National Synchrotron Light Source II

Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

The future of advanced automation at the NSLS-II AMX beamline

Jean Jakoncic (jjakoncic@bnl.gov)

& Dale Kreitler, Shekar Venkateswaran, Edwin Lazo

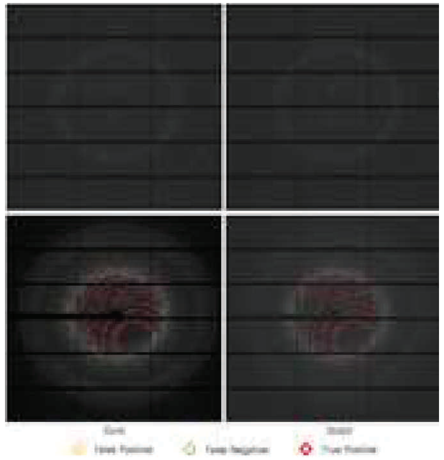
November 21st, Trieste, Italy.

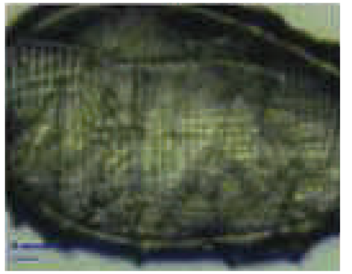
NIH NIGMS

BER Structural Biology Resources

Outline

- NSLS-II AMX & FMX beamlines
- Current Access Modes
- “Current” state of Automated Data collection
- Motivations
- New Access Mode
- Work in Progress and impact of AI / ML tools
- Planned development using AI / ML
- NSLS-II implementation of pilot Fragment Screening
- New beamline proposal & NSLS-IIU
- Conclusions





What It Is That We Do ?

2 independent MX beamlines (NIH, NIH-NIGMS/DOE-BER): in General User operation since 2017-01

Support most challenging structural biology projects: small membrane protein Xtals in opaque medium / very large complexes …

Each beamline sees up to 3 (5) grou

Remote / Automated / **Local / Stand**

Raster factory (find crystal(s), region

Rotation (standard collection)

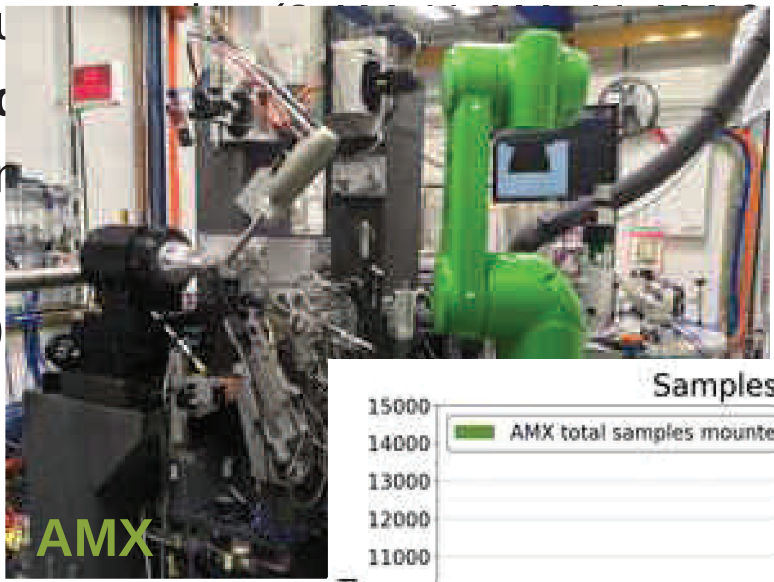
Vector (dose along linear trajectory)

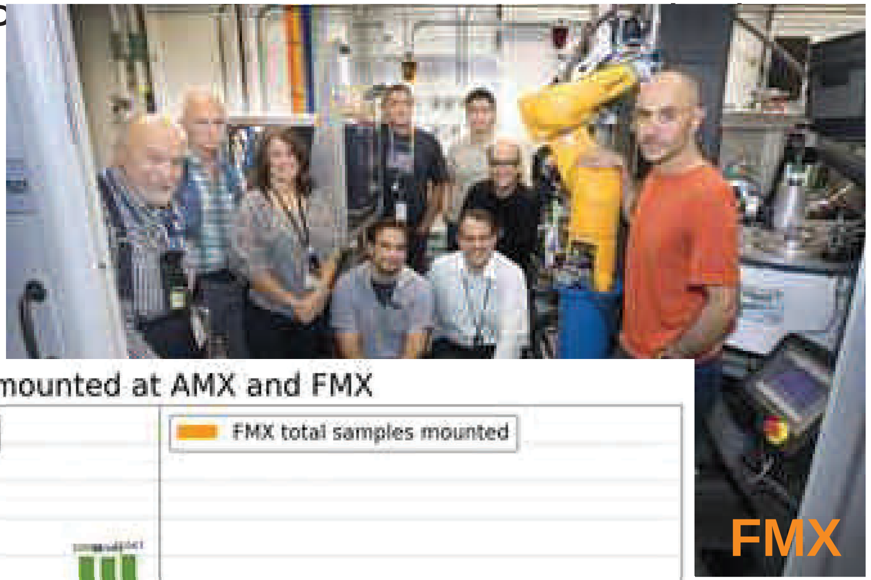
Multiple (small crystals in mesh)

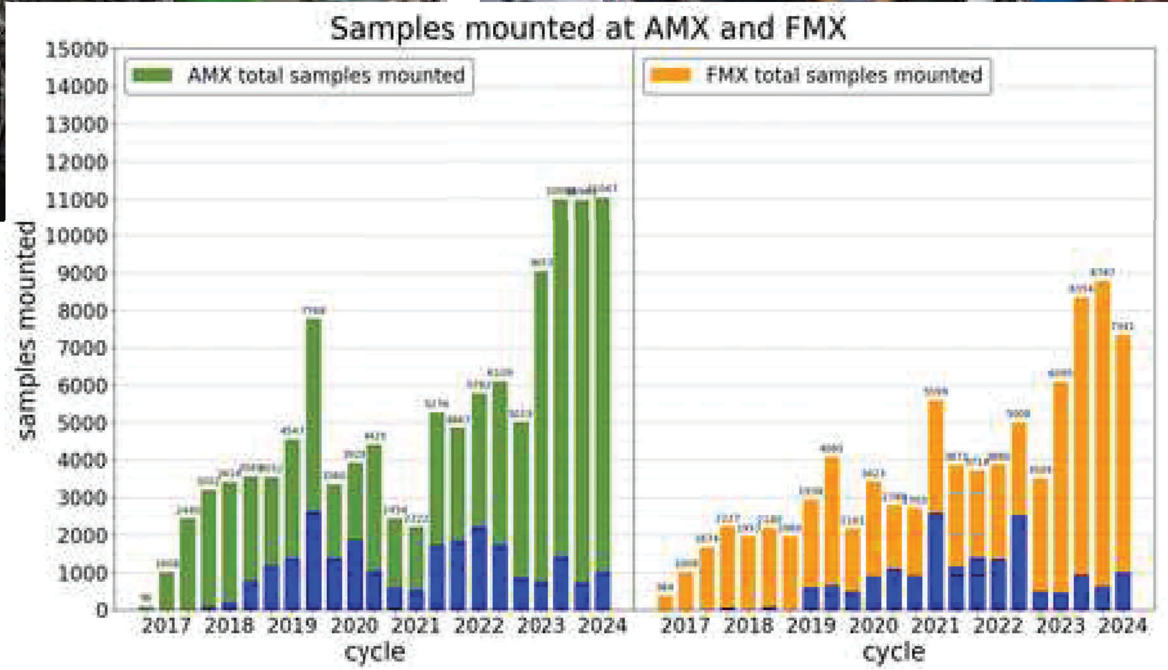
Serial (loop / mesh)

Screening protocols

GU / BAG / Rapid Access / Proprietary





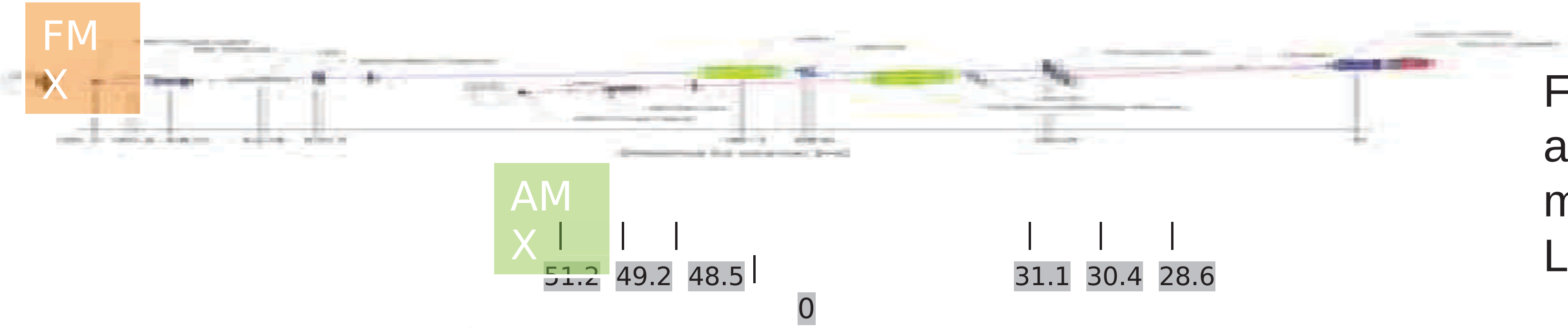


<u>Capacity:</u>

24 **Unipucks**
384 **<u>Spine</u>** samples

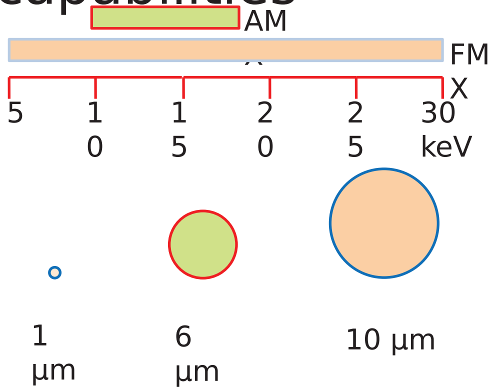
3

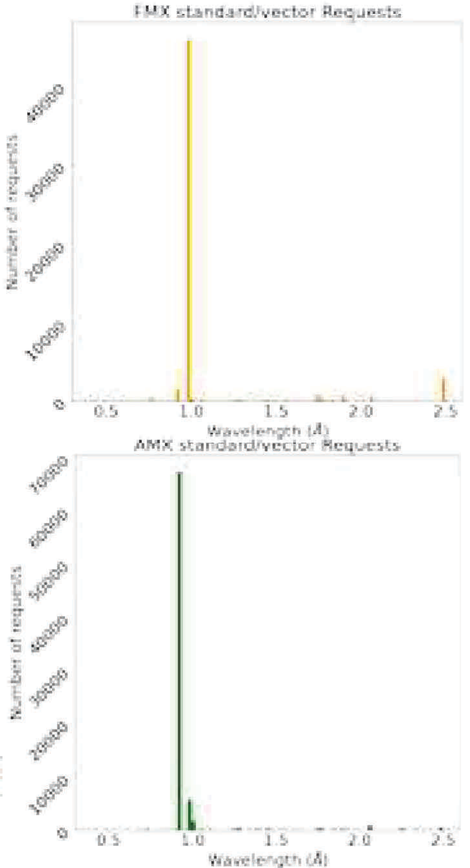
AMX & FMX Beamlines Characteristics



FMX:
ambient T time resolve
metals
Low E phasing (He)

Two independent beamlines
with overlapping and complementary capabilities





Specifications	AMX	FMX
Energy range	9.5 – 18 keV	5 – 30 keV
Wavelength range	0.7 – 2.5 Å	0.4 – 2.5 Å
Flux at focus at 12.7 keV	5.4×10^{12} ph/s	3.5×10^{12} ph/s
Focal spot min (H×V)	7 × 5 µm^2	1.5 × 1 µm^2
Focal spot range	NA	1 – 20 µm
Detector FPS Data Format	Eiger X 9M [illegible]	Eiger X 16M ≤ 133 Hz [illegible]

AMX “native” energy: Br absorption K edge 13.475 keV.
FMX “native” energy: Se absorption K edge 12.658 keV.

