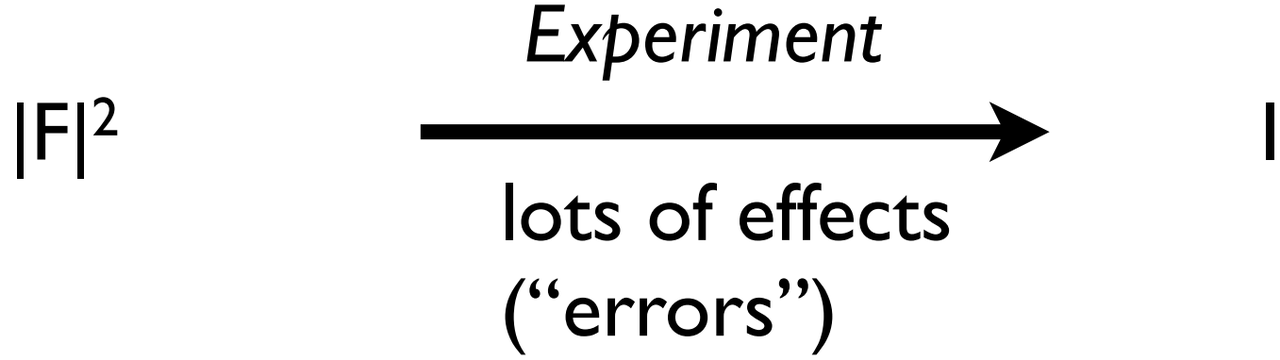
Data Reduction

Space Group Determination, Scaling and Intensity Statistics

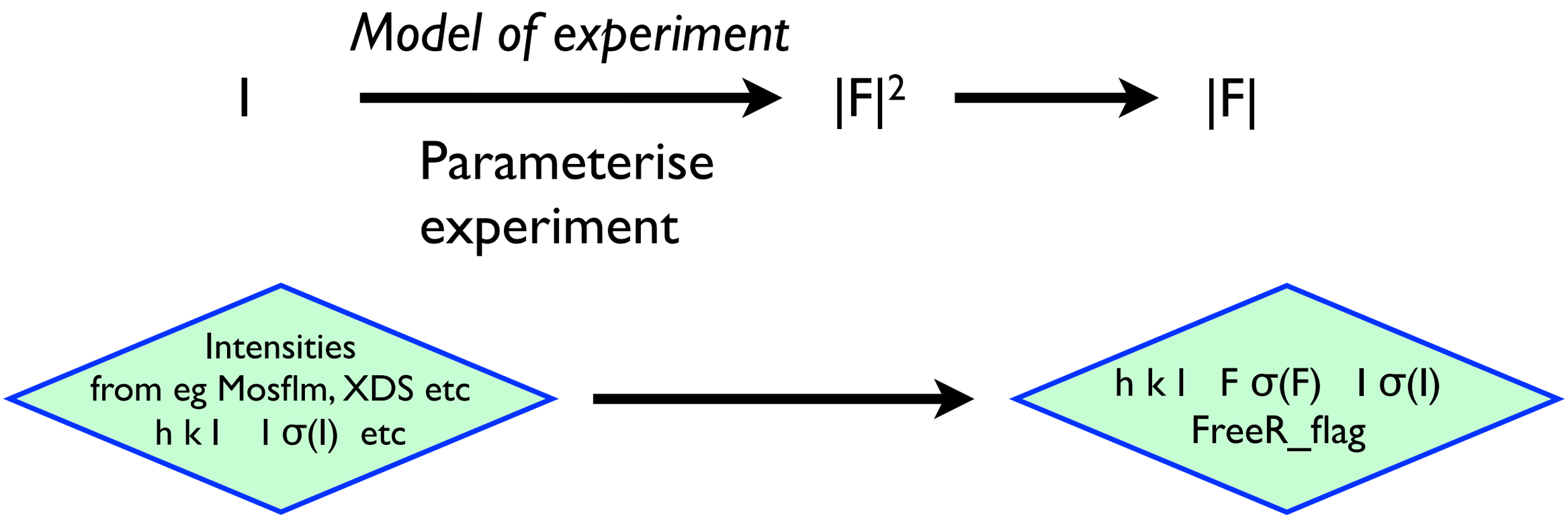
Phil Evans Tsukuba November 2014

MRC Laboratory of Molecular Biology
Cambridge UK

Scaling and Merging



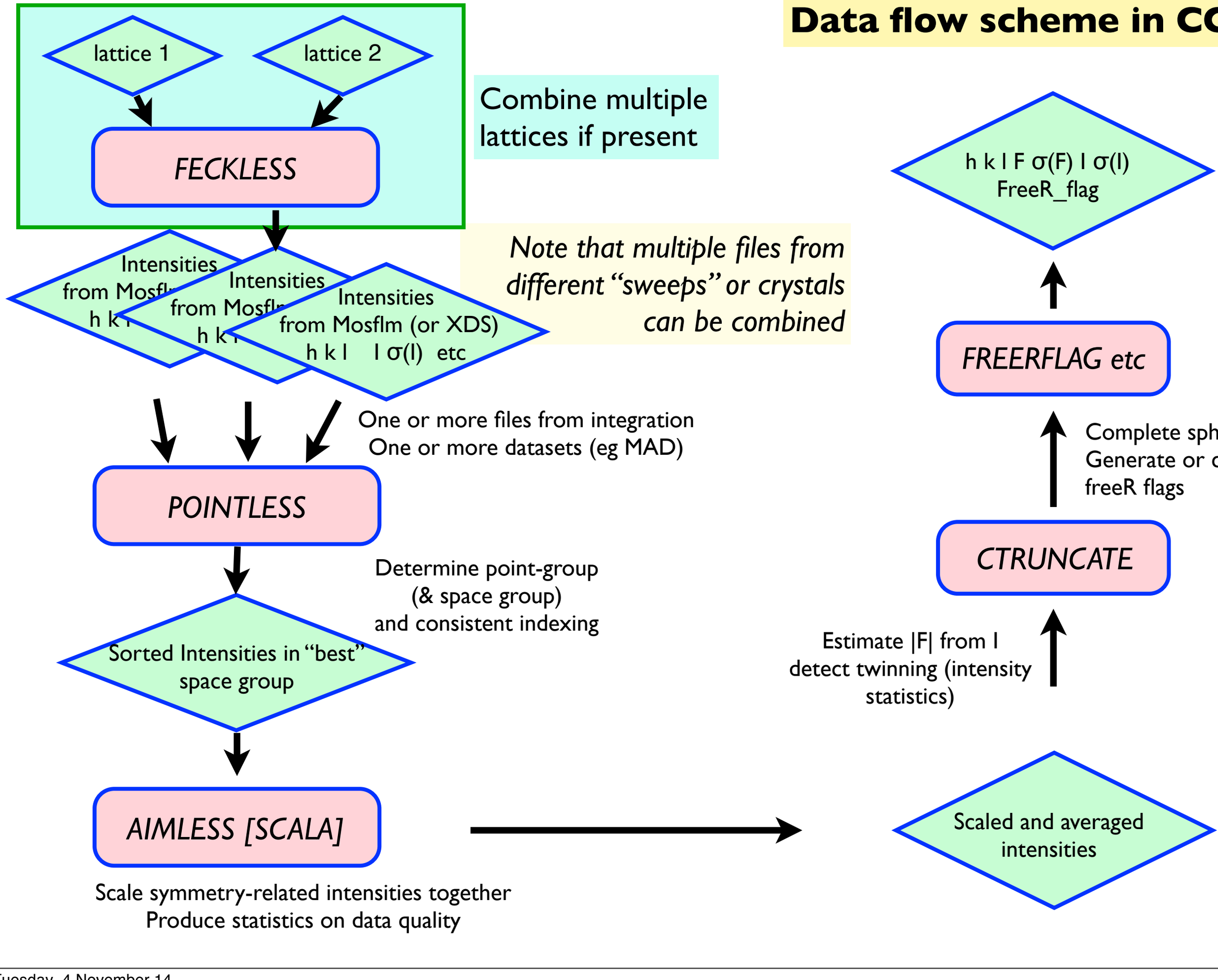
Our job is to invert the experiment: we want to *infer* |F| from our measurements of intensity I



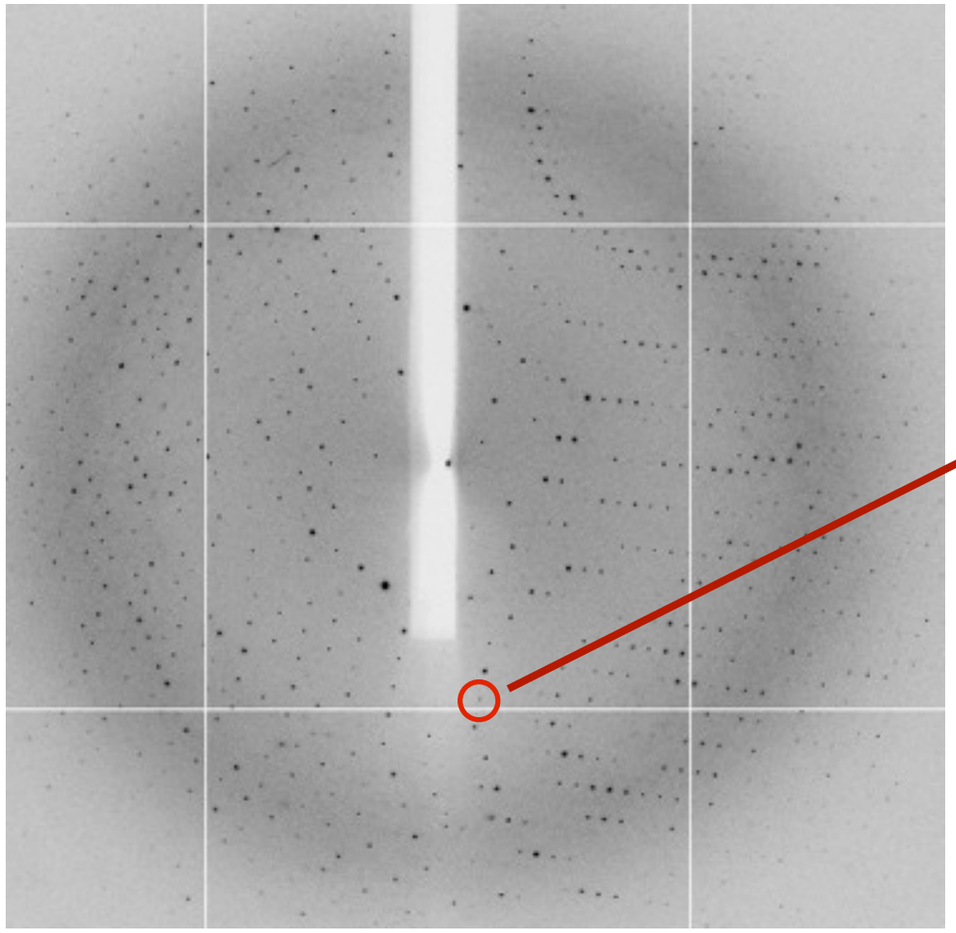
Integration and data reduction can be done in an automated pipeline such as XIA2 (Graeme Winter): this goes from images to a list of hkl F ready for structure determination)

Automation works pretty well, but in difficult cases you may need finer control over the process

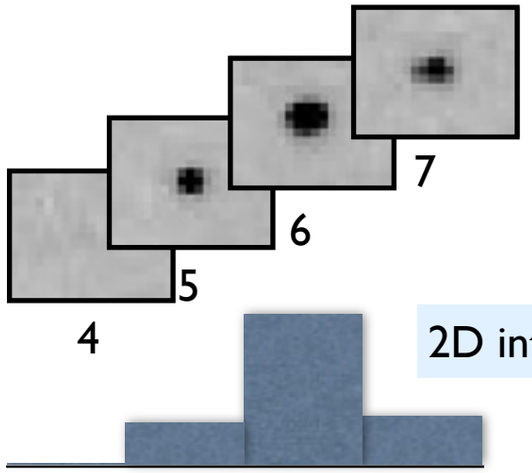
Data flow scheme in CCP4



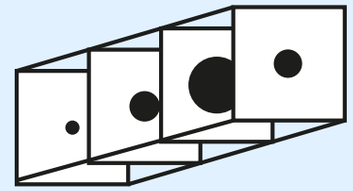
Track one reflection through the process



Spot over 4 images



Spot profile



3D integration eg 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

Partial

Partial

-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	202						1.00
-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

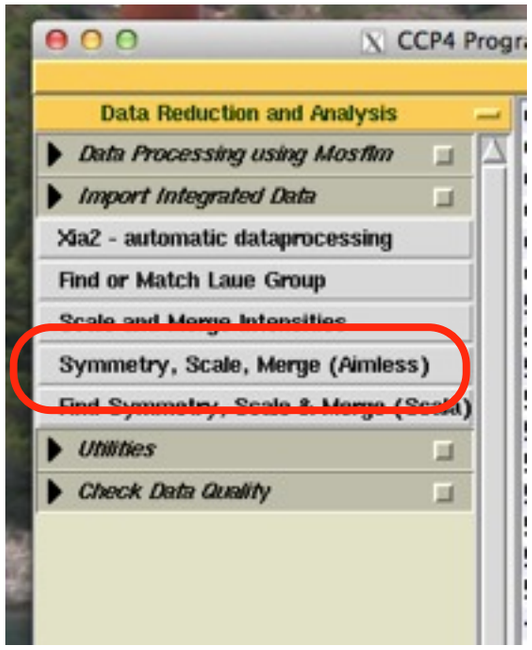
Infer |F| from I

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

Data reduction from ccp4i

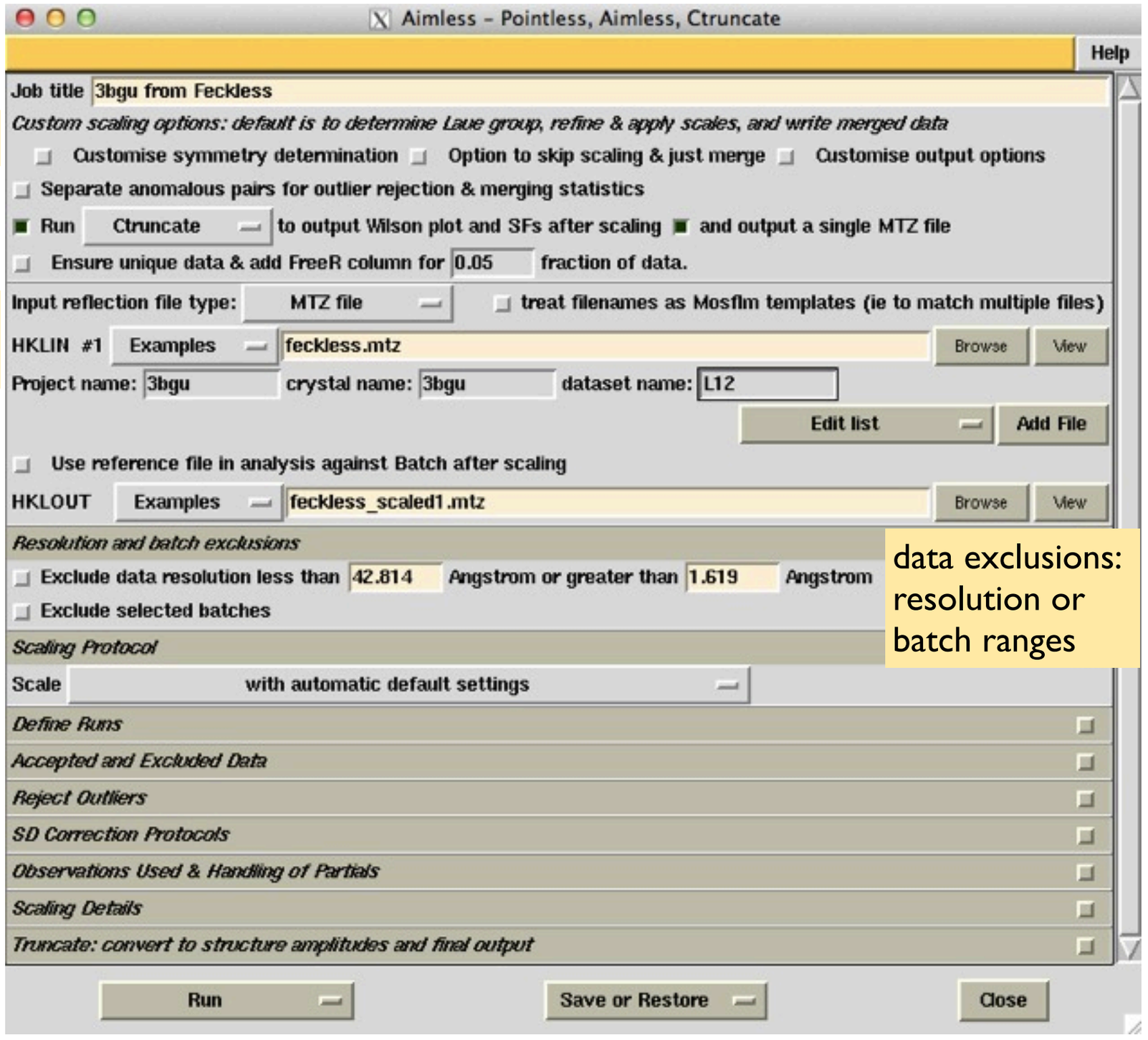
alternative to QuickScale



Note that ccp4i does not (yet) run FECKLESS on multilattice data

main options

input file specifications



data exclusions: resolution or batch ranges

POINTLESS can read one or more files, which may belong to multiple datasets (eg for MAD)

