XFEL crystallography: Imaging Molecular Reactions

What? Ultimate goal: imaging real-time molecular dynamics of biomolecules at the atomic-scale.

How? Time-resolved serial crystallography using pulsed X-ray Free-Electron Laser (XFEL) and synchrotron sources.

When? XFEL crystallography is young (10 years). First time-resolved SFX experiments ~ 5 years ago.

Within the next 5-10 years we aim to capture the first atomic-scale molecular movies from single biomolecules.

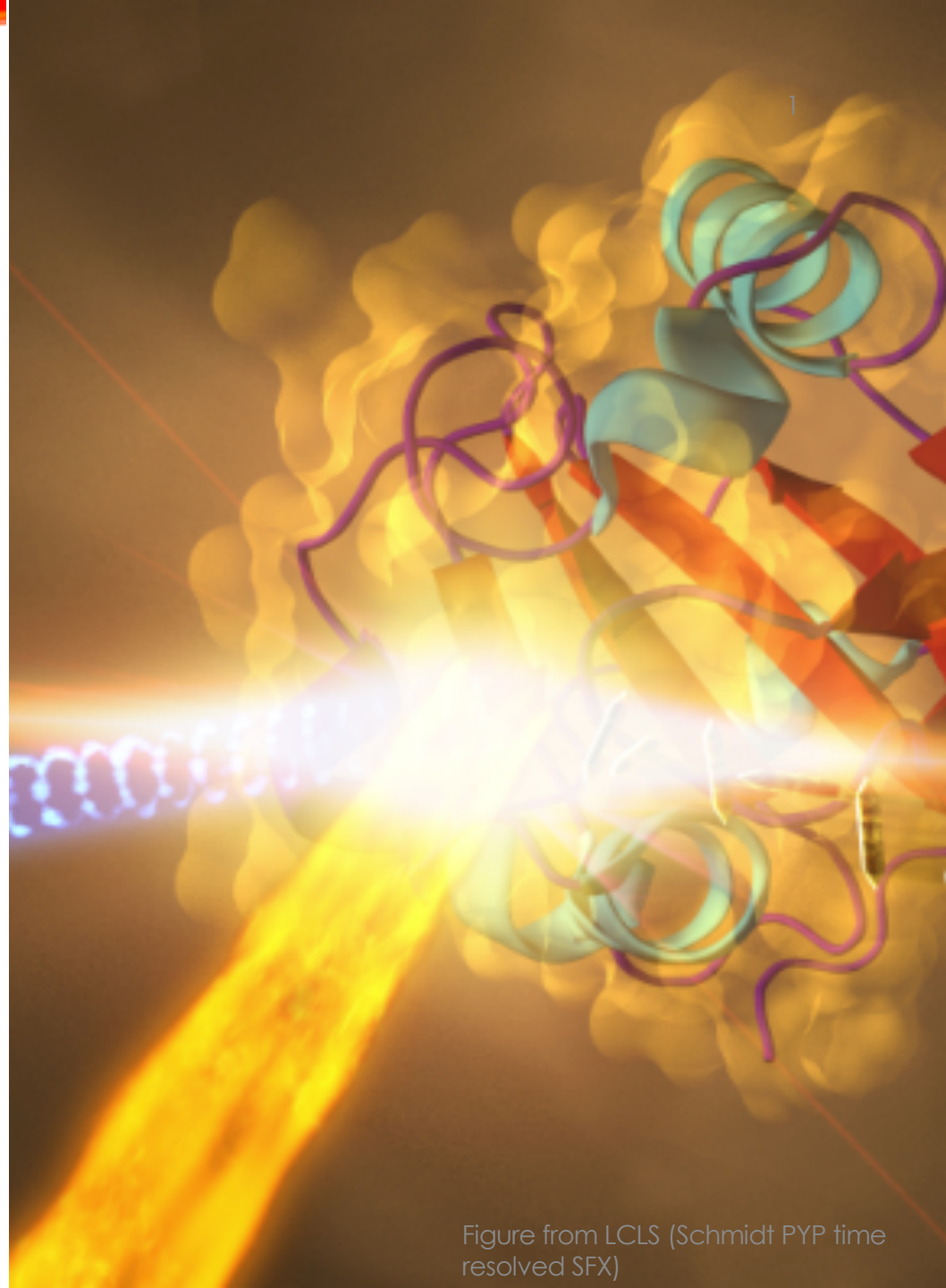
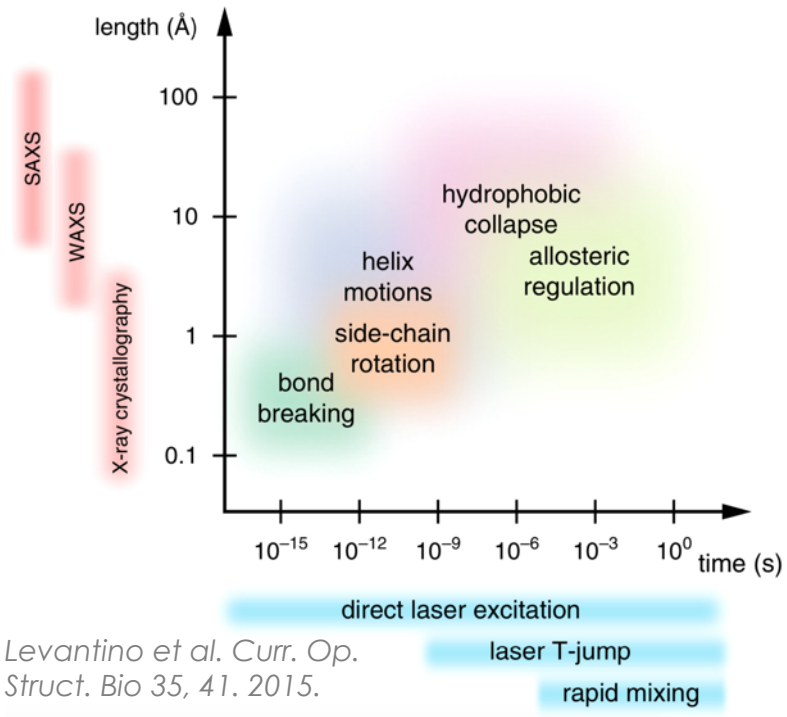


