U.S. DEPARTMENT OF
ENERGY

UNIVERSITY OF
CALIFORNIA

BERKELEY LAB

LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Using DIALS for XFEL data

Aaron Brewster

DIALS 6 Workshop

May 27th, 2014

cctbx.xfel + DIALS

- *cctbx.xfel* overall goals:
  - Still-specific indexing and integration algorithms
  - Massively parallelize reduction of millions of frames
  - Handle the multipanel detectors (CSPAD)
- Project goal: implement *cctbx.xfel* algorithms for indexing and integrating still images in the DIALS framework

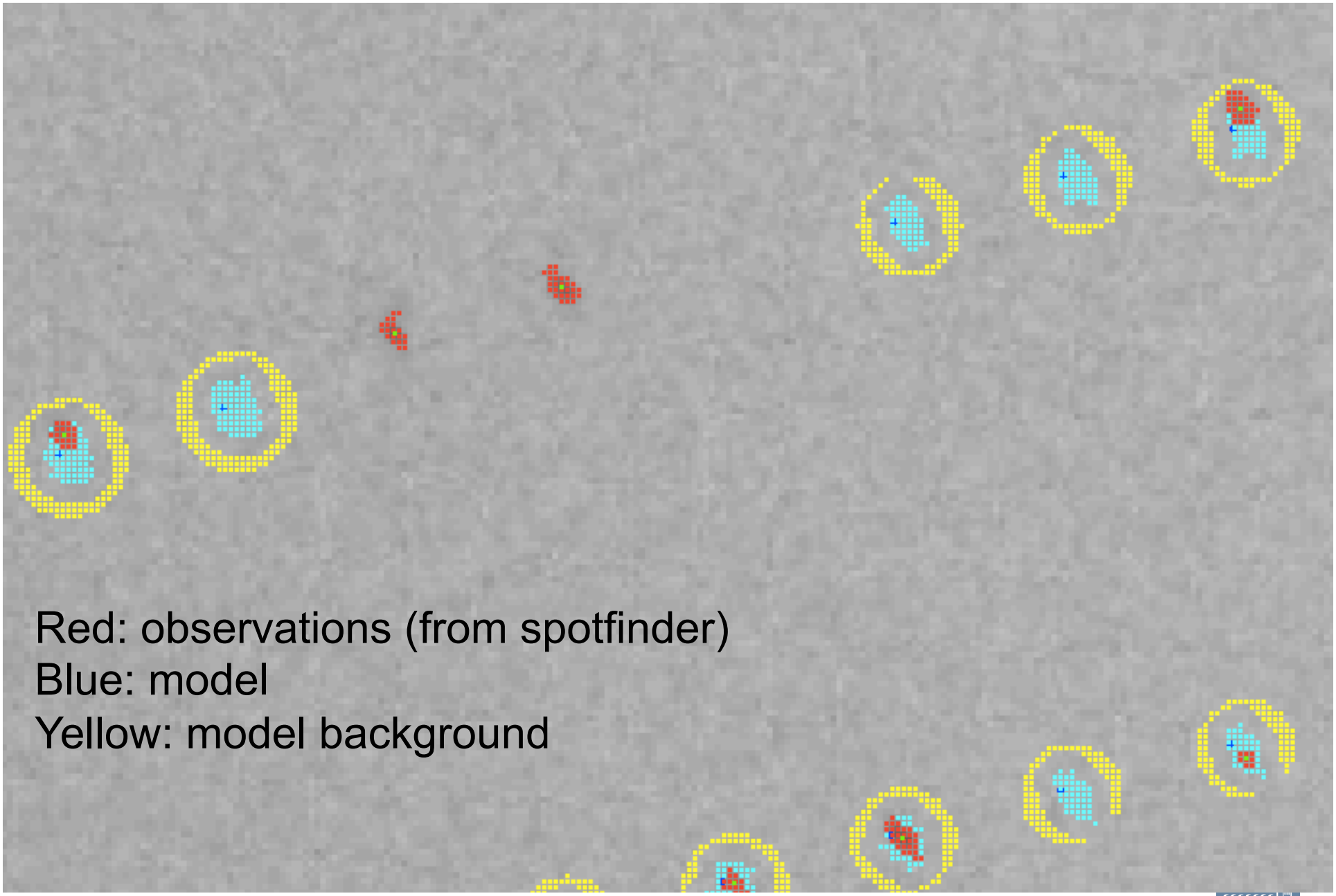
Stills integration: thesis

Accurately determining which weak reflections are in the diffraction condition is one of the most difficult steps in processing still images

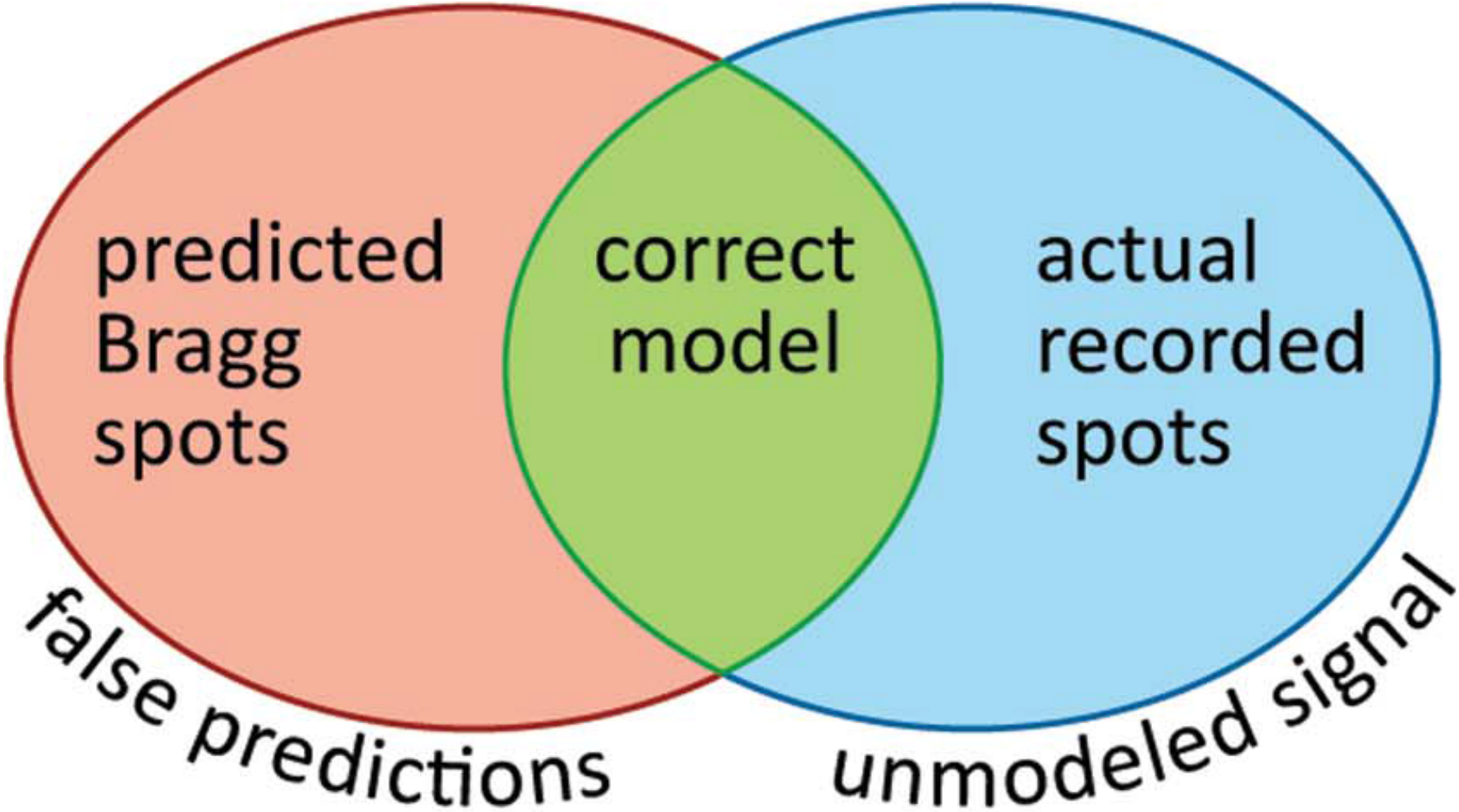
Primary citation:

Sauter NK, Hattne J, Brewster AS, Echols N, Zwart PH, Adams PD (1 Dec 2014): “Improved crystal orientation and physical properties from single-shot XFEL stills.” Acta Crystallogr. D Biol. Crystallogr. 70, 3299-309

