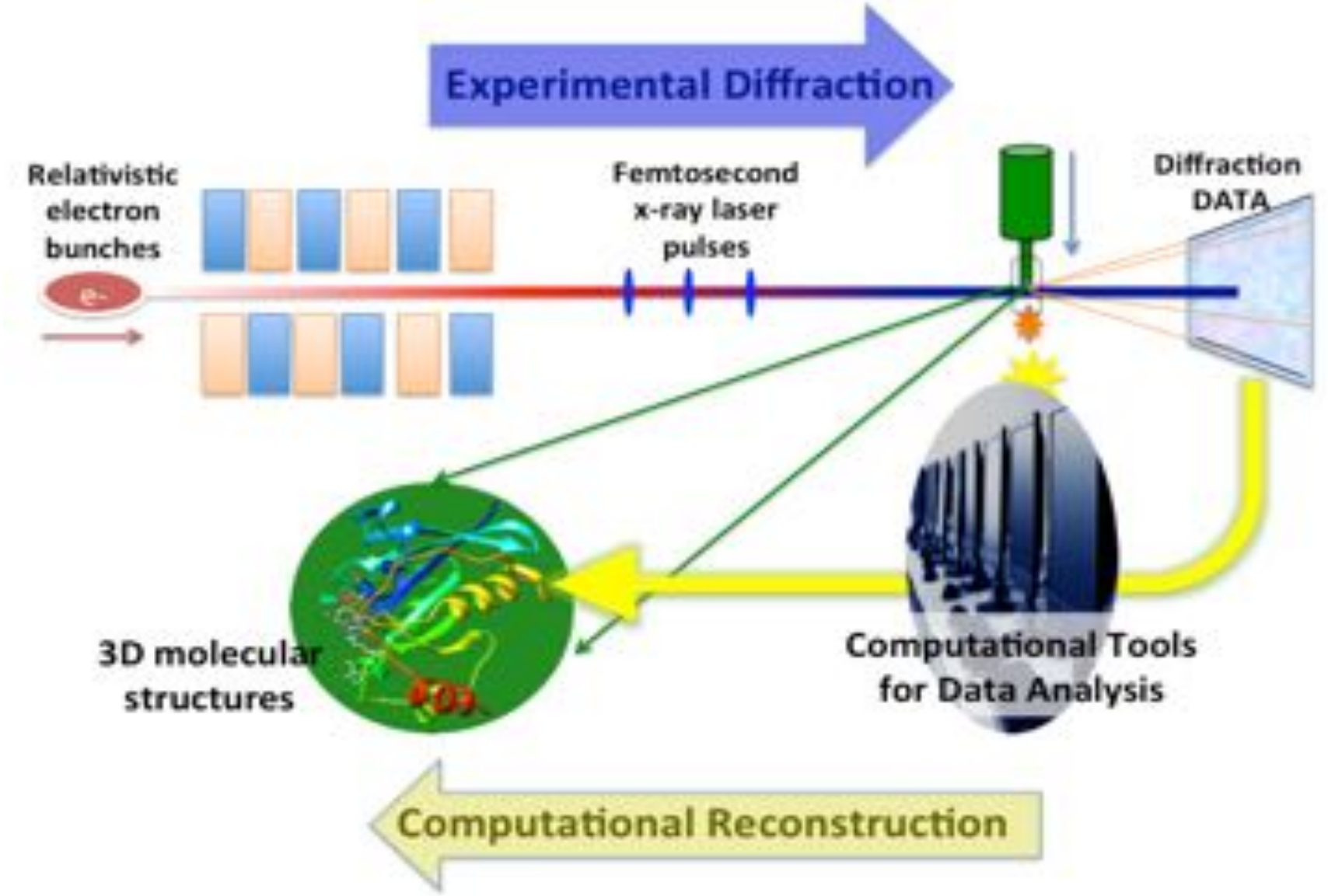
Workshop on Computational Methods in Bio-imaging Sciences

XFEL single particle scattering data classification and assembly

Haiguang Liu

Beijing Computational Science Research Center

XFEL imaging & Model Reconstruction

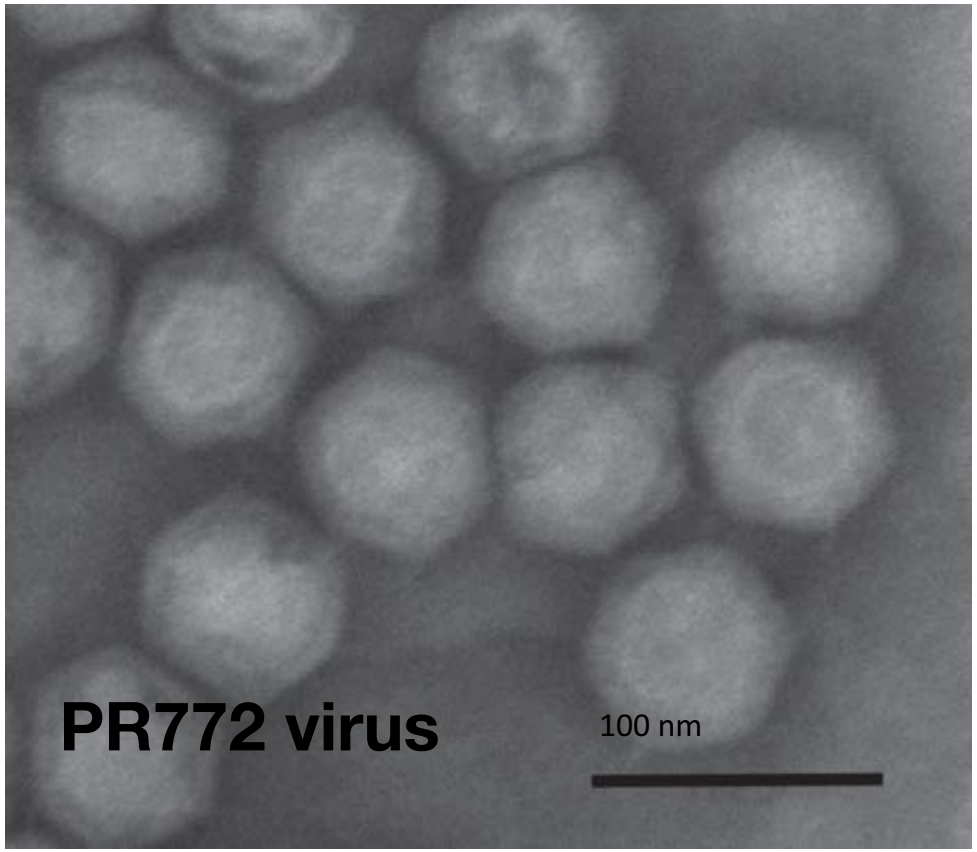


Translate experimental data to 3D models

Outline

1. Single Particle Scattering Imaging using XFEL
2. Correlation Function and its applications
   - Extracting structure information
   - Facilitating the recovery of orientations
3. Sample Heterogeneity
   - Nanoparticle (size variation)

Single Particle Scattering using X-ray Free electron laser with virus samples



LCLS single particle imaging consortium effort: using **PR772 virus** as one of the model systems to investigate the application of XFEL in biological particle structure determination

Phage DNA virus
diameter ~70nm

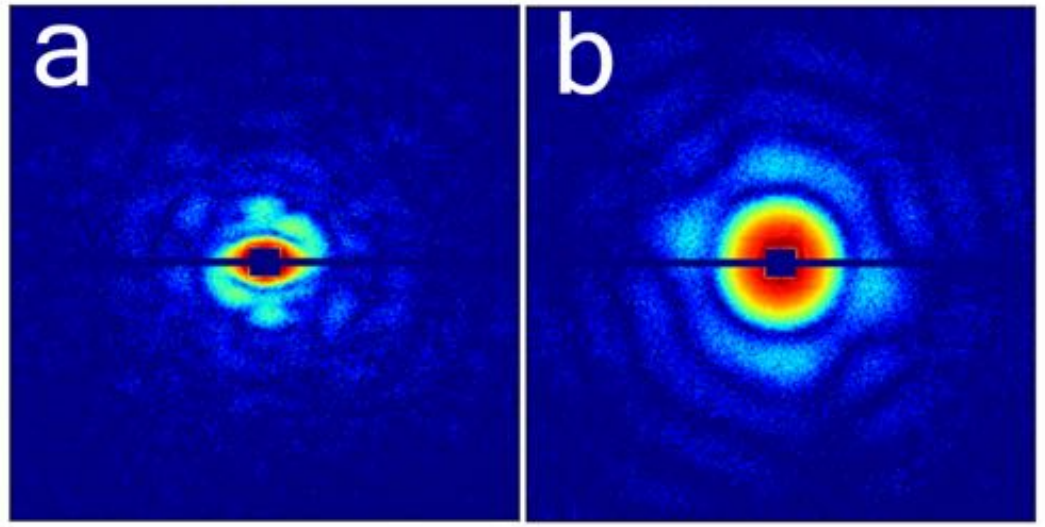
Reddy et al., Scientific Data 4, Article number: 170079 (2017)

XFEL single particle data for PR772

The complete dataset we used contains 64,511 diffraction patterns (260x257) pixels, after hit-finding and down-sampling using Cheetah.

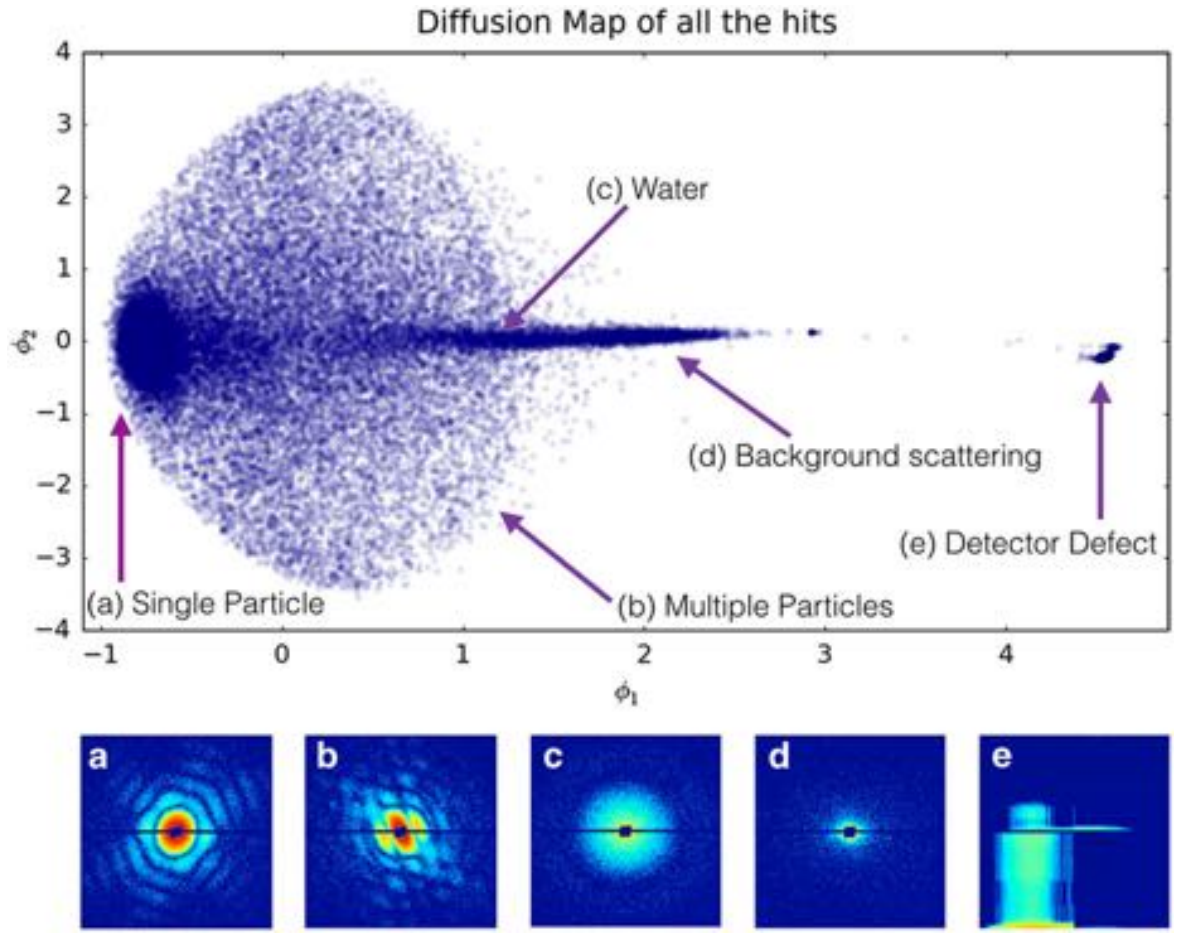
Four steps to complete the data processing:

1). Convert electron signals (ADU) to photon counts
2). Select single hits: the patterns resulted from single PR772 particles
3). Recover orientations and merge patterns to 3D diffraction volume
4). Phasing



(a)Non-single hit and (b) single hit patterns.

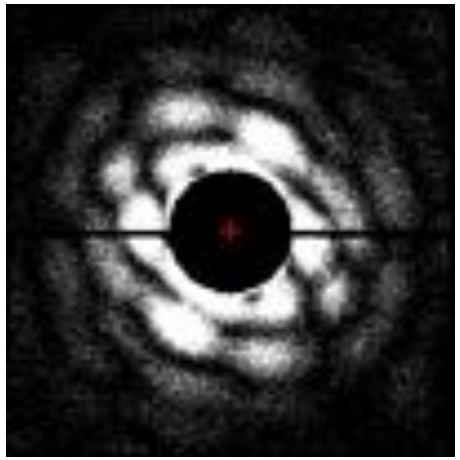
PR772 Single particle scattering data classification



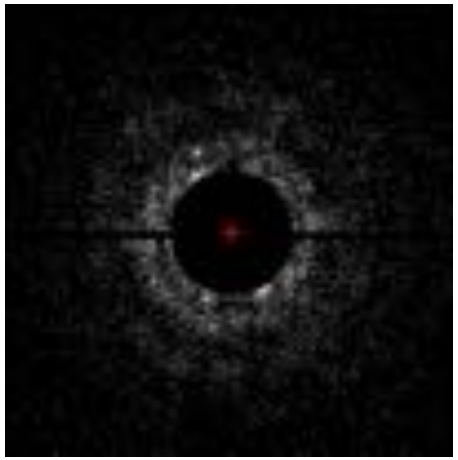
- Signal is weak!
- **unwanted data**
  - multiple particles
  - water droplet
  - unknown source
  - defective images

Reddy et al. Coherent Soft X-ray Diffraction Imaging of Coliphage PR772 at the Linac Coherent Light Source. *Scientific Data* (2017)

The diffusion map method identified 14,722 patterns from single particles, but we can observe these:



multiple particle



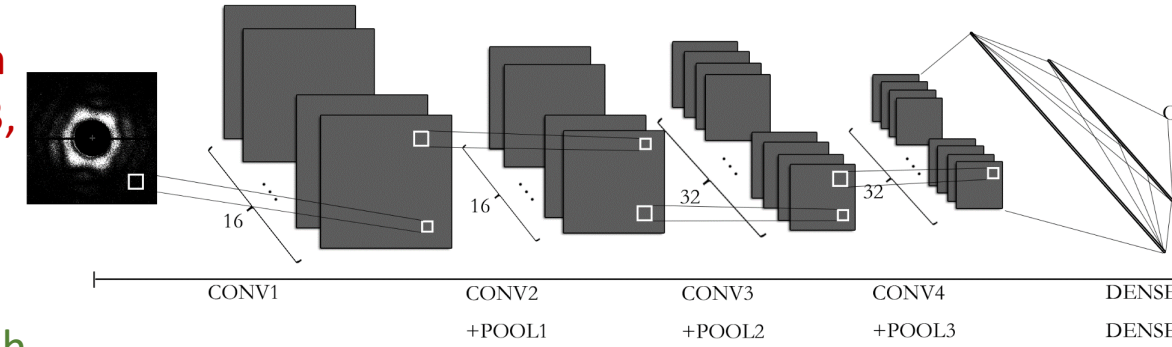
water droplet

low SNR

Patterns classification

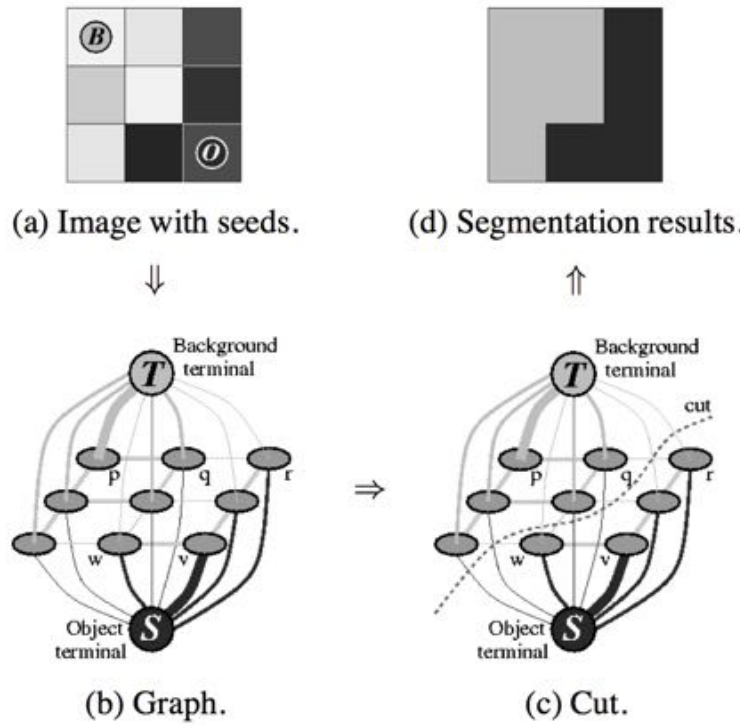
Methods

• Convolutional neural network (CNN) : The network is based on VGG16 net. We reduced the number of convolutional layers to 3, added batch normalization and drop-out layers after every convolutional layer. Total parameters ~300.



• Graph-cut method (GC) : We build a weighted undirected graph whose vertices represent the patterns and the edge weights are based on the Euclidean distance of pairwise patterns. A variational method based on Potts model is proposed for the partitioning of the graph, where each vertex is assigned a score between 0 and 1, indicating the likelihood of belonging to a specific cluster.

• Manifold - Diffusion map (DM) : Diffraction data published in CXIDB



Training set

Randomly select **200** patterns from whole data set. **79** of them are single particle scattering patterns.
(A little over-fitted for CNN as there are more parameters than images, but we add drop out layers to relax that.)

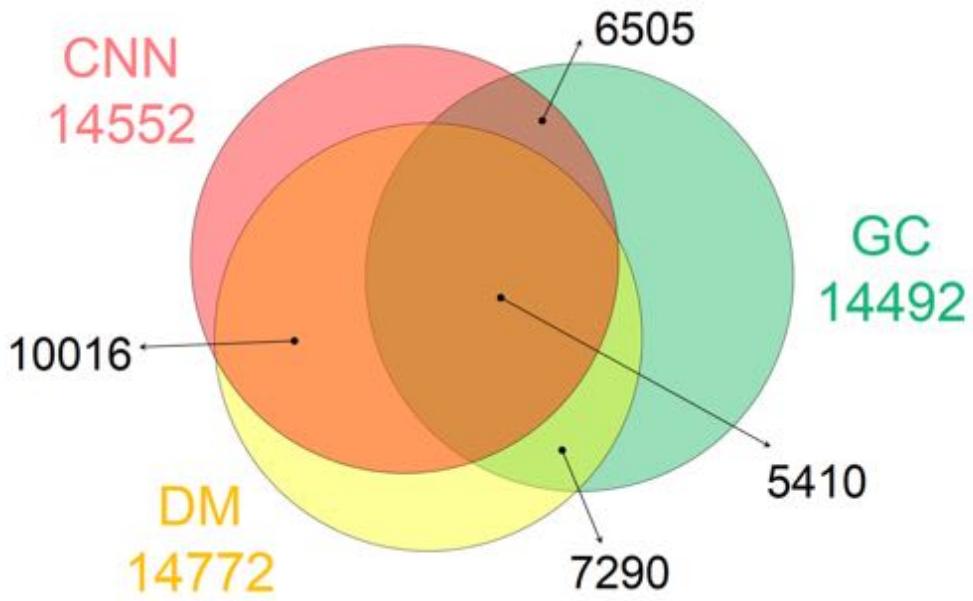
Single hit identification Summary

Methods	CNN	Graph cut	Diffusion map
Number of output single-hits	14552	14492	14772

Time consumed

Algorithms	Hardware	Time
CNN	K80 GPU	~10 min
GC	Xeon CPU / Ten Cores*	~15 min
DM	Xeon CPU / Ten Cores*	~70 min

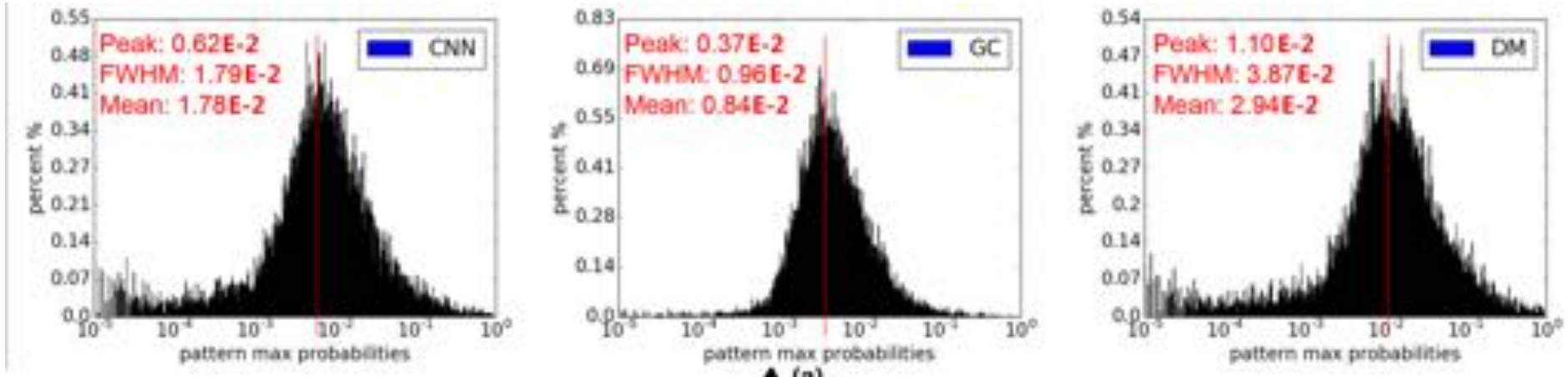
*The table lists a maximum number of cores available for work.
CPU utilization depends on algorithms.