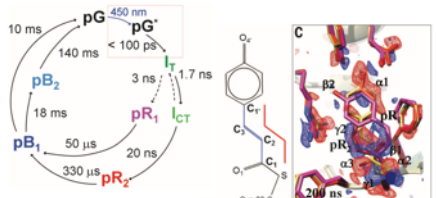
Figure from LCLS (Schmidt PYP time resolved SFX)

Time-resolved serial femtosecond crystallography

What?



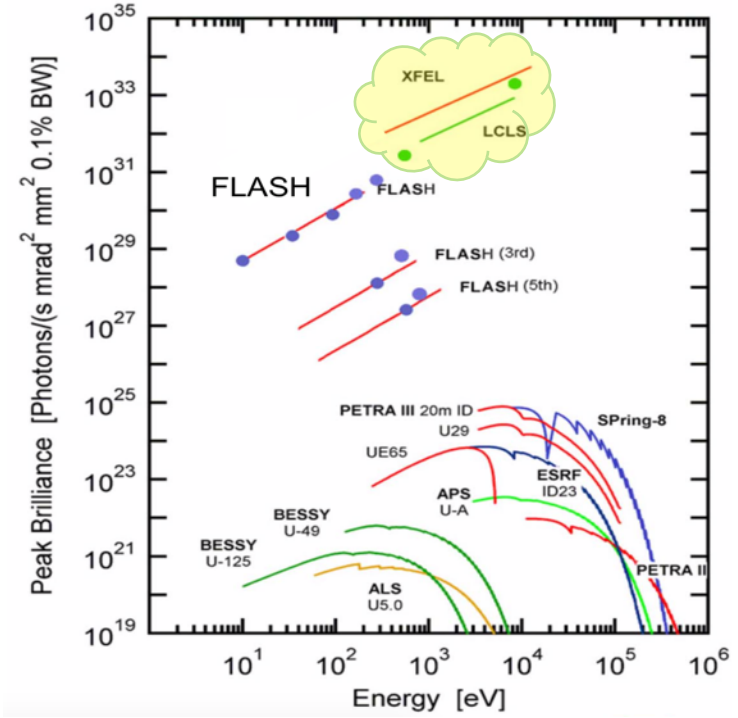
Levantino et al. Curr. Op. Struct. Bio 35, 41. 2015.

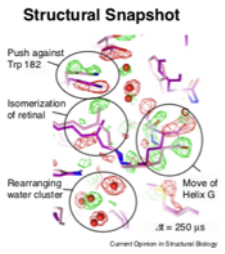


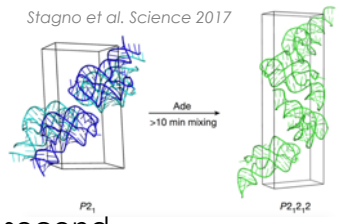
fs - ms

PYP photocycles
Pande et al. Science 2016.

How?







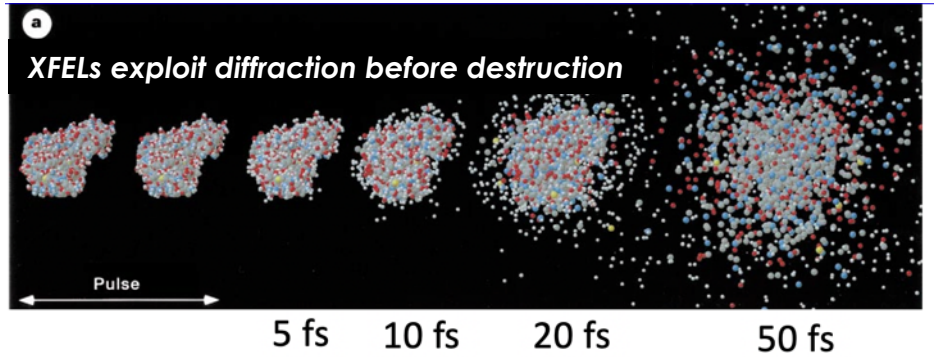
second

State of the art

TR-SFX with XFELs enabled the unprecedented:

atomic-resolution macromolecular structures of **undamaged**, room temperature, **reaction intermediates** from tiny crystals **with fs to second time resolution**

Time-resolved serial femtosecond crystallography



Neutze et al. Potential for biomolecular imaging with fs X-ray pulses. 2000. Nature 406, 752.

