

# **Data Reduction**

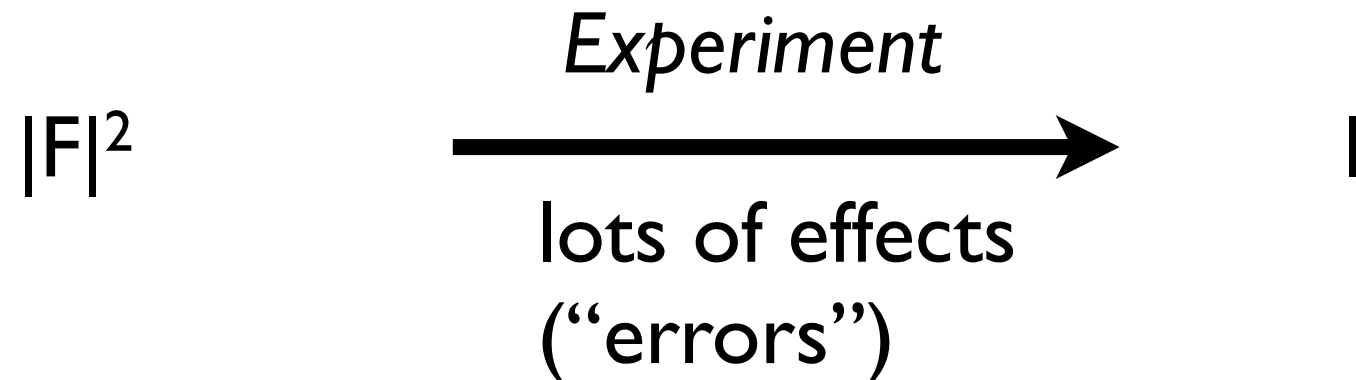
## **Space Group Determination, Scaling and Intensity Statistics**

*Phil Evans    Fukuoka October 2012*

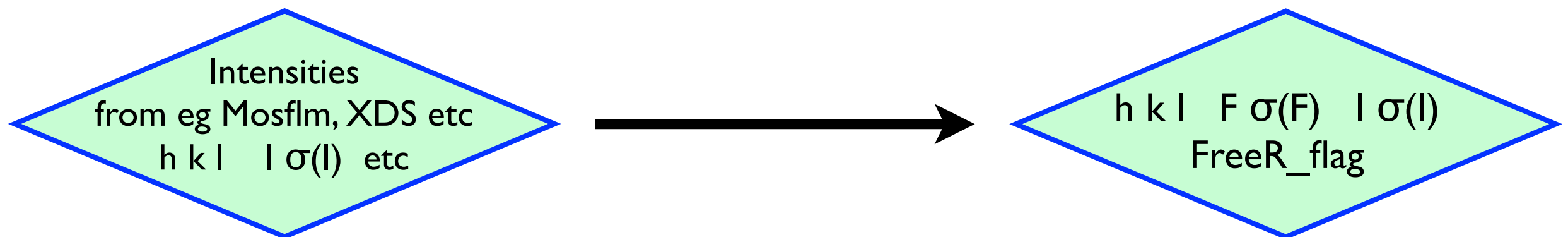
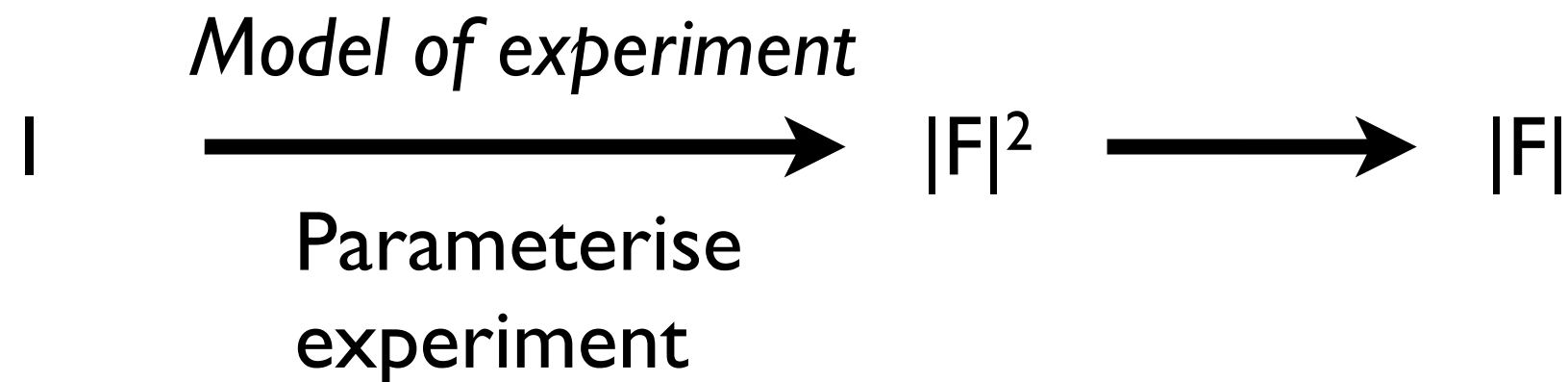
*MRC Laboratory of Molecular Biology  
Cambridge UK*



# Scaling and Merging



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



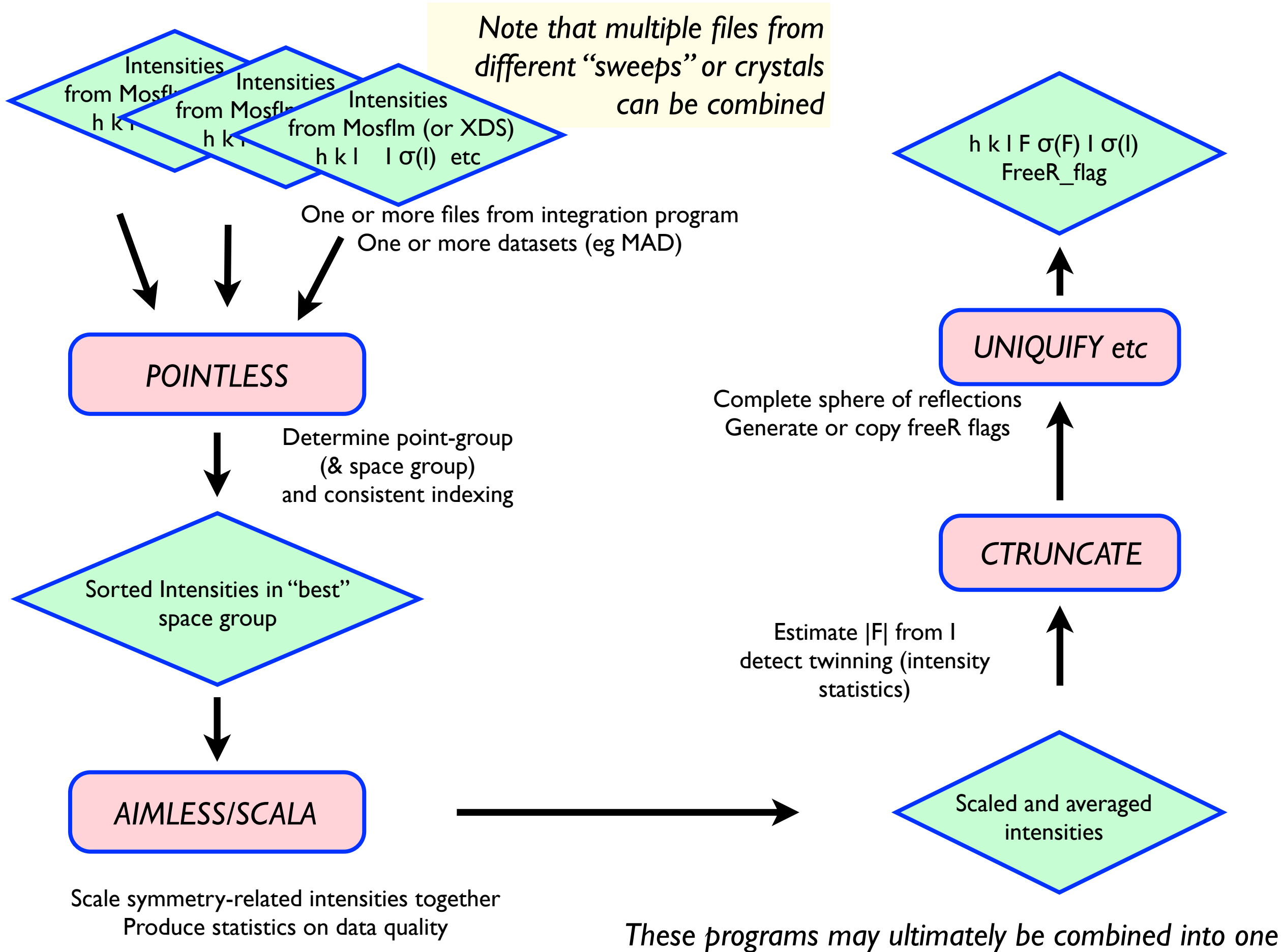


Integration and data reduction can be done in an automated pipeline such as XIA2 (Graeme Winter): this goes from images to a list of hkl F ready for structure determination)

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

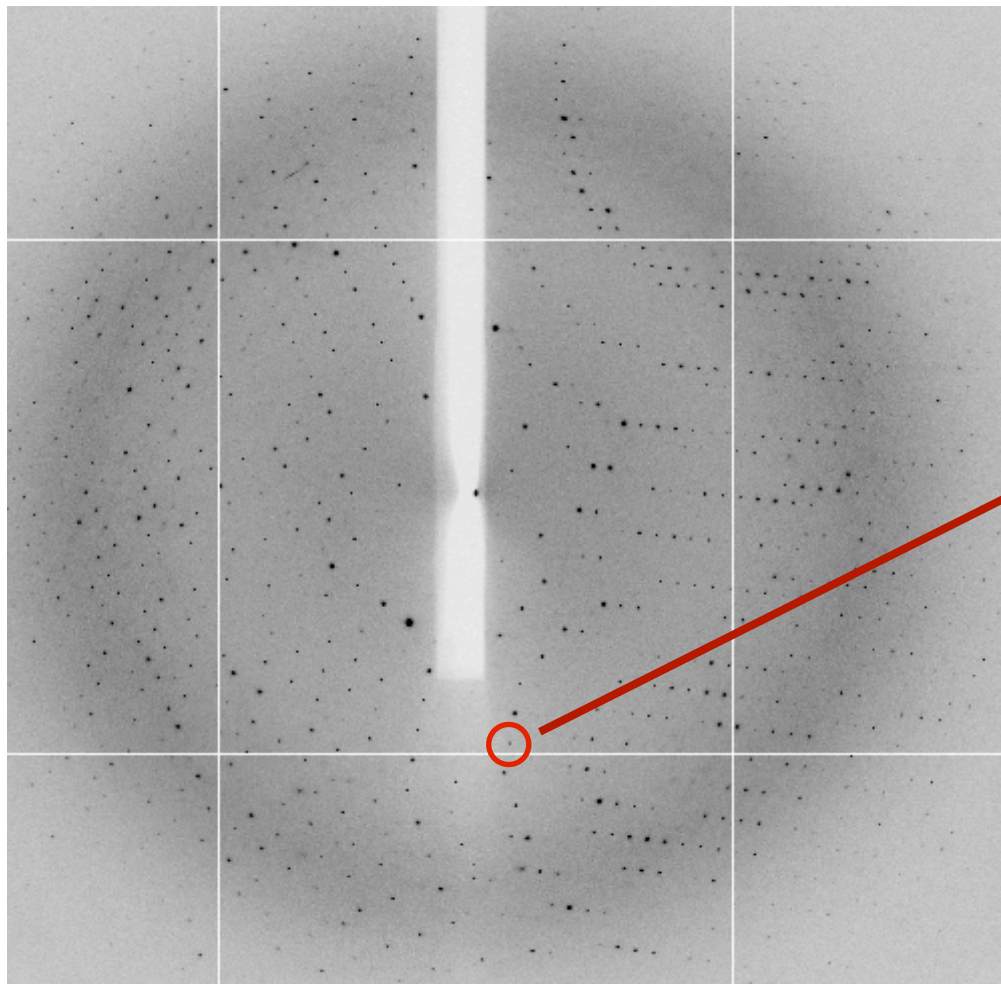


# Data flow scheme in CCP4

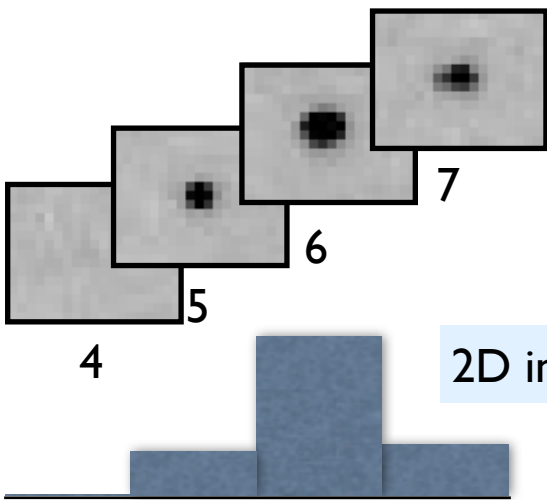




# Track one reflection through the process



Spot over 4 images



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



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

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

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

AIMLESS

scale and merge

Merged file, one line for each hkl

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

Optional unmerged output  
Partials summed, scaled, outliers rejected

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

CTRUNCATE

Infer |F| from I

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



# 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

## Stage 1: lattice symmetry

From the unit cell dimensions we can deduce the maximum possible lattice symmetry, ie the crystal system: this the only information available to the integration program (Mosflm)

Systems are :

cubic, hexagonal/trigonal, tetragonal, orthorhombic, monoclinic, or triclinic,  
+ lattice centring P, C, I, R, or F

For example if  $a = b = c$  and  $\alpha = \beta = \gamma = 90^\circ$  (within some tolerance eg  $2^\circ$ ) then the maximum lattice symmetry is cubic



# Stage 2: individual rotational symmetry operators

Ignore symmetry from integration program

Test all rotation operators consistent with the lattice symmetry

Score agreement between reflections related by each operator, by R-factor ( $R_{\text{meas}}$ ) and correlation coefficient (CC) on normalised intensities  $|E|^2$

Calculate a “probability” based on the CC

## Pseudo-cubic example

Cell: 79.15 81.33 81.15 90.00 90.00 90.00  $a \approx b \approx c$

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



## What score to use?

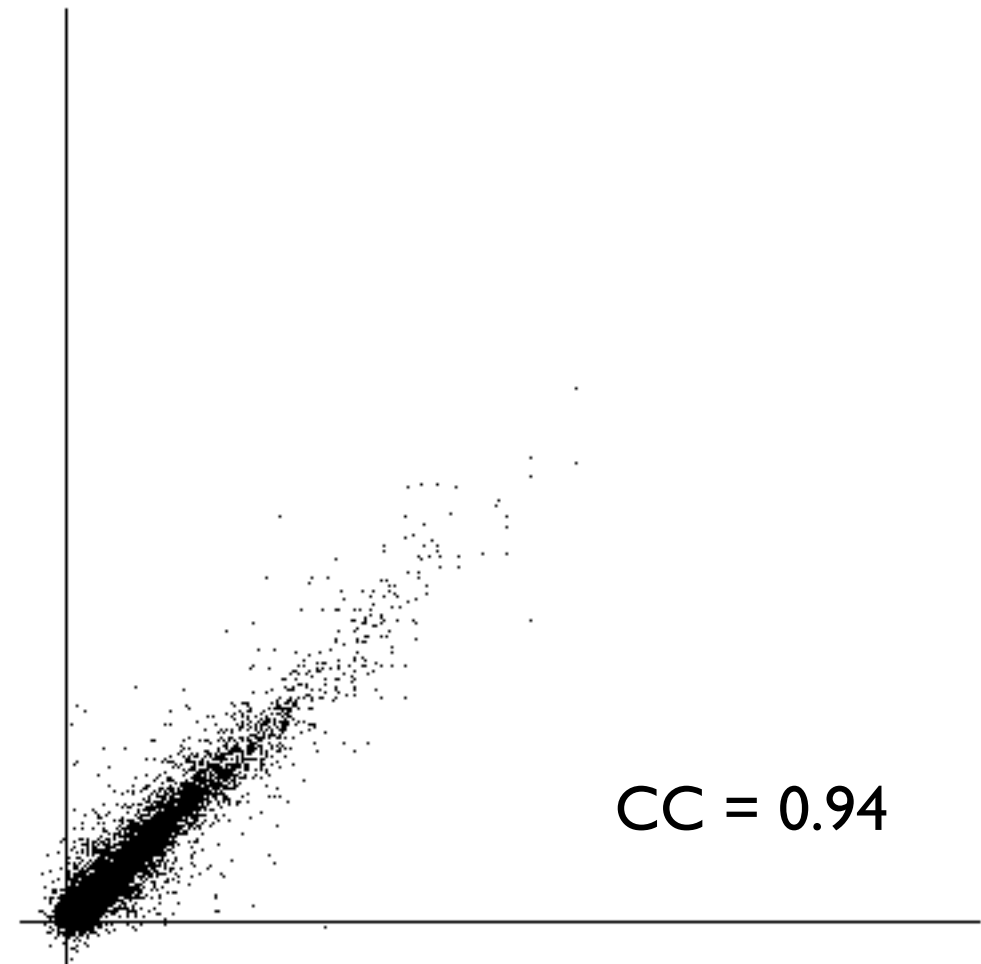
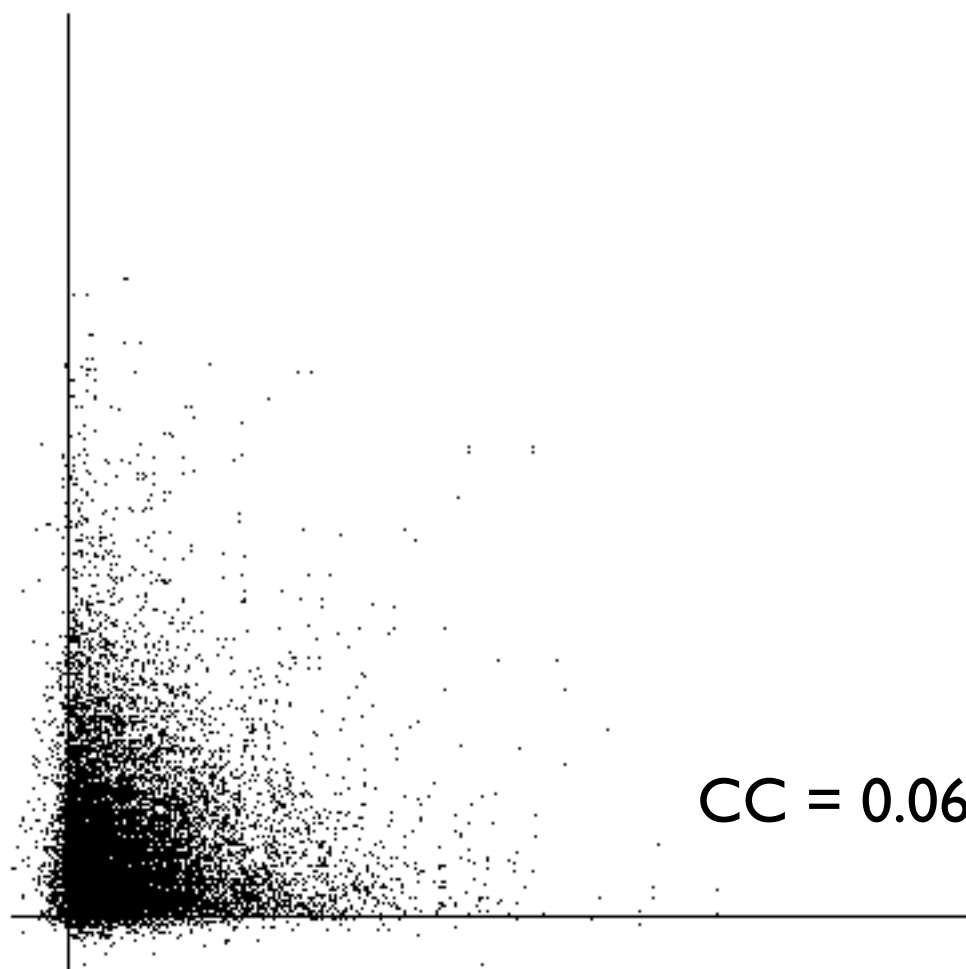
### Linear correlation coefficient

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 3: point group

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]



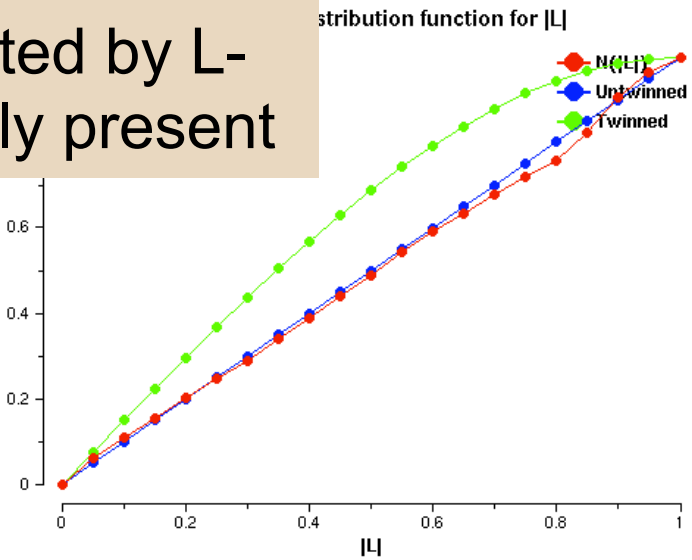
# What can go wrong? A bad case

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!)

