Data reduction: scaling and merging data

James Beilsten-Edmands
Diamond Light Source, U.K.

ACKNOWLEDGEMENTS

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

How good are my data and what is the resolution?

Philip R. Evans* and Garib N. Murshudov

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail: pre@mrc-lmb.cam.ac.uk

Received 7 September 2012
Accepted 2 January 2013

Following integration of the observed diffraction spots the process of 'data reduction' initially aims to determine the point-group symmetry of the data and the likely space group. This can be performed with the program *POINTLESS*. The scaling program then puts all the measurements onto a common scale, averages measurements of symmetry-related reflections (using the symmetry determined previously) and produces many statistics that provide the first important measures of data quality. A new scaling program, *AIMLESS*, implements

CrossMark

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Breaking the indexing ambiguity in serial crystallography

Wolfgang Brehm and Kay Diederichs*

Department of Biology, Universität Konstanz, Box 647, 78457 Konstanz, Germany

Correspondence e-mail: kay.diederichs@uni-konstanz.de

Received 7 August 2013
Accepted 13 September 2013

In serial crystallography, a very incomplete partial data set is obtained from each diffraction experiment [a 'snapshot']. In some space groups, an indexing ambiguity exists which requires that the indexing mode of each snapshot needs to be established with respect to a reference data set. In the absence of such re-indexing information, crystallographers have thus far resorted to a straight merging of all snapshots, yielding a perfectly twinned data set of higher symmetry which is poorly suited for structure solution and refinement. Here, two algorithms have been designed for assembling complete data

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Scaling and assessment of data quality

Philip Evans

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail: pre@mrc-lmb.cam.ac.uk

Received 6 June 2005
Accepted 8 November 2005

The various physical factors affecting measured diffraction intensities are discussed, as are the scaling models which may be used to put the data on a consistent scale. After scaling, the intensities can be analysed to set the real resolution of the data set, to detect bad regions (*e.g.* bad images), to analyse radiation damage and to assess the overall quality of the data set. The significance of any anomalous signal may be assessed by probability and correlation analysis. The algorithms used by the *CCP*4 scaling program *SCALA* are described. A requirement for the scaling and merging of intensities is

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

XDS

Wolfgang Kabsch

Max-Planck-Institut für Medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany

Correspondence e-mail: wolfgang.kabsch@mpimf-heidelberg.mpg.de

Received 31 August 2009
Accepted 9 November 2009

The usage and control of recent modifications of the program package *XDS* for the processing of rotation images are described in the context of previous versions. New features include automatic determination of spot size and reflecting range and recognition and assignment of crystal symmetry. Moreover, the limitations of earlier package versions on the number of correction/scaling factors and the representation of pixel contents have been removed. Large program parts have been restructured for parallel processing so that the quality and completeness of collected data can be assessed soon after

A version of this paper will be published as a chapter in the new edition of Volume F of *International Tables for Crystallography*.