- XFEL pulses outrun radiation damage
- Reaction initiation is flexible, i.e. broadly applicable
- Time resolution demonstrated:
  - fs pump probe SFX (light sensitive proteins)
  - ms to minutes for mix-and-inject experiments (most of biology)

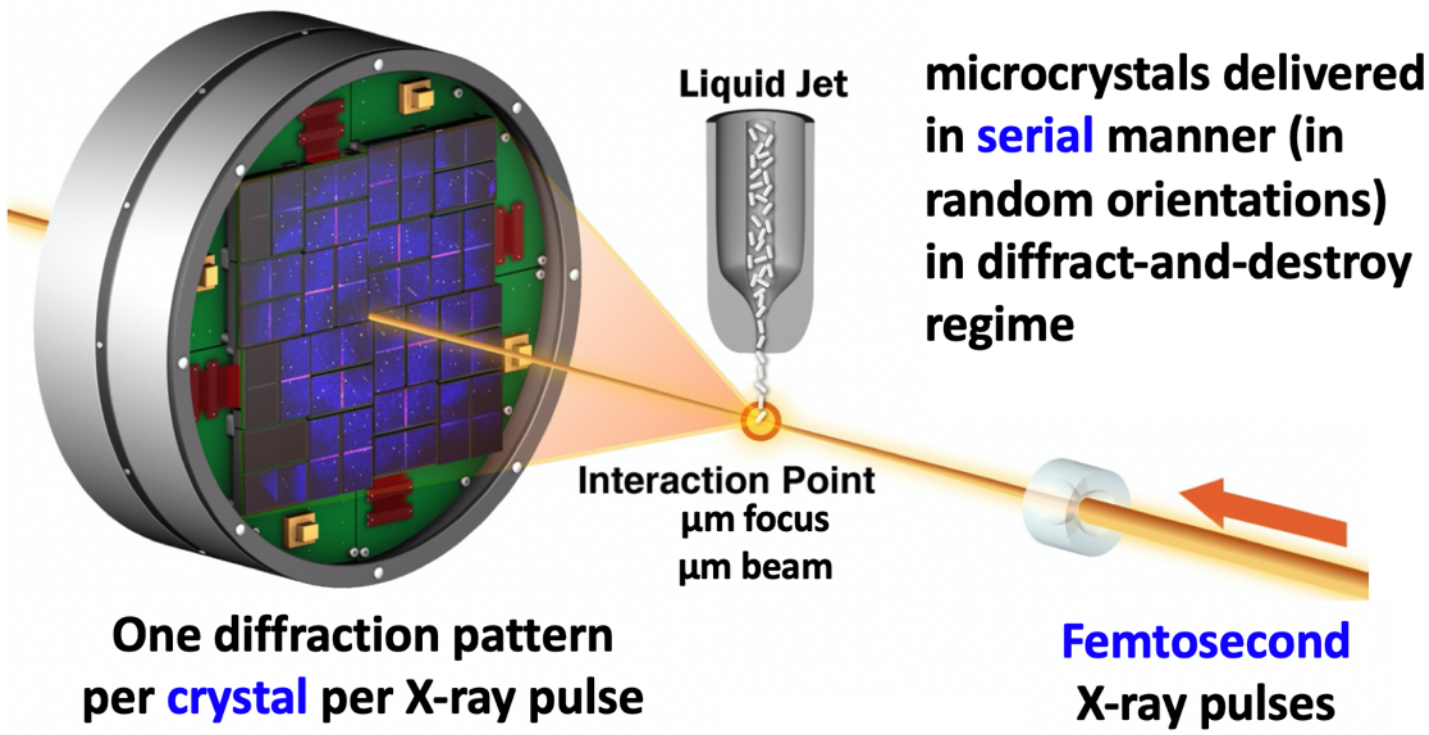
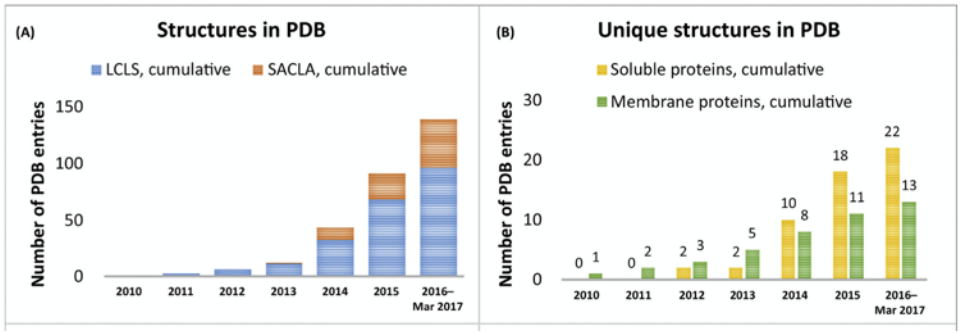


Fig: adapted from Chapman et al. 2011 Nature "Serial fs crystallography"

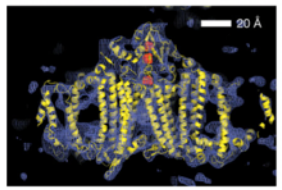
Serial femtosecond crystallography applications

- Protein structure → function
- Applications:
  - understanding fundamental processes like photosynthesis;
  - Biomedically important proteins
  - Structure-based drug design



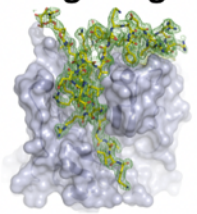
Johansson et al. 2017. Trends in Biochem. Sci. 42(9).

photosynthesis



3PCQ
Photosystem I (the first SFX experiment)
Chapman et al., Nature 2011

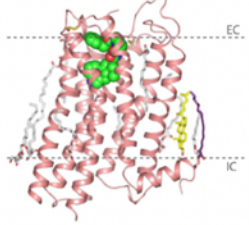
Drug design



4HWY
Natively inhibitied Cathepsin B
Redecke, Nass et al., Science 2013

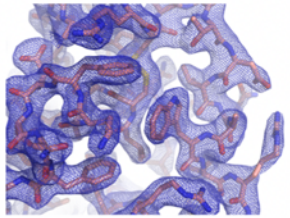
Human GPCRs

4O9R
Smoothened receptor using LCP injector
Weierstall et al., Nature Communications 2014



