GΦL
Global Phasing Limited

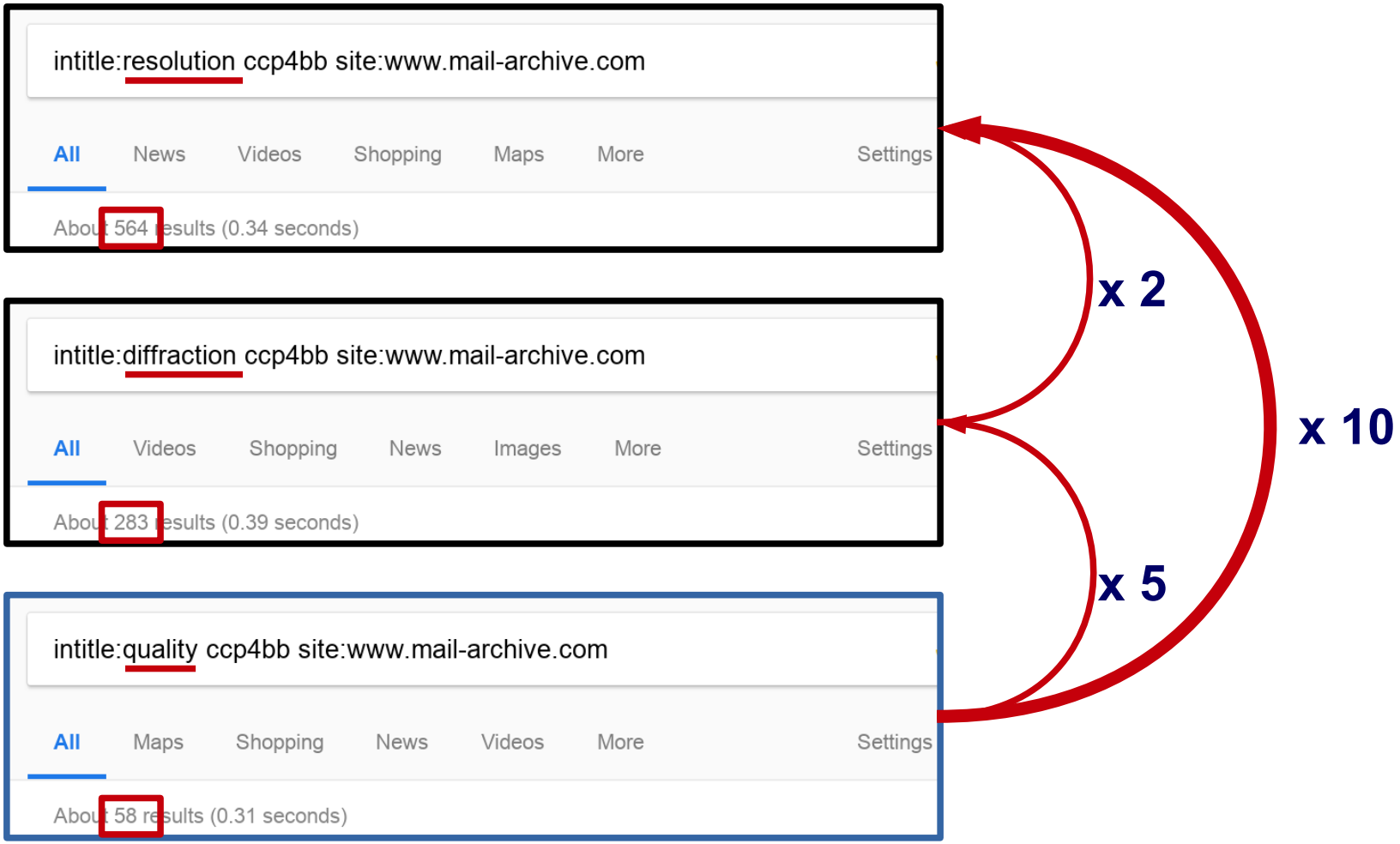
Resolution, data quality and all that ...

Clemens Vonrhein
Global Phasing Ltd.

DLS-CCP4 Workshop 2021

GΦL
Global Phasing Limited

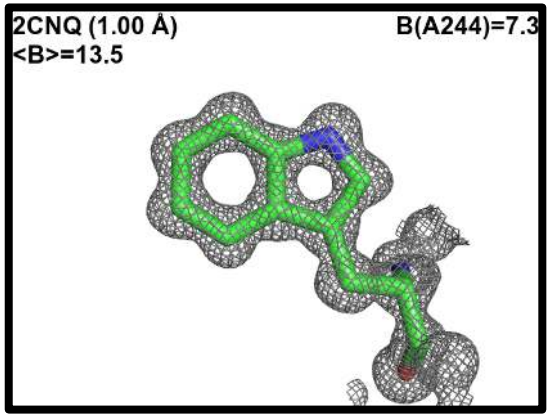
Too much focus on single number (resolution) to describe model quality?



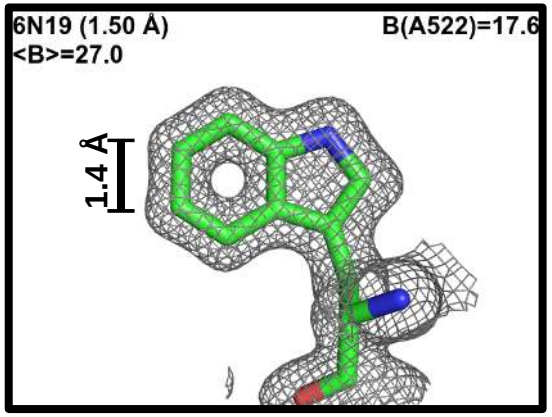
GΦL
Global Phasing Limited

Resolution = ability to resolve detail

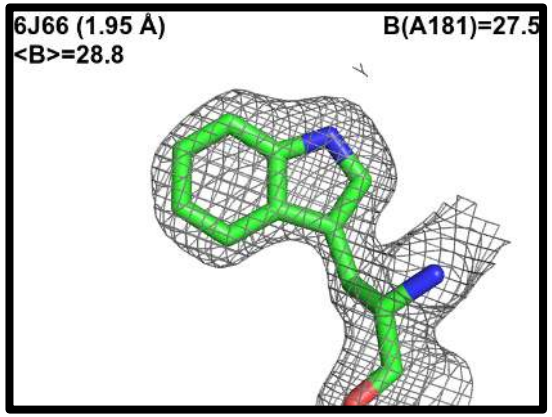
1.0 Å



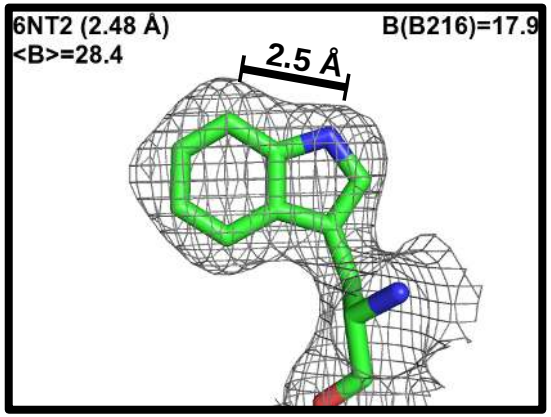
1.5 Å



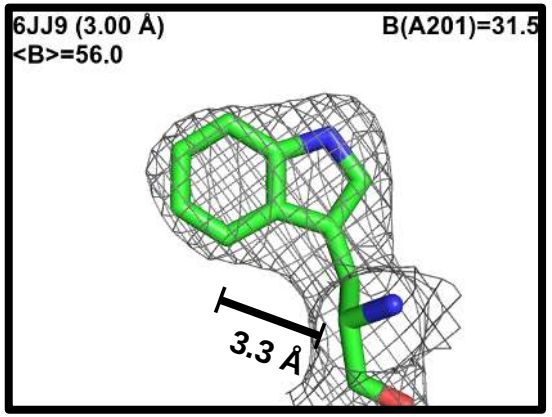
2.0 Å



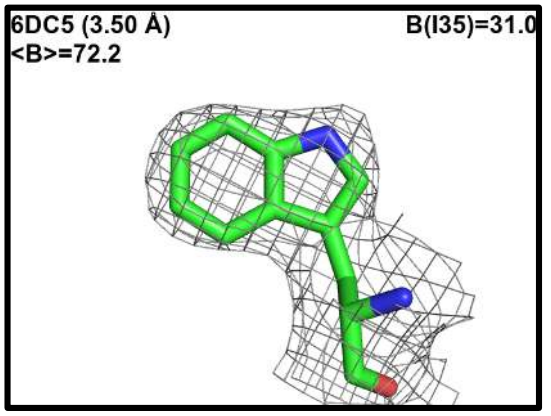
2.5 Å



3.0 Å



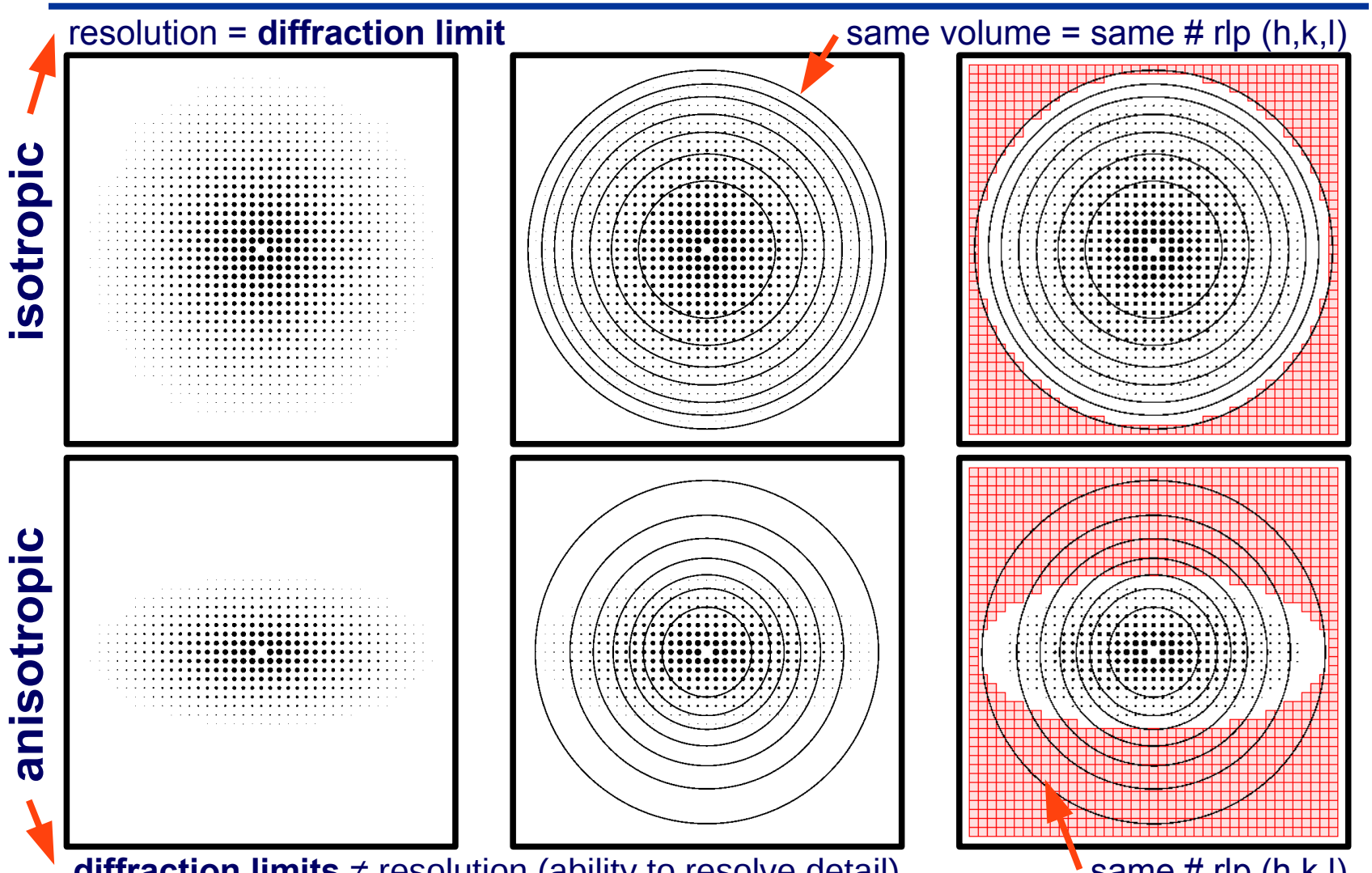
3.5 Å



2mFo-DFc maps after BUSTER refinement at 1.0 rms

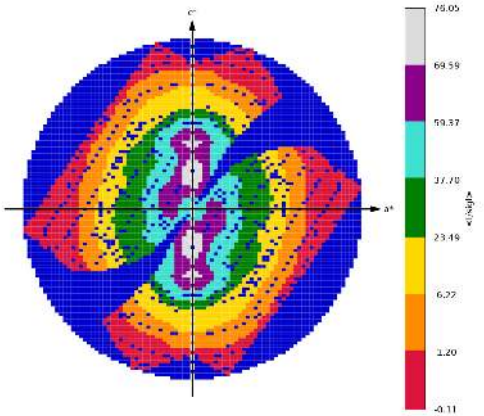
GΦL
Global Phasing Limited

Resolution, diffraction limit and binning



rlp = reciprocal lattice point

STARANISO – http://staraniso.globalphasing.org/



- main author: Ian Tickle (Global Phasing)
- **anisotropic** cut-off of merged intensity data
  - avoids problems of
    - isotropic resolution cut-offs
    - throwing away good data (in one direction)
    - including poor data (in another direction)
- Bayesian estimation of structure factor amplitudes
  - more rigorous treatment of prior and anomalous data
  - for more details see: **https://staraniso.globalphasing.org/**staraniso_about.html
- (optional) anisotropic correction of data
- Part of autoPROC ('process', 'aP_scale')
  - user is given classical, isotropic (TRUNCATE) output
  - but also data after **anisotropic** analysis (STARANISO)

merged or unmerged data (MTZ, XDS)

Anisotropy is not necessarily ellipsoidal!

