Data Reduction

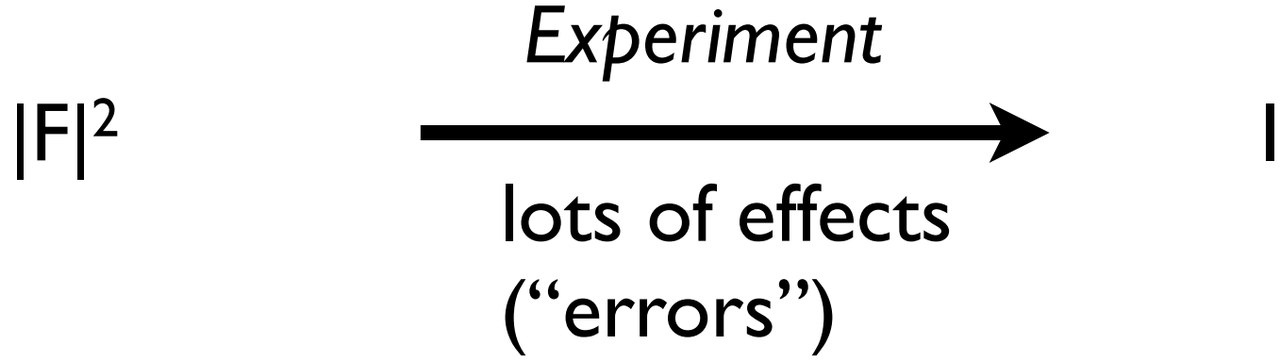
Space Group Determination, Scaling and Intensity Statistics

Phil Evans

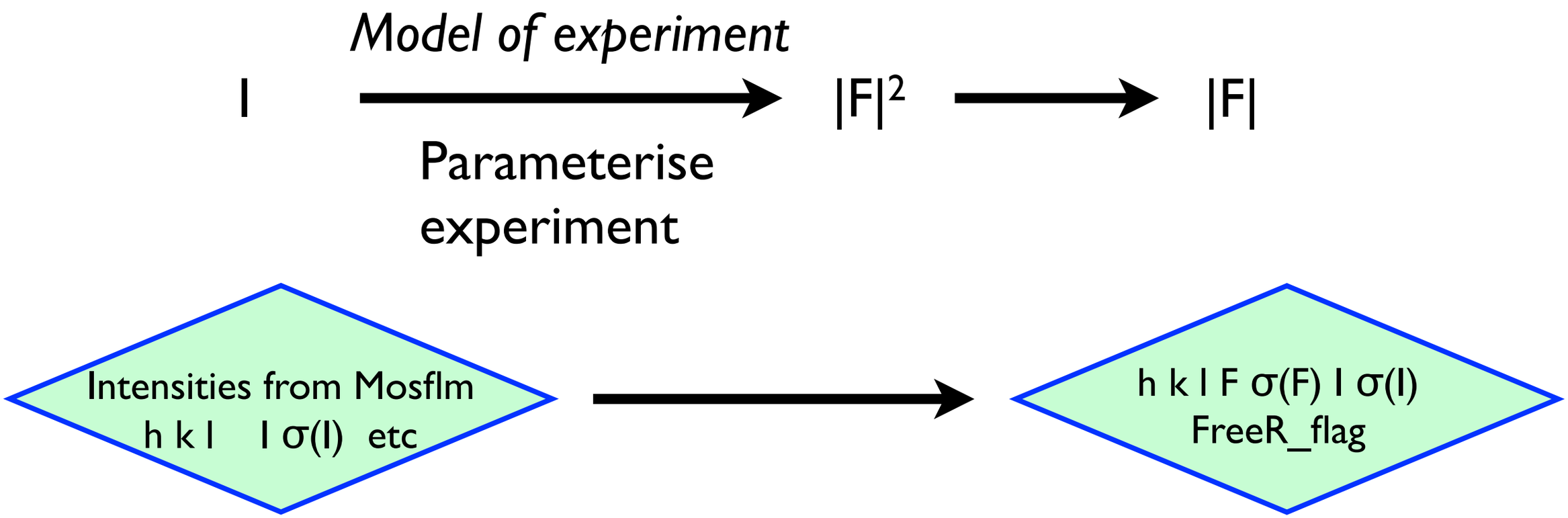
CCP4 Study Weekend January 2010

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



Data reduction can be done in an automated pipeline such as XIA2 (along with integration, *ie* go from images to a list of hkl F ready for structure determination)

XIA2 is used at Diamond (see Graeme Winter for more information)

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

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

POINTLESS

Determine point-group
(& space group)

Sorted Intensities in
"best" space group

SCALA

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

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

POINTLESS

Determine point-group
(& space group)

Sorted Intensities in
"best" space group

SCALA

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

Determination of Space group

The space group symmetry is only a **hypothesis** until the structure is solved, since it is hard to distinguish between true crystallographic and approximate (non-crystallographic) symmetry.

By examining the symmetry of the diffraction pattern we can get a good idea of the likely space group

It is also useful to find the likely symmetry as early as possible, since this affects the data collection strategy

Lattice symmetry imposes constraints on the cell dimensions (eg $\alpha=\beta=\gamma=90°$ for an orthorhombic lattice), but the converse is not true: cell dimensions can have special relationships accidentally. Indexing in *eg* Mosflm only considers lattice *geometry* not symmetry (cubic, hexagonal/trigonal, tetragonal, orthorhombic, monoclinic, or triclinic, + lattice centring P, C, I, R, or F)

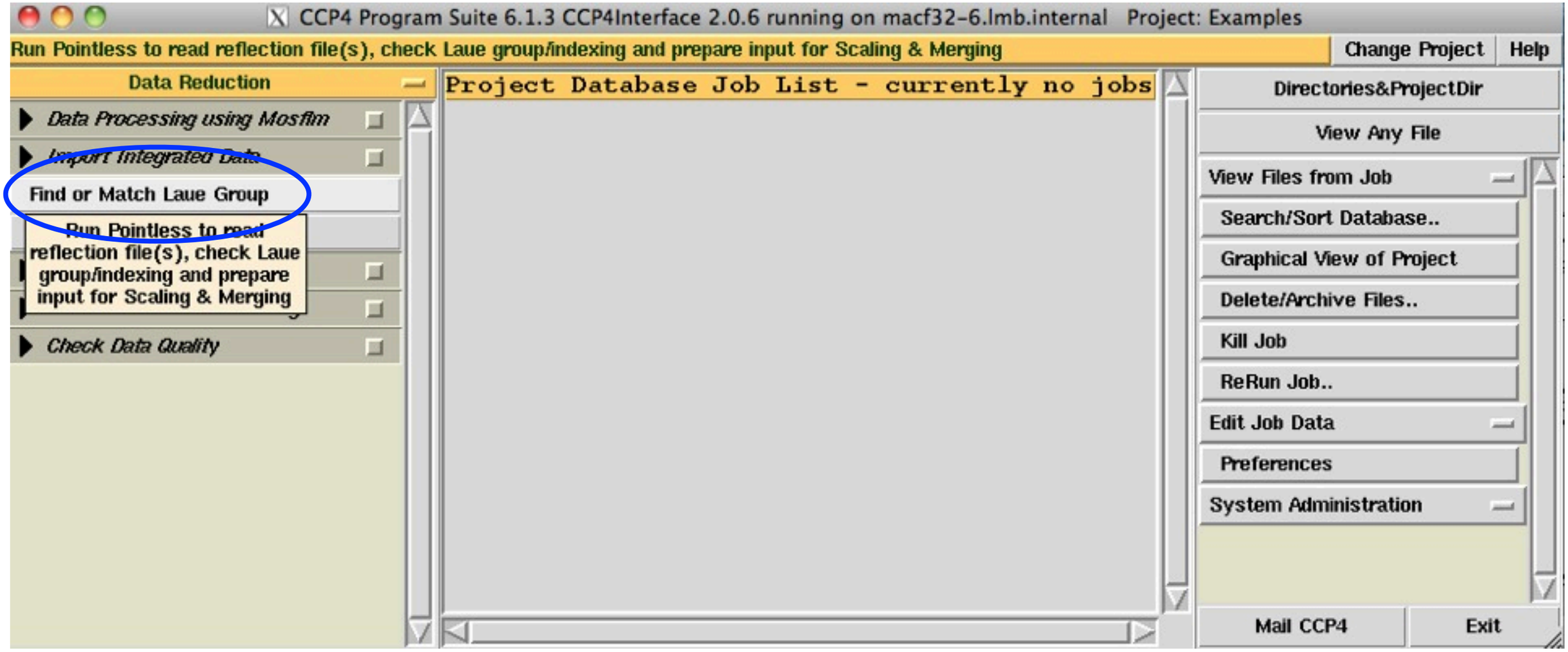
The Laue group (Patterson group) is the symmetry of the diffraction pattern, so can be determined from the observed intensities. It corresponds to the space group without any translations, and with an added centre of symmetry from Friedel's law.

The space group is the point group + lattice centring + translations (*eg* screw dyad rather than pure dyad). Only visible in diffraction pattern as systematic absences along axes – these are not very reliable indicators as there are few axial reflections and there may be accidental absences.

Protocol for space group determination (program *POINTLESS*)

1. From the unit cell dimensions, find the highest compatible lattice symmetry (within a tolerance)
2. Score each symmetry element (rotation) belonging to lattice symmetry using all pairs of observations related by that element
3. Score combinations of symmetry elements for all possible sub-groups (Laue groups) of lattice symmetry group.
4. Score possible space groups from axial systematic absences

Scoring functions for rotational symmetry based on ***correlation coefficient***, since this relatively independent of the unknown scales. R_{meas} values are also calculated


CCP4 Program Suite 6.1.3 CCP4Interface 2.0.6 running on macf32-6.lmb.internal Project: Examples
Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging
Change Project
Help
Data Reduction
Data Processing using Mosflm
Find or Match Laue Group
Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging
Check Data Quality
Project Database Job List - currently no jobs
Directories&ProjectDir
View Any File
View Files from Job
Search/Sort Database..
Graphical View of Project
Delete/Archive Files..
Kill Job
ReRun Job..
Edit Job Data
Preferences
System Administration
Mail CCP4
Exit

CCP4 Program Suite 6.1.3 CCP4Interface 2.0.6 running on macf32-6.lmb.internal Project: Examples

Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging

Change Project | Help

Data Reduction

- *Data Processing using Mosflm*
- *Import Integrated Data*
- Find or Match Laue Group
- Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging
- *Check Data Quality*

Project Database Job List - currently no jobs

Directories&ProjectDir

View Any File

View Files from Job

Search/Sort Database..

Graphical View of Project

Delete/Archive Files..

Kill Job

ReRun Job..

Edit Job Data

Exit

Pointless: prepare intensity data for scaling

Main program option

Help

Job title

Determine Laue group | Match index to reference | Choose a previous solution | Just combine input files

Input reflection file type: MTZ file

Project name: | crystal name: | dataset name:

Select input file

MTZ #1 Examples — Browse | View

Edit list | Add File

Write output reflections in the best space/pointgroup

Output MTZ Examples — Browse | View

Always set primitive orthorhombic groups in cell length order (a<b<c) & allow monoclinic I2 setting of C2

Excluded Data

Lattice Symmetry Determination

Criteria For Accepting Partials

Additional Options

Run | Save or Restore | Close

CCP4 Program Suite 6.1.3 CCP4Interface 2.0.6 running on macf32-6.lmb.internal Project: Examples

Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging

Change Project | Help

Data Reduction

- Data Processing using Mosflm
- Import Integrated Data
- Find or Match Laue Group
- Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging
- Check Data Quality

Project Database Job List - currently no jobs

Directories&ProjectDir

View Any File

View Files from Job

Search/Sort Database..

Graphical View of Project

Delete/Archive Files..

Kill Job

ReRun Job..

Edit Job Data

Exit

Pointless: prepare intensity data for scaling

Use the title!

Help

Job title

Determine Laue group | Match index to reference | Choose a

Specify Project/crystal/dataset names

Input reflection file type: MTZ file

Project name: New crystal name: New dataset name: New

MTZ #1 Examples GearXe001a_001.mtz Browse View

Edit list Add File

Write output reflections in the best space/pointgroup

Specify output file name

Output MTZ Examples Browse View

Always set primitive orthorhombic groups in cell length order (a<b<c) & allow monoclinic I2 setting of C2

Excluded Data

Lattice Symmetry Determination

Criteria For Accepting Partials

Additional Options

Run | Save or Restore | Close

CCP4 Program Suite 6.1.3 CCP4Interface 2.0.6 running on macf32-6.lmb.internal Project: Examples

Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging

Change Project | Help

Data Reduction

- Data Processing using Mosflm
- Import Integrated Data
- Find or Match Laue Group
- Run Pointless to read reflection file(s), check Laue group/indexing and prepare input for Scaling & Merging
- Check Data Quality

Project Database Job List - currently no jobs

Directories&ProjectDir

View Any File

View Files from Job

Search/Sort Database..

Graphical View of Project

Delete/Archive Files..

Kill Job

ReRun Job..

Edit Job Data

Exit

Pointless: prepare intensity data for scaling

Help

Job title: Find spacegroup and sort

Determine Laue group | Match index to reference | Choose a previous solution | Just combine input files

Input reflection file type: MTZ file

Project name: Gamma crystal name: Xe1 dataset name: Xe1

MTZ #1 Examples GearXe001a_001.mtz Browse View

Edit list | Add File

Write output reflections in the best space/pointgroup

Output MTZ Examples gamma_xe1.mtz Browse View

Always set primitive orthorhombic groups in cell length order (a<b<c) & allow monoclinic I2 setting of C2

