Automated analysis workflows

James Wrigley for DA/MID/FXE/lots of lovely users

whoami

- Data analysis contact for MID.
- Spends lots of time preparing analysis for experiments and helping users.

Common pain points :(

(from an analysis perspective)

1. Many experiments produce large amounts of data that require significant work to get something scientifically meaningful. Examples:
   - XPCS: 404 runs, 628TB
   - SFX: 286 runs, 161TB
   - Solution scattering: 435 runs, 441TB
2. The raw data recorded by the facility is often difficult to understand (many data sources, confusing names, etc) and in a different representation than what's desired. Examples:
   - Nozzle temperature: `MID_EXP_UPP/CTRL/LSHORE.inputB.krdg`
   - Monochromator position: `FXE_XTD9_MONO-1/MOTOR/ACCM_PITCH.actualPosition`

Key takeaway

With *high-level analysis* and *automation* it's possible to get a synchrotron analysis experience[1]!

[1] This is a lie, but it's close to being true.

That synchrotron feeling

Concrete steps that we suggest doing:

That synchrotron feeling

Concrete steps that we suggest doing:

- Most importantly, talk to us. Tell the instrument contact or local data contact about what you want to do or send an email to `da@xfel.eu`.

That synchrotron feeling

Concrete steps that we suggest doing:

- Most importantly, talk to us. Tell the instrument contact or local data contact about what you want to do or send an email to `da@xfel.eu`.
- Some technical steps:
    - Use extra-data's aliases for sensible names to do e.g. `run.alias["mono-position"]`:

extra-data-aliases.yml

```yaml
nozzle-temperature: [MID_EXP_UPP/CTRL/LSHORE, inputB.krdg]
mono-position:      [FXE_XTD9_MONO-1/MOTOR/ACCM_PITCH, actualPosition]
```

That synchrotron feeling

Concrete steps that we suggest doing:

- Most importantly, talk to us. Tell the instrument contact or local data contact about what you want to do or send an email to `da@xfel.eu`.
- Some technical steps:
  - Use extra-data's aliases for sensible names to do e.g. `run.alias["mono-position"]`:

    extra-data-aliases.yml

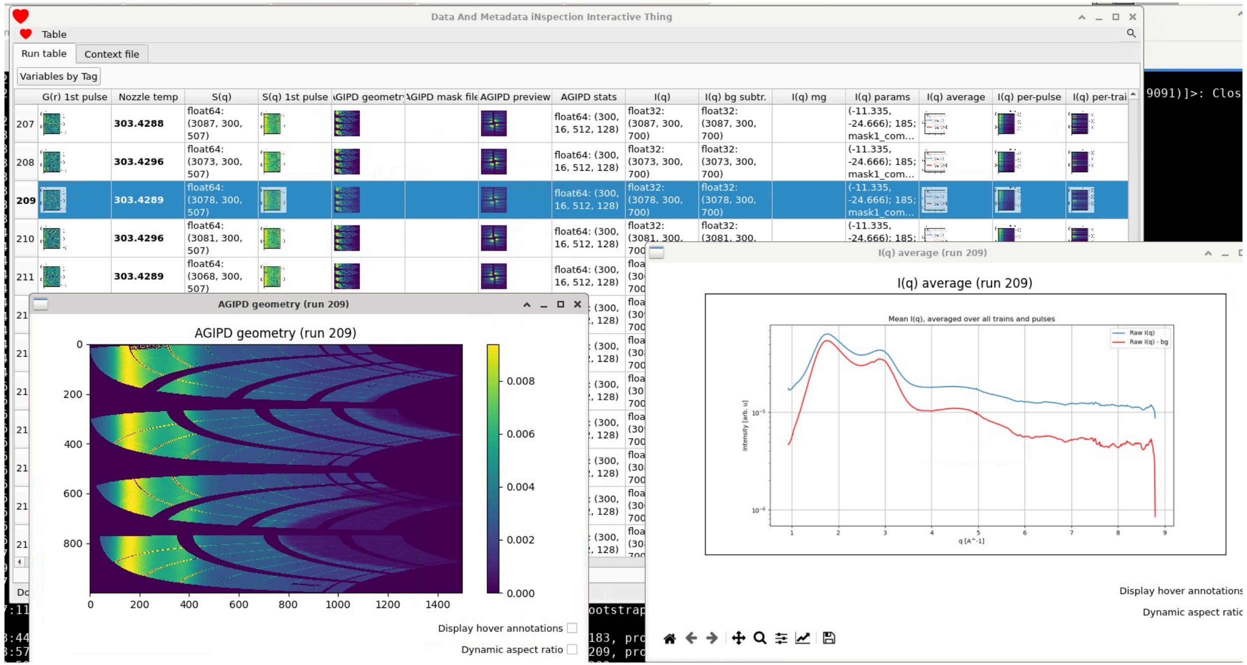
    ```
    nozzle-temperature: [MID_EXP_UPP/CTRL/LSHORE, inputB.krdg]
    mono-position:      [FXE_XTD9_MONO-1/MOTOR/ACCM_PITCH, actualPosition]
    ```

  - Move code from notebooks into an automated pipeline with DAMNIT.