GΦL
Global Phasing Limited

Isotropic vs anisotropic

GLOSSARY of terminology related to anisotropy

ABOUT ANISOTROPY

Anisotropy is the property of being directionally dependent, implying a variation of certain physical properties as a function of direction. It is the opposite of ***isotropy***, that is, possessing exactly the same properties in whatever direction one looks. Here we are mainly concerned with **two** potentially anisotropic properties of the measured X-ray diffraction data: the ***diffraction limit*** (also known as the '***resolution limit***'), and the ***attenuation of the diffracted intensity*** due to thermal motion and/or disorder.

The diffraction limit

Beyond the diffraction limit of the sample, the merged intensities of unique reflections fall, on average, below the level of statistical significance and should therefore be considered to be non-informative (inclusion of such data in further procedures, scaling in particular, may in fact be positively detrimental).

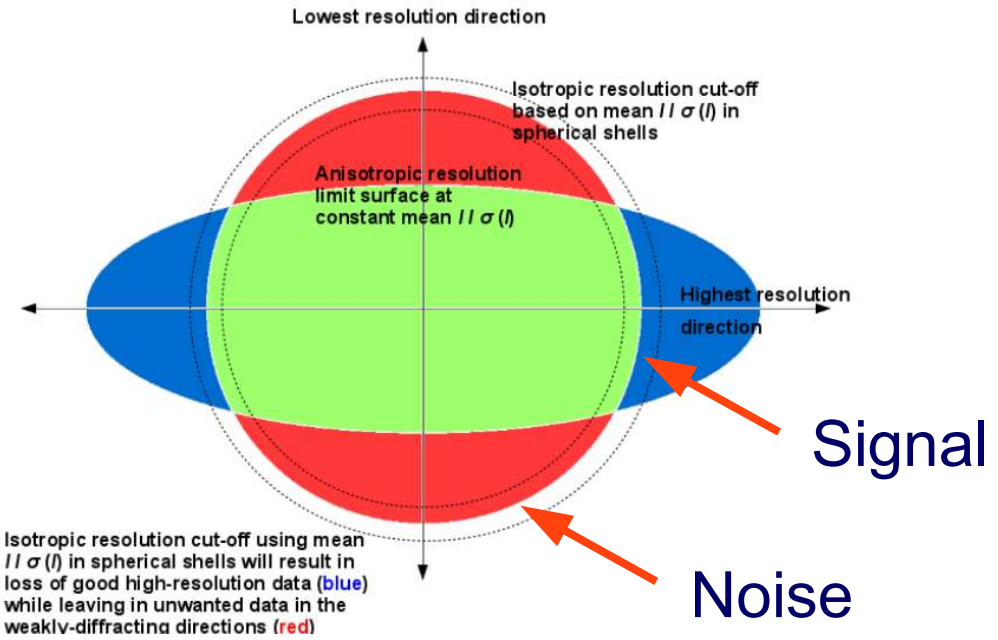
The statistical significance of the intensity data after merging is commonly expressed as either the value of the mean signal/noise ratio $I/\sigma(I)$, or the half-dataset correlation coefficient $CC_{½}$ [1], computed in spherical shells of reciprocal space. The statistical significance, in whatever form one chooses to express it, is used as a criterion for applying a ***resolution cut-off*** to the reflection data that should match the diffraction limit as closely as possible. Clearly if the applied resolution cut-off falls inside the diffraction limit, then useful data containing information about the structure will be lost; if it falls outside, then uninformative data will be kept.

Since the diffraction limit may be anisotropic, in STARANISO the locally-averaged value of $I/\sigma(I)$, rather than its spherical average, is used to define an anisotropic resolution cut-off - that is, a cut-off varying with direction in reciprocal space. The boundary between statistically significant and insignificant data therefore defines a surface, the ***anisotropic resolution cut-off surface*** (see diagram on right: click image to enlarge).

No assumption is made about the shape or orientation of the anisotropic resolution cut-off surface, other than that it has at least the symmetry of the Laue class. This means that in orthorhombic and higher symmetry space groups the principal axes of the surface shape must coincide with the crystal axes. In monoclinic space groups only one of the principal axes of the surface shape is constrained to lie along a crystal axis, namely ***b***, but there is no symmetry constraint on the second principal axis: it can lie anywhere in the ***ac*** plane since the choices of the ***a*** and ***c*** axes are completely arbitrary. Similarly, in triclinic space groups the choices of all three crystal axes are arbitrary so there are no symmetry constraints whatsoever on the directions of the principal axes of the surface shape.

Note that it is ***not*** assumed that the shape of the anisotropic cut-off surface is an ellipsoid (the contours in the figure below are clearly not ellipsoidal). The principal cause of distortion of the diffraction limit surface away from a perfect ellipsoid is the variation of the reflection redundancy through its effect on the $\sigma(I)$ values of the merged data (*i.e.* inversely proportional to the square root of the redundancy), and therefore on the mean $I / \sigma(I)$. The variation of redundancy in reciprocal space is largely arbitrary, being a consequence of both the experimental setup and the data collection strategy (whether a standard protocol or a deliberate user choice).

However, for the purposes of estimating the resolution limit as a function of direction an ellipsoid is fitted to the anisotropic cut-off surface. Measured data with d^* greater than the largest semi-axis length of the fitted ellipsoid are considered to be outside the anisotropic cut-off surface. Then all measured intensities (excluding any systematic absences) that fall inside this anisotropic cut-off surface are considered to be '***observed***', that is they are included in the reflection statistics and are written to the reflection output file, while all those that fall outside are '***unobserved***' and are not used further. Unmeasured data (except systematic absences) that fall inside the fitted ellipsoid (for example in a cusp) are deemed to be '***observable***' (*i.e.* they had the potential to be observed if only they had been measured), whereas those unmeasured reflections falling outside the fitted ellipsoid are deemed to be '***unobservable***' (*i.e.* not worth measuring).



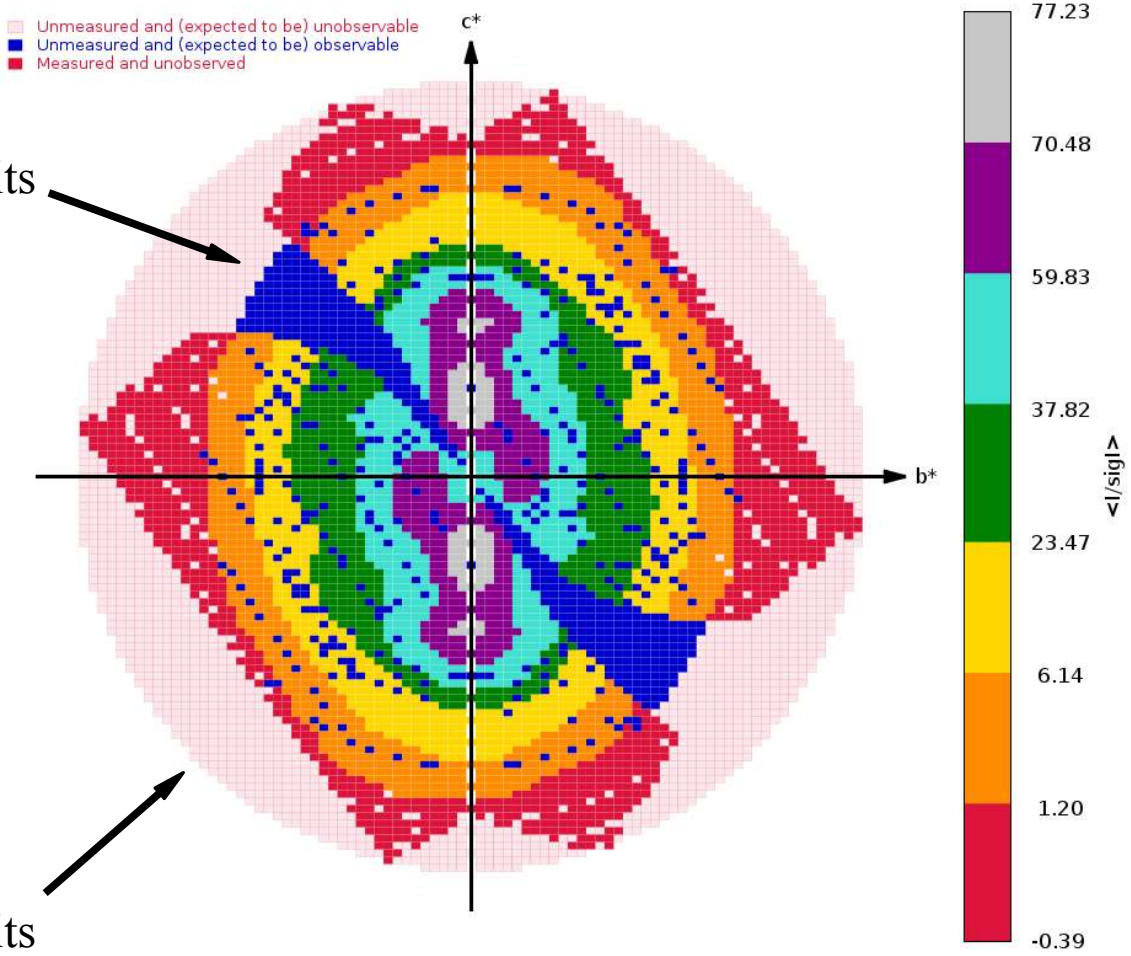
Global Phasing Limited

autoPROC: STARANISO plots

rec. unit cell

Diffraction limits observable

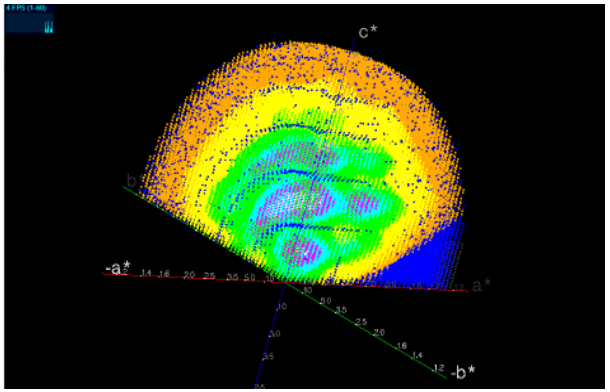
Diffraction limits recorded



GΦL
Global Phasing Limited

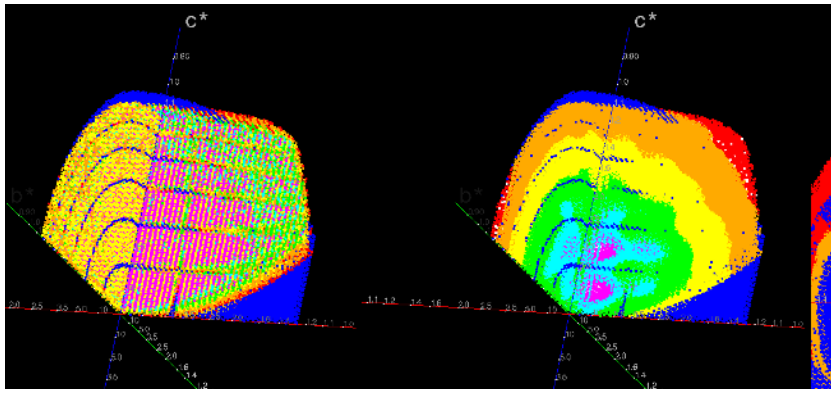
3D WebGL visualisation in STARANISO

https://staraniso.globalphasing.org/cgi-bin/PDBpeep.cgi
https://staraniso.globalphasing.org/



4YUO $P2_12_12_1$
a* ~ 7°
single sweep

beam centre in module gap



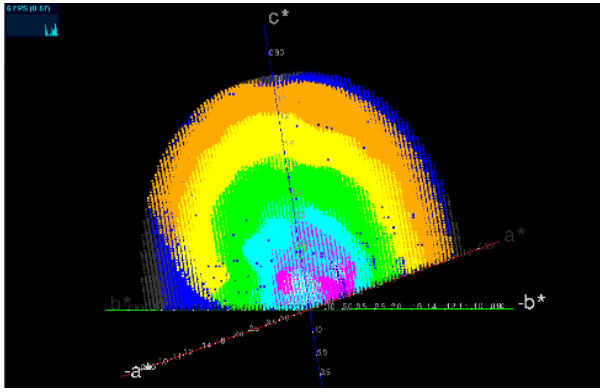
local <I/σI>

merged data

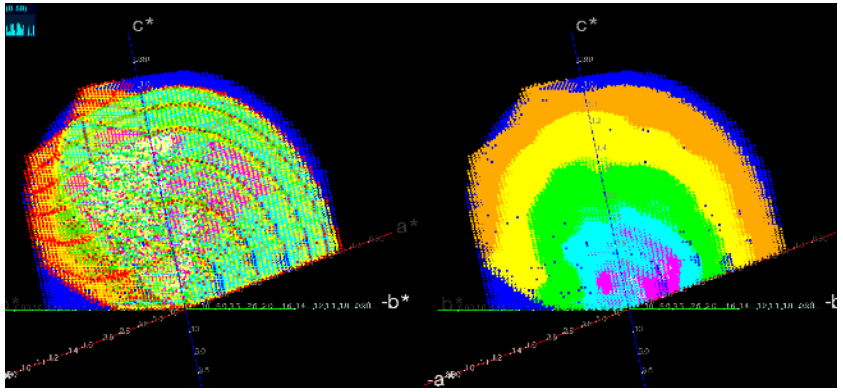
unmerged data

redundancy

local <I/σI>



4ZC9 $P2_12_12_1$
b* ~ 14°
2 sweeps
(distance)