PROJECT SUPPORT

DIALS
Diffraction Integration for Advanced Light Sources

diamond

CCP4

BioStruct-X

wellcome

BERKELEY LAB
Lawrence Berkeley National Laboratory

UKRI
Science and Technology Facilities Council

EU FP7: #283570
NIH: GM095887 & GM102520

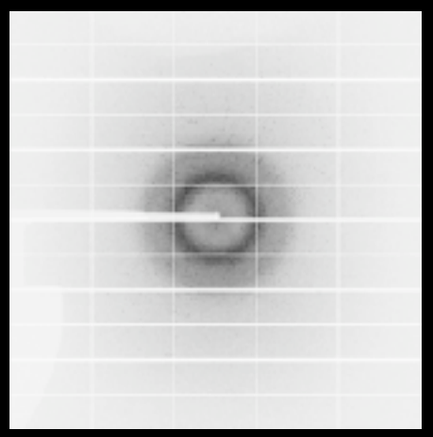
DATA REDUCTION

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

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

These intensities should then be proportional to F^2

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



$\{I_i^{h,k,l}\}$

data integration

$\{\mathbf{I}^{h,k,l}\}$

data reduction

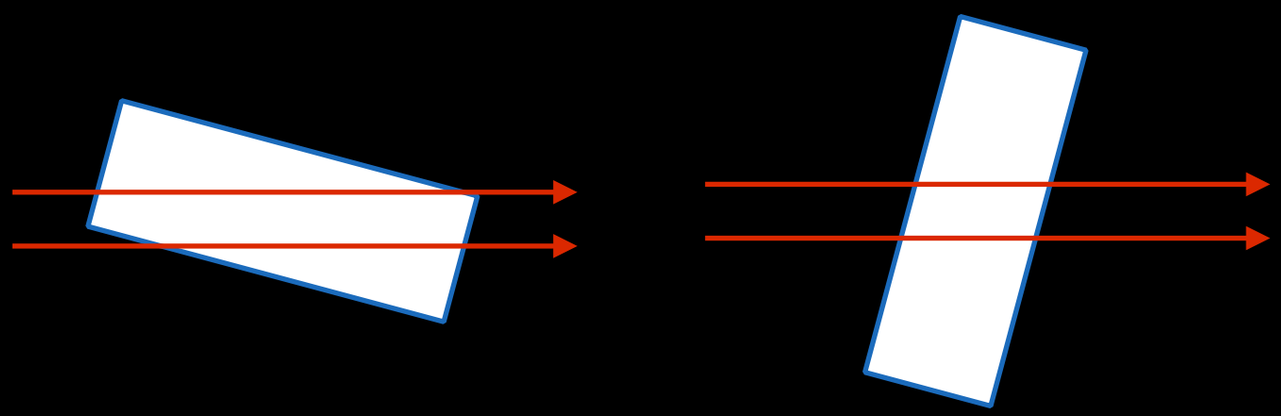
structure determination

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?

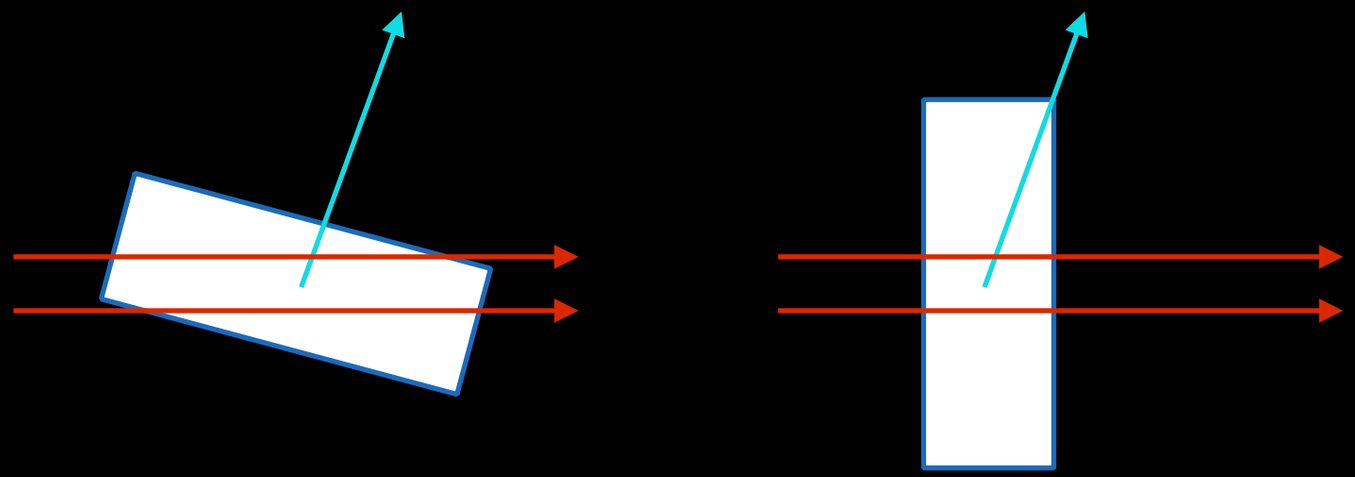


Beam effects - changes in beam intensity (slow), illuminated volume.