Recognised formats (unmerged data):

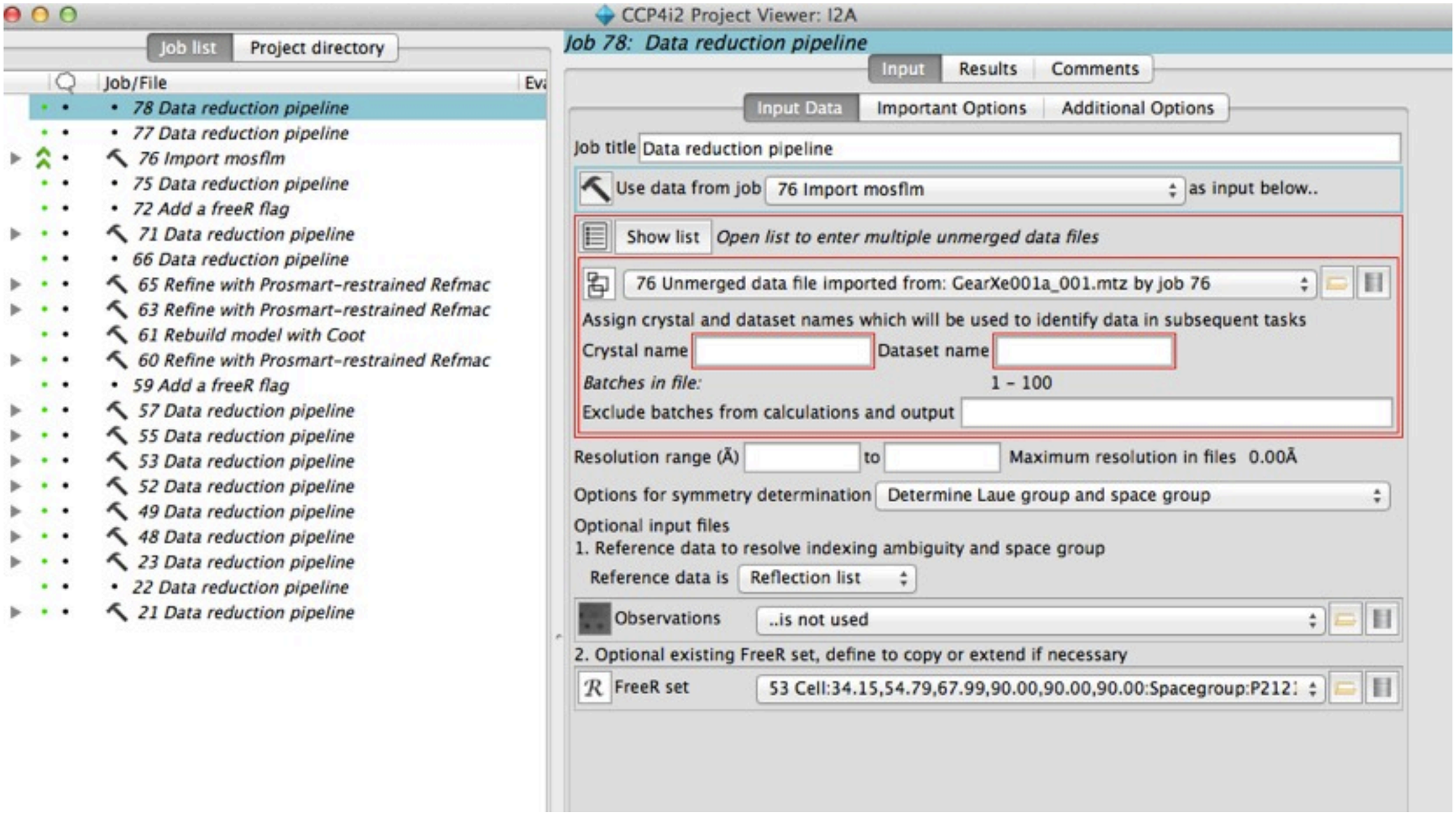
- MTZ (Mosflm)
- XDS
- SAINT

Formats with limited scaling options

- Scalepack
- Shelx

ccp4i2

Currently under test for release



ccp4i2 Results

CCP4i2 Project Viewer: I2A

Job 78: Data reduction pipeline

Input | Results | Comments

Key summary iverall summar er merging gra pace group de :tails of mergi tensity statisti ow job run det

▼ Hide Key summary

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

Solution probability: 0.704, Confidence 0.629

Key statistics for Dataset: I2A/Xe/Xe

Resolution of input data: 1.78A, resolution estimate: beyond 1.78A
Rmeas: overall 0.048, inner bin 0.069
In outer bin: Mean(I/sdI) 4.0 CC(1/2) 0.893
Anomalous CC(1/2) in inner bin 0.068

► Show Overall summary

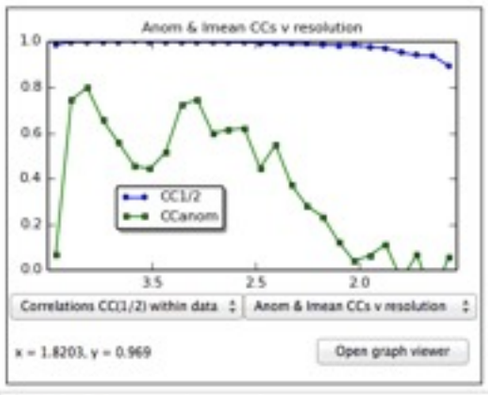
► Show Other merging graphs

► Show Details of space group determination

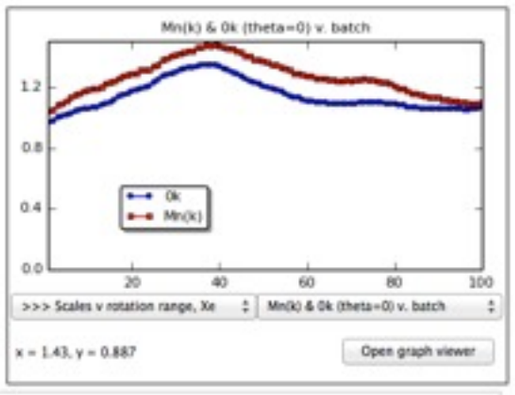
► Show Details of merging

► Show Intensity statistics

Analysis as a function of resolution

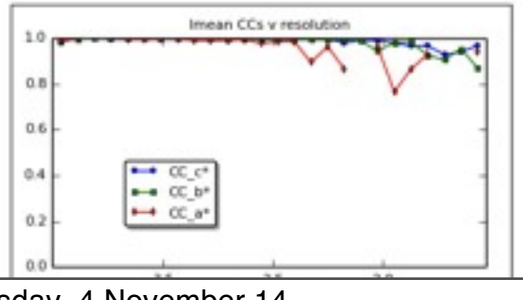


Analysis as a function of batch



▼ Hide Other merging graphs

Other merging statistics graphs



Scatter plots etc

Normal probability | Anomalous difference

DelAnom scatterplot | Outlier positions

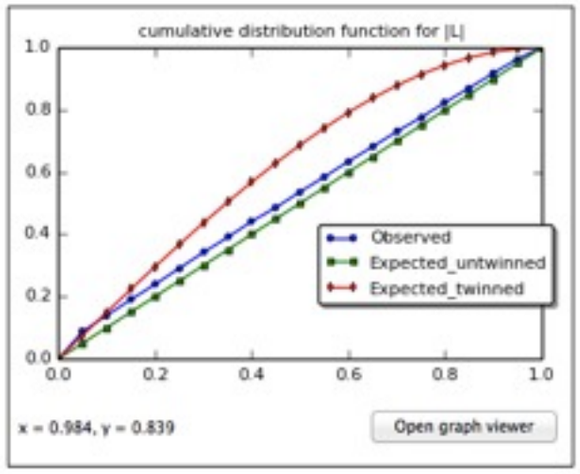
Space group determination

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	0.629
Laue group confidence	0.980
Laue group probability	0.983
Systematic absence probability	0.716

L-test for twinning



Data internal consistency statistics

Summary of merging statistics for dataset
I2A/Xe/Xe

	Overall	Inner	Outer
Low resolution limit	34.00	34.00	1.82
High resolution limit	1.78	9.08	1.78
Rmerge(within I+/I-)	0.036	0.050	0.205
Rmerge(all I+ and I-)	0.053	0.075	0.242
Rmeas (within I+/I-)	0.048	0.069	0.275
Rmeas (all I+ & I-)	0.062	0.090	0.286
Rpim (within I+/I-)	0.032	0.047	0.181
Rpim (all I+ & I-)	0.031	0.049	0.149
Rmerge in top intensity bin	0.021		
Number of observations	44778	300	1168
Number unique	12161	108	363
Mean(I)/sd(I)	18.4	33.5	4.0
Half-set correlation CC(1/2)	0.998	0.985	0.893
Completeness	95.4	88.6	50.9
Multiplicity	3.7	2.8	3.2
Anomalous completeness	89.0	89.3	41.5
Anomalous multiplicity	1.9	2.0	1.9
DelAnom CC(1/2)	0.502	0.068	0.056
Mid-Slope of Anom Probability	1.275		

Download

There appears to be a significant anomalous signal so anomalous flag was switched ON

Symmetry determination, point group and space group (POINTLESS)

Indexing in eg MOSFLM only gives the possible lattice symmetry, ie constraints of unit cell dimensions. Crystal classes: cubic, hexagonal/trigonal, tetragonal, orthorhombic, monoclinic, or triclinic, + lattice centring P, C, I, R, or F

1. from the cell dimensions, determine the maximum possible lattice symmetry (ignoring any input symmetry)
2. for each possible rotation operator, score 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 maximum down to P1)
4. score axial systematic absences to detect screw axes, hence space group (note that axial observations are sometimes unobserved)

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

Pseudo-cubic example

Only orthorhombic symmetry operators are present

```
Analysing rotational symmetry in lattice group P m -3 m
-------------------------------------------------
Scores for each symmetry element

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}
```

What score to use?

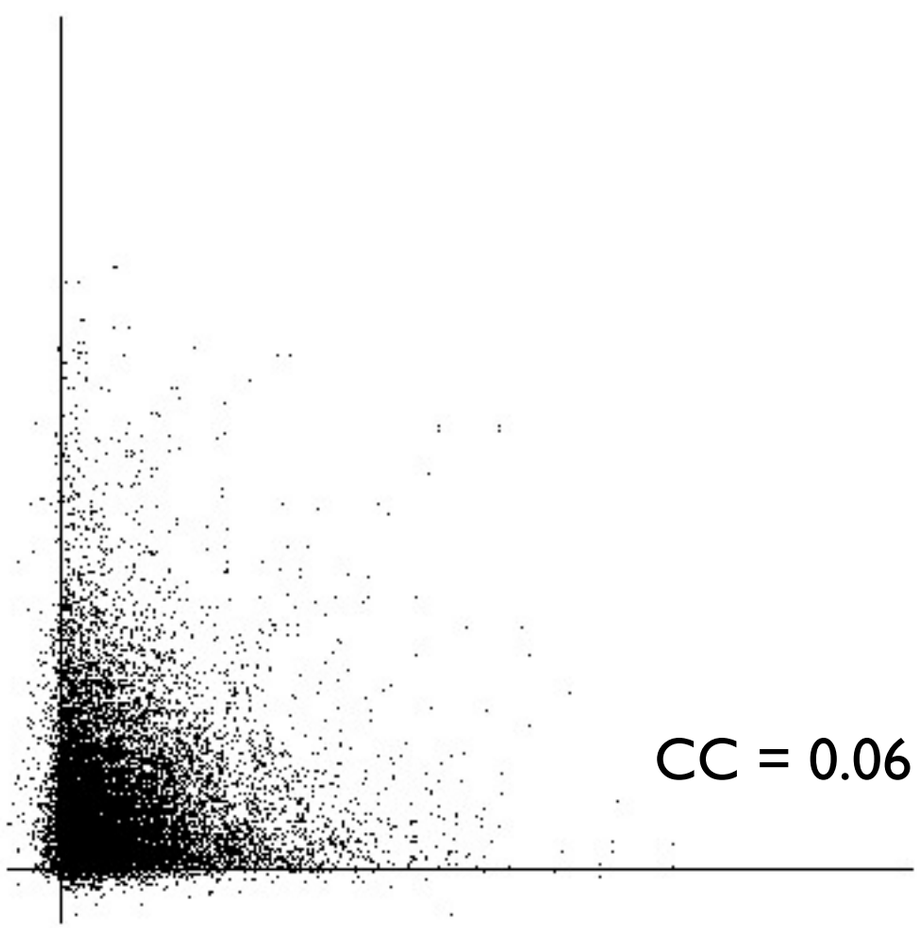
Linear correlation coefficient

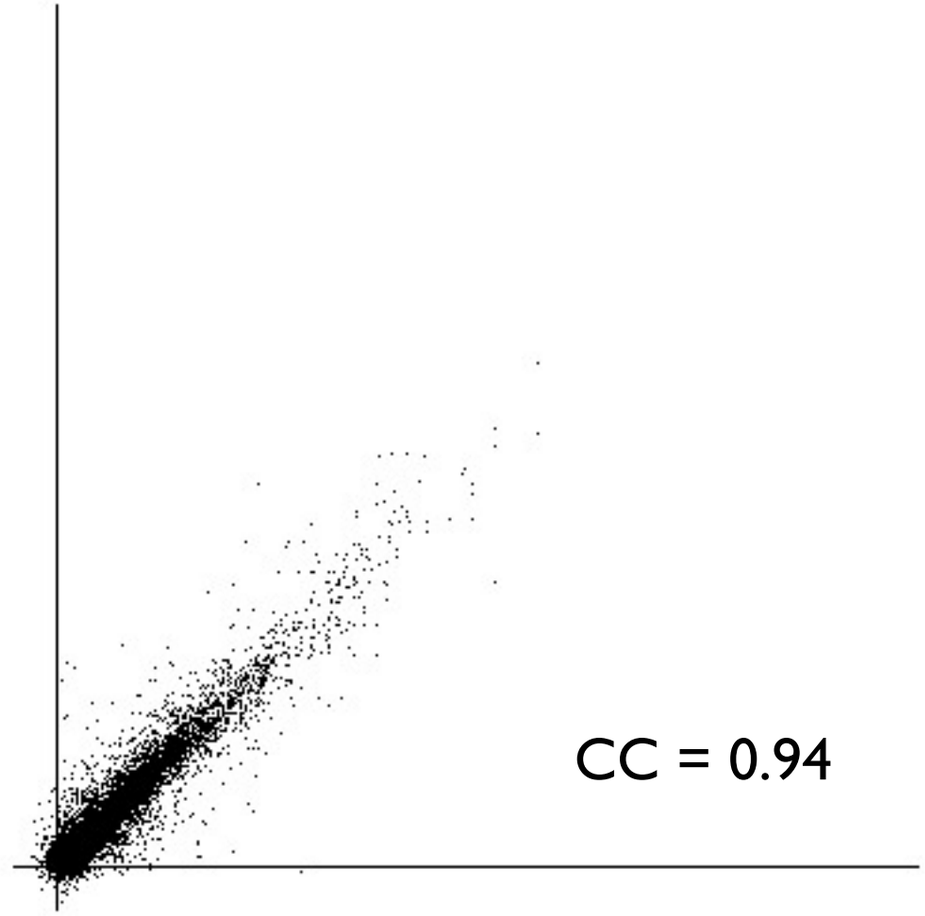
For equal axes, the correlation coefficient (CC) is the slope of the "best" (least-squares) straight line through the scatter plot

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

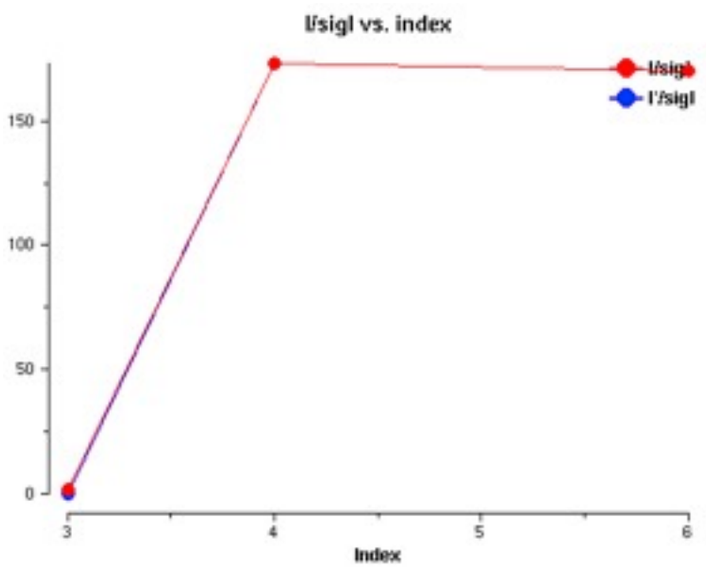
	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
= 1	C m m m	***	0.989	9.45	9.62	0.17	0.96	0.02	0.08	0.76	0.0	[h,k,l]
2	P 1 2/m 1		0.004	7.22	9.68	2.46	0.97	0.25	0.06	0.56	0.0	[-1/2h+1/2k,-l,-1/2h-1/2k]
3	C 1 2/m 1		0.003	7.11	9.61	2.50	0.96	0.25	0.08	0.55	0.0	[h,k,l]
4	C 1 2/m 1		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

