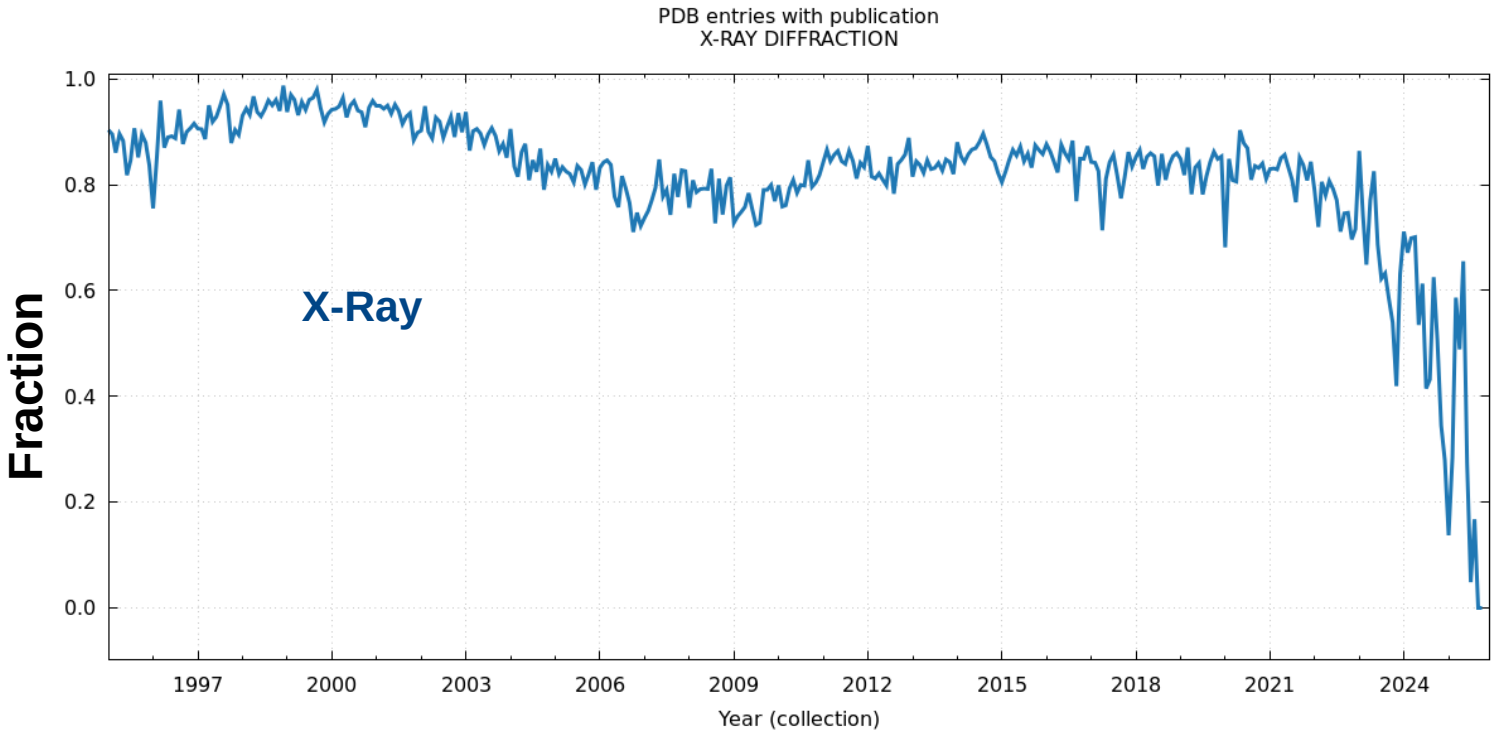
Metadata quality assurance in (group) depositions

ISPyB/MXCuBE meeting

Nov 17 - 19 2025

Fewer depositions come with primary publication

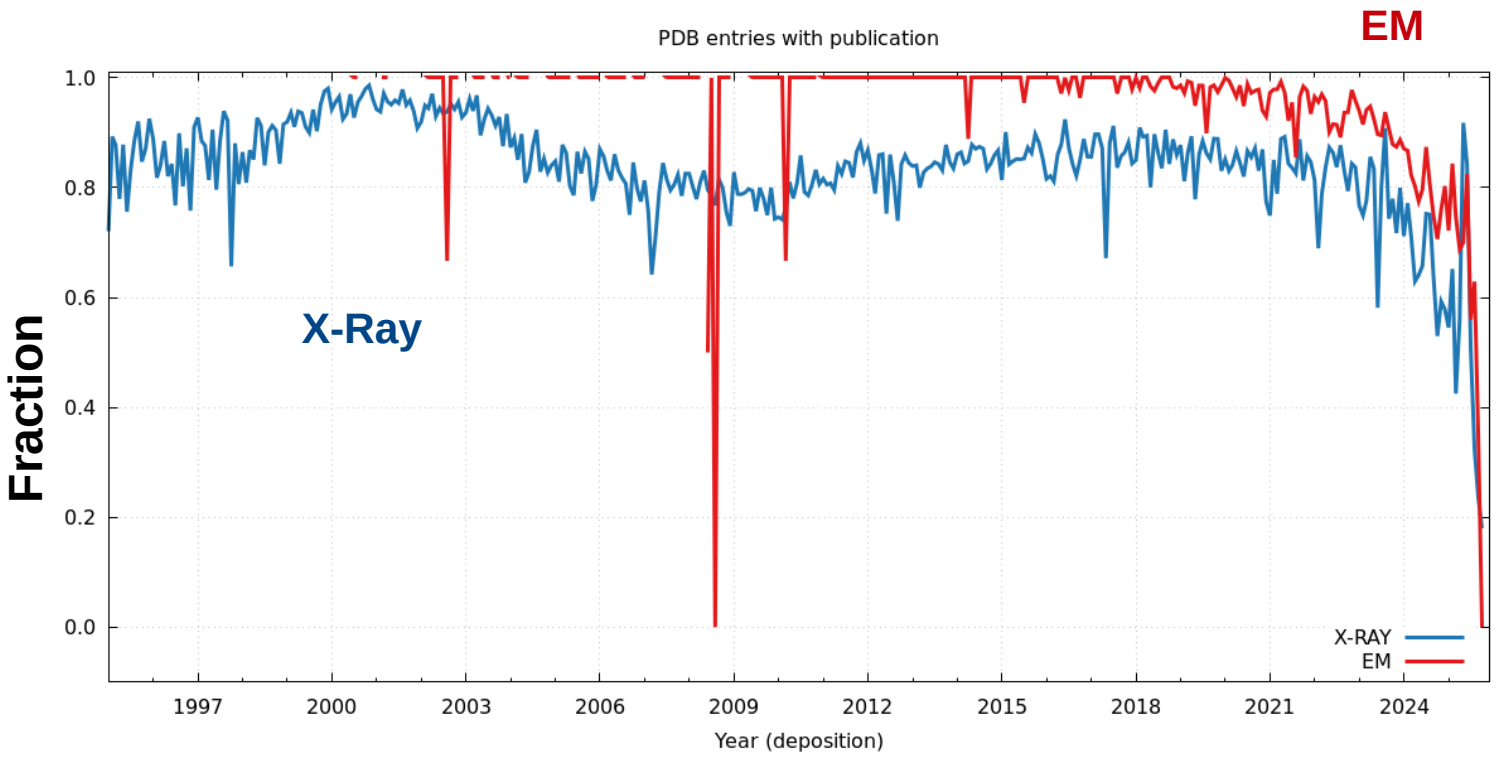
Collection Date



If we don't have a paper to help us understand the model in the context of the underlying data, the metdata has to fulfil that role.

