BERKELEY LAB
Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY

UNIVERSITY OF CALIFORNIA

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

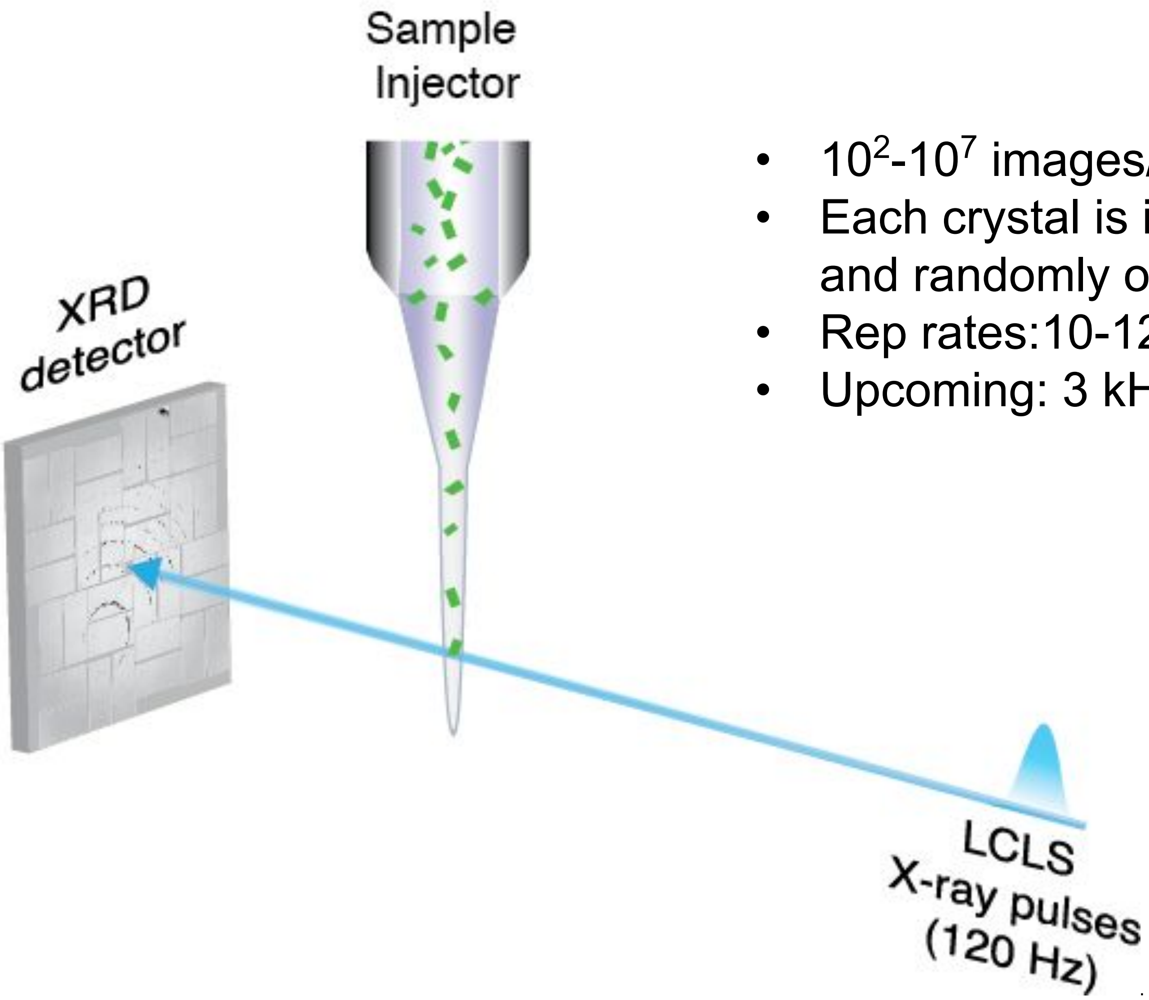
U.S. DEPARTMENT OF ENERGY

The *cctbx.xfel* GUI: processing serial crystallographic data at massive scales

Aaron S. Brewster

March 17th, 2021

Serial crystallography



- 10^2-10^7 images/dataset
- Each crystal is independent and randomly oriented
- Rep rates:10-120 Hz
- Upcoming: 3 kHz

How much data do I need?