DAMNIT

- Automatically creates a run table from custom, user-defined functions.
- Results saved into a database for display and accessible through a Python API.

```python
from extra.damnit import Damnit

db = Damnit(1234)
i_q = db[200]["i_q"].read()
```



Data And Metadata iNspection Interactive Thing

Table

Run table | Context file

Variables by Tag

	G(r) 1st pulse	Nozzle temp	S(q)	S(q) 1st pulse	AGIPD geometry	AGIPD mask file	AGIPD preview	AGIPD stats	I(q)	I(q) bg subtr.	I(q) mg	I(q) params	I(q) average	I(q) per-pulse	I(q) per-trai
207		**303.4288**	float64: (3087, 300, 507)					float64: (300, 16, 512, 128)	float32: (3087, 300, 700)	float32: (3087, 300, 700)		(-11.335, -24.666); 185; mask1_com...			
208		**303.4296**	float64: (3073, 300, 507)					float64: (300, 16, 512, 128)	float32: (3073, 300, 700)	float32: (3073, 300, 700)		(-11.335, -24.666); 185; mask1_com...			
209		**303.4289**	float64: (3078, 300, 507)					float64: (300, 16, 512, 128)	float32: (3078, 300, 700)	float32: (3078, 300, 700)		(-11.335, -24.666); 185; mask1_com...			
210		**303.4296**	float64: (3081, 300, 507)					float64: (300, 16, 512, 128)	float32: (3081, 300, 700	float32: (3081, 300,		(-11.335, -24.666); 185;			
211		**303.4289**	float64: (3068, 300, 507)					float64: (300, 16, 512, 128)	floa (30 700						

9091)]>: Clos

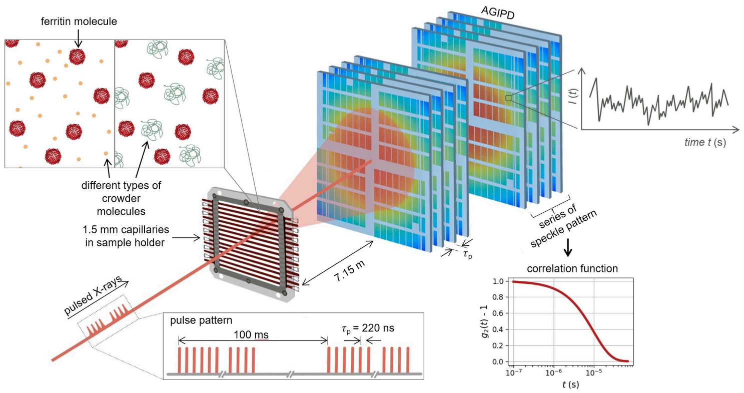
AGIPD geometry (run 209)

I(q) average (run 209)

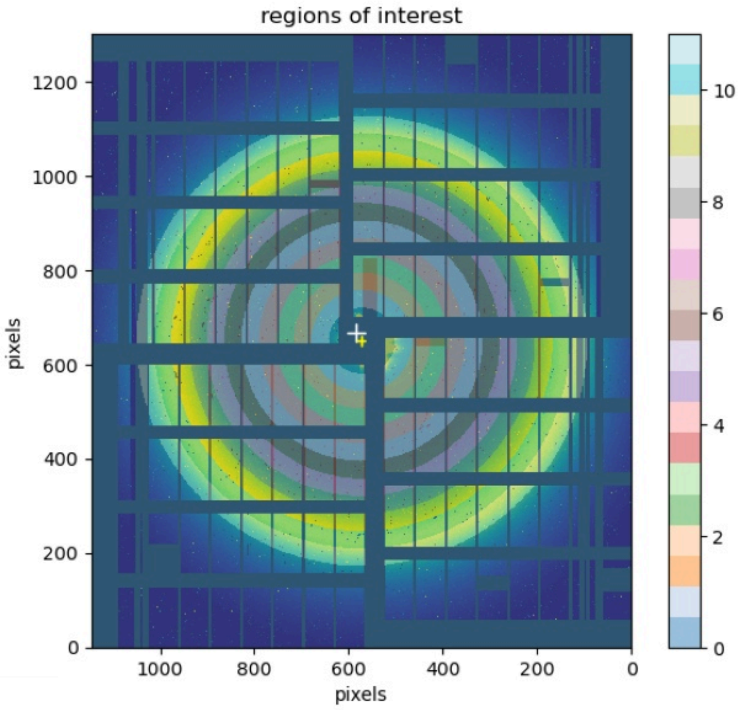
Also see poster 58, “DAMNIT, a tool for Automated Experiment Overview” by T. Michelat.

