

Data reduction: scaling and merging data

DLS-CCP4 workshop
James Beilsten-Edmands
Diamond Light Source, U.K.



ACKNOWLEDGEMENTS



PROJECT SUPPORT



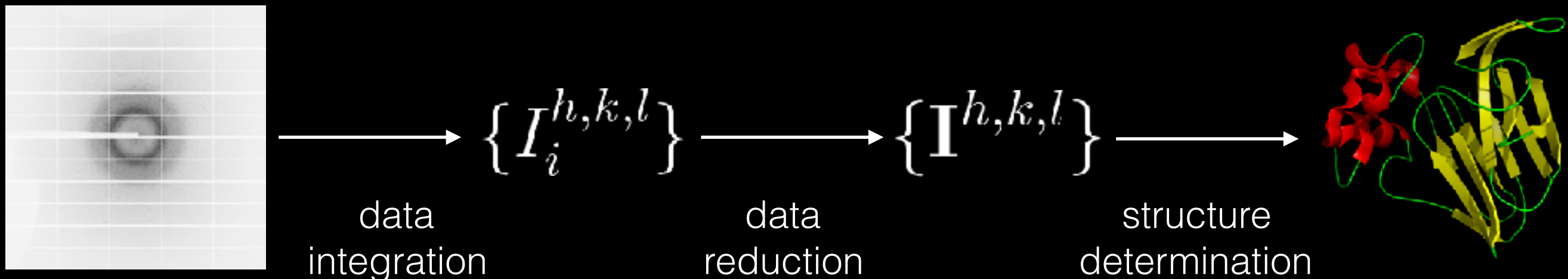
DATA REDUCTION

- Definition: reducing our observations, a set of integrated intensities, to a set of symmetry-unique intensities (and sigmas).

$$I_1^{(1,1,1)}, \sigma_1^{(1,1,1)}, I_2^{(1,1,1)}, \sigma_2^{(1,1,1)}, \dots \rightarrow \mathbf{I}^{(1,1,1)}, \boldsymbol{\sigma}^{(1,1,1)}$$

- These intensities should then be proportional to F^2

$$F_{hkl} = \sum_j^{atoms} f_j e(2\pi i(hx_j + ky_j + lz_j))$$

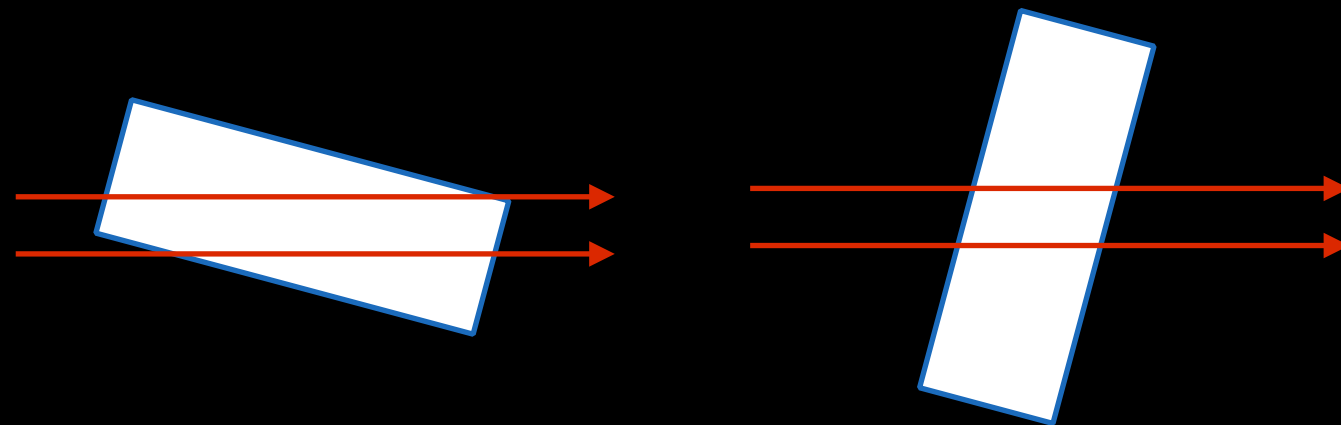


EXPERIMENTAL EFFECTS

Integrated intensities are a combination of F^2 , square of structure factor amplitudes, and experimental effects.

$$F_{hkl} = \sum_j^{atoms} f_j e(2\pi i(hx_j + ky_j + lz_j))$$

Why do symmetry-equivalent observations have different intensities?

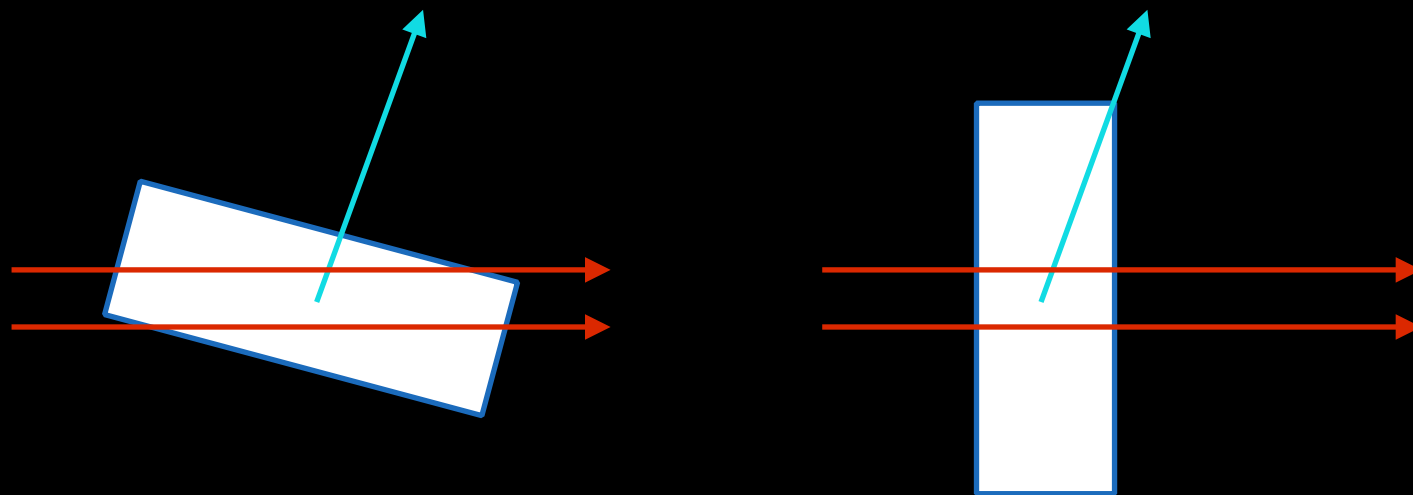


Beam effects - changes in beam intensity (slow), illuminated volume.

EXPERIMENTAL EFFECTS

Why do symmetry-equivalent observations have different intensities?

Crystal effects - absorption of the incident & scattered beam, radiation damage.

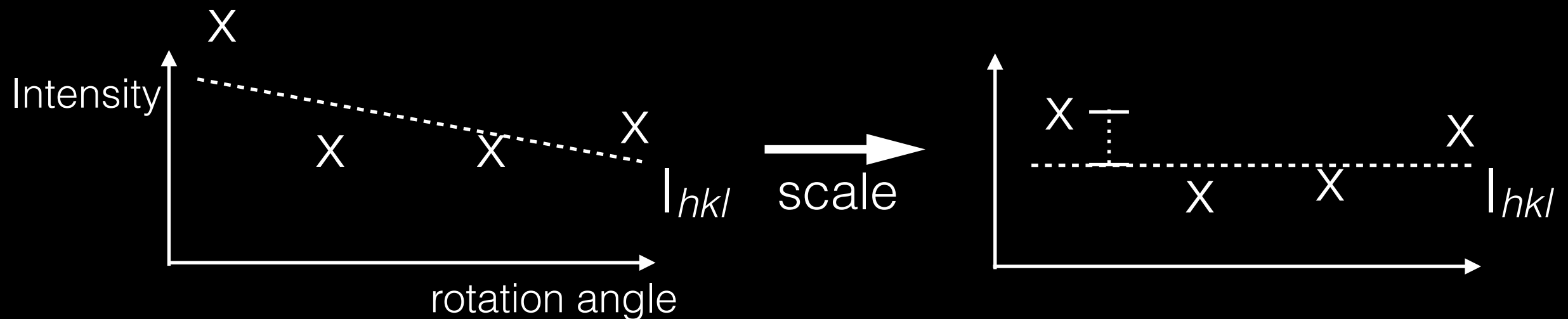


Detector effects - gain & calibration issues, edge effects (CCDs).

SCALING AND MERGING

Use the redundancy of multiple measurements (of symmetry equivalent reflections) to determine a model to account for experimental effects.

Determine scale factors for the reflections, so that symmetry equivalent reflections have the best agreement - least deviation from a weighted average for the group.



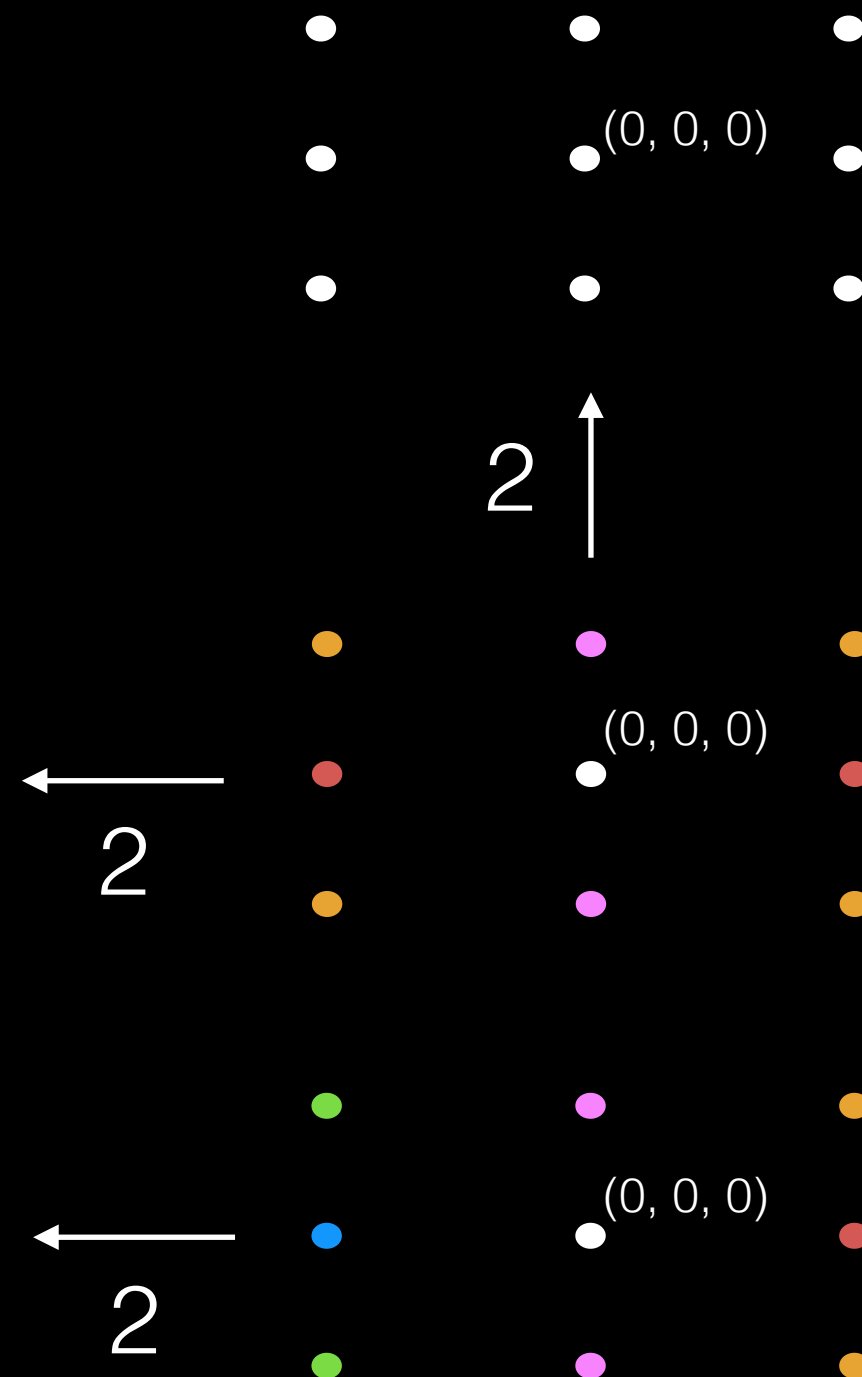
$$\delta = I_{hkl}^i - g^i \langle I_{hkl} \rangle$$

PRECURSOR: SYMMETRY ANALYSIS

Before integration, only know symmetry of the lattice (spot positions), not the diffraction pattern (intensities).

Point group symmetry should be evaluated, so that intensities are correctly grouped in scaling.

Can scale in lower symmetry point group, but then not using full equivalence of certain observations!



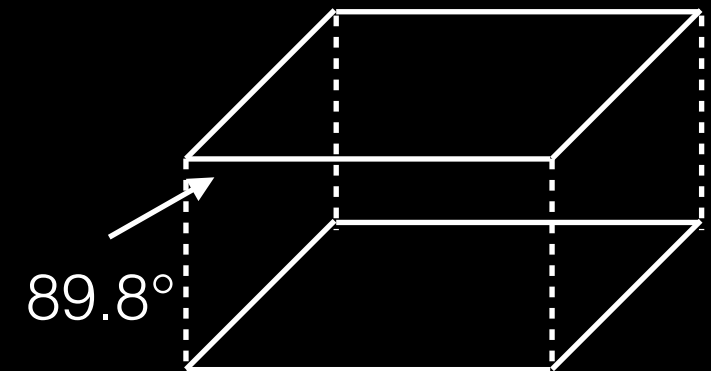
PRECURSOR: SYMMETRY ANALYSIS

Lattice symmetry $\neq$ diffraction pattern symmetry

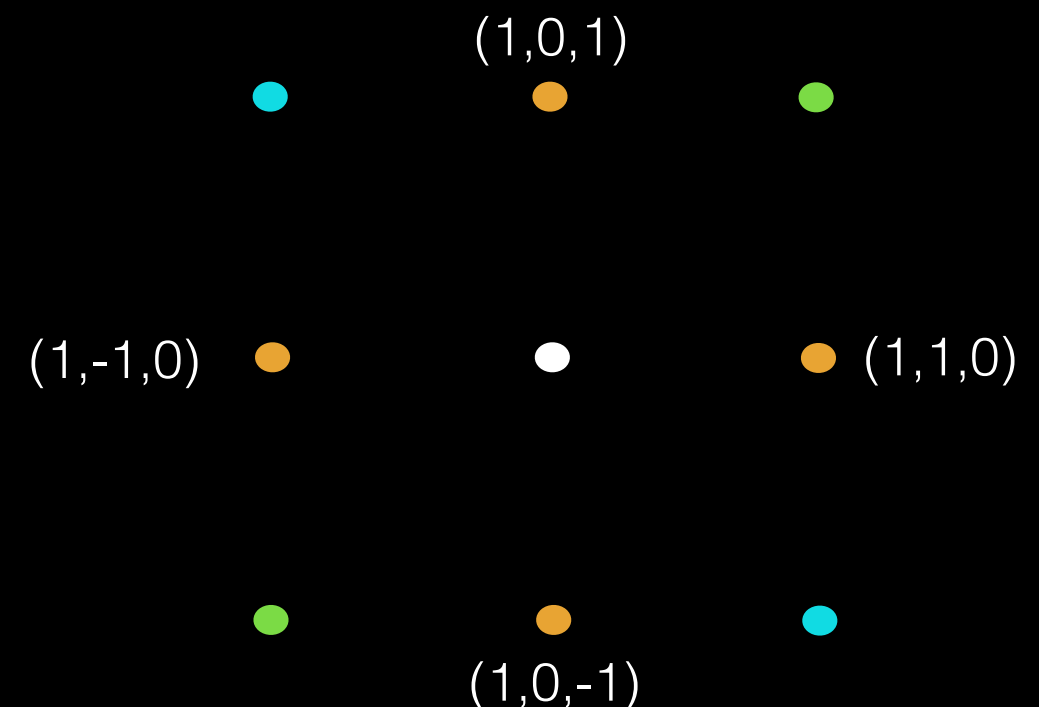
May have `accidental` symmetry of the lattice, e.g.
P1 symmetry with angle close to 90° looks like P2.