Excluded Data

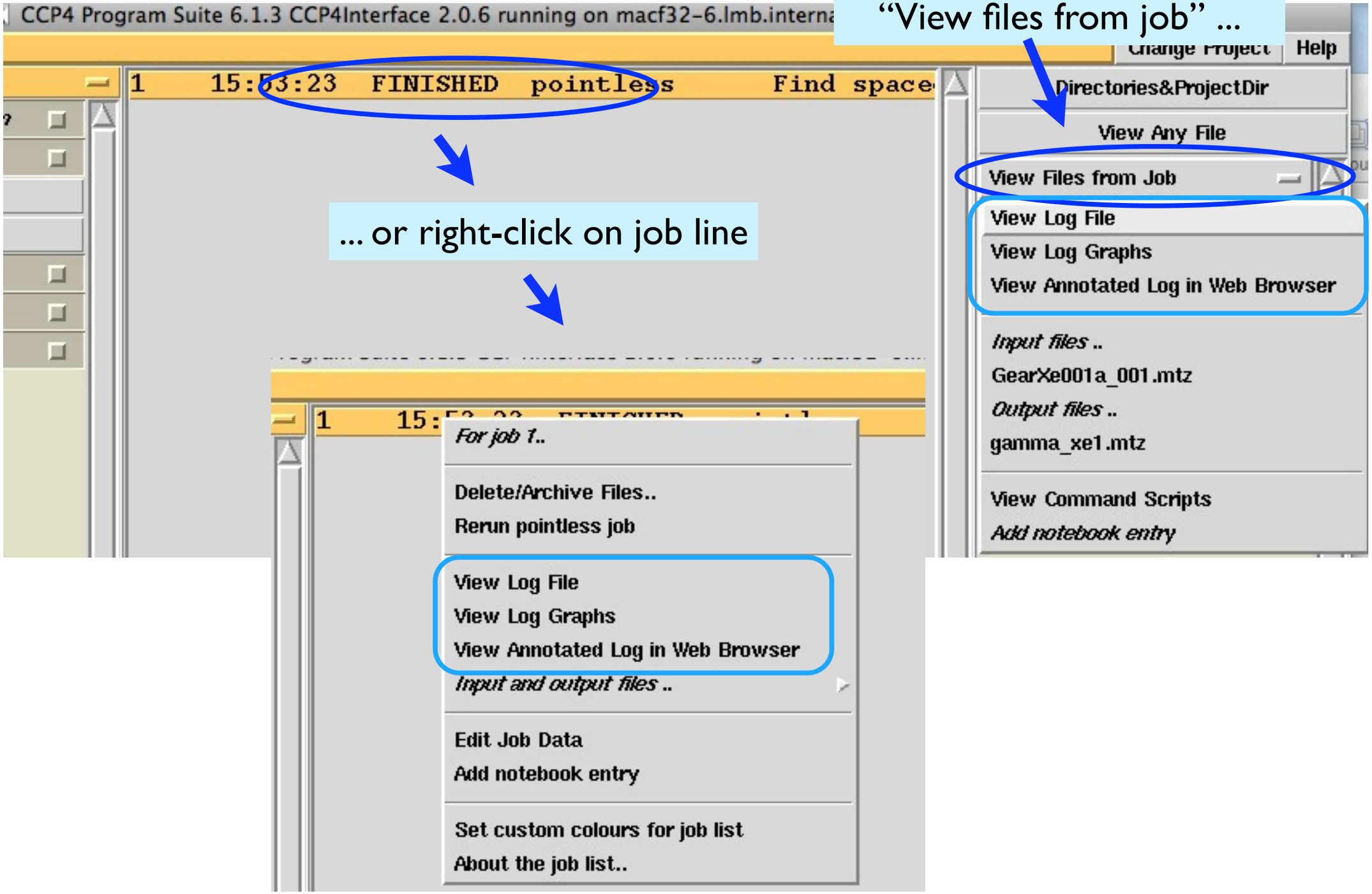
Lattice Symmetry Determination

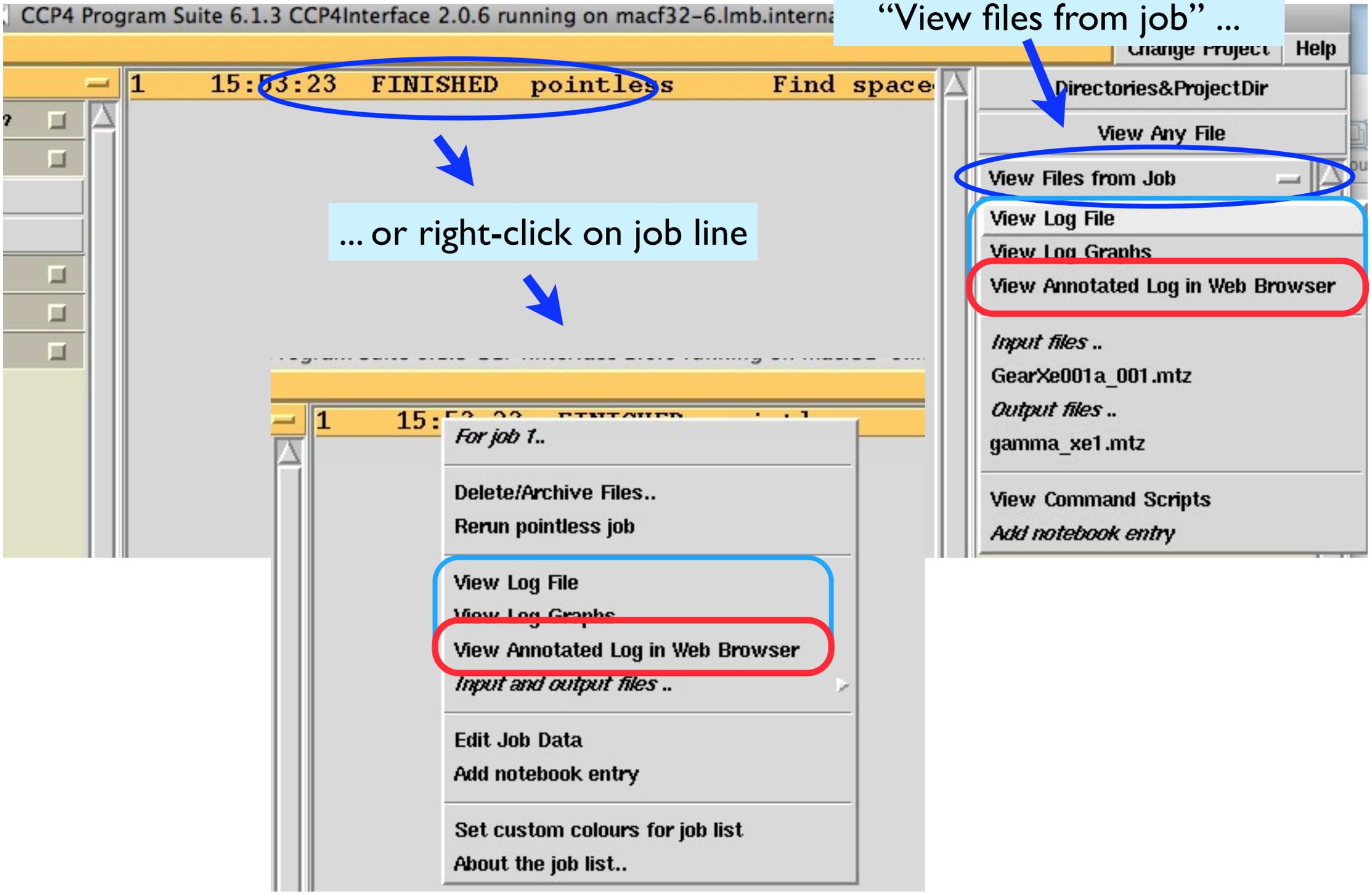
Criteria For Accepting Partials

Additional Options

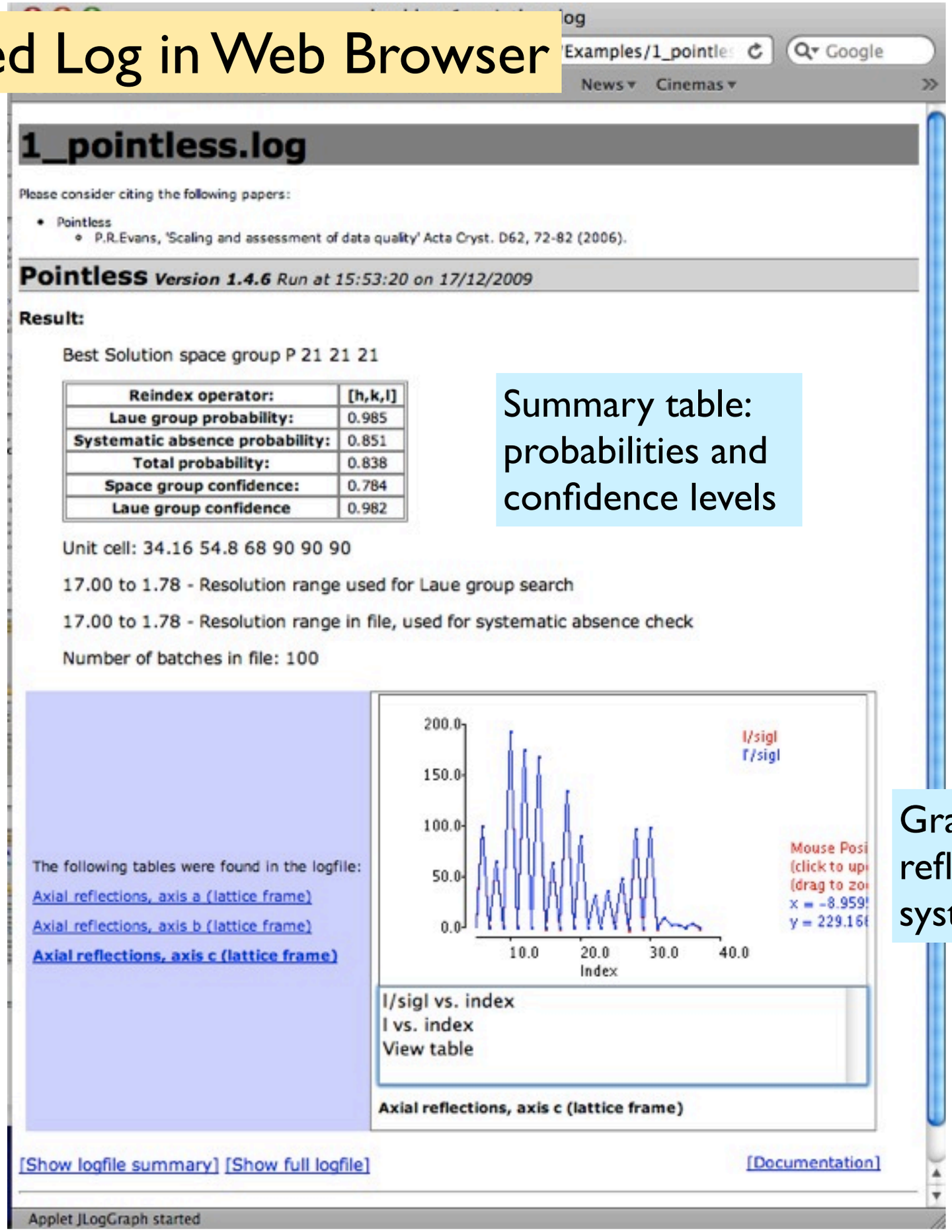
Run | Save or Restore | Close

Examine output either from "View files from job" ...




Examine output either from "View files from job" ...
CCP4 Program Suite 6.1.3 CCP4Interface 2.0.6 running on macf32-6.lmb.interna
1 15:53:23 FINISHED pointless Find space
Change Project
Help
Directories&ProjectDir
View Any File
View Files from Job
View Log File
View Log Graphs
View Annotated Log in Web Browser
Input files ..
GearXe001a_001.mtz
Output files ..
gamma_xe1.mtz
View Command Scripts
Add notebook entry
... or right-click on job line
For job 1..
Delete/Archive Files..
Rerun pointless job
View Log File
View Log Graphs
View Annotated Log in Web Browser
Input and output files ..
Edit Job Data
Add notebook entry
Set custom colours for job list
About the job list..

View Annotated Log in Web Browser



Summary table: probabilities and confidence levels

Graphs of axial reflections for systematic absences

A straightforward orthorhombic case

Scoring the symmetry operators separately sometimes allows detection of pseudo-symmetry, *eg* if some rotation operators are much weaker than others

Analysing rotational symmetry in lattice group P m m m

Probability | Correlation coefficient | R-factor

Nelmt	Lklhd	Z-cc	CC	N	Rmeas		Symmetry & operator (in Lattice Cell)	
1	0.948	9.54	0.95	12122	0.097		identity	
2	0.942	9.44	0.94	18346	0.121	***	2-fold l (0 0 1)	{-h,-k,+l}
3	0.949	9.58	0.96	30259	0.097	***	2-fold h (1 0 0)	{+h,-k,-l}
4	0.912	9.15	0.92	17427	0.120	***	2-fold k (0 1 0)	{-h,+k,-l}

Separate scores for each symmetry operator in maximum possible lattice symmetry

Probability | Correlation coefficient | R-factor

	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
= 1	P m m m	***	0.985	9.35	9.35	0.00	0.94	0.00	0.11	0.00	0.0	[h,k,l]
2	P 1 2/m 1		0.006	0.38	9.56	9.18	0.96	0.92	0.10	0.12	0.0	[-k,-h,-l]
3	P 1 2/m 1		0.005	-0.01	9.38	9.39	0.94	0.94	0.11	0.11	0.0	[-h,-l,-k]
4	P 1 2/m 1		0.003	-0.13	9.31	9.44	0.93	0.94	0.11	0.11	0.0	[h,k,l]
5	P -1		0.000	0.22	9.54	9.32	0.95	0.93	0.10	0.11	0.0	[h,k,l]

Combined scores for all possible Laue (point) groups down to P1

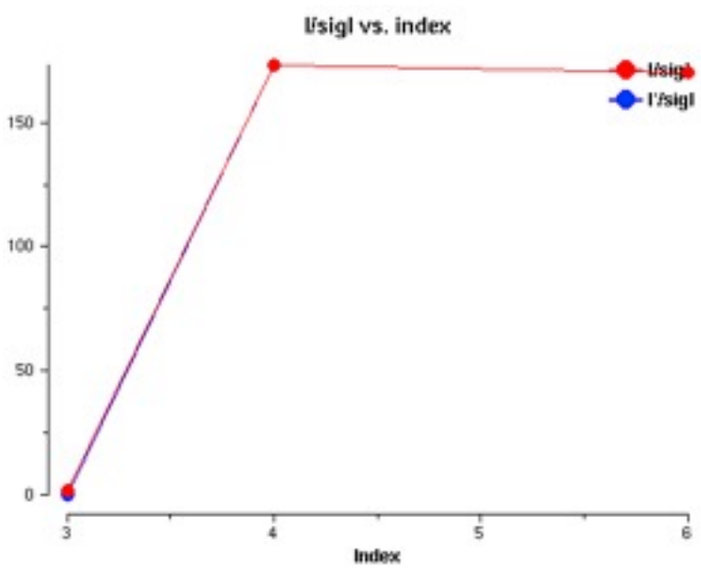
A clear indication that the Laue group is Pmmm (P222)

Possible axial systematic absences to determine space group

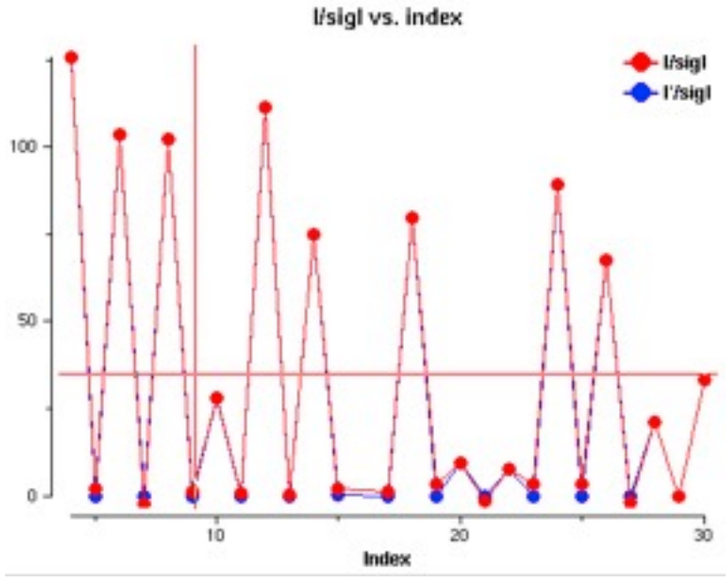
Zone	Number	PeakHeight	SD	Probability	ReflectionCondition
Zones for Laue group P m m m					
1 screw axis 2(1) [a]	3	1.000	0.296	** 0.889	h00: h=2n
2 screw axis 2(1) [b]	26	1.000	0.142	*** 0.971	0k0: k=2n
3 screw axis 2(1) [c]	46	0.997	0.097	*** 0.986	00l: l=2n

Fourier analysis of I/σ(I)

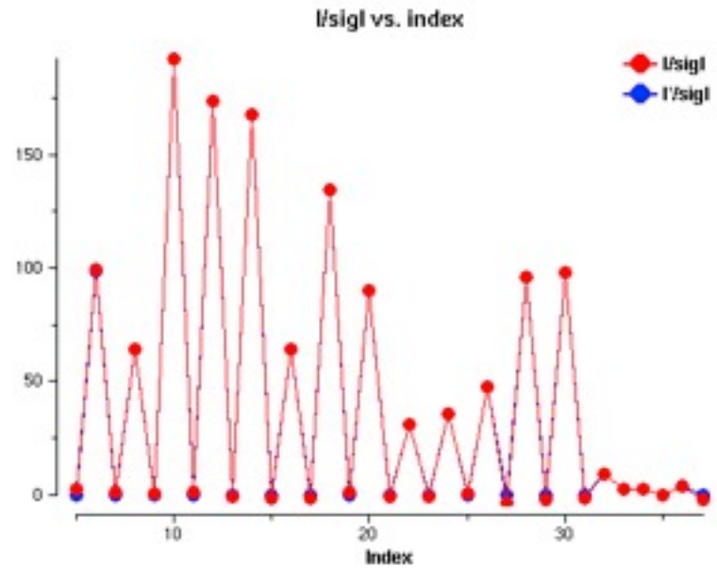
There are indications of 2_1 screw symmetry along all principle axes
(though note there are only 3 observations on the *a* axis (h00 reflections))



Possible 2_1 axis along a



Clear 2_1 axis along b



Clear 2_1 axis along c

```
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

Pseudo-cubic example

Cell: 79.15 81.33 81.15 90.00 90.00 90.00 a ≈ b ≈ c