Example: XPCS

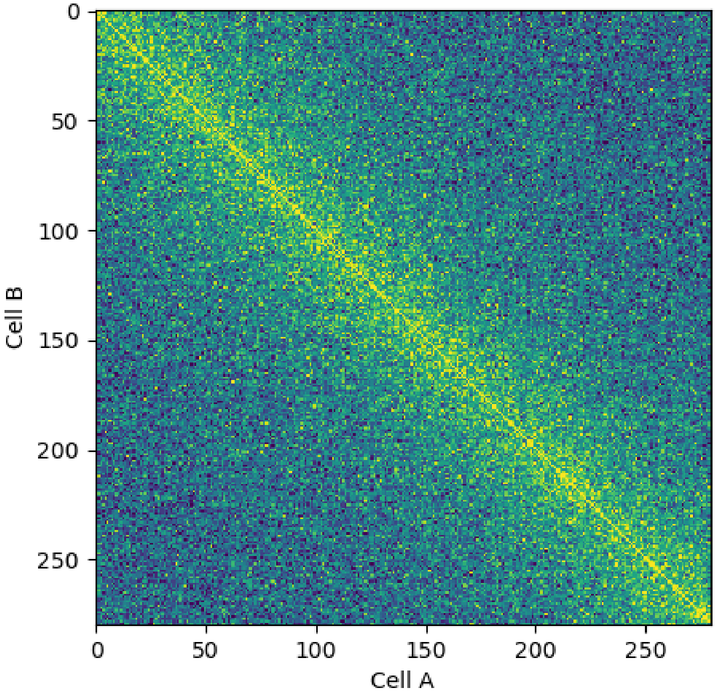
- X-ray Photon Correlation Spectroscopy: technique for investigating sample dynamics based on correlating laser speckle across time.
- Key quantities are Two-Time Correlation Functions (TTCFs) and g_2 functions.



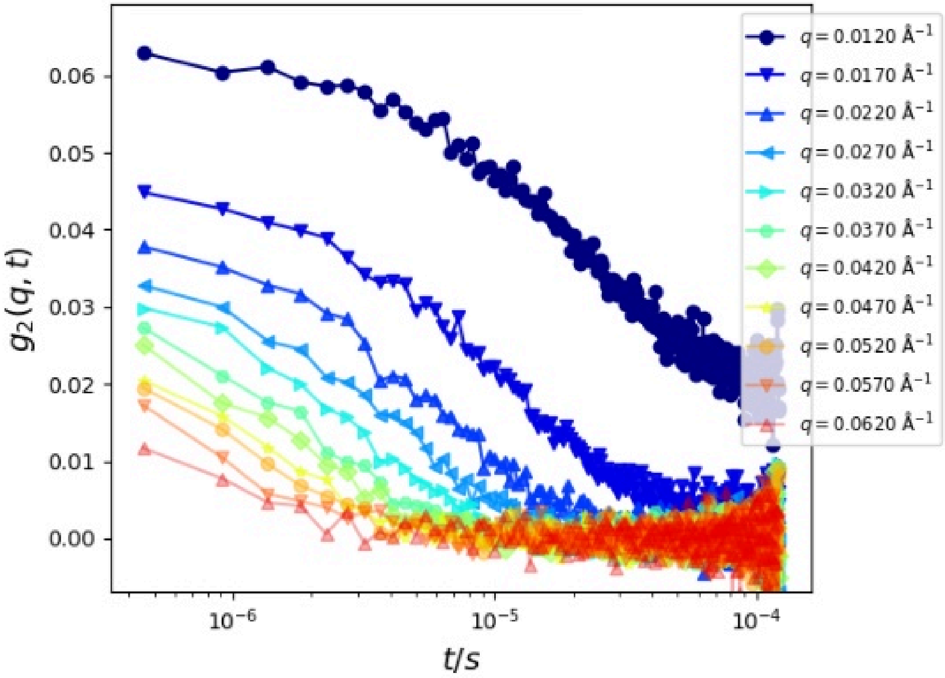
XPCS experiment diagram by M. Dargasz.



Example: XPCS



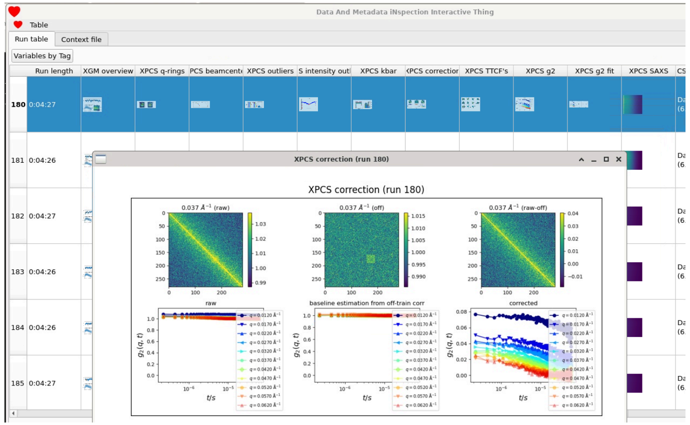
Example TTCF



g_2's for multiple TTCFs at different q values

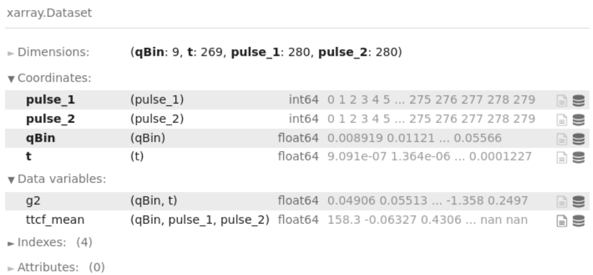
Example: XPCS

- Mature offline pipeline, developed over a few years by multiple people from DA/MID/University of Siegen.
- Performance: processing a 5 minute run takes ~20 minutes using multiple processes.
- Fully integrated with DAMNIT