Twinning not detected by L-test though probably present

Correct operators indicated for 321 point group



Analysing rotational symmetry in lattice group P 6/m m m

-----

Scores for each symmetry element

NeImt	Lklhd	Z-cc	CC	N	Rmeas	Symmetry & operator (in Lattice Cell)			
1	0.932	9.94	0.99	1299	0.033	identity			
2	0.794	8.19	0.82	1328	0.168	**	2-fold	l	( 0 0 1) {-h,-k,l}
3	0.815	8.29	0.83	1346	0.168	**	2-fold	k	( 0 1 0) {-h,h+k,-l}
4	0.625	7.55	0.76	1340	0.171	*	2-fold	h	( 1 0 0) {h+k,-k,-l}
5	0.653	7.65	0.77	1362	0.173	*	2-fold		( 1-1 0) {-k,-h,-l}
6	0.934	9.83	0.98	1510	0.048	***	2-fold		( 2-1 0) {h,-h-k,-l}
7	0.934	9.85	0.99	1486	0.045	***	2-fold		(-1 2 0) {-h-k,k,-l}
8	0.936	9.72	0.97	1450	0.055	***	2-fold		( 1 1 0) {k,h,-l}
9	0.936	9.66	0.97	3007	0.065	***	3-fold	l	( 0 0 1) {k,-h-k,l}{-h-k,h,l}
10	0.503	7.09	0.71	2790	0.186	*	6-fold	l	( 0 0 1) {h+k,-h,l}{-k,h+k,l}



# 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!)

... but POINTLESS selects the wrong Laue group in this case

Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
= 1	P 6/m m m ***	0.981	8.74	8.74	0.00	0.87	0.00	0.11	0.00	0.0	[h,k,l]
2	P -3 m 1	0.018	2.10	9.80	7.70	0.98	0.77	0.05	0.17	0.0	[h,k,l]
3	C m m m	0.000	0.60	9.10	8.50	0.91	0.85	0.10	0.11	0.0	[h,h+2k,l]
4	P -3 1 m	0.000	-0.12	8.68	8.80	0.87	0.88	0.11	0.10	0.0	[h,k,l]
5	C m m m	0.000	0.29	8.92	8.62	0.89	0.86	0.10	0.11	0.0	[h+k,-h+k,l]
6	C m m m	0.000	0.33	8.94	8.61	0.89	0.86	0.10	0.11	0.0	[-k,2h+k,l]
7	P 6/m	0.000	-0.09	8.69	8.78	0.87	0.88	0.11	0.11	0.0	[h,k,l]
8	P -3	0.000	1.38	9.81	8.43	0.98	0.84	0.05	0.13	0.0	[h,k,l]
9	C 1 2/m 1	0.000	1.39	9.84	8.45	0.98	0.84	0.04	0.13	0.0	[h-k,h+k,l]
10	C 1 2/m 1	0.000	1.45	9.89	8.44	0.99	0.84	0.04	0.13	0.0	[h+2k,-h,l]
11	C 1 2/m 1	0.000	1.48	9.90	8.43	0.99	0.84	0.04	0.13	0.0	[2h+k,k,l]
12	C 1 2/m 1	0.000	0.60	9.21	8.61	0.92	0.86	0.09	0.11	0.0	[h,h+2k,l]
13	P 1 2/m 1	0.000	0.55	9.17	8.62	0.92	0.86	0.09	0.11	0.0	[k,l,h]
14	C 1 2/m 1	0.000	0.20	8.90	8.70	0.89	0.87	0.09	0.11	0.0	[h+k,-h+k,l]
15	C 1 2/m 1	0.000	0.16	8.86	8.70	0.89	0.87	0.09	0.11	0.0	[-k,2h+k,l]
16	P -1	0.000	1.37	9.94	8.58	0.99	0.86	0.03	0.12	0.0	[l,-h,-k]

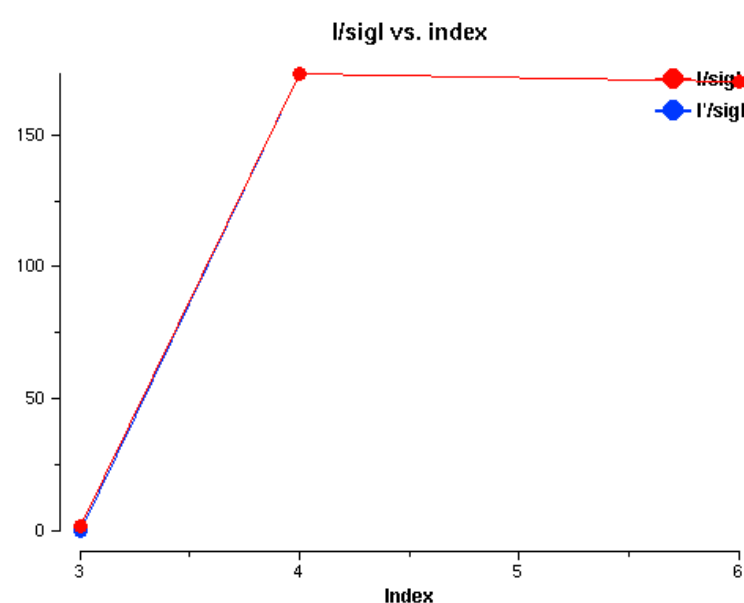


# Stage 4: space group from axial systematic absences

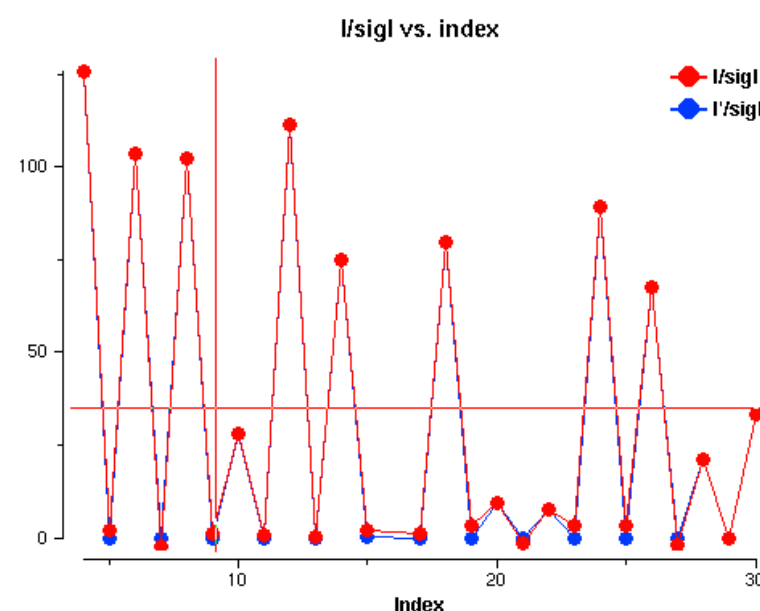
Zone	Number	PeakHeight	SD	Probability	ReflectionCondition
Zones for Laue group $P\ m\ m\ m$					
1 screw axis $2\{1\}$ [a]	3	1.000	0.296	** 0.889	$h00: h=2n$
2 screw axis $2\{1\}$ [b]	26	1.000	0.142	*** 0.971	$0k0: k=2n$
3 screw axis $2\{1\}$ [c]	46	0.997	0.097	*** 0.986	$00l: l=2n$

## Fourier analysis of $I/\sigma(I)$

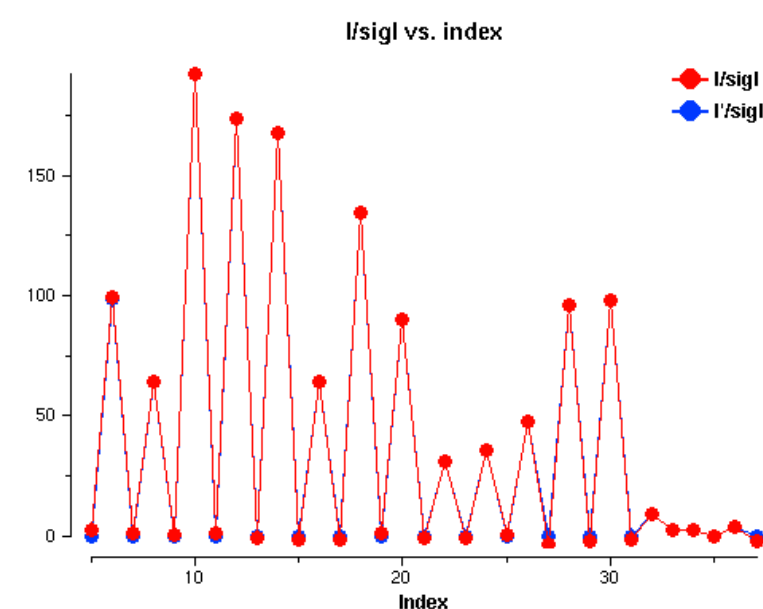
There are indications of  $2_1$  screw symmetry along all principle axes (though note there are only 3 observations on the  $a$  axis ( $h00$  reflections))



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, 00l: l=2n (zones 1,2,3)
<P 2 21 21> ( 18)	0.104	0.106		0k0: k=2n, 00l: l=2n (zones 2,3)
<P 21 2 21> ( 18)	0.025	0.026		h00: h=2n, 00l: 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

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

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



# Pseudo-cubic example

Cell: 79.15 81.33 81.15 90.00 90.00 90.00       $a \approx b \approx c$

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  $a \approx b \approx c$

	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<sub>1</sub> 2<sub>1</sub> 2<sub>1</sub>)



# Combining multiple files (and multiple MAD datasets)

Pointless: prepare intensity data for scaling

Job title: pk ip rm Se34

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

Input reflection file type: MTZ file

Project name: Brap crystal name: Se34 dataset name: pk

MTZ #1 Full path.. /Amb/home/pre/Projects/Brap/Se34/pk\_1\_001.mtz Browse View

MTZ #2 Full path.. /Amb/home/pre/Projects/Brap/Se34/pk\_2\_001.mtz Browse View

☒ Assign to the same dataset as the previous file

MTZ #3 Full path.. /Amb/home/pre/Projects/Brap/Se34/pk\_180\_1\_001.mtz Browse View

☒ Assign to the same dataset as the previous file

MTZ #4 Full path.. /Amb/home/pre/Projects/Brap/Se34/ip\_1\_001.mtz Browse View

☐ Assign to the same dataset as the previous file

Project name: Brap crystal name: Se34 dataset name: ip

MTZ #5 Full path.. /Amb/home/pre/Projects/Brap/Se34/rm\_1\_001.mtz Browse View

☐ Assign to the same dataset as the previous file

Project name: Brap crystal name: Se34 dataset name: Rm

Edit list Add File

☒ Write output reflections in the best space/pointgroup

Output MTZ Brap se34\_pk\_ip\_rm.mtz Browse View

☐ Test Laue group of 1st file before reading rest ☐ Assume all files have same indexing (faster)

☐ 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 Partial ☐

Additional Options ☐

Run Save or Restore Close

3 files  
assigned to  
same dataset

Dataset 1, pk, 3 files

Dataset 2, ip, 1 file

Dataset 3, rm, 1 file



# Combining multiple files (and multiple MAD datasets)

Alternative index test relative to first file

Alternative reindexing	CC	R(E <sup>2</sup> )	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 <sup>2</sup> )	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 <sup>2</sup> )	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 <sup>2</sup> )	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)

The screenshot shows the 'Pointless: prepare intensity data for scaling' window. The 'Job title' field contains 'N15 get reference indexing in space group H3'. Under the 'Determine Laue group' section, the 'Match index to reference' checkbox is checked and circled in blue. The 'Input reflection file type' is set to 'MTZ file'. The 'Project name' is 'AP2', 'crystal name' is 'N15', and 'dataset name' is 'N15'. The 'MTZ #1' field shows the full path to '/Users/PhilStuff/Projects/Xtal/src/MtzUnmrg/Data/N15\_5\_001\_cut.mtz'. The 'Reference MTZ' field, also circled in blue, shows the full path to '/Users/PhilStuff/Projects/Xtal/src/MtzUnmrg/Data/N15\_5\_001\_F.mtz'. The 'I or F label' is set to 'IMEAN\_N15'. The 'Write output reflections in the space group from the reference file' checkbox is checked. The 'Output MTZ' field shows 'MtzUnmrg' and 'junk.mtz'. The 'Excluded Data' section at the bottom has a checked checkbox.



# 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)

```
Space group from HKLIN file : R 3 :H
Cell:  257.89 257.89 144.72  90.00  90.00 120.00
```

Alternative index test relative to reference file

Alternative reindexing	CC	R(E <sup>2</sup> )	Number	Cell_deviation
[h,k,l]	0.885	0.157	6537	0.81
[k,h,-l]	-0.001	0.511	6084	0.81

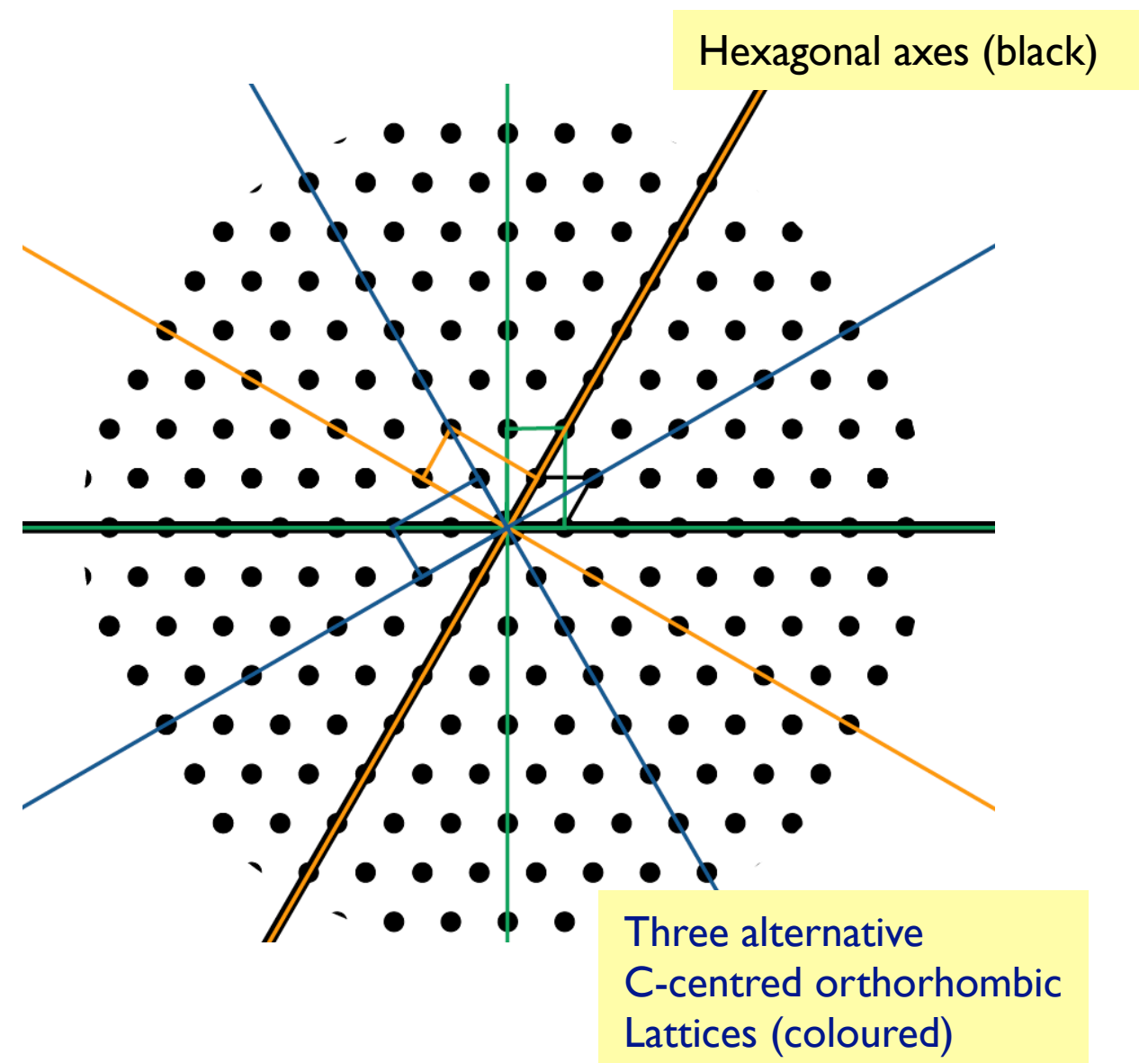
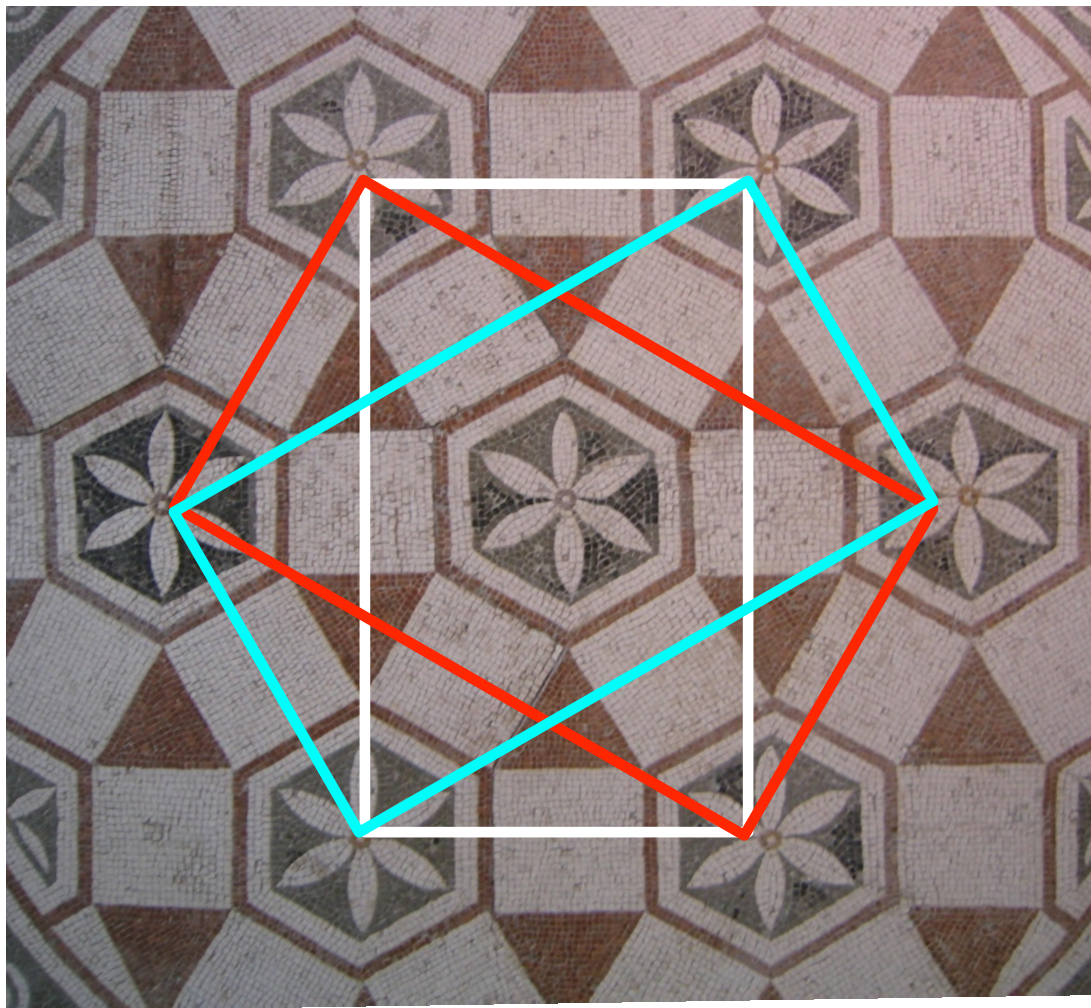