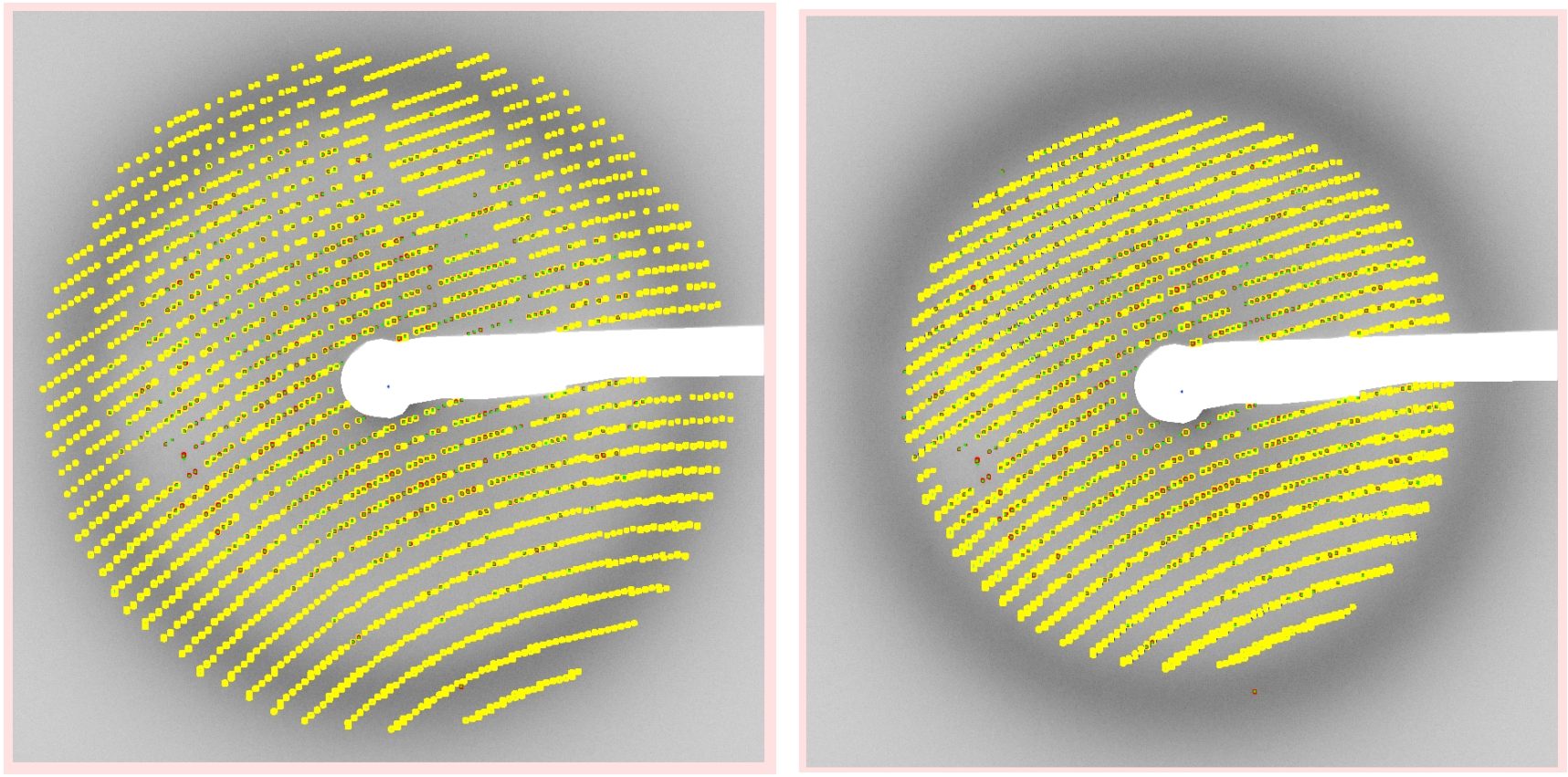
BERKELEY LAB



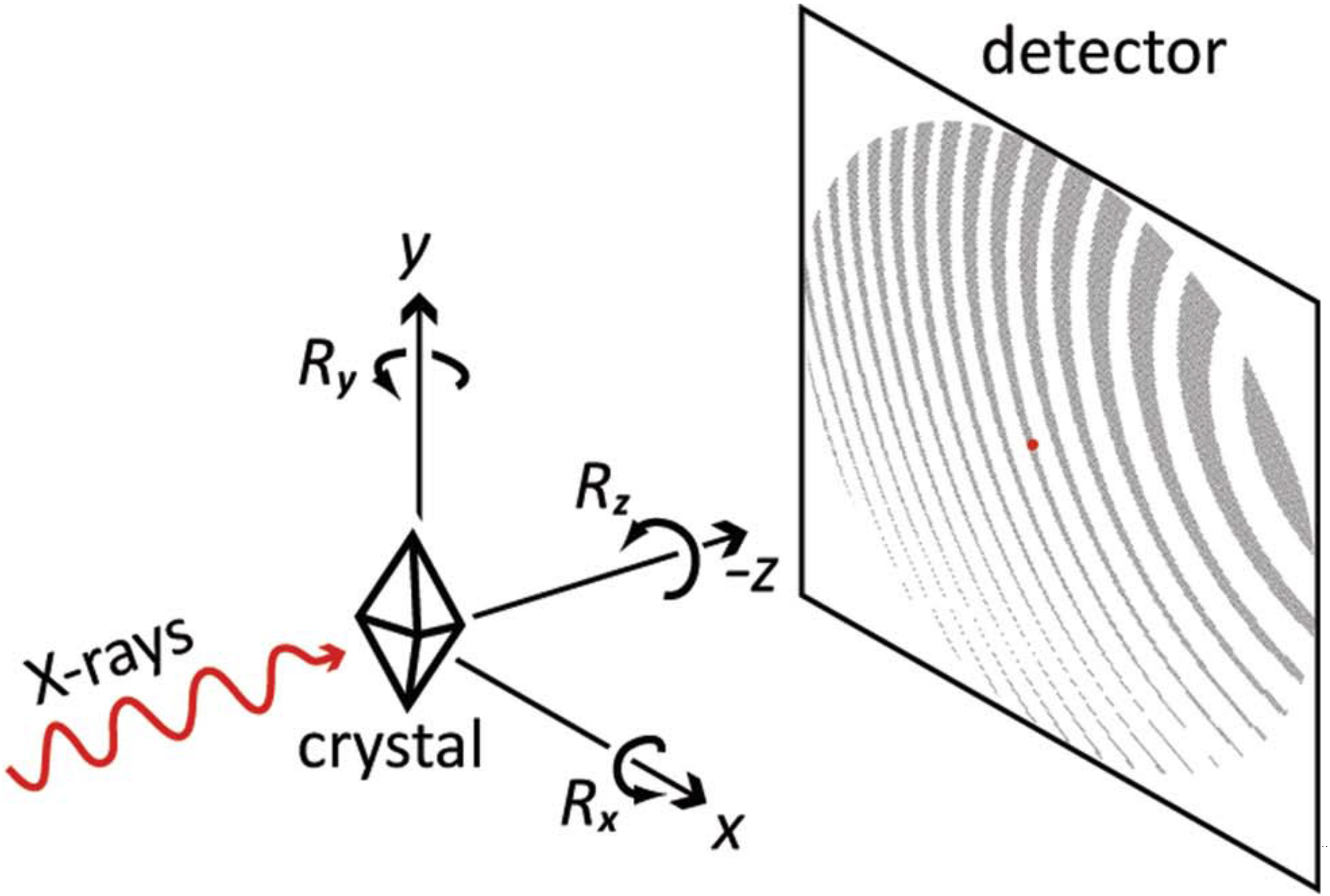
BERKELEY LAB


predicted
Bragg
spots
correct
model
actual
recorded
spots
false predictions
unmodeled signal

Problem case



P1 P222


detector
y
R_y
R_z
-z
X-rays
crystal
R_x
x

Target functions

$$F = \sum_{spots} \left(r_{obs} - r_{\mathrm{mod}\,el} \right)^2$$

$$F = \sum_{spots} \left(\Delta\psi_{calc} \right)^2$$

Three parameter model of reciprocal lattice point shape

Juers DH, Lovelace J, Bellamy HD, Snell EH, Matthews BW and Borgstahl GE (2007). "Changes to crystals of Escherichia coli beta-galactosidase during room-temperature/low-temperature cycling and their relation to cryo-annealing." *Acta Crystallogr. D Biol. Crystallogr.* **63**: 1139-1153.

Nave C (2014). "Matching X-ray beam and detector properties to protein crystals of different perfection." *Journal of synchrotron radiation* **21**.

Sauter NK, Hattne J, Brewster AS, Echols N, Zwart PH, Adams PD (1 Dec 2014): “Improved crystal orientation and physical properties from single-shot XFEL stills.” Acta Crystallogr. D Biol. Crystallogr. 70, 3299-309

