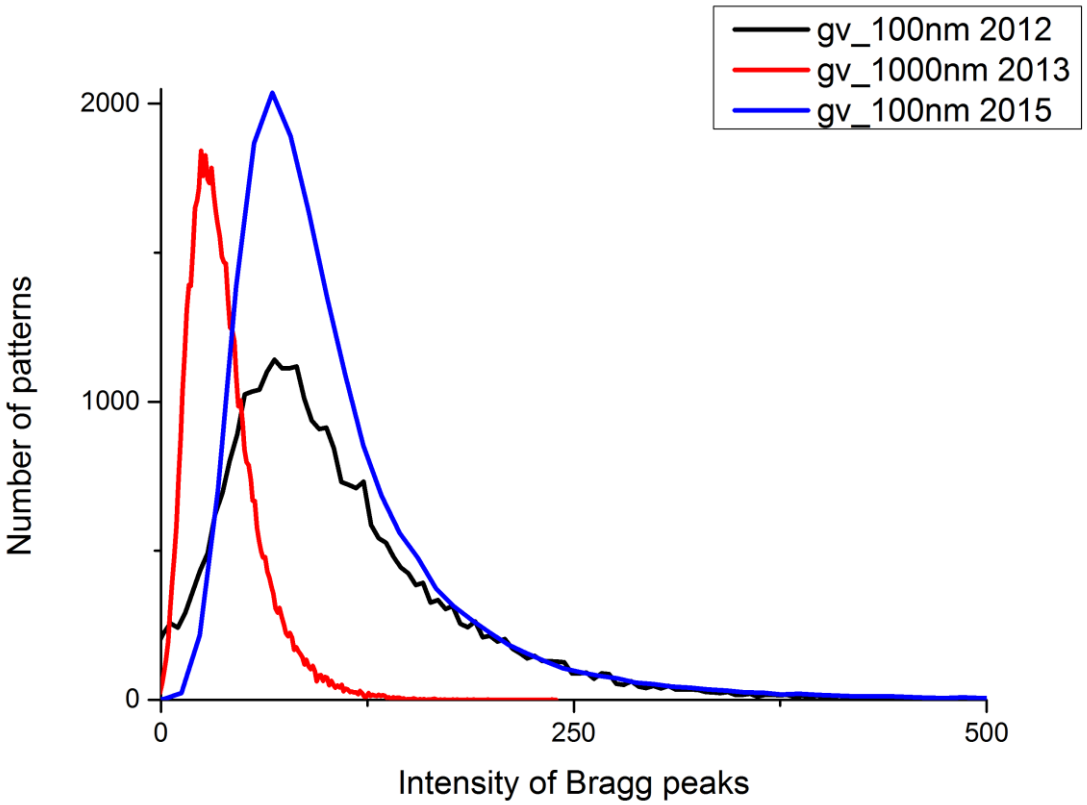

Figure 3. Intensity of Bragg peaks in three experiments.

B. Simulation of the GV patterns

Diffraction patterns were simulated for a GV structure using program 'Moltrans' with the flux density 10^{13} photons in 1 um^2. Scattering from each atom in a unit cell was calculated at each pixel of the detector and summed coherently for all atoms. Then the scattered complex value at each pixel was modified taking into account relative positions of all unit cells in the crystal.

Simulation parameters: central part of CS-PAD detector (just 400x400 pixels 110um each), distance to the detector is 100 mm.

The example of the cross section in reciprocal space and one Bragg peak can be seen on the Figure 4.