Also able to have point group symmetry lower
than the lattice symmetry.

Programs can use: *Pointless*, *dials.symmetry*, *XDS*
(*CORRECT* step)



Point group 23, (1,0,0) plane
Lattice symmetry 432



PRECURSOR: SYMMETRY ANALYSIS

Score symmetry element based on correlation between intensities.

Score different Patterson groups based on combination of symmetry elements of the lattice.



●



Cubic
insulin

Scoring individual symmetry elements

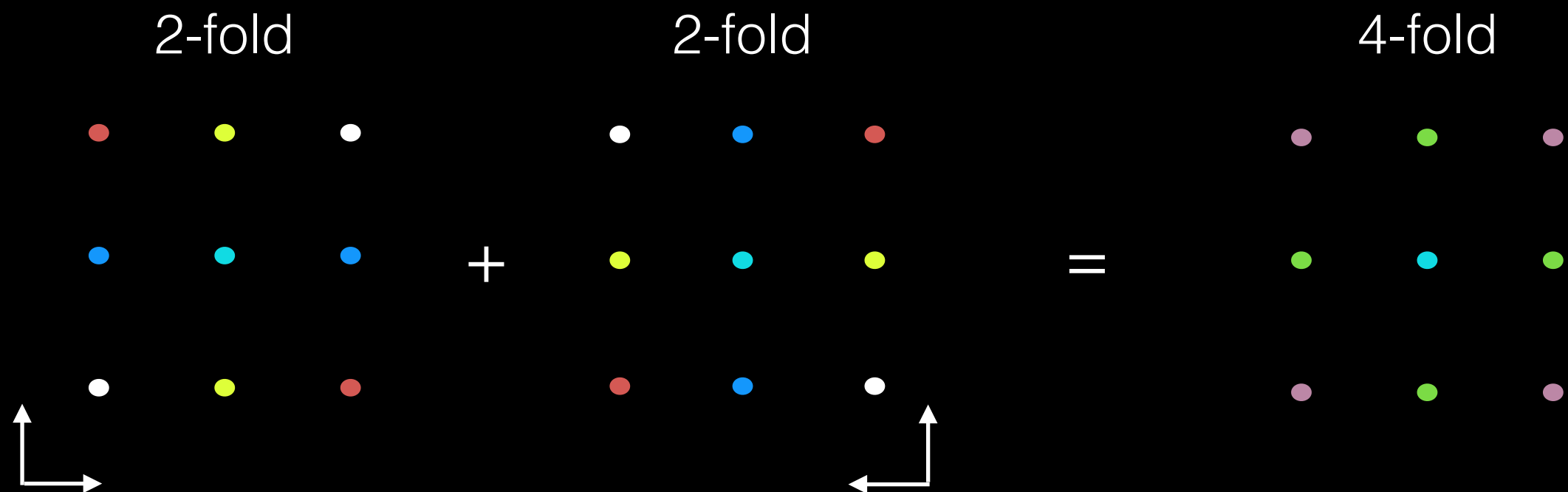
likelihood	Z	CC	CC	N		Operator
0.943	9.90	9.99	905647	***	1	(2, 0, 0)
0.196	5.02	9.98	1694945		4	(1, 1, 0)
0.196	5.02	9.98	1693679		4	(1, 0, 1)
0.196	5.03	9.98	1694485		4	(0, 1, 1)
0.931	9.67	9.97	1713055	***	3	(1, 0, 0)
0.939	9.80	9.98	1715415	***	3	(2, 1, 0)
0.939	9.81	9.98	1715539	***	3	(2, 0, 1)
0.932	9.68	9.97	1712444	***	3	(1, 1, 1)
0.933	9.70	9.97	856927	***	2	(1, 1, 0)
0.197	5.04	9.98	847218		2	(0, 1, 1, 0)
0.938	9.78	9.98	857342	***	2	(1, 0, 1)
0.197	5.05	9.98	851699		2	(0, 1, 1)
0.926	9.59	9.96	876485	***	2	(2, 1, 1)
0.195	5.02	9.98	862772		2	(0, 0, 1, 1)
0.197	5.05	9.98	847708		2	(1, 1, 2)
0.197	5.05	9.98	847688		2	(1, 2, 1)
0.198	5.06	9.91	855659		2	(2, 1, 1)

Scoring all possible sub-groups

Patterson group		Likelihood	NetZee	Zee1	Zee	CU	CO	delta	Reindex operator
I 11 -3	***	1.000	4.70	5.74	5.84	0.97	2.60	0.0	-a,b,-c
I 11 11 11		0.000	2.91	5.74	6.84	0.97	2.70	0.0	-a,b,-c
I 11 -3 11		0.000	7.62	7.62	8.82	0.74	2.60	0.0	-a,b,-c
I 11 11 11 11		0.000	0.00	0.00	0.00	0.74	2.60	0.0	-a,b,-c

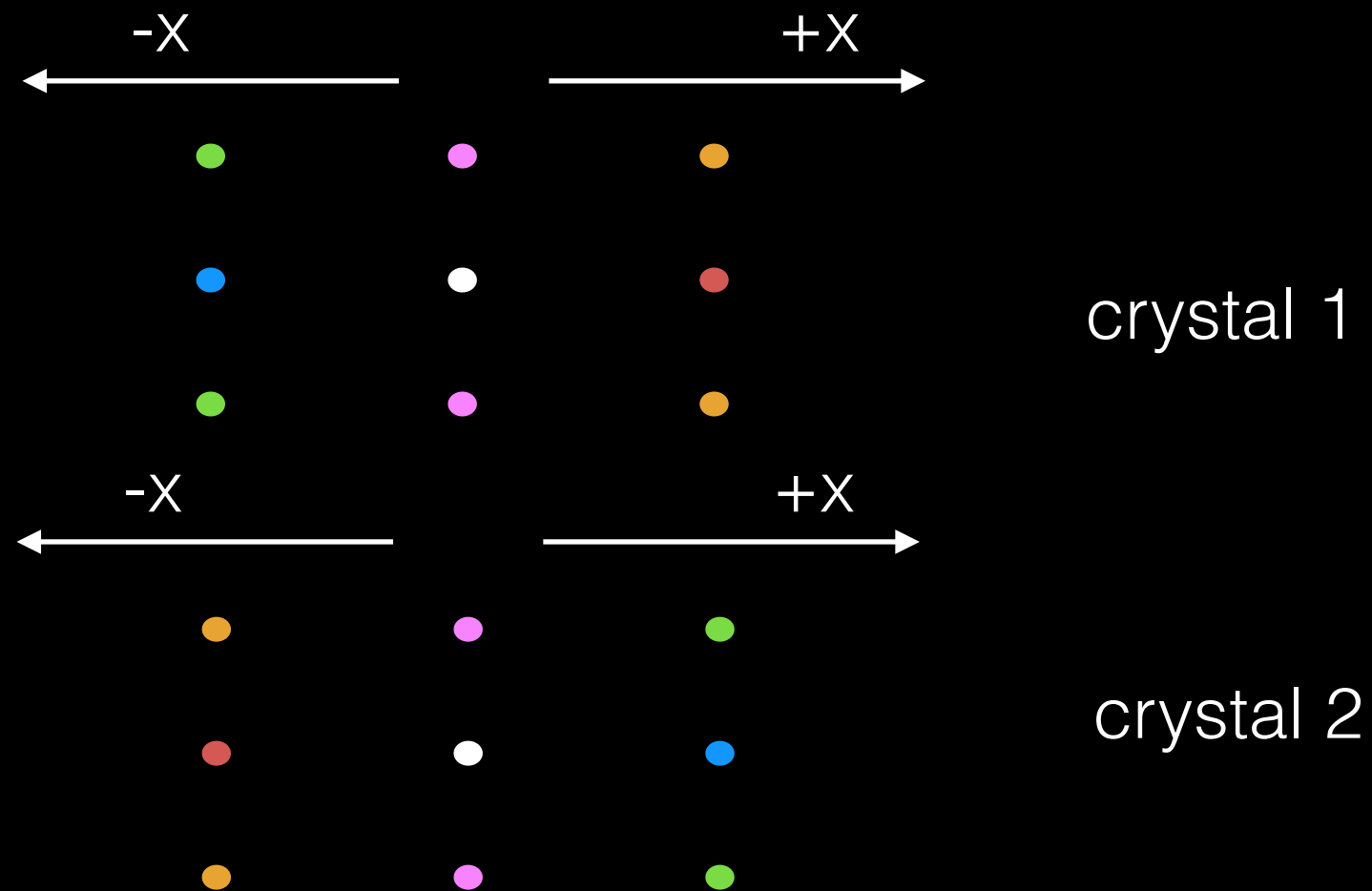
SYMMETRY ANALYSIS DIFFICULTIES

Effects of twinning & pseudo-symmetry, can give appearance of higher symmetry — see other talks! Can not be certain that you know the symmetry until you have solved the structure!



SYMMETRY ANALYSIS DIFFICULTIES

Indexing ambiguities in multi-crystal experiments, for certain point groups, different ways to index.



Need consistent indexing in order to scale. A particular issue for sparse multi-crystal datasets - can use *dials.cosym* [1] to solve symmetry and reindex. Or use *pointless/dials.reindex/XDS* to reindex against reference.

[1] Determination of Patterson group symmetry from sparse multi-crystal data sets in the presence of an indexing ambiguity. Gildea, R. J. & Winter, G. (2018)

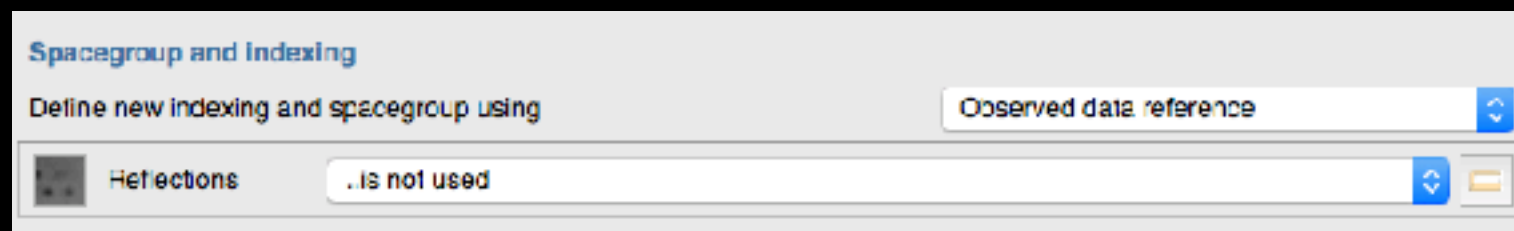
SYMMETRY ANALYSIS DIFFICULTIES

Indexing ambiguities

dials.cosym

```
Space groups:  
R 3 :H  
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,  
 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44,  
 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,  
 66, 67, 68, 69]  
Reindexing operators:  
-x,-y,-z  
[1, 3, 6, 7, 8, 9, 10, 13, 15, 16, 19, 20, 21, 22, 24, 25, 26, 32, 33, 36, 37, 38, 39,  
 40, 44, 45, 48, 49, 50, 51, 55, 56, 57, 60, 61, 62, 63, 64, 65, 66, 67, 69]  
z,x,y  
[0, 2, 4, 5, 11, 12, 14, 17, 18, 23, 27, 28, 29, 30, 31, 34, 35, 41, 42, 43, 46, 47,  
 52, 53, 54, 58, 59, 68]
```

Pointless - ccp4i2



XDS - option REFERENCE_DATA_SET=

RECAP: DATA REDUCTION STEP 1

Assess point group (space group) symmetry*.
Beware of difficulties.

Integrated reflections



Integrated reflections, in correct point group

*Unless have prior knowledge for this system, or want to process in $P1$.

SCALING

Determine scale factors for the reflections, so that symmetry equivalent reflections have the best agreement. Makes the data *internally consistent*.

Create a parameterised scaling model and refine the model to give most internally consistent data.

Once data is scaled, can assess the internal consistency with a number of measures [1] - note that this assesses the *precision*, not the *accuracy*, need to be aware of possibility for large systematic errors. Recommended metrics: $CC_{1/2}$ (for resolution cutoff), R_{pim} , R_{meas} , $\langle I/\sigma(I) \rangle$ (signal to noise)

Programs can use: Aimless, XDS/XSCALE, dials.scale, scalepack

[1] Assessing and maximising data quality in macromolecular crystallography, Karplus & Diederichs, 2015.

SCALING PROCESS COMPONENTS

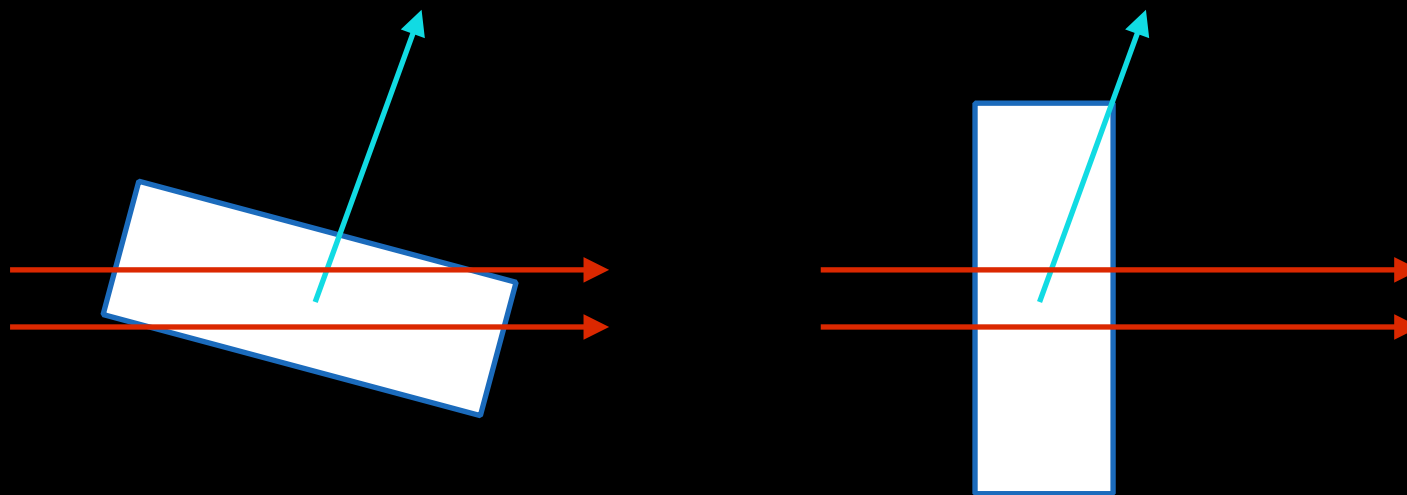
- rounds of outlier rejections (right reflections)
- rounds of scaling model refinement (best scales)
- model the error distribution (sigmas) (better errors)
- combine profile and summation intensity estimates (best integrated intensities)

$$I, \sigma \rightarrow g \rightarrow I', \sigma'$$

SCALING MODELS

Principle - experimental effects multiplicative in their effect on intensities - scale factor given by product of corrections.

