Data Reduction

Space groups, scaling and data quality

CCP4-BGU workshop 2020

Ed Lowe, University of Oxford
With thanks to Phil Evans

Department of Biochemistry

UNIVERSITY OF OXFORD

Scaling and Merging

$|F|^2$ —*Experiment*→ I

lots of effects
("errors")

Our job is to invert the experiment: we want to *infer* $|F|^2$ and $|F|$ from our measurements of intensity I

I —*Model of experiment*→ $|F|^2$ - - - → $|F|$

Parameterise
experiment

Intensities
from eg Mosflm, XDS, DIALS
h k l I σ(I) etc

→

h k l F σ(F) I σ(I)
FreeR_flag

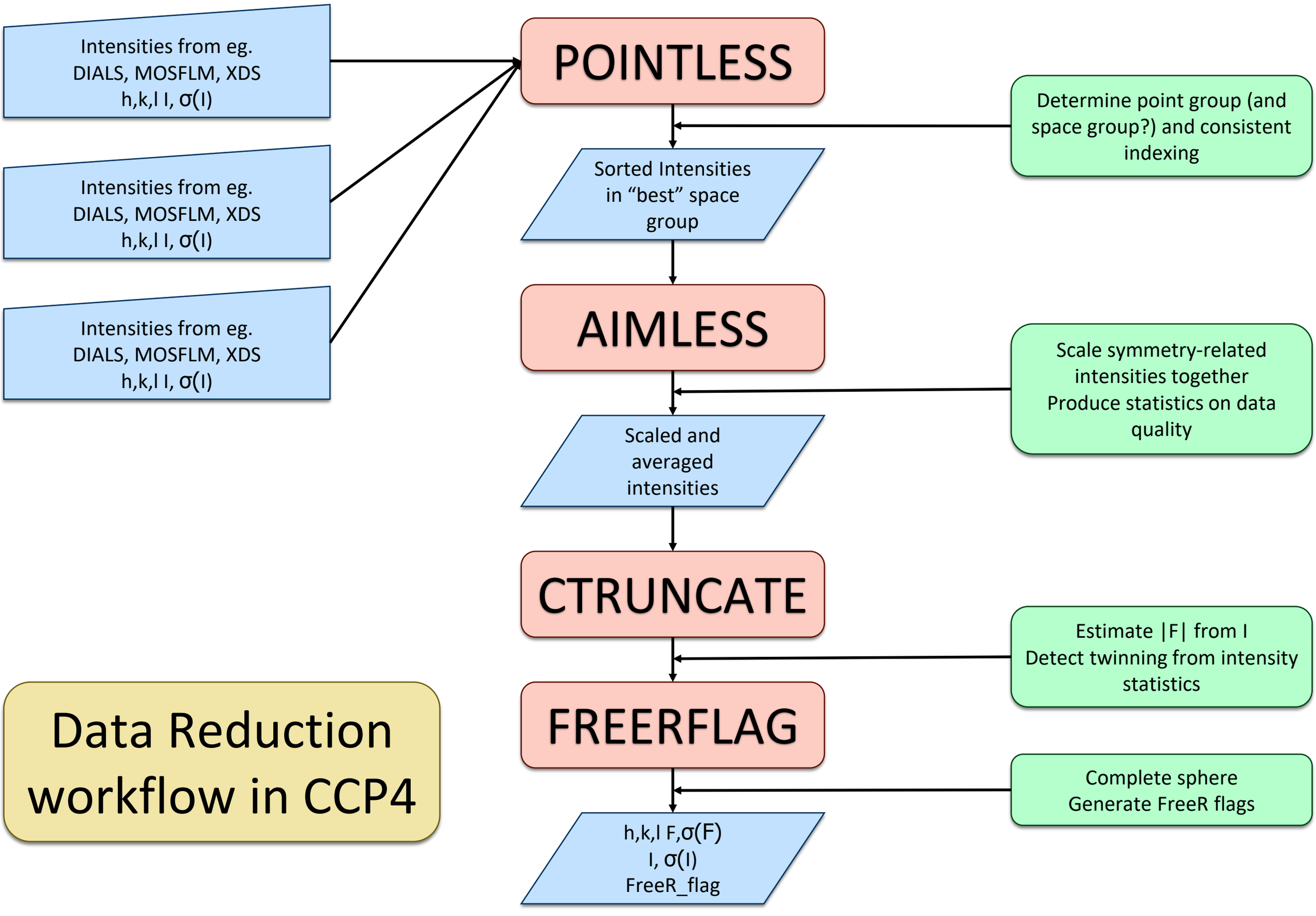
Overview of data reduction process

1. Determine point group and if possible space group
   - *we need the point group to scale the data*
   - *too low symmetry makes solving the structure harder, (though not impossible)*
2. Scale data to make it internally consistent

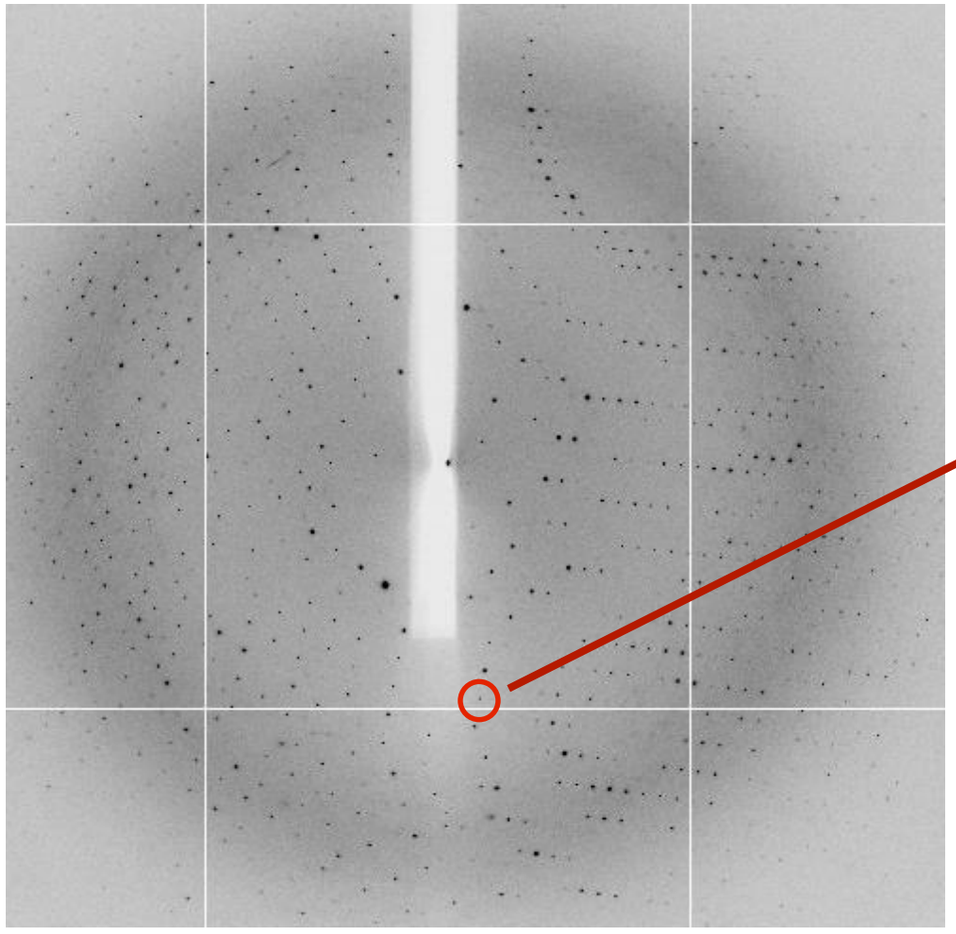
   analyse for:-
   - *maximum resolution*
   - *radiation damage*
   - *data quality*
3. Analyse for pathologies, and estimate amplitude
   - *twinning*
   - *translational non-crystallographic symmetry*

NB I am discussing data from one or a few crystals, not from hundreds of crystals, not serial crystallography

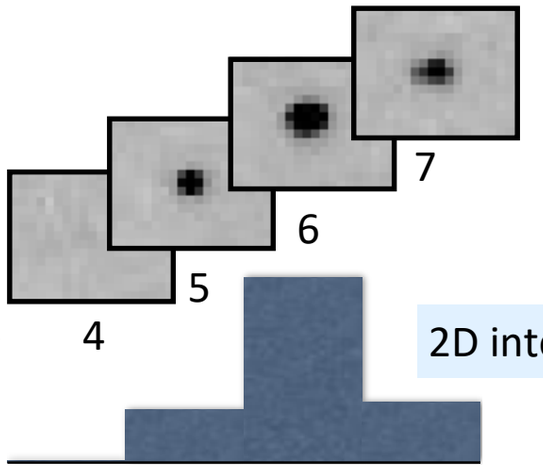
Data Reduction workflow in CCP4



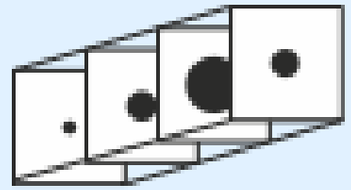
Track one reflection through the process



Spot over 4 images



Spot profile



3D integration *eg* DIALS, XDS, d*trek

2D integration *eg* Mosflm, hkl2000

In MTZ file from Mosflm, ordered by image (BATCH) number
Entries spread through file

h	k	l	M/ISYM	BATCH	I	SIGI	IPR	SIGIPR
...								
-20	12	10	258	4	13	3	7	3
...								
-20	12	10	258	5	304	24	322	24
...								
-20	12	10	258	6	1072	84	1101	84
...								
-20	12	10	258	7	349	27	324	27

Summation integration

Profile fit

After POINTLESS: Possibly reindexed observation parts grouped by reduced hkl (sorted)

Symmetry-related observations

Reduced hkl			*Original hkl*			*Full/part+ Symmetry number*	*Image number*	*Intensities & σ(I)*				*Fraction*	*Detector pixel*		*Rotation*	*Width*	*LP*	*Partial serial*		*Flags*
-20	12	10	-20	-12	10	4	36	1566.08	126.54	1682.27	53.25	2.26	2013.67	1406.74	295.54	0.41	0.17	0	0	5211.00
-20	12	10	20	-12	-10	258	4	13.33	3.88	7.84	3.88	0.00	1605.33	2028.12	265.50	3.01	0.03	501	0	11.00
						258	5	304.72	24.30	322.00	24.30	0.27	1605.29	2028.11	265.46	2.98	0.03	402	0	11.00
						258	6	1072.06	84.15	1101.98	84.15	0.51	1605.31	2028.10	265.48	2.95	0.03	302	0	12.00
						258	7	349.08	27.75	324.41	27.75	0.21	1605.33	2028.07	265.43	2.93	0.03	404	0	11.00
-20	12	10	20	12	-10	259	46	1253.10	102.71	1381.90	102.71	0.99	1049.15	1664.63	305.83	0.40	0.17	201	0	11.00
						259	47	7.35	27.23	28.40	27.23	0.01	1049.21	1664.61	305.83	0.39	0.17	202	0	11.00

Unmerged file from Pointless, multiple entries for each unique hkl
(note that we need to know the point group to connect these)

Full	-20	12	10	-20	-12	10	4	36	1566.08	126.54	1682.27	53.25	2.26	2013.67	1406.74	295.54	0.41	0.17	0	0	5211.00
	-20	12	10	20	-12	-10	258	4	13.33	3.88	7.84	3.88	0.00	1605.33	2028.12	265.50	3.01	0.03	501	0	11.00
Partial							258	5	304.72	24.30	322.00	24.30	0.27	1605.29	2028.11	265.46	2.98	0.03	402	0	11.00
							258	6	1072.06	84.15	1101.98	84.15	0.51	1605.31	2028.10	265.48	2.95	0.03	302	0	12.00
							258	7	349.08	27.75	324.41	27.75	0.21	1605.33	202						1.00
Partial	-20	12	10	20	12	-10	259	46	1253.10	102.71	1381.90	102.71	0.99	1049.15	166						1.00
							259	47	7.35	27.23	28.40	27.23	0.01	1049.21	166						1.00

Three symmetry-related ***observations*** for one ***reflection***

AIMLESS

scale and merge

Merged file, one line for each hkl

h	k	l	IMEAN	SIGIMEAN	I(+)	SIGI(+)	I(-)	SIGI(-)
-20	12	10	1773.74	74.64	1633.04	179.11	1803.31	82.11

Optional unmerged output
Partials summed, scaled, outliers rejected

h	k	l	Orig. H	Orig. K	Orig. L	M/ISYM	BATCH	I	SIGI	SCALEUSED	SIGSCALEUSED	NPART	FRACTIONCALC	XDET	YDET	ROT	WIDTH	LP
-20	12	10	-20	-12	10	4	36	1890.60	142.80	1.14	0.00	1	2.26	2013.67	1406.74	295.54	0.41	0.17
-20	12	10	20	-12	-10	2	6	1760.21	100.35	1.01	0.00	4	0.99	1605.32	2028.10	265.47	2.97	0.03
-20	12	10	20	12	-10	3	46	1633.04	179.11	1.17	0.00	2	1.00	1049.18	1664.62	305.83	0.39	0.17

CTRUNCATE

Infer |F| from I

In ccp4i2, stored as I(+) and I(-)

Merged file, one line for each *hkl*, intensities and amplitudes F

h	k	l	F	SIGF	DANO	SIGDANO	F(+)	SIGF(+)	F(-)	SIGF(-)	ISYM	IMEAN	SIGIMEAN	I(+)	SIGI(+)	I(-)	SIGI(-)
-20	12	10	485.95	14.21	-24.77	28.43	473.57	26.06	498.34	11.35	0	1773.74	74.64	1633.04	179.11	1803.31	82.11

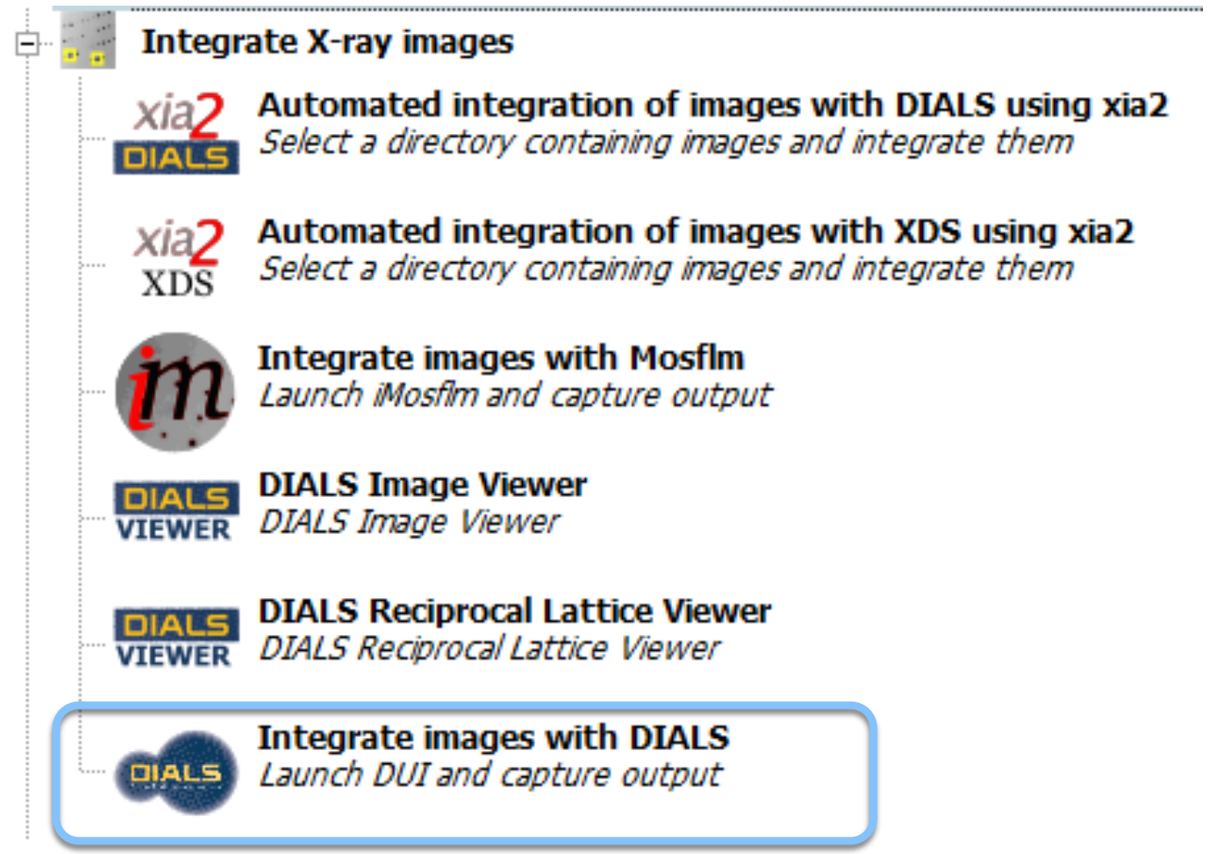
How to start from ccp4i2

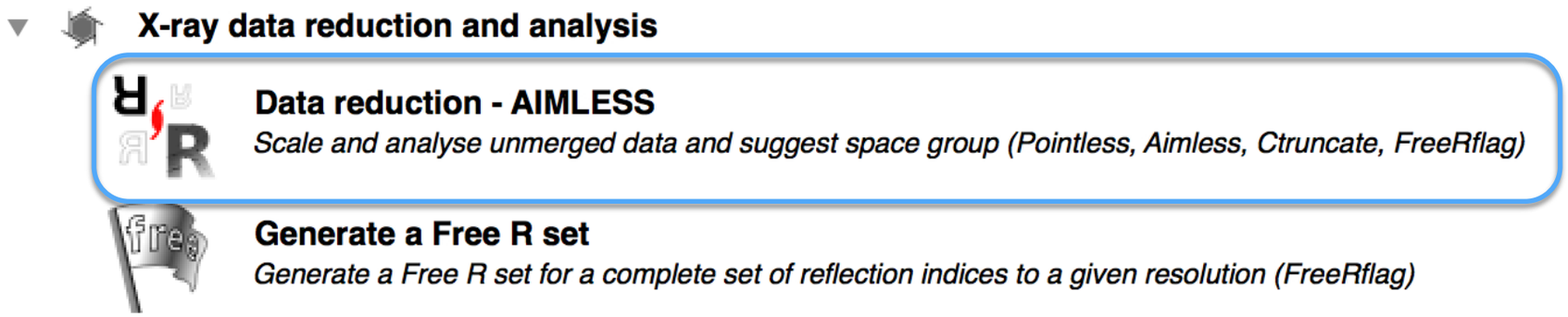
Run xia2 with DIALS or XDS

Start DUI (or iMosflm)

Follow-on to run data reduction

Just click “Run”