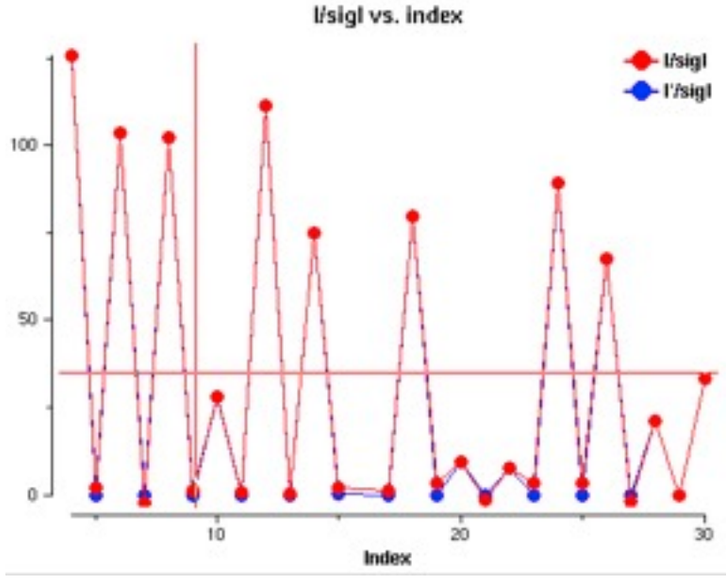
Zone	Number	PeakHeight	SD	Probability	ReflectionCondition
Zones for Laue group P m m m					
1 screw axis 2(1) [a]	3	1.000	0.296	** 0.889	h00: h=2n
2 screw axis 2(1) [b]	26	1.000	0.142	*** 0.971	0k0: k=2n
3 screw axis 2(1) [c]	46	0.997	0.097	*** 0.986	00l: l=2n

Fourier analysis of I/σ(I)

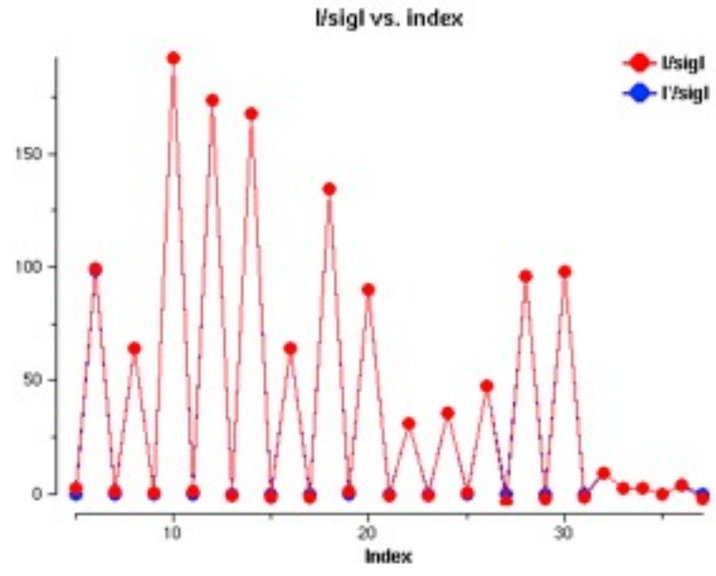
There are indications of 2_1 screw symmetry along all principle axes
(though note there are only 3 observations on the *a* axis (h00 reflections))



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!)

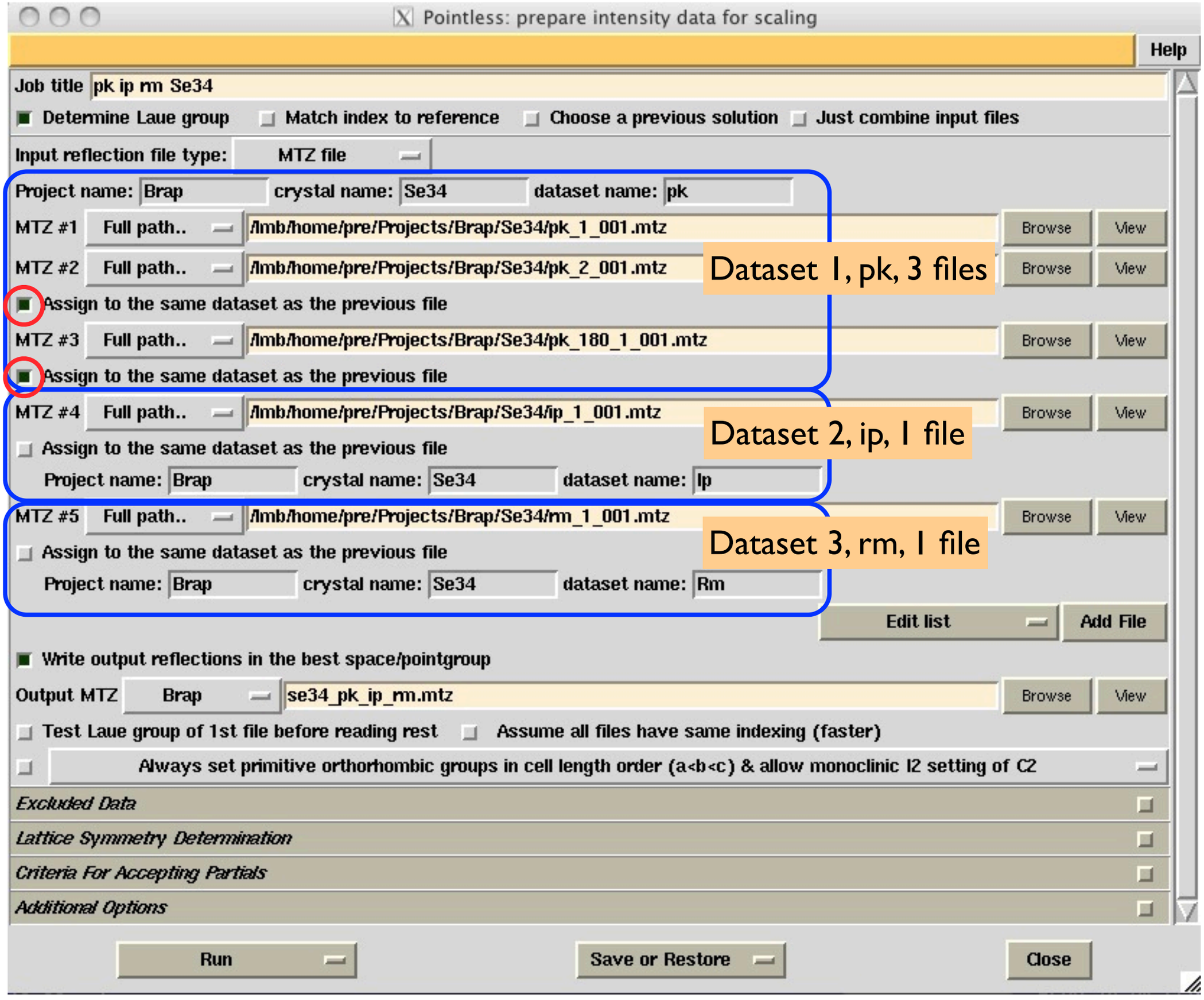
POINTLESS works (usually) with unscaled data (hence use of correlation coefficients), so data with a large range of scales, including a dead crystal, may give a too-low symmetry.
In bad cases either just use the first part of the data, or scale in P1 and run POINTLESS on the scaled unmerged data

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

Combining multiple files (and multiple MAD datasets)

3 files assigned to same dataset



Combining multiple files (and multiple MAD datasets)

Alternative index test relative to first file

Alternative reindexing	CC	R(E^2)	Number	Cell_deviation
[h,k,l]	0.965	0.086	23592	0.00
[-k,h,l]	0.789	0.205	22755	0.30
[l,k,-h]	0.102	0.438	21060	0.76
[k,l,h]	0.055	0.459	22714	0.66
[-h,l,k]	0.048	0.461	23282	0.46
[l,h,k]	0.043	0.457	21194	0.66

Alternative index test relative to files so far

Alternative reindexing	CC	R(E^2)	Number	Cell_deviation
[h,k,l]	0.933	0.124	40670	0.14
[-k,h,l]	0.610	0.283	40494	0.43
[l,k,-h]	0.061	0.463	40338	0.84
[-h,l,k]	0.045	0.470	40635	0.43
[l,h,k]	0.027	0.477	40352	0.68
[k,l,h]	0.020	0.479	40461	0.77

Alternative index test relative to files so far

Alternative reindexing	CC	R(E^2)	Number	Cell_deviation
[h,k,l]	0.933	0.124	40670	0.14
[-k,h,l]	0.610	0.283	40494	0.43
[l,k,-h]	0.061	0.463	40338	0.84
[-h,l,k]	0.045	0.470	40635	0.43
[l,h,k]	0.027	0.477	40352	0.68
[k,l,h]	0.020	0.479	40461	0.77

Alternative index test relative to files so far

Alternative reindexing	CC	R(E^2)	Number	Cell_deviation
[h,k,l]	0.960	0.095	22712	0.07
[-k,h,l]	0.706	0.241	22712	0.36
[l,k,-h]	0.084	0.455	22690	0.80
[k,l,h]	0.050	0.468	22698	0.67
[-h,l,k]	0.046	0.465	22701	0.44
[l,h,k]	0.025	0.472	22693	0.72

Alternative indexing relative to first file(s):

	Reindex operator	CC	File name
2	[h,k,l]	0.965	pk_2_001.mtz
3	[h,k,l]	0.933	pk_180_1_001.mtz
4	[h,k,l]	0.960	ip_1_001.mtz
5	[h,k,l]	0.958	rm_1_001.mtz

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

Alternative indexing

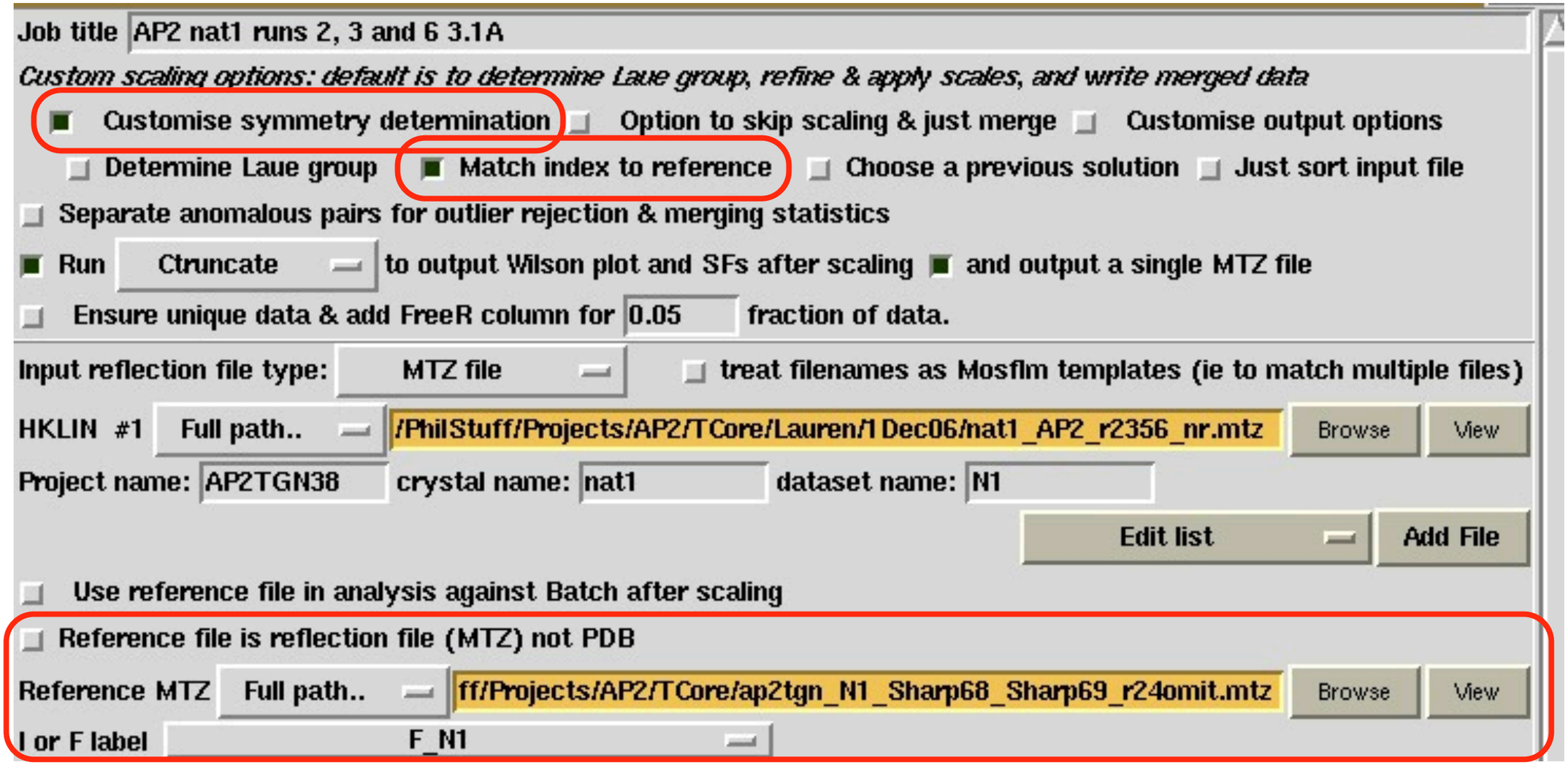
If the true point group is lower symmetry than the lattice group, alternative valid but non-equivalent indexing schemes are possible, related by symmetry operators present in lattice group but not in point group *(note that these are also the cases where merohedral twinning is possible)*