# A confusing case in C222:

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





# Score each symmetry operator in P622

“Likelihood”		Correlation coefficient on E <sup>2</sup>		Rfactor (multiplicity weighted)			
Z-score(CC)							
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}

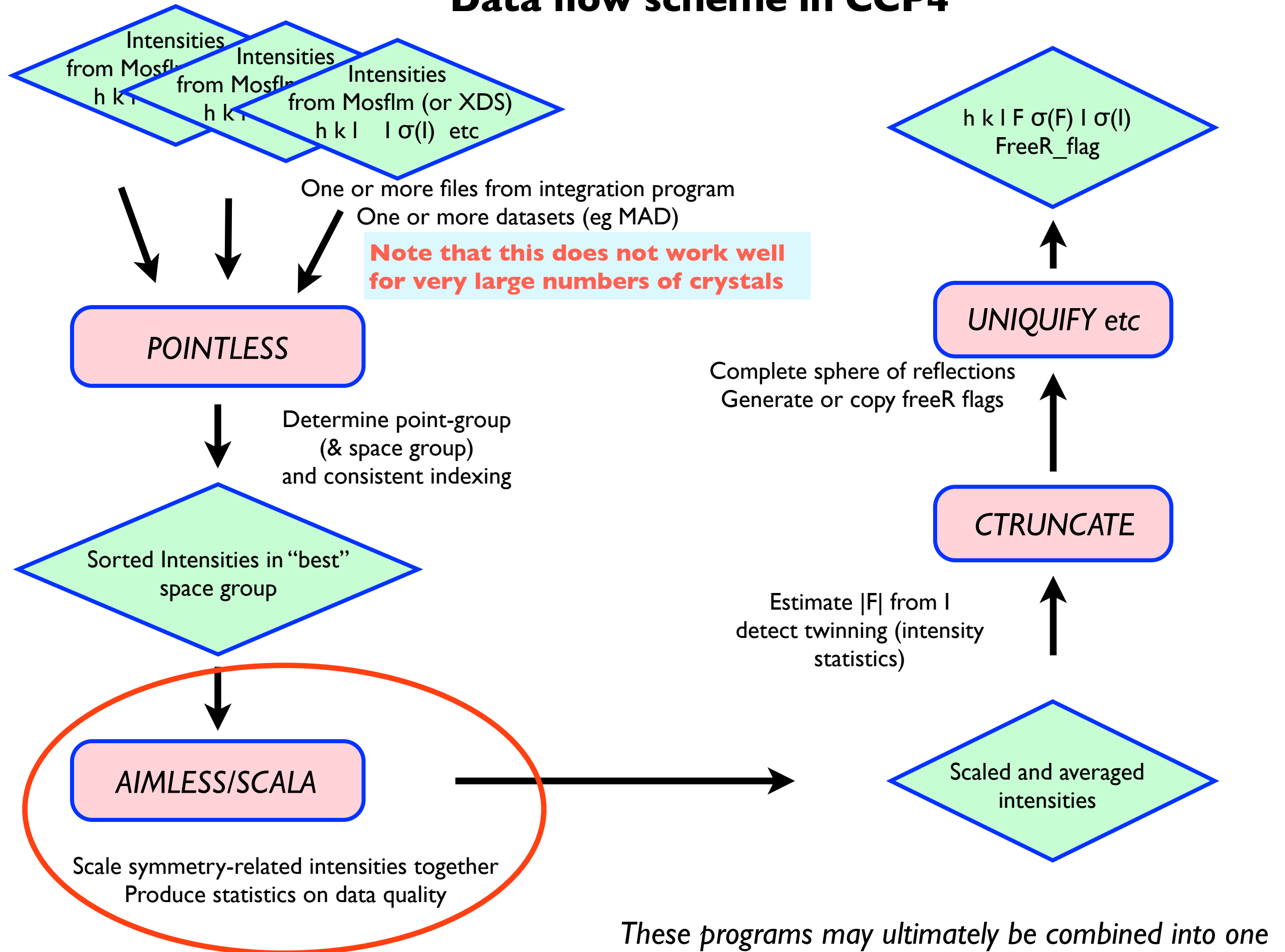
Only the orthorhombic symmetry operators are present



# Scaling and Data Quality



# Data flow scheme in CCP4





# Choices

- What scaling model?
  - the scaling model should reflect the experiment
  - considerations of scaling may affect **design** of experiment*
- Is the dataset any good?
  - should it be thrown away immediately?
  - what is the real resolution?
  - are there bits which should be discarded (bad images)?



# Why are reflections on different scales?

Various physical factors lead to observed intensities being on different scales. Some corrections are known eg Lorentz and polarisation corrections, but others can only be determined from the data

Scaling models should if possible parameterise the experiment so different experiments may require different models

Understanding the effect of these factors allows a sensible design of correction and an understanding of what can go wrong

- (a) Factors related to incident beam and the camera
- (b) Factors related to the crystal and the diffracted beam
- (c) Factors related to the detector



## 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  $\varphi$  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  $\varphi$ . (*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  $\text{CuK}\alpha$ )

(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)



# 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**.

Note that we do not know the true intensities and an internally-consistent dataset is not necessarily correct. Systematic errors which are the same for symmetry-related reflections will remain

$$\text{Minimize } \Phi = \sum_{hl} w_{hl} (I_{hl} - 1/k_{hl} \langle I_h \rangle)^2$$

$I_{hl}$   $l$ '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

$g_{hl} = g(\varphi \text{ rotation/image number}) \cdot g(\text{time}) \cdot g(s) \quad \dots \text{other factors}$   
*Primary beam  $s_0$                       B-factor      Absorption*



# AIMLESS

Performs the same function as SCALA (only better of course!)

- Read file from eg POINTLESS, sort data if necessary (SORTMTZ no longer needed)
- Initial scale estimates from average intensities
- First round scaling with small selection of strong reflections, chosen on  $I/\sigma(I)$
- First outlier rejection
- Optimise the combination of profile-fitted (for weak spots) or summation integration intensities (for strong spots) from MOSFLM
- First optimisation of  $\sigma(I)$  estimates
- Main scaling on relatively strong reflections, chosen on normalised intensity  $E^2$   
(eg  $0.8 < E^2 < 5$ )
- Second outlier rejection
- Final optimisation of  $\sigma(I)$  estimates
- Final outlier rejection
- Final statistics
- Output of merged or unmerged data (MTZ or Scalepack format, or both)



# How can we measure data quality?

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

We can also compare the intensities with their estimated errors to get signal/noise ratio

## A. Measures of statistical significance:

$I/\sigma(I)$  after averaging symmetry-related observations and after “correcting”  $\sigma(I)$  estimates (“Mn(I/sd)” in AIMLESS and SCALA output) is a measure of signal/noise

– but is sensitive to problems in the estimate of  $\sigma(I)$



## B. Measures of internal consistency:

### 1. *R*-factors

$$R_{\text{merge}} = \sum | I_{hl} - \langle I_h \rangle | / \sum | \langle I_h \rangle | \quad \text{a.k.a } R_{\text{sym}} \text{ or } R_{\text{int}}$$

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

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

multiplicity-weighted, better (but larger)

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

“Precision-indicating R-factor” gets better (smaller) with increasing multiplicity, ie it estimates the precision of the merged  $\langle I \rangle$

### 2. *correlation coefficients*

Half-dataset correlation coefficient:

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

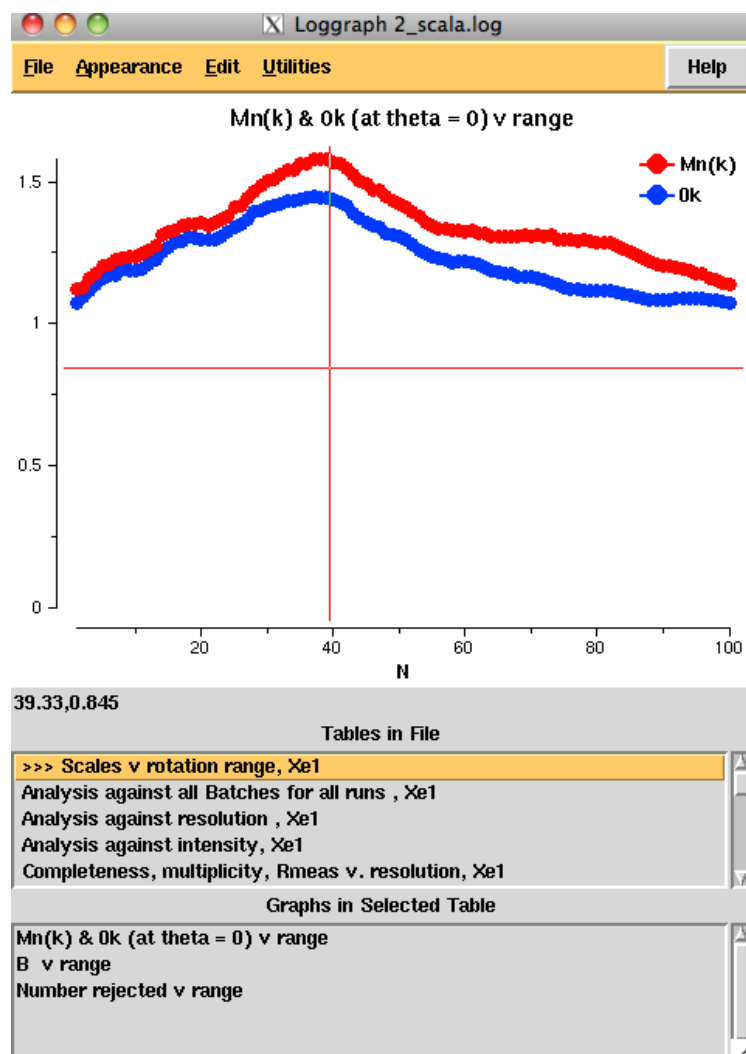
As a general rule , R-factors are good for measuring the difference between things that agree well, while correlation coefficients are better for measuring things which agree poorly



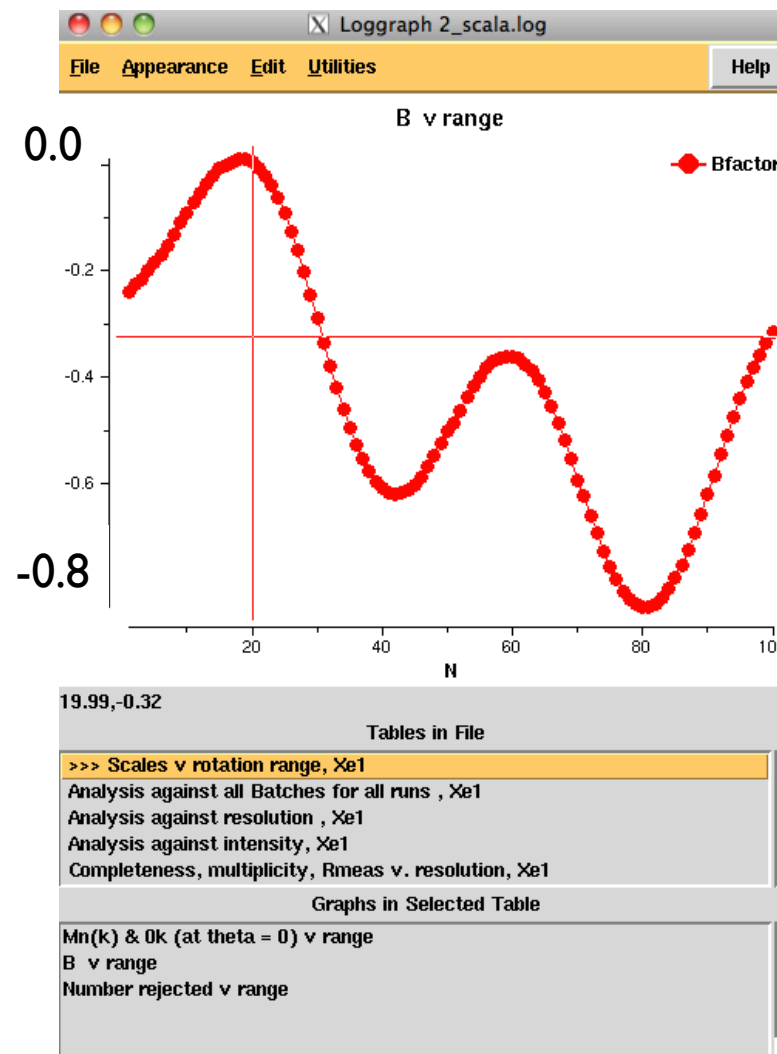
# Directions of analyses: analyses against “batch” (image number or “time”)

- 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

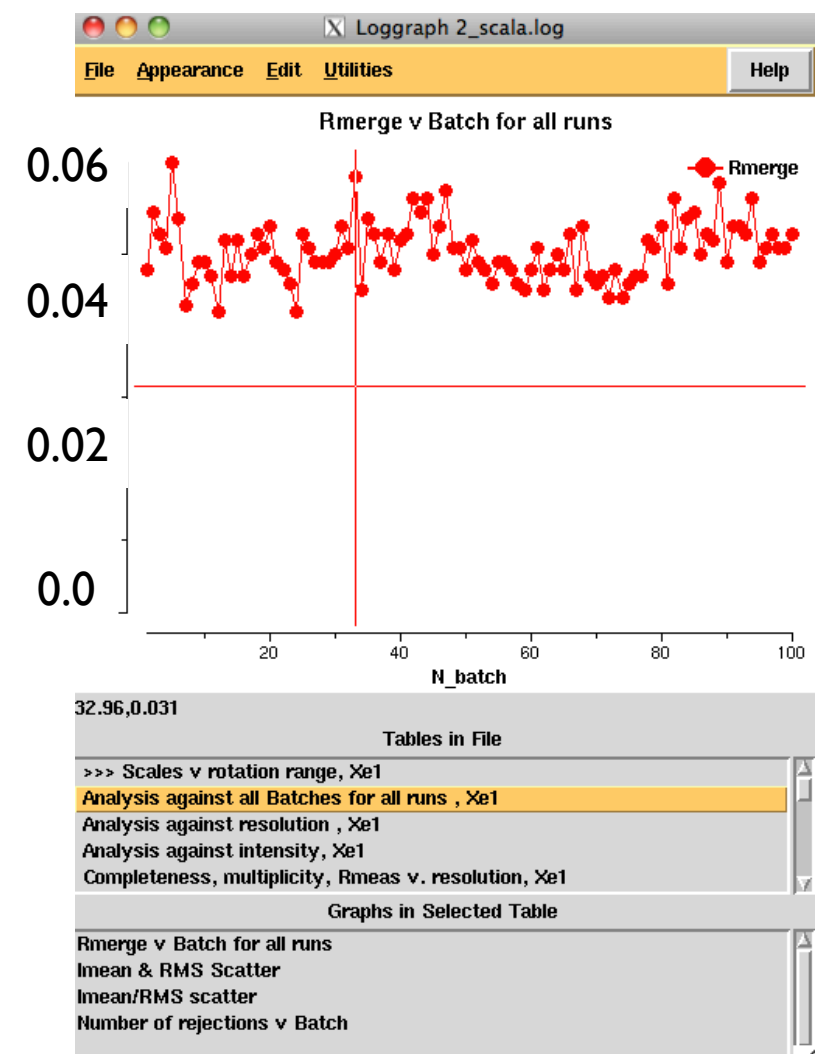
## A good case



No great difference  
between average scale  
Mn(k) & scale at  $\theta=0$



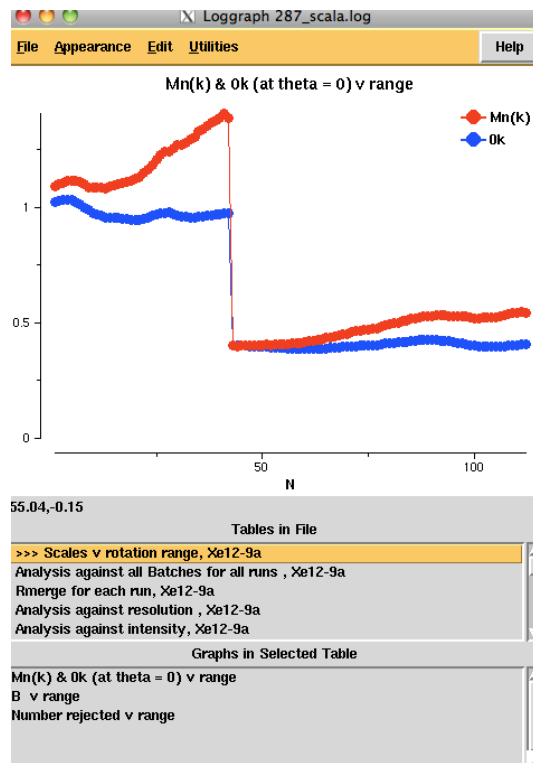
Small variation in  
relative B-factor



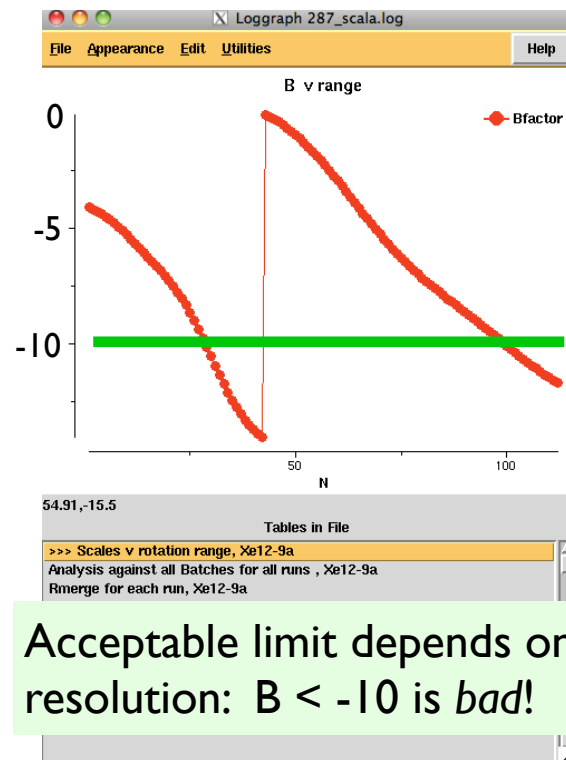
Uniform and low  $R_{\text{merge}}$



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

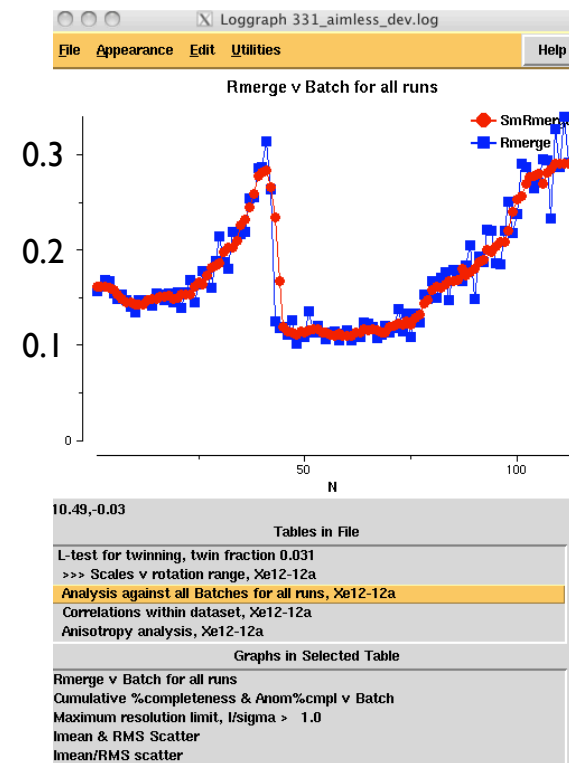


Increasing difference between average scale Mn(k) & scale at  $\theta=0$

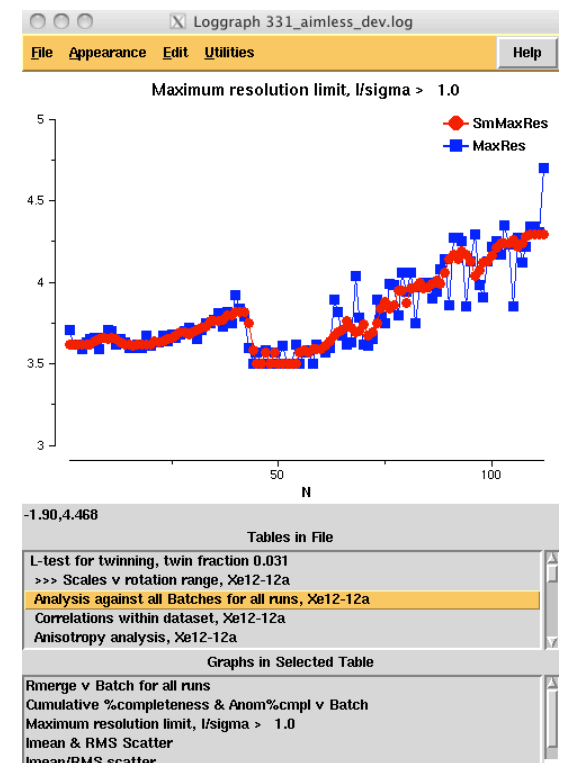


Acceptable limit depends on resolution:  $B < -10$  is *bad*!