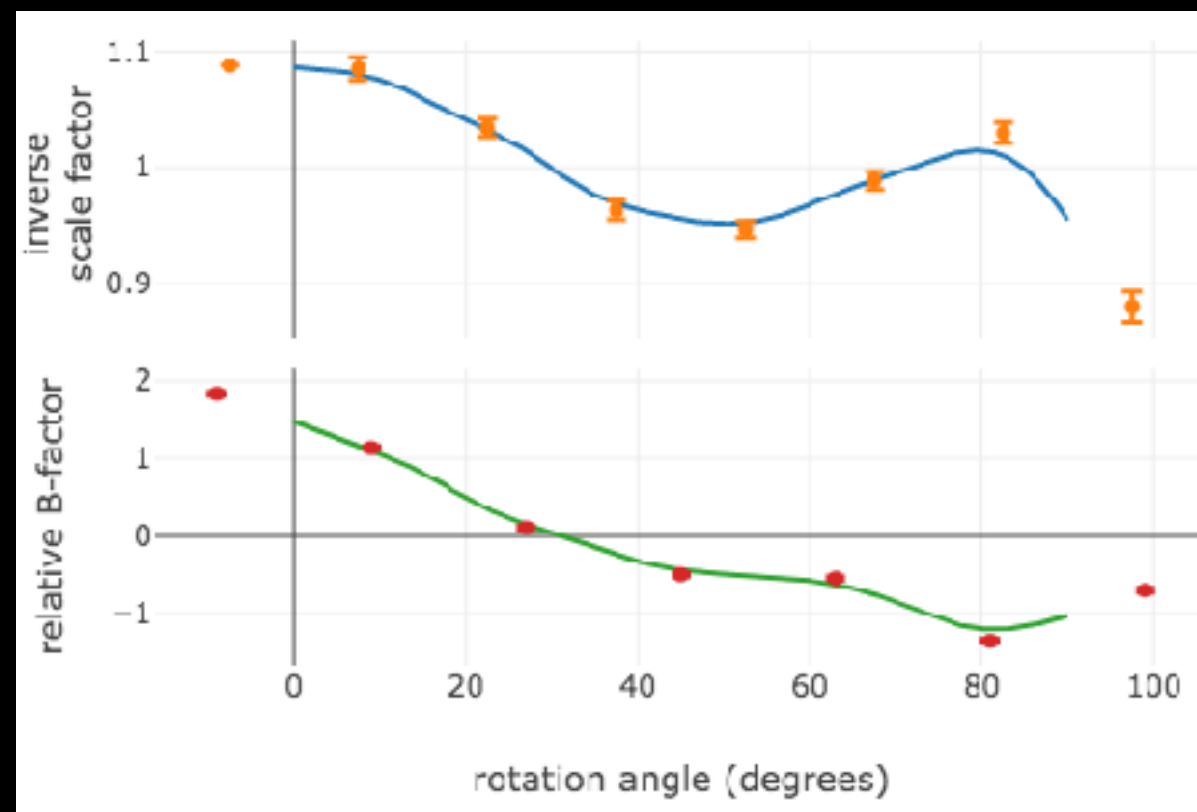
$$g_{hl} = S_{hl}(p_1) \times T_{hl}(p_2) \times \dots$$



PHYSICAL SCALING MODEL

Three component model. Used in Aimless & DIALS

- Smoothly varying scale factor
 - Smoothly varying isotropic B-factor $e(B(t) / 2d^2)$
 - Absorption correction (relative)
- } smooth weighting between model parameters



PHYSICAL SCALING MODEL

Absorption correction

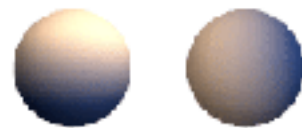
Correct for relative variations in absorption in different directions in space - based on incoming and outgoing scattering vectors. Typically small for MX ($\sim 1\%$), unless at long wavelength/absorption edge.

Define a correction surface, correction depends on direction relative to crystal frame. Use spherical harmonics as a basis to define a smoothly varying surface.

$l = 0$



$l = 1$



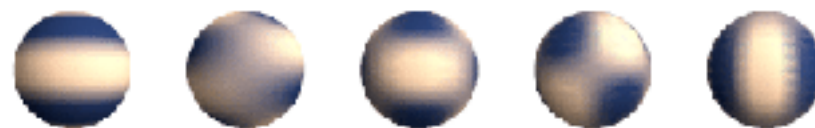
$l = 2$



$l = 3$



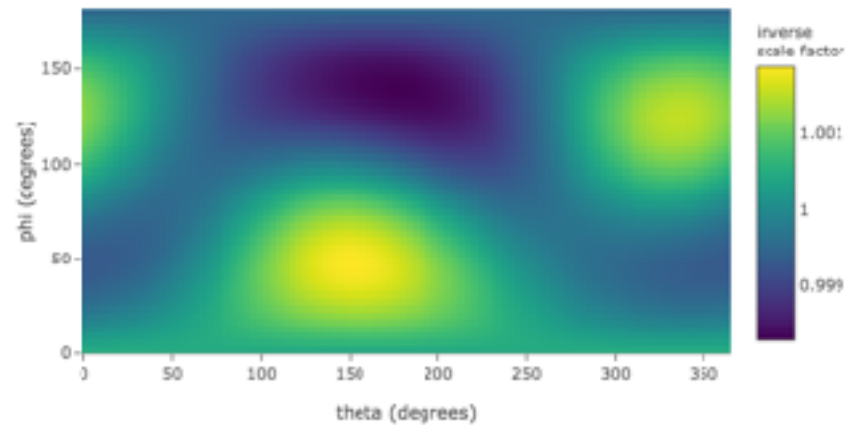
$l = 4$



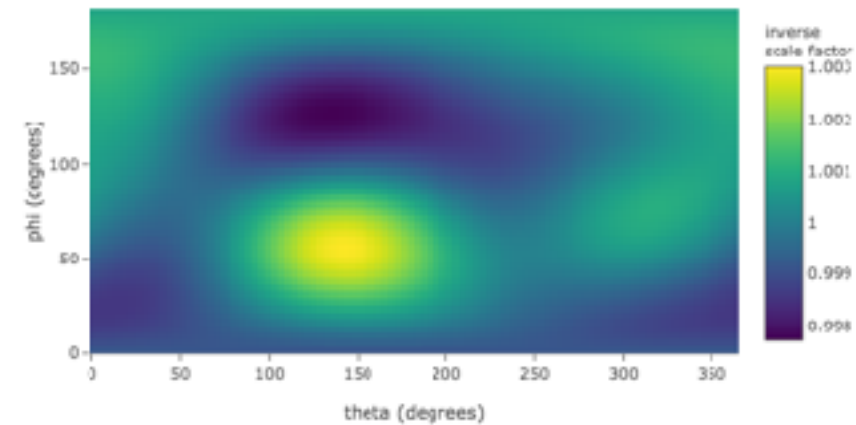
PHYSICAL SCALING MODEL

Absorption correction

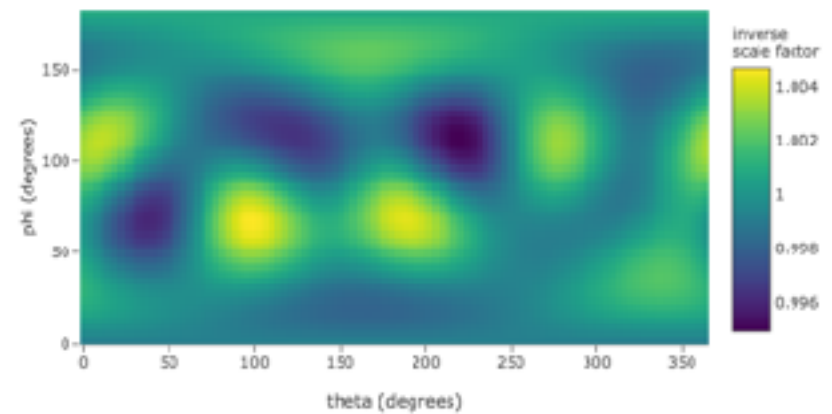
$L = 2$



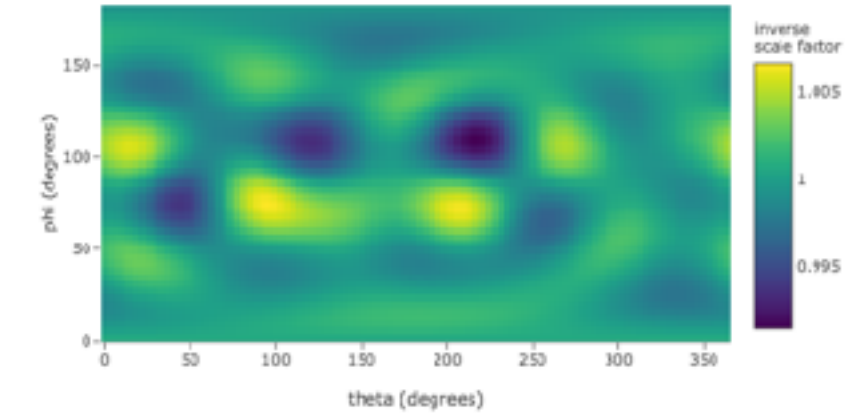
$L = 3$



$L = 4$



$L = 6$



ARRAY SCALING MODEL

More generalised approach - just define what the corrections depend on. e.g d-value, rotation angle, position on the detector. XDS/XSCALE approach.

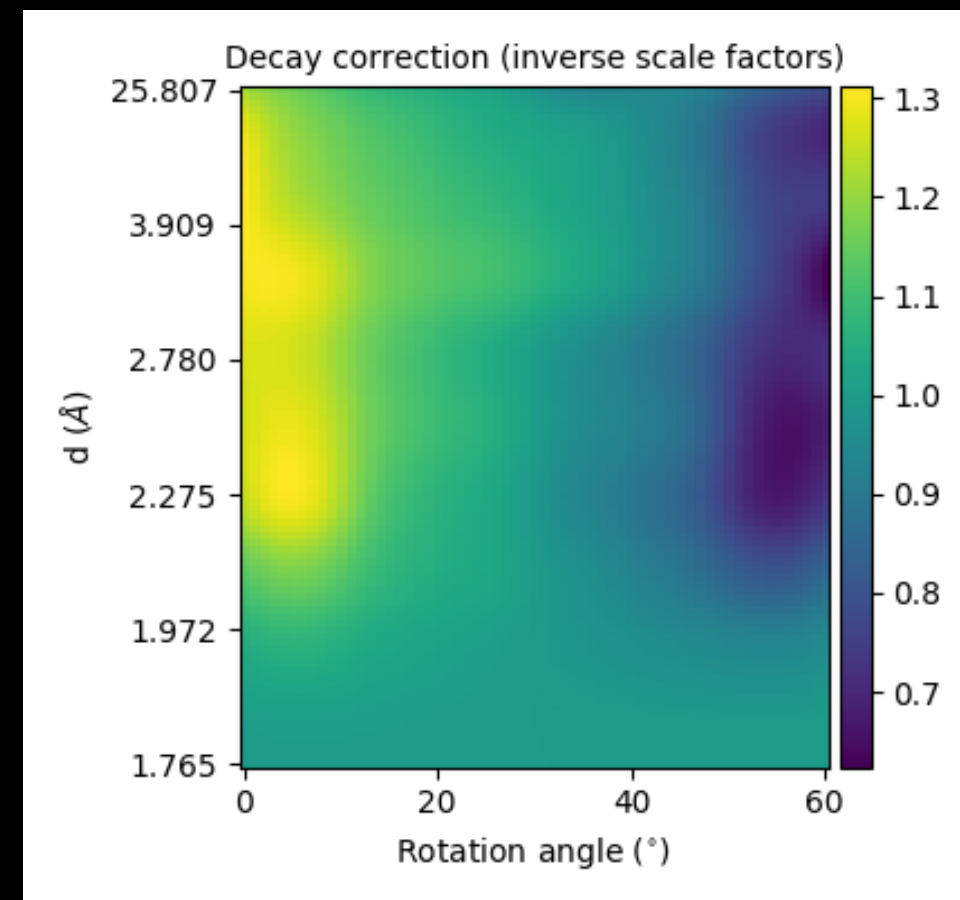
Array model; dials.scale implementation of this approach. Smoothly weight a grid of parameters to get correction for a reflection.

Decay correction - 2D grid of parameters; d-values & rotation angle.

Absorption correction - 3D grid of parameters; reflection x, y coordinate on detector, rotation angle.

Advantage: less restrained by assumptions.

Disadvantage: physical interpretation not always clear. Many more parameters than physical model (~5-10x)



OUTLIER REJECTION

Outliers can have a disproportionate effect on scaling - largest RMSD.

Test the normalised deviation from the symmetry group, reject those above a threshold (default 6 normalised deviations)

$$\delta_{hl} = (I_{hl} - g_{hl} \langle I_h \rangle') / \sigma_d$$

$$\sigma_d^2 = \sigma^2(I_{hl}) + g^2 \sigma^2(\langle I_h \rangle')$$

Several rounds of outlier rejection - at each point retest *all* reflections.