EXPERIMENTAL EFFECTS

Why do symmetry-equivalent observations have different intensities?

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



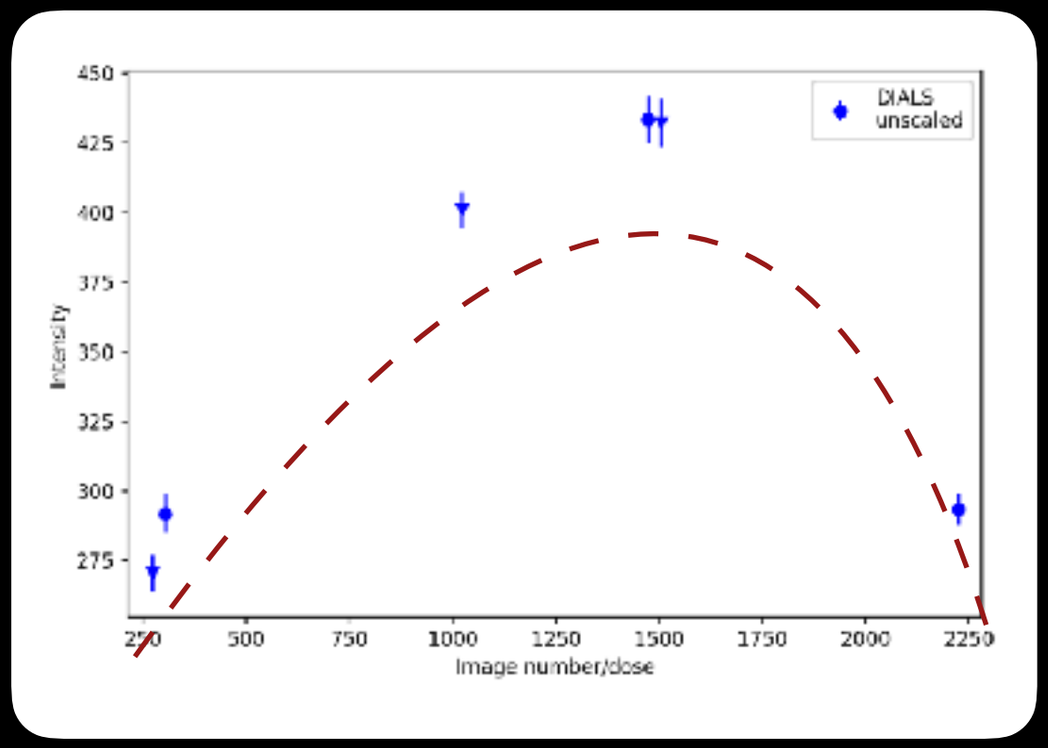
Detector effects - gain & calibration issues, edge effects (CCDs).