relative B gets more negative with radiation damage



High and increasing  $R_{\text{merge}}$

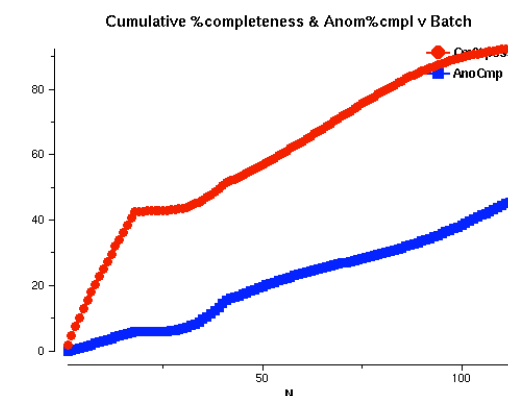


“Resolution limit” where  $I/\sigma$  falls below 1.0 getting worse

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

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

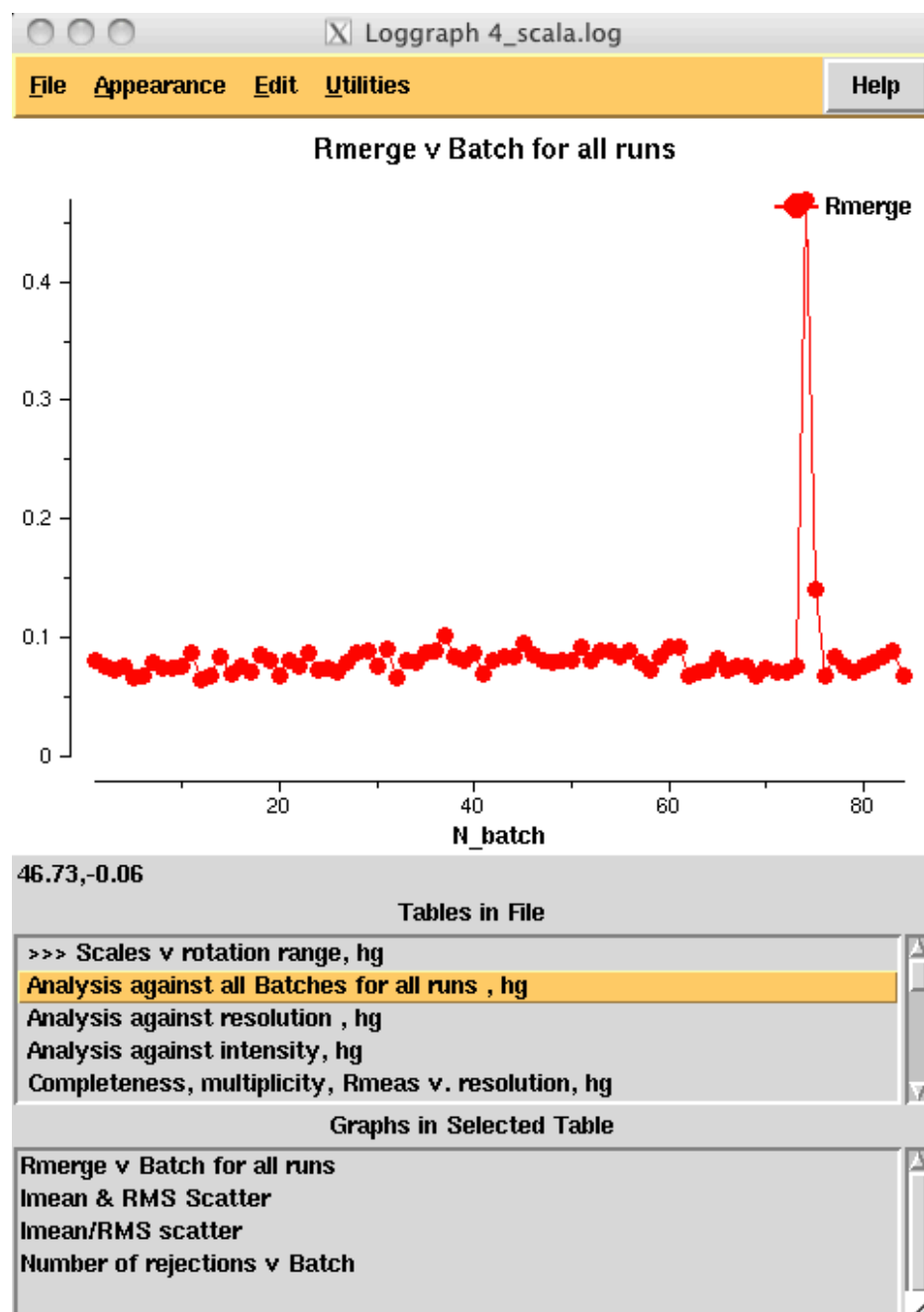
## Cumulative completeness



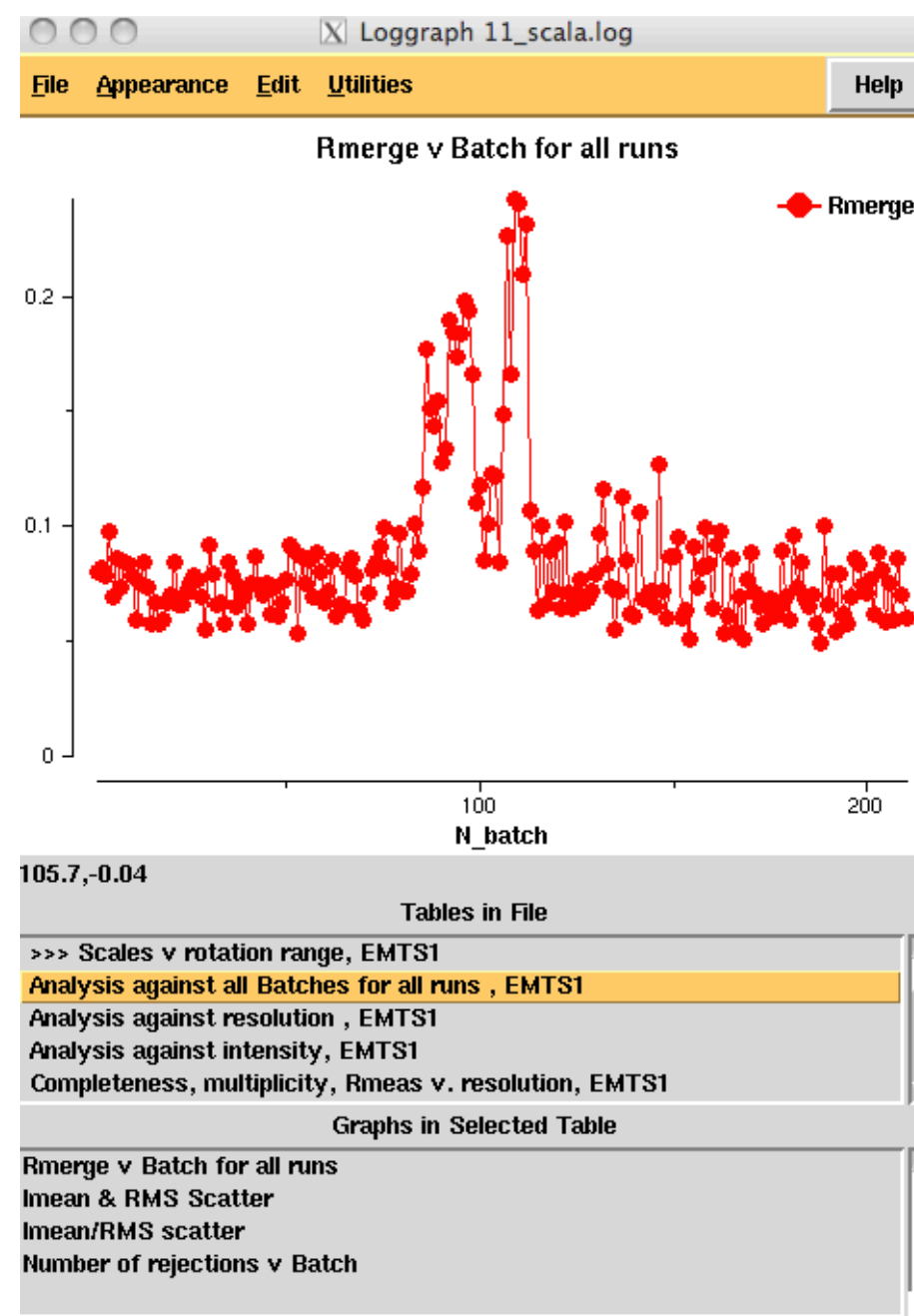
still very incomplete!



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



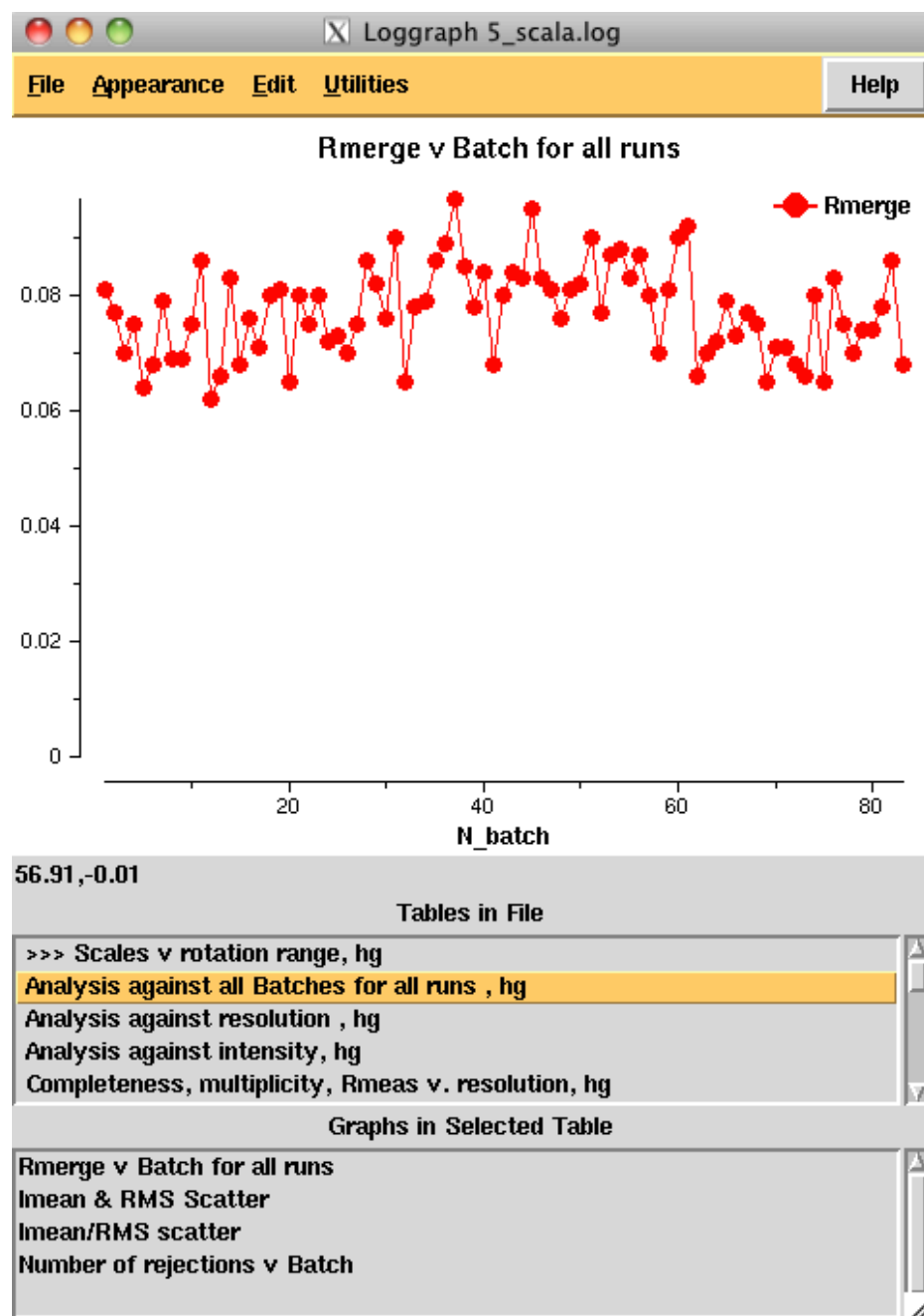
One bad (weak) image



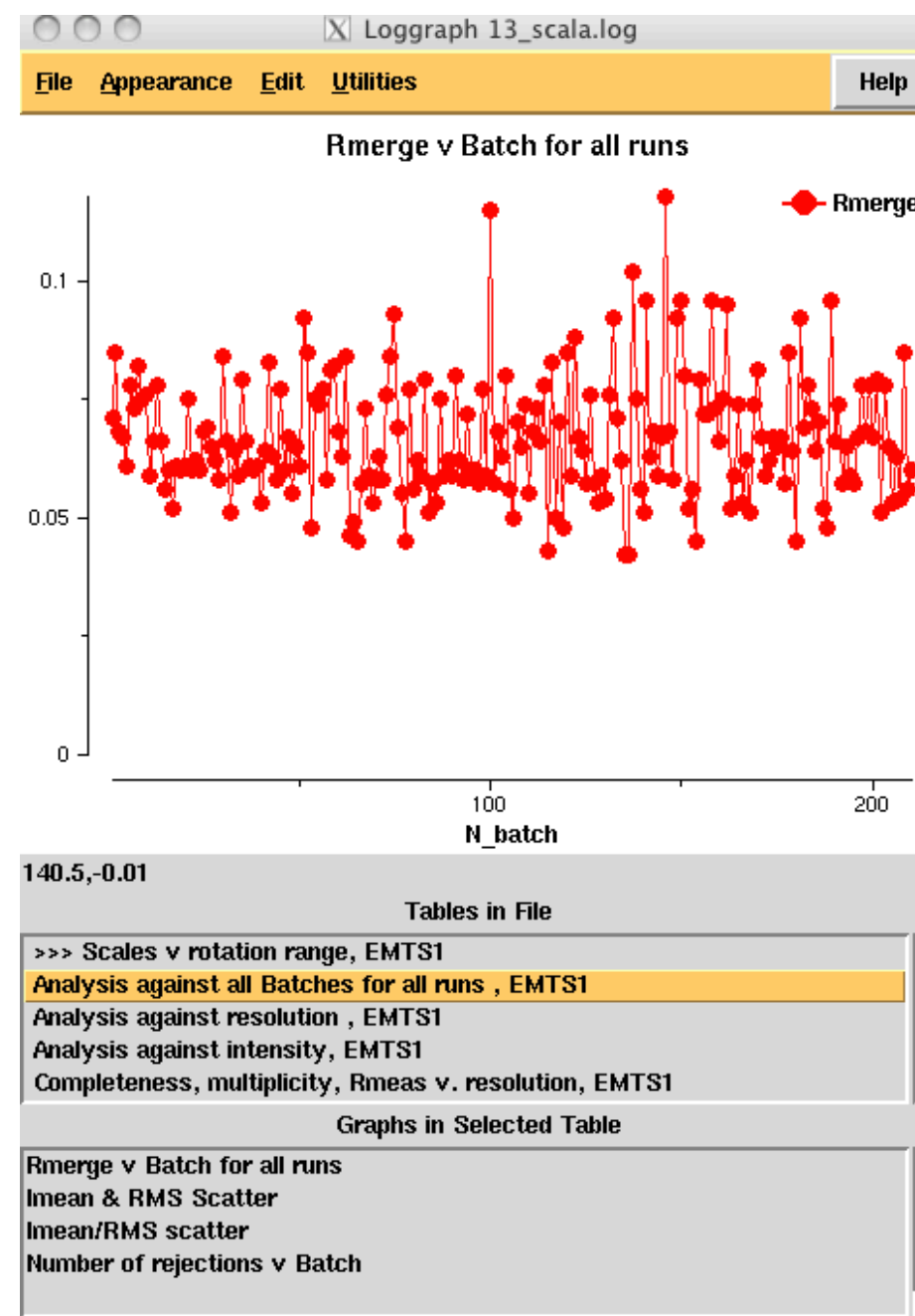
Bad region where integration had gone wrong



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



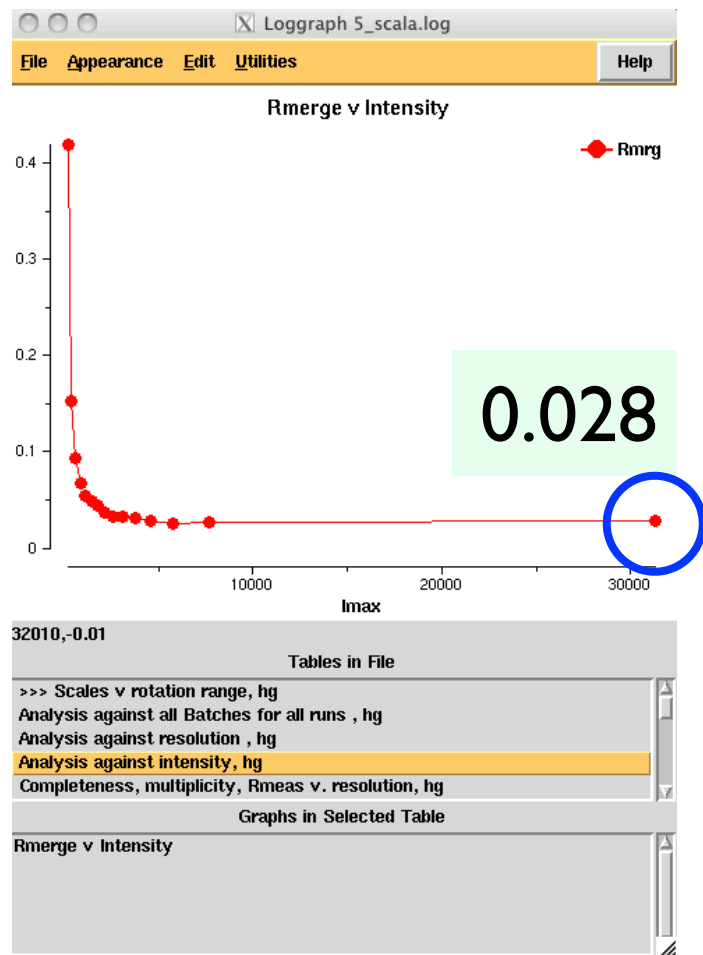
Omitting bad image



Reprocessed



# Analyses against intensity



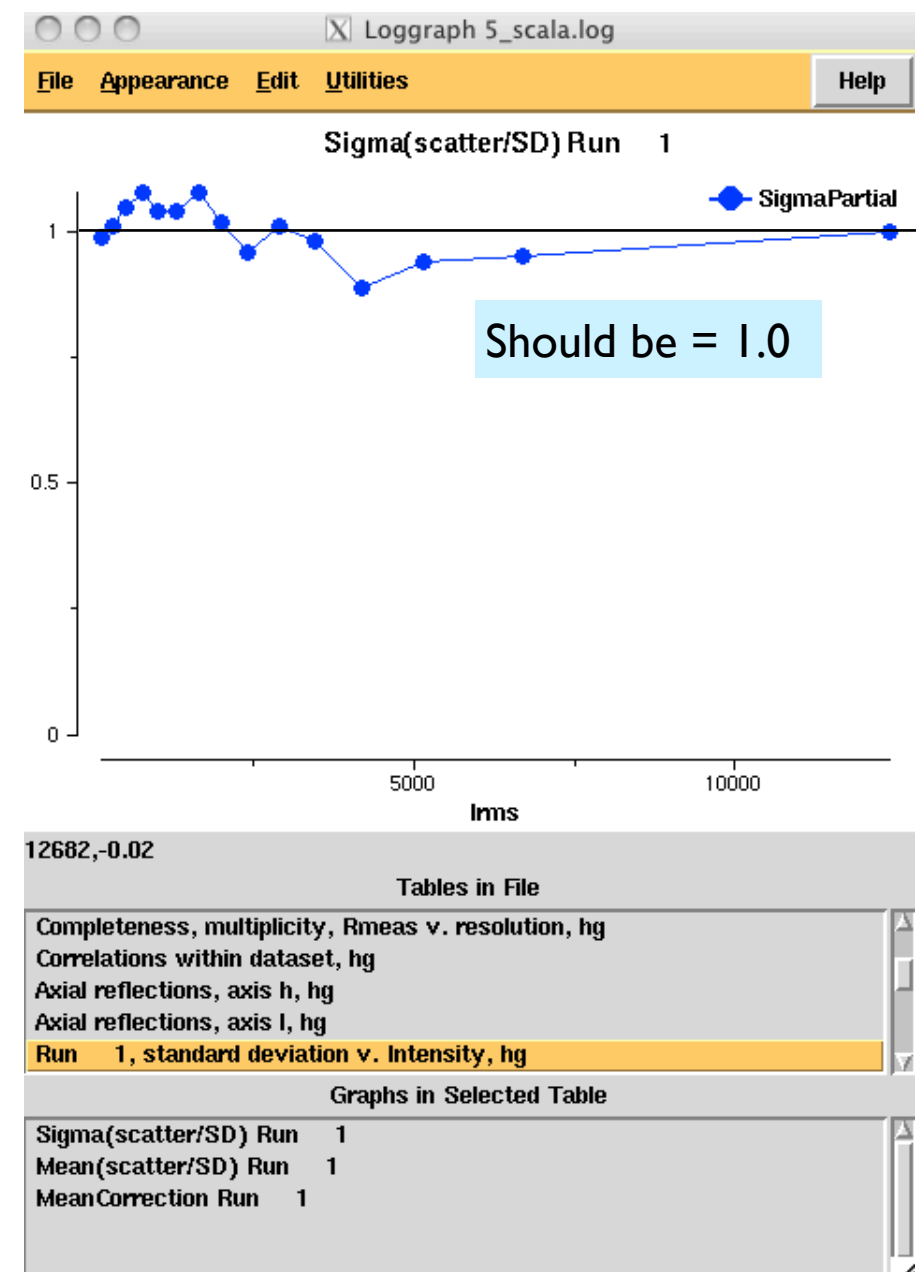
$R_{\text{merge}}$  vs.  $I$  not generally useful (since  $R$  is a fractional measure, it will always be large for small  $I$ ), but the value in the top intensity bin should be small