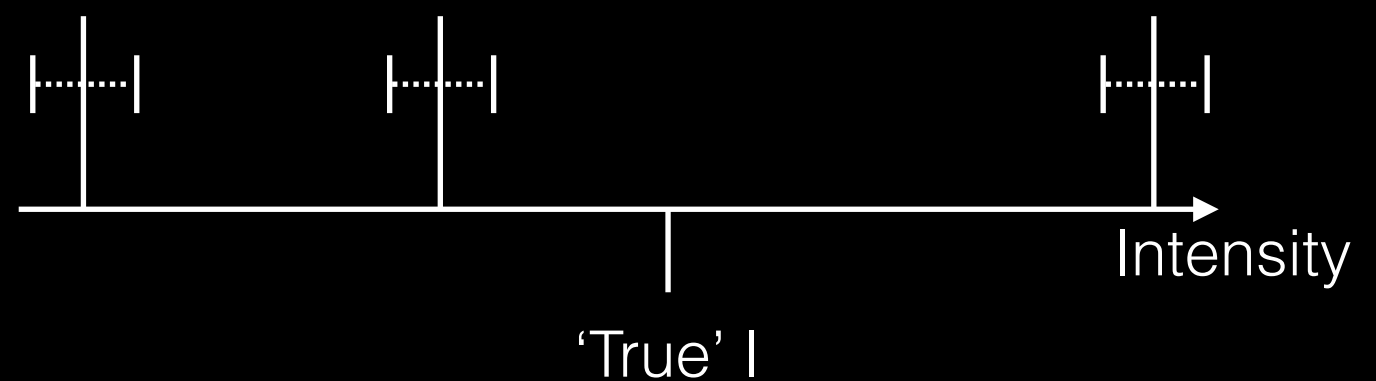
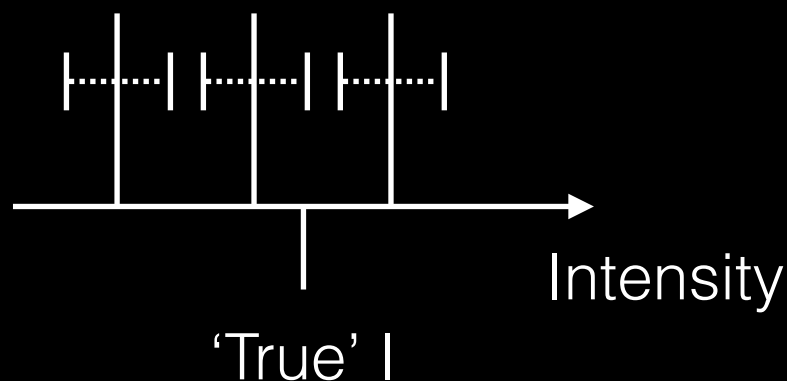
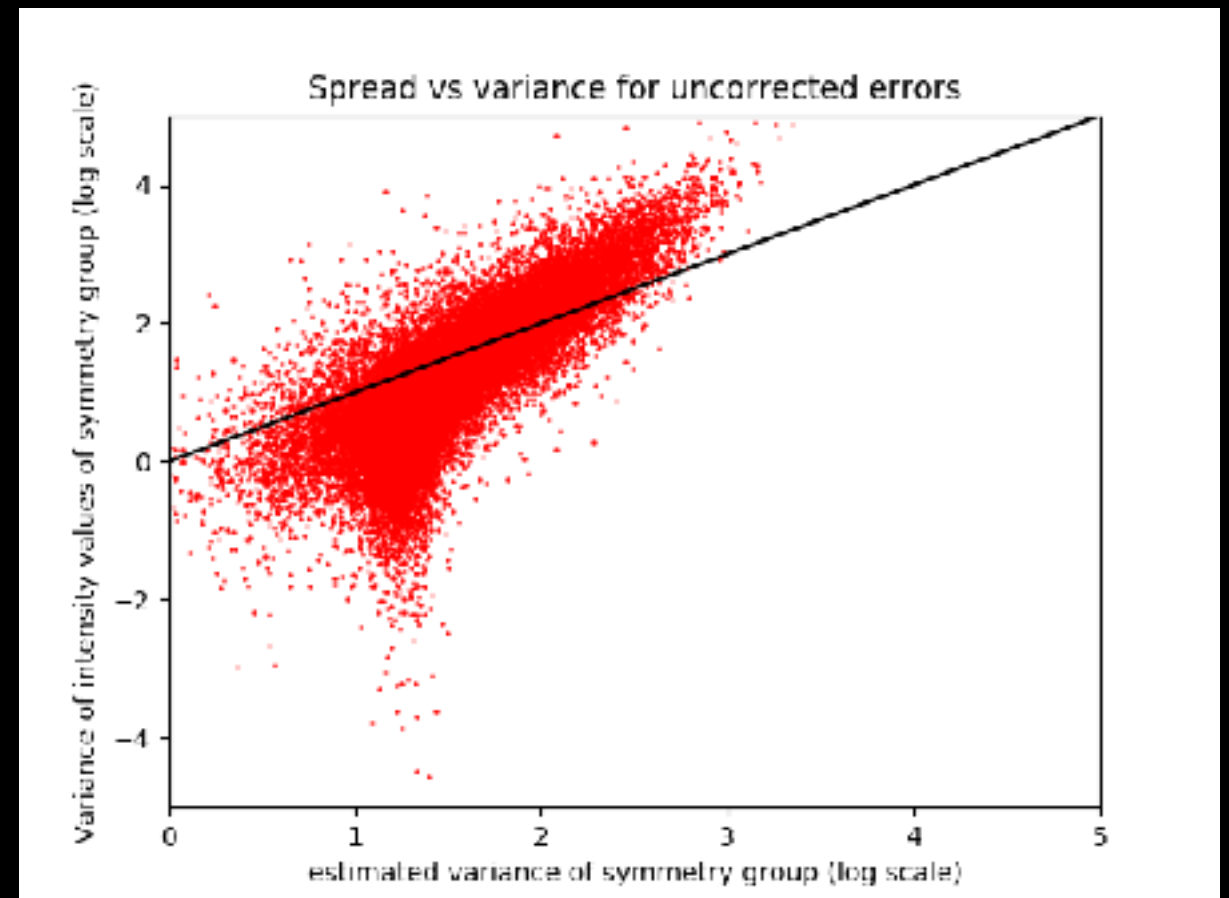


ERROR (SIGMA) DISTRIBUTION

At scaling step, able to investigate the distribution of intensities and errors (uncertainties) throughout the dataset.

Assumption; if error values are correct, then *on average* the distribution of observations should equal the estimated standard deviation. Typically find that for the most intense reflections the estimated standard deviation is lower - error is underestimated!

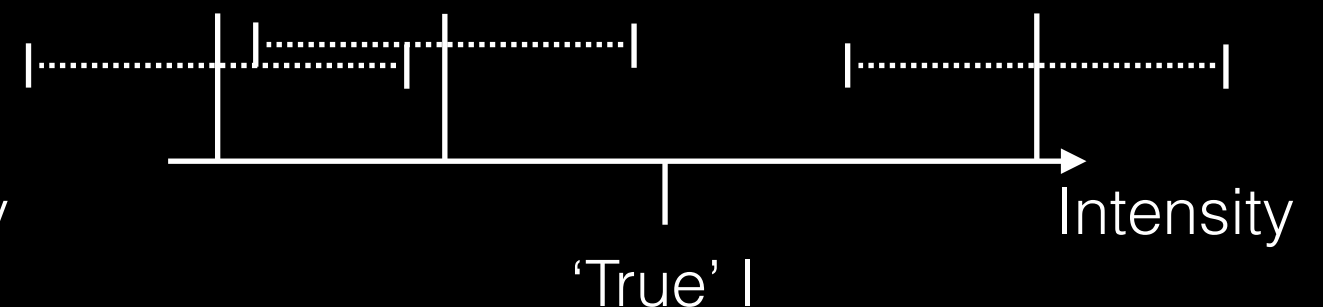
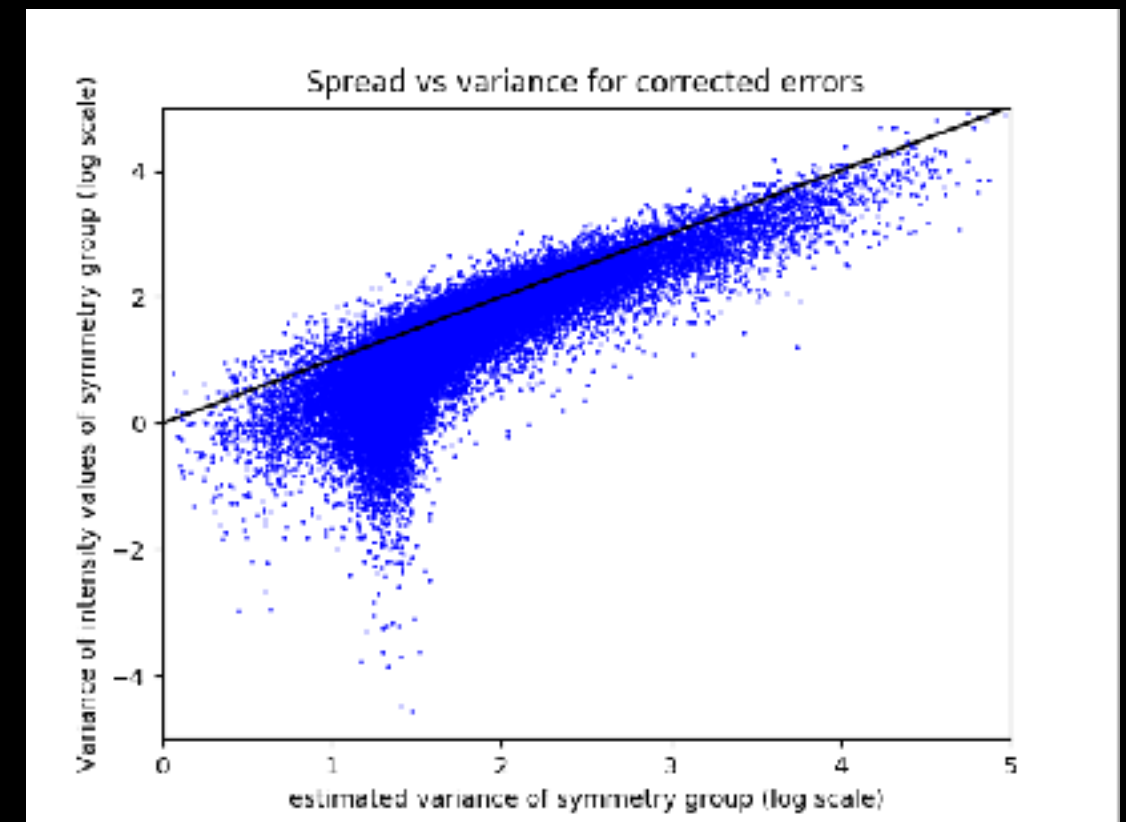
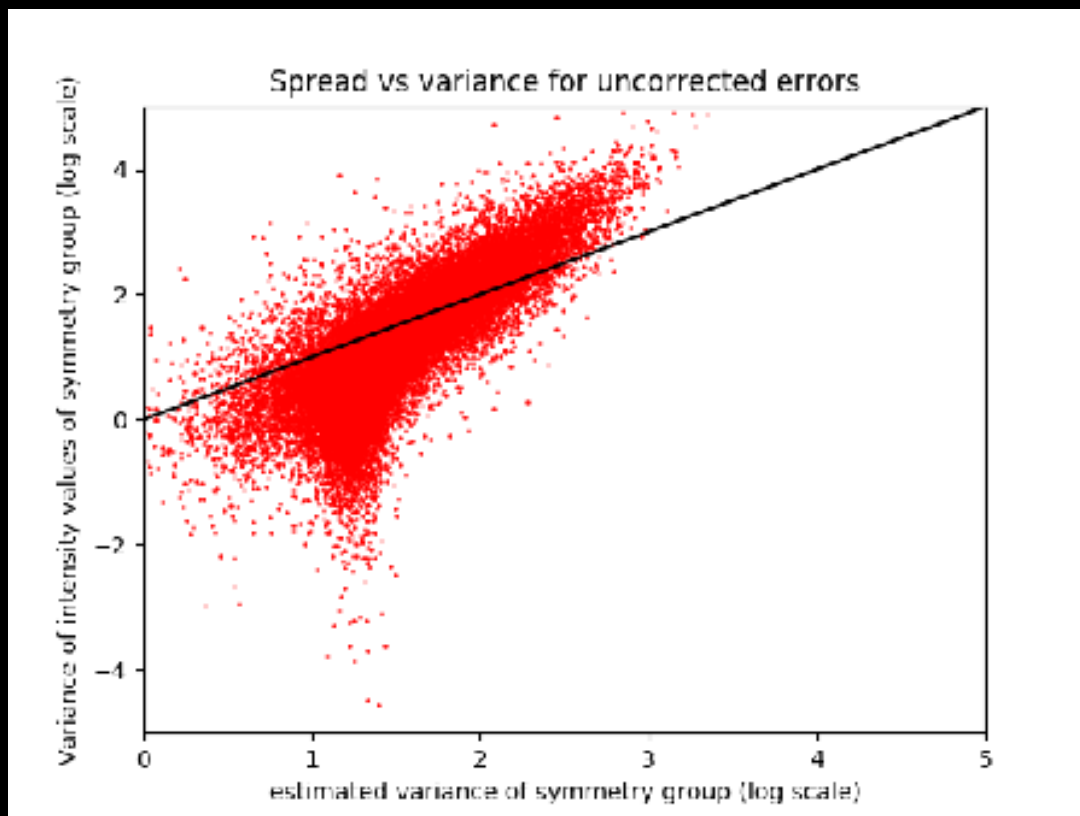
Systematic errors may not have been fully accounted for, alongside deficiencies in experimental models (including scaling).



INTENSITY ERROR DISTRIBUTION

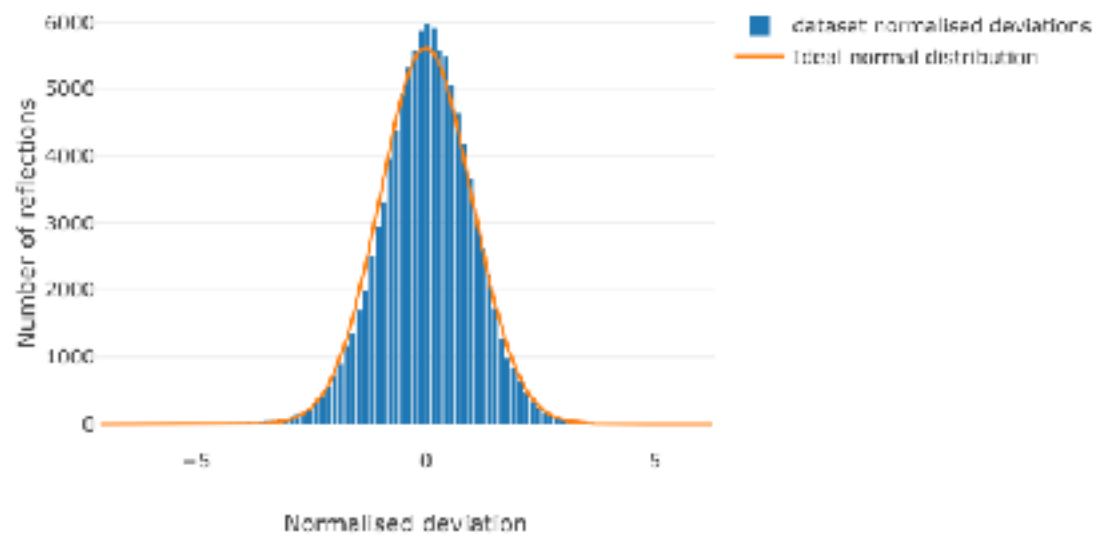
Two component error model, adjust errors as a function of intensity, by looking at distribution of intensities.

$$\sigma'^2 = a^2(\sigma^2 + (bI)^2)$$

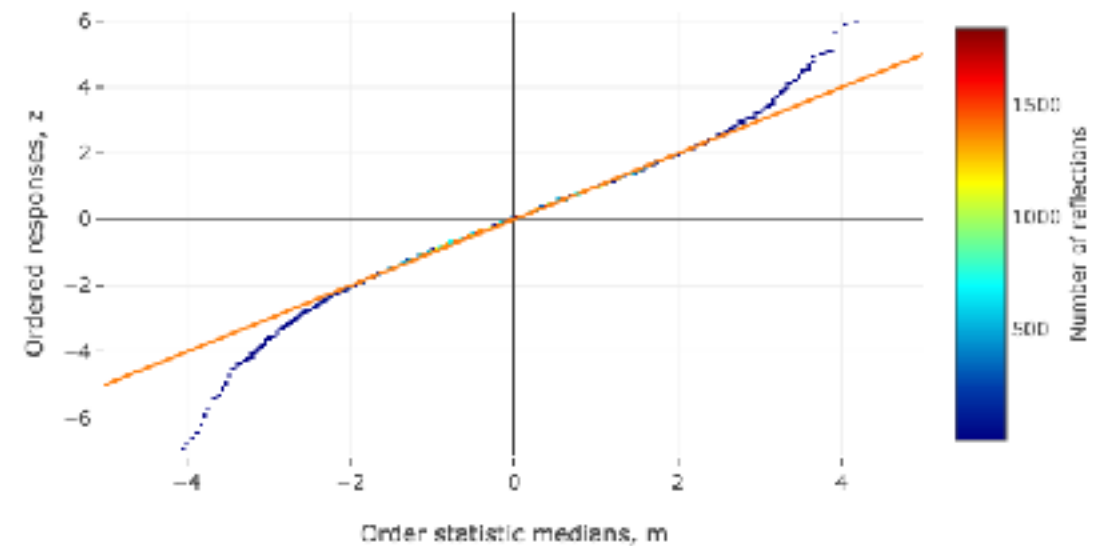


INTENSITY ERROR DISTRIBUTION

Normal deviations with error model applied



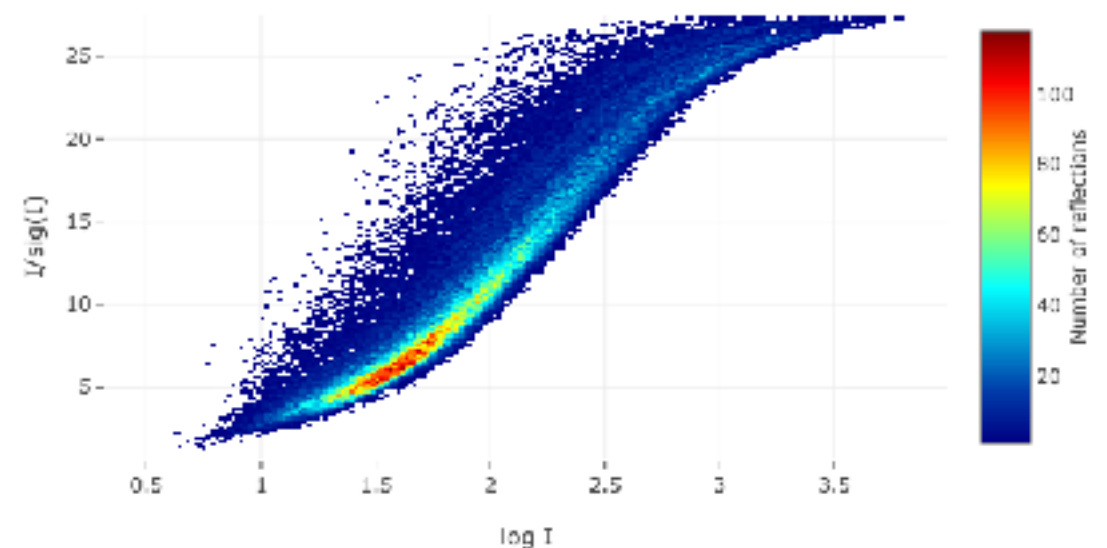
Normal probability plot with error model applied



$$I / \sigma_{asymptotic} = 1 / (a \times b)$$

$a \times b$ is measure of fraction of systematic error (2% very good)