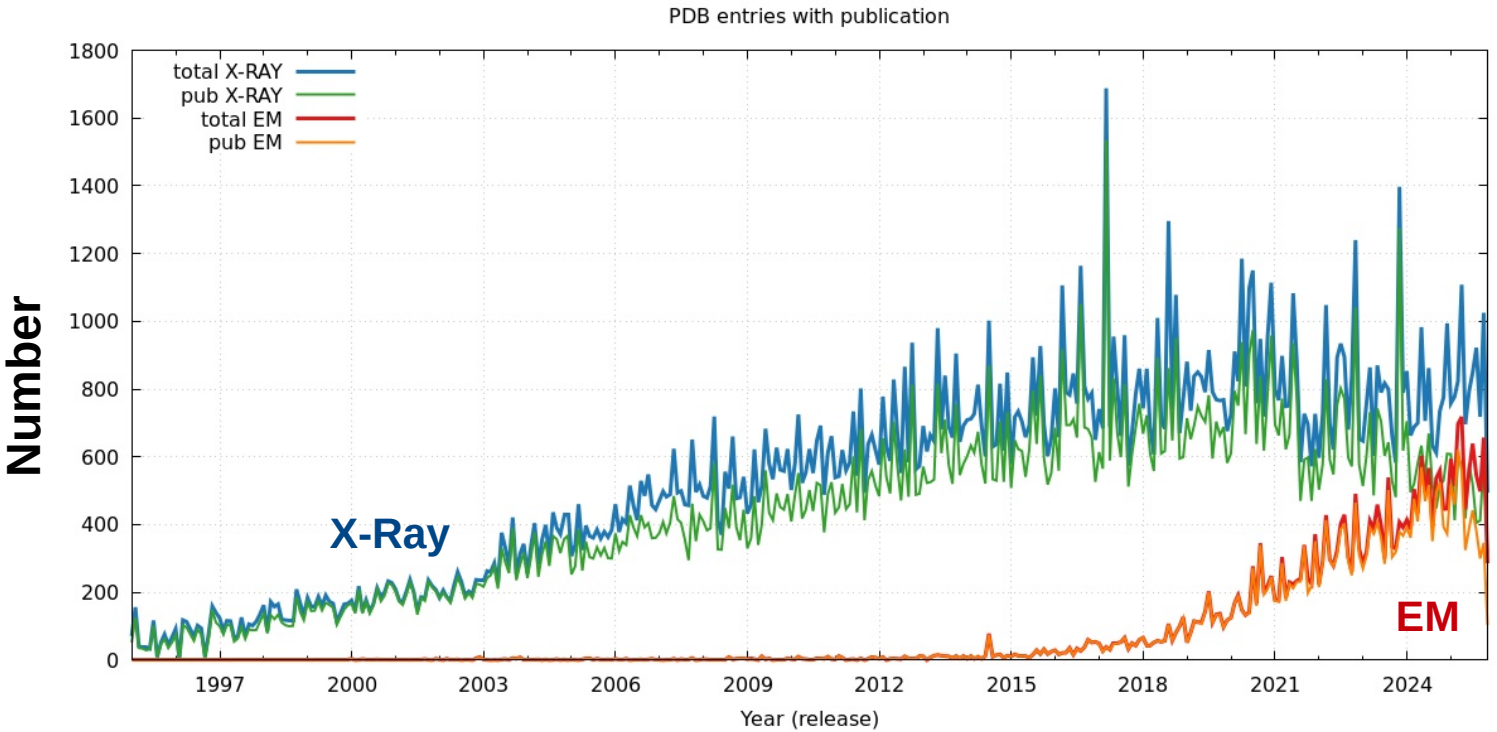
... also for EM

Deposition Date

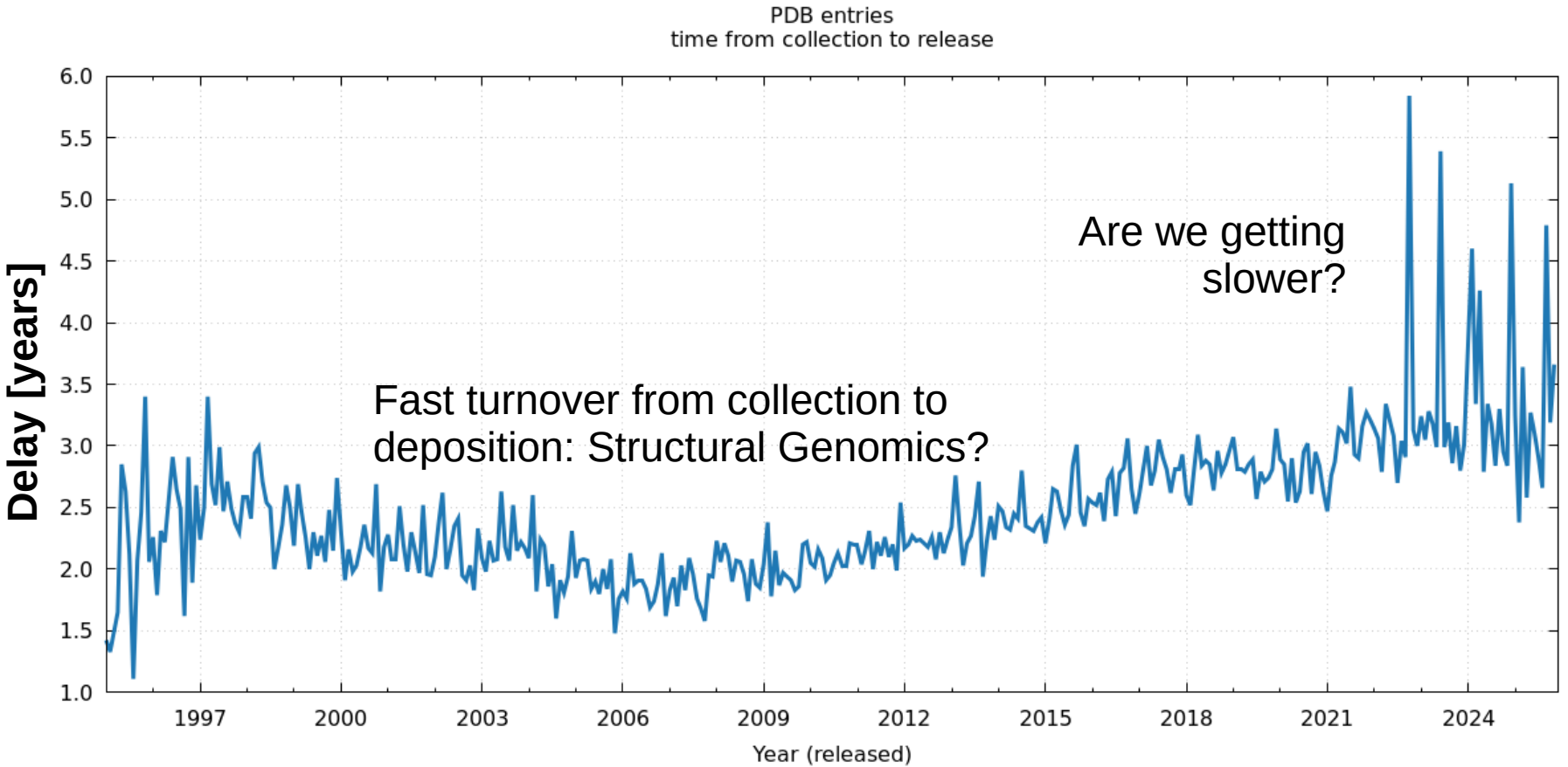


Fewer depositions come with primary publication

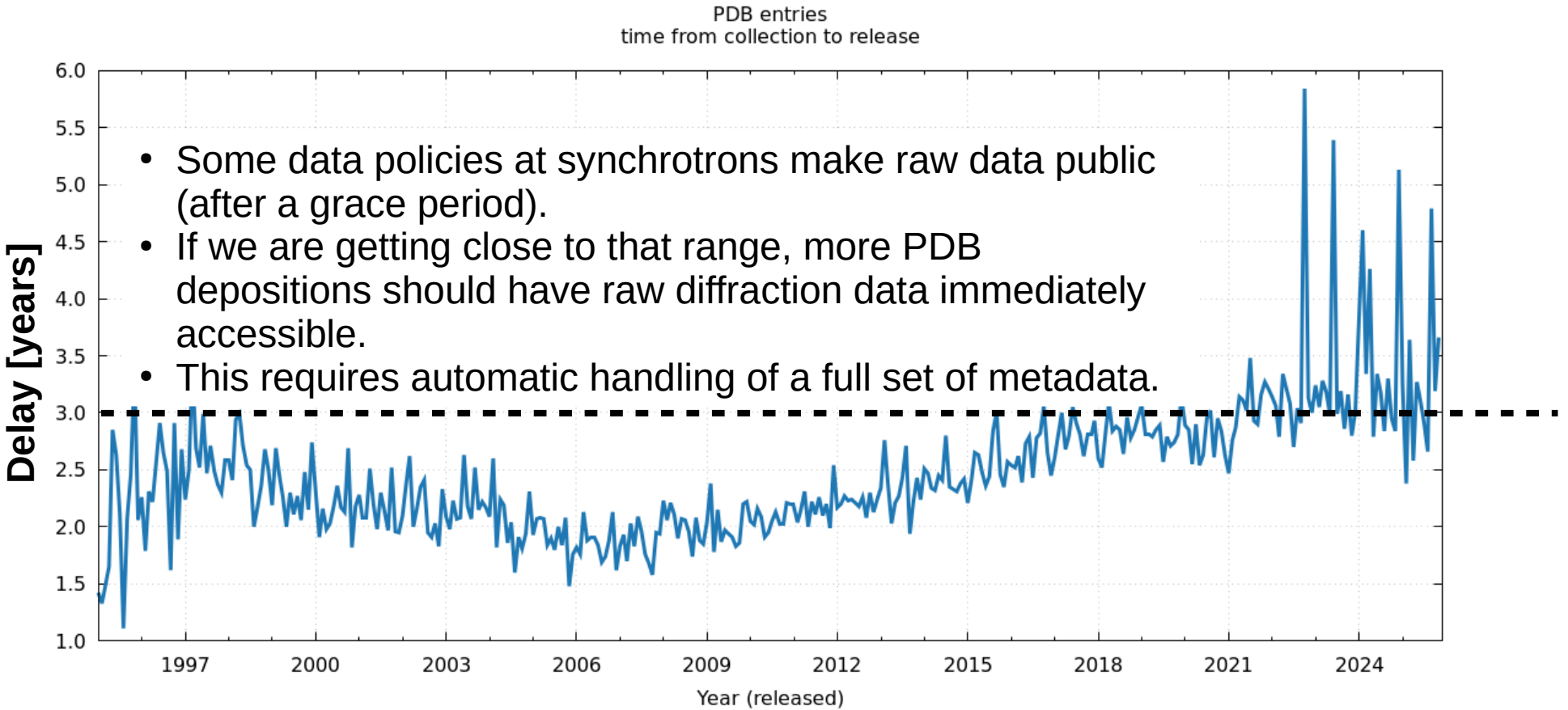
Release Date



Gap between data collection and deposition is increasing



Problem #2: Gap between data collection and deposition is increasing



- Some data policies at synchrotrons make raw data public (after a grace period).
- If we are getting close to that range, more PDB depositions should have raw diffraction data immediately accessible.
- This requires automatic handling of a full set of metadata.

Metadata helps to understand (model | data)

- Looking at "X-RAY DIFFRACTION" PDB entries where the structure title contains "PanDDA" or "fragment screen" or that are part of a group deposition.
  - In those we check if "_diffrn.details" (reflection data) contains the strings "PanDDA", "event" or "evidence" (all case insensitive).
- This gives us **3413 reflection datablocks in 2948 PDB entries** (PDB archive as of 6th February 2025).
  - This should constitute a set of PDB entries where we have access to the original compound/fragment evidence.
- 2670 (**78 %**) datablocks have only FWT/PHWT columns (map coefficients)
- 726 (21 %) have FP, SIGFP and PHWT columns (probably FP is the amplitude to use for map calculation - not sure what SIGFP represents)
- remaining (0.5 %) look mostly like incorrect descriptions (and data is actually something else entirely)

Map computation possible?

Metadata helps to understand (model | data)

- 114 (3 %) datablocks describe ground-state: "data for reference ground-state (PanDDA style mean map)"
- 110 (3 %) with "data for z-score based deviation from reference data (PanDDA style z-map)"
- 3183 (93 %) with "data for bound-state (PanDDA style event map)" or "data for ligand evidence map (PanDDA event map)"

Type of map?

- **These 3183 "PanDDA event map" datablocks (2940 PDB entries) include 84 group depositions (2939 PDB entries), where 51 (61 %) have no publication associated with them (1267 PDB entries, 43 %).**
- Fragment screening campaigns (PanDDA-style) are nearly always deposited via "**group depositions**" - or: when deposited using "group depositions" the event maps are also included.
- Nearly half of PDB entries associated with fragment screening and providing event maps come without associated publication.
- For those 1267 PDB entries (1319 event maps), the **deposited data/metadata are the only available evidence** to check the presence, placement and modeling of ligands.

