Bio-XFEL Data Analysis Workshop
August 21, 2014, LBNL
Bldg 15-253

How good are my data?

Nicholas K. Sauter
Lawrence Berkeley National Laboratory

Software tutorial: http://cci.lbl.gov/xfel
Location at pslogin.slac.stanford.edu: /reg/g/cctbx

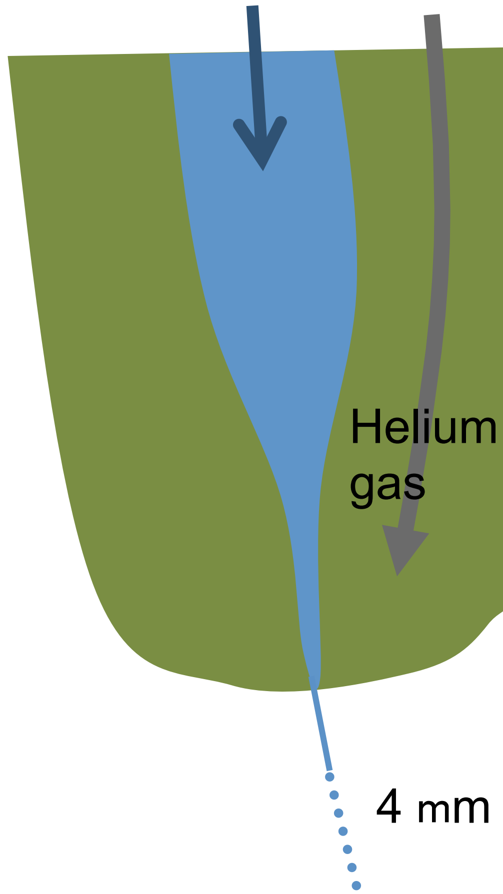
National Institutes of Health
National Institute of General Medical Sciences
BASIC DISCOVERIES FOR BETTER HEALTH

BERKELEY LAB
Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY

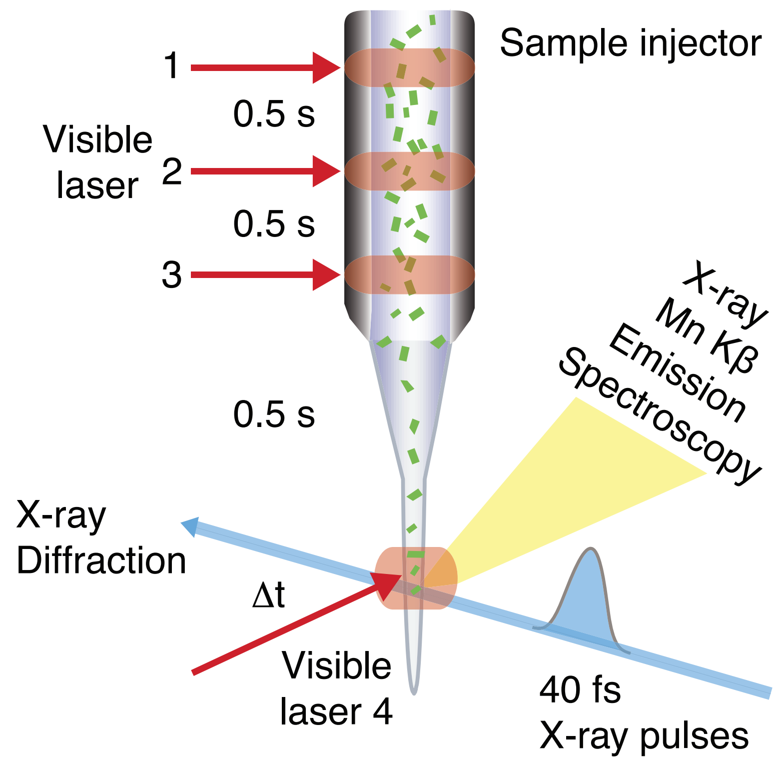
XFEL diffraction results processed with *cctbx.xfel*

Gas dynamic virtual nozzle



- L220 lysozyme (in tutorial)
- Gd-lysozyme (in tutorial)
- *Bacillus* Cry3A δ-endotoxin *in vivo* crystals
- Polyhedrin crystals (submitted)

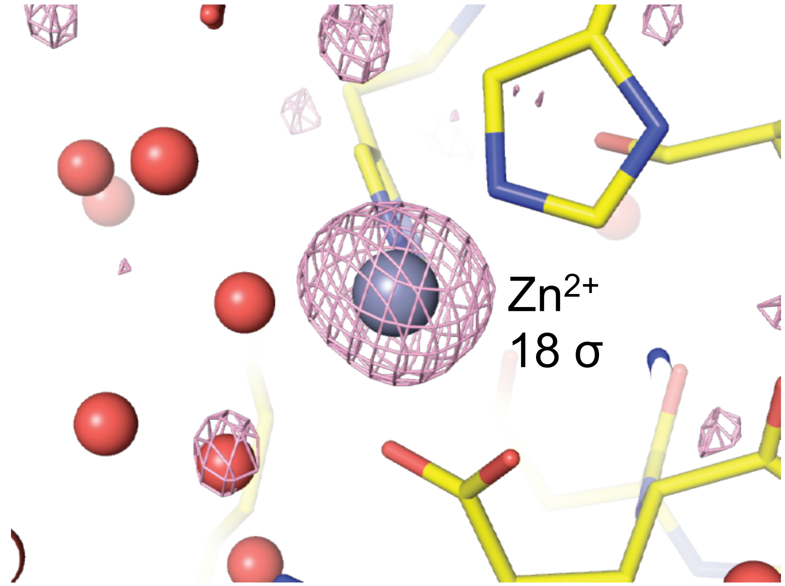
Electrospun microjet



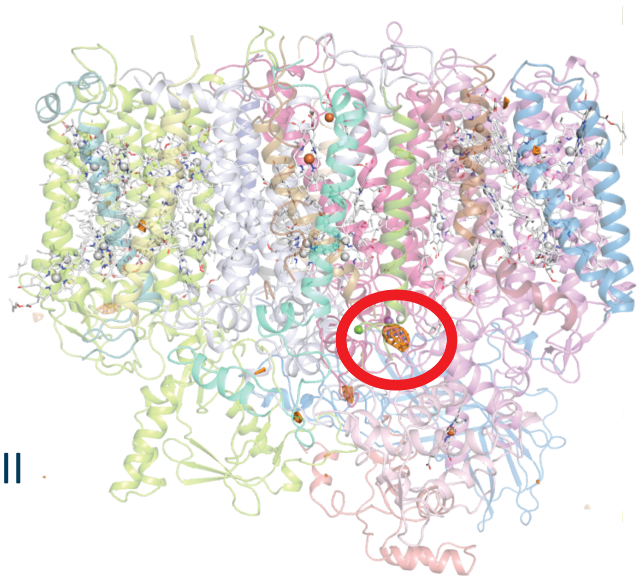
- Photosystem II redox states
- Thermolysin (in tutorial)

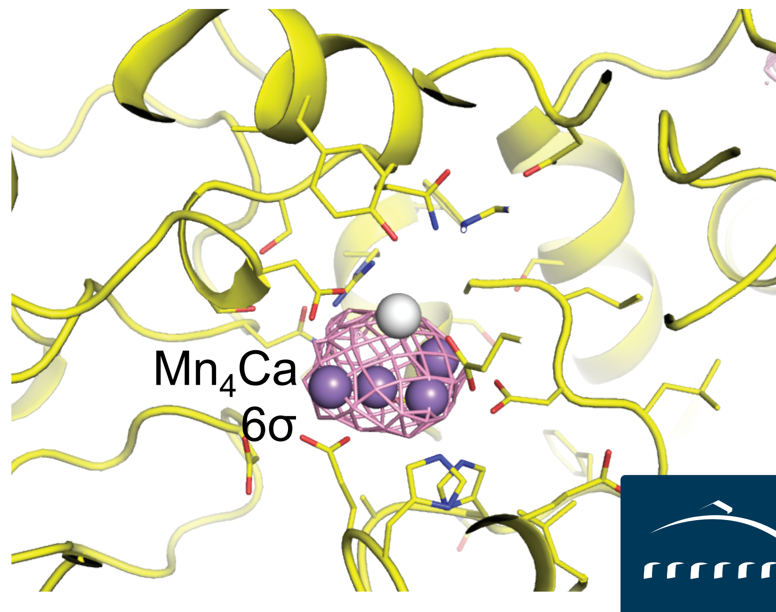
BERKELEY LAB
Lawrence Berkeley National Laboratory

Anomalous difference maps detect Zn, Mn



thermolysin





photosystem II
S_0 state

Kern (2014) Nature Communications 5: 4371

BERKELEY LAB
Lawrence Berkeley National Laboratory

The *cctbx.xfel* workflow:

------------------------------------SLAC computing---------------——————----------------Laptop

pyana / cctbx module

```
~$USER/myrelease/cxi84914
  /*.cfg
  /*.phil
cxi.lsf → batch queues
```

Original data XTC format → *Index and integrate* → *Merging* → *Structure solution*

Original data:

```
/reg/d/psdm/cxi/cxi84914/xtc
  e157-r0021-s00-c00.xtc
                -c01.xtc
            -s01-c00.xtc
```

Index and integrate:

```
/reg/d/psdm/cxi/cxi84914/scratch
  /r0021/000/stdout/s00.log
            /out/idx-s00-*.pickle
            /integration/int-*.pickle
```

Merging:

```
cxi.merge
  → *.mtz
```

Structure solution:

```
phenix
```

cctbx standalone

```
any directory
cxi.index idx*.pickle *.phil
cctbx.image_viewer
```

Software tutorial: http://cci.lbl.gov/xfel

cctbx.xfel at pslogin.slac.stanford.edu: /reg/g/cctbx

This powerpoint: http://cci.lbl.gov/xfel/2014_workshop

BERKELEY LAB
Lawrence Berkeley National Laboratory

Cctbx.xfel goals for the Bio-XFEL workshop

- Work through L498 thermolysin data (or a subset)
  - First worked example from XTC to final maps
  - Detect anomalous signal?
- Try the Gd-lysozyme data
- Start tonight
  - 2-hour work session insufficient to actually work the data
  - Discover student questions not anticipated
  - Application to students' own data
  - Gd-lysozyme not yet documented on the wiki
- Extended topics
  - LCLS instruments CXI or XPP; SACLA (Aaron Brewster)
  - Space groups with a lattice ambiguity (Wolfgang Brehm)
  - Sparse data (Oliver Zeldin)
  - Migration to psana (Chris O'Grady)

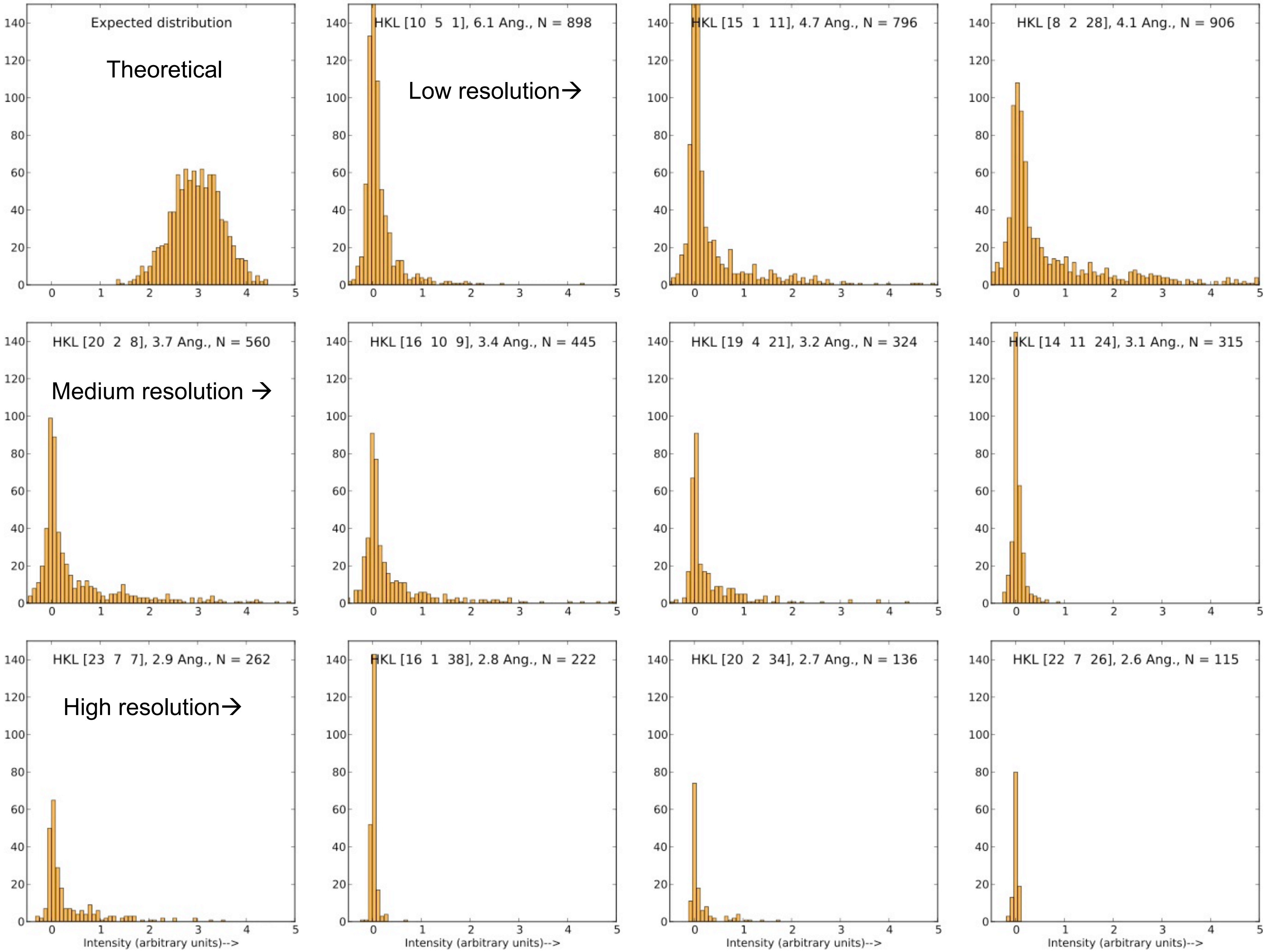
Software tutorial: http://cci.lbl.gov/xfel