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

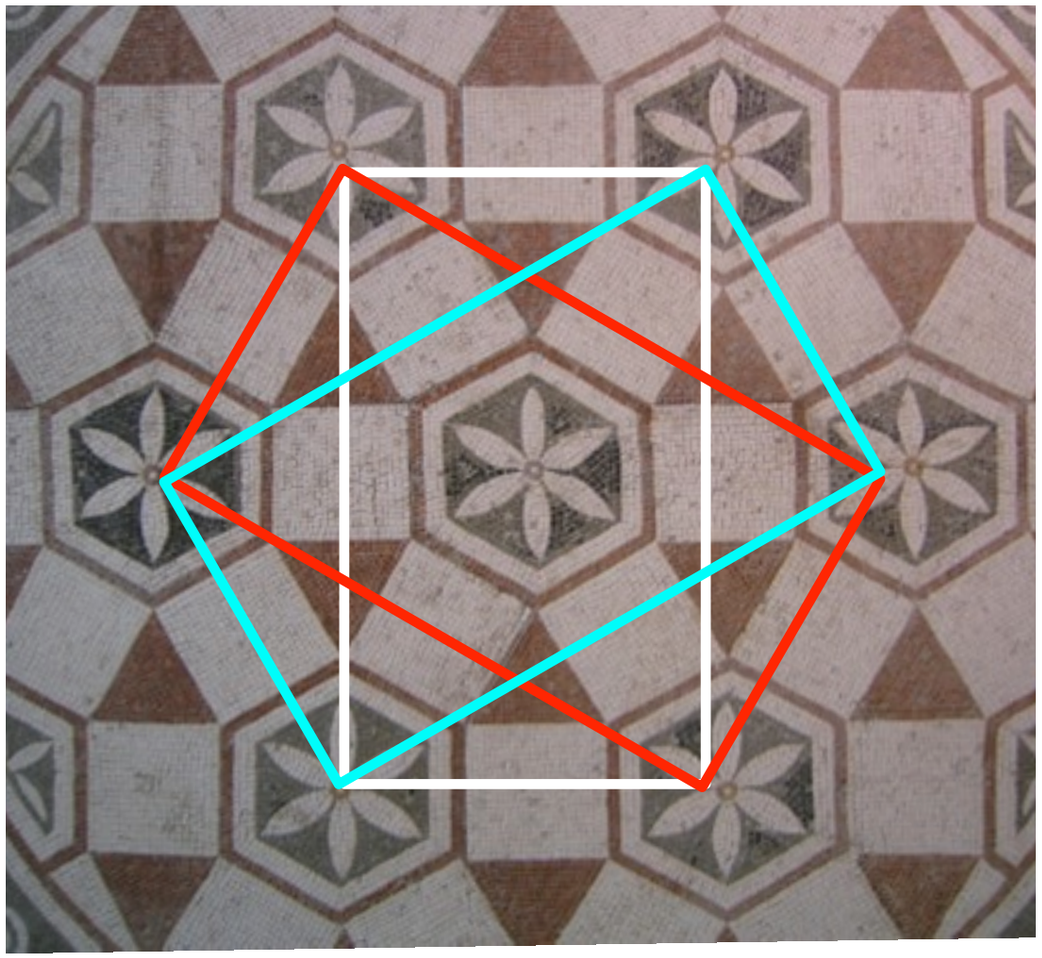


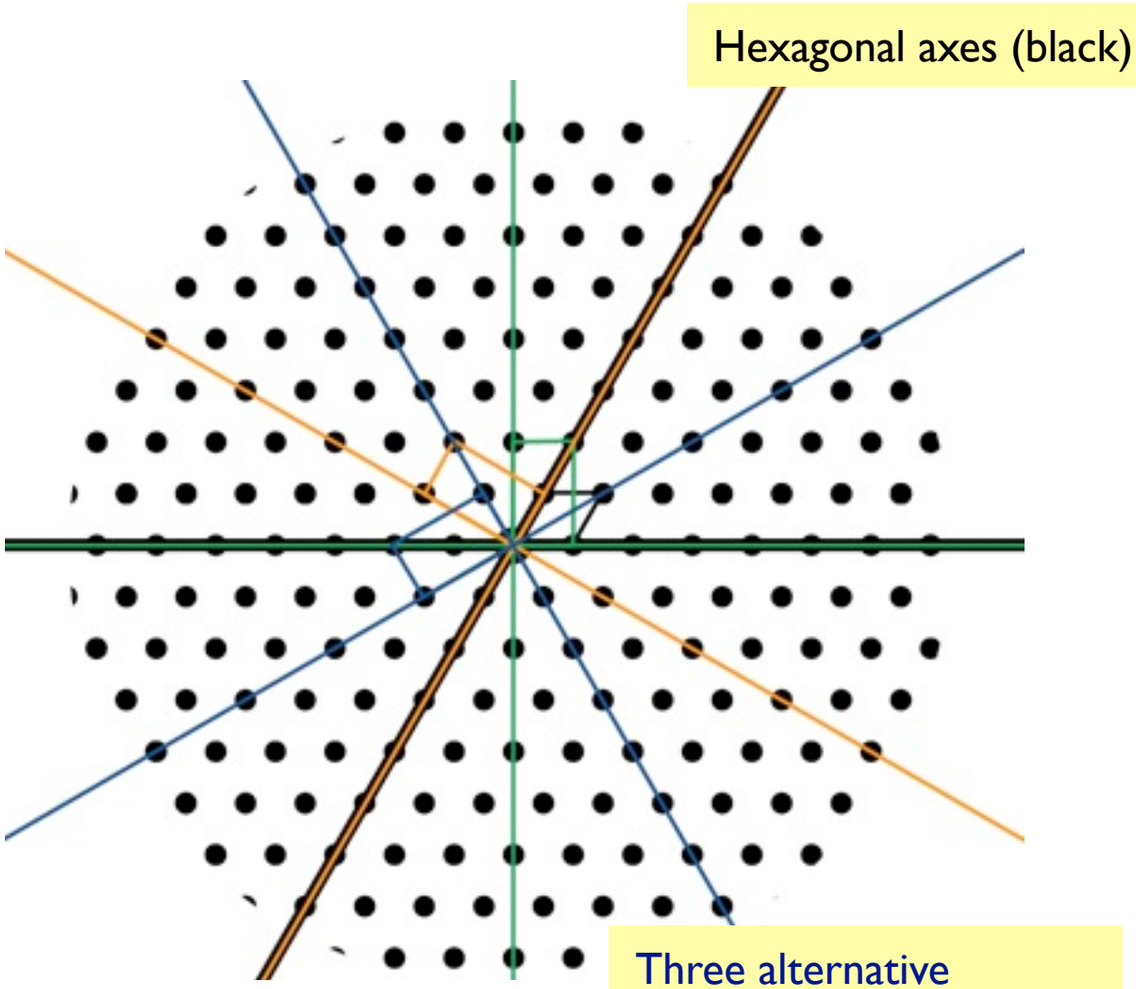
A confusing case in C222:

Unit cell 74.72 129.22 184.25 90 90 90

This has $b \approx \sqrt{3}\, a$ 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





Three alternative
C-centred orthorhombic
Lattices (coloured)

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

Scaling and Data Quality

Data flow scheme in CCP4

Intensities
from Mosflm
h k l

Intensities
from Mosflm
h k l

Intensities
from Mosflm (or XDS)
h k l I σ(I) etc

One or more files from integration program
One or more datasets (eg MAD)

Note that this does not work well for very large numbers of crystals

POINTLESS

Determine point-group
(& space group)
and consistent indexing

Sorted Intensities in "best" space group

AIMLESS/SCALA

Scale symmetry-related intensities together
Produce statistics on data quality

Scaled and averaged intensities

Estimate |F| from I
detect twinning (intensity statistics)

CTRUNCATE

Complete sphere of reflections
Generate or copy freeR flags

UNIQUIFY etc

h k l F σ(F) I σ(I)
FreeR_flag

These programs may ultimately be combined into one

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

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*

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 detector) avoids synchronisation errors (though variation in rotation speed could still cause trouble, and there is a dead time during readout)

Factors related to crystal and diffracted beam

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

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

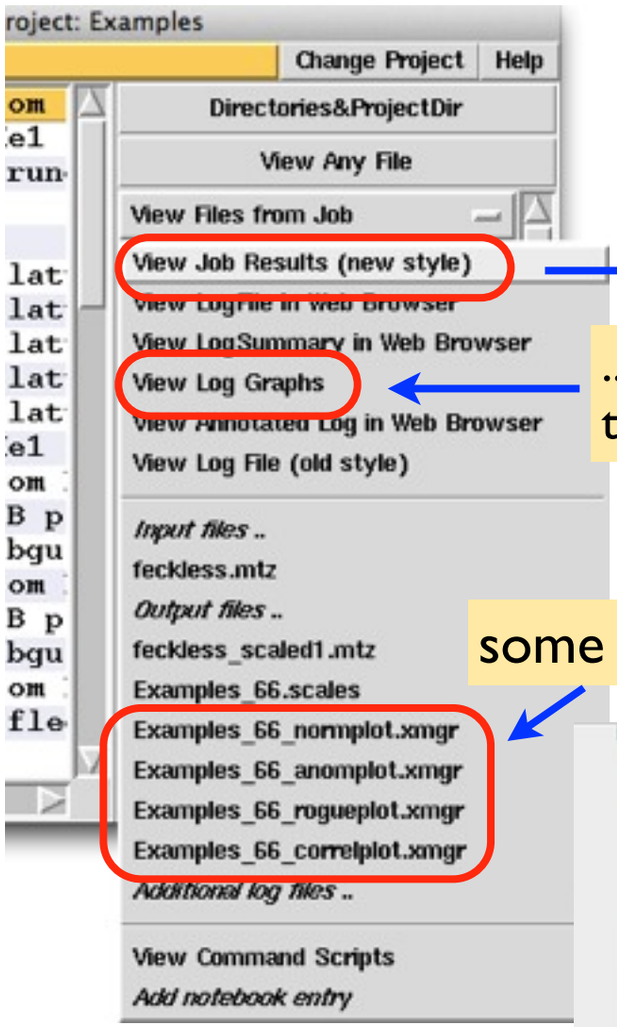
Maybe extrapolate back to zero time? (but this needs high multiplicity)

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

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

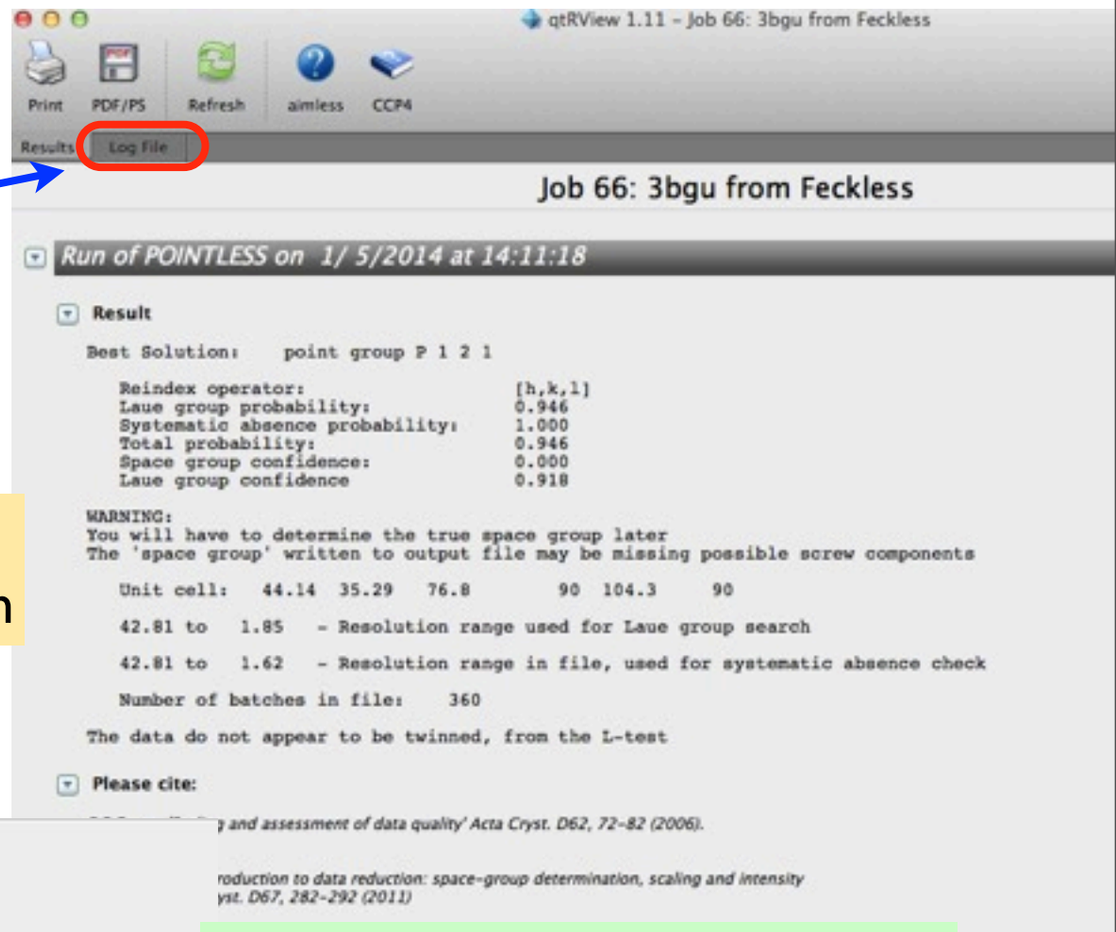


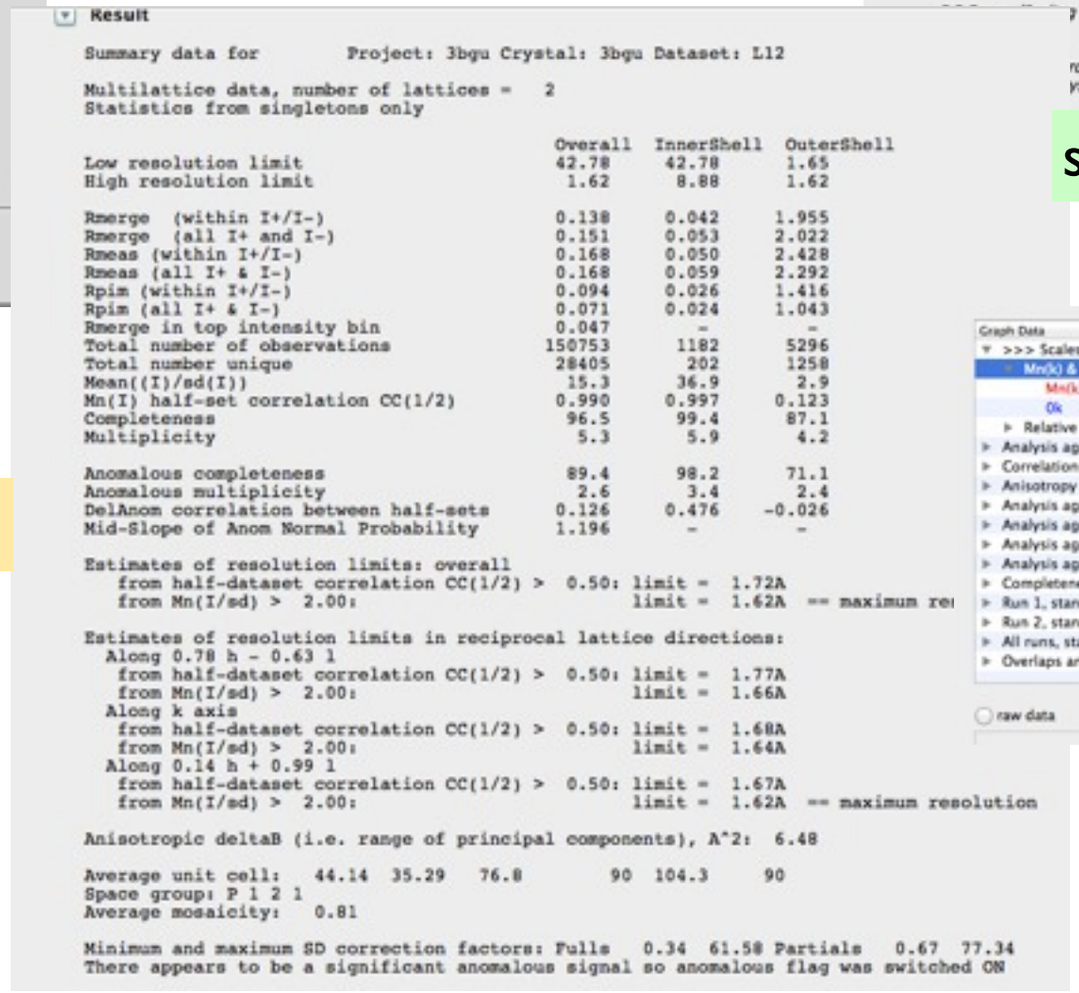
view log file

... or just the graphs

some other plots

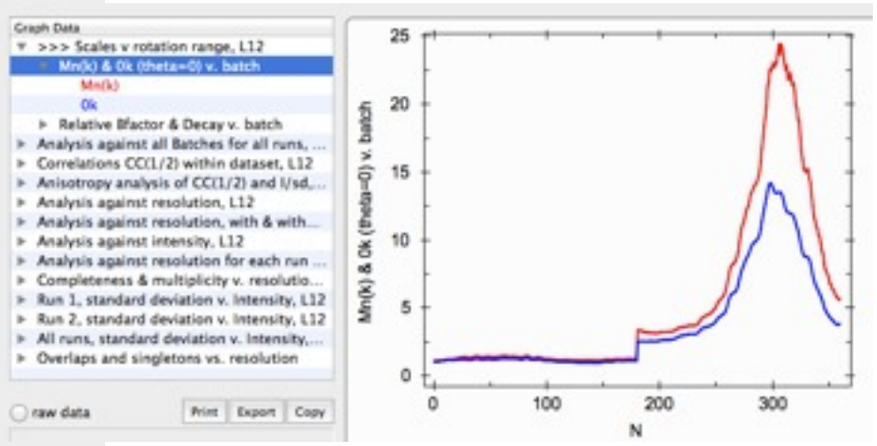
output from each program





"Table 1"

same output from QuickScale



graphs

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 $\sigma(I)$ 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*

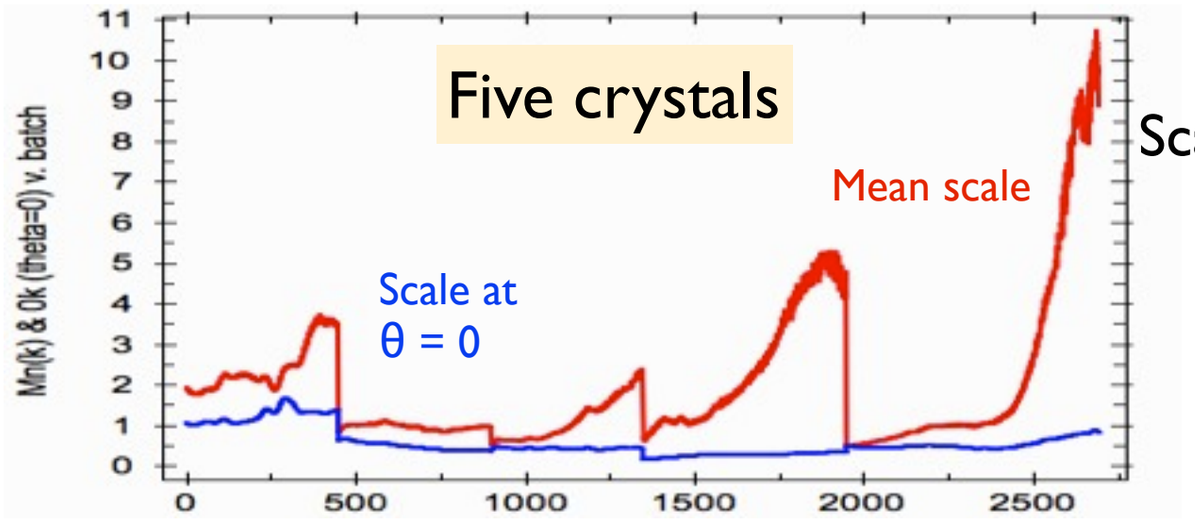
Half-dataset correlation coefficient:

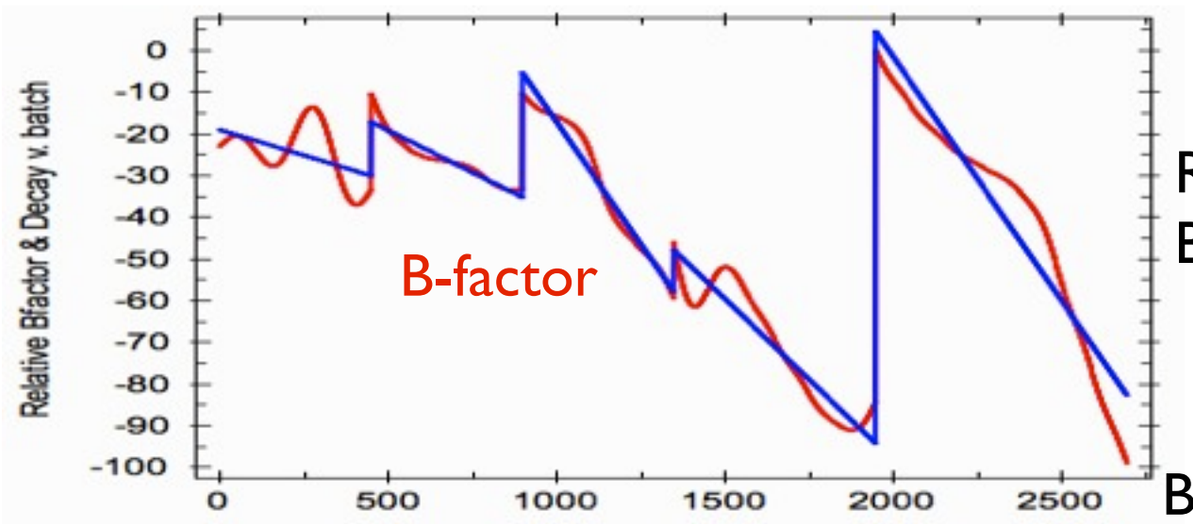
Split observations for each reflection data randomly into 2 halves, and calculate the correlation coefficient between them

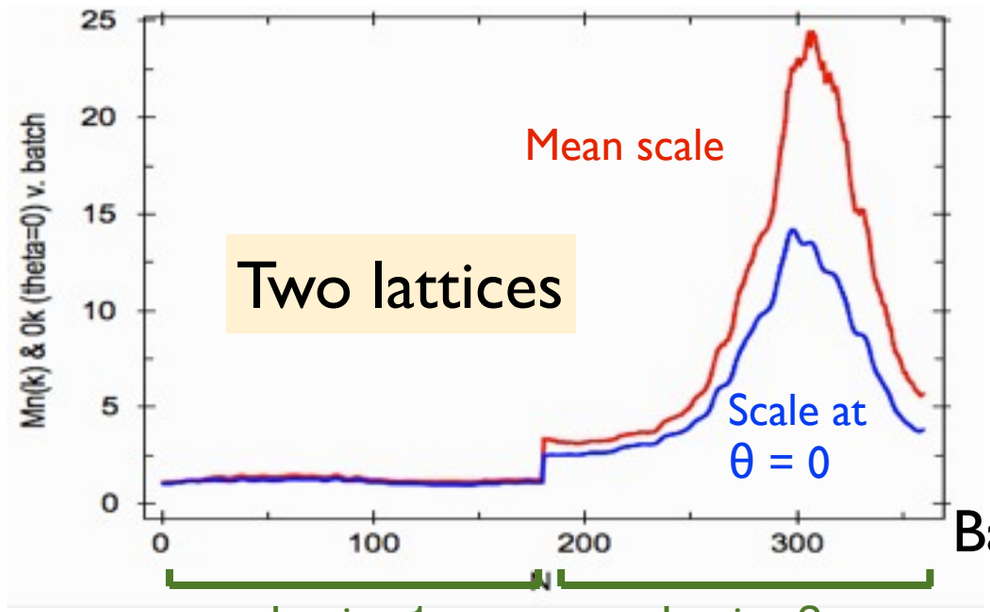
What should you look at?

Analyses as a function of "batch" (ie image number)

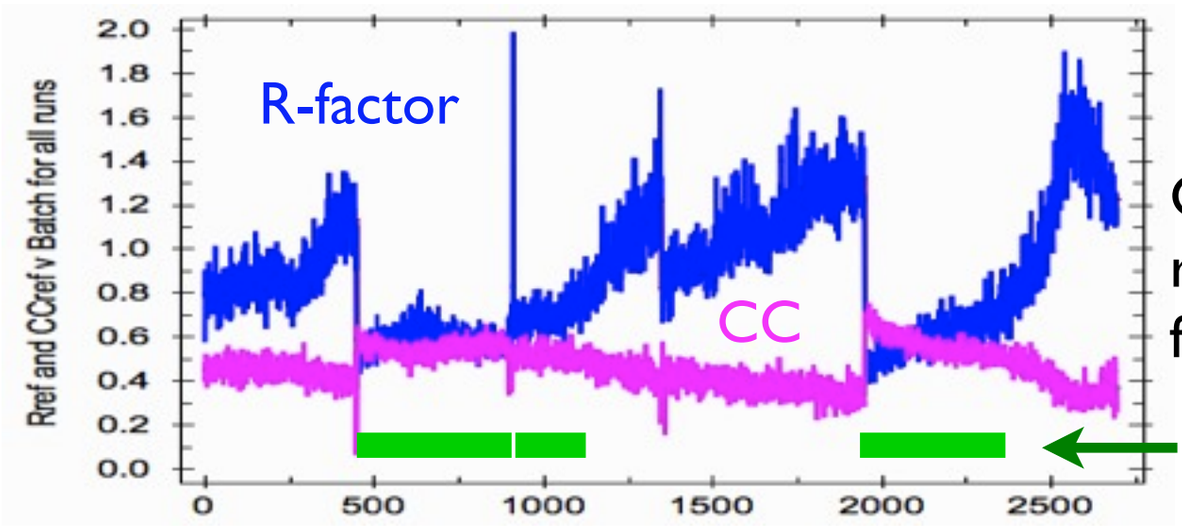
Look at :
- scales
- relative B-factor (overall radiation damage)
- cumulative completeness
- [comparison to reference]





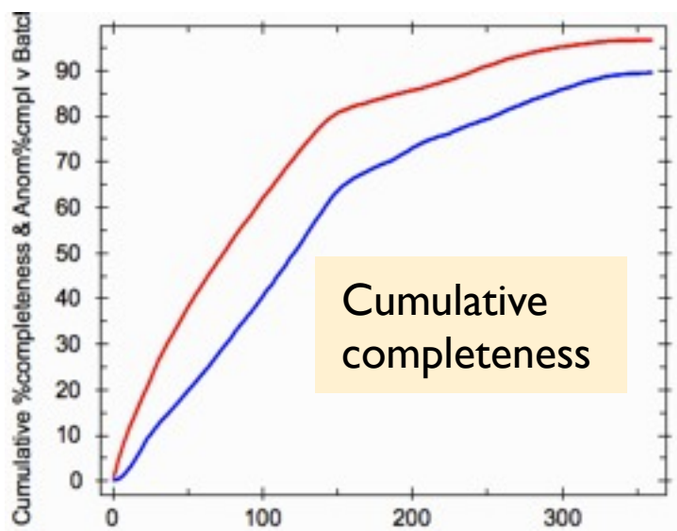


Lattice 2 is much weaker in the middle



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?

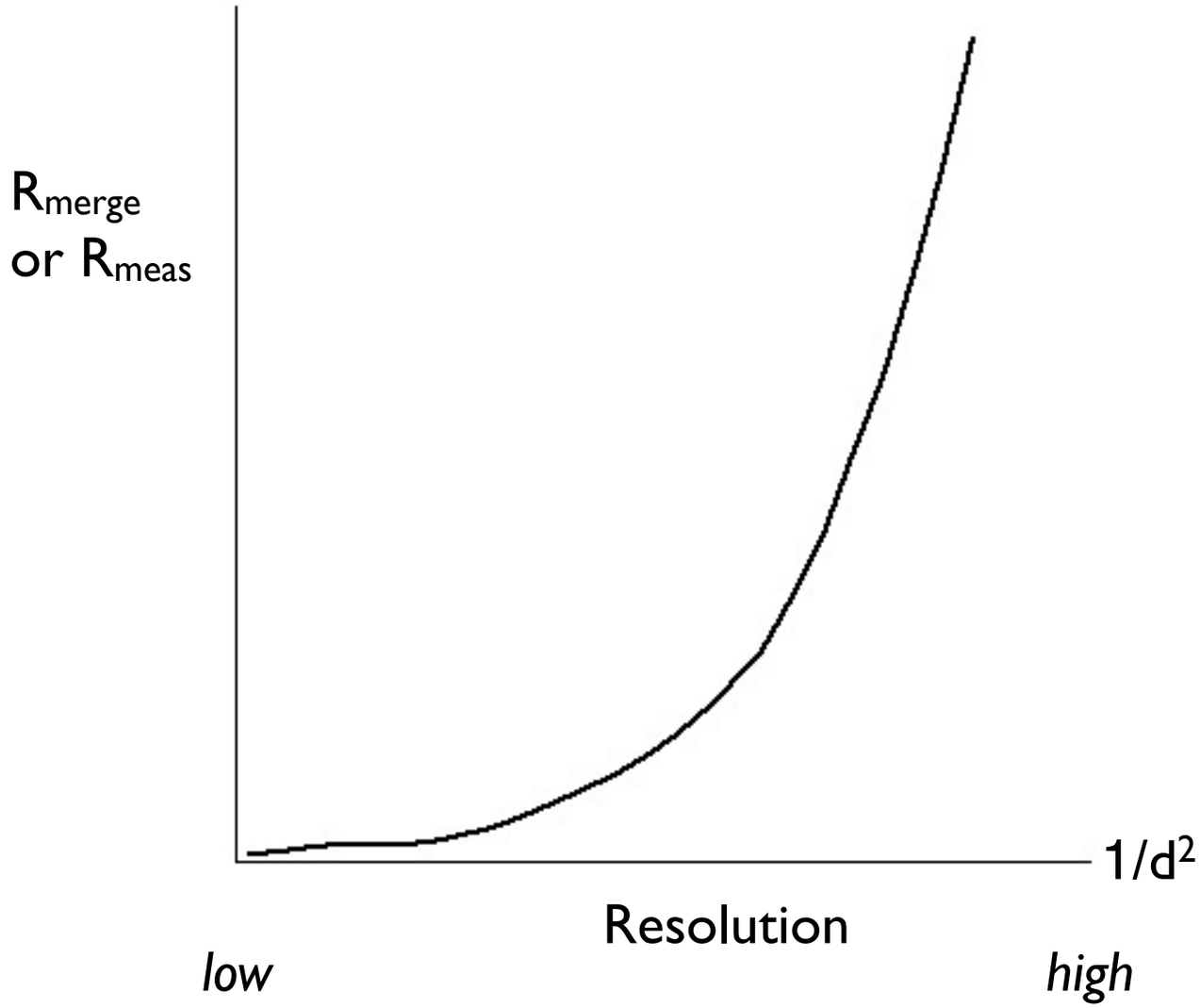
Resolution is a contentious issue, often with referees, eg:

> "The crystallographic Rmerge and Rmeas values are not acceptable, although, surprisingly the I/sigmaI and R-factors for these shells are OK. **I cannot ever recall seeing such high values of Rmerge - 148% !** The discrepancy between the poor R_{merge} values and vs. the other statistics is highly unusual- generally a high R_{merge} corresponds a low I/sigma-I would have expected a value closer to 1.0 than 2.0 here - and no explanation is offered. The authors may want to decrease the claimed resolution such that **acceptable** values are obtained for all metrics."

> "The crystallographic structure determination is reported to be at 2.7Å resolution. However, the data statistics presented in table S1 show <I>/<<sigma>> =1.2 in the highest resolution shell and thus represent marginally significant reliability. **Consequently, the resolution may be overstated**."

What scores can we use?

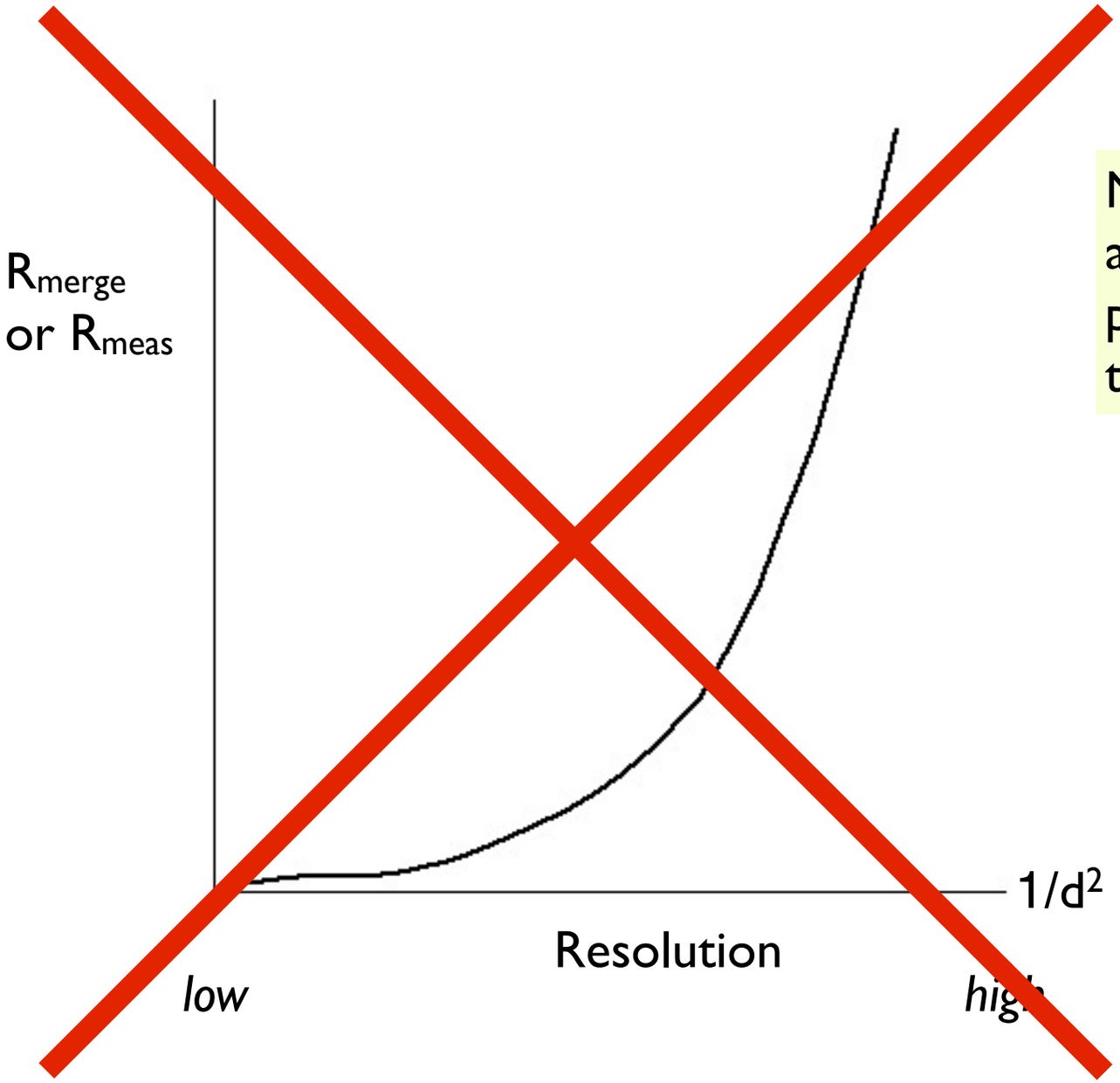
What about R-factors?



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

What about R-factors?