SCALING AND MERGING

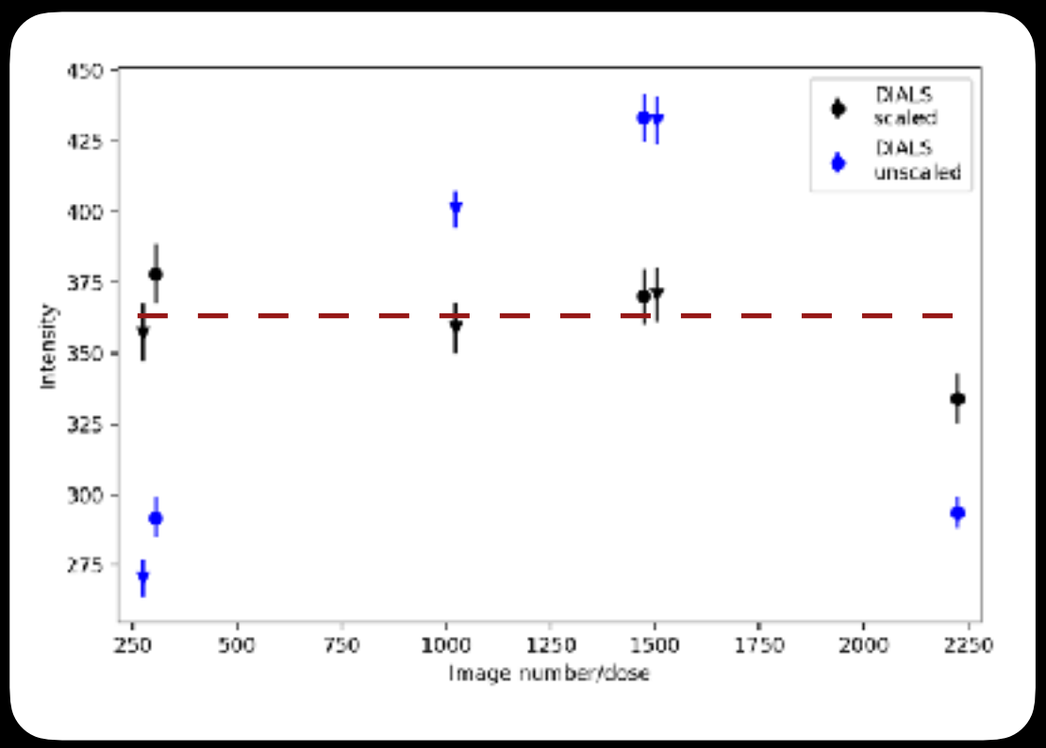
Use the redundancy within the dataset, *i.e.* repeated measurements of equivalent reflections, to determine a model to account for experimental effects.

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

Then merge (average) the intensities in each group.