	Run length	XGM overview	XPCS q-rings	PCS beamcent	XPCS outliers	S intensity outl	XPCS kbar	XPCS correctior	XPCS TTCF's	XPCS g2	XPCS g2 fit	XPCS SAXS	CS
180	0:04:27												Da (6
181	0:04:26												Da (6
182	0:04:27												Da (6
183	0:04:26												Da (6
184	0:04:26												Da (6
185	0:04:27												Da (6

```python
from extra.damnit import Damnit

db = Damnit(1234)
xpcs = db[300]["xpcs_mean_dataset"].read()
```



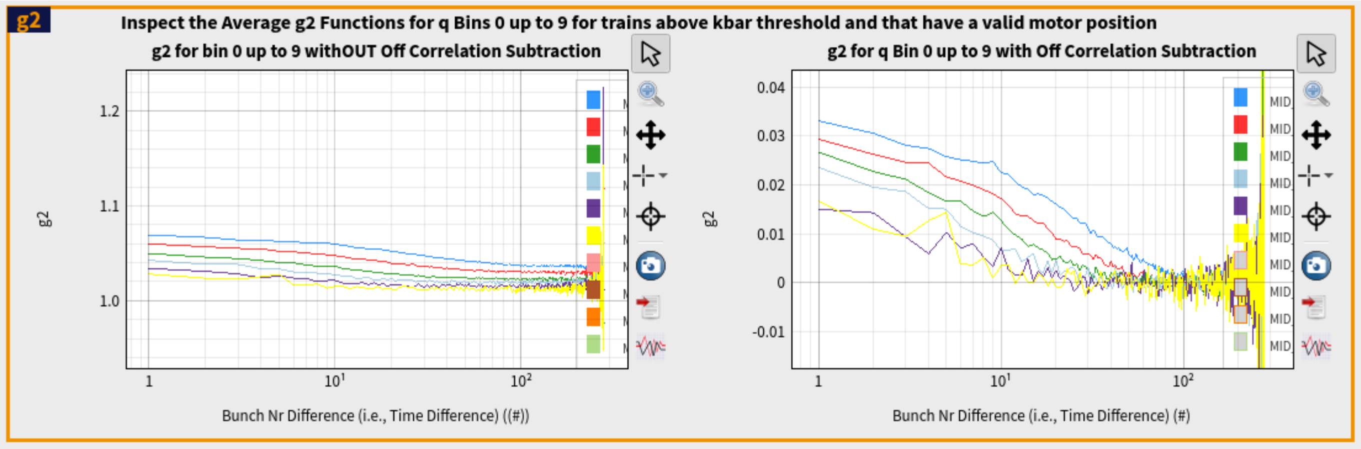
```
xarray.Dataset

► Dimensions:     (qBin: 9, t: 269, pulse_1: 280, pulse_2: 280)
▼ Coordinates:
    pulse_1       (pulse_1)                  int64    0 1 2 3 4 5 ... 275 276 277 278 279
    pulse_2       (pulse_2)                  int64    0 1 2 3 4 5 ... 275 276 277 278 279
    qBin          (qBin)                   float64    0.008919 0.01121 ... 0.05566
    t             (t)                      float64    9.091e-07 1.364e-06 ... 0.0001227
▼ Data variables:
    g2            (qBin, t)                float64    0.04906 0.05513 ... -1.358 0.2497
    ttcf_mean     (qBin, pulse_1, pulse_2) float64    158.3 -0.06327 0.4306 ... nan nan
► Indexes:  (4)
► Attributes:  (0)
```

Also see poster 11, “A High-Throughput Data Pipeline for MHz-XPCS: Offline Analysis” by A. Leonau.

Also see the “A High-Throughput Data Pipeline for MHz XPCS: Analyzing Protein Dynamics via Solution Scattering” talk by A. Leonau in the Data Science session on the 22nd.

Example: XPCS (online)

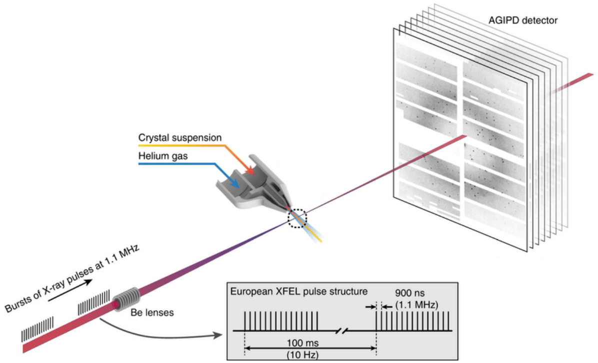
- Quantities are computed in the same way as offline, modulo certain filtering methods.
- Provides real-time feedback at 10Hz.
- Also integrated with DAMNIT to make online results available as soon as the run finishes.