cctbx.xfel at pslogin.slac.stanford.edu: /reg/g/cctbx

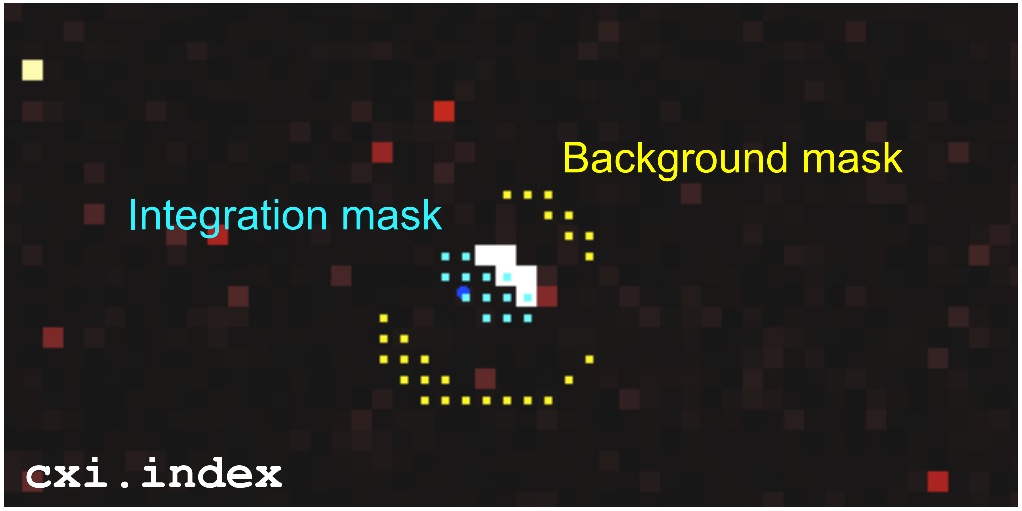
This powerpoint: http://cci.lbl.gov/xfel/2014_workshop

BERKELEY LAB
Lawrence Berkeley National Laboratory

Ensemble distribution of measurements of single Bragg spots



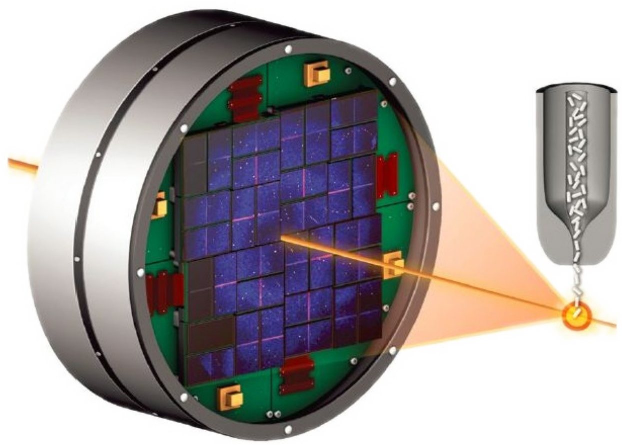
Signal to noise and model accuracy are major issues



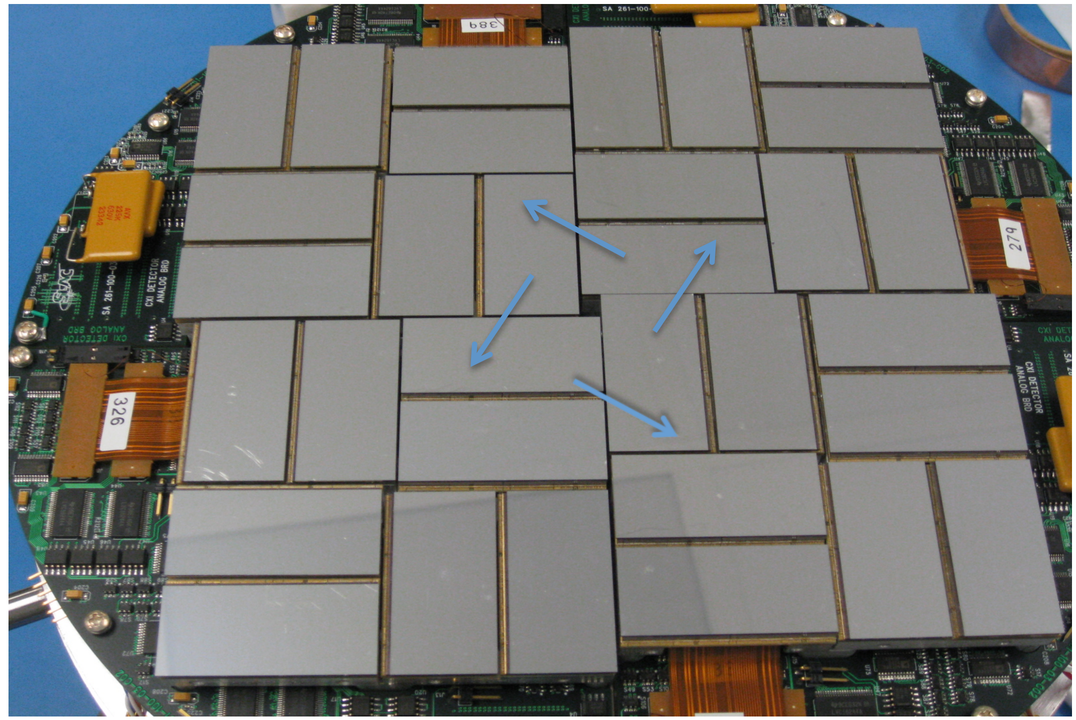
- Well conforming spot shape models intended to enhance signal to noise
- But are susceptible to positional error

Exact relationship of sensor tiles is crucial for modeling reflections

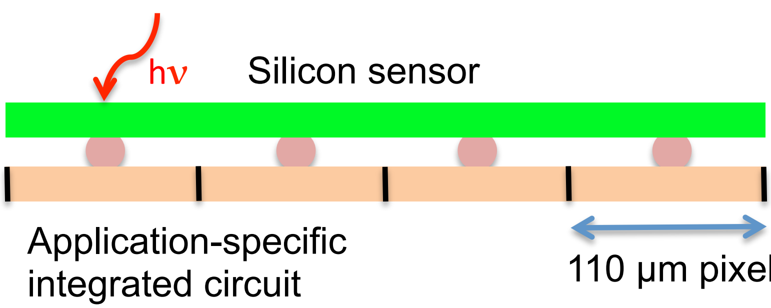
[Hart *et al.*, 2012, *Proc. of SPIE*]