Real space

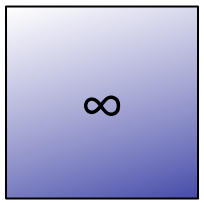
Reciprocal space

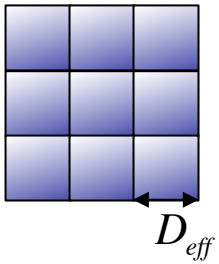
Perfect, infinite crystal

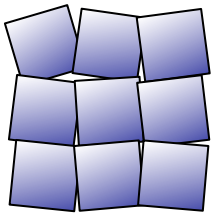
Finite domain size

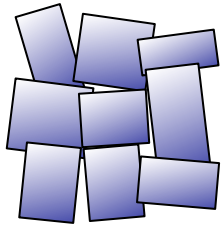
Mosaic spread

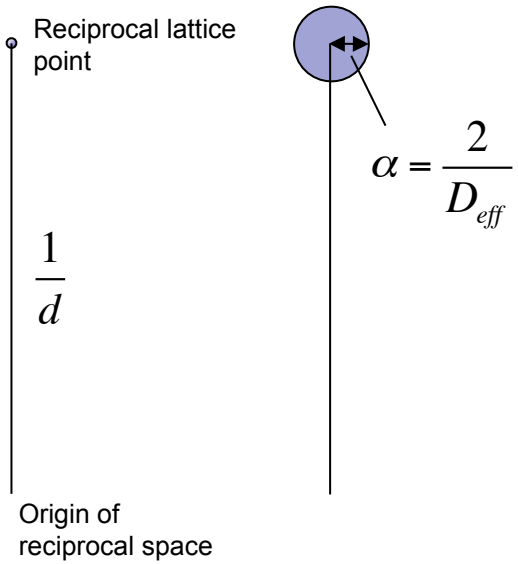
Unit cell variation

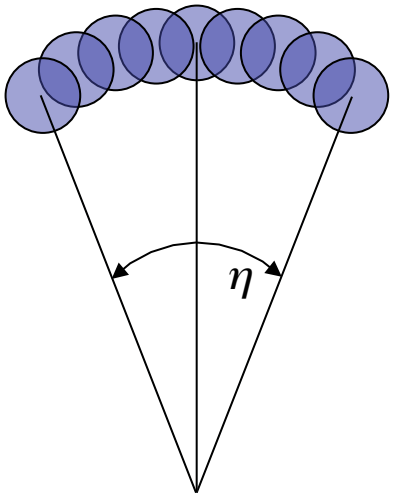


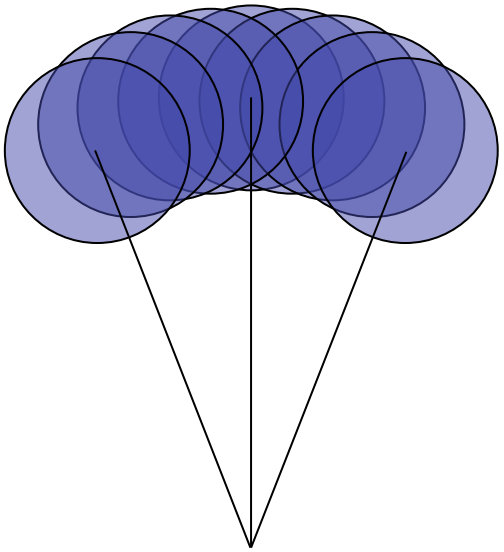








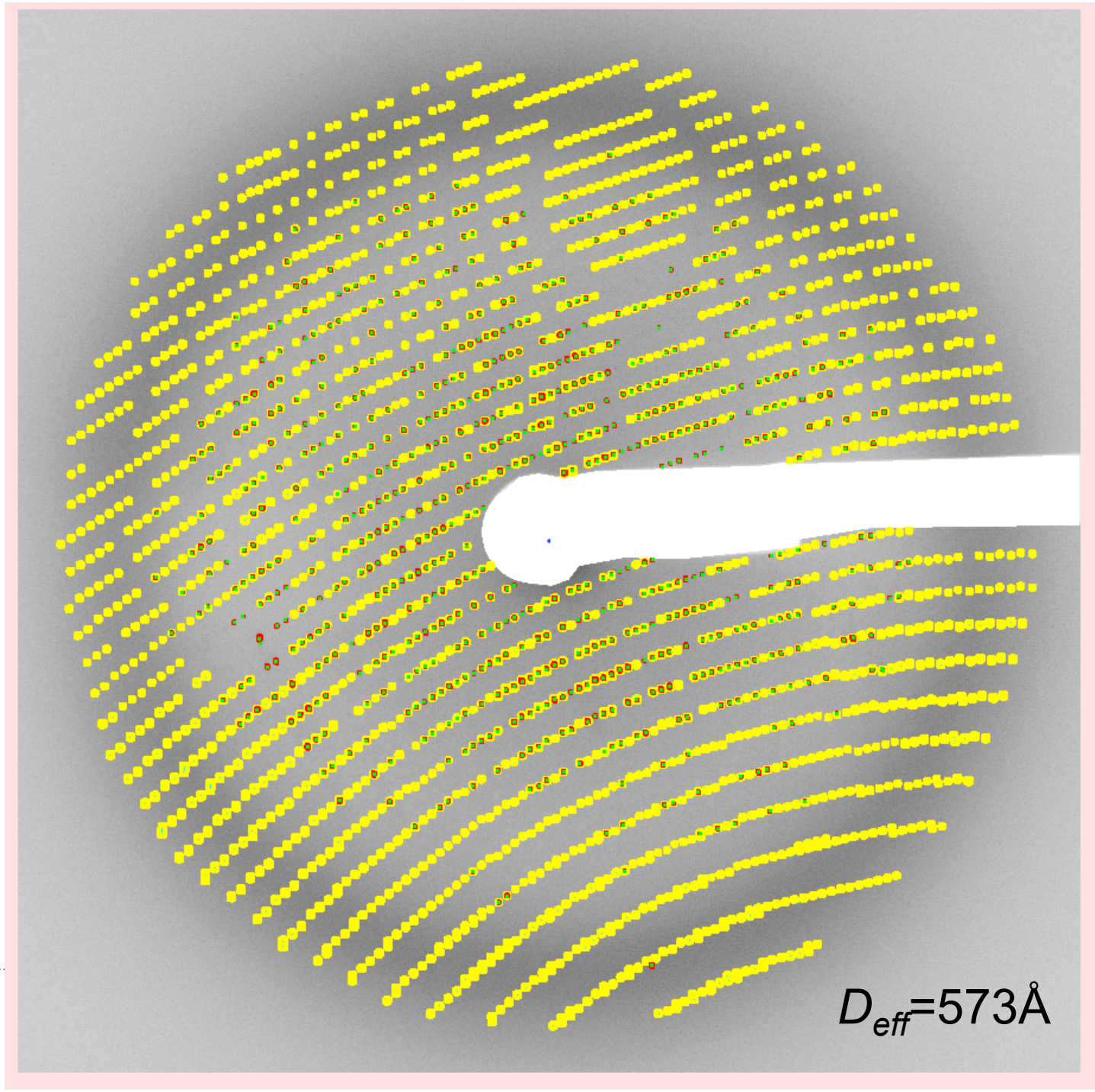




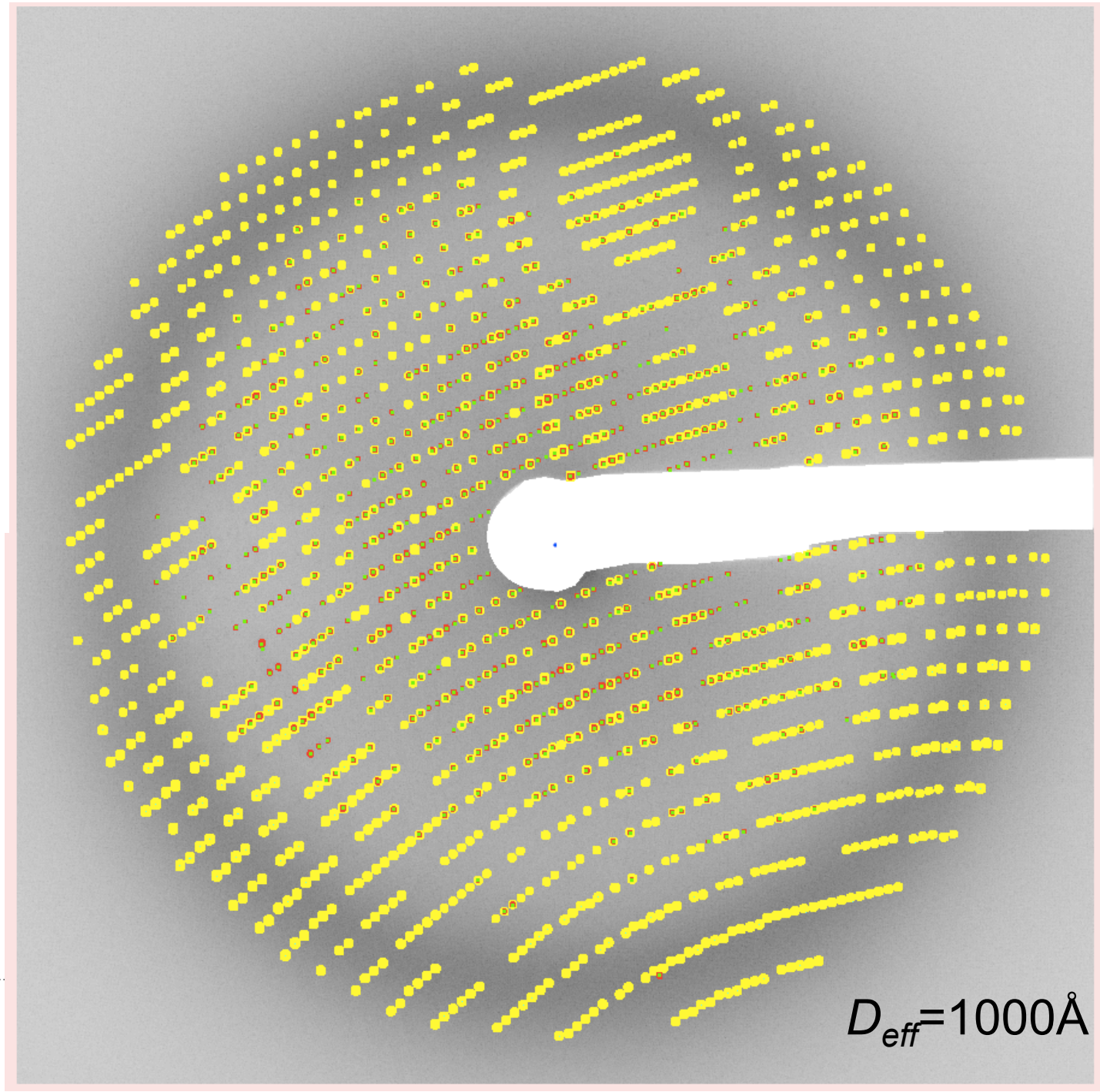
Resolution independent

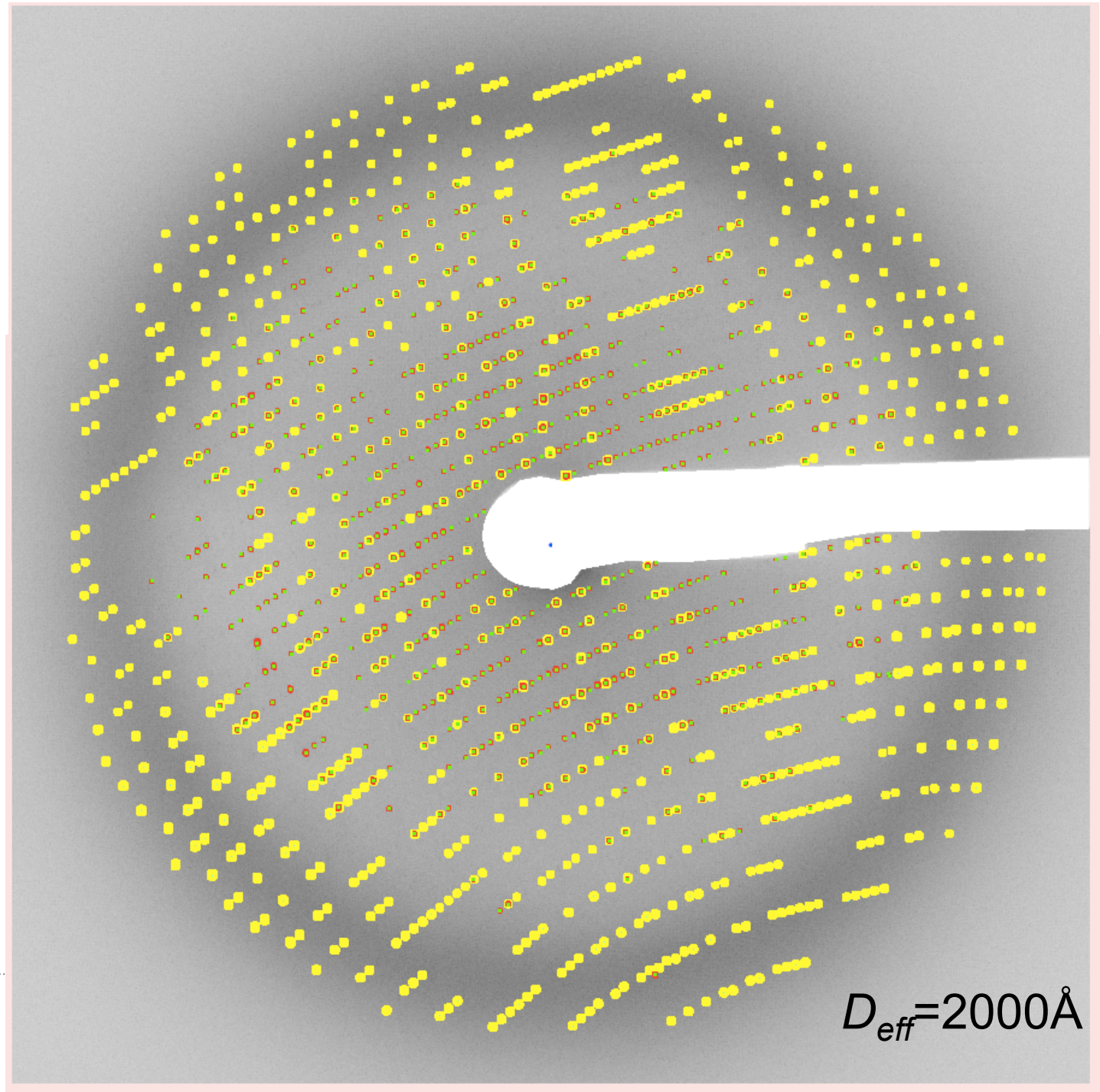
Resolution dependent

BERKELEY LAB

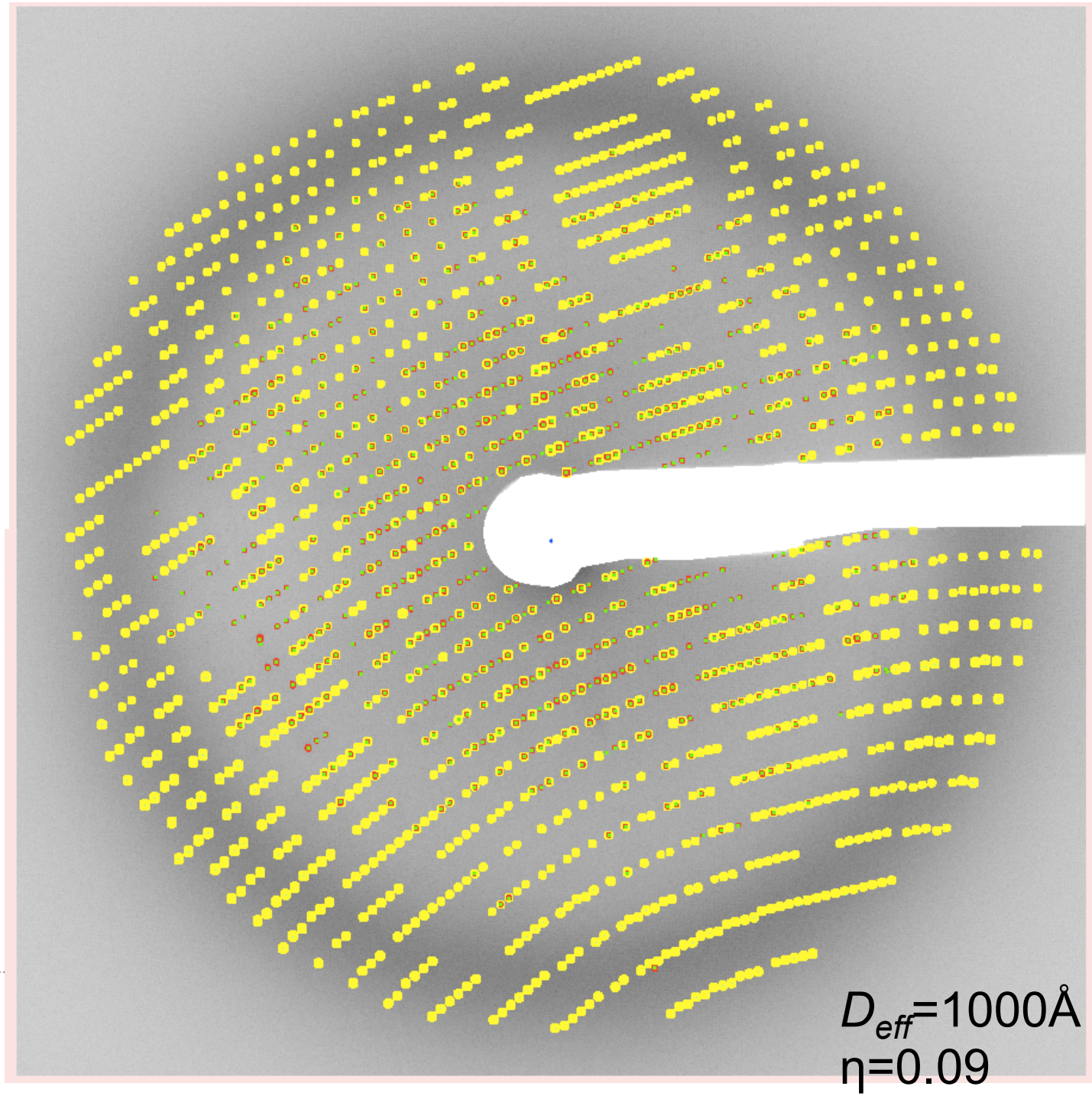