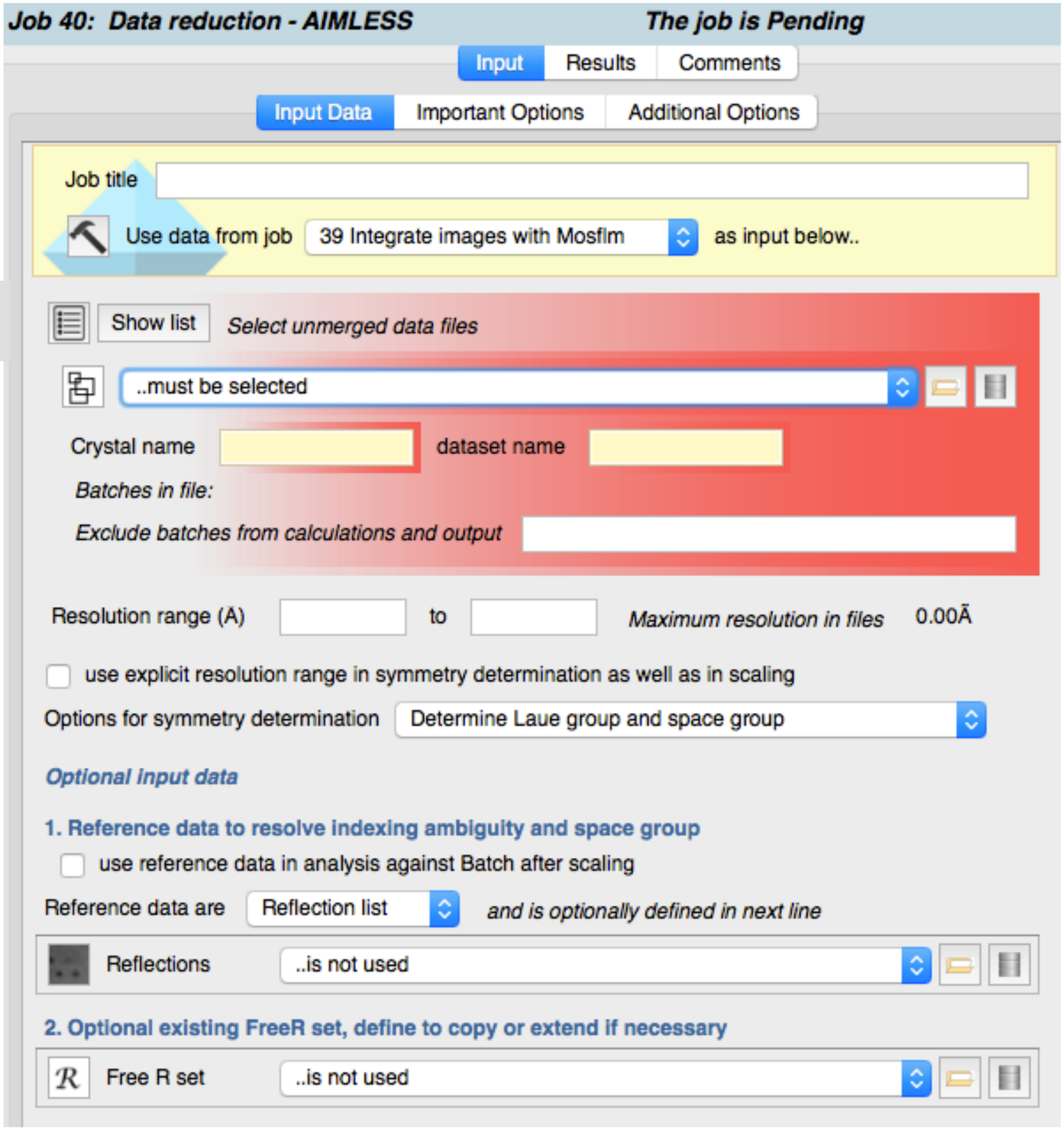




Data reduction task

Import one or more files

Identify dataset
(short names without spaces)



Symmetry determination, point group and space group (POINTLESS)

The crystal symmetry may impose constraints on the unit cell dimensions, according to the crystal class (the Bravais lattice): cubic, hexagonal/trigonal, tetragonal, orthorhombic, monoclinic, or triclinic, + lattice centring P, C, I, R, or F. For example, in the tetragonal system a=b, and all angles = 90°

Indexing in MOSFLM, XDS, DIALS, etc only gives a unit cell, which implies possible lattice symmetry, due to the constraints of unit cell dimensions. But to determine the point group we need to look at the intensities, as rotational and screw symmetry in real space leads to rotational symmetry in reciprocal space

Note that POINTLESS (and other programs) will find **symmetry** in the diffraction pattern, but this symmetry may or may not be *crystallographic* (rather than non-crystallographic pseudo symmetry)

Stages of space group determination in POINTLESS

1. from the cell dimensions, determine the maximum possible lattice symmetry, with some tolerance (ignoring any input symmetry)
2. for each possible rotation operator, score potentially related observations pairs for agreement (correlation coefficients and R-factor)
3. score all possible combinations of operators to determine the point group (point groups from the maximum down to P1)
4. score axial systematic absences to detect screw axes, hence space group (note that axial observations are sometimes unobserved)

Symmetry determination, point group and space group (POINTLESS)

Stage 1: score individual symmetry operators in the maximum lattice group

Maximum possible lattice symmetry determined from cell dimensions

pseudo-cubic example, a ≈ b ≈ c, angles ≈ 90°

Compare pairs of observations related by each possible rotational operator, using correlation coefficients and R-factors on normalised intensities $|E|^2$

```
Analysing rotational symmetry in lattice group P m -3 m
-------------------------------------------------------
Scores for each symmetry element
                                                    0.950
 Nelmt   Lklhd   Z-cc     CC        N   Rmeas     Symmetry & operator (in Lattice Cell)

   1    0.955    9.70    0.97    13557   0.073      identity
   2    0.062    2.66    0.27    12829   0.488      2-fold    ( 1 0 1)   {+l,-k,+h}
   3    0.065    2.85    0.29    10503   0.474      2-fold    ( 1 0-1)   {-l,-k,-h}
   4    0.056    0.06    0.01    16391   0.736      2-fold    ( 0 1-1)   {-h,-l,-k}
   5    0.057    0.05    0.00    17291   0.738      2-fold    ( 0 1 1)   {-h,+l,+k}
   6    0.049    0.55    0.06    13758   0.692      2-fold    ( 1-1 0)   {-k,-h,-l}
   7    0.950    9.59    0.96    12584   0.100  *** 2-fold k  ( 0 1 0)   {-h,+k,-l}
   8    0.049    0.57    0.06    11912   0.695      2-fold    ( 1 1 0)   {+k,+h,-l}
   9    0.948    9.57    0.96    16928   0.136  *** 2-fold h  ( 1 0 0)   {+h,-k,-l}
  10    0.944    9.50    0.95    12884   0.161  *** 2-fold l  ( 0 0 1)   {-h,-k,+l}
  11    0.054    0.15    0.01    23843   0.812      3-fold    ( 1 1 1)   {+l,+h,+k}  {+k,+l,+h}
  12    0.055    0.11    0.01    24859   0.825      3-fold    ( 1-1-1)   {-l,-h,+k}  {-k,+l,-h}
  13    0.055    0.14    0.01    22467   0.788      3-fold    ( 1-1 1)   {+l,-h,-k}  {-k,-l,+h}
  14    0.055    0.12    0.01    27122   0.817      3-fold    ( 1 1-1)   {-l,+h,-k}  {+k,-l,-h}
  15    0.061   -0.10   -0.01    25905   0.726      4-fold h  ( 1 0 0)   {+h,-l,+k}  {+h,+l,-k}
  16    0.060    2.53    0.25    23689   0.449      4-fold k  ( 0 1 0)   {+l,+k,-h}  {-l,+k,+h}
  17    0.049    0.56    0.06    25549   0.653      4-fold l  ( 0 0 1)   {-k,+h,+l}  {+k,-h,+l}
```

Only orthorhombic symmetry operators are present

High CC, low R_{meas}

What score to use?

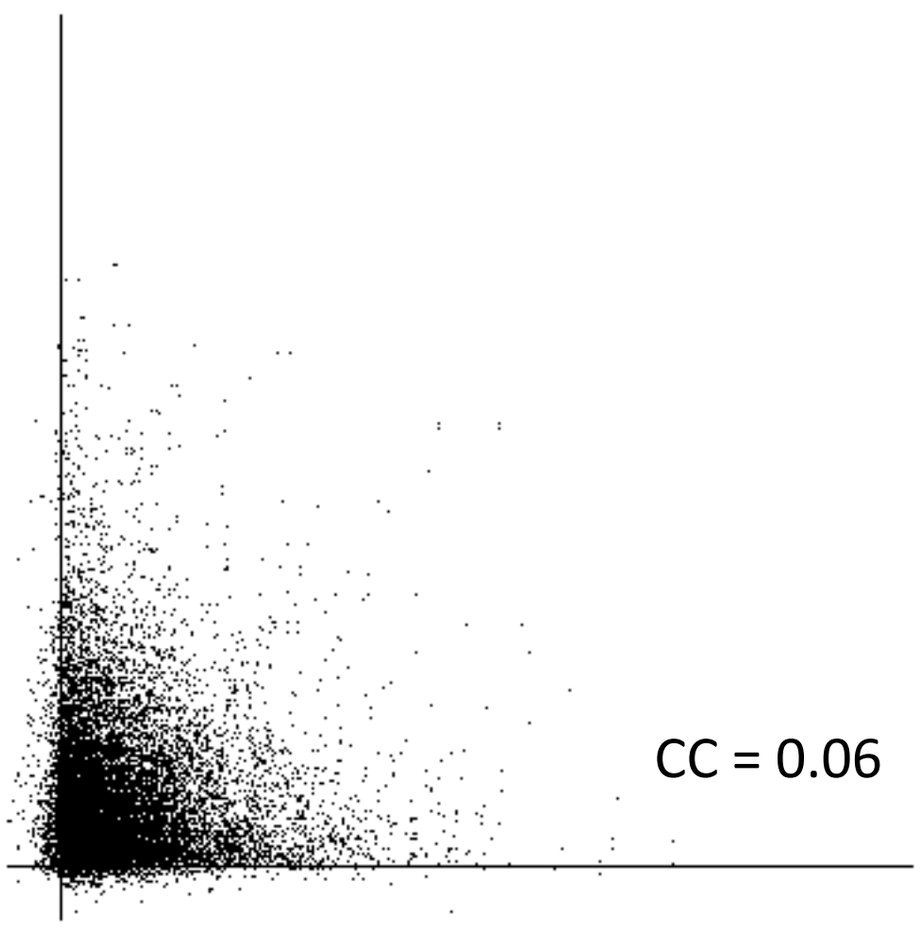
Linear correlation coefficient

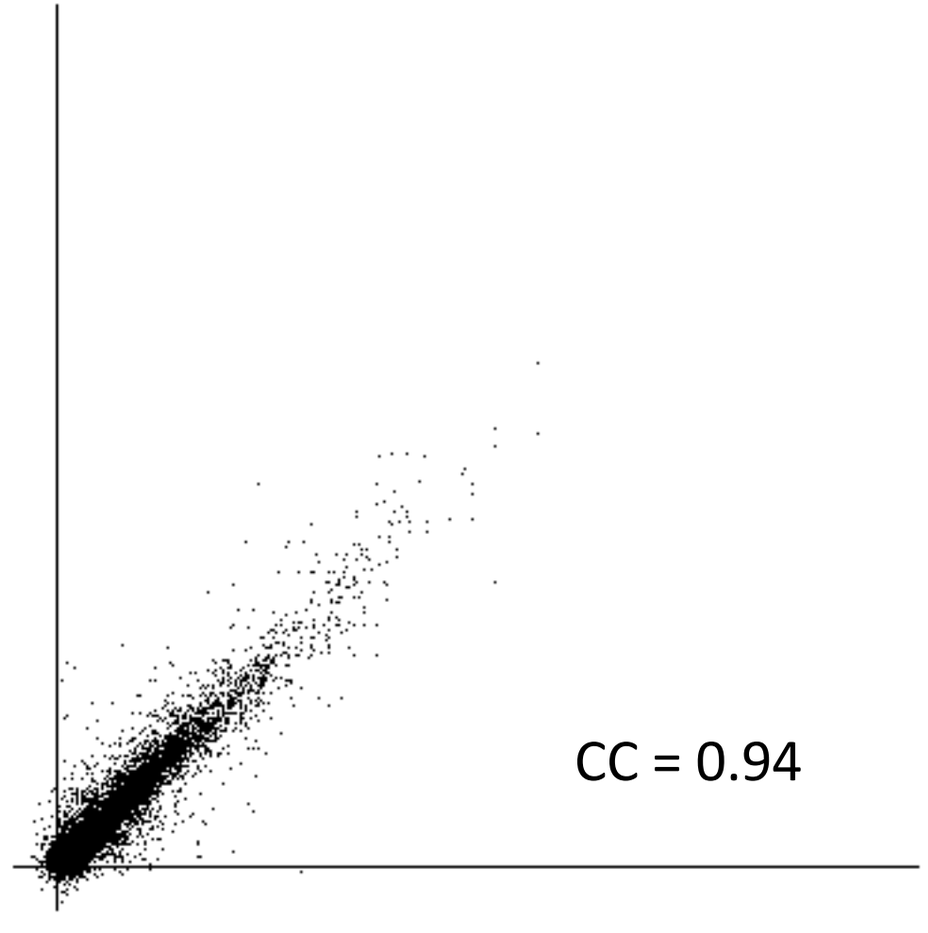
For equal axes, the correlation coefficient (CC) is the slope of the “best” (least-squares) straight line through the scatter plot

CCs have the advantage over eg R-factors in being relatively insensitive to incorrect scales

... but they are more sensitive to outliers

... and CCs need to correlate values that come from the same distribution, ie in this case $|E|^2$ rather than I





Stage 2: score possible point groups

All possible combinations of rotations are scored to determine the point group.

Good scores in symmetry operations which are absent in the sub-group count against that group.

Example: C-centred orthorhombic which might been hexagonal

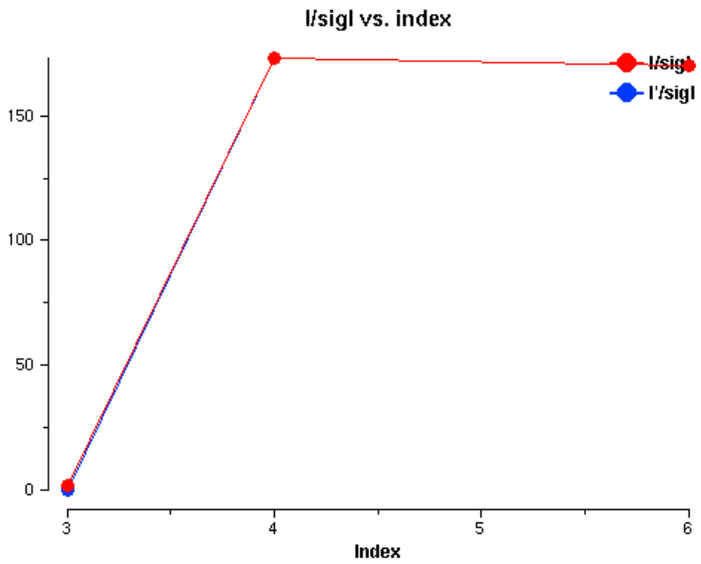
	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
= 1	C m m m	***	0.989	9.45	9.62	0.17	0.96	0.02	0.08	0.76	0.0	[h,k,l]
2	P 1 2/m 1		0.004	7.22	9.68	2.46	0.97	0.25	0.06	0.56	0.0	[-1/2h+1/2k,-l,-1/2h-1/2k]
3	C 1 2/m 1		0.003	7.11	9.61	2.50	0.96	0.25	0.08	0.55	0.0	[h,k,l]
4	C 1 2/m 1		0.003	7.11	9.61	2.50	0.96	0.25	0.08	0.55	0.0	[-k,-h,-l]
5	P -1		0.000	6.40	9.67	3.27	0.97	0.33	0.06	0.49	0.0	[1/2h+1/2k,1/2h-1/2k,-l]
6	C m m m		0.000	1.91	5.11	3.20	0.51	0.32	0.34	0.51	2.5	[1/2h-1/2k,-3/2h-1/2k,-l]
7	P 6/m		0.000	1.16	4.59	3.43	0.46	0.34	0.41	0.46	2.5	[-1/2h-1/2k,-1/2h+1/2k,-l]
8	C 1 2/m 1		0.000	1.51	5.15	3.64	0.52	0.36	0.33	0.47	2.5	[1/2h-1/2k,-3/2h-1/2k,-l]
9	C 1 2/m 1		0.000	1.51	5.15	3.64	0.51	0.36	0.33	0.47	2.5	[-3/2h-1/2k,-1/2h+1/2k,-l]
10	P -3		0.000	1.04	4.75	3.71	0.48	0.37	0.40	0.45	2.5	[-1/2h-1/2k,-1/2h+1/2k,-l]
11	C m m m		0.000	2.13	5.23	3.10	0.52	0.31	0.32	0.52	2.5	[-1/2h-1/2k,-3/2h+1/2k,-l]
12	C 1 2/m 1		0.000	1.64	5.25	3.61	0.53	0.36	0.32	0.47	2.5	[-1/2h-1/2k,-3/2h+1/2k,-l]
13	C 1 2/m 1		0.000	1.67	5.27	3.60	0.53	0.36	0.32	0.47	2.5	[-3/2h+1/2k,1/2h+1/2k,-l]
14	P -3 1 m		0.000	0.12	4.00	3.87	0.40	0.39	0.44	0.44	2.5	[-1/2h-1/2k,-1/2h+1/2k,-l]
15	P -3 m 1		0.000	0.14	4.00	3.86	0.40	0.39	0.44	0.44	2.5	[-1/2h-1/2k,-1/2h+1/2k,-l]
16	P 6/m m m		0.000	3.93	3.93	0.00	0.39	0.00	0.44	0.00	2.5	[-1/2h-1/2k,-1/2h+1/2k,-l]

Stage 3: space group from axial systematic absences

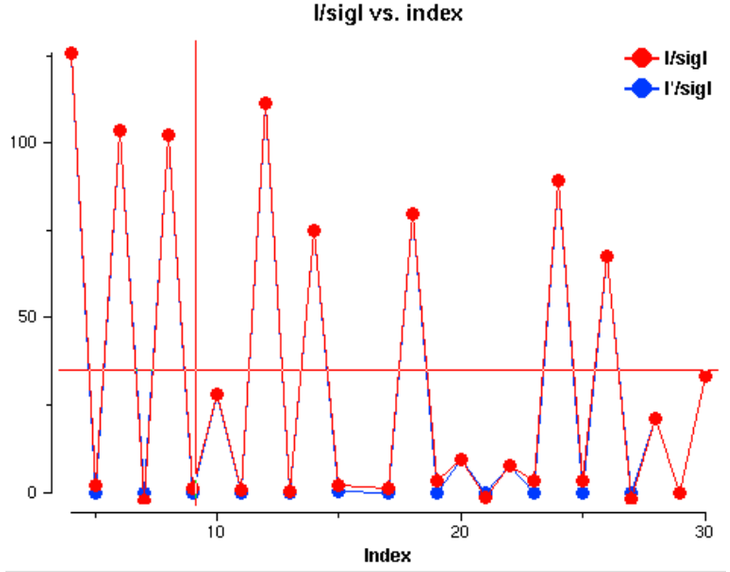
Zone	Number	PeakHeight	SD	Probability	ReflectionCondition
Zones for Laue group P m m m					
1 screw axis 2(1) [a]	3	1.000	0.296	** 0.889	h00: h=2n
2 screw axis 2(1) [b]	26	1.000	0.142	*** 0.971	0k0: k=2n
3 screw axis 2(1) [c]	46	0.997	0.097	*** 0.986	00l: l=2n

Fourier analysis of I/σ(I)

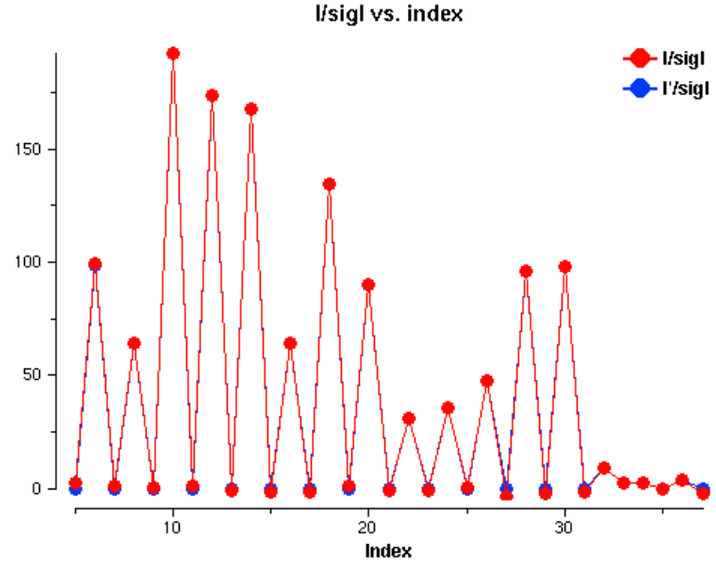
There are indications of 2_1 screw symmetry along all principle axes (though note there are only 3 observations on the *a* axis (h00 reflections))



Possible 2_1 axis along a



Clear 2_1 axis along b



Clear 2_1 axis along c

... BUT “confidence” in space group may be low due to sparse or missing information
Always check the space group later in the structure solution!

```
Possible spacegroups:
---------------------
Indistinguishable space groups are grouped together on successive lines

'Reindex' is the operator to convert from the input hklin frame to the standard spacegroup frame.

'TotProb' is a total probability estimate (unnormalised)

'SysAbsProb' is an estimate of the probability of the space group based on
the observed systematic absences.

'Conditions' are the reflection conditions (absences)


   Spacegroup          TotProb SysAbsProb     Reindex         Conditions
  <P 21 21 21> ( 19)     0.838  0.851                        h00: h=2n, 0k0: k=2n, 001: l=2n (zones 1,2,3)
   ..........
   <P 2 21 21> ( 18)     0.104  0.106                        0k0: k=2n, 001: l=2n (zones 2,3)
   ..........
   <P 21 2 21> ( 18)     0.025  0.026                        h00: h=2n, 001: l=2n (zones 1,3)
   ..........
   <P 21 21 2> ( 18)     0.012  0.012                        h00: h=2n, 0k0: k=2n (zones 1,2)
```

```
Best Solution     space group P 21 21 21

   Reindex operator:                        [h,k,l]
   Laue group probability:                  0.985
   Systematic absence probability:          0.851
   Total probability:                       0.838
   Space group confidence:                  0.784
   Laue group confidence                    0.982

   Unit cell:   34.16   54.8     68         90     90     90

   17.00 to   1.78   - Resolution range used for Laue group search

   17.00 to   1.78   - Resolution range in file, used for systematic absence check

   Number of batches in file:    100
```

Note high confidence in Laue group, but lower confidence in space group

What can go wrong?

