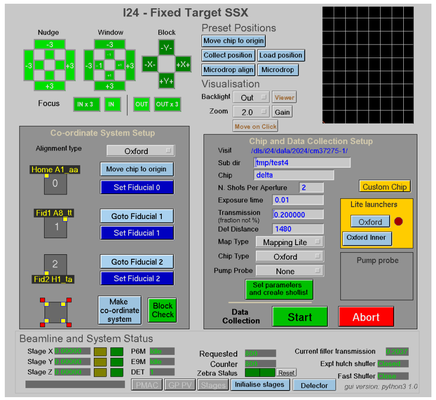
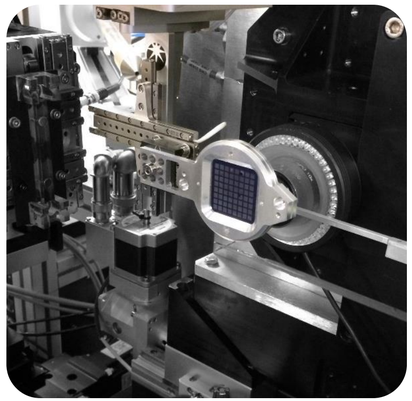
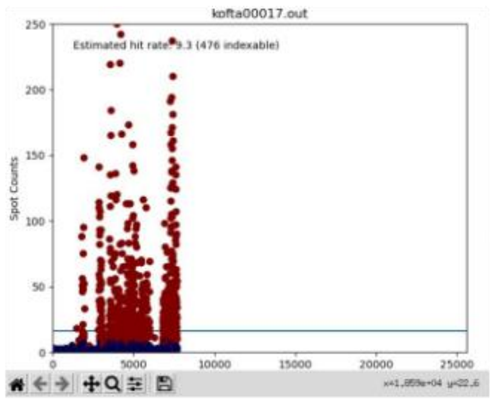
Automation for serial crystallography data collection and processing at Diamond Light Source







---------- Merging ----------
7850 crystals in total scaled in space group P 43 21 2
Median cell: 79.226, 79.226, 38.161, 90.0, 90.0, 90.0
Merged mtz file: merged.mtz

	Suggested	Low	High	Overall
High resolution limit		1.60	4.34	1.60	1.32
Low resolution limit		56.02	56.06	1.63	56.02
Completeness		100.0	100.0	100.0	94.6
Multiplicity		94.1	169.9	39.1	61.5
I/sigma		20.9	69.7	1.4	12.6
Rmerge(I)		0.964	0.662	20.751	1.038
Rmerge(I+/-)		0.962	0.661	20.711	1.035
Rmeas(I)		0.968	0.664	21.022	1.045
Rmeas(I+/-)		0.971	0.665	21.231	1.048
Rpim(I)		0.094	0.052	3.333	0.112
Rpim(I+/-)		0.127	0.068	4.640	0.152
Rsplit(I)		0.077	0.046	1.253	0.195
CC half		0.996	0.995	0.316	0.959
Wilson B factor		9.883			
Anomalous completeness			100.0	100.0	100.0	88.6
Anomalous multiplicity			50.9	103.2	20.6	33.8
Anomalous correlation			-0.017	-0.089	-0.031	-0.004
Anomalous slope		1.515			
dF/F		0.090			
dI/s(dI)		1.572			
Total observations			1558766	158714	31224	1710375
Total unique			16566	934	799	27793

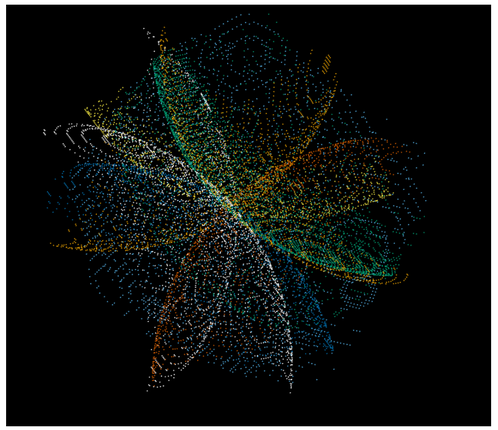
Dr James Beilsten-Edmands, 21/11/24

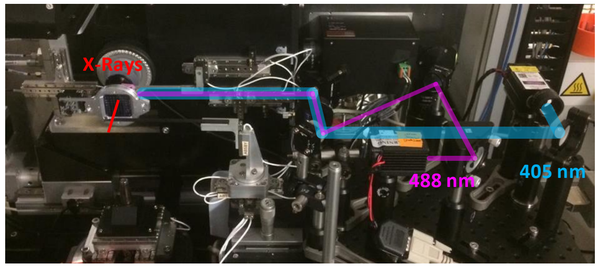
Serial crystallography

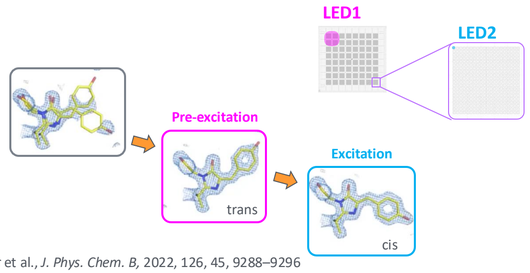
Serial still-shot crystallography is the measurement of 1 to N(<~10) diffraction patterns from given crystal, with no rotation. A complete, high-resolution dataset requires measurements of 1000s of crystals.

This technique is of particular importance for the study of dynamic biology at room temperature with minimal radiation damage, where processes on the timescale of sub-ms to seconds can be investigated at a synchrotron, or down to femtosecond timescales at an XFEL.

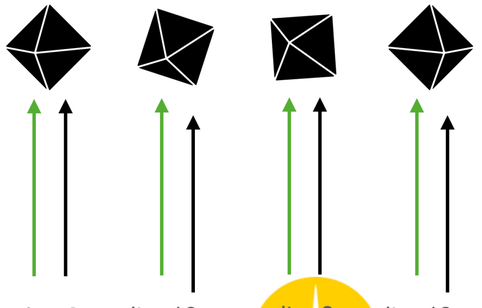
Typical experiments include the investigation of light-activated processes, chemical mixing/binding dynamics, studying of the effect of dose on metalloproteins.