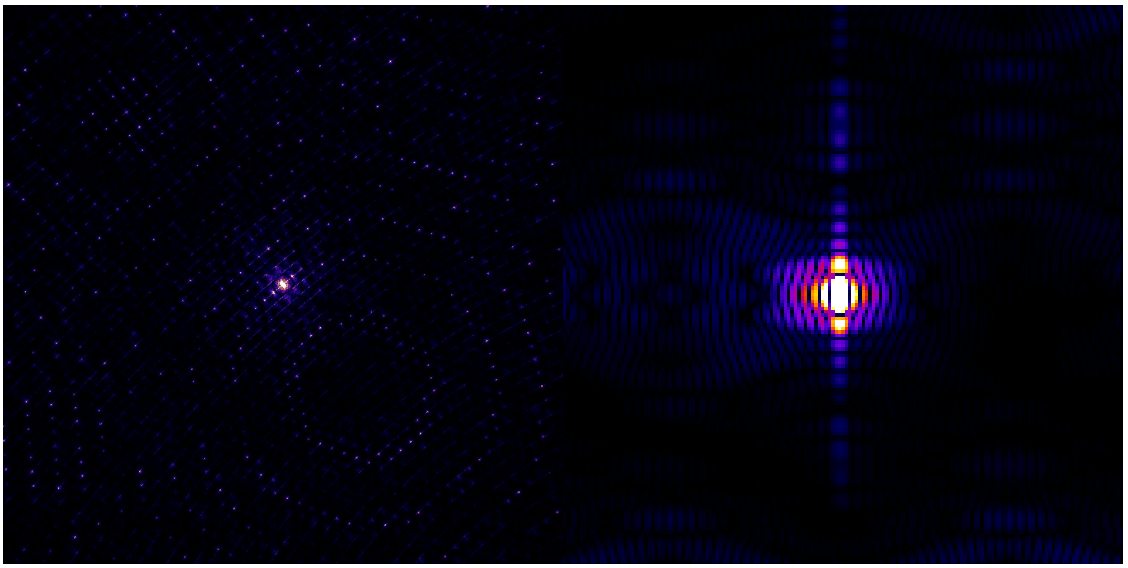

Figure 4. GV cross section and one Bragg peak.

The simulated curve is added to the graph on the Figure 5. It looks a little bit strange – too wide. However we expected the peak like experimental curves. The probable reason could be the following. The intensity in each pixel in simulation corresponds to the center of the pixel, while in the experiment it corresponds to the total signal measured in that pixel.