SG/cell differences

- **Event maps are not crystallographic objects**, i.e. they don't obey symmetry and cell repeat. In that sense they are much more like EM maps.
- Storing map coefficients (only available method to provide a 3D map for "X-RAY DIFFRACTION" PDB entries) requires care so that standard map computation (e.g. in Coot or CCP4/FFT) will show the full map over the whole molecule(s).
- This implies the generation of a **P1 cell** covering the deposited model (which is a crystallographic object) and the event map - including some boundary buffer. Transforming that P1 cell map into map coefficients (amplitude and phase) would give the required reflection data for one particular event.
- Originally, **each event map is associated with a particular contouring value** to show the evidence as used by the authors. When using a reconstituted event map (via map coefficients from an artificial and enlarged P1 cell), that level is not obvious to users: the **map "rms" is now meaningless** since it takes e.g. also the buffer region into account and covers something different from an asymmetric unit or unit cell.
- A deposition needs to provide additional information so that one can see the same event map at the same level for each modeled ligand instance: one can’t use a standard 3·rms level or such.

Taking the first PDB entry (7HLA) of the group deposition (G_1002320) with the most entries (109) for which no publication is available:

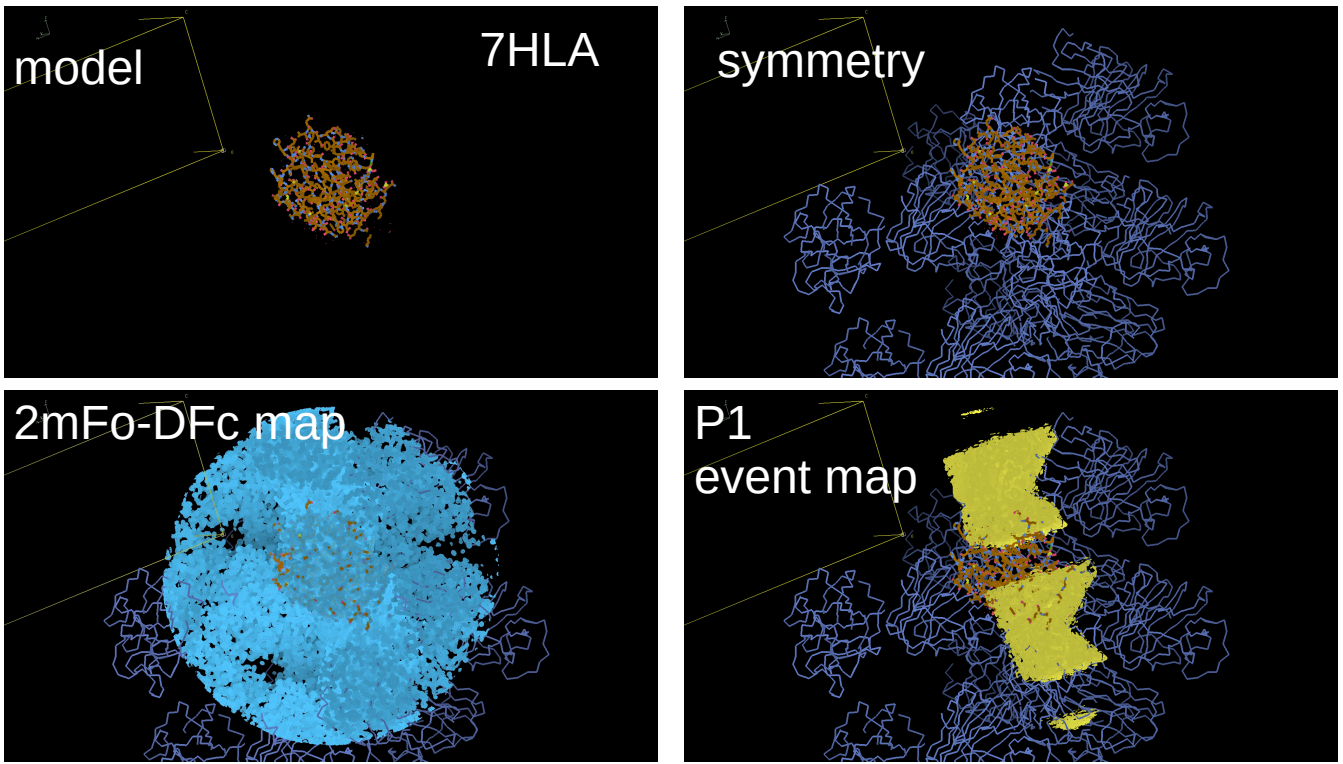
- Fetching model and reflection data and converting into MTZ ("fetch_PDB_gemmi")

```
r7hlasf    95.716 95.716 45.745 90.0 90.0 90.0 I4   data from final refinement with ligand, final.mtz
r7hlaAsf   95.716 95.716 45.745 90.0 90.0 90.0 I4   data from original reflections, data.mtz
r7hlaBsf   95.716 95.716 45.745 90.0 90.0 90.0 P1   data for ligand evidence map (PanDDA event map), event_map_1.mtz
```

- Can we visualise the event map? Either directly using the MTZ file in Coot, or by computing the map (FFT: LABIN F1=FWT PHI=PHWT) and extending it over the molecule (MAPMASK: BORDER 5)?

Event maps from reflection data

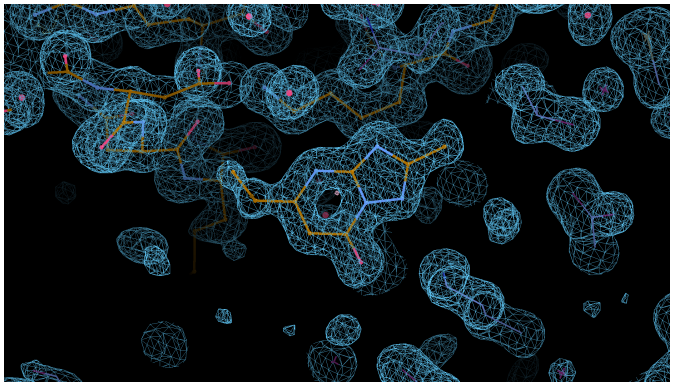
- 786 (60 %) of event maps are in P1
  - 747 (95 %) have same cell dimensions as model data