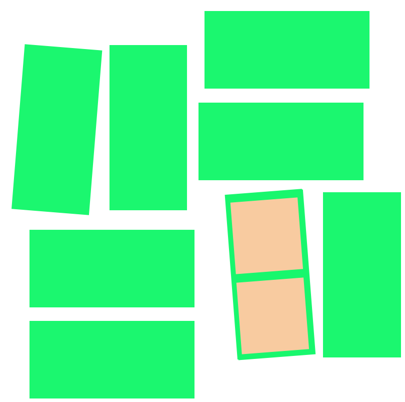


Cornell-SLAC pixel array detector (CS-PAD) at LCLS CXI endstation



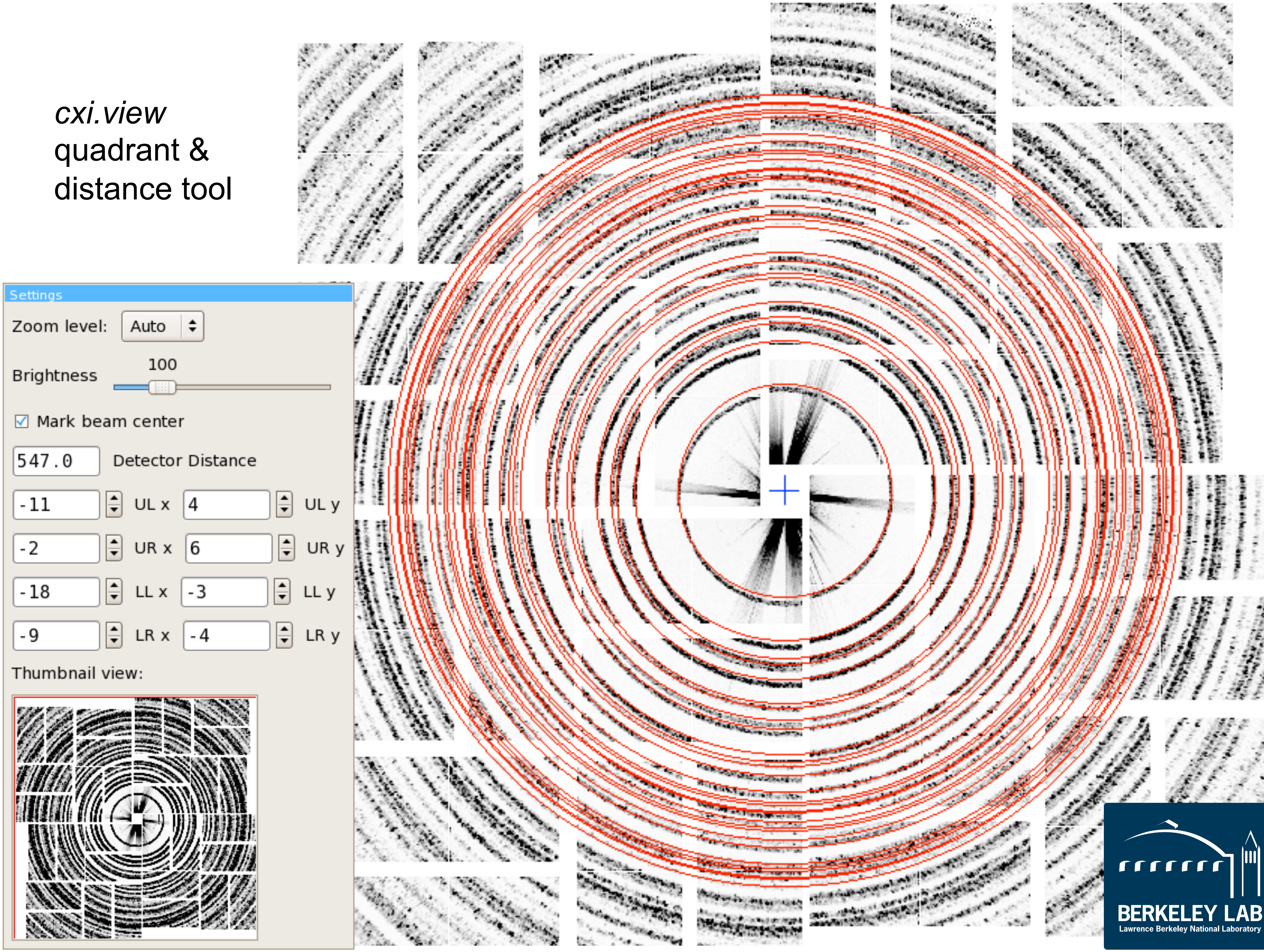
Quadrants are on movable rails



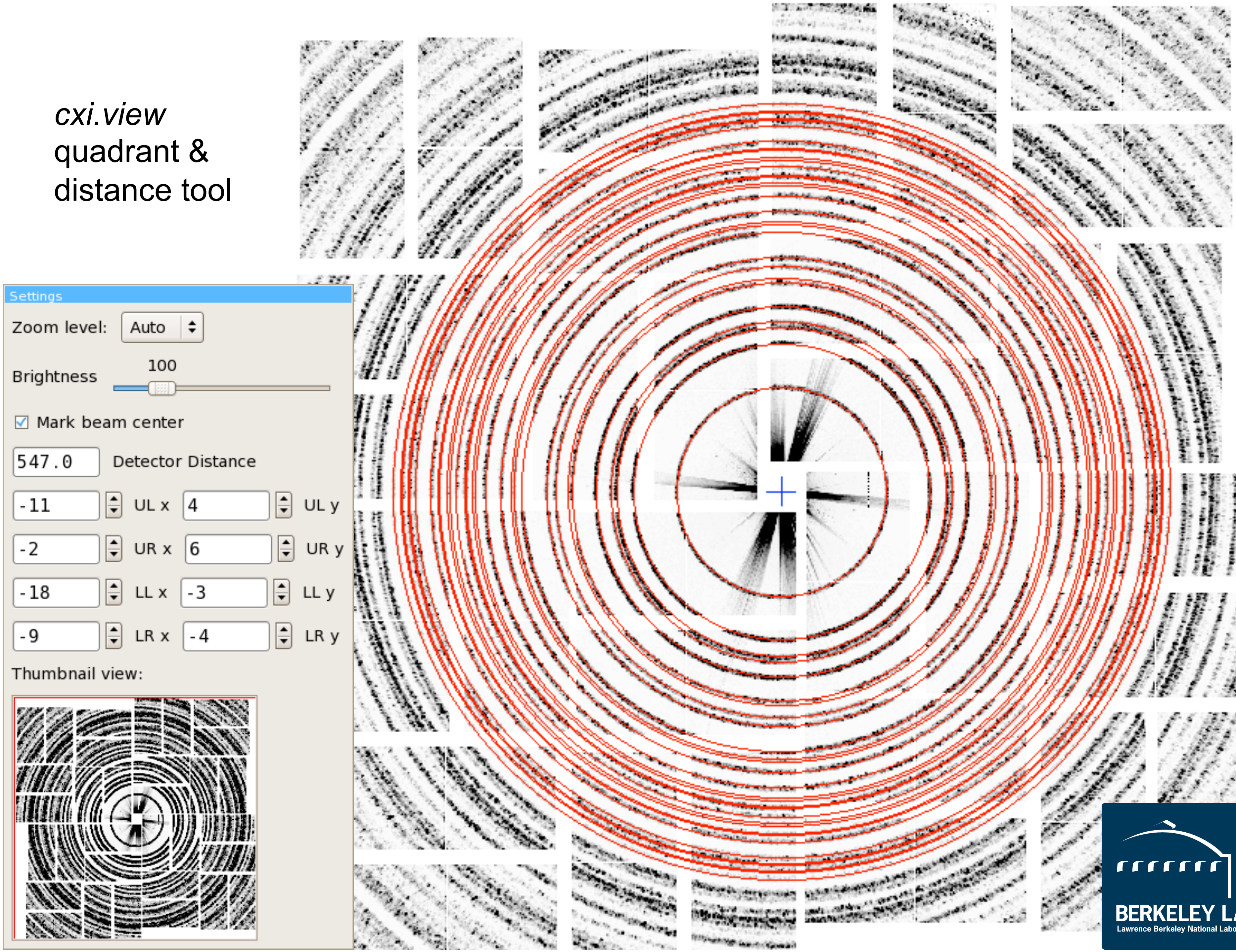


Sensors are field-serviceable

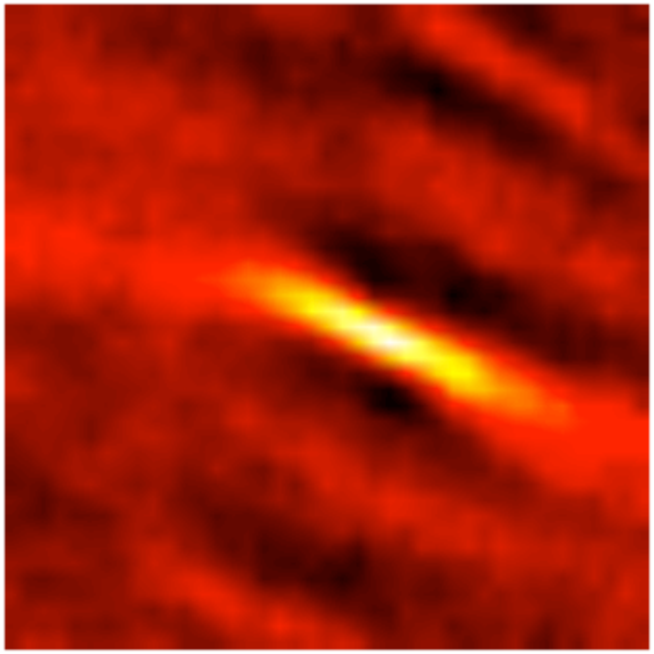
cxi.view
quadrant &
distance tool



cxi.view
quadrant &
distance tool



cspad.quadrants tool for automatic detection



- As a function of quadrant position (x & y), autocorrelation of the ring pattern with itself upon 45° rotation.
- Do the same with all 4 quadrants.

Toolbox is now refactored; the user now places the result x,y quadrant displacements directly in the phil file:

```
distl {
  quad_translations=3 -4 -3 -7 -12 0 -7 -5
}
```

BERKELEY LAB
Lawrence Berkeley National Laboratory

Detector versioning

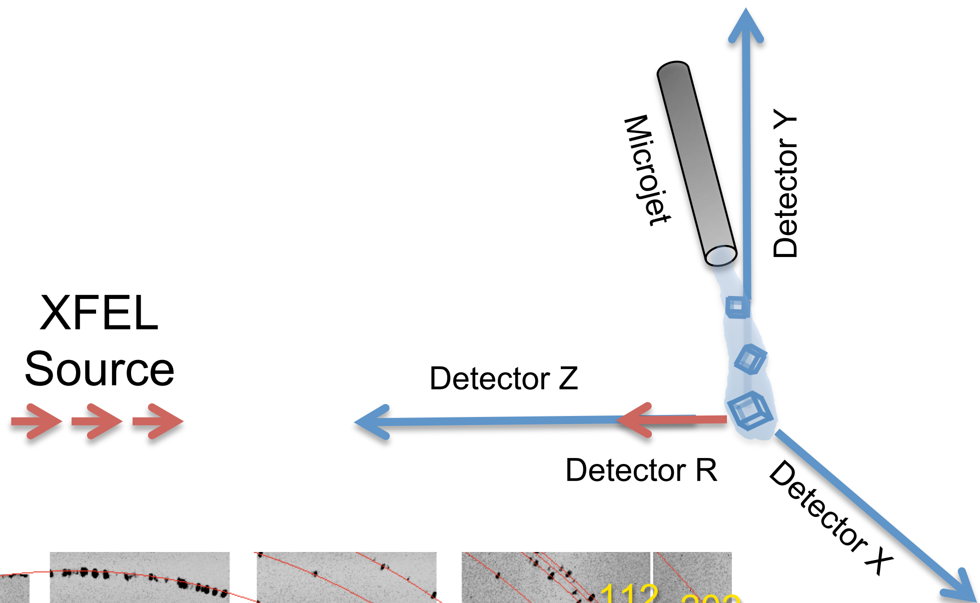
The new system brings metrology under user rather than developer control.
Default detector geometries are still hard-coded based on date stamp.

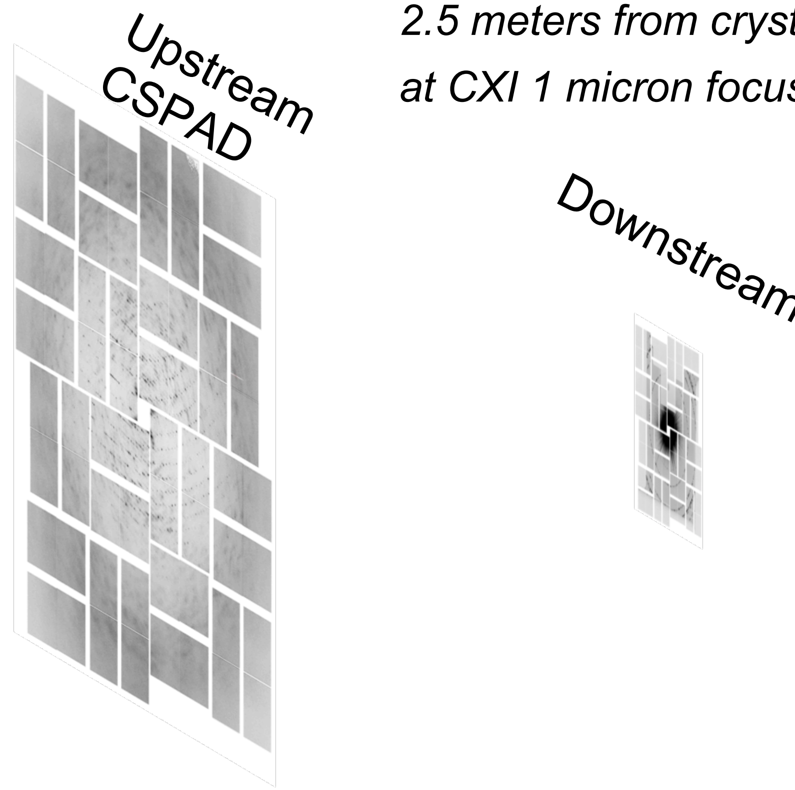
```
> cxi.detector_format_versions

Format version   Det. address      Start time   End time
       CXI 3.2   CxiDs1-0|Cspad-0 Sep 2010     Mar 2011
       CXI 4.1   CxiDs1-0|Cspad-0 Mar 2011     Oct 2011
       CXI 5.1   CxiDs1-0|Cspad-0 Oct 2011     May 2012
       CXI 6.1   CxiDs1-0|Cspad-0 May 2012     Jan 2013
       CXI 7.1   CxiDs1-0|Cspad-0 Jan 2013     Aug 2013
       CXI 7.d   CxiDsd-0|Cspad-0 Jan 2013     Aug 2013
       CXI 8.1   CxiDs1-0|Cspad-0 Aug 2013     Jan 2014
       CXI 8.2   CxiDs1-0|Cspad-0 Jan 2014     Mar 2014
       CXI 8.d   CxiDsd-0|Cspad-0 Aug 2013     Mar 2014
       CXI 9.1   CxiDs2-0|Cspad-0 Mar 2014     Aug 2014
   Sacla.MPCCD        Sacla.MPCCD None         None
       XPP 7.1   XppGon-0|Cspad-0 Jan 2013     Aug 2013
  XPP 7.marccd  XppGon-0|marccd-0 Jan 2013     Aug 2013
       XPP 8.1   XppGon-0|Cspad-0 Aug 2013     Mar 2014
  XPP 8.marccd  XppGon-0|marccd-0 Aug 2013     Mar 2014
       XPP 9.1   XppGon-0|Cspad-0 Mar 2014     Aug 2014
```

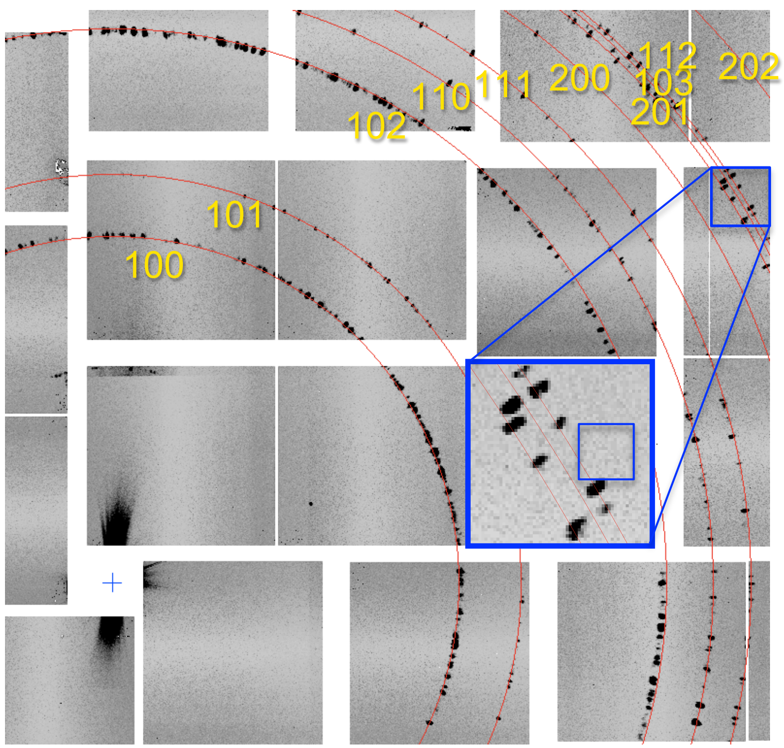
BERKELEY LAB
Lawrence Berkeley National Laboratory