Current Access Modes and APS-U impact

All proposals (PRP reviewed) with allocated time scheduled with a minimum of 0.5 shift (mini-shift).

Groups (BAG, GU, RA, PR, DT): access time, mostly with remote access these days.

2023-2024

User time
75 -> 85 %
APS_U

To accommodate multiple visits from these many groups ”standby” access was introduced in 2023_3.

In standby: users fill in an automated form and ship up to 7 pucks

Staff load and collect data using automated workflow when time become available.

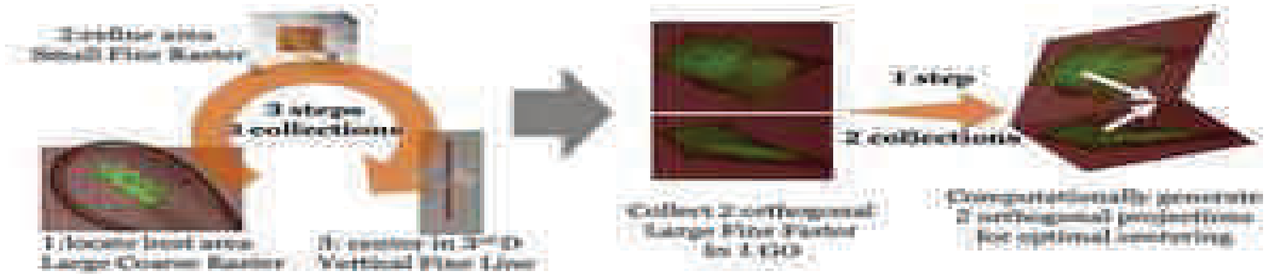
Users finishes earlier, beamtime cancellation.

Users access data instantly with Globus.

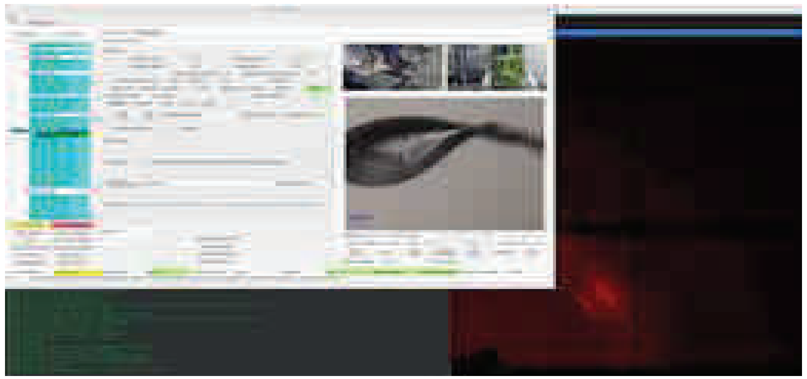
Staff / Users communication: Slack

Current Automated Data Collection

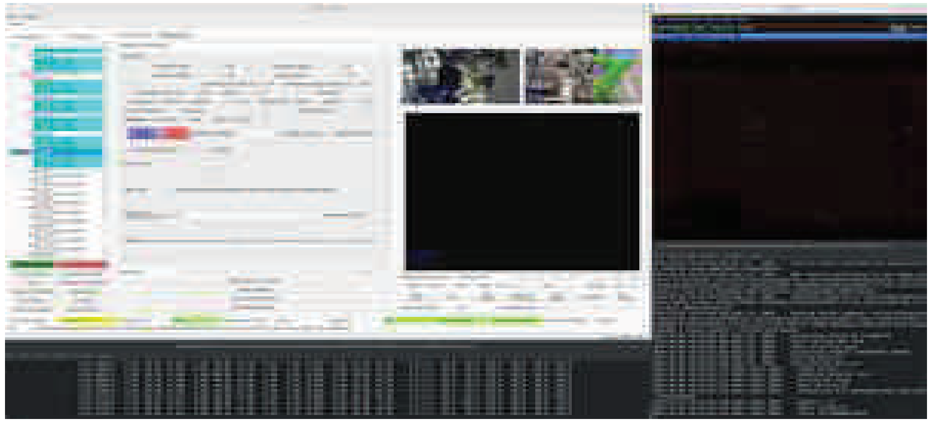
2 protocols tested and offered to user groups: Auto Collect and rasterScreen
crystal centering (~170 sec): **24 samples per hour (including data collection): samples 20 µm and more**



From XREC to New loop centering model (Dale & Shekar)



Movie recorded in 07/2022 : ~16 samples/H



National Synchrotron Light Source II

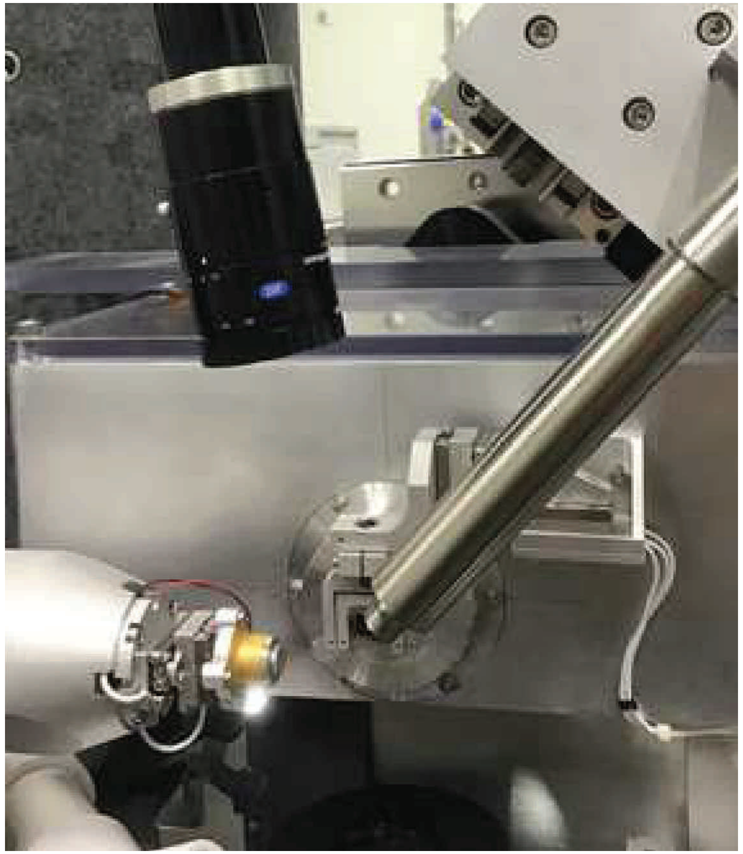
Movie recorded in 03/2024: 24 samples / H

Future improvements:
Better spot finder (weak reflections)
Better raster scoring (new scoring function: WIP)
To: auto vector / Sorting / Strategy / multiple Collections

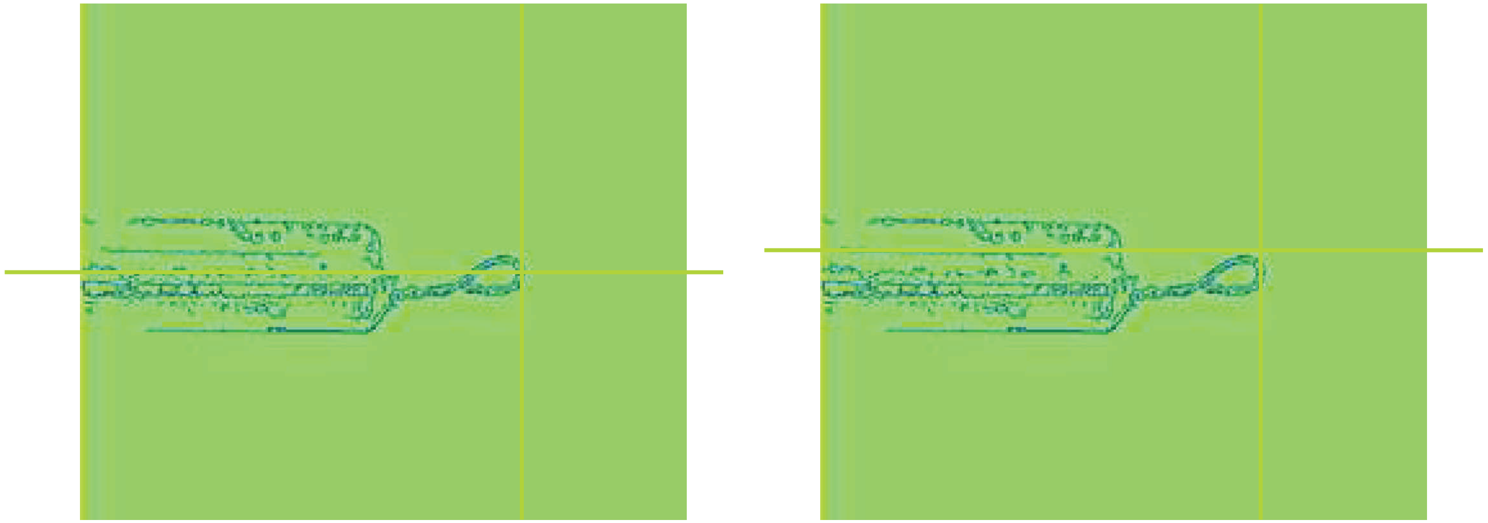
Including automated decision-making steps

Note that the 2 rasters are now done by default after each sample mount for manual operation. *(0 hkls > 2nd skipped)*

Off-Axis Camera Alignment

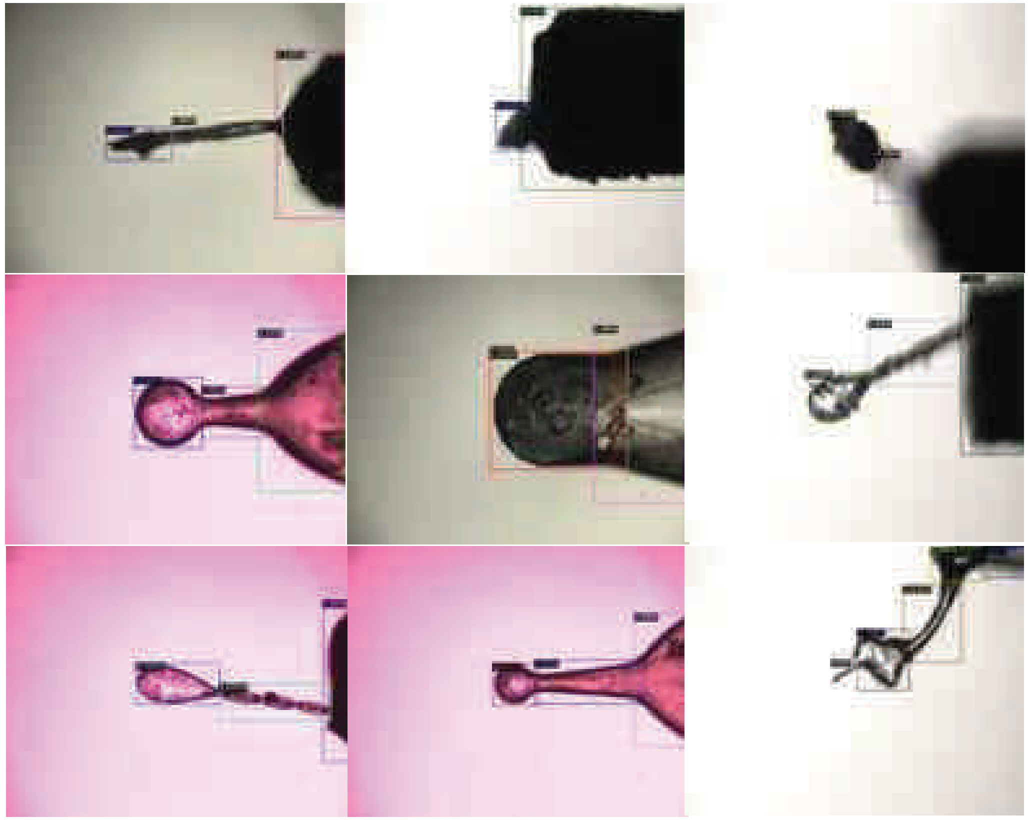


- First step in alignment after robot mounting
- Motion and acquisition during state transition (SE – SA): 2 frames
- Necessary to bring sample into focal plane of on-axis microscope (FOV: 1.3 mm for the low mag and 0.35 mm for the high mag)
- Canny edge detection: classic computer vision (image processing ~1 ms)
- Current image resolution: 7 µm/pixel