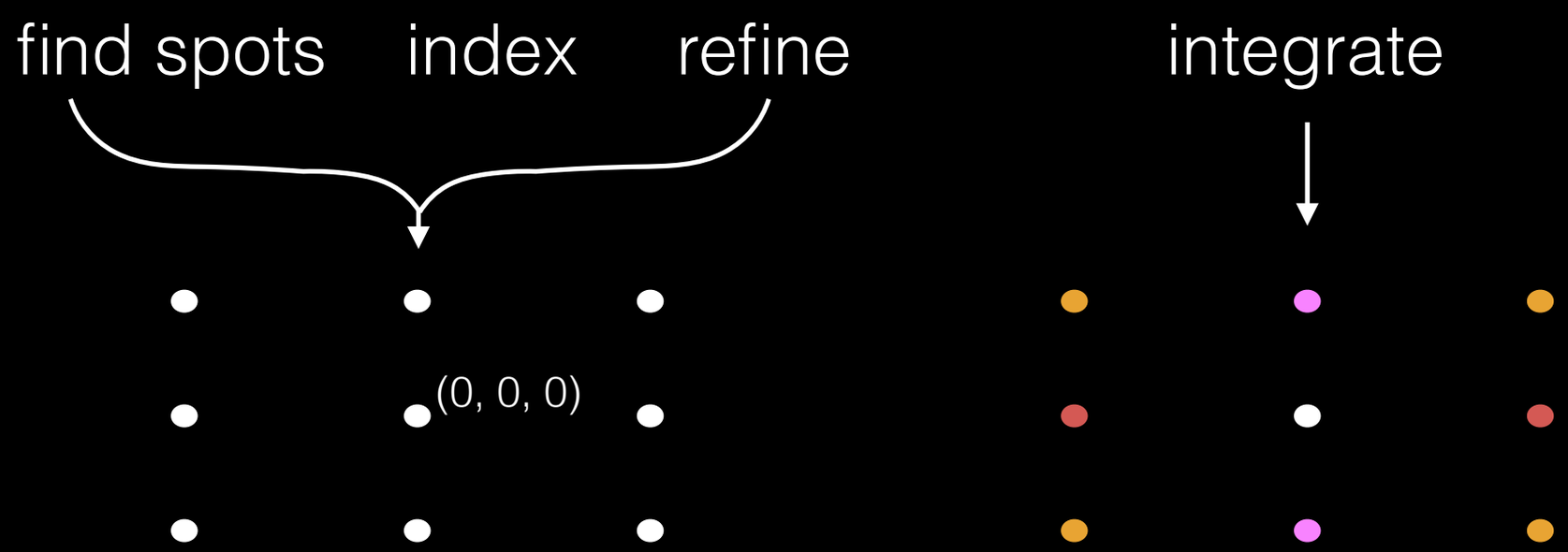
scale



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

PRECURSOR: SYMMETRY ANALYSIS

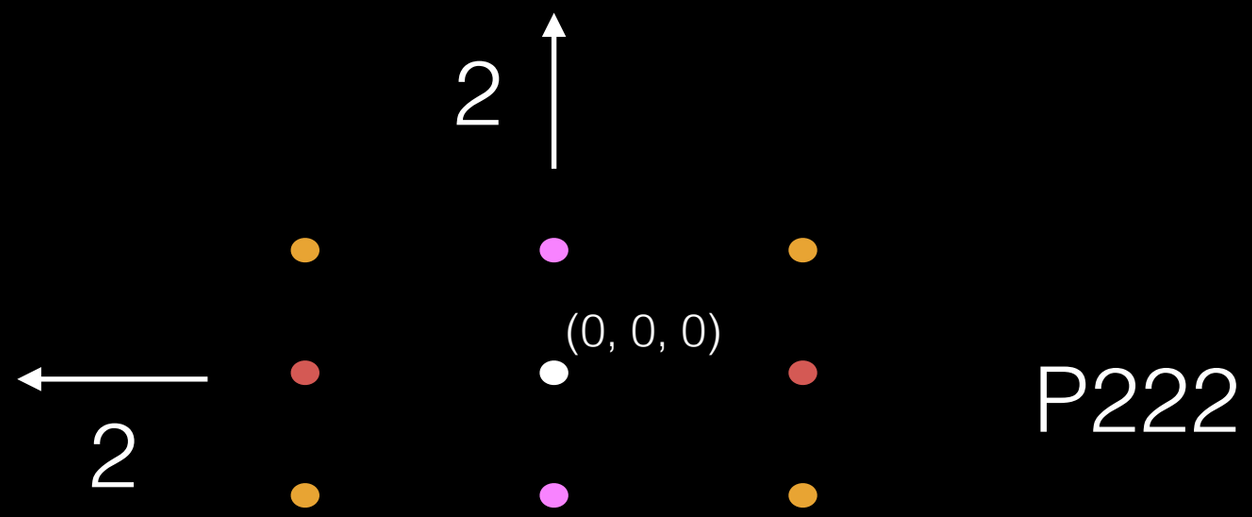
We need to know the symmetry of the diffraction pattern, so that observations are correctly grouped for scaling and merging.



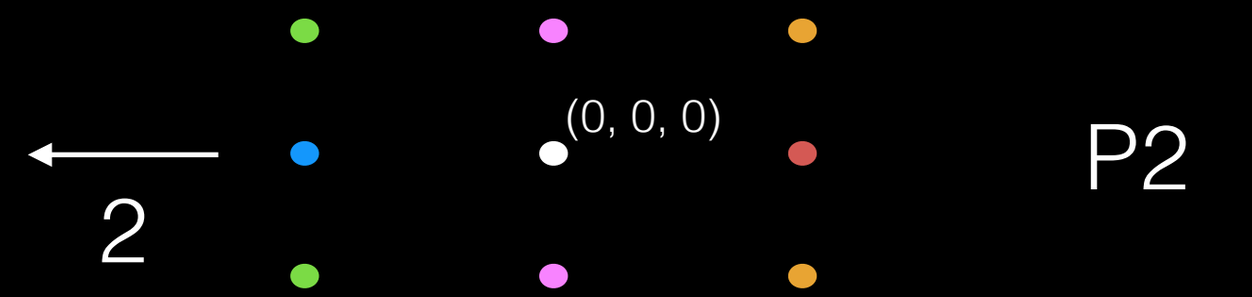
Before integration, only know symmetry of the lattice (spot positions), not the diffraction pattern (intensities). Lattice symmetry != diffraction pattern symmetry

