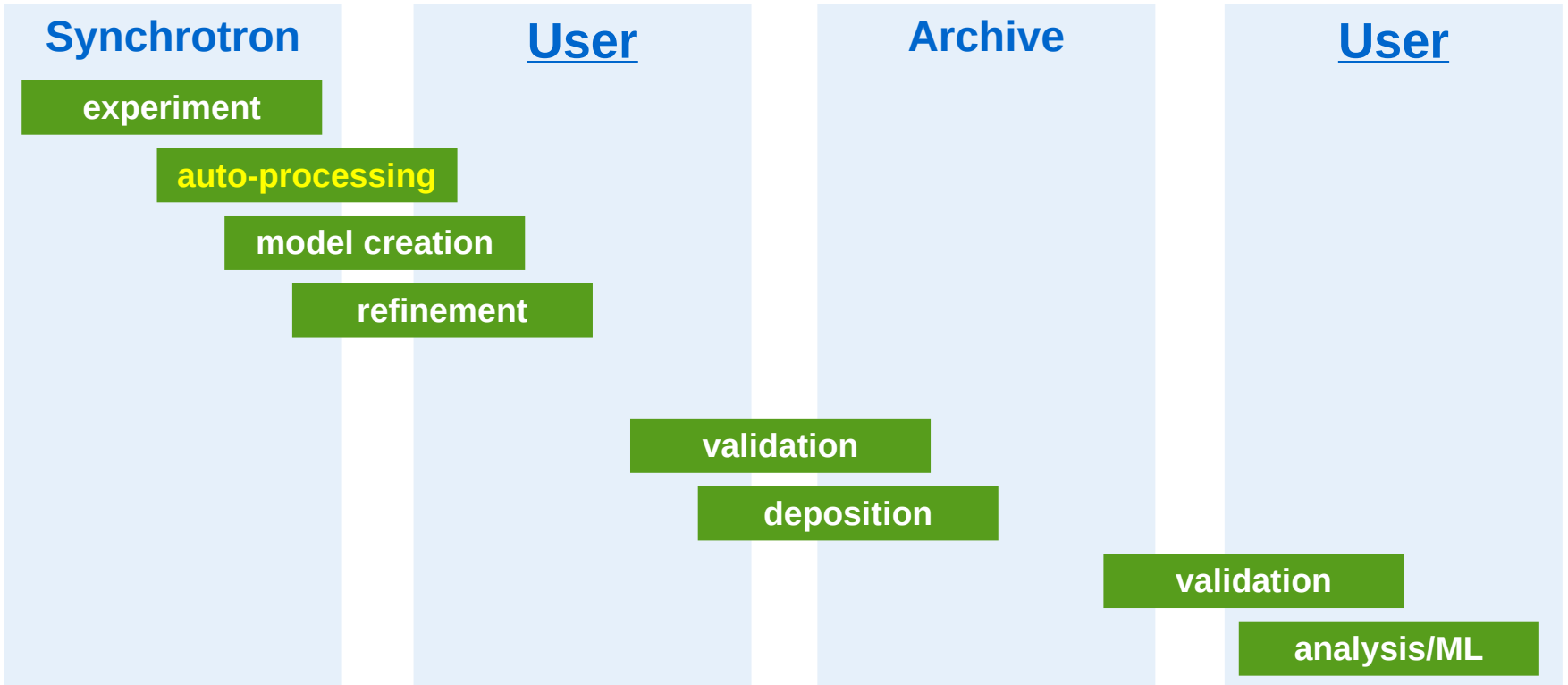
GΦL
Global Phasing Limited

Standardising (auto-)processing and access to its results while avoiding Procrustean beds