Pseudo-symmetry or twinning (often connected) can suggest a point group symmetry which is too high. Careful examination of the scores for individual symmetry operators may indicate the truth (the program is not foolproof!)

POINTLESS works (usually) with unscaled data (hence use of correlation coefficients), so data with a large range of scales, including a dead crystal, may give a too-low symmetry.
In bad cases either just use the first part of the data, or scale in P1 and run POINTLESS on the scaled unmerged data

Potential axial systematic absences may be absent or few, so it may not be possible to determine the space group. In that case the output file is labelled with the “space group” with no screw axes, eg P2, P222, P622 etc, and the space group will have to be determined later

NOTE that the space group is only a **hypothesis** until the structure has been determined and satisfactorily refined

What can go wrong?

Pseudo symmetry example

Monoclinic, pseudo-orthorhombic (from NCS), β ≈ 90°

Unit cell 107.99 270.51 155.96 90.00 **90.36** 90.00

Nelmt	Lklhd	Z-cc	CC	N	Rmeas	Symmetry & operator (in Lattice Cell)
1	0.925	9.13	0.91	14115	0.126	identity
2	0.928	9.16	0.92	6811	0.176 ***	2-fold l (0 0 1) {-h,-k,l}, along original k
3	0.659	7.96	0.80	31850	0.252 *	2-fold k (0 1 0) {-h,k,-l}, along original l
4	0.678	8.02	0.80	6841	0.245 *	2-fold h (1 0 0) {h,-k,-l}, along original h

one 2-fold is stronger than the other two, but not enough to give the right answer

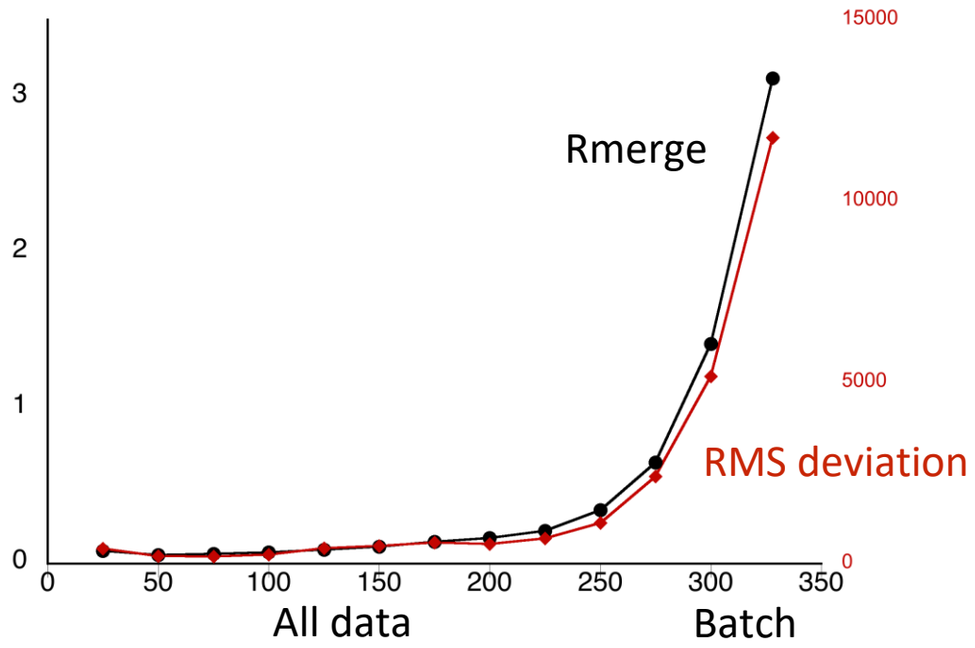
	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
> 1	P m m m	**	0.745	8.33	8.33	0.00	0.83	0.00	0.20	0.00	0.4	[h,l,-k]
= 2	P 1 2/m 1		0.183	1.20	9.14	7.94	0.91	0.79	0.14	0.25	0.0	[h,k,l]
3	P 1 2/m 1		0.030	0.59	8.75	8.16	0.88	0.82	0.16	0.24	0.4	[-l,-h,k]
4	P 1 2/m 1		0.028	-0.32	8.27	8.59	0.83	0.86	0.20	0.21	0.4	[h,l,-k]
5	P -1		0.014	1.01	9.13	8.12	0.91	0.81	0.13	0.24	0.0	[-h,-l,-k]

Best Solution: point group P 2 2 2

Reindex operator:	[h,l,-k]
Laue group probability:	0.745
Systematic absence probability:	0.832
Total probability:	0.620
Space group confidence:	0.000
Laue group confidence	0.647

Note low confidence in Laue (point) group

What can go wrong?

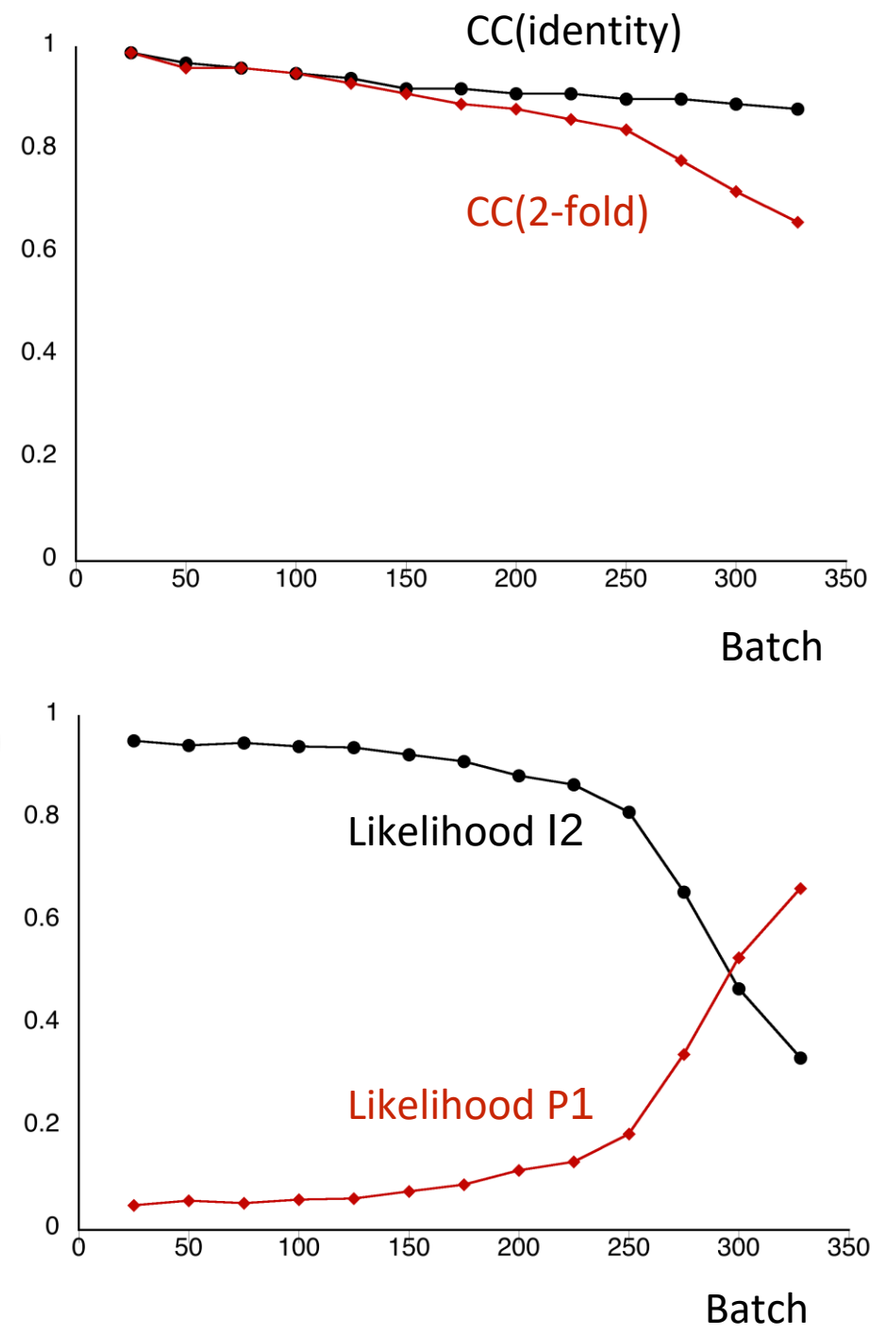


Severe radiation damage after image 250

True space group is I2 (== C2),
i.e. it has a crystallographic dyad
(2-fold rotation)

Radiation damage obscures the dyad, giving the wrong lower symmetry P1

Radiation damage example



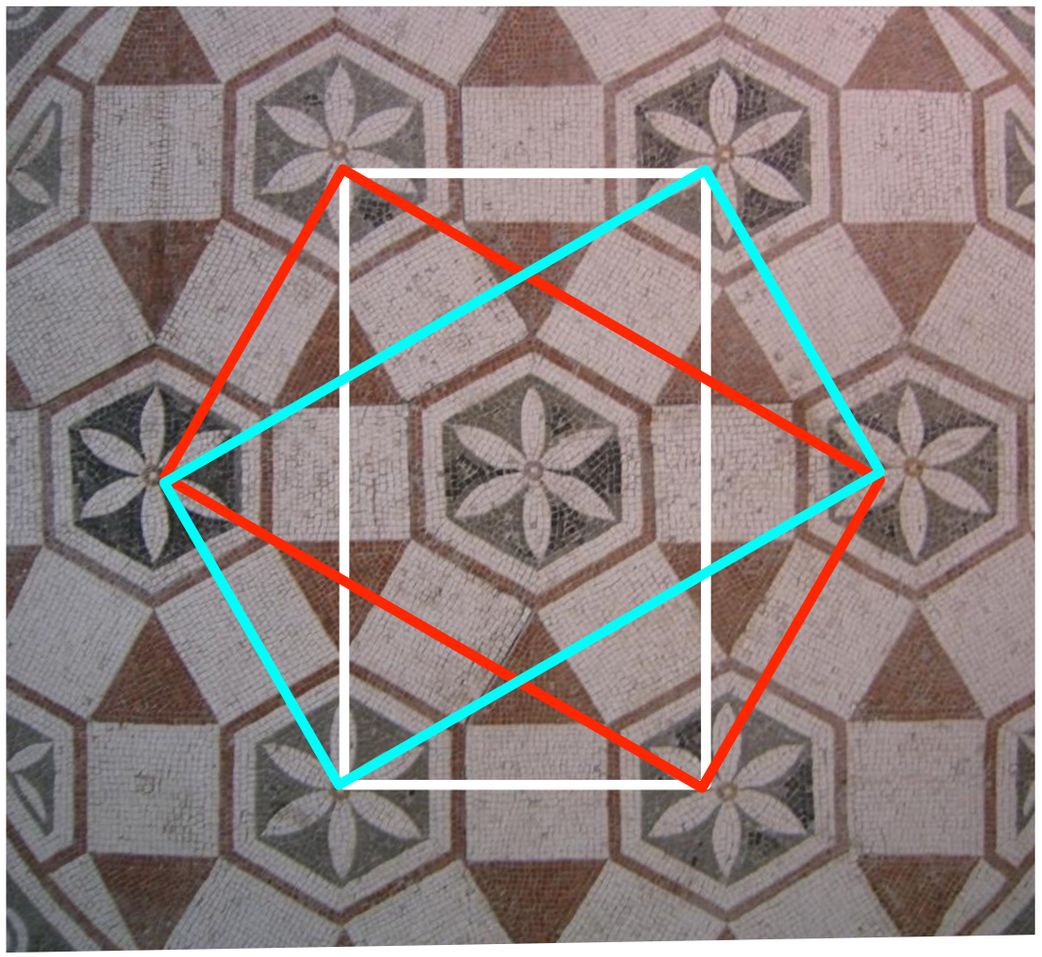
Scores from cumulative batch groups from the start, i.e. 1-25, 1-50, 1-75, … etc

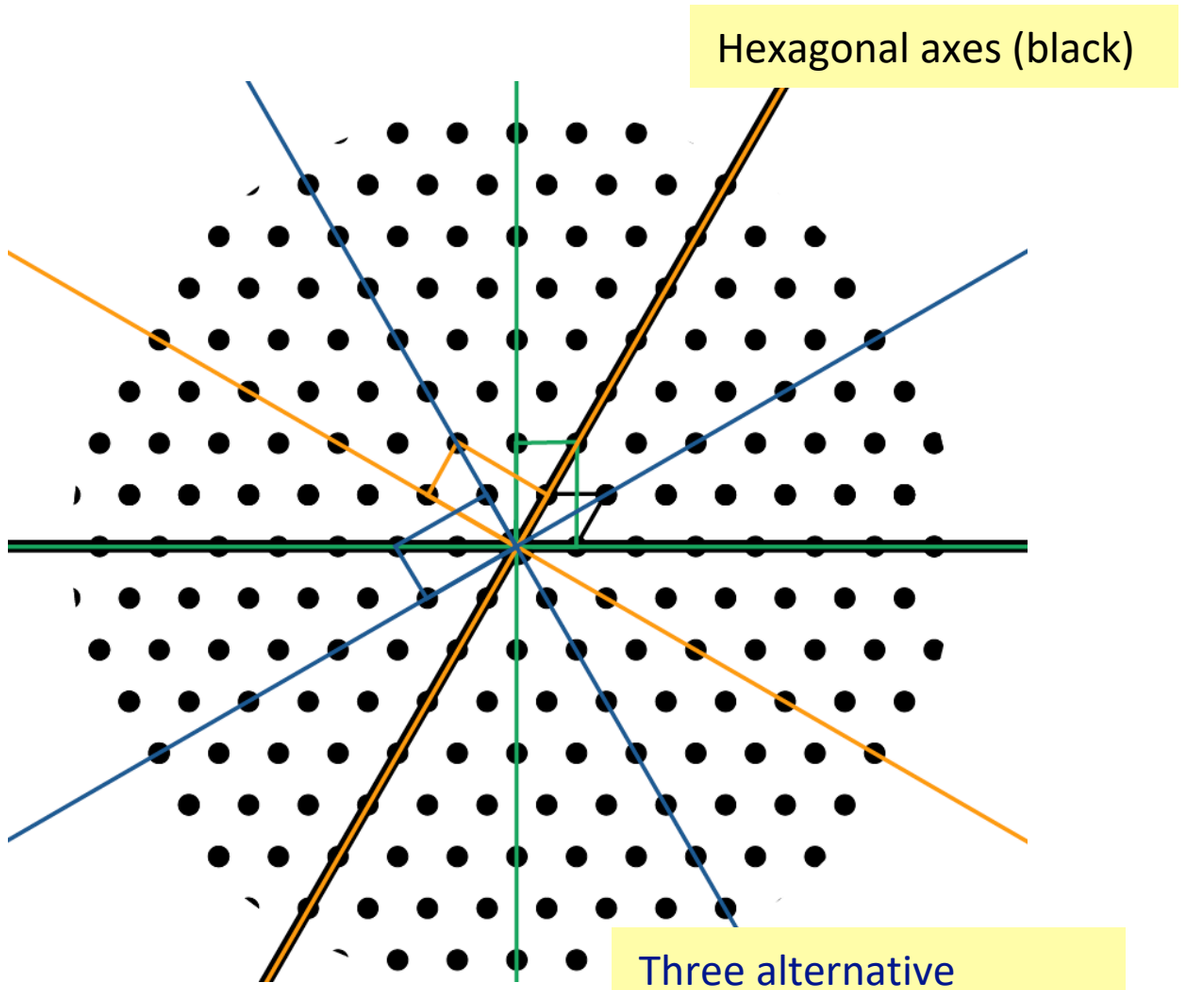
A confusing case in C222:

Unit cell 74.72 129.22 184.25 90 90 90

This has b ≈ √3a so can also be indexed on a hexagonal lattice,
lattice point group P622 (P6/mmm), with the reindex operator: h/2+k/2, h/2-k/2, -l

Conversely, a hexagonal lattice may be indexed as C222 in three distinct ways, so there is a 2 in 3 chance of the indexing program choosing the wrong one





Score each symmetry operator in P622

"Likelihood"

Z-score(CC)

Correlation coefficient on E^2

Rfactor (multiplicity weighted)

Nelmt	Lklhd	Z-cc	CC	N	Rmeas		Symmetry & operator (in Lattice Cell)		
1	0.808	5.94	0.89	9313	0.115		identity		
2	0.828	6.05	0.91	14088	0.141	***	2-fold l (0 0 1)	{-h,-k,+l}	
3	0.000	0.06	0.01	16864	0.527		2-fold (1-1 0)	{-k,-h,-l}	
4	0.871	6.33	0.95	10418	0.100	***	2-fold (2-1 0)	{+h,-h-k,-l}	
5	0.000	0.53	0.08	12639	0.559		2-fold h (1 0 0)	{+h+k,-k,-l}	
6	0.000	0.06	0.01	16015	0.562		2-fold (1 1 0)	{+k,+h,-l}	
7	0.870	6.32	0.95	2187	0.087	***	2-fold k (0 1 0)	{-h,+h+k,-l}	
8	0.000	0.55	0.08	7552	0.540		2-fold (-1 2 0)	{-h-k,+k,-l}	
9	0.000	-0.12	-0.02	11978	0.598		3-fold l (0 0 1)	{-h-k,+h,+l}	{+k,-h-k,+l}
10	0.000	-0.06	-0.01	17036	0.582		6-fold l (0 0 1)	{-k,+h+k,+l}	{+h+k,-h,+l}