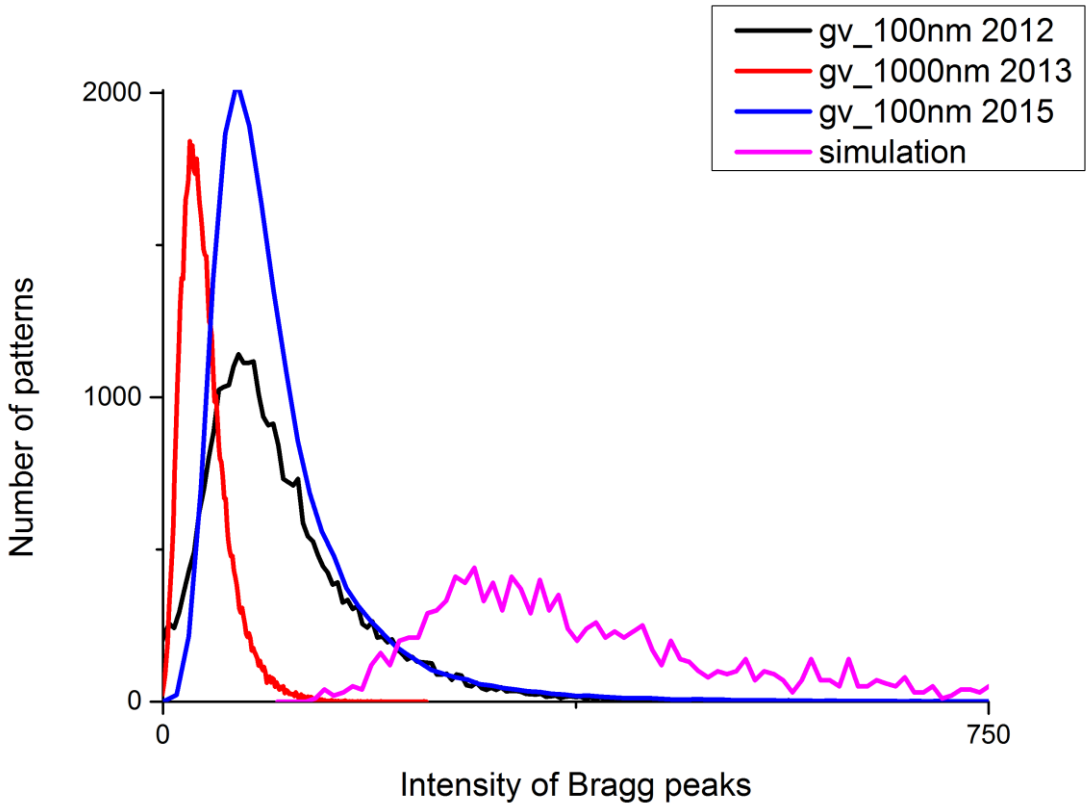

Figure 5. Intensity of Bragg peaks in three experiments and simulation.

It is better to see it on the Figure 6. First two graphs show the intensity distribution corresponds to the horizontal axis (Figure 4) and second two graphs show the intensity distribution corresponds to the vertical axis (Figure 4).

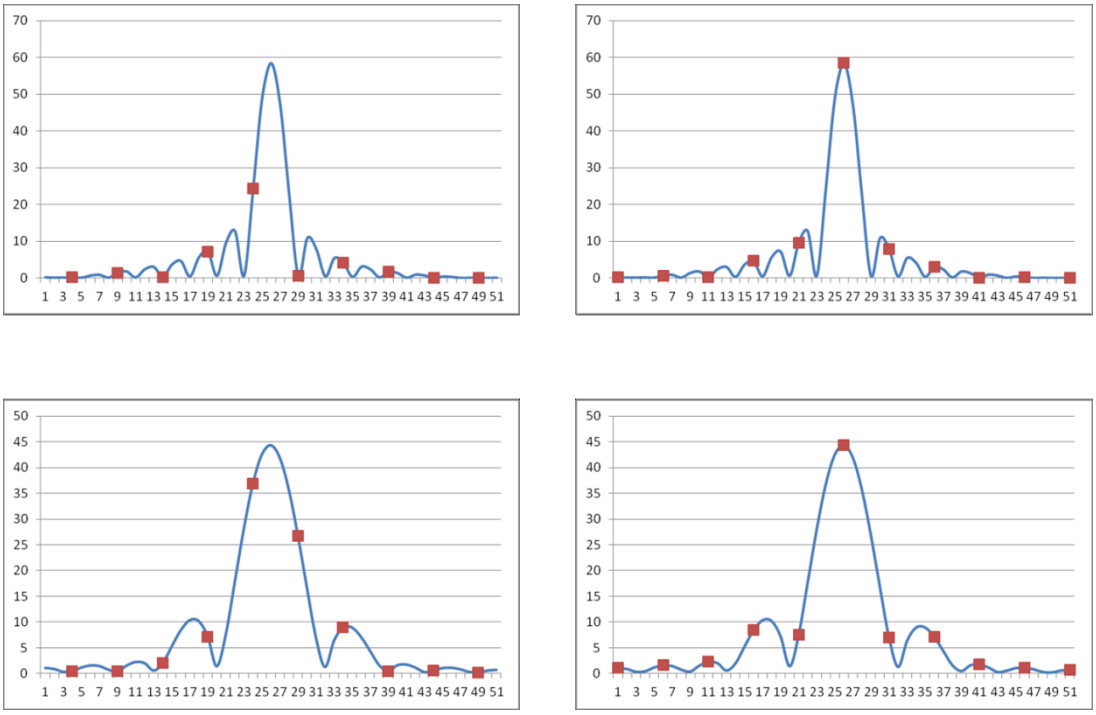

Figure 6. Simulated (blue curve) and experimental (red points) plots for the GV.