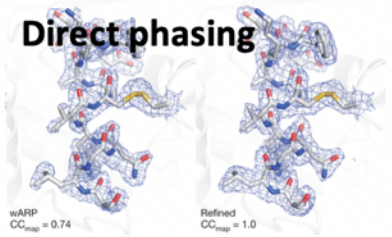
4NC3
Serotonin receptor bound to ergotamine
Liu et al., Science 2013

synchrotron

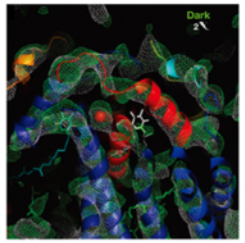


4O34
Serial crystallography using a synchrotron beamline
Stellato et al., IUCrJ 2014

Direct phasing



4N5R
Lysozyme (Gd derivative) ab initio phasing using SAD
Barends et al., Nature 2013

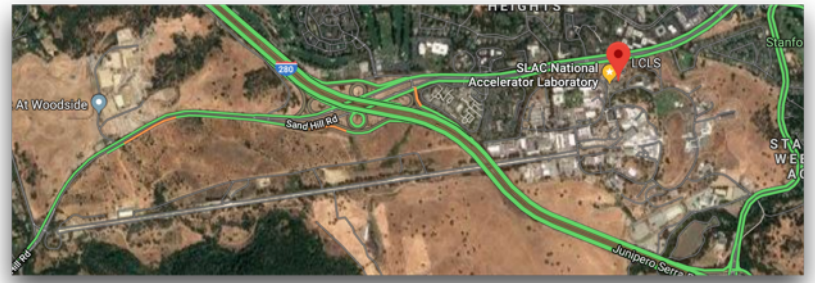


4Q54
Photosystem II in S_3 excited state
Kupitz, Basu et al., Nature 2014

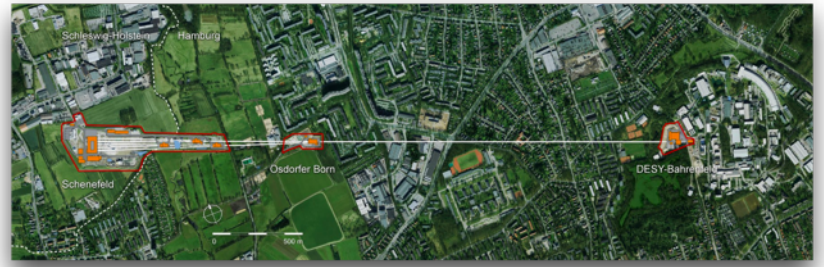
Pump probe

& more

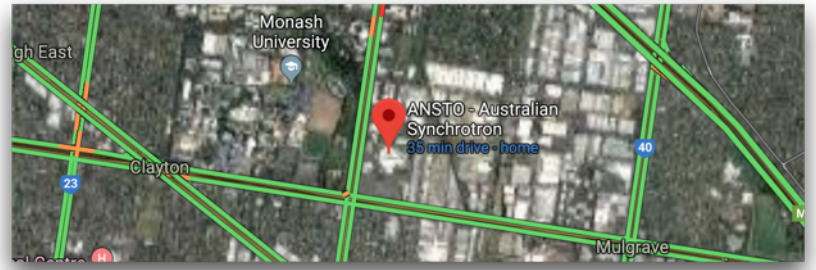
5 XFELs available in 2020



LCLS
LCLS-II
LCLS-II HE (high energy upgrade)