## Improved estimate of $\sigma(I)$

The error estimate  $\sigma(I)$  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  $\sigma'(I)$ , in all intensity ranges

$$\text{Corrected } \sigma'(I)^2 = \text{SDfac}^2 [\sigma^2 + \text{SdB} \langle I_h \rangle + (\text{SdAdd} \langle I_h \rangle)^2]$$

**SDfac**, **SdB** and **SdAdd** are adjustable parameters





## Directions of analyses: 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:

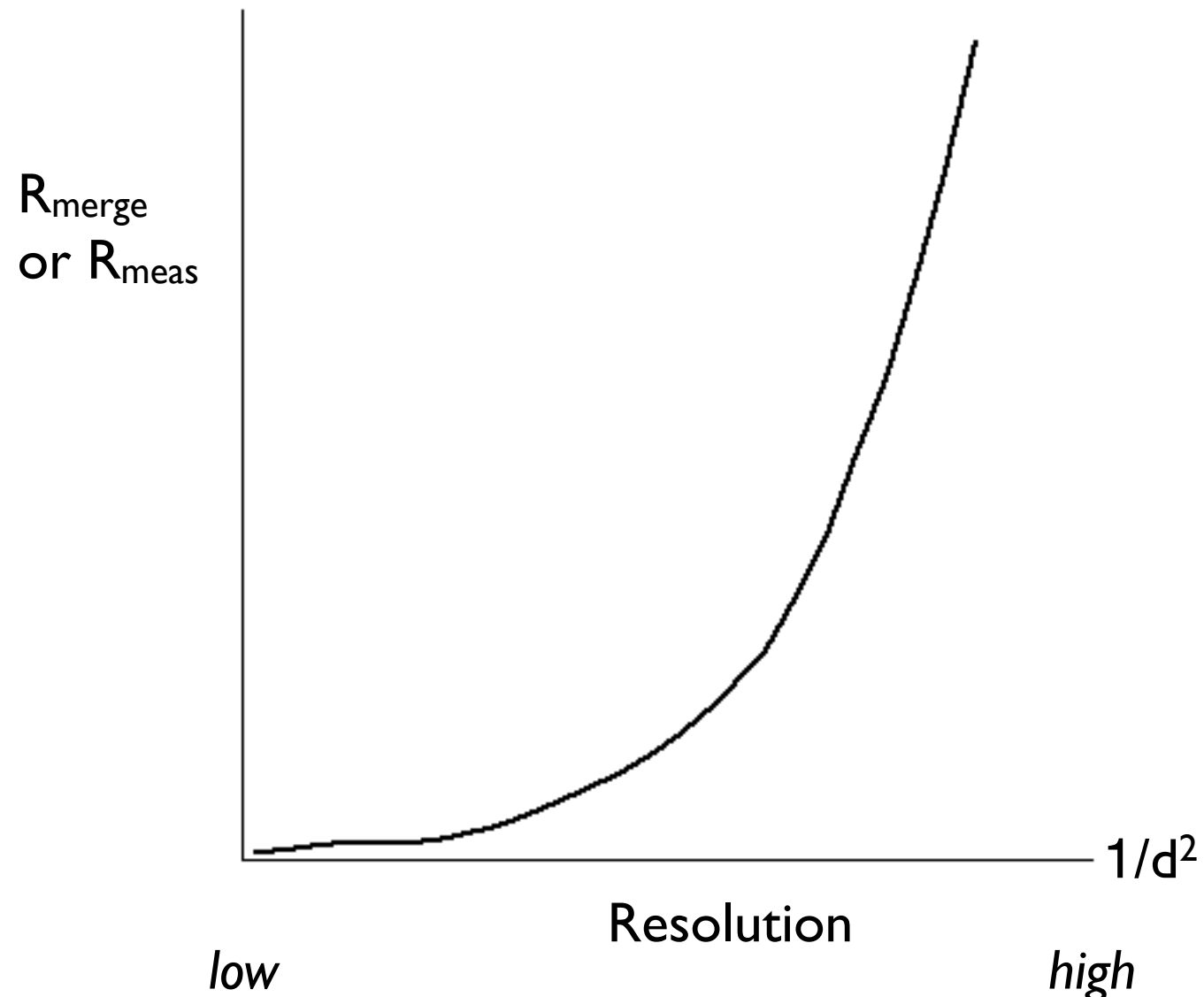
“The crystallographic  $R_{\text{merge}}$  and  $R_{\text{meas}}$  values are not acceptable, although, surprisingly the  $I/\sigma(I)$  and  $R$ -factors for these shells are OK. **I cannot ever recall seeing such high values of  $R_{\text{merge}}$  - 148% !** The discrepancy between the poor  $R_{\text{merge}}$  values and vs. the other statistics is highly unusual- generally a high  $R_{\text{merge}}$  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  $\langle I \rangle / \langle \sigma \rangle = 1.2$  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?

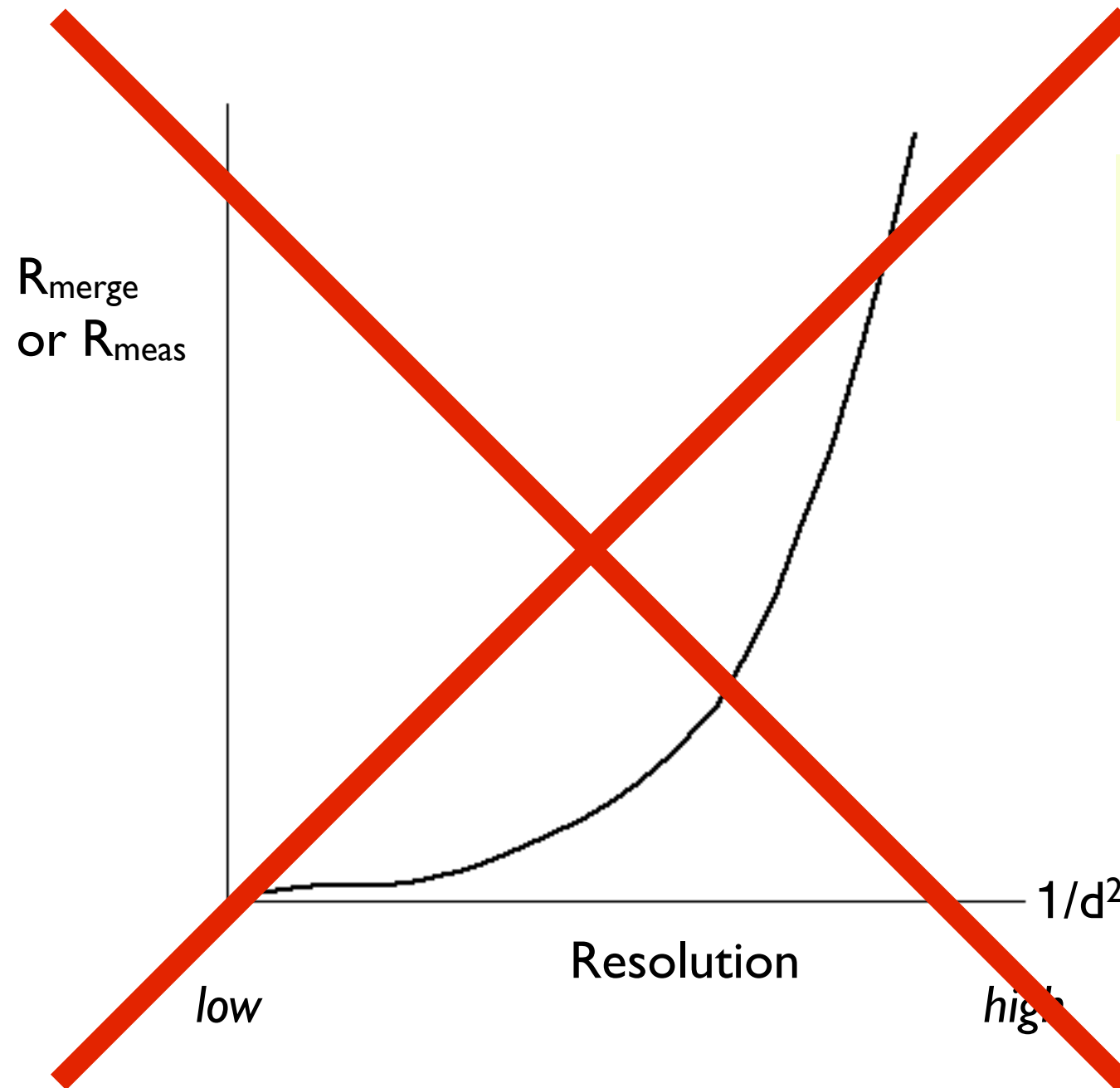


Where is the cut-off point?

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



## What about R-factors?



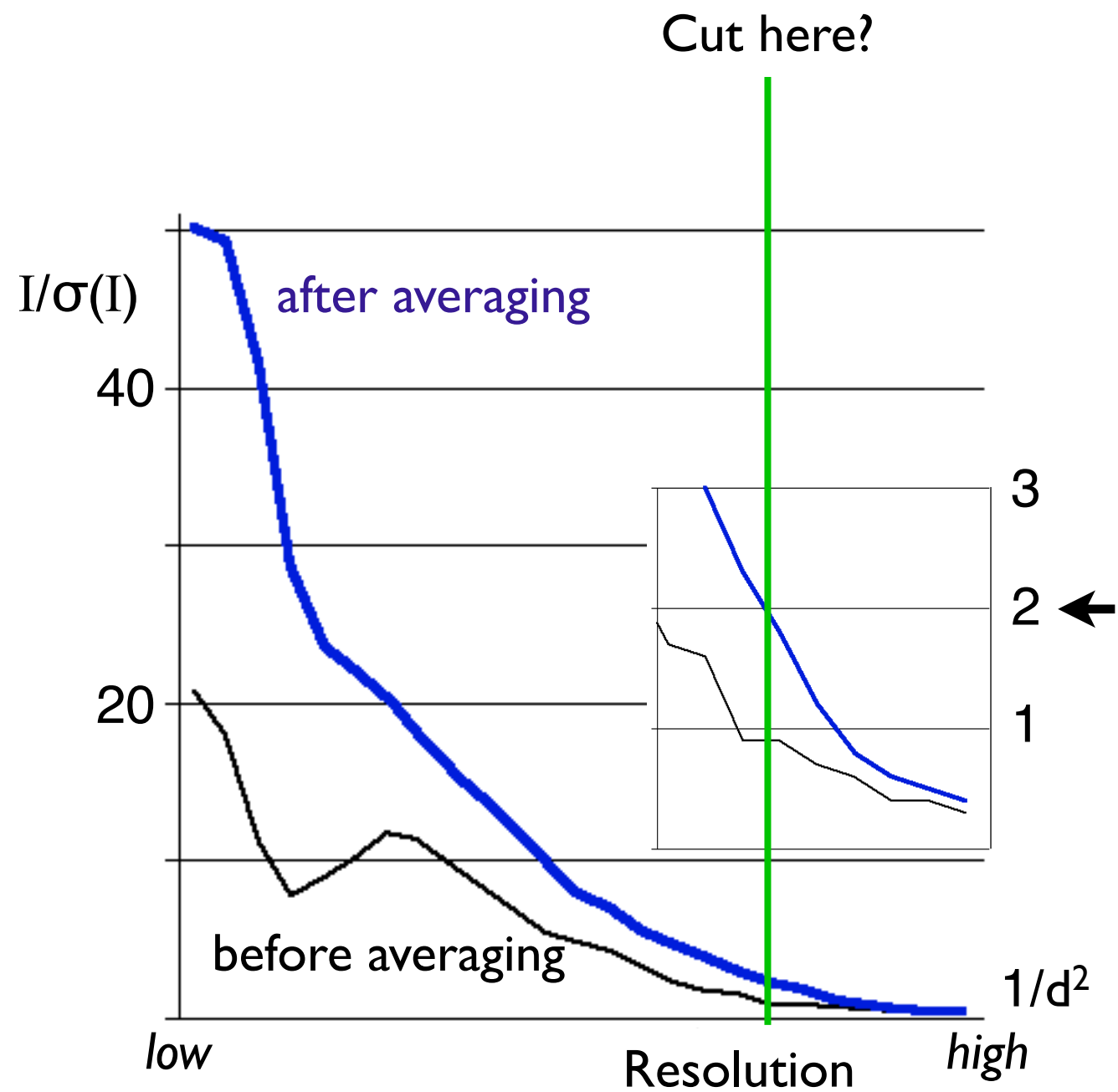
Note that  $R_{\text{merge}}$  and  $R_{\text{meas}}$  are useful for other purposes, but not for deciding the resolution cutoff

Where is the cut-off point?

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



## What about $I/\sigma(I)$ (signal/noise)?

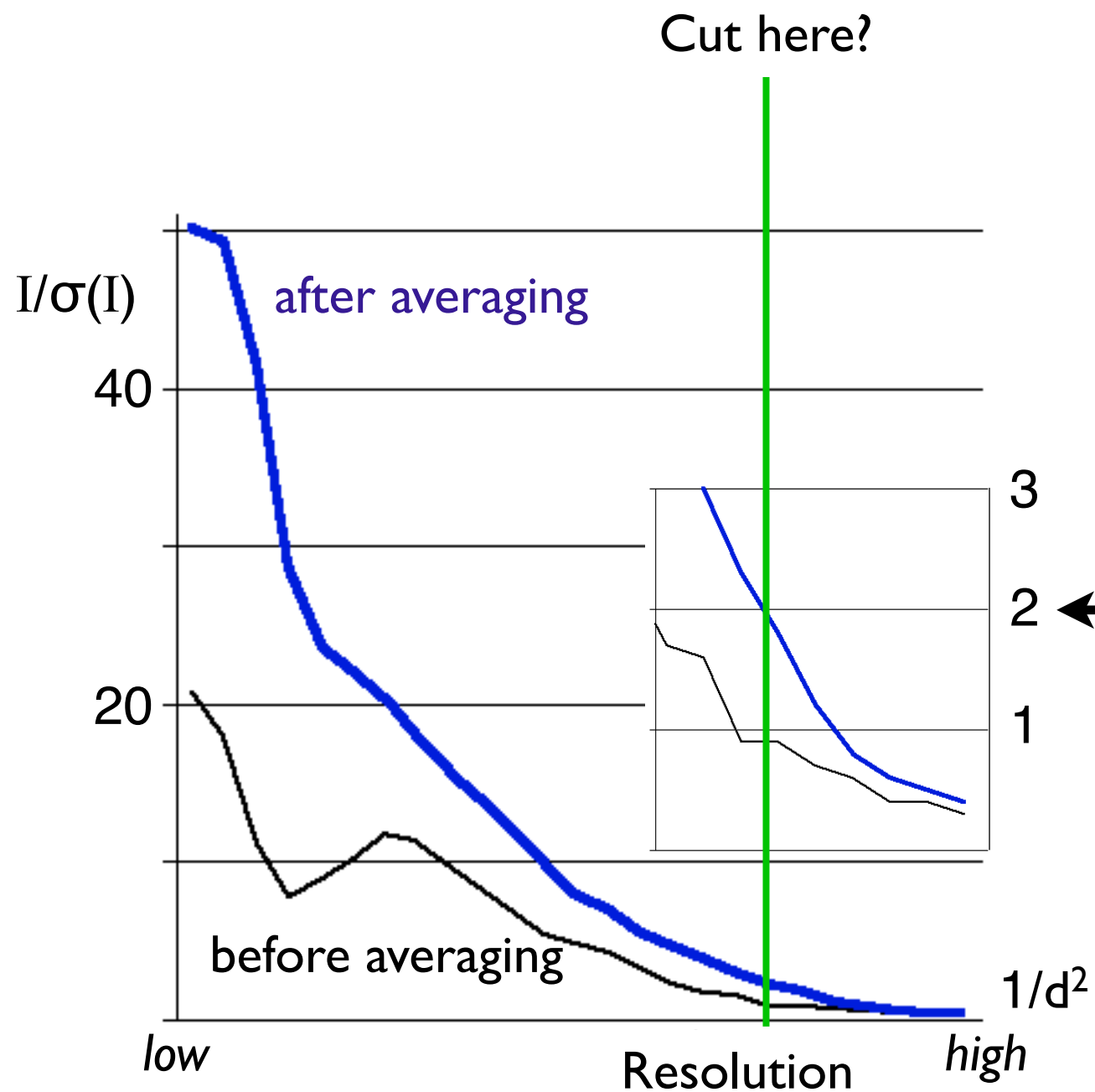


Cut resolution at  $I/\sigma(I)$  after averaging  
( $M_n(I/sd) = 2$  (or maybe 1?))

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



## What about $I/\sigma(I)$ (signal/noise)?



?OK

Cut resolution at  $I/\sigma(I)$  after averaging  
( $M_n(I/sd) = 2$  (or maybe 1?))