Only the orthorhombic symmetry operators are present

Alternative indexing

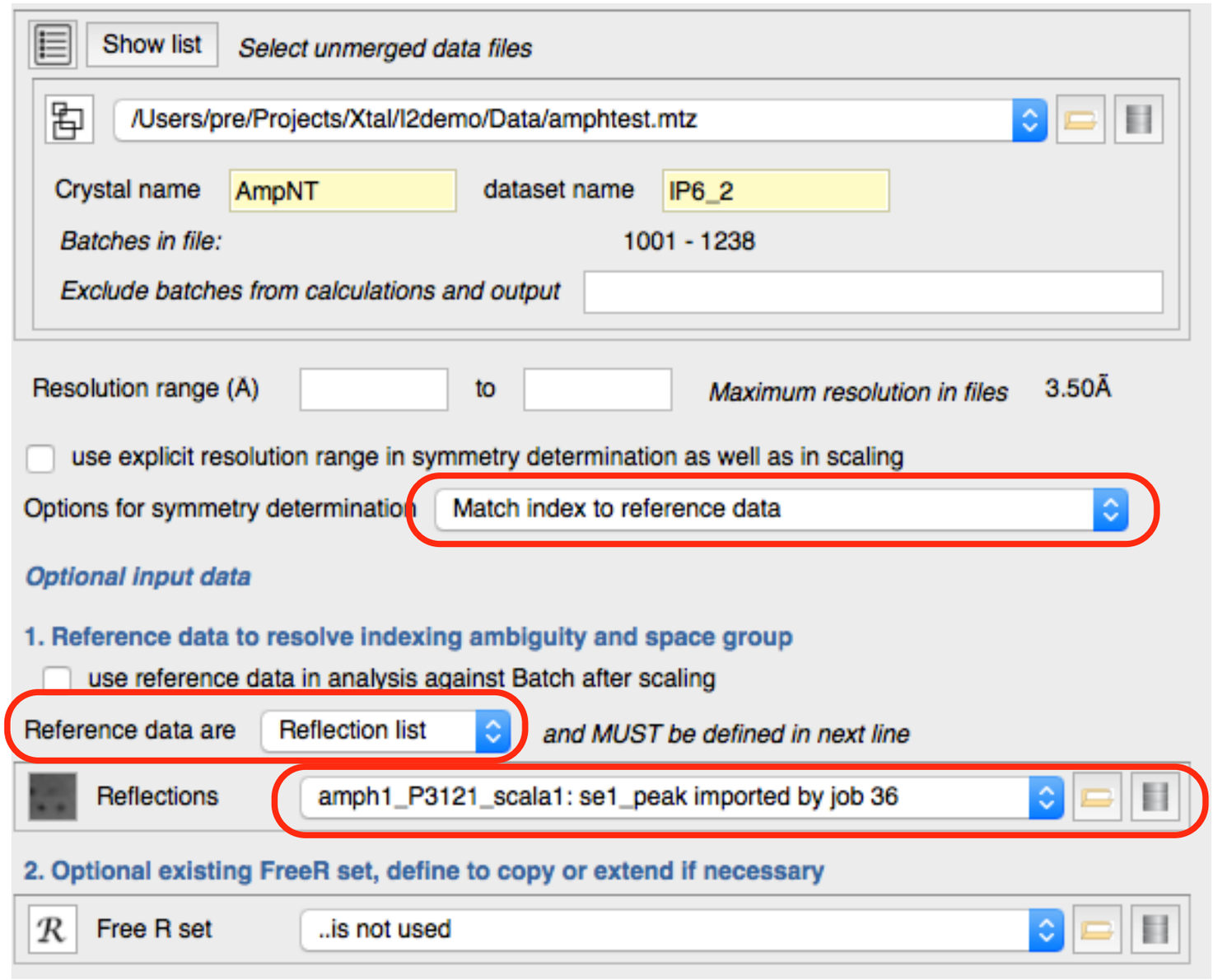
If the true point group is lower symmetry than the lattice group, alternative valid but non-equivalent indexing schemes are possible, related by symmetry operators present in lattice group but not in point group *(note that these are also the cases where merohedral twinning is possible)*

eg if in space group P3 (or $P3_1$) there are 4 different schemes
(h,k,l) or (-h,-k,l) or (k,h,-l) or (-k,-h,-l)

For the first crystal, you can choose any scheme

For subsequent crystals, the autoindexing will randomly choose one setting, and we need to make it consistent: *POINTLESS* will do this for you by comparing the unmerged test data to a reference dataset (merged or unmerged, or coordinates)

Note that the space group from the reference will be assumed to be correct



Combining multiple files

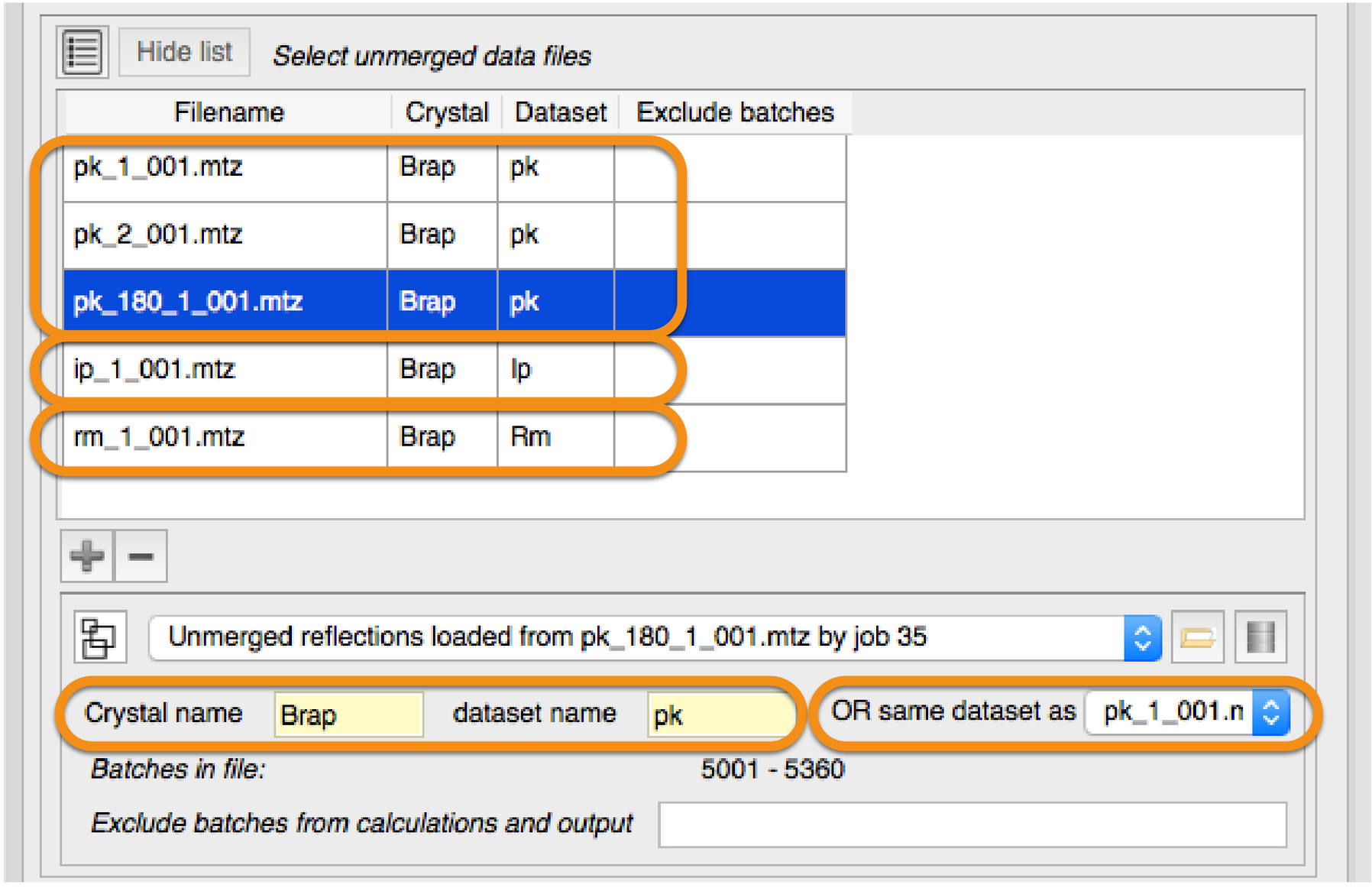
Multiple “sweeps” or datasets (eg MAD)

Peak, 3 files

Inflection, 1 file

Remote, 1 file

Use the dataset names



or assign files to the same dataset

▼ Alternative index scores

Possible reindex operators: [h,k,l], [-k,h,l], [l,k,-h], [-h,l,k], [l,h,k], [k,l,h]

Reindex operator	Likelihood	CC
[h,k,l]	0.664	0.64
[-k,h,l]	0.236	0.50
[l,k,-h]	0.029	0.17
[-h,l,k]	0.024	0.11
[l,h,k]	0.024	0.11
[k,l,h]	0.023	0.11

Reindex operator	Likelihood	CC
[h,k,l]	0.755	0.65
[-k,h,l]	0.153	0.45
[l,k,-h]	0.026	0.12
[-h,l,k]	0.022	0.07
[l,h,k]	0.022	0.07
[k,l,h]	0.021	0.06

Reindex operator	Likelihood	CC
[h,k,l]	0.873	0.74
[-k,h,l]	0.067	0.43
[l,k,-h]	0.018	0.13
[-h,l,k]	0.015	0.07
[k,l,h]	0.014	0.05
[l,h,k]	0.014	0.04

Reindex operator	Likelihood	CC
[h,k,l]	0.757	0.65
[-k,h,l]	0.153	0.45
[l,k,-h]	0.026	0.12
[l,h,k]	0.022	0.06
[-h,l,k]	0.021	0.06
[k,l,h]	0.021	0.06

Because of an indexing ambiguity (pseudo-cubic orthorhombic), we must check for consistent indexing between files

Note also some ambiguity with the operator [-k,h,l] due to pseudo-merohedral twinning

Scaling, merging and Data Quality

Put observations on a common scale

Analyse to:-

- estimate resolution
- check for radiation damage
- reject outliers
- improve error estimates

Why are reflections on different scales?

(a) Factors related to incident beam and the camera
incident beam intensity; illuminated volume; primary beam absorption
(b) Factors related to the crystal and the diffracted beam
absorption; radiation damage (worse at high resolution)
(c) Factors related to the detector
miscalibration; corners of fibre-optic tapers for CCDs
Beam-stop shadow etc (Important)

Scaling tries to make symmetry-related and duplicate measurements of a reflection equal, by modelling the diffraction experiment, principally as a function of the incident and diffracted beam directions in the crystal. This makes the data internally consistent (not necessarily correct)

$$\text{Minimize } \Phi = \Sigma_{hl}\, w_{hl}\, (I_{hl} - g_{hl}\langle I_h\rangle)^2$$

I_{hl} l'th intensity observation of reflection h k_{hl} scale factor for I_{hl}
$\langle I_h\rangle$ current estimate of I_h

$g_{hl} = 1/k_{hl}$ is a function of the parameters of the **scaling model**

g_{hl} = g(ϕ rotation/image number) . g(time) . g(s) … other factors
Primary beam s_0 *B-factor* *Absorption*

The scale model should reflect the data collection strategy

Data collection strategy should be designed to get good scaling and analysis

high multiplicity (low dose) gives:-

- good scaling
- good outlier rejection
- the opportunity to reject radiation damaged parts of the data without losing completeness

For example, in the extreme case of serial crystallography, with small rotation (or zero) range per crystal and many crystals, use one scale & B-factor / crystal

Average radiation damage
(scales up high resolution observations)

g_{hl} = g(ϕ rotation/image number) . g(time) . g(s2) … other factors

Primary beam s_0 *B-factor* *Absorption*

Illuminated volume etc

Important with big crystals at long wavelength

$\exp(-2B(\sin\theta/\lambda)^2)$

Factors related to incident Xray beam

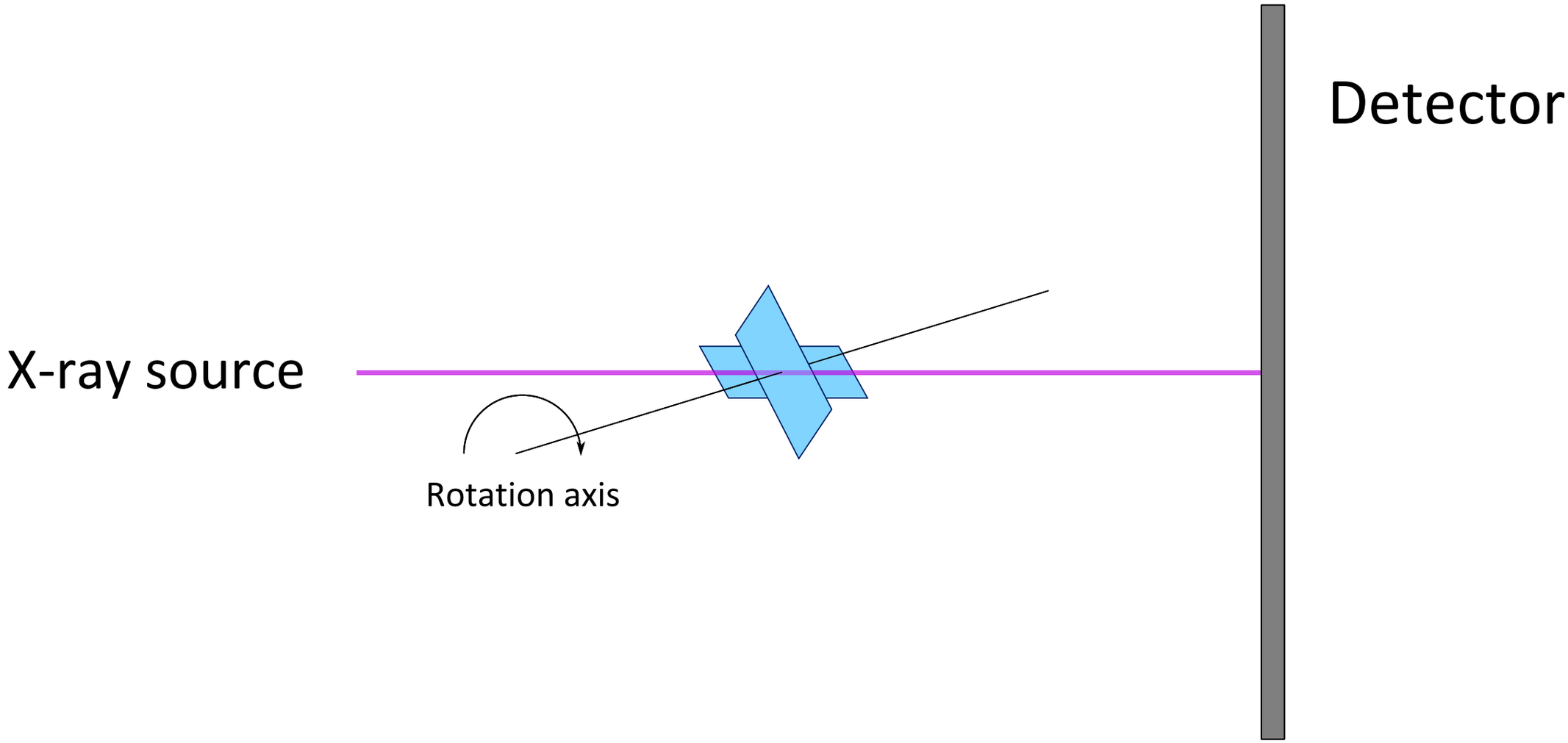
(a) incident beam intensity: variable on synchrotrons and not normally measured. Assumed to be constant during a single image, or at least varying smoothly and slowly (relative to exposure time). If this is not true, the data will be poor

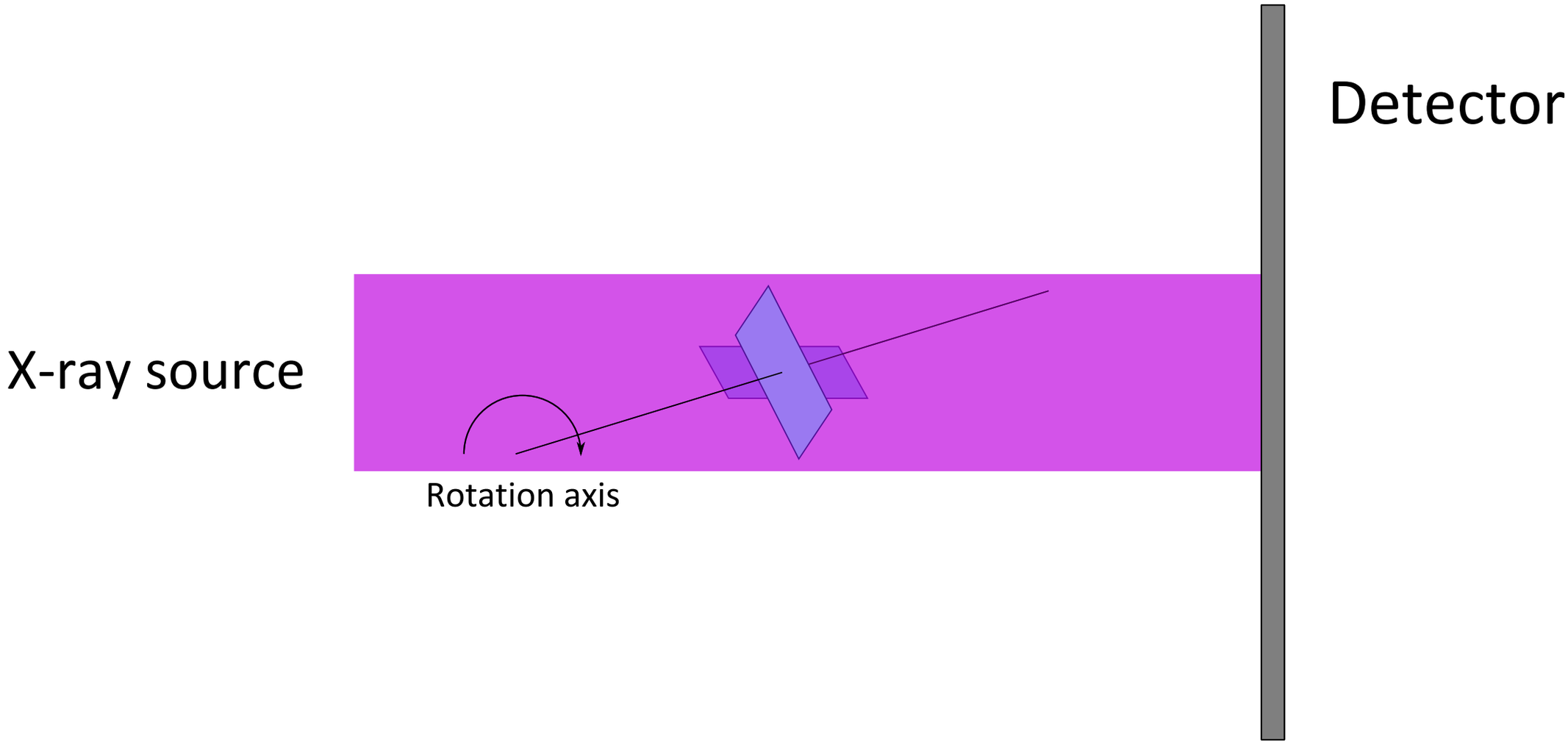
(b) illuminated volume: changes with ϕ if beam smaller than crystal

(c) absorption in primary beam by crystal: indistinguishable from (b)

(d) variations in rotation speed and shutter synchronisation. These errors are disastrous, difficult to detect, and (almost) impossible to correct for: we **assume** that the crystal rotation rate is constant and that adjacent images exactly abut in ϕ. *(Shutter synchronisation errors lead to partial bias which may be **positive**, unlike the usual negative bias)*

Data collection with open shutter (eg with Pilatus or Eiger detector) avoids synchronisation errors