$I/\sigma(I)$ vs I



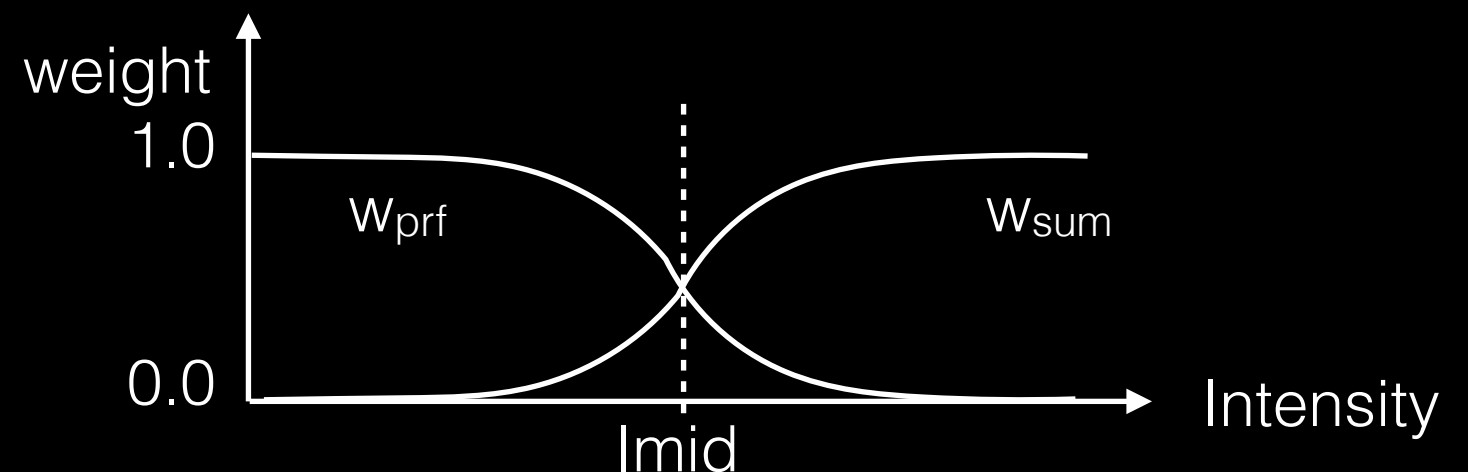
COMBINING SUMMATION/PROFILE INTENSITIES.

Integration programs give summation and profile fitted intensities. Profile intensity estimates usually best - but summation estimates can be best for most intense observations.

Try to find best combination - smooth cross over between profile and summation, optimise crossover point.

$$I_{scale} = w_{prf} I_{prf} + w_{sum} I_{sum}$$

$$w_{sum} = 1 - w_{prf}$$



Score based on R_{meas}

Performing profile/summation intensity optimisation.

Combination	CC1/2	Rmeas
prf only	0.99959	0.21789
sum only	0.99933	0.33332
Imid = 20.02	0.99954	0.22257
Imid = 2003.81	0.99959	0.21789
Imid = 200.33	0.9996	0.21773
Imid = 20.04	0.99955	0.22256
Imid = 2.0	0.99945	0.25957

Combined intensities with Imid = 200.380823531 determined to be best for scaling.

RECAP: DATA REDUCTION STEP 2

Determine corrections to make the data most internally consistent. Reject outliers, improve sigmas.

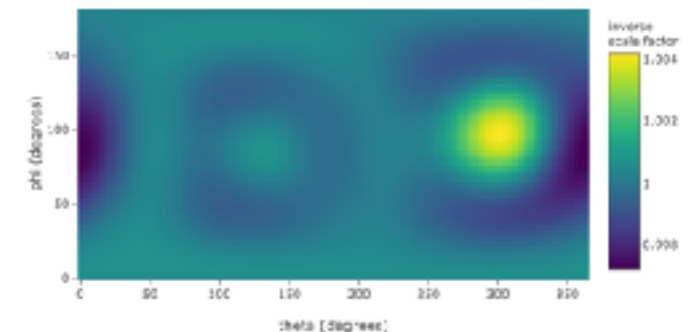
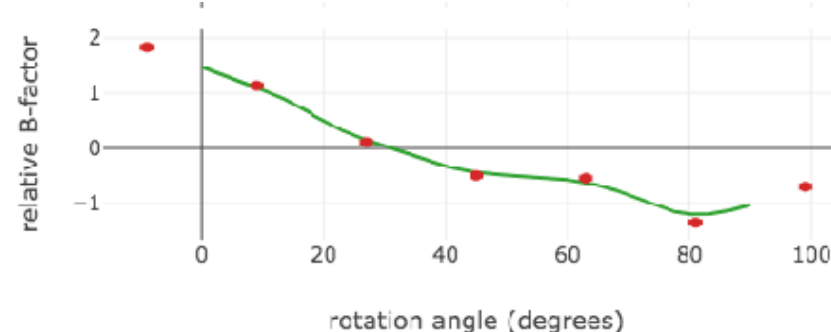
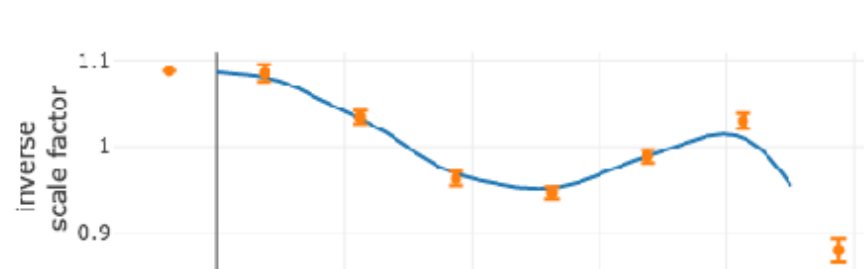
Integrated reflections



Integrated reflections, in correct point group



Scaled (corrected) intensities and sigmas



ASSESSING SCALING RESULTS

Resolution dependent plots

Summary statistics table

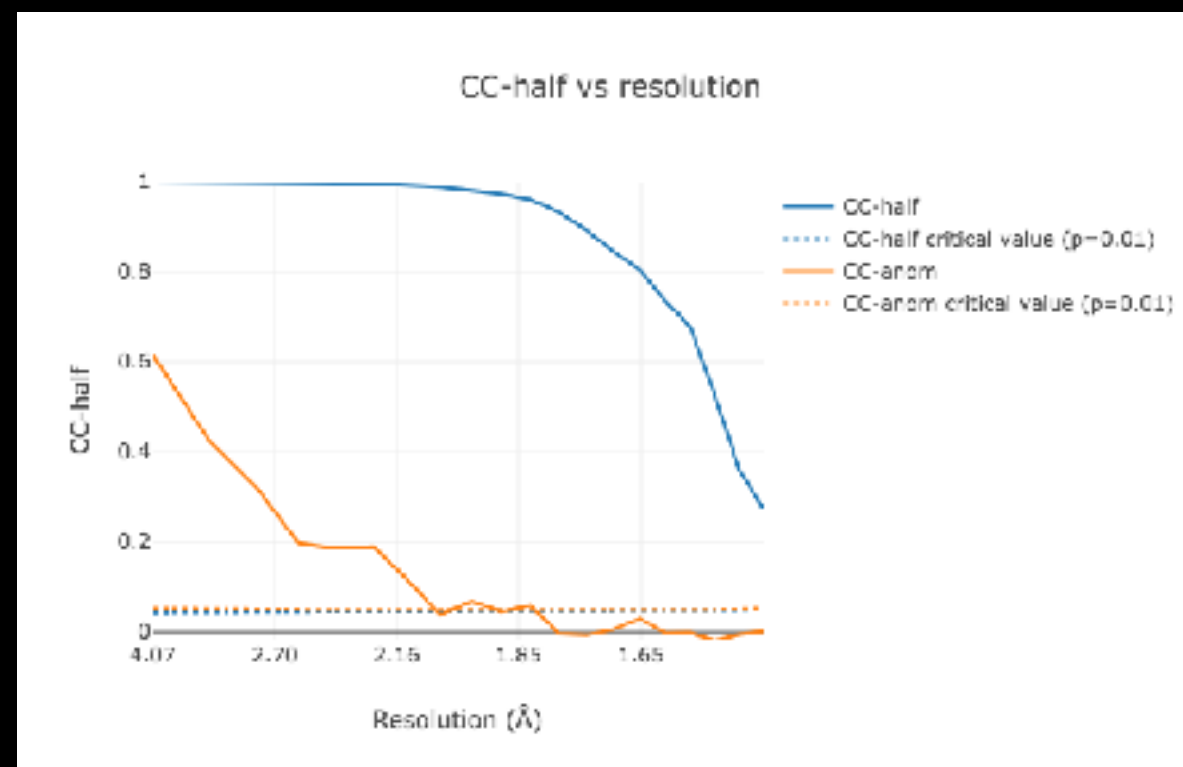
Overall merging statistics (non-anomalous)

Resolution: 79.97 - 1.59
Observations: 3715748
Unique reflections: 50690
Redundancy: 73.1
Completeness: 97.56%
Mean intensity: 4.6
Mean I/sigma(I): 12.1
R-merge: 0.217
R-meas: 0.216
R_pim: 0.025

Statistics by resolution bin:

d_max	d_min	N_obs	N_uniq	mult.	%comp	<I>	<I/sig>	r_mrg	r_meas	r_pim	cc1/2	cc_anom
89.99	4.97	205574	2989	71.23	100.00	28.4	69.1	0.075	0.076	0.009	1.000+	0.617+
4.97	3.23	195440	2785	72.52	99.99	23.9	49.9	0.098	0.099	0.011	0.999+	0.426+
3.23	2.82	285423	2650	77.52	99.74	18.6	31.3	0.156	0.157	0.018	0.999+	0.322+
2.82	2.56	167915	2618	75.68	99.62	6.2	20.8	0.217	0.218	0.025	0.998+	0.197+
2.56	2.38	195432	2611	75.23	99.35	3.9	14.9	0.291	0.293	0.033	0.996+	0.185+
2.38	2.24	289312	2595	77.16	99.24	3.8	12.9	0.343	0.345	0.039	0.995+	0.198+
2.24	2.13	289978	2565	78.35	99.83	2.2	9.2	0.431	0.434	0.049	0.992+	0.216+
2.13	2.04	189980	2562	73.76	99.77	1.6	5.9	0.525	0.529	0.068	0.988+	0.098
2.04	1.96	199713	2552	74.73	98.65	1.1	5.1	0.664	0.668	0.076	0.988+	0.068+
1.96	1.89	199254	2543	74.83	98.69	0.8	3.8	0.836	0.841	0.096	0.973+	0.045
1.89	1.83	195822	2538	77.55	98.41	0.6	2.8	1.084	1.091	0.121	0.968+	0.059+
1.83	1.78	192136	2584	76.73	98.20	0.4	2.2	1.321	1.330	0.149	0.933+	-0.003
1.78	1.73	189983	2529	74.73	97.87	0.3	1.5	1.735	1.747	0.199	0.892+	-0.007
1.73	1.69	183456	2505	73.23	97.74	0.2	1.2	2.121	2.125	0.244	0.845+	0.006
1.69	1.65	165845	2474	75.12	97.52	0.2	1.0	2.545	2.549	0.291	0.805+	0.030
1.65	1.62	194388	2516	77.23	97.68	0.1	0.8	3.055	3.075	0.343	0.735+	0.002
1.62	1.58	187931	2483	75.68	97.19	0.1	0.6	3.672	3.696	0.417	0.678+	-0.002
1.58	1.55	165350	2468	67.68	95.75	0.1	0.5	4.371	4.493	0.527	0.625+	-0.018
1.55	1.53	137840	2369	68.97	93.21	0.1	0.3	5.165	5.210	0.665	0.552+	-0.006
1.53	1.50	189836	2229	65.14	89.86	0.1	0.3	5.968	6.028	0.811	0.275+	0.002
79.97	1.50	3715748	50690	73.14	97.56	4.5	12.1	0.217	0.216	0.025	1.000+	0.376+