So as it can be seen that the integrated intensity using the experimental data can be lower than the integrated intensity using the simulated data especially with the horizontal axis.

To check this out it was decided to simulate the data with the unit cell. Multiplied the intensities in the expected Bragg peaks pixels by squared N2 (as it was claimed in the Introduction) we got the result presented on the slide. It appeared to be very close to the most intensive patterns in the first simulation. It can be seen on the Figure 7 – the intensity-q graph.

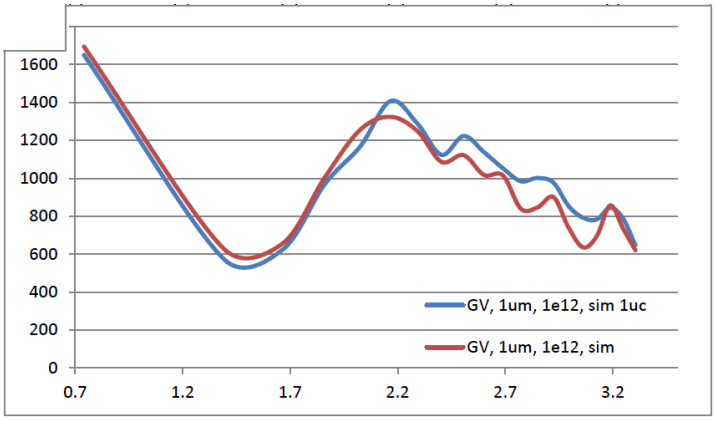

Figure 7. Comparison of the GV simulation and ‘unit cell’ simulation.

C. Comparison of experimental and simulated results

Finally using the simulation parameters (10^{13} incident photons, 1 um^2 focus, 8 keV energy) the graph (Figure 5) can be rescaled.

So all these analysis shows that we can take the flux density distribution from the June'15 experiment with the maximum flux 10^{13} photons in $1um^2$ (blue curve at Figure 8).