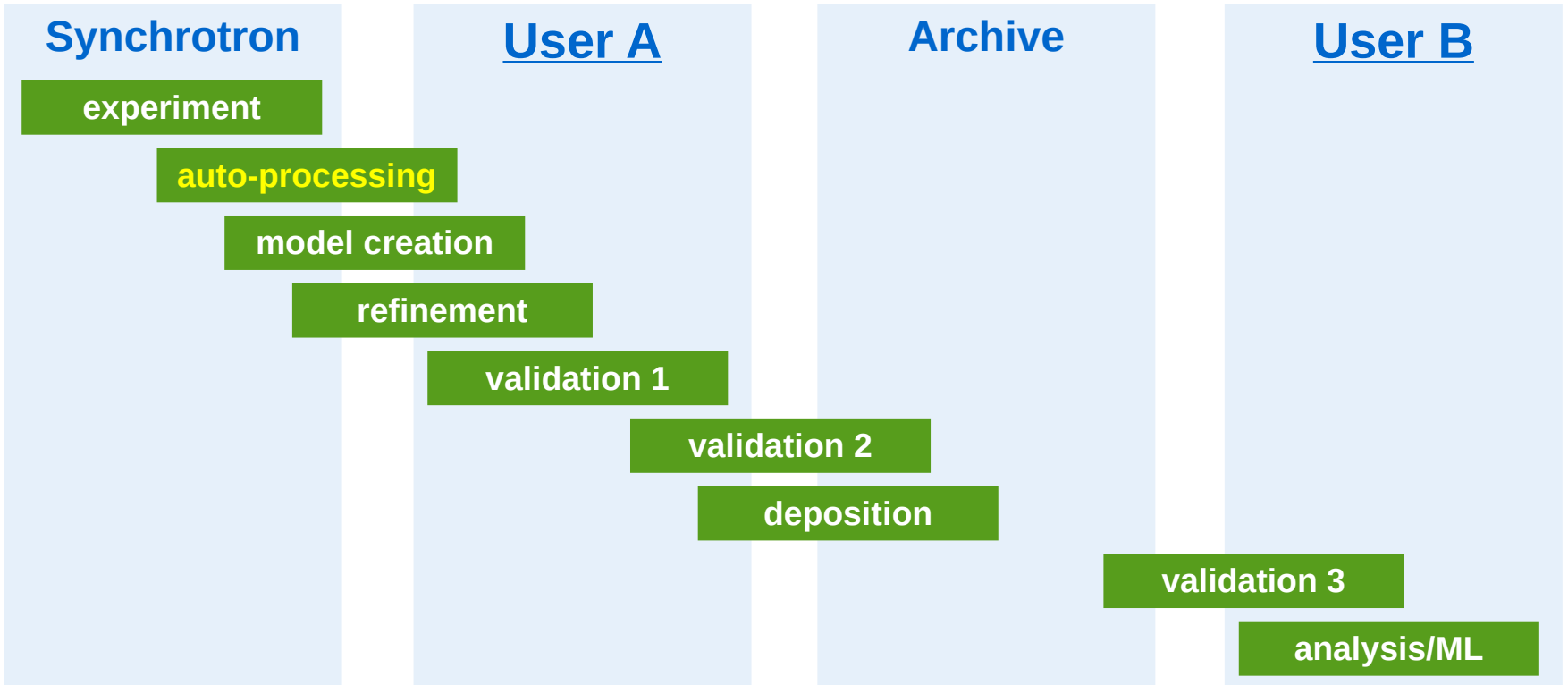
MXCuBE/ISPyB Scientific Day

Nov 18, 2025

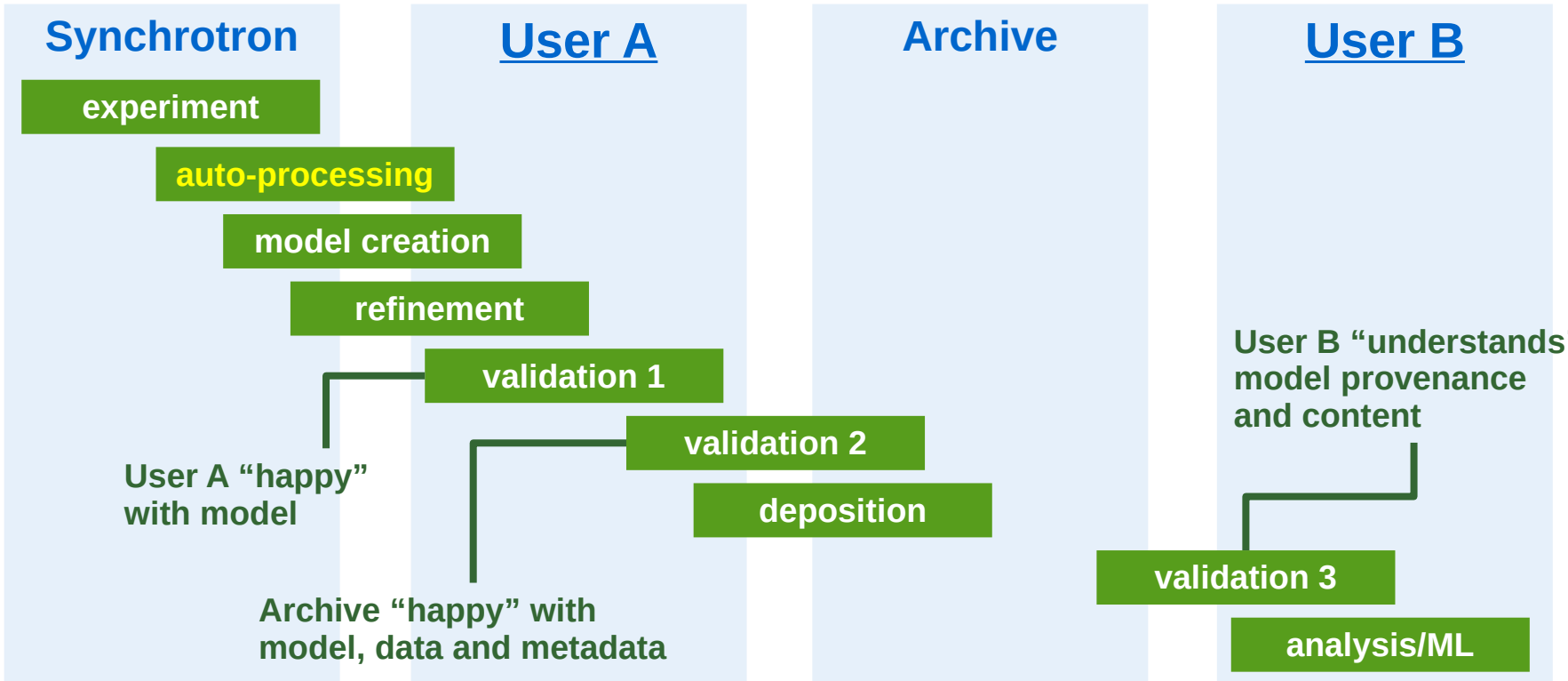
A possible (simple) view from user(s) perspective