Downstream detector





2.5 meters from crystal
at CXI 1 micron focus

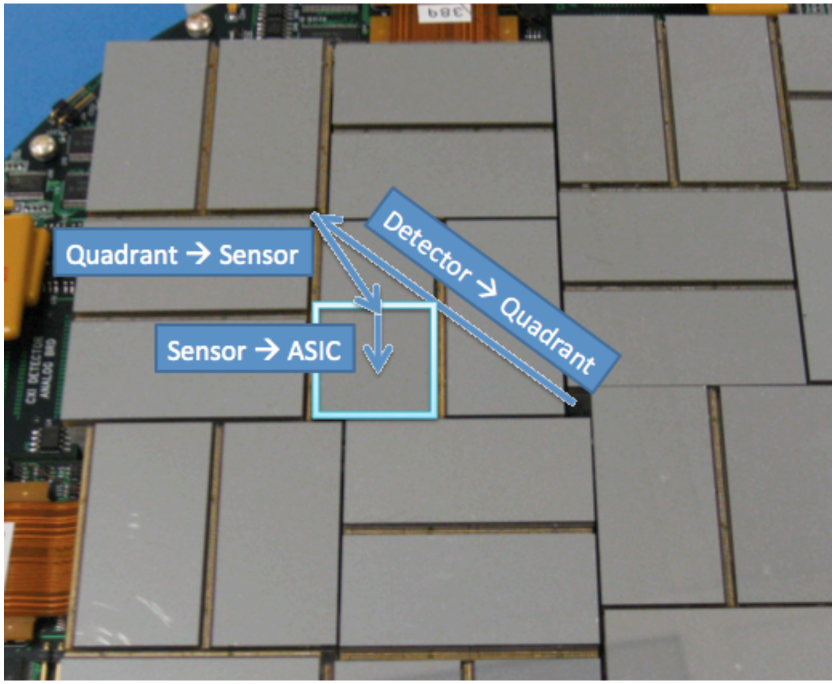


Indexed the lowest-resolution thermolysin powder rings

BERKELEY LAB
Lawrence Berkeley National Laboratory

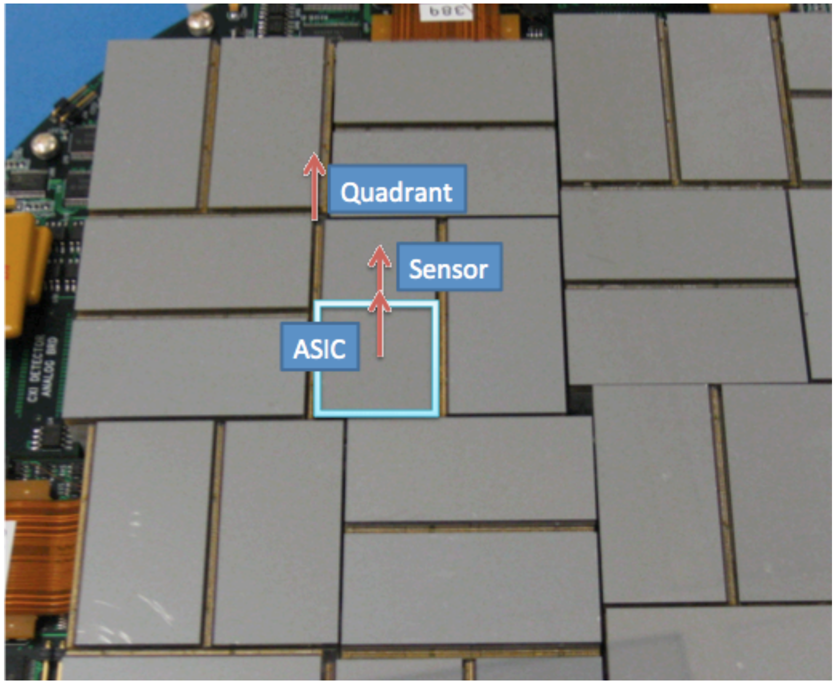
We are converging on a data format to capture the hierarchical design

- ImageCIF/ CBF: international standard file format
- Formal language for describing the experiment (and the data)
- Herbert Bernstein – Dowling College; critical collaborator
- DIALS software package will handle this standard format
- Will use HDF5 containers



Hierarchical organization:

Detector -> Quadrant -> Sensor -> ASIC

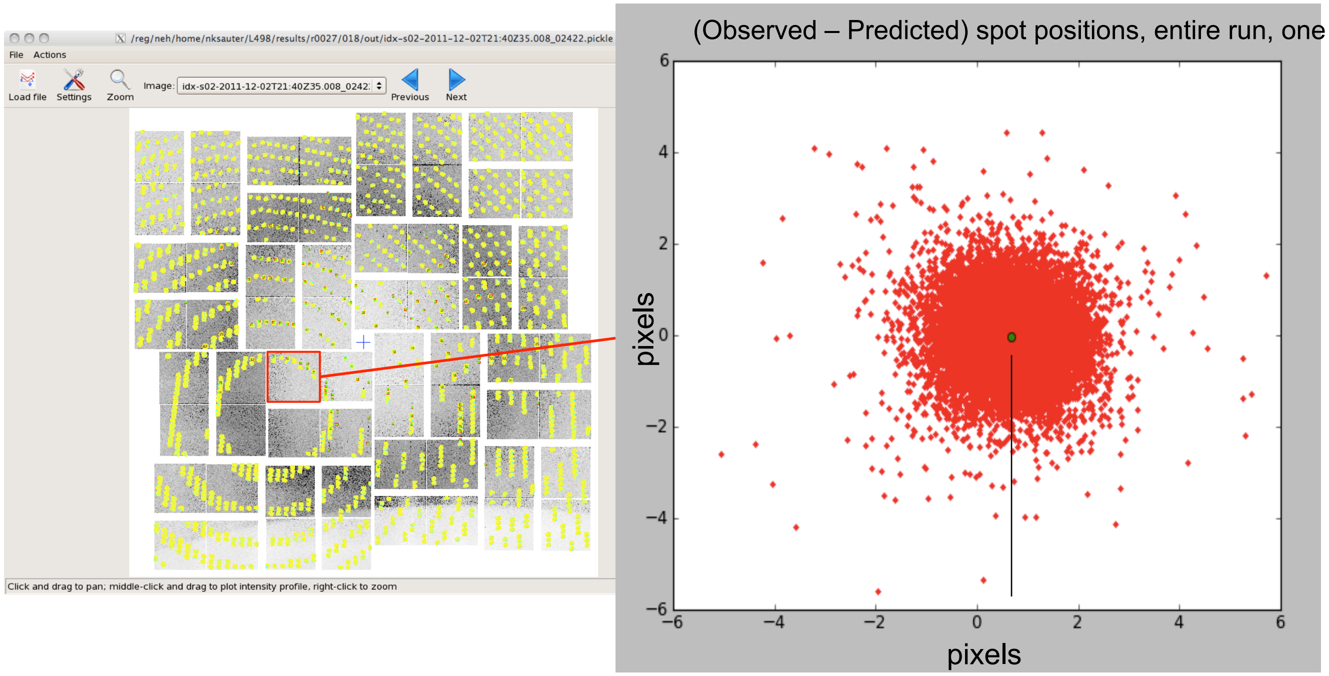


Each level:

Relative 3D translation and 3D rotation, as well as readout fast & slow axes

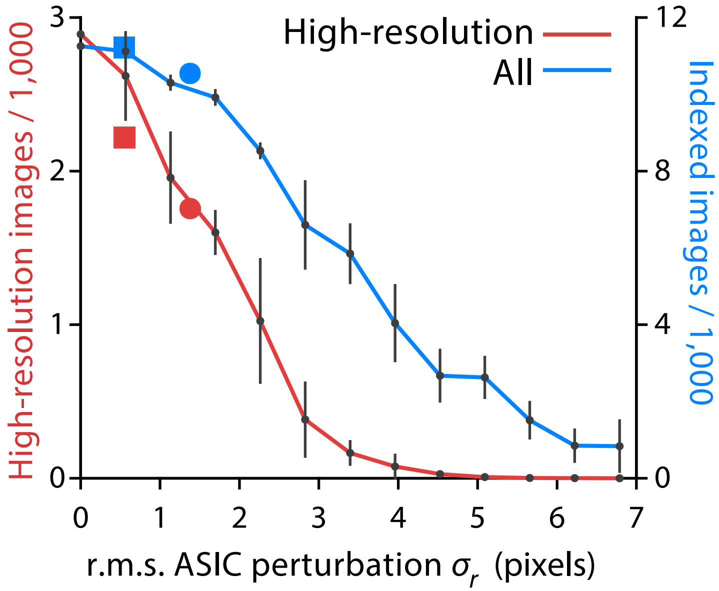
Brewster AS et. al (2014). "XFEL Detectors and ImageCIF." Comp. Crystallogr. Newsl. **5**:19–24

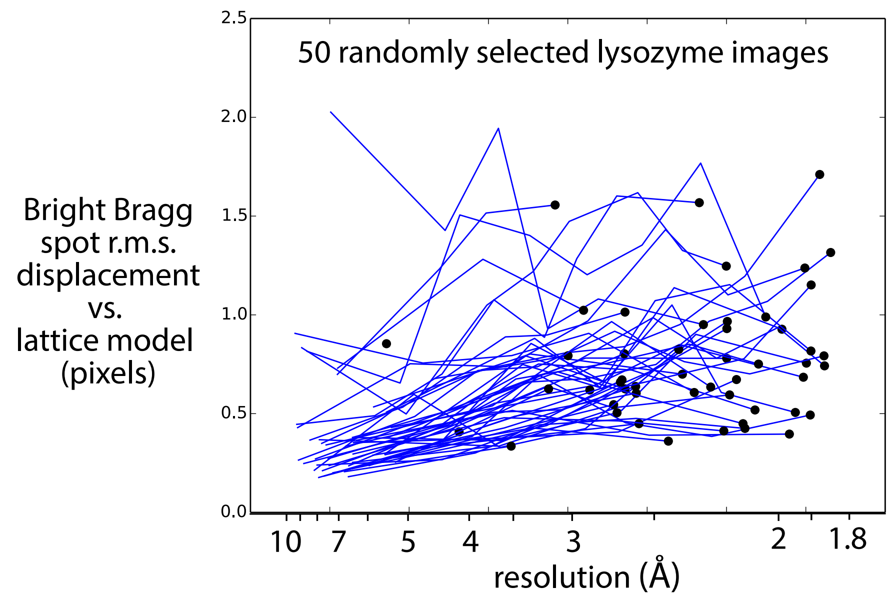
cspad.metrology tool: whole-pixel and sub-pixel corrections



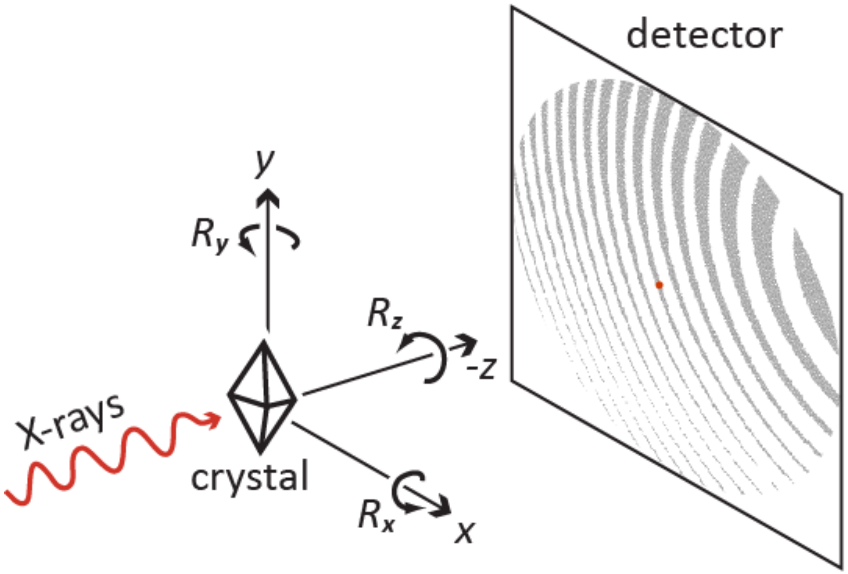
	Overall RMSD, all Bragg spots	Average Tile Displacement
As-given with manual quad placement `distl {quad_translations= }`	2.15 pixels	1.35 pixels
Nearest-pixel tile placement `distl {tile_translations= }`	1.86 pixels	0.40 pixels
Sub-pixel tile translations `integration {subpixel_joint_model{translations}}`	1.39 pixels	0.23 pixels
Rotations too `integration {subpixel_joint_model{rotations}}`	0.65 pixels	Near zero

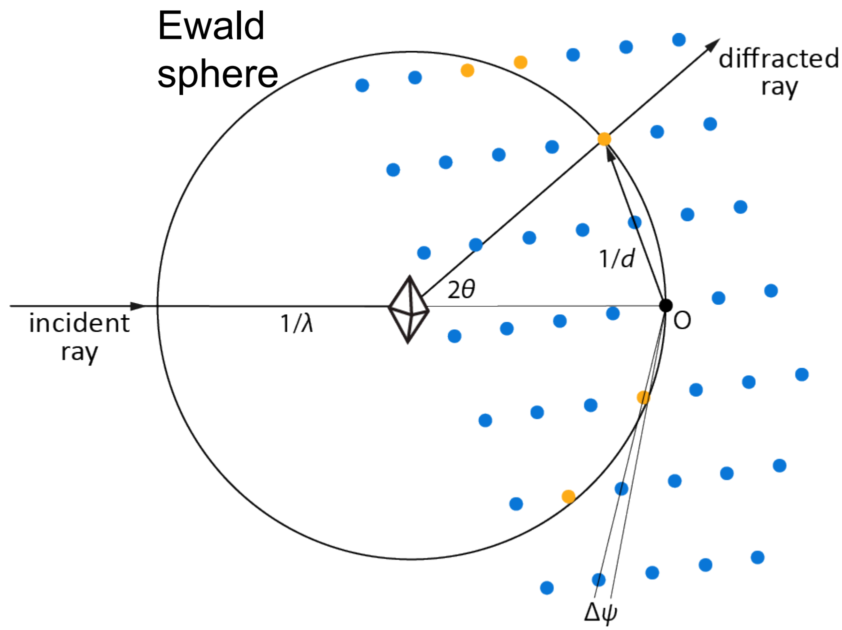
Varying the resolution cutoff for each image