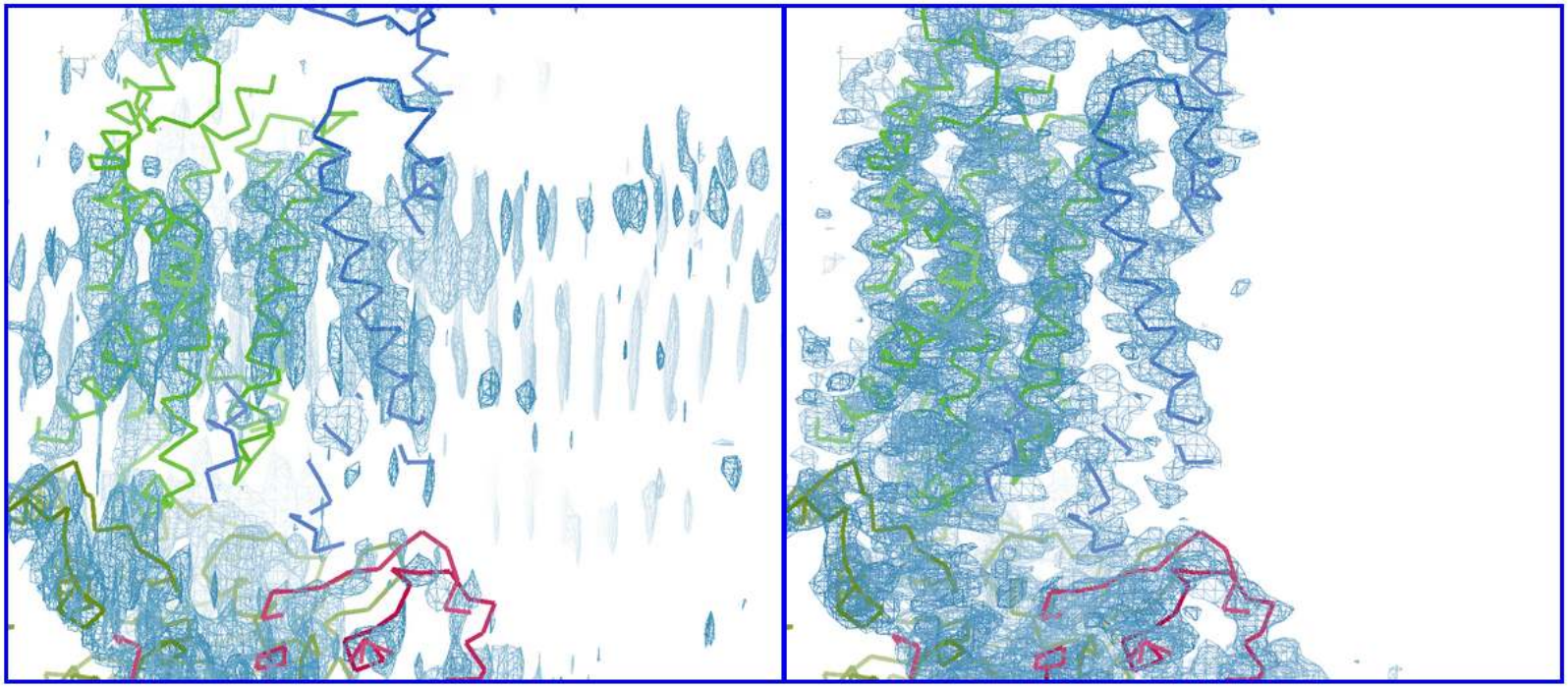


staraniso.globalphasing.org

- **The following results were communicated to us by Neil Paterson at Diamond Light Source.**

The panels show the maps after density modification (CCP4 DM program) using (left) the standard protocol with an isotropic resolution cut-off, and (right) the STARANISO protocol with anisotropic resolution cut-off and correction. Compare in particular the density in the more mobile loop regions at the ends of the helices, and also the flatness of the solvent region (right-hand half of each map).

Neil comments: "STARANISO has been key to solving our structure. We've got a severely anisotropic case that goes from around 7 Å along a^* to 3.0-3.5 Å along b^*/c^*, with lots of artifacts in the maps. We can find the substructure using Shel-X without correction but the maps are pretty bad; once corrected it works a treat (77% solvent!)."



GΦL
Global Phasing Limited

Evolution of diffraction data metrics - “Table 1”

R_{merge} (R_{sym})

Completeness

Redundancy (multiplicity)

<I/σI>

R_{meas}

R_{pim}

$CC_{1/2}$

Table I. *X-ray Data Reduction and Crystallographic Refinement Statistics*

(*A*) X-ray data reduction statistics	
Space group	$P2_12_12_1$
Unit cell dimensions	
a, *b*, *c*	33.5 Å, 68.2 Å, 137.6 Å
Resolution	68.8 Å–2.25 Å
(last shell)	(2.26 Å–2.25 Å)
Total measurements (last shell)	109,559 (1,168)
Number of unique reflections	15,637
(last shell)	(163)
Wavelength	1 Å
R-merge[a] (last shell)	0.065 (0.592)
I/σ(I) (last shell)	16.4 (3.3)
Completeness (last shell)	0.997 (1)
Redundancy (last shell)	7 (7.2)

As function of resolution in spherical shells (equal spacing or equal volume): overall and outer shell

R_{merge} : Arndt et al (1958)
R_{meas} : Diederichs & Karplus (1997), Weiss & Hilgenfeld (1997)
R_{pim} : Weiss (2001)
$CC_{1/2}$: Karplus & Diederichs (2012)

GΦL
Global Phasing Limited

Binning

- **Equally spaced** in $(d^*)^2$
  - Number of measurements used for computing bin averages increases with resolution
  - Low-resolution statistics can be unreliable (too few measurements)
  - High-resolution statistics not finely enough sampled?
  - Used by XDS, AIMLESS

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENE OF DATA
3.13	75353	14441	14448	100.0%
2.22	136754	26193	26370	99.3%
1.81	178487	34010	34139	99.6%
1.57	207171	40372	40378	100.0%
1.40	214727	45702	45725	99.9%
1.28	183266	50533	50658	99.8%
1.18	142566	54470	55100	98.9%
1.11	57895	43932	59148	74.3%
1.04	15808	15171	63013	24.1%
total	1212027	324824	388979	83.5%

- **Equal volume** (to have same number of reciprocal lattice points in each bin)
  - Allows different sampling (coarse or fine) depending on requirements
  - Remember ice-rings (3.89 3.66 3.44 2.67 2.25 2.07 1.95 1.91 1.88 1.72 1.52 1.47 1.44 1.37 1.36 1.30 1.27 1.26 1.22 1.17 1.12)!
  - Adequate for isotropic data that is complete with homogeneous multiplicity
- **Equal number** of (actual) observations
  - Can be seen as generalisation of idea behind “equal volume” binning
  - Automatically self-adjusting for anisotropic (STARANISO) and incomplete data (serial crystallography, LCP, in-plate, ...)

GΦL
Global Phasing Limited

Binning: decision making, data description

Binning is boring - but an important aspect here:

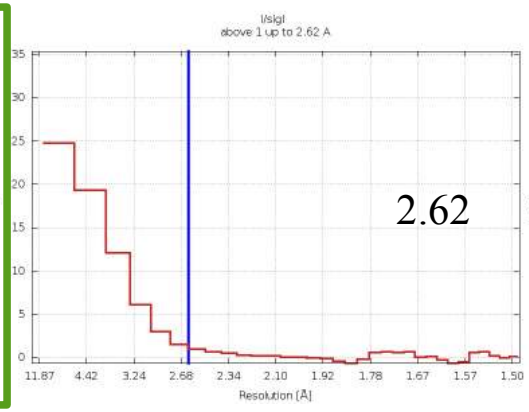
- Decision making:
  - Which images/datasets/reflections to include
  - Which reflections to include: isotropic/anisotropic diffraction limit
- Comparisons:
  - Between programs and pipelines
  - Between different processing options
- Basis of decision making
  - Raw data (binned statistics)
  - Smoothed (spline, Bezier, …)
  - Ice-ring resolution ranges included/excluded/smoothed?

GΦL
Global Phasing Limited

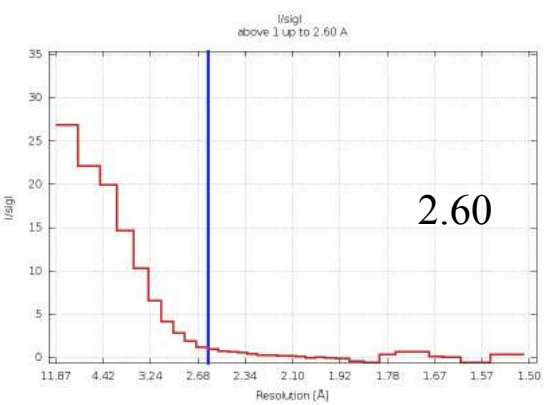
What is my high-resolution limit?
… depending on binning method

d*^2 (20 bins)

equal-volume

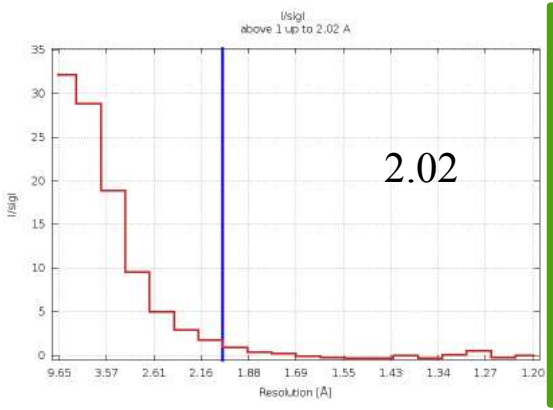


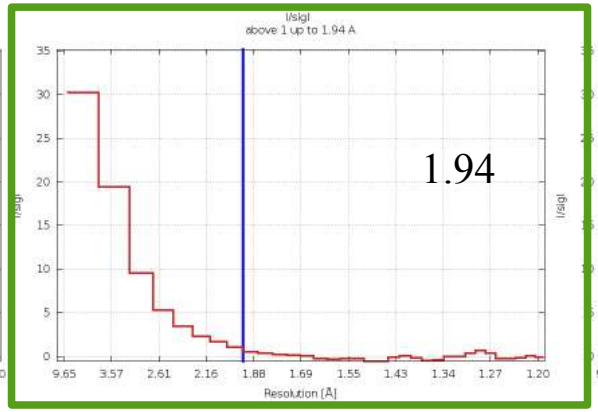
equal-Nobs

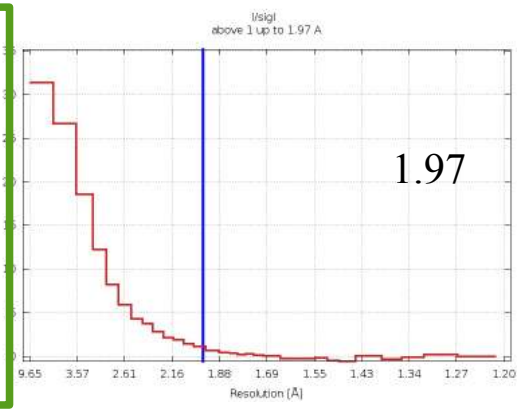


I/sigI>1

looking from the low-resolution end



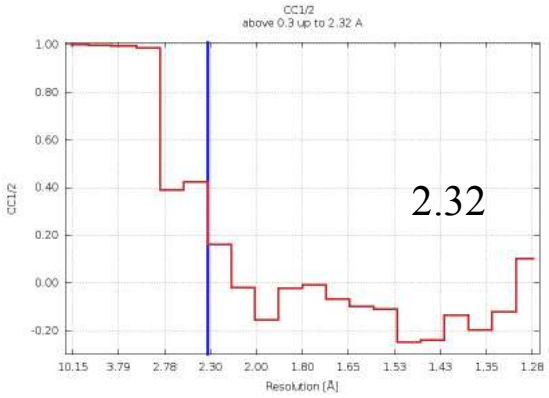




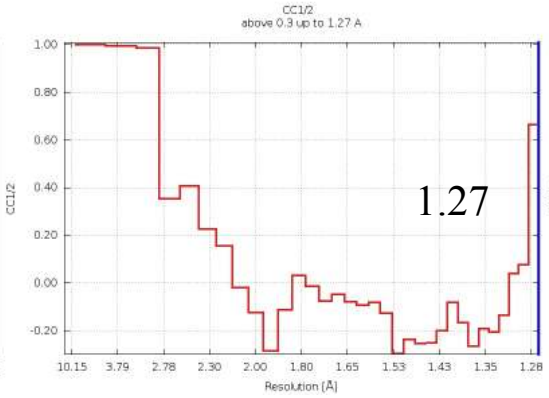
GΦL
Global Phasing Limited

How do we look at data/statistics: from low- or high-resolution end?