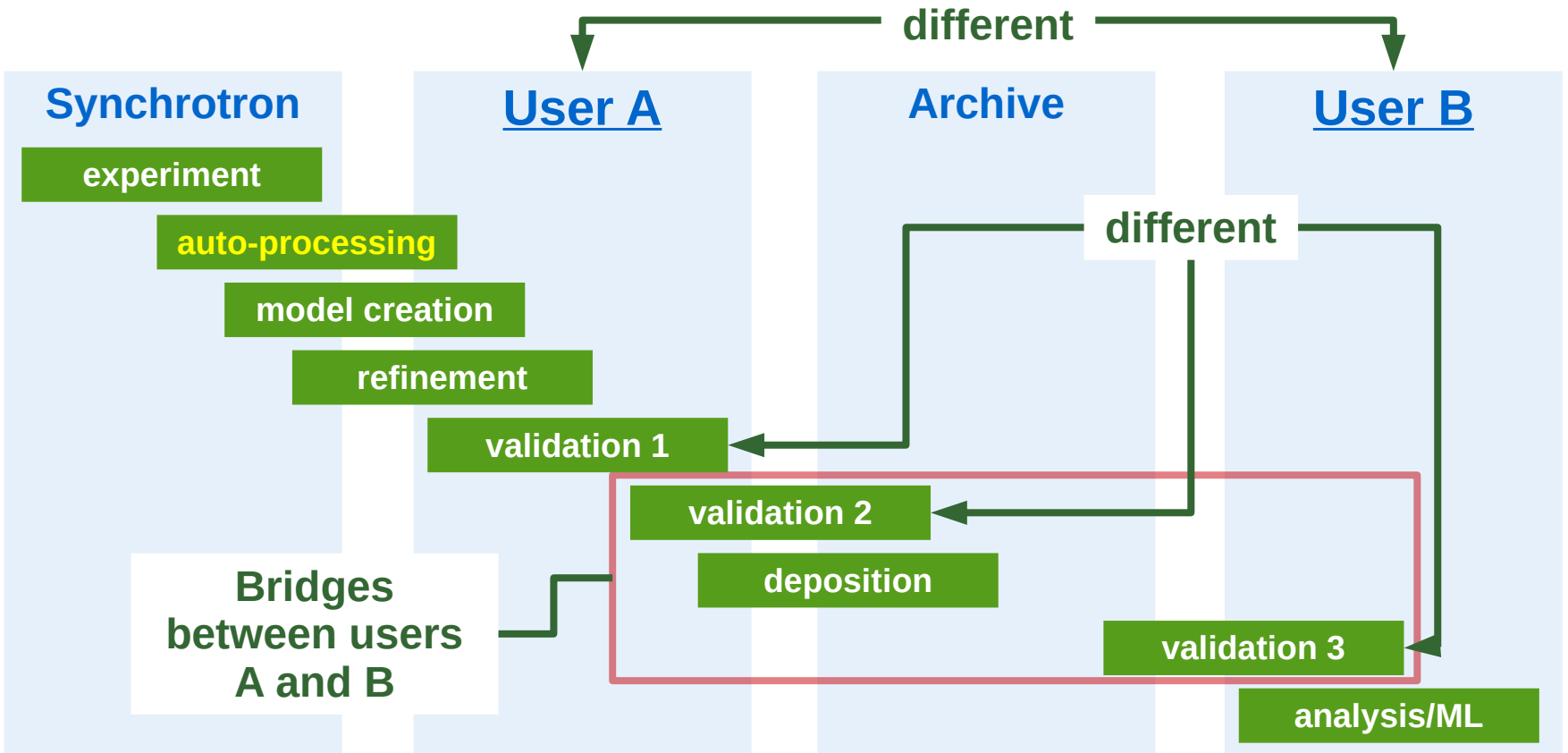
Some small complication



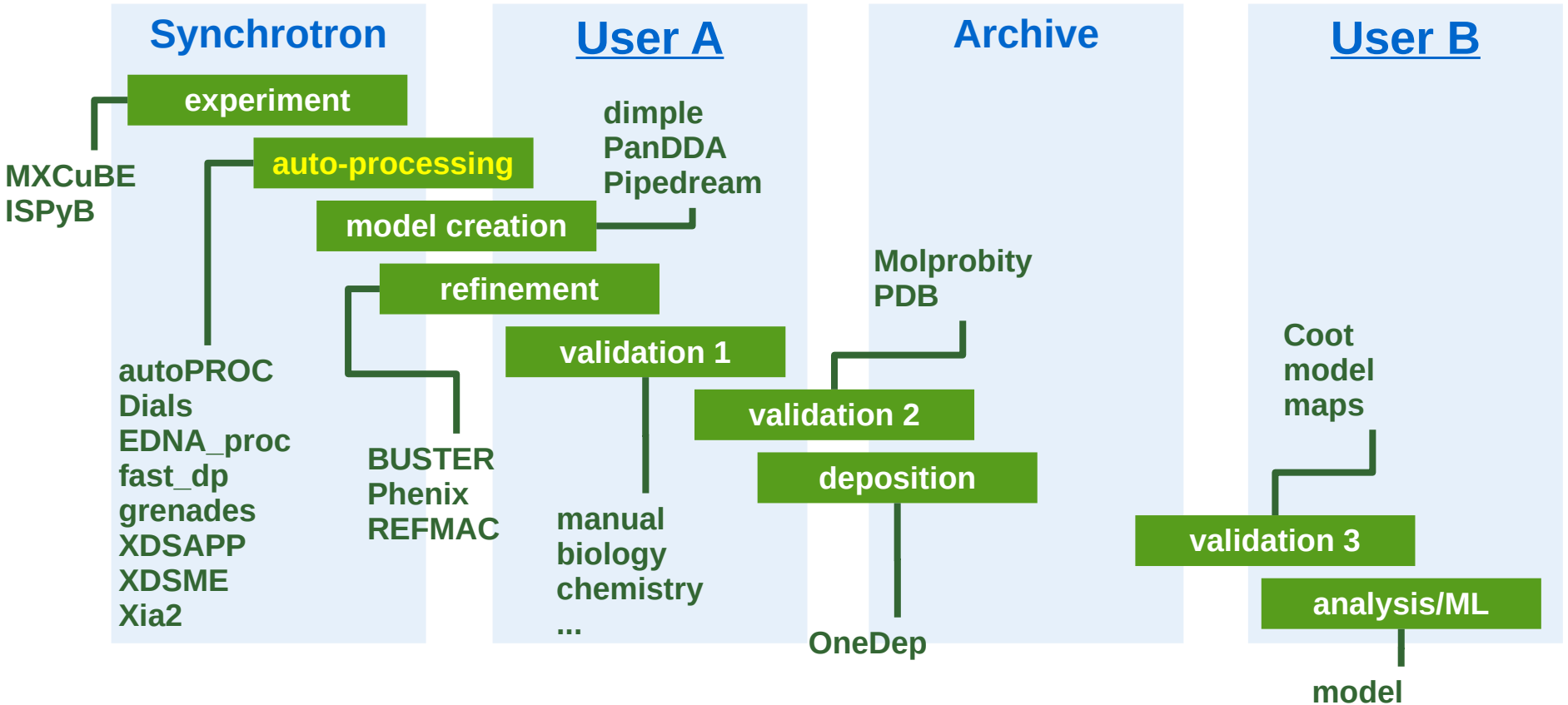
Some small complication



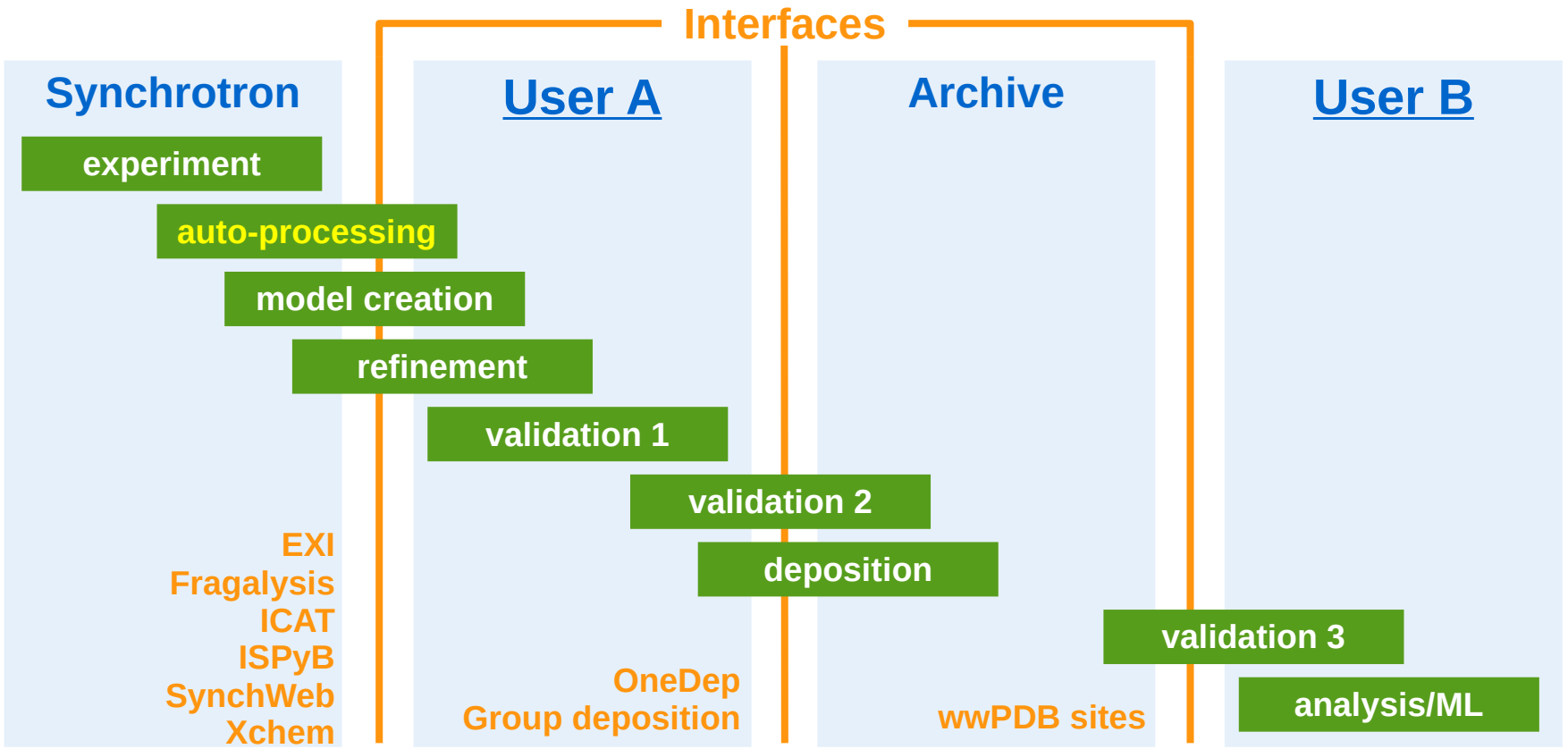
Some small complication



Diverse software ecosystem



Multiple systems and players



Standardisations ...

- Standardisation of raw data:
  - "**transferability of processing**"
  - data producer (**user A**) can process data seemlessly at home ... N years later
  - archive user (**user B**) can re-process data at any time in the future
- Standardisation of auto-processing pipelines run at synchrotrons:
  - running each pipeline optimally
  - "**does exactly what it says on the tin**"
  - at least for those processing packages that have generic releases available to anyone
- Standardising presentation of pipeline results to users:
  - **taking advantage of each pipeline's strength**
  - data producer (**user A**) can make informed decision during the experiment (fast feedback)
  - data producer (**user A**) can decide to take auto-processed reflection data (or not)
  - if at some point also processing results are made public: **user B** can see full details of auto-processing

Transferability of processing

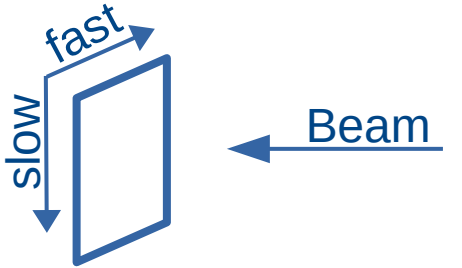
It's all in the metadata:

- **mini-cbf** and **HDF5** the de-facto standards for MX diffraction data
  - **mini-cbf** are simple (**one file per image**) for processing, with a short ASCII header - but lack organisation of multiple related datasets apart from file naming conventions or directory structures
    - great working format
  - **HDF5** are more complicated for processing (XDS plugins, separate pixel mask, various compression filters) - but **rich in metadata** possibilities (e.g. ASCII/UTF8 variable/fixed size strings null-padded or not)
    - great metadata and archiving format
- For "transferability" of processing we require a **complete description of the instrument and experiment** in a format widely supported by different processing and helper packages.