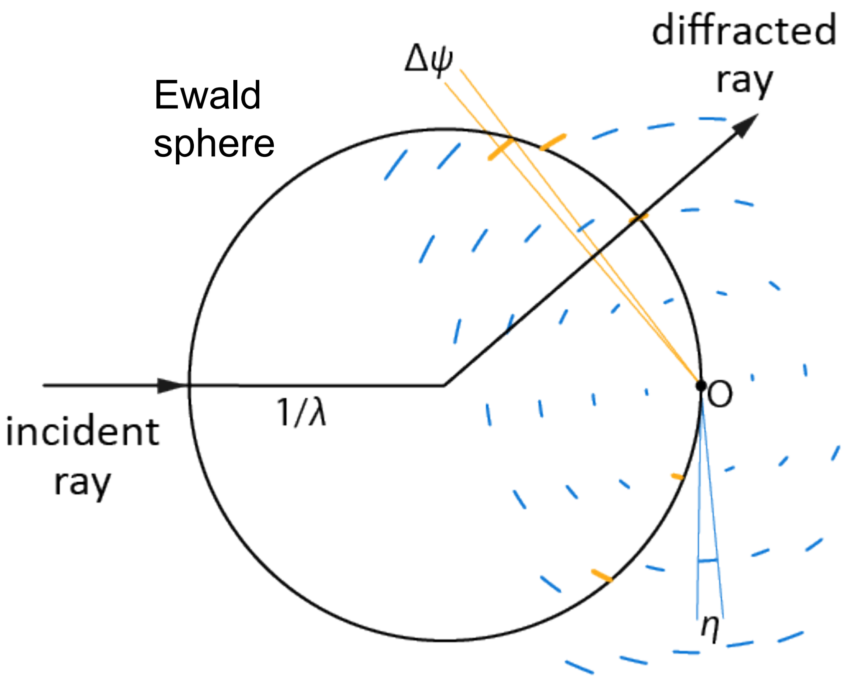
It is difficult to get crystal orientation from a still shot



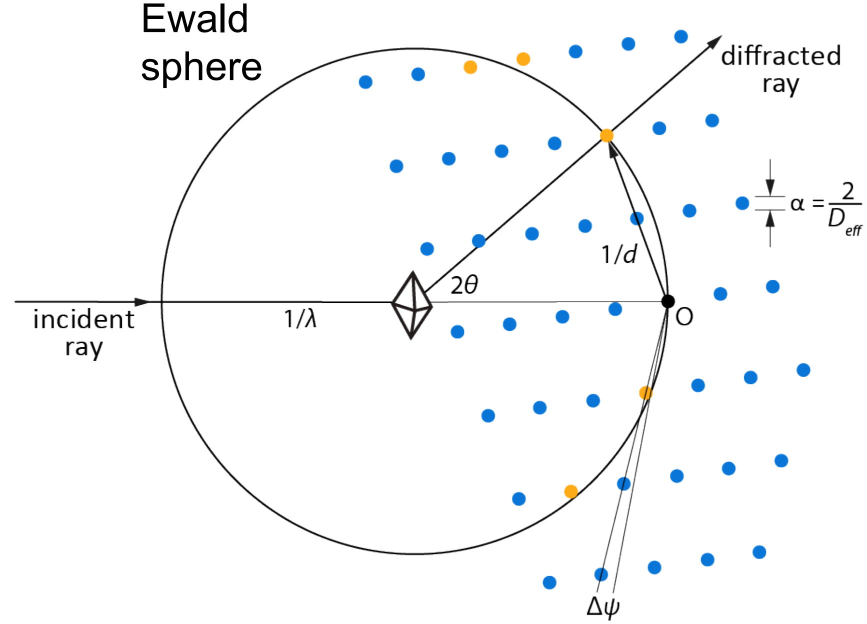


$$f = \sum_{\substack{spotfinder \\ spots}} \left(\mathbf{r}_{obs} - \mathbf{r}_{calc}\right)^2 + \left(\Delta\psi\left[rotx, roty\right]\right)^2$$

Crystal disorder is a combination of mosaicity and block size



Effective mosaicity η

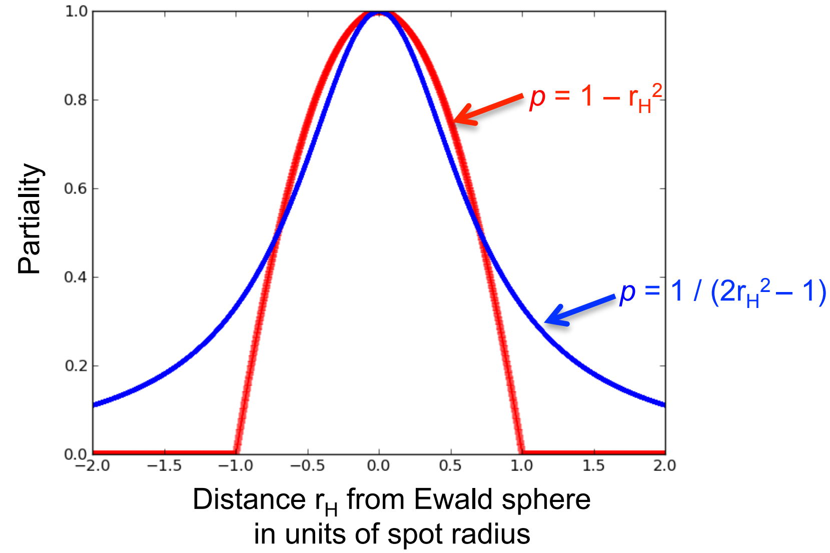


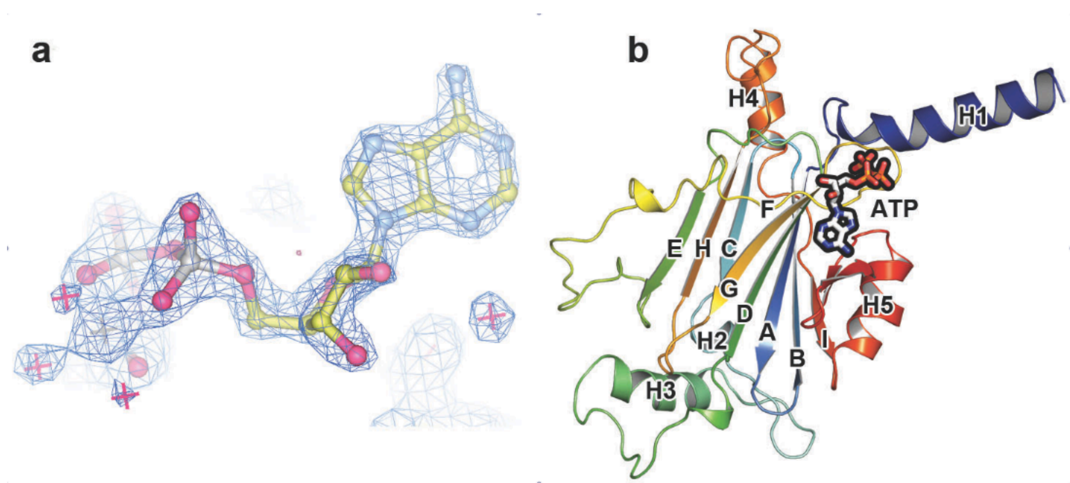
Effective size D_{eff} of the coherently scattering mosaic block

Best-parameter fitting in cctbx.xfel; algorithm defined in phil file:
integration.mosaic.refinement_target=ML

BERKELEY LAB
Lawrence Berkeley National Laboratory

Rough mathematical expression for spot partiality





Cypovirus polyhedron crystal/ David Stuart group

- Molecular replacement, 10% sequence identity

Δψ angle offers a correction for partiality, supported by intensity statistics

N(Z) test

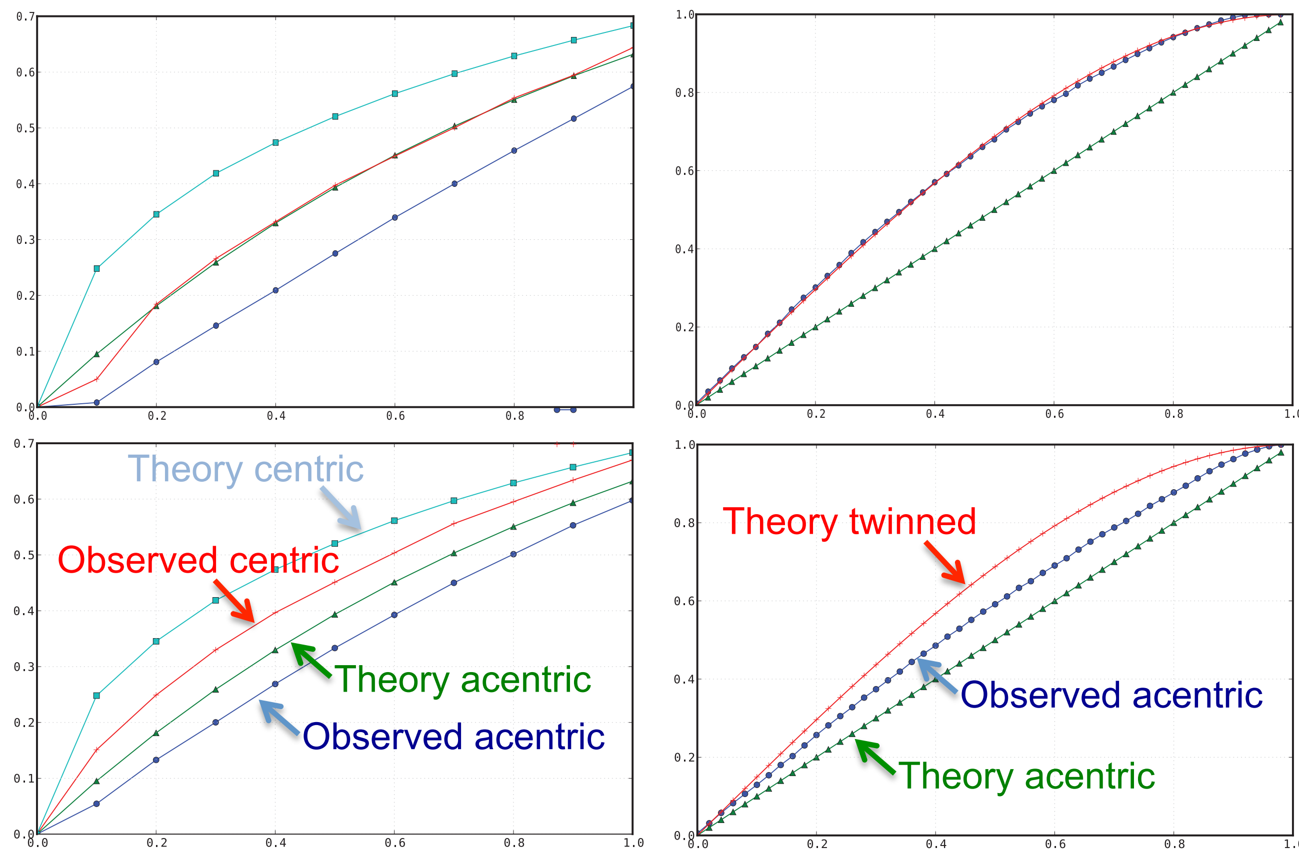
L test

Initial scaling & merging:

$\langle I^2 \rangle / \langle I \rangle^2 = 1.5$

Rough partiality correction:

$\langle I^2 \rangle / \langle I \rangle^2 = 1.7$ (theoretical 2)



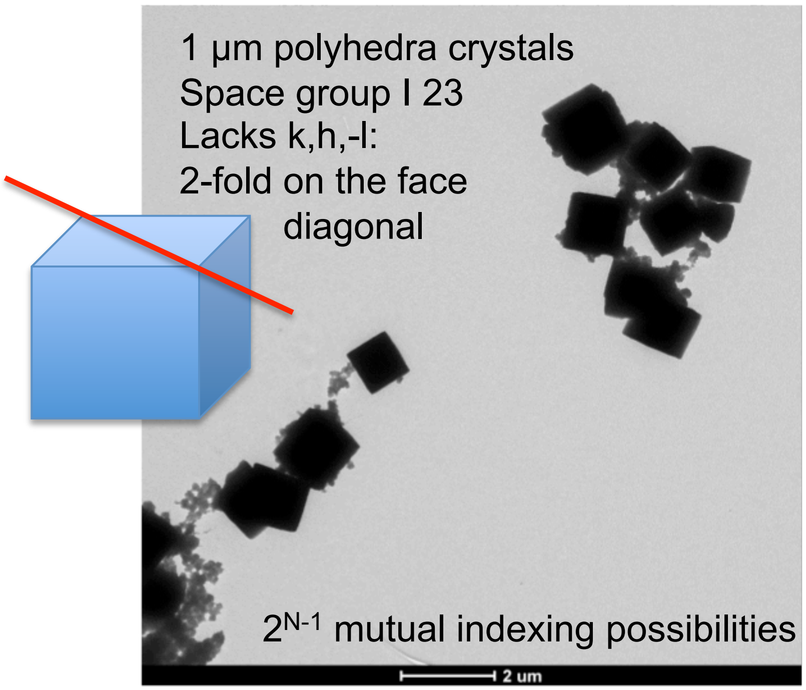
$Z = I / \langle I \rangle$

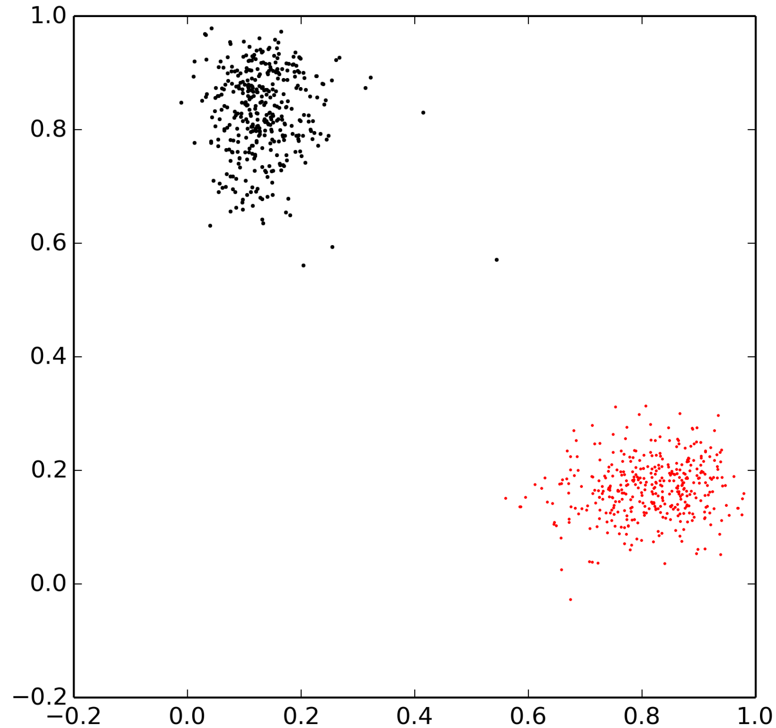
$|L| = |\ (I_1 - I_2) / (I_1 + I_2)\ |$

Thermolysin data

Resolving an indexing ambiguity

- 28% of crystal structures have space groups where the lattice symmetry is higher than the space group symmetry