Dale Kreitler: dkreitler@bnl.gov

“Two click” Centering with Loop Detection Model



~1 µm / pixel

Detectron 2 trained with AMX microscope images (diverse set)

Centering sequence:

1. “Face on” center moved to camera center
2. Rotate 90 deg, cluster pixel intensities and go to center of mass
3. These boxes define raster boundaries

~0.5 sec / image

Run twice with 20 images each.
HW GPU optimization underway

The 2 orthogonal rasters are drawn and executed for each sample mounted on AMX and FMX and will be used for the new crystal centering

Dale Kreitler: dkreitler@bnl.gov

Meta Open Source

https://ai.meta.com/tools/detectron2/

Other improvements

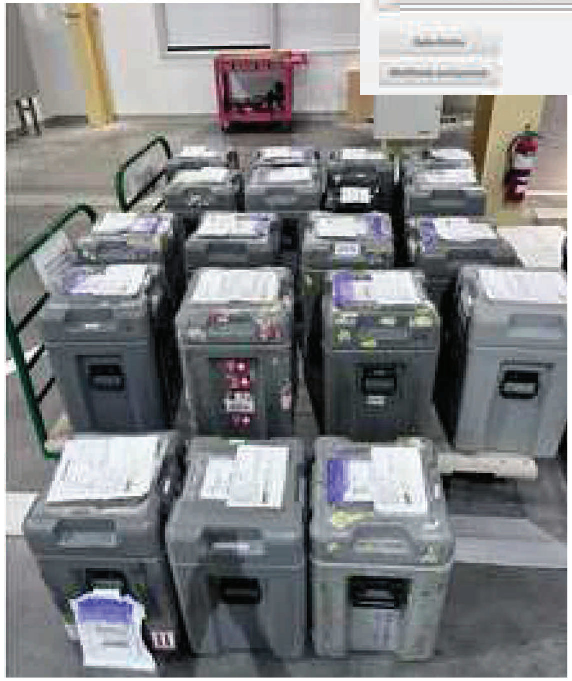
Puck etching with data matrix and readout at the beamline

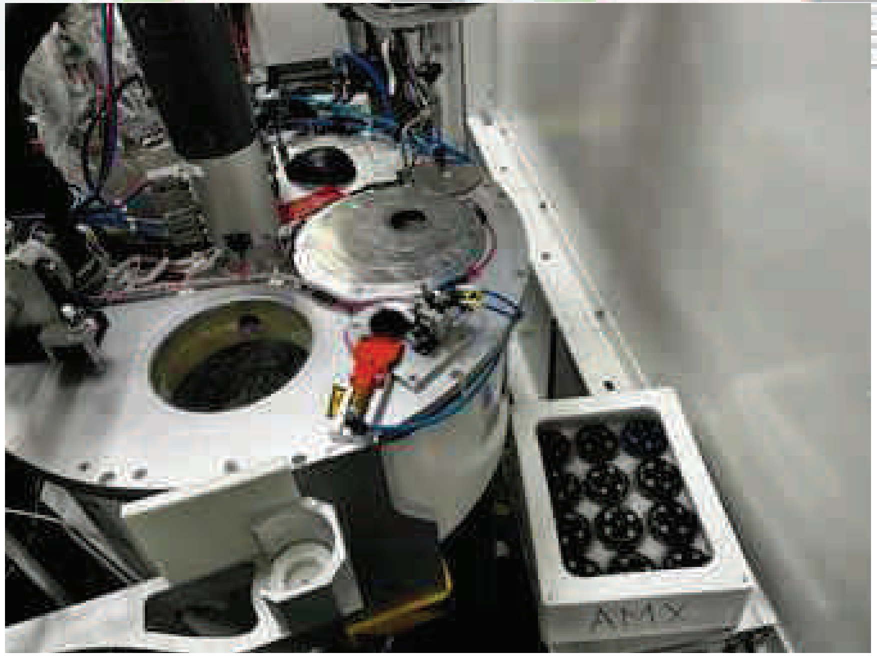
Puck pre-staging (rapid unloading)

New augmented spreadsheet (priority, workflow, dose, …)

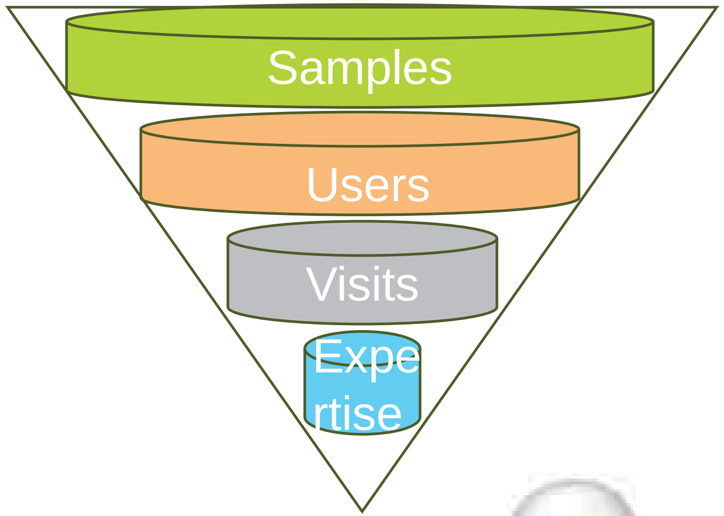
3.5 PB / >2000 Cores (NIH supplement APS_U)

Sentinel: automated recovery from recurring “bugs”, workflow





Motivations for new access mode



We are witnessing that more and more users are less and less experts

We also see that the number of samples per user group usually increases

Users love standby, and for most it delivers equal results.

Some APS users collect “sub-optimal” data (0.5 sec / 0.5 deg / 50 % !!)

We are approaching this :

- By increasing the achievable sample throughput
- By implementing better decision-making algorithms
- By implementing a new hybrid access mode : Automated Then Manual
- By maintaining dedicated training sessions: workbench / 1 on 1 training sessions

A significant fraction of these are from the APS-U, with the ALS-U in the horizon we anticipate the same demand for our MX beamlines.

Hybrid Mode: Automated Collection for All

Samples from groups with scheduled time one of a given day are pre-screened (auto collection or rasterScreen protocol) the night before.

Enhances user sample spreadsheet to be used as a stop gap (priority, collection workflow, dose, resolution …)

Users review their beamtime report containing all necessary information: we currently plan to generate a comprehensive pdf report with figures and sorter table; once reconfigured, users will access ISPyB/SynchWeb

Each of the 3 schedule groups will have the opportunity to manually collect data for up to 3 h during their schedule time.

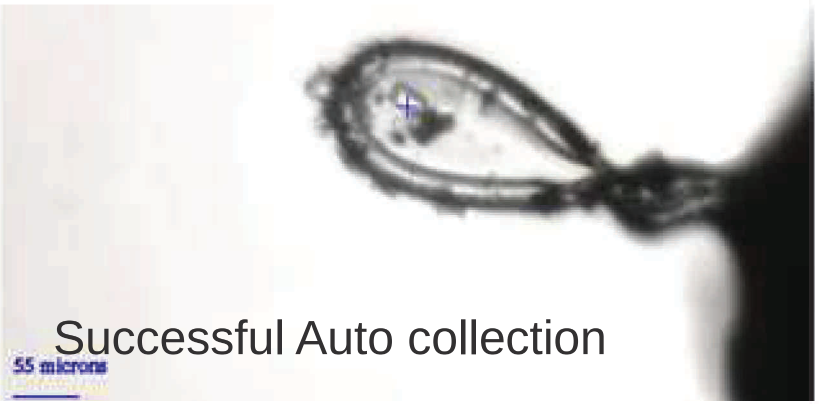
Users opting out for manual collection will free time to unload-pre-load pucks for additional automated collection.

At least 1 weekend day reserved for remote manual collections for groups requiring this access mode.

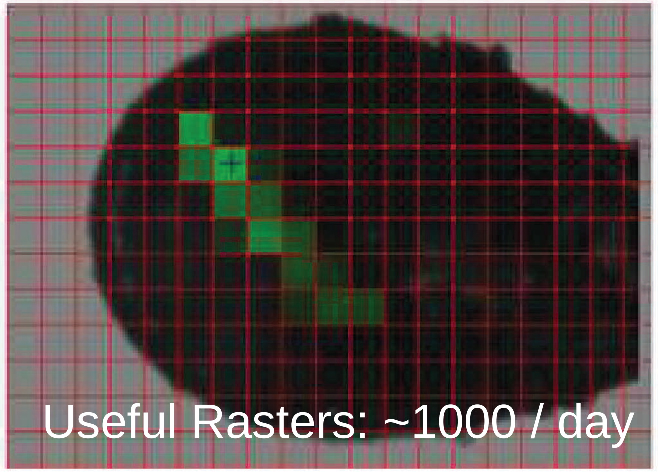
Special experiments, such as fragment-based drug discovery and ambient temperature time resolve will be pre-scheduled with 1 or 2 continuous days of uninterrupted beamtime.

Not Everything Can be Automated

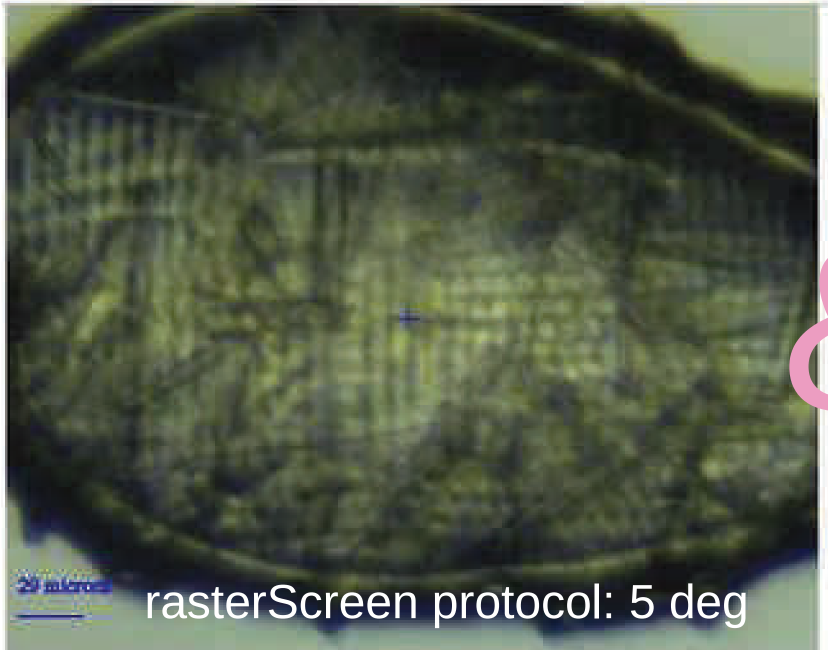
Successful Vector collection



Successful Auto collection

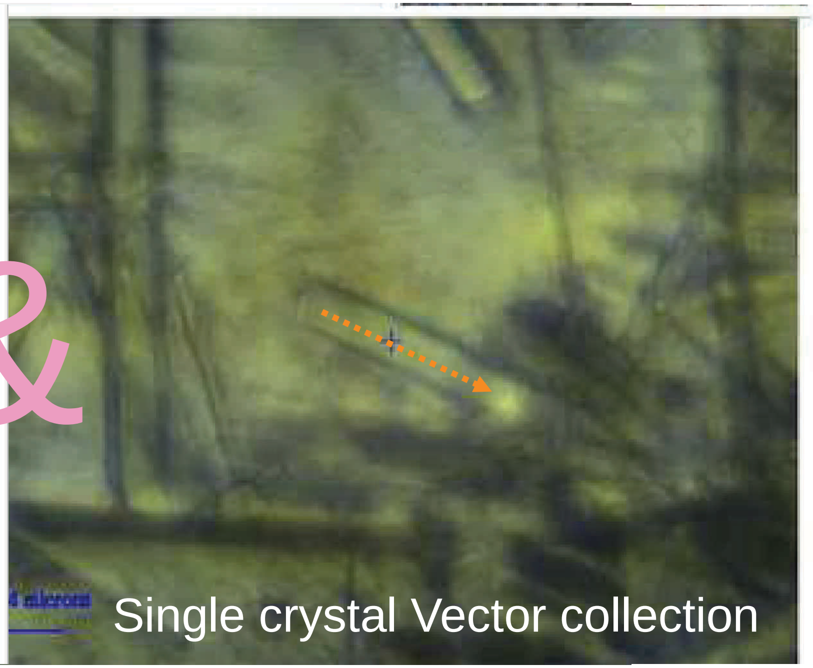

Useful Rasters: ~1000 / day

Good enough “fishing”