EuXFEL

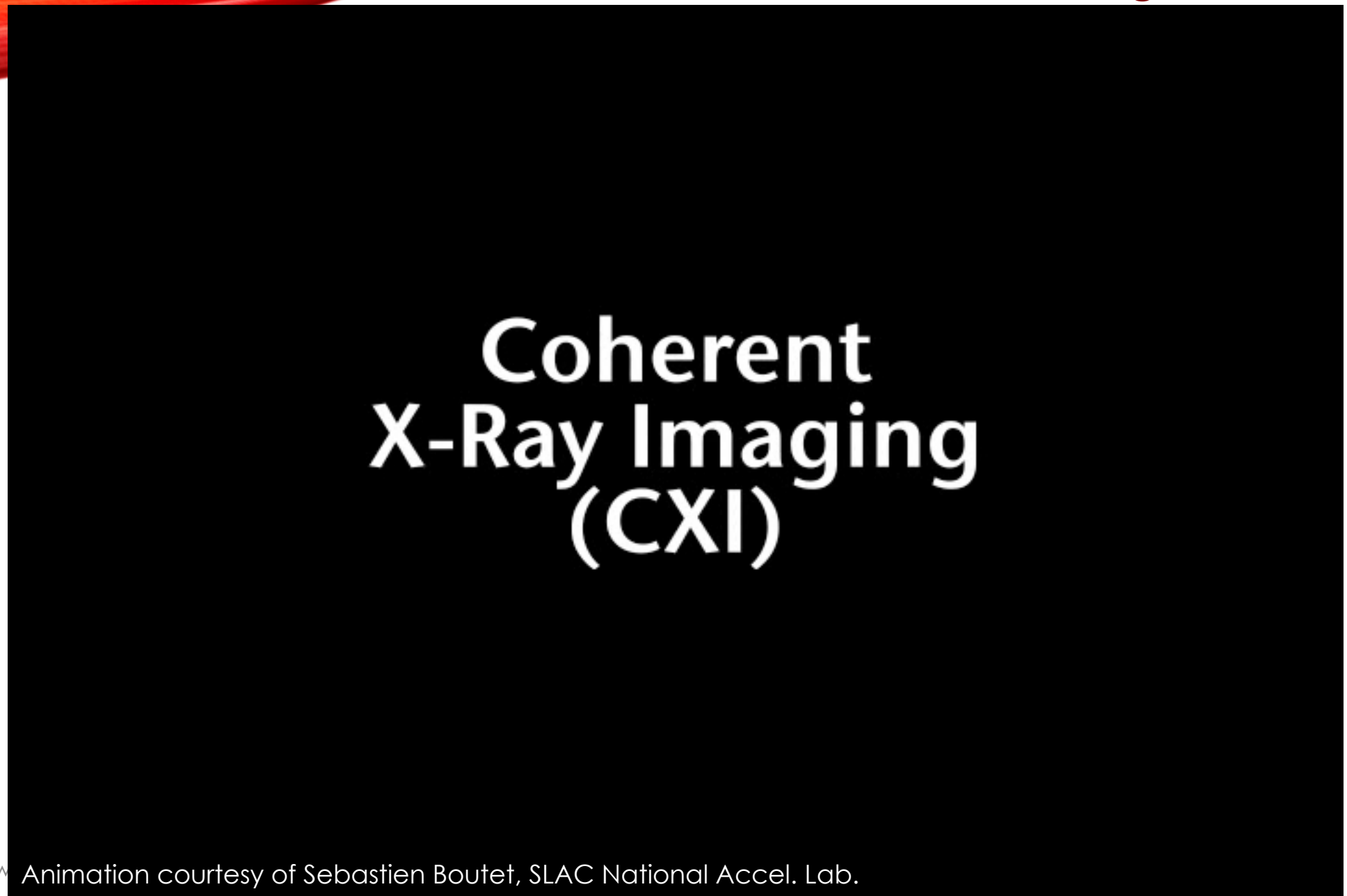


Australian Synchrotron

Facility	Year of first experiments	X-ray pulses per second	Minimum X-ray wavelength (nm)	Peak brilliance
EuXFEL	2017	27,000	0.05	5 x 10^33
LCLS	2009	120	0.15	2 x 10^33
PAL-XFEL	2017	60	0.06	1.3 x 10^33
SACLA	2011	60	0.08	1 x 10^33
SwissFEL	2018	100	0.1	1 x 10^33

Source: Source: European XFEL

LCLS: Linac Coherent Light Source, SLAC



Animation courtesy of Sebastien Boutet, SLAC National Accel. Lab.

Data rates in serial crystallography

XFEL crystal data collection

- One shot per crystal
- Need to sample reciprocal space fully (all orientations of crystals)
- Not every XFEL pulse hits a crystal
- Not every crystal hit is indexable
- 10-50 thousand indexed patterns needed per structure, per time point

Large data demands are due to shot-to-shot fluctuations in

- Crystal sizes
- Crystal quality
- Crystal isomorphism
- X-ray pulse bandwidth
- X-ray pulse spectrum
- X-ray pulse intensity
- Position of crystal in the beam
- Partially recorded reflections

Example of dataset size

2 % hit rate
50 % indexing rate
25,000 frames per dataset
4 datasets (100K patterns)
Total: **10E6 patterns** for 4 time points.