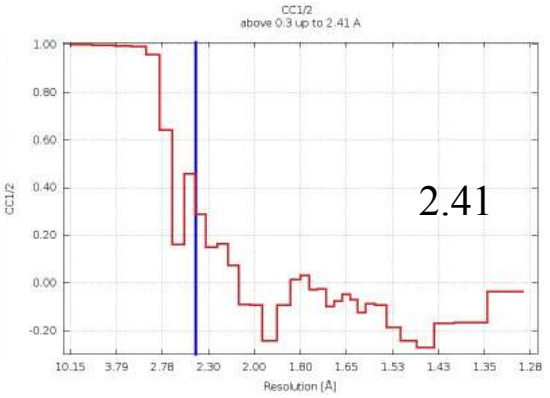
d*^2 (20 bins)



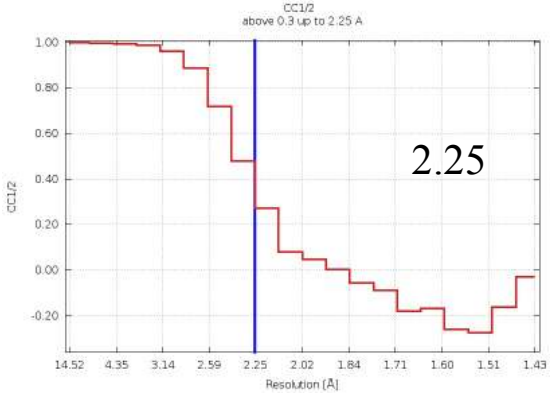
equal-volume

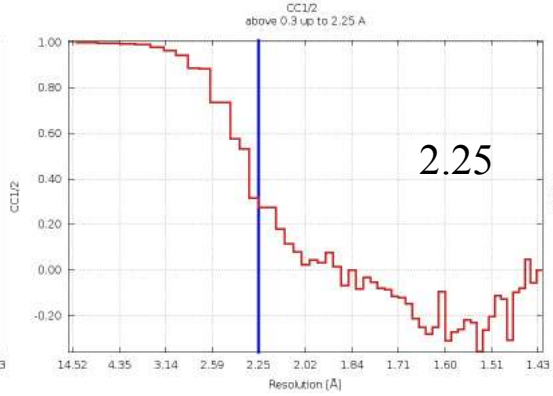


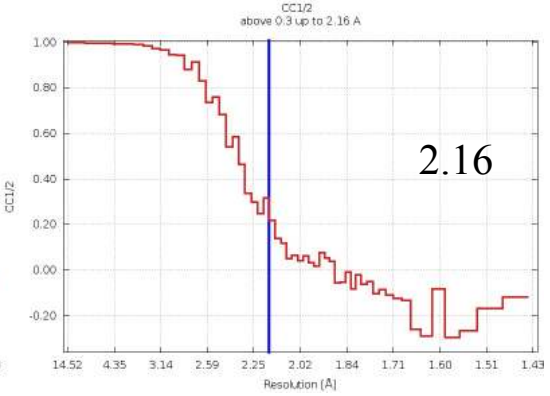
equal-Nobs



CC(1/2)>0.3 looking from the high-resolution end

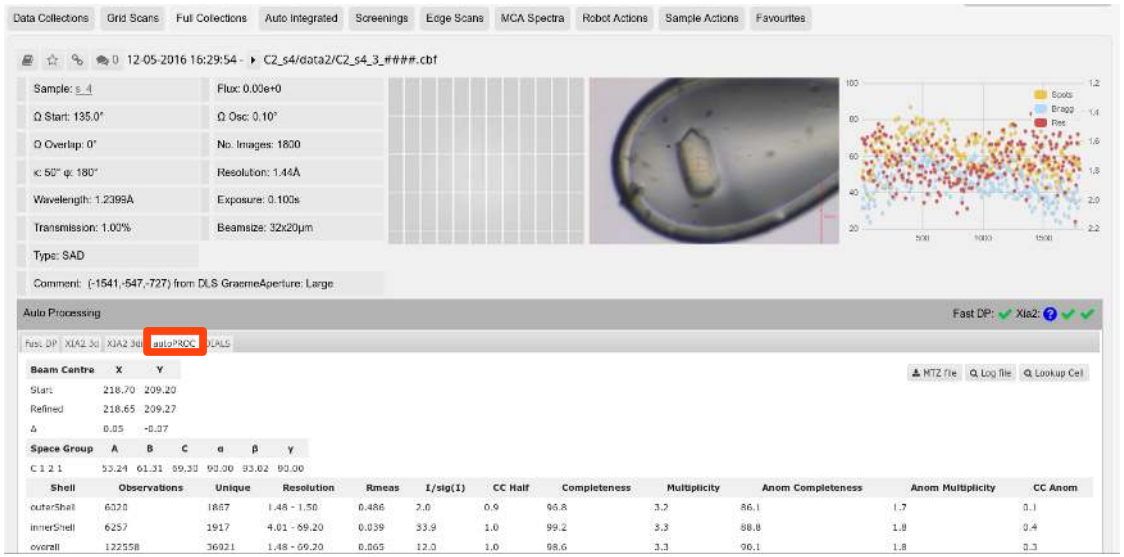






GΦL
Global Phasing Limited

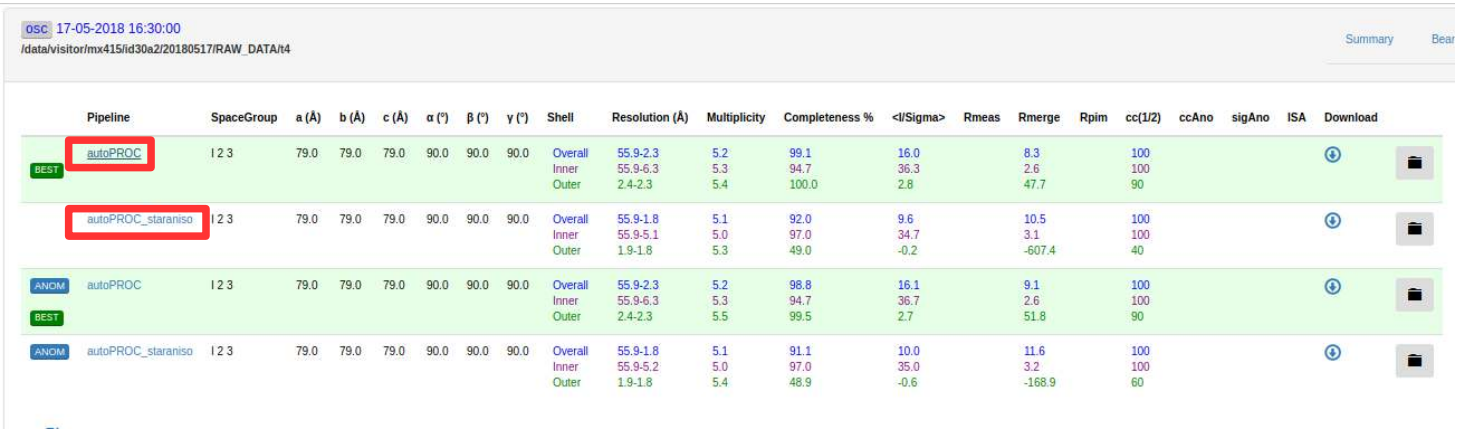
automatic processing at synchrotrons



Some generalisation required to provide similar/identical information for different pipelines.

Be aware of different binning (ie don't over-interpret small differences)

Diamond (DLS), UK



ESRF, F

also: ALBA (ES), APS (US), MaxIV (S), PETRA (D), SLS (CH), SSRF, …

GΦL
Global Phasing Limited

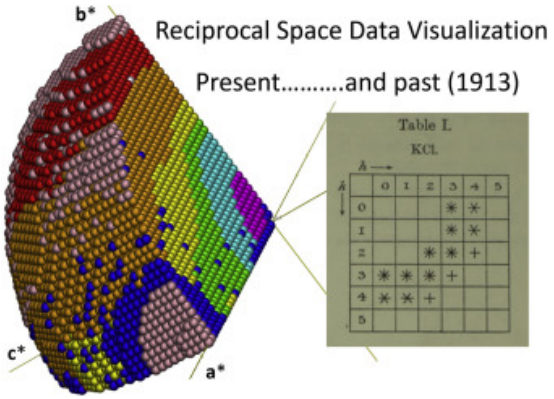
Data presentation

Rupp, B., 2018. Against Method: Table 1—Cui Bono?. Structure, 26(7), pp.919-923.

Structure
Perspective

Against Method: Table 1—Cui Bono?

Bernhard Rupp[1,2,*]
[1]k.-k.Hofkristallamt, San Diego, CA 92084, USA
[2]Division of Genetic Epidemiology, Medical University Innsbruck, Schöpfstraße 41, Innsbruck, Tyrol 6020, Austria
*Correspondence: bernhard.rupp@i-med.ac.at
https://doi.org/10.1016/j.str.2018.04.013



GΦL
Global Phasing Limited

Diffraction data quality metrics of PDB entries harvested from archived mmCIF files - from depositions (20211028 PDB release, aB_cif2table1 from BUSTER refinement package, Global Phasing Ltd.)

PDB identifier (4-letter code): Go

- by PDB id
- by data collection year
- by Synchrotron Beamline
- by Data scaling software
- by Refinement software

You can get to a particular entry (e.g. 5UHO) via the URL https://staraniso.globalphasing.org/table1/uh/5uho.html. This resource is automatically updated on a weekly basis. For any questions, please contact us at **proc-develop<AT>globalphasing.com.**

Last modification: Thu Oct 28 05:38:42 CEST 2021

General information	
Spacegroup name	P 32 2 1
Unit cell parameters	47.914 47.914 200.488 90.0 90.0 120.0
Wavelength	0.97932 Å

Data quality metrics	Overall	OuterShell
Low resolution limit [Å]	40.630	1.940
High resolution limit [Å]	1.900	1.900
Rmerge	0.050	0.433
Rmeas	0.062	0.540
Rpim	0.036	0.318
Total number of observations	-	-
Total number unique	22045	1386
<I/σ(I)>	14.30	3.10
Completeness [%]	99.9	99.9
Multiplicity	4.8	4.9
CC(1/2)	0.997	0.913

https://staraniso.globalphasing.org/table1/

Global Phasing Limited

“Table 1” - 5E74 : isotropic (standard)

```
Spacegroup name          C2221
Unit cell parameters     80.694 96.593 57.922 90.0 90.0 90.0
Wavelength               0.97942 A


Number of active ice-rings within this resolution range =  7

Criteria used in determination of resolution cut:
-------------------------------------------------
   Rpim           <=    0.6000
   I/sigI         >=    2.00
   CC(1/2)        >=    0.3000
```

	Overall	InnerShell	OuterShell
Low resolution limit	42.302	42.302	1.686
High resolution limit	1.657	4.497	1.657
Rmerge (all I+ & I-)	0.035	0.020	0.819
Rmerge (within I+/I-)	0.035	0.019	0.814
Rmeas (all I+ & I-)	0.038	0.021	0.873
Rmeas (within I+/I-)	0.039	0.022	0.927
Rpim (all I+ & I-)	0.013	0.007	0.297
Rpim (within I+/I-)	0.018	0.010	0.438
Total number of observations	204143	12396	11170
Total number unique	23978	1462	1325
Mean(I)/sd(I)	25.2	74.7	2.3
Completeness	88.0	99.9	99.9
Multiplicity	8.5	8.5	8.4
CC(1/2)	1.000	1.000	0.931
Anomalous completeness	87.4	99.9	99.8
Anomalous multiplicity	4.5	4.9	4.4
CC(ano)	-0.245	-0.208	-0.025
|DANO|/sd(DANO)	0.651	0.689	0.572

“Table 1” from autoPROC

assuming isotropic data

additional items:

- **active ice-rings**
- **criteria for cut-off decision**
- **... others?**

“Table 1” - 5E74 : anisotropic (STARANISO)

Spacegroup name C2221
Unit cell parameters 80.694 96.593 57.922 90.0 90.0 90.0
Wavelength 0.97942 A

Diffraction limits & eigenvectors of ellipsoid fitted to diffraction cut-off surface:

2.186	**1.0000**	**0.0000**	**0.0000**	**_a_***
1.442	**0.0000**	**1.0000**	**0.0000**	**_b_***
1.379	**0.0000**	**0.0000**	**1.0000**	**_c_***

Number of active ice-rings within this resolution range = 7

Criteria used in determination of diffraction limits:
---
local(I/sigI) >= 1.20

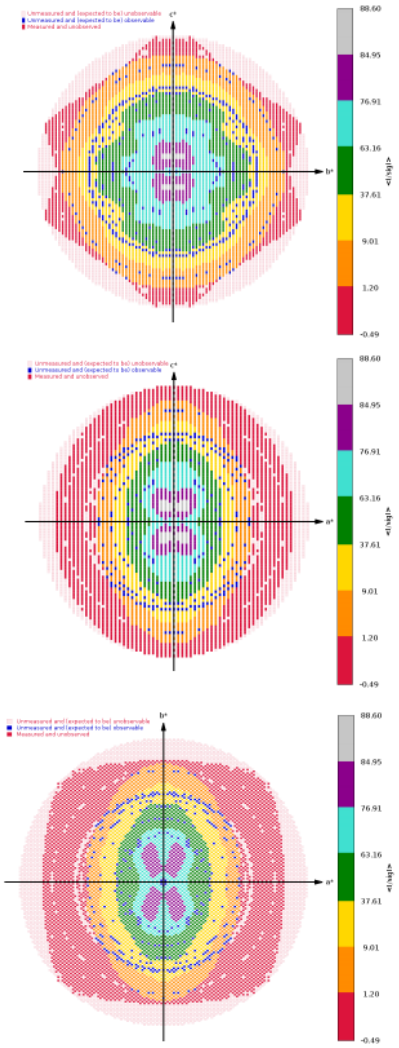
	Overall	InnerShell	OuterShell
Low resolution limit	42.302	42.302	1.504
High resolution limit	1.386	4.824	1.386
Rmerge (all I+ & I-)	0.035	0.019	1.146
Rmerge (within I+/I-)	0.034	0.019	1.074
Rmeas (all I+ & I-)	0.037	0.020	1.230
Rmeas (within I+/I-)	0.038	0.021	1.234
Rpim (all I+ & I-)	0.012	0.007	0.435
Rpim (within I+/I-)	0.018	0.009	0.596
Total number of observations	202888	10343	9127
Total number unique	24057	1202	1203
Mean(I)/sd(I)	25.5	75.5	1.4
Completeness (spherical)	**52.0**	**100.0**	**12.1**
Completeness (ellipsoidal)	**84.1**	**100.0**	**72.9**
Multiplicity	8.4	8.6	7.6
CC(1/2)	1.000	1.000	0.645
Anomalous completeness (spherical)	50.8	99.9	11.5
Anomalous completeness (ellipsoidal)	83.3	99.9	72.3
Anomalous multiplicity	4.5	5.0	4.0
CC(ano)	-0.213	-0.229	-0.015
|DANO|/sd(DANO)	0.656	0.679	0.629