PRECURSOR: SYMMETRY ANALYSIS

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



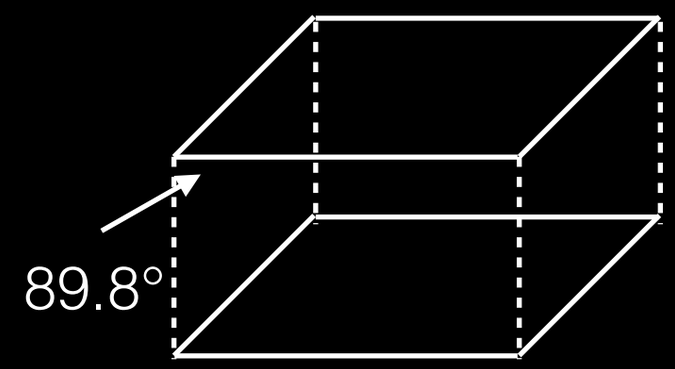
Can scale in lower symmetry point group, but then not using full equivalence of certain observations!



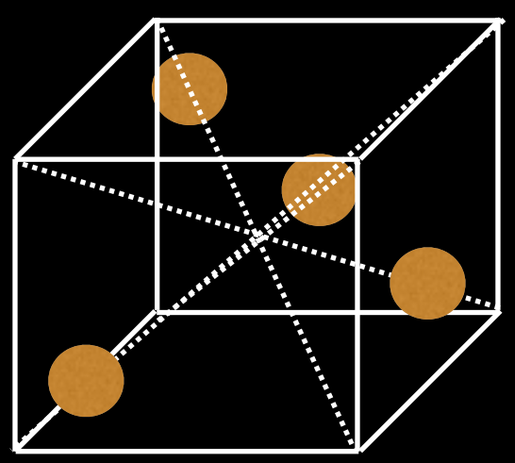
PRECURSOR: SYMMETRY ANALYSIS

Why is lattice symmetry != diffraction pattern symmetry?

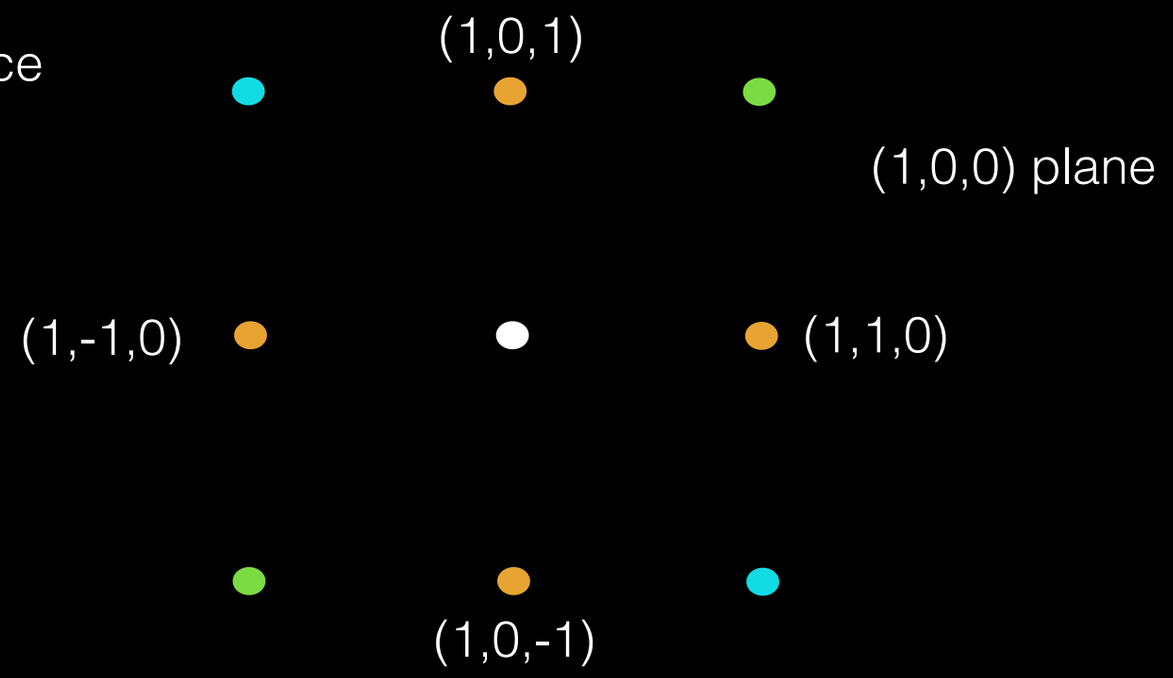
May have `accidental` symmetry of the lattice, e.g. P1 symmetry with angle close to 90° looks like P2.