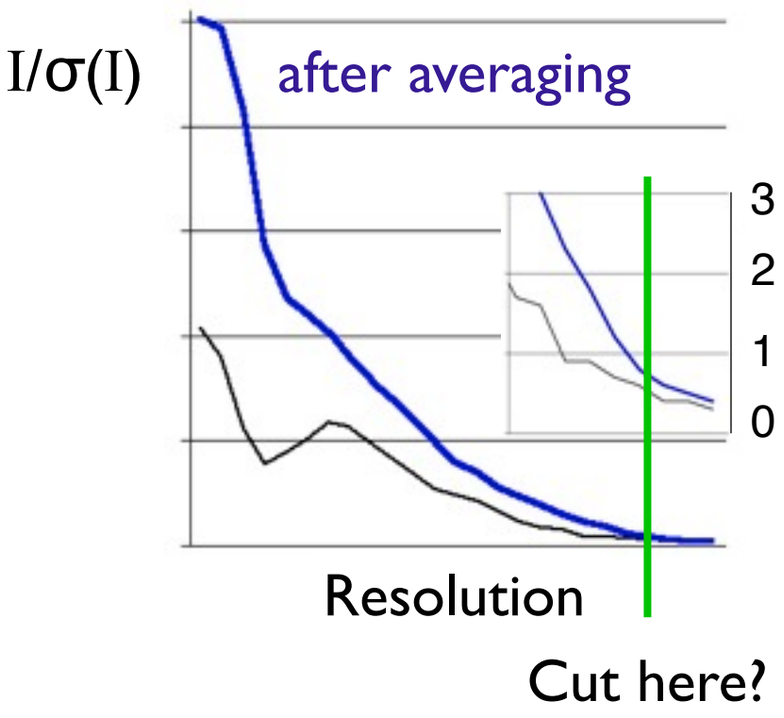


Note that R_{merge} and R_{meas} are useful for other purposes, but not for deciding the resolution cutoff

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>



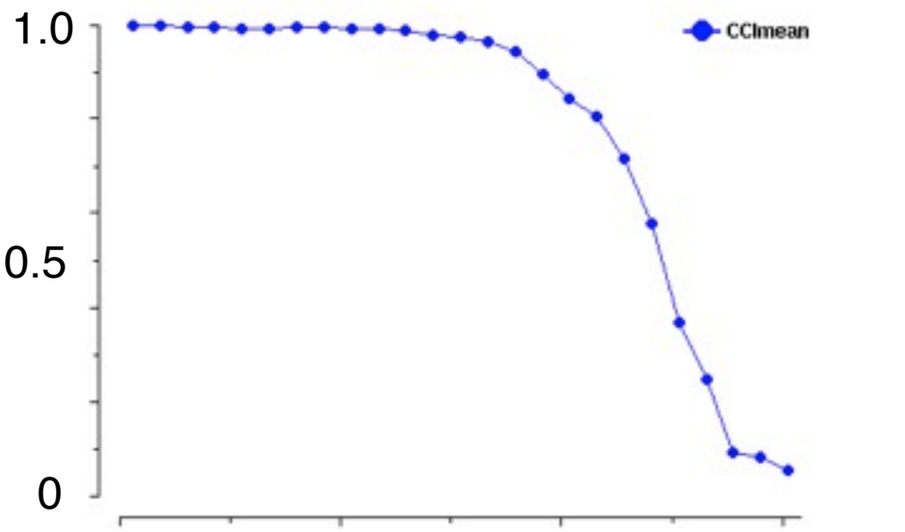
A reasonably good criterion, but it relies on σ(I), which is not entirely reliable

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



Advantages:

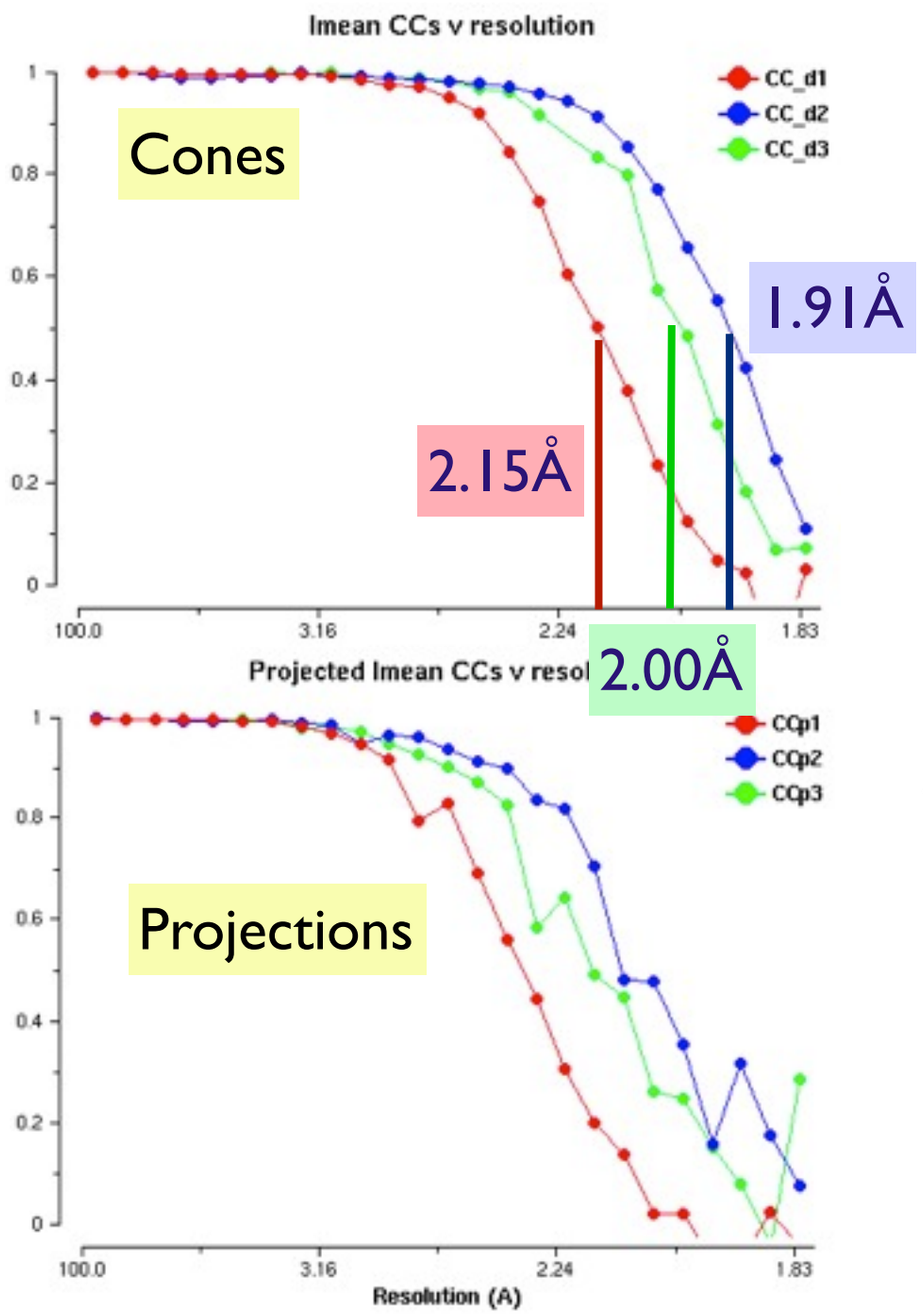
- Clear meaning to values (1.0 is perfect, 0 is no correlation) , known statistical properties
- Independent of σ(I)

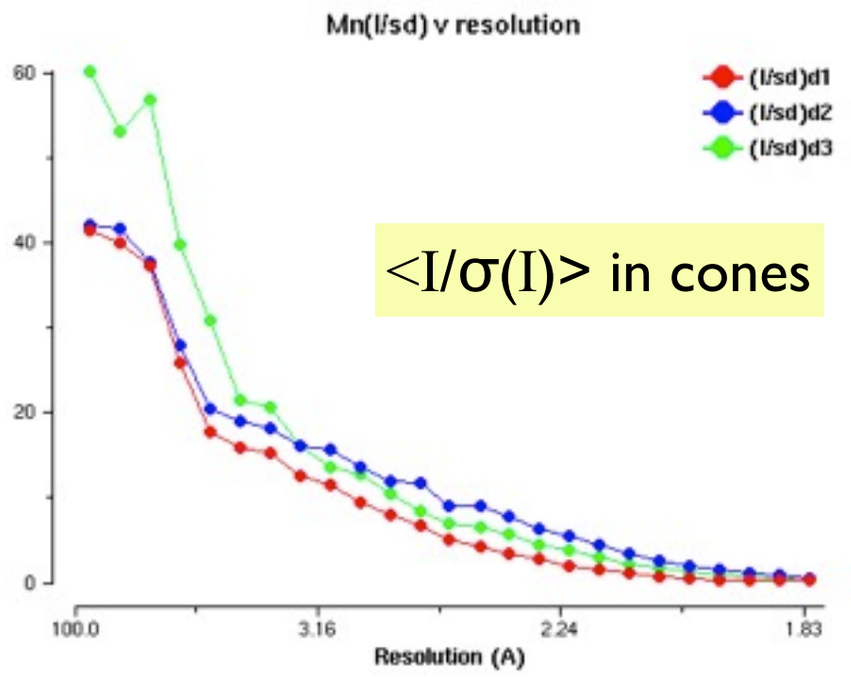
cut resolution at CC ~= 0.3 – 0.5

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

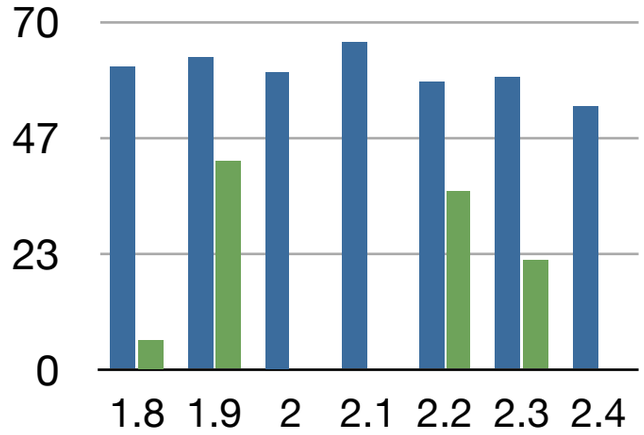
- Molecular replacement: Phaser uses likelihood weighting, but there is probably no gain in using the very weak high resolution data

- 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!

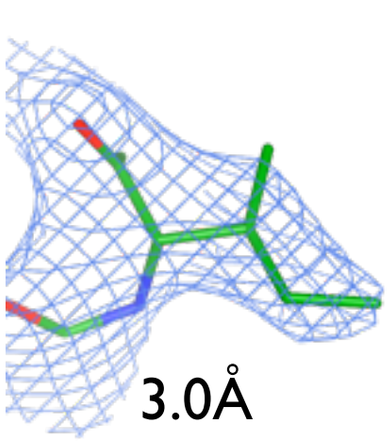
Future developments may improve treatment of weak noisy data

Example:
3 molecules/asu, omit 22/276 residues from each molecule, model build with Arp/warp at different resolutions



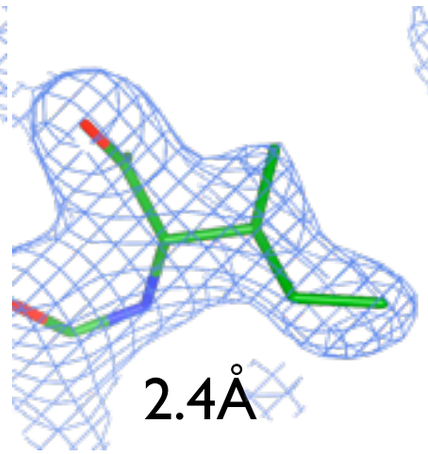
Number of residues built and sequenced

figures made with ccp4mg



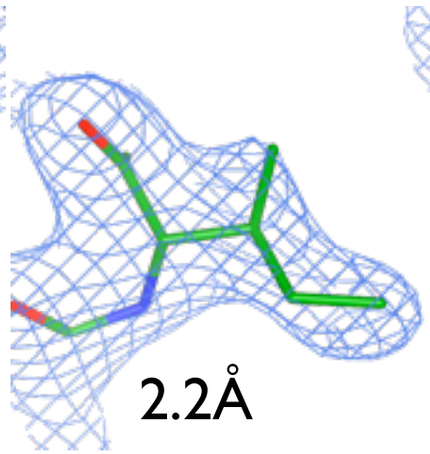
3.0Å

R_{free} 0.294



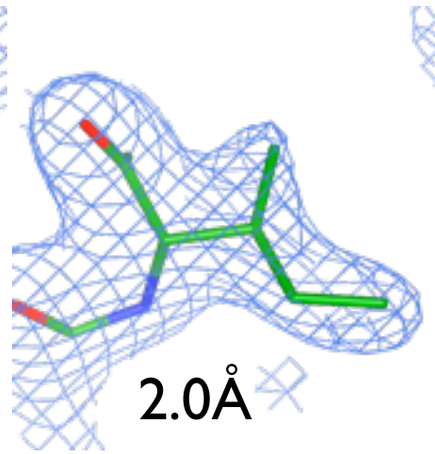
2.4Å

R_{free} 0.282



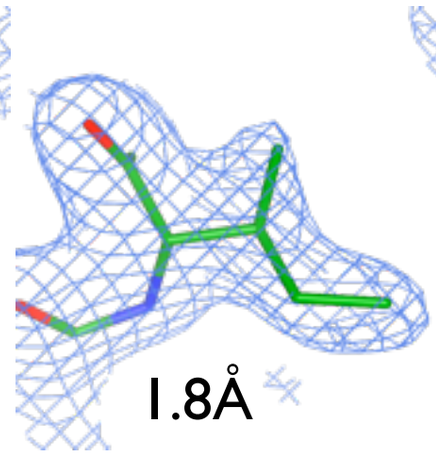
2.2Å

R_{free} 0.284



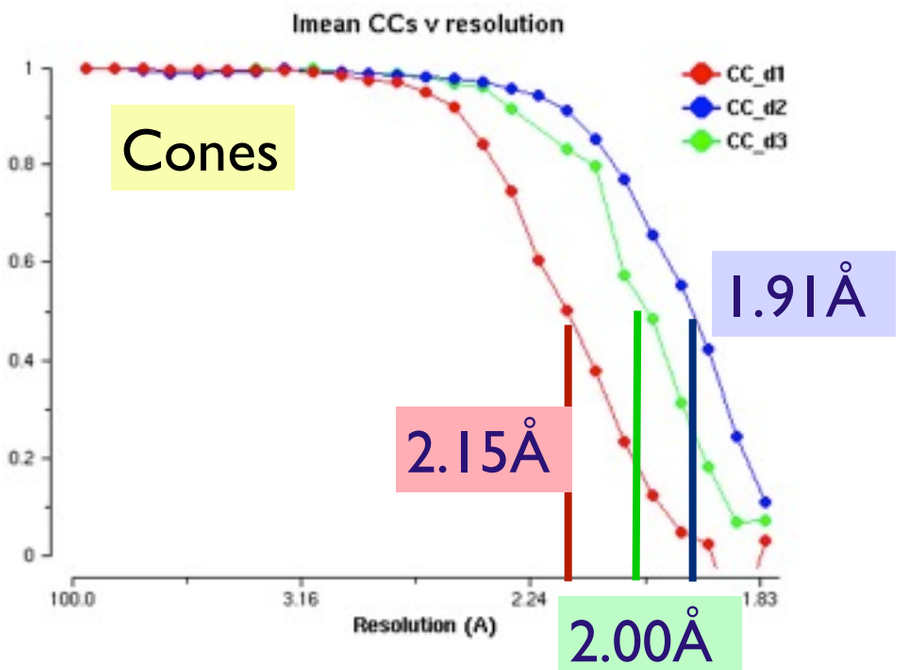
2.0Å

R_{free} 0.285

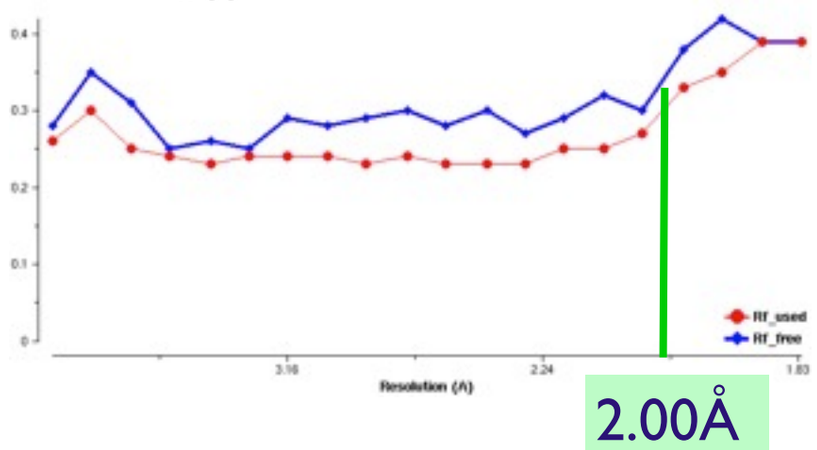


1.8Å

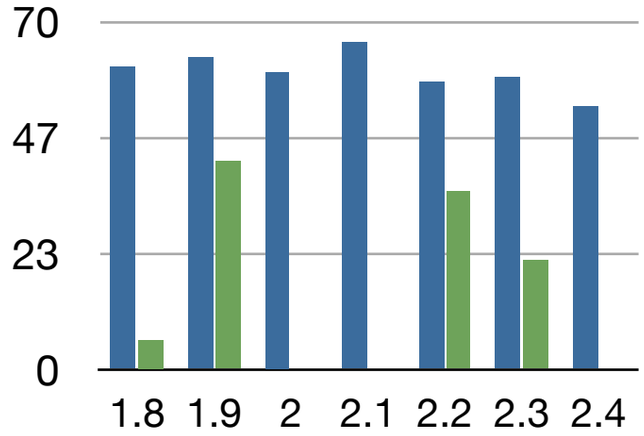
R_{free} 0.290



R & R_{free} after initial refinement

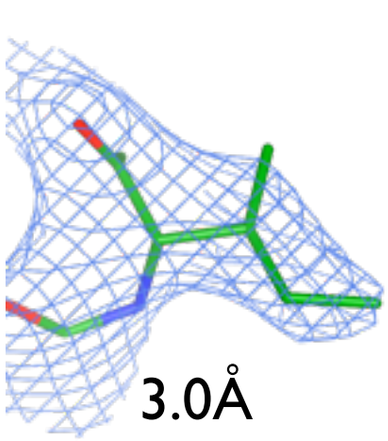


Example:
3 molecules/asu, omit 22/276 residues from each molecule, model build with Arp/warp at different resolutions