Baxter et al., *J. Phys. Chem. B,* 2022, 126, 45, 9288–9296



Serial crystallography

Challenges from an automation perspective

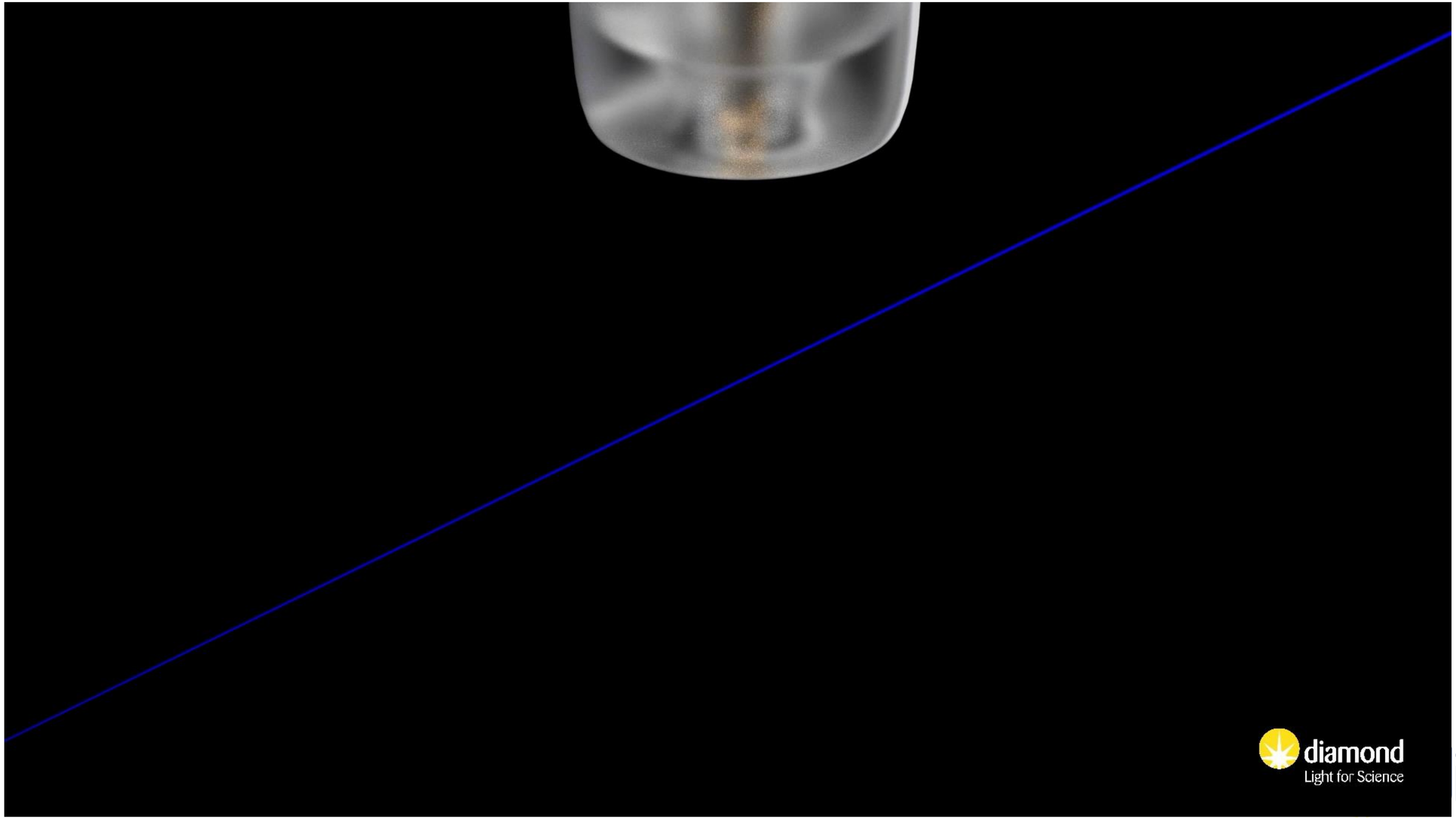
Speed

In theory you could collect new crystals at the detector collection frequency – 500Hz for Eiger, >kHz for newer detectors. Challenge is then the automation of the sample delivery process to achieve high crystal refresh rate. Following on from that, fast data processing and feedback are required to guide the experiment and make best use of available resources.