Detector
X-ray source
Rotation axis

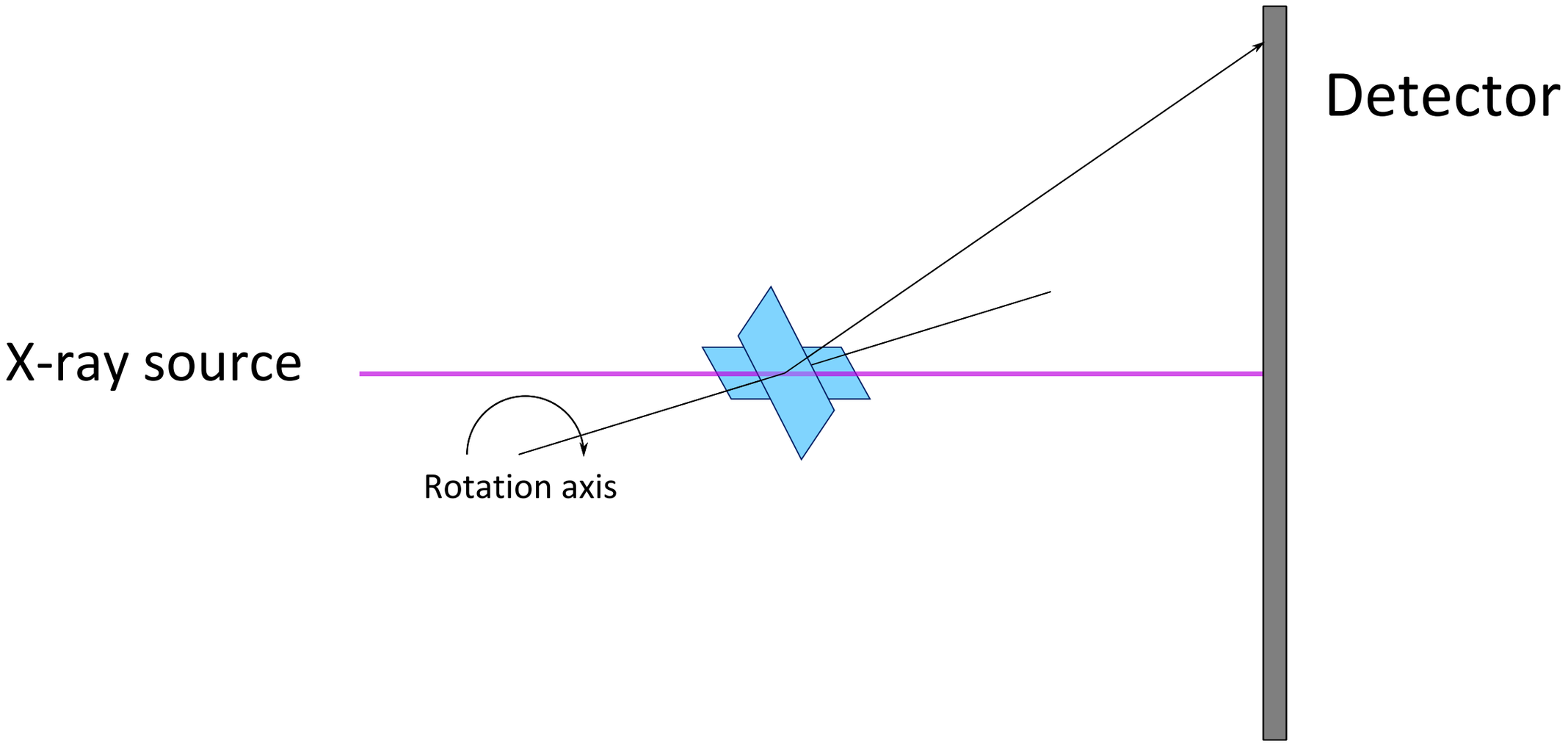

Detector
X-ray source
Rotation axis

Factors related to crystal and diffracted beam

(e) Absorption in secondary beam - serious at long wavelength (including CuKα)

(f) radiation damage - serious. Not easily correctable unless small as the structure is changing

The relative B-factor is largely a correction for the average radiation damage


Detector
X-ray source
Rotation axis

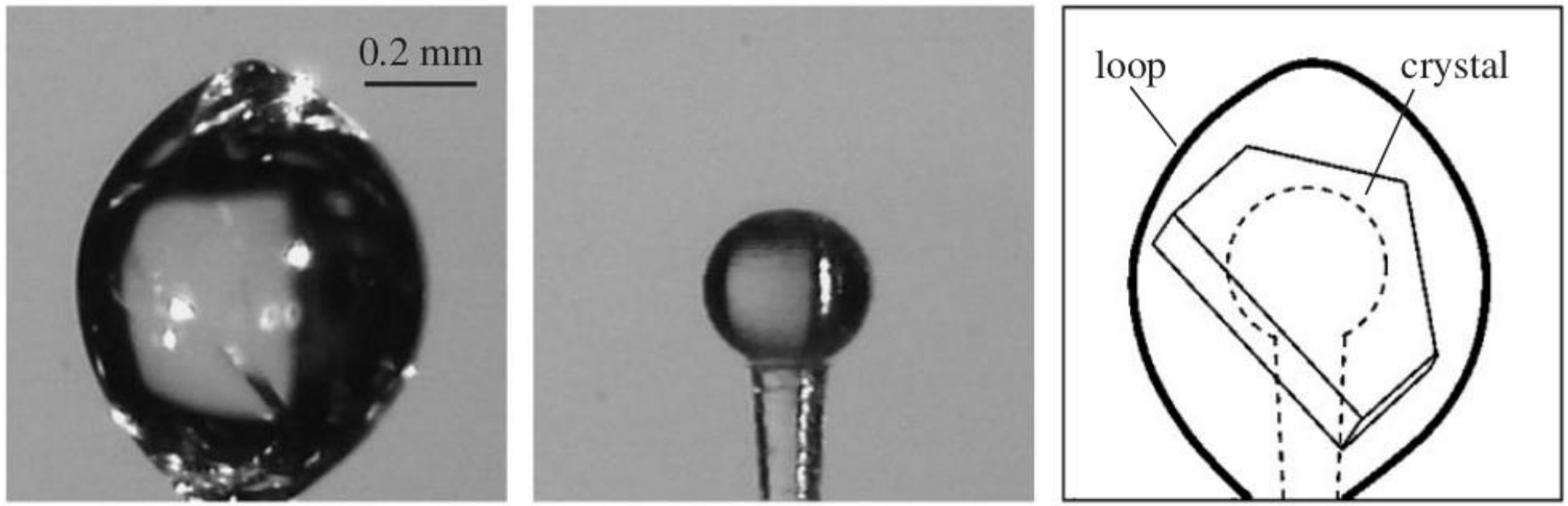


Fig. 2. A protein crystal ball. The HEWL crystal was modified into a spherical shape by laser irradiation. (a) The loop-mounted crystal before laser irradiation. The crystal was flash-cooled after immersion in a cryoprotectant. (b) The laser-processed crystal. A diameter of the spherical part was 0.3 mm. (c) Corresponding illustration of the photographs. The dashed line indicates a contour of the sample after laser irradiation.

H. Kitano et al. Jpn. J. Appl. Phys., **44,** 2

Factors related to the detector

- The detector should be properly calibrated for spatial distortion and sensitivity of response, and should be stable. Problems with this are difficult to detect from diffraction data. There are known problems in the tile corners of CCD detectors (corrected for in XDS)

- The useful area of the detector should be calibrated or told to the integration program

  – Calibration should flag defective pixels (hot or cold) and dead regions eg between tiles

  – The user should tell the integration program about shadows from the beamstop, beamstop support or cryocooler (define bad areas by circles, rectangles, arcs etc)

Viewing the output statistics (job report from ccp4i2)

1. key summary

▼ Key summary

Selecting space group P 21 21 21
as there is a single space group with the highest score

Solution probability: 0.853, Confidence 0.835 (high resolution limit for symmetry testing 3.252A)

NOTE: the final selected symmetry and cell have alternative indexing schemes, but no reference data has been given
Possible alternative indexing operators (with cell differences in A): [h,k,l] (0.00), [-h,l,k] (0.71), [-k,h,l] (0.73), [l,h,k] (1.25), [k,l,h] (1.25), [l,k,-h] (1.44)
If you already have a matching dataset, you should choose it as a reference set to get consistent indexing

Key statistics for Dataset: I2demo/Brap/pk

Resolution of input data: 2.79A, resolution estimate 2.87A
Rmeas: overall 0.151, inner bin 0.069
In outer bin: Mean(I/sdI) 0.8 CC(1/2) 0.266
Anomalous CC(1/2) in inner bin 0.753
Significant anomalous signal extends to a resolution of 3.74A (above CCanom threshold 0.15)

Warnings:
red, bad;
orange, maybe OK;
green, OK

Warning: Possible twinning, twin fraction estimates from Britton plot 0.20, from H-test 0.23

No evidence of possible translational non-crystallographic symmetry

Some anisotropy detected. This may have an effect on statistics.

Warning: Completeness test shows some issues.

No ice rings found.

Viewing the output statistics (job report)

"Table 1"

2. main summary

Space group determination

scores for individual symmetry elements may detect pseudo-symmetry …

… or suggest twinning

Space group determination

WARNING: the L-test suggests that the data may be twinned, so the indicated Laue symmetry may be too high
Rough estimated twin fraction: 0.096

Selecting space group P 21 21 21
as there is a single space group with the highest score

Solution type: space group

Group name	P 21 21 21
Reindex	[h,k,l]
Space group confidence	0.835
Laue group confidence	0.938
Laue group probability	0.948
Systematic absence probability	0.900

Scores for each symmetry element

Lattice group name P 4 3 2

Reindex operator from input to lattice: [h,k,l]

Likelihood	CC	R		Symmetry
0.913	0.88	0.097		identity
0.901	0.87	0.109	***	2-fold l (0 0 1) {-h,-k,l}
0.917	0.88	0.090	***	2-fold k (0 1 0) {-h,k,-l}
0.913	0.88	0.103	***	2-fold h (1 0 0) {h,-k,-l}
0.214	0.56	0.191		2-fold (1-1 0) {-k,-h,-l}
0.051	0.06	0.588		2-fold (0 1-1) {-h,-l,-k}
0.052	0.13	0.494		2-fold (1 0-1) {-l,-k,-h}
0.223	0.57	0.193		2-fold (1 1 0) {k,h,-l}
0.051	0.11	0.512		2-fold (1 0 1) {l,-k,h}
0.051	0.06	0.562		2-fold (0 1 1) {-h,l,k}
0.053	0.04	0.720		3-fold (1-1-1) {-k,l,-h}
0.054	0.04	0.712		3-fold (1 1-1) {-l,h,-k}
0.053	0.04	0.701		3-fold (1-1 1) {l,-h,-k}
0.053	0.04	0.704		3-fold (1 1 1) {k,l,h}
0.203	0.55	0.198		4-fold l (0 0 1) {-k,h,l}
0.050	0.10	0.509		4-fold k (0 1 0) {l,k,-h}
0.052	0.05	0.565		4-fold h (1 0 0) {h,l,-k}

Data internal consistency statistics

Summary of merging statistics for dataset
I2demo/Brap/pk

	Overall	Inner	Outer
Low resolution limit	57.82	57.82	2.94
High resolution limit	2.79	8.81	2.79
Rmerge(within I+/I-)	0.140	0.064	2.085
Rmerge(all I+ and I-)	0.154	0.076	2.218
Rmeas (within I+/I-)	0.151	0.069	2.336
Rmeas (all I+ & I-)	0.161	0.080	2.352
Rpim (within I+/I-)	0.057	0.026	1.013
Rpim (all I+ & I-)	0.045	0.023	0.754
Rmerge in top intensity bin	0.063		
Number of observations	178940	5691	18047
Number unique	14045	501	1955
Mean((I)/sd(I))	9.7	33.5	0.8
Half-set correlation CC(1/2)	0.998	0.997	0.266
Completeness %	99.5	97.6	96.9
Multiplicity	12.7	11.4	9.2
Anomalous completeness %	98.5	98.1	90.8
Anomalous multiplicity	6.4	6.9	4.7
DelAnom CC(1/2)	0.664	0.753	-0.026
Mid-Slope of Anom Probability	1.020		

Download

Download as CSV file

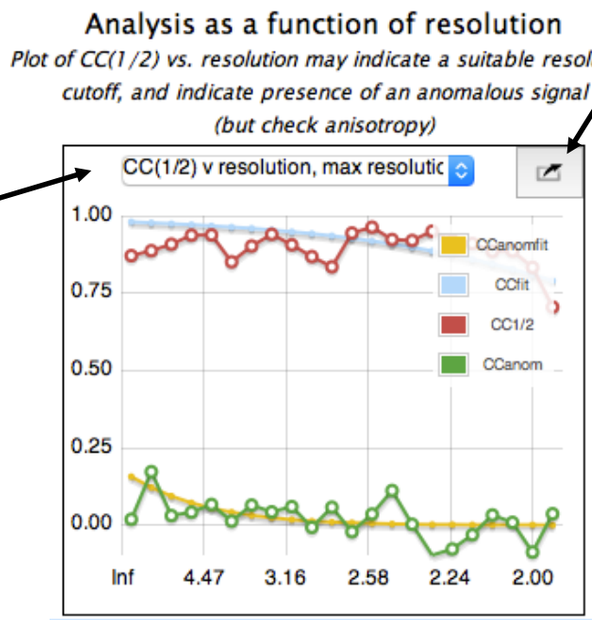
…rs to be a significant anomalous signal so
…nomalous flag was switched ON

Viewing the output statistics (job report)

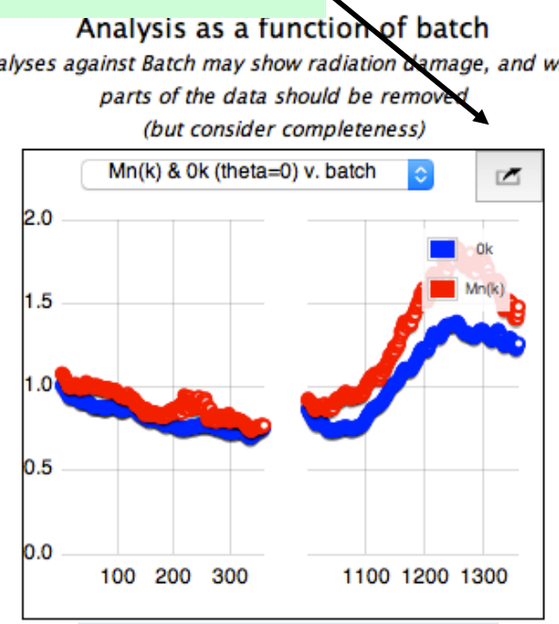
3. The most important graphs

pull-down to change graph

pop out separate graph viewer



Analyses by resolution



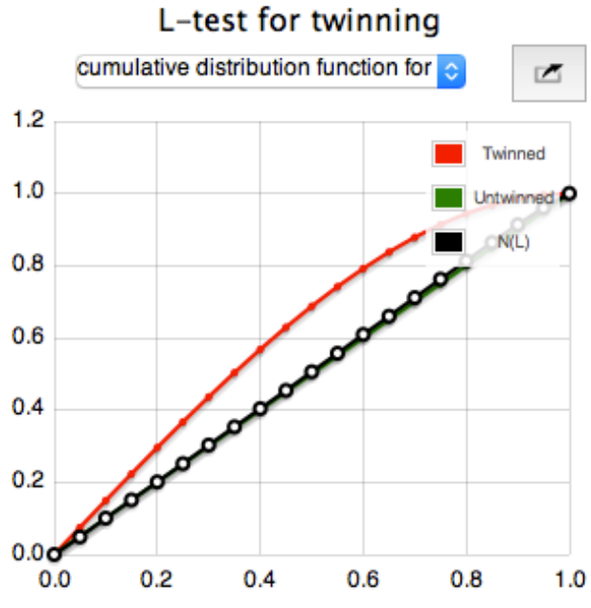
Analyses by batch

ing etc, more

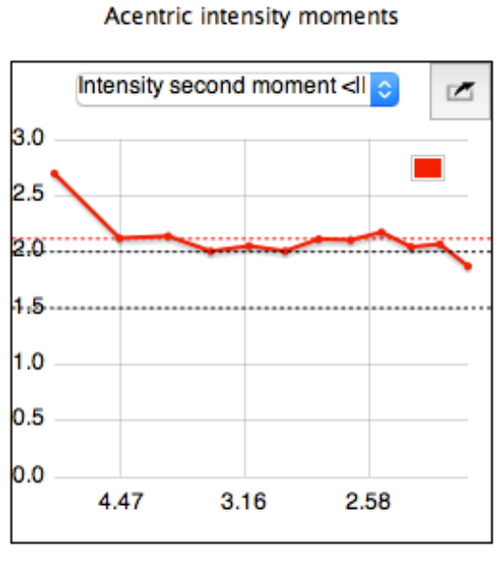
4. more details in folders, closed by default

- ▶ Details of space group determination
- ▶ Other merging graphs
- ▶ Details of merging
- ▶ Intensity statistics: twinning tNCS etc

This dataset is probably NOT twinned



Analyses for twinning

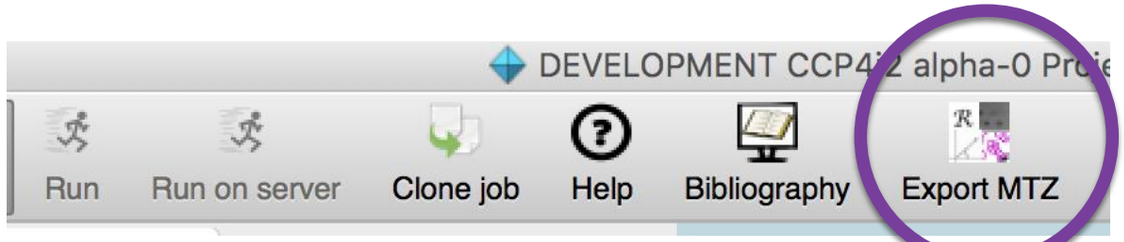


Values for these data, and for ideal data (untwinned or twinned)

Operator	Value	Untwinned	Perfect twin
<I^2>/<I>^2	2.117	2.000	1.500
<I^3>/<I>^3	7.082	6.000	3.000
<I^4>/<I>^4	32.680	24.000	7.500

Export of processed data from I2 for e.g. I1

Easiest way is to choose ExportMTZ from data reduction task



What should you look at? What are the questions?

Are there some parts of the data which much worse than the best parts? Maybe these should be omitted (subject to completeness)

Should you apply a resolution cutoff?

Measures of quality:

Signal/noise estimates $\langle I/\sigma(I) \rangle$ *note* $\neq \langle I \rangle / \langle \sigma(I) \rangle$

but σ(I) estimates are not perfect

Measures of internal consistency:

(1) R-factors

$R_{merge} = \Sigma \mid I_{hl} - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$ a.k.a R_{sym} or R_{int}

traditional overall measures of quality, but increases with multiplicity although the data improves

$R_{meas} = R_{r.i.m.} = \Sigma \sqrt{(n/n-1)} \mid I_{hl} - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$

multiplicity-weighted, better (but larger)

$R_{p.i.m.} = \Sigma \sqrt{(1/n-1)} \mid I_{hl} - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$

"Precision-indicating R-factor" gets better (smaller) with increasing multiplicity, ie it estimates the precision of the merged <I>

(2) *correlation coefficients*

Half-dataset correlation coefficient $CC_{1/2}$:

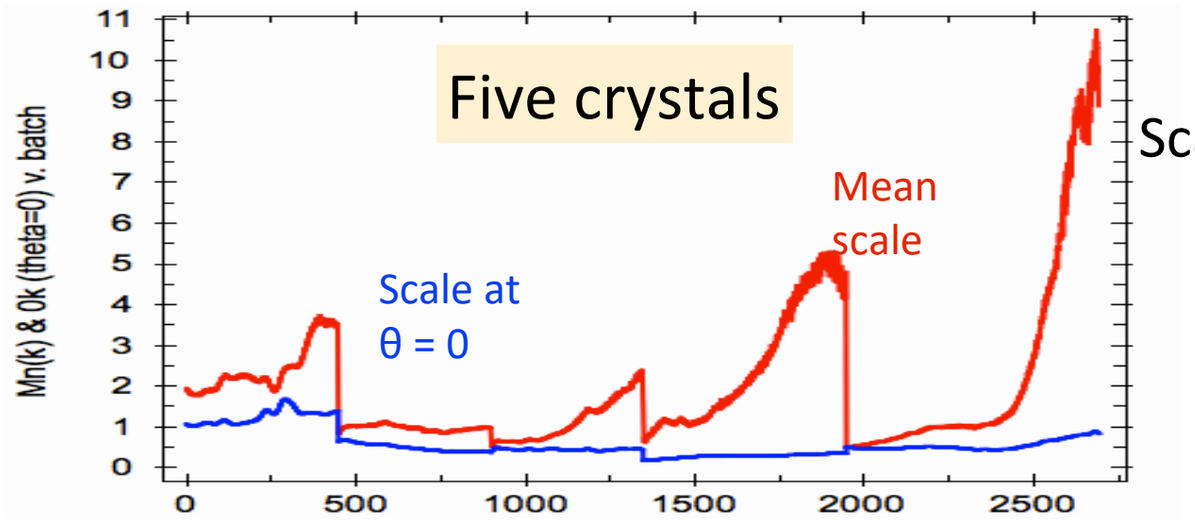
Split observations for each reflection data randomly into 2 halves, and calculate the correlation coefficient between them (essentially comparing the dispersion of individual observations with the dispersion of the data)

What should you look at?