At 3520 frames/s, this is 50 minutes data collection at EuXFEL

Light/dark interleaving of data → 100 minutes at EuXFEL

100s of TBs per experiment

XFEL serial fs crystallography data analysis pipeline

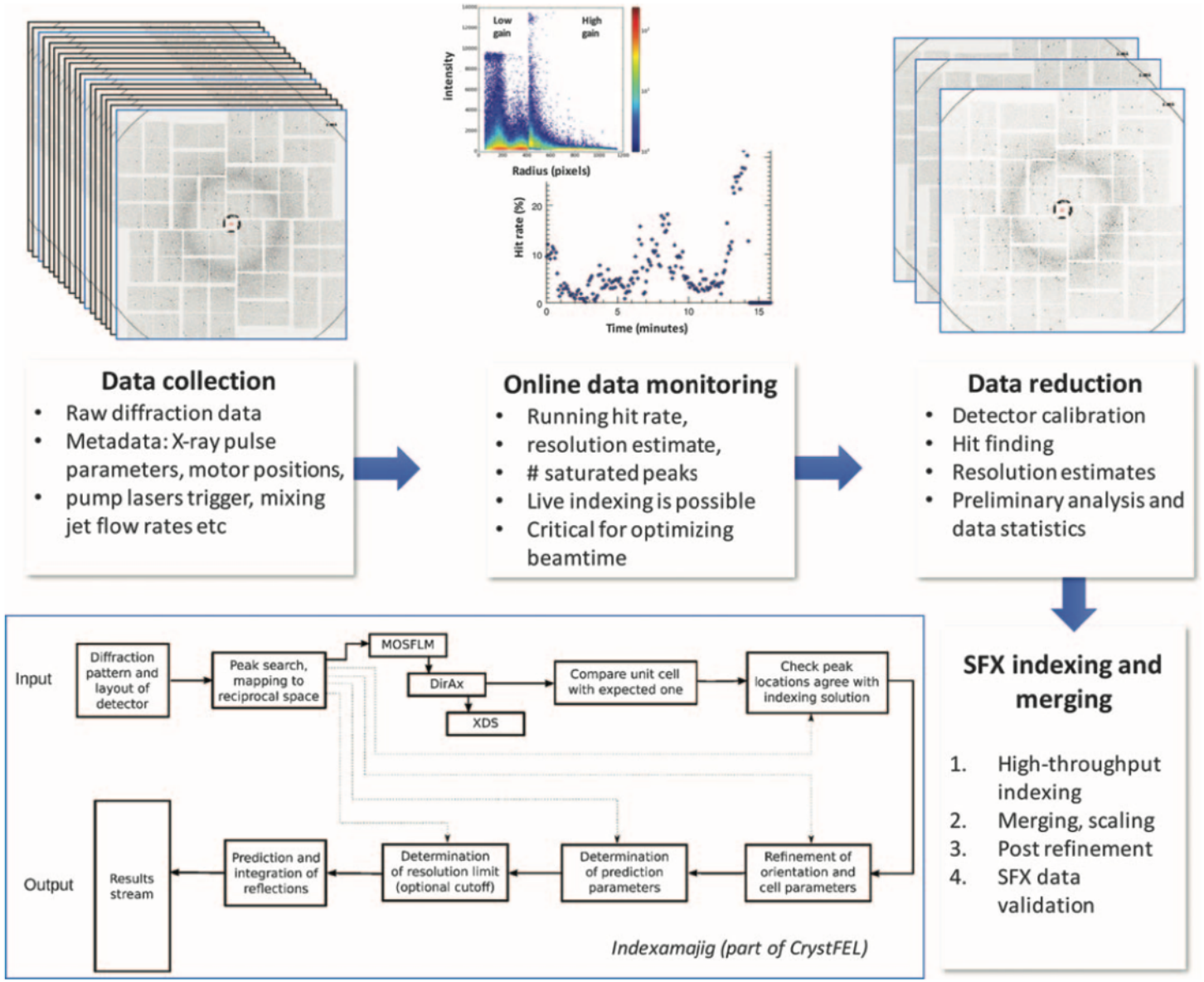
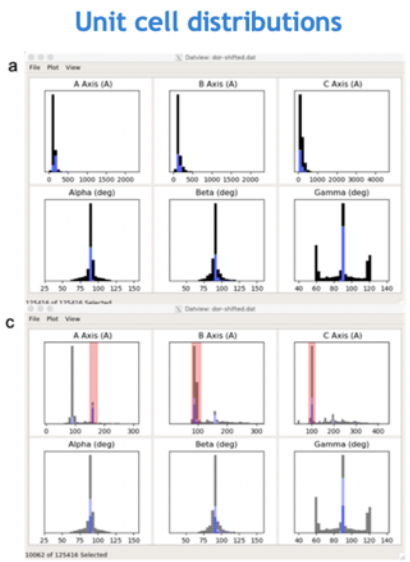
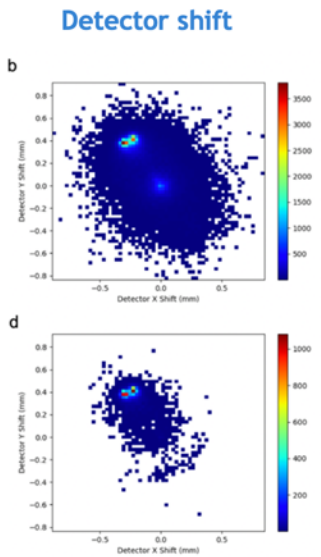


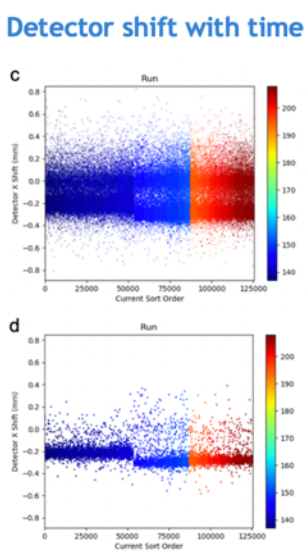
Fig: Zatsepin "Crystallography with X-ray free-electron lasers" in Protein Crystallography: Challenges and Practical Solutions Roy. Soc. Chem. 2018.

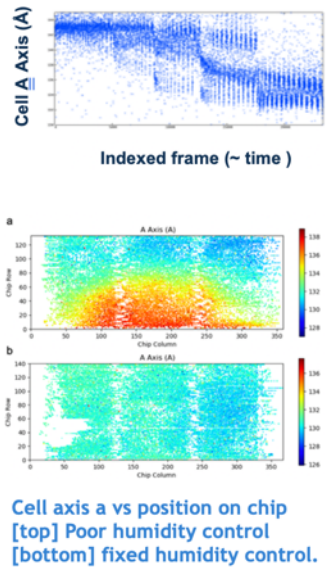
Data processing – highly parallelizable

- 100's thousands patterns need to be reprocessed multiple times for optimization of data processing – basics done during beamtime, mostly done in months after beamtimes.
- This is highly parallelizable – ideal for batch processing on CPUs.

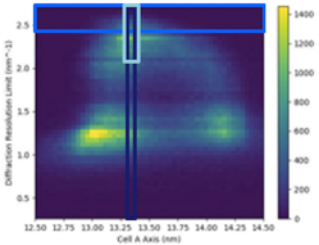


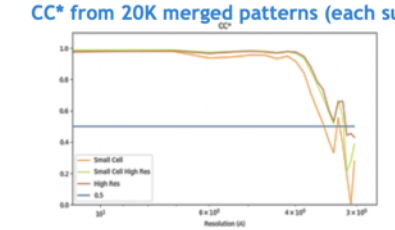






Cell axis a vs position on chip
[top] Poor humidity control
[bottom] fixed humidity control.

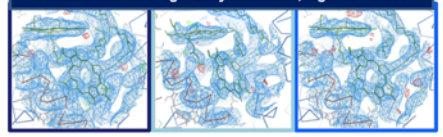




Phenix.phaser with PDB 3wu2

Dataset	a >=	a <=	Res >	Count
Small Cell	13.28	13.32	-	20,373
Small Cell, High Res	13.25	13.36	2.09	20,291
High Res	12.5	14.5	2.42	19,889

Phenix.refine with Rigid Body Monomers, high res = 3.3 Å



Small Cell
R-Free = 0.3290

Small Cell, High Res
R-Free = 0.2987

High Res
R-Free = 0.2883

Stander, Fromme, Zatsepin. "DatView: a graphical user interface for visualizing and querying large data sets in serial femtosecond crystallography." 2019 IUCrJ 52.