```
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}
```

Only orthorhombic symmetry operators are present

Pseudo-cubic example

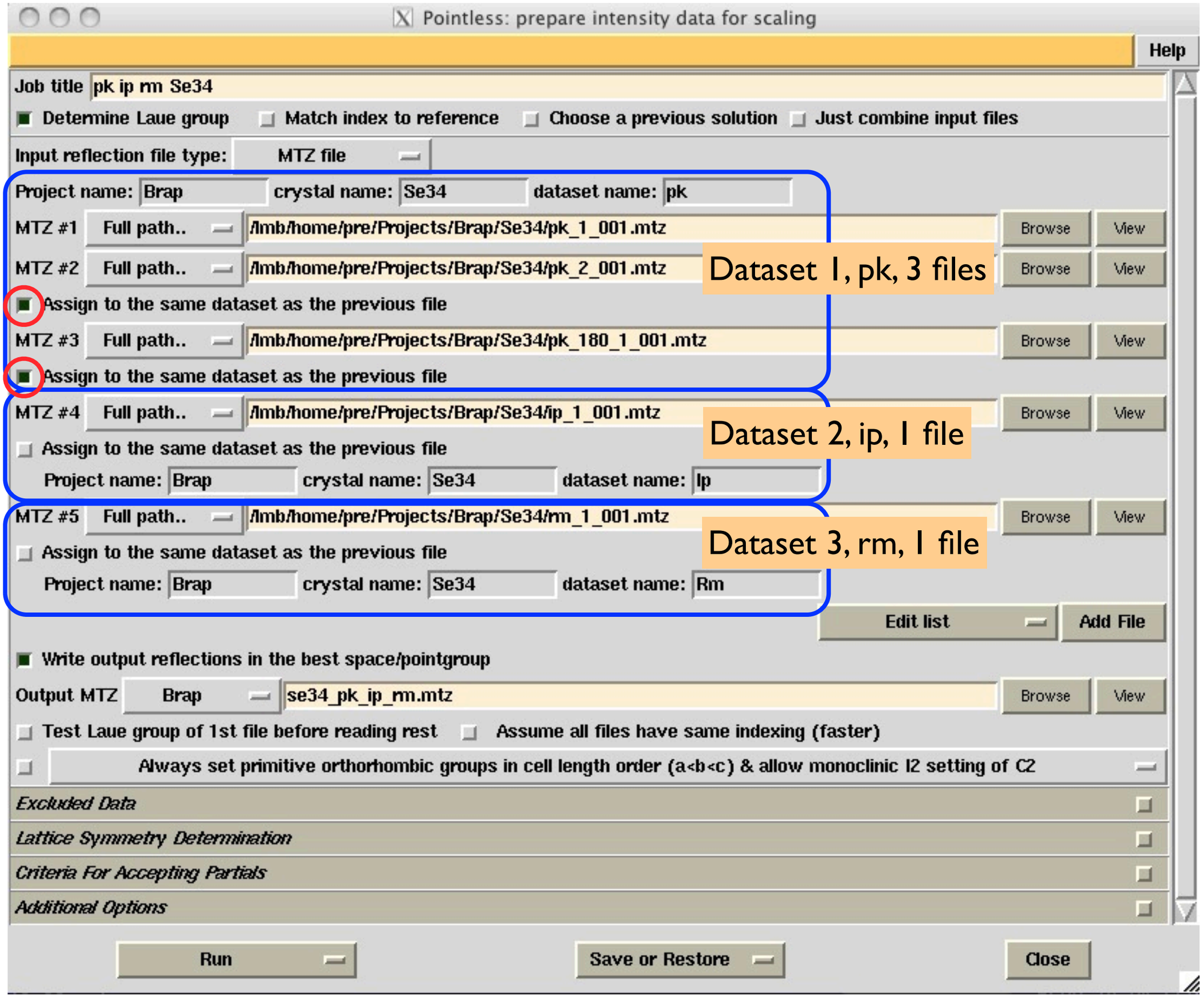
Cell: 79.15 81.33 81.15 90.00 90.00 90.00 a ≈ b ≈ c

	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
= 1	P m m m	***	0.989	8.93	9.59	0.66	0.96	0.07	0.12	0.69	0.0	[-h,-l,-k]
2	P 1 2/m 1		0.003	7.85	9.65	1.80	0.97	0.18	0.09	0.60	0.0	[-h,-l,-k]
3	P 1 2/m 1		0.003	7.95	9.63	1.68	0.96	0.17	0.10	0.61	0.0	[l,h,k]
4	P 1 2/m 1		0.003	7.80	9.61	1.81	0.96	0.18	0.11	0.60	0.0	[h,k,l]
5	P 4/m m m		0.000	6.69	6.90	0.21	0.69	0.02	0.24	0.75	1.5	[-k,-h,-l]
6	P 4/m m m		0.000	4.55	5.41	0.85	0.54	0.09	0.34	0.68	0.1	[-l,-k,-h]
7	P 4/m		0.000	5.45	7.20	1.75	0.72	0.18	0.20	0.62	1.5	[-k,-h,-l]
8	P 4/m		0.000	4.72	6.53	1.81	0.65	0.18	0.25	0.60	0.1	[-l,-k,-h]
9	P -1		0.000	7.48	9.70	2.22	0.97	0.22	0.07	0.57	0.0	[-h,-l,-k]
10	P 4/m		0.000	4.03	5.96	1.92	0.60	0.19	0.29	0.59	1.4	[-h,-l,-k]
11	P 4/m m m		0.000	4.93	5.63	0.69	0.56	0.07	0.32	0.69	1.4	[-h,-l,-k]
12	C m m m		0.000	4.97	6.67	1.70	0.67	0.17	0.24	0.62	1.5	[h-k,-h-k,-l]
13	C 1 2/m 1		0.000	4.80	6.99	2.19	0.70	0.22	0.21	0.57	1.5	[-h-k,-h+k,-l]
14	C 1 2/m 1		0.000	4.51	6.71	2.20	0.67	0.22	0.23	0.58	1.5	[h-k,-h-k,-l]
15	C m m m		0.000	3.08	5.01	1.93	0.50	0.19	0.36	0.59	0.1	[-k-l,-k+l,-h]
16	P m -3		0.000	3.35	4.32	0.97	0.43	0.10	0.44	0.63	1.5	[h,k,l]
17	C 1 2/m 1		0.000	2.58	4.95	2.36	0.49	0.24	0.35	0.56	0.1	[k-l,-k-l,-h]
18	C 1 2/m 1		0.000	2.65	5.01	2.36	0.50	0.24	0.34	0.56	0.1	[-k-l,-k+l,-h]
19	H -3		0.000	2.17	4.56	2.39	0.46	0.24	0.40	0.55	1.5	[-k+l,-h-l,h-k-l]
20	H -3		0.000	2.09	4.48	2.39	0.45	0.24	0.40	0.55	1.5	[h-l,-h-k,-h+k-l]
21	H -3		0.000	2.15	4.54	2.39	0.45	0.24	0.39	0.55	1.5	[-h+k,-k-l,-h-k+l]
22	H -3		0.000	2.20	4.59	2.38	0.46	0.24	0.39	0.55	1.5	[k-l,h-k,-h-k-l]
23	C 1 2/m 1		0.000	3.10	5.42	2.32	0.54	0.23	0.31	0.56	1.4	[-h-l,h-l,-k]
24	C 1 2/m 1		0.000	3.36	5.67	2.31	0.57	0.23	0.30	0.56	1.4	[-h+l,-h-l,-k]
25	C m m m		0.000	3.32	5.29	1.97	0.53	0.20	0.34	0.59	1.4	[-h-l,h-l,-k]
26	H -3 m		0.000	-0.01	2.66	2.67	0.27	0.27	0.52	0.54	1.5	[-h+k,-k-l,-h-k+l]
27	H -3 m		0.000	-0.03	2.65	2.68	0.26	0.27	0.52	0.54	1.5	[k-l,h-k,-h-k-l]
28	H -3 m		0.000	-0.13	2.58	2.71	0.26	0.27	0.53	0.53	1.5	[h-l,-h-k,-h+k-l]