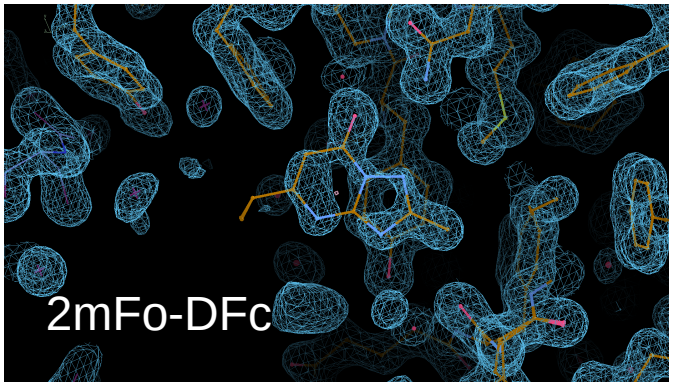


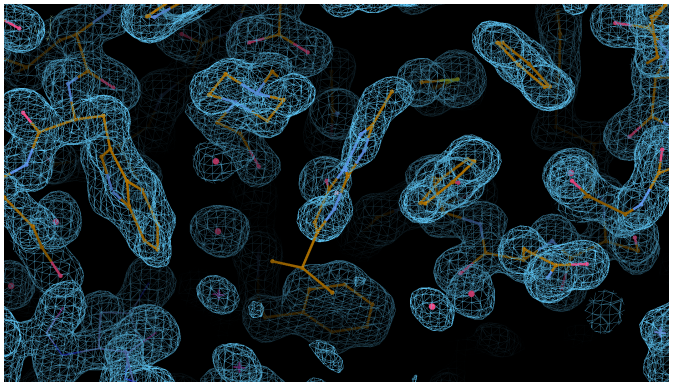
GΦL
Global Phasing Limited

Event maps from reflection data

7HLA

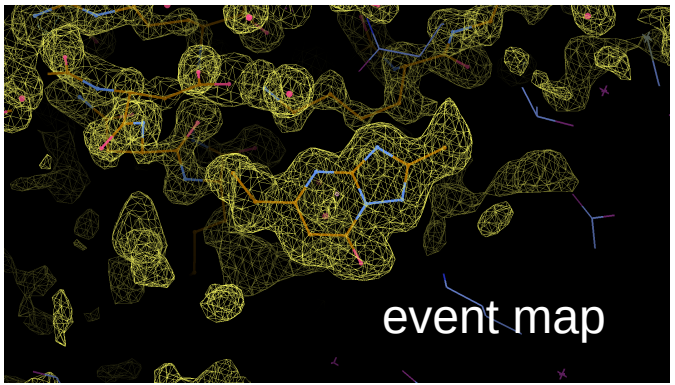


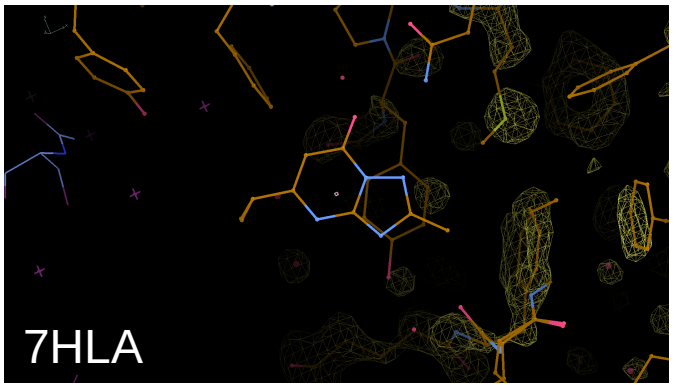


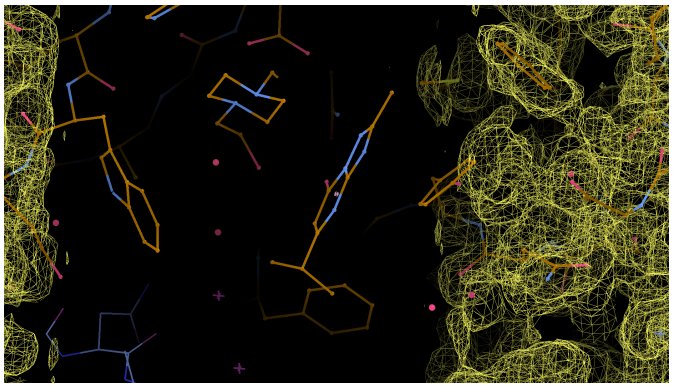


Binding site 1

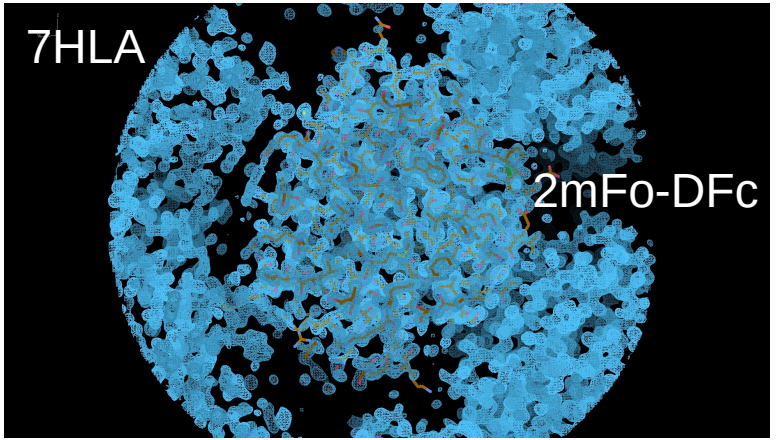
Binding site 2

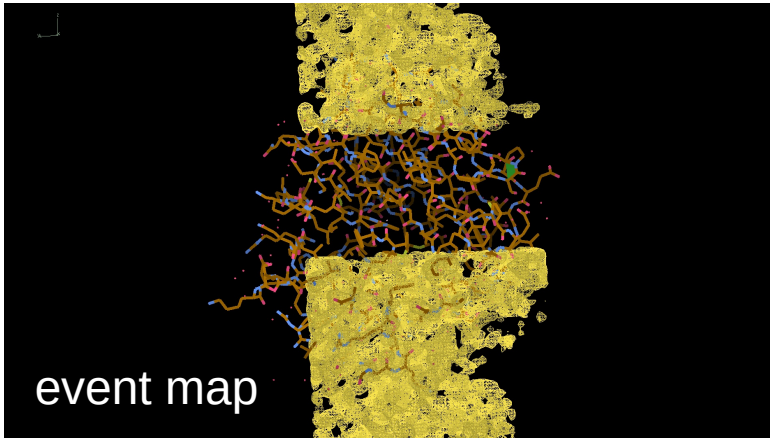






Event maps from reflection data



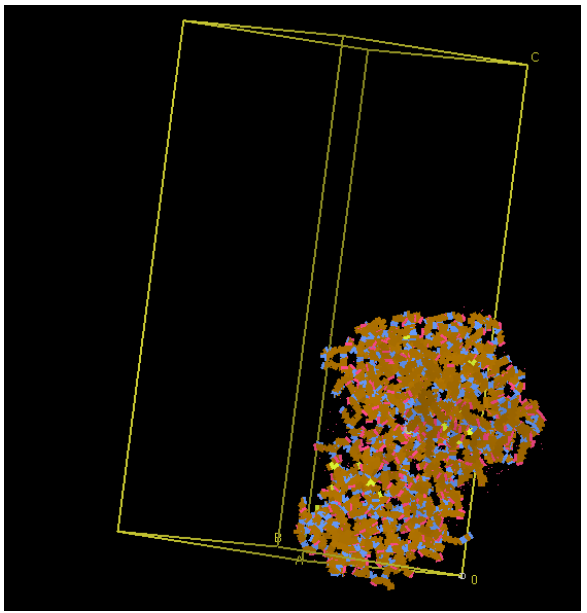


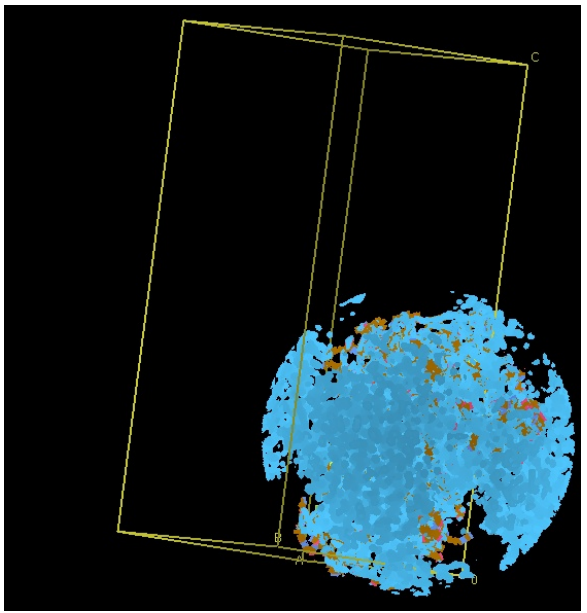
- **Reflection data is intrinsically a crystallographic object.**
- 2mFo-DFc map behaves as expected.

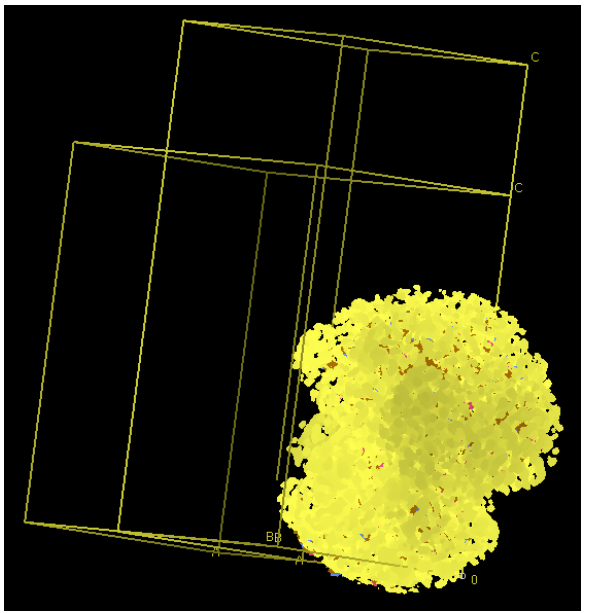
- **Event maps are not crystallographic** and to represent them as reflection data requires care.
- This is very much like NCS or cross-crystal averaging and the know-how of experts in that field should be made better use of.

Event maps from reflection data

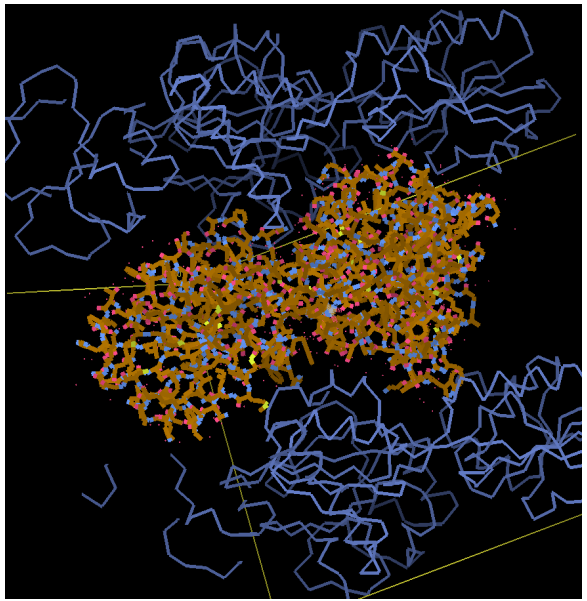
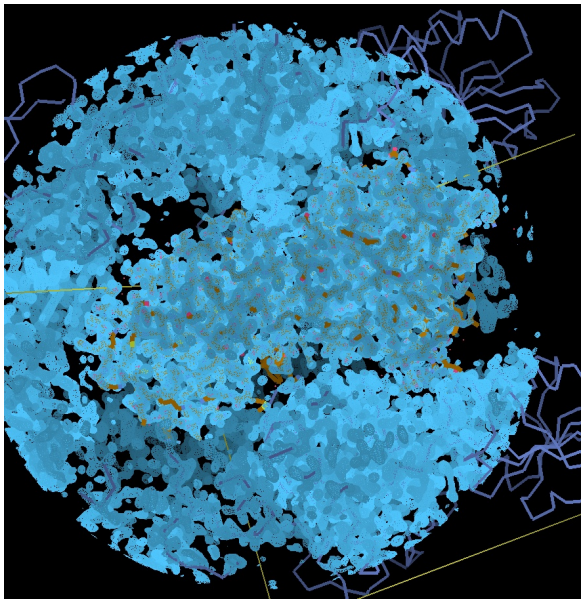
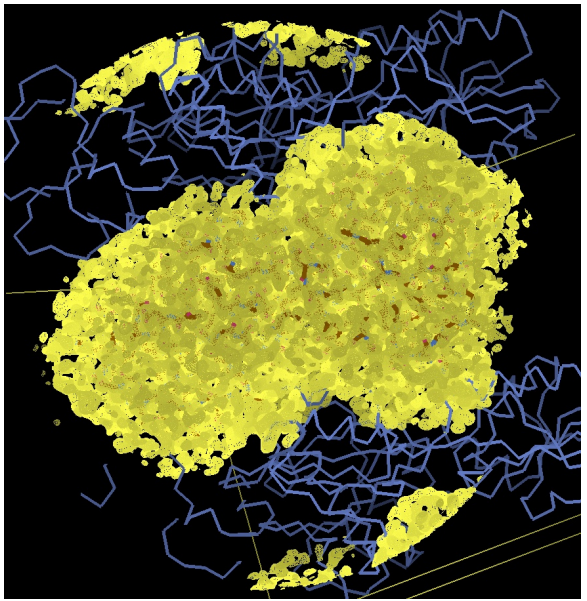
- 786 (60 %) of event maps are in P1
  - 747 (95 %) have same cell dimensions as model data
  - 40 (5 %) have cell different (>5A in each axis) from model - looking at 5Q1J:







Event maps from reflection data

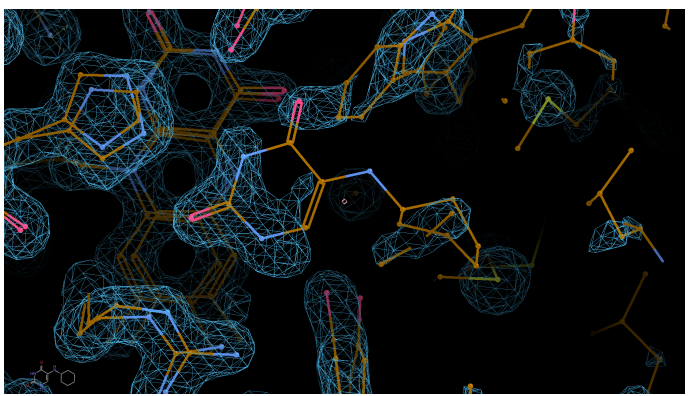




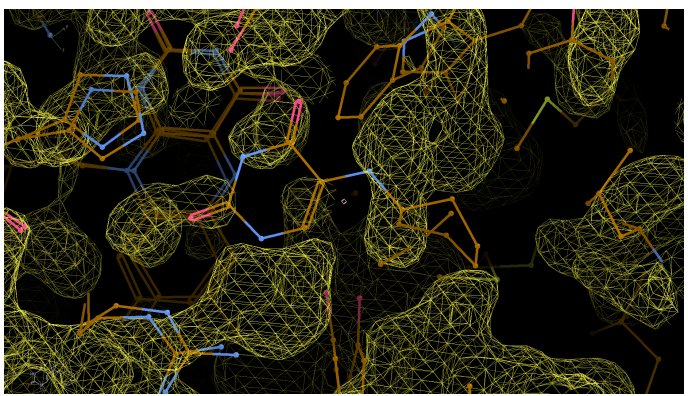
- Event map covers molecule (good), but doesn't follow symmetry of model (unavoidable).
- A change of atom position by symmetry/cell-repeat is allowed and equivalent - but reflection data representing event map would have to be re-created.

Event maps from reflection data

- 786 (60 %) of event maps are in P1
  - 747 (95 %) have same cell dimensions as model data
  - 40 (5 %) have cell different (>5A in each axis) from model
- 533 (40 %) are not in P1
  - 532 have same cell as model data
  - Original event map transformed into asymmetric unit of model data?
  - Mostly correct ...

5QIB





Overview

- About half of fragment-screening (PanDDA-style) group depositions have event maps provided in a way that makes it impossible to recreate them from the deposited data.
- Ca 40% of PDB entries for those group depositions have no associated publication.
- So for only about 30% of fragment-screening (PanDDA-style) group depositions can we (probably) see the event map in the publication **and** recreate it for inspection.
- However, the **correct level for looking at those recomputed maps is not provided** at all (as far as we know).

- At the moment one has to take those ligand structures mostly on trust - unless they are strong binders and are visible in standard 2mFo-DFc maps (or mFo-DFc omit maps) anyway. This does not necessarily mean the interpretation of weakly binding ligands is incorrect or doubtful: just that there is missing (meta)data to allow for independent validation and reproducibility.

But even a publication is not always a help

8CN1: https://doi.org/10.1016/j.antiviral.2023.105675

X-ray diffraction data were collected at 100 K using the Beamline Proxima 2 at the Soleil synchrotron (Gif-sur-Yvette, France) and the Beamline iO4 at Diamond Light Source (Didcot, UK). The diffraction data were processed with AutoProc (Clemens et al., 2011). The crystals of hDLG1 PDZ2-ETEV presented high levels of anisotropy, necessitating further data processing with Staraniso (Global Phasing suite). The crystal structures were determined by molecular replacement with phaser (Bunkóczi et al., 2013; Liebschner et al., 2019) and PDB 3RL7 as the search model. The structures were refined through iterative cycles of manual model building with COOT (Emsley and Cowtan, 2004) reciprocal space refinement with phenix.refine (Afonine et al., 2012) and Buster (Smart et al., 2012). The crystallographic statistics are shown in Supplementary Table 1. Atomic coordinates and structure factors of hDLG1 PDZ1 and PDZ2 complexes with ETEV peptide have been deposited in the protein data bank (PDB) under the accession codes 8CN1 and 8CN3, respectively.

https://staraniso.globalphasing.org/table1/cn/8cn1.html

Data quality metrics extracted from 8cn1.cif.gz by **aB_cif2table1** from BUSTER (Global Phasing Ltd.).

Experimental information for **8CN1** at RCSB, PDBe, PDBj

Experiment	
Source type	SYNCHROTRON
Source details	DIAMOND BEAMLINE I04-1
Synchrotron site	Diamond
Beamline	I04-1
Temperature [K]	100
Detector	DECTRIS PILATUS3 S 6M
Wavelength(s) [Å]	0.979
Software	
Data reduction	XDS
Data scaling	Aimless
Phasing	PHASER
Refinement	BUSTER

Getting to a publication - can we do better?
User A to Archive interface

I04-1 Specification

Energy	fixed, monochromatic: 0.920Å / 13.53 keV

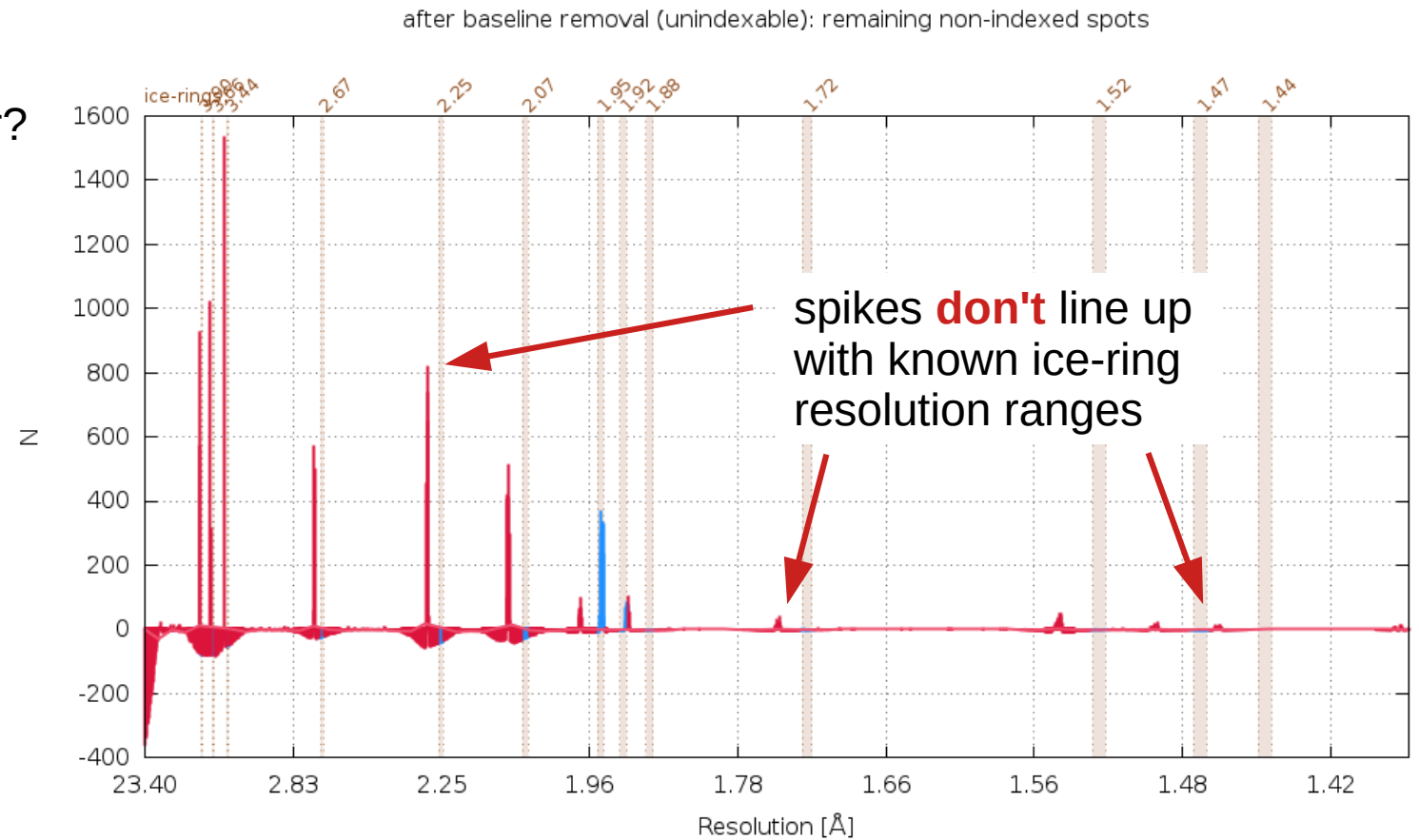
Beamline	Detector	Wavelength [Å]	#PDB
"I04-1"	"DECTRIS"	0.91-0.93	5735
"I04-1"	"DECTRIS"	**<0.91 || >0.93**	**223**
"I04"	"DECTRIS"	0.91-0.93	179
"I04"	"DECTRIS"	<0.91 || >0.93	2258

This could be checked if deposition software could connect to an up-to-date record of beamline configurations (maintained by beamline itself).

Probably "just" a typo by User A upon deposition? But: very confusing to User B!