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

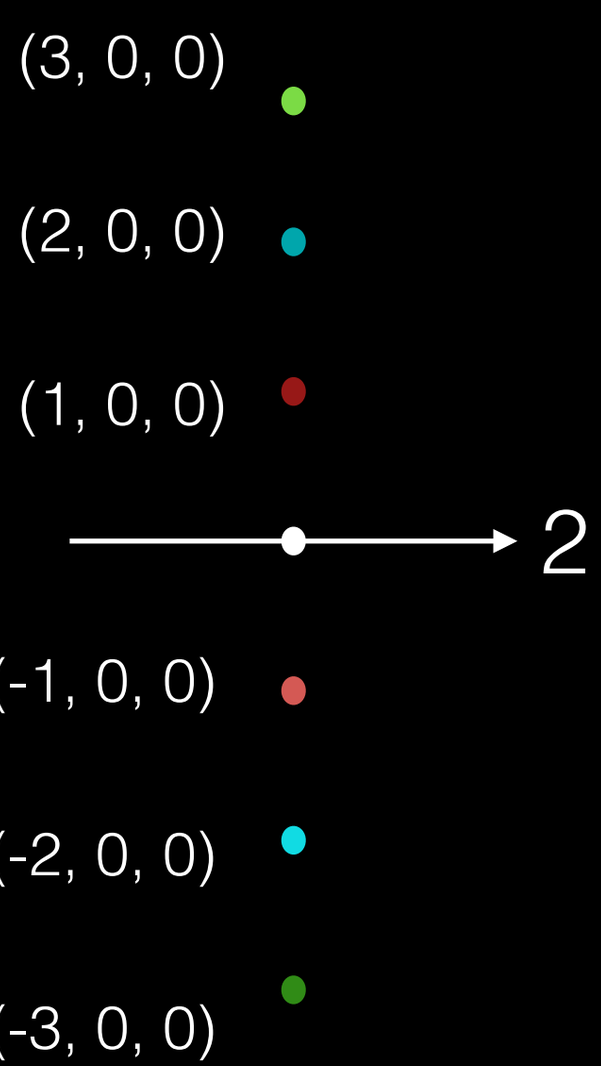


Point group 23, Lattice symmetry 432



Programs for symmetry assessment: *Pointless, dials.symmetry, XDS (CORRECT step)*