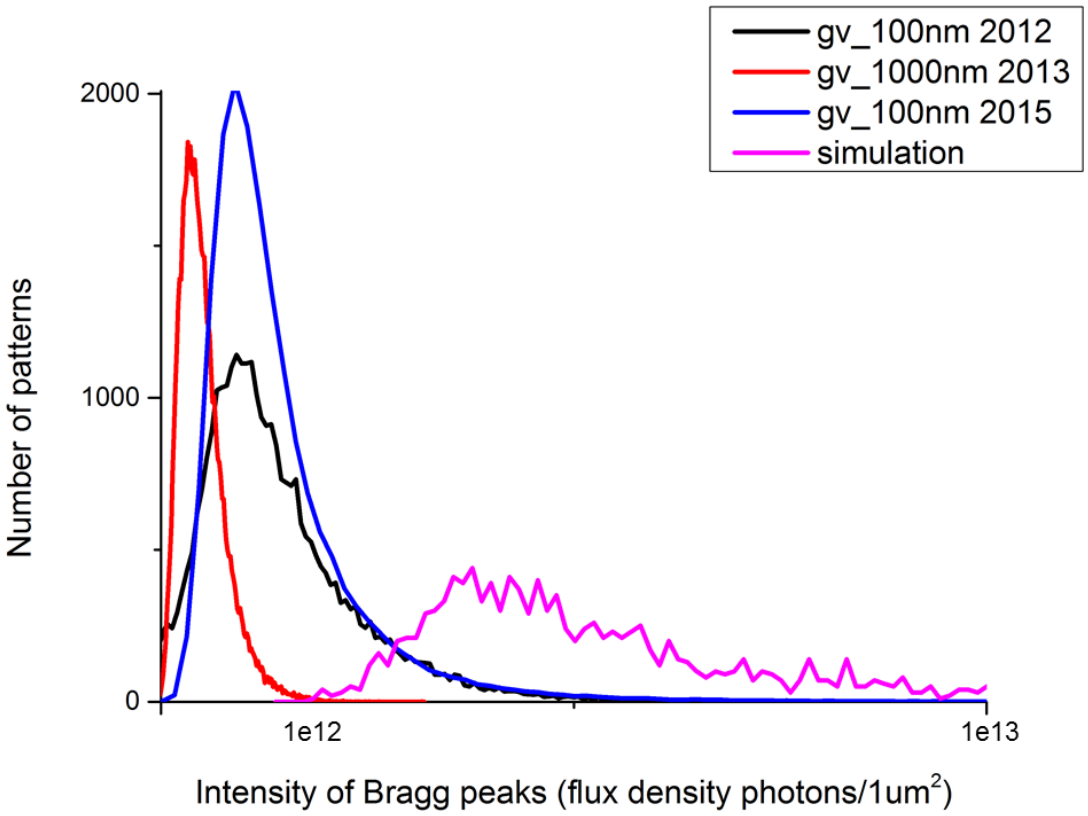

Figure 8. Flux density for three experiments and simulation graph.

Another thing that should be checked is how much we can win using the 100 nm focus instead of 1000 nm focus.

Using another parts of CrystFEL program the graph 'intensity-q' for each experiments can be plotted (Figure 9).

We split the data to "all" (all data) and "best" (small portion of the most intense patterns). The numbers for different experiments are presented on the Table 1.

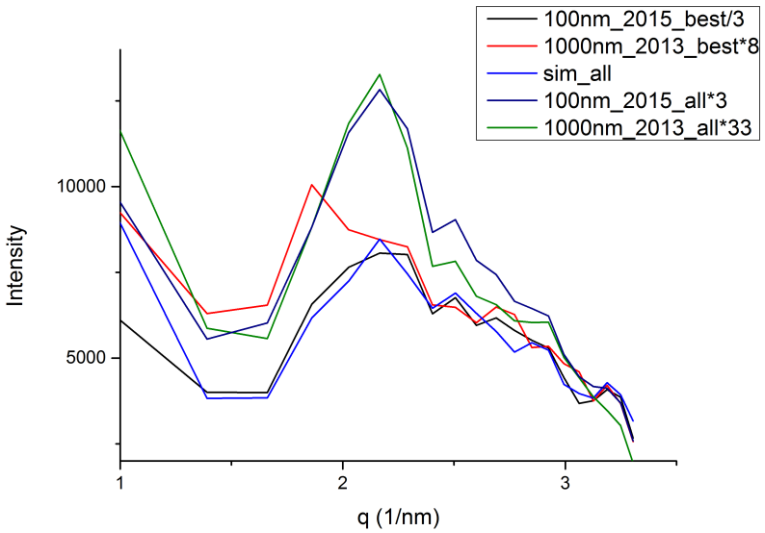

Figure 9. Intensity-q graph for 100 nm and 1 um experiments. 'Best' are the most intensive patterns.

To compare the results we were trying to match curves 100nm_2015_all and 1000nm_2013_all (and the same with the 'best' patterns) multiplying them with different factors. As it can be seen on the graph (Figure 9) the gain is not so big. The order is only 10 instead of the expected 100. This can be explained by the fact that focusing mirrors for 100 nm focus are not as good and for 1um focus. Also probably designed focus 100 nm is in fact close to 200 nm (factor of 4 already).

3 Simulation of single particle experiment

A. Applying flux distribution and noise from SPI LCLS experiment

More than 60000 diffraction patterns were simulated using the program 'Moltrans'. The next step was to apply realistic photon flux parameters with Poisson noise. This was done using the following parameters: flux density distribution (from June'15 experiment) with highest flux corresponding to 10^{13} photons, focus size = 0.1um by 0.1um. A typical diffraction pattern is shown in Figure 10.