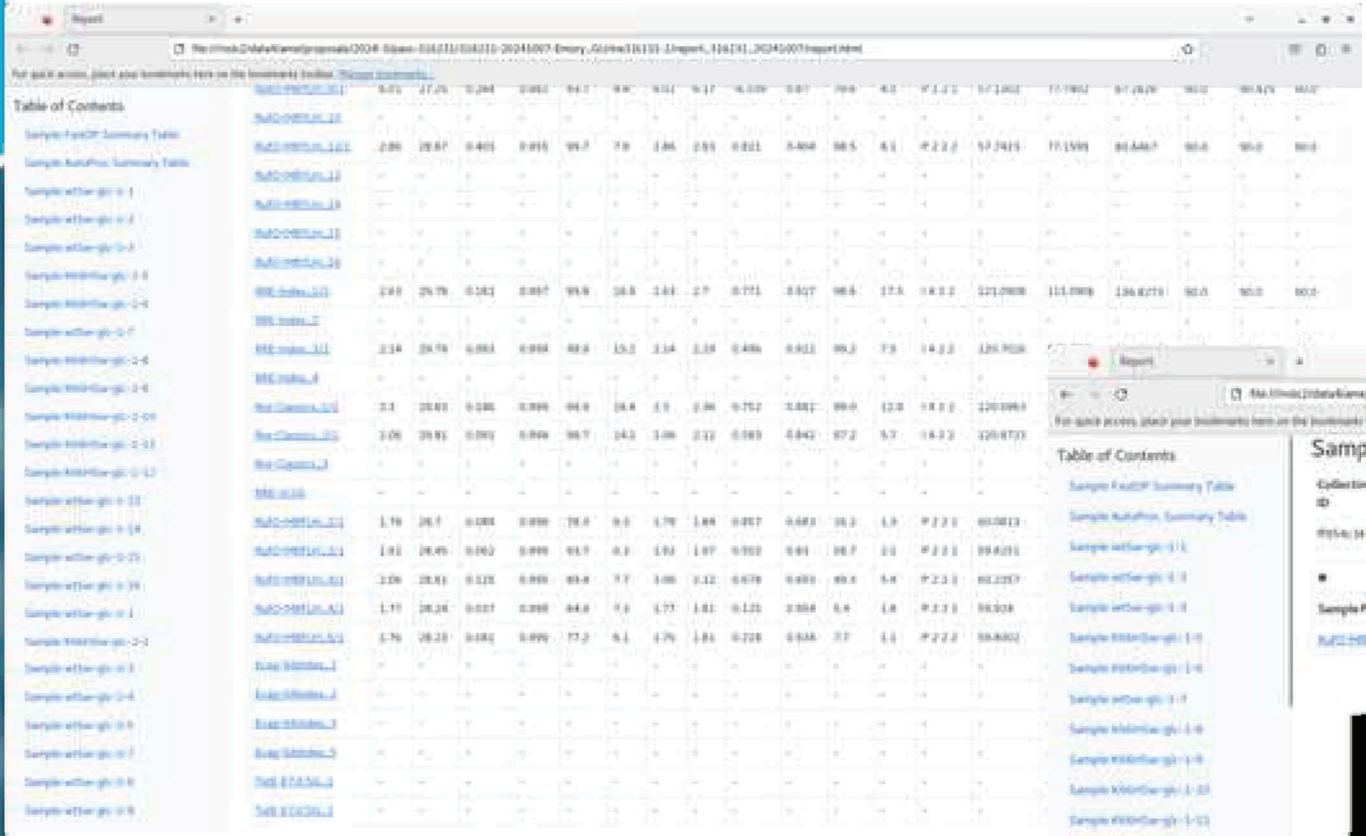


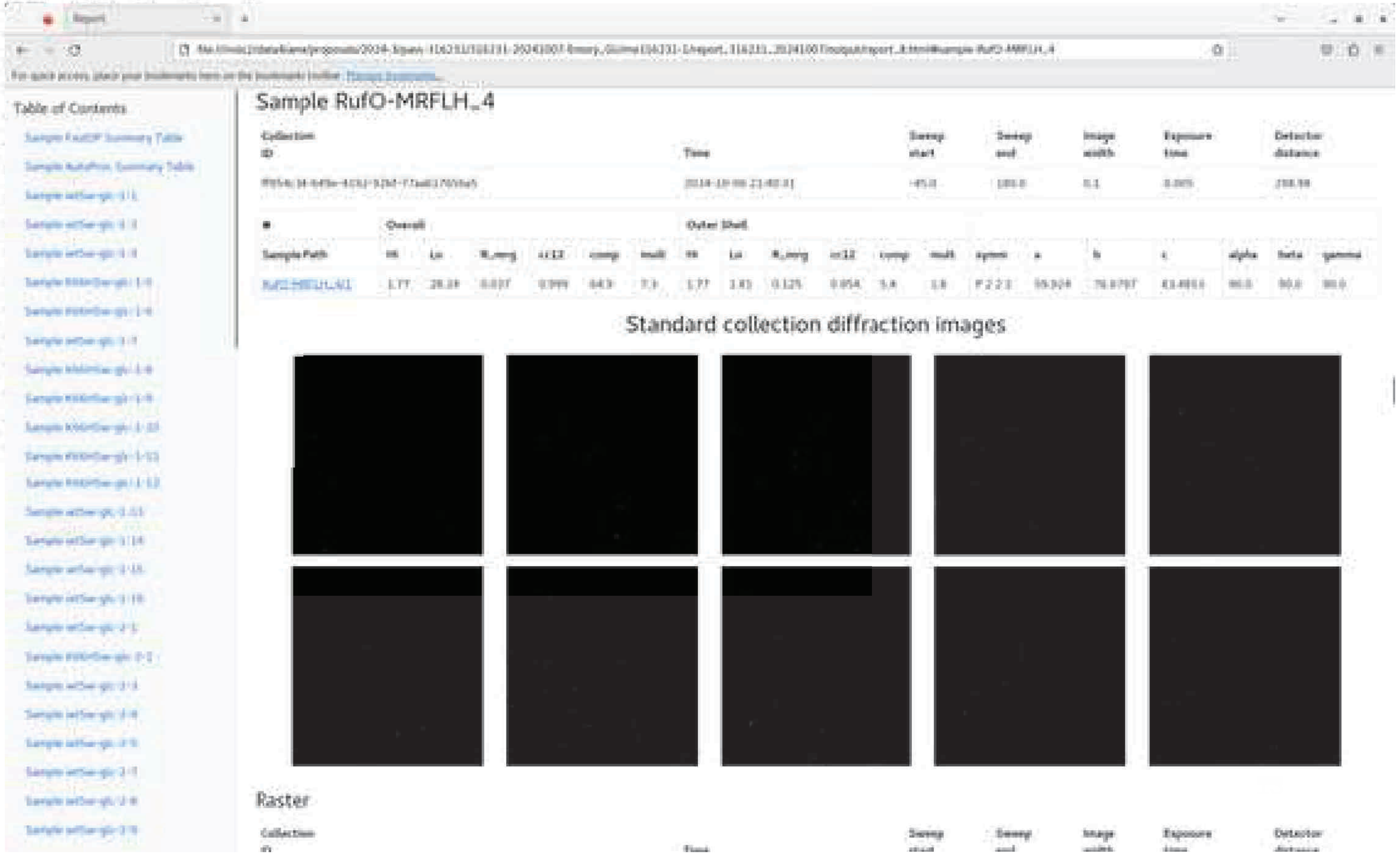
rasterScreen protocol: 5 deg

&


Single crystal Vector collection

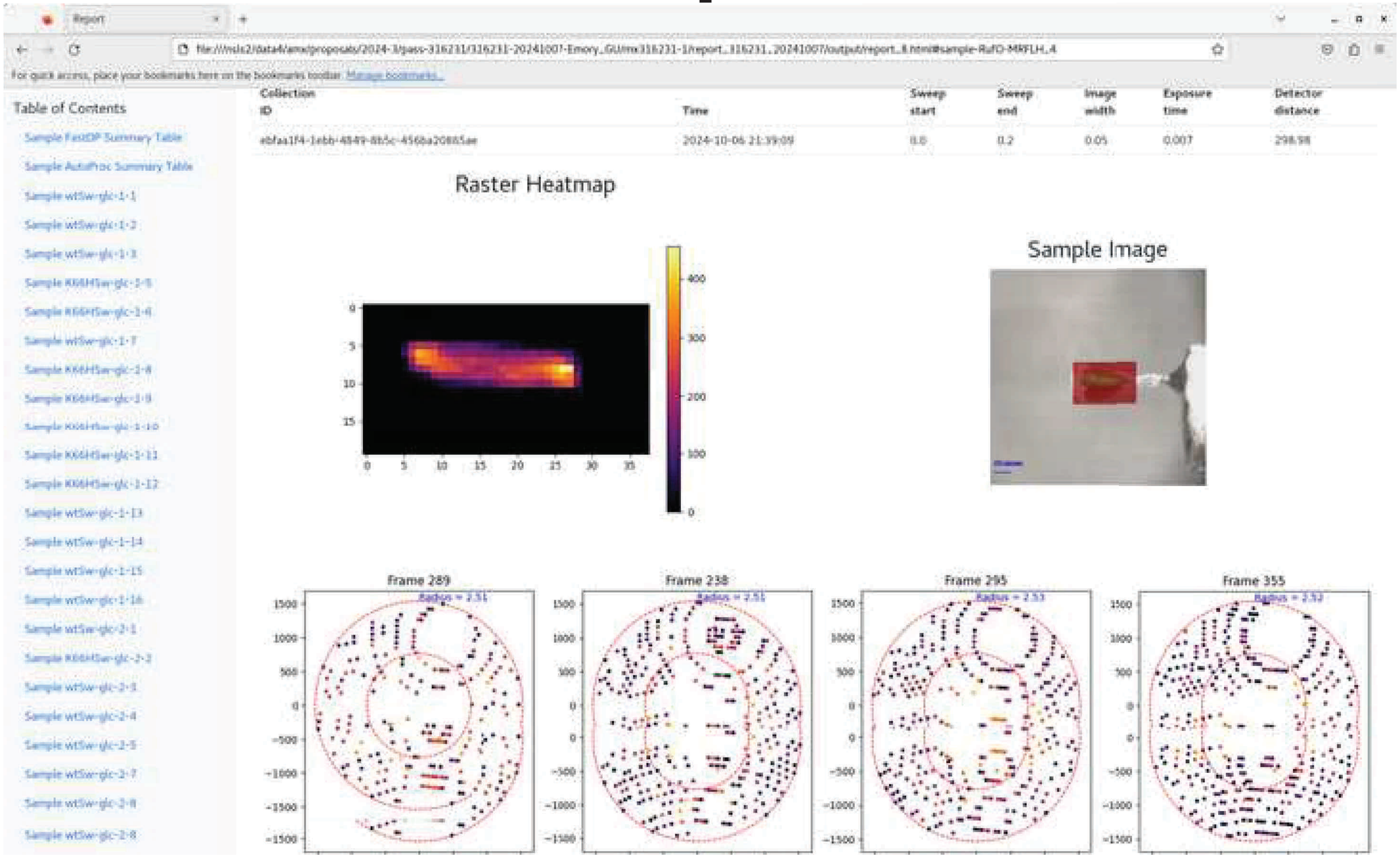