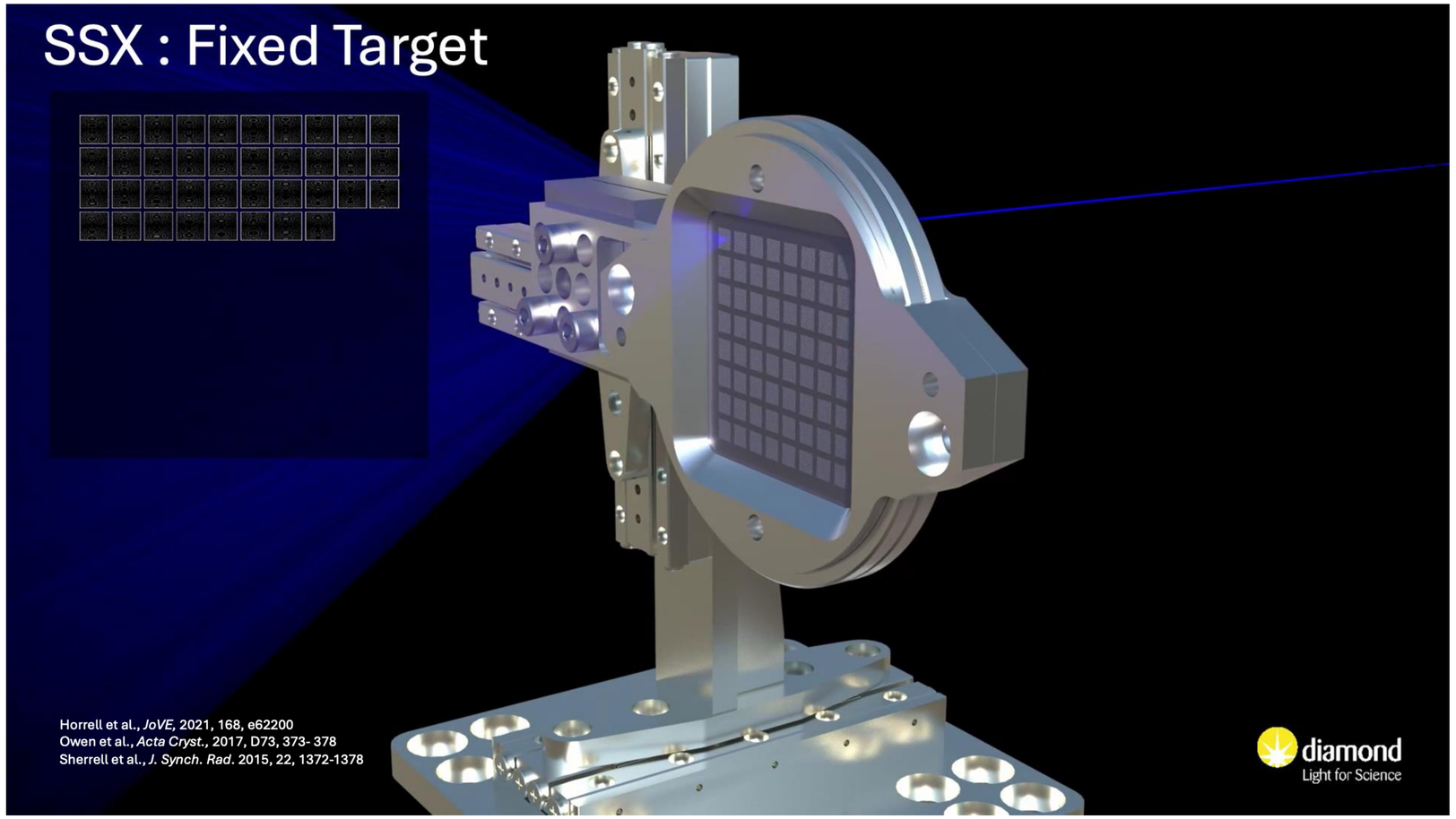
Complexity

Inherently more complex experiment type compared to rotation crystallography, different hardware options depending on biological question of interest, this extra metadata needs to be handled in processing and display of results.

Serial crystallography: extruder



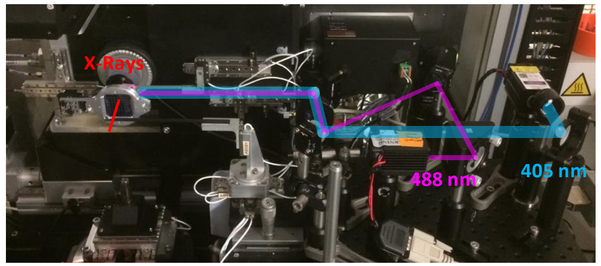
Serial crystallography: Fixed target

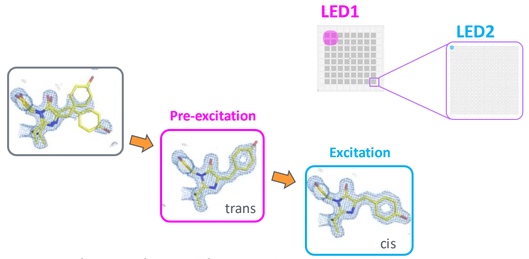


Chip 30 x 30 mm
8 x 8 city blocks
400 apertures on each city block
Aperture size 7 – 20 µm
25600 positions
~ 8 min full data collection

diamond

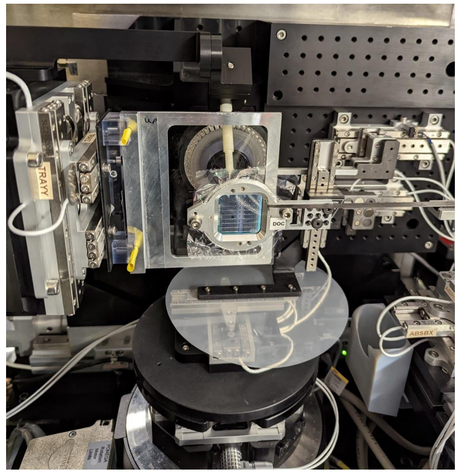
Serial crystallography: add-ons

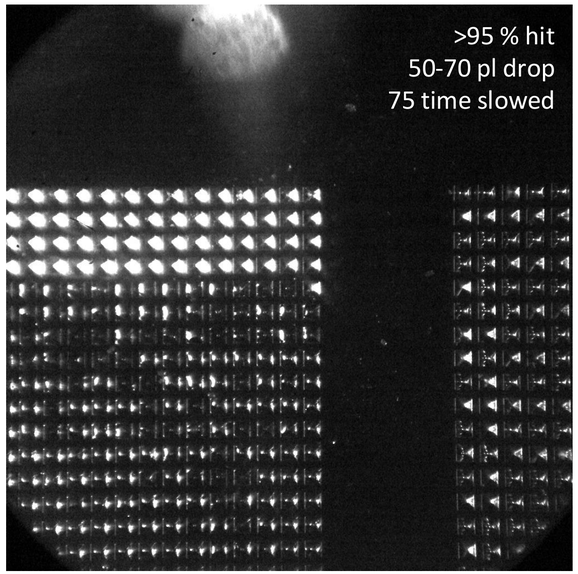




Baxter et al., *J. Phys. Chem. B,* 2022, 126, 45, 9288–9296

Laser excitation.

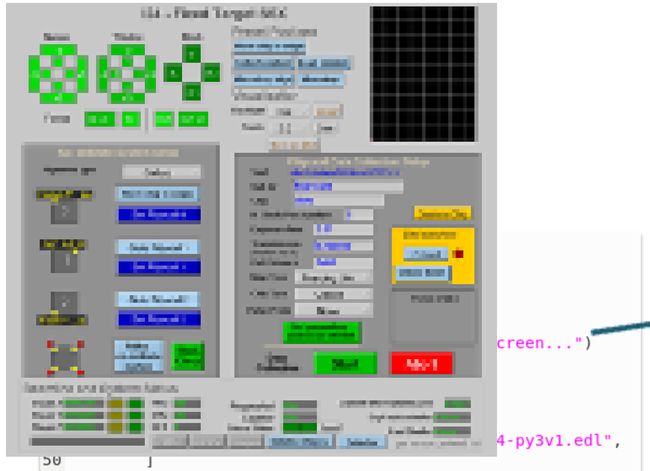




Rapid mixing with micro-drop ejector.

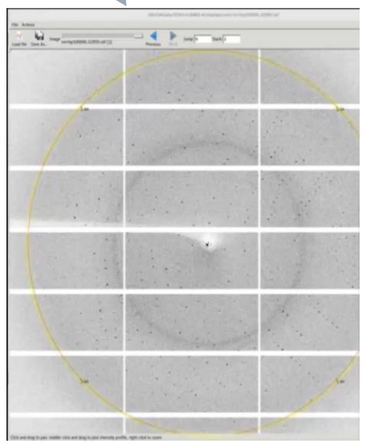
diamond

Automated data collection: I24 experiment control



```
creen...")

4-py3v1.edl",
50          ]
51      )
52      logger.info("Edm screen closed.")
```

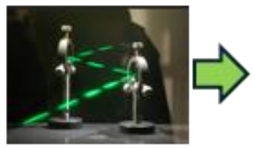


Motion controller

X ray shutter

Trigger control (Zebra)