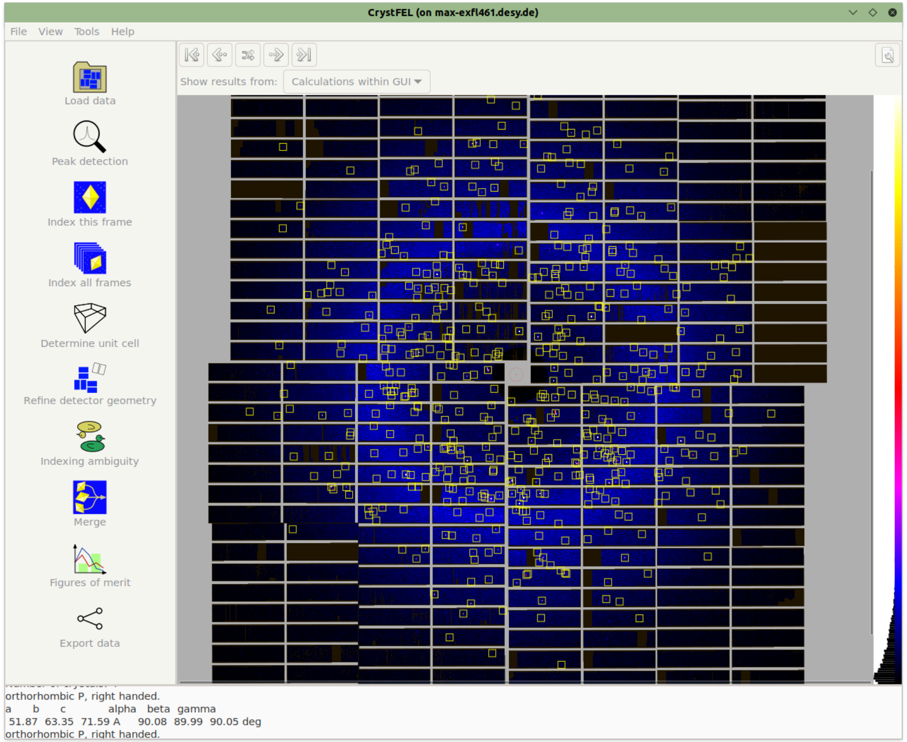

g2
Inspect the Average g2 Functions for q Bins 0 up to 9 for trains above kbar threshold and that have a valid motor position
g2 for bin 0 up to 9 withOUT Off Correlation Subtraction
g2
1.2
1.1
1.0
1
10^1
10^2
Bunch Nr Difference (i.e., Time Difference) ((#))
g2 for q Bin 0 up to 9 with Off Correlation Subtraction
g2
0.04
0.03
0.02
0.01
0
-0.01
1
10^1
10^2
MID
Bunch Nr Difference (i.e., Time Difference) (#)

Also see poster 10, "A high throughput Data Pipeline for MHz XPCS: Online Analysis" by M. Jakobsen.

Example: SFX

- Serial Femtosecond Crystallography: technique for finding the atomic structure of a sample.
- Users come with many small crystals (e.g. from a protein), which are fed into the beam to diffract onto the detector.
- Analysis requires multiple steps (e.g. hit finding, indexing) to end up with an electron density map.



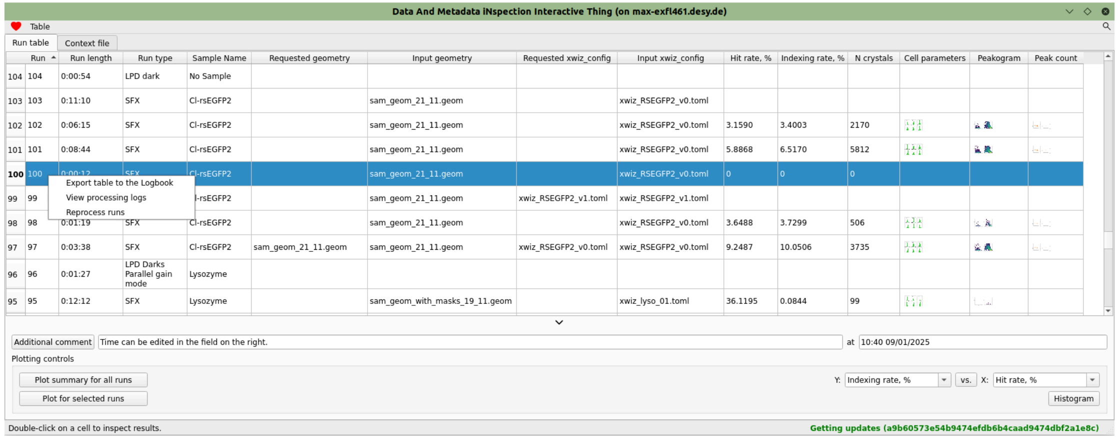


```
orthorhombic P, right handed.
a      b      c           alpha   beta  gamma
 51.87  63.35  71.59 A     90.08  89.99  90.05 deg
orthorhombic P, right handed.
```

SFX experiment diagram adapted from Wiedorn et al. (2018), Nat. Comm. 9.

Example: SFX

- Extra-Xwiz: a wrapper around CrystFEL.
- Goal is to simplify parallelization, data handling, and configuration. End result is the structure factors.
- Integrated with DAMNIT to analyze runs automatically.