- What is my hitrate?
  - Solvent rate
  - Crystal rate
  - Indexing rate
  - High quality rate
- What is my completeness?
- Can I stop collecting data?

Beamtime is scarce! Fast feedback is critical!

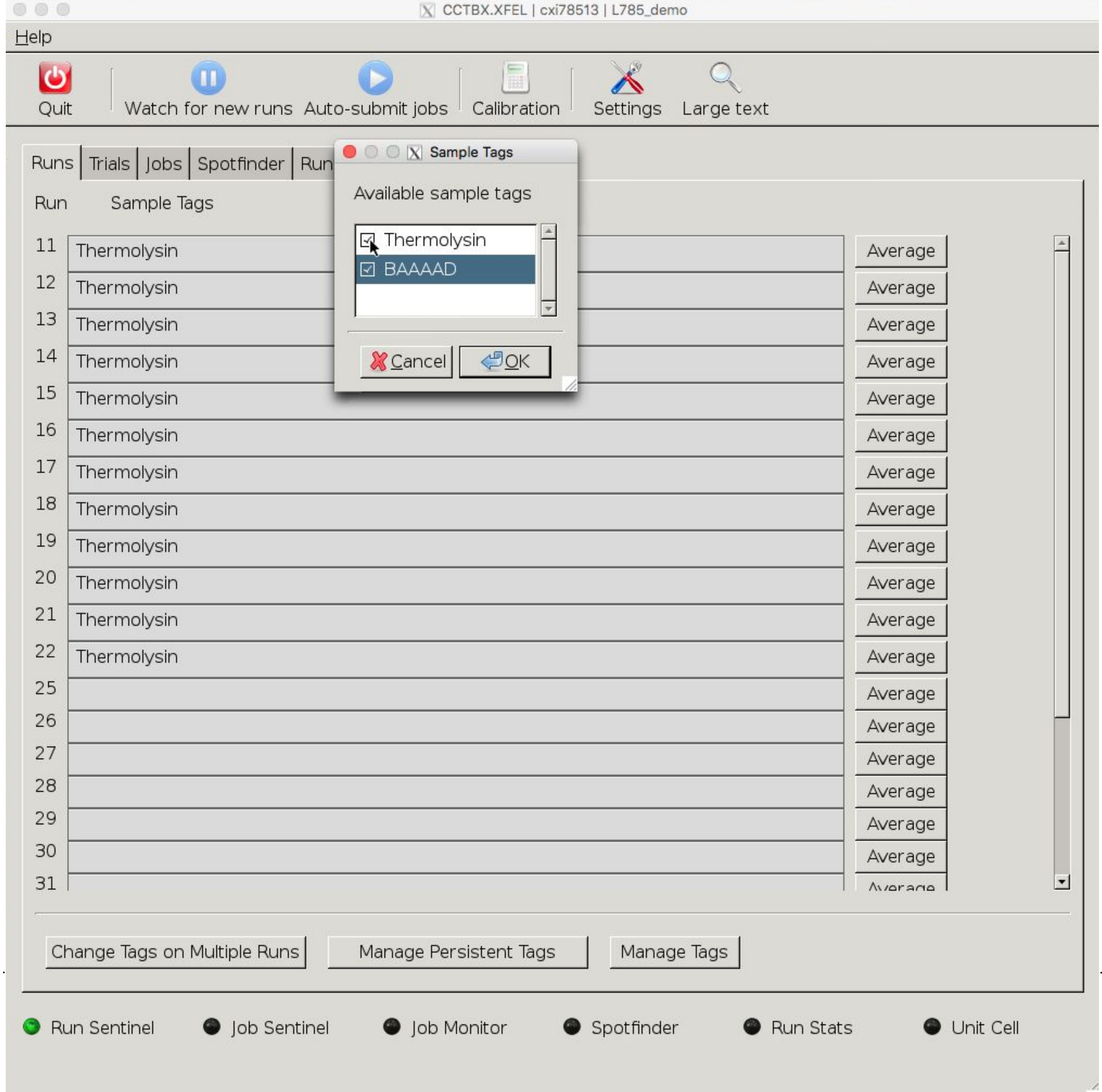
cctbx.xfel GUI

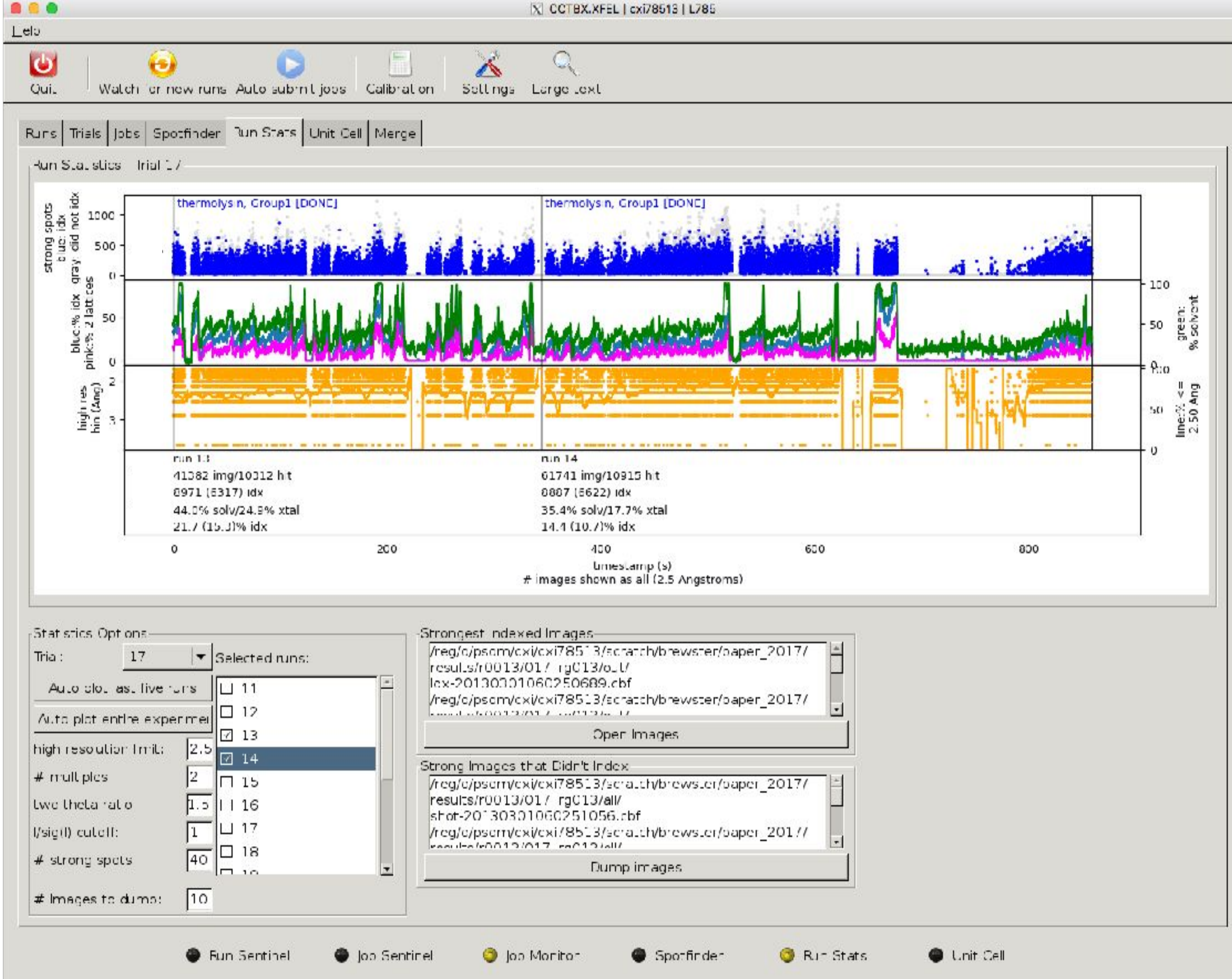
- Built on *cctbx* and DIALS

Winter G. (2018). DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr D Struct Biol* **74**, 85-97