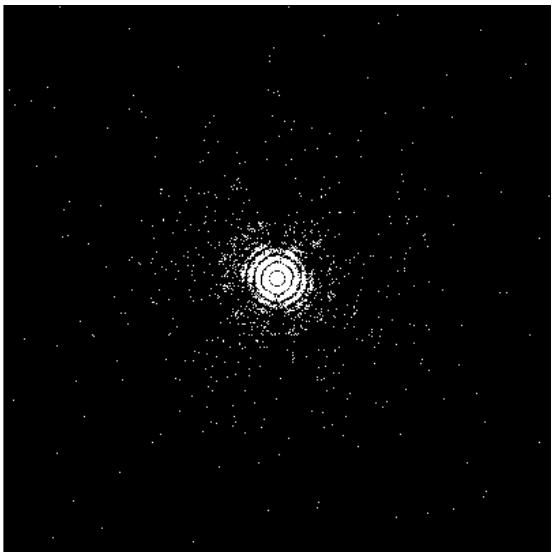

Figure 10. Simulation with the realistic flux density distribution.

The next step was to create a mask - detector file to find out which pixels were really usable and which pixels did not need to take into account. The idea of the generation was rather simple: while the simulations used flux density distribution from the June'15 experiment, for the mask file the intensities of all pixels were defined as '1'. So the detector area stayed with the value '0' and all other pixels turned to '1'. After that the values were inverted. Thus the detector file with 'good' pixels as '0' and bad pixels as '1'was created (Figure 11).

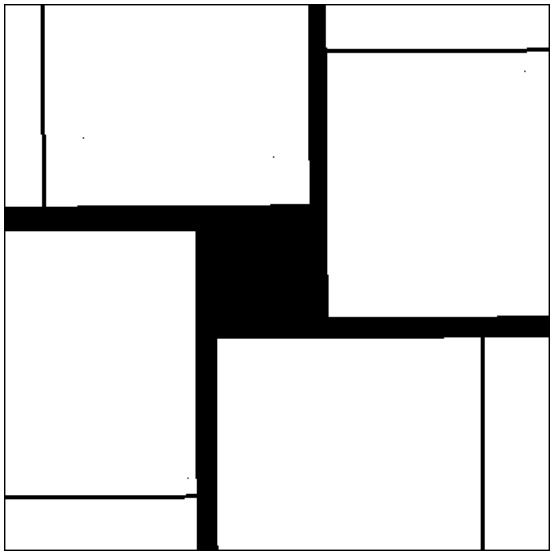

Figure 11. Mask (detector file).

The last step was to get experimental background from 2015 SPI experiment. All measured data is kept in '.h5' format. Each '.h5' file corresponds to one run and contained more than 10000 frames with different background. Most of the measured frames contained only beamline and detector noise due to very low hit rate.

Each simulated diffraction pattern (with flux scaled according to flux density distribution found from GV experiment) was summed with the background extracted from a randomly chosen frame from randomly chosen '.h5' file. After adding to the simulation background the intensities in each pixel were multiplied with the mask. Combined all these three components we finally got the dataset that can be used in EMS reconstruction algorithm. The examples of the 'modified' diffraction patterns are shown on the Figure 12.

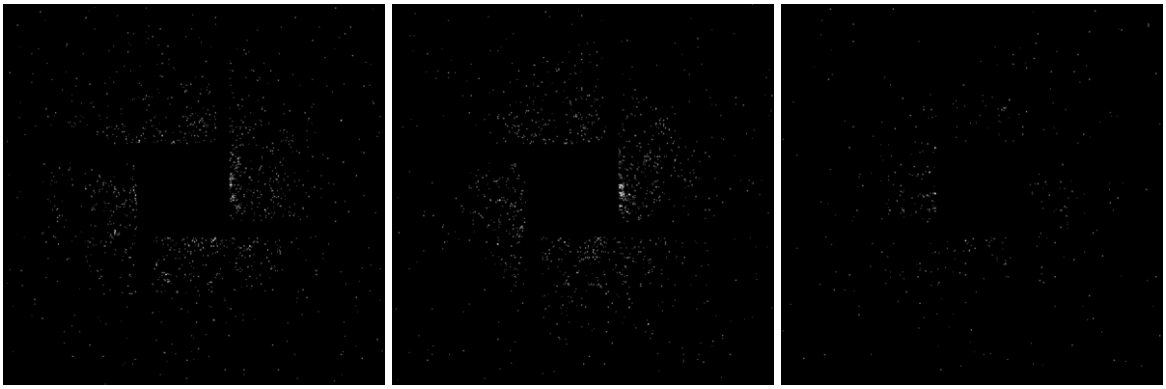

Figure 12. Examples of the 'modified' diffraction patterns.