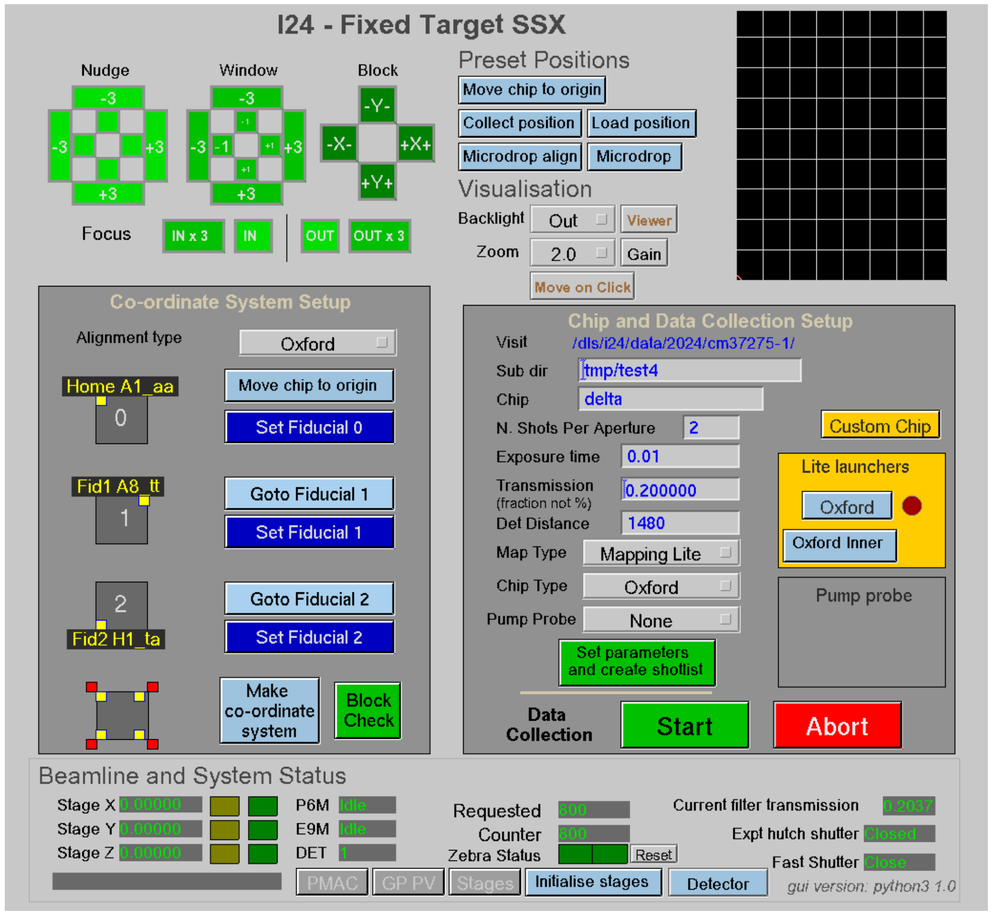
Micro-pipette

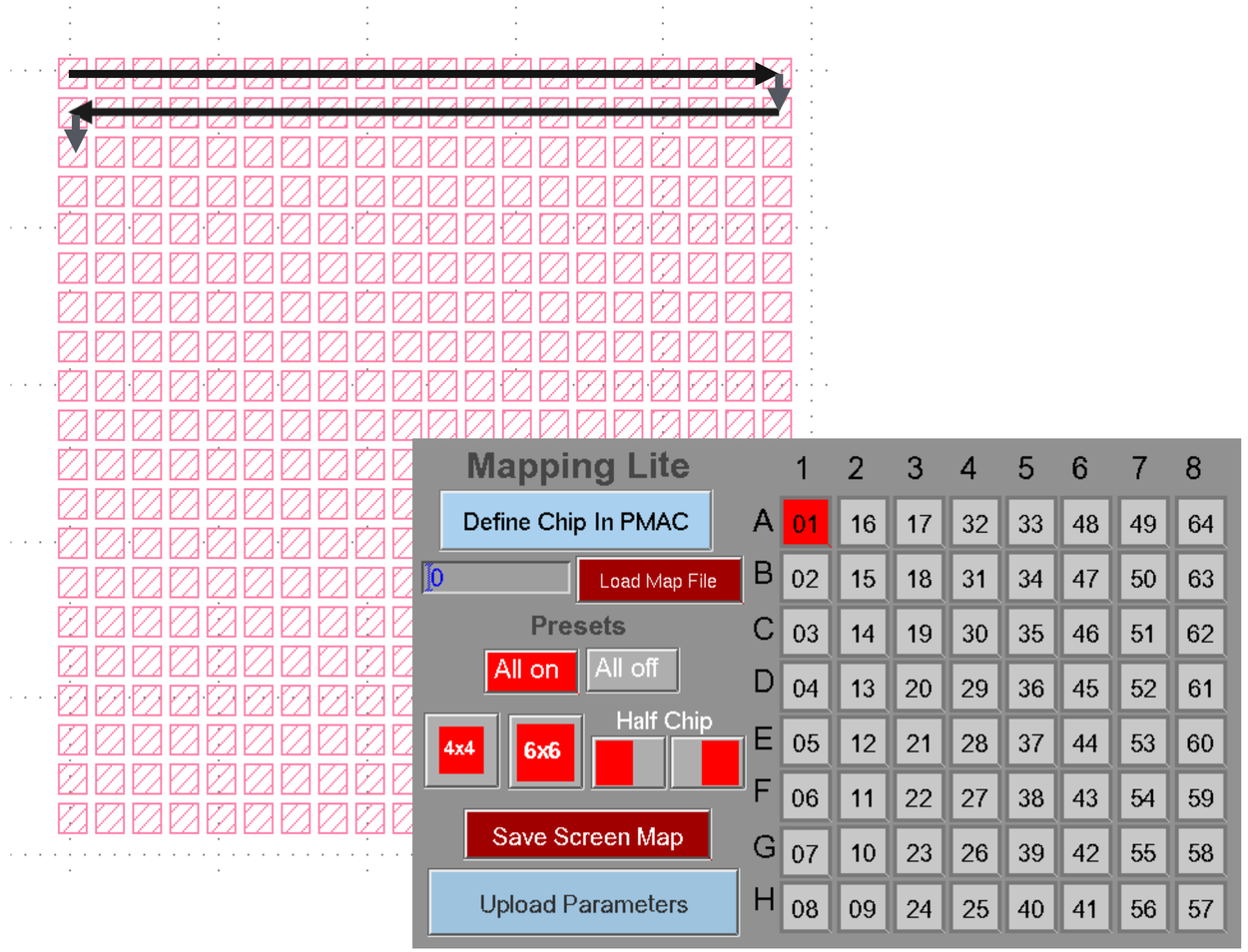


Laser

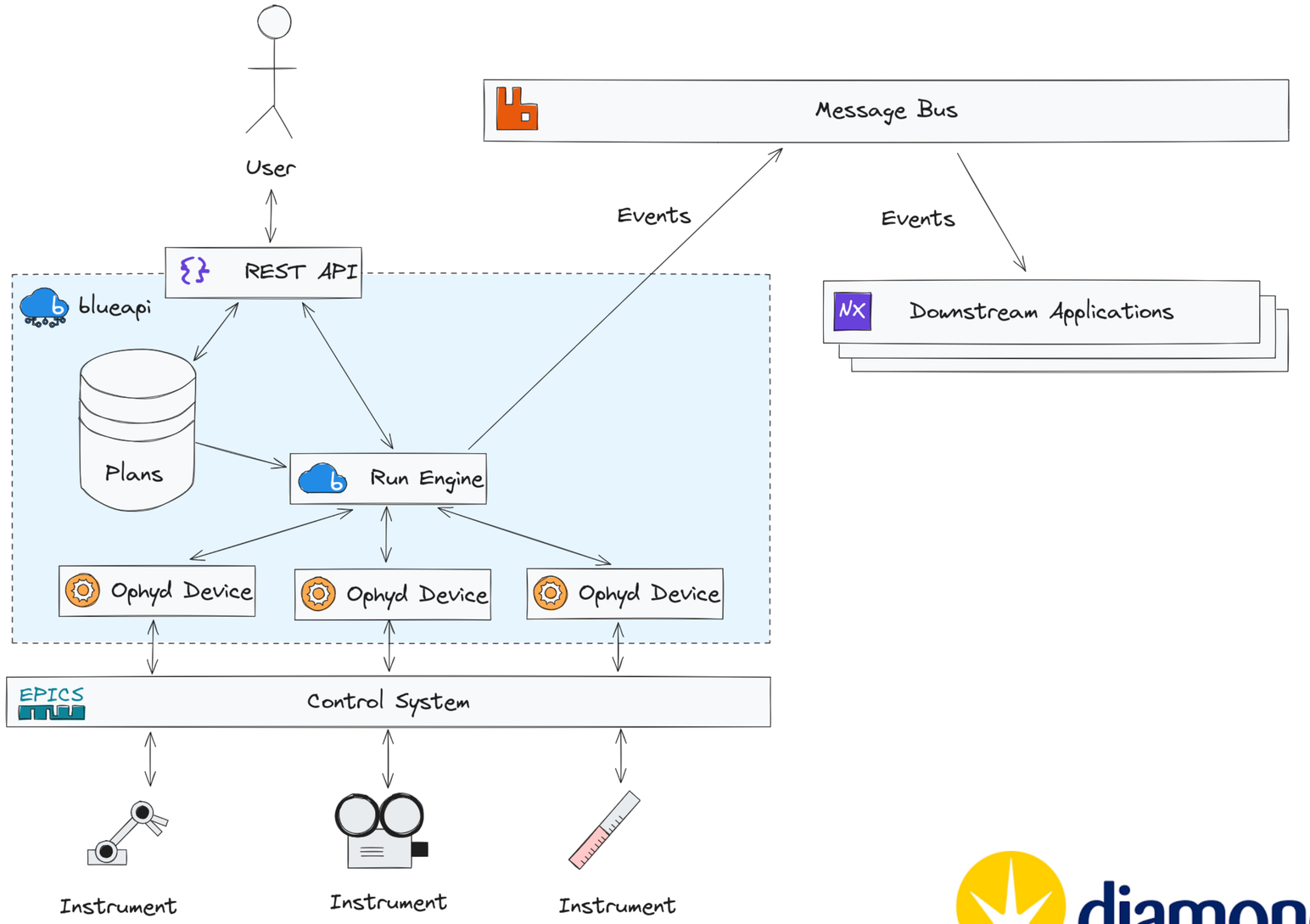