Hybrid Mode: the report



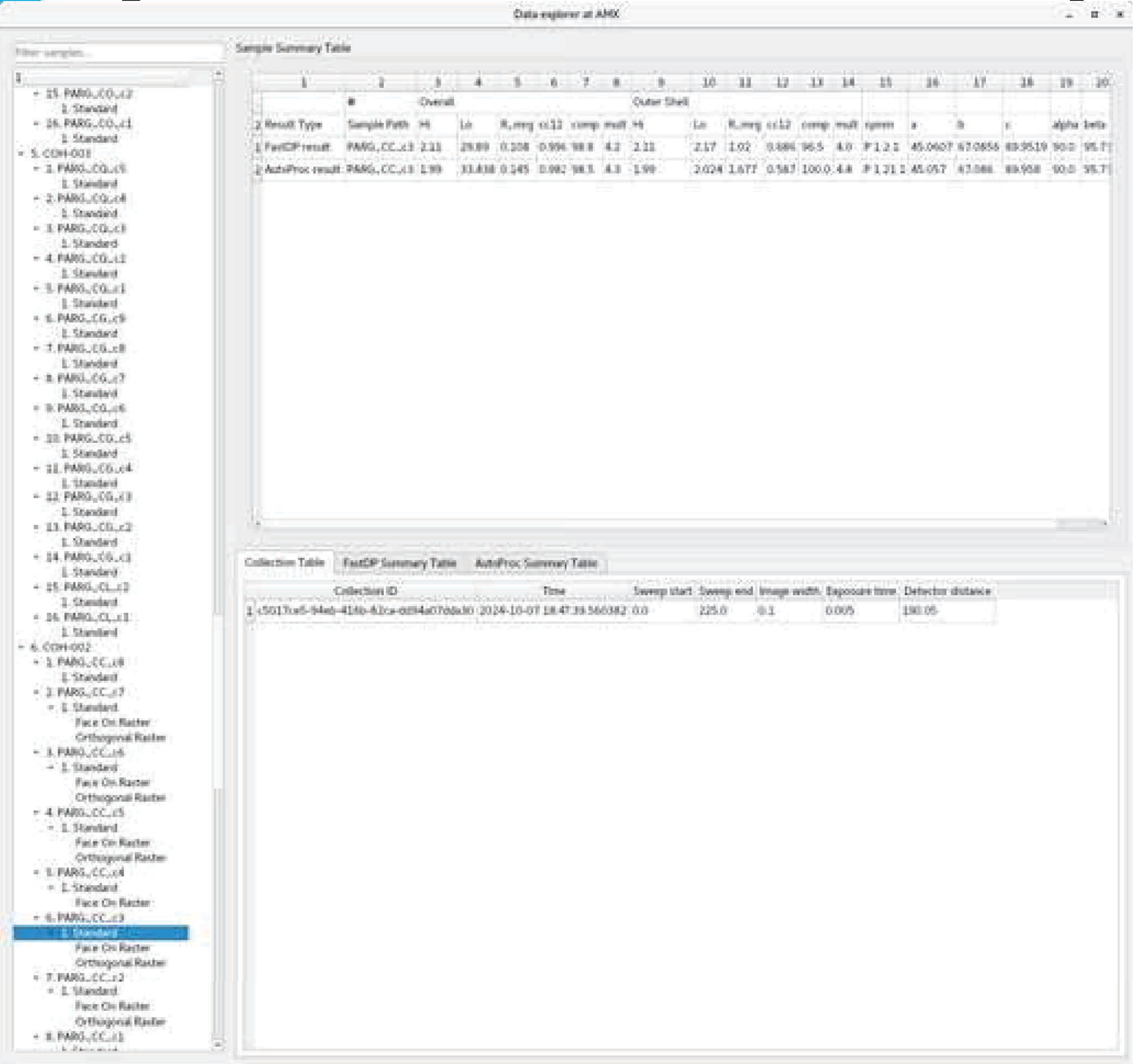


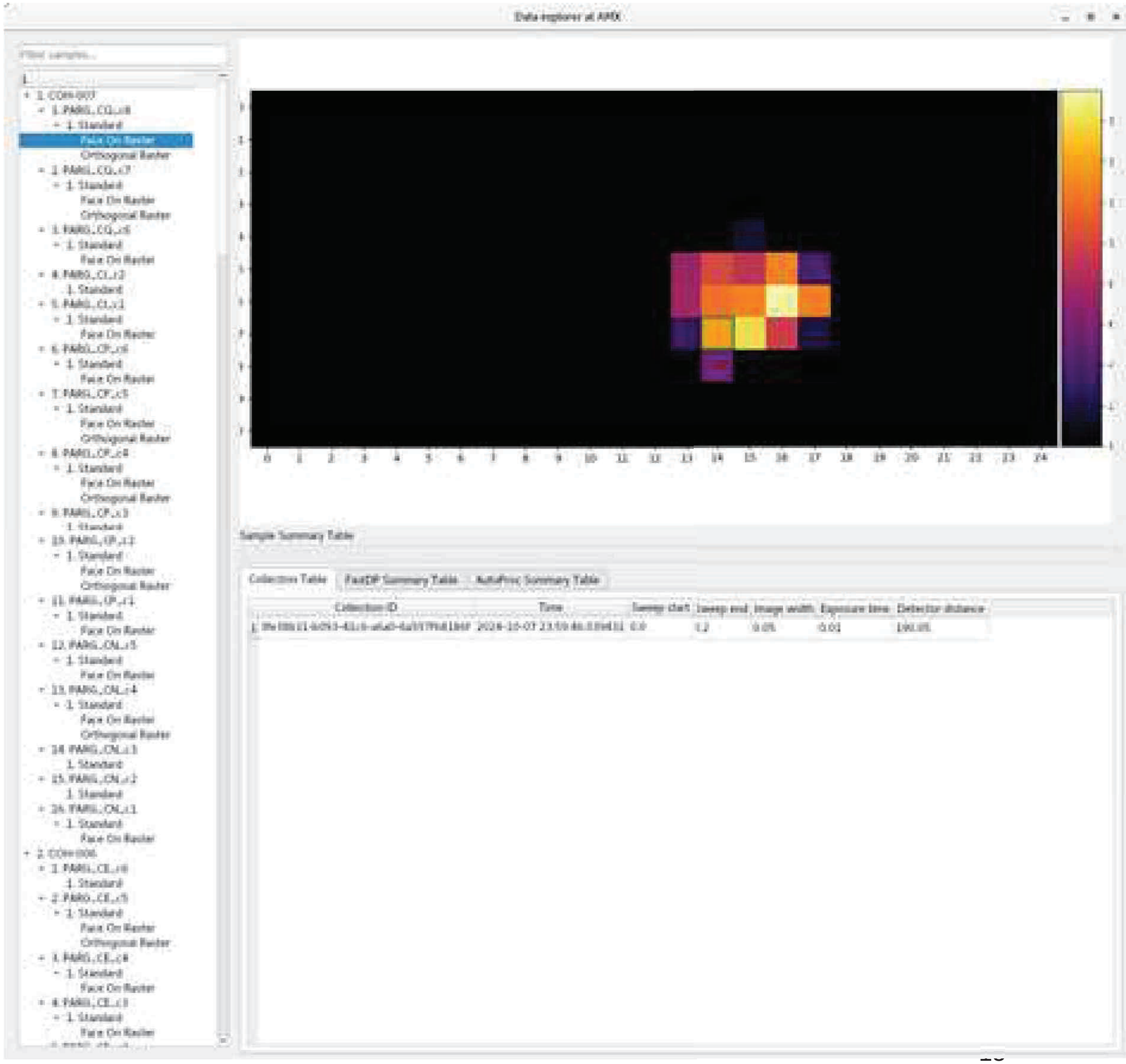
Html file
Uploaded on private slack channel