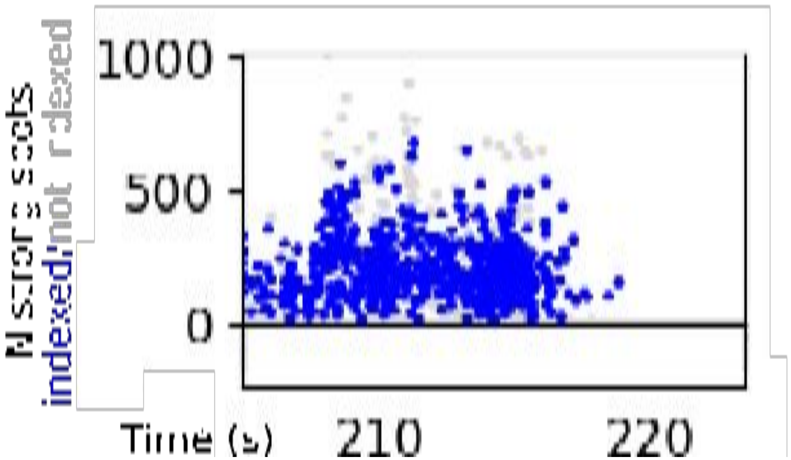
- Full SX workflow
  - Each image: import, spot-find, index, integrate, scale, post-refine, merge
  - Monitor for new data and submit jobs
  - Organize processing many datasets
  - Massively parallel multiprocessing on the biggest computing clusters
  - MySQL backend allows scaling to the kHz regime

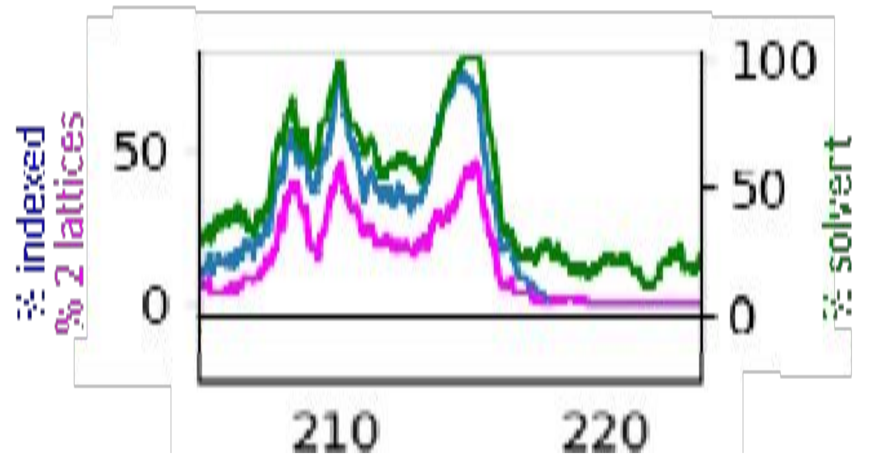
BERKELEY LAB

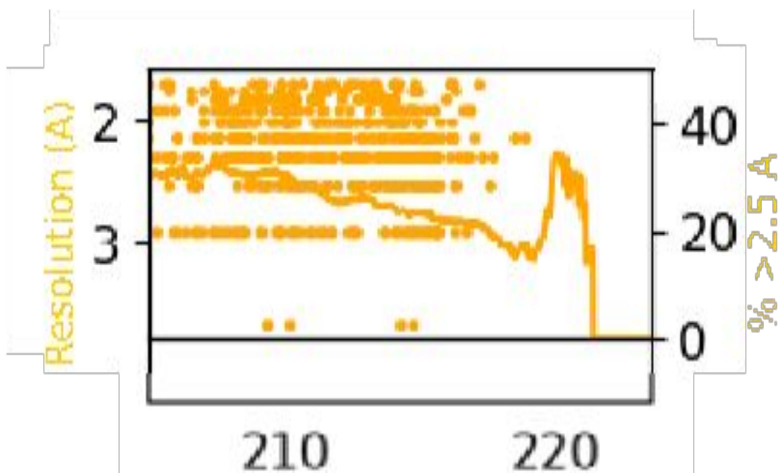

CCTBX.XFEL | cxi78513 | L785_demo
Help
Quit
Watch for new runs
Auto-submit jobs
Calibration
Settings
Large text
Runs
Trials
Jobs
Spotfinder
Run
Sample Tags
Available sample tags
Thermolysin
BAAAAD
Cancel
OK
Run
Sample Tags
11 Thermolysin
12 Thermolysin
13 Thermolysin
14 Thermolysin
15 Thermolysin
16 Thermolysin
17 Thermolysin
18 Thermolysin
19 Thermolysin
20 Thermolysin
21 Thermolysin
22 Thermolysin
25
26
27
28
29
30
31
Average
Change Tags on Multiple Runs
Manage Persistent Tags
Manage Tags
Run Sentinel
Job Sentinel
Job Monitor
Spotfinder
Run Stats
Unit Cell