Automated data collection: I24 experiment control

Fixed target GUI





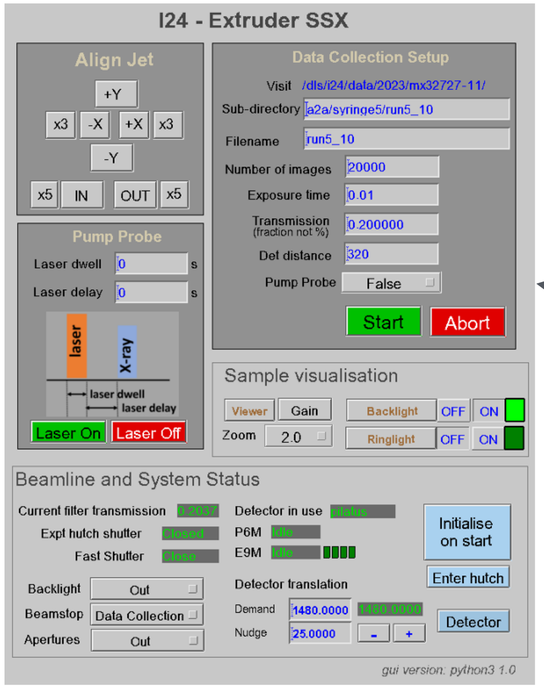
Automated data collection on I24: Bluesky