Instrument description

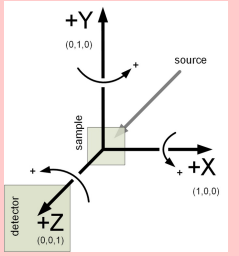
Right-hand rules for coordinate system and rotation axis

A “natural” coordinate system could start with the **detector axes** and then define **goniostat** accordingly:



detector X axis = (1, 0, 0)
detector Y axis = (0, 1, 0)
incident beam = (0, 0, 1)

Any reference coordinate system could do (e.g. in XDS), but **HDF5 files following the “HDRMX Gold Standard” (i.e. NeXus NXmx) need to use the McSTAS coordinate system**.



https://doi.org/10.1107/S2052252520008672
https://manual.nexusformat.org/design.html#the-nexus-coordinate-system
https://manual.nexusformat.org/classes/applications/NXmx.html

mini-cbf headers

```
# Wavelength 1.3418 A
# Detector_distance 0.07000 m
# Beam_xy (698.0,548.0) pixels
# Detector_2theta -0.00000 deg.
# Phi 114.43561 deg.
# Phi_increment 0.00000 deg.
# Omega 214.43735 deg.
# Omega_increment -0.20000 deg.
# Kappa -70.52043 deg.
# Kappa_increment 0.00000 deg.
# Oscillation_axis OMEGA
# Rotation_axis_vector 0.0 1.0 0.0
# Start_angle 214.43735 deg.
# Angle_increment -0.20000 deg.
# Detector_fast_axis_vector 1.0 0.0 0.0
# Detector_slow_axis_vector 0.0 1.0 0.0
# Incident_beam_vector 0.0 0.0 1.0
# Omega_axis_vector 0.0 1.0 0.0
# Kappa_axis_vector 0.0 0.64279 0.76604
# Phi_axis_vector 0.0 1.0 0.0
# 2Theta_axis_vector 0.0 1.0 0.0
```

- **v2.0 (26th Oct 2021)** specification added full instrument definition to mini-cbf headers: work between Rigaku, Global Phasing and Dectris
- This allows e.g. autoPROC to read a full instrument and experiment specification (as needed for processing) - similar to what HDF5 (“Gold Standard”, NXmx) provides.
- If providing mini-cbf: think of moving to v2.0 if possible!
- Avoids guesswork ...

Running each pipeline optimally

- Some are **intended for fast feedback**, e.g. fast_dp:
  - spot search using a small subset of images
  - allows 50% unindexed spots in single IDXREF
  - runs single INTEGRATE in P1
  - merges P1-integrated data in most likely SG
- Some are trying to get **best data with as much analysis as possible**, e.g. autoPROC:
  - spot search using all images
  - iterative indexing (detect multiple lattices, ice-rings)
  - first INTEGRATE in P1, SG determination and re-running INTEGRATE with most likely SG, updated parameters, better mosaicity estimate etc
  - Scaling in AIMLESS and analysis with STARANISO
  - HTML, PDF, PDBx/mmCIF, lost of plots and explanations
- **Different purposes at different times** (decision making while collecting data, or taking data into refinement and ultimately OneDep deposition)

Avoiding lowest common denominator

- Different auto-processing pipelines have different emphasis, different feature sets and different “added value”:
  - **also useable offline at home**: autoPROC, Dials, fast_dp, XDSAPP, XDSme, Xia2, ...
  - purely onsite: EDNA-proc, grenades, ...
- Each pipeline should ideally be **run as intended by the pipeline developers**:
  - running in non-default or not recommended mode will give wrong impression to users
  - if things behave poorly, users will blame the pipeline and their developers (and not the synchrotron/beamline/IT)
- Scraping logfiles is to be avoided - especially for pipelines that produce rich metadata in standard formats (**ISPyB-compatible XML**, **PDBx/mmCIF**)
  - if something is missing/incorrect: better to fix at source
  - looking at logfiles error-prone and potentially completely wrong
  - rushed patches have a tendency to stay for decades

ANOM/NOANOM - confused historical baggage

- One can **always** output anomalous reflection data I(+)/SIGI(+), I(-)/SIGI(-) and DANO/SIGDANO alongside IMEAN/SIGIMEAN at the final merging step

- Special treatment of anomalous data during scaling (to maximise the anomalous signal) and/or outlier rejection (to avoid rejecting large difference measurements) only makes sense with **very large anomalous signal**, **high multiplicity and an explicit phasing experiment**:
  - as a default for any experiment it never made a lot of sense to me
  - in the age of AlphaFold (and MR) this should **definitely not be a default**

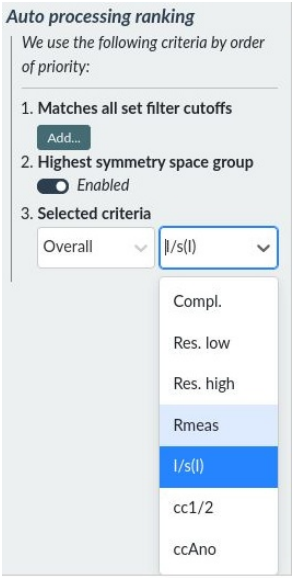
- Beware: FRIEDEL'S_LAW= FALSE in XDS changes the definition of a "unique reflection" for correction factors as well as completeness, R-values,(CC1/2) statistics etc (in CORRECT and XSCALE):
  - we might get lower completeness, lower I/sigI (merged reflections), lower R-values, higher ISa (unmerged reflections) in CORRECT.LP
  - the statistics in CORRECT/XSCALE pretend that reciprocal space has no inversion centre
  - what we are interested in: describing the data used downstream - ultimately in refinement, i.e. IMEAN/SIGIMEAN. And those will be more accurate when FRIEDEL'S_LAW= TRUE.
  - we are not trying to push one or several metrics into a more favourable region (high ISa deemed good, lower Rmeas better etc)
  - MRFANA unaffected: the definitions are not changed
- Solution:
  - **always use FRIEDEL'S_LAW= FALSE** in XDS/XSCALE pipelines (apart from XDSCONV if that is used to merge data and go from intensities to amplitudes)
  - or: use a program like MRFANA to compute all merging statistics consistently (well defined definitions, control over binning etc)