29	H -3 m		0.000	-0.02	2.66	2.68	0.27	0.27	0.52	0.53	1.5	[-k+l,-h-l,h-k-l]
30	P m -3 m		0.000	2.67	2.67	0.00	0.27	0.00	0.53	0.00	1.5	[h,k,l]

... symmetry is actually orthorhombic (P 2_1 2_1 2_1)

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

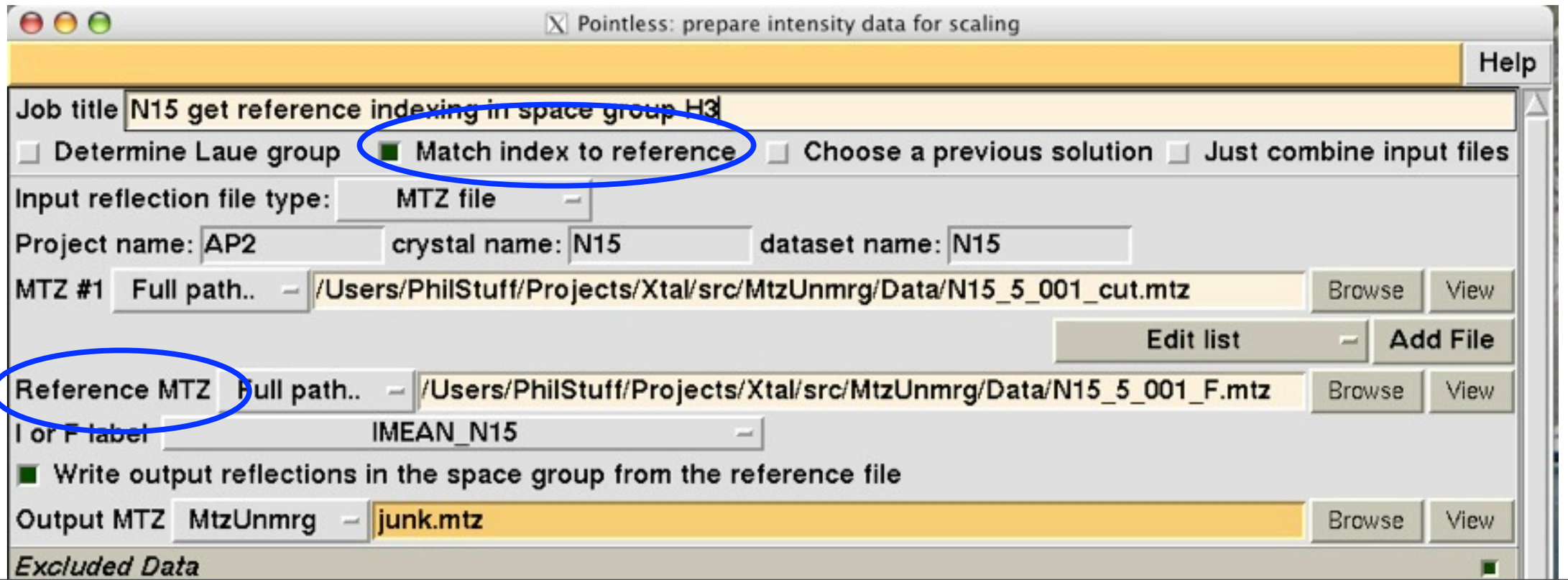
Alternative indexing

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)



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

POINTLESS

Determine point-group
(& space group)

merged

Sorted Intensities in
"best" space group

SCALA

Scale symmetry-related intensities together
Produce statistics on data quality

Scaled and averaged
intensities

CTRUNCATE

Estimate |F| from I
detect twinning

UNIQUIFY etc

Complete sphere of reflections
Generate or copy freeR flags

h k l F σ(F) I σ(I)
FreeR_flag

Scaling

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.

After scaling, the remaining differences between observations can be analysed to give an *indication* of data quality, though not necessarily of its absolute correctness.

Measures of internal consistency:

R-factors & correlation coefficients:

R_{merge} (R_{sym}) = $\Sigma \mid I_{hl} - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$

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>

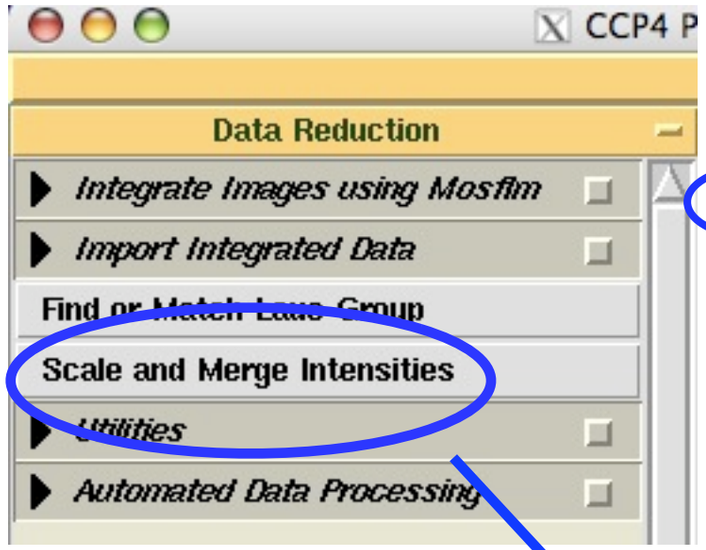
CC pairwise correlation coefficients (see later)

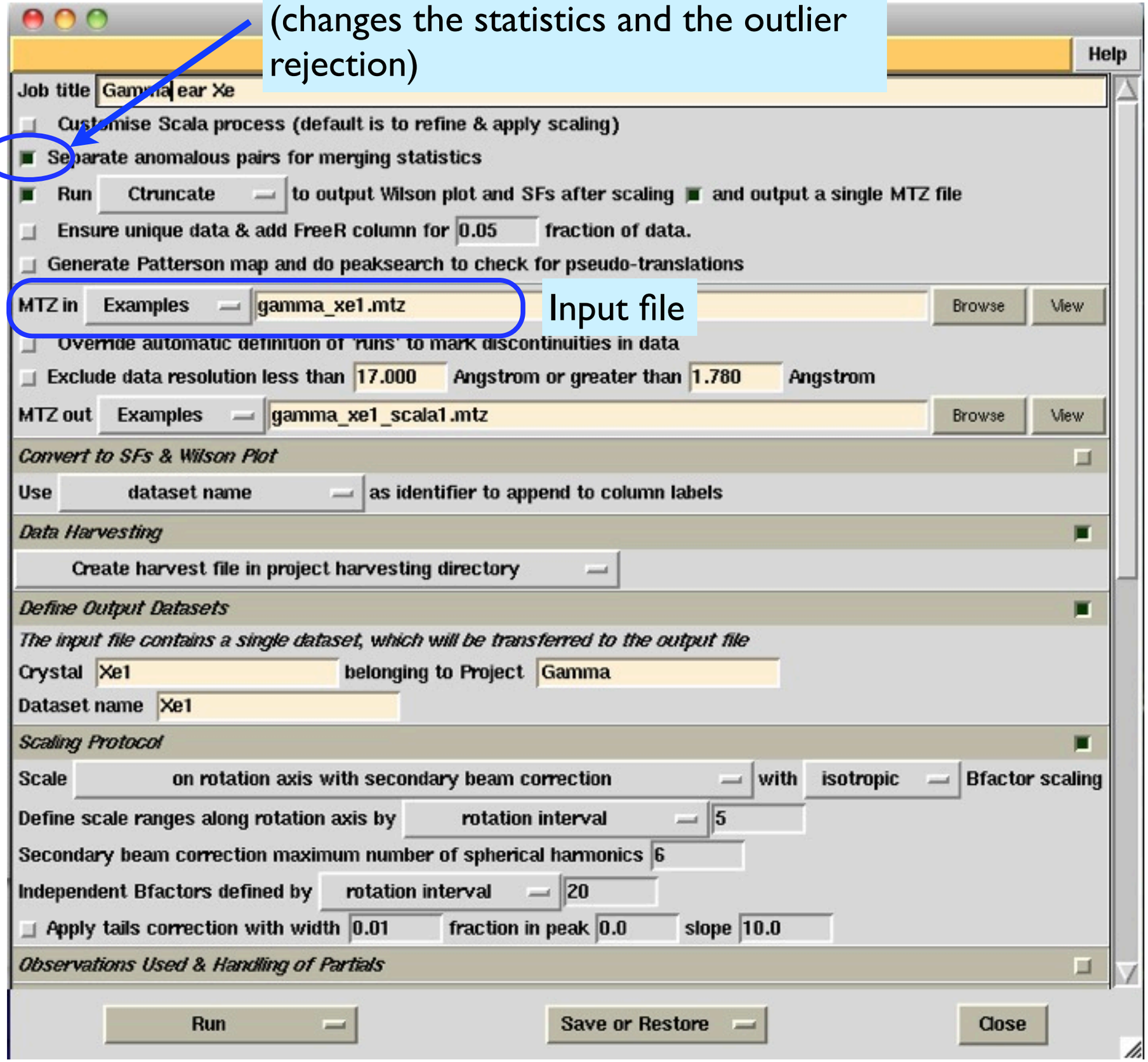
Running *SCALA* from *ccp4i* interface

Click if you have anomalous scattering (changes the statistics and the outlier rejection)

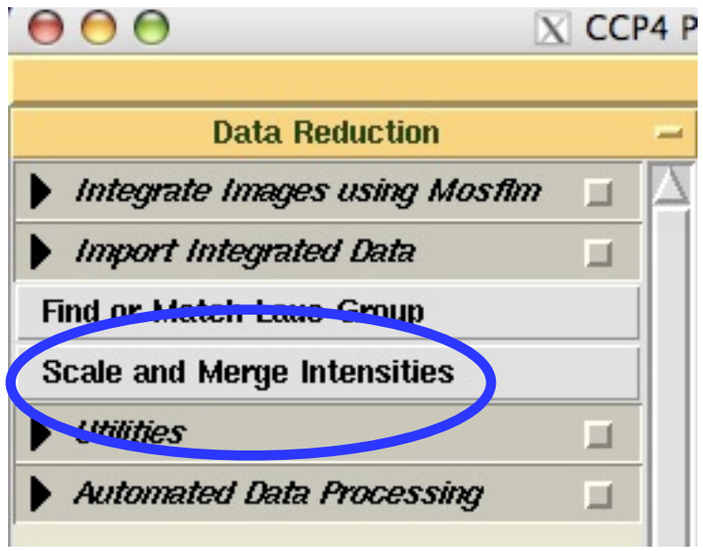
Input file

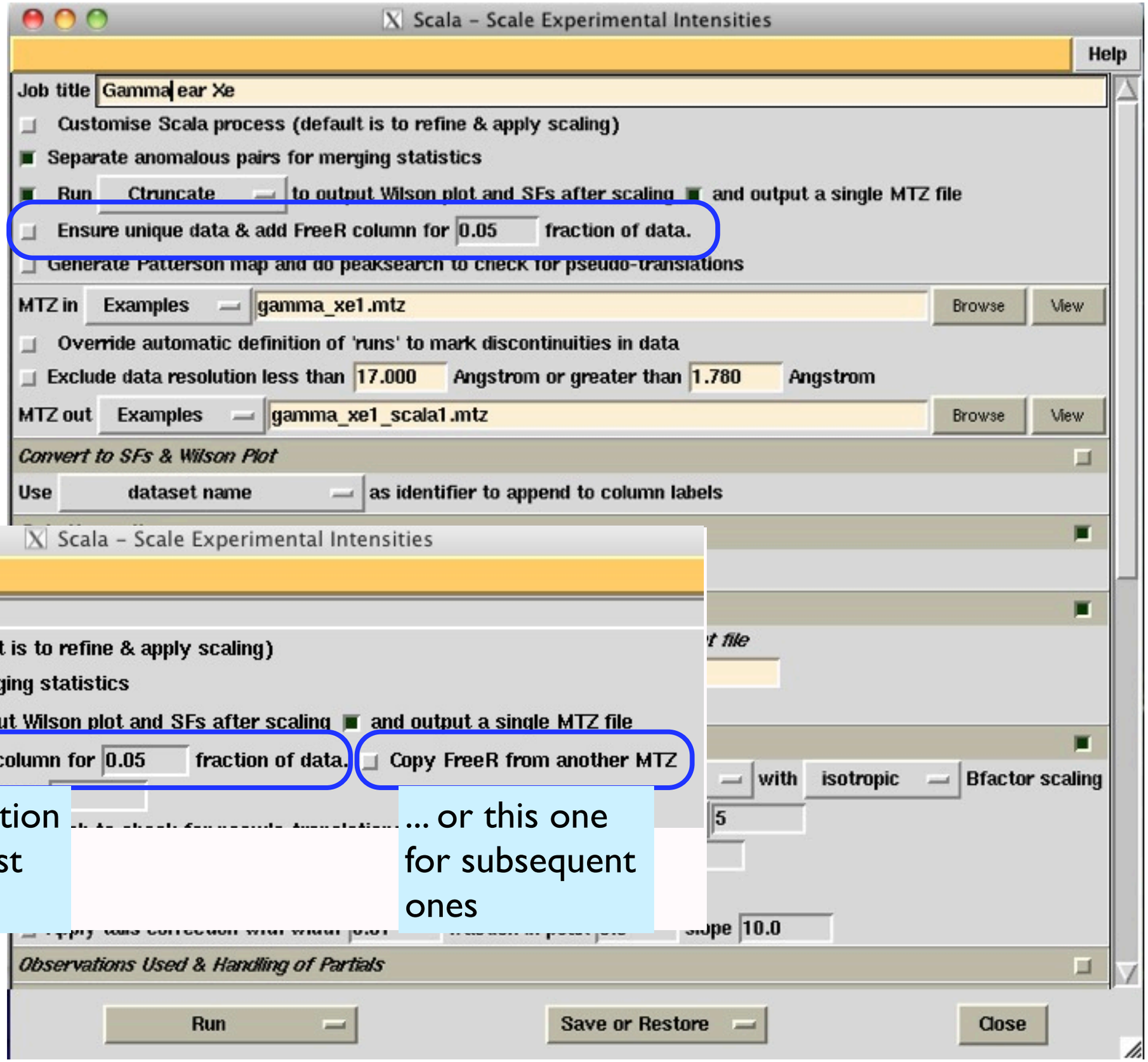
Usually use the default scaling options





Running *SCALA* from *ccp4i* interface





Use this option for your first dataset ...

... or this one for subsequent ones

What to look at?

View Annotated Log in Web Browser

Summary

"Table 1"

Scala ***Version 3.3.15*** *Run at 17:31:58 on 21/12/2009* ***Finished with: ** Normal termination ****

Result:

Summary data for Project: Gamma Crystal: Xe1 Dataset: Xe1

	Overall	InnerShell	OuterShell
Low resolution limit	17.00	17.00	1.88
High resolution limit	1.78	5.63	1.78
Rmerge	0.034	0.025	0.196
Rmerge in top intensity bin	0.021	-	-
Rmeas (within I+/I-)	0.046	0.034	0.261
Rmeas (all I+ & I-)	0.059	0.056	0.264
Rpim (within I+/I-)	0.030	0.023	0.171
Rpim (all I+ & I-)	0.029	0.030	0.133
Fractional partial bias	-0.003	-0.002	-0.010
Total number of observations	44572	1443	4824
Total number unique	12130	435	1403
Mean((I)/sd(I))	18.0	30.0	5.6
Completeness	95.1	93.9	77.4
Multiplicity	3.7	3.3	3.4
Anomalous completeness	88.5	92.2	65.1
Anomalous multiplicity	2.1	2.1	2.0
DelAnom correlation between half-sets	0.539	0.762	-0.024
Mid-Slope of Anom Normal Probability	1.399	-	-

Outlier rejection and statistics assume that there is anomalous scattering, ie I+ differs from I-

Average unit cell: 34.16 54.80 68.00 90.00 90.00 90.00

Space group: P 21 21 21

Average mosaicity: 0.98

Minimum and maximum SD correction factors: Fulls 1.09 2.60 Partials 1.73 16.73

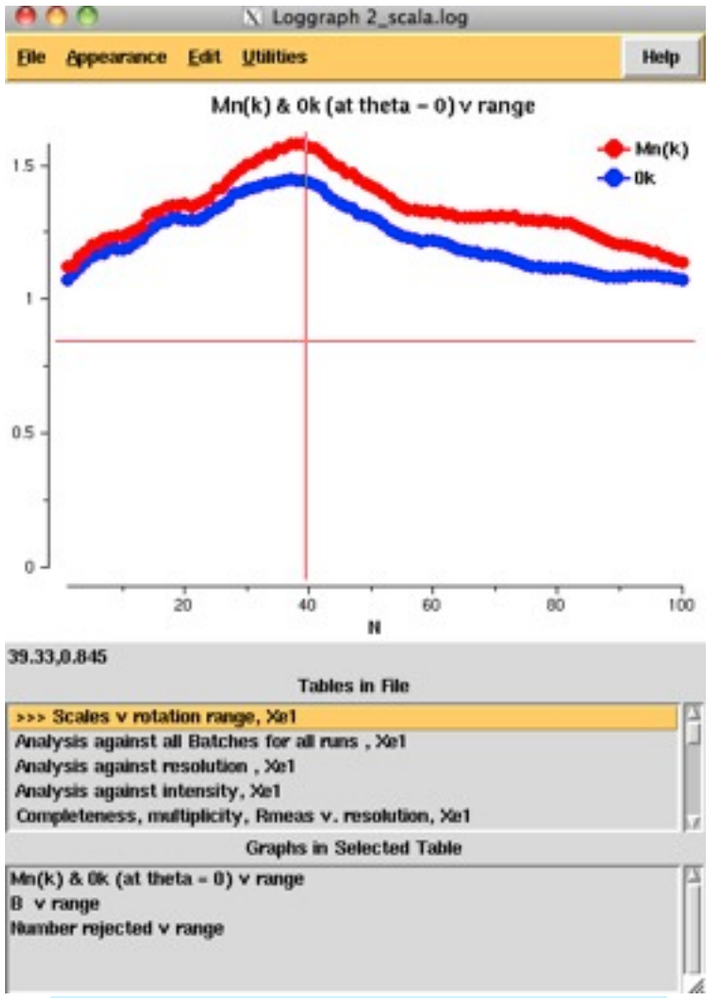
Dataset: Gamma/Xe1/Xe1
written as averaged data to output file /Users/pre/Projects/Xtal/Temp/Examples_2_1_mtz.tmp

Maximum resolution: 1.78A

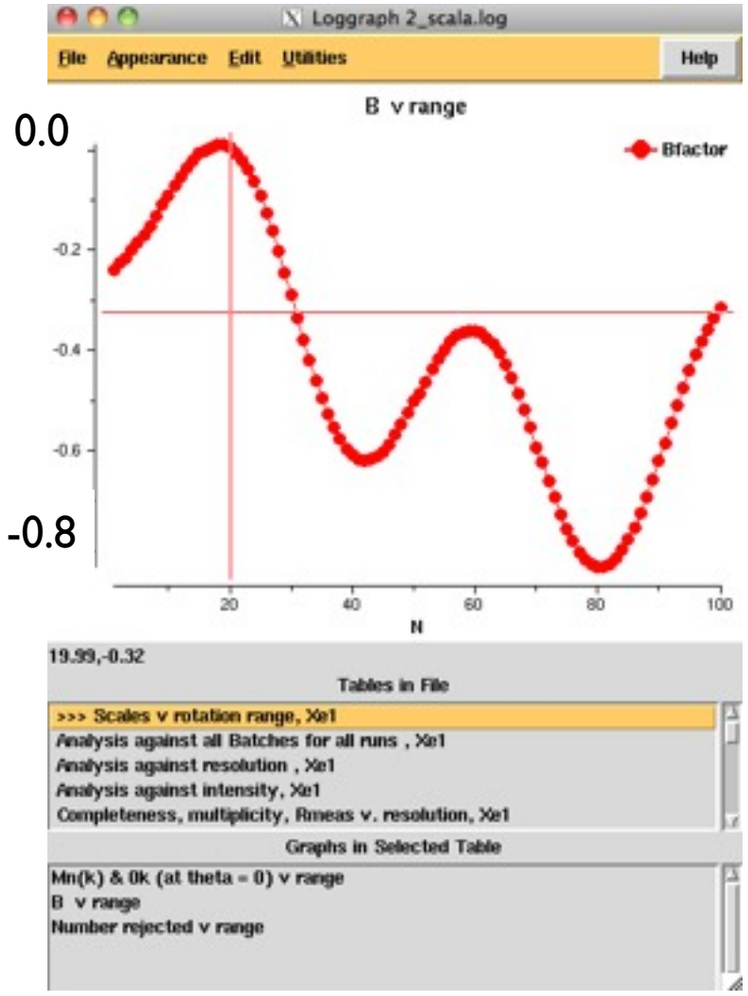
Graphs: Analyses against "batch" (image number or "time")

- check for level of radiation damage
  if you cut back from the end, there is a trade-off between damage and completeness
- check for bad images or regions

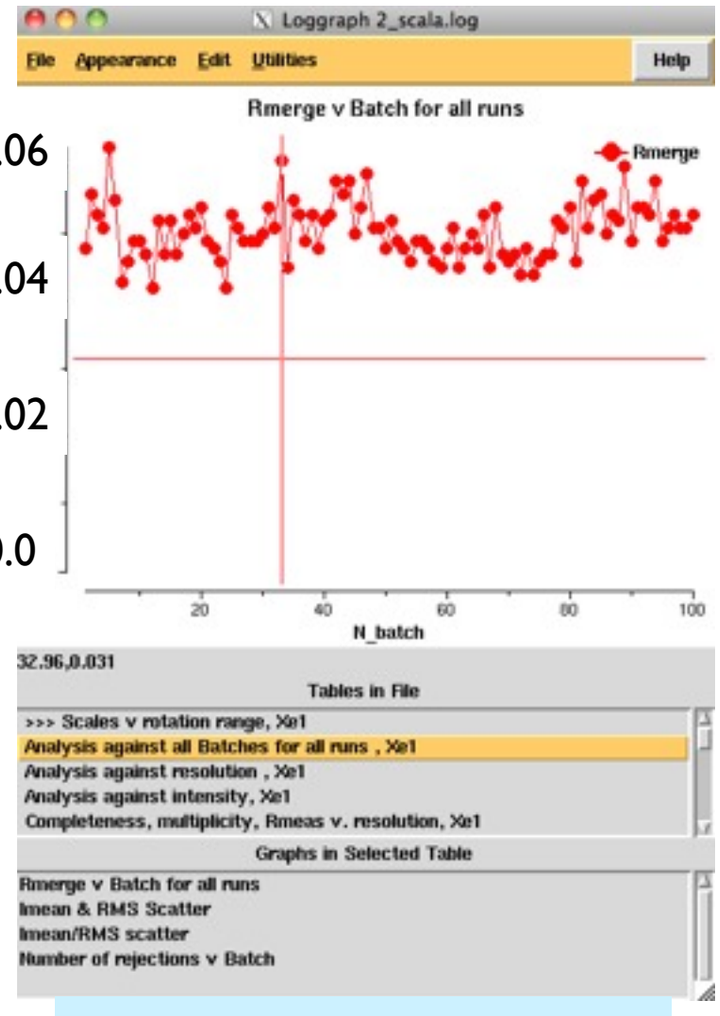
A good case



No great difference between average scale Mn(k) & scale at θ=0

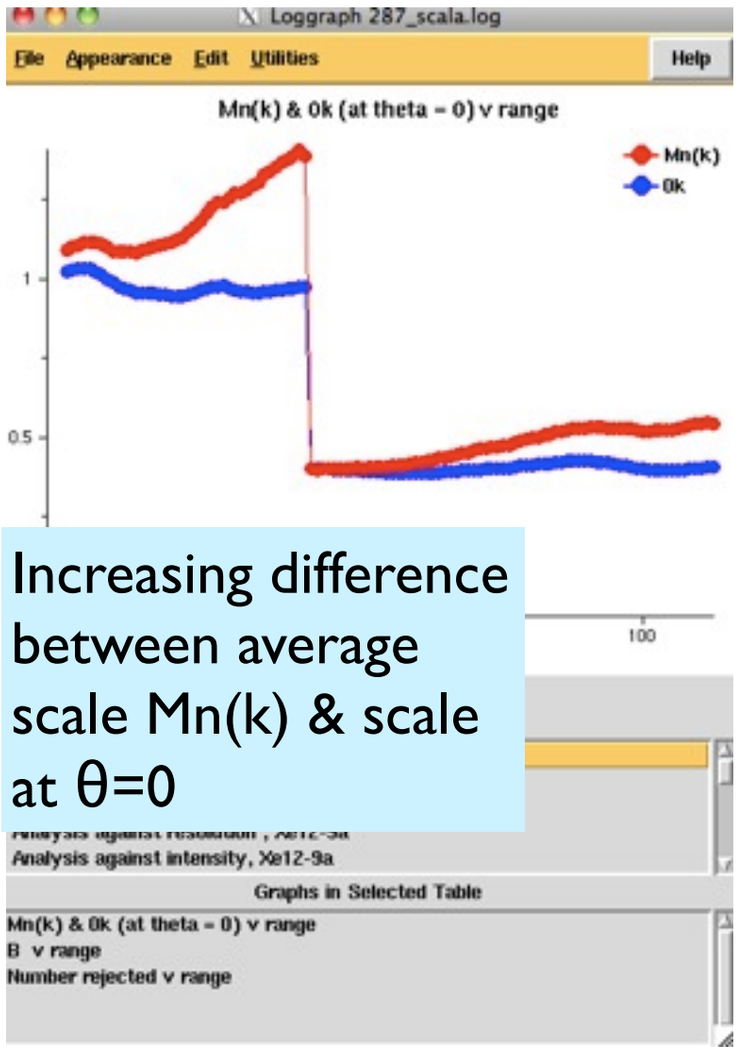


Small variation in relative B-factor



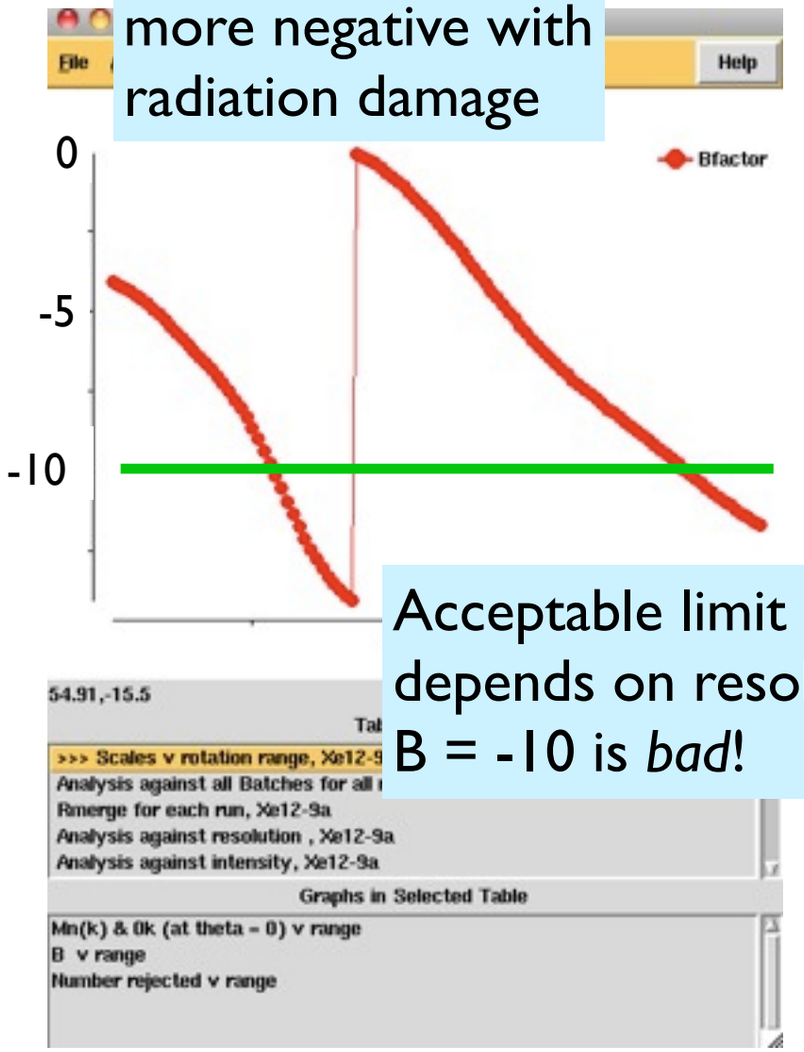
Uniform and low R_{merge}

A bad case: two crystals, both dying, both incomplete

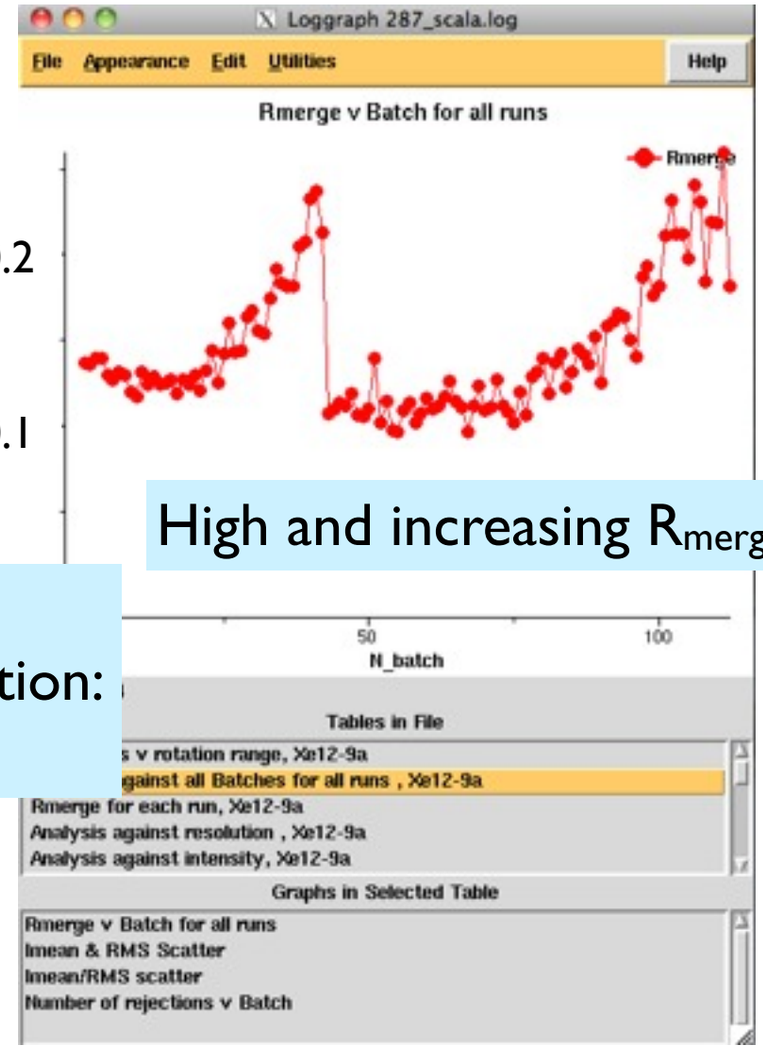


Increasing difference between average scale Mn(k) & scale at θ=0

relative B gets more negative with radiation damage



Acceptable limit depends on resolution: B = -10 is *bad*!

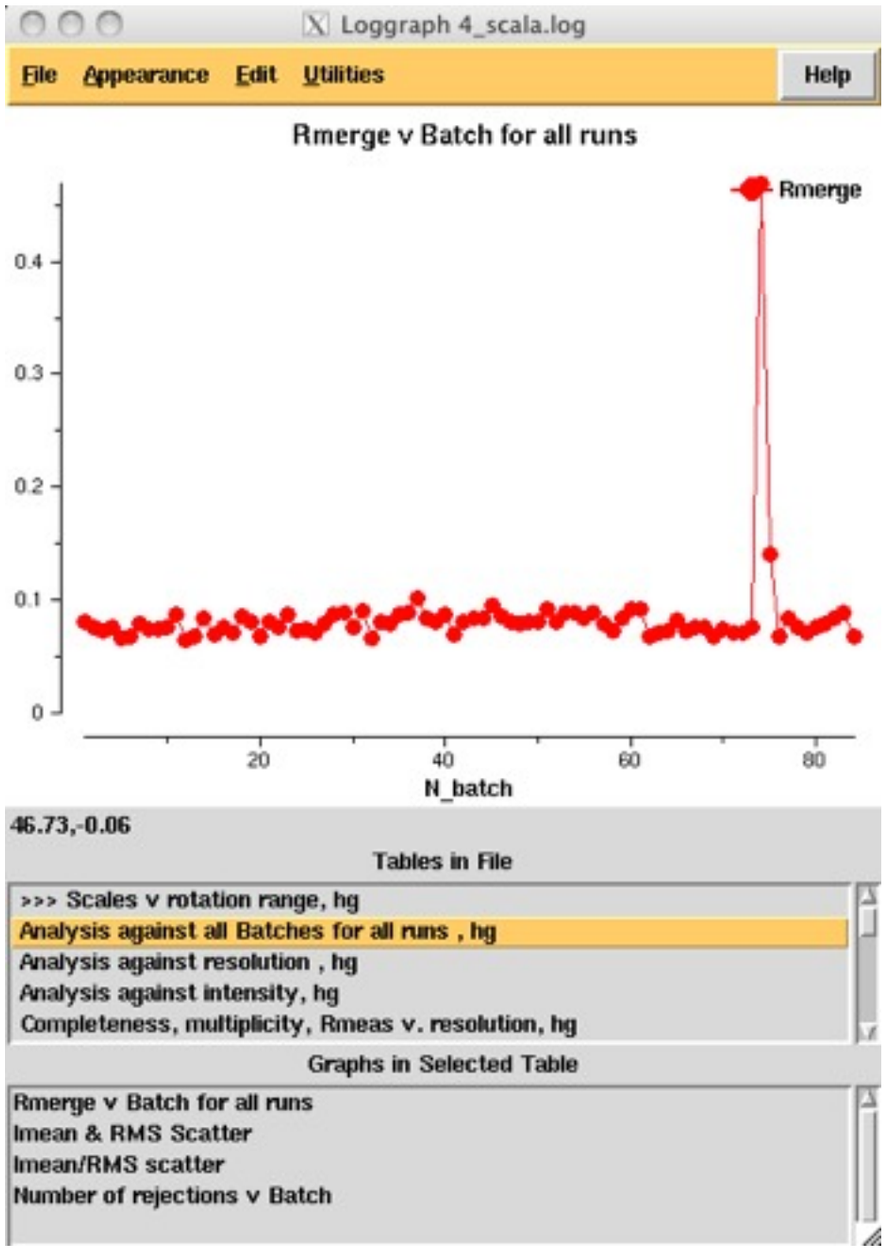


High and increasing R_{merge}

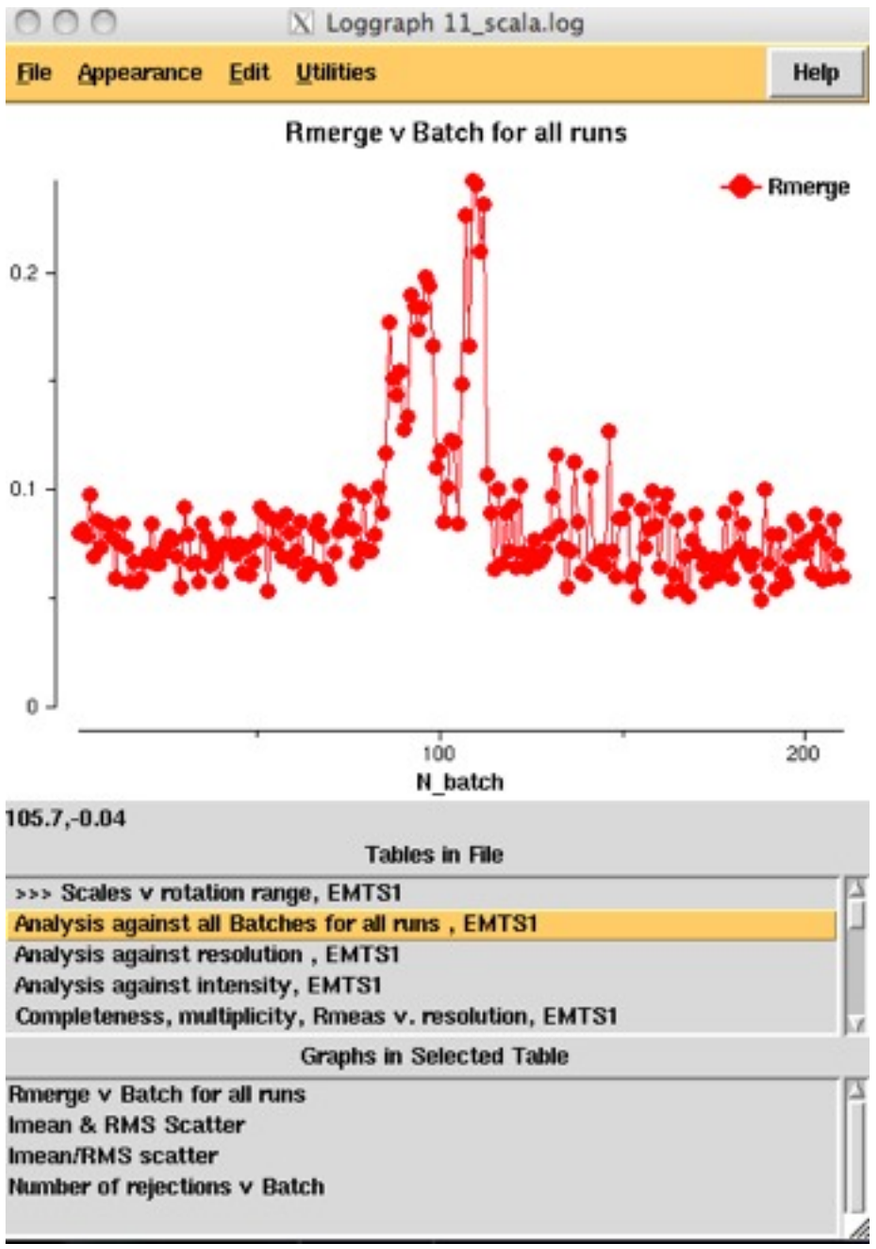
The relative B-factor gives a resolution-dependent scale factor as a function of "time" (dose): average radiation damage decay is greater at high resolution

$$k(time) = \exp[-2B(time)\ \sin^2\theta/\lambda^2]$$

Graph of R_{merge} vs batch may also detect individual bad images, or bad regions, that should be investigated or rejected

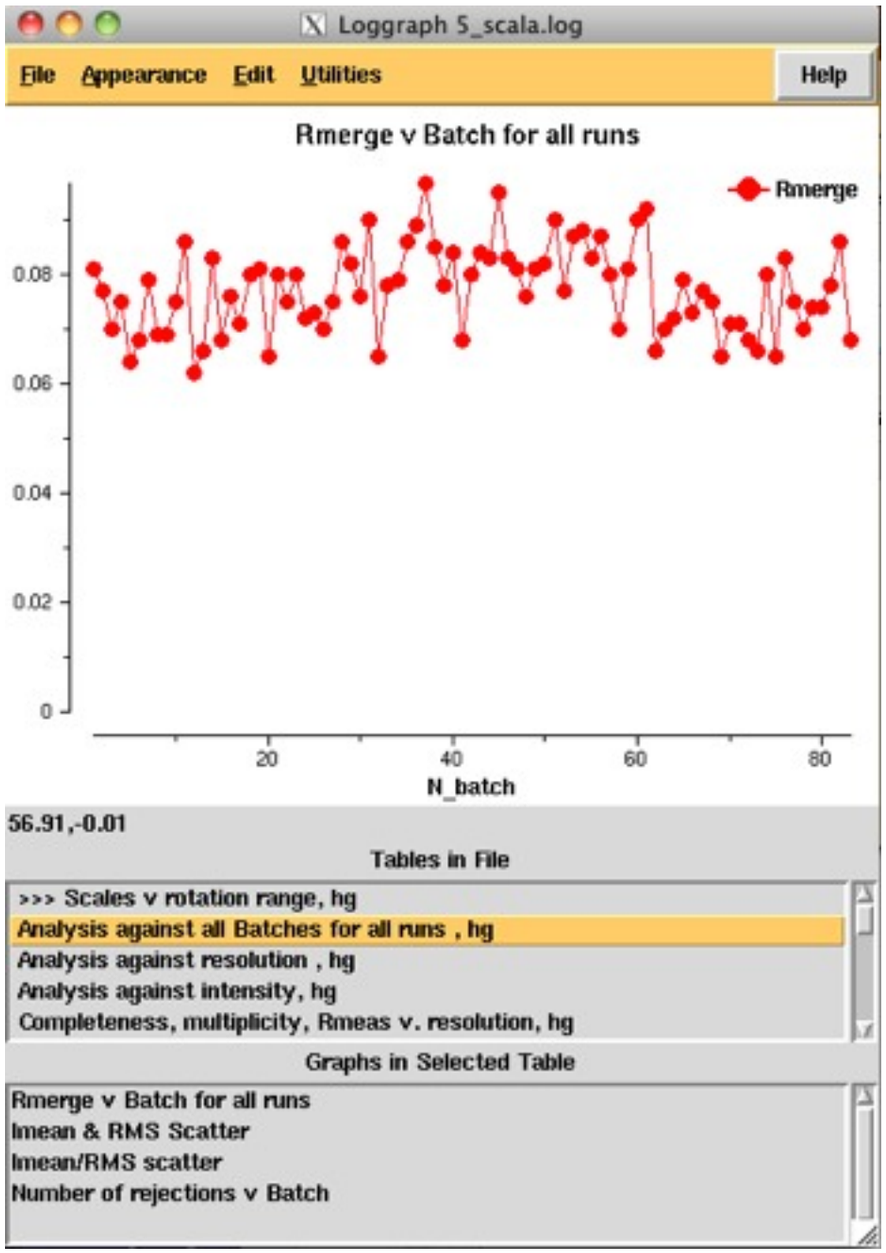


One bad (weak) image

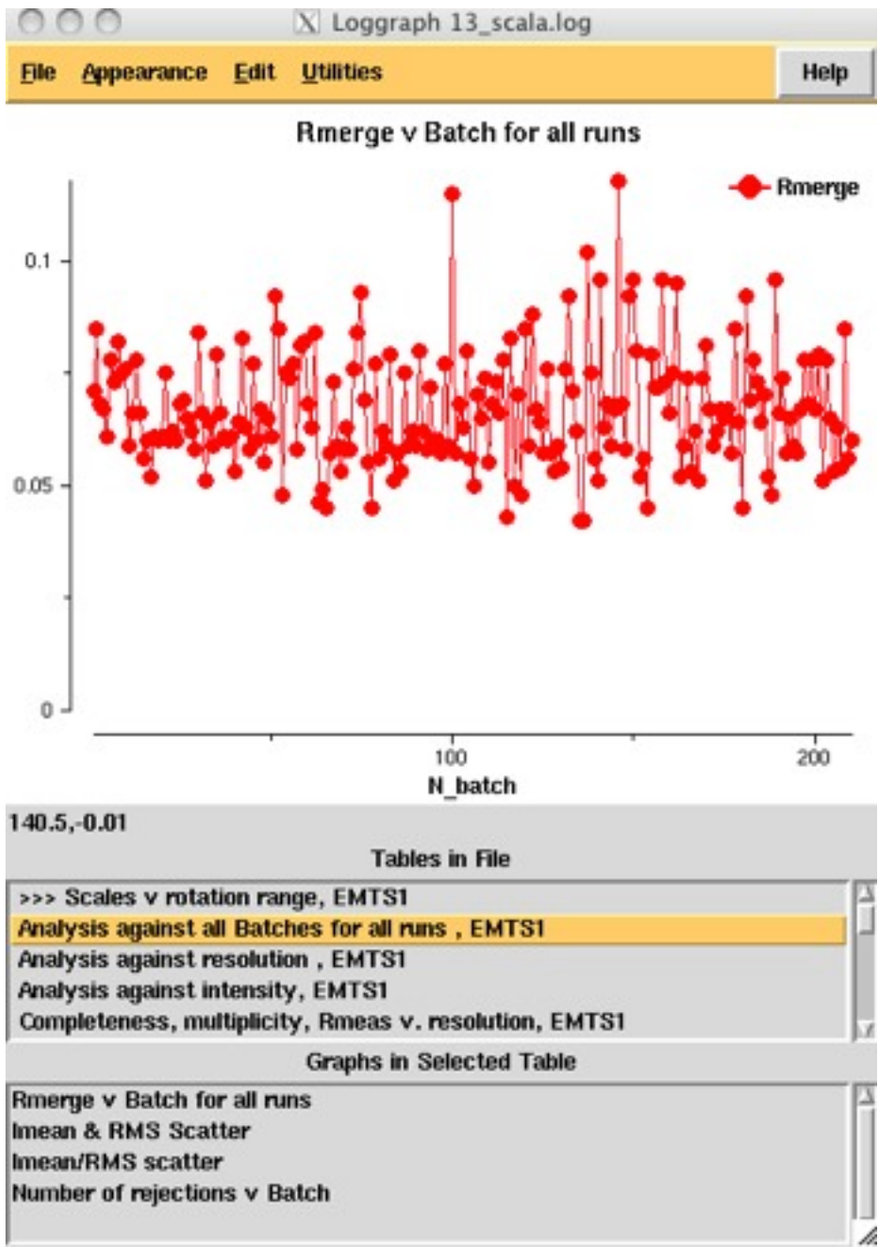


Bad region where integration had gone wrong

Graph of R_{merge} vs batch may also detect individual bad images, or bad regions, that should be investigated or rejected

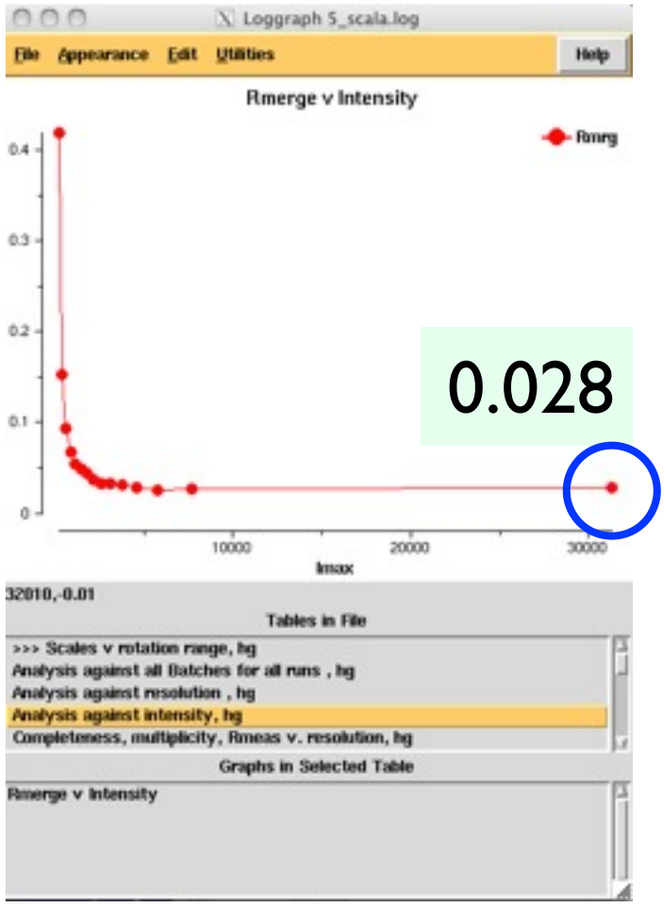


Omitting bad image

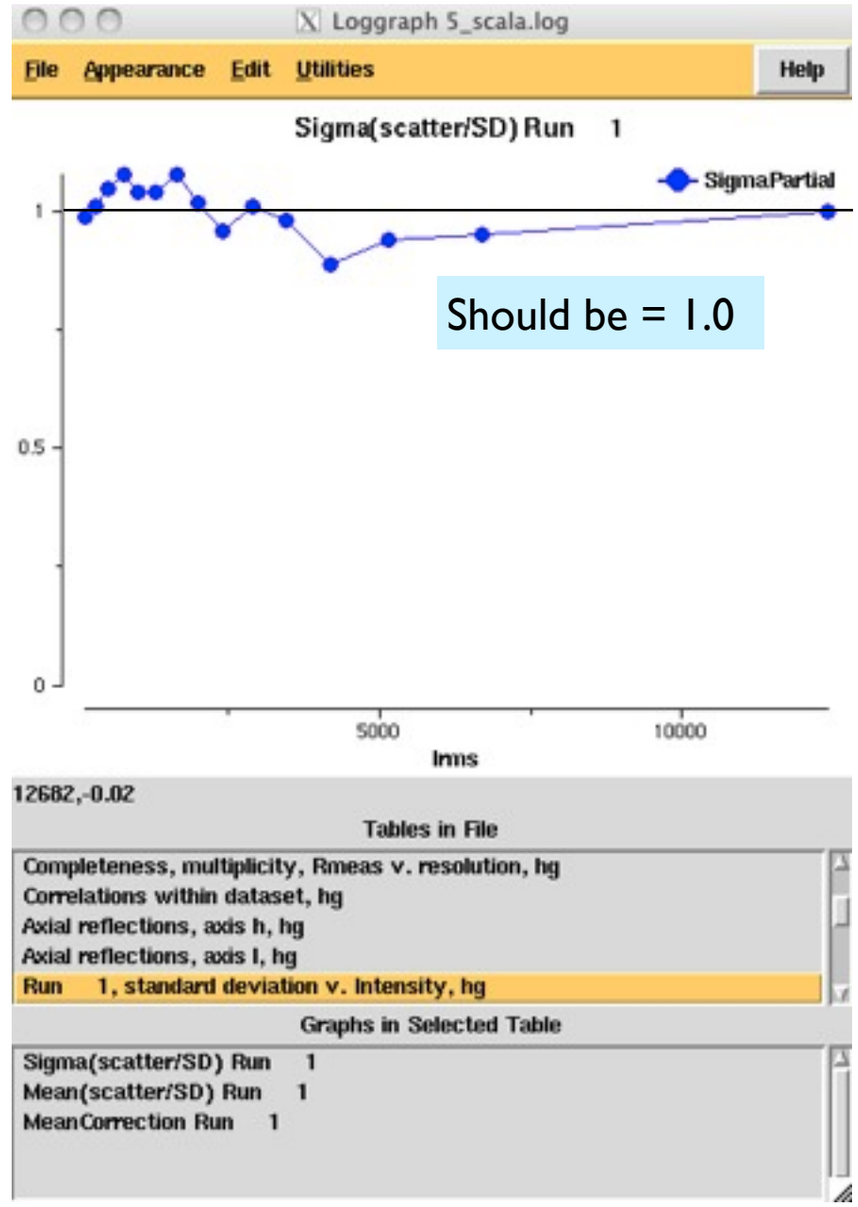


Reprocessed

Analyses against intensity



R_{merge} vs. I not generally useful (since R is a fractional measure, it will always be large for small I), but the value in the top intensity bin should be small



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

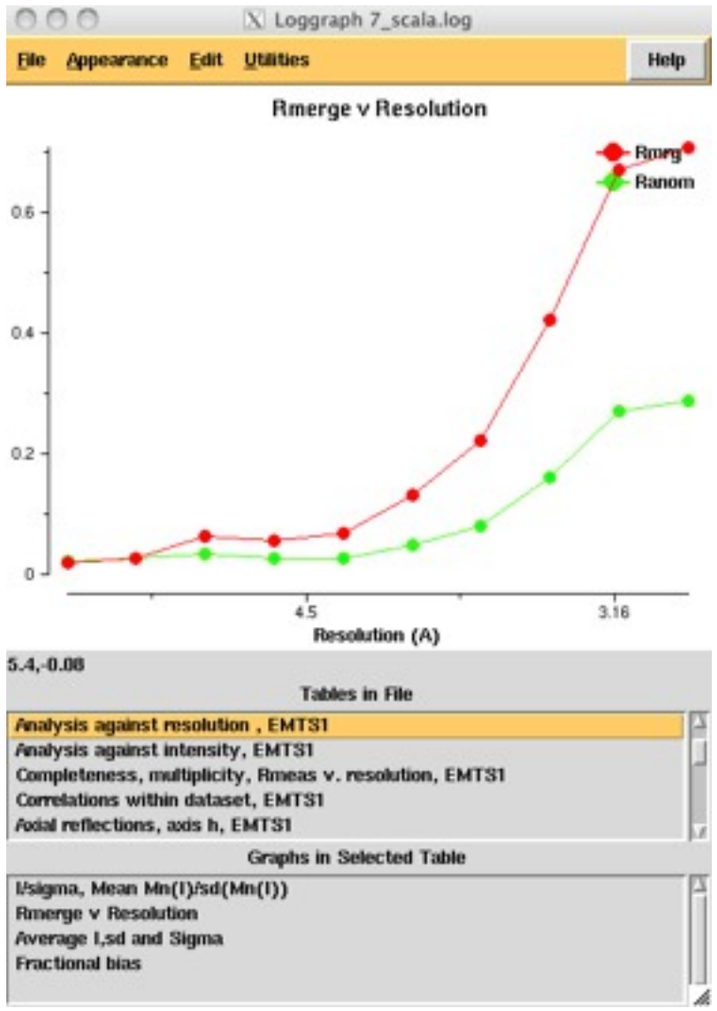
SDfac, SdB and SdAdd are adjustable parameters

Analyses against resolution

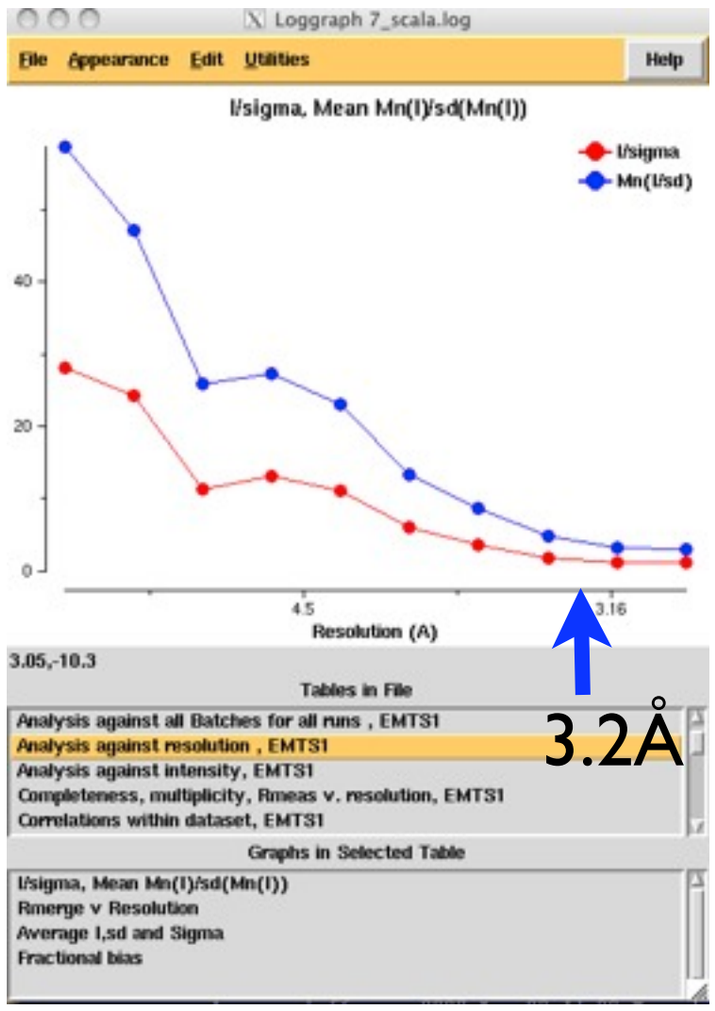
What is the real resolution? not an easy question to answer

May depend on what you want the data for: more stringent for experimental phasing than for refinement

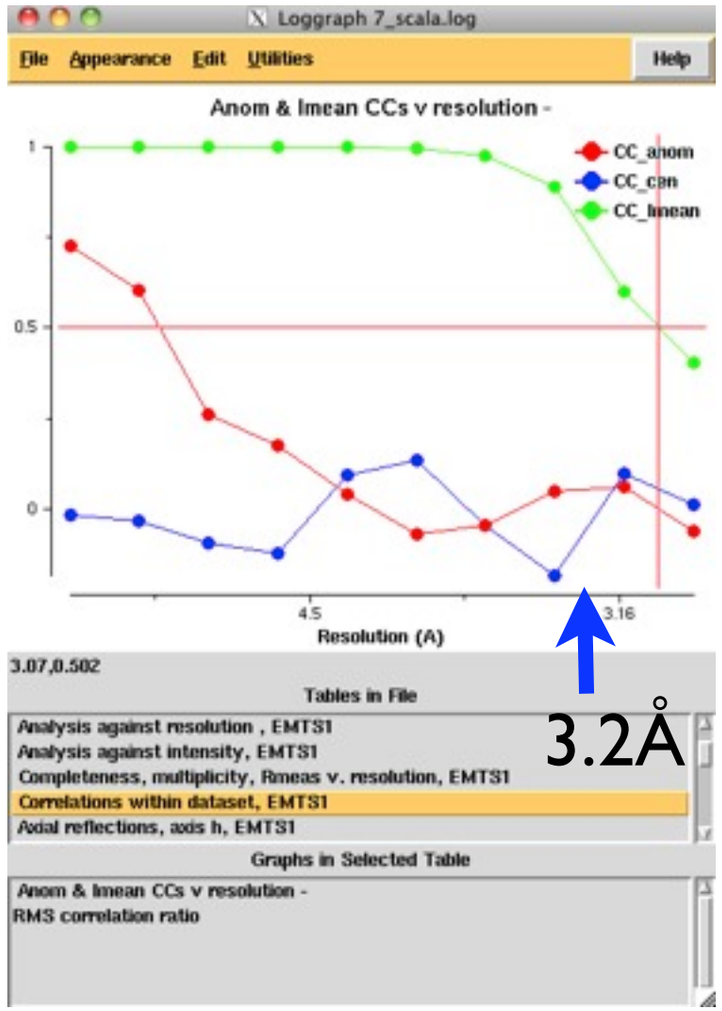
Anisotropic data needs a less stringent overall cut-off to keep best data



R_{merge} is not particularly useful: it gets higher at high resolution



<I/σ(I)> after merging (blue line) should be > ~ 1– 2

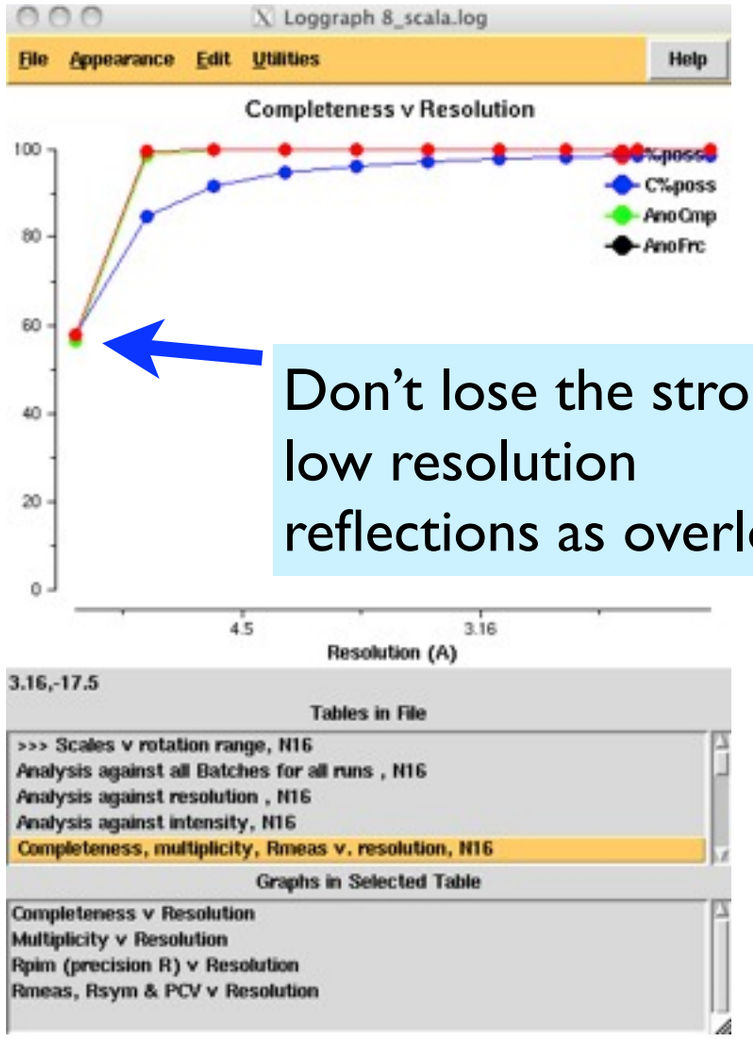


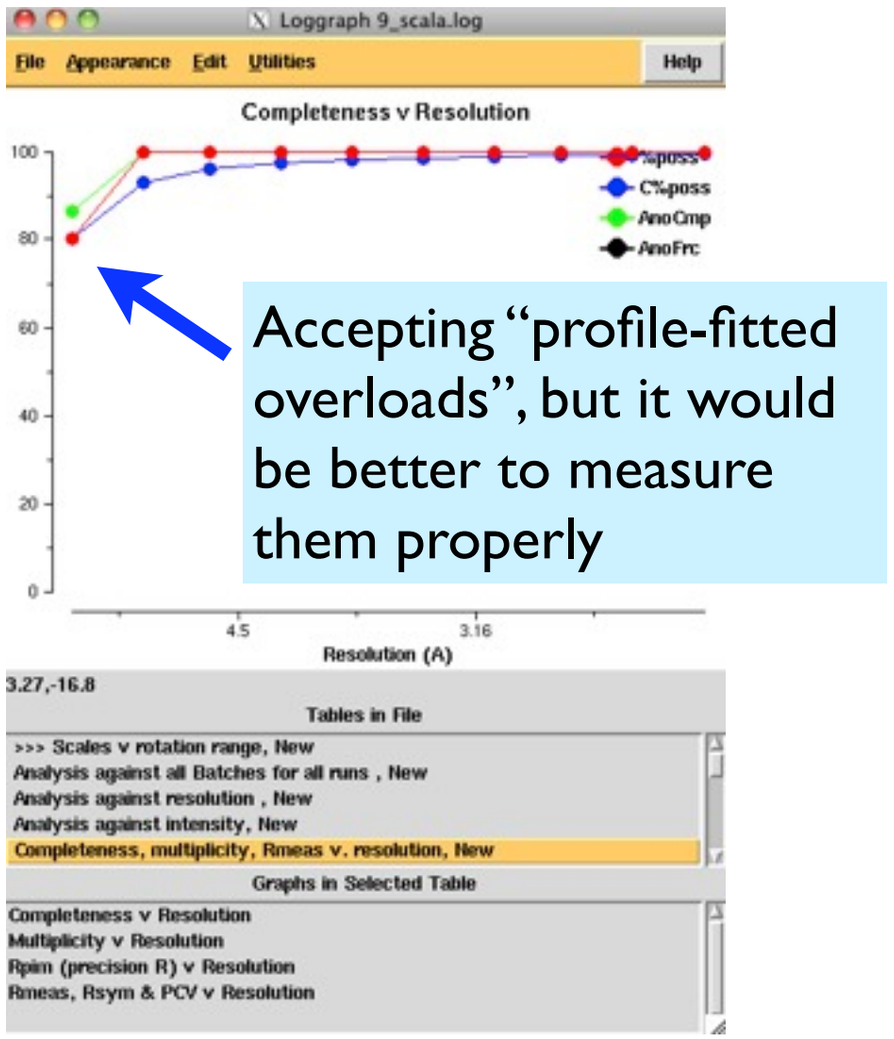
CC between random half-datasets should be > ~ 0.5

Completeness

Data completeness is important, preferably in all resolution shells, though you can probably get away with some incompleteness at the outer edge.

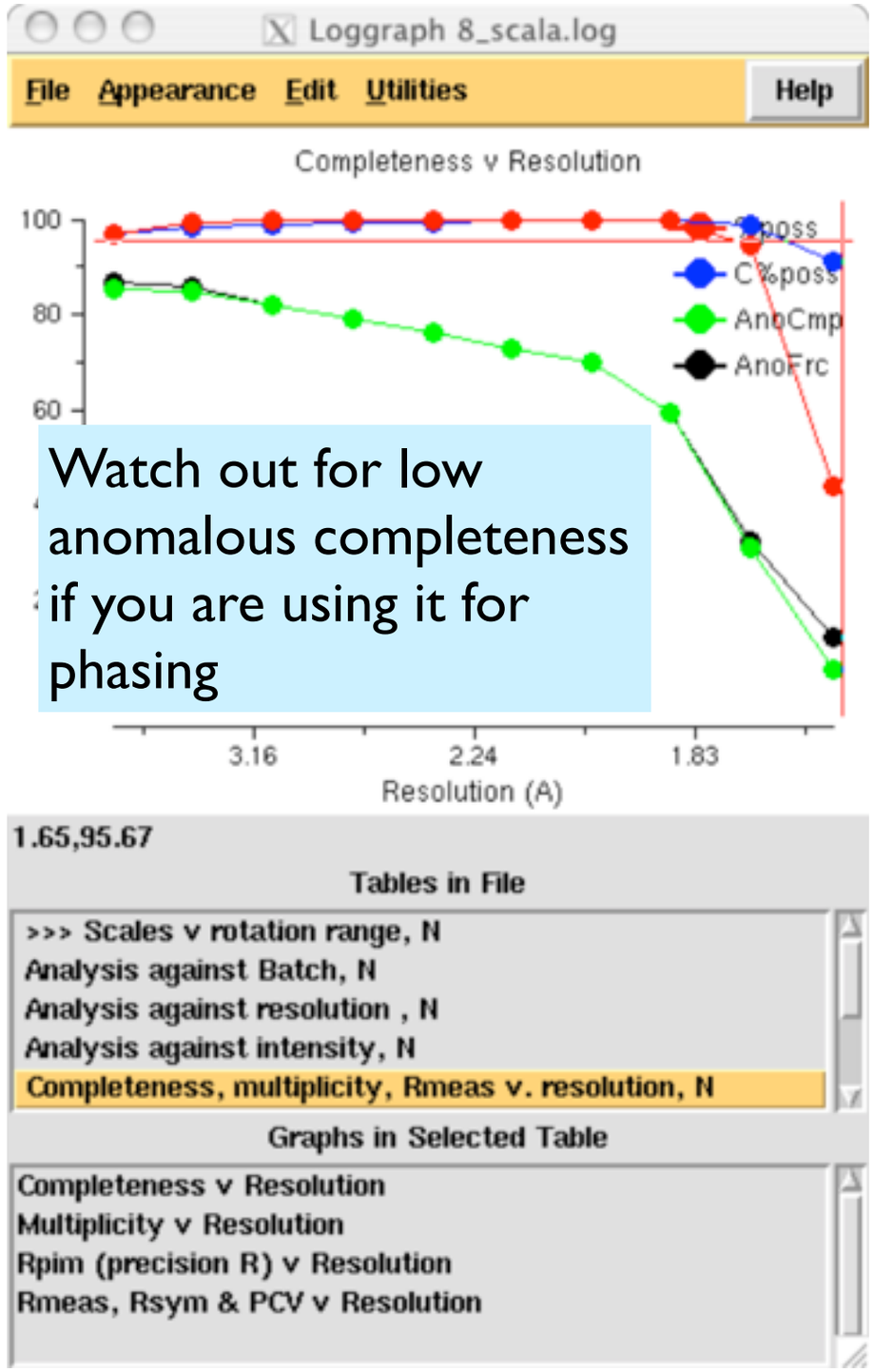
See James Holton's movies for an illustration of the importance of completeness http://ucxray.berkeley.edu/~jamesh/movies/



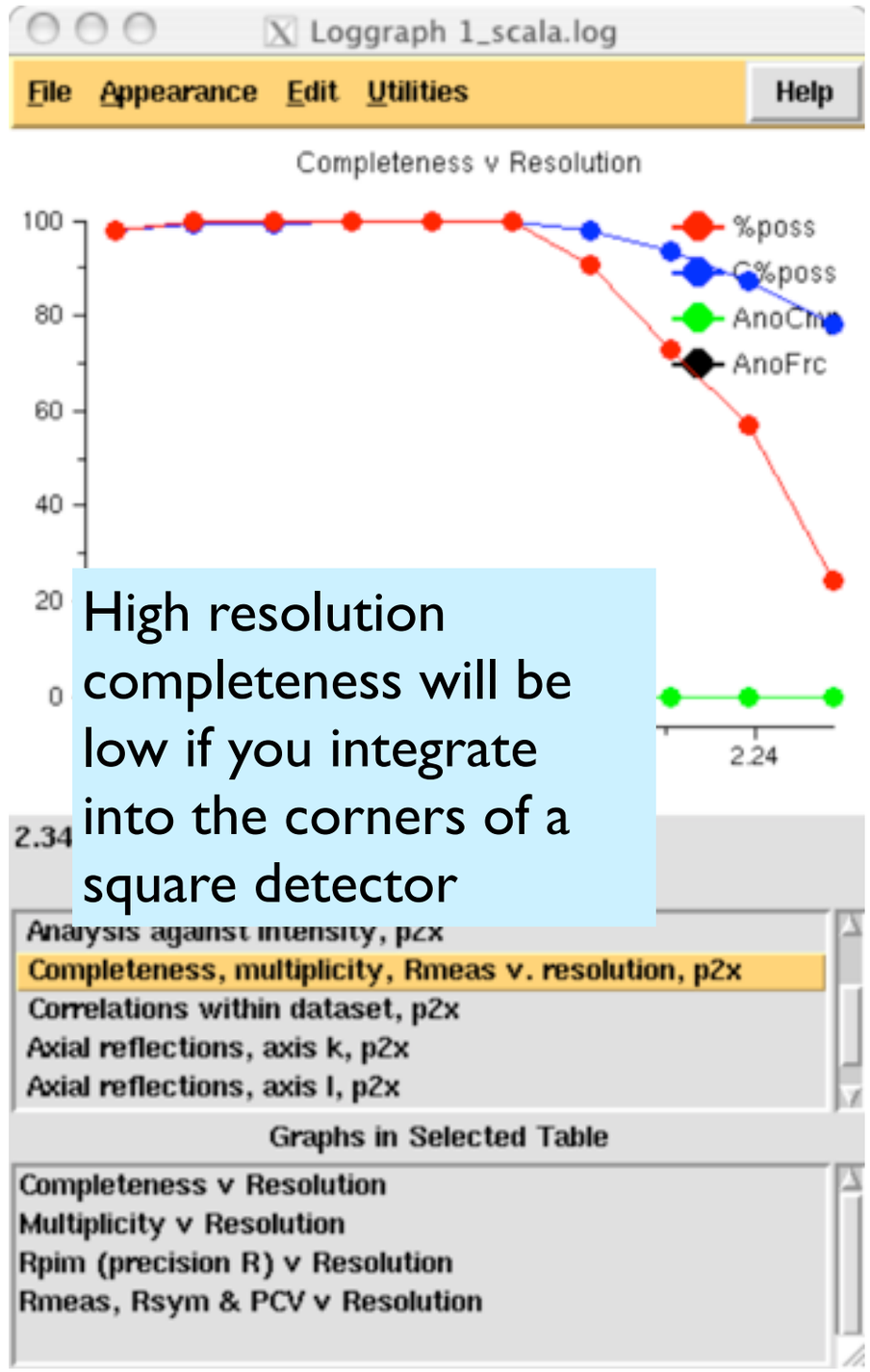


Completeness

Data completeness is important, preferably in all resolution shells, though you can probably get away with some incompleteness at the outer edge.



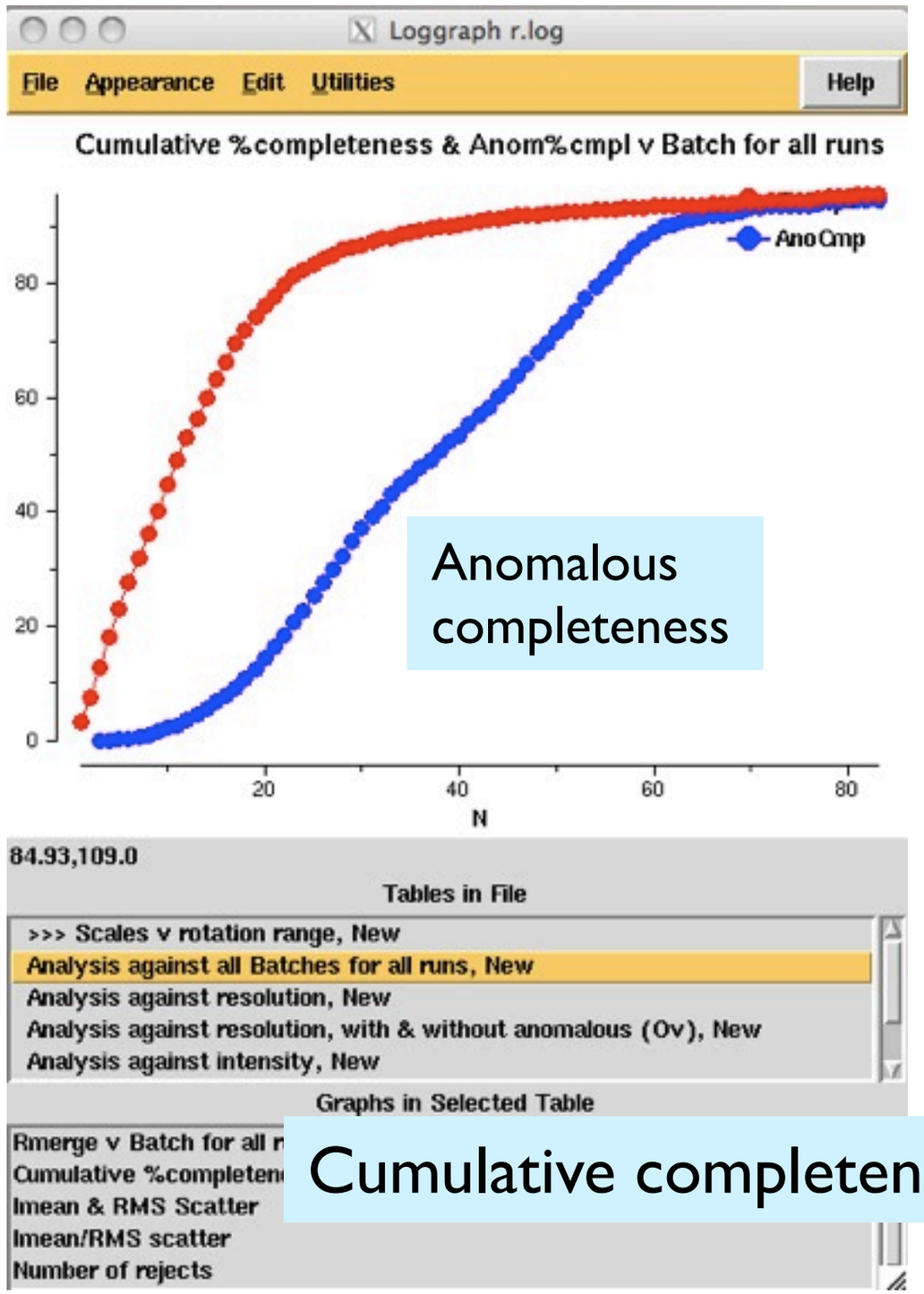
Watch out for low anomalous completeness if you are using it for phasing

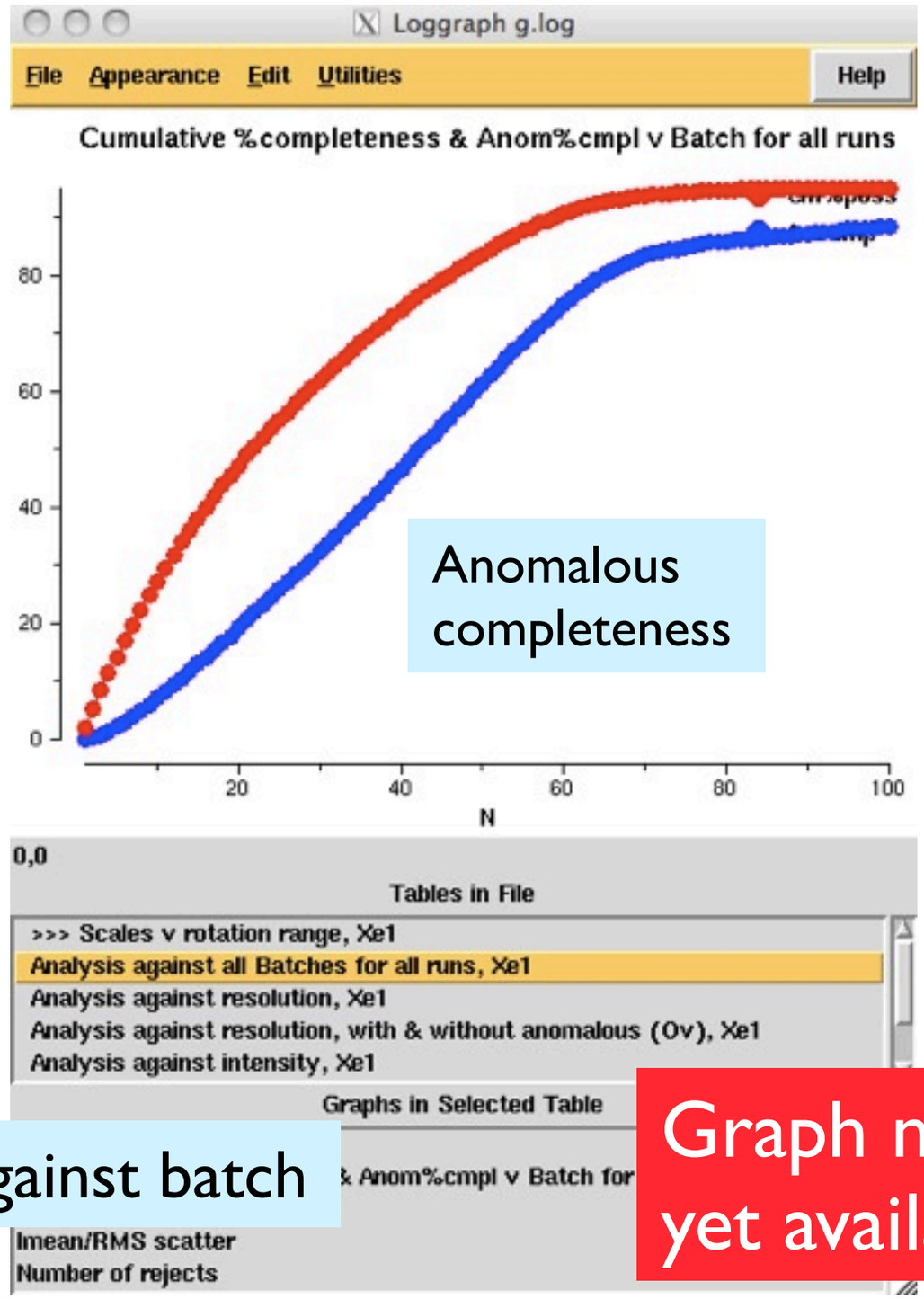


High resolution completeness will be low if you integrate into the corners of a square detector

Completeness

Data completeness is important, preferably in all resolution shells, though you can probably get away with some incompleteness at the outer edge.





Cumulative completeness against batch

Graph not yet available!

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

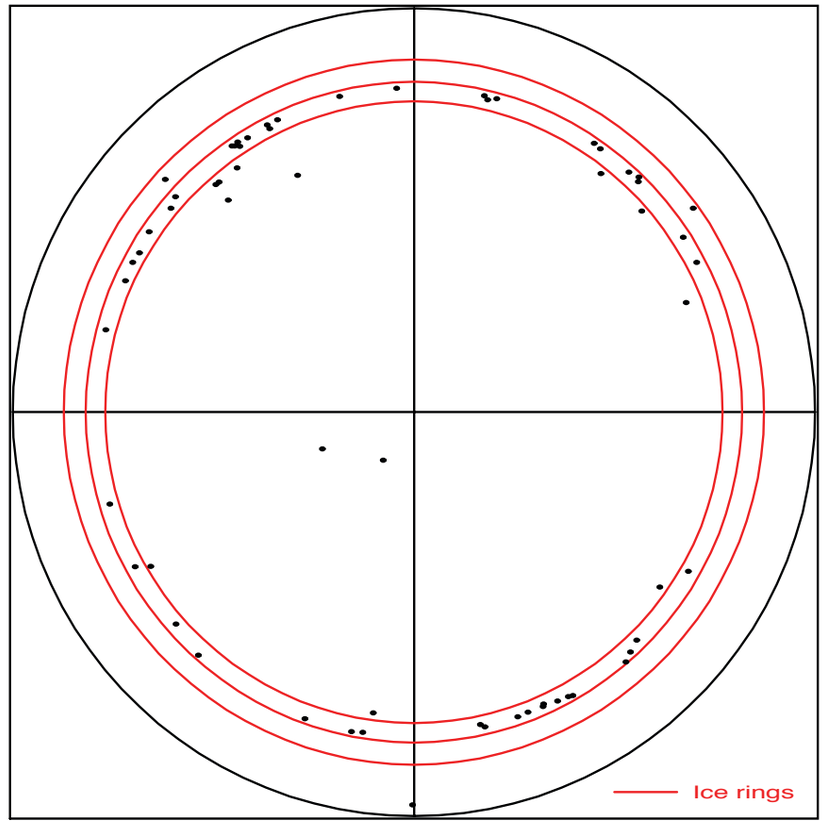
- zingers
- bad prediction (spot not there)

  improve prediction

- spot overlap

  lower mosaicity, smaller slice, move detector back

  deconvolute overlaps

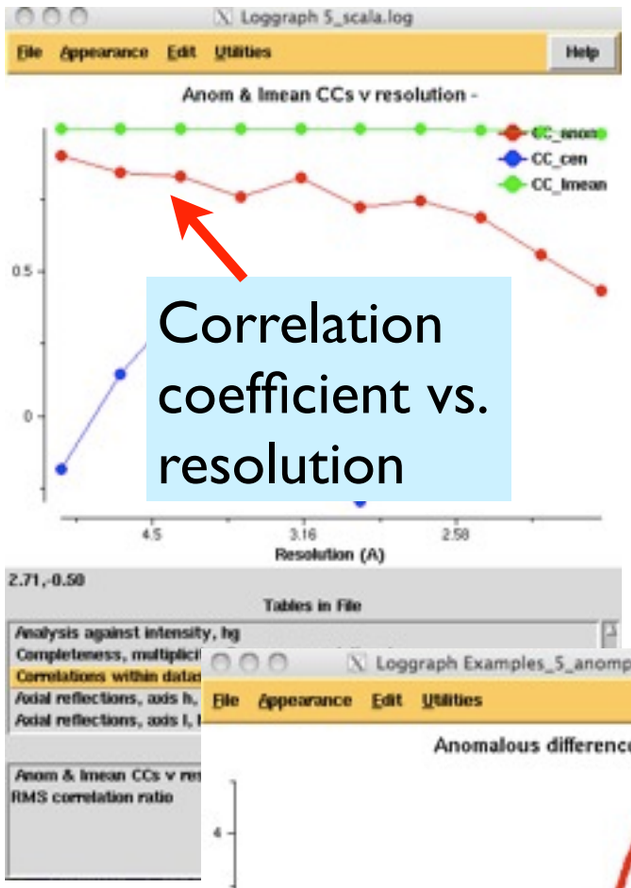


Rejects lie on ice rings (red) (ROGUEPLOT in Scala)

Position of rejects on detector

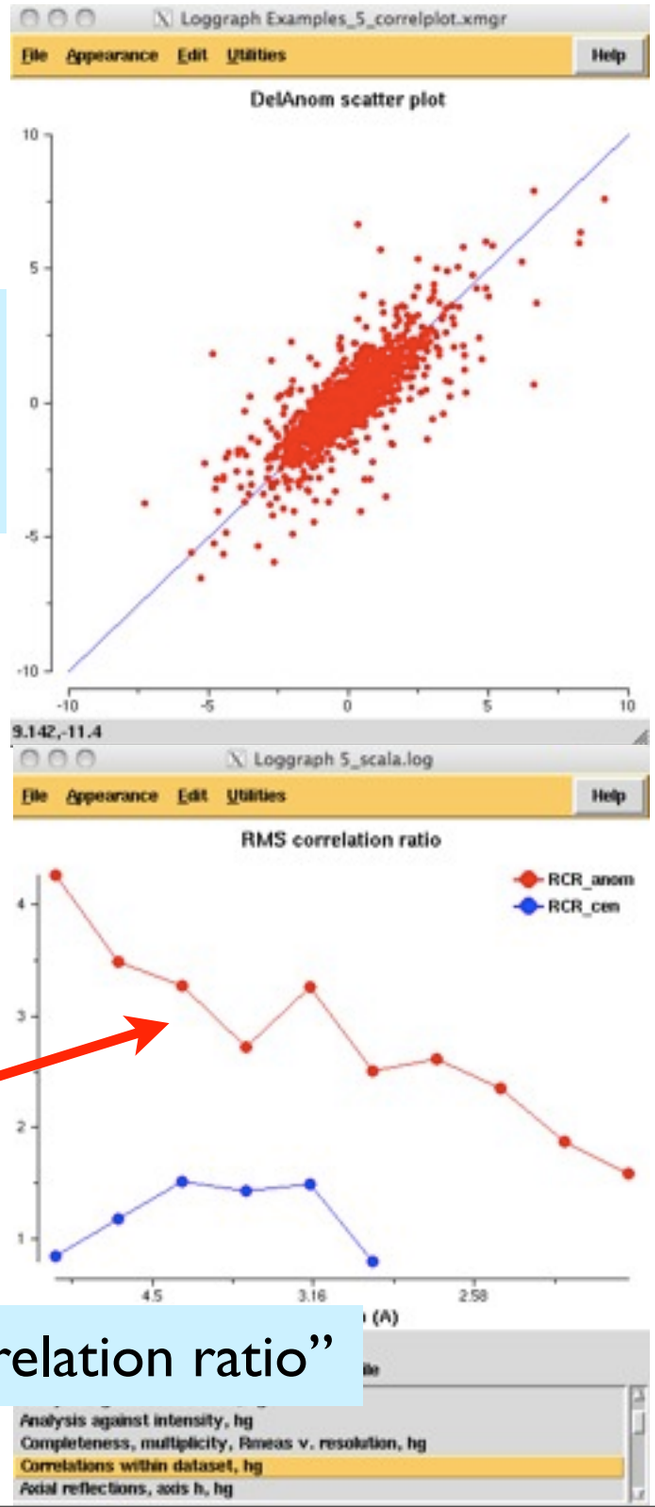
Detecting anomalous signals

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

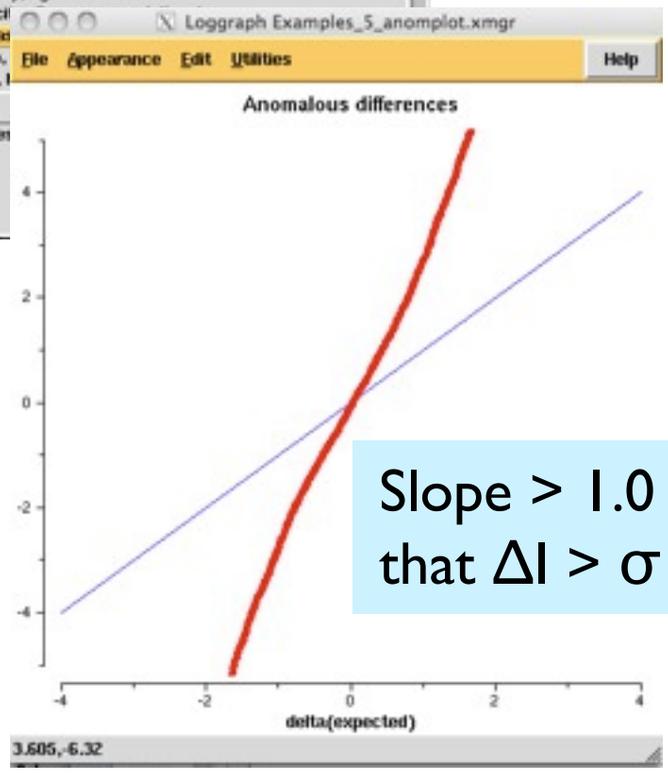


Correlation coefficient vs. resolution

Split one dataset randomly into two halves, calculate correlation between the two halves



Plot ΔI_1 against ΔI_2 should be elongated along diagonal



Strong anomalous signal

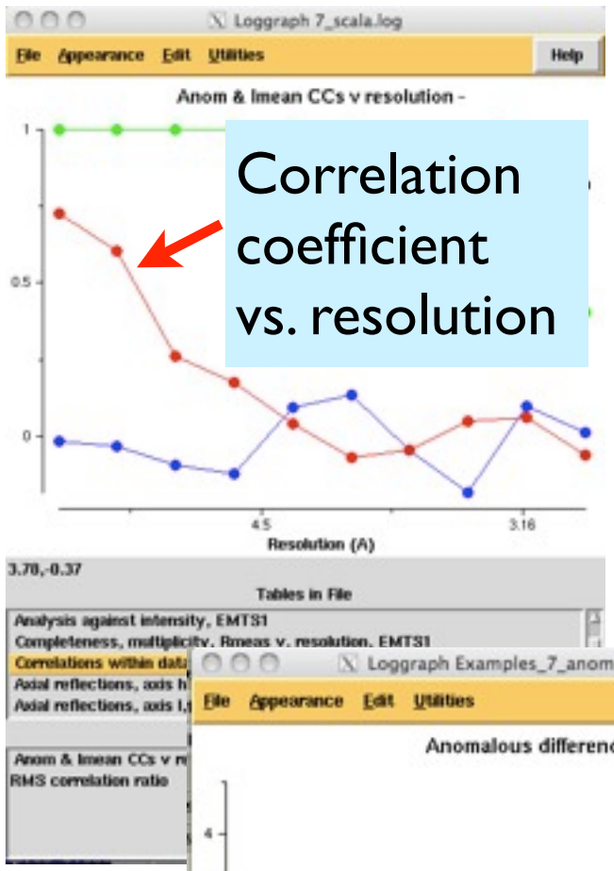
Slope > 1.0 means that $\Delta I > \sigma$

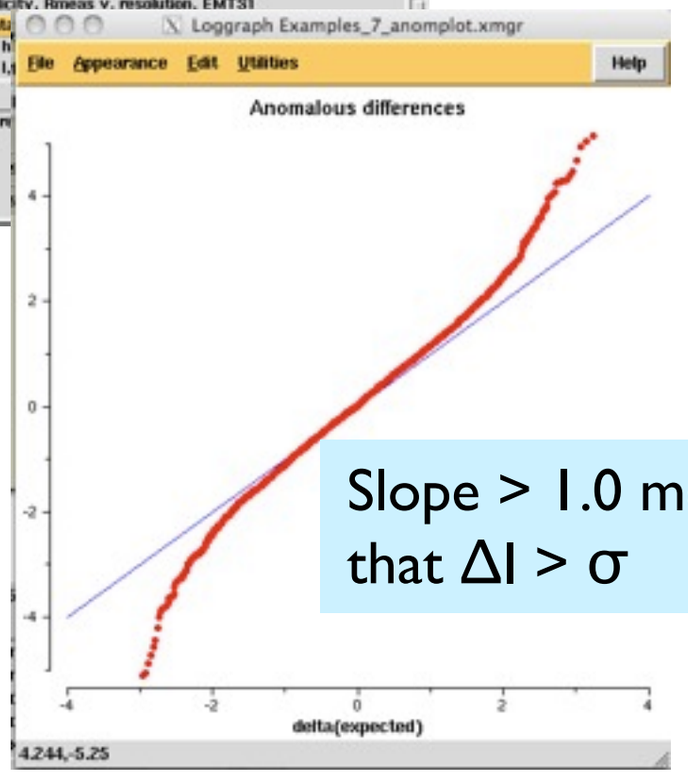
Ratio of width of distribution along diagonal to width across diagonal

"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.





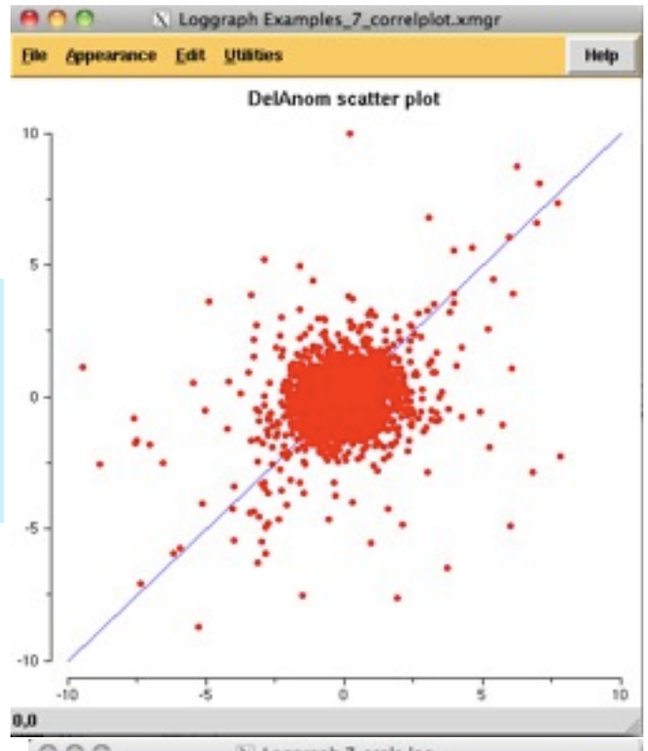
Slope > 1.0 means that $\Delta I > \sigma$

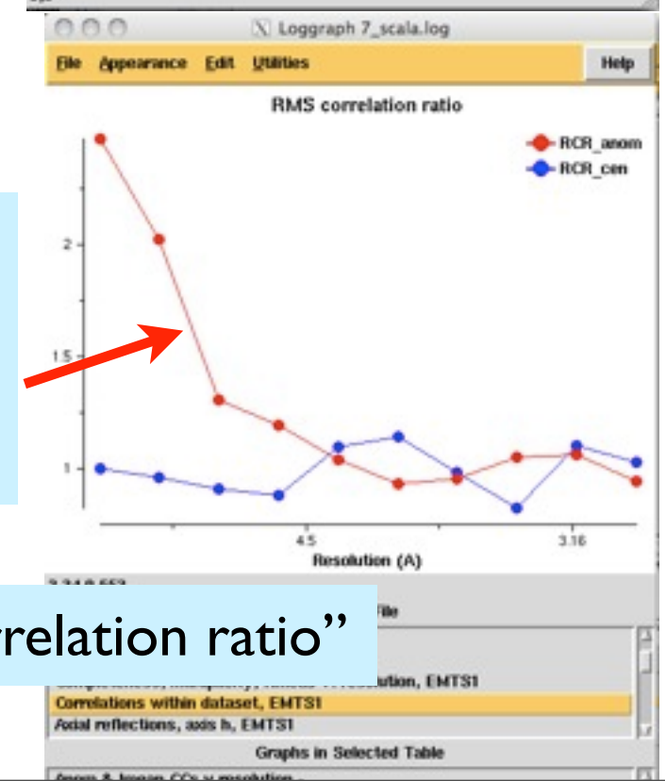
Split one dataset randomly into two halves, calculate correlation between the two halves

Plot ΔI_1 against ΔI_2 should be elongated along diagonal

Weak but useful anomalous signal

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

Scale symmetry-related intensities together
Produce statistics on data quality

merged

Scaled and averaged
intensities

CTRUNCATE

Estimate |F| from I
detect twinning

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

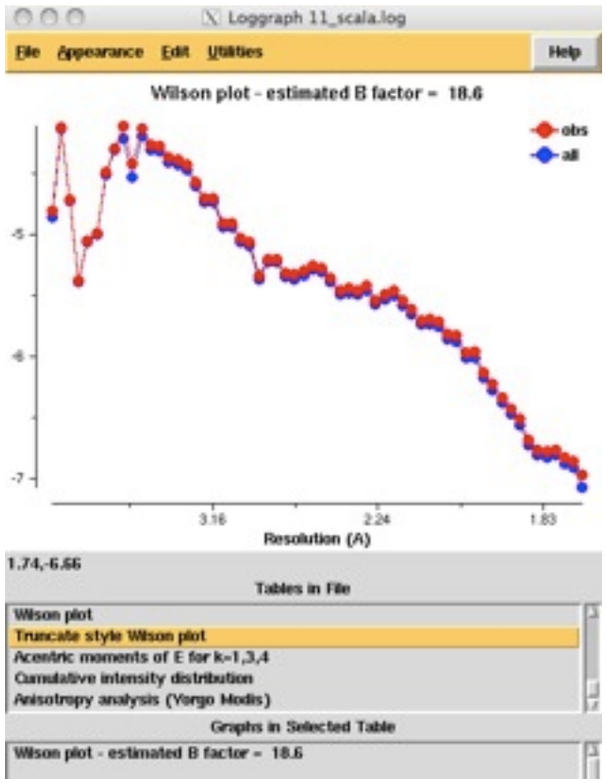
Intensity statistics

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 $s = \sin\theta/\lambda$



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

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

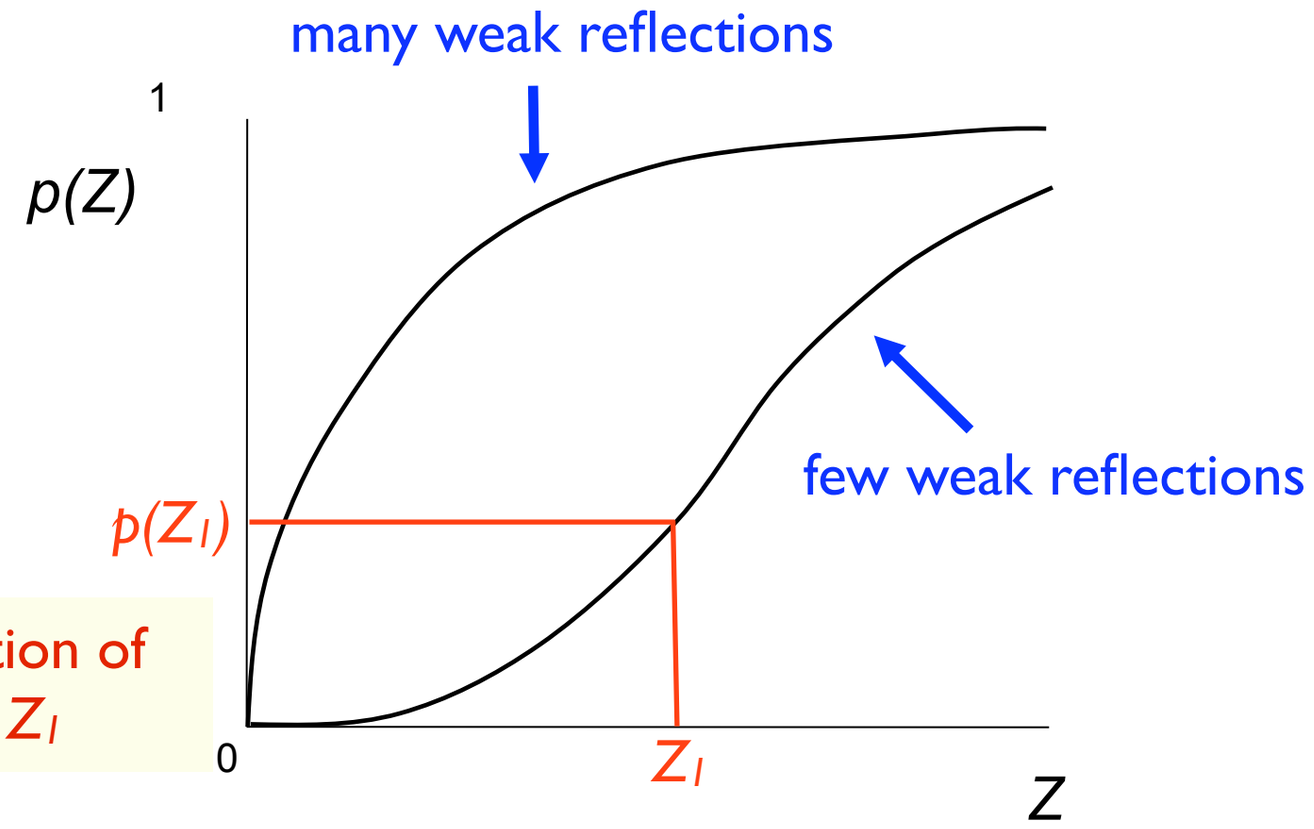
For the purposes of looking for crystal pathologies, we are not interested in the variation with resolution, so we can use "normalised" intensities which are independent of resolution

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)

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



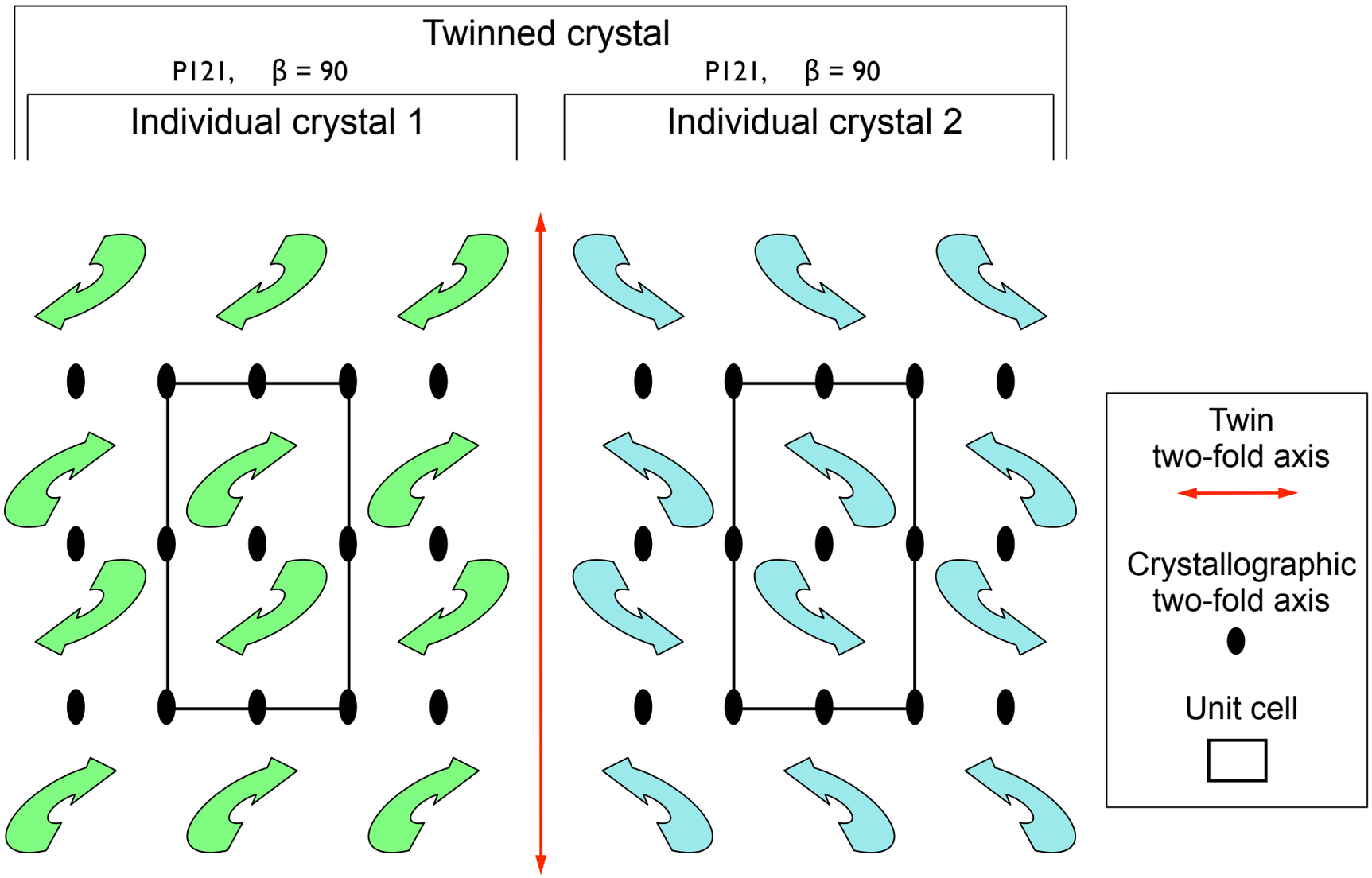
$p(Z_I)$ is the proportion of reflections with $Z < Z_I$

Twinning by (pseudo)merohedry

Two crystals whose lattices overlap (nearly) exactly: this can happen when the true symmetry is lower than the lattice symmetry

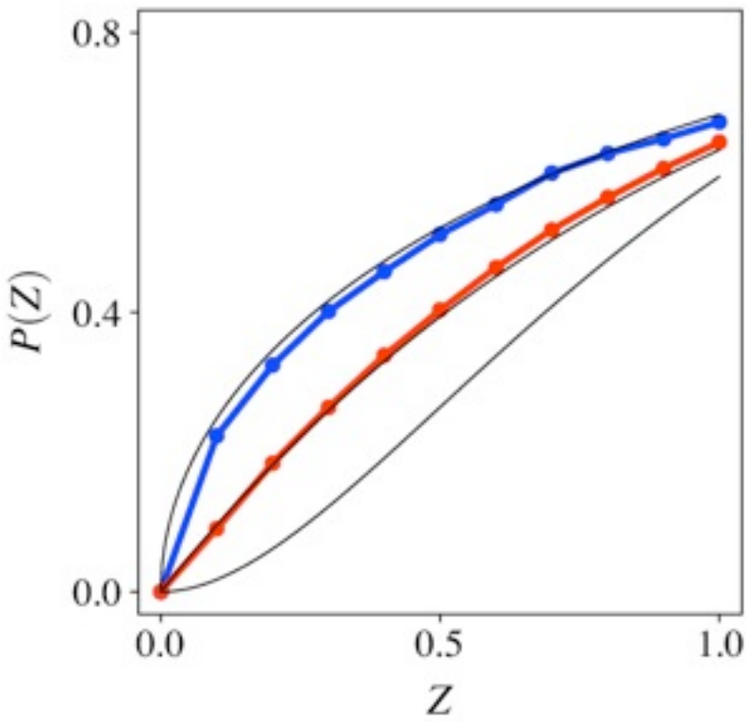
Measured intensities are the *sum* of two different reflections related by the twin operator, so a weak intensity is likely to be inflated by a stronger one

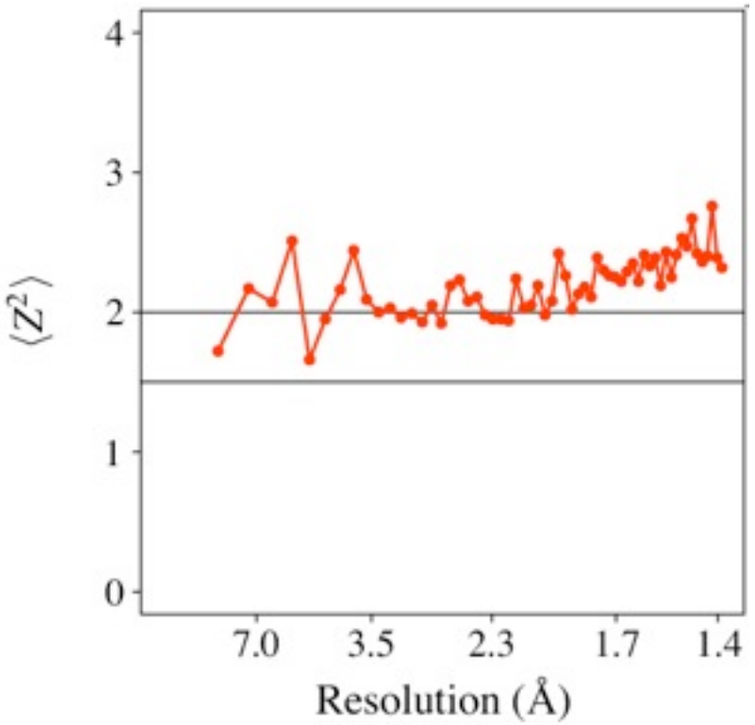
too few weak intensities



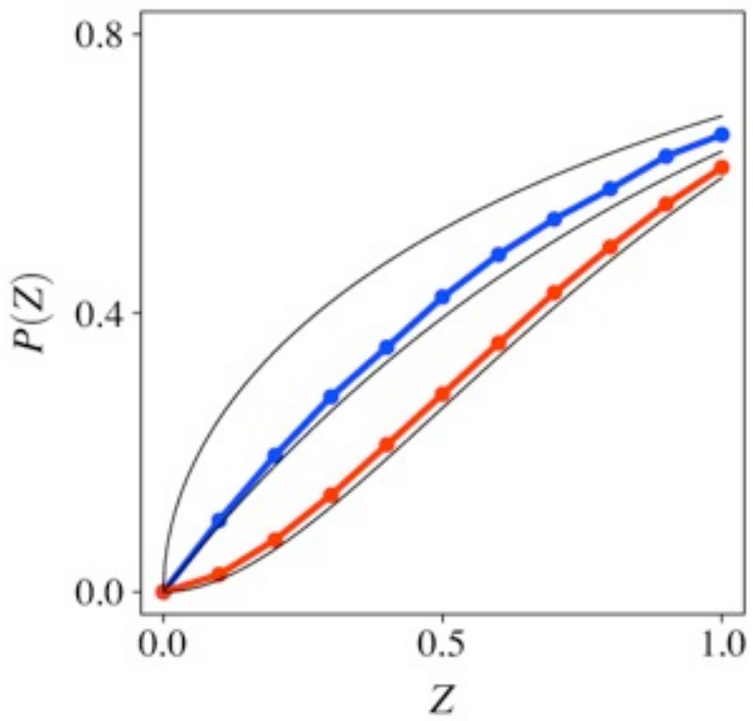
Andrey Lebedev

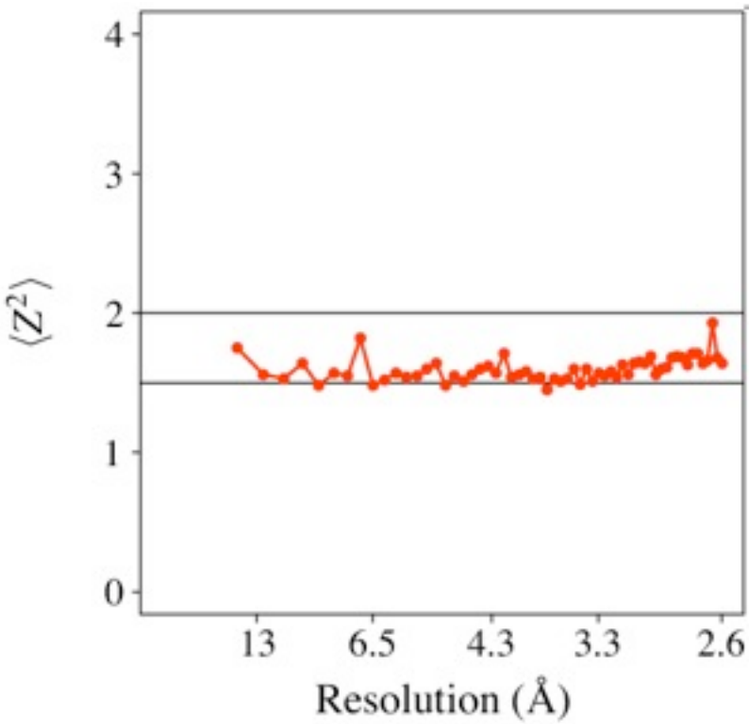
Examples





PDB entry 1i1j
single crystal





C-terminal domain of gp2 protein from phage SPP1 (unpublished)
perfect twin

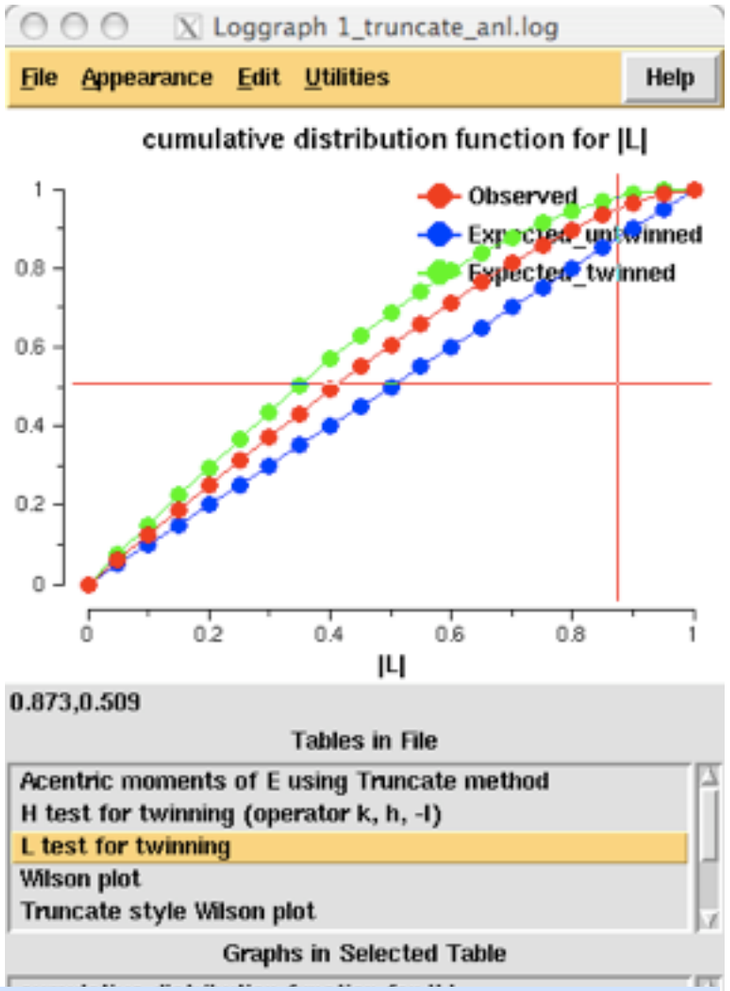
Cumulative intensity distribution

2nd moment of Z or $\langle E^4 \rangle$

Andrey Lebedev

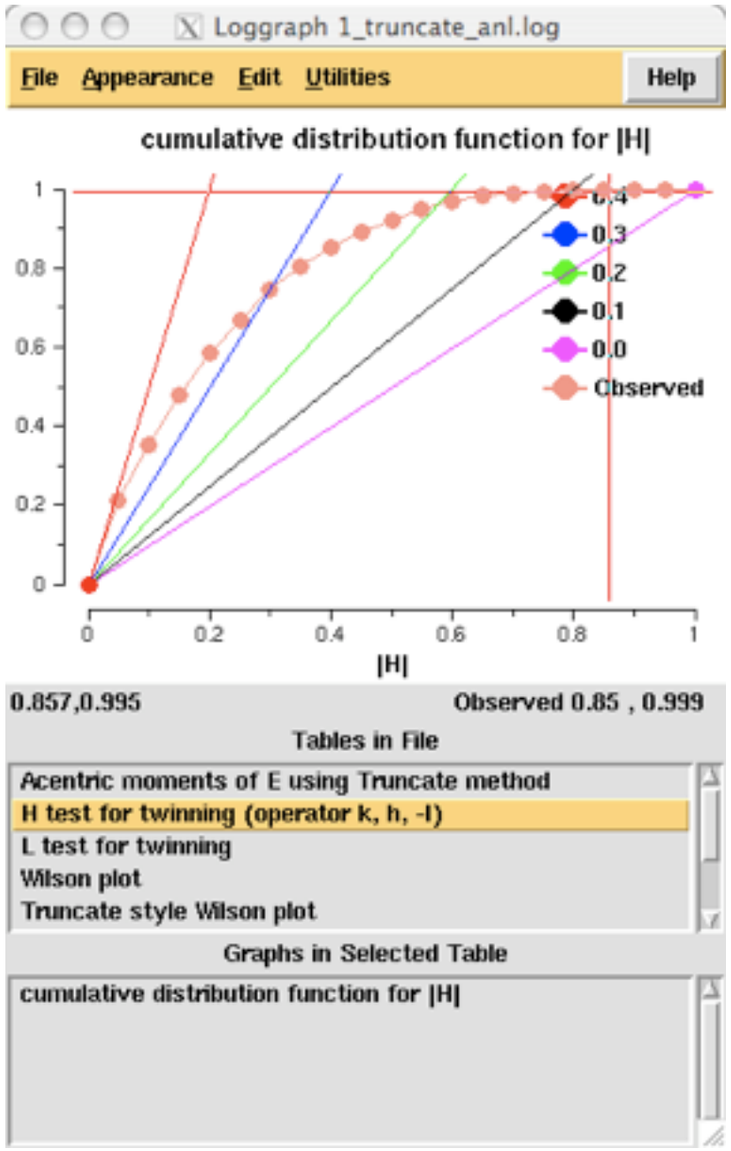
Ctruncate: L- and H-tests

Cumulative distribution of L
(L-test)



L-test is probably the most reliable test for twinning

Cumulative distribution of H
(H-test)
(Partial twinning test)



L = (I1 - I2)/(I1 + I2)
I1 & I2 close in reciprocal space

H = (I1 - I2)/(I1 + I2)
I1 & I2 related by twin symmetry

Andrey Lebedev

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

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

Summary

Questions & Decisions

- What is the point group (Laue group)?
- What is the space group?
- Is there radiation damage: should data be cut away from the end (possibly at the expense of resolution)?
- What is the best resolution cut-off?
- Is there anomalous signal (if you expect one)?
- Are the data twinned?
- Is this dataset better or worse than ones you have already?

Acknowledgements

Andrew Leslie	many discussions
Harry Powell	many discussions
Ralf Grosse-Kunstleve	cctbx
Kevin Cowtan	clipper, simplex minimiser, C++ advice
Martyn Winn & CCP4 gang	ccp4 libraries
Peter Briggs	ccp4i
Airlie McCoy	C++ advice, code etc
Graeme Winter	testing & bug finding
Clemens Vonrhein	testing & bug finding
Eleanor Dodson	many discussions
Andrey Lebedev	intensity statistics & twinning
Norman Stein	ctruncate
George Sheldrick	discussions on symmetry detection