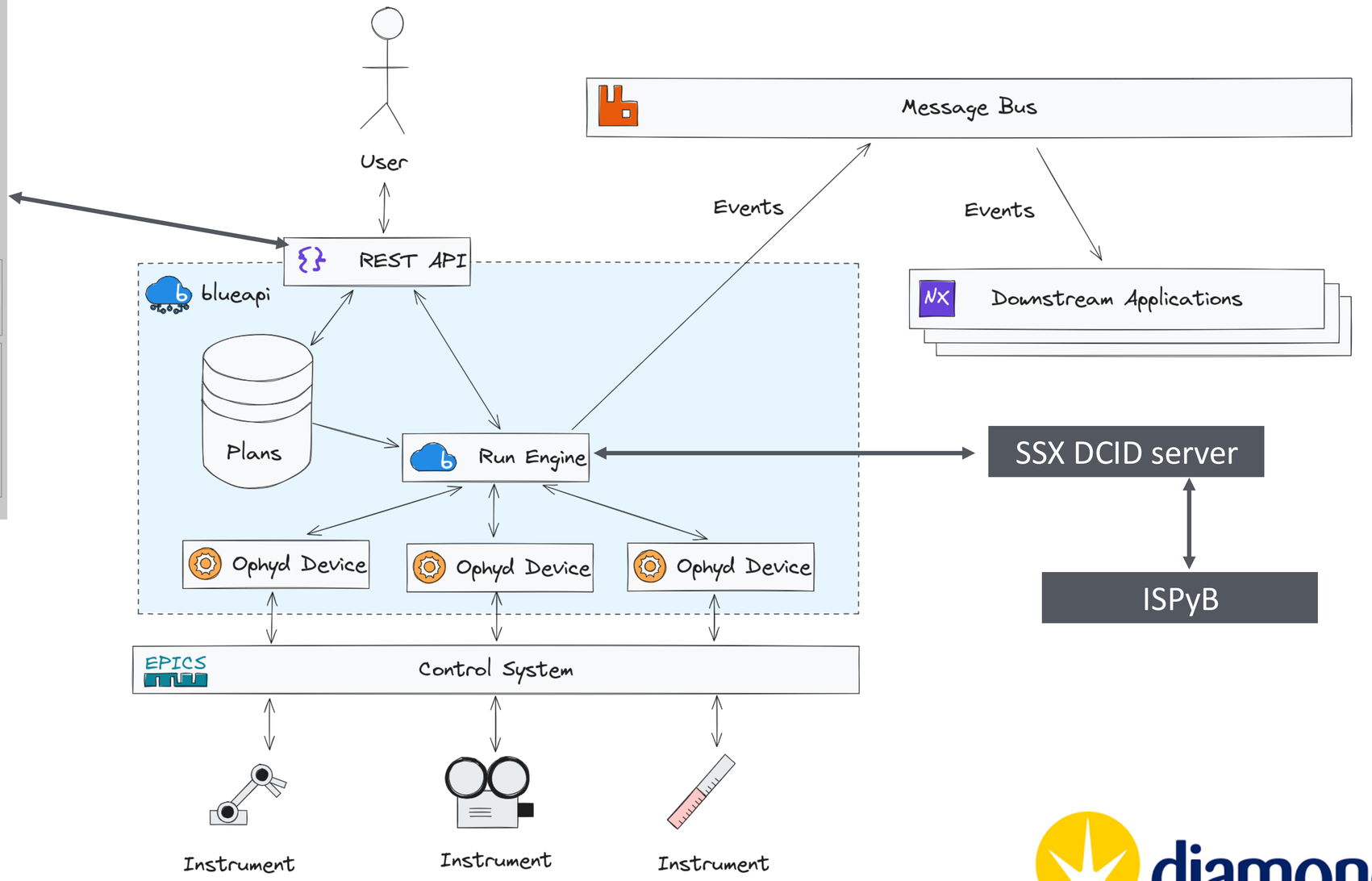


github.com/DiamondLightSource/blueapi

diamond

Automated data collection on I24: Bluesky





github.com/DiamondLightSource/blueapi

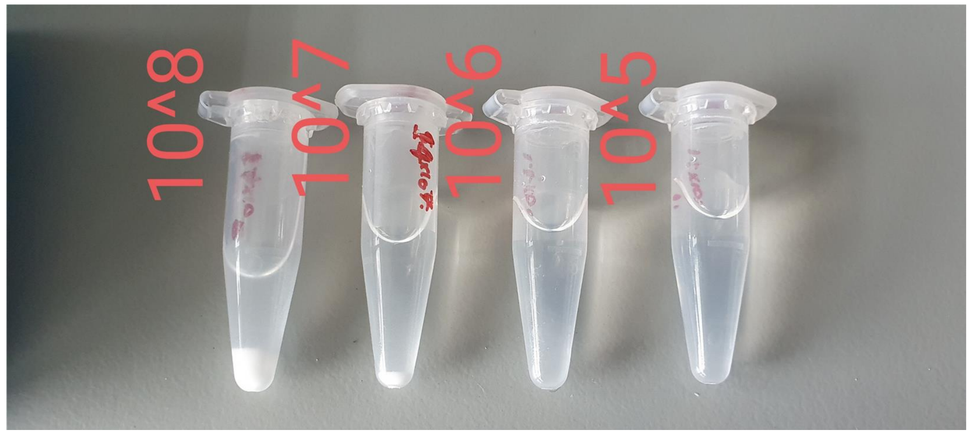
diamond

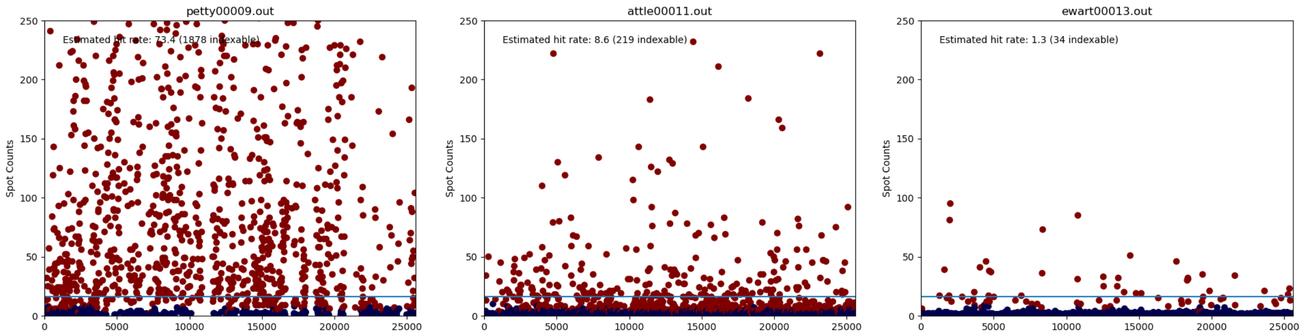
Automated data processing

Fast feedback essential for understanding success of sample preparation and data quality.

Hit rate -> indexing success rate -> data quality statistics.