Standardised data (quality) descriptors

- based on **scaled+unmerged** reflection data
  - **after** outlier/misfits removal
  - measurements that go into inverse-variance weighted merging
  - **XDS_ASCII.HKL**, XSCALE.HKL, **unmerged MTZ** from AIMLESS or dials.scale
- Traditionally done in **bins** (resolution shells):
  - **correct comparison between pipelines would require identical binning**
    - not possible for Overall and Outer shells if using scaled+unmerged data after applying a data cutoff (since each pipeline might employ a different method for deciding on those cut-offs - for very good reasons)
    - always possible for low-resolution bin (but be aware of beamstop masking differences): could “standardise” on a resolution range?
    - some tricky details (resolution depends on unit cell - and since each processing will result in a slightly different unit cell, slightly different Miller indices will make it into a specific bin ... or not)
  - even if overcoming those difficulties: one can always sort pipeline results (numerical comparisons are neutral) - but this requires a single value to sort on ... and assigning a preference to one over the other is misleading:
    - is (**CC1/2=0.999, <I/sigI=22.4>**) better than (**CC1/2=0.998, <I/sigI=22.5>**)?
- Sorting/labeling pipeline results is extremely complicated:
  - probably better to follow the “neutral” DLS approach: first come, first serve (i.e. sorted by “speed of results”)

Global Phasing Limited

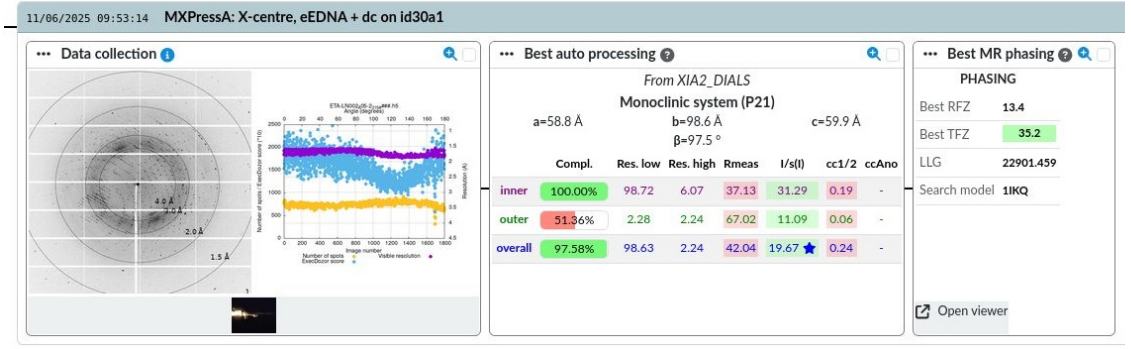
Synchrotron to User A interfaces - ESRF

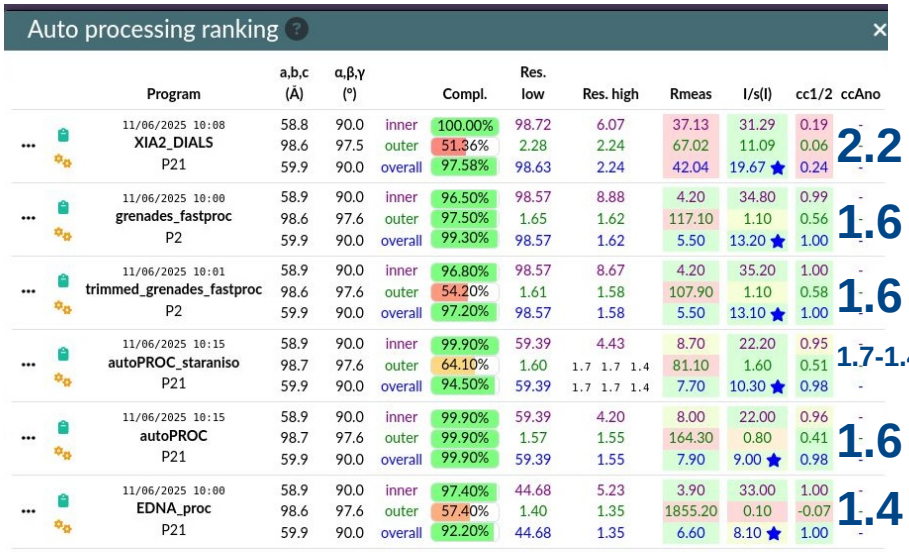


Is the need to provide a “Best auto processing” annotation driven by user request or by internal accounting needs ... or just historical baggage?

Overall, inner- or outer-shell statistics often influenced e.g. by binning, smoothing, ice-rings and anisotropy.

As far as we can see, a user can’t select a combination of criteria (as has been possible in MRFANA since 2010).