Diffraction limits and directions

Completeness:

#(HKL with value) / #(possible HKL)

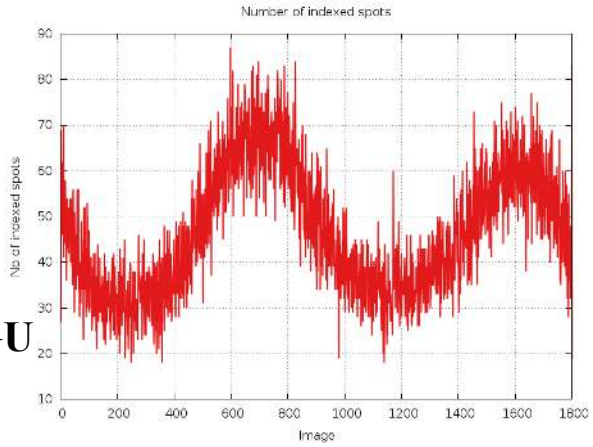
#observed / #observable



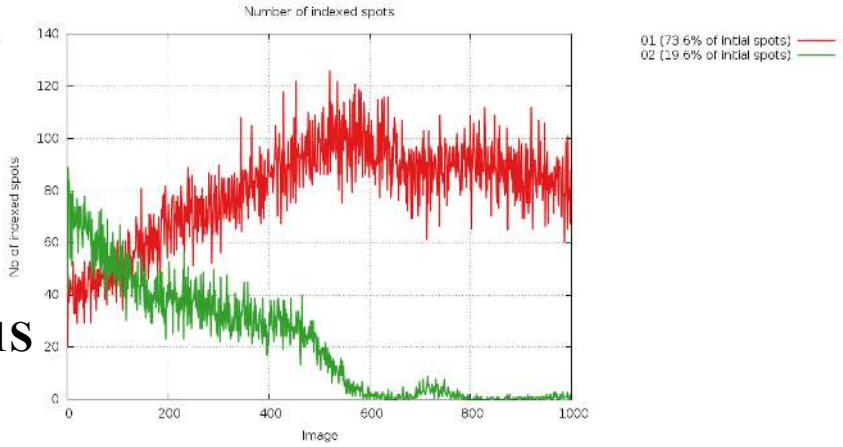
GΦL
Global Phasing Limited

Spots indexed vs image number per unique orientation ("iterative indexing")

4XGU

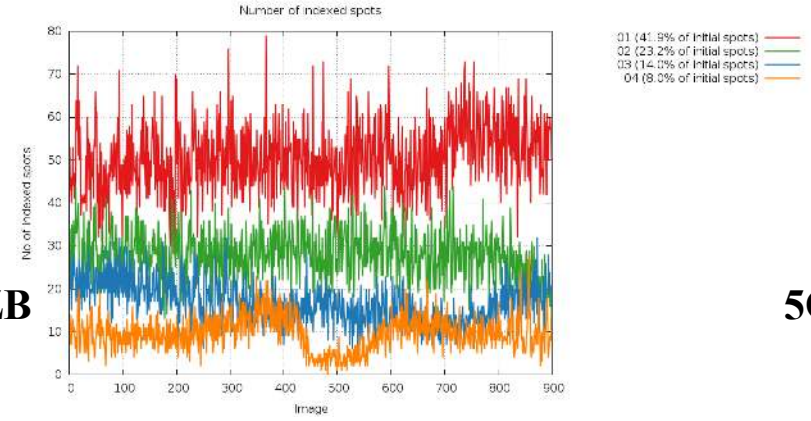


5J1S

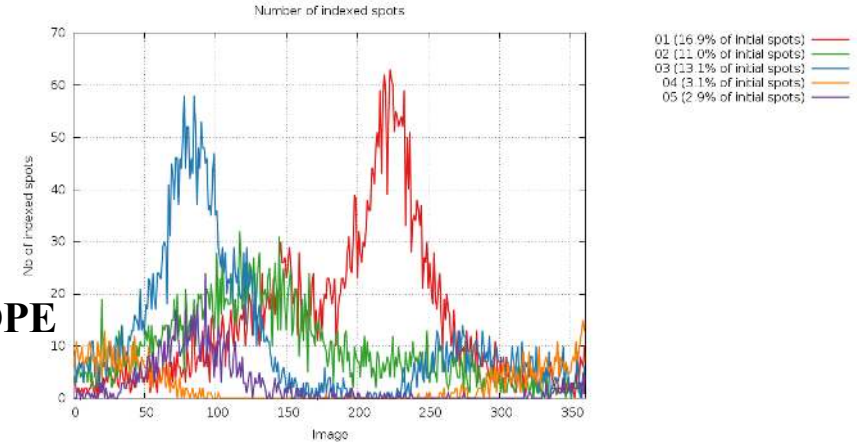


How to report this in 'Table 1'?

5SZB

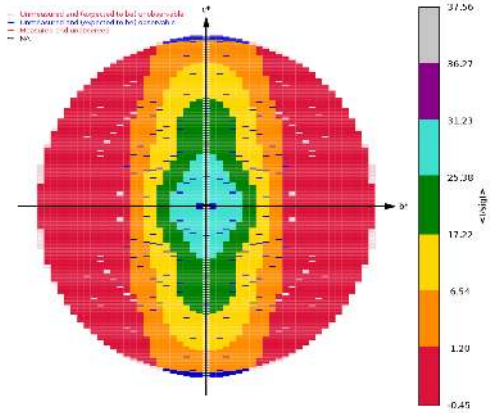


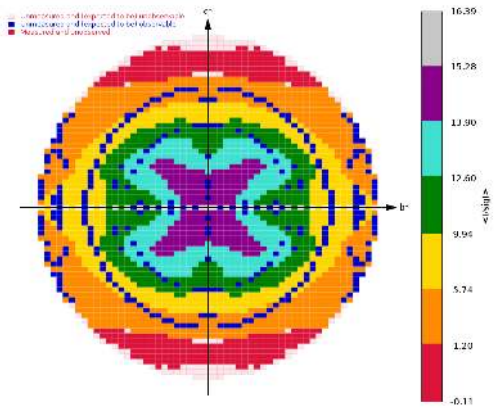
5OPE



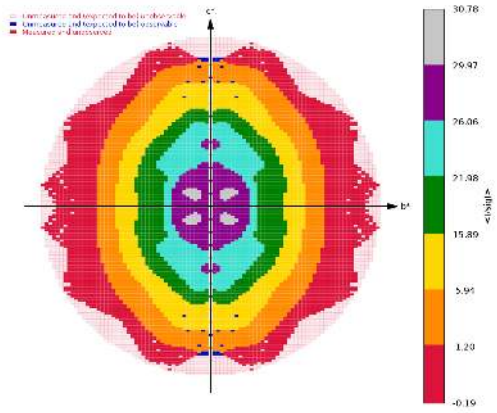
GΦL
Global Phasing Limited

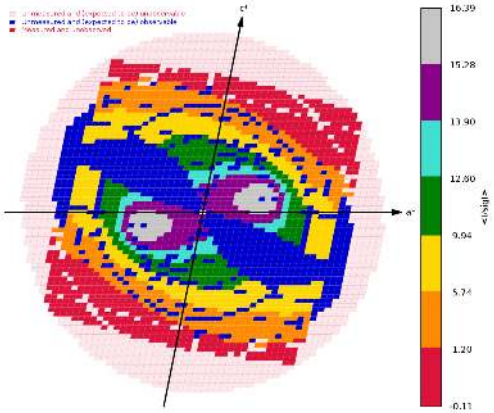
STARANISO within autoPROC – 2D plots



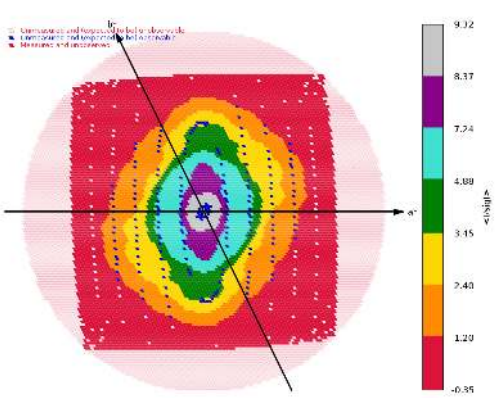


ice

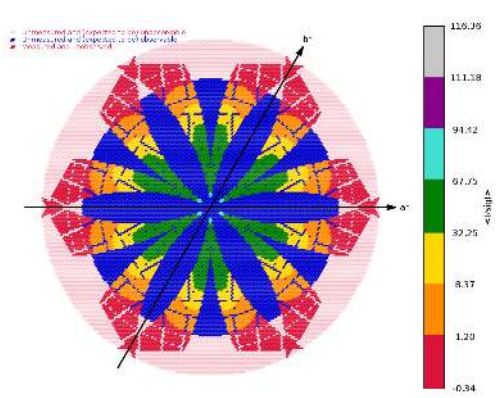




cusp + ice



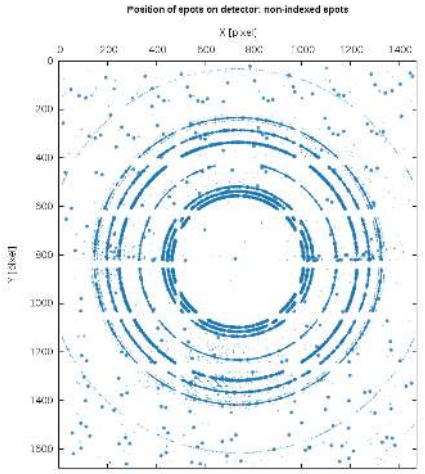
gaps



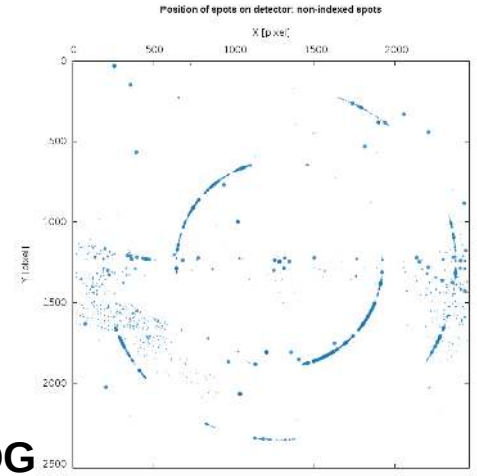
cusp

GΦL
Global Phasing Limited

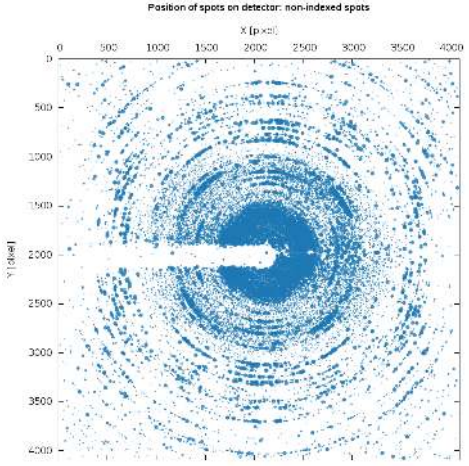
Unindexed spots on detector surface ice or powder rings



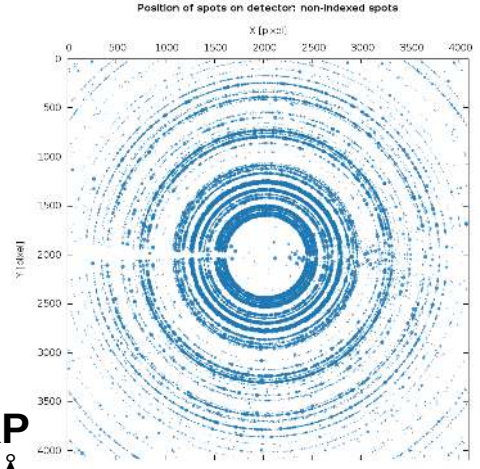
5E73
1.71Å



5HQG
2.00Å



5USC
2.20Å



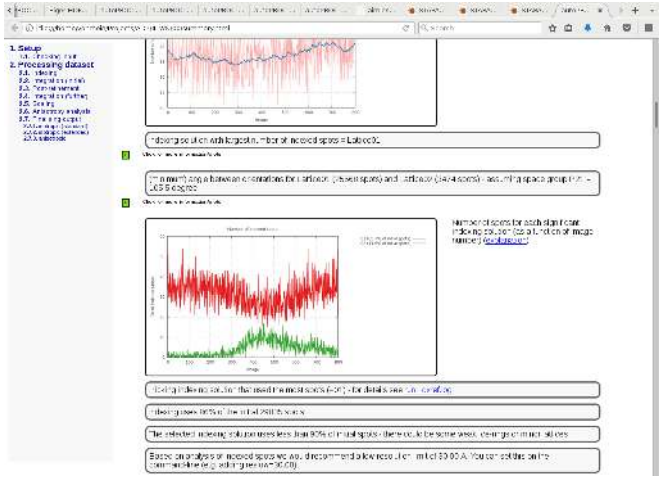
6CKP
1.15Å

How do we report and record this:

- **to the user (to be aware of e.g. cryo-cooling problems)**
- **upon deposition (to alert user of structure to potential contamination of data)**

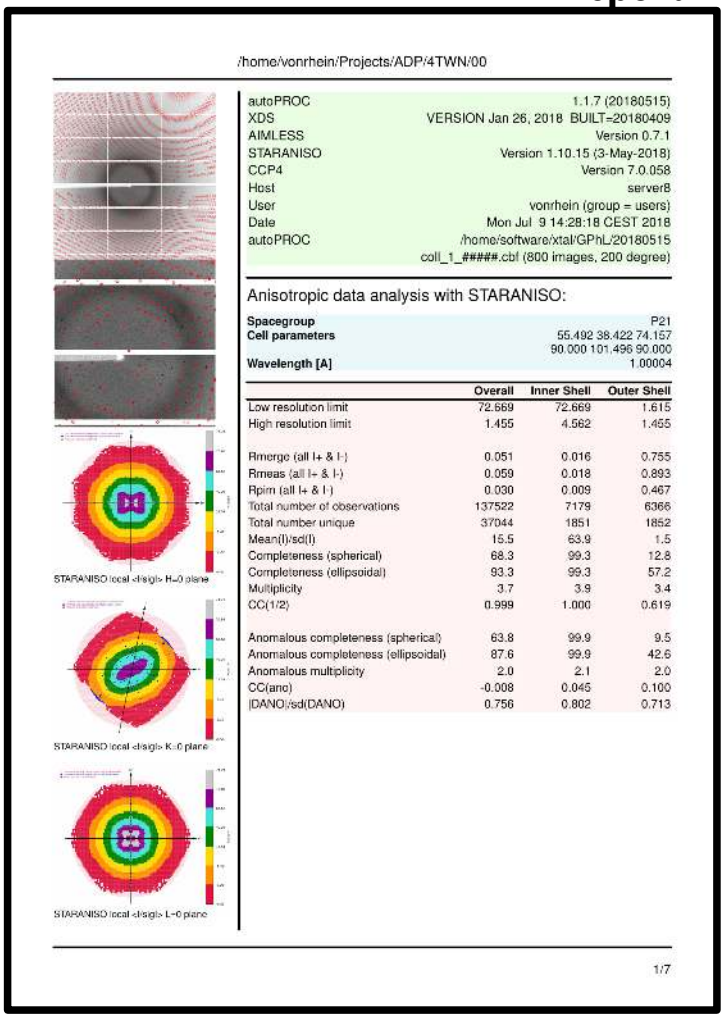
autoPROC reporting

Summary HTML (stages, decision making)



ISPyB XML file

PDF report



/home/vonrhein/Projects/ADP/4TWN/00

autoPROC	1.1.7 (20180515)
XDS	VERSION Jan 26, 2018 BUILT=20180409
AIMLESS	Version 0.7.1
STARANISO	Version 1.10.15 (3-May-2018)
CCP4	Version 7.0.058
Host	server8
User	vonrhein (group = users)
Date	Mon Jul 9 14:28:18 CEST 2018
autoPROC	/home/software/xtal/GPhL/20180515
	coll_1_#####.cbf (800 images, 200 degree)

Anisotropic data analysis with STARANISO:

Spacegroup	P21
Cell parameters	55.492 38.422 74.157
	90.000 101.496 90.000
Wavelength [A]	1.00004

	Overall	Inner Shell	Outer Shell
Low resolution limit	72.669	72.669	1.615
High resolution limit	1.455	4.562	1.455
Rmerge (all I+ & I-)	0.051	0.016	0.755
Rmeas (all I+ & I-)	0.059	0.018	0.893
Rpim (all I+ & I-)	0.030	0.009	0.467
Total number of observations	137522	7179	6366
Total number unique	37044	1851	1852
Mean(I)/sd(I)	15.5	63.9	1.5
Completeness (spherical)	68.3	99.3	12.8
Completeness (ellipsoidal)	93.3	99.3	57.2
Multiplicity	3.7	3.9	3.4
CC(1/2)	0.999	1.000	0.619
Anomalous completeness (spherical)	63.8	99.9	9.5
Anomalous completeness (ellipsoidal)	87.6	99.9	42.6
Anomalous multiplicity	2.0	2.1	2.0
CC(ano)	-0.008	0.045	0.100
|DANO|/sd(DANO)	0.756	0.802	0.713

STARANISO local <I/sigI> H=0 plane

STARANISO local <I/sigI> K=0 plane

STARANISO local <I/sigI> L=0 plane

1/7

```
staraniso_alldata-unique.mtz
staraniso_alldata-unique.table1
Data_1_autoPROC_STARANISO_all.cif
```

GΦL
Global Phasing Limited

Summary

- Is there such a thing as “the resolution”?
- We should probably start talking about diffraction limit(s).

- Describing diffraction data via a single number makes life easy ... but doesn’t really do the variation between different datasets justice: multiple lattices, ice, detector gaps, cusp, split spots etc.

- Thinking about binning is boring - but important when assuming isotropic diffraction and making decisions.

- The STARANISO approach (local <I/sig(I)> to define anisotropic cut-off surface) excludes noise and includes all significant data while avoiding (isotropic) binning headache.
  - Completeness: fraction of data actually observed relative to potentially observable data (i.e. significant, not just Miller indices).

GΦL
Global Phasing Limited

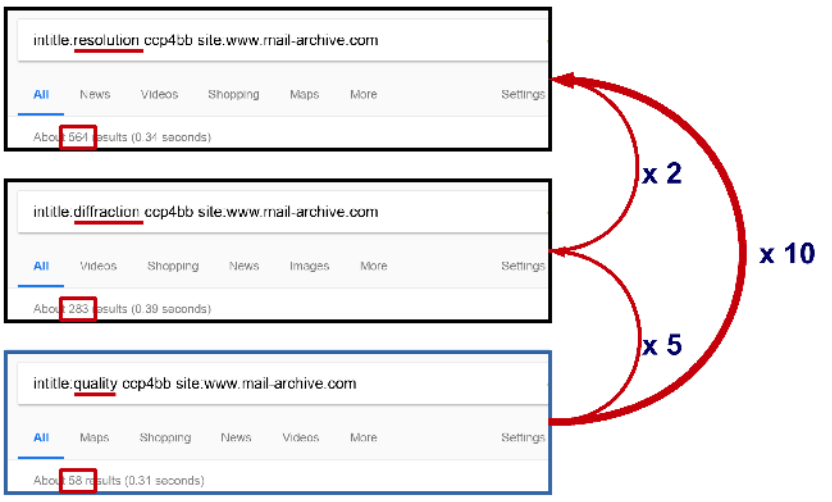
Resolution, data quality and all that ...

Clemens Vonrhein
Global Phasing Ltd.

DLS-CCP4 Workshop 2021

GΦL
Global Phasing Limited

Too much focus on single number (resolution) to describe model quality?



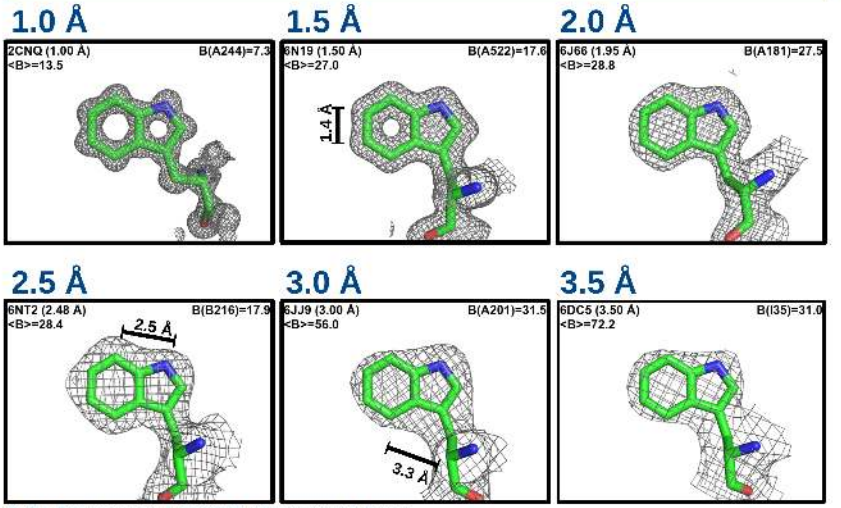
Especially resolution seems to preoccupy users a lot: it is mentioned 10 times more often in the subject line of messages to the CCP4BB than e.g. the term “quality”.

Why is that?

Because we traditionally equate data resolution with model quality: the higher the resolution, the better the model.

GΦL
Global Phasing Limited

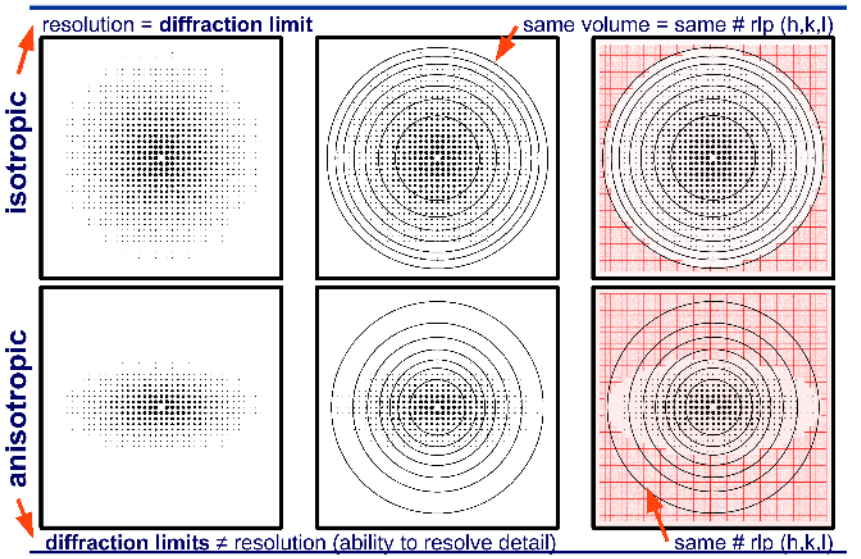
Resolution = ability to resolve detail



2mFo-DFc maps after BUSTER refinement at 1.0 rms

GΦL
Global Phasing Limited

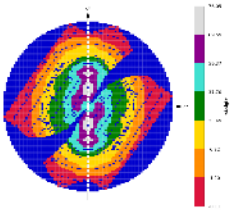
Resolution, diffraction limit and binning



rlp = reciprocal lattice point

GΦL
Global Phasing Limited

STARANISO – http://staraniso.globalphasing.org/



- main author: Ian Tickle (Global Phasing)
- **anisotropic** cut-off of merged intensity data
  - avoids problems of
    - isotropic resolution cut-offs
    - throwing away good data (in one direction)
    - including poor data (in another direction)
- Bayesian estimation of structure factor amplitudes
  - more rigorous treatment of prior and anomalous data
  - for more details see:
    https://staraniso.globalphasing.org/staraniso_about.html
- (optional) anisotropic correction of data
- Part of autoPROC ('process', 'aP_scale')
  - user is given classical, isotropic (TRUNCATE) output
  - but also data after **anisotropic** analysis (STARANISO)

merged or unmerged data (MTZ, XDS)

Anisotropy is not necessarily ellipsoidal!

GΦL
Global Phasing Limited

Isotropic vs anisotropic

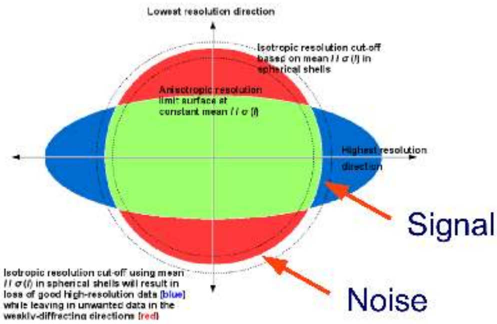
GLOSSARY of terminology related to anisotropy

ABOUT ANISOTROPY

[illegible]

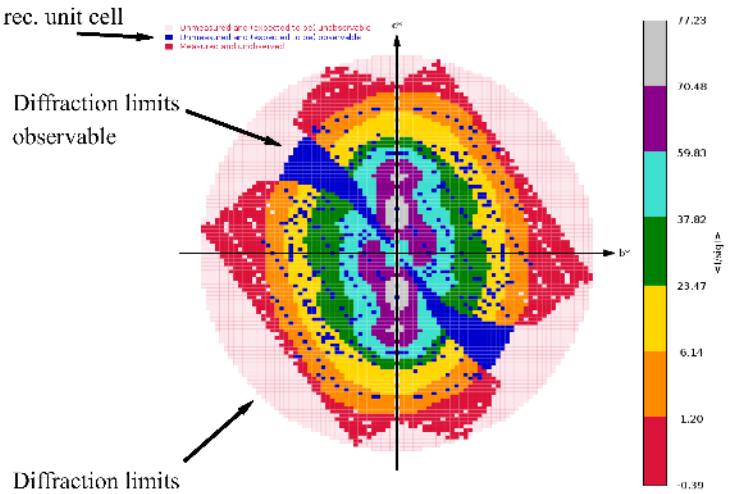
The diffraction limit

[illegible]