Data And Metadata iNspection Interactive Thing (on max-exfl461.desy.de)

Table

Run table | Context file

	Run	Run length	Run type	Sample Name	Requested geometry	Input geometry	Requested xwiz_config	Input xwiz_config	Hit rate, %	Indexing rate, %	N crystals	Cell parameters	Peakogram	Peak count
104	104	0:00:54	LPD dark	No Sample										
103	103	0:11:10	SFX	Cl-rsEGFP2		sam_geom_21_11.geom		xwiz_RSEGFP2_v0.toml						
102	102	0:06:15	SFX	Cl-rsEGFP2		sam_geom_21_11.geom		xwiz_RSEGFP2_v0.toml	3.1590	3.4003	2170			
101	101	0:08:44	SFX	Cl-rsEGFP2		sam_geom_21_11.geom		xwiz_RSEGFP2_v0.toml	5.8868	6.5170	5812			
100	100	0:00:12	SFX	Cl-rsEGFP2		sam_geom_21_11.geom		xwiz_RSEGFP2_v0.toml	0	0	0			
99	99			l-rsEGFP2		sam_geom_21_11.geom	xwiz_RSEGFP2_v1.toml	xwiz_RSEGFP2_v1.toml						
98	98	0:01:19	SFX	Cl-rsEGFP2		sam_geom_21_11.geom		xwiz_RSEGFP2_v0.toml	3.6488	3.7299	506			
97	97	0:03:38	SFX	Cl-rsEGFP2	sam_geom_21_11.geom	sam_geom_21_11.geom	xwiz_RSEGFP2_v0.toml	xwiz_RSEGFP2_v0.toml	9.2487	10.0506	3735			
96	96	0:01:27	LPD Darks Parallel gain mode	Lysozyme										
95	95	0:12:12	SFX	Lysozyme		sam_geom_with_masks_19_11.geom		xwiz_lyso_01.toml	36.1195	0.0844	99			

Additional comment: Time can be edited in the field on the right. at 10:40 09/01/2025

Plotting controls

Double-click on a cell to inspect results.

Getting updates (a9b60573e54b9474efdb6b4caad9474dbf2a1e8c)

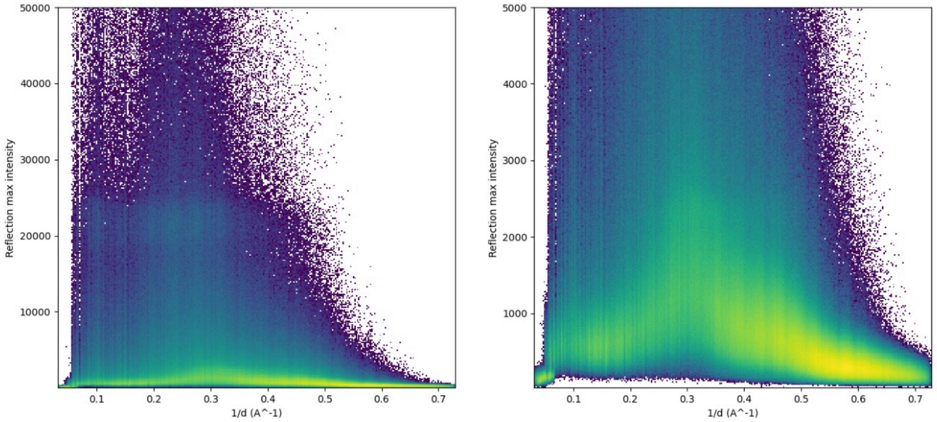
xwiz_lysozyme.toml

```
[data]
proposal = 1234
runs = [28, 33, 34]
list_prefix = "r0028_t12"

[geom]
file_path = "geom_with_masks.geom"

[proc_coarse]
resolution = 0.4
peak_method = "peakfinder8"
peak_threshold = 300
peak_snr = 7.0
peak_min_px = 1
peak_max_px = 20
peaks_hdf5_path = "entry_1/result_1"
index_method = "xgandalf"
```

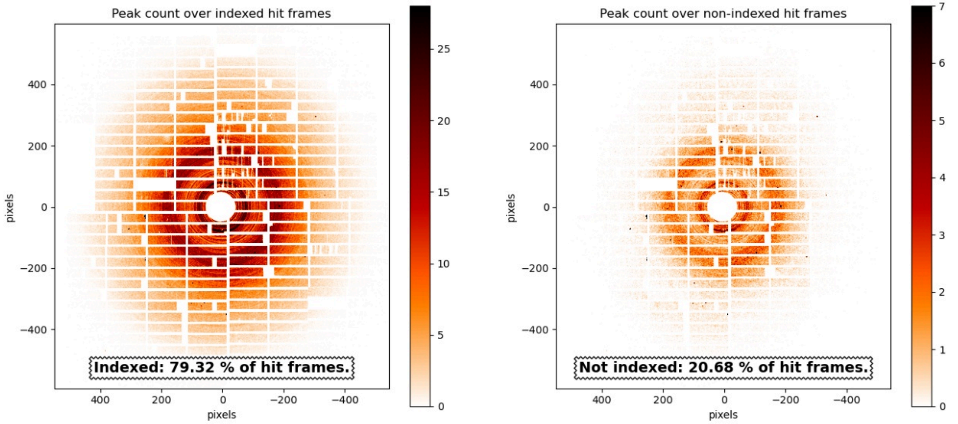
Also see poster 40, “Automatic data processing and results overview during SFX experiments” by O. Turkot.