CCTBX.XFEL | cxi78513 | L785
Quit
Watch for new runs
Auto submit jobs
Calibration
Settings
Large text
Runs
Trials
Jobs
Spotfinder
Run Stats
Unit Cell
Merge
thermolysin, Group1 [DONE]
strong spots
blue: idx
gray: did not idx
blue:% idx
pink:% 2 lattices
green:
% solvent
high res
bin (Ang)
run 13
41382 img/10312 hit
8971 (6317) idx
44.0% solv/24.9% xtal
run 14
61741 img/10915 hit
8887 (6622) idx
35.4% solv/17.7% xtal
timestamp (s)
images shown as all (2.5 Angstroms)
Statistics Options
Trial:
Selected runs:
Auto plot last five runs
Auto plot entire experiment
high resolution limit:
multiples
two theta ratio
I/sig(I) cutoff:
strong spots
Images to dump:
Strongest indexed Images
Open Images
Strong Images that Didn't Index
Dump images
Run Sentinel
Job Sentinel
Job Monitor
Spotfinder
Run Stats
Unit Cell

Rate tracking

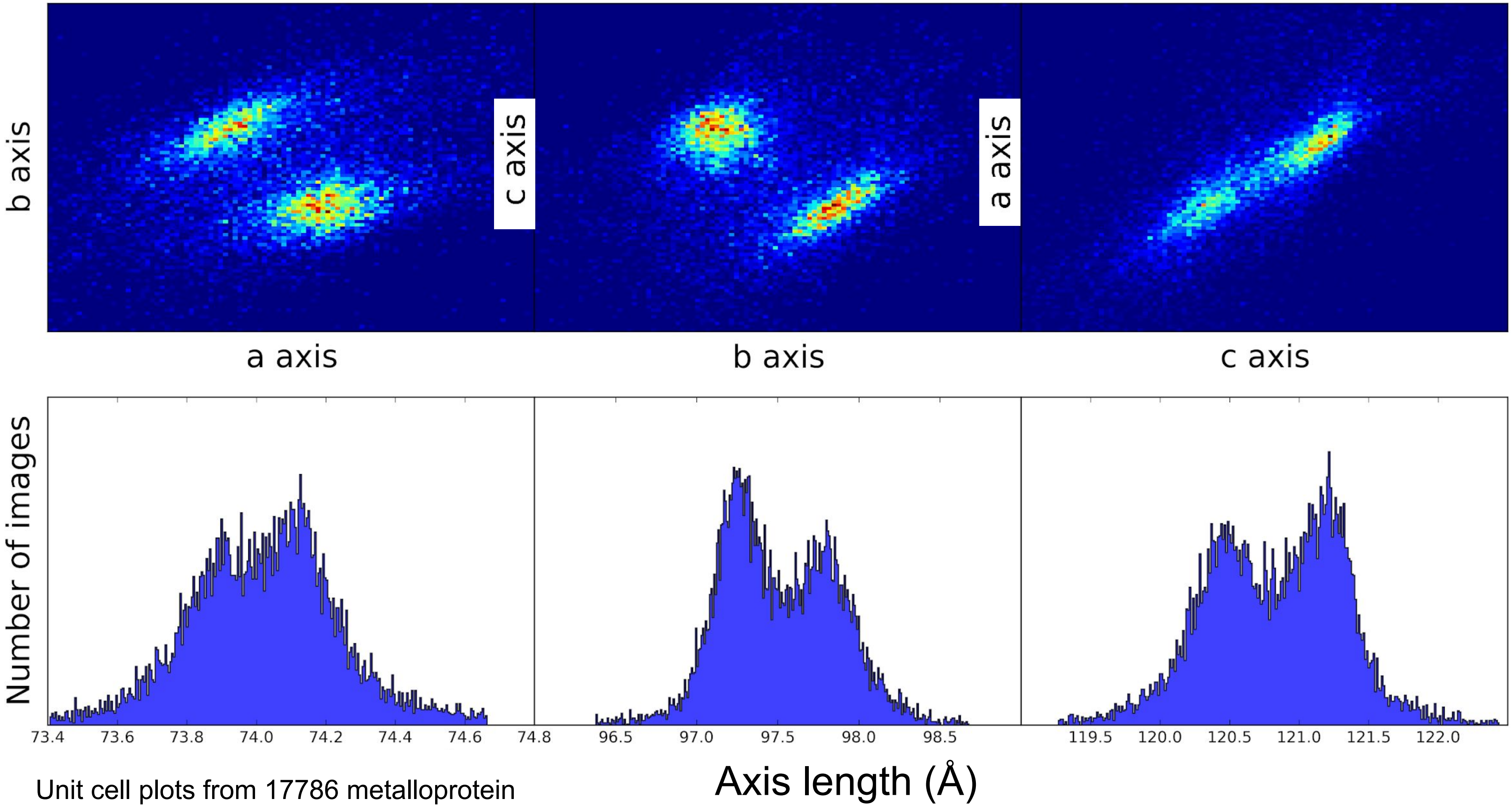






run 13
41382 img/10312 hit
8971 (6317) idx
44.0% solv/24.9% xtal
21.7 (15.3)% idx

Non-isomorphism: two clusters



Unit cell plots from 17786 metalloprotein crystals (space group $P2_12_12_1$) collected at LCLS beamline MFX.

Heuristics

- How much data do I need?
  - What is my hitrate?
  - What is my completeness?
  - Can I stop collecting data?

Easy answer (rule of thumb):

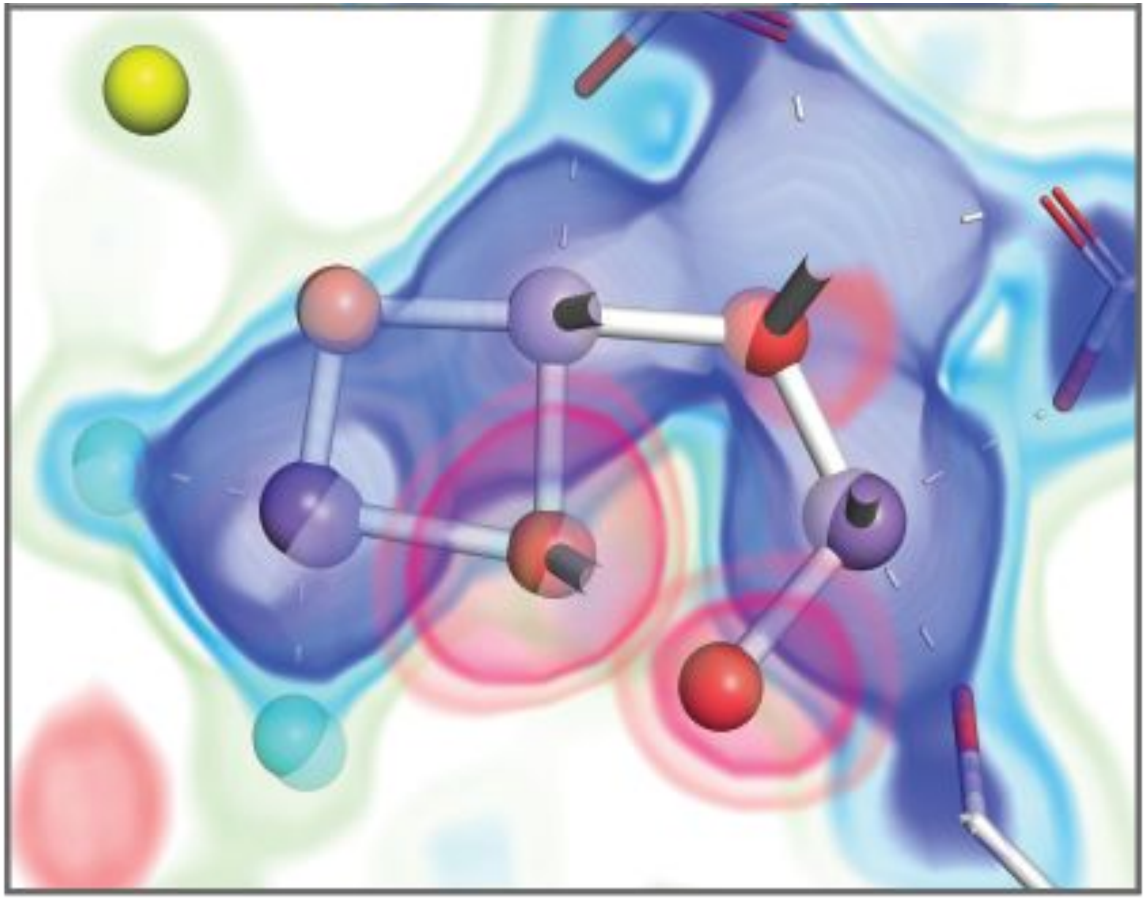
- 10-fold multiplicity at the resolution of interest
- Monotonically decreasing $CC_{1/2}$ to the resolution of interest