GΦL
Global Phasing Limited

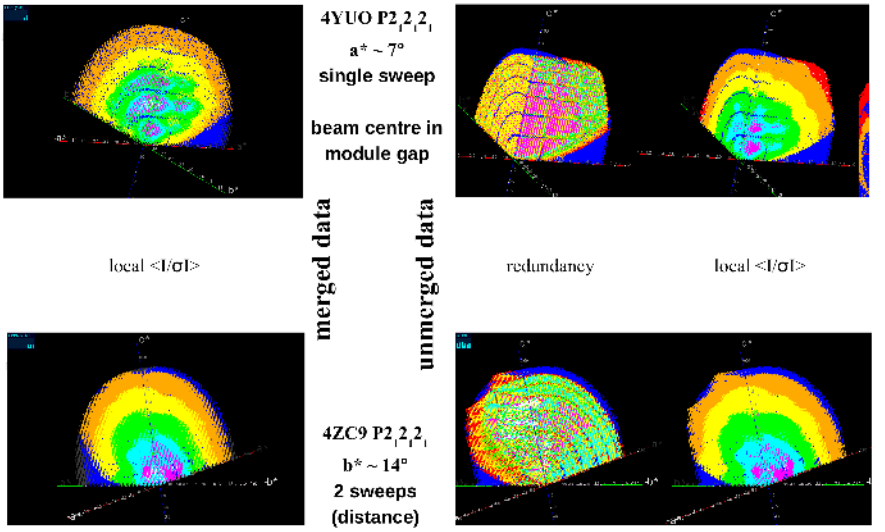
autoPROC: STARANISO plots



GΦL
Global Phasing Limited

3D WebGL visualisation in STARANISO

https://staraniso.globalphasing.org/cgi-bin/PDBpeep.cgi
https://staraniso.globalphasing.org/



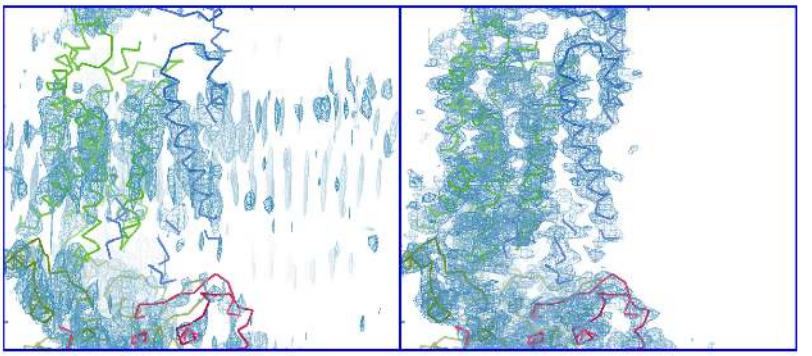
GΦL
Global Phasing Limited

staraniso.globalphasing.org

- **The following results were communicated to us by Neil Paterson at Diamond Light Source.**

The panels show the maps after density modification (CCP4 DM program) using (left) the standard protocol with an isotropic resolution cut-off, and (right) the STARANISO protocol with anisotropic resolution cut-off and correction. Compare in particular the density in the more mobile loop regions at the ends of the helices, and also the flatness of the solvent region (right-hand half of each map).

Neil comments: "STARANISO has been key to solving our structure. We've got a severely anisotropic case that goes from around 7 Å along a^* to 3.0-3.5 Å along b^*/c^*, with lots of artifacts in the maps. We can find the substructure using Shel-X without correction but the maps are pretty bad; once corrected it works a treat (77% solvent)."



GΦL
Global Phasing Limited

Evolution of diffraction data metrics - "Table 1"

R_{merge} (R_{sym})
Completeness
Redundancy (multiplicity)
<I/σI>
R_{meas}
R_{pim}
$CC_{1/2}$

Table I. *X-ray Data Reduction and Crystallographic Refinement Statistics*

(A) X-ray data reduction statistics	
Space group	$P2_12_12_1$
Unit cell dimensions	
a, *b*, *c*	33.5 Å, 68.2 Å, 137.6 Å
Resolution	68.8 Å–2.25 Å
(last shell)	(2.26 Å–2.25 Å)
Total measurements (last shell)	109,559 (1,168)
Number of unique reflections	15,637
(last shell)	(163)
Wavelength	1 Å
R-merge[a] (last shell)	0.065 (0.592)
I/σ(I) (last shell)	16.4 (3.3)
Completeness (last shell)	0.997 (1)
Redundancy (last shell)	7 (7.2)

As function of resolution in spherical shells (equal spacing or equal volume): overall and outer shell

R_{merge} : Arndt et al (1958)
R_{meas} : Diederichs & Karplus (1997), Weiss & Hilgenfeld (1997)
R_{pim} : Weiss (2001)
$CC_{1/2}$: Karplus & Diederichs (2012)

C. Vonrhein, Global Phasing Ltd
DLS-CCP4 2021

The metrics for describing data quality have evolved over the last few years

New ariants of R-values have appeared and the use of correlation coefficients is now standard.

Our picture of how to best describe data has clearly evolved – so maybe it is time to also look at our notion of resolution and completeness?

GΦL
Global Phasing Limited

Binning

- **Equally spaced** in $(d^*)^2$
  - Number of measurements used for computing bin averages increases with resolution
  - Low-resolution statistics can be unreliable (too few measurements)
  - High-resolution statistics not finely enough sampled?
  - Used by XDS, AIMLESS

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA
3.13	75353	14441	14448	100.0%
2.22	136754	26193	26370	99.3%
1.81	178487	34019	34139	99.6%
1.57	207171	40372	40378	100.0%
1.40	214727	45702	45725	99.9%
1.28	193266	50533	50658	99.8%
1.18	142566	54473	55100	98.9%
1.11	57895	43932	59148	74.3%
1.04	15808	15171	63013	24.1%
total	1212027	324824	388979	83.5%

- **Equal volume** (to have same number of reciprocal lattice points in each bin)
  - Allows different sampling (coarse or fine) depending on requirements
  - Remember ice-rings (3.89 3.66 3.44 2.67 2.25 2.07 1.95 1.91 1.88 1.72 1.52 1.47 1.44 1.37 1.36 1.30 1.27 1.26 1.22 1.17 1.12)!
  - Adequate for isotropic data that is complete with homogeneous multiplicity
- **Equal number** of (actual) observations
  - Can be seen as generalisation of idea behind "equal volume" binning
  - Automatically self-adjusting for anisotropic (STARANISO) and incomplete data (serial crystallography, LCP, in-plate, ...)

GΦL
Global Phasing Limited

Binning: decision making, data description

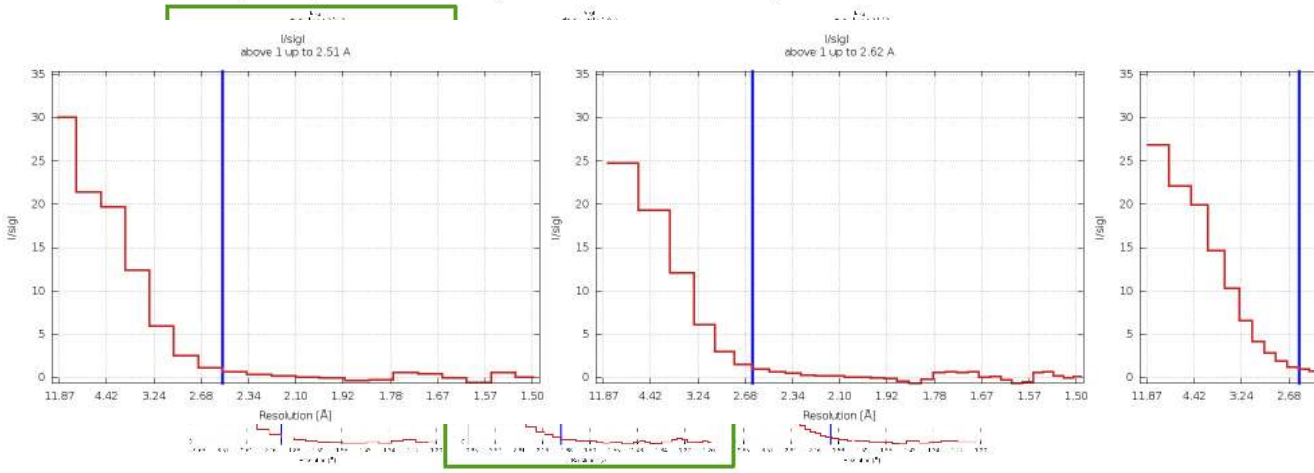
Binning is boring - but an important aspect here:

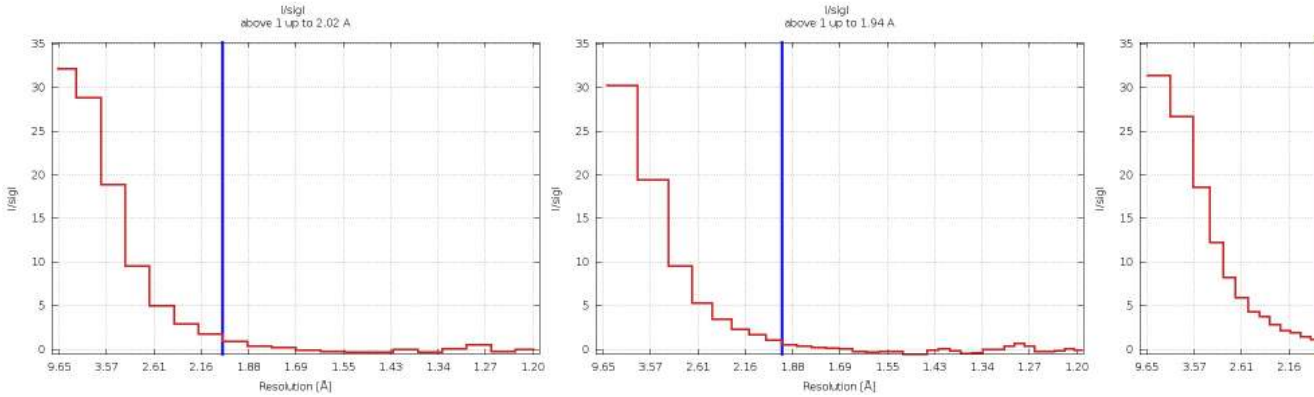
- Decision making:
  - Which images/datasets/reflections to include
  - Which reflections to include: isotropic/anisotropic diffraction limit
- Comparisons:
  - Between programs and pipelines
  - Between different processing options
- Basis of decision making
  - Raw data (binned statistics)
  - Smoothed (spline, Bezier, …)
  - Ice-ring resolution ranges included/excluded/smoothed?

GΦL
Global Phasing Limited

What is my high-resolution limit?
… depending on binning method

d*^2 (20 bins) | equal-volume | equal-Nobs

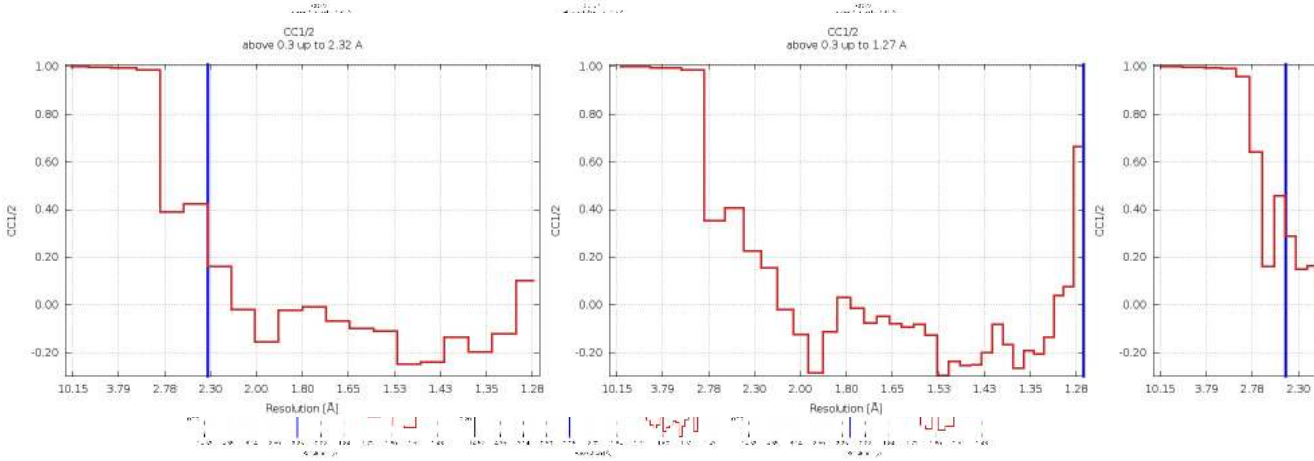


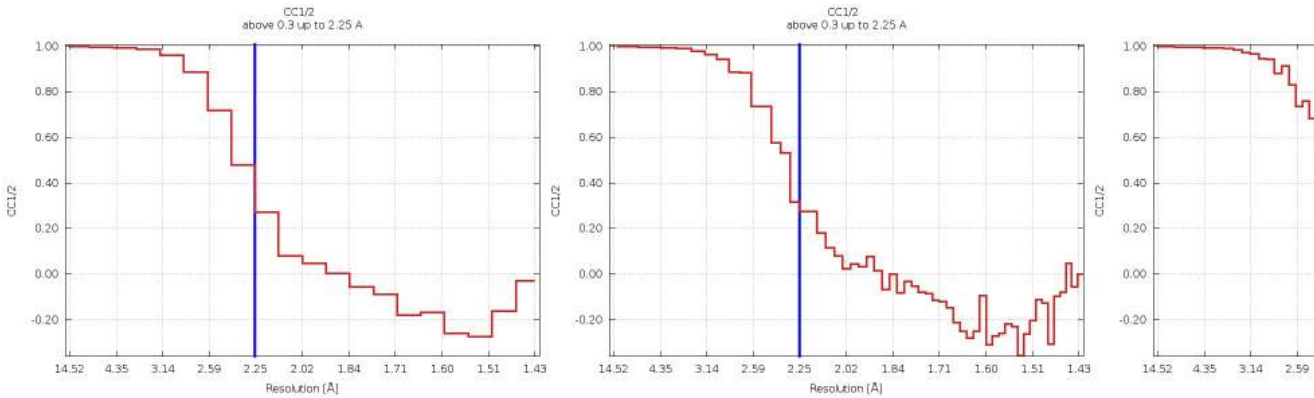


GΦL
Global Phasing Limited

How do we look at data/statistics: from low- or high-resolution end?

d*^2 (20 bins) equal-volume equal-Nobs





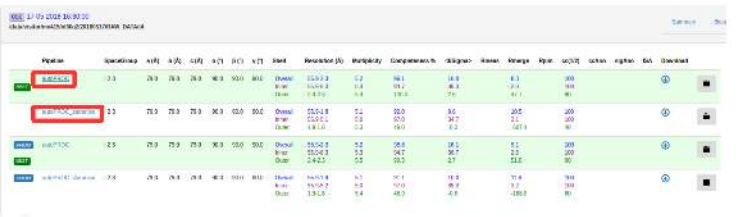
GΦL
Global Phasing Limited

automatic processing at synchrotrons

Some generalisation required to provide similar/identical information for different pipelines.