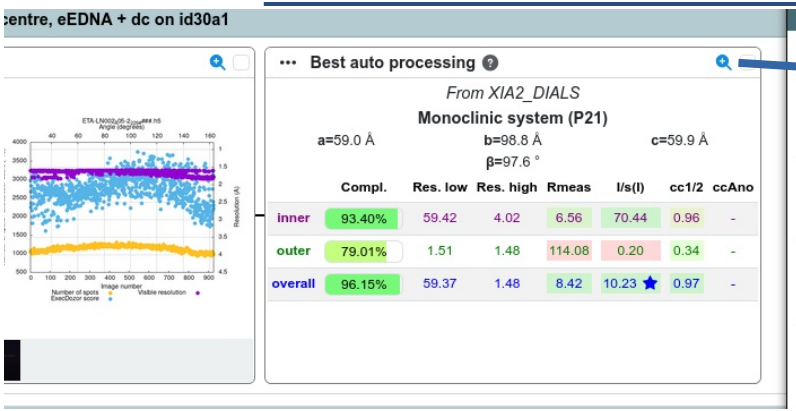
The “best” auto processing is the one that extracts all available signal accurately.

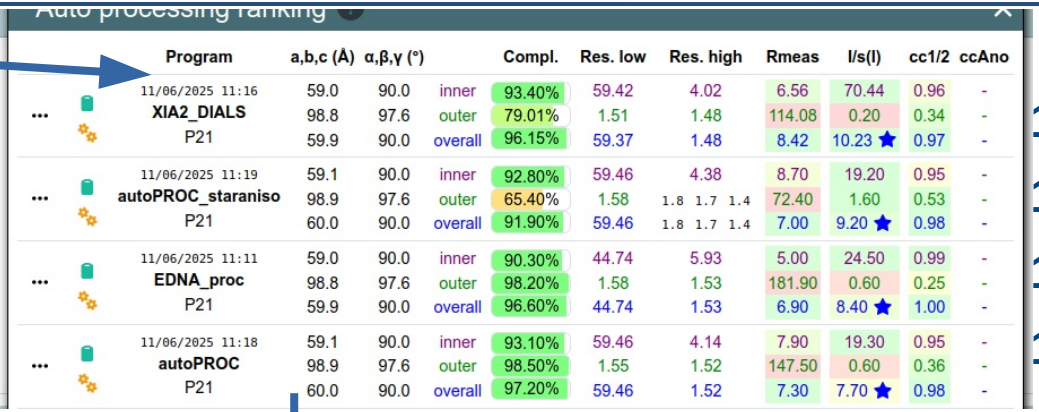
How does one measure that?

We have access to a small number of synchrotron interfaces - thanks for that possibility!

GΦL
Global Phasing Limited

User guidance: "best" data processing





1.5 Å
1.8 - 1.4 Å
1.5 Å
1.5 Å

"Operational resolution"

- How many merged reflections with signal (I/sig(I)>=2)?
- What sphere in reciprocal space would they fill (for given crystal symmetry)?
- What is the radius of that sphere?

#	pipeline	opres
01	XIA2_DIALS	2.039
02	autoPROC_staraniso	1.809
03	EDNA_proc	1.831
04	autoPROC	**1.799**

R/Rfree

#	no LL-removal	with LL-removal
01	0.2082/0.2397	0.1952/0.2251
02	**0.1806/0.2082**	**0.1790/0.2065**
03	0.1835/0.2150	0.1828/0.2141
04	0.1820/0.2108	0.1803/0.2083

BUSTER/aB_autorefine with same (sub)set of reflections

MX & CryoEM - Complimentary methods

total number of PDB entries = 241922 (Sep 2025)

X-Ray crystallography	**= 197707 (81.7%)**
Cryo-EM	**= 28918 (12.0%)**
Electron diffraction	= 273 (0.1%)
NMR	= 14421 (6.0%)

total number of PDB entries with the concept of "resolution" (X-Ray, cryo-EM and ED) = 226898

		X-Ray			cryo-EM			ED		
Resolution	**#PDB**	**#PDB**	**%total**	**%method**	**#PDB**	**%total**	**%method**	#PDB	%total	%method
- 4.0	7212	1304	18.1	0.7	5882	81.6	20.3	26	0.4	9.5
4.0 - 3.0	29339	13940	47.5	7.1	15363	52.4	53.2	36	0.1	13.2
3.0 - 2.5	38694	32709	84.5	16.5	5945	15.4	20.6	40	0.1	14.7
2.5 - 2.0	61089	59534	97.5	30.1	1510	2.5	5.2	45	0.1	16.5
2.0 - 1.5	68286	68055	99.7	34.4	195	0.3	0.7	36	0.1	13.2
1.5 - 1.0	21090	21011	99.6	10.6	11	0.05	0.04	68	0.3	24.9
1.0 -	1173	1153	98.3	0.6	0	0.0	0.0	20	1.7	7.3
Total	226883	197706 (=87.1%)			28906 (=12.7%)			271 (=0.001%)		

MX & CryoEM - Complimentary methods

total number of PDB entries = 241922 (Sep 2025)

X-Ray crystallography	**= 197707 (81.7%)**
Cryo-EM	**= 28918 (12.0%)**
Electron diffraction	= 273 (0.1%)
NMR	= 14421 (6.0%)

total number of PDB entries with the concept of "resolution" (X-Ray, cryo-EM and ED) = 226898

		X-Ray			cryo-EM			ED		
Resolution	**#PDB**	**#PDB**	**%total**	**%method**	**#PDB**	**%total**	**%method**	#PDB	%total	%method
- 4.0	7212	1304	18.1	0.7	5882	81.6	20.3	26	0.4	9.5
4.0 - 3.0	29339	13940	47.5	7.1	15363	52.4	53.2	36	0.1	13.2
3.0 - 2.5	38694	32709	84.5	16.5	5945	15.4	20.6	40	0.1	14.7
2.5 - 2.0	61089	59534	97.5	30.1	1510	2.5	5.2	45	0.1	16.5
2.0 - 1.5	68286	68055	99.7	34.4	195	0.3	0.7	36	0.1	13.2
1.5 - 1.0	21090	21011	99.6	10.6	11	0.05	0.04	68	0.3	24.9
1.0 -	1173	1153	98.3	0.6	0	0.0	0.0	20	1.7	7.3
Total	226883	197706 (=87.1%)			28906 (=12.7%)			271 (=0.001%)		