Example: SFX

Example debugging plots:



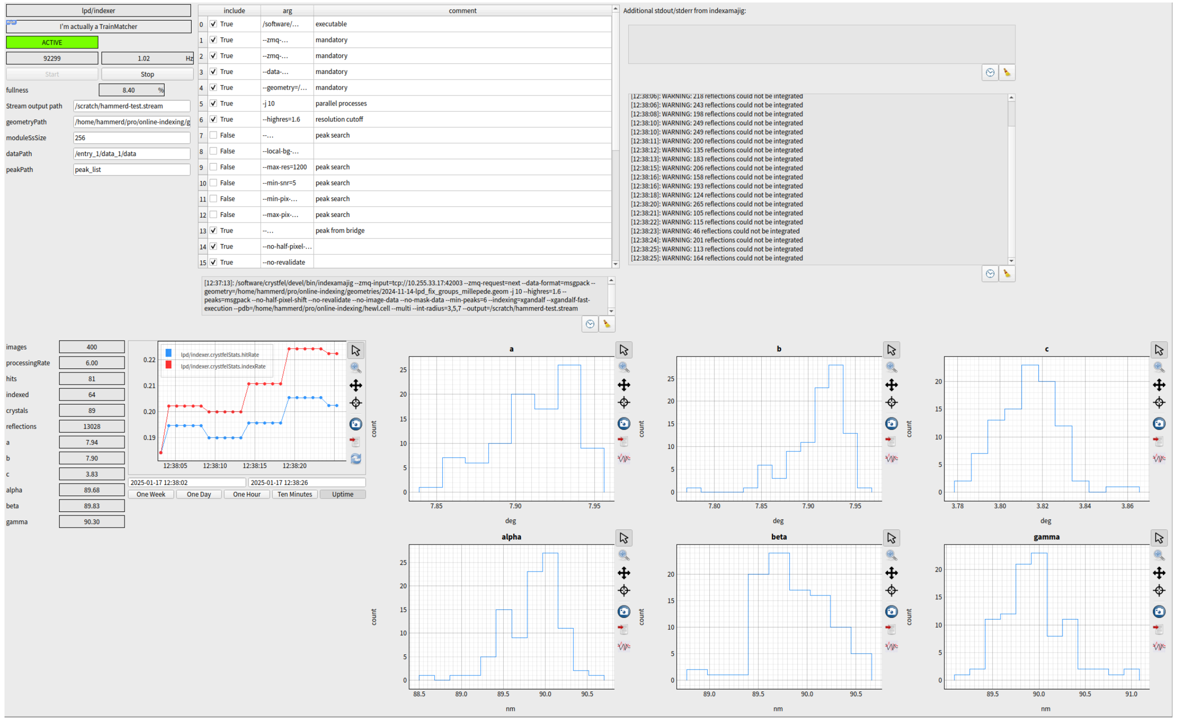
Peakogram



Peak counts for indexed/non-indexed frames

Example: SFX (online)

What we can currently do online: hit finding and indexing.



lpd/indexer

I'm actually a TrainMatcher

ACTIVE

92299 | 1.02 Hz

Start | Stop

fullness	8.40 %
Stream output path	/scratch/hammerd-test.stream
geometryPath	/home/hammerd/pro/online-indexing/g
moduleSsSize	256
dataPath	/entry_1/data_1/data
peakPath	peak_list

	include	arg	comment
0	☑ True	/software/…	executable
1	☑ True	--zmq-…	mandatory
2	☑ True	--zmq-…	mandatory
3	☑ True	--data-…	mandatory
4	☑ True	--geometry=/…	mandatory
5	☑ True	-j 10	parallel processes
6	☑ True	--highres=1.6	resolution cutoff
7	☐ False	--…	peak search
8	☐ False	--local-bg-…	
9	☐ False	--max-res=1200	peak search
10	☐ False	--min-snr=5	peak search
11	☐ False	--min-pix-…	peak search
12	☐ False	--max-pix-…	peak search
13	☑ True	--…	peak from bridge
14	☑ True	--no-half-pixel-…	
15	☑ True	--no-revalidate	

```
[12:37:13]: /software/crystfel/devel/bin/indexamajig --zmq-input=tcp://10.255.33.17:42003 --zmq-request=next --data-format=msgpack --
geometry=/home/hammerd/pro/online-indexing/geometries/2024-11-14-lpd_fix_groups_millepede.geom -j 10 --highres=1.6 --
peaks=msgpack --no-half-pixel-shift --no-revalidate --no-image-data --no-mask-data --min-peaks=6 --indexing=xgandalf --xgandalf-fast-
execution --pdb=/home/hammerd/pro/online-indexing/hewl.cell --multi --int-radius=3,5,7 --output=/scratch/hammerd-test.stream
```