Be aware of different binning (ie don't over-interpret small differences)

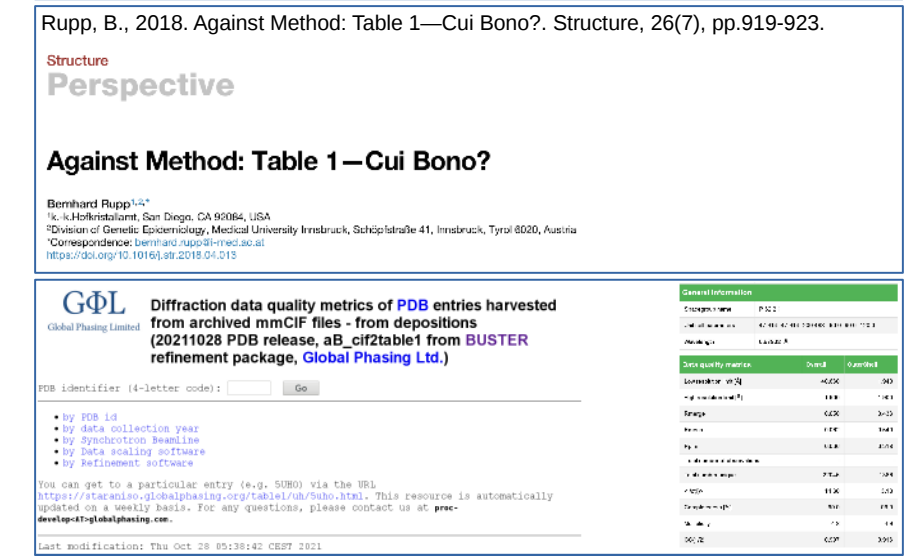
Diamond (DLS), UK



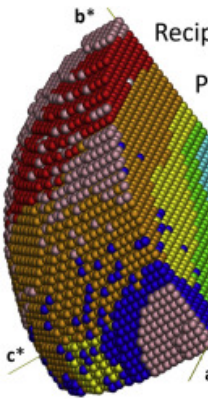
ESRF, F

also: ALBA (ES), APS (US), MaxIV (S), PETRA (D), SLS (CH), SSRF, …

Data presentation



https://staraniso.globalphasing.org/table1/



When it comes to archiving, it is not only the storing of a rich set of meta-data that is important, but also the way of displaying the stored data quality metrics to the user of a particular structural model. No one will have the time to look through long columns of raw numbers. At the same time, an assessment of data quality to judge the quality of the archived model in the context of the experimental data is crucial - especially for users that will have difficulty interpreting the details of an actual electron density map.

In his 2018 paper, Bernhard Rupp argued for a revamping of the traditional “Table 1” and cited the way STARANISO visualises 3D data quality metrics as an alternative. As a way of standardising the already available data quality metrics information of existing PDB entries, we provide a web-service at the URL shown below. This might show a “Table 1”

GΦL
Global Phasing Limited

"Table 1" - 5E74 : isotropic (standard)

Spacegroup name C2221
Unit cell parameters 88.894 93.693 57.922 90.0 90.0 90.0
Wavelength 0.97942 A

Number of active ice-rings within this resolution range = 3

Criteria used in determination of resolution cuts

Rpim <~ 0.6000
I/sigI >~ 2.00
CC(1/2) >~ 0.3000

	Overall	InnerShell	OuterShell
Low resolution limit	47.892	47.987	1.696
High resolution limit	1.657	1.693	1.657
Average I(all I+ & I-)	0.035	0.020	0.819
Average I(within I+/I-)	0.035	0.018	0.814
Error I(all I+ & I-)	0.038	0.021	0.873
Error I(within I+/I-)	0.039	0.022	0.927
Rpim (all I+ & I-)	0.011	0.006	0.297
Rpim (within I+/I-)	0.010	0.010	0.430
Total number of observations	200143	12595	11178
Total number unique	23978	1462	1325
Mean(I)/Sd(I)	26.2	74.7	2.3
Completeness	88.0	59.5	99.9
Multiplicity	8.5	8.5	8.4
CC(1/2)	1.000	1.000	0.831
Anomalous completeness	87.4	59.5	99.8
Anomalous multiplicity	4.5	4.5	4.6
CC(ano)	-0.245	-0.285	-0.025
SigAno|Sd|DANO	0.651	0.588	0.572

"Table 1" from autoPROC

assuming isotropic data

additional items:
- active ice-rings
- criteria for cut-off decision
- ... others?

GΦL
Global Phasing Limited

"Table 1" - 5E74 : anisotropic (STARANISO)

```
Diffraction limits & eigenvectors of ellipsoid fitted to diffraction cut-off surface:
  2.186        1.0000   0.0000   0.0000        a_*
  1.682        0.0000   1.0000   0.0000        b_*
  1.379        0.0000   0.0000   1.0000        c_*

Number of active low-limits within this resolution range =  0

Criteria used in determination of diffraction limits:
-----------------------------------------------------
   Local(I/sig(I)  >=    1.20
```

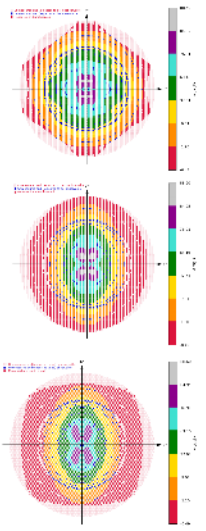
	Overall	InnerShell	OuterShell
Low resolution limit	42.382	42.382	1.404
High resolution limit	1.386	1.824	1.386
Rmerge (all I+ & I-)	0.055	0.019	1.116
Rmerge (within I+/I-)	0.034	0.019	1.074
Rmeas (all I+ & I-)	0.037	0.022	1.230
Rmeas (within I+/I-)	0.038	0.021	1.234
Rpim (all I+ & I-)	0.012	0.008	0.435
Rpim (within I+/I-)	0.010	0.008	0.396
Total number of observations	202000	1030	9167
Total number unique	26057	1202	1209
Mean(I)/sd(I)	25.5	75.0	1.6
Completeness (spherical)	**52.0**	**100.0**	**12.1**
Completeness (ellipsoidal)	**84.1**	**100.0**	**72.9**
Multiplicity	8.4	8.6	7.6
CC(1/2)	1.000	1.000	0.668
Anomalous completeness (spherical)	50.8	99.9	11.5
Anomalous completeness (ellipsoidal)	83.9	99.9	72.9
Anomalous multiplicity	4.5	5.0	4.0
CC(anom)	-0.213	-0.225	-0.015
SigAno(Ano:DANO)	0.686	0.875	0.629

Diffraction limits and directions

Completeness:

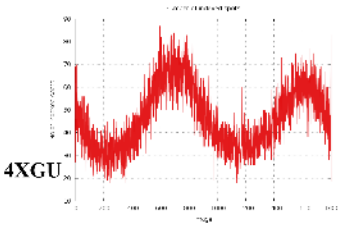
#(HKL with value) / #(possible HKL)

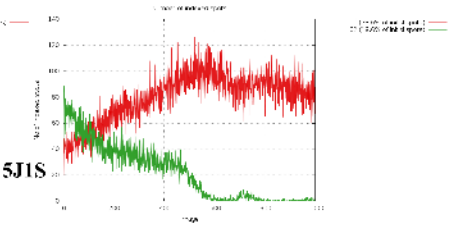
#observed / #observable



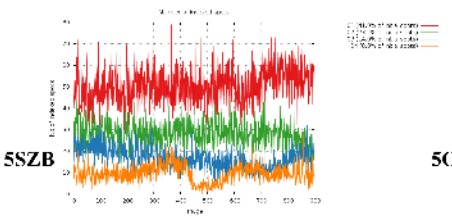
GΦL
Global Phasing Limited

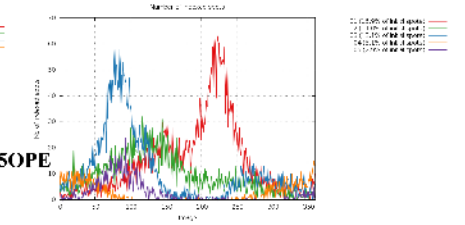
Spots indexed vs image number per unique orientation ("iterative indexing")





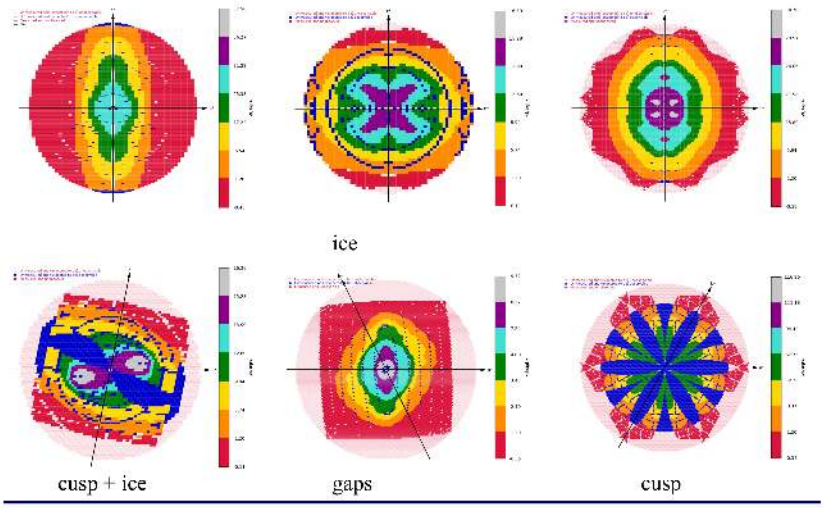
How to report this in "Table 1"?





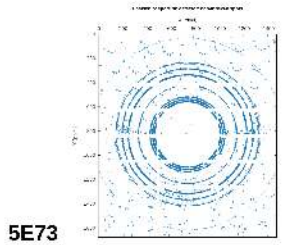
GΦL
Global Phasing Limited

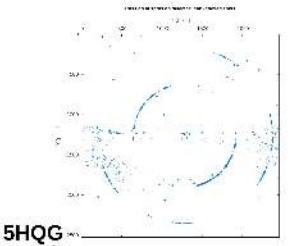
STARANISO within autoPROC – 2D plots

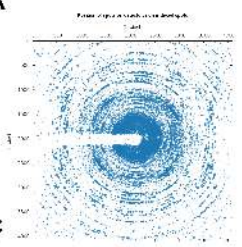


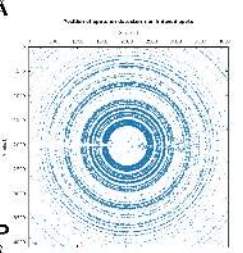
GΦL
Global Phasing Limited

Unindexed spots on detector surface ice or powder rings


5E73
1.71Å


5HQG
2.00Å


5USC
2.20Å


6CKP
1.15Å

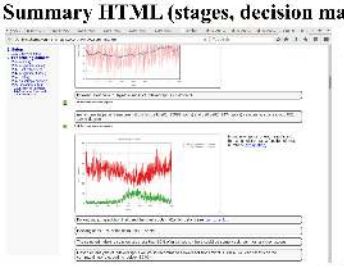
How do we report and record this:

- **to the user (to be aware of e.g. cryo-cooling problems)**
- **upon deposition (to alert user of structure to potential contamination of data)**

GΦL
Global Phasing Limited

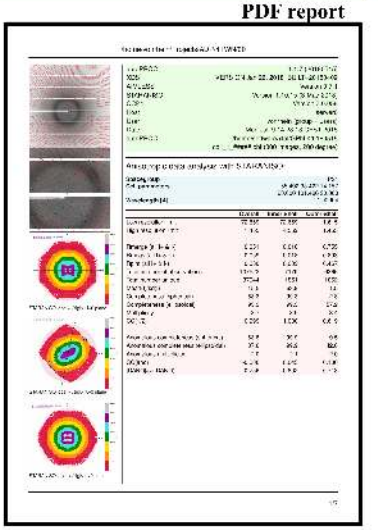
autoPROC reporting

Summary HTML (stages, decision making)



ISPyB XML file

PDF report



```
staraniso_alldata-unique.mtz
staraniso_alldata-unique.table1
Data_1_autoPROC_STARANISO_all.cif
```

GΦL
Global Phasing Limited

Summary

- Is there such a thing as “the resolution”?
- We should probably start talking about diffraction limit(s).

- Describing diffraction data via a single number makes life easy ... but doesn’t really do the variation between different datasets justice: multiple lattices, ice, detector gaps, cusp, split spots etc.

- Thinking about binning is boring - but important when assuming isotropic diffraction and making decisions.

- The STARANISO approach (local <I/sig(I)> to define anisotropic cut-off surface) excludes noise and includes all significant data while avoiding (isotropic) binning headache.
  - Completeness: fraction of data actually observed relative to potentially observable data (i.e. significant, not just Miller indices).