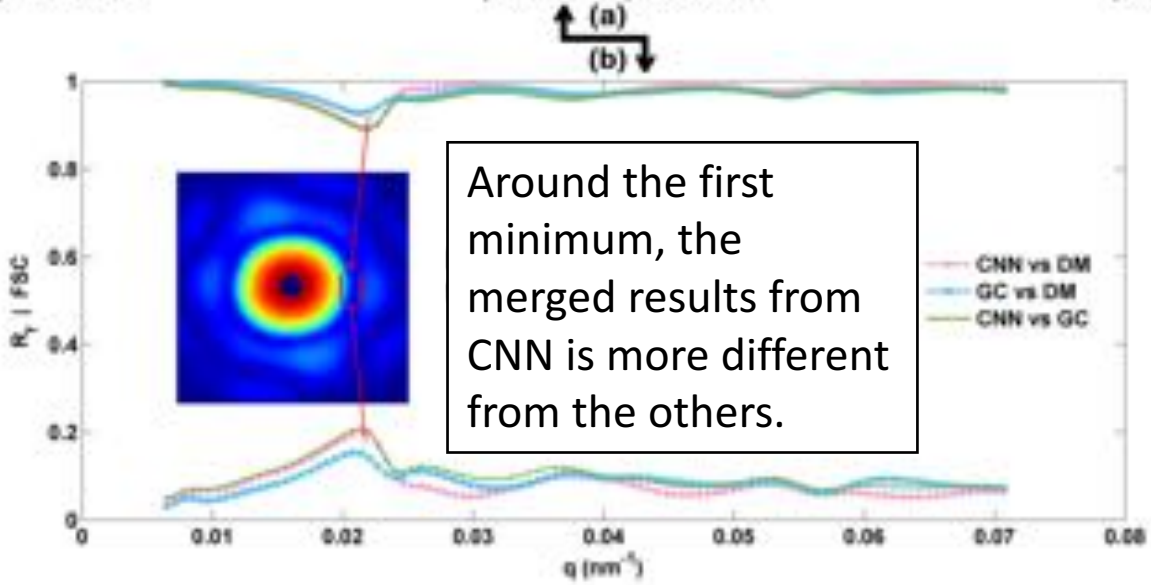
3D Merging --- Dragonfly

Parameters:
For every 10 step, beta=beta*1.414 (beta_init=0.006).
No symmetry assumption.
Run 30 steps.

(a) Orientation Specifity:
The maximum probability (among all orientations) of every pattern
Comparable!



(b) Pairwise FSC/R-factor:
Results from CNN are slightly different from the other two.
Overall, very similar!



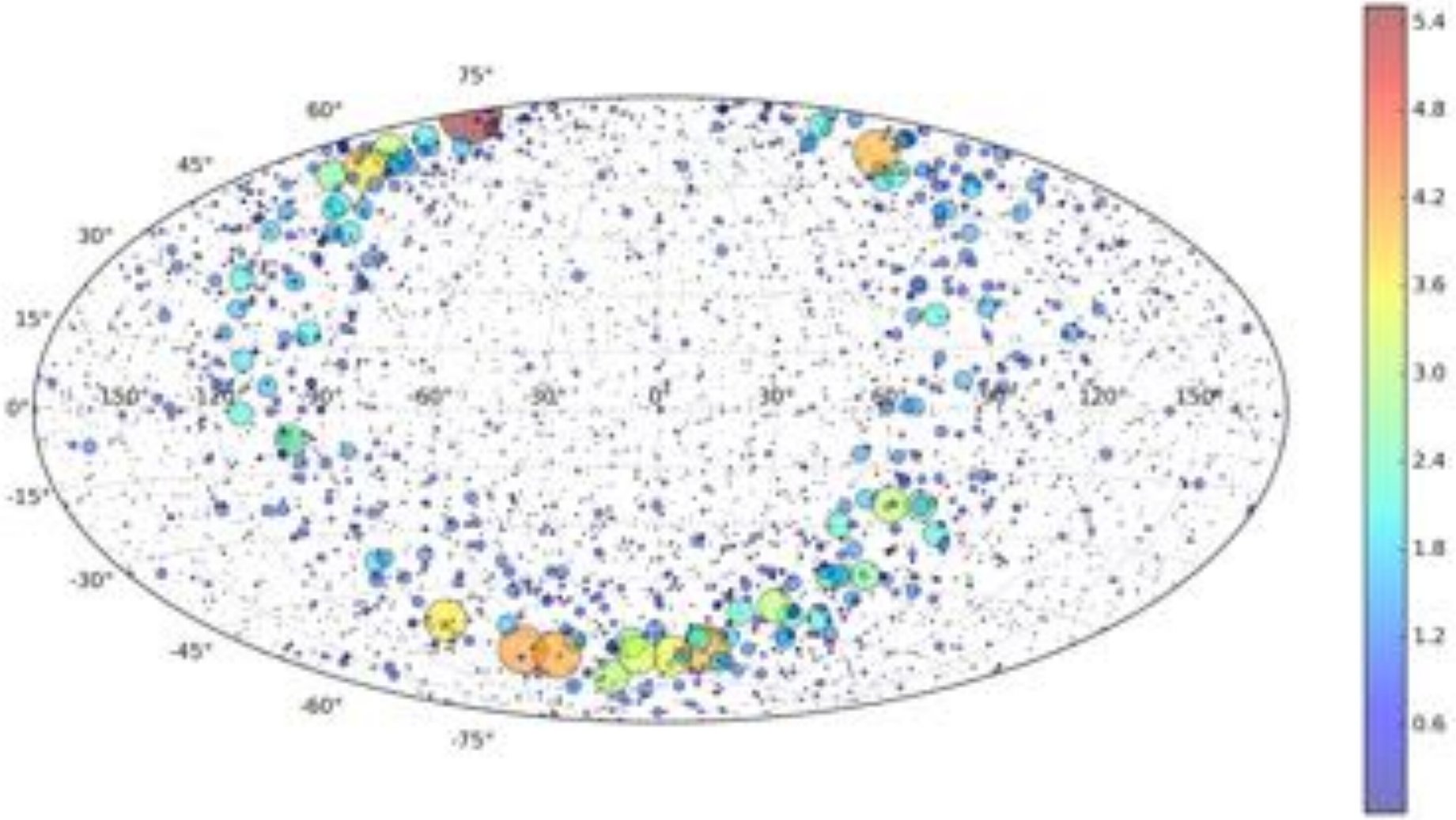
Around the first minimum, the merged results from CNN is more different from the others.

Orientation bias?

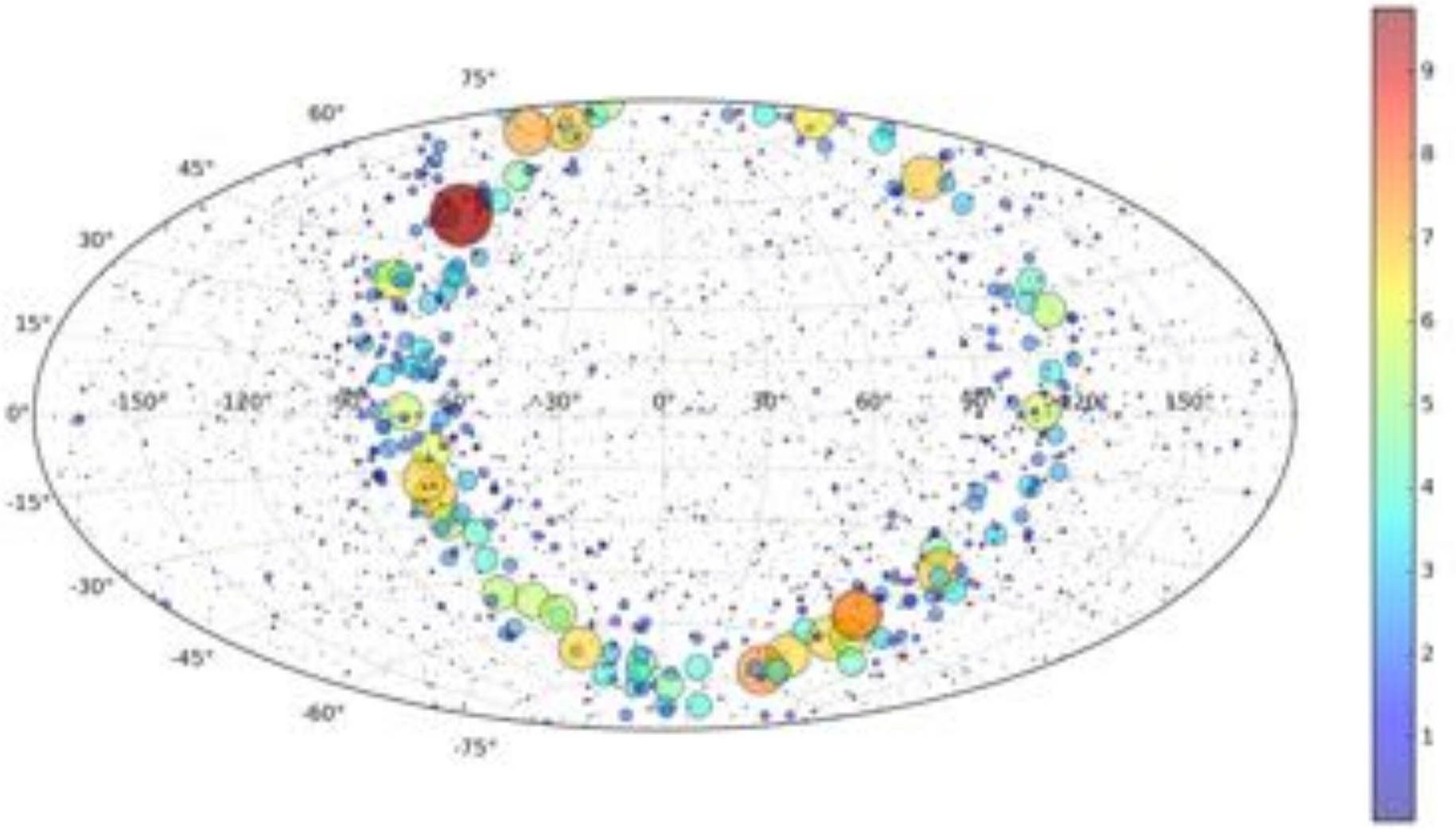
- Orientation distribution estimation
    - For each pattern, 10 most probable orientations and their probabilities are aggregated by accumulating **possibilities** for any orientation bin
    - In-plane rotation not distinguished (ONLY looking at the ray directions)

- The orientations are specified using quaternions, who are used to rotate a vector located at (longitude=0,latitute=0) to its destination.

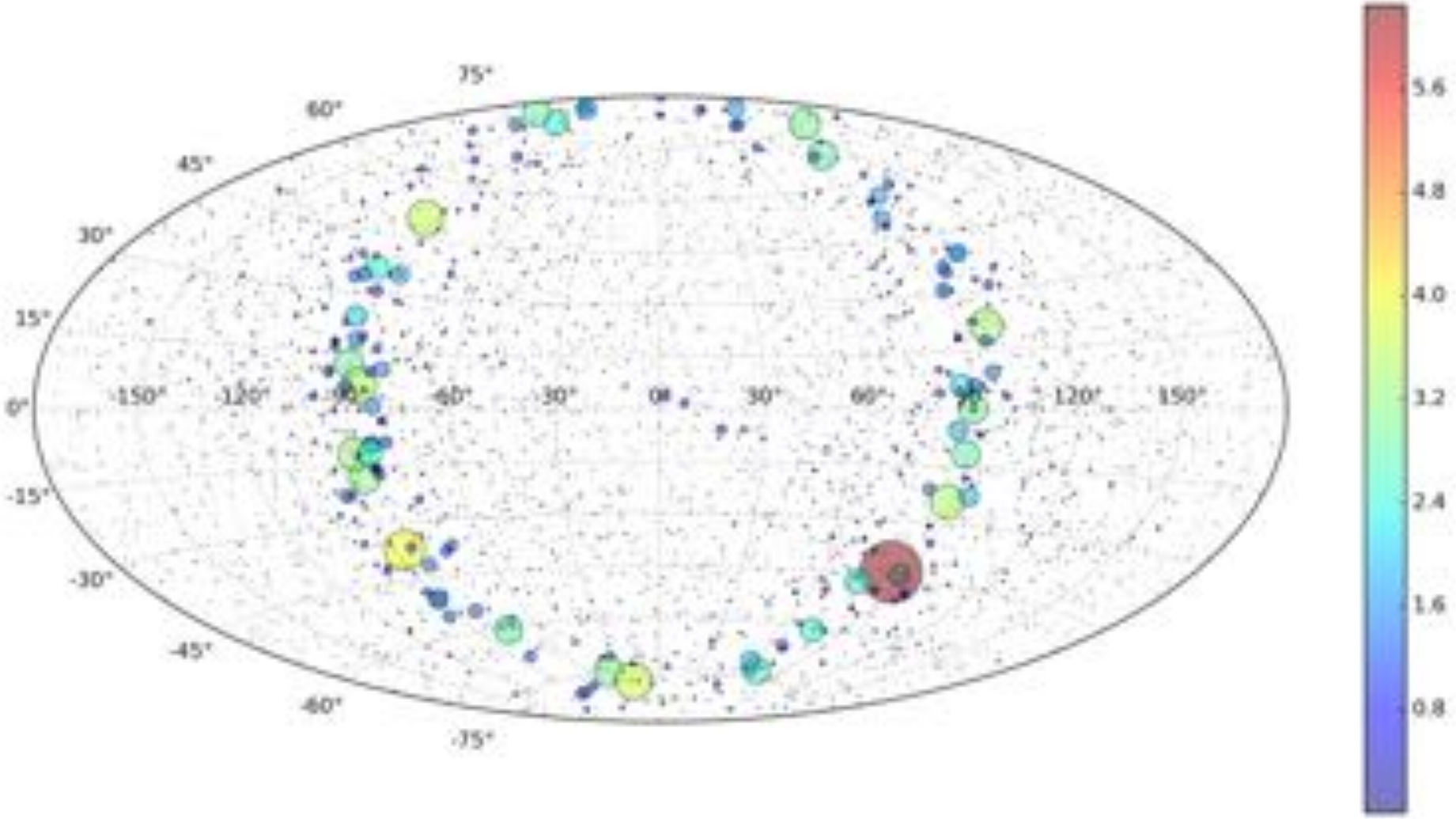
orientation distribution (CNN)



orientation distribution (DM)



orientation distribution (Graph-Cut)

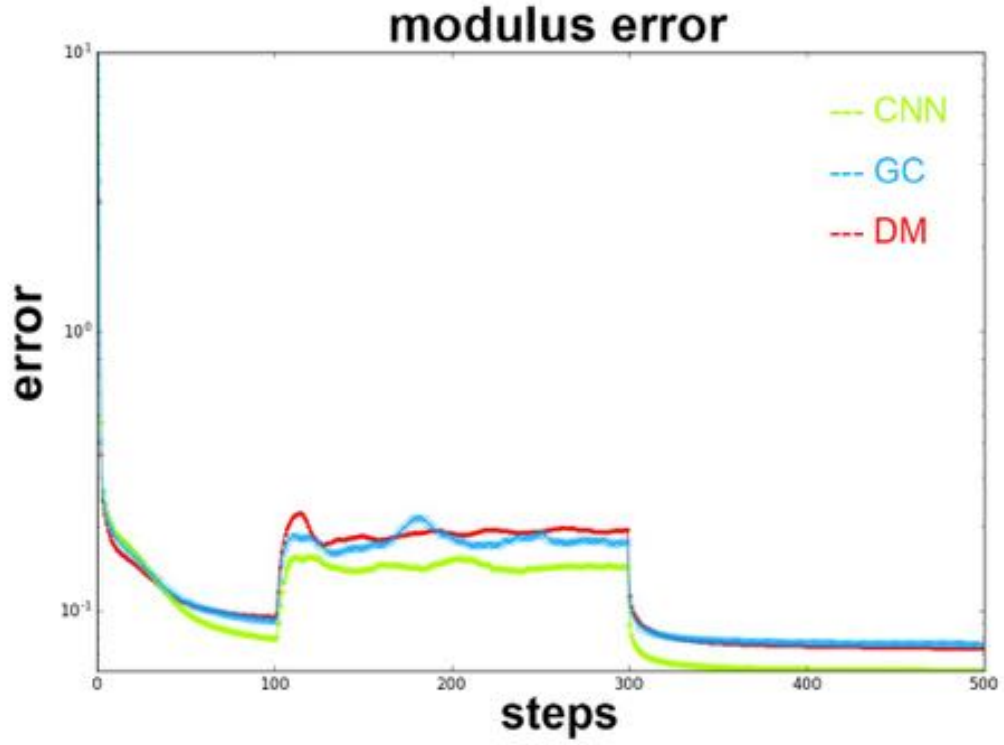


Phase retrieval

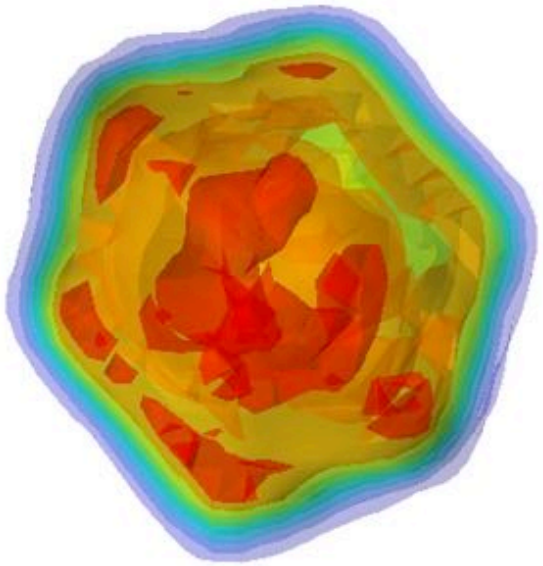
Iterative methods (Andrew Morgan's 3D phasing):

100 error reduction steps +
200 difference map steps +
200 error reduction steps

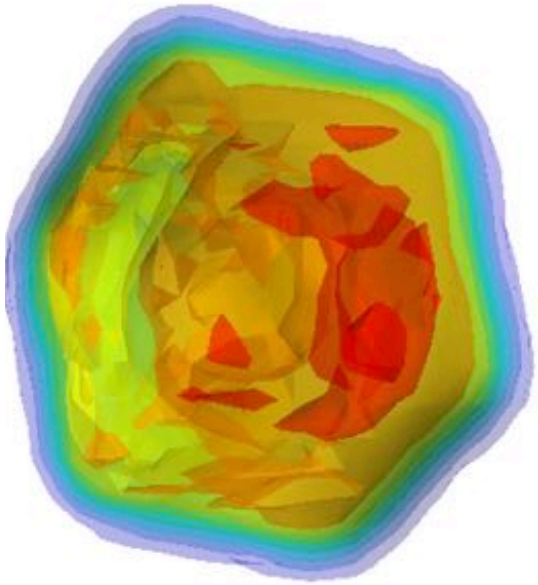
Average models from 40 runs for the final model.