Additional stdout/stderr from indexamajig:

```
[12:38:06]: WARNING: 218 reflections could not be integrated
[12:38:06]: WARNING: 243 reflections could not be integrated
[12:38:08]: WARNING: 198 reflections could not be integrated
[12:38:10]: WARNING: 249 reflections could not be integrated
[12:38:10]: WARNING: 249 reflections could not be integrated
[12:38:11]: WARNING: 200 reflections could not be integrated
[12:38:12]: WARNING: 135 reflections could not be integrated
[12:38:13]: WARNING: 183 reflections could not be integrated
[12:38:15]: WARNING: 206 reflections could not be integrated
[12:38:16]: WARNING: 158 reflections could not be integrated
[12:38:16]: WARNING: 193 reflections could not be integrated
[12:38:18]: WARNING: 124 reflections could not be integrated
[12:38:20]: WARNING: 265 reflections could not be integrated
[12:38:21]: WARNING: 105 reflections could not be integrated
[12:38:22]: WARNING: 115 reflections could not be integrated
[12:38:23]: WARNING: 46 reflections could not be integrated
[12:38:24]: WARNING: 201 reflections could not be integrated
[12:38:25]: WARNING: 113 reflections could not be integrated
[12:38:25]: WARNING: 164 reflections could not be integrated
```

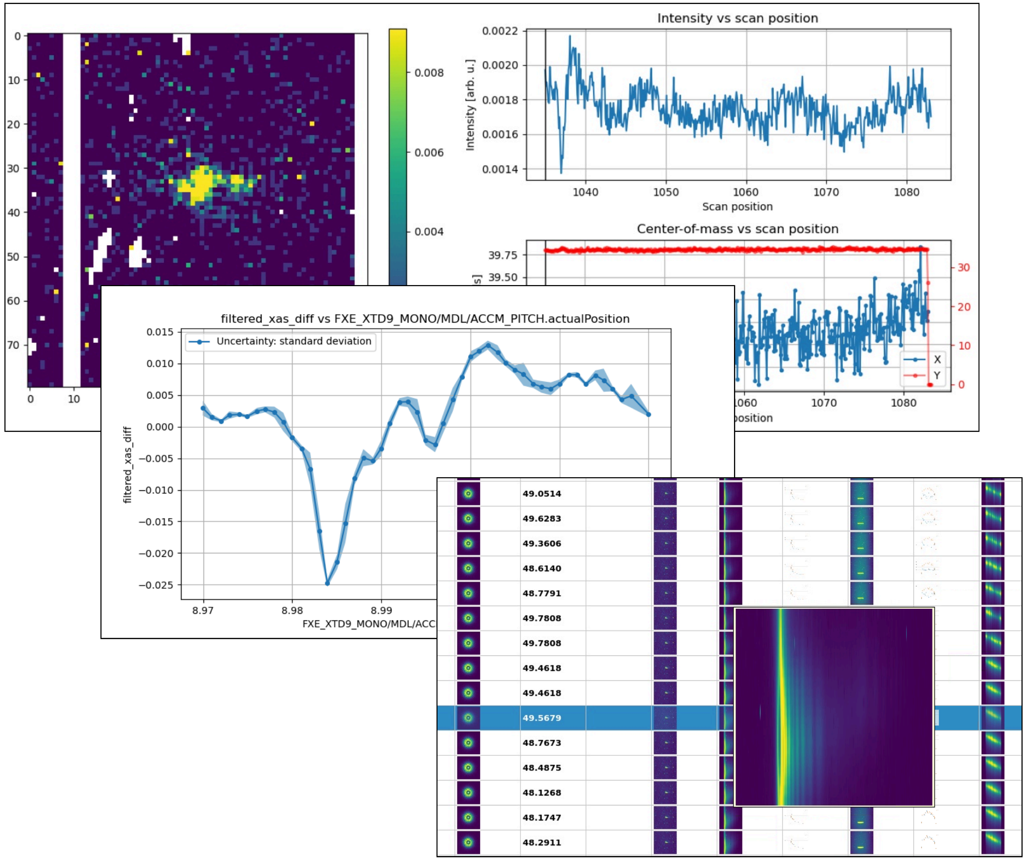
images	400
processingRate	6.00
hits	81
indexed	64
crystals	89
reflections	13028
a	7.94
b	7.90
c	3.83
alpha	89.68
beta	89.83
gamma	90.30

2025-01-17 12:38:02 | 2025-01-17 12:38:26

Others

These techniques are in various stages of maturity and coverage:

- 🟧 Bragg XPCS
- 🟧 Single particle imaging
- 🟧 Single crystal diffraction
- 🟧 X-ray absorption spectroscopy
- 🟧 X-ray emission spectroscopy
- 🟧 X-ray scattering



49.0514
49.6283
49.3606
48.6140
48.7791
49.7808
49.7808
49.4618
49.4618
49.5679
48.7673
48.4875
48.1268
48.1747
48.2911

Conclusion

- This only works by collaborating with amazeballs users and instrument scientists <3

Conclusion

- This only works by collaborating with amazeballs users and instrument scientists <3
- Analysis at our facility is probably never going to be quite as easy as at a synchrotron, but we can get close.

Conclusion

- This only works by collaborating with amazeballs users and instrument scientists <3
- Analysis at our facility is probably never going to be quite as easy as at a synchrotron, but we can get close.

Questions?

(or contact us at `da@xfel.eu`)