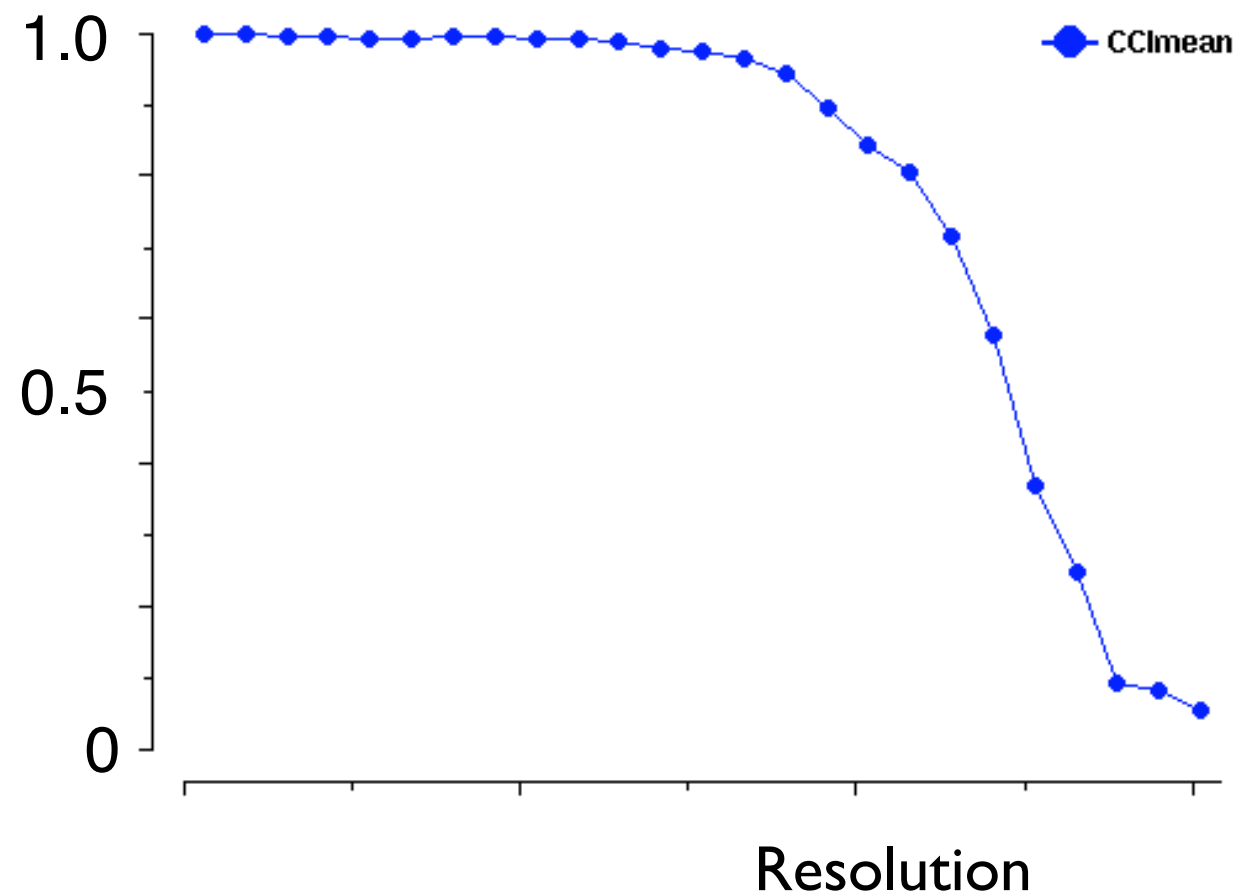
A reasonably good criterion, but it  
relies on  $\sigma(I)$ , which is not entirely  
reliable



# What about correlation coefficients?

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  $\sigma(I)$

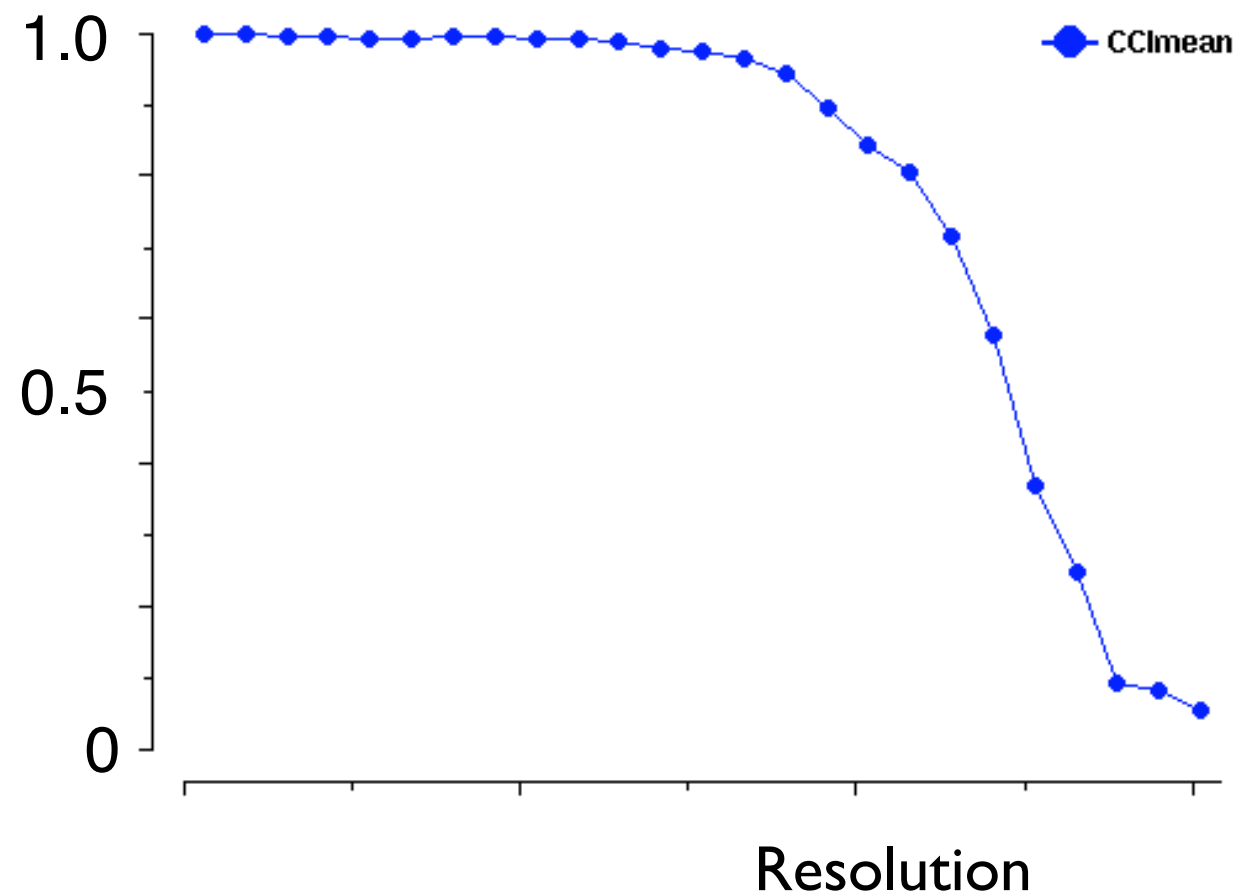
**Maybe** cut data at CC  $\approx$  0.5



## What about correlation coefficients?

Half-dataset correlation coefficient:

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



OK  
good

Advantages:

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

**Maybe** cut data at CC  $\approx$  0.5

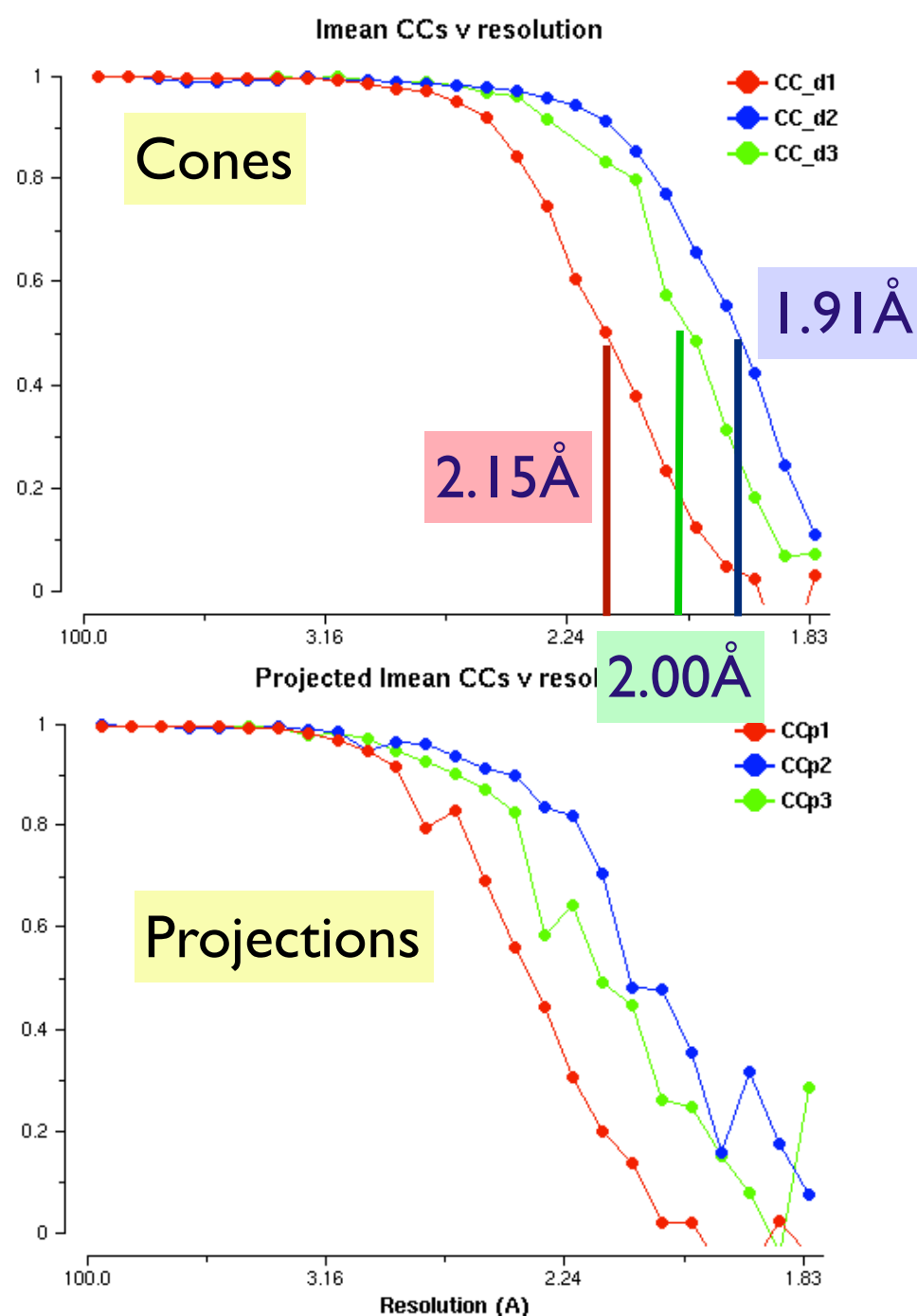


# Anisotropy

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  $I/\sigma(I)$  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 ...

“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

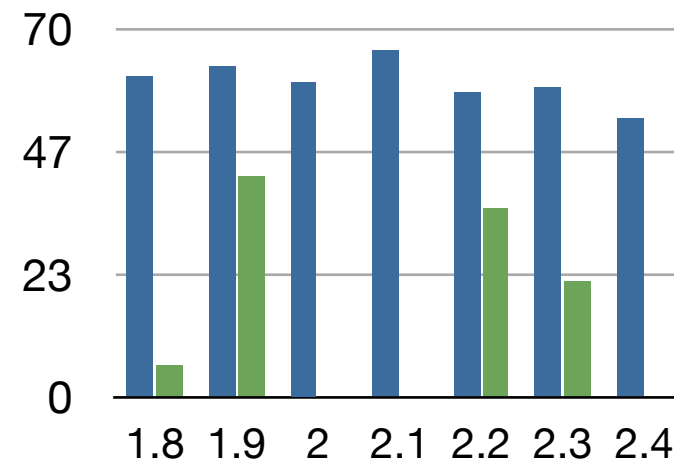
There is no reason to suppose that cutting back the resolution to satisfy referees will improve your model!

Future developments may improve treatment of weak noisy data



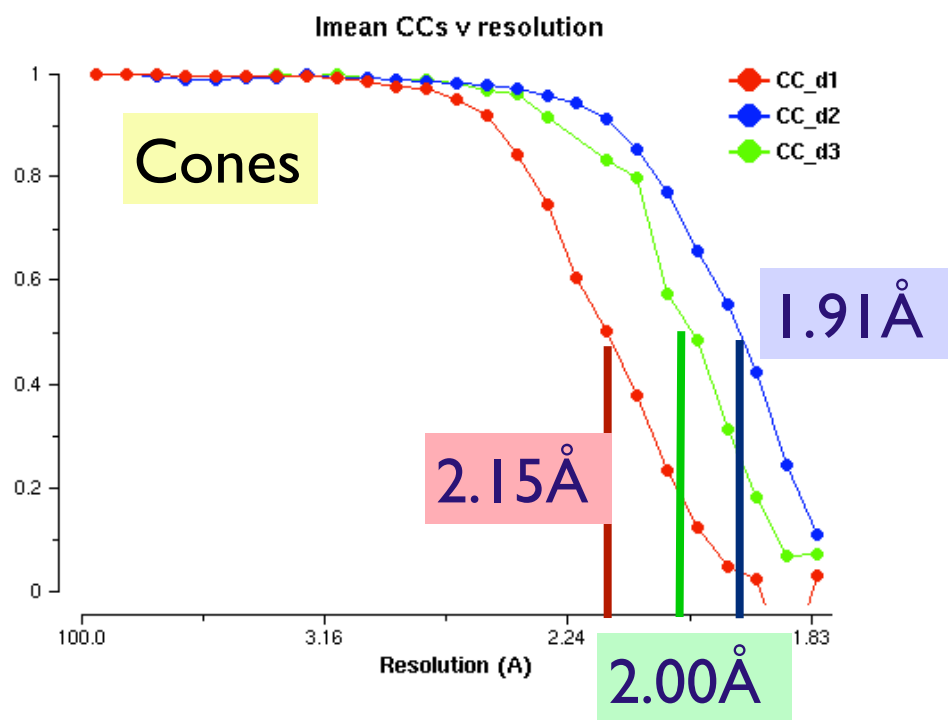
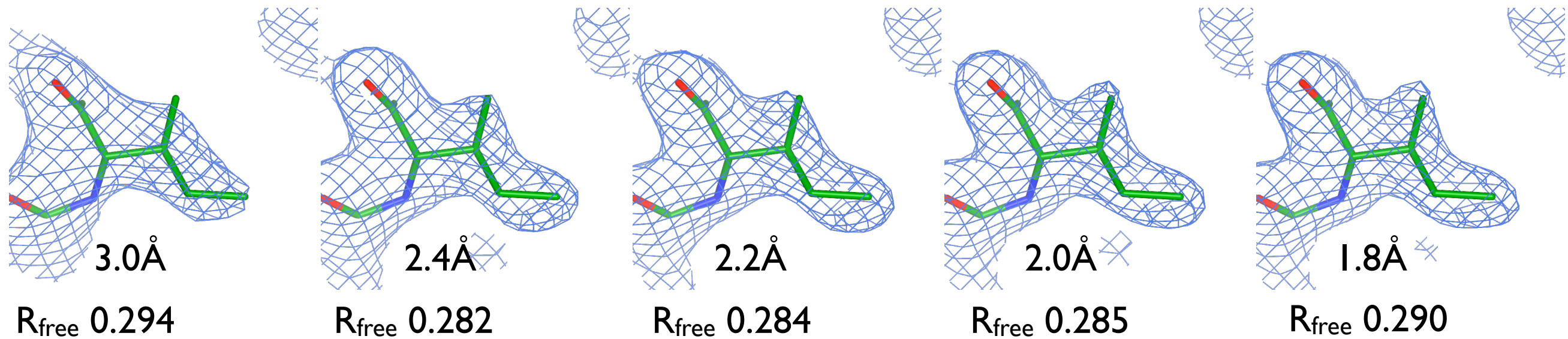
Example:

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

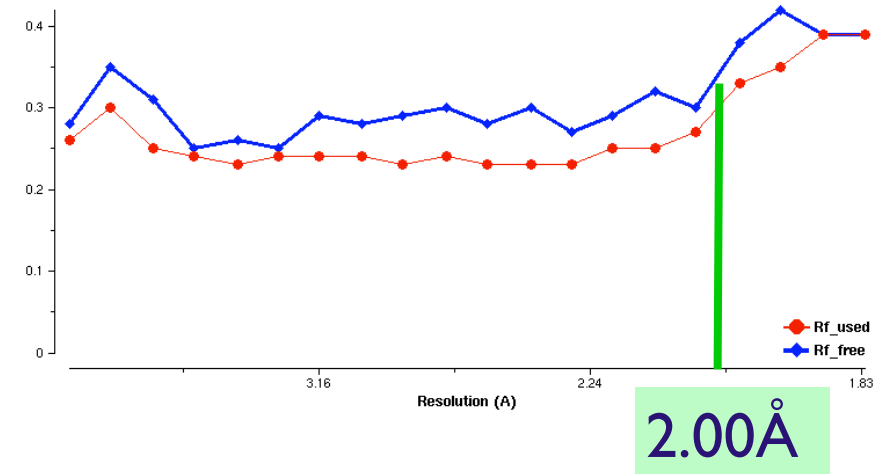


Number of residues **built** and **sequenced**

figures made with ccp4mg

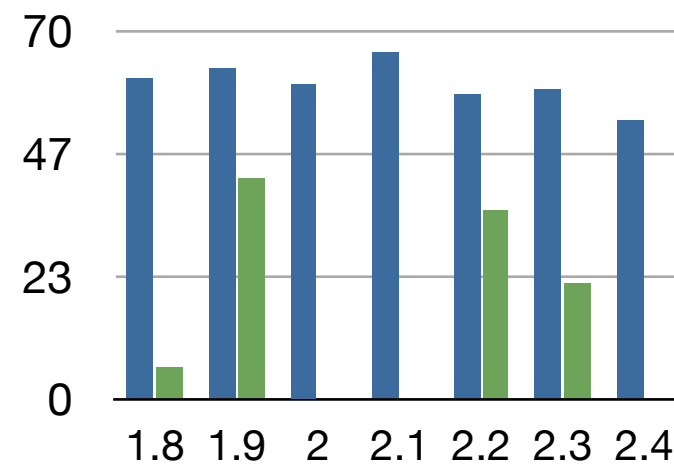


R & R<sub>free</sub> after initial refinement



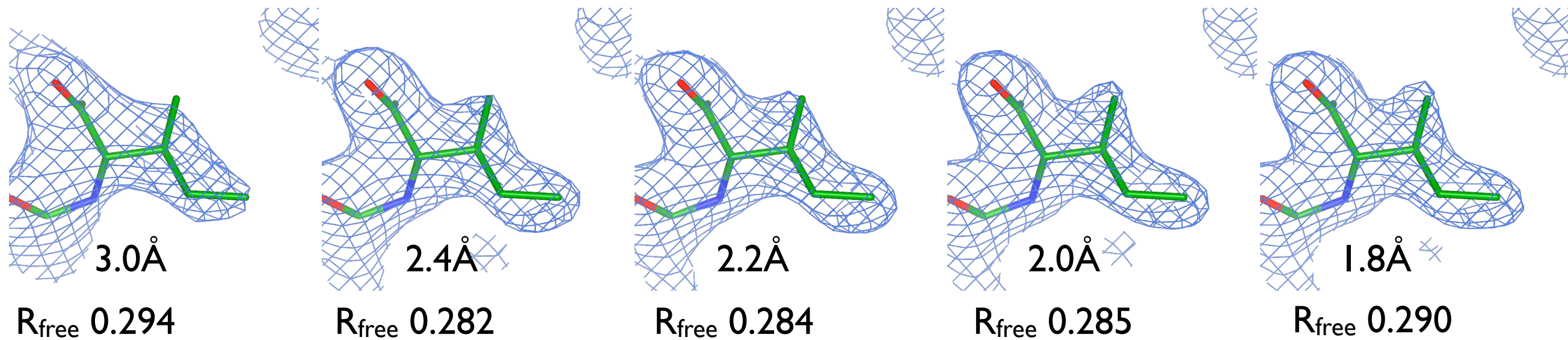


Example:  
3 molecules/asu, omit 22/276 residues  
from each molecule, model build with  
Arp/warp at different resolutions