XFELs for structural biology

- XFEL use for structural biology includes serial crystallography (SFX), solution scattering (uncrystallized) and single particle imaging (SPI)
- HPC needs differ: e.g. SPI needs larger data volumes and benefits more from GPUs than SFX

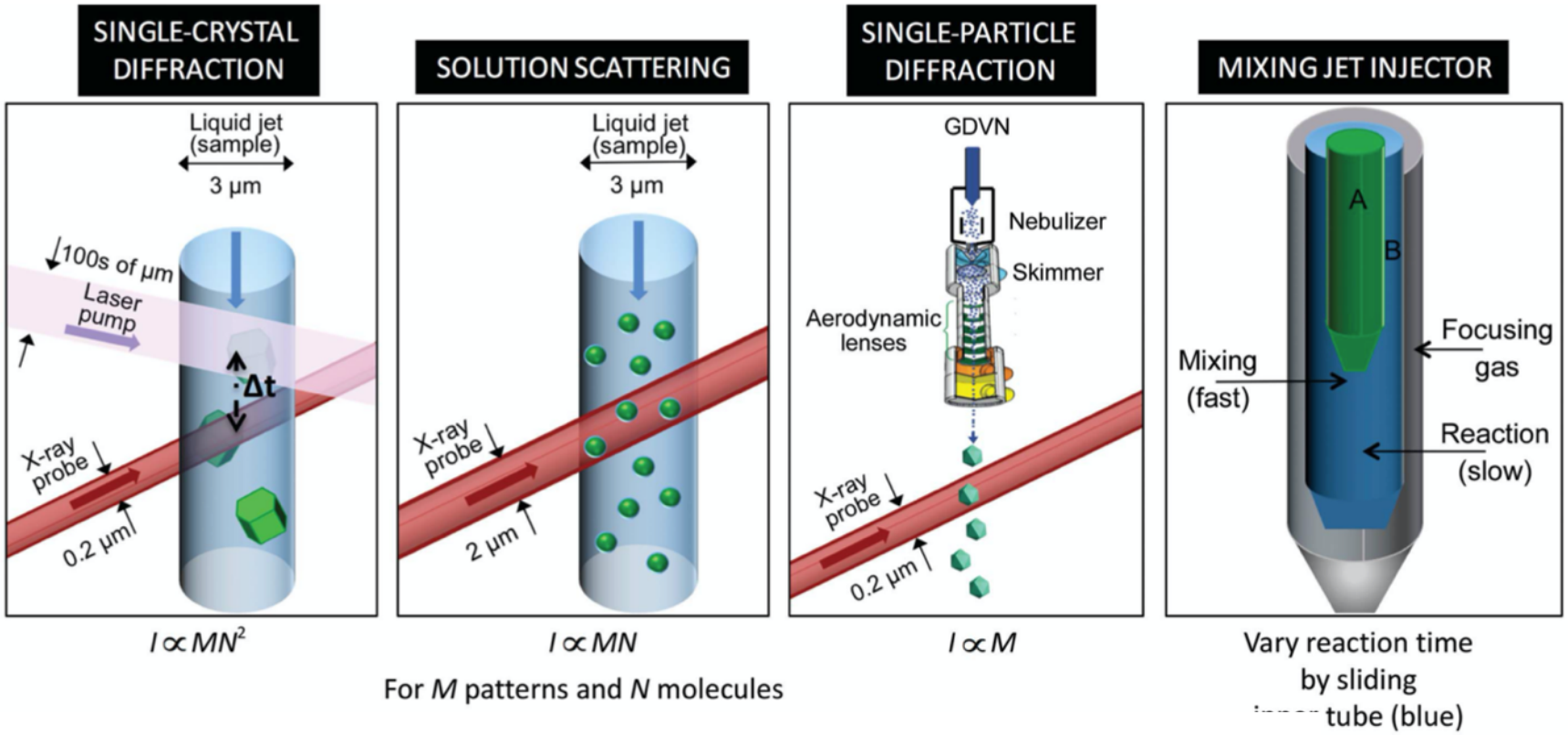


Fig: Spence XFELs for structure and dynamics in biology. 2017 IUCrJ 4, 322.

Data rates in serial crystallography

XFELs

- EuXFEL: 27,000 Hz (at rate of 4.5 MHz in bunches at 10 Hz)
- Detectors. **Up to 32 GB/s**
  - AGIPD: 1 MP, up to 3520 frames/s
  - Large Pixel Detector: 1 MP, 512 frame memory depth. 5110 frames/s. **Up to ~10 GB/s** per megapixel.
  - DSSC: 1 MP, 8000 frames/s.
- Other XFELs: 10 to 120 Hz
- LCLS-II: 1 MHz planned…

Synchrotron serial crystallography

- EIGER detector: 133 Hz, 18 MP
- EIGER2 XE: 400 - 550 Hz,
  - **up to 20 GB/s** before compression
  - Best lossless compression offers at best 10x reduction (data dependent)

- Time-resolved crystallography is also being developed at synchrotrons (for slower time points).
- Time-resolved serial crystallography at the Australian synchrotron could be possible in a few years at MX3
  - We need to plan now for data processing and storage

Serial crystallography data curation

Who should store the data? Users? Facilities? Who should pay?