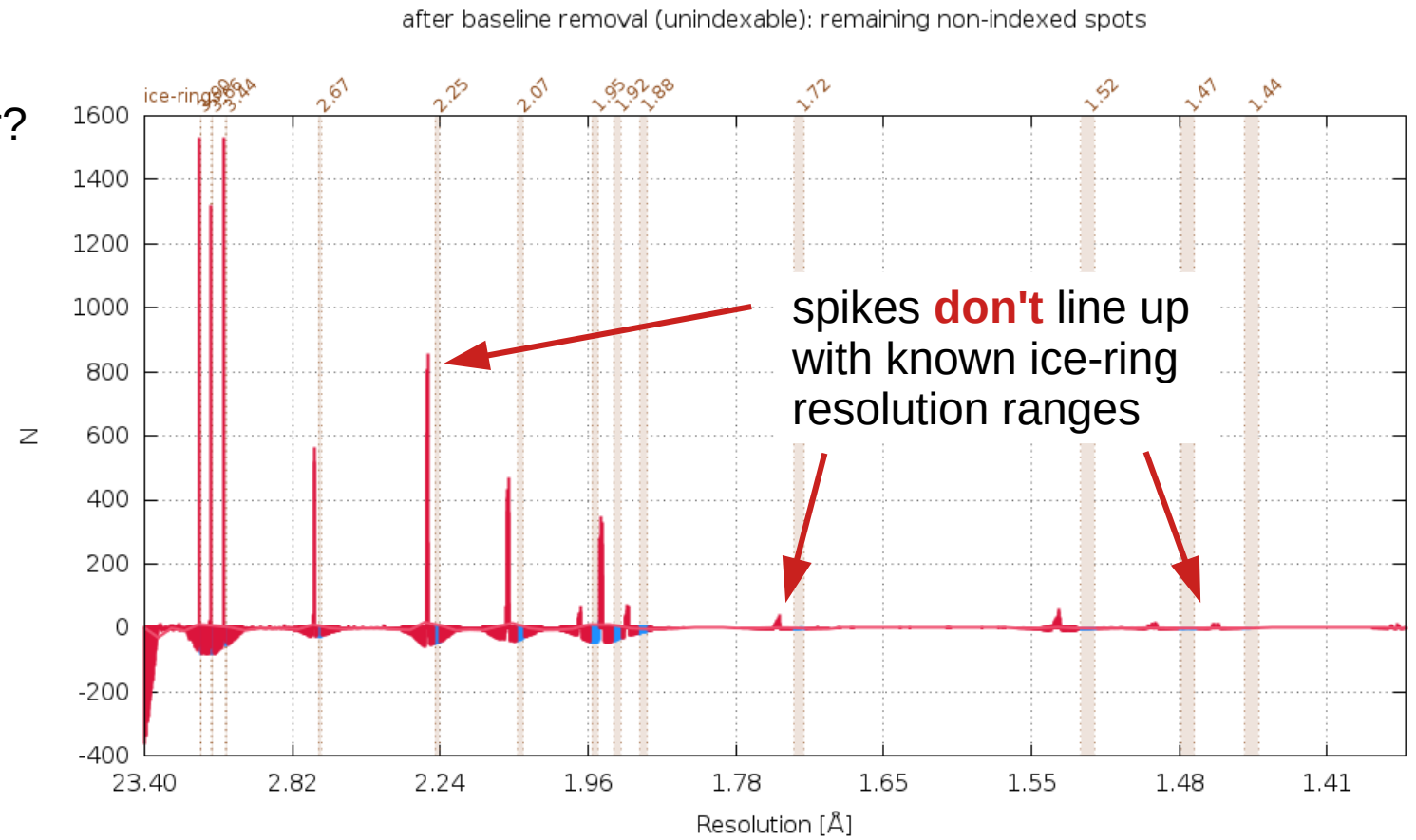
Incorrect distance and incorrect wavelength

Does wavelength value matter?



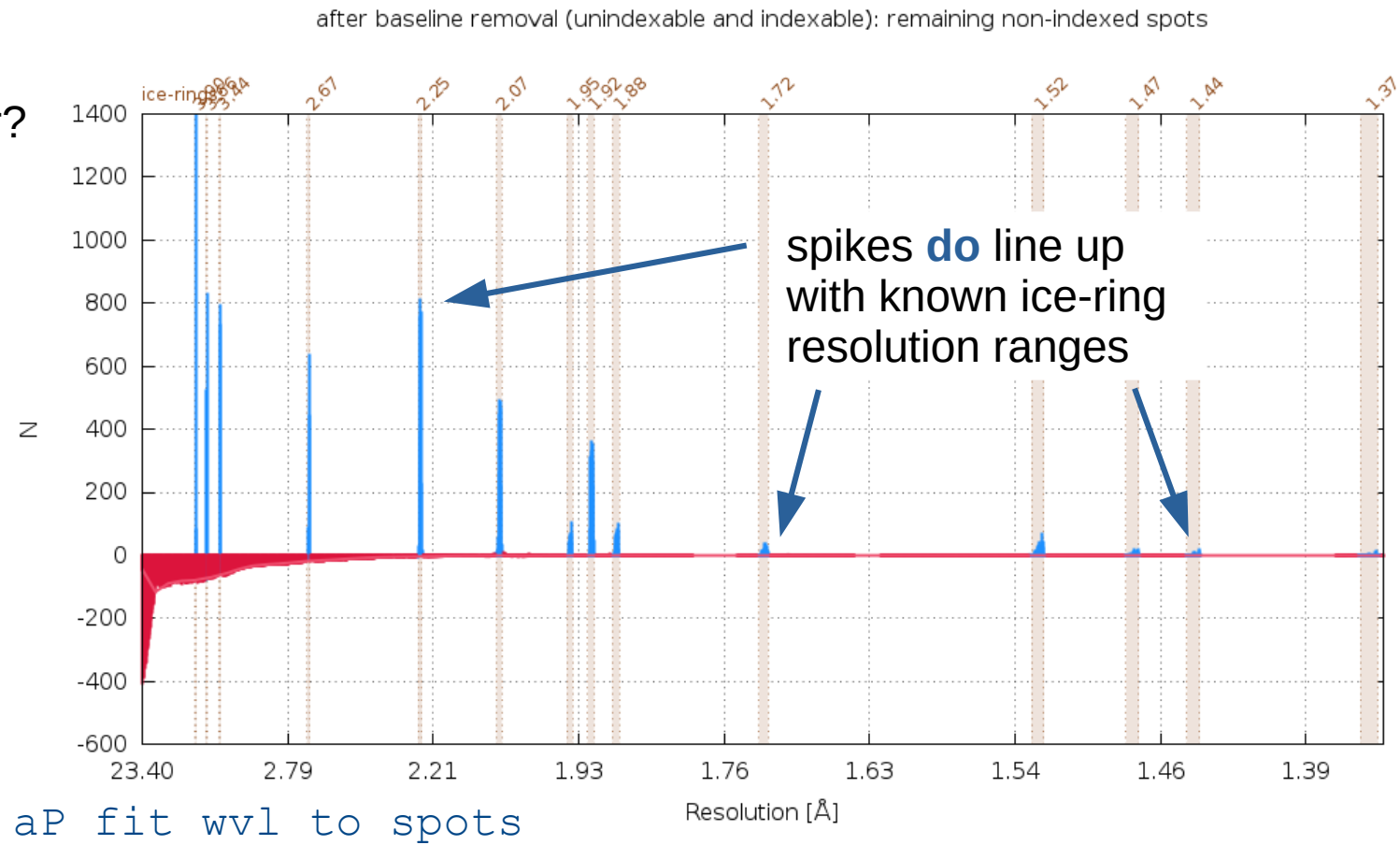
Correct distance and incorrect wavelength

Does wavelength value matter?



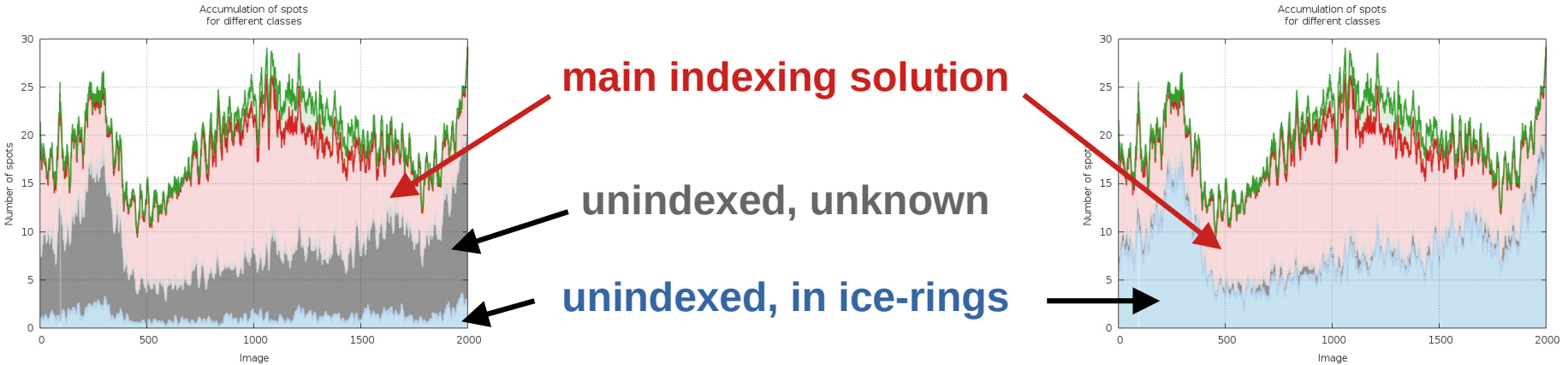
Correct distance and correct wavelength

Does wavelength value matter?



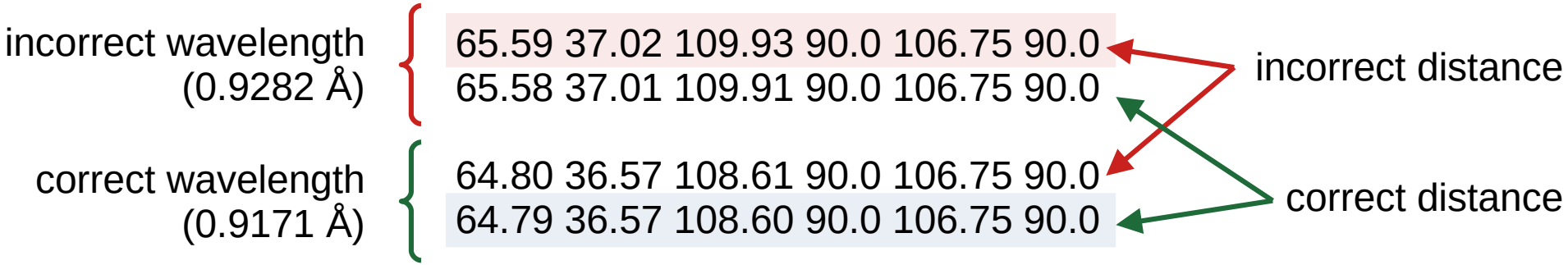
autoPROC: `aP_fit_wvl_to_spots`

Incorrect distance and/or wavelength: does it matter?



- Correct distance and wavelengths allows for **clear identification of ice-rings** in diffraction data
  - as information for user, e.g. **subsequent improvements** in crystal handling or cryo-cooling protocols
  - adjustment of data processing procedure, e.g. **exclusion of ice-ring resolution shells** during integration
- **Integration** of intensities in reciprocal space unaffected by incorrect wavelength.
- **But** what happens in real space (even if no ice-rings are present in data)?