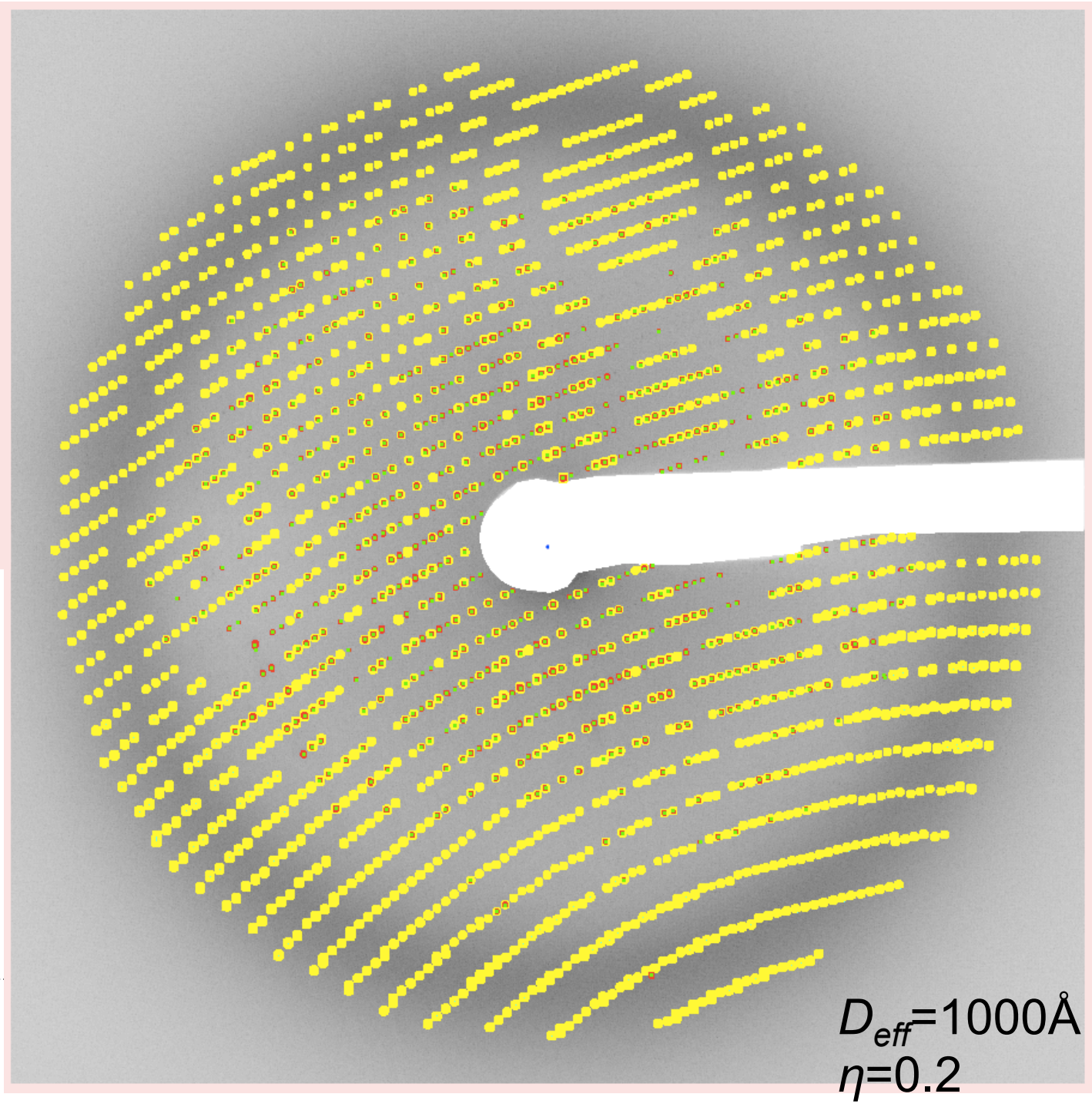
BERKELEY LAB


D_{eff}=1000Å

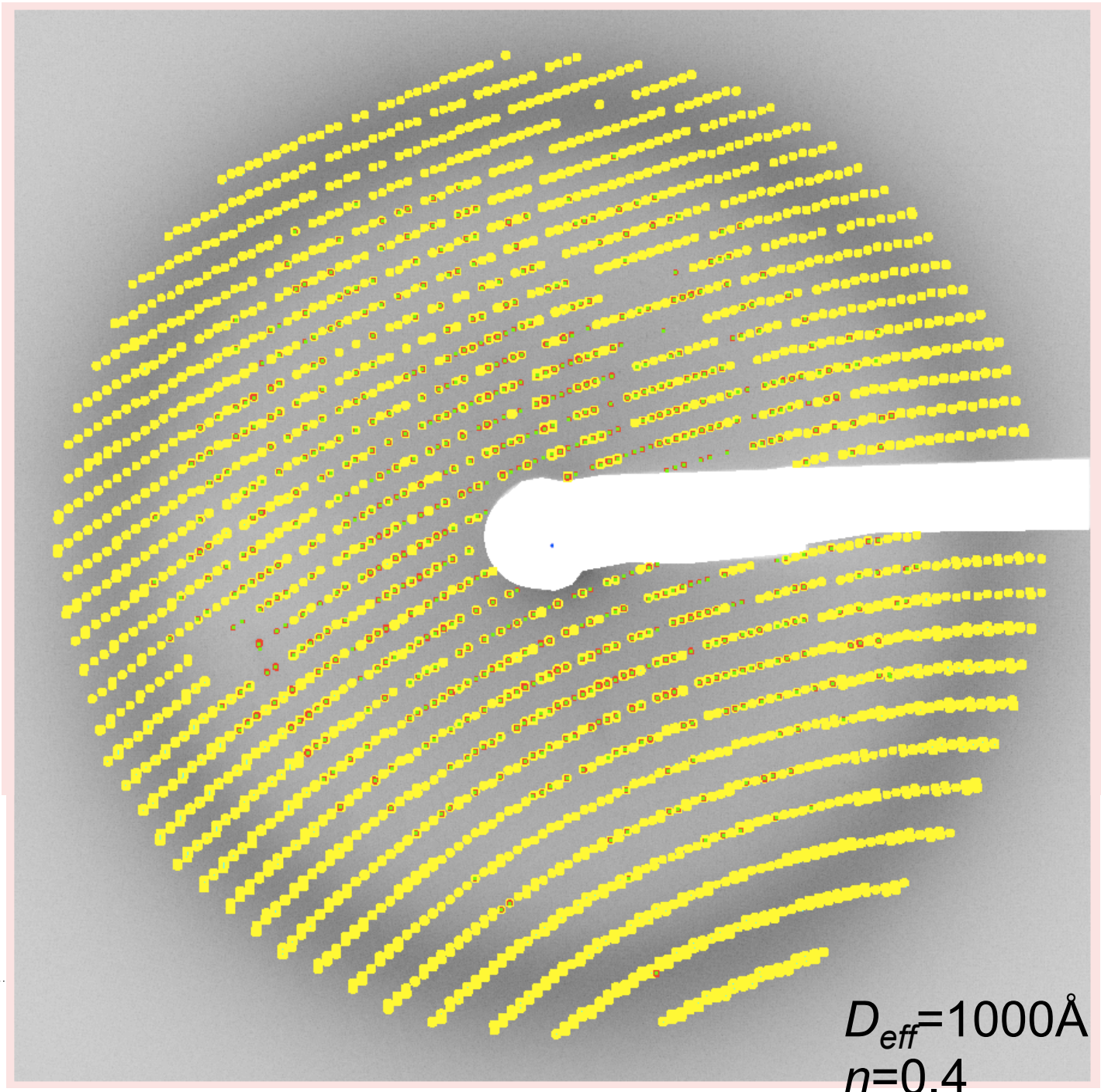


BERKELEY LAB

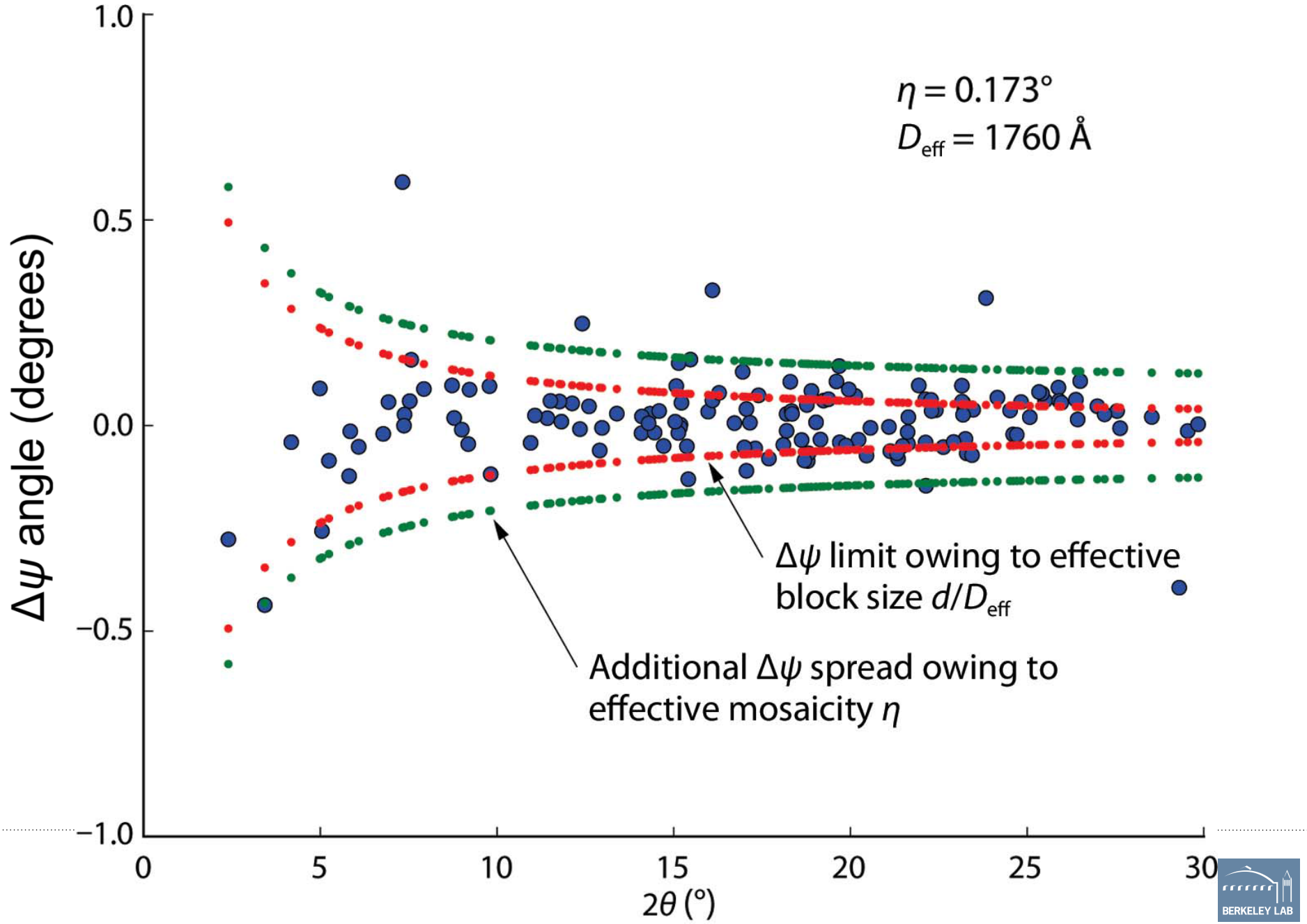

D_{eff}=1000Å
η=0.09



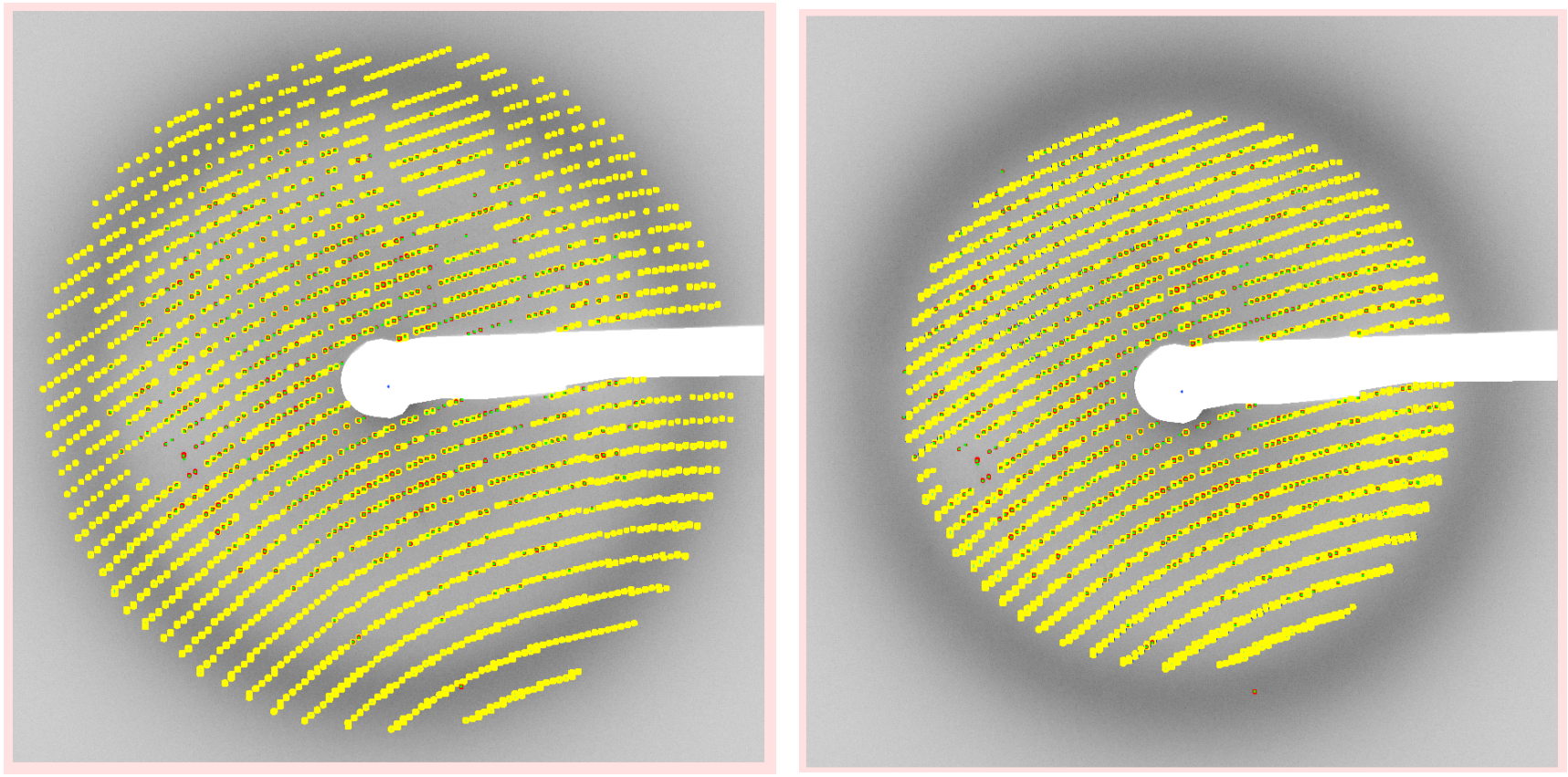
BERKELEY LAB


D_{eff}=1000Å
η=0.4

Effects of D_{eff} and η in reciprocal space

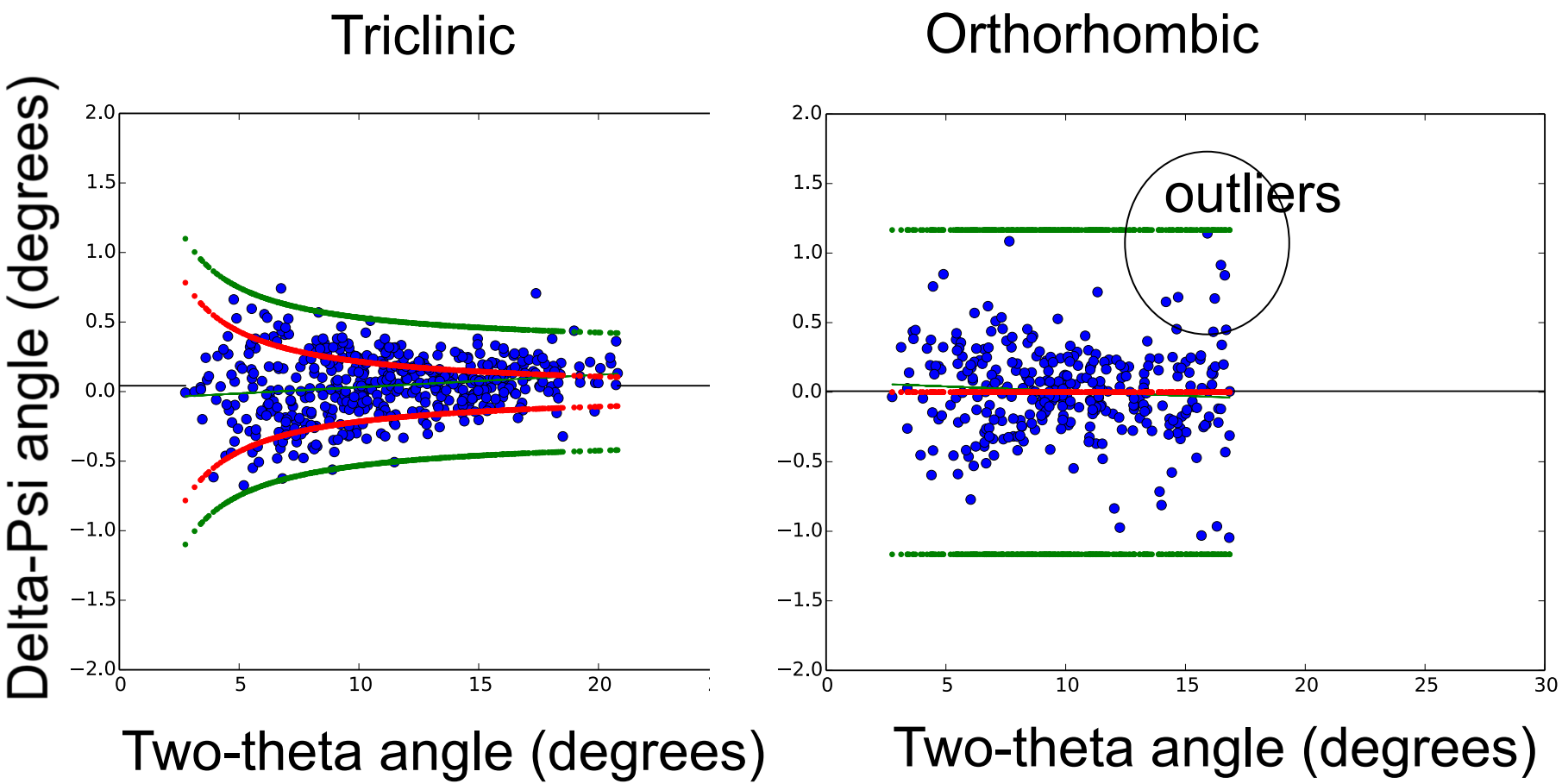

1.0
0.5
0.0
−0.5
−1.0
Δψ angle (degrees)
0
5
10
15
20
25
30
2θ (°)
η = 0.173°
D_eff = 1760 Å
Δψ limit owing to effective block size d/D_eff
Additional Δψ spread owing to effective mosaicity η

Problem case

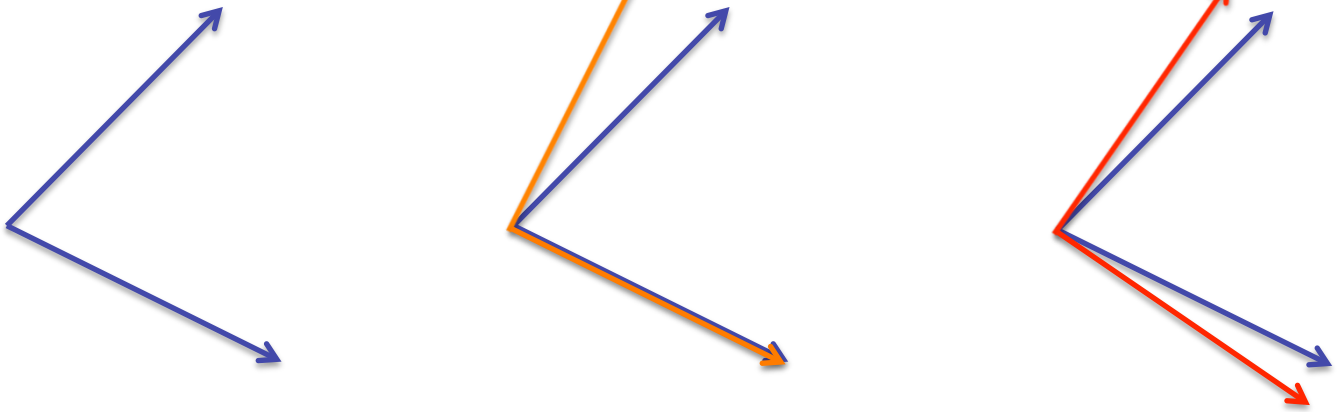


P1 P222

Use the delta-psi vs. 2θ plot as an assay