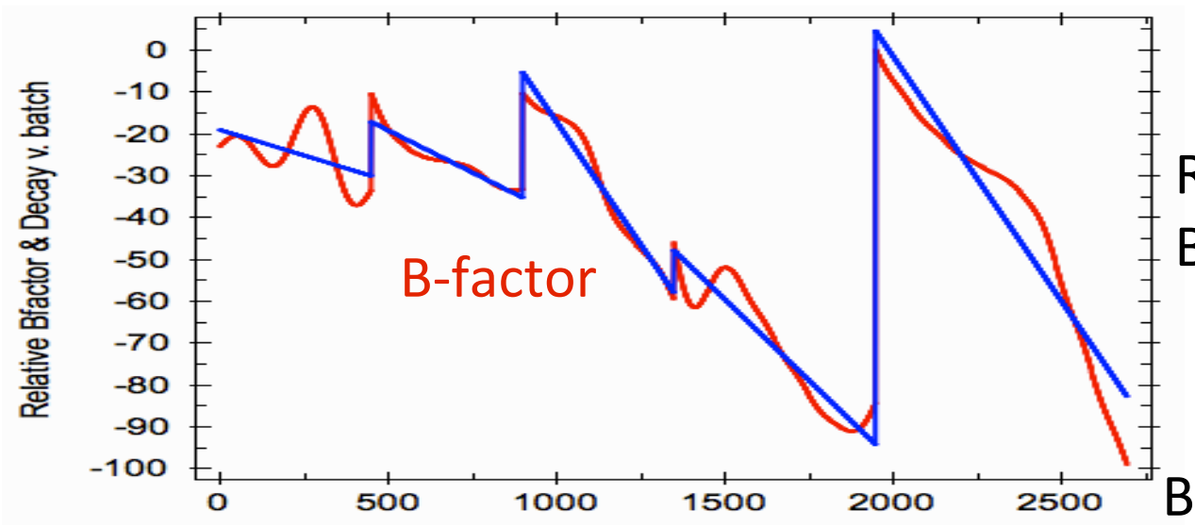
Analyses as a function of “batch” (ie image number)

Look at :

- scales
- relative B-factor (overall radiation damage)
- cumulative completeness
- maybe comparison to reference

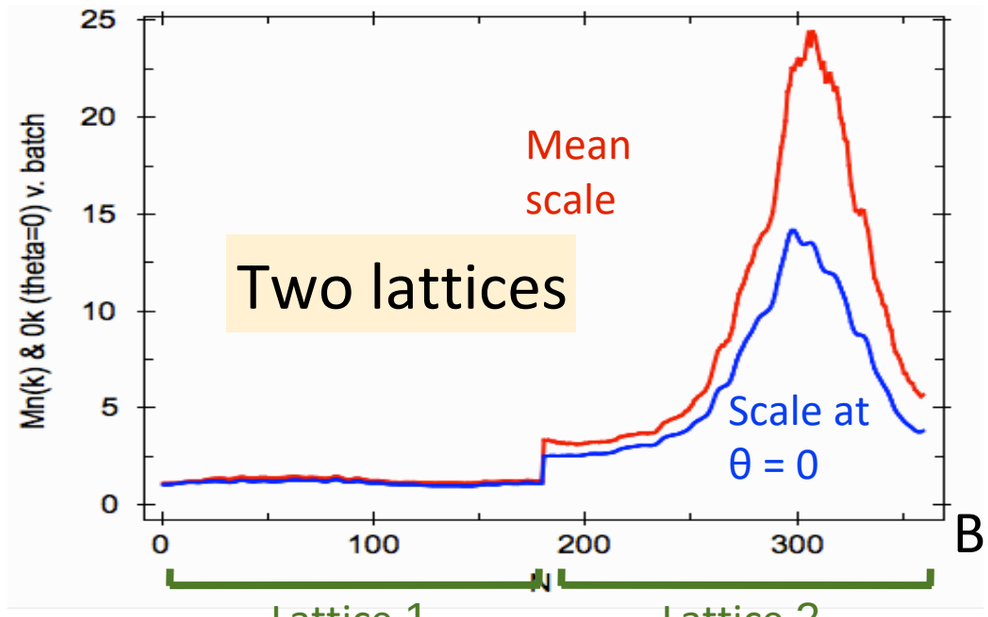


Scales



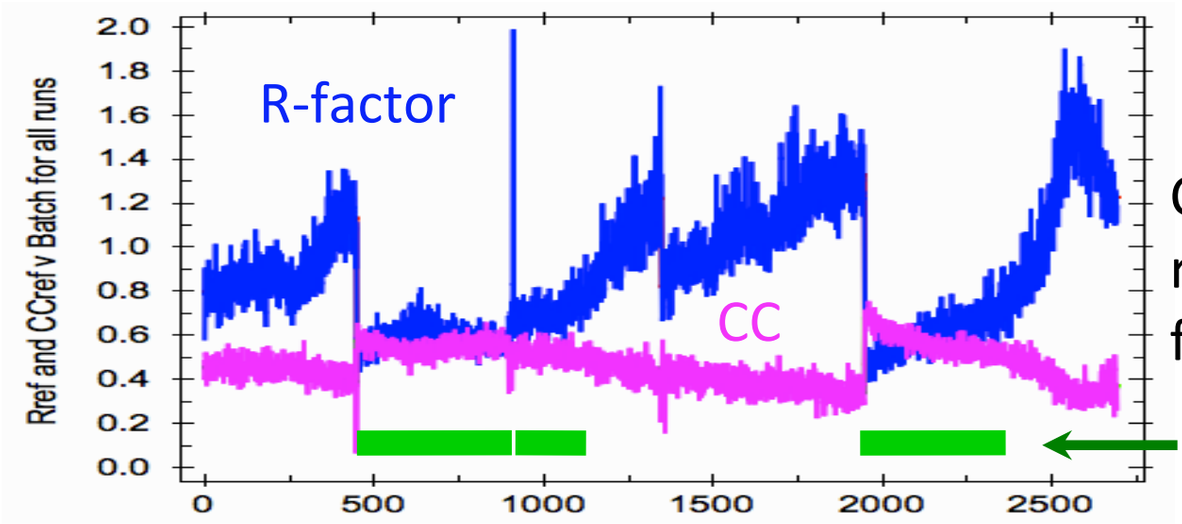
Relative B-factor

Batch



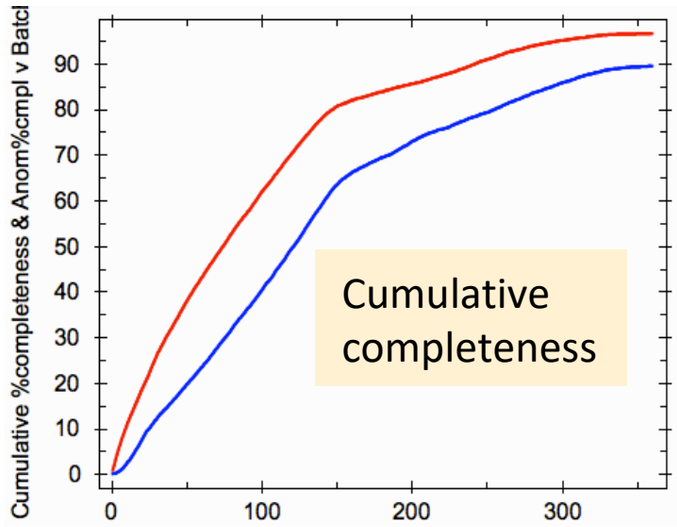
Batch

Lattice 2 is much weaker in the middle



Comparison to reference calculated from model

Good parts (least bad)
CC higher, R-factor lower



Analyses as a function of resolution

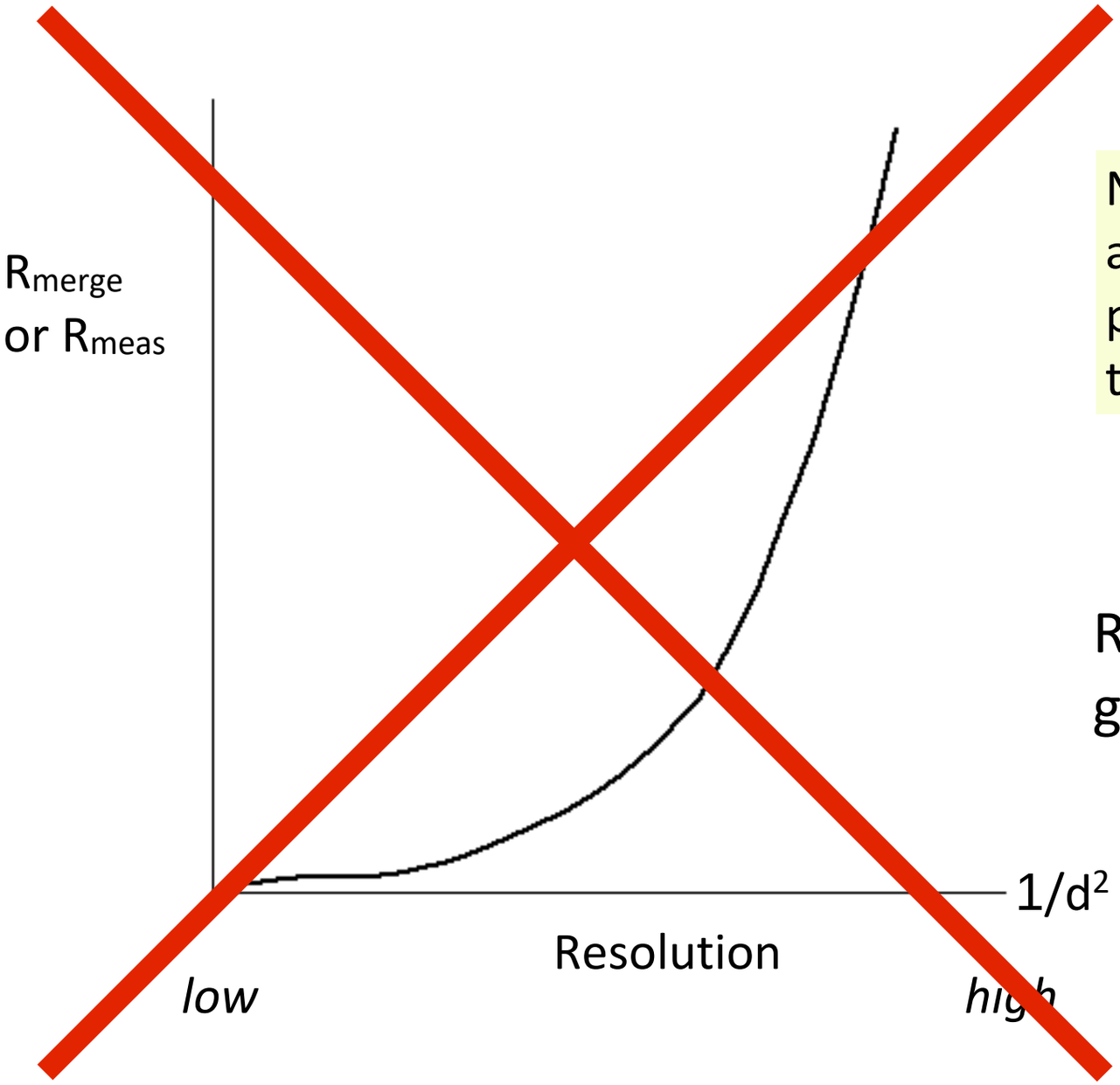
We can plot various statistics against resolution to determine where we should cut the data, allowing for anisotropy.

What do we mean by the “resolution” of the data? We want to determine the point at which adding another shell of data does not add any “significant” information, but how do we measure this?

Resolution is a contentious issue, often with referees:

What scores can we use?

What about R-factors?



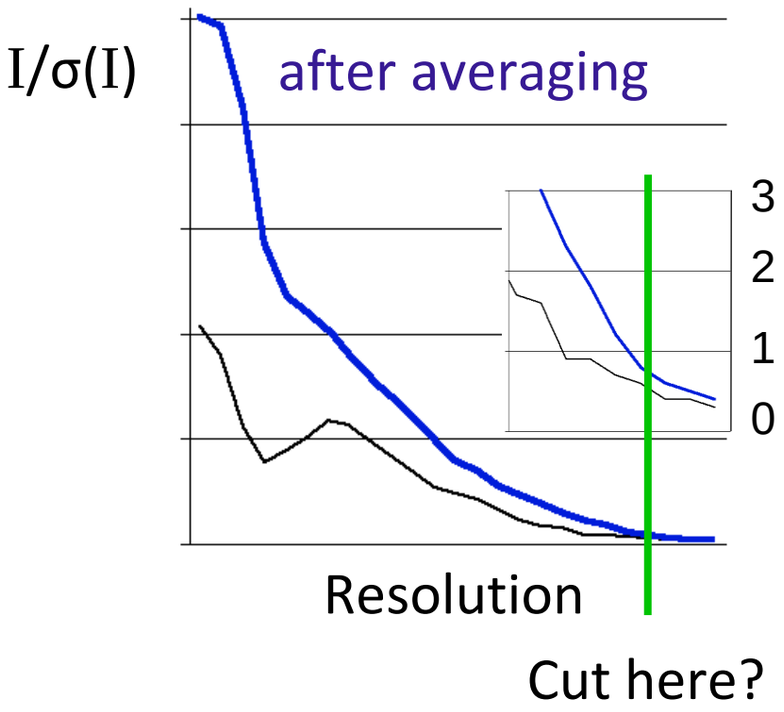
Note that R_{merge} and R_{meas} are useful for other purposes, but not for deciding the resolution cutoff

R_{merge} tends to infinity as data gets weaker

Where is the cut-off point?

Note that the crystallographic R-factor behaves quite differently: at higher resolution as the data become noisier, R_{cryst} tends to a constant value, not to infinity

1. <I/σ(I)> ≈ <signal/noise>



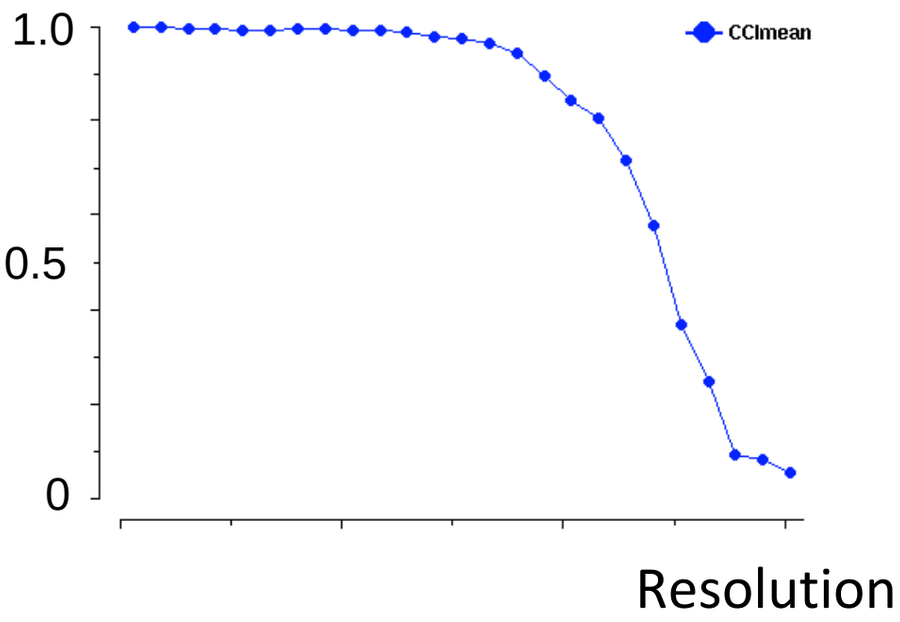
A reasonably good criterion, but it relies on σ(I), which is not entirely reliable

Cut resolution at
<I/σ(I)> after averaging
Mn(I/sd) = 1 – 2

2. $CC_{1/2}$

Half-dataset correlation coefficient:

Split observations for each reflection randomly into 2 halves, and calculate the correlation coefficient between them (or equivalent calculation)



Advantages:

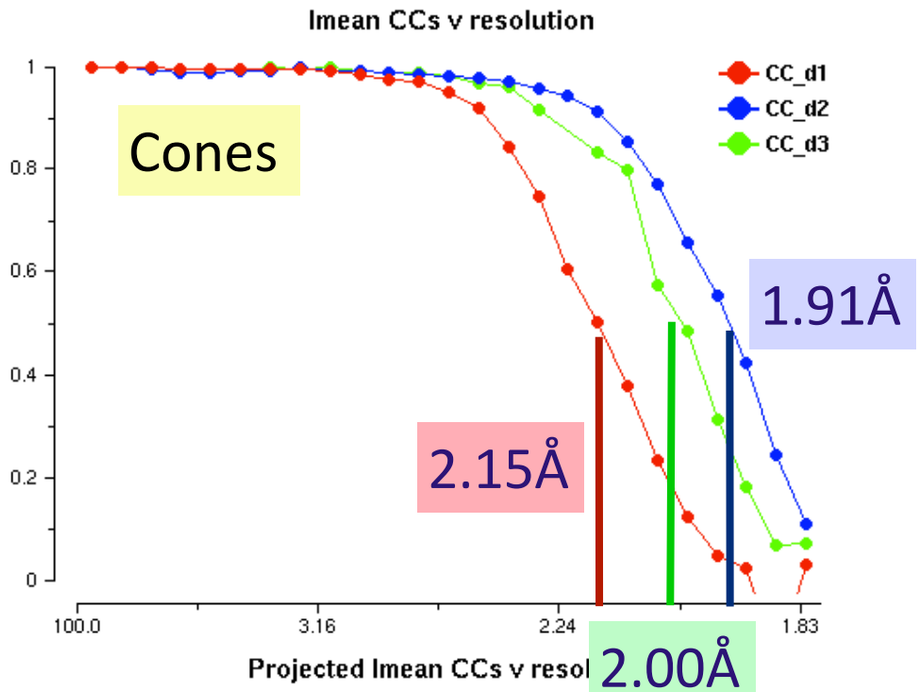
- Clear meaning to values (1.0 is perfect, 0 is no correlation) , known statistical properties
- Independent of σ(I)