Normal and anomalous CC_{1/2}

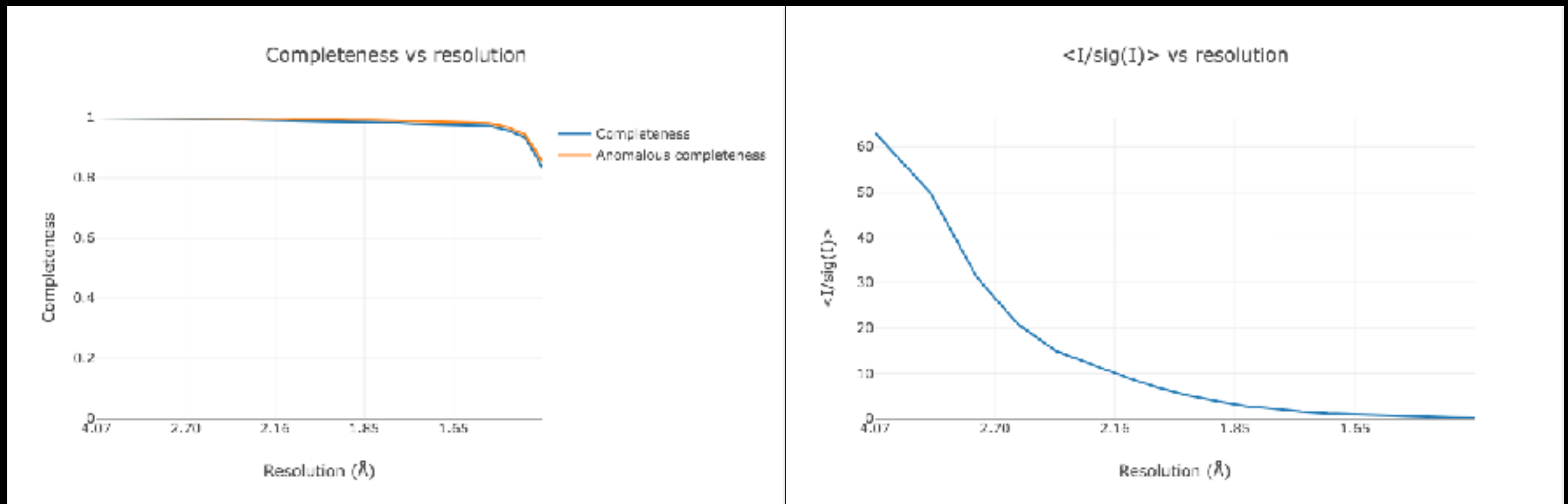


ASSESSING SCALING RESULTS

Resolution dependent plots

Completeness, multiplicity etc

Average $I/\sigma(I)$



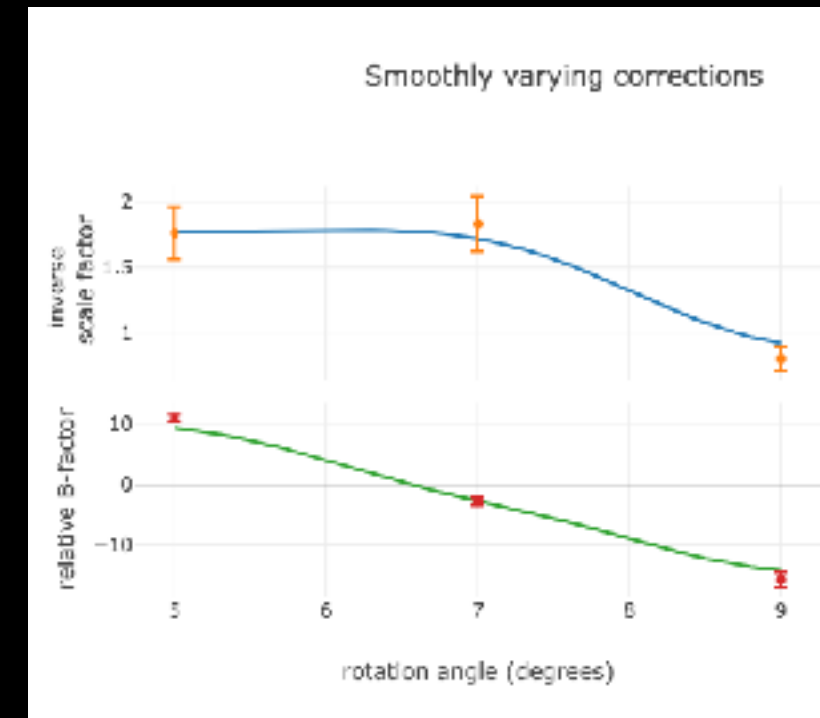
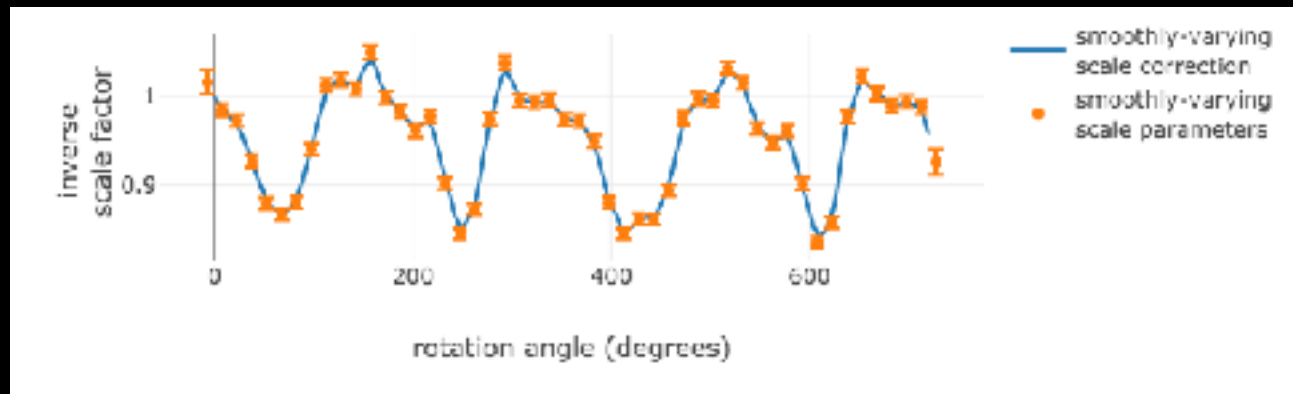
RESOLUTION LIMITS OF THE DATA

- $CC_{1/2}$ is best metric [1]. Clear interpretation of values & known statistical properties. Not dependent on $\sigma(I)$. Common value for cutoff is 0.3 or 0.5.
- R-factors. Shouldn't be used to determine resolution cutoffs. Historically used, so useful to compare to previous practises. Rise towards infinity as signal decreases. R_{pim} depends on multiplicity. Useful indicators.
- $\langle I/\sigma(I) \rangle$. Shouldn't be used to determine resolution cutoffs. Relies on $\sigma(I)$ which may be inaccurate. Useful indicator of quality of low resolution data.

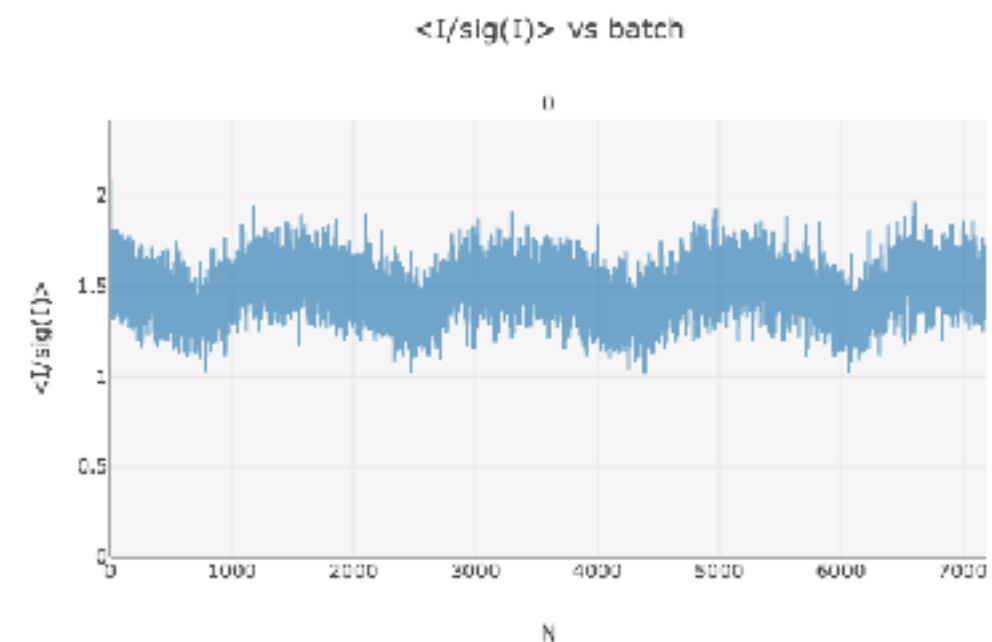
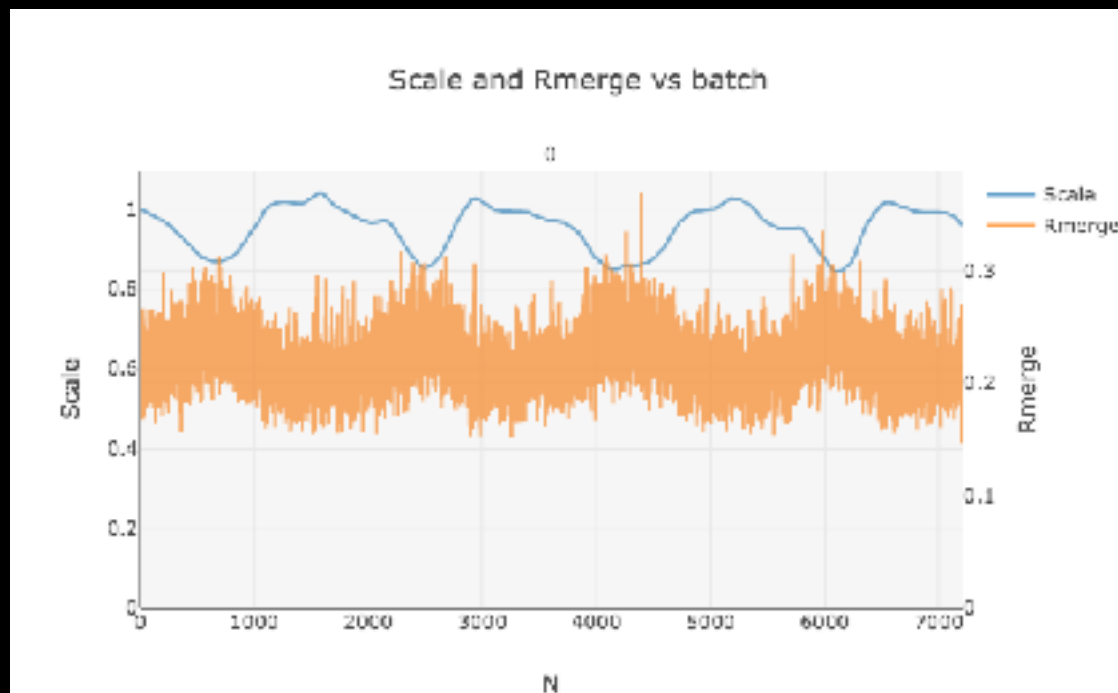
[1] Assessing and maximising data quality in macromolecular crystallography, Karplus & Diederichs, 2015.

PLOTS AS A FUNCTION OF IMAGE NUMBER

Individual dataset

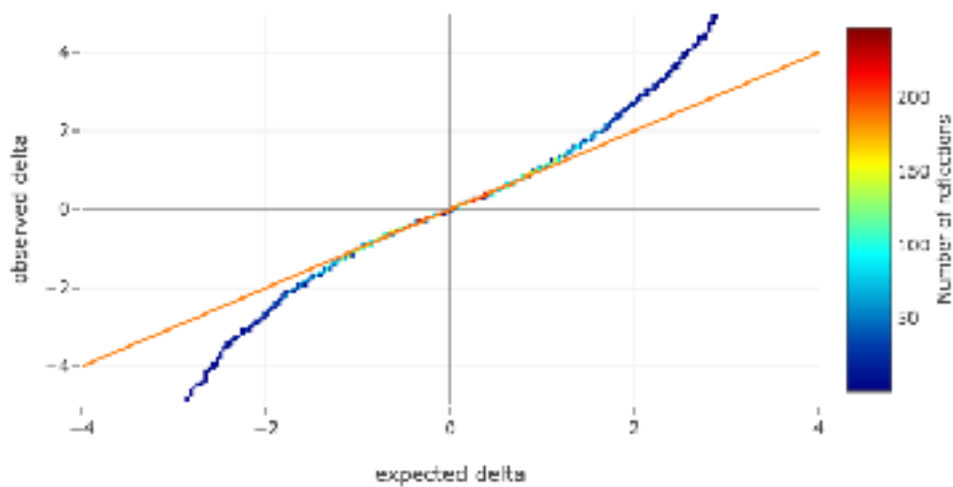


Individual/all datasets

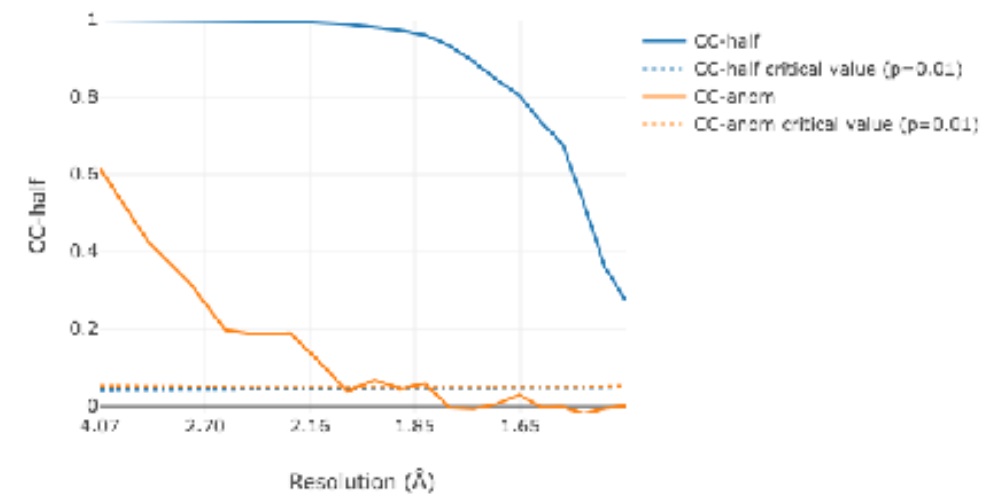


ANOMALOUS PLOTS

Normal probability plot of anomalous differences ($d > 2.13$)

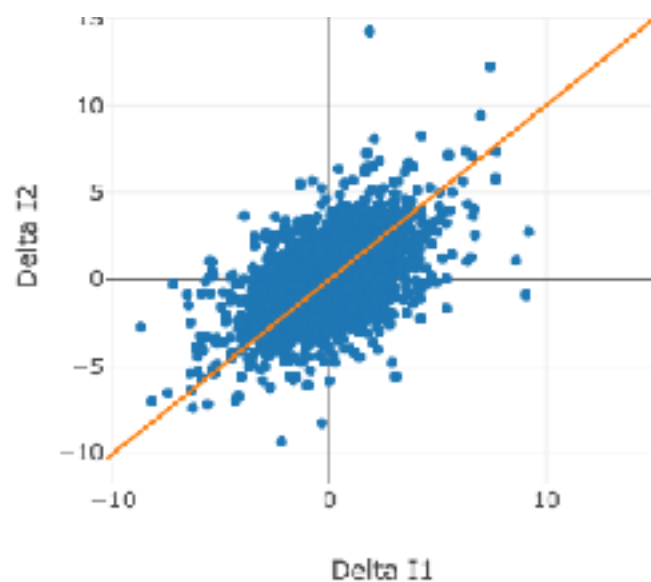


CC-half vs resolution

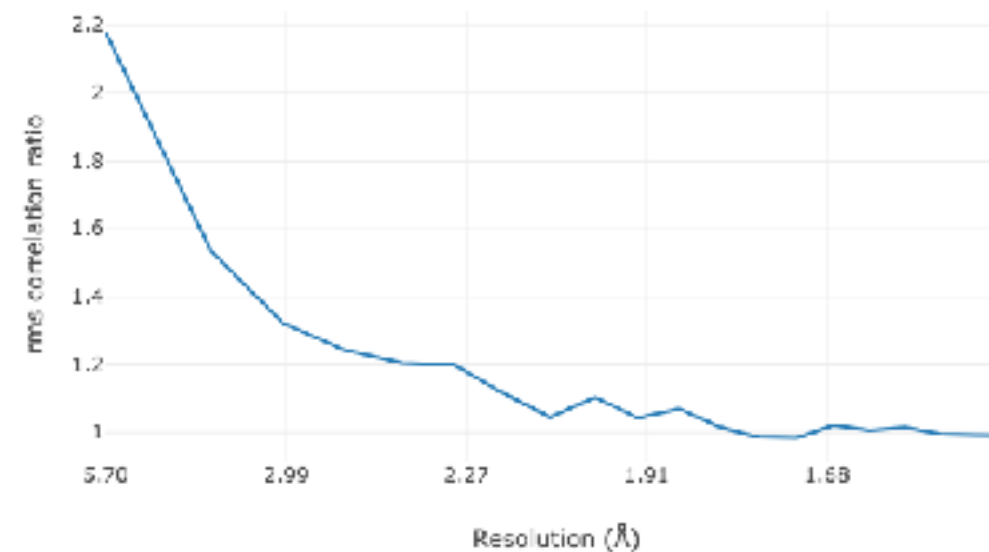


$$\delta_{\text{anom}} = (I^+ - I^-) / (\sigma^2(I^+) + \sigma^2(I^-))^{\frac{1}{2}}$$