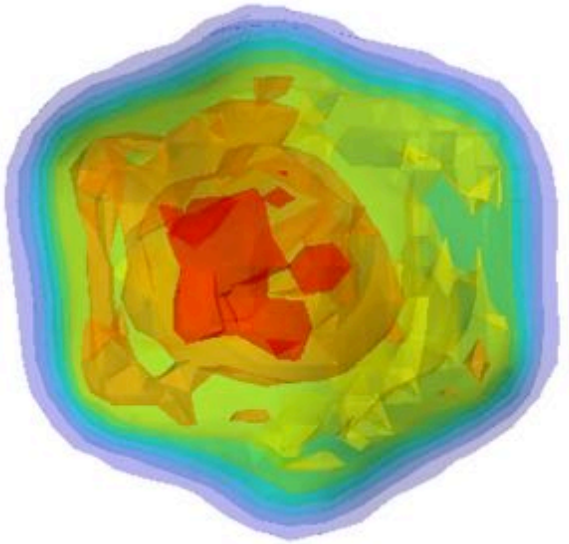
3D maps after Phasing



CNN Structure
14.7 nm



DM Structure
11.6nm



GC Structure
15.1nm

Outline

1. Single Particle Scattering Imaging using XFEL

2. **Correlation Function and its applications**

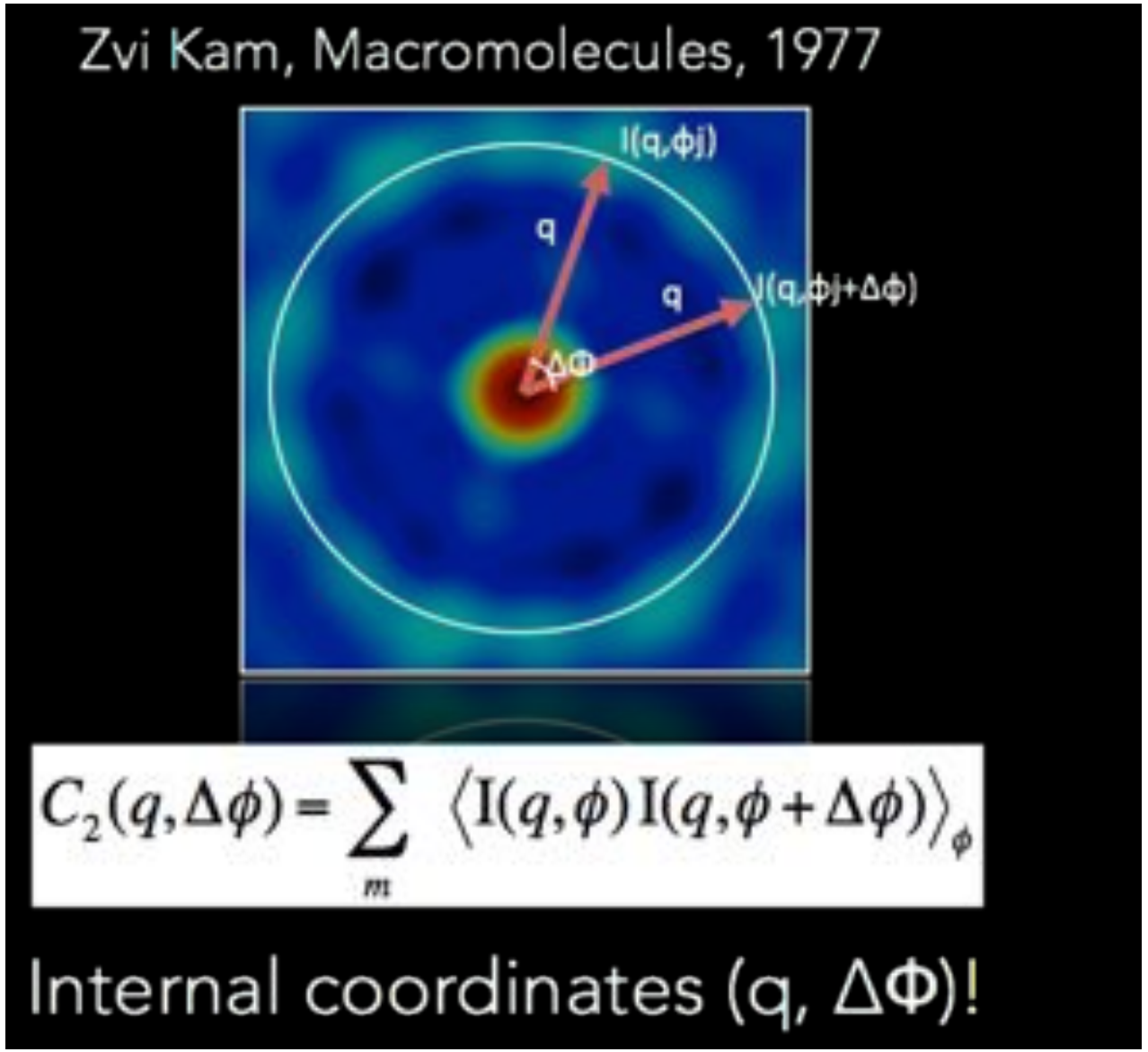
   - Extracting structure information
   - Facilitating the recovery of orientations

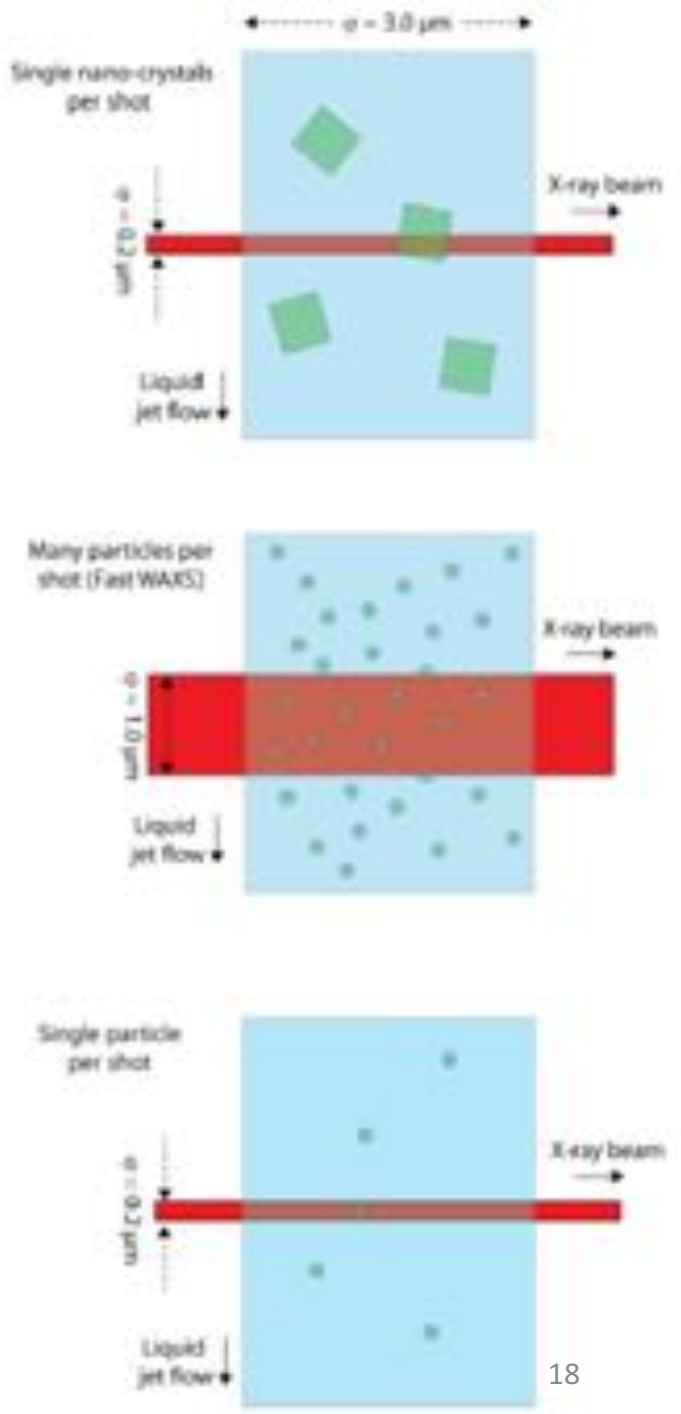
3. Sample Heterogeneity

   - Nanoparticle (size variation)

Correlation function

Structure information embedded in internal coordinates





Correlation Scattering, Flucutation Scattering, Snapshot Scattering, Fast Solution Scattering, etc.

exposure time << rotation time

Intensity Correlations

Correlation between intensities at (q1,ϕ1) and (q2,ϕ2)

$$C_2(q_1,q_2,\phi_1,\phi_2) = Cor(I(q_1,\phi_1),I(q_2,\phi_2))$$

For unknown orientations, the correlation needs to be integrated over angular parameter

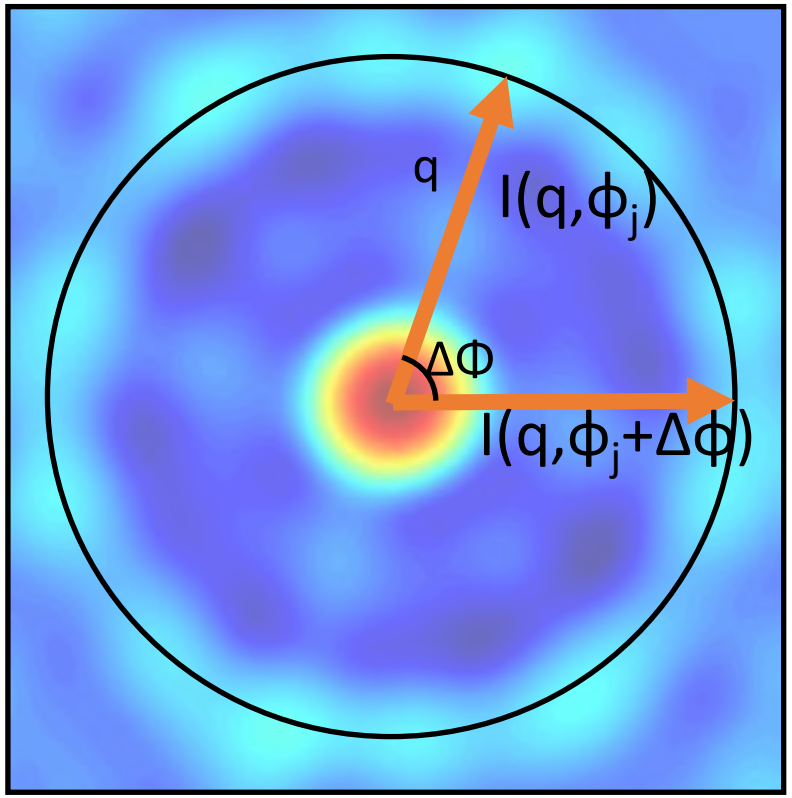
$$C_2(q_1,q_2,\Delta\phi) = \left\langle I(q_1,\phi)I(q_2,\phi+\Delta\phi)\right\rangle_\phi$$

Auto-correlation: a special case

$$C_2(q,\Delta\phi) = \left\langle I(q,\phi)I(q,\phi+\Delta\phi)\right\rangle_\phi$$

In 3D, hoping N patterns samples all orientations evenly

$$C_2(q,\Delta\phi) = \frac{1}{N}\sum_i C_2^i(q,\Delta\phi)$$



PROFILE AND REAL SPACE MODEL

$$\rho(\mathbf{r}) = \sum_{n=0}^{N_{max}} \sum_{l=0}^{n} \sum_{m=-l}^{l} c_{nlm} R_{nl}(r) Y_{lm}(\omega)$$

$$A_q(\omega_q) = 4\pi \sum_{l}^{n_{max}} \sum_{m=-l}^{+l} a_{lm} Y^*_{lm}(\omega_q)$$

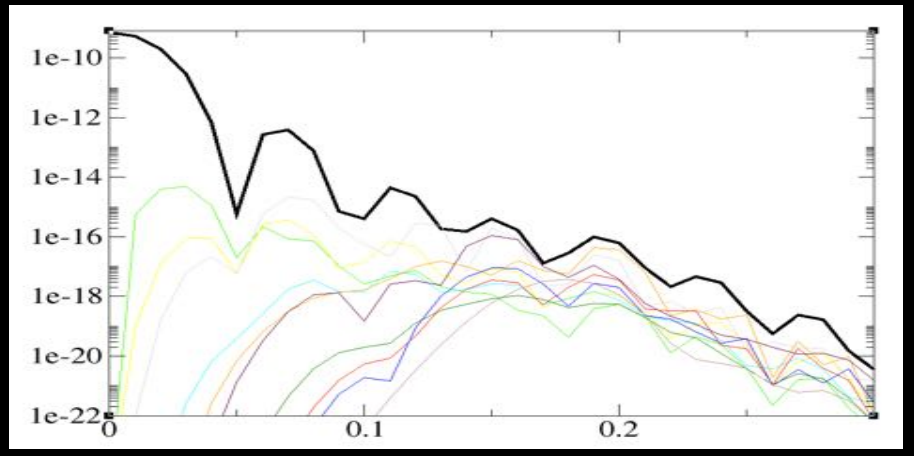
$$a_{lm} = \sum_{n}^{n_{max}} w_{nl}(q) c_{nlm}$$

$$I_{lm} = \sum_{l'} \sum_{l''} \sum_{m'} \sum_{m''} a_{l'm'} a_{l''m''} G \begin{pmatrix} l & l' & l'' \\ m & m' & m'' \end{pmatrix}$$

$$B_l(q) = \sum_{m} |I_{lm}(q)|^2$$

Now, straightforward to go from real space to profile.
Reverse Modelling.

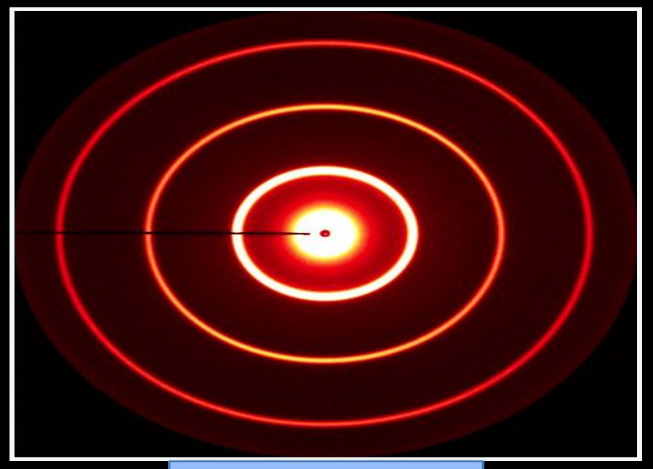
$B_l(q)$



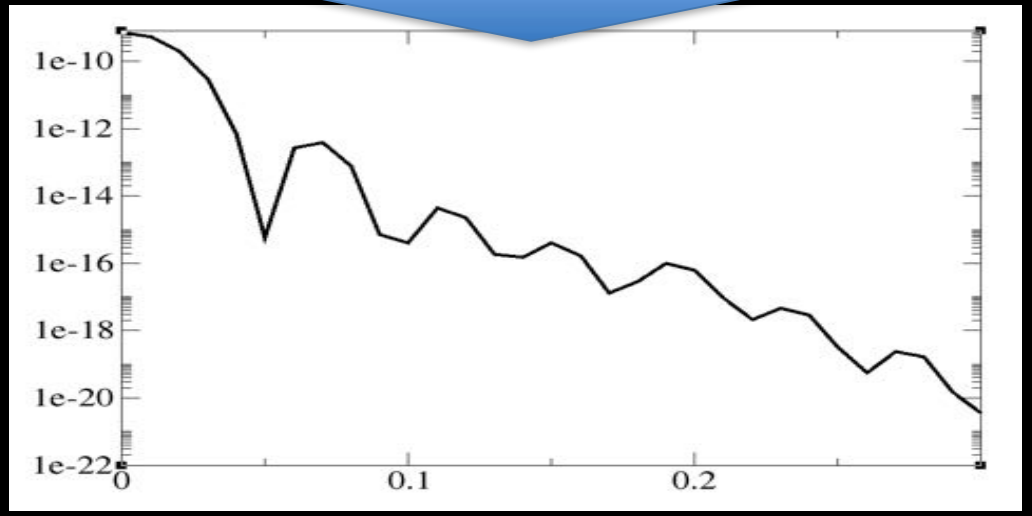
Liu *et al.* Acta Cryst. A **68**, 561-7 (2012)