Sam Horrell





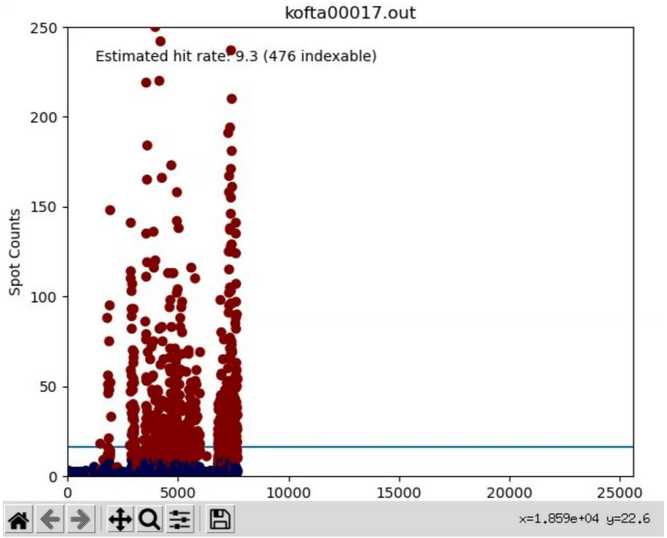
75% 9% 1%

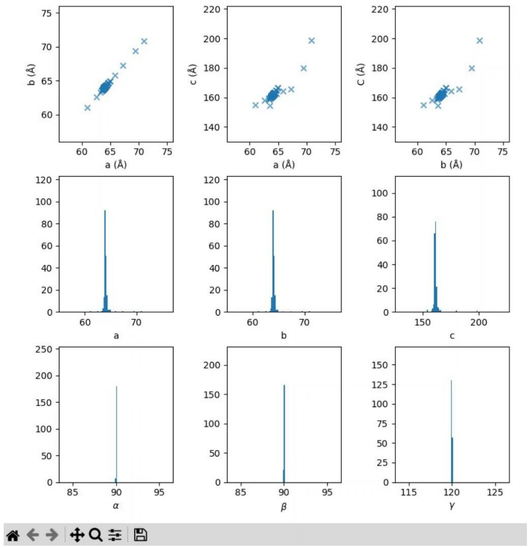
Didn't bother to collect it

Automated data processing: fast feedback

Live feedback from DIALS processing – dials.find_spots, dials.index

dials.github.io





Per-image spot-finding, hit rate estimate

Per-hit indexing – cell parameter distributions

diamond

Automated data processing: infrastructure

Analysis infrastructure at Diamond

Data analysis managed by zocalo software. github.com/DiamondLightSource/python-zocalo

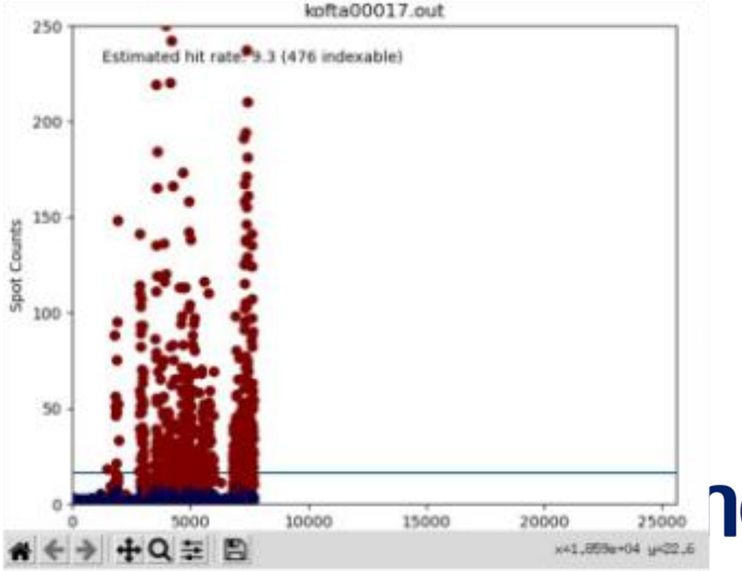
Zocalo defines a set of services (run on Kubernetes cluster at DLS) – these run continuously and wait for work – via messages managed by RabbitMQ server.

Zocalo also defines wrappers – run processing commands on demand (e.g. xia2 processing job).

Finally have a set of recipes which define workflows based on these services/wrappers as a flowchart/graph.

Recipe for live hit-rate analysis

1. Filewatcher service watches for files appearing on disk in collection directory
2. Per-Image-Analysis service
3. Hit rate plotter service.



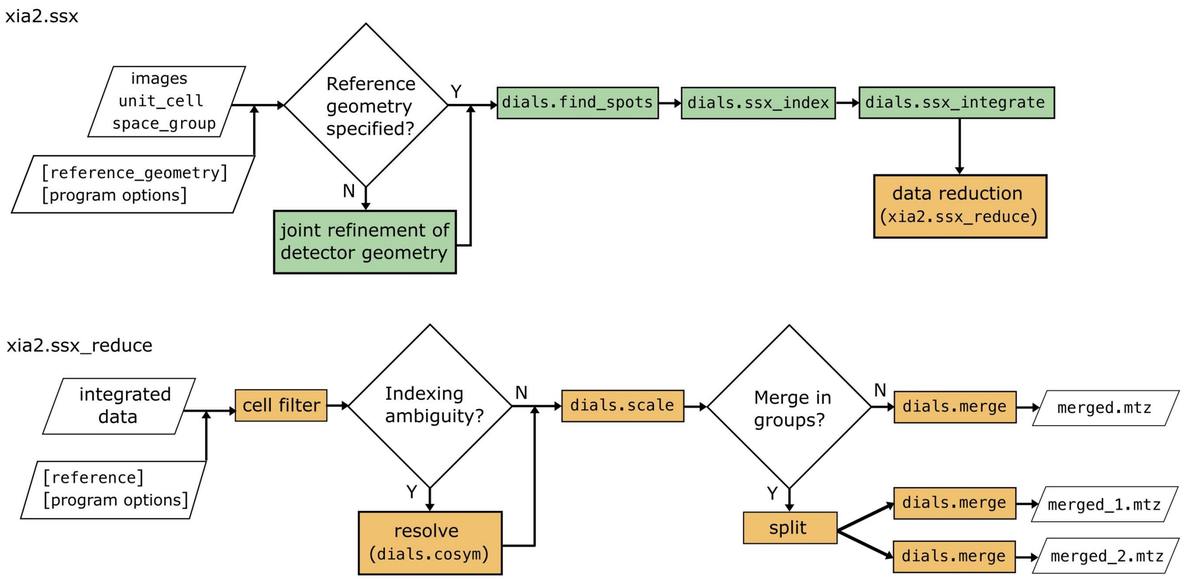
Automated data processing: full processing

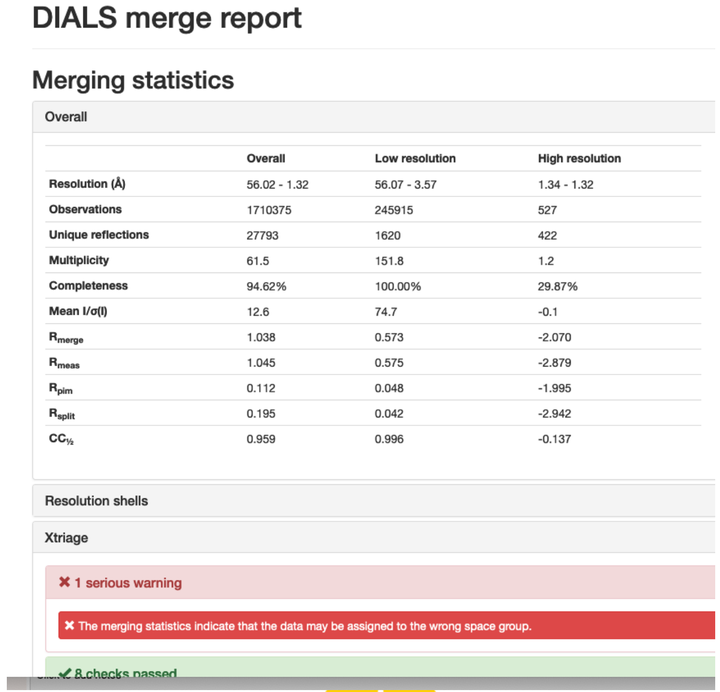
Full post-processing with xia2.ssx triggered automatically after data collection.