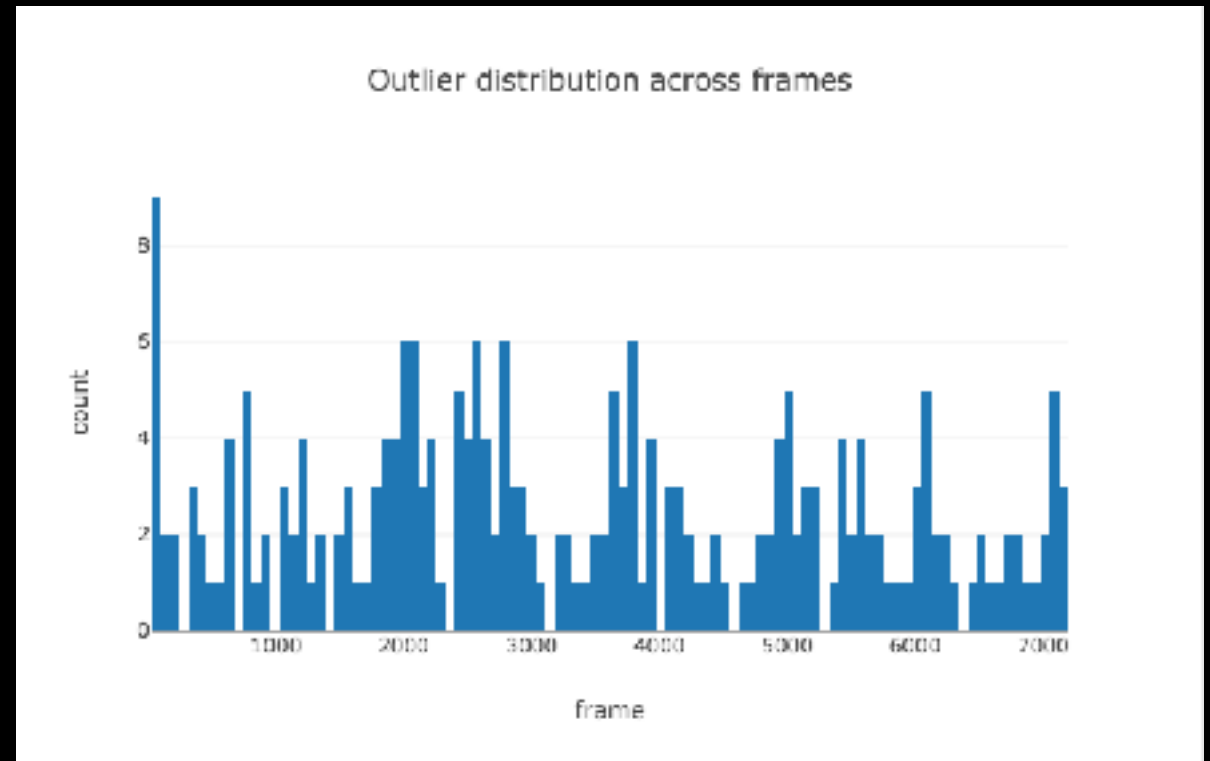
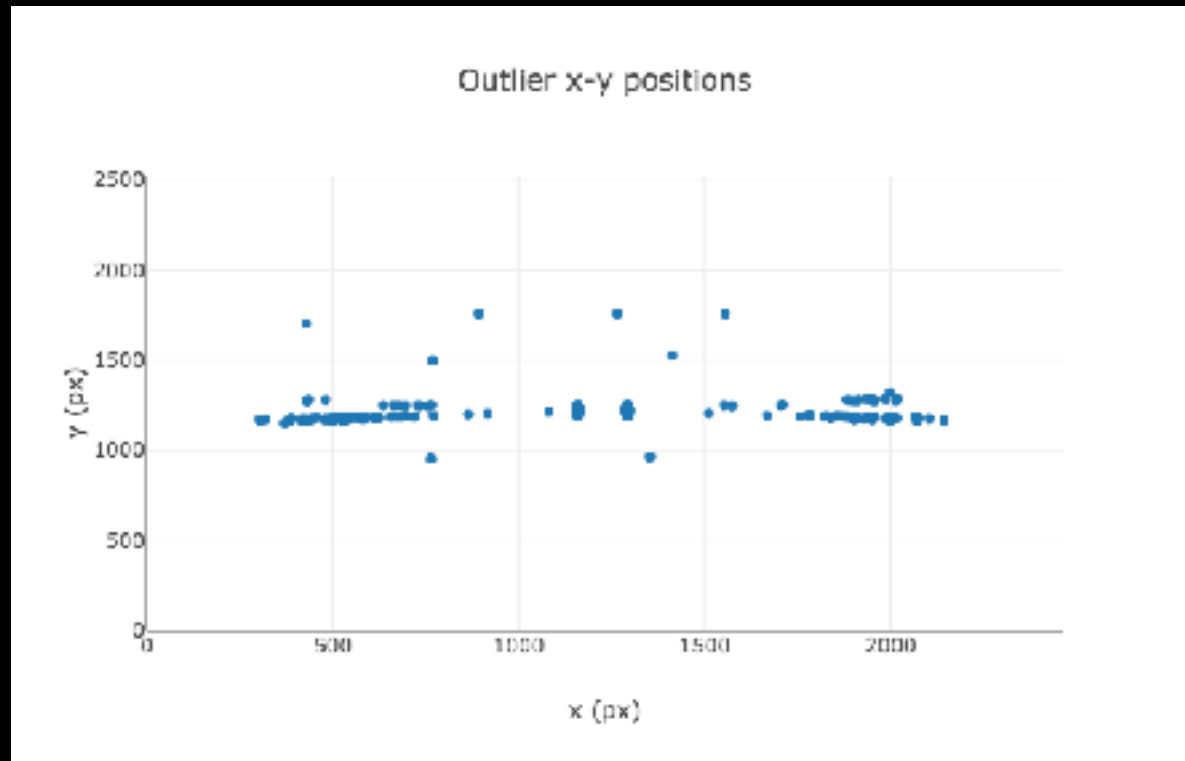
Correlation of half-set differences ($d > 2.13$)



Anomalous R.M.S. correlation ratio (acentric reflections)



OUTLIER PLOTS



Were bad areas masked, low angle well integrated? Are there bad frames?

ASSESSING SCALING RESULTS IN DIALS.SCALE

- Is RMSD_I close to, or lower than, 1.0?
- Do the error model parameters look reasonable? (a ~ 1.0, ~0.015 < b < ~0.06?)
- How many outliers were rejected, too many?
- Were profile/summation intensities combined?

```
Performing a round of error model refinement.

Error model details:
Type: basic
Current parameters: a = 0.91327, b = 0.03893
Error model formula:  $\sigma^2 = a^2(\sigma^2 + (bI)^2)$ 
estimated 1/sigma asymptotic limit: 28.125
```

```
Performing a round of scaling with an LBFGS minimizer.
```

```
Time taken for refinement 1.23618006706
```

```
Refinement steps:
```

Step	Nref	RMSD_I (a.u)
0	21207	1.0333
1	21207	1.0302
2	21207	1.0245
3	21207	1.0175
4	21207	1.0155
5	21207	1.0133
6	21207	1.0129
7	21207	1.0127
8	21207	1.0123
9	21207	1.0121
10	21207	1.0119
11	21207	1.0118

```
Performing profile/summation intensity optimisation.
```

Combination	CC1/2	Rmeas
prf only	0.99959	0.21789
sum only	0.99933	0.33332
Imid = 20.02	0.99956	0.22257
Imid = 2003.31	0.99959	0.21789
Imid = 200.33	0.9996	0.21773
Imid = 20.04	0.99956	0.22256
Imid = 2.0	0.99946	0.25967

```
Combined intensities with Imid = 200.330803531 determined to be best for scaling.
```

AFTER SCALING; MERGING AND TRUNCATING

Merge symmetry equivalent reflections to get the best estimate (optionally keeping anomalous pairs separate).

Most structure determination programs require structure factor amplitudes:

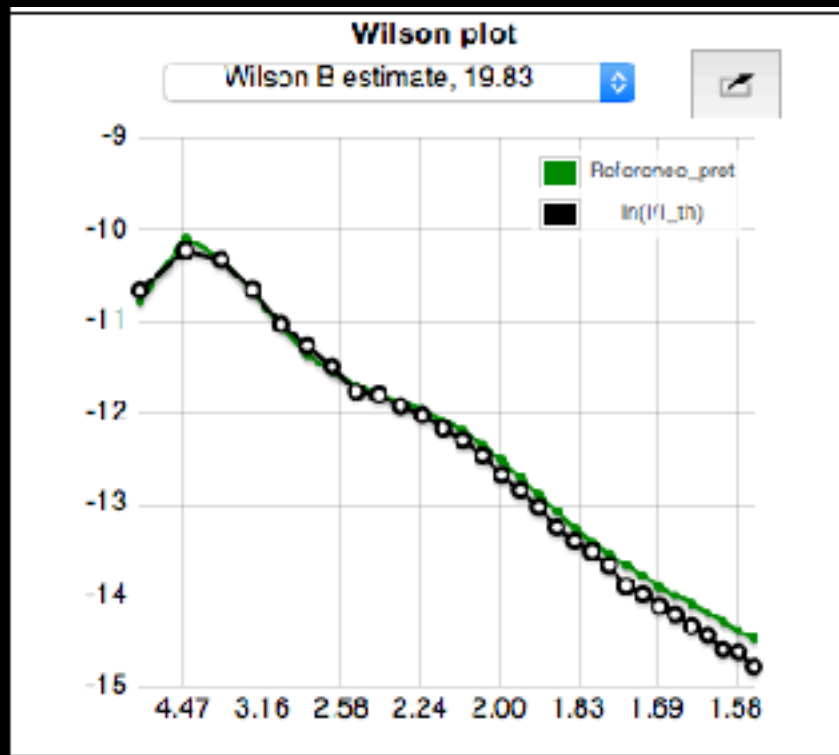
$$|F| = \sqrt{I}$$

Weak reflections can have negative intensities due to background subtraction!

French Wilson algorithm (1978), uses Bayesian statistics to calculate positive structure factors for all intensities (based on I and $\sigma(I)$).

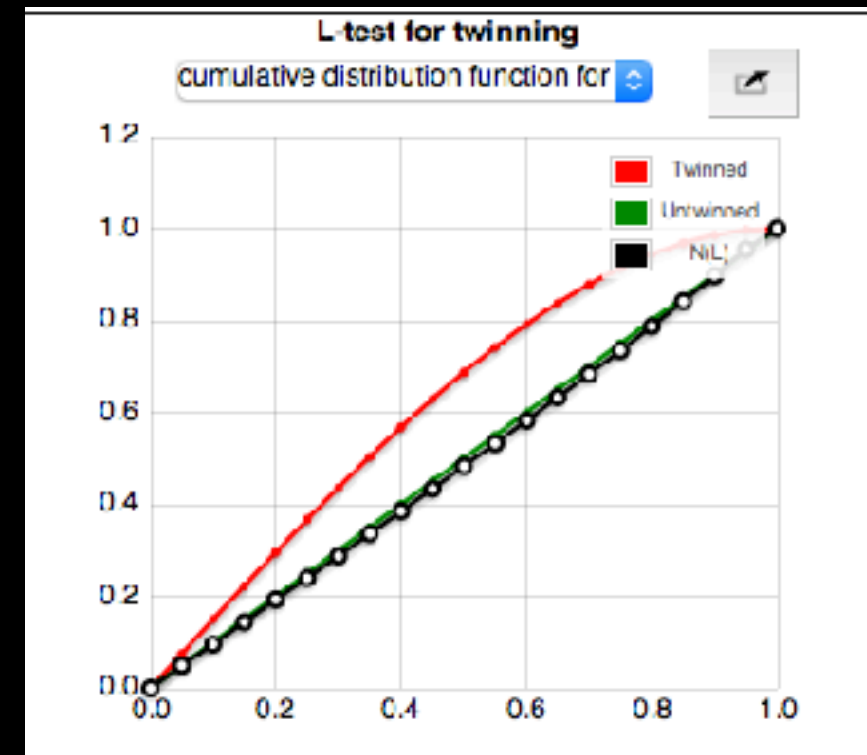
Program can use: ctruncate (ccp4), dials.merge

INTENSITY STATISTICS USING CTRUNCATE



$$\langle I \rangle = C \exp\left(-\frac{B}{2} \frac{1}{d^2}\right)$$

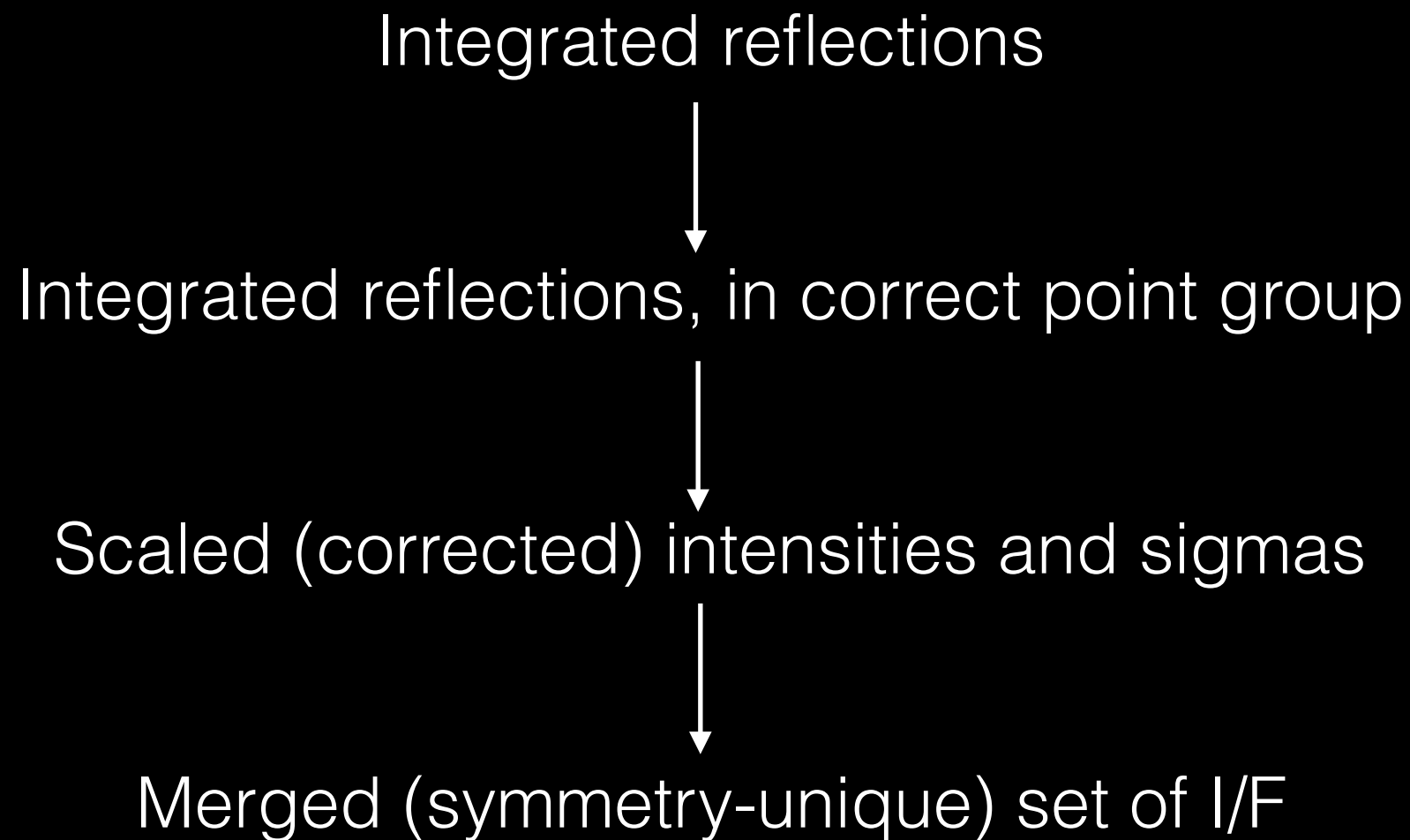
Average intensity falls with resolution due to atomic motions (B-factors).