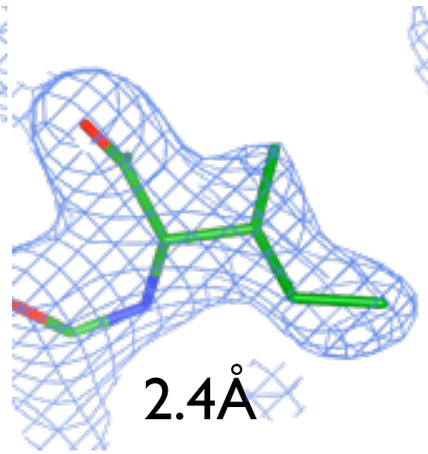
Number of residues built and sequenced

figures made with ccp4mg



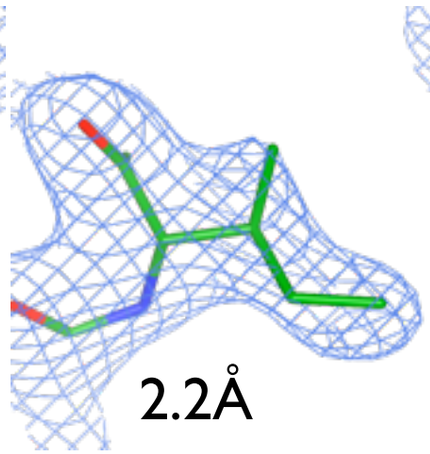
3.0Å

R_{free} 0.294



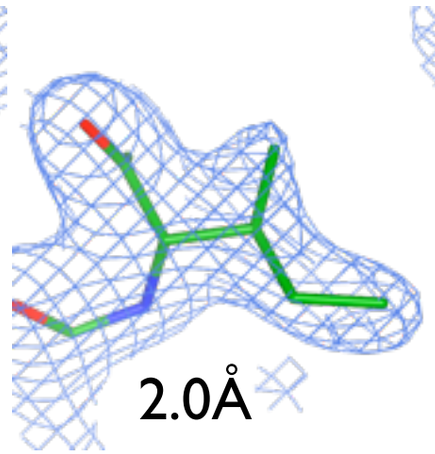
2.4Å

R_{free} 0.282



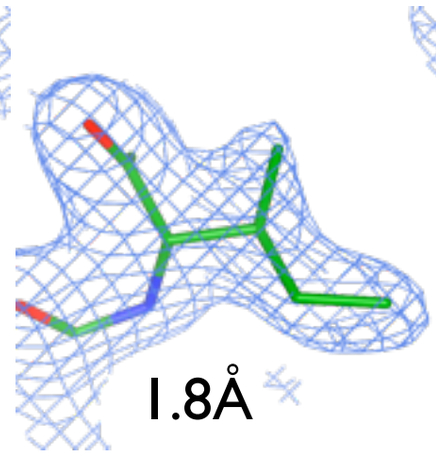
2.2Å

R_{free} 0.284



2.0Å

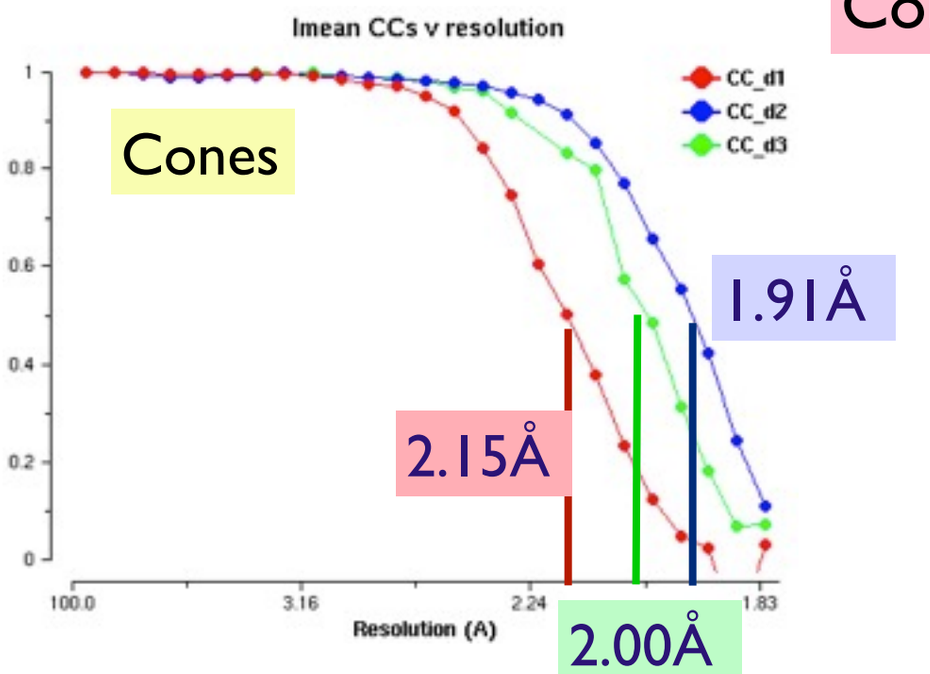
R_{free} 0.285



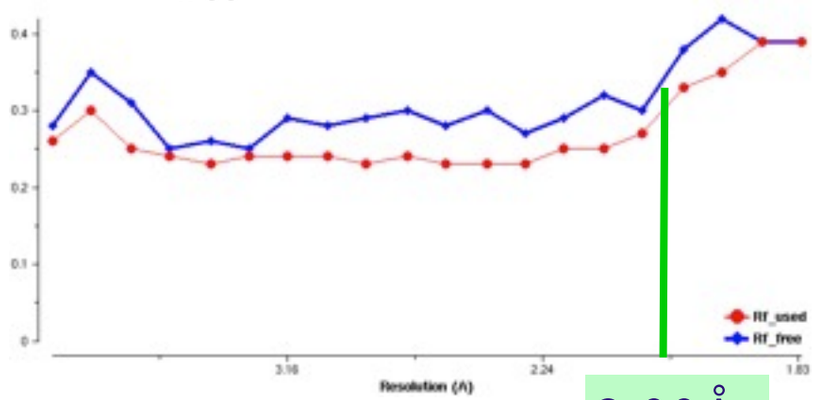
1.8Å

R_{free} 0.290

Conclusion: there is not a huge difference

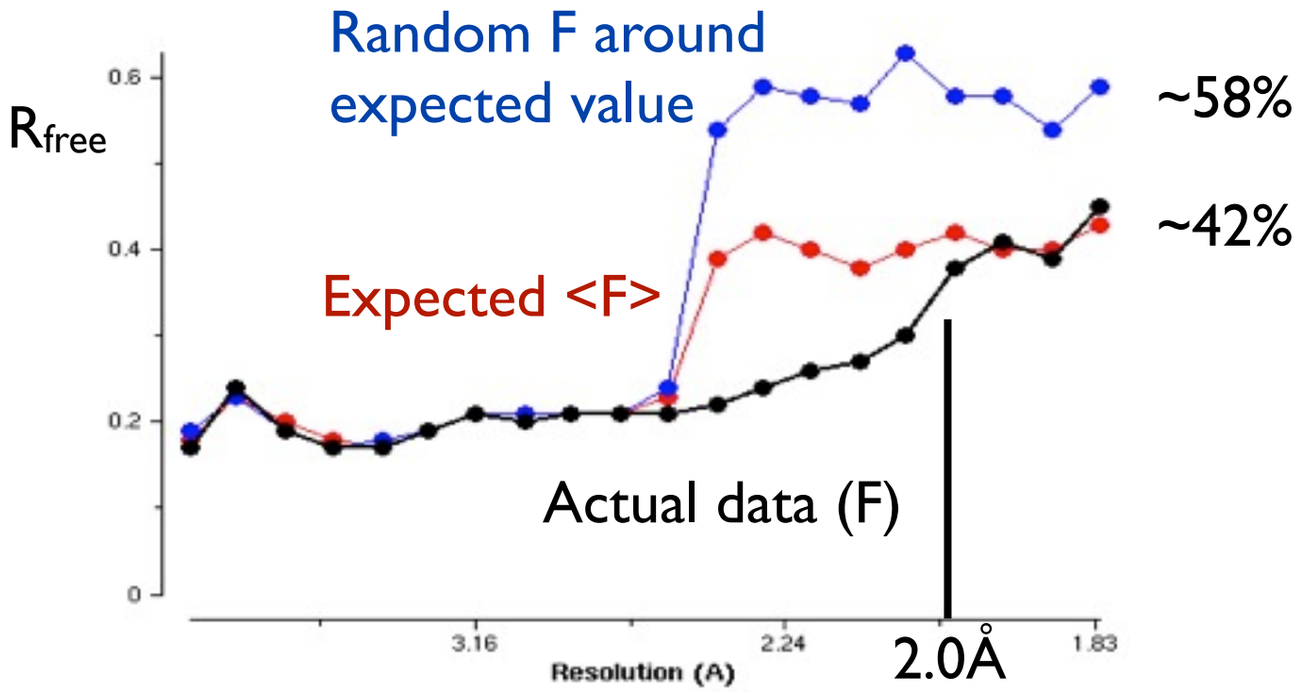


R & R_{free} after initial refinement

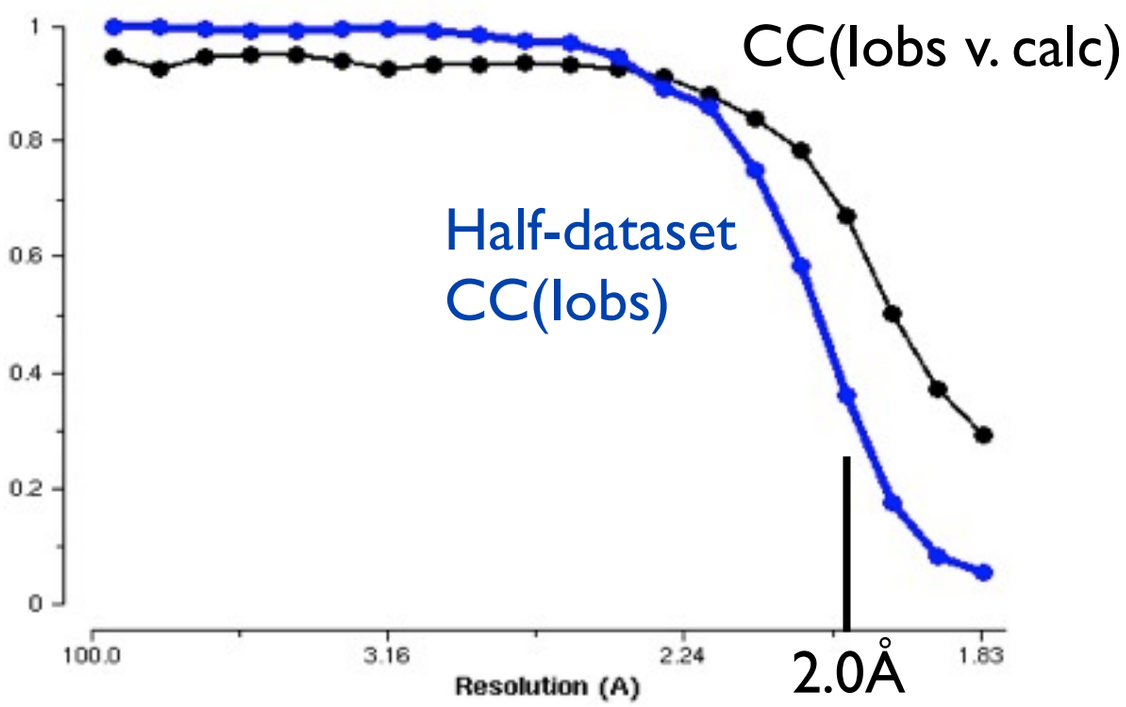


Example continued: refinement against real data or simulated data

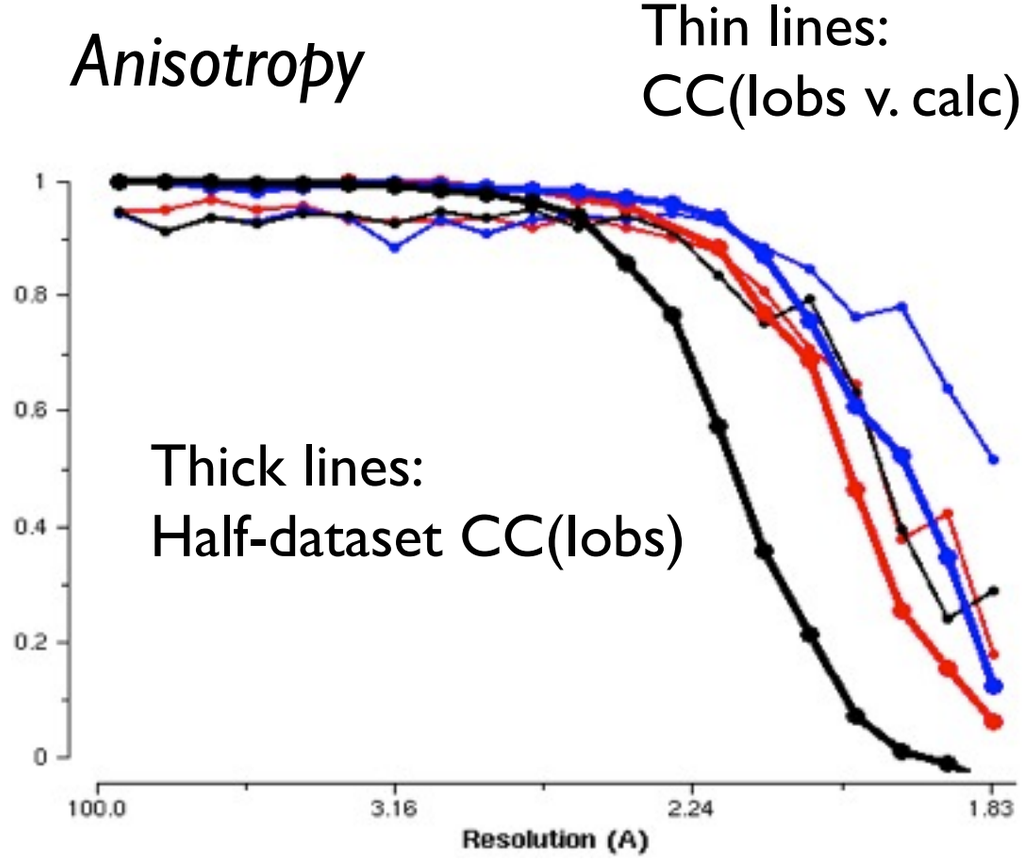
thanks to Garib Murshudov



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

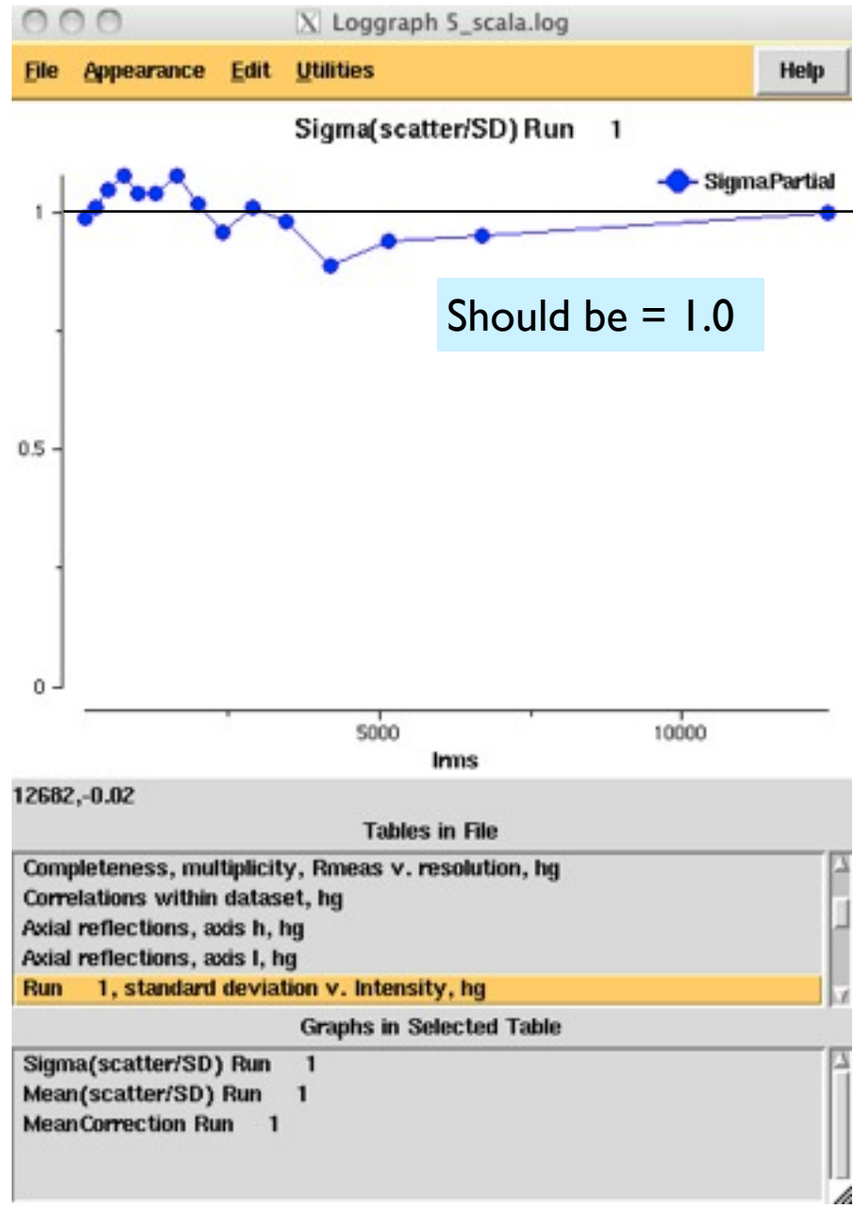


Improved estimate of σ(I)

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



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 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