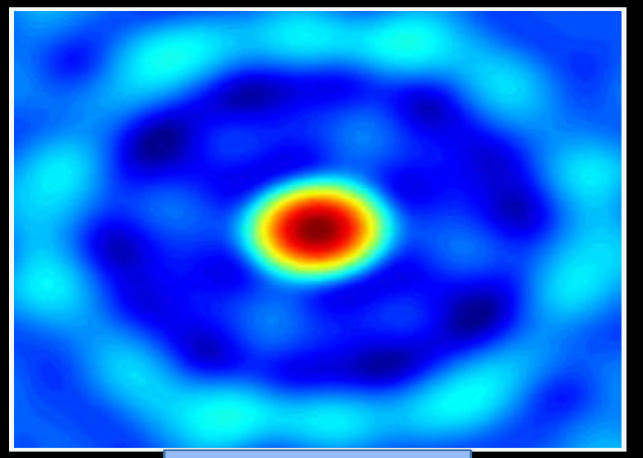
×

Snapshot Scattering profile has much more information than SAXS profile

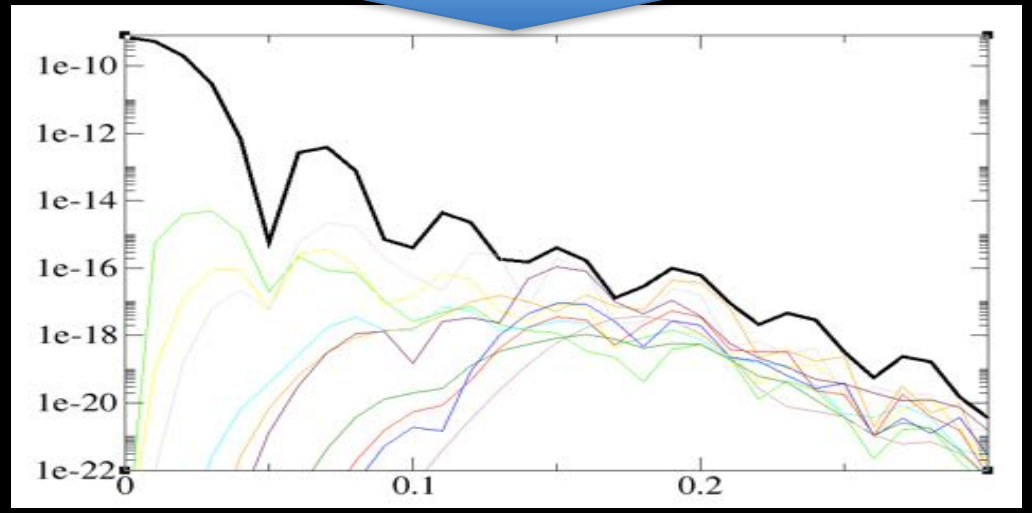


SAXS



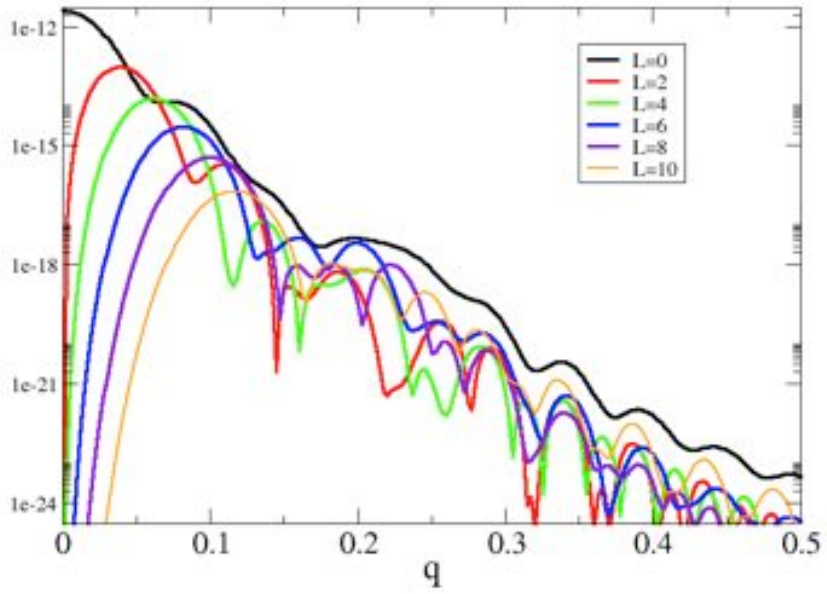


SnapS

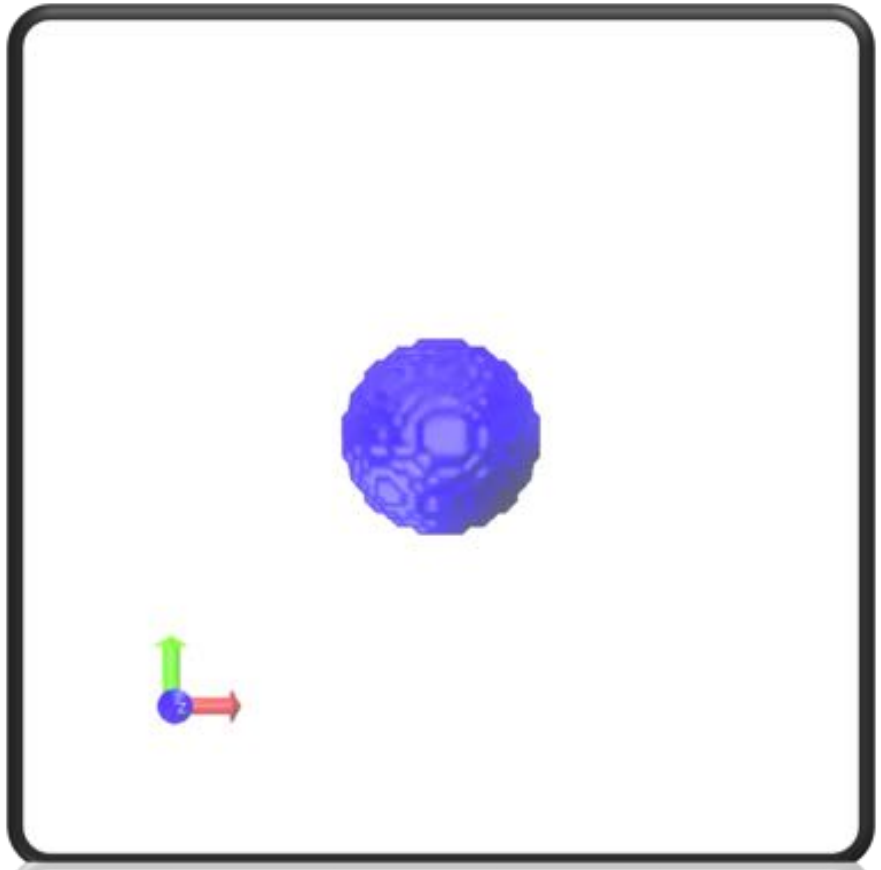


X

Snapshot Scattering profile to Model: Sculpturing

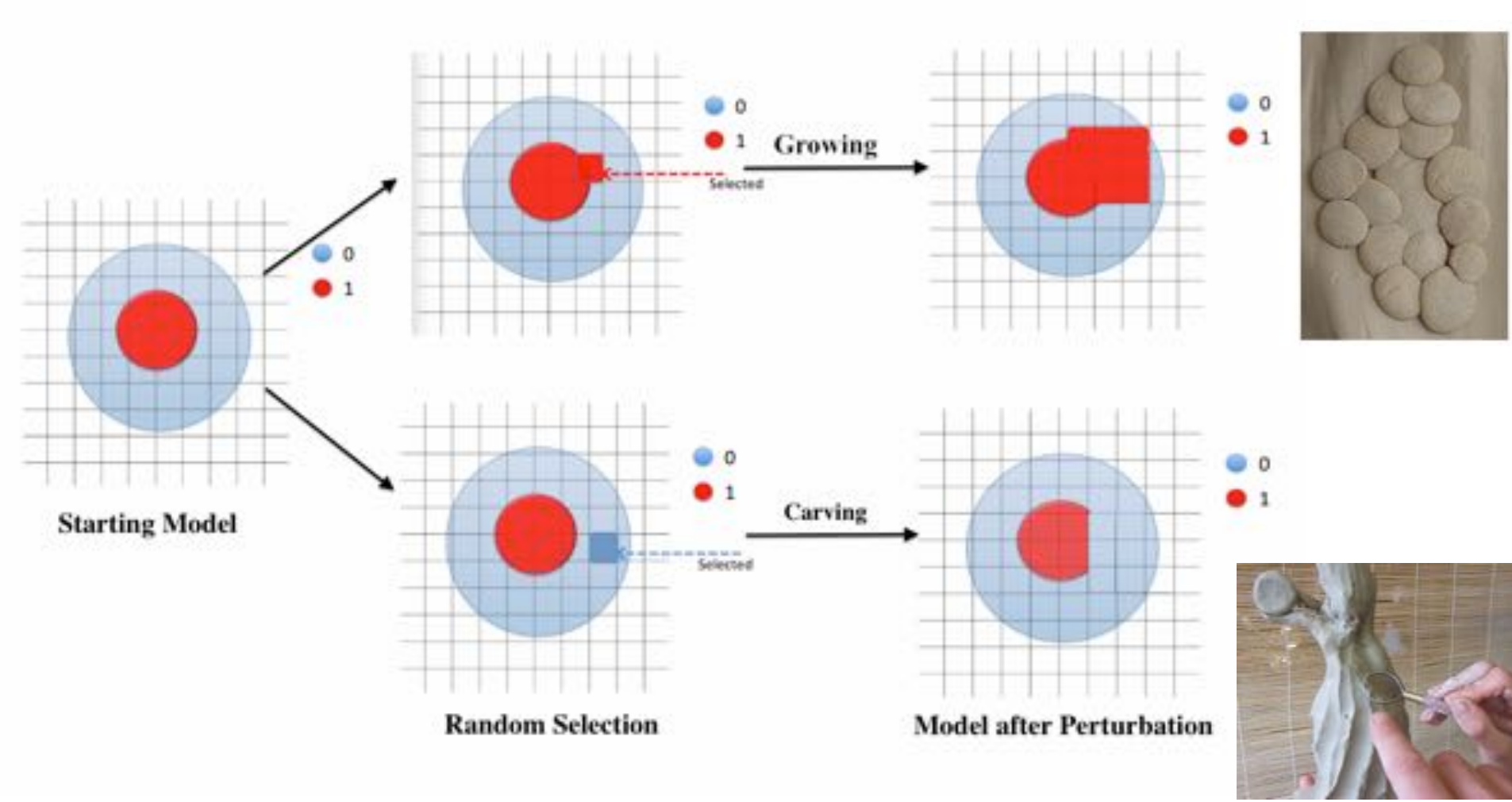


Liu et al., Acta A68, 561-567 (2013)

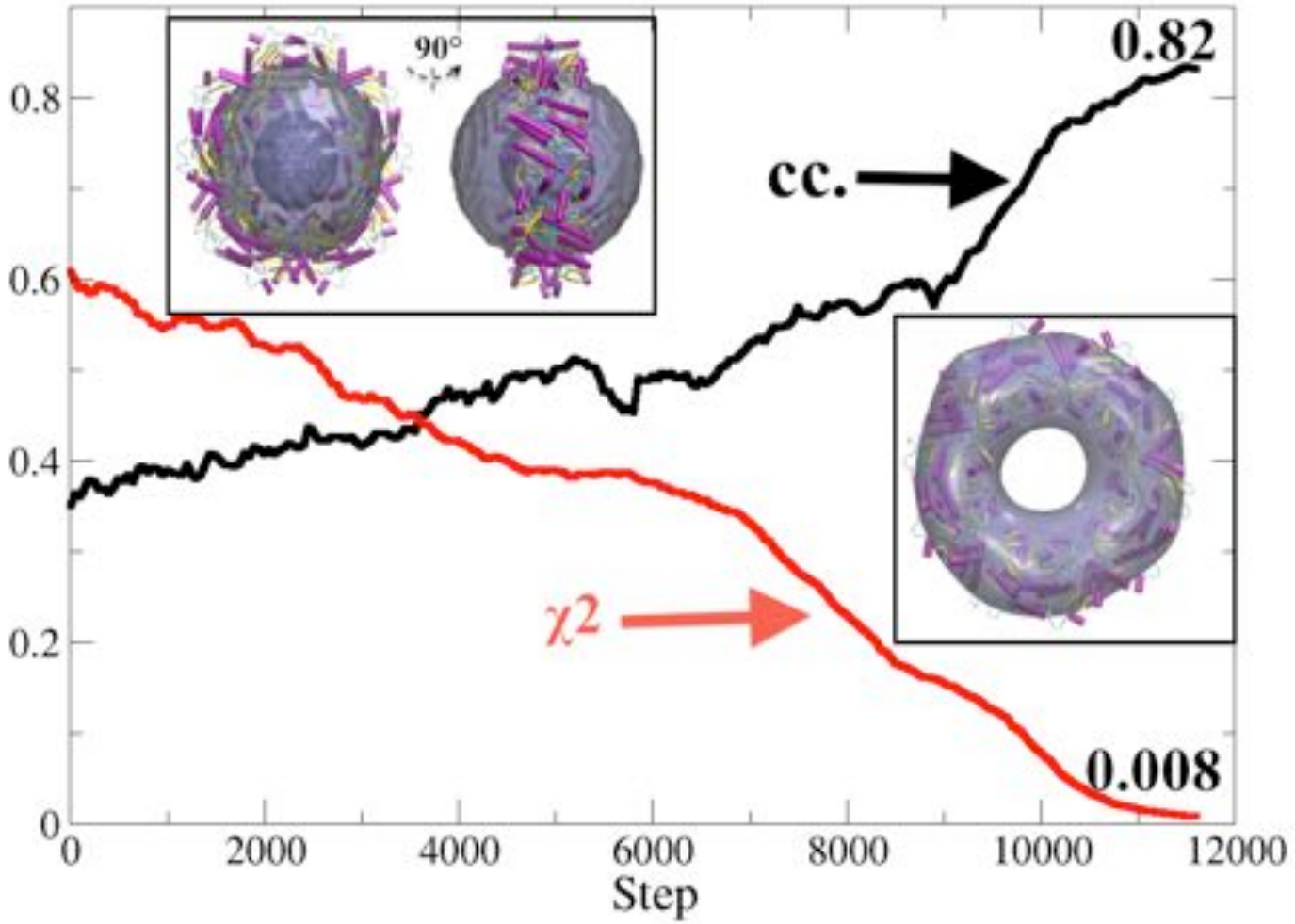


Liu et al., Acta A69, 365-373 (2013)

Perturbations: Growing/Carving



From sphere (or something else) to donut

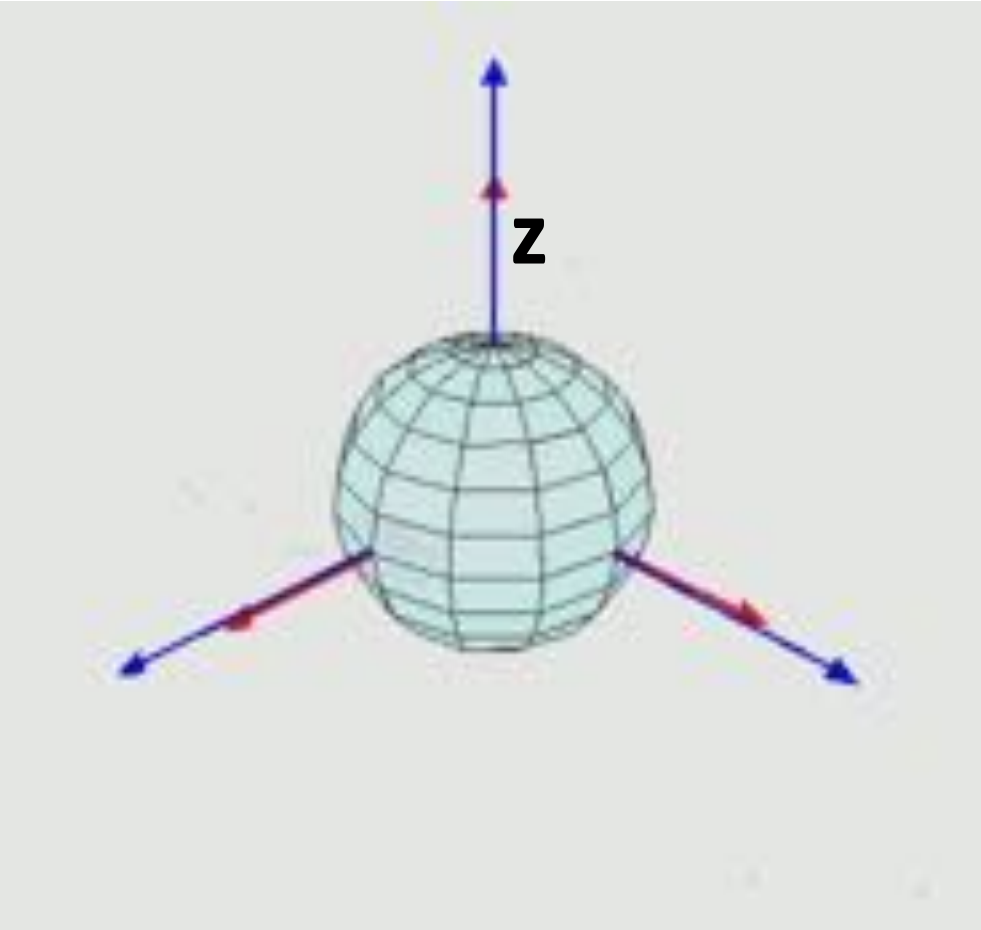


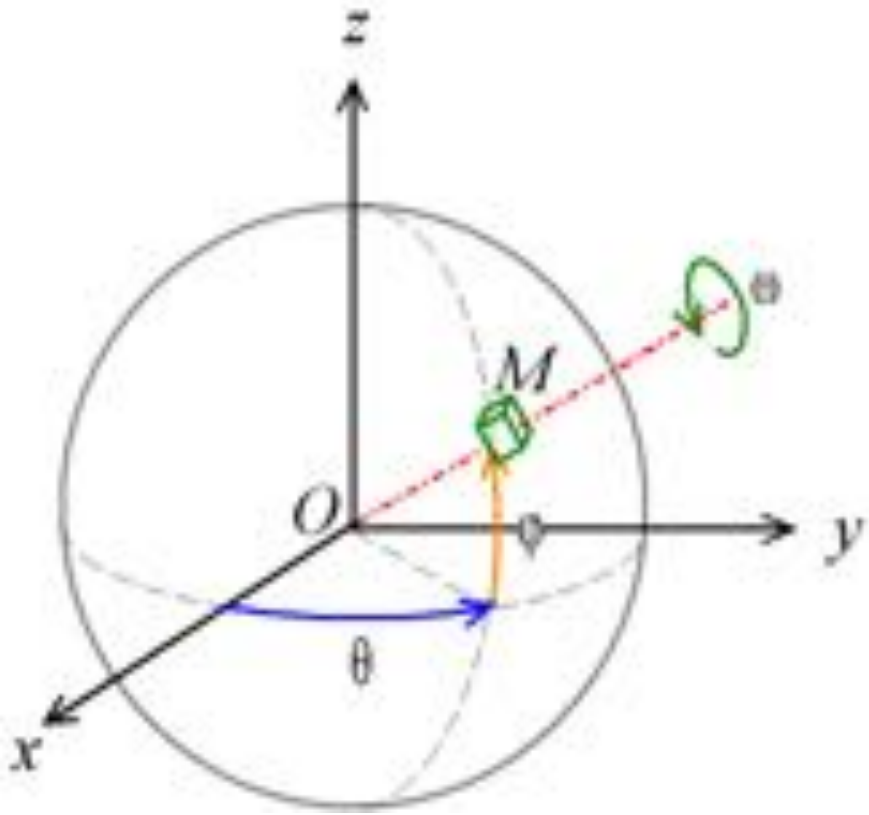
Optimized using **Simulated Annealing** Algorithm

Outline

1. Single Particle Scattering Imaging using XFEL
2. Correlation Function and its applications
   - Extracting structure information
   - **Facilitating the recovery of orientations**
3. Sample Heterogeneity
   - Nanoparticle (size variation)

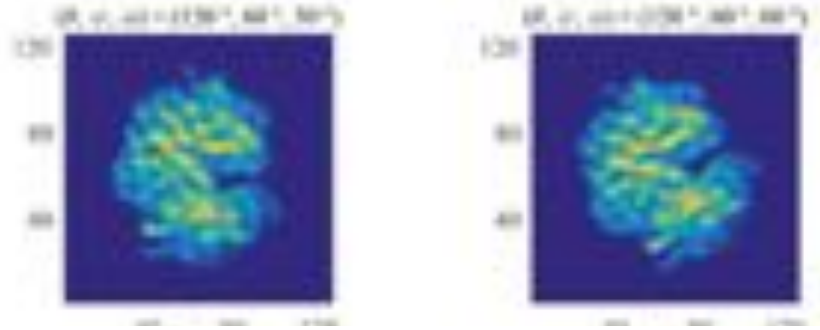
Orientation Recovery with auto-correlations





Euler Angles (θ,ϕ,ω)

Speed up orientation recovery



- Three Euler Angles requires ***O(N³)***, let N be number of grids representing each rotation degree.

- Angular Correlation function (ACF) is independent of in-plane rotation, the angle decoupling simplifies the complexity to ***O(N²).***

- If *M* directions were chosen from ACF, then the complexity is ***O(MN²)***
  - ***Gain speed if M<<N***

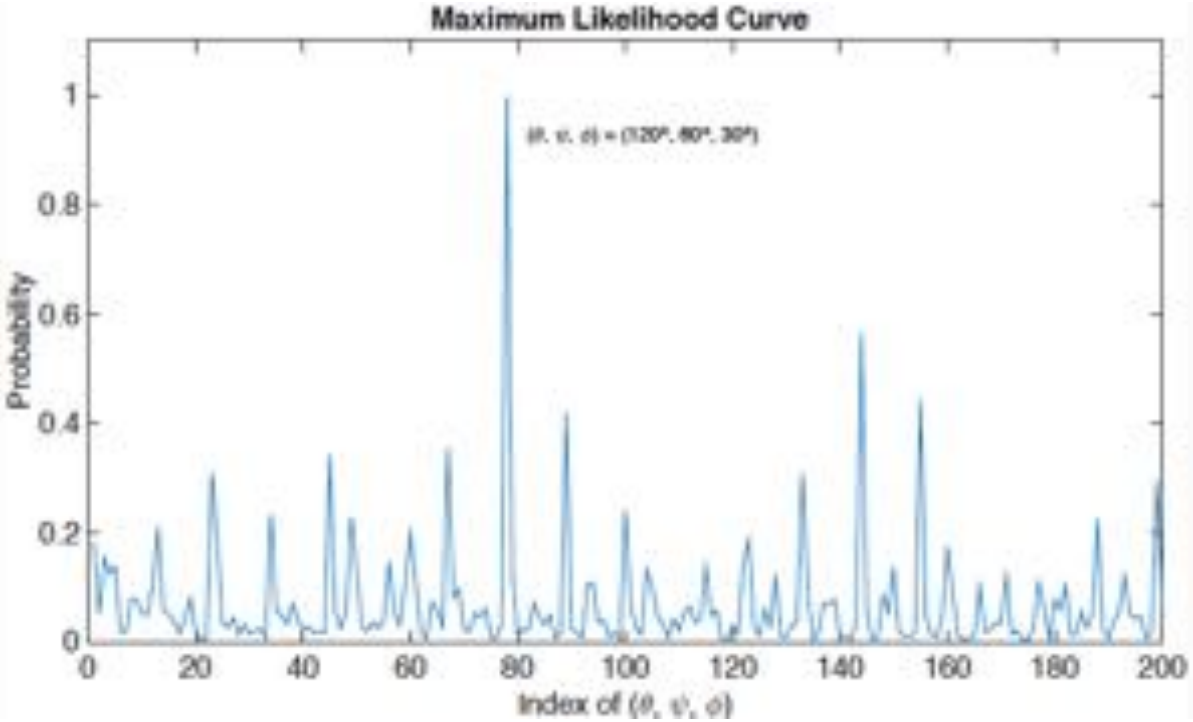
Step-wise Orientation Recovery: SOR

$$P(\boldsymbol{X}_{ij}|\theta,\psi,\phi,\Theta,\boldsymbol{Y}_{ij}) = P_1(\boldsymbol{X}_{ij}|\theta,\psi,\mathscr{Y})P_2(\boldsymbol{X}_{ij}|\phi,\mathscr{Y}(\theta,\psi))$$

$$P_1(\boldsymbol{X}_{ij}|\theta,\psi,\mathscr{Y}) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} \exp\left(\frac{|C_2(\boldsymbol{X}_{ij}) - C_2(\boldsymbol{Y}_{ij})|^2}{-2\sigma(C_2)_{ij}^2}\right)$$