Yeates-Padilla test (L-test)
Different expected cumulative intensity distribution depending on twin fraction.

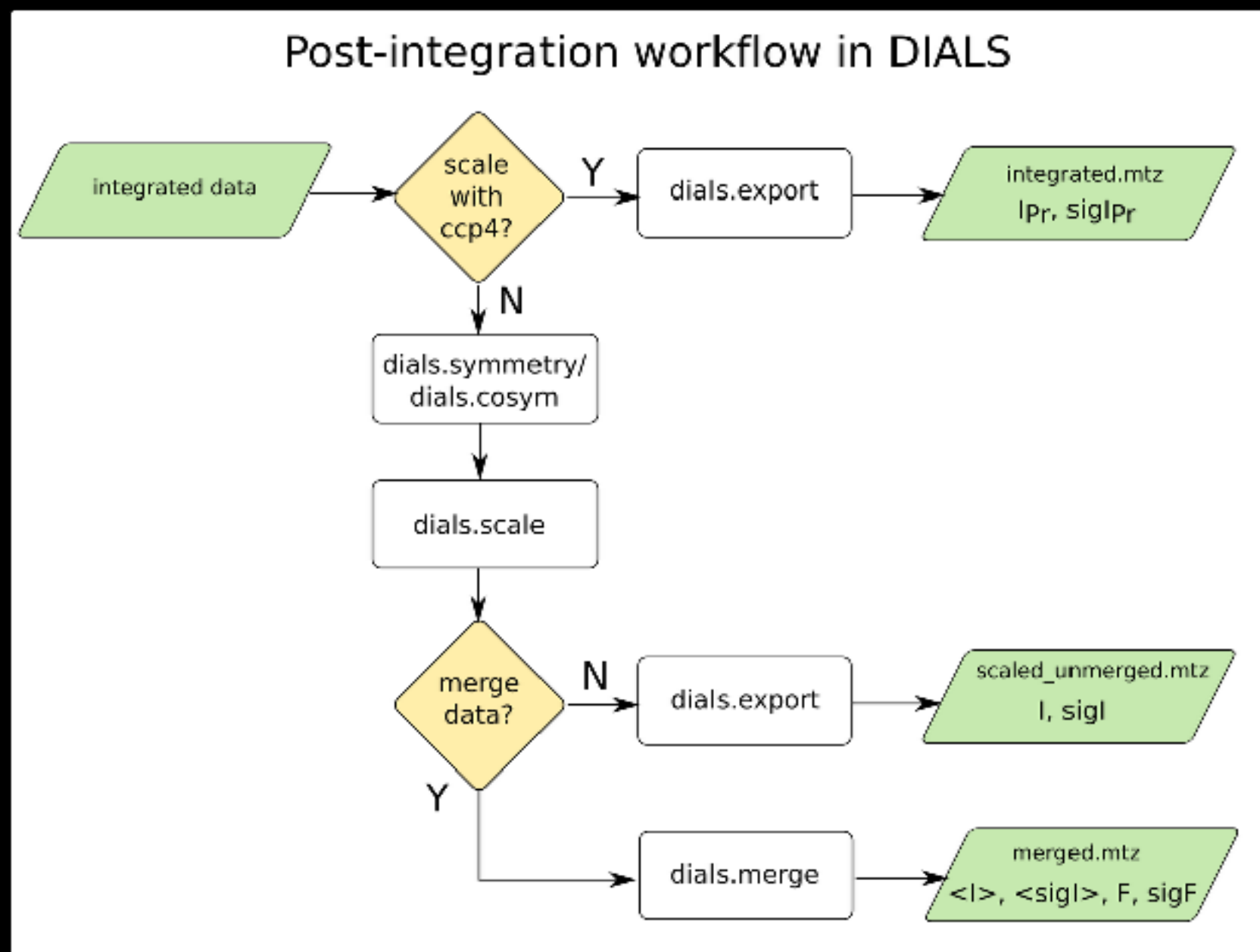
RECAP: DATA REDUCTION STEP3

Merge equivalent reflections, convert to F_s^*



*Some move away from converting I to F - e.g. Phaser can use I

DIALS DATA REDUCTION WORKFLOW



SCALING WORKFLOWS IN DIALS

Standard single sweep

dials.symmetry integrated.refl integrated.expt

dials.scale symmetrized.refl symmetrized.expt

dials.scale scaled.refl scaled.expt d_min=2.1

N.B. Reuse output scaling files for rescaling!

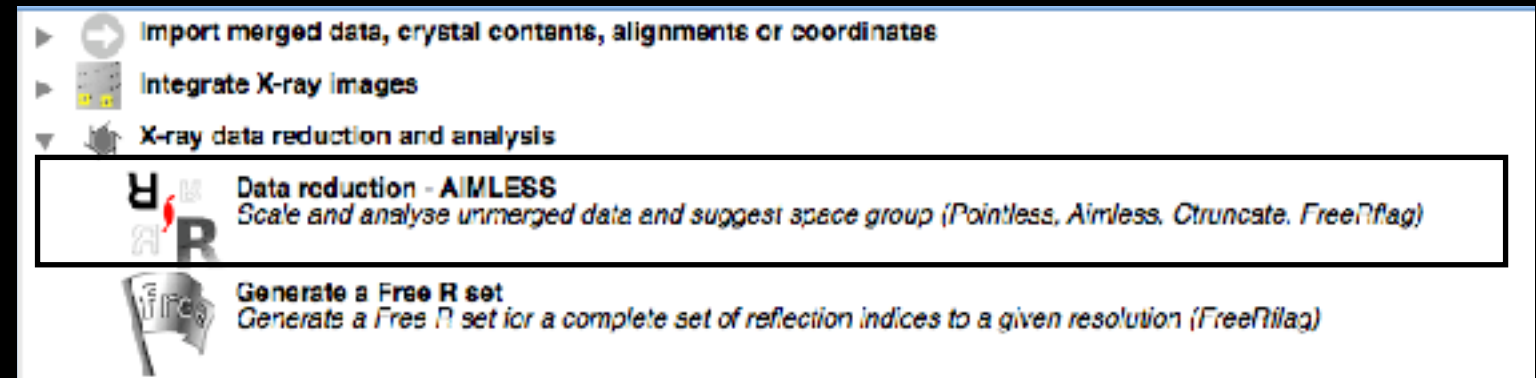
dials.merge scaled.refl scaled.expt

Routine options for single sweep

- d_min = 2.1 Set resolution cutoff
- exclude_images="100:120" Exclude a set of bad images

DATA REDUCTION WORKFLOWS IN CCP4I2

Integrated MTZ (from DIALS, Mosflm, xia2(XDS, DIALS))



Pointless → Aimless → Ctruncate

Add multiple datasets

Set resolution range

The screenshot shows the 'Job 56: Data reduction - AIMLESS' configuration window. The 'Input Data' tab is selected. The 'Job title' is 'Data reduction'. The 'Use data from job' dropdown is set to 'No'. The 'Show list' button is highlighted. The 'Crystal name' and 'dataset name' fields are empty. The 'Resolution range (Å)' is set to '1.22 Å'. The 'Options for symmetry determination' dropdown is set to 'Determine Locus group and space group'. The 'Optional input data' section includes a checkbox for 'use explicit resolution range in symmetry determination as well as in scaling', a checkbox for 'use reference data in analysis against Batch after scaling', and a dropdown for 'Reference data on' set to 'Reflection list'. The 'Reflections' dropdown is set to 'is not used'. The 'Free R set' dropdown is set to 'is not used'. The 'Fraction of reflections in generated free R set' is set to 'Default fraction is 0.05'.

SCALING WORKFLOWS IN DIALS

Standard multi sweep

dials.symmetry/dials.cosym integrated_1.refl integrated_1.expt

dials.scale symmetrized.refl symmetrized.expt

dials.scale scaled.refl scaled.expt d_min=2.1

dials.merge scaled.refl scaled.expt

SCALING WORKFLOWS IN DIALS

Single sweep, controlling model (advanced)

```
dials.scale symmetrized.refl symmetrized.expt \  
  model=physical physical.absorption_correction=False
```

Extra options for controlling models

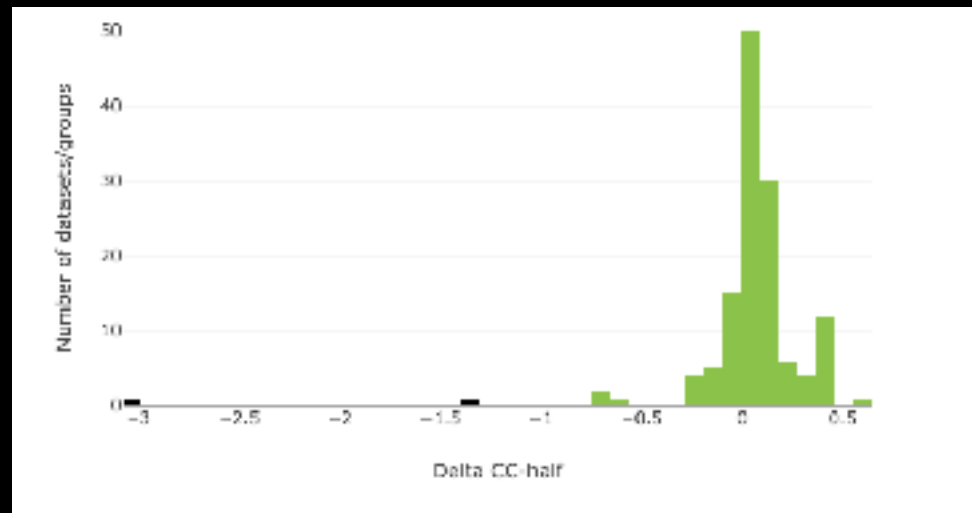
- `model=array, model=physical` Explicitly chose a model
- `model=physical physical.absorption_correction=False \
 physical.decay_correction=False` Disable a correction
- `model=physical physical.scale_interval=30.0` Choose parameter spacing

MULTI-CRYSTAL SCALING AND MERGING

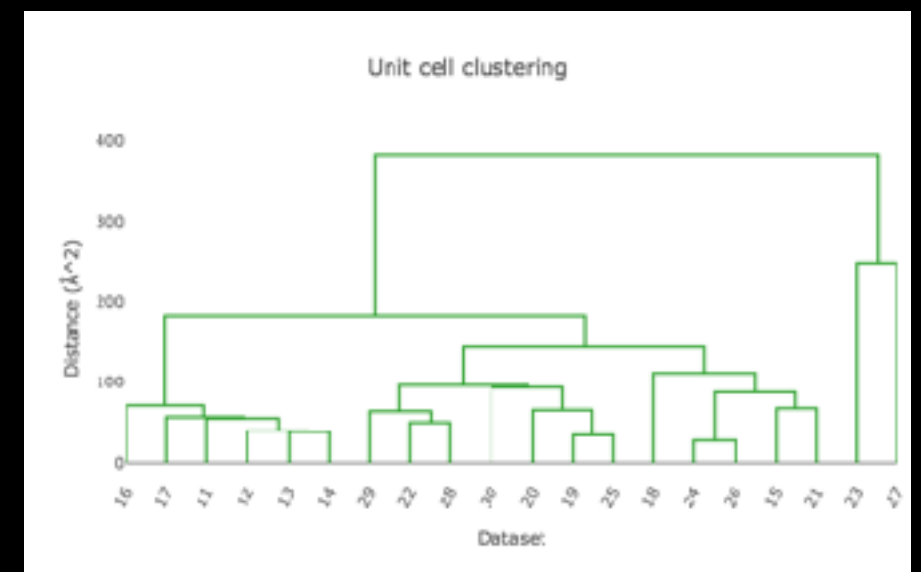
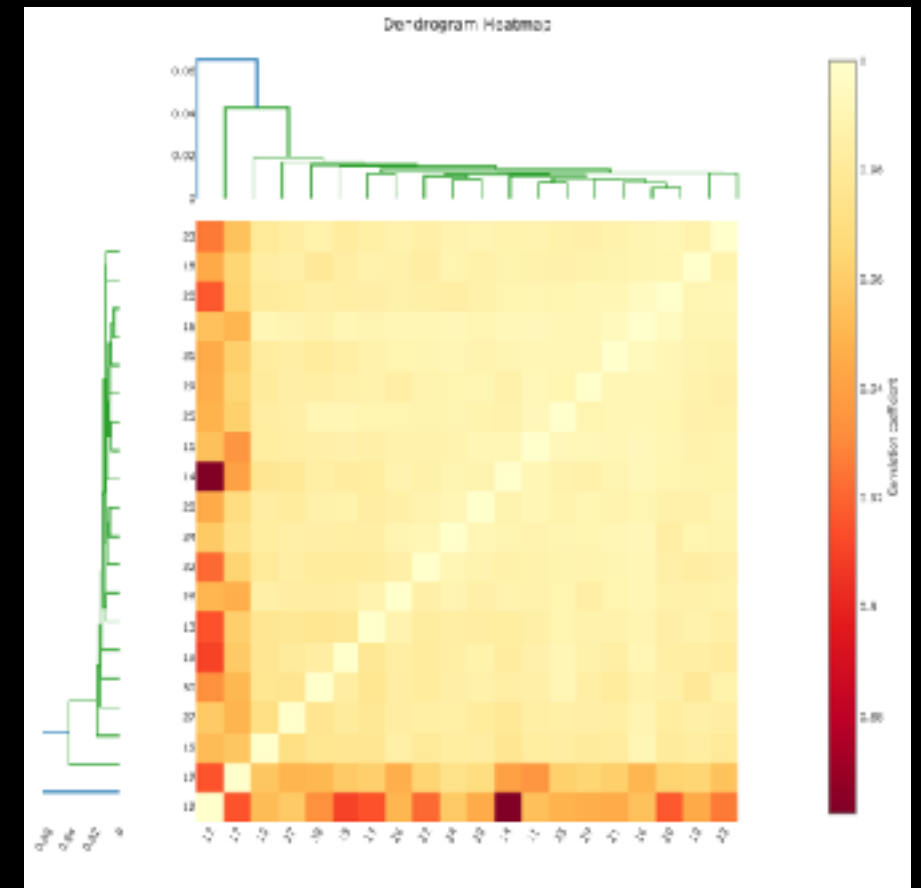
When scaling & merging multi-crystal data, need to decide which subset of the data to keep due to inhomogeneity.

Clustering/dataset rejection:

- unit cell clustering
- delta cc1/2 clustering



Tools available: xia2.multiplex, dials.cluster_unit_cell, dials.scale, dials.compute_delta_cchalf, XDS, ..



SUMMARY

- Data reduction process of scaling and merging integrated intensities.
- Scaling attempts to correct for experimental effects to make the data internally consistent.
- Critically evaluate the merging statistics and plots after scaling. Is the dataset complete, was radiation damage significant? Should the resolution be cut back? Weak data is not necessarily bad data.
- Are many crystals being merged - which data should be included?
- Merging statistics *indication* of data quality. Try to solve the structure: you might decide you need to reprocess.