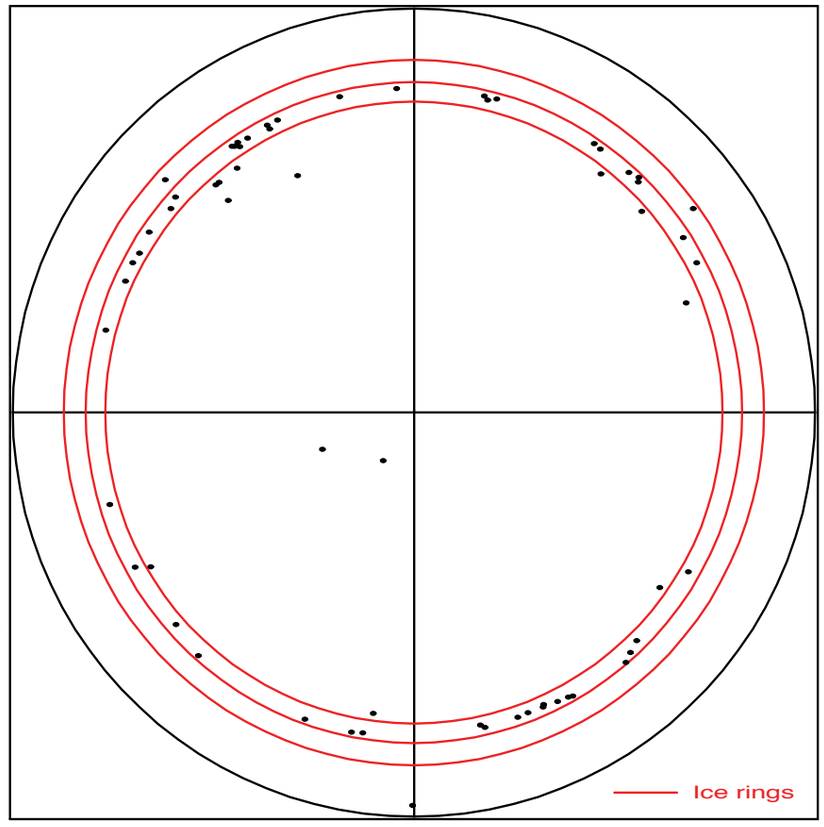
- zingers
- bad prediction (spot not there)

  improve prediction

- spot overlap

  lower mosaicity, smaller slice, move detector back

  deconvolute overlaps

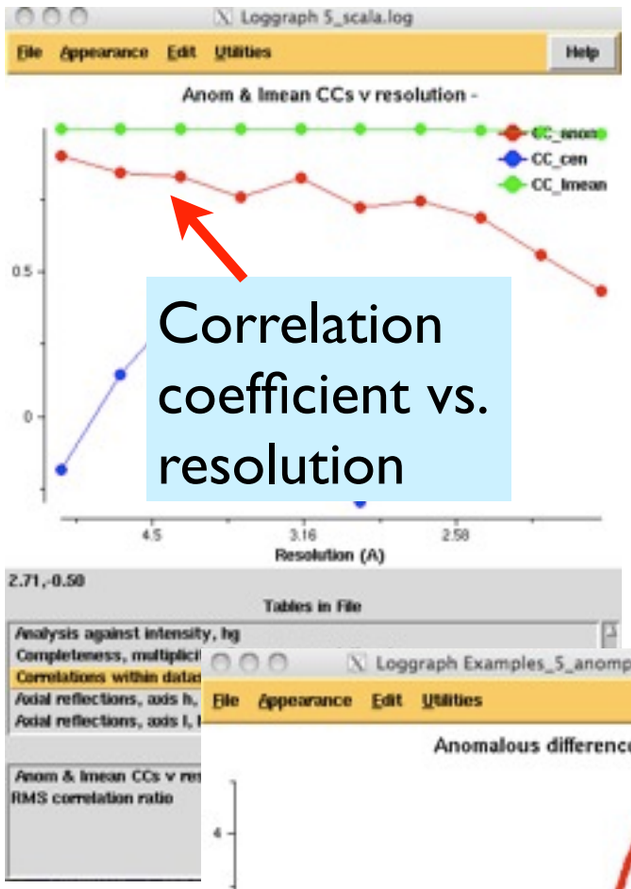


Rejects lie on ice rings (red) (ROGUEPLOT in Scala)

Position of rejects on detector

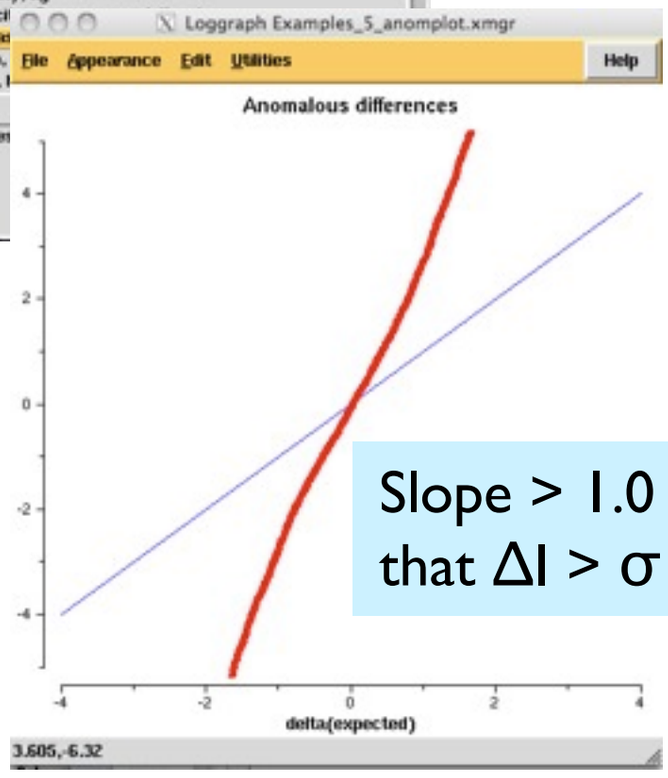
Detecting anomalous signals

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



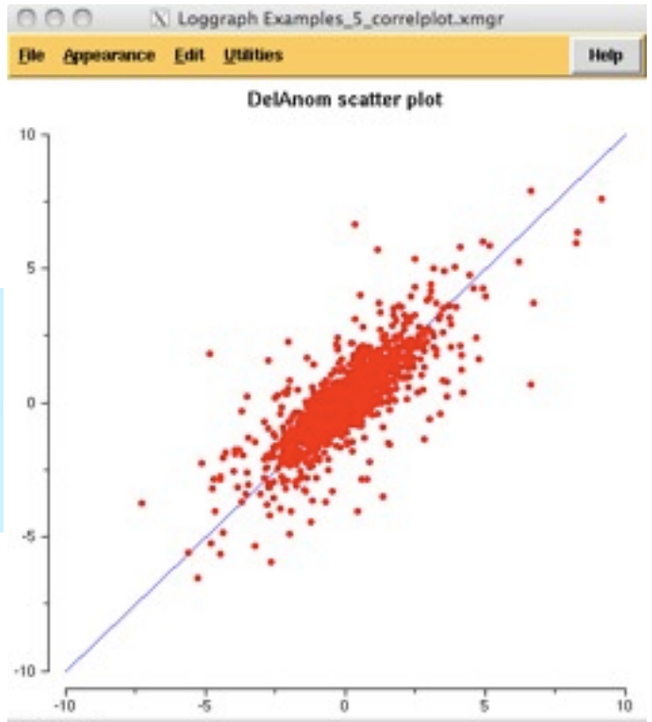
Correlation coefficient vs. resolution

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

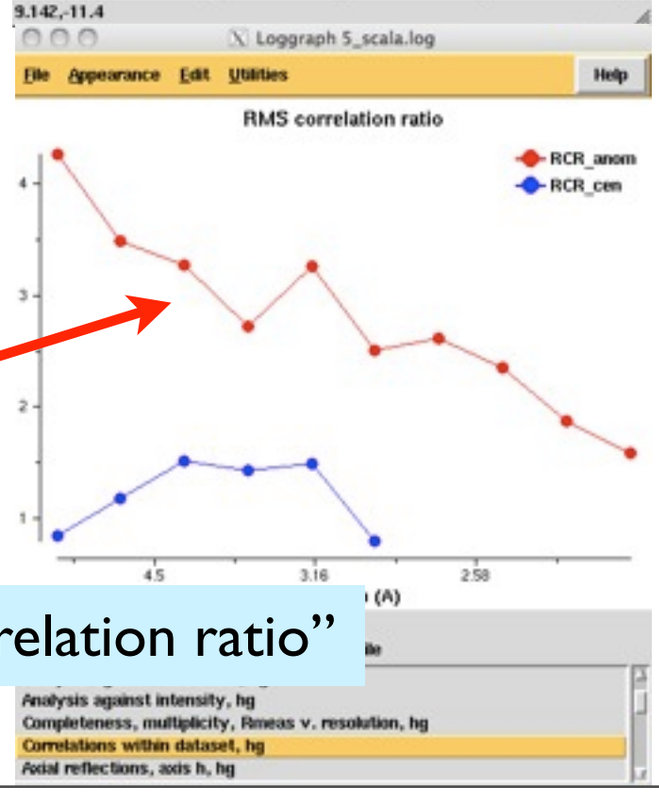


Slope > 1.0 means that $\Delta I > \sigma$

Strong anomalous signal



Plot ΔI_1 against ΔI_2 should be elongated along diagonal

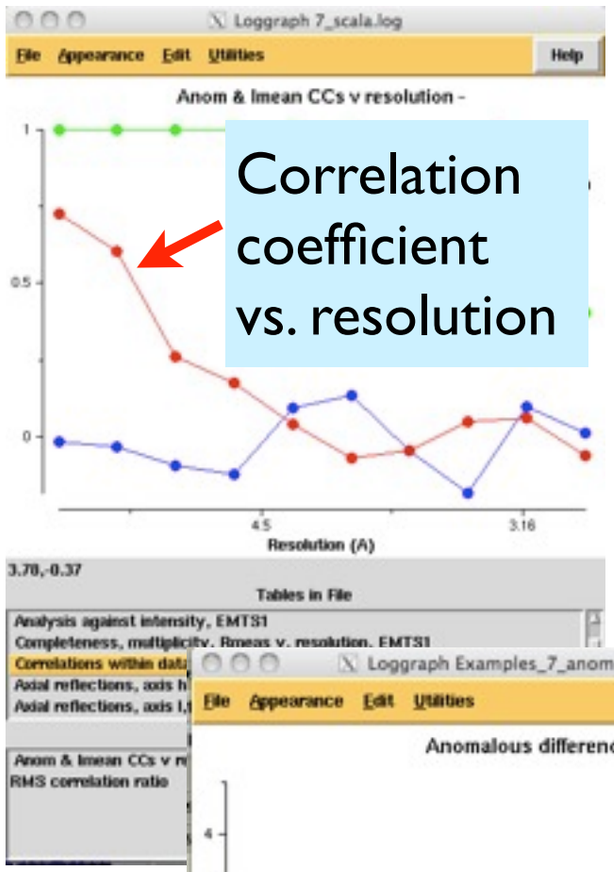


Ratio of width of distribution along diagonal to width across diagonal

"RMS correlation ratio"

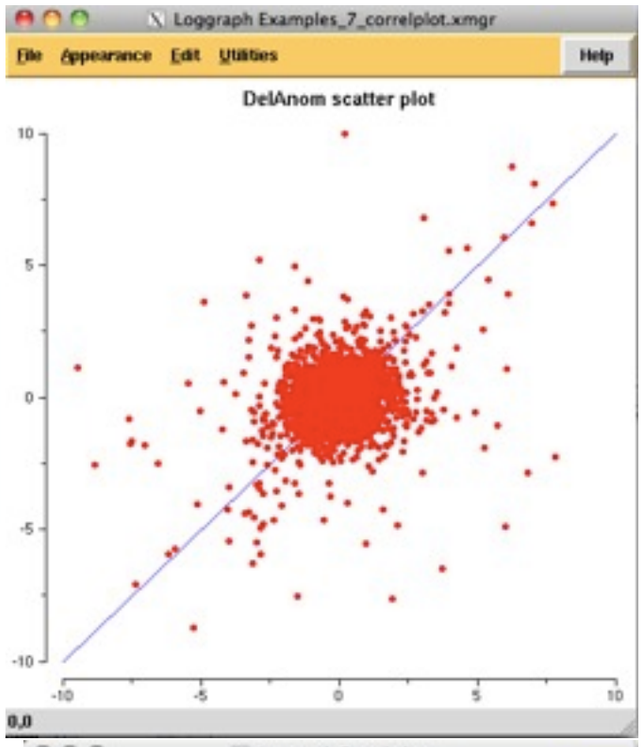
Detecting anomalous signals

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

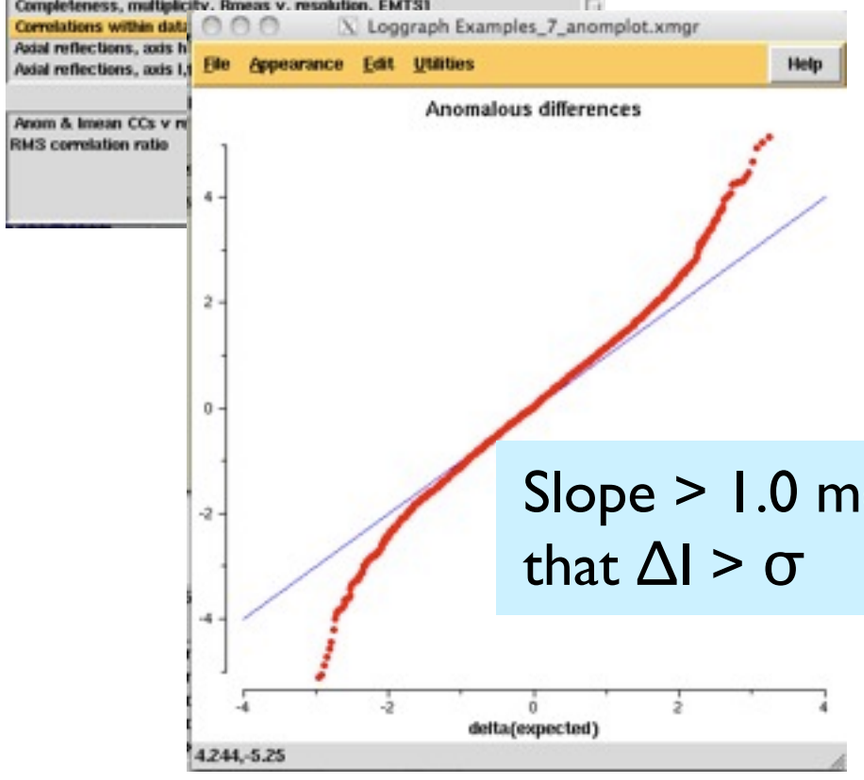


Split one dataset randomly into two halves, calculate correlation between the two halves or
compare different wavelengths (MAD)

Plot ΔI_1 against ΔI_2 should be elongated along diagonal

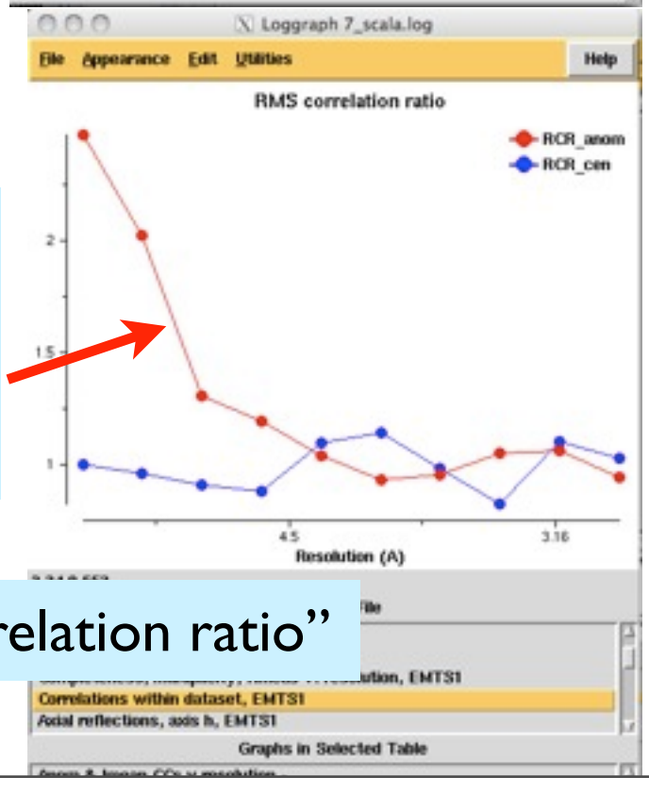


Weak but useful anomalous signal



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

Ratio of width of distribution along diagonal to width across diagonal



"RMS correlation ratio"

Intensities from Mosflm
h k l I σ(I) etc

POINTLESS

Determine point-group
(& space group)

Sorted Intensities in
"best" space group

SCALA/AIMLESS

Scale symmetry-related intensities together
Produce statistics on data quality

Scaled and averaged
intensities

CTRUNCATE

Estimate |F| from I
detect twinning
(intensity statistics)

UNIQUIFY etc

Complete sphere of reflections
Generate or copy freeR flags

h k l F σ(F) I σ(I)
FreeR_flag

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 $\sqrt{I}$ 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

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