The easy answer lends itself to sampling the data

How much data do I need?

- How much data do I need to answer my biological questions?

How much data do I need?



PSII 2Fo-Fc and Fo-Fc at 2.07 Å, S3 state

Kern et al., 2018 Nature 563:421

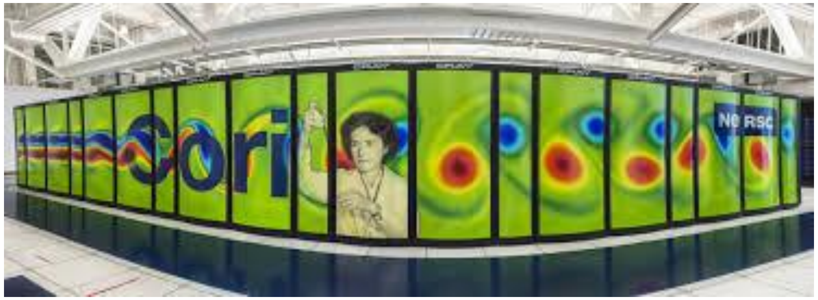
Data analytics at the petascale for free electron lasers

- Processing 1 image takes 30s per hit, 1s per miss, on average
- Vastly varies depending on protocol:
  - Multiple indexing methods
  - Multiple lattices
  - Large unit cell
  - Hit rate
- 30 Hz Rayonix at LCLS: only 288 cores available. Cannot keep up with good hitrates
  - 100% hitrate = 30s/image * 30 images/s = 900s to process 1s of data collection
  - Need 900 cores

Data analytics at the petascale for free electron lasers

Solution:

- Sample the data at LCLS for GUI (process 1 in 5 images). At the same time, transfer the data over 100 Gb-ESnet to NERSC, and process all images on Cori
  - 2388 Intel Xeon Haswell nodes
  - 9688 Intel Xeon Phi “Knight’s Landing” nodes (68 cores)
  - 1.8 PB burst buffer (keep the data on solid state drives)
- 9 TB example experiment; 861969 diffraction images (2.0 hrs’ data collection @ 120 Hz)
  - 150466 useful crystal patterns (17% hit rate)
  - Required 140 “Knight’s Landing” node-hours to process
  - Not counting I/O, 70 nodes in continuous use during data collection
  - About 1% of Cori



Four year Exascale Computing Project funding to scale this for

- bigger algorithms
- faster LCLS-II pulse rate
- exascale machines in a few years’ time
- Collaboration with Amedeo Perrazo, SLAC

Data collection at LCLS / MFX December 2018

- Four 12-hour shifts at MFX (LU50 / P143)
- 20 – 30 Hz overall collection rate (Rayonix 340)
- 594,230 total indexed images
- > 8.4 indexed images s^{-1} (31% of all shots)
- Overall 48% collection duty cycle
- ~ 54 tr-SFX + tr-XES data sets
- ~ 12 types of enzymes
- Typically ~ 1 mg protein per dataset
- Some data beyond 1.3 Å resolution

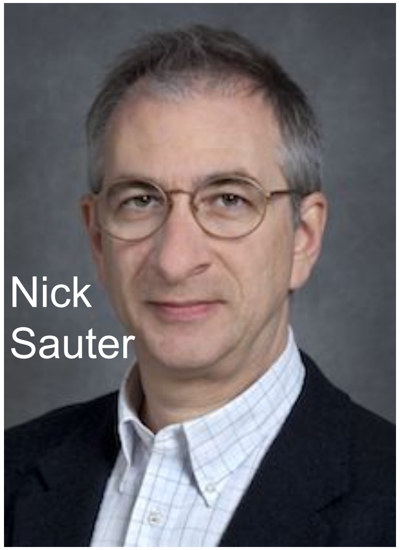
data se						x 340 r rate
1						lz
2						lz
3						lz
4						lz
5						lz
6						lz
7						lz
8						lz
9						lz
10						lz
11						lz
12						
13						lz
14						lz
15	Enz 01 oxidized | no run 20, no data in run	1.6	1.76	16851	45	30 Hz