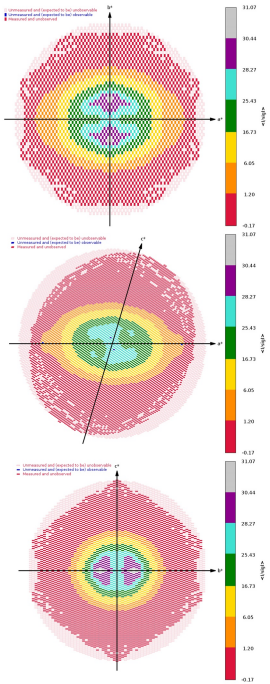
Incorrect distance and/or wavelength in real-space



- A ~1.2% error in wavelength gives a ~1.2% error in cell dimensions

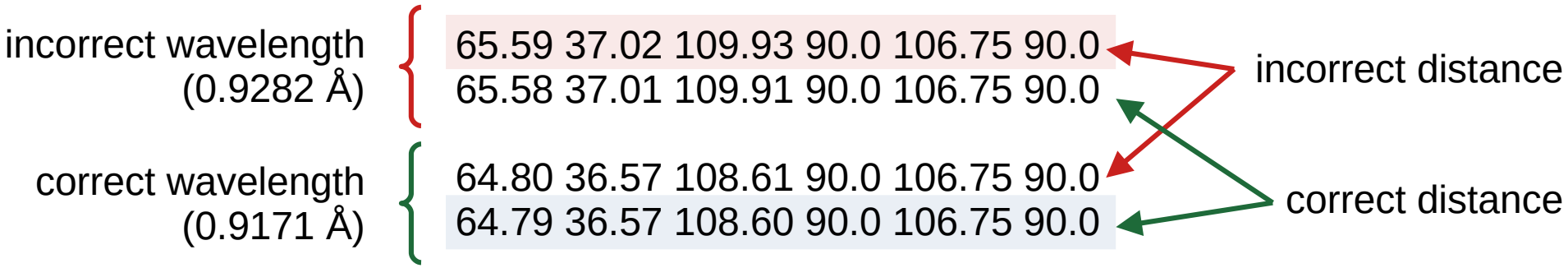
Incorrect distance and/or wavelength in reciprocal-space

	reprocessed autoPROC+STARANISO								
	deposited		incorrect wvl			correct wvl			
	Overall	Outer	Overall	Inner	Outer	Overall	Inner	Outer
Low res. limit [A]	34.870	2.120	62.355	62.355	1.821	61.582	61.582	1.808
High res. limit [A]	2.060	2.060	1.509	5.808	1.509	1.479	5.764	1.479
Rmerge	0.071	0.919	0.083	0.069	0.837	0.083	0.070	0.721
Rmeas	0.075	0.963	0.088	0.073	0.904	0.087	0.073	0.785
Rpim	0.023	0.286	0.027	0.022	0.335	0.027	0.022	0.301
Total number of obs.	-	-	165190	7894	5573	156586	7764	5190
Total number unique	15862	1217	15289	764	764	15048	752	752
<I/sig(I)>	15.00	1.80	11.2	26.4	1.4	11.4	26.6	1.5
Completeness [%]	99.9	100.0	82.0	99.9	41.5	81.8	99.9	36.8
Multiplicity	11.0	11.4	10.8	10.3	7.3	10.4	10.3	6.9
CC(1/2)	0.999	0.969	0.999	0.998	0.821	0.999	0.998	0.918
	65.7 37.3 108.5 90.0 106.7 90.0		65.6 37.1 109.1 90.0 106.6 90.0			64.8 36.7 107.8 90.0 106.6 90.0			



- Main difference is treatment of highly anisotropic diffraction (1.4, 2.0, 2.5 Å)
- Data quality metrics hardly affected by incorrect wvl (as expected)

Incorrect distance and/or wavelength in real-space



- A ~1.2% error in wavelength gives a ~1.2% error in cell dimensions
- Refinement of a model in the incorrect cell (here: too large) against external restraints (standard geometries for amino-acids, DNA/RNA etc):
  - given increased unit cell parameters, all bonds would like to be a bit longer
  - need **stronger weight on geometry** (or lower weight on X-Ray) to keep geometry close to external restraints
  - increase in R/Rfree (because of **lower X-Ray weight**)
  - with fixed X-Ray weight: **increased rms(bond)** at incorrect unit cell lengths

GΦL
Global Phasing Limited

Balancing X-Ray vs Geometry weight

fixed geom

max X-Ray weight

<bond length>:
- 0.0015 Å
- 0.04 σ

<bond length>:
+ 0.0126 Å
+ 0.61 σ

Scale	X-Ray Weight	Rwork	Rfree	**rms(bond)**
0.960	1.00	0.1555	0.1637	0.019
0.962	1.00	0.1544	0.1628	0.017
0.964	1.16	0.1490	0.1612	0.017
0.966	1.43	0.1446	0.1547	0.017
0.968	1.79	0.1394	0.1538	0.017
0.970	2.30	0.1356	0.1501	0.017
0.972	2.87	0.1355	0.1491	0.017
0.974	4.08	0.1320	0.1469	0.017
0.976	5.26	0.1302	0.1435	0.017
0.978	5.58	0.1295	0.1421	0.017
0.980	5.73	0.1291	0.1424	0.017
0.982	6.32	0.1287	0.1432	0.017
0.984	6.41	0.1295	0.1442	0.017
0.986	6.35	0.1296	0.1426	0.017
0.988	5.87	0.1286	0.1420	0.017
0.990	5.15	0.1307	0.1436	0.016
0.992	4.55	0.1311	0.1446	0.016
0.994	3.89	0.1314	0.1467	0.016
0.996	3.23	0.1330	0.1460	0.016
0.998	2.60	0.1345	0.1471	0.016
1.000	2.17	0.1371	0.1465	0.016
1.002	1.77	0.1404	0.1504	0.016
1.004	1.50	0.1428	0.1544	0.016
1.006	1.26	0.1453	0.1557	0.016
1.008	1.08	0.1495	0.1584	0.016
1.010	1.00	0.1509	0.1591	0.017
1.012	1.00	0.1521	0.1621	0.018
1.014	1.00	0.1514	0.1608	0.019
1.016	1.00	0.1529	0.1619	0.020
1.018	1.00	0.1538	0.1636	0.021
1.020	1.00	0.1555	0.1646	0.022

fixed X-Ray

min rms(bond)

<bond length>:
- 0.0015 Å
- 0.04 σ

<bond length>:
+ 0.0160 Å
+ 0.80 σ

Scale	**X-Ray Weight**	Rwork	Rfree	rms(bond)
0.960	6.40	0.1299	0.1452	0.031
0.962	6.40	0.1277	0.1435	0.029
0.964	6.40	0.1281	0.1433	0.028
0.966	6.40	0.1288	0.1431	0.026
0.968	6.40	0.1286	0.1433	0.024
0.970	6.40	0.1275	0.1419	0.023
0.972	6.40	0.1289	0.1419	0.021
0.974	6.40	0.1286	0.1408	0.020
0.976	6.40	0.1284	0.1428	0.019
0.978	6.40	0.1294	0.1432	0.018
0.980	6.40	0.1286	0.1432	0.017
0.982	6.40	0.1289	0.1429	0.017
0.984	6.40	0.1291	0.1428	0.017
0.986	6.40	0.1292	0.1441	0.017
0.988	6.40	0.1282	0.1426	0.017
0.990	6.40	0.1287	0.1423	0.018
0.992	6.40	0.1288	0.1428	0.019
0.994	6.40	0.1275	0.1413	0.020
0.996	6.40	0.1283	0.1419	0.021
0.998	6.40	0.1279	0.1426	0.022
1.000	6.40	0.1291	0.1424	0.024
1.002	6.40	0.1293	0.1421	0.025
1.004	6.40	0.1290	0.1418	0.027
1.006	6.40	0.1289	0.1429	0.029
1.008	6.40	0.1289	0.1434	0.031
1.010	6.40	0.1297	0.1418	0.032
1.012	6.40	0.1285	0.1429	0.034
1.014	6.40	0.1285	0.1424	0.036
1.016	6.40	0.1296	0.1425	0.038
1.018	6.40	0.1285	0.1416	0.040
1.020	6.40	0.1292	0.1433	0.042

Cell scale

Fully automatic refinement with BUSTER ("aB_autorefine")

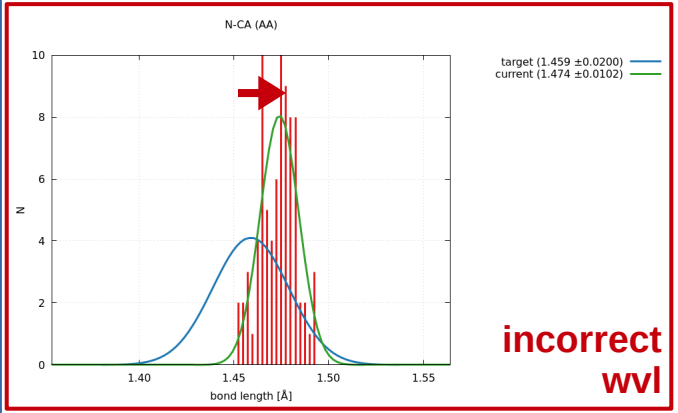
https://doi.org/10.1021/acs.jmedchem.7b00933

Incorrect energy

- Leads to **incorrect unit cell** - which in turn leads to **incorrect weighting** of X-Ray versus geometry terms:
  - to keep sensible geometry we need more weight on geometry (i.e. less on X-Ray) since incorrect unit cell will force e.g. "stretching" of bond distances
- We can **analyse observed bond distances** in deposited models to find instances of **likely cell scaling** (similar to WHAT_CHECK feature):
  - look at all N-CA, CA-C and C-O bonds of fully-occupied atoms in standard amino-acids
  - find that cell scale that will give the smallest deviation of the observed mean from the expected mean (Engh&Huber values).