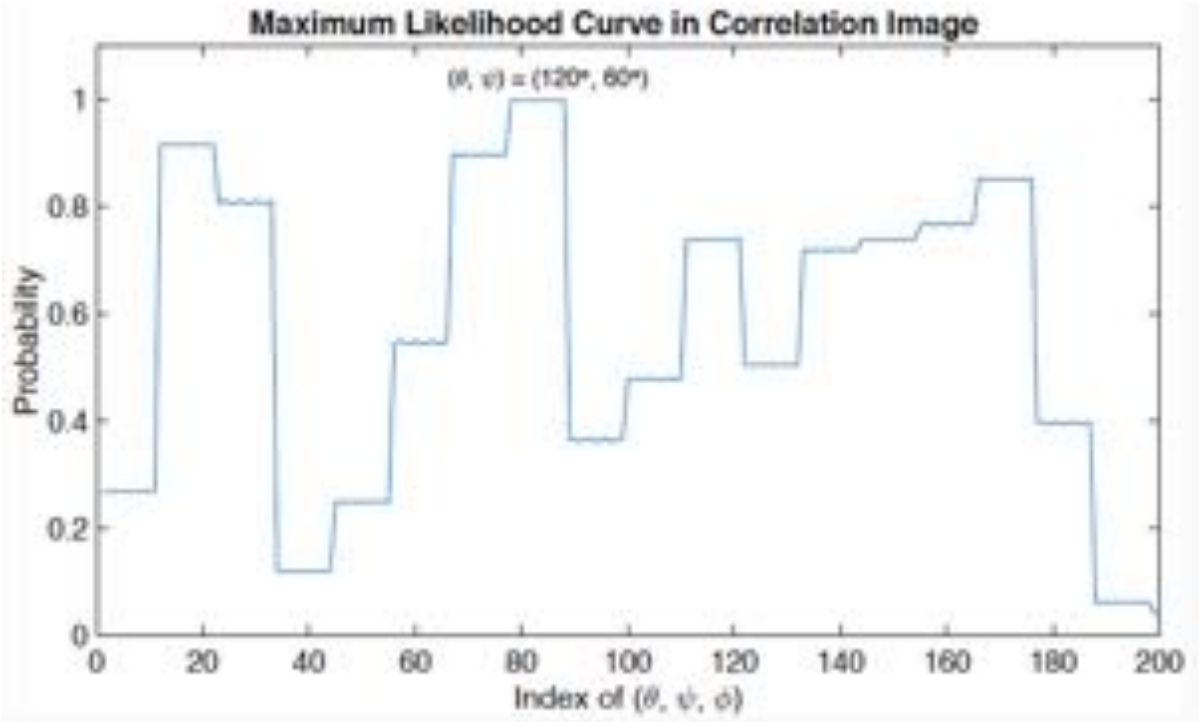
$$P_2(\boldsymbol{X}_{ij}|\phi,\mathscr{Y}(\theta,\psi)) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} \exp\left(\frac{|\boldsymbol{X}_{ij} - \boldsymbol{Y}_{ij}|^2}{-2\sigma_{ij}^2}\right)$$

Most likely **orientation** from raw data



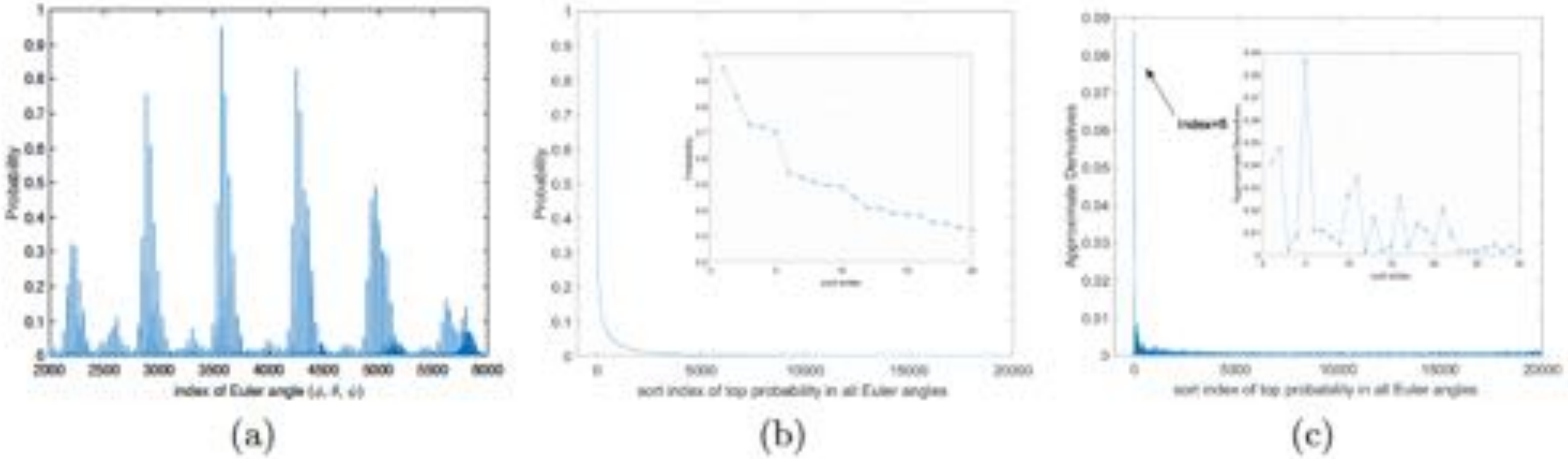
6 x 3 x **12** reference patterns

Most likely **direction** from ACF



Steps of **12** in-plane rotations

Improve robustness of SOR:
choosing M best directions for further searching



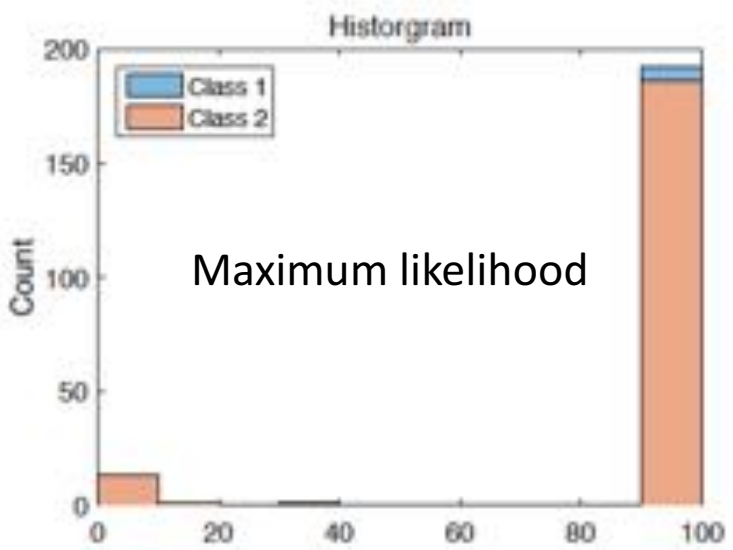
(a) (b) (c)

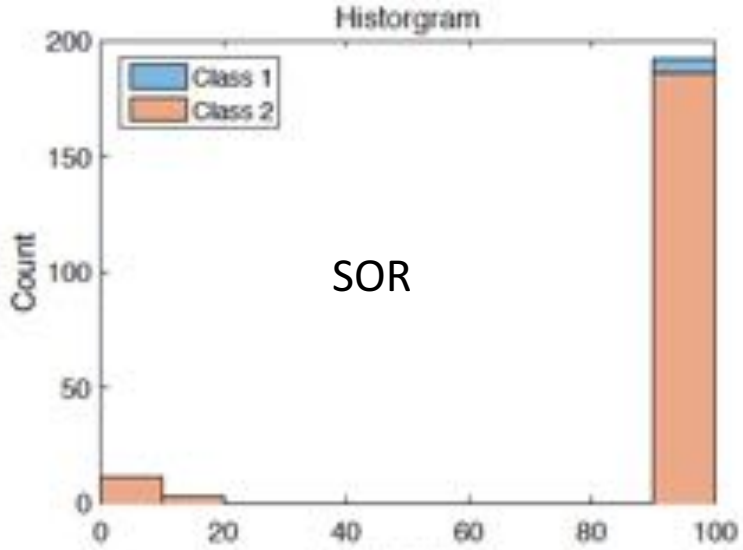
(a) Probability being at each orientation;
(b) Sorted based on the probabilities;
(c) Derivative of probability curve in (b);

The number of possible directions subjected to further analysis is set based on the derivative trend.

Performance of SOR:

Faster, without compromising accuracy.





Approach \ Reference grids	Space	10°(24,624)	3°(878,400)
Maximum likelihood	real-space	4.7089s	116.8273s
	fourier-space	3.9285s	111.6007s
SOR	real-space	0.3334s	1.7484s
	fourier-space	0.4556s	1.7365s

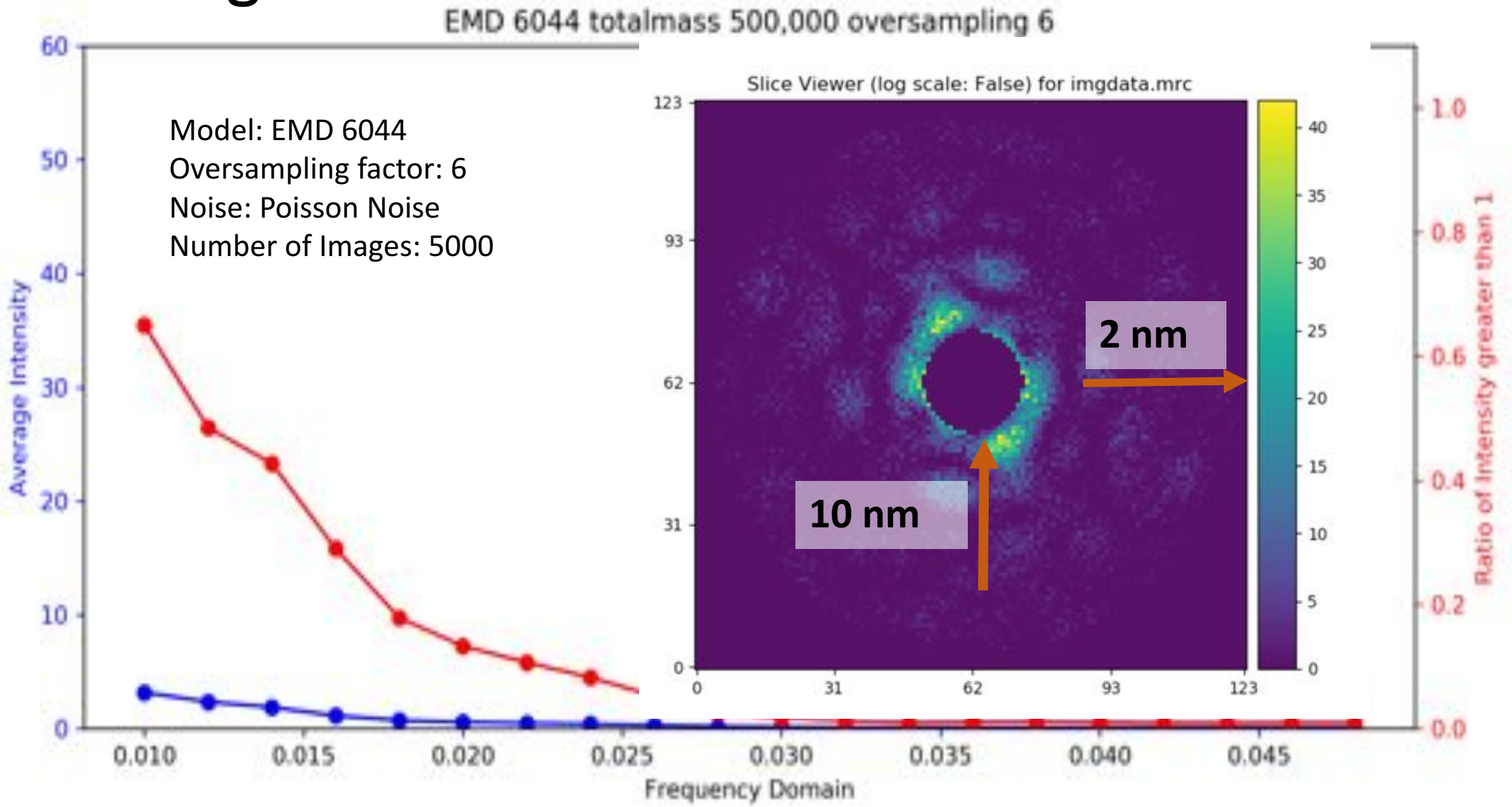
Class1: ω = 0
Class2: ω nq. 0

Maximum likelihood approach and SOR method give similar accuracy in this case.

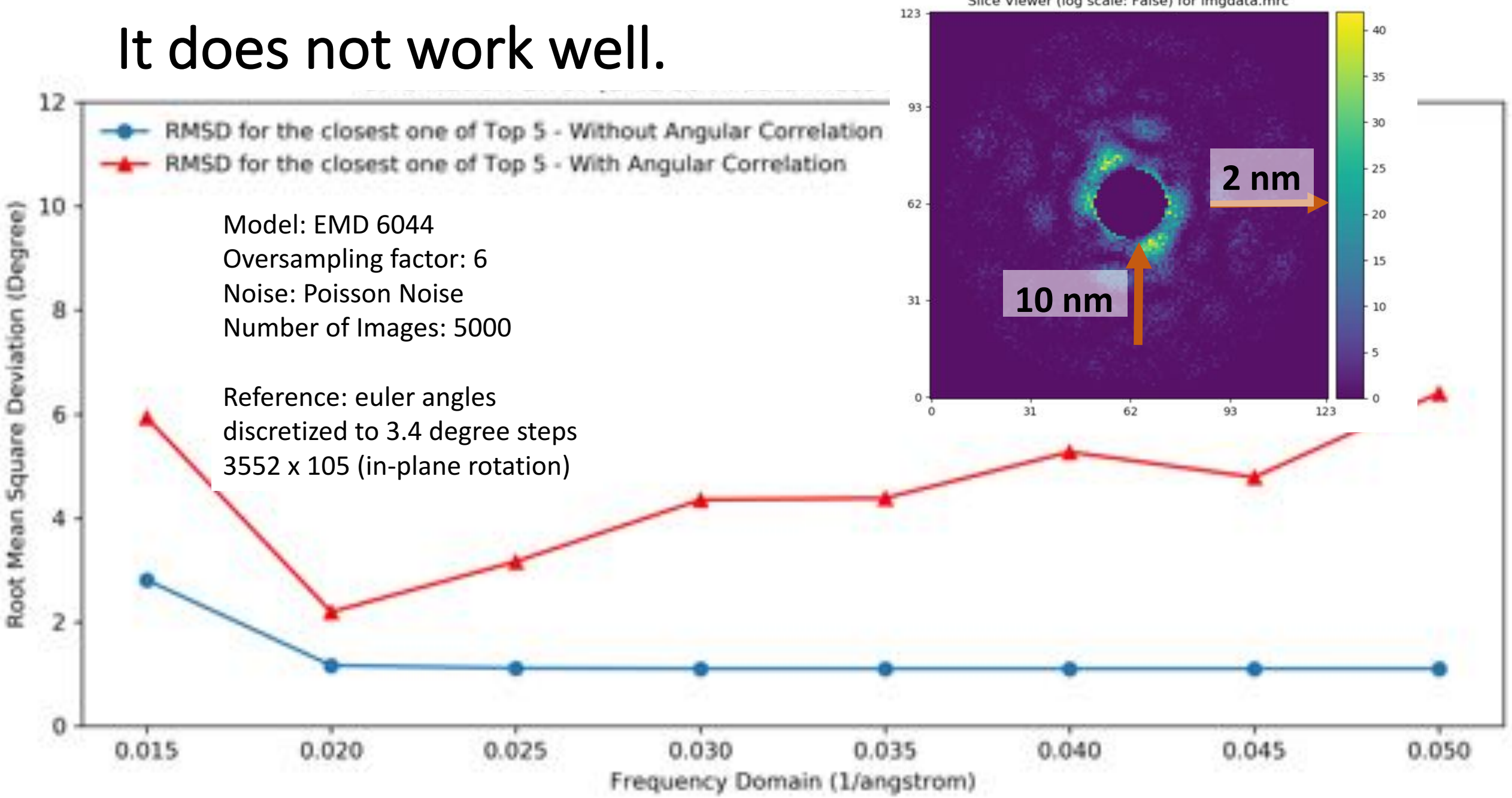
Depending on the number of in-plane rotation, the speed gain is different.

8.5x -- 64.5x

Low signals?


EMD 6044 totalmass 500,000 oversampling 6
Model: EMD 6044
Oversampling factor: 6
Noise: Poisson Noise
Number of Images: 5000
Slice Viewer (log scale: False) for imgdata.mrc
2 nm
10 nm
Average Intensity
Ratio of Intensity greater than 1
Frequency Domain

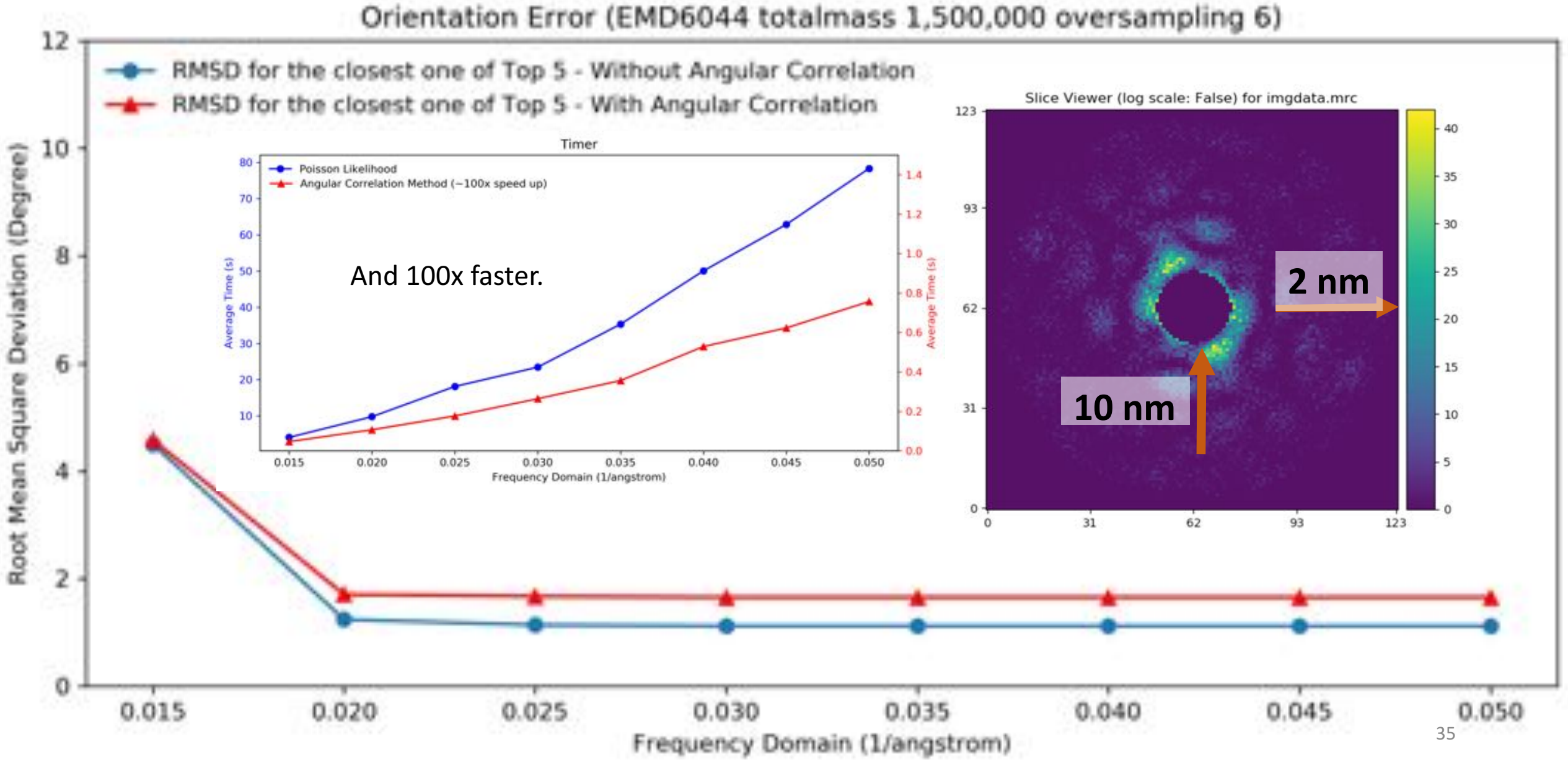
It does not work well.