LCLS

Policy by Folder

Space	Quota	Backup	Lifetime	Comment
xtc	None	Tape archive	4 months	Raw data
usrdaq	None	Tape archive	4 months	Raw data from users' DAQ systems
hdf5	None	Tape archive	4 months	Data translated to HDF5
scratch	None	None	4 months	Temporary data (lifetime not guaranteed)
results	4TB, 10K files	Tape backup	2 years	Analysis results ⭐
calib	None	Tape backup	2 years	Calibration data
User home	20GB	Disk + tape	Indefinite	User code
Tape archive	-	-	10 years	Raw data (xtc, hdf5, usrdaq)
Tape backup	-	-	Indefinite	User home, results and calib folder
Disk backup	-	-	Indefinite	Accessible under ~/.zfs/

⭐ *For older experiments this folder is called res instead of results*

European XFEL

Data retention policy for the European XFEL

Storage class	Quota	Safety	Lifetime	Comment
dcache.raw	None	Tape Archive	6 months	Raw data on commodity disks
raw	None	None	2 months	Very fast accessible raw data, lifetime not guaranteed
usr	5TB	Snapshots + Tape Backup	24 months	User data, results
proc	None	None	6 months	Processed data (e.g. calibrated)
scratch	None	None	6 months	Temporary data (lifetime not guaranteed)
dcache.cal	None	Tape Archive	10 years	Calibration constants on commodity disks
cal	None	Tape archive	6 months	Very fast accessible calibration data
user home	20GB	Snapshots + Tape Backup	Lifetime of the account	Home folder for user account
archive.raw	None	-	Long-term	"Long term" means 5 years and XFEL will strive for 10 years
archive.cal	None	-	10 years	

"European XFEL GmbH will strive for 10 years (***storage period***)… The precise period will depend on the type and volume of data concerned and the economic consequences associated to long-term data storage. Thus, the European XFEL GmbH reserves the right to restrict the storage periods of data sets **in consultation with the respective communities of high data-rate inst**ruments.

Serial crystallography data curation

Who should store the data? Users? Facilities?
Who should pay?

AUSTRALIAN RESEARCH COUNCIL

2019 July

The ARC does not require full, detailed data management plans (such as those required by some funding agencies internationally) and does not mandate open access to data.

This might (should?) change

National Science Foundation (USA)

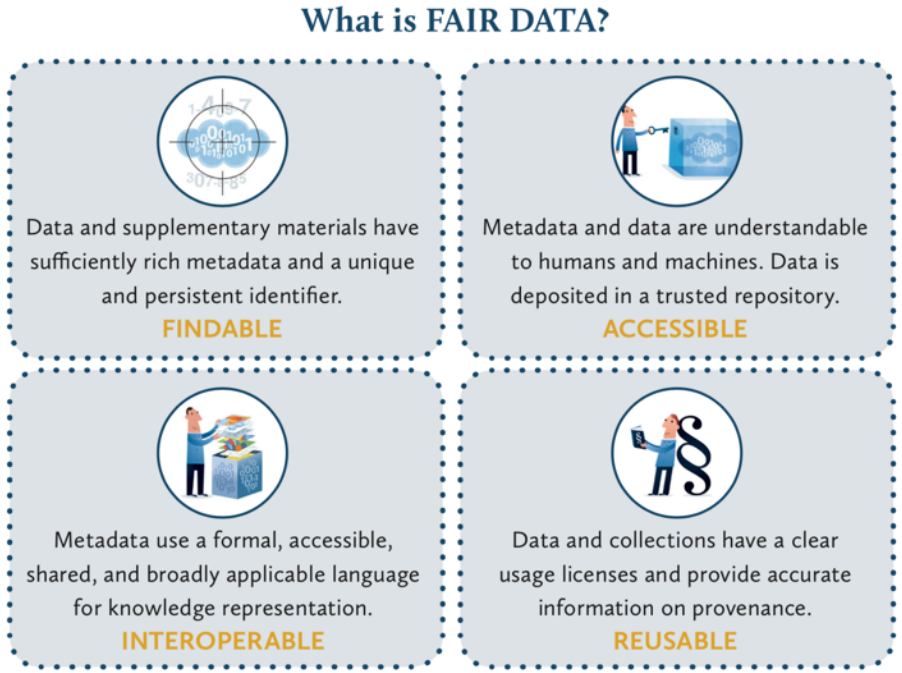
NSF DATA SHARING POLICY

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

Further considerations about serial crystallography data

- Things to consider: FAIR data
- Data storage long term
- Data sharing – to maximise impact of publicly funded research

- Consider: MX3 users at Aust. Synchrotron could apply for HPC resources alongside applying for beamtime: forces a data analysis and management plan and best use of HPC. Would need collaboration with ANSTO…



https://libereurope.eu/blog/2018/07/13/fairdataconsultation/

XFEL data sharing


Depositing Data

CXIDB Coherent X-ray Imaging Data Bank

To deposit a new dataset in CXIDB please send an email to cxidb@cxidb.org with the subject 'CXIDB initial deposition' and the following content (example from ID 59):

```
# Deposition Summary
- Depositor name:  Kartik Ayyer
- Depositor email: kartik@example.edu

# Publication Details
- Title:  Macromolecular diffractive imaging using imperfect crystals
- Authors: Kartik Ayyer et al.
- Journal: Nature
- Year: 2016
- DOI: 10.1038/nature16949

# Experimental Conditions
- Method: Serial Femtosecond Crystallography
- Sample: Photosystem-II
- Wavelength: 1.31 Å (9.48 keV)
- Lightsource: LCLS
- Beamline: CXI

# Description
Please check the README file for more information about the dataset.
```

If you don't know or cannot, at the time, disclose certain fields just leave them empty.

You will need to complete all fields before the entry is made available. All deposited data and metadata are made available under the CC0 waiverto promote maximum reuse.

Maia, F.R.N.C. The Coherent X-ray Imaging Data Bank. *Nat. Methods* **9**, 854–855 (2012).