Change the algorithm for indexing the high-symmetry spots

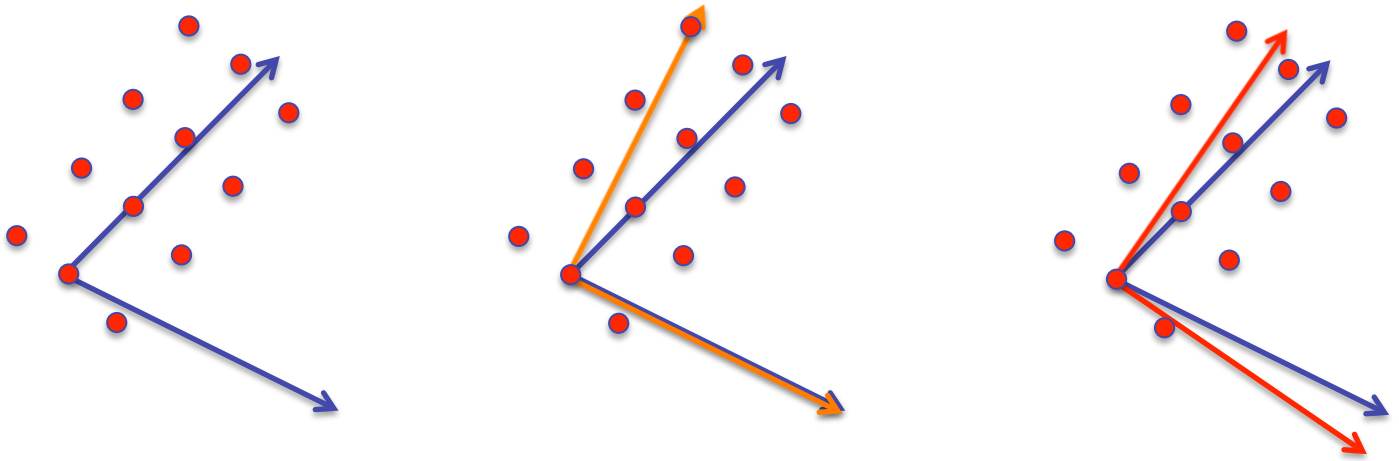


Best triclinic model

Constrained orthorhombic model

Refined orthorhombic model

Change the algorithm for indexing the high-symmetry spots



Best triclinic model

Constrained orthorhombic model

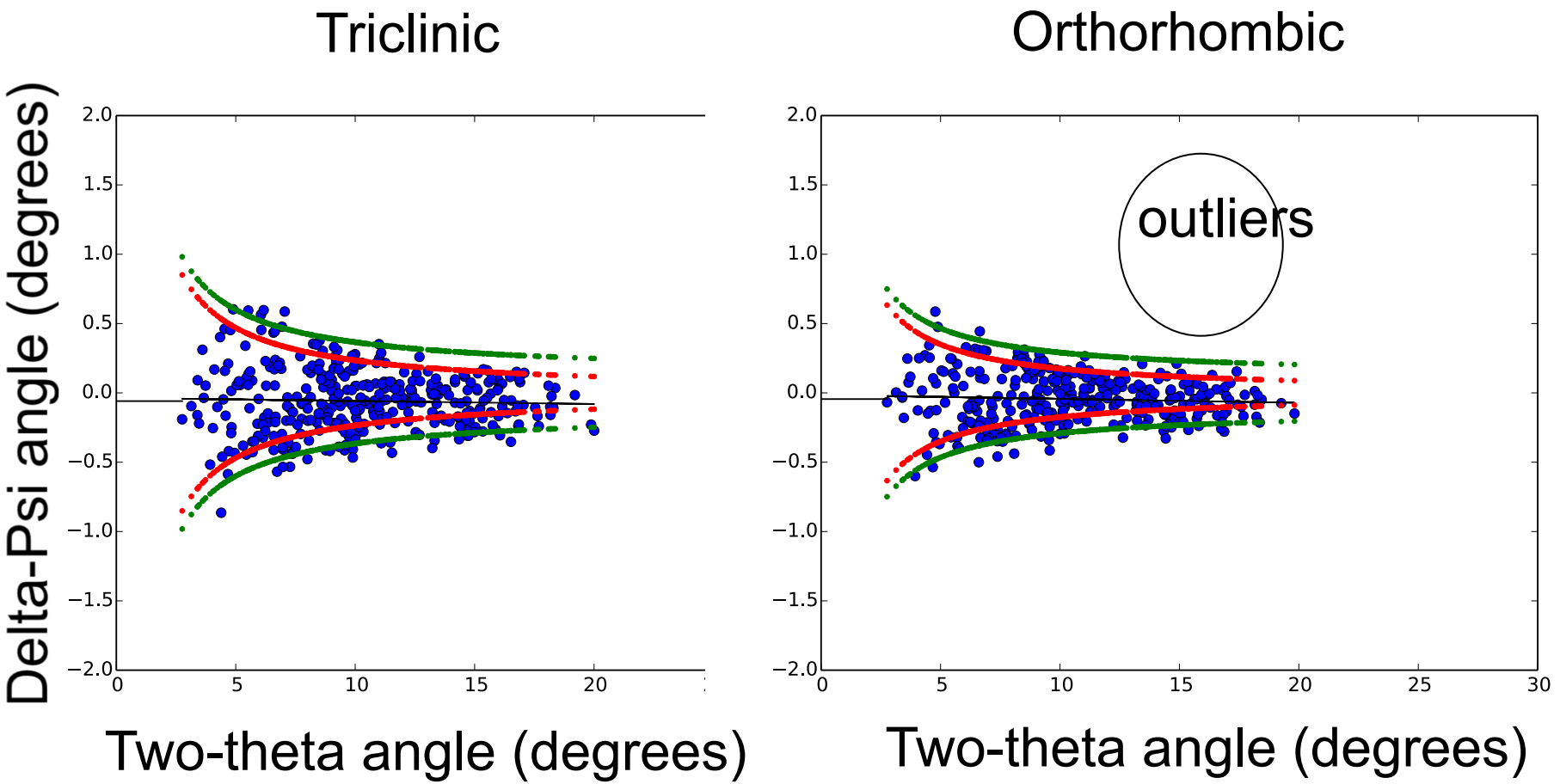
Refined orthorhombic model

Target functions

$$F = \sum_{spots} \left(r_{obs} - r_{\mathrm{mod}\,el} \right)^2 \qquad F = \sum_{spots} \left(\Delta\psi_{calc} \right)^2$$

$$F = \sum_{spots} \left[\left(r_{obs} - r_{\mathrm{mod}\,el} \right)^2 + \left(\Delta\psi_{calc} \right)^2 \right]$$