Relies on a small configuration (.yaml) file detailing expected unit cell and space group.

xia2.ssx – processes images to merged MTZ running DIALS programs.

xia2.ssx_reduce – reduces integrated together from multiple collections.





DIALS merge report

Merging statistics

Overall

	Overall	Low resolution	High resolution
Resolution (Å)	56.02 - 1.32	56.07 - 3.57	1.34 - 1.32
Observations	1710375	245915	527
Unique reflections	27793	1620	422
Multiplicity	61.5	151.8	1.2
Completeness	94.62%	100.00%	29.87%
Mean I/σ(I)	12.6	74.7	-0.1
R_{merge}	1.038	0.573	-2.070
R_{meas}	1.045	0.575	-2.879
R_{pim}	0.112	0.048	-1.995
R_{split}	0.195	0.042	-2.942
$CC_{½}$	0.959	0.996	-0.137

Resolution shells

Xtriage

✖ 1 serious warning

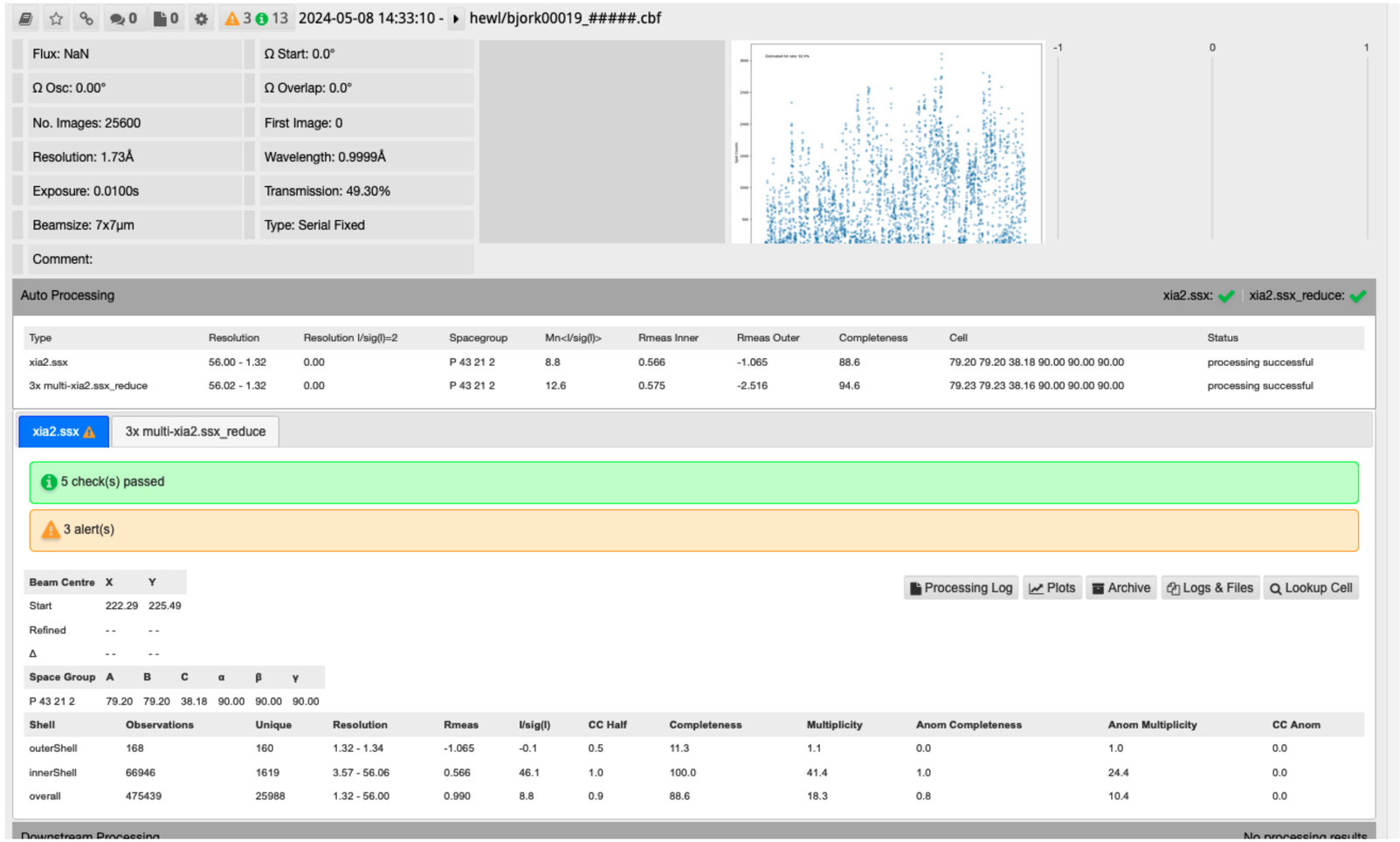
✖ The merging statistics indicate that the data may be assigned to the wrong space group.

✔ 8 checks passed

J. Beilsten-Edmands et al., Methods in Enzymology 709, (2024).

Automated data processing: full processing

Results displayed in SynchWeb on a per-chip basis. Multiple chips of the same condition merged together automatically. Access to log files, html report and merged MTZ file.



Timescale ~15-30 mins

Automated serial data processing: future directions

- Acceleration of live processing through migration to GPU devices.
- Extension of live processing to cover the full processing, enabling complete metrics on fast timescale.
- Completion of metadata models to cover all serial experiment types (dose/time series).
- Web-based GUI for live display of results
- Migration from reliance on file systems to stream-based workflows.

Acknowledgements

Data acquisition team, in particular Noemi Frisina, David Perl, Dominic Oram

I24 Beamline team – Robin Owen, Danny Axford, Sofia Jaho, Do-Heon Gu + many others over last 10+ years

Data Analysis/DIALS team, in particular Nick Devenish, James Parkhurst, Graeme Winter

Grant funding from Wellcome trust, CCP4/UKRI, EU grant funding.

github.com/DiamondLightSource/blueapi

nsls-ii.github.io/bluesky/

dials.github.io. xia2.github.io

github.com/DiamondLightSource/python-zocalo

Slide material: Danny Axford, Robin Owen, Sam Horrell