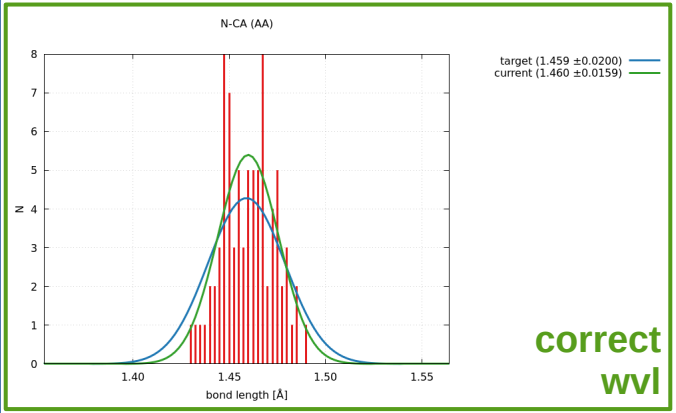
Global Phasing Limited

Incorrect wavelength impact on distributions

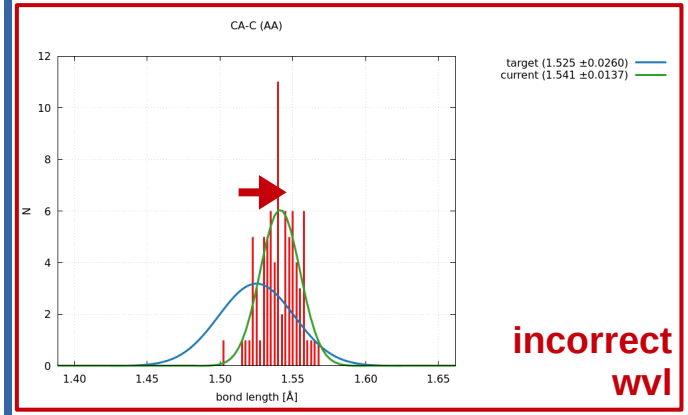


incorrect wvl

N - CA

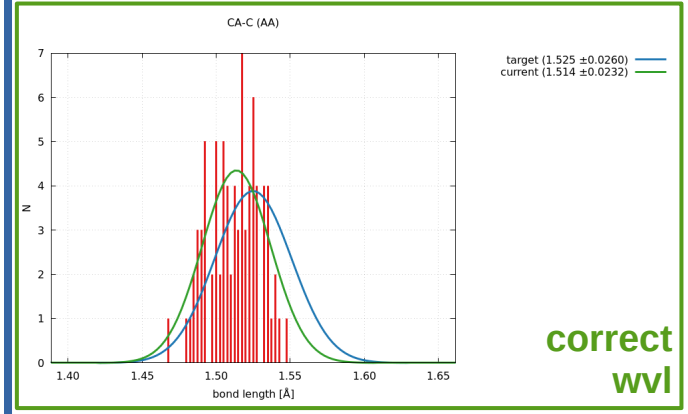


correct wvl

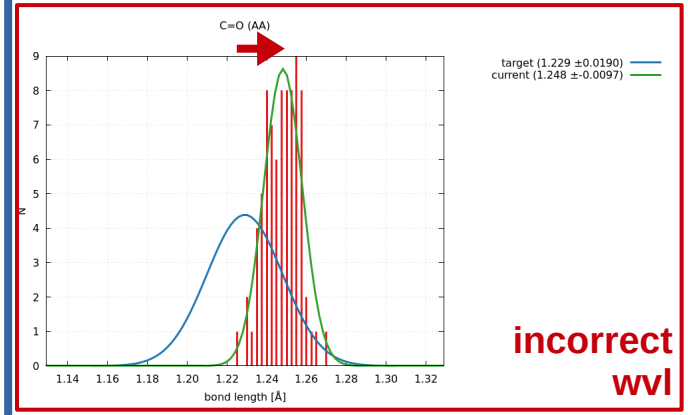


incorrect wvl

CA - C

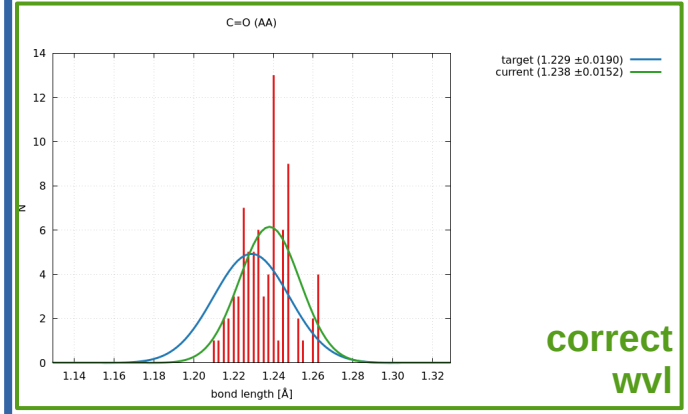


correct wvl



incorrect wvl

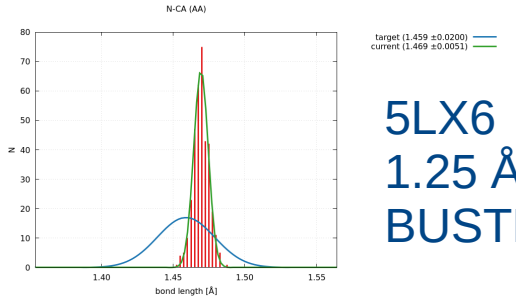
C - O



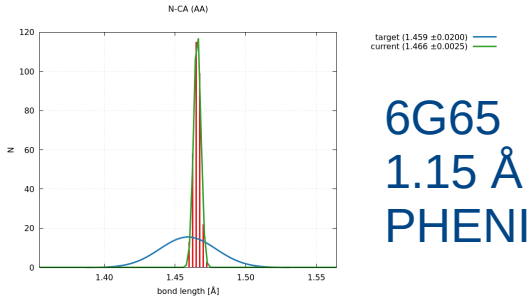
correct wvl

Global Phasing Limited

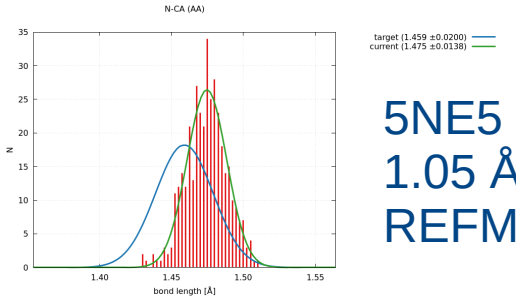
Incorrect wavelength, high-resolution refinements



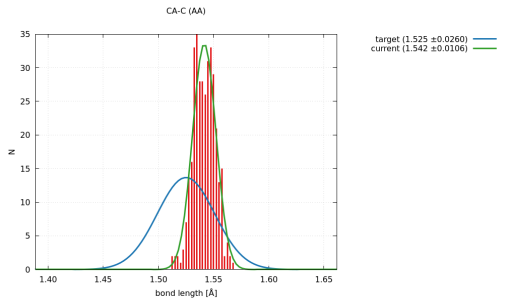
5LX6
1.25 Å
BUSTER

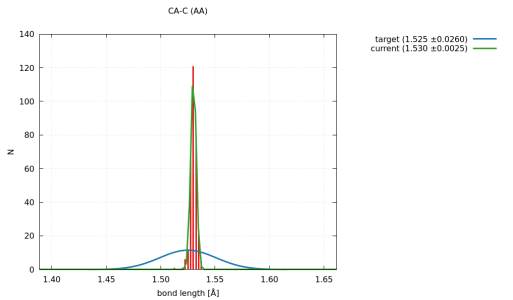


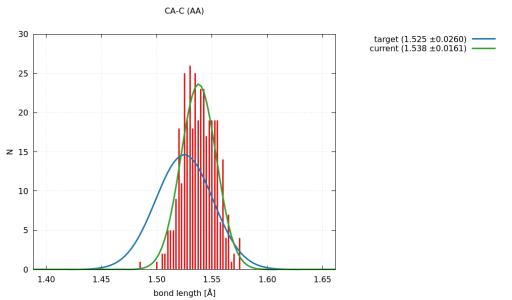
6G65
1.15 Å
PHENIX

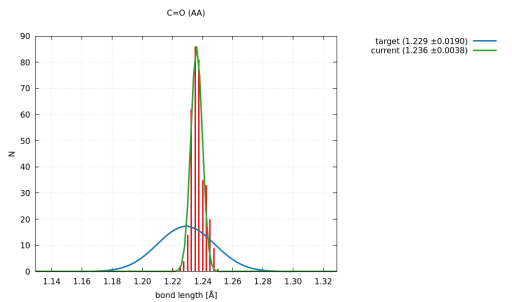


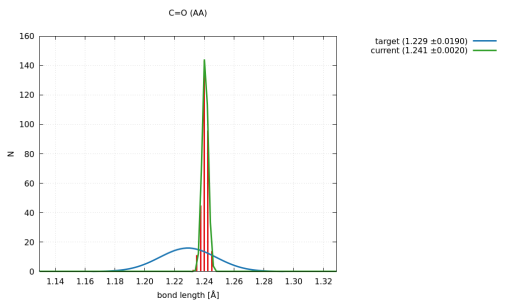
5NE5
1.05 Å
REFMAC

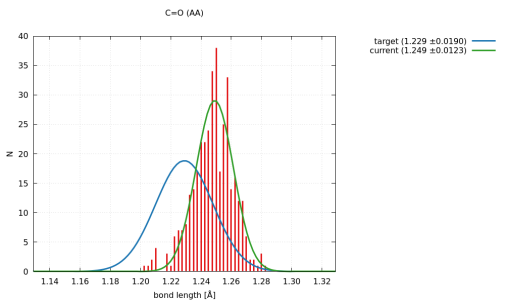












Challenges (?)

- Working in crystallographic spaces (real or reciprocal) can be confusing:
  - depositors should check that they can use the reflection data in the archive as intended
  - wwPDB sites should start looking at datablocks N>1:
    - not just archiving them
    - ideally: provide maps for all reflection datablocks
    - consult with external experts to have a robust procedure in place
  - non-crystallographic maps (like PanDDA event maps) require additional information to become useful:
    - additions to PDBx/mmCIF dictionary initiated by PanDDA team?
- Small mistakes during metadata creation or data analysis can have a significant/visible impact
- How do we annotate archived models for AI/ML? Do we need to look at:
  - Is there an associated publication?
  - Does a model provide reproducibility and validation?
- To improve our provision and handling of metadata, fast turnaround is required:
  - between **beamlines** (where experiments produce raw data), **pipeline developers** (collate metadata in machine-readable form), **users** (transfering metadata from initial generation through various systems) and **wwPDB systems** (final archiving and making available to community).