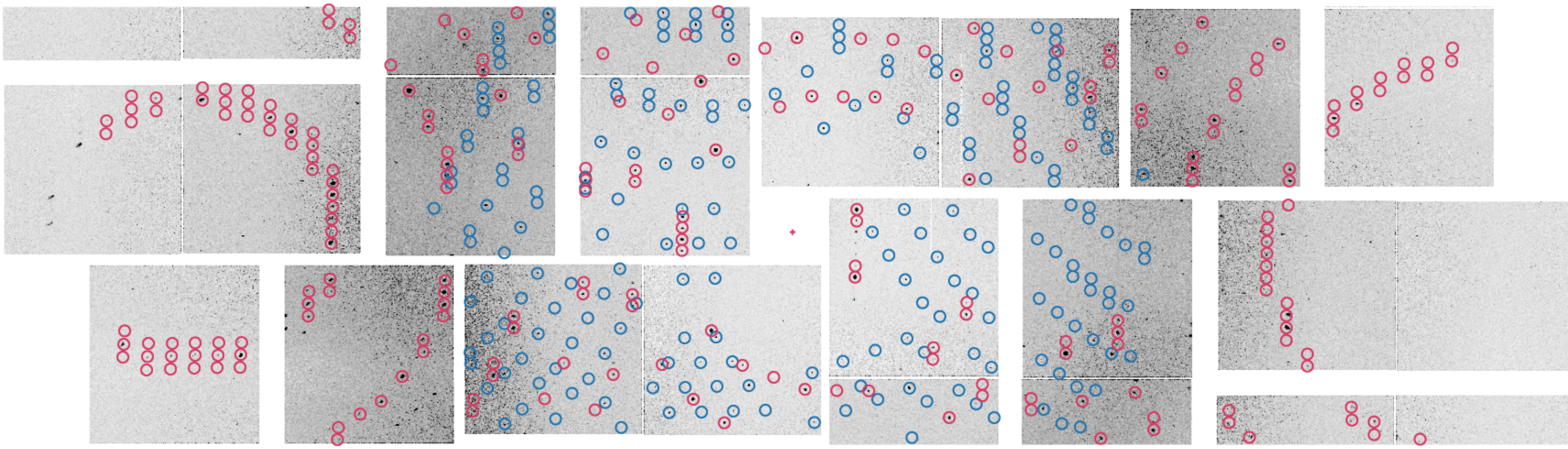


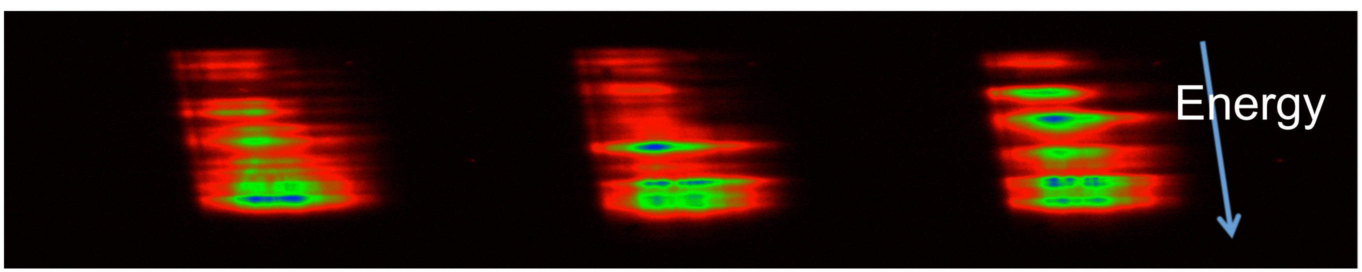
- Brehm & Diederichs (2014) Acta D70: 101
- Break the ambiguity by pairwise comparison of image-to-image intensity correlation coefficient
- 768 lattices sorted into two piles

cxi.brehm_diederichs tool is documented at
http://cci.lbl.gov/xfel/index.php/Resolving_an_Indexing_Ambiguity

BERKELEY LAB
Lawrence Berkeley National Laboratory

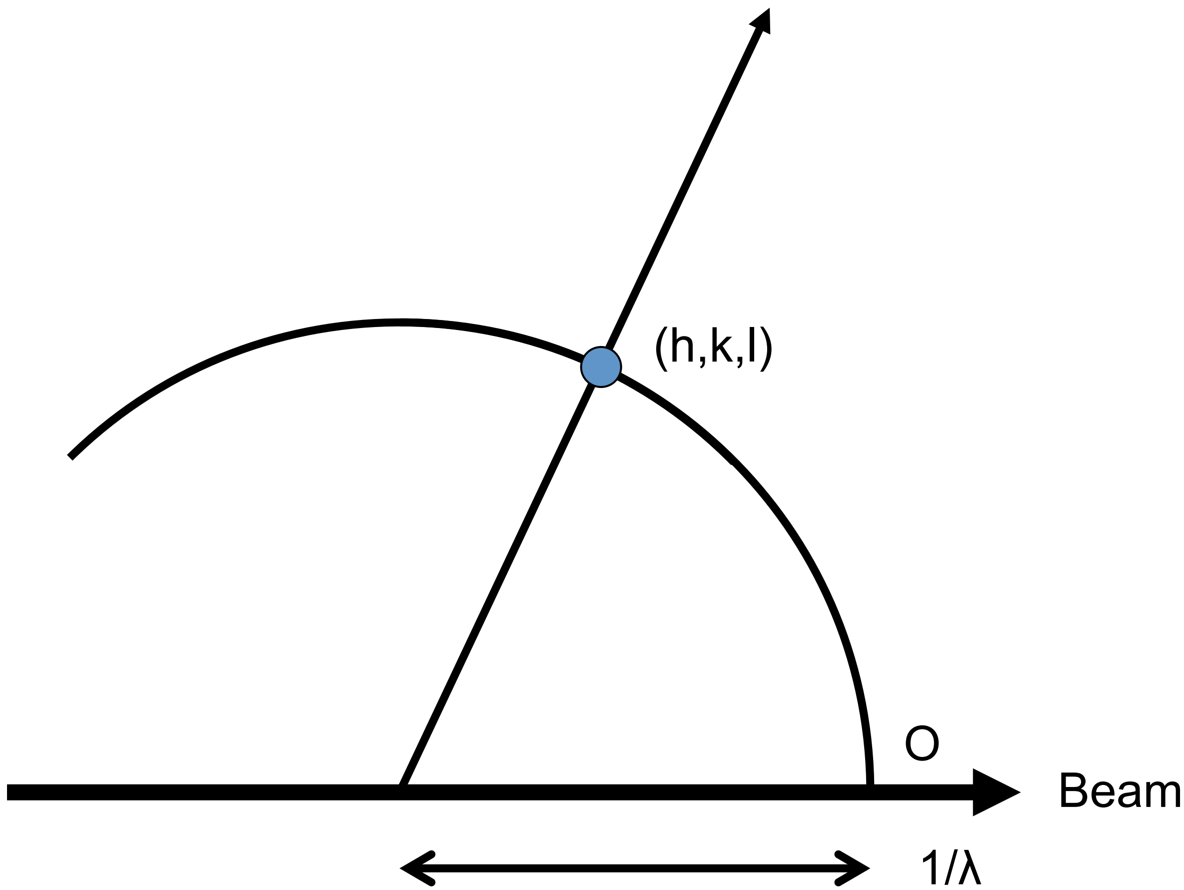
Challenges: Radial streaking





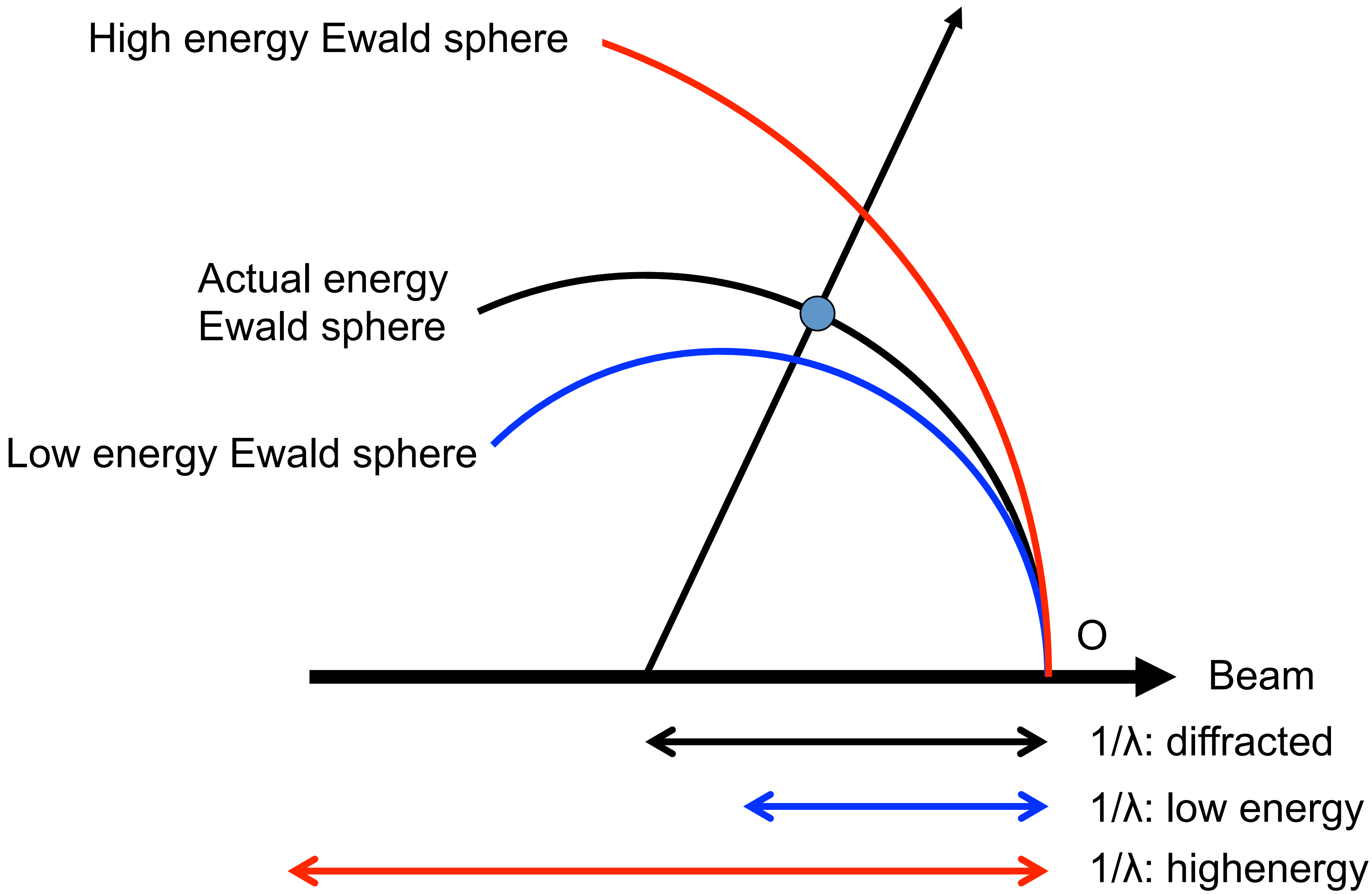
Shot-to-shot measurement of the X-ray incident pulse