MX & CryoEM - Complimentary methods

total number of PDB entries = 241922 (Sep 2025)

X-Ray crystallography	**= 197707 (81.7%)**
Cryo-EM	**= 28918 (12.0%)**
Electron diffraction	= 273 (0.1%)
NMR	= 14421 (6.0%)

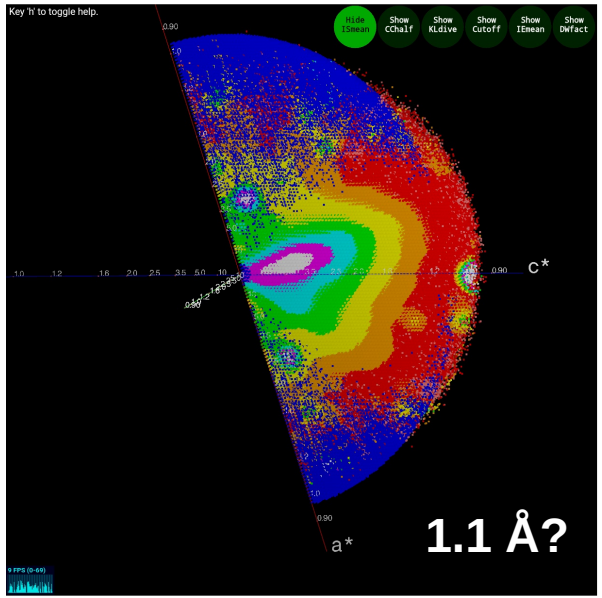
Snapshot of *current* state of *depositions*!

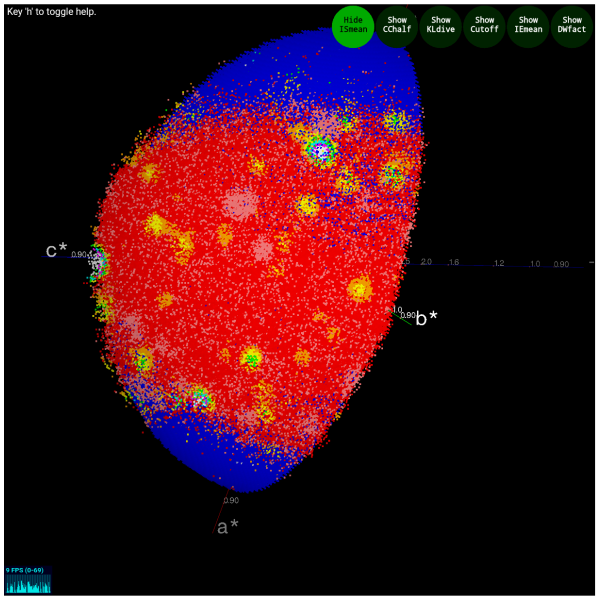
total number of PDB entries with the concept of "resolution" (X-Ray, cryo-EM and ED) = 226898

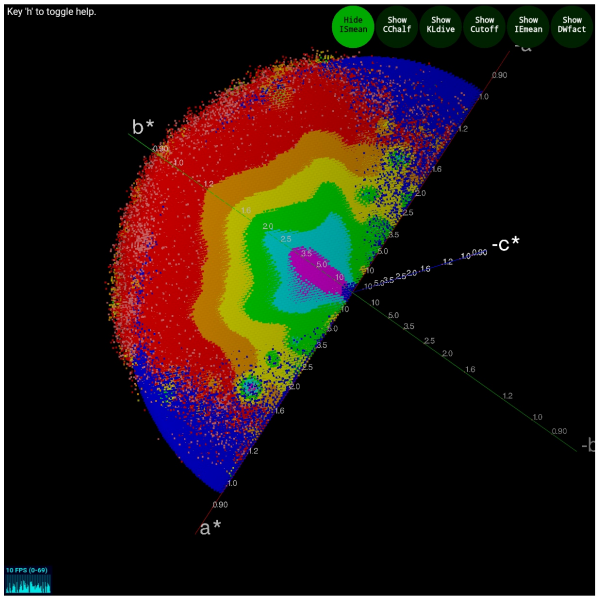
		X-Ray			cryo-EM			ED		
Resolution	**#PDB**	**#PDB**	**%total**	**%method**	**#PDB**	**%total**	**%method**	#PDB	%total	%method
- 4.0	7212	1304	18.1	0.7	5882	81.6	20.3	26	0.4	9.5
4.0 - 3.0	29339	13940	47.5	7.1	15363	52.4	53.2	36	0.1	13.2
3.0 - 2.5	38694	32709	84.5	16.5	5945	15.4	20.6	40	0.1	14.7
2.5 - 2.0	61089	59534	97.5	30.1	1510	2.5	5.2	45	0.1	16.5
2.0 - 1.5	68286	68055	99.7	34.4	195	0.3	0.7	36		
1.5 - 1.0	21090	21011	99.6	10.6	11	0.05	0.04	68		
1.0 -	1173	1153	98.3	0.6	0	0.0	0.0	20		
Total	226883	197706 (=87.1%)			28906 (=12.7%)			271		

In X-Ray crystallography (MX) we are looking for detailed chemical information with high accuracy.

Latest “high resolution” micro ED SSX (9FY7)







Everything is tuned towards achieving that “high resolution” label:

- As a proxy for “high quality”?
- Is that the tail wagging the dog?

Impacting standardisation target

- Auto-processing **pipelines are different**
  - for very good reasons
  - standardisation does not mean making them similar again
  - each pipeline should behave as intended by its developers
- **Associating a label** (“best”) to auto-processing results **is complicated**
  - neutral sorting seems better
  - “operational resolution”
- Chasing the “high resolution” **badge**
  - much more complex than just scraping a value out of a logfile
  - introduces a lot of bias (and tendency to brush the ugly bits under the carpet)
- Devil in the **details**
  - Synchrotron-agnostic developers/experts can provide added value