Hybrid Mode: the report



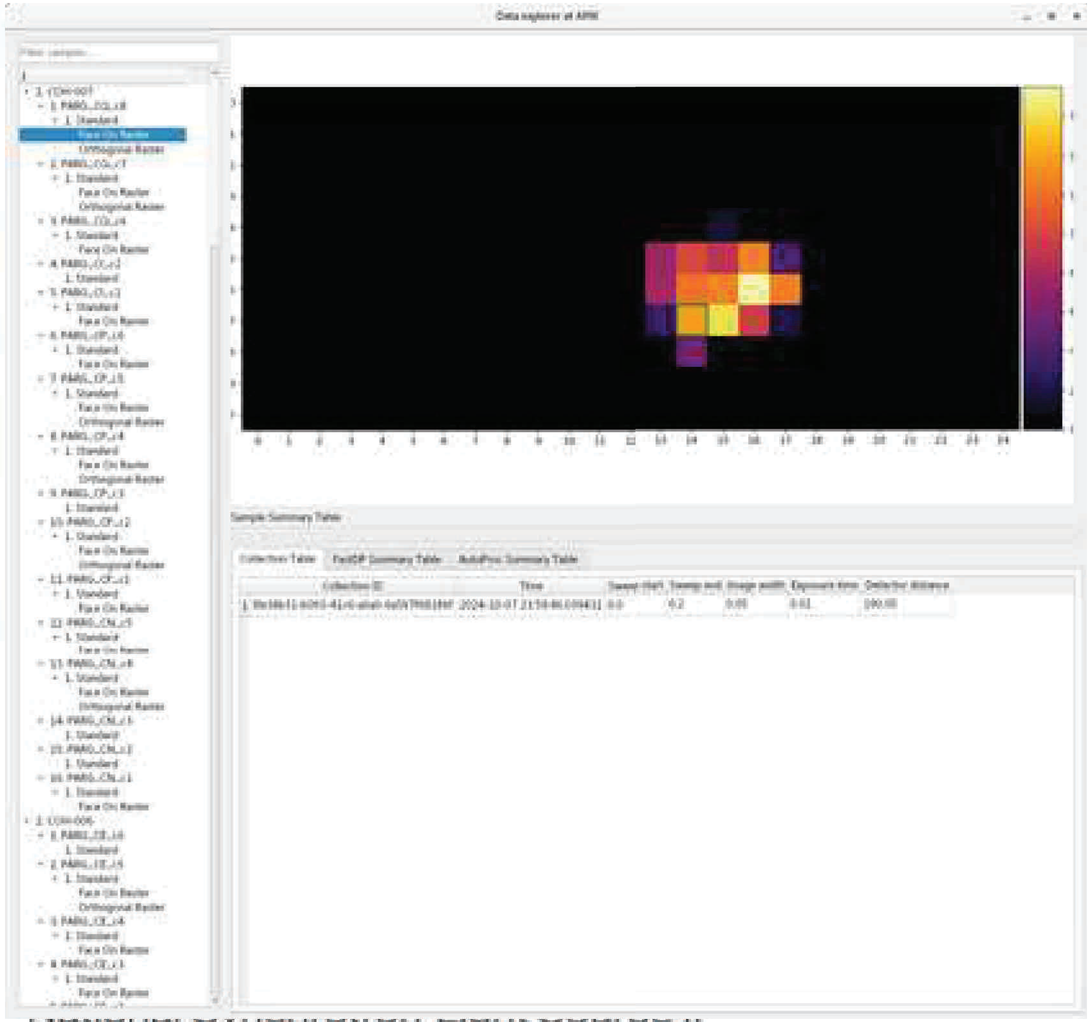
Hybrid mode: rasterXplor GUI

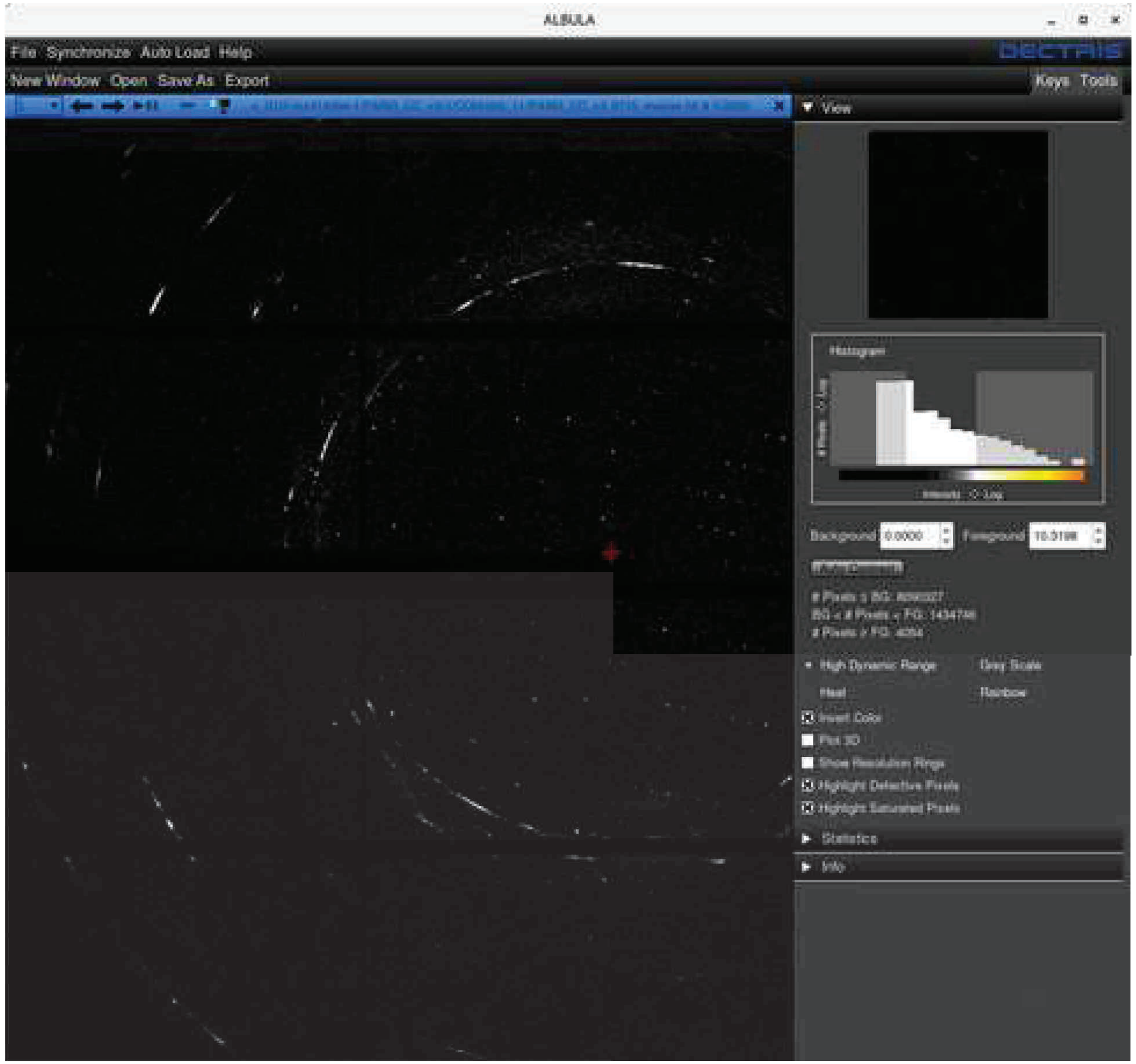




Hybrid mode: rasterXplor GUI

NX to beamline WS or VMWare Horizon)





Upcoming Planned Upgrades

FMX new detector: Eiger 2 XE 9M: from 10 msec to 2 msec exposure time in 9 M mode.
AMX new detector: Eiger 2 XE CdTe 9M: from 5 msec to 2 msec exposure time in 9 M mode
current standard energy at AMX: 13.5 toward 15 keV: DQE from 65-60 to 95 % (a 50 % increase).

Commercial dual grippers delivering under 10 seconds sample exchange time in automated mode.

Will enable higher achievable sample throughputs, from the automated collection
from currently 24 samples per hour to 32 samples per hour (2 pucks) for standard collection.

These upgrades will also improve throughput from user-controlled operation.

Accompanied with software and workflow improvements to deliver higher reliability.

And unleash fully automated / autonomous data collection and or crystal screening from 8 PM to 8 AM for all samples before user access to mini-shifts the next day (9 AM to 6 PM) at higher throughput (3 x 3 H)

Toward 2 full dewars per day ~ 800 samples / day at the NSLS-II in all standby mode
Toward 500 samples per day in future hybrid mode: (3 x 2 H beamtime for user controlled exps)

BraggSpotFinder

The challenge: existing applications are fast and mostly efficient

With upcoming faster detectors and brighter beamlines we expect at least a 10x in rastering speed and more for specialized experiments

Today ”state of the art” MX beamlines achieve ~ 1 000 FPS, toward ~ 20 000 in the future

Not necessarily translates in 20x increased fold in speed of rastering, unless using a sub-micron beam and large sample mount.

Existing applications may need upgrades to keep up, beyond hardware upgrades (cluster: cores and freq).

Most spot finder applications benchmarked with rastering data from AMX deliver reliable results

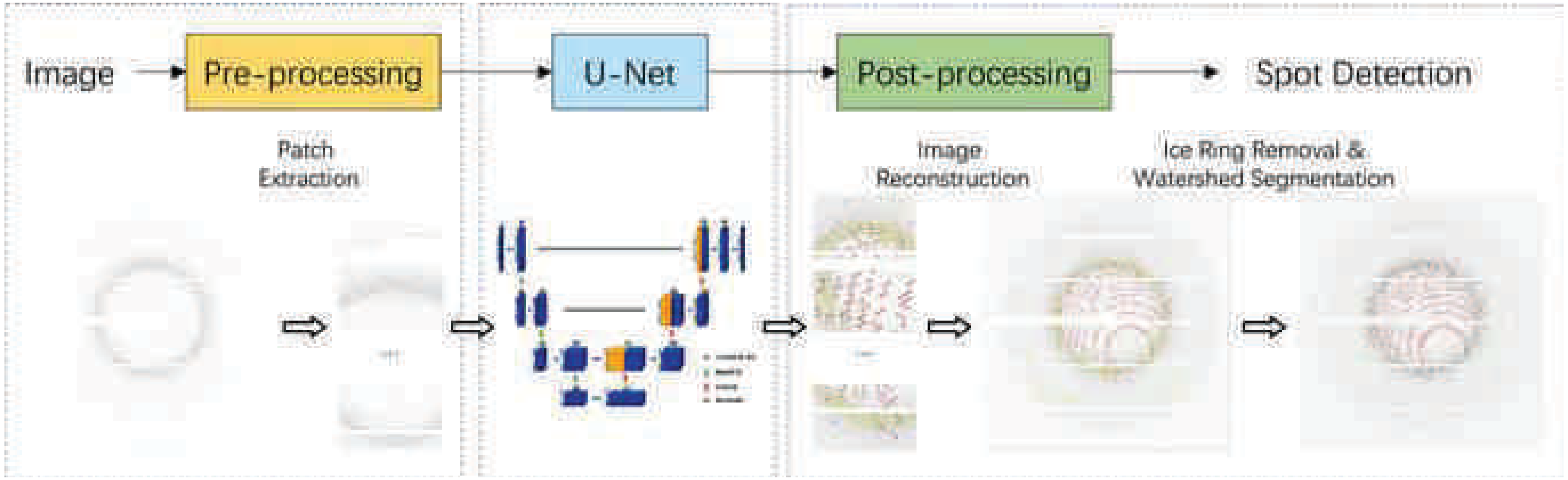
However they tend to fail detecting weakest reflections, and extra computation requires multiple pass of the data slowing down the achievable throughput:~ 250 frames per second on a dual EPYC 9554 (64 cores @ 3.1 GHz)

BraggSpotFinder

We assembled a divers data set with a variety of diffraction frames from rastering experiments.

Data was annotated by 2 crystallographers using: dozor output and adxv to add / remove reflections.

Supervised machine learning : U-Net, a convolutional neural network for biomedical image segmentation & Watershed algorithm to isolate nearby reflections



Collaboration with Zhaozheng Yin, CS Stony Brook University, NY.

J. Appl. Cryst., **57**(3), 670-680 (2024). DOI: 10.1107/S1600576724002450

BSD data set on zenodo
309 frames from 49 crystals
10.5281/zenodo.10667263

Planned Work: AI, ML, and more

- Lossy compression: develop an optimized algorithms for MX & data reconstruction using Generative Adversarial Network for extreme compression (>1000/1)
- Trained model to draw raster area around protein crystal(s) (new training set:10 000 images assembled)

    labelling using: pin / stem / loop / crystal on 2 frames per sample

Trained model using 10 000s of processed data sets available at the NSLS-II beamlines to potentially

Use a few random diffraction patterns and the scaled statistics from each of these data sets

Using a few random diffraction patterns of data collected with no scaled results too.

To train a model predicting: quality of data

Investigate potential use of indexing during rastering experiment (rastering: 0.02 0.05 deg / frame)

Implement near neighbor cluster analysis during rastering to reduce data collection from multiple lattices part of the new scoring function.

High Throughput Fragment Screening


Chems library storage


crystalShifter



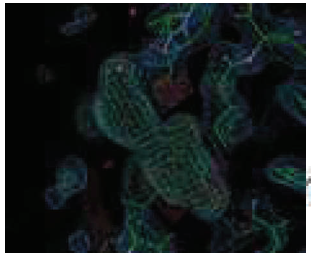
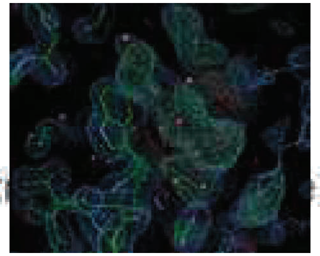
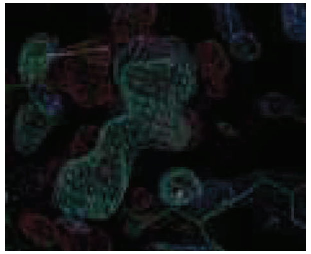
Dedicated uniPucks and > 1000 Spine Pins

Pilot users from:
California, Connecticut, New York
Pilot project areas:
Infectious disease, cancer
Lead compounds in development

Available Libraries @ NSLS-II:

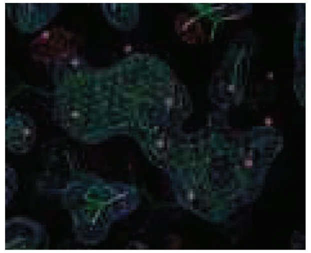
- DSI Poised library (860 compounds)
- LifeChemicals Fragment Diversity Set #3 (320 compounds)
- FragLites (31 compounds)

Future Libraries:
Natural Compounds Fraction: NCI
….