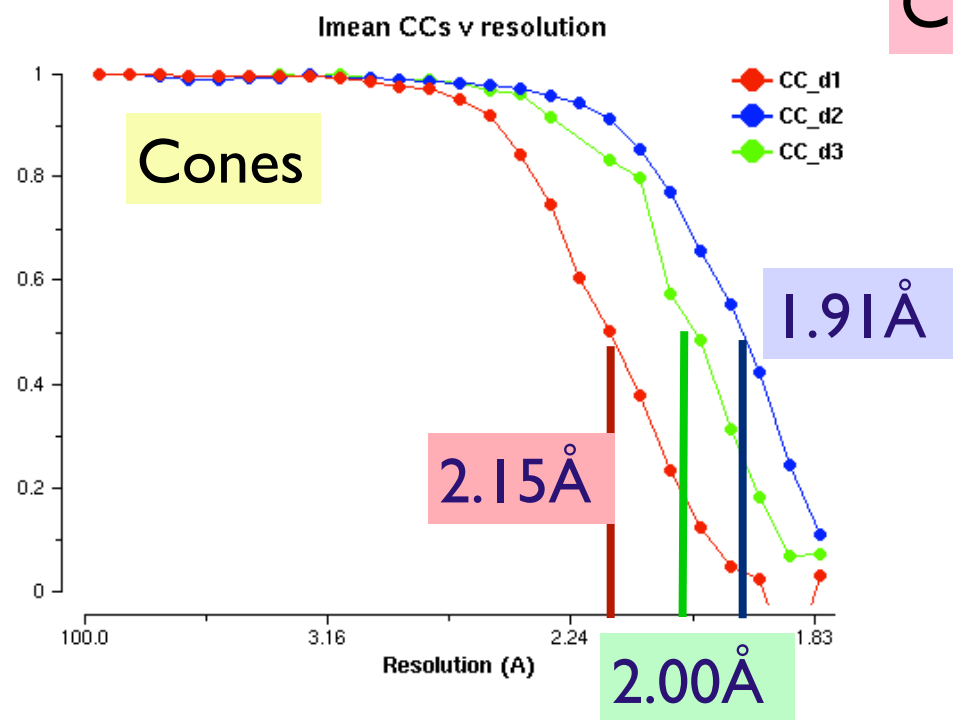


Number of  
residues **built**  
and **sequenced**

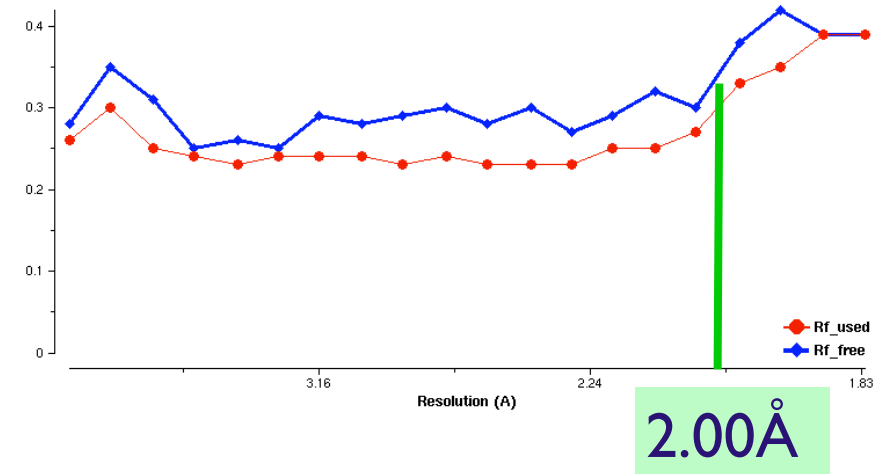
*figures made with ccp4mg*



Conclusion: there is not a huge difference



R &  $R_{\text{free}}$  after initial refinement

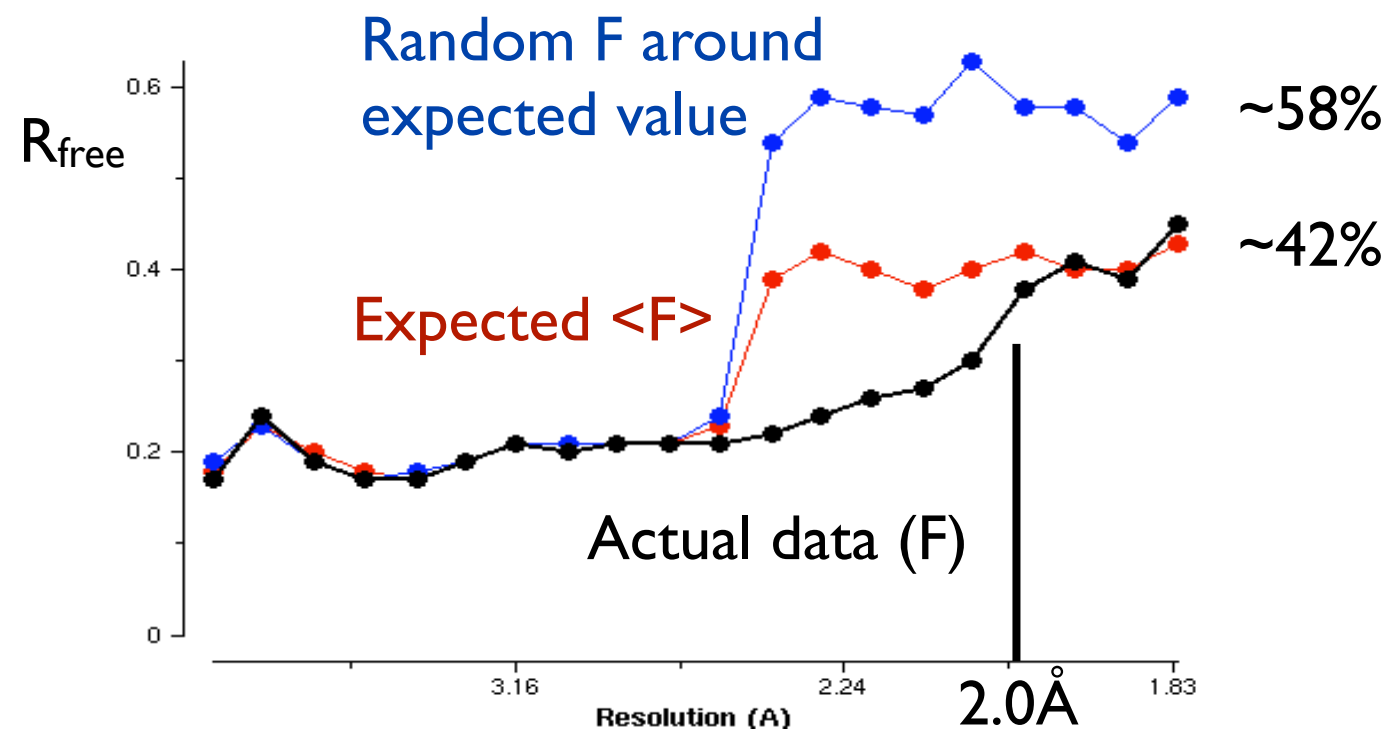




# Example continued: refinement against real data or simulated data



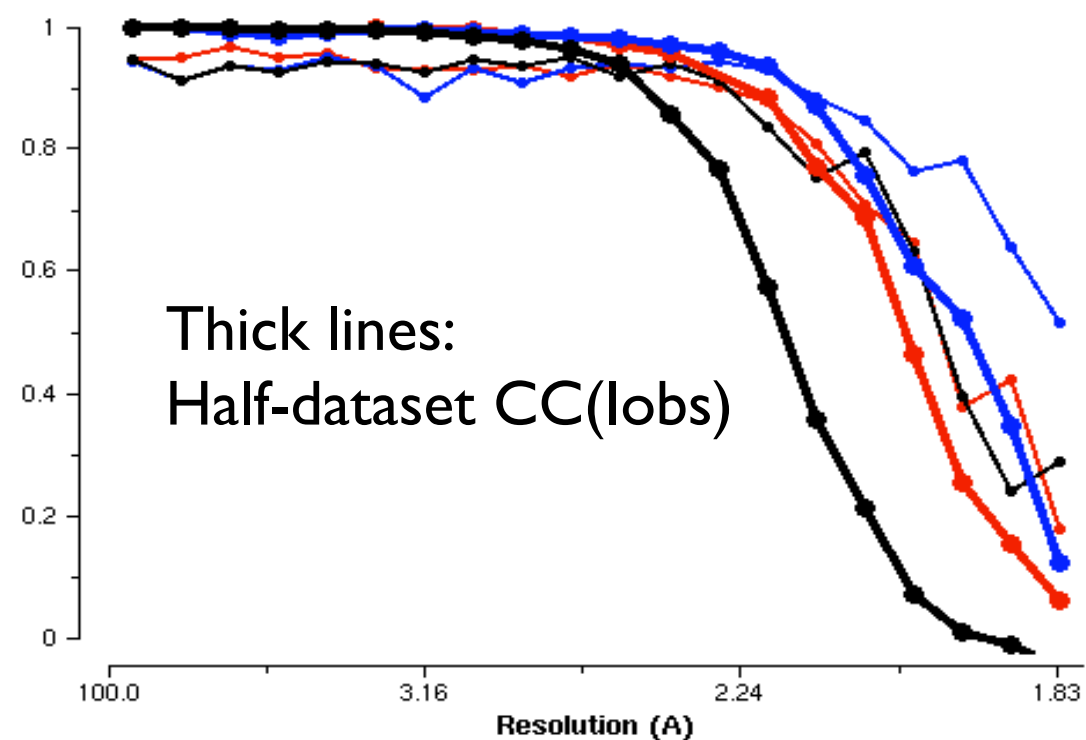
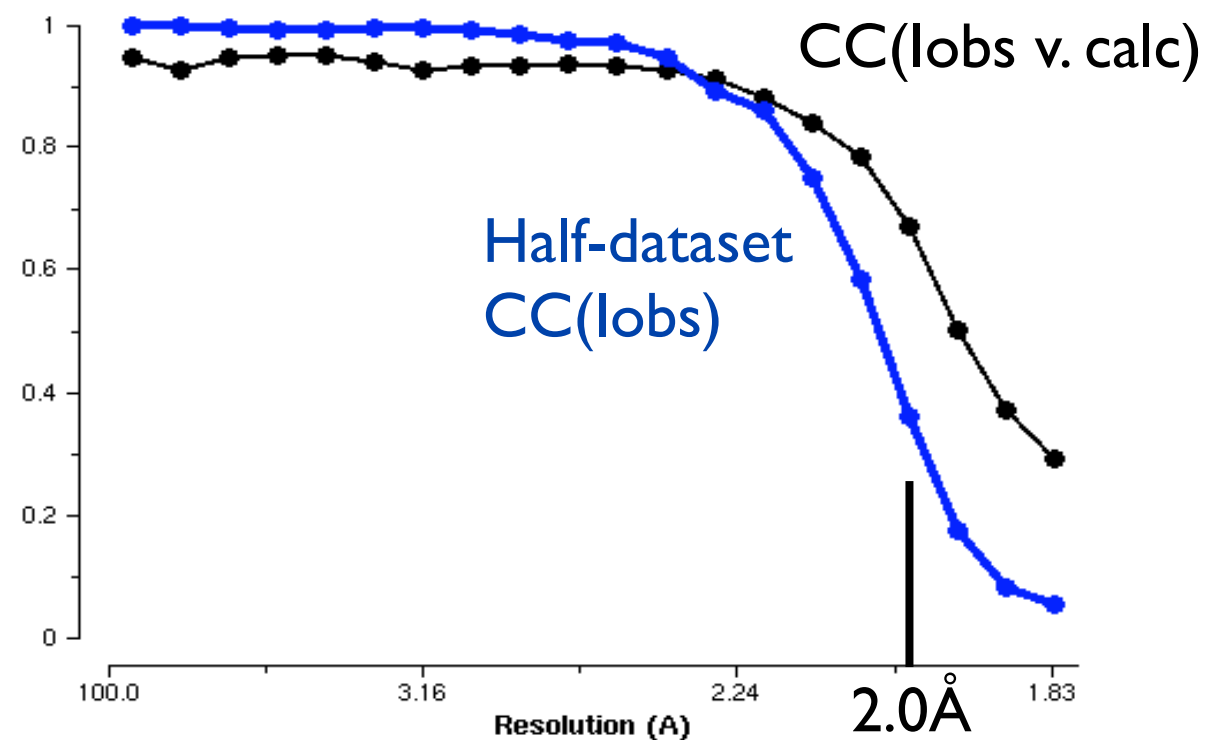
thanks to Garib Murshudov



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

Thin lines: CC(lobs v. calc)





# 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)

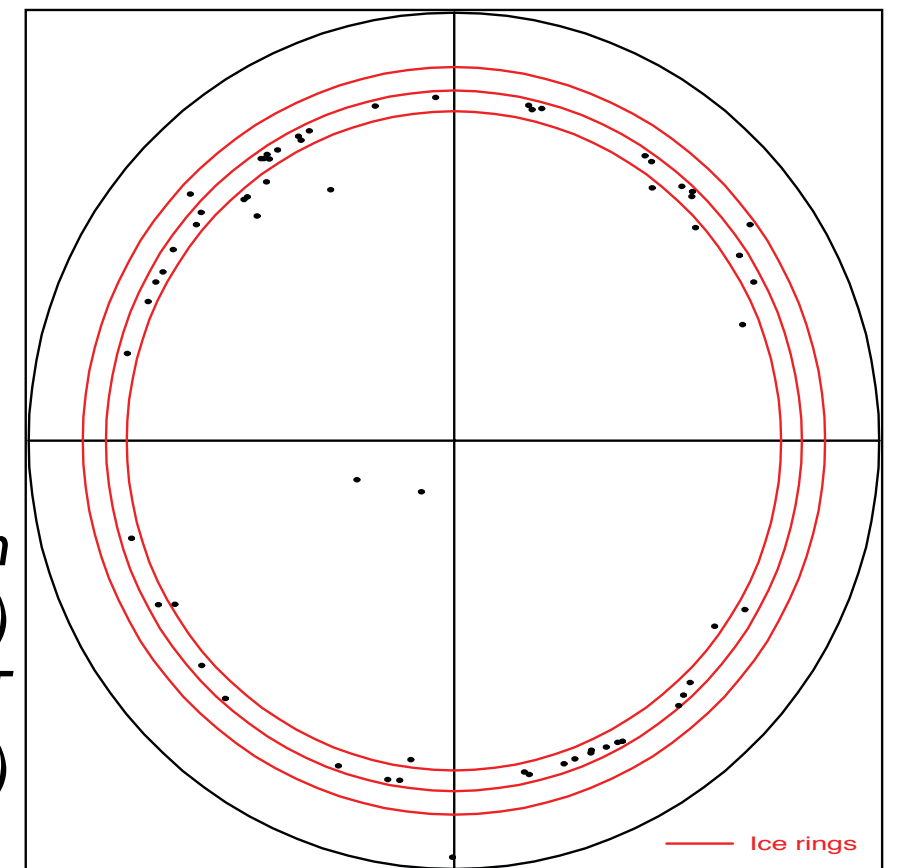
improve prediction

- spot overlap

lower mosaicity, smaller slice, move detector back

deconvolute overlaps

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

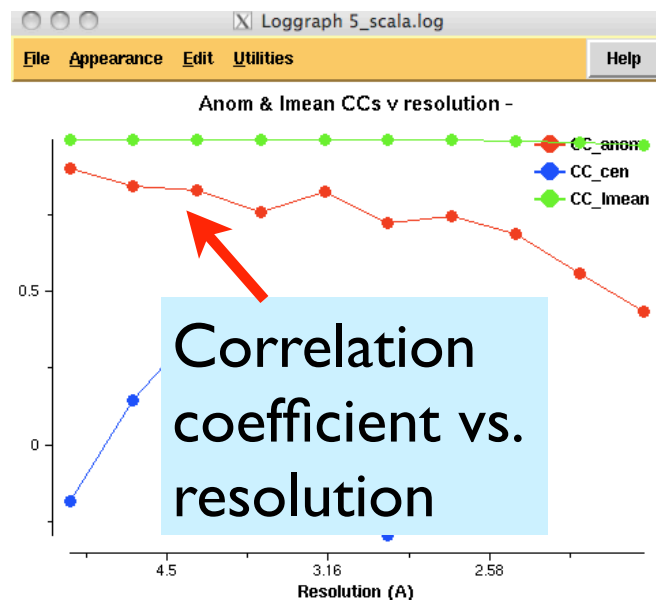


*Position of rejects on detector*



# Detecting anomalous signals

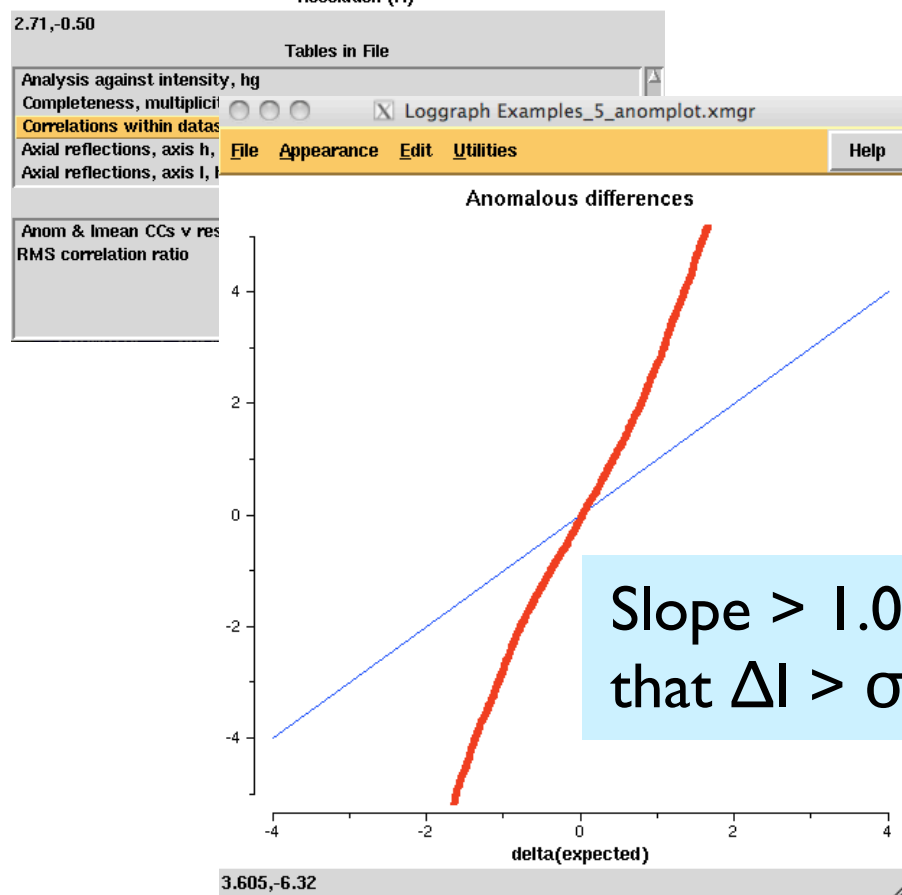
The data contains both I+ (hkl) and I- (-h-k-l) observations and we can detect whether there is a significant difference between them.



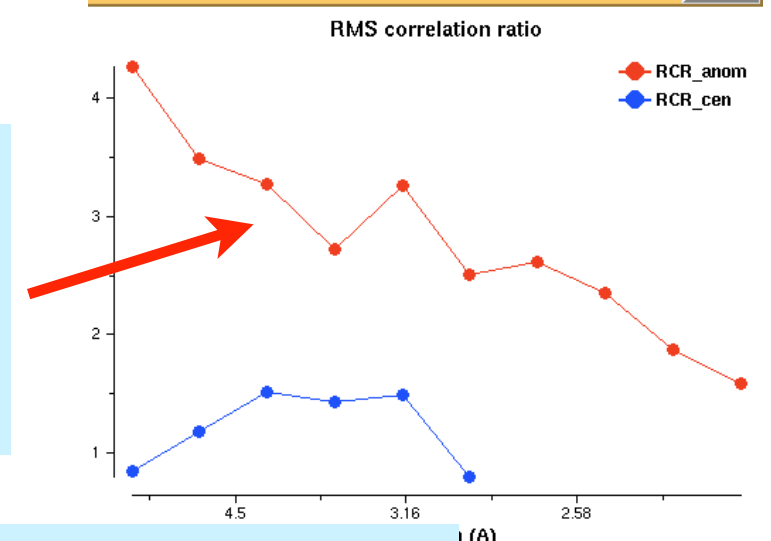
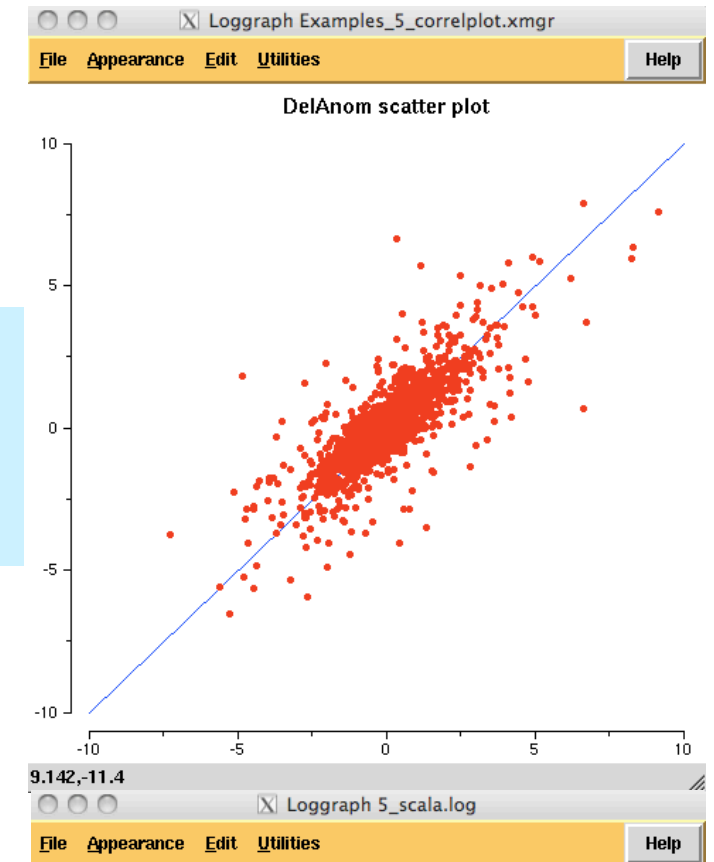
Split one dataset randomly into two halves, calculate correlation between the two halves or compare different wavelengths (MAD)

Plot  $\Delta I_1$  against  $\Delta I_2$  should be elongated along diagonal

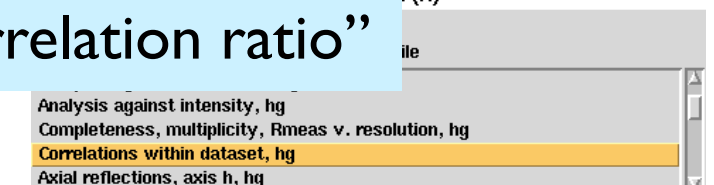
Strong anomalous signal



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.