Deployment: every XFEL

	LCLS	SACLA	PAL-XFEL	SwissFEL	EuXFEL
Location	Stanford, CA USA	Harima, Japan	Pohang, South Korea	Aargau, Switzerland	Schenefeld, Germany
Computing	NERSC or SDF	Local	KISTI	PSI	Maxwell
Detectors	CSPAD, ePix, Jungfrau, Rayonix	Octal-8 MPCCD, Rayonix	Rayonix	Jungfrau 16M	AGIPD

Supports LSF, SGE, Slurm, HTCondor, Docker/Shifter
New use of GUI: DLS beamline I24 (Pilatus3 6M)
Best current result: LCLS + NERSC, 5 minute turnaround for fully merged dataset

XFEL @ LBL


Nick
Sauter


Asmit Bhowmick


Elyse
Schriber


Aaron Brewster


Derek Mendez


Dan
Paley

Acknowledgements

Berkeley National Lab
Nicholas Sauter
Asmit Bhowmick
Derek Mendez
Dan Paley
Elyse Schriber
Paul Adams
Peter Zwart
Vittal Yachandra
Junko Yano
Jan Kern
Ruchira Chatterjee
James Holton
Johannes Blaschke
In-Sik Kim
Cindy C. Pham
Joshua Heinemann
Louise Lasalle
Billy K. Poon

LLNL
Michael Wall

Brookhaven
Herbert Bernstein

Nebraska
Limei Zhang

IBS
Jacques-Philippe Colletier

LCLS
Uwe Bergmann
Alberto Lutman
Franklin Fuller
Roberto Alonso-Mori
Sebastien Boutet
Mengning Liang
Amedeo Perazzo
Chuck Yoon
Chris O'Grady
Mark Hunter
Jason Koglin
Ray Sierra
Mona Uervirojnangkoorn
Alexander Batyuk
Diling Zhu
Anton D. Loukianov

UCSF
James Fraser
Iris Young
Michael Thompson
Alex Wolff
Rahel Woldeyes

Harvard
Kevin Dalton
Doeke Hekstra

Diamond Light Source
Gwyndaf Evans
Graeme Winter
David Stuart
Jonathan Grimes
James Parkhurst
Luis Fuentes-Montero
Markus Gerstel
Nicholas Devenish
Allen Orville
Pierre Aller
Agata Butryn
Danny Axford
Tiankun Zhou
Jos Kamps
Robert Bosman

CCP4
David Waterman

UCLA
David Eisenberg
Duilio Cascio
Michael Sawaya
Jose Rodriguez
Luki Goldschmidt
Johan Hattne

Stanford University
Axel Brunger
Bill Weis
Soichi Wakatsuki
Henry van dem Bedem

SSRL
Mike Soltis
Ana Gonzalez
Ashley Deacon
Aina Cohen
Artem Lyubimov
Scott McPhillips

Charité-Universitätsmedizin Berlin
Andrea Schmidt
Michal Szczepek
Patrick Scheerer

Alums
Nat Echols
Tara Michels-Clark
Lee O'Riordan
Muhamed Amin
Oliver Zeldin
Sheraz Gul
Robert Bolotovsky
Kyle D. Sutherlin

NIH/NIGMS grants 1R01GM117126
DOE/Office of Science contract DE-AC02-05CH11231

Zoom background: "Black Lotus" from Magic the Gathering

References

Brewster AS, et. al (2019): Processing serial crystallographic data from XFELs or synchrotrons using the *cctbx.xfel* GUI. *Computational Crystallography Newsletter* **10**, 22-39.

http://www.phenix-online.org/newsletter

Brewster AS, et. al (2018): Improving signal strength in serial crystallography with DIALS geometry refinement. *Acta Crystallogr D Struct Biol* 74, 877-894.

http://cci.lbl.gov/xfel

http://dials.github.io