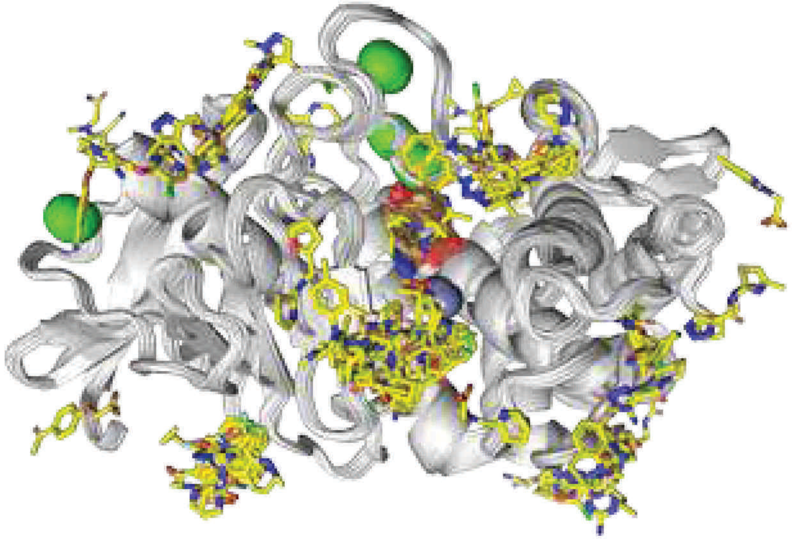
6 pilot projects through workflow (so far)
More are underway; RNA, Cancer Targets, Plant Biology …
A few in-house targets selected;
Users come to NSLS-II lab to execute workflow

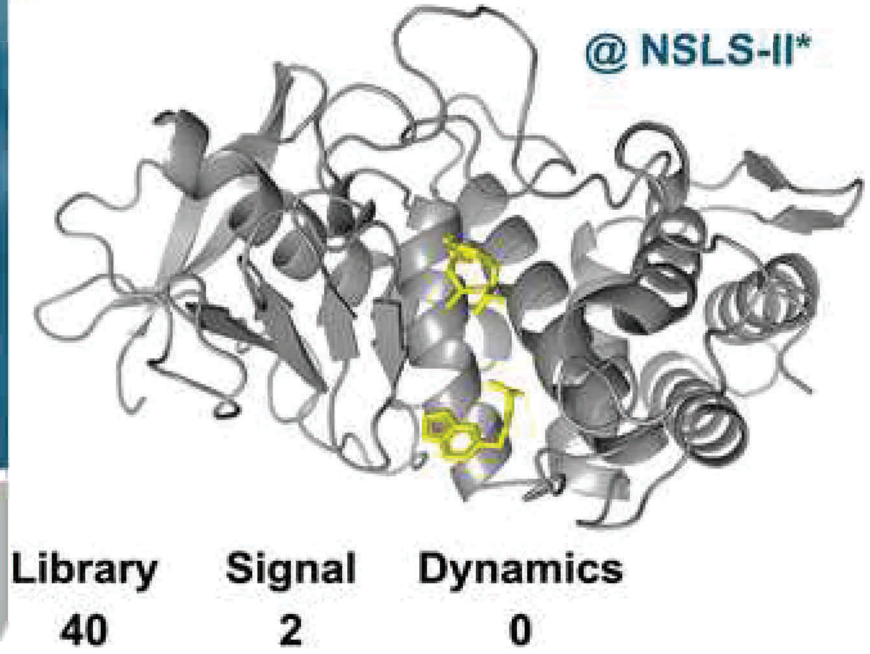
Up to ~ 1000 data sets from ~ 1000 crystals (~50 for solvent test)
Data collection: fully automated: 8 shifts
Automated data processing: fast_dp & autoProc
For academia only for now
PanDDA analysis: ID of structural events and ligand binding
>> bound and unbound states + ground state
Using NSLS-II cluster; 5 % hit rate !!



Dale Kreitler: dkreitler@bnl.gov

Thermolysin Fragment Hits





- Multiple allosteric sites observed
- Dipeptide occlude some portion of active site

Today we can achieve ~ 600 samples / day

Interest from: plant science & biofuels

Interest from industry for the complete workflow

- Toward 1200 samples / day (fragment projects)
- Toward > 5000 collections / day (>5 Xtals/mount)

MAX Beamline Proposal

Mission:

Tackling structural biology science projects requiring access to extremely high throughput, unattended, autonomous, automated data collection and advanced analysis, without compromising data quality, for accelerated discoveries

Source	3m IVU optimized for 18.5 keV
Energy	Fixed: 18.5 keV (0.67 Å)
Beam Size	Selectable: 2-10-20 um (2 banks of CRLs)
Optics	Si DCM + CRLs + harmonic rejection Rh coated mirror
Flux:	> 2 10^{12} ph/s (4 10^{12} ph/s at NSLS-IIU) [5 10^{14} ph/s wide bandpass multilayer]
Detector	Eiger 2 XE 4M CdTe: 2250 FPS
Robot	NSLS-II Style with double-gripper
Sample storage	2 x Dewars: ~ 70 pucks

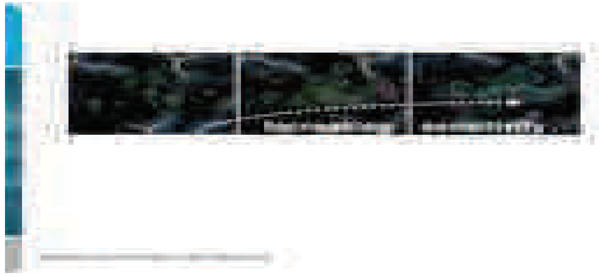
Enzymatic reactions
and electron transfer

Accelerate Therapeutics
development
Ligand binding studies
Fragment screening

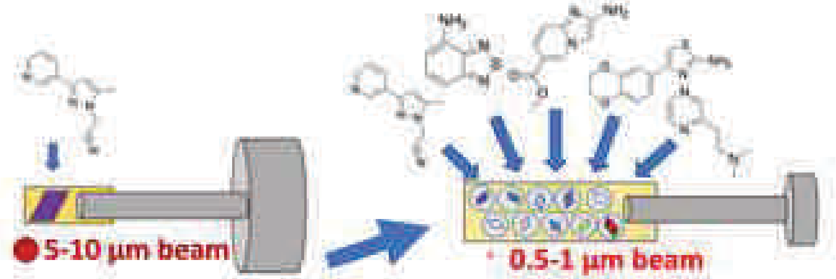
Metabolite screening
and **bio-fuels**

Protein **Dynamics**
Crystal Screening
Development of **predictive** ligand /
protein interactions models

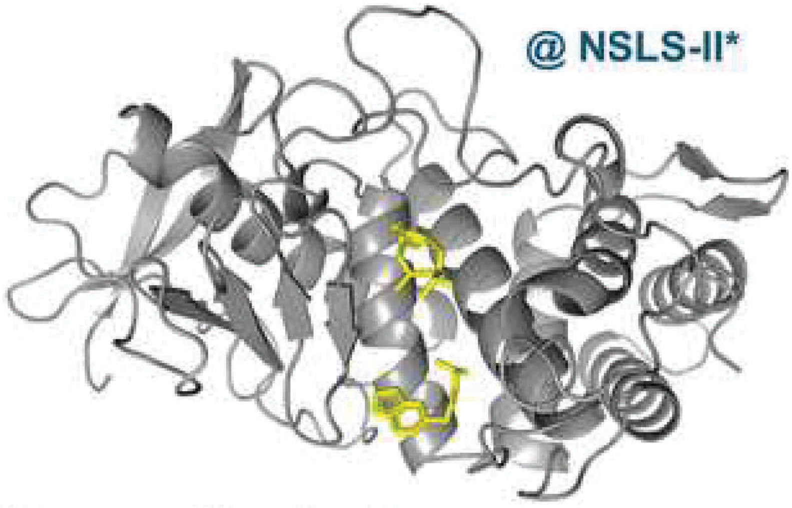
High Sensitivity Chemical Biology



Enabling new science	Understanding and then controlling biological processes at the molecular level to transform our understanding and treatment of many diseases.
How?	By determining the structures of 1000s of complexes between macromolecules and small-molecules. Collecting 100s data sets on each crystal using NSLS-IIU beam will reveal weakest features.
Why?	Develop next generation drugs: cancer, plant (food security), pain medications, antibiotics and bio-fuels production
The Future: NSLS-IIU	The improved emittance and flux enables 10,000 fold increased data compared to NSLS-II, revealing new science and biomolecular dynamics. Massive data will unleash drug discovery AI engines.