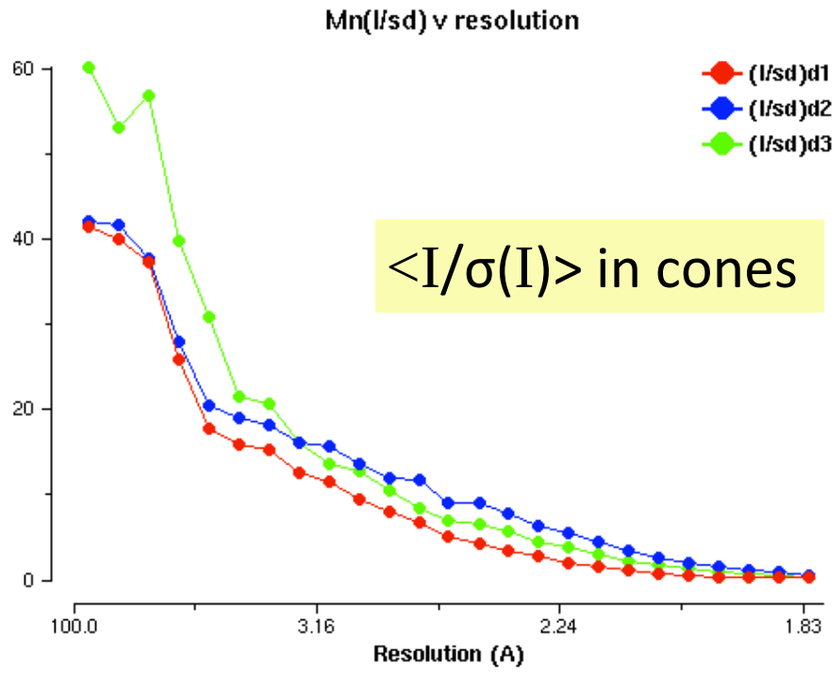
cut resolution at CC ~= 0.3 – 0.5

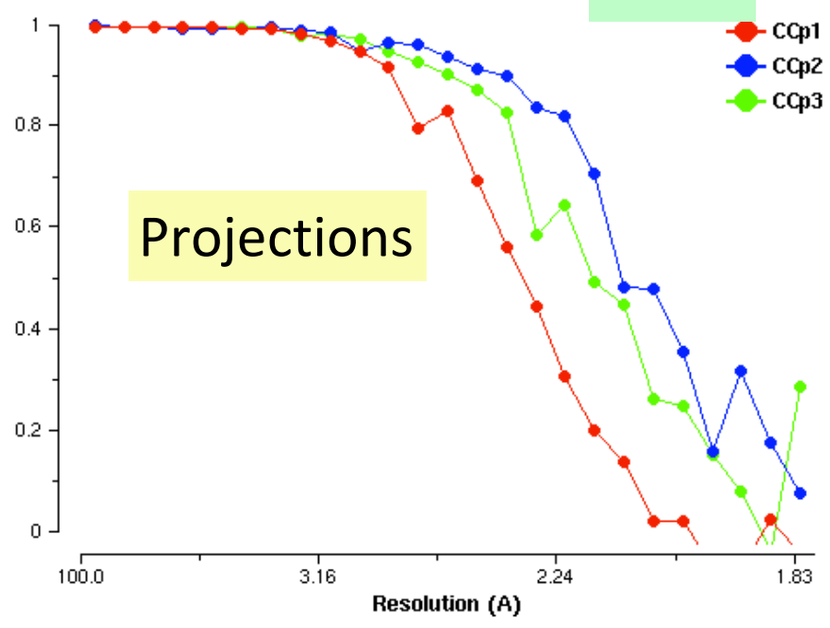
Anisotropy

Many (perhaps most) datasets are *anisotropic*
The principal directions of anisotropy are defined by symmetry (axes or planes), except in the monoclinic and triclinic systems, in which we can calculate the orthogonal principle directions

We can then analyse half-dataset CCs or <I/σ(I)> in cones around the principle axes, or as projections on to the axes







Anisotropic cutoffs are probably a Bad Thing, since it leads to strange series termination errors and problem with intensity statistics

So where should we cut the data?
Maybe at some compromise point

How should we decide the resolution of a dataset?

I don't know, but ...

Look at CC1/2, <I/σ(I)>, and anisotropy

"Best" resolution is different for different purposes, so don't cut it too soon

- Experimental phasing
  - substructure location is generally unweighted, so cut back conservatively to data with high signal/noise ratio
  - for phasing, use all "reasonable" data

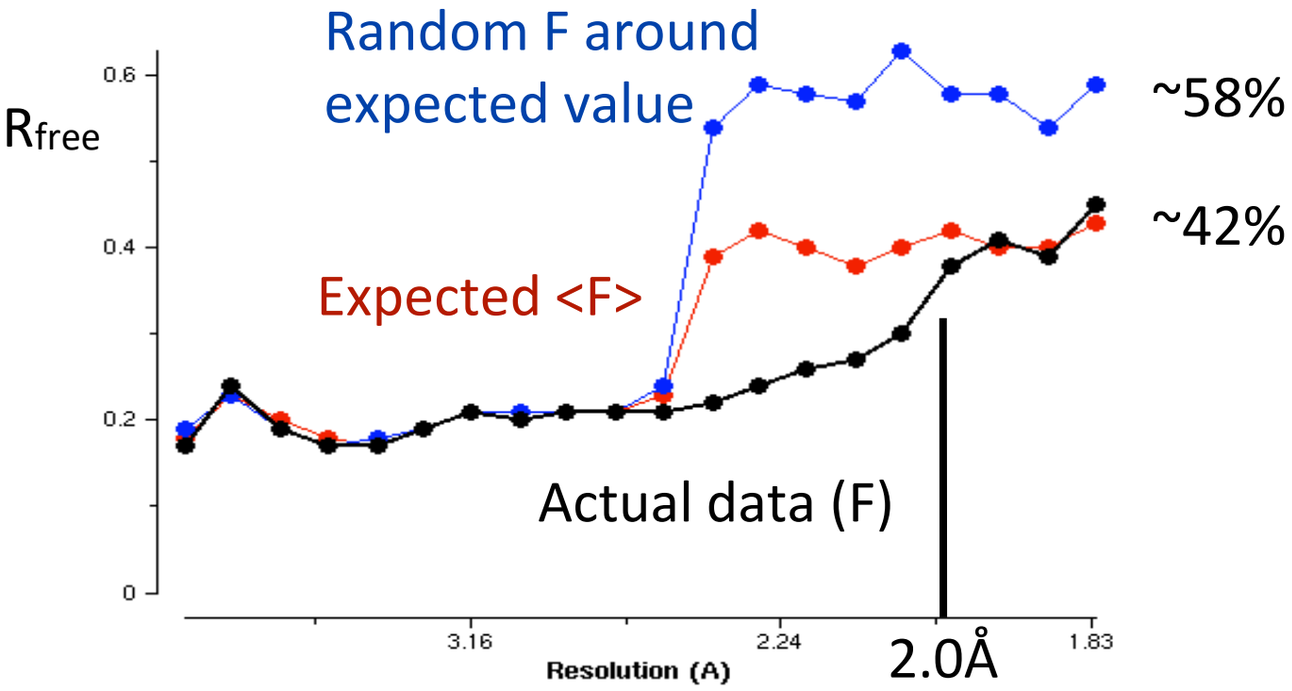
- Molecular replacement: Phaser uses likelihood weighting, but there is probably no gain in using the very weak high resolution data

- Model building and refinement: if everything is perfectly weighted (perfect error models!), then extending the data should do no harm and may do good

There is no reason to suppose that cutting back the resolution to satisfy referees will improve your model!

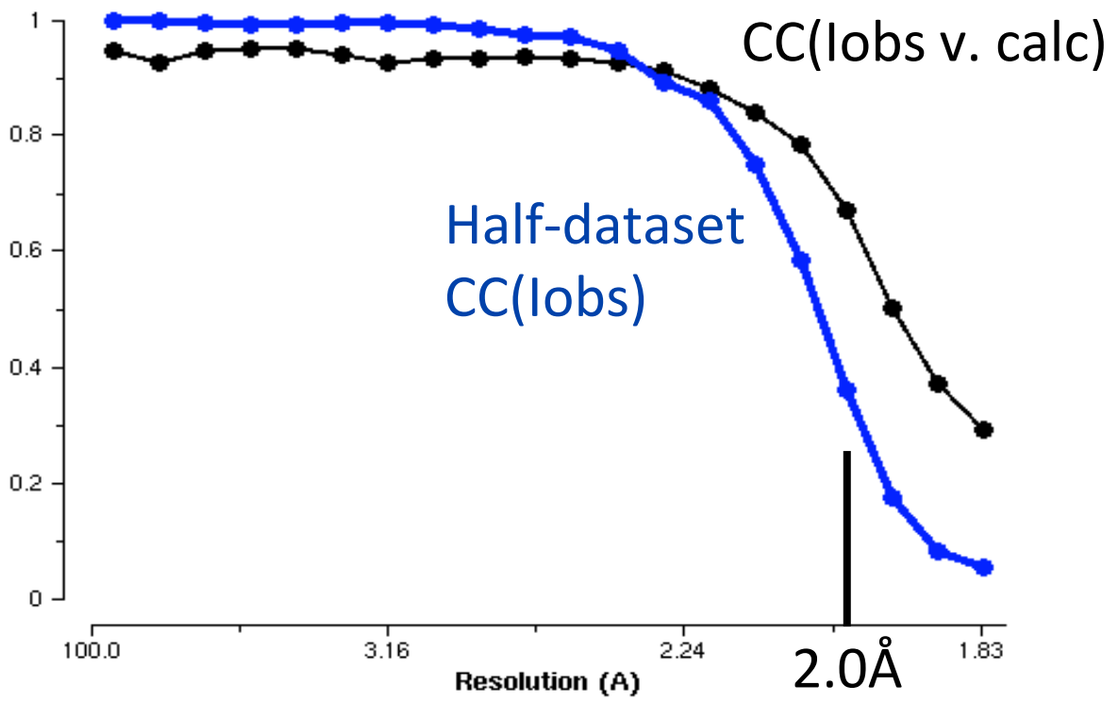
Future developments may improve treatment of weak noisy data

Example continued: refinement against real data or simulated data

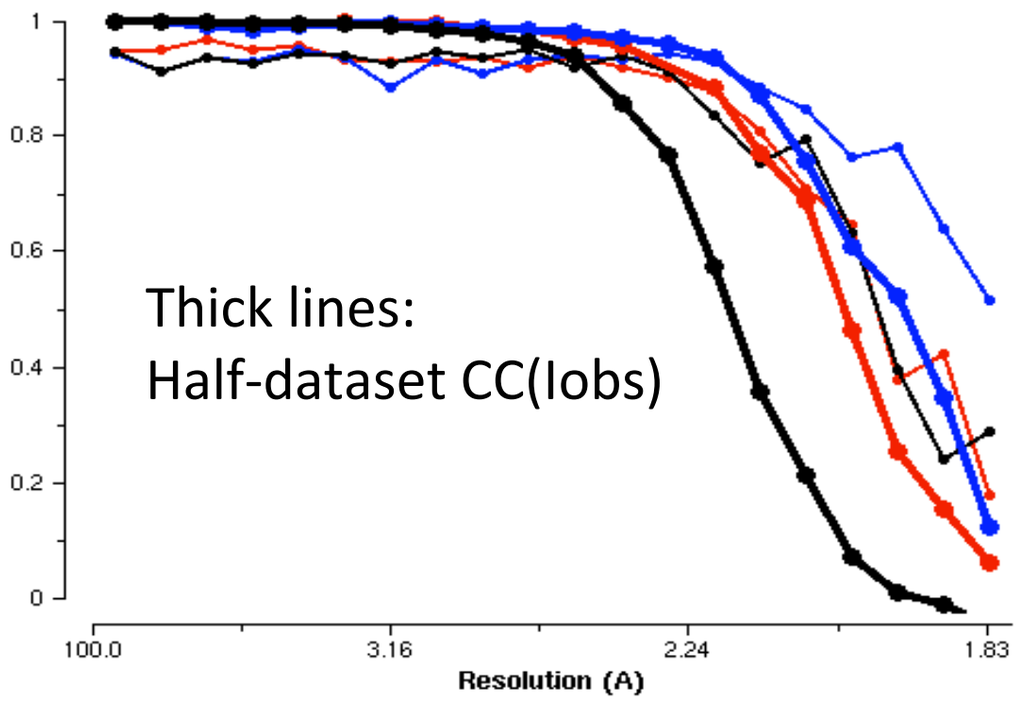


thanks to Garib Murshudov

All these indicators are roughly consistent that a suitable resolution cutoff is around 2.0Å, but that anything between 1.9Å and 2.1Å can be justified, **with current technologies**



Anisotropy



Improved estimate of σ(I)

The error estimate σ(I) from the integration program is too small particularly for large intensities. A “corrected” value may be estimated by increasing it for large intensities such that the mean scatter of scaled observations on average equals σ’(I), in all intensity ranges

Corrected $\sigma'(Ihl)^2 = SDfac^2 [\sigma^2 + SdB \langle I_h \rangle + (SdAdd \langle I_h \rangle)^2]$

SDfac, SdB and SdAdd are automatically adjusted parameters

Sigma(scatter/SD) and mean(χ^2) should ≈ 1.0

… but error estimation is difficult

▼ SD analysis

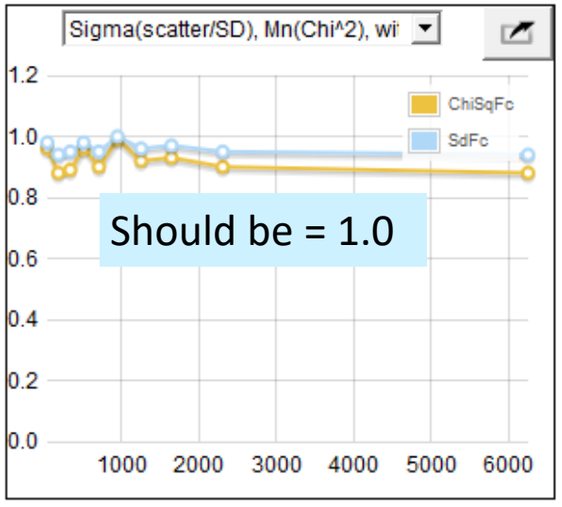
Analysis of sd(I)

Error estimates sd(I) are analysed as functions of intensity (and resolution and batch),
as sigma(delta(I)/sd(I)), where delta(I) = Ihl−<Ih>, and as reduced chi^2 (goodness of fit), ([delta(I)/sd(I)]^2)/(n−m)

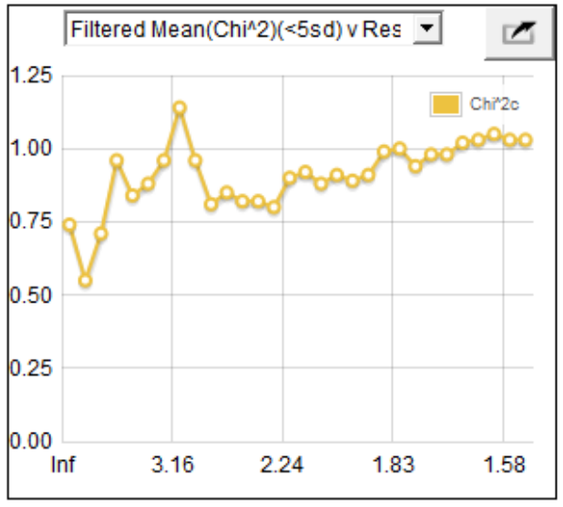
Parameters for improvement of sd(I) estimates: sd'(I) = SdFac * Sqrt[sd(I)^2 + SdB I + (SdAdd I)^2]
SDcorrection parameters are flagged with moderate (*) or severe (x) warning, if SDfac > 2.0 or 3.0,
or if SDadd is negative or > 0.06 or 0.1 respectively
ISa is the asymptotic maximum I/sig(I) = 1/(SdFac*SdAdd)

	SdFac	flag	SdB	SdAdd	flag	ISa
Run 1	0.92		2.7	0.0424		25.6

Analysis against intensity



Analysis against resolution



Outliers

Detection of outliers is easiest if the multiplicity is high

Removal of spots behind the backstop shadow does not work well at present: usually it rejects all the good ones, so tell integration program (eg Mosflm) where the backstop shadow is.

Reasons for outliers

- outside reliable area of detector (eg behind shadow)

  specify backstop shadow, calibrate detector

- ice spots

  do not get ice on your crystal!

- multiple lattices

  find single crystal

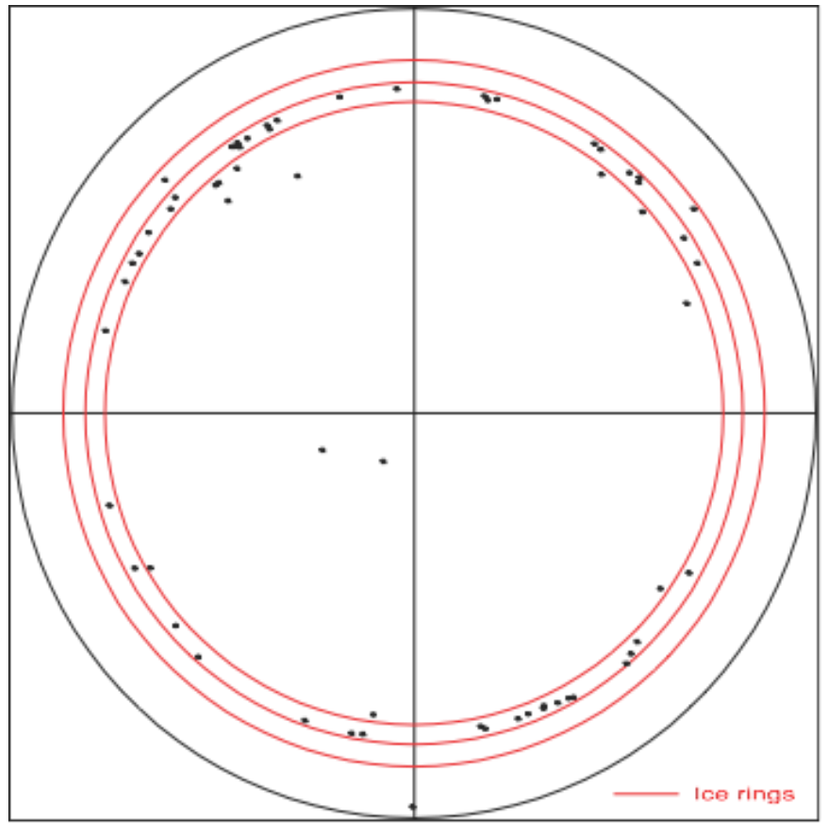
- zingers
- bad prediction (spot not there)

  improve prediction

- spot overlap

  lower mosaicity, smaller slice, move detector back

  deconvolute overlaps

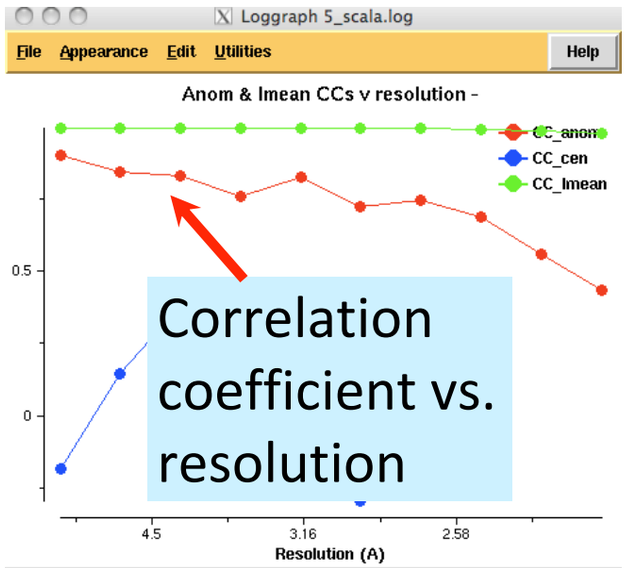


Rejects lie on ice rings (red) (ROGUEPLOT in Scala)

Position of rejects on detector

Detecting anomalous signals

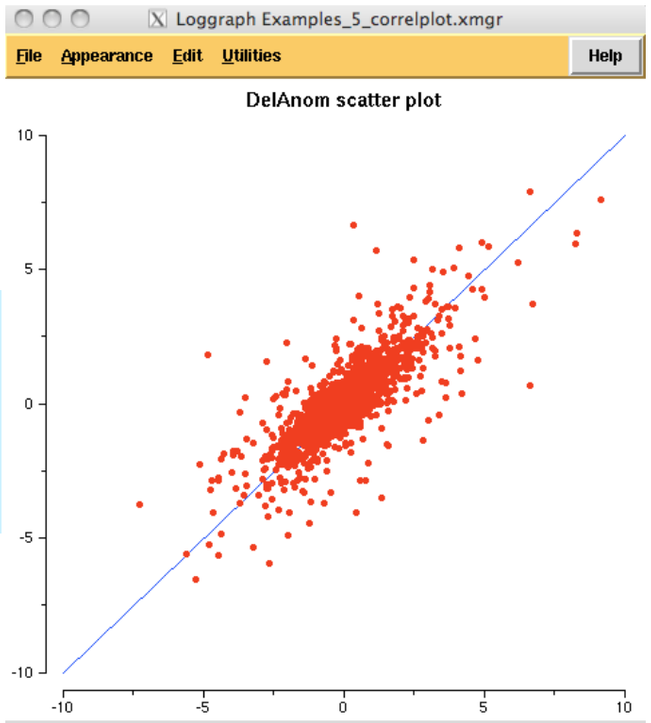
The data contains both I+ (hkl) and I- (-h-k-l) observations and we can detect whether there is a significant difference between them.