New refinement methods, fix the indexing problem

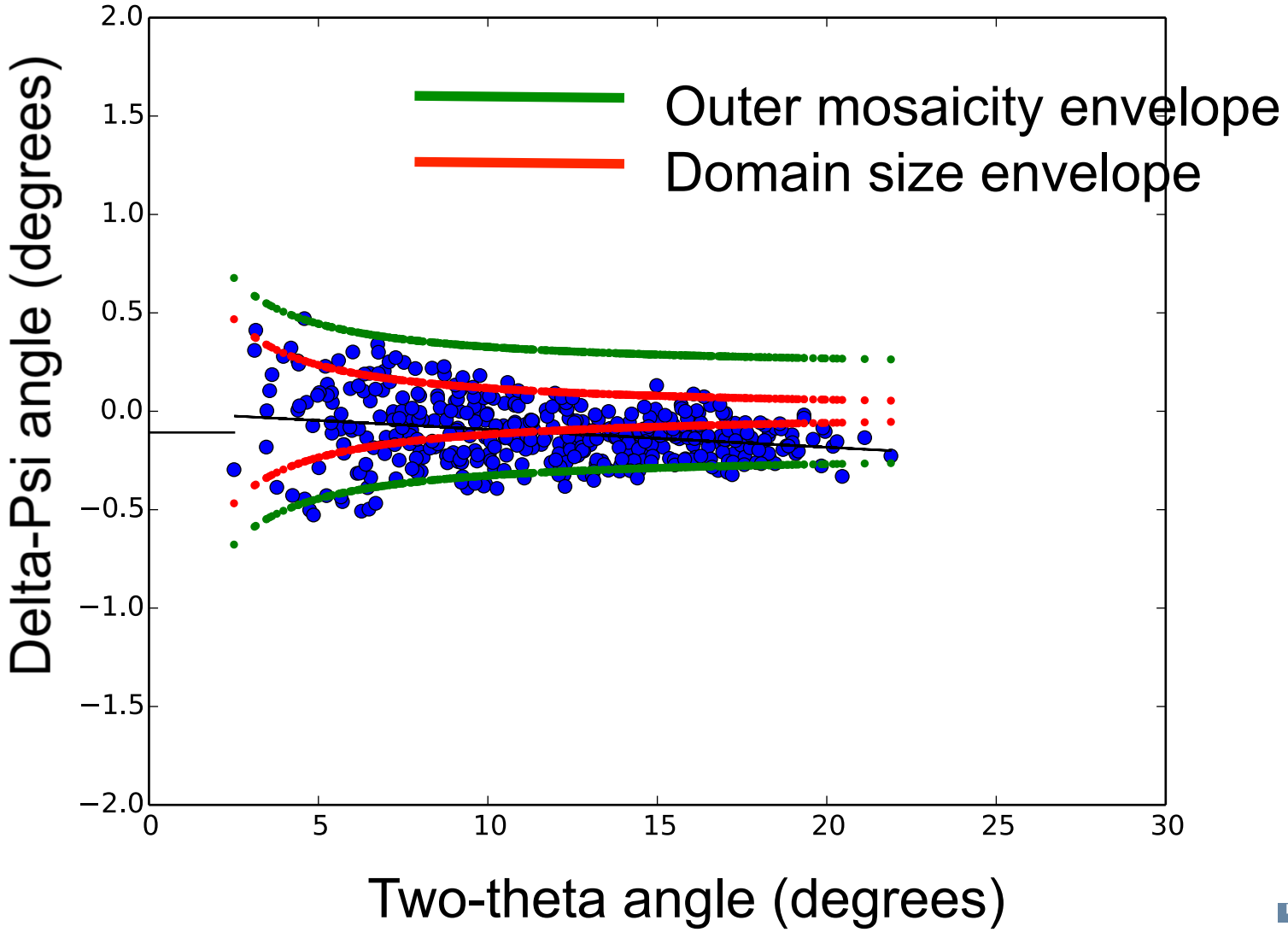


Outer mosaicity envelope

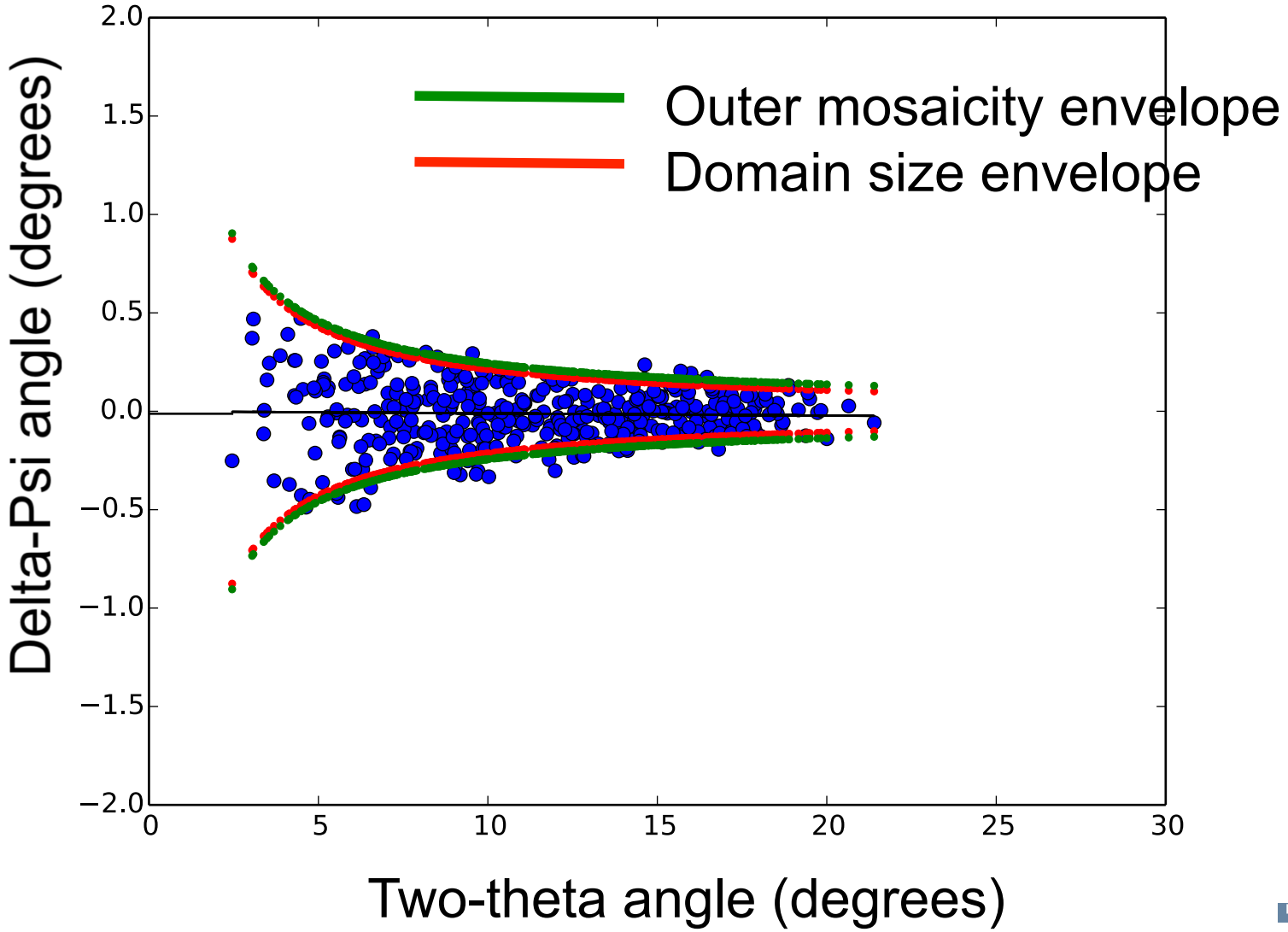
Domain size envelope

BERKELEY LAB

Skewed distribution of delta-Psi from another image

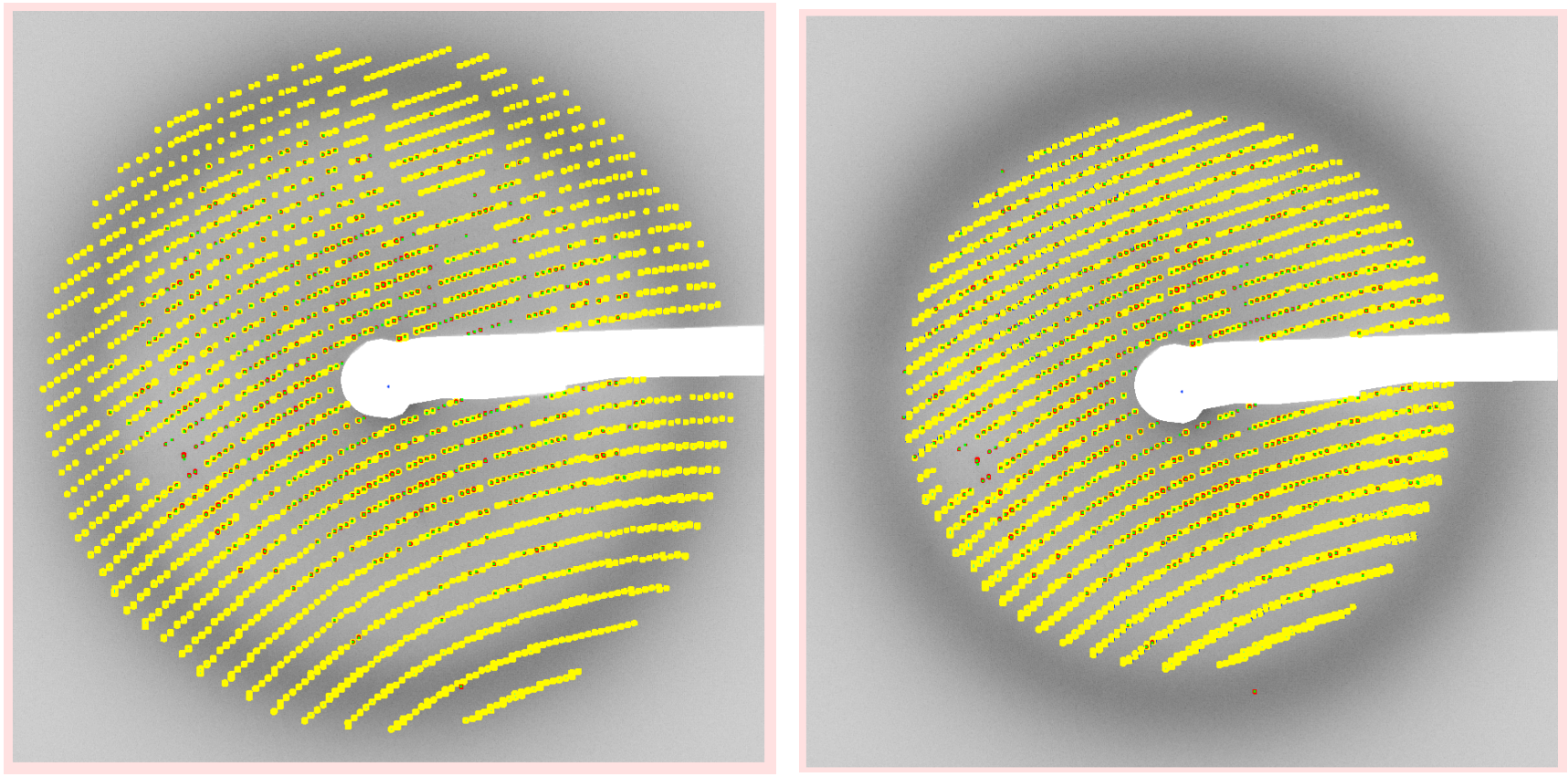


Asymmetrical plot is fixed by allowing the detector distance to refine



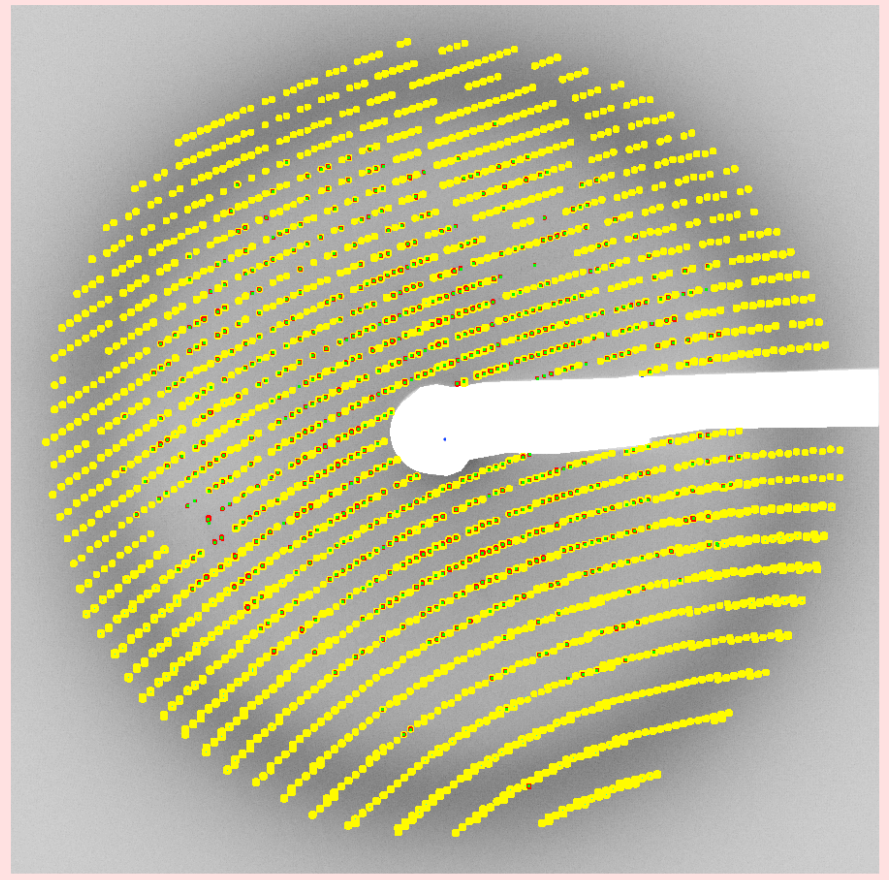
BERKELEY LAB

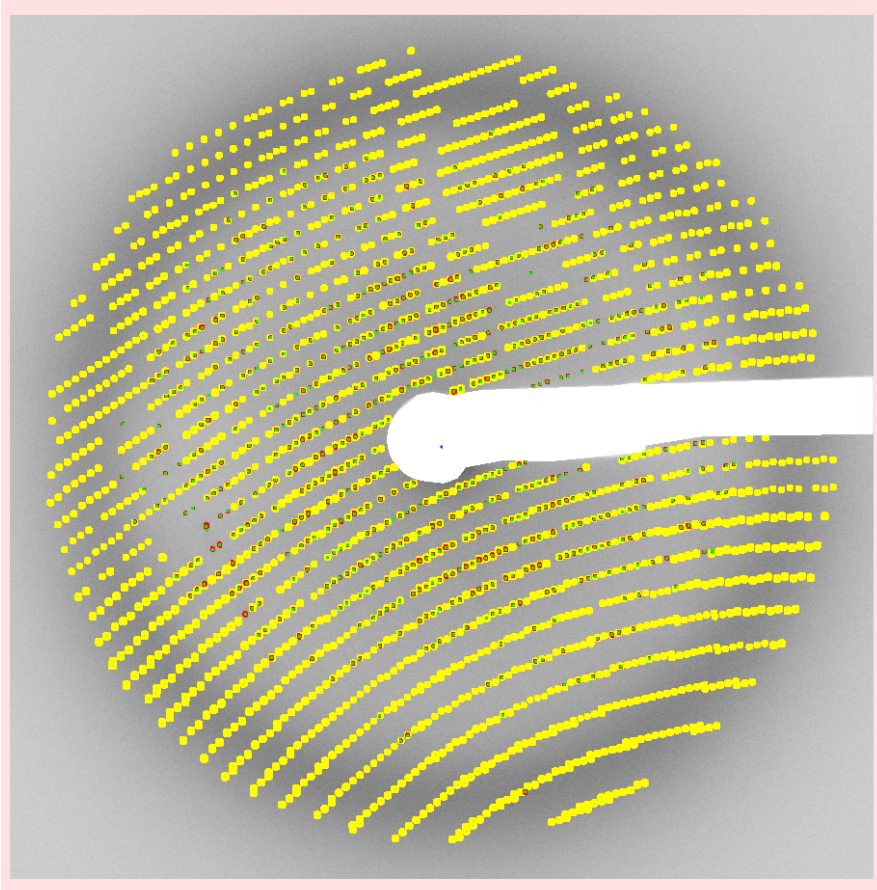
Problem case



P1 P222

Improvements in orthorhombic setting





P1

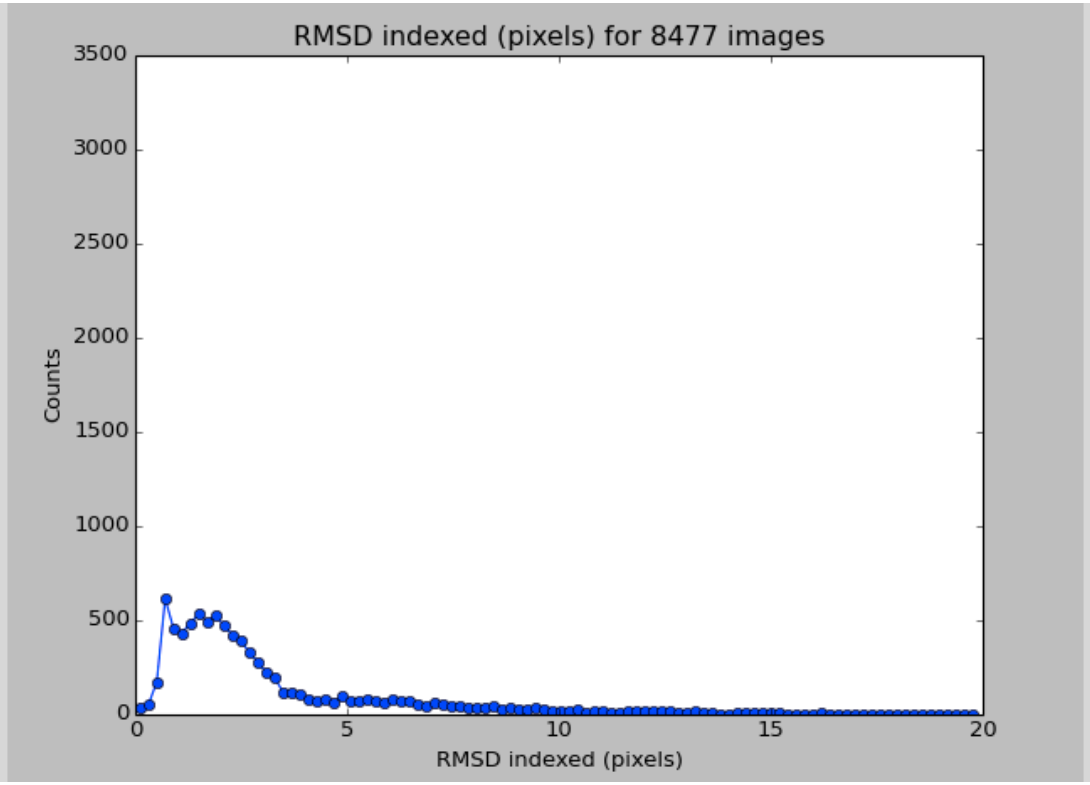
P222

BERKELEY LAB

DIALS refinement

- Refinement of models (rotation images)
  - Target function: x, y, and delta phi RMSD
  - Tukey outlier rejection
- During indexing, each candidate is refined before choosing likely basis vectors

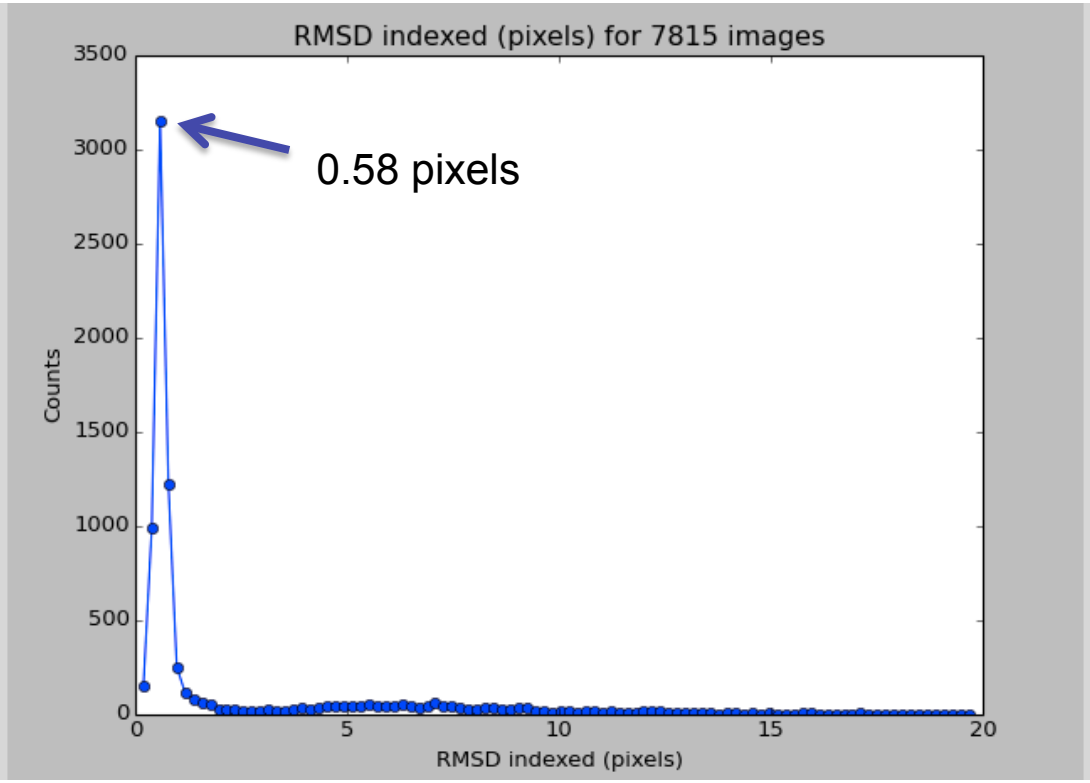
Lysozyme images

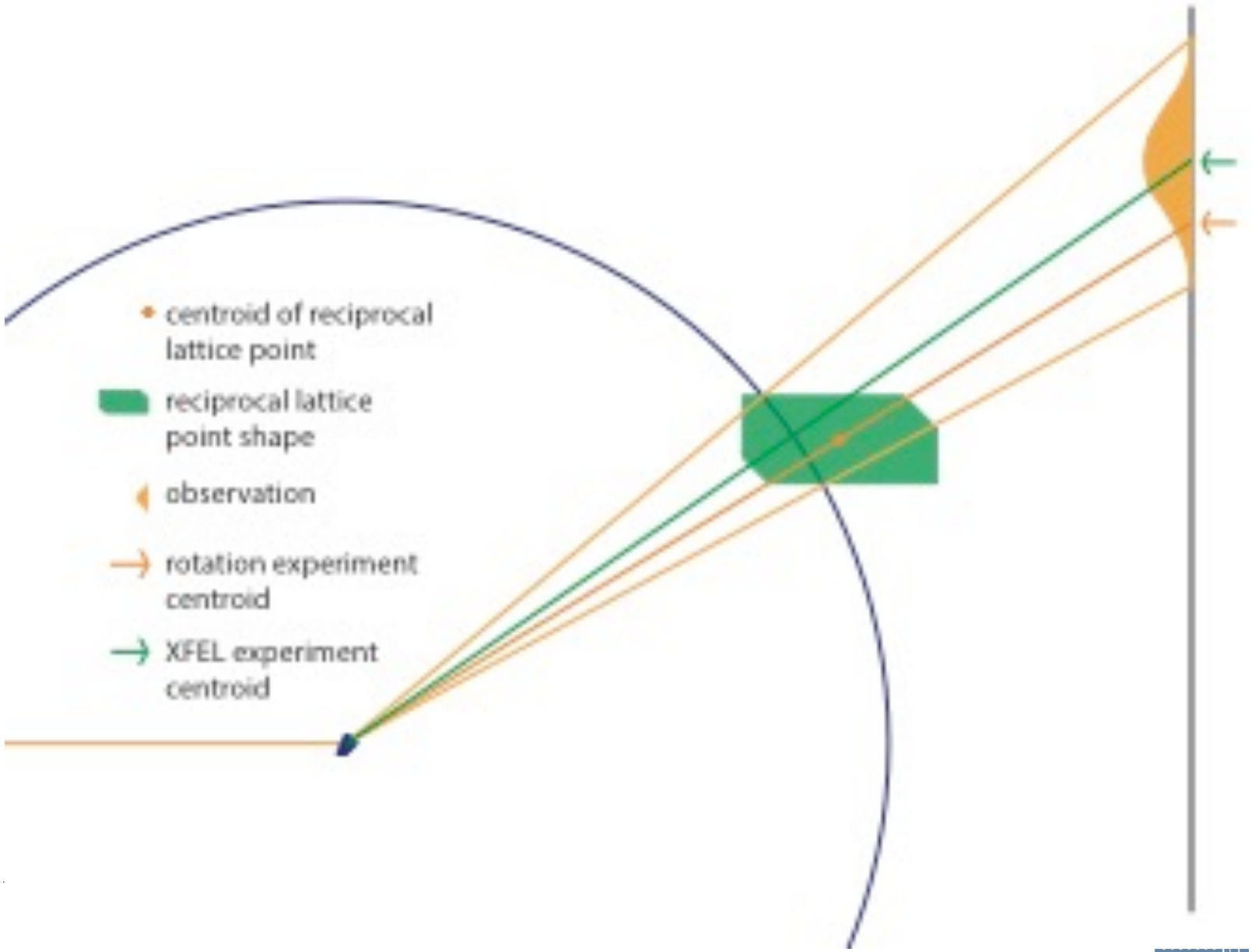


Still indexing/integration in DIALS

- Refinement of models (rotation images)
  - Target function: x, y, and delta phi RMSD
  - Tukey outlier rejection
- Stills refinement:
  - Target function: x, y, and delta psi RMSD
  - Two rounds of refinement with different outlier rejection techniques:
    - Sauter and Poon (2010)
    - Nave parameters (reflections outside of green curve)

Lysozyme images: new refinement for stills




centroid of reciprocal lattice point
reciprocal lattice point shape
observation
rotation experiment centroid
XFEL experiment centroid

Dials index, refine and integrate (fully detailed)

- Index
  - Init: use only strong refls, determine primitive setting of target
  - Discover better experimental model
  - Loop:
    - find_lattices (indexer specific: fft3d, fft1d, real_space_grid_search)
      - » Determine set of candidate vectors
      - » Choose the best candidate
    - Loop over number of cycles requested:
      - » Index reflections, convert to target setting