NSLS-II to NSLS-IIU:
100x collections/crystal & 100x crystals: **10,000x data**

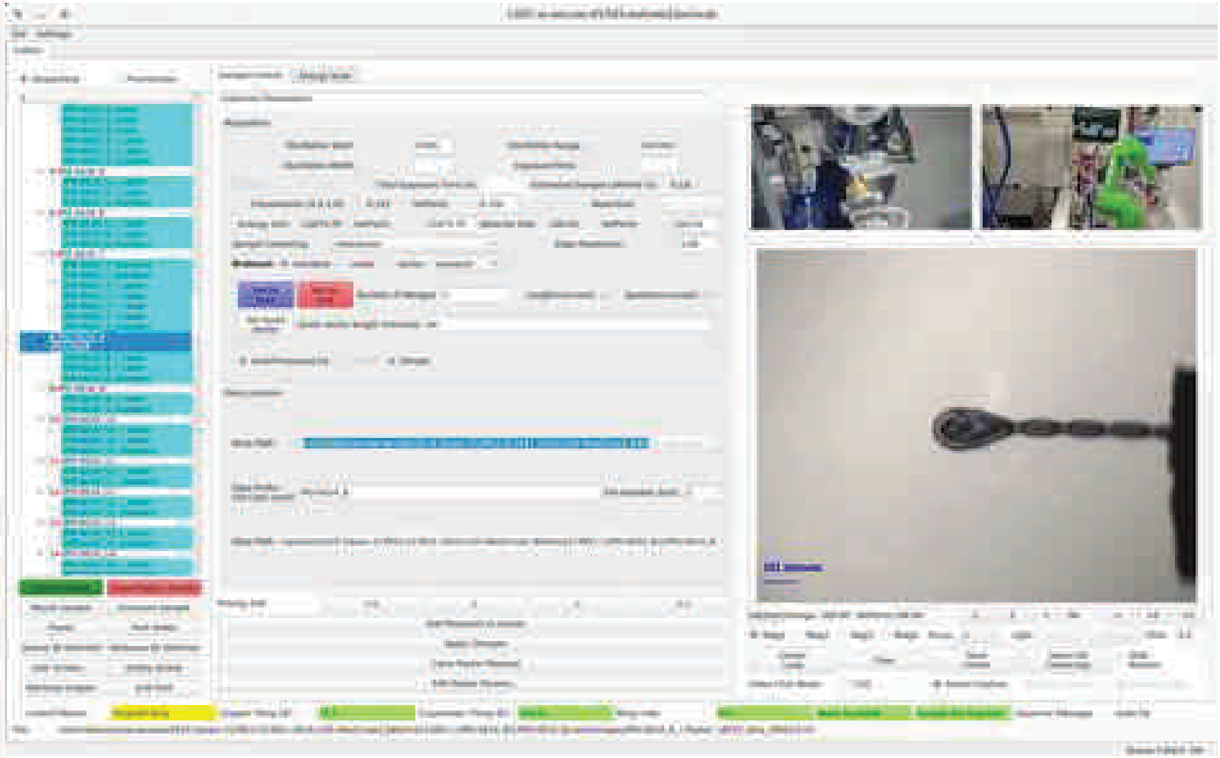


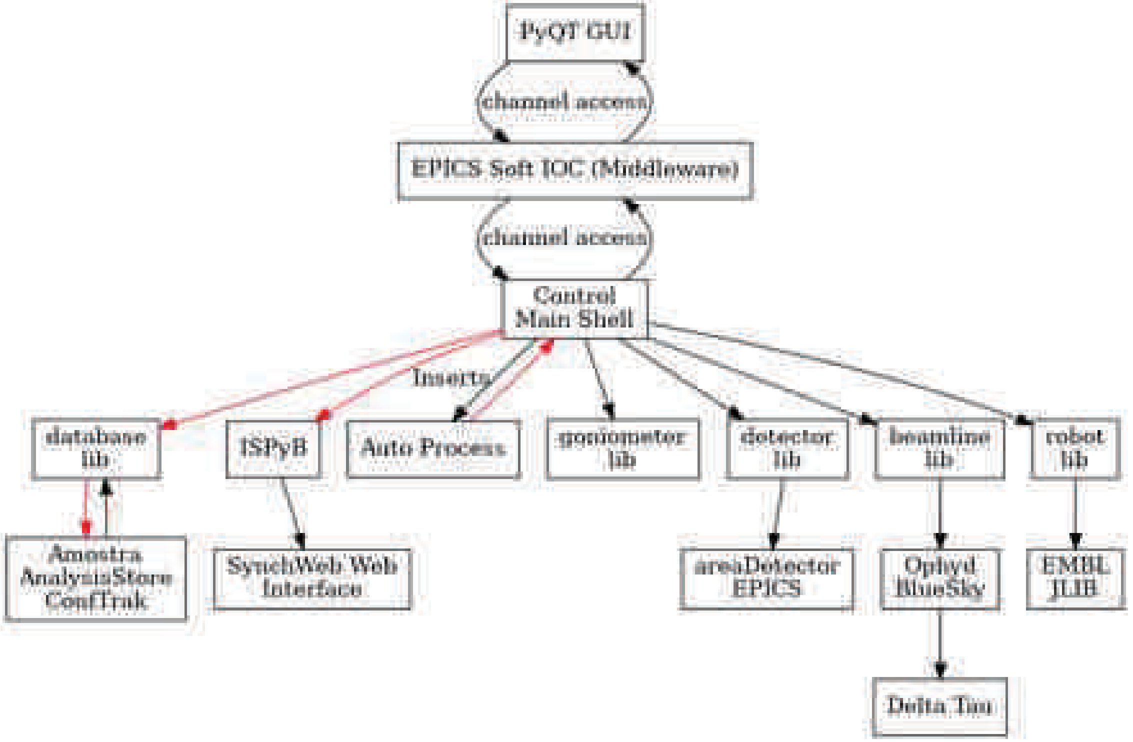
Library	**Signal**	**Dynamics**
40	**2**	**0**

*manual experiment

LSDC @ all NSLS-II MX beamlines

MXCuBE Py-QT inspired GUI
Developed during AMX/FMX
Construction projects (2013-2016)



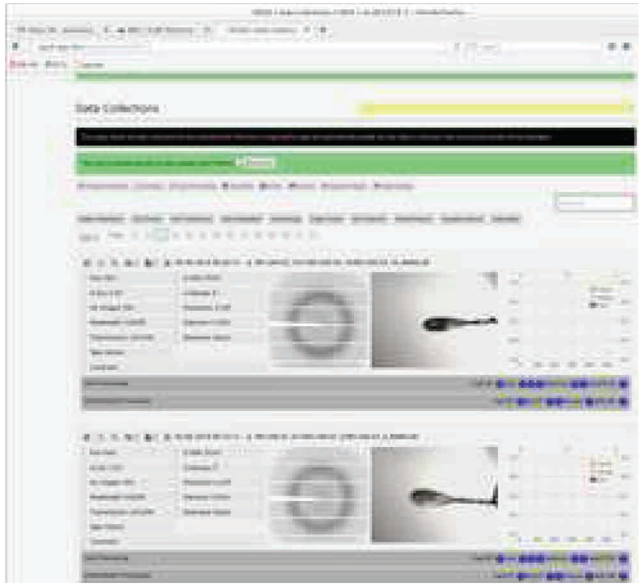


Rasters / vectors / standard were all developed at the delta tau controller level
& converted into Ophyd objects for optimal integration

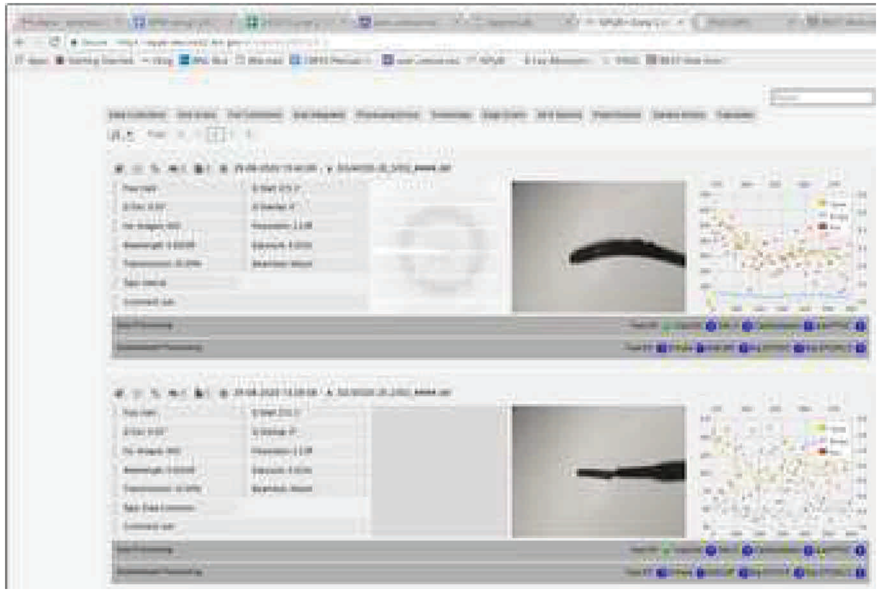
ISPyB/SynchWeb at NSLS-II MX beamlines

ISPyB / SynchWeb installed and running in 2019 with “basic” features

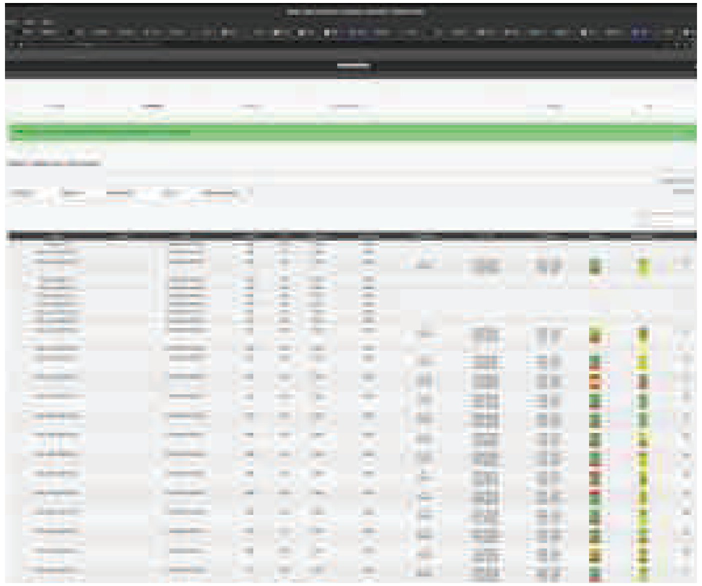
1st frame from rasters



1st frame from data sets + spot plots



Data collection summary



2021: Re-IP and major cyber security work preventing ISPyB/SynchWeb at NSLS-II
2023: resources allocated and 11/2024: some basic features available in 2024

Short term at NSLS-II: puck shipping / sample sheet import ISPyB/SynchWeb while users rely on provided report

Long term: users rely only on ISPyB/SynchWeb + HCA module and Fragment Module ? / better report + post processing preset workflows : will wait and see development from here

Conclusions and Future Outlook

Use existing tools as often, collaborate when needed and develop tools if required.

Sometimes modest investment have huge impact (top view, new loop centering)

Working on advanced data analysis: users are overwhelmed with amount of data (reprocessing too)

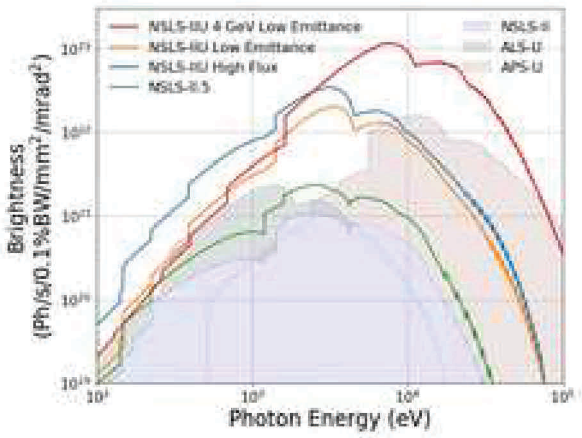
MAX beamline proposal: extreme throughput with autonomous operation and advanced analysis

NSLS-IIU in the horizon:

best strategy ?: work with AI/ML experts BUT make sure formulation of project s optimal
VS crystallographers using AI/ML tools

We are transitioning to VMWare horizon for remote access and remote post-processing

Investigating potential use of event-based detector



Acknowledgements

Dale Kreitler, Shekar Venkateswaran, Edwin Lazo, Stuart Myers, Herbert Bernstein, Martin Fuchs, Sean McSweeney

NSLS-II CBMS team & DSSI team

SDCC team

Jianxiang Dong and Zhaozheng Yin, CS Stony Brook University

The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011).

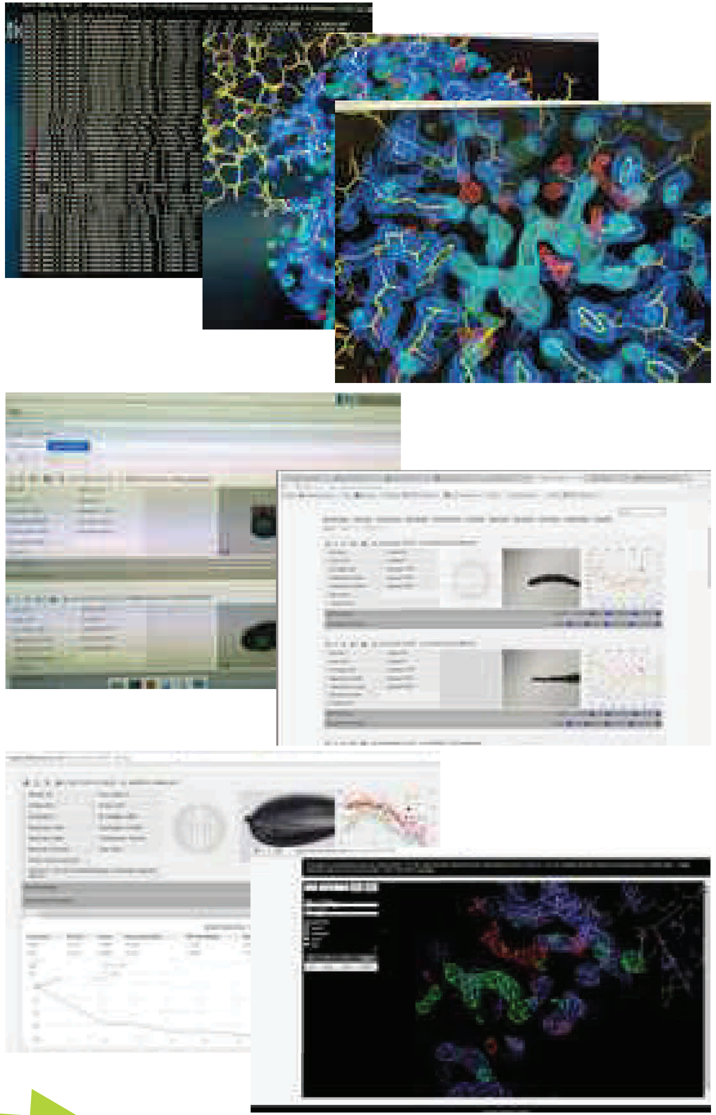
As part of NSLS-II, a national user facility at Brookhaven National Laboratory, work performed at the CBMS is supported in part by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Program under contract number DE-SC0012704.

Users send 100s to 1000s of crystals or rely on the FBDD platform at NSLS-II

Fragment Based Drug Discovery

Data are automatically and autonomously collected
Up to 3000 samples /day
NO GUI

extreme and relentless



Advanced pipelines and workflows are executed for each science case and curated results presented to users within 1-2 days