PRECURSOR: SYMMETRY ANALYSIS WITH *DIALS.SYMMETRY / POINTLESS / XDS*

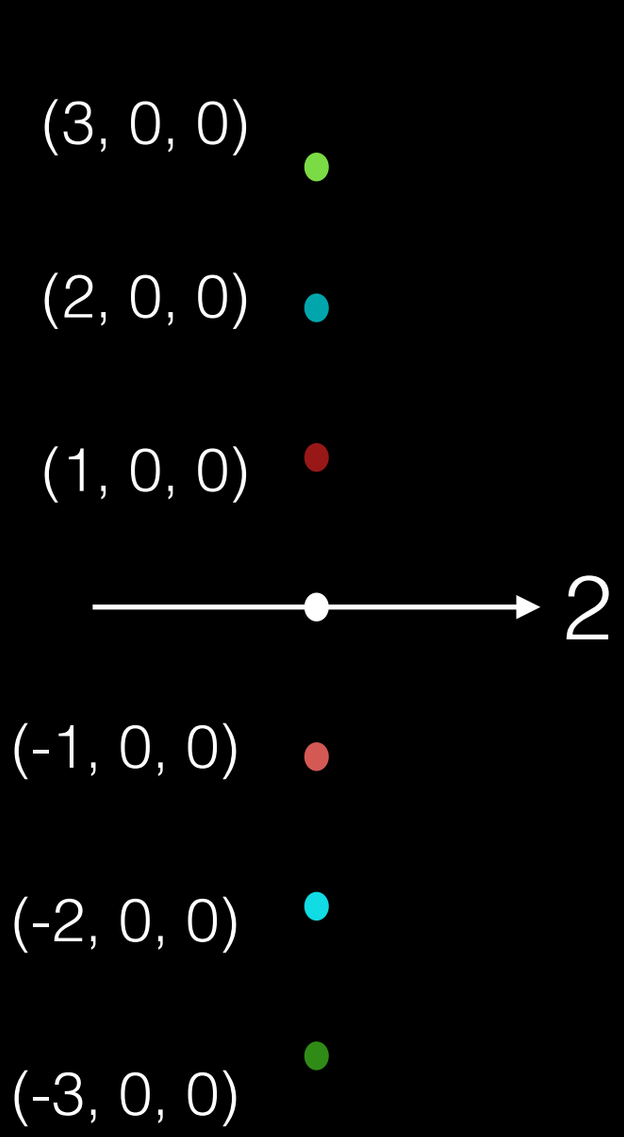


Score symmetry element based on correlation between intensities.

Cubic insulin

```
Scoring individual symmetry elements
-----------------------------------------------------
likelihood  Z-CC  CC    N              Operator
-----------------------------------------------------
0.943       9.90  0.99  905647   ***   1 |(0, 0, 0)
0.196       5.02  0.50  1694845        4 |(1, 1, 0)
0.196       5.02  0.50  1693679        4 |(1, 0, 1)
0.196       5.03  0.50  1694455        4 |(0, 1, 1)
0.931       9.67  0.97  1713055  ***   3 |(1, 0, 0)
0.939       9.80  0.98  1715415  ***   3 |(0, 1, 0)
0.939       9.81  0.98  1715530  ***   3 |(0, 0, 1)
0.932       9.68  0.97  1712444  ***   3 |(1, 1, 1)
0.933       9.70  0.97  856921   ***   2 |(1, 1, 0)
0.197       5.04  0.50  847218         2 |(-1, 1, 0)
0.938       9.78  0.98  857342   ***   2 |(1, 0, 1)
0.197       5.05  0.50  851699         2 |(-1, 0, 1)
0.926       9.59  0.96  876485   ***   2 |(0, 1, 1)
0.195       5.02  0.50  862772         2 |(0, -1, 1)
0.197       5.05  0.50  847298         2 |(1, 1, 2)
0.197       5.05  0.50  847688         2 |(1, 2, 1)
0.198       5.06  0.51  855659         2 |(2, 1, 1)
-----------------------------------------------------
```

PRECURSOR: SYMMETRY ANALYSIS WITH *DIALS.SYMMETRY / POINTLESS / XDS*



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