Ewald sphere

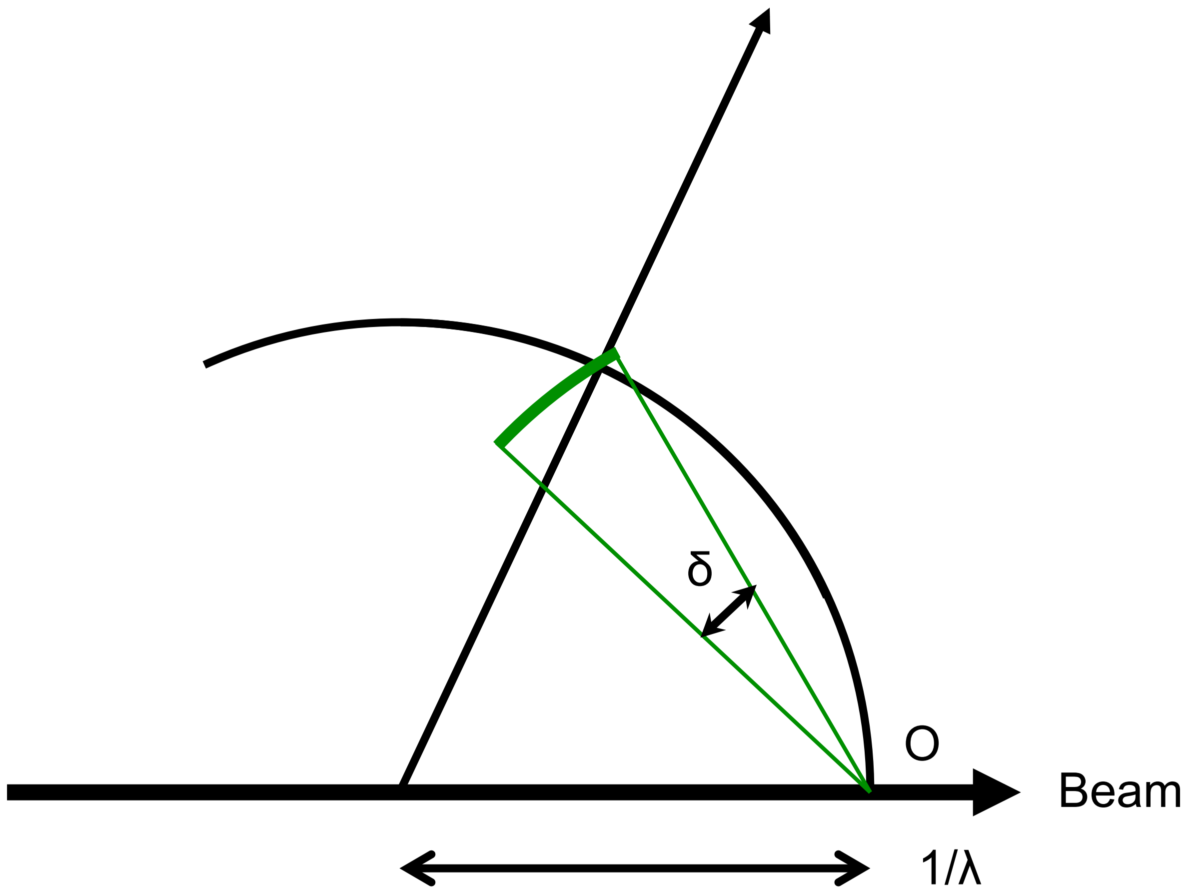


Slide sequence: Aaron Brewster

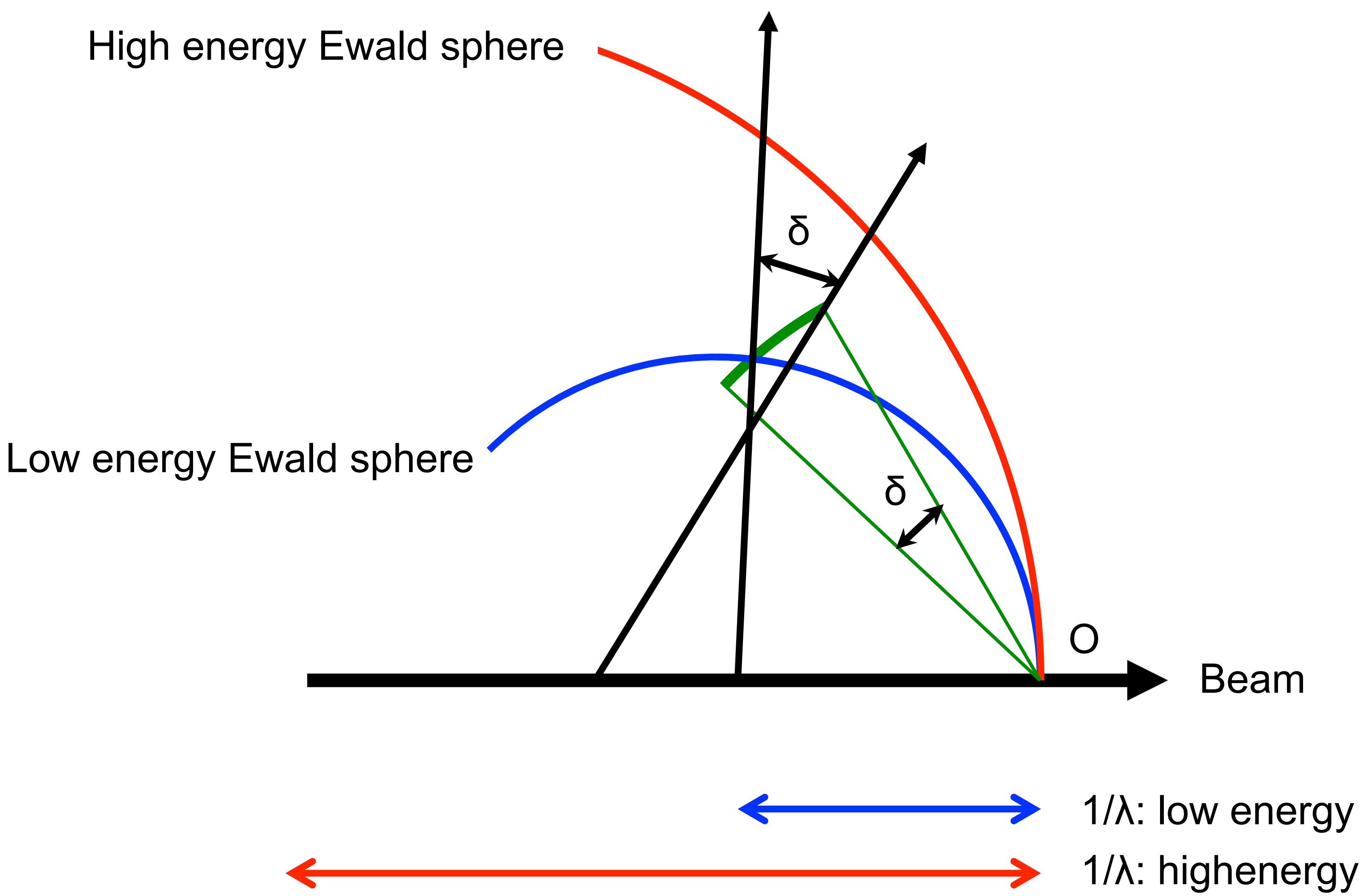
Ewald sphere paradigm with high/low bandpass limits