B. Reconstruction of the 3D reciprocal space for samples

The EMC reconstructing algorithm is using a big amount of two-dimensional measurements. As it was said at the beginning the particle orientation is unknown. To do reconstruction of the single molecule it is needed to collect correct data with acceptable noise and resolution.

The EMC consists of three steps: expansion (E), maximization (M) and compression (C). E-step: expand the grid intensities into the tomographic representation, M-step: update the tomographic intensities by expectation maximization, C-step: compress the tomographic model back into a grid model[5].

One of the doctors in CFEL Coherent Imaging Team Kartic Ayyer successfully run this algorithm but only with the simulation and the detector file without background. So the next step in studying single particle imaging is to add the real background and get the 3D electron density to the molecules.

We're still working on accomplishing this step because it is crucial for new SPI experiments.

4 Conclusions

We have analyzed data from three experiments with Granulovirus: from October 2012, February 2013 and June 2015. Using CrystFEL program the Bragg peaks were predicted and three histograms were plotted (number of patterns vs intensity of Bragg peaks). To define flux density diffraction patterns from Granulovirus were simulated and matched with the simulations for only one unit cell. Then they were compared with the experimental data. It was also checked that using 100 nm focus does not bring so many advantages instead of the 1 um focus as it was expected.

Next part of the work was to simulate many (>60000) diffraction patterns for a single biological specimen (RDV) to test reconstruction algorithm. Obtained from GV experiments flux density distribution was applied to the simulated patterns and noise from SPI LCLS experiment was added. The resulting realistic diffraction patterns were multiplied by the detector mask to have only the information we can currently measure at LCLS.

Now we are in process of running reconstruction EMC algorithm. First of all we will trying to check EMC algorithm only with the simulation and the mask. If it works correctly we can do it with the real background from the June'15 experiment. When it works with RDV we're going to apply the same procedure to smaller artificial biological samples (20-30 um in diameter).

Finally we will get all information we need: the size of the dataset and the number of the iterations we need to reconstruct the structure of the single particles. And the most important we will be able to answer the question "What is the smallest size of the molecule we can measure at LCLS and reconstruct using this EMC?" This is really needed for the future SPI experiment to understand current limitations and what has to be done to overcome these limitations.

5 Acknowledgements

I would like to thank my supervisors Oleksandr Yefanov and Kartik Ayyer for explaining all the details and answering all my questions. Their patience and help during this project have supported me all the time during the summer school, their ideas and good mood were really inspiring.

I also would like to thank summer school organizing team. They made this time unforgettable and also provide a lot of support for each of the summer students.

To sum up the DESY summer school was a unique opportunity and great experience. I would recommend it to any student!

6 References

[1] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, J. Hajdu, Potential for biomolecular imaging with femtosecond X-ray pulses, *Nature, vol 406, 2000.*

[2] H. N. Chapman, Femtosecond X-ray protein nanocrystallography, *Nature, vol 470, 2011.*

[3] M. M. Seibert, Single mimivirus particles intercepted and imaged with an X-ray laser, *Nature, vol 470, 2011.*

[4] Thomas A. White, Richard A. Kirian, Andrew V. Martin, Andrew Aquila, Karol Nass, Anton Barty and Henry N. Chapman, CrystFEL: a software suite for snapshot serial crystallography, *Applied Crystallography, 45, 2012.*

[5] Ne-Te D. Loh, V. Elser, Reconstruction algorithm for single-particle diffraction imaging experiments, *Phys. Rev. E80, 026705, 2009.*