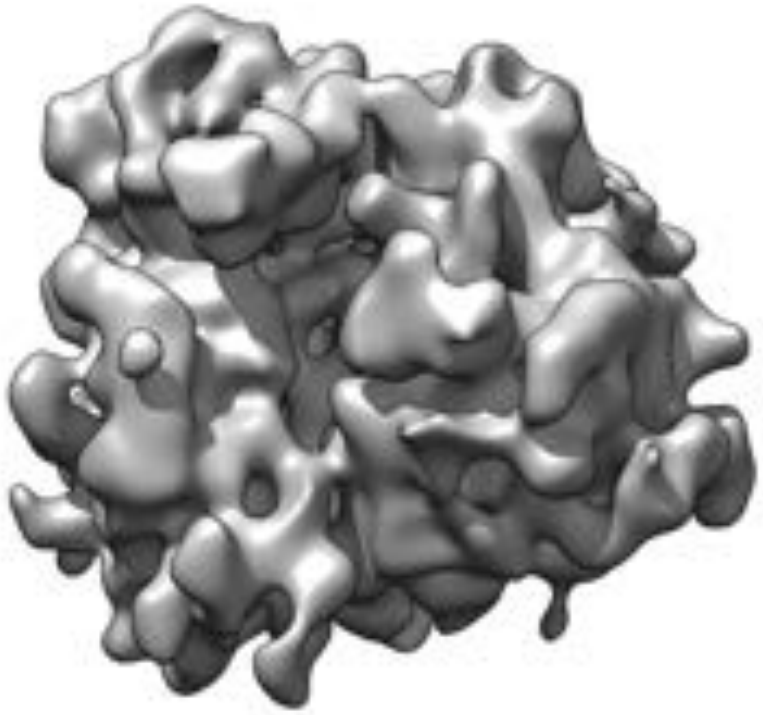
Model: EMD 6044
Oversampling factor: 6
Noise: Poisson Noise
Number of Images: 5000

Reference: euler angles
discretized to 3.4 degree steps
3552 x 105 (in-plane rotation)

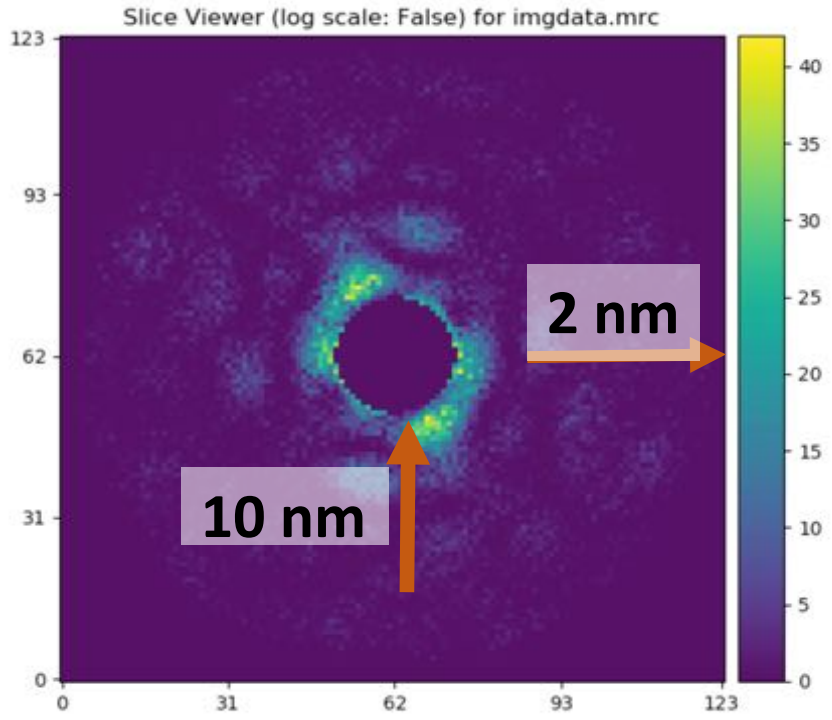
It does work well, if fluence is 3 times higher.



EMD 6044: Yeast 80s Ribosome



25 – 30 nm diameter, smaller than PR772



More photons are needed.

Outline

1. Single Particle Scattering Imaging using XFEL

2. Correlation Function and its applications

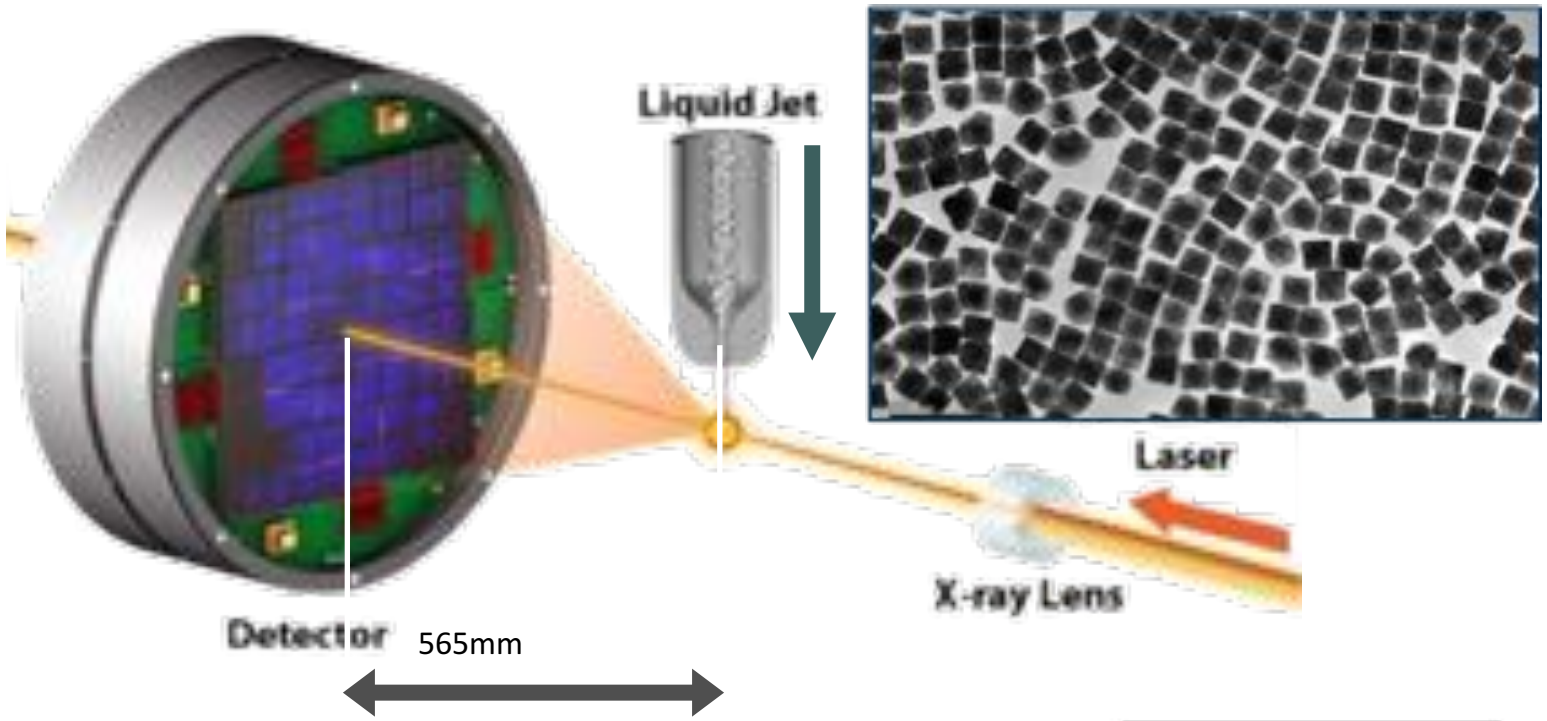
   - Extracting structure information

   - Facilitating the recovery of orientations

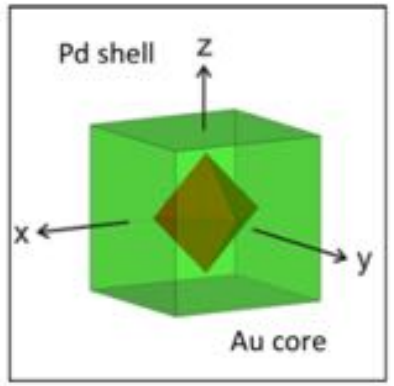
3. **Sample Heterogeneity**

   - Nanoparticle (size variation)

Proof of principle experiment

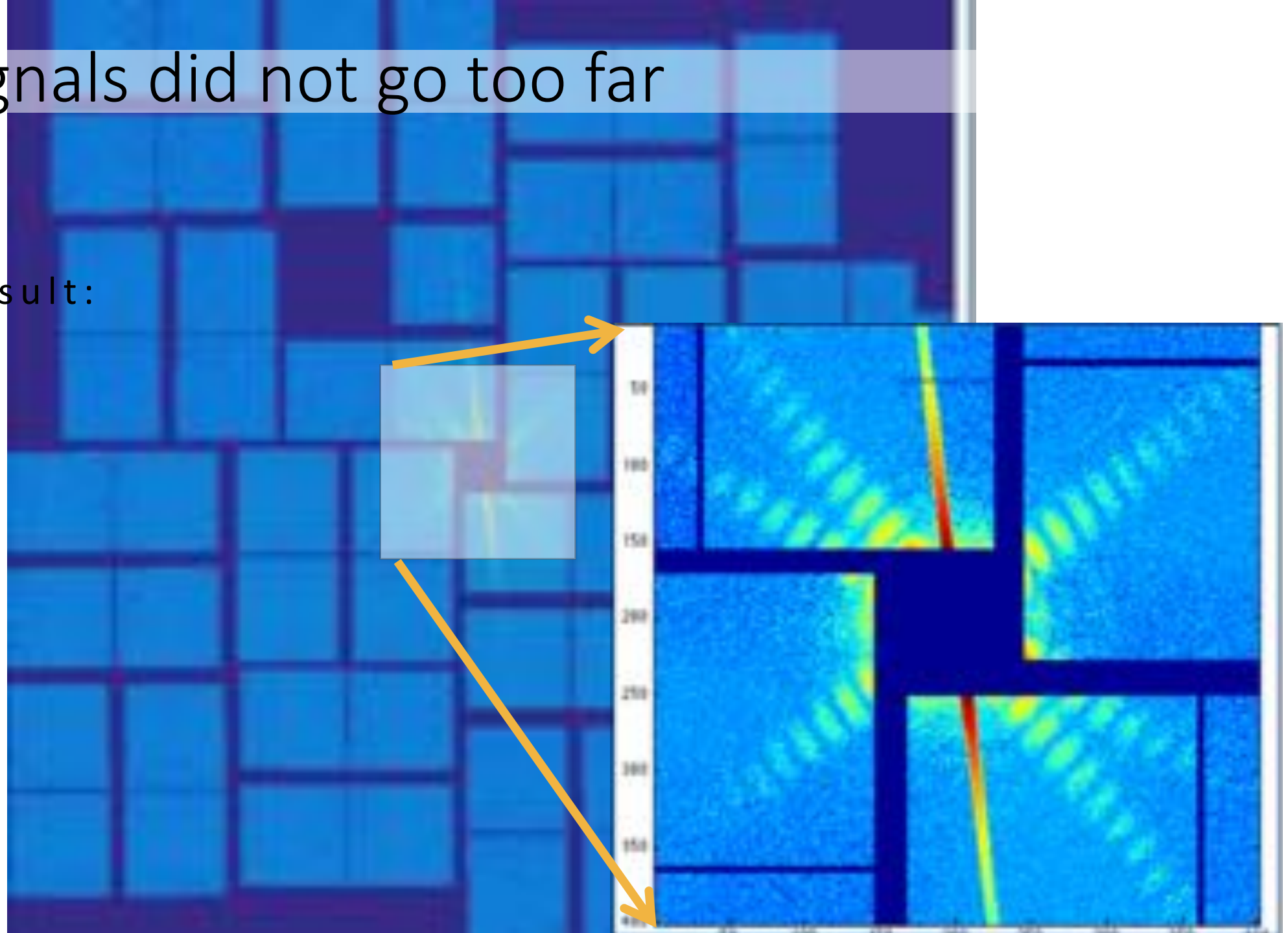


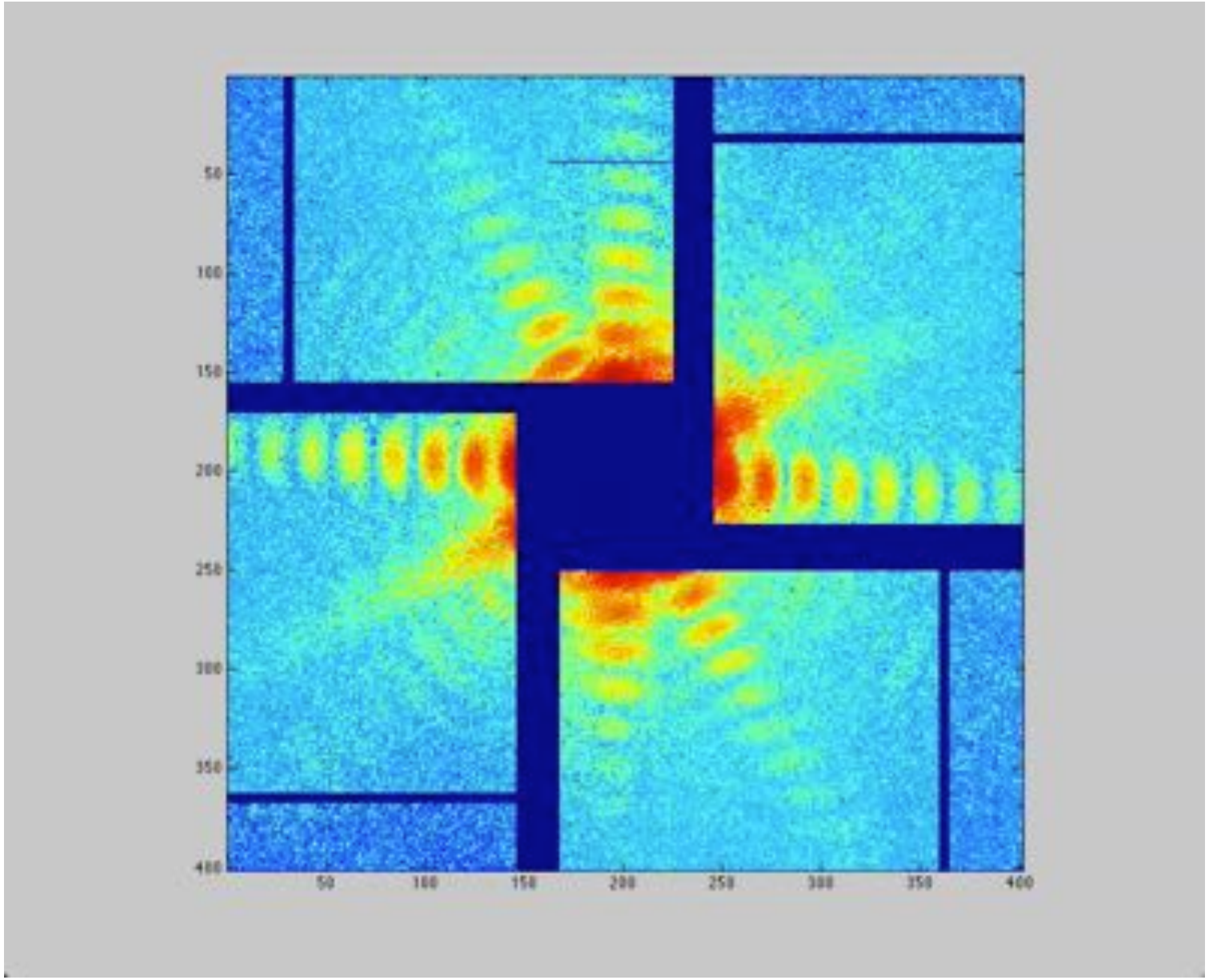
- Question to be answered: How far can the signals go?
- Let's focus x-rays to strong scatters
  - 100 nm beam on Gold particles (gold core, palladium shell)

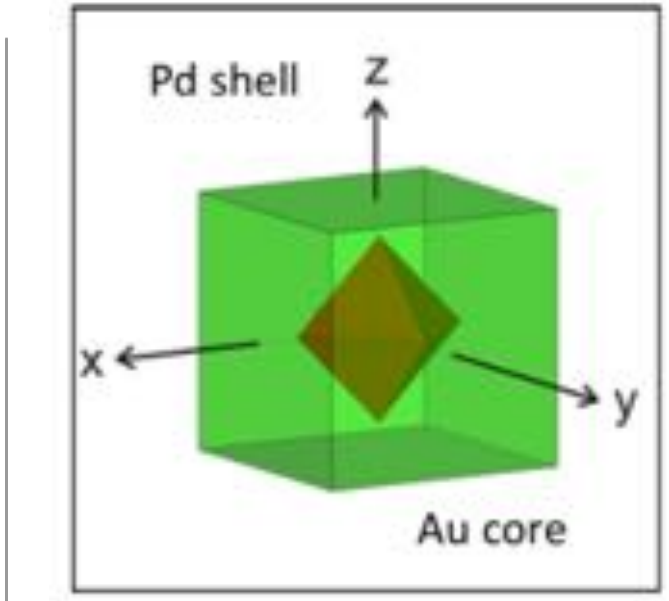


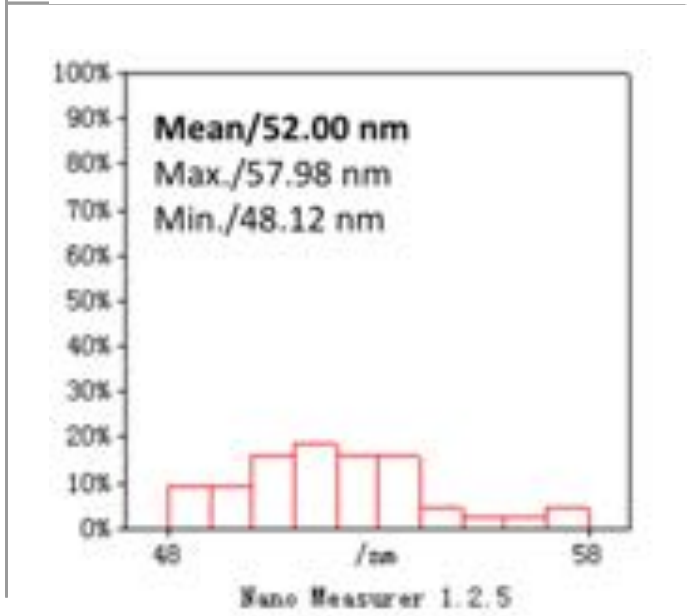
Signals did not go too far

- Result:



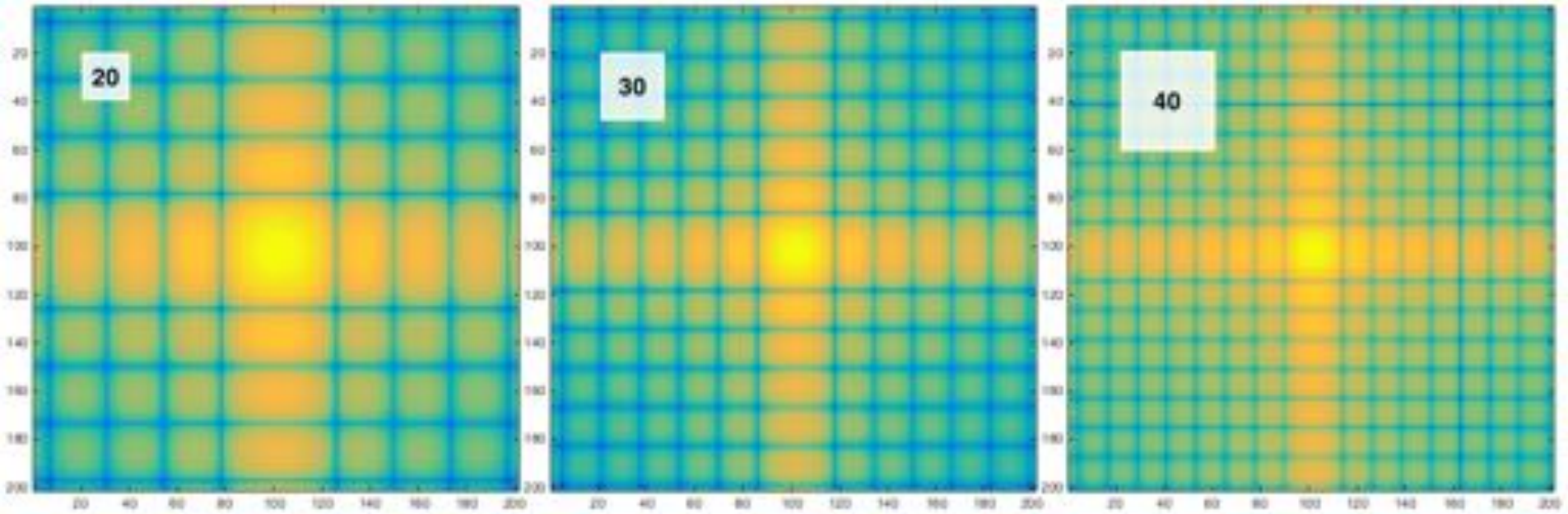






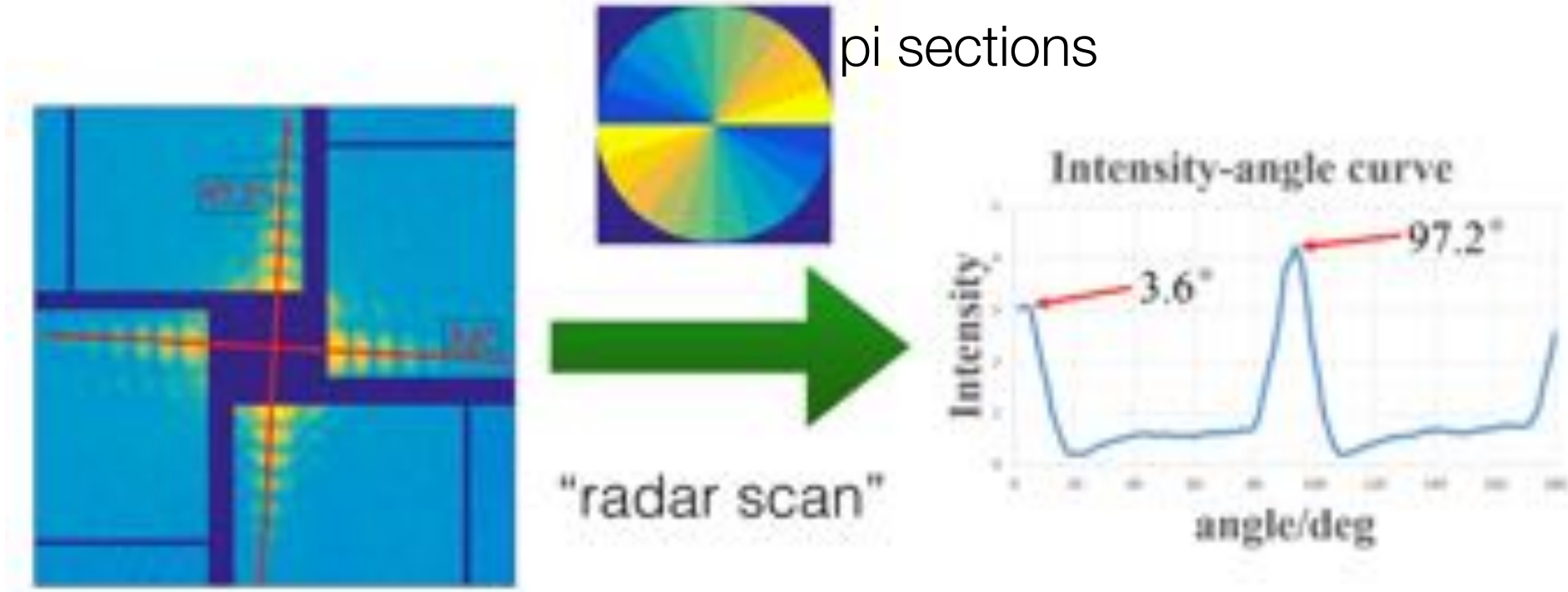
A representative set of patterns (Au:Pd, core-shell particle)

Sample size variation makes it hard to conduct pixel-level comparison



It does not make sense to directly compare intensities at each pixel.

Integration over |q|.

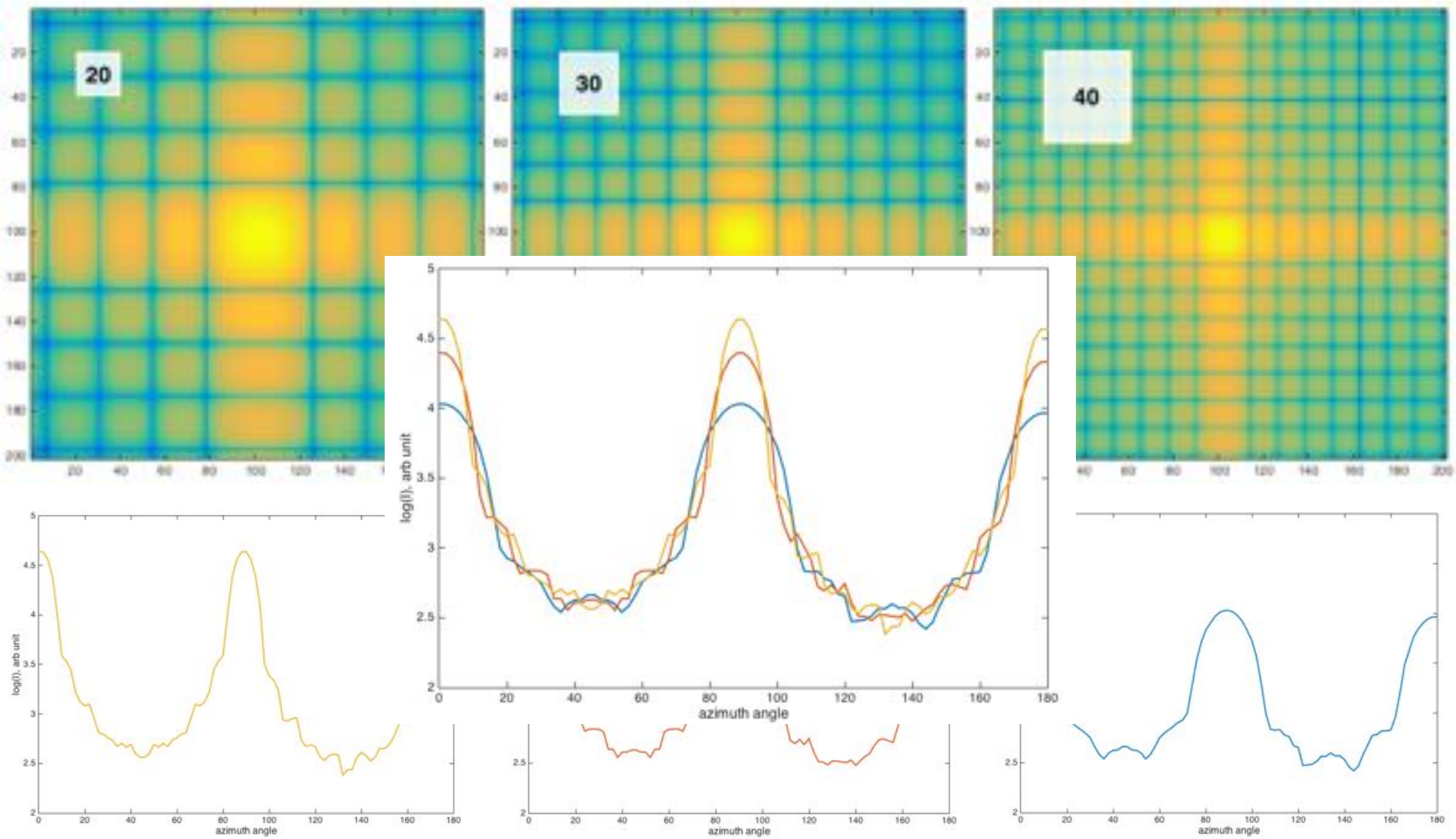


This function depends on four parameters: particle size + orientations (three euler angles)

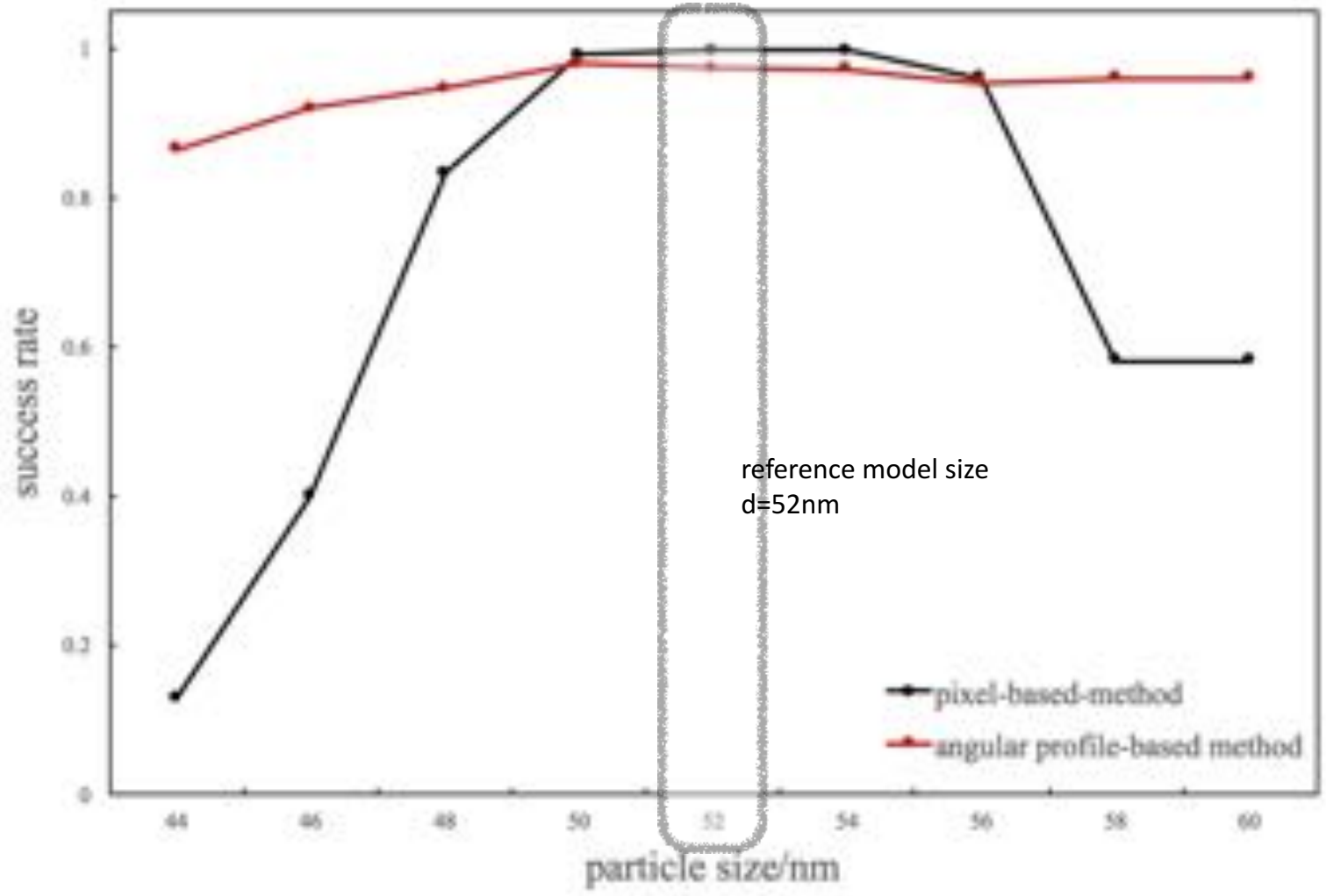
Convert patterns to profiles as **a function of azimuth angle**, then compare the profile to reference set and found the most likely orientation.

$$I(\theta) = \int_{q_{min}}^{q_{max}} I(q,\theta)dq$$

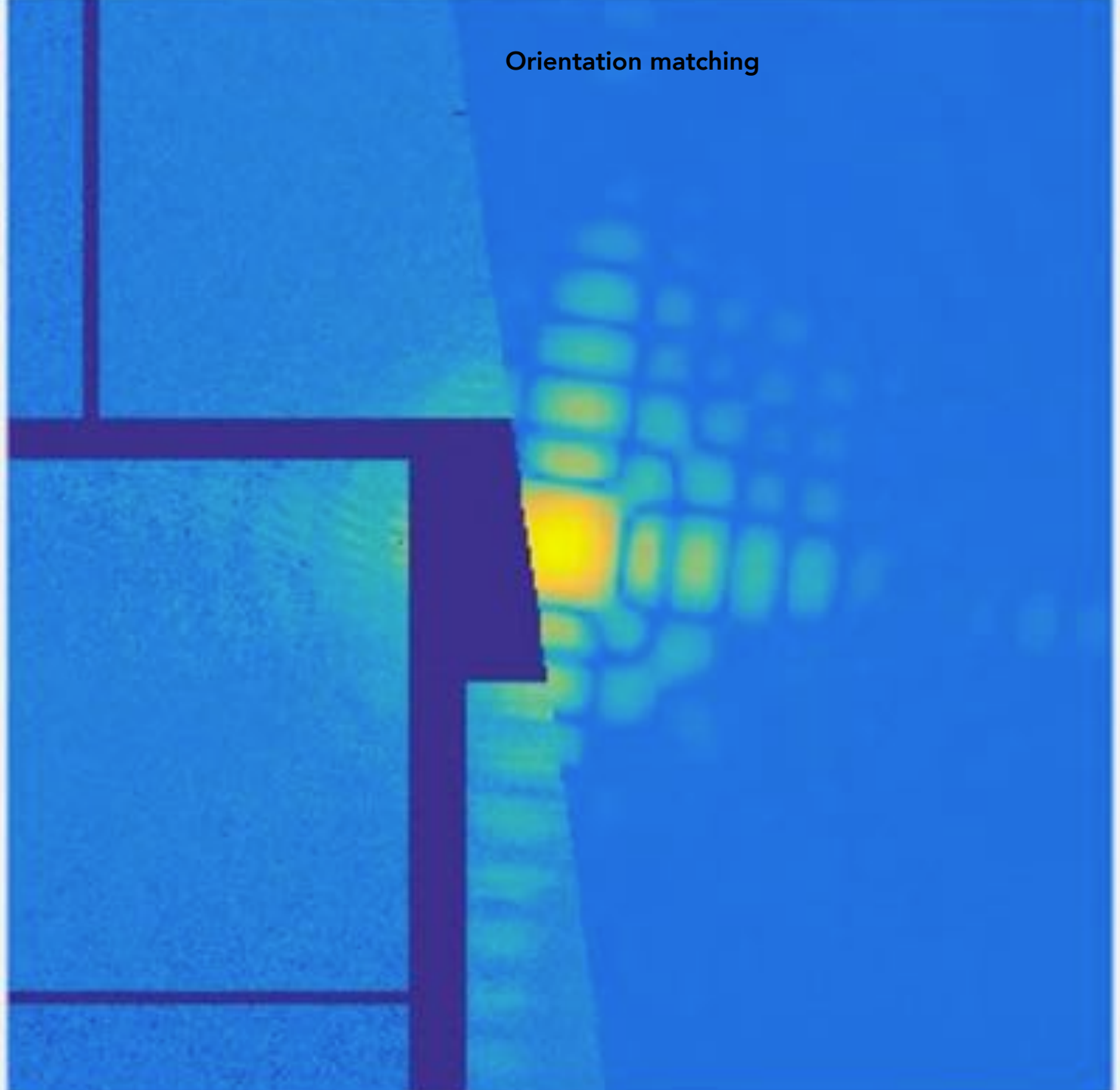
Angular intensity profile is less sensitive to particle size variations



Success rate of orientation recovery (Simulation)

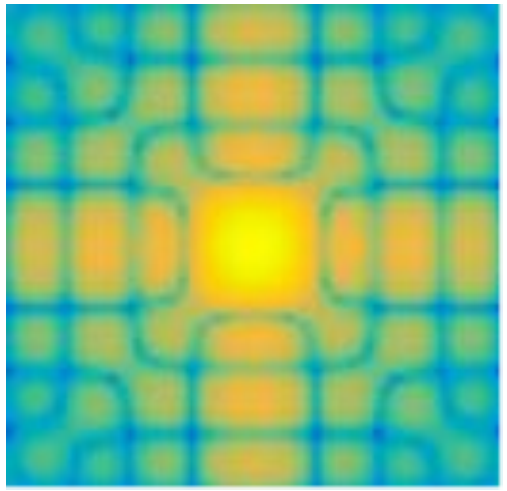


Angular profile is a better signature for orientations, much less sensitive than raw pattern, when comparing patterns


Orientation matching

size variations

- two scenarios:
- (1) normal incidence:
  - inverse of q-spacing



Δx

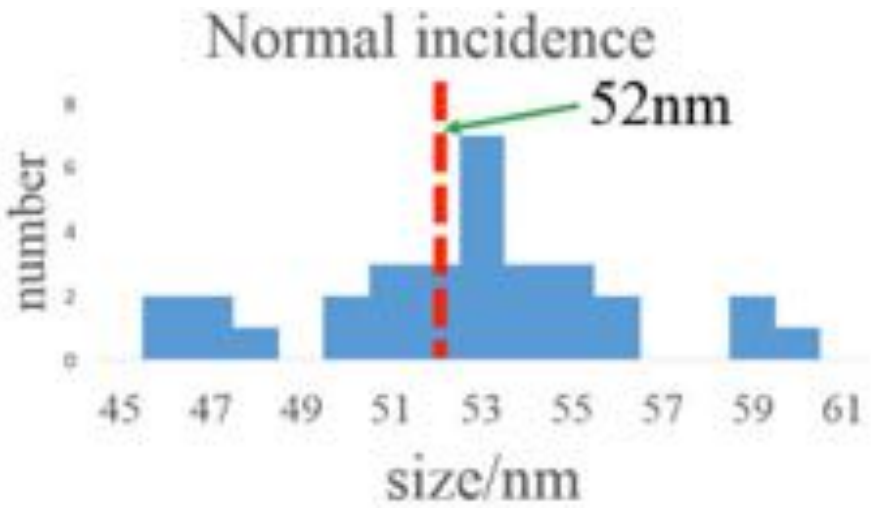
- (2) tilted incidence:
  - fit to the best size

$$particle\_size = \frac{\lambda R}{\Delta x}$$

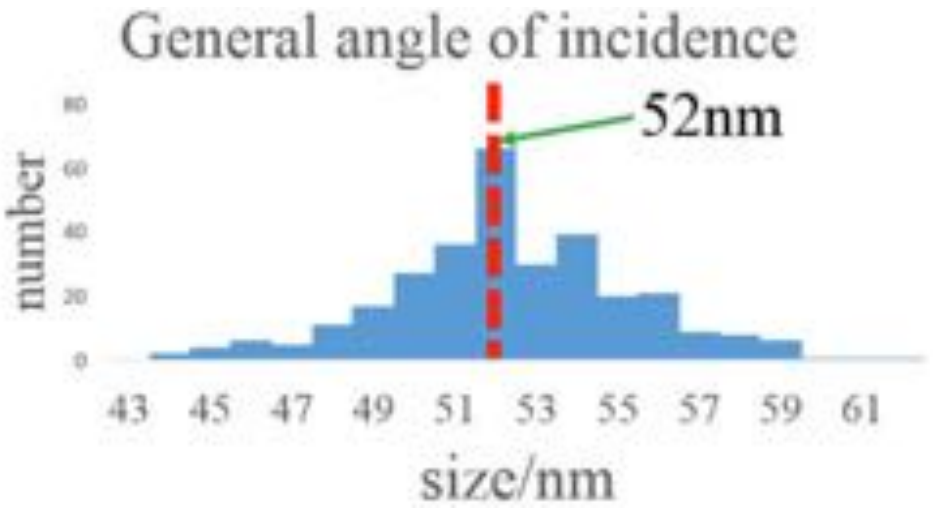
R: distance from sample to detector
λ: wavelength

$$particle\_size = simulated\_particle\_size \frac{\Delta x_{exp}}{\Delta x_{sim}}$$