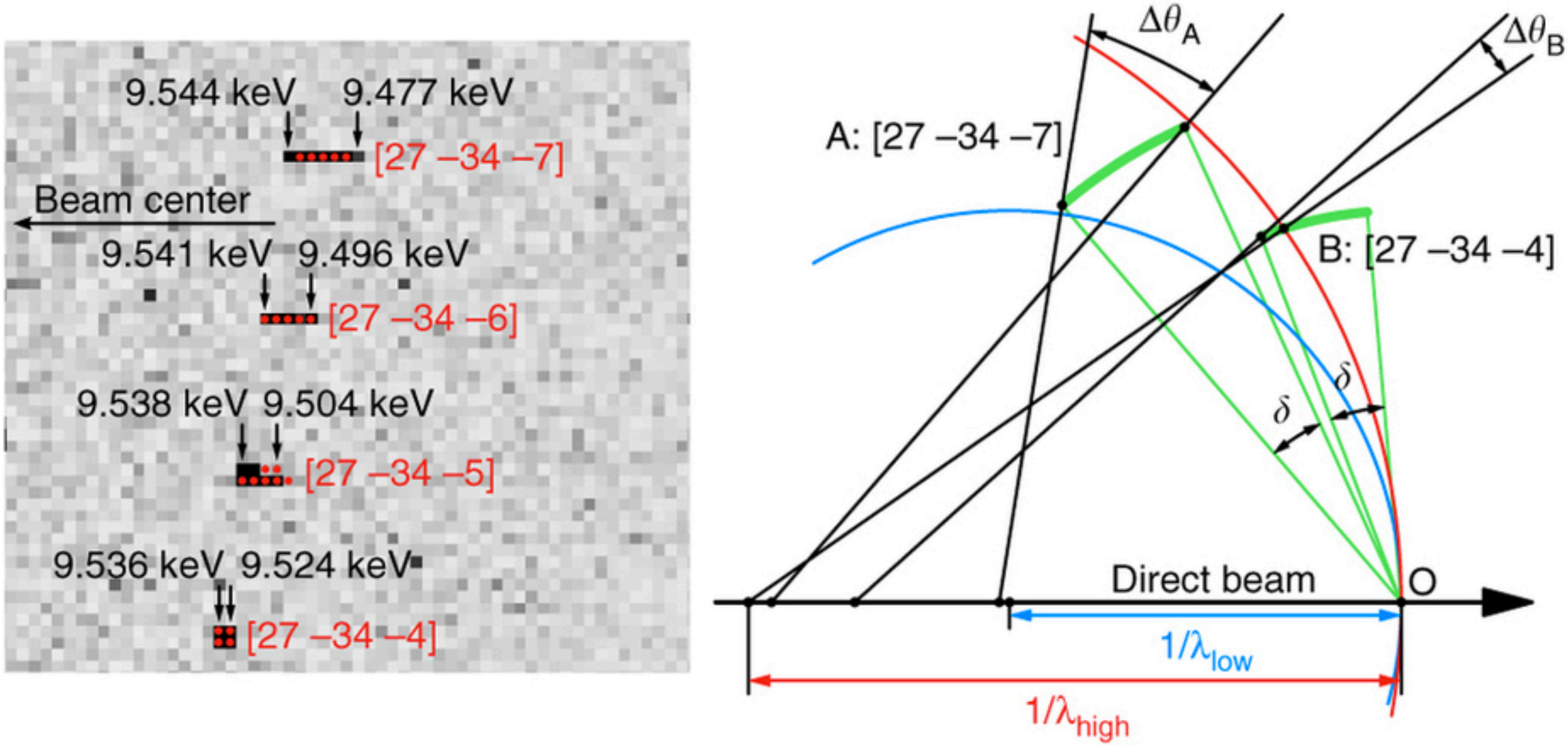
Ewald sphere with mosaicity arclets



Full model: bandpass + mosaicity

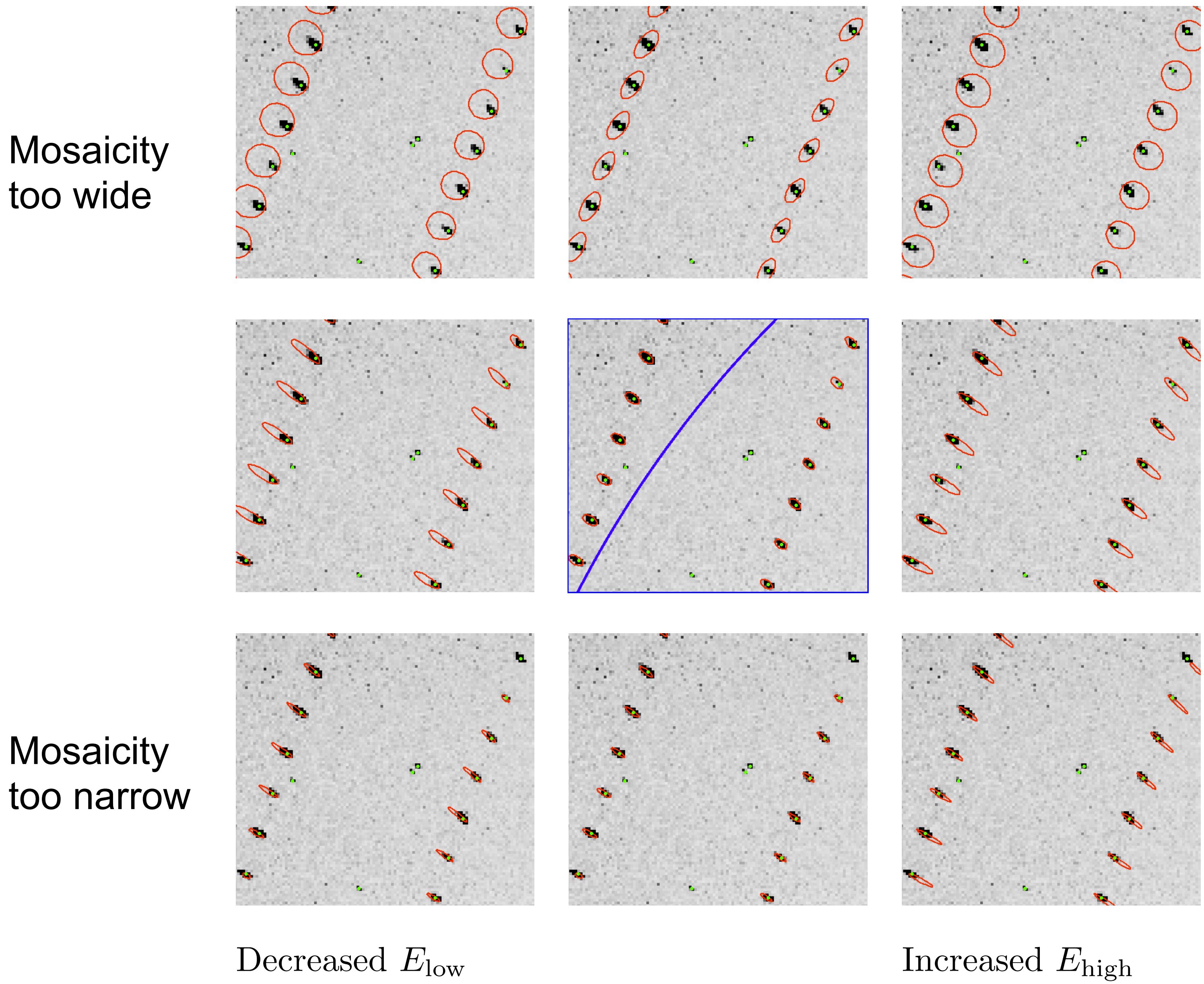


Radial dispersion: modeling each pixel

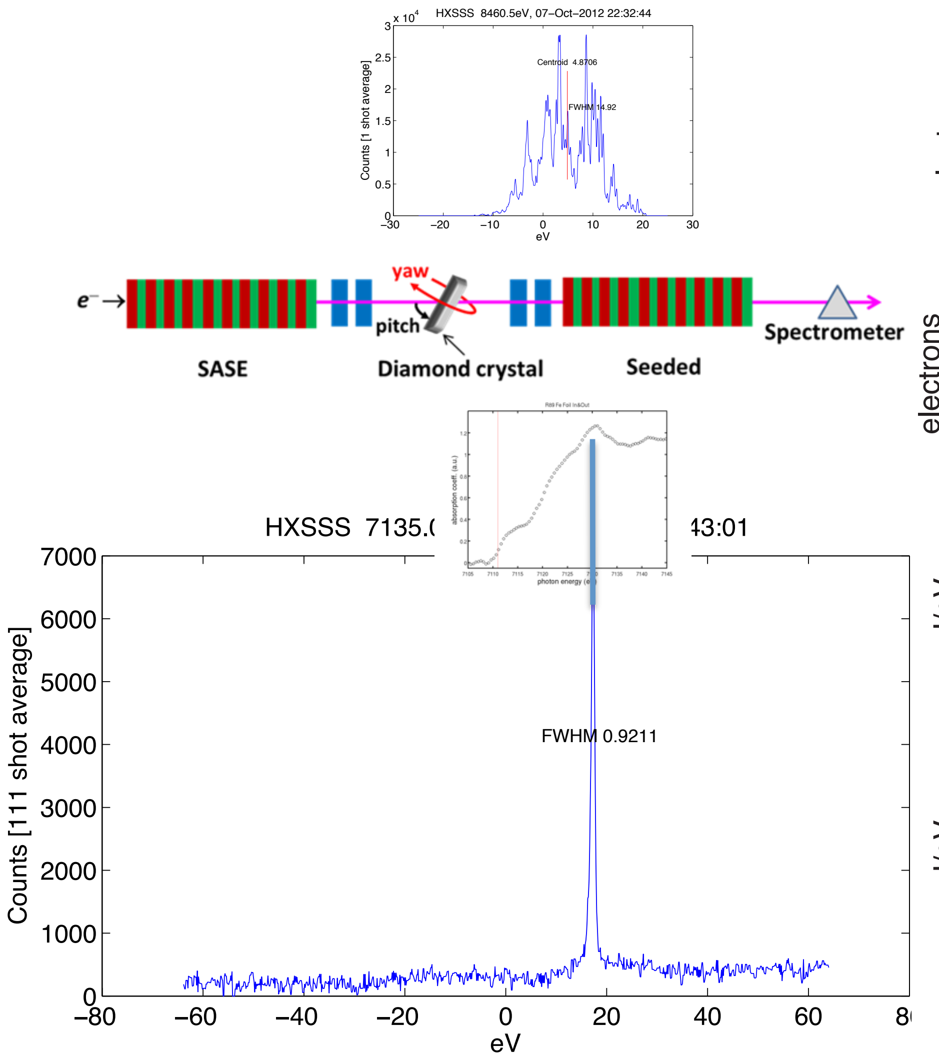


Hattne J et. al. (2014): *Nature Methods* **11**, 545-548.

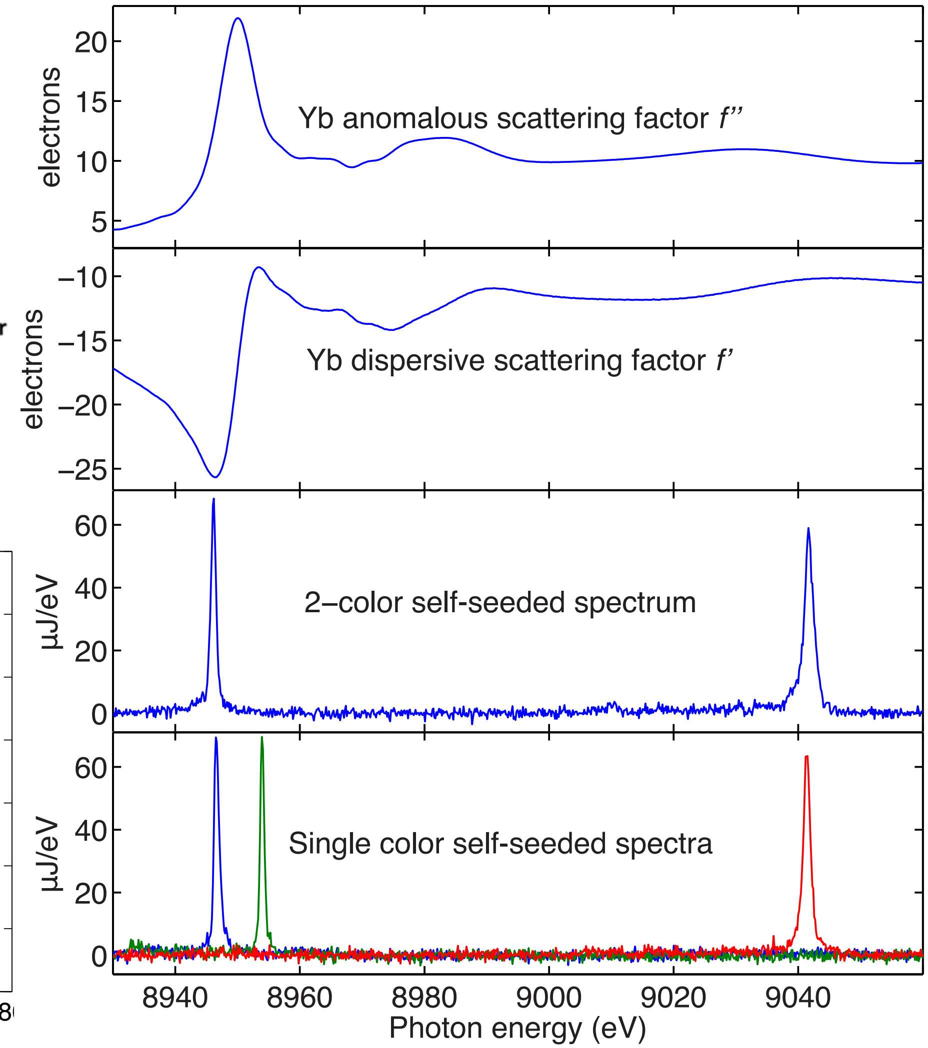
Optimal combination of dispersion + mosaicity to model spots



Single and two-color self seeded XFEL for SAD/MAD *de novo* phasing



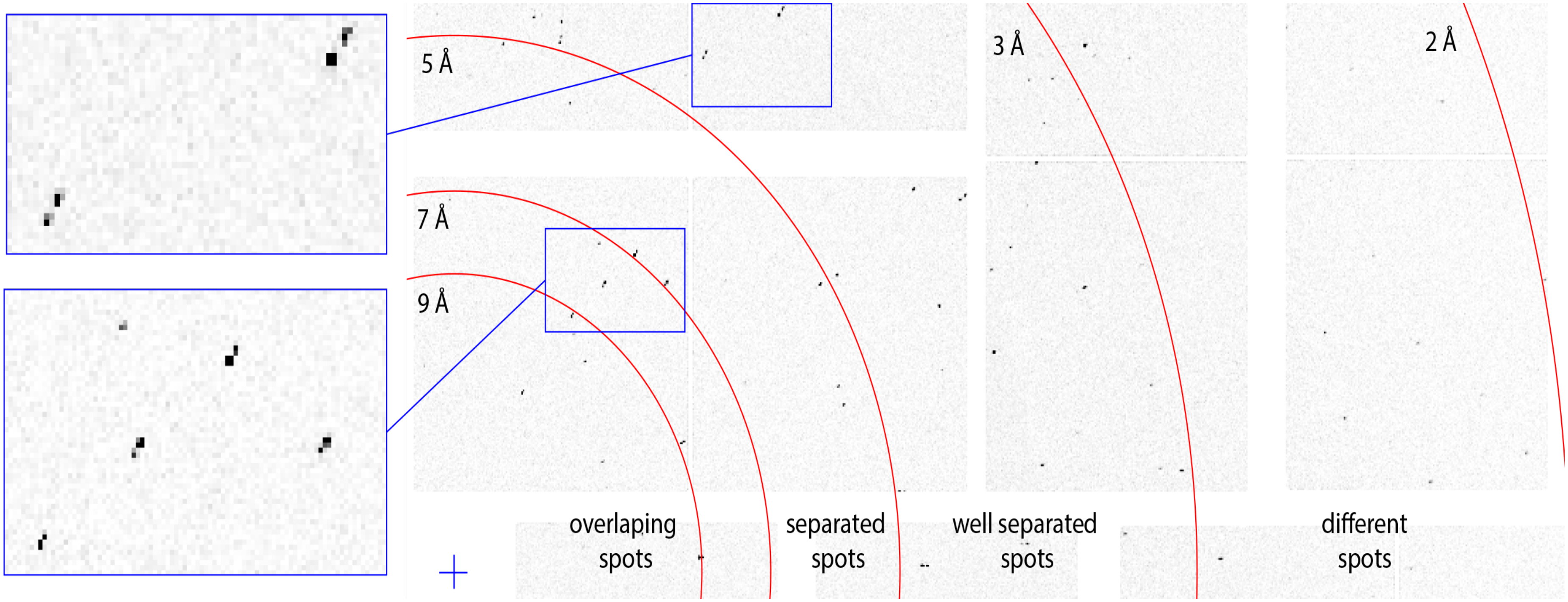
Optimized self seeded XFEL for SAD phasing

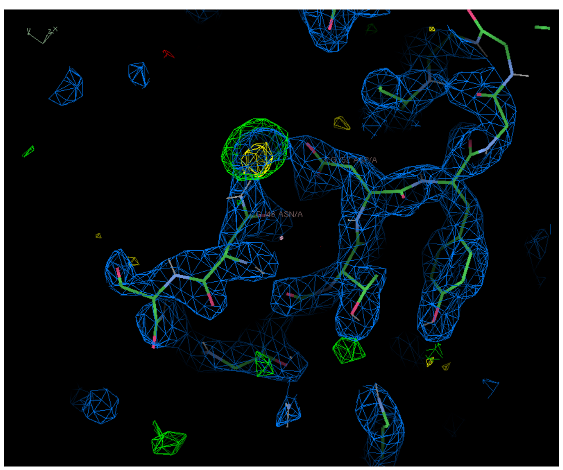


Two color self seeded XFEL with DE=95 eV
For Yb MAD phasing

Soichi Wakatsuki, Bill Weis, Axel Brunger (Stanford)

Two-color spectrum causes doubled Bragg spots





June 2014 preliminary results

Work in progress:

- Muhamed Amin
- Tara Michels-Clark
- Monarin Uervirojnangkoorn
- Ulf Lundstrom

So how good are my data?

Sources of systematic error:

- Variable crystal volume hit by beam
- Variable flux
- Detector metrology
- Inability to fully constrain the orientation model of still shots
- Varying crystal quality & internal disorder
- Stochastic SASE spectrum
- Non-isomorphism (Oliver Zeldin)

Measures of success:

- Bright spots r.m.s. displacement lattice model vs. observation
- Intensity statistics
- Wilson B factor
- Rfactors
- Peak height in the anomalous Fourier map

Acknowledgements

LBNL
Computational Methods
Johan Hattne
Aaron Brewster
Muhamed Amin
Nat Echols
Paul Adams
Peter Zwart
James Holton

Photosystem II
Vittal Yachandra
Junko Yano
Jan Kern

SLAC/PULSE
Mike Bogan
Ray Sierra

LCLS
Spectroscopy
Uwe Bergmann

SSRL
Goniometry Methods
Mike Soltis
Ana Gonzalez
Ashley Deacon
Aina Cohen
Yingssu Tsai
Scott McPhillips

UCLA
BT toxin; amyloid
David Eisenberg
Duilio Cascio
Michael Sawaya
Jose Rodriguez
Luki Goldschmidt

BNL
Sample Delivery
Allen Orville
Christian Roessler

UCLA
BT toxin; amyloid
David Eisenberg
Duilio Cascio
Michael Sawaya
Jose Rodriguez
Luki Goldschmidt

Dowling College
Computational Methods
Herbert Bernstein

DIALS Development
Graeme Winter
Richard Gildea
James Parkhurst
Luis Fuentes-Montero

NERSC / ESNet
David Skinner
Gregory Bell

Diamond Light Source
Polyhedra crystals
David Stuart
Gwyndaf Evans
Jonathan Grimes
Helen Ginn
Daniel Axford

Stanford University
Computational Methods
Axel Brunger
Mona Uervirojnangkoorn
Oliver Zeldin

Phasing Methods
Soichi Wakatsuki
Bill Weis

MPI Heidelberg
Data Processing
Ilme Schlichting
Karol Nass

NIH/NIGMS grants 1R01GM095887 and 1R01GM102520
DOE/Office of Science contract DE-AC02-05CH11231