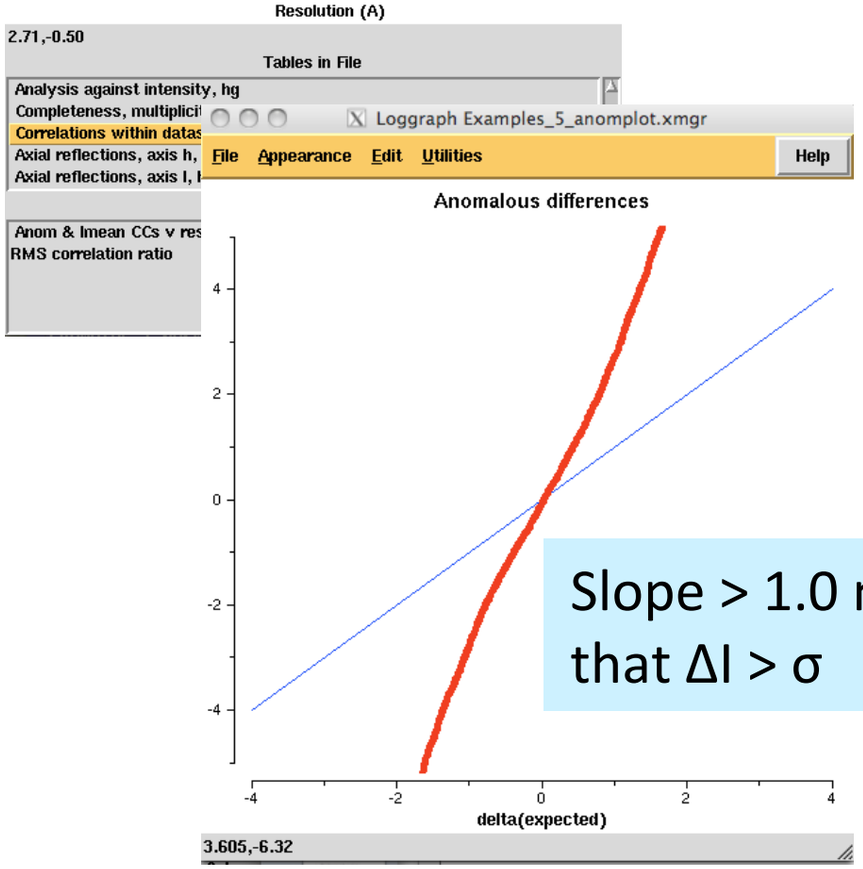
Correlation coefficient vs. resolution

Split one dataset randomly into two halves, calculate correlation between the two halves or
compare different wavelengths (MAD)

Plot ΔI_1 against ΔI_2 should be elongated along diagonal



Strong anomalous signal



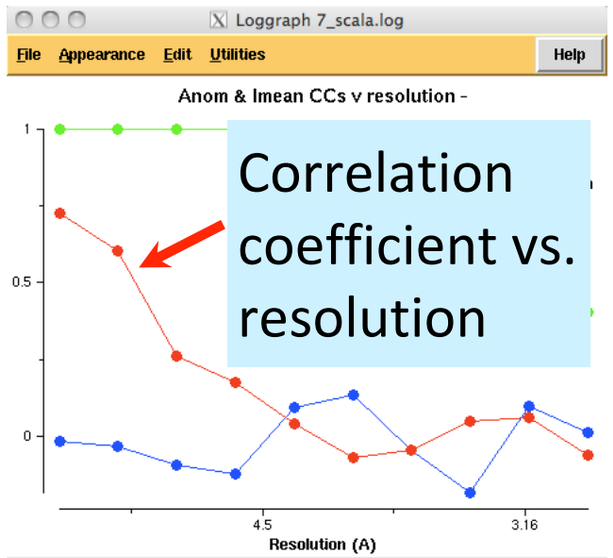
Slope > 1.0 means that $\Delta I > \sigma$

Ratio of width of distribution along diagonal to width across diagonal

"RMS correlation ratio"

Detecting anomalous signals

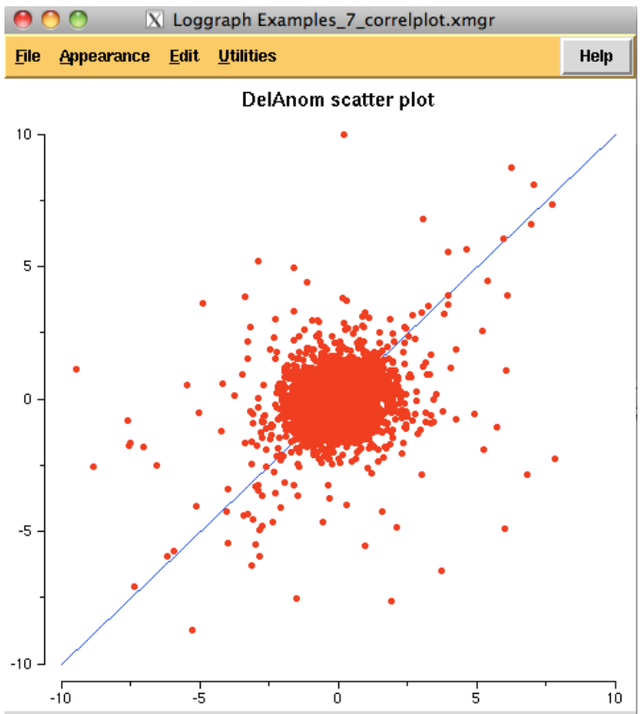
The data contains both I+ (hkl) and I- (-h-k-l) observations and we can detect whether there is a significant difference between them.



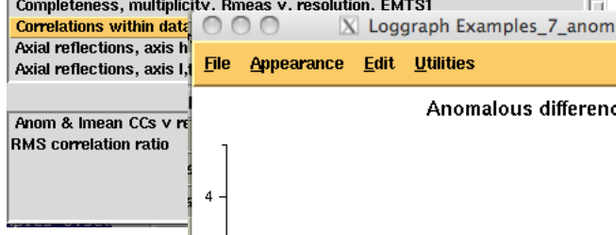
Correlation coefficient vs. resolution

Split one dataset randomly into two halves, calculate correlation between the two halves or
compare different wavelengths (MAD)

Plot ΔI_1 against ΔI_2 should be elongated along diagonal

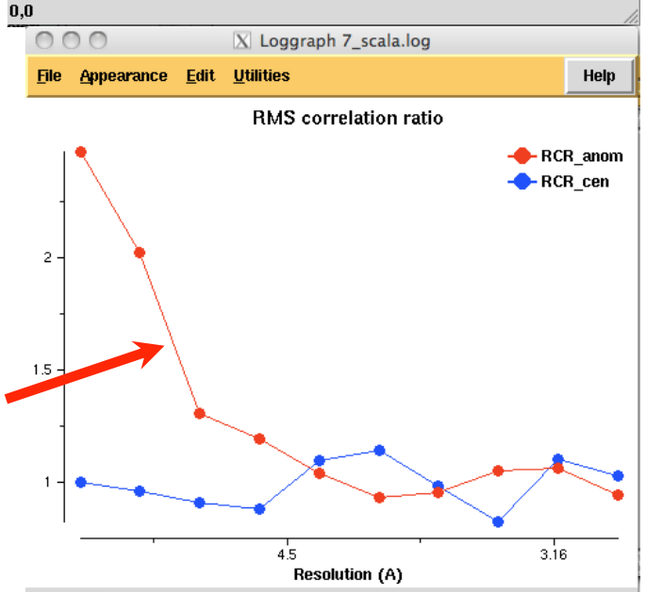


Weak but useful anomalous signal

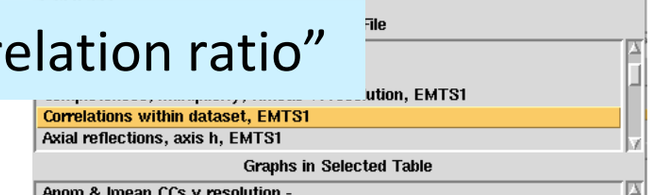


Slope > 1.0 means that $\Delta I > \sigma$

Ratio of width of distribution along diagonal to width across diagonal



"RMS correlation ratio"



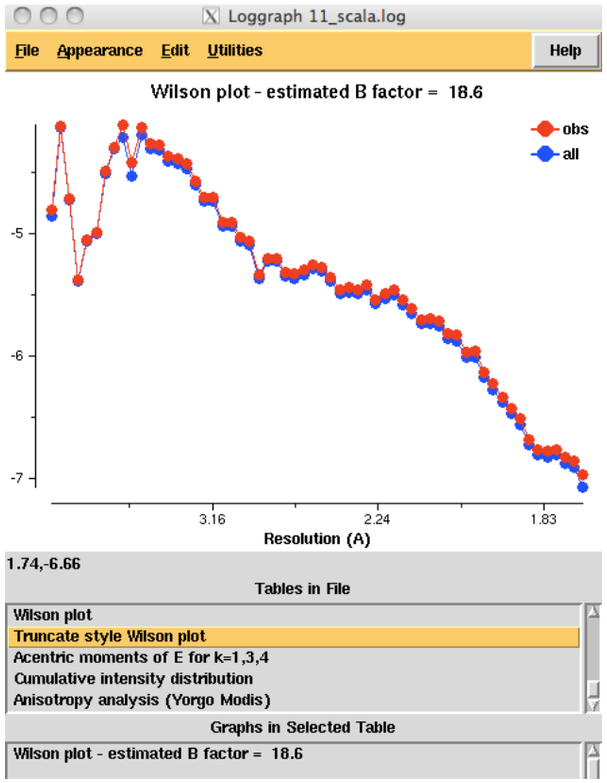
Intensity statistics

We need to look at the distribution of intensities to detect twinning

Assuming atoms are randomly placed in the unit cell, then

$$\langle I \rangle(s) = \langle F\, F^* \rangle(s) = \Sigma_j\, g(j, s)^2$$

where g(j, s) is the scattering from atom j at $s = \sin\theta/\lambda$



Average intensity falls off with resolution, mainly because of atomic motions (B-factors)

$\langle I \rangle(s) = C \exp(-2\, B\, s^2)$

Wilson plot: $\log(\langle I \rangle(s))$ vs s^2

This would be a straight line if all the atoms had the same B-factor

For the purposes of looking for crystal pathologies, we are not interested in the variation with resolution, so we can use "normalised" intensities which are independent of resolution

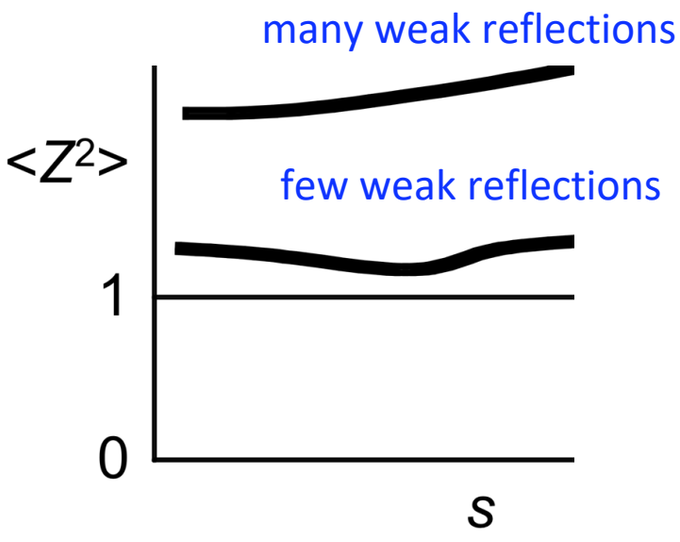
Normalised intensities: relative to average intensity at that resolution

$Z(h) = I(h)/\langle I(s)\rangle \approx |E|^2$

$\langle Z(s)\rangle = 1.0$ by definition

$\langle Z^2(s)\rangle > 1.0$ depending on the distribution

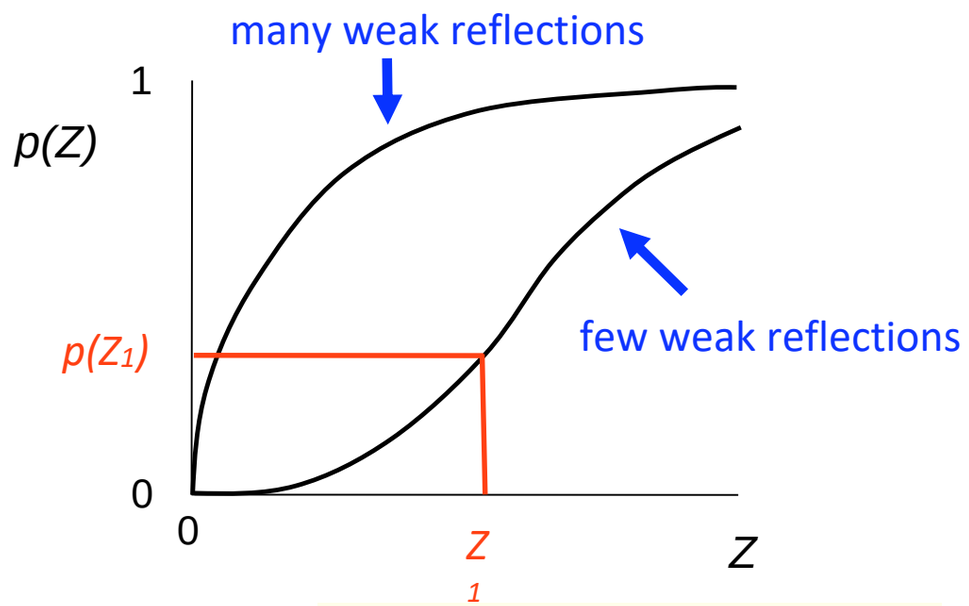
$\langle Z^2(s)\rangle$ is larger if the distribution of intensities is wider: it is the 2nd moment
ie the *variance* (this is the 4th moment of E)



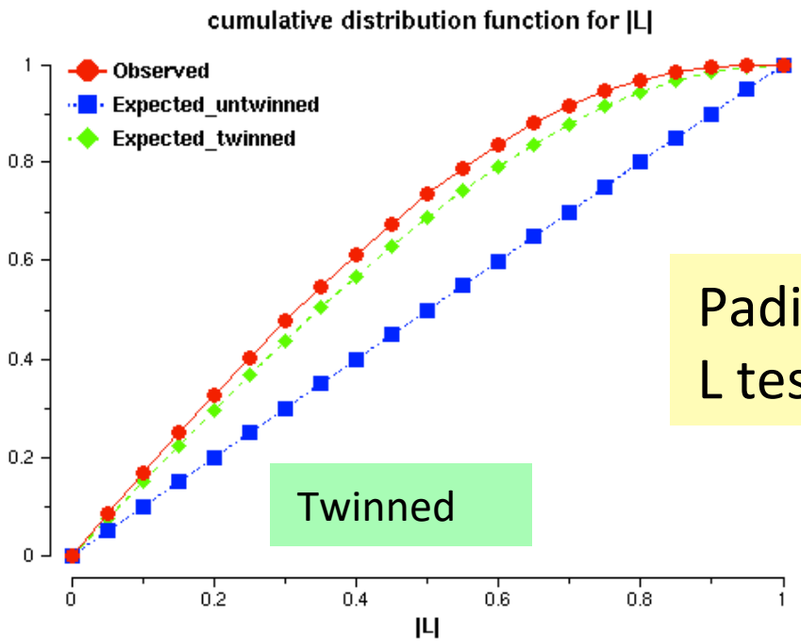
Moment

s

Cumulative distribution of Z: p(Z) vs. Z



$p(Z_1)$ is the proportion of reflections with $Z < Z_1$



Twinned

Padilla – Yeates
L test

Other features of the intensity distribution which may obscure or mimic twinning

Translational non-crystallographic symmetry:

whole classes of reflections may be weak

eg h odd with a NCS translation of ~1/2, 0 0

<I> over all reflections is misleading, so Z values are inappropriate

The reflection classes should be separated (not yet done)

Anisotropy: <I> is misleading so Z values are wrong

ctruncate applies an anisotropic scaling before analysis

Weak data: the ideal statistics are based on perfect data.

If the signal/noise ratio is small, then the statistics may falsely suggest twinning

Overlapping spots: a strong reflection can inflate the value of a weak neighbour, leading to too few weak reflections

this mimics the effect of twinning

Estimation of amplitude |F| from intensity *I*

If we knew the true intensity *J* then we could just take the square root

$$|F| = \sqrt{J}$$

But measured intensities *I* have an error σ(*I*) so a small intensity may be measured as negative.

The "best" estimate of |F| larger than √*I* for small intensities (<~ 3 σ(*I*)) to allow for the fact that we know than |F| must be positive

[c]truncate estimates |F| from *I* and σ(*I*) using the average intensity in the same resolution range: this give the prior probability *p(J)*

$$\mathrm{E}(F\,;\,I,\sigma(I)) = \int_0^\infty F\;p(I\,;\,J,\sigma(I))\;p(J)\;dJ$$

French & Wilson 1978

BUT best to use intensities I rather than amplitude F wherever possible

Summary: Questions & Decisions

- *Do look critically at the data processing statistics*

  - What is the point group (Laue group)?
  - What is the space group?
  - Was the crystal dead at the end?
  - Is the dataset complete?
  - Do you want to cut back the resolution?
  - Is this the best dataset so far for this project?
  - Should you merge data from multiple crystals?
  - Is there anomalous signal (if you expect one)?
  - Are the data twinned?

Try alternative processing strategies: different choices of cutoffs, merging crystals, etc

test with MR (log-likelihood gain) or refinement (R_{free}, map quality)

Data processing is not necessarily something you just do once

Acknowledgements

Andrew Leslie	many discussions
Harry Powell	many discussions
Ralf Grosse-Kunstleve	cctbx
Kevin Cowtan	clipper, C++ advice
Airlie McCoy	C++ advice, code, useful suggestions, etc
Randy Read & co.	minimiser
Graeme Winter	testing & bug finding
Clemens Vonrhein	testing & bug finding
Eleanor Dodson	many discussions
Andrey Lebedev	intensity statistics & twinning
Norman Stein	ctruncate
Charles Ballard	ctruncate
George Sheldrick	discussions on symmetry detection
Garib Murshudov	intensity statistics
Martyn Winn & CCP4 gang	ccp4 libraries
Peter Briggs	ccp4i
Liz Potterton	ccp4i2
Martin Noble	ccp4i2
Kay Diederichs	discussions and papers