32 patterns

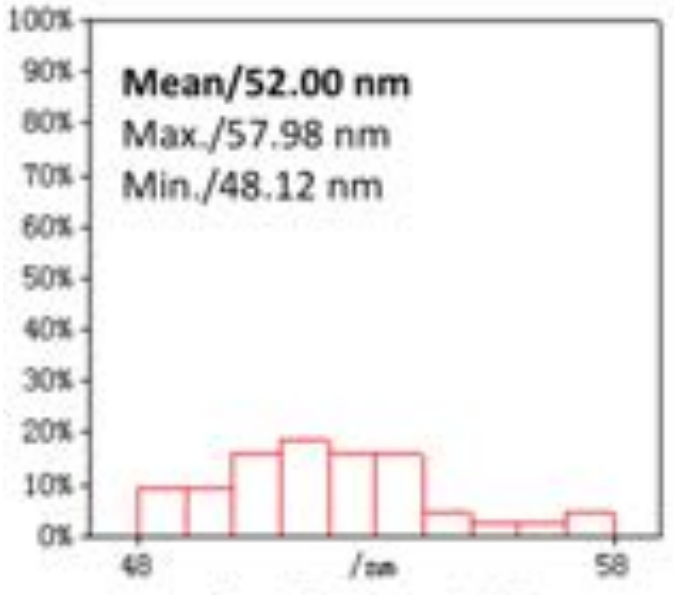


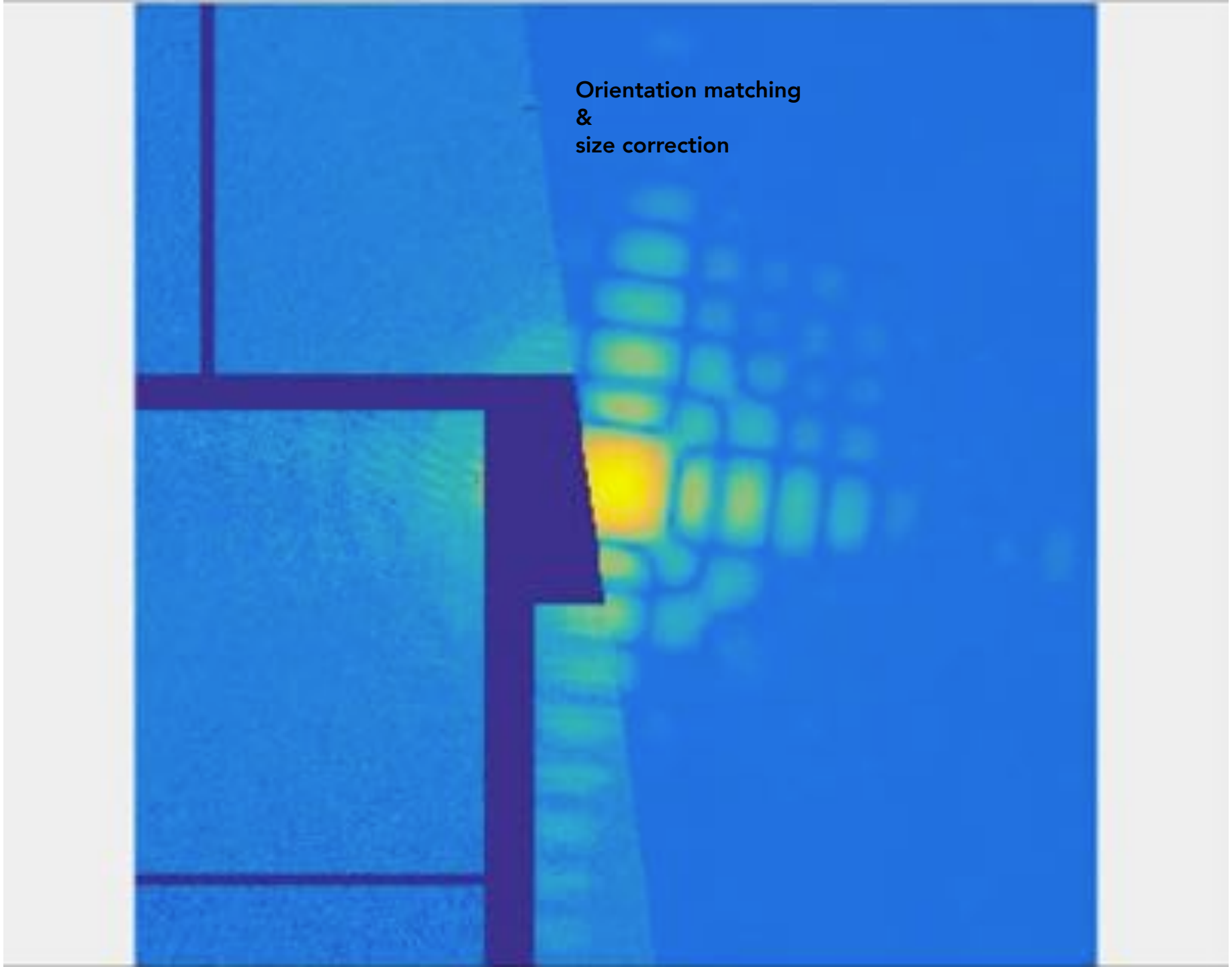
527 patterns with fringes



Experimental data

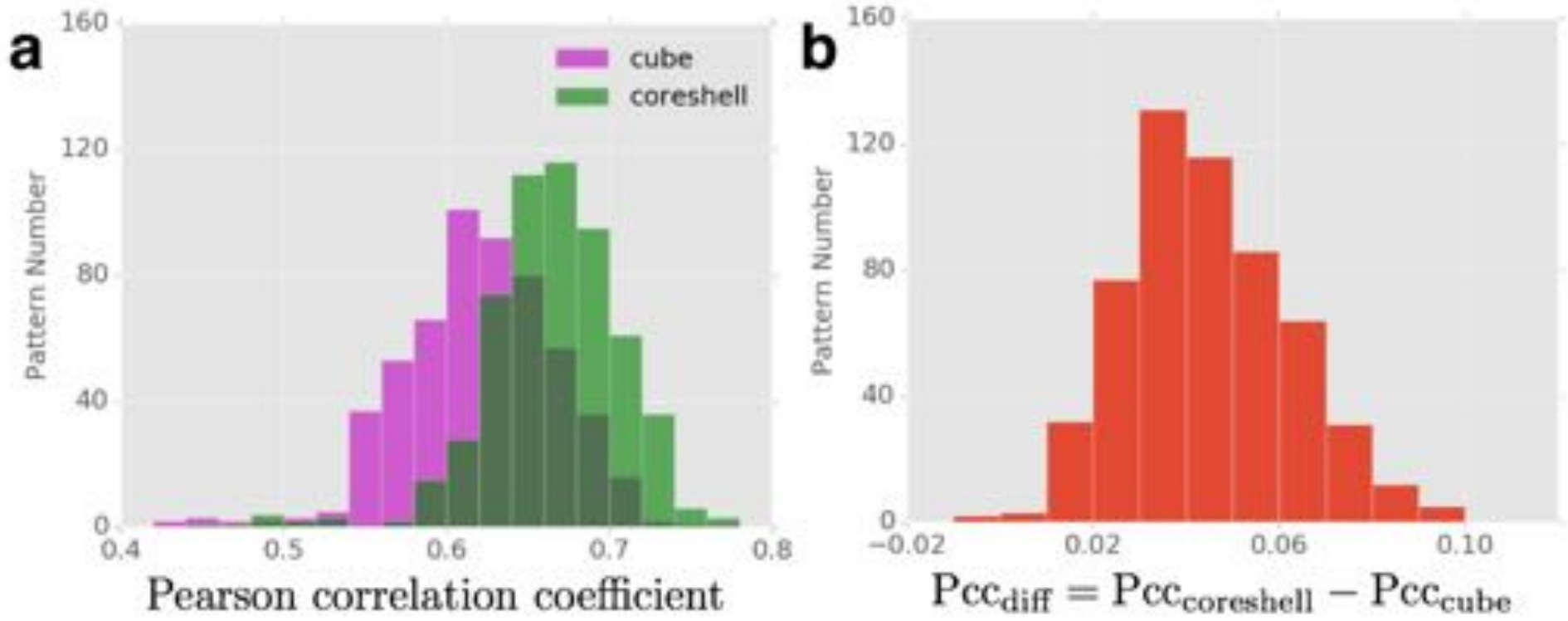
- consistent with TEM data




Orientation matching
&
size correction

The core-shell model is preferred over cubic model

- Pearson correlation between expt data & simulation patterns



Data Descriptor: Li *et al.*, Scientific Data **4**: 170048 (2017)
Analysis: Li *et al.*, IUCrJ **4**, 741-750 (2017)


X-RAY LASER GUNS
Four operational facilities worldwide fire bright, X-ray laser light that can determine structures at atomic resolution. Each X-ray flash lasts around 100 femtoseconds — short enough to capture molecular motions.
With 27,000 pulses per second, the European X-ray Free Electron Laser has a firing rate around 200 times higher than other lasers.
LCLS
United States
First experiments: 2009
Eu-XFEL
Germany
2017
SwissFEL
Switzerland
Expected: 2018
PAL-XFEL
South Korea
2017
SACLA
Japan
2011
The Linac Coherent Light Source is planning an upgrade that, by the 2020s, will allow it to fire X-rays at 1 million pulses per second.
Chinese XFELs will start soon.
superconducting, 1MHz hard X-rays
Expected: 2024
©nature

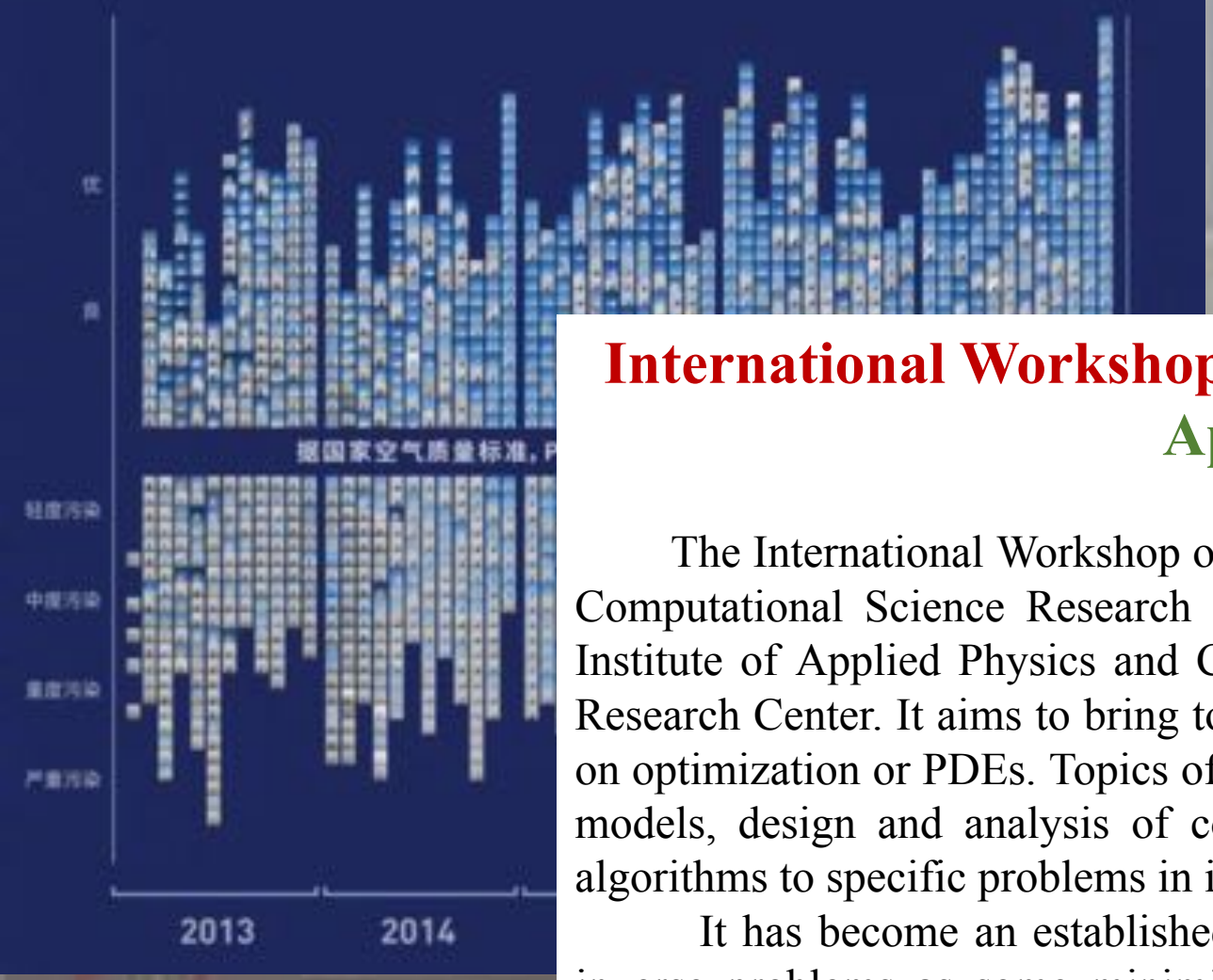
Acknowledgements

Yingchen Shi

Lanqing Huang

Xuanxuan Li

Experimental Team (ASU, MPI, DESY, LCLS), June 2013



www.csrc.ac.cn

Liulab.csrc.ac.cn

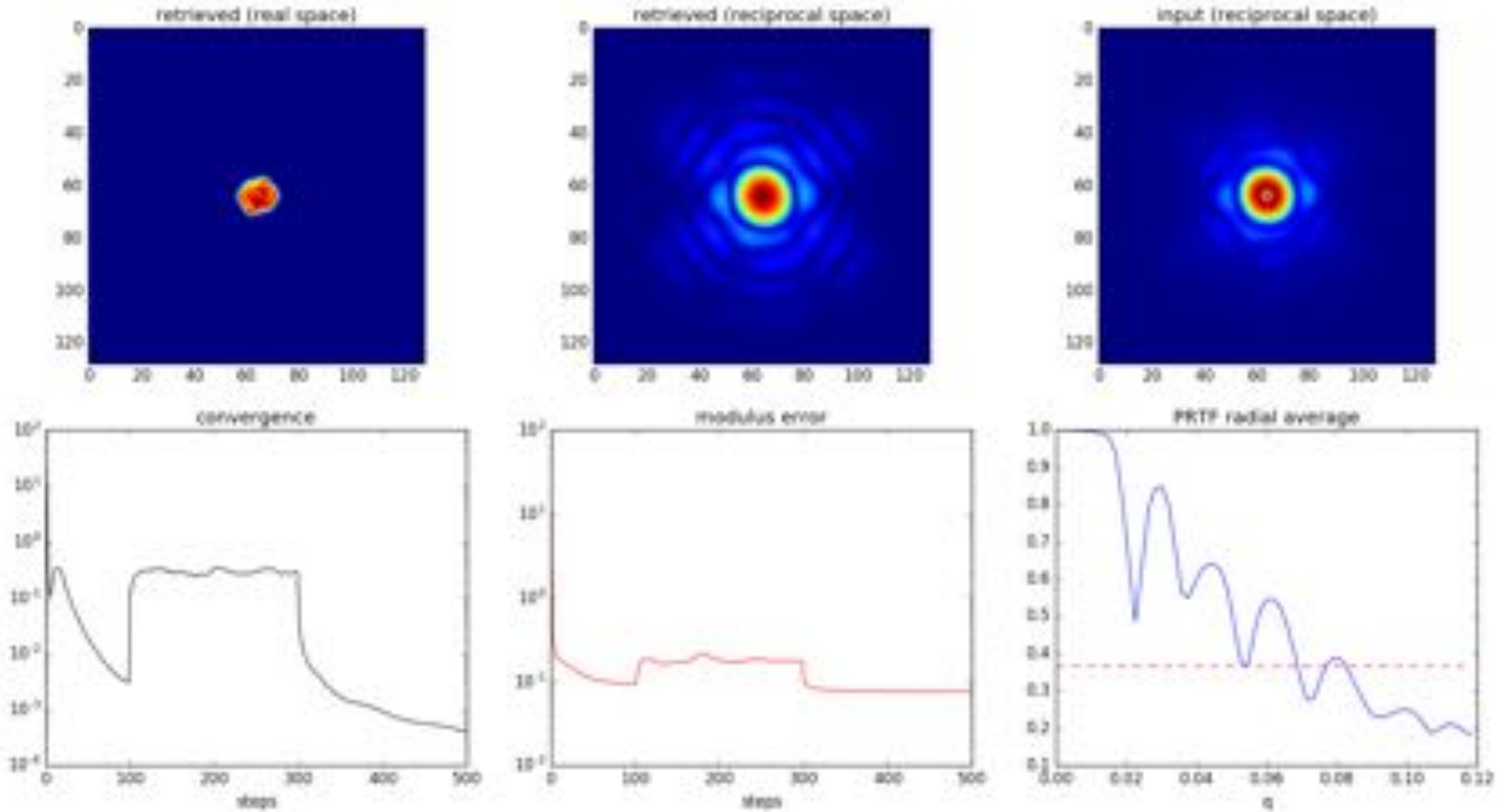
International Workshop On Image Processing and Inverse Problems

April 21-24, 2018 Beijing

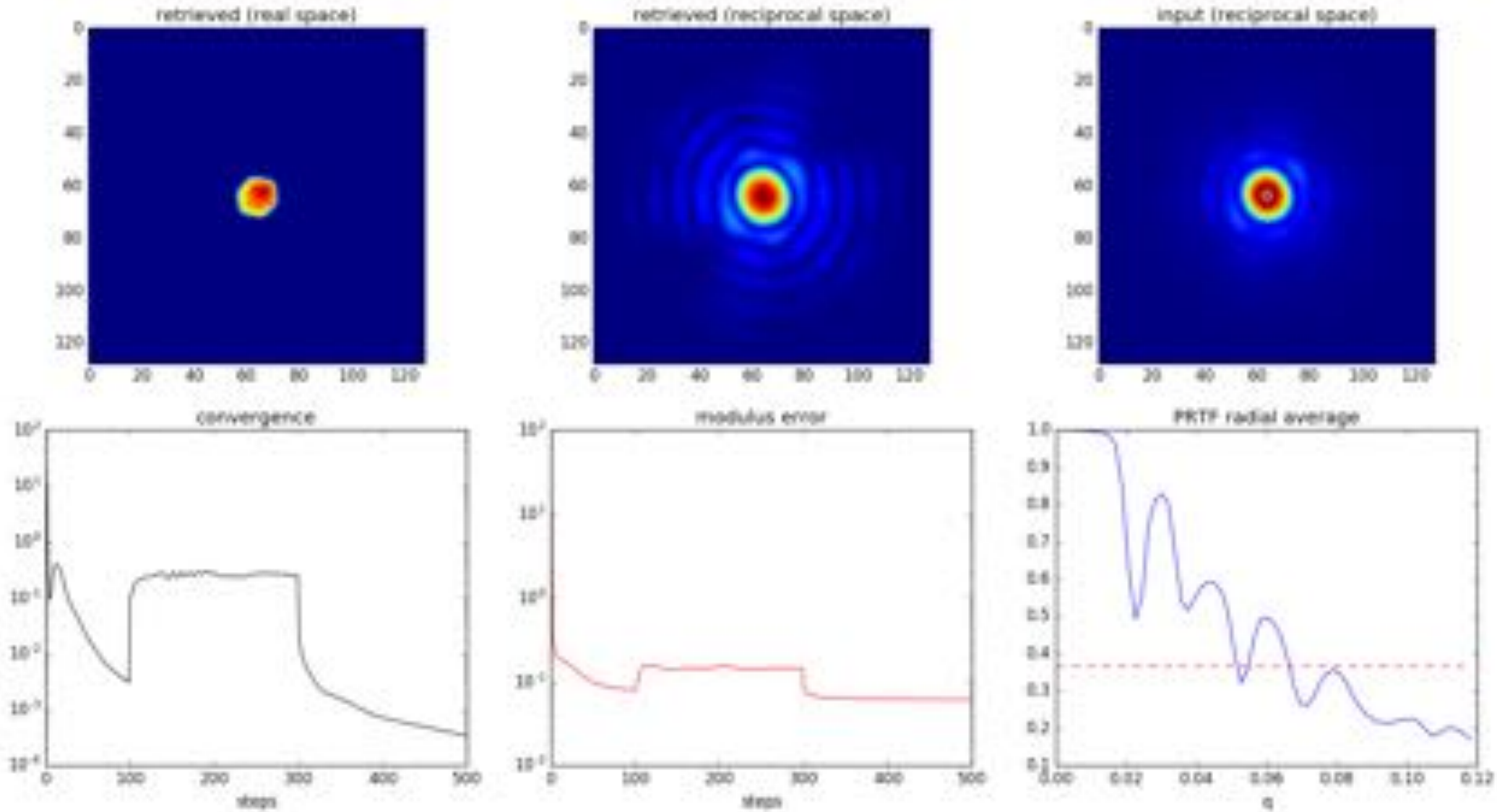
The International Workshop on Image Processing and Inverse Problems will be held at Beijing Computational Science Research Center, Beijing, China. The workshop is jointly organized by Institute of Applied Physics and Computational Mathematics and Beijing Computational Science Research Center. It aims to bring together experts on image processing and inverse problems based on optimization or PDEs. Topics of presentation are devoted to design and analysis of mathematical models, design and analysis of computational algorithms and applications of such models and algorithms to specific problems in industry.

It has become an established paradigm to formulate problems within image processing and inverse problems as some minimization problems with properly chosen energy functional. The solutions of these minimization problems leads to some variational problems or finite dimensional optimization problems. This compact, yet expressive framework makes it possible to incorporate a range of desired properties of the solutions and to design algorithms based on well-founded mathematical theory. An increasing amount of research has also approached more general problems within data analysis, real world imaging and reconstruction, and demonstrated the advantages over earlier, more established algorithms.

CNN phasing result. Estimated resolution from PRTF : 14.7nm



Graph-cut phasing result. Estimated resolution from PRTF : 15.1nm



Diffusion-map phasing result. Estimated resolution from PRTF : 11.6nm