      - » Refine models
      - » Increase resolution
- Refine:
  - Detector position/orientation, crystal cell, orientation
  - Target function: x, y, delta phi RMSD
- Integrate
  - Generate indices
  - Predict reflections
  - Generate integration masks
  - Integrate

Dials index, refine and integrate

- Index
  - Discover better experimental models
  - Loop:
    - find_lattices (indexer specific: fft3d, fft1d, real_space_grid_search)
      - » Determine set of candidate vectors
      - » Choose the best candidate
    - Loop:
      - » Index reflections, convert to target setting
      - » Refine models
      - » Increase resolution
- Refine models
- Integrate
  - Generate indices
  - Predict reflections
  - Generate integration masks
  - Integrate

2D indexing/integration of stills

- Index
  - Discover better experimental models
  - Loop:
    - find_lattices (indexer specific: ~~fft3d~~, fft1d, real_space_grid_search)
      - » Determine set of candidate vectors
      - » Choose the best candidate
    - Loop:
      - » Index reflections, convert to target setting
      - » Refine models
      - » Increase resolution
- Refine models
- Integrate
  - Generate indices
  - Predict reflections
  - Generate integration masks
  - Integrate

Refinement of crystallographic models

- Refinement for rotation images:
  - Detector position/orientation, crystal cell and crystal orientation
  - Target function: x, y, delta phi RMSD
  - Outlier rejection using Tukey criteria
- Refinement for still images:
  - Detector position/orientation, crystal cell and crystal orientation
  - Target function: x, y, delta psi RMSD
  - Outlier rejection using Sauter and Poon (2010) for x, y RMSDs
  - Further outlier rejection: compute Nave parameters and remove reflections outside of green curve
  - Refine twice

Changes: choose best triclinic orientation matrix:

- Base indexer, for each candidate:
  - Index strong reflections for each candidate
  - Predict reflections
  - Refine models
  - Choose based on positional x, y RMSDs
- Stills indexer, for each candidate:
  - Index strong reflections for each candidate
  - Predict reflections
  - Refine models
  - Choose based on positional x, y RMSDs
  - Throw out result if RMSD or green curve volume is too high

Integration

- Spot prediction:
  - Index generation. Which reflections are in the diffracting condition?
  - $|\Delta\psi_i| \leq \Delta\psi_{model}$ where $\Delta\psi_{model} = \frac{d\alpha}{2} + \frac{\eta}{2}$
  - Ray prediction
- Integration mask: pixel-wise union of nearest 10 bright, indexed reflections

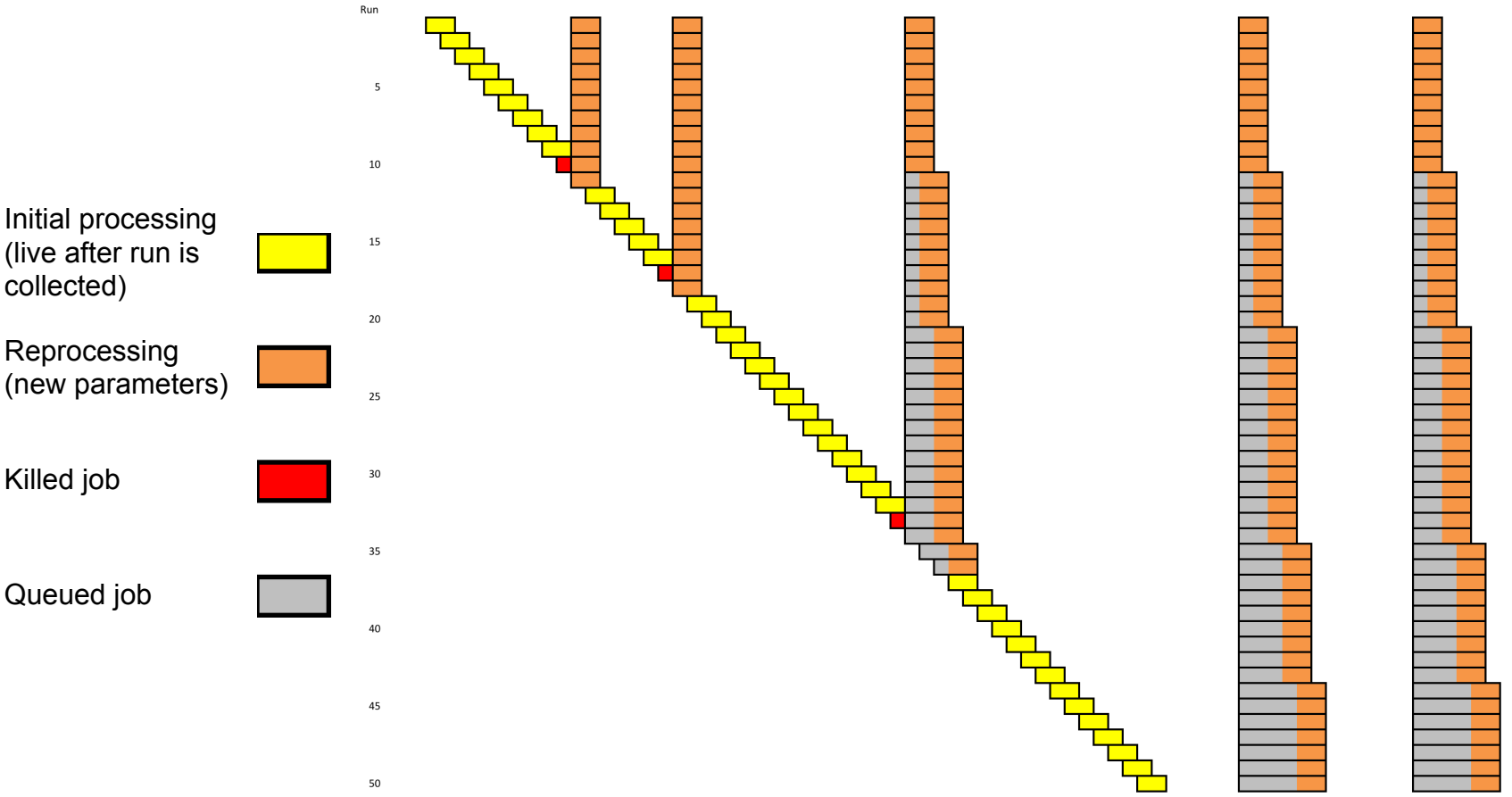
Use case: LH80

- LCLS XFEL experiment early May
- 5 days, night shifts, last day 24 hours
- CSPAD detector, CXI end station, GVDN injection system

XFEL timings

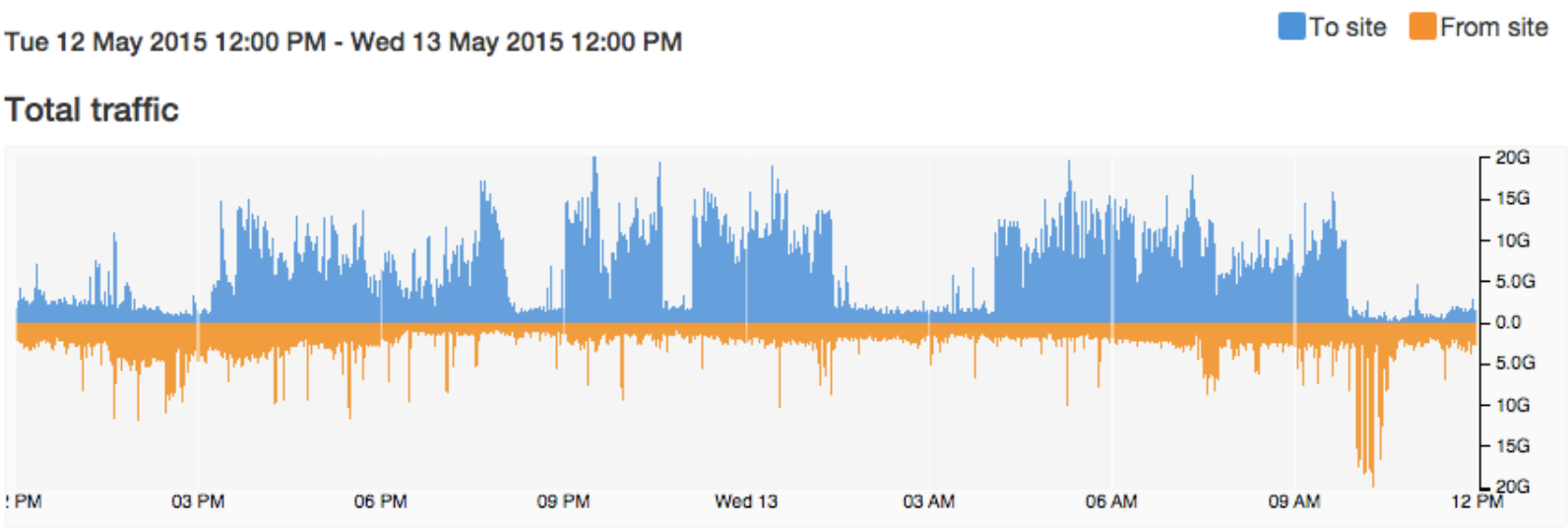
- Typical run: 72000 events, 10 minutes real time
- Hit rate: 10%, indexing rate: 5%
- Time on 192 cores at LCLS: 28 minutes
- Current cores available at LCLS: 480 (public) + 180 (current experiment)

Use case: 5-day XFEL experiment



BERKELEY LAB

Data rates: copy 77.5 GB over ES.NET to NERSC (National Energy Research Scientific Computing Center)



XFEL timings: NERSC

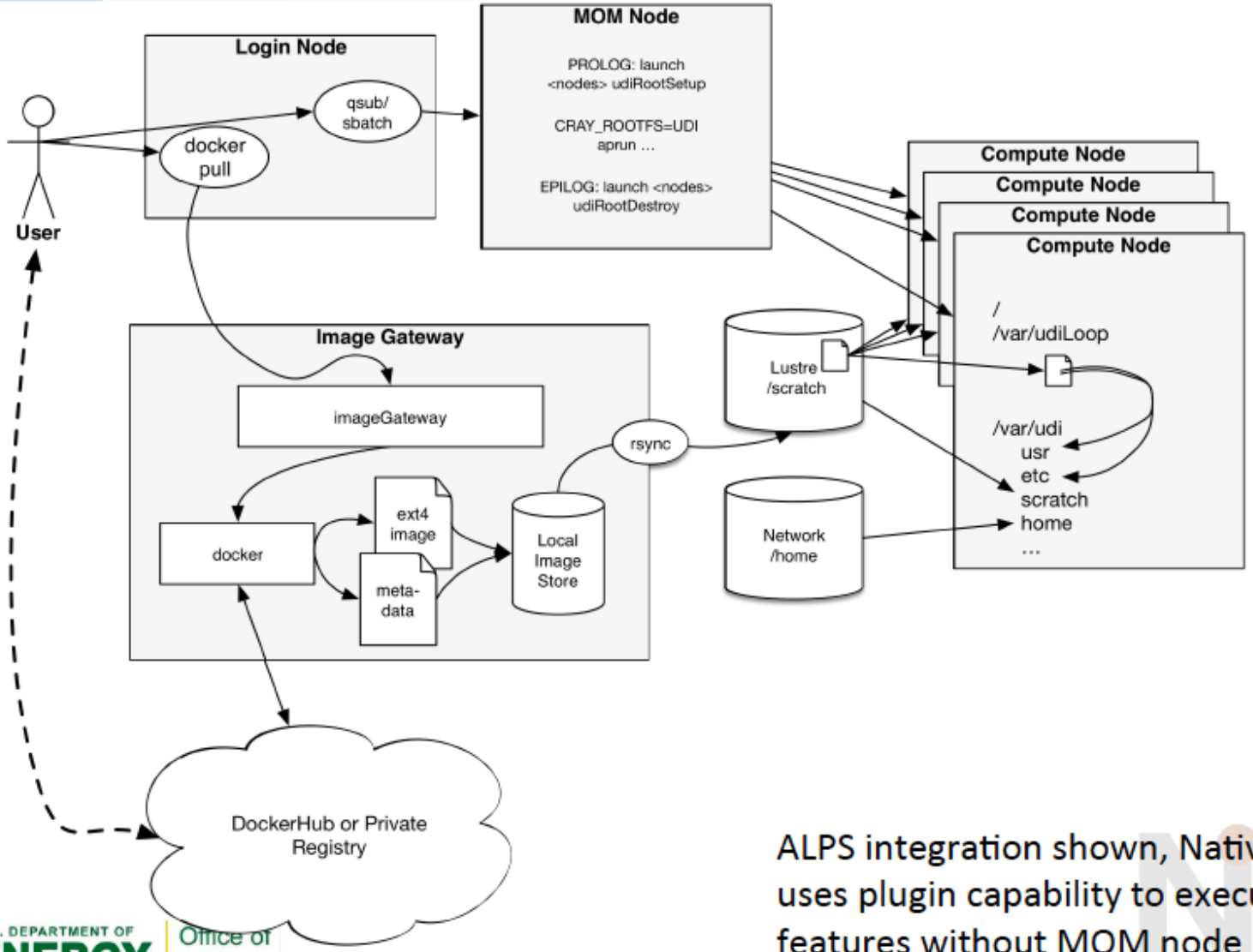
- Edison: Cray supercomputer, 133K cores available
- Got a priority boost during the experiment
- One run, 240 cores: 12 minutes, most of which was startup time.
- 24 runs submitted at once, 240 cores each. Startup time for each: 40 minutes, almost 5K computer hours.
- Culprit: python modules on a GPFS file system
- Solution: Docker

Slide with image on luster, showing metadata copy to computer node memory.

Goal: reprocess entire experiment using 20000 cores.

Goal: upcoming experiment, process live at NERSC.

Prototype Implementation: "Shifter"



ALPS integration shown, Native Slurm uses plugin capability to execute features without MOM node

Slide from Doug Jacobsen, NERSC

Nersc: Docker

Mention AWS?

Putting it all together

- Challenges for XFEL data reduction:
  - Multi-panel detectors (CSPAD)
  - Huge amounts of data
  - Still images
- Solutions: DIALS implementations of *cctbx.xfel*
  - CSPAD CBF
  - Multiprocessing at NERSC (next test: 20000 cores)
  - New dials indexer/integrator specifically for stills

Acknowledgements

Berkeley National Lab
Nicholas Sauter
Muhamed Amin
Tara Michels-Clark
Iris Young
Nat Echols
Paul Adams
Peter Zwart
Vittal Yachandra
Junko Yano
Jan Kern
James Holton

Janelia Farm
Johan Hattne

LCLS
Uwe Bergmann
Alberto Lutman
…and many others

Diamond Light Source
David Stuart
Gwyndaf Evans
Graeme Winter
Jonathan Grimes
Richard Gildea
James Parkhurst
Luis Fuentes-Montero

CCP4
David Waterman

UCLA
David Eisenberg
Duilio Cascio
Michael Sawaya
Jose Rodriguez
Luki Goldschmidt

IBS
Jacques-Philippe Colletier

Stanford University
Axel Brunger
Bill Weis
Mona Uervirojnangkoorn
Artem Lyubimov
Oliver Zeldin

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

BNL
Allen Orville

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

BERKELEY LAB