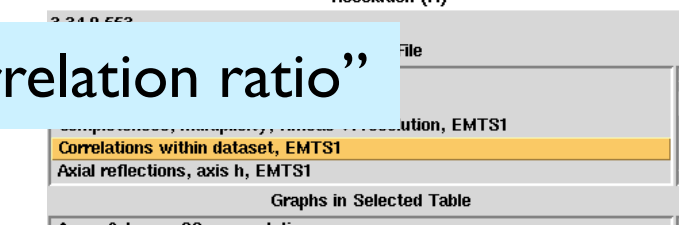
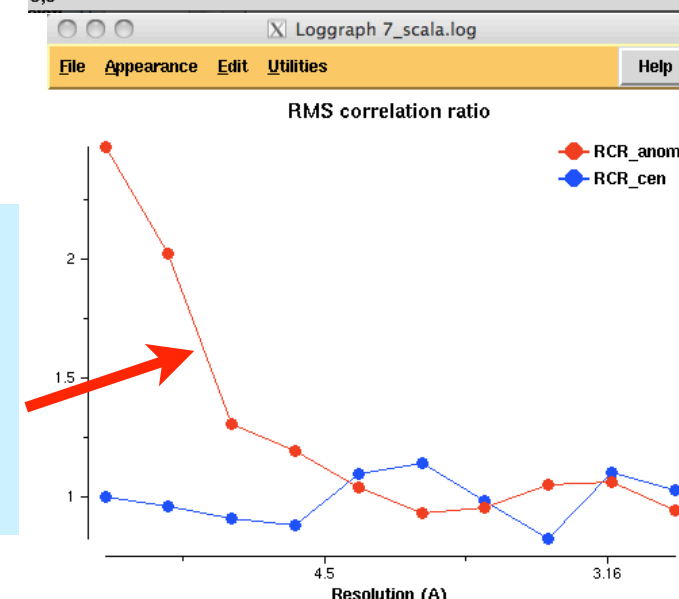
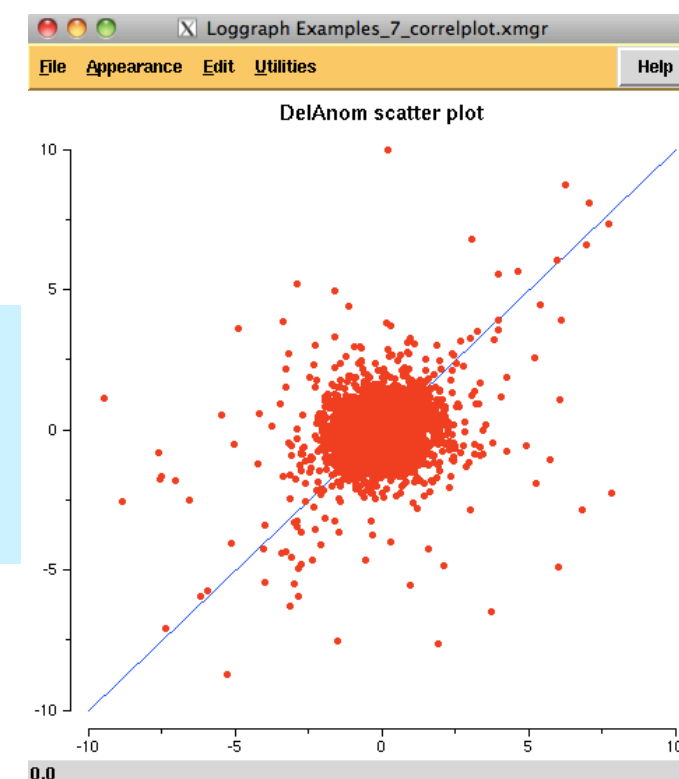
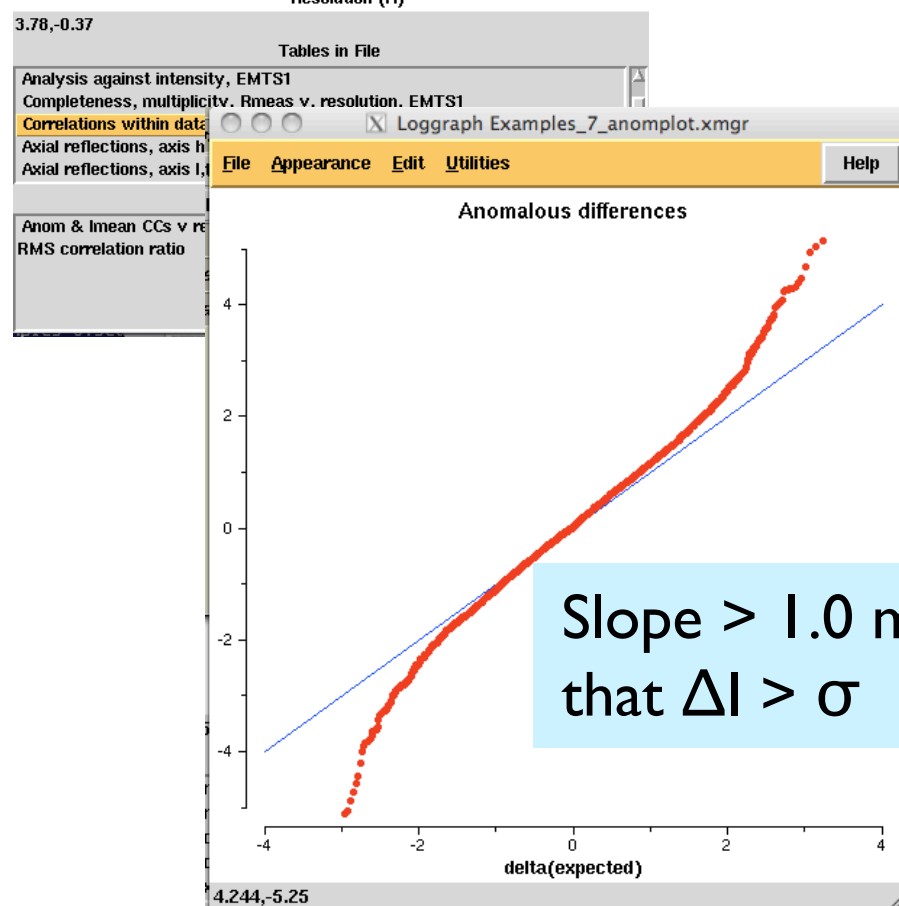
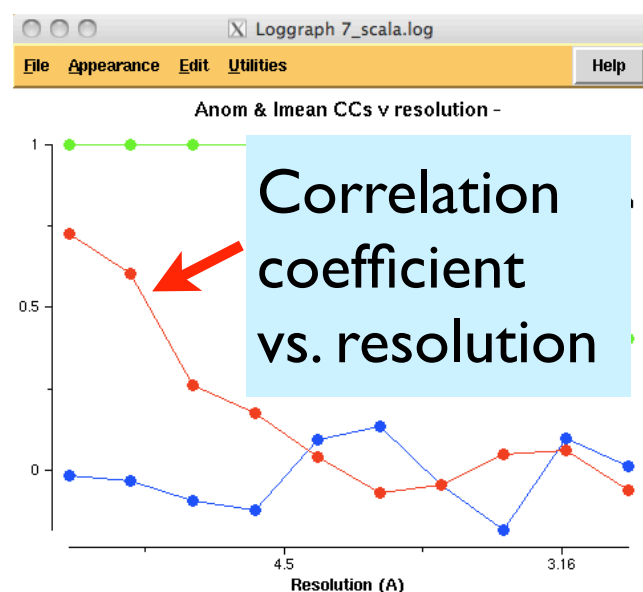
Split one dataset randomly into two halves, calculate correlation between the two halves or compare different wavelengths (MAD)

Plot  $\Delta I_1$  against  $\Delta I_2$  should be elongated along diagonal

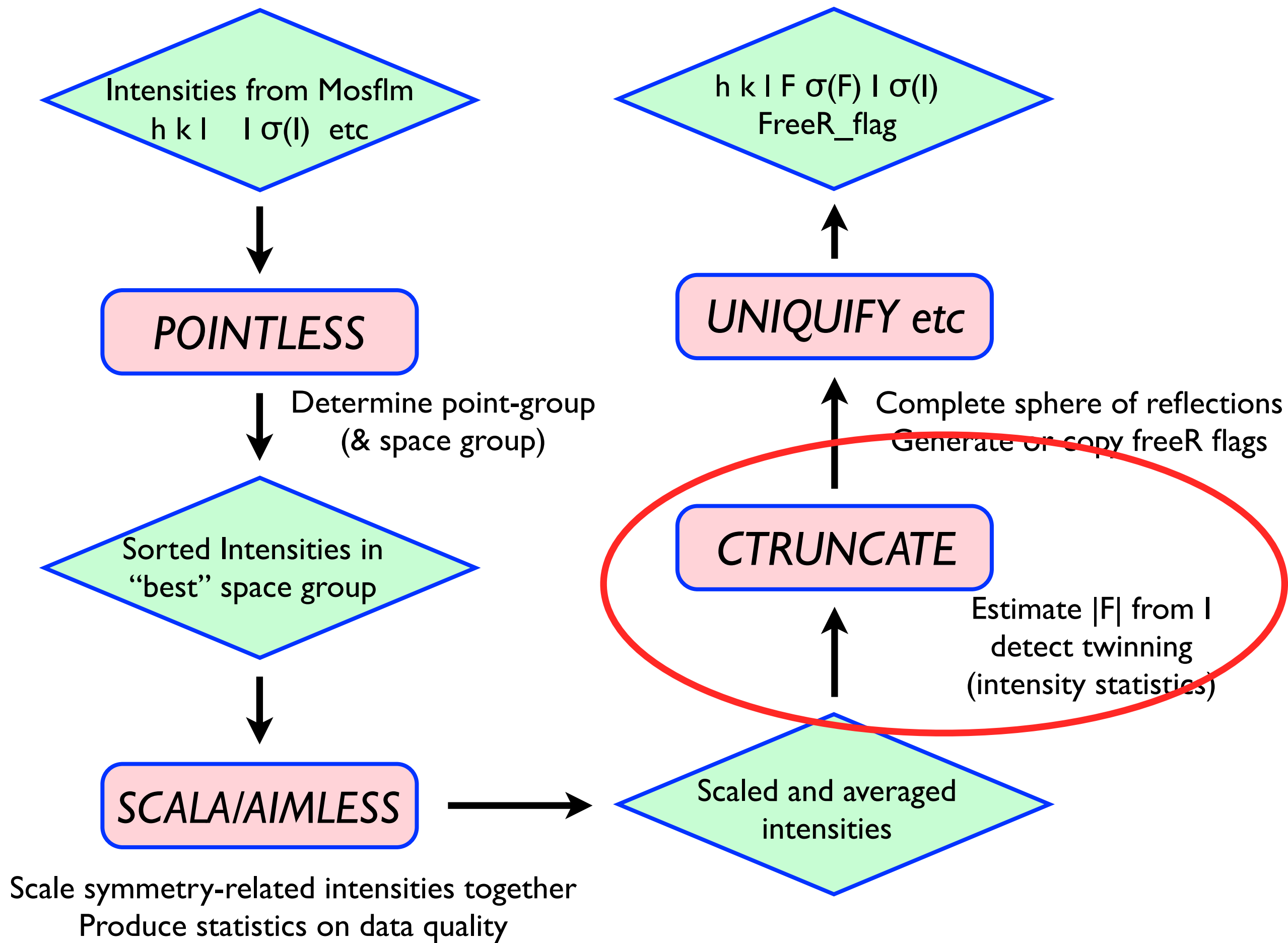
Weak but useful anomalous signal

Ratio of width of distribution along diagonal to width across diagonal

“RMS correlation ratio”









# 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  $\sigma(I)$  so a small intensity may be measured as negative.

The “best” estimate of  $|F|$  larger than  $\sqrt{I}$  for small intensities ( $< \sim 3 \sigma(I)$ ) to allow for the fact that we know that  $|F|$  must be positive

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

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

*French & Wilson 1978*



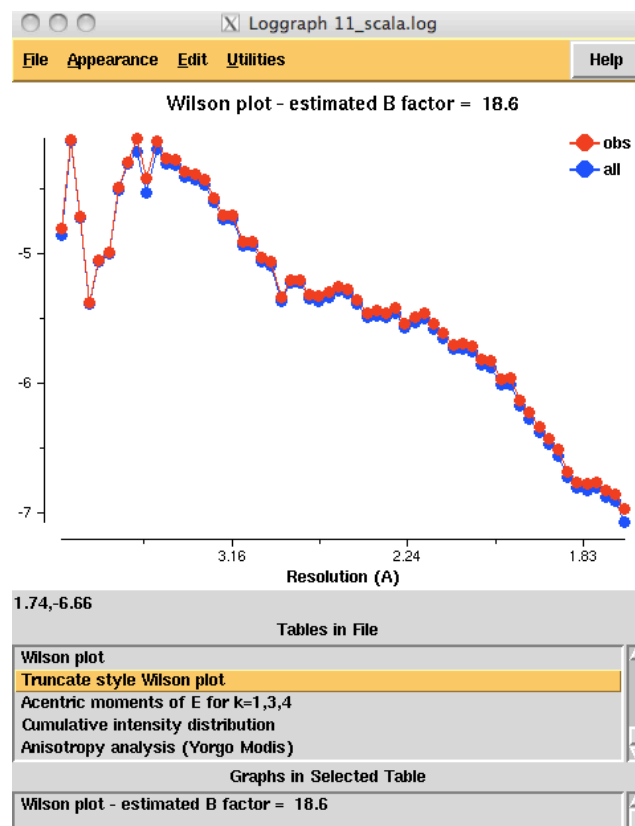
# Intensity statistics

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

Assuming atoms are randomly placed in the unit cell, then

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

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



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

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

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

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



*Normalised intensities*: relative to average intensity at that resolution

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

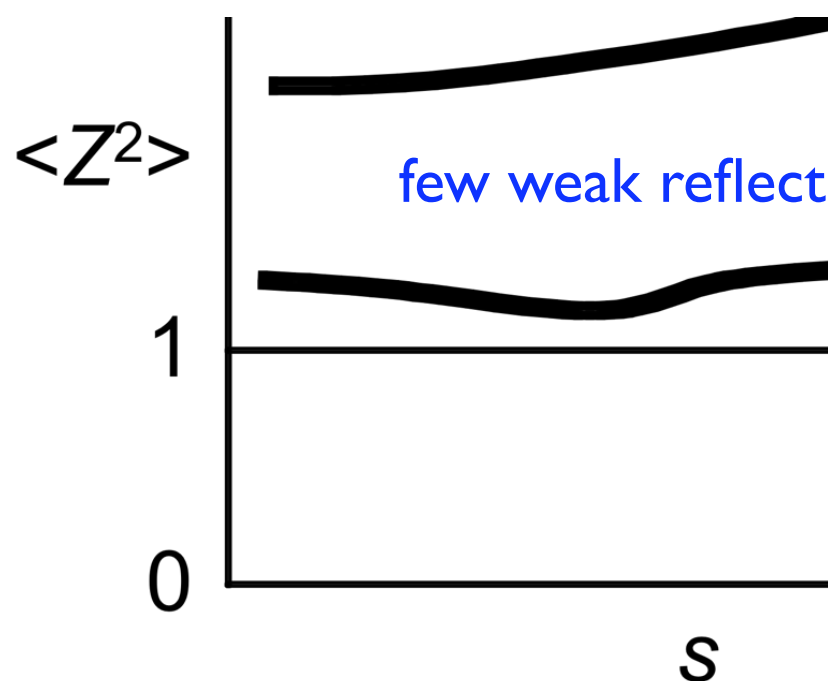
$$\langle Z(s) \rangle = 1.0 \text{ by definition}$$

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

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

many weak reflections

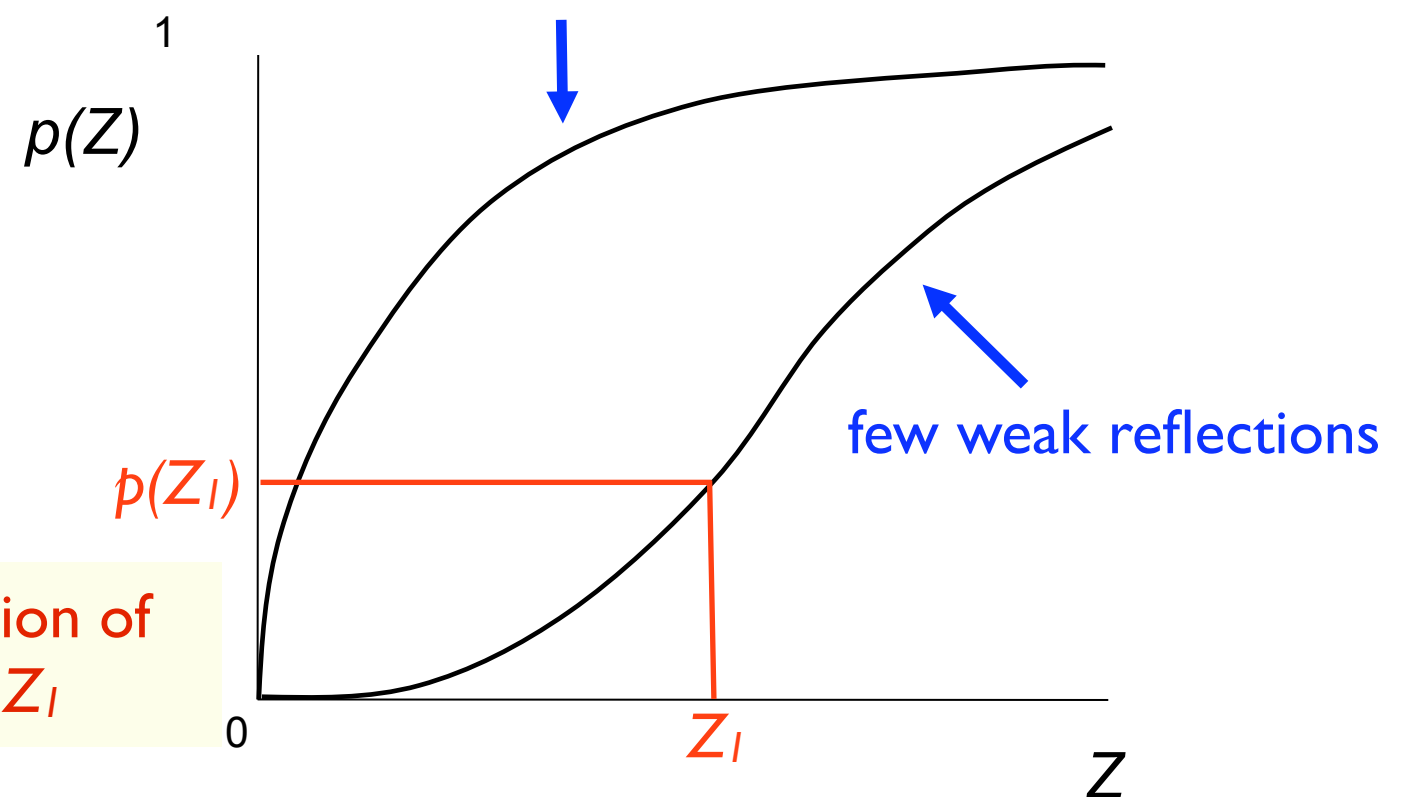
few weak reflections



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

many weak reflections

few weak reflections



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



## 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  $\sim 1/2, 0, 0$

$\langle I \rangle$  over all reflections is misleading, so  $Z$  values are inappropriate

The reflection classes should be separated (not yet done)

Anisotropy:  $\langle I \rangle$  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

- *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?

*Future developments may improve extraction of extra information from weak data at the resolution edge*

To help these developments, it would be useful to:

- Deposit data to higher resolution than you might use
- Deposit intensities as well as F
- Deposit unmerged data (is this possible?)
- Maybe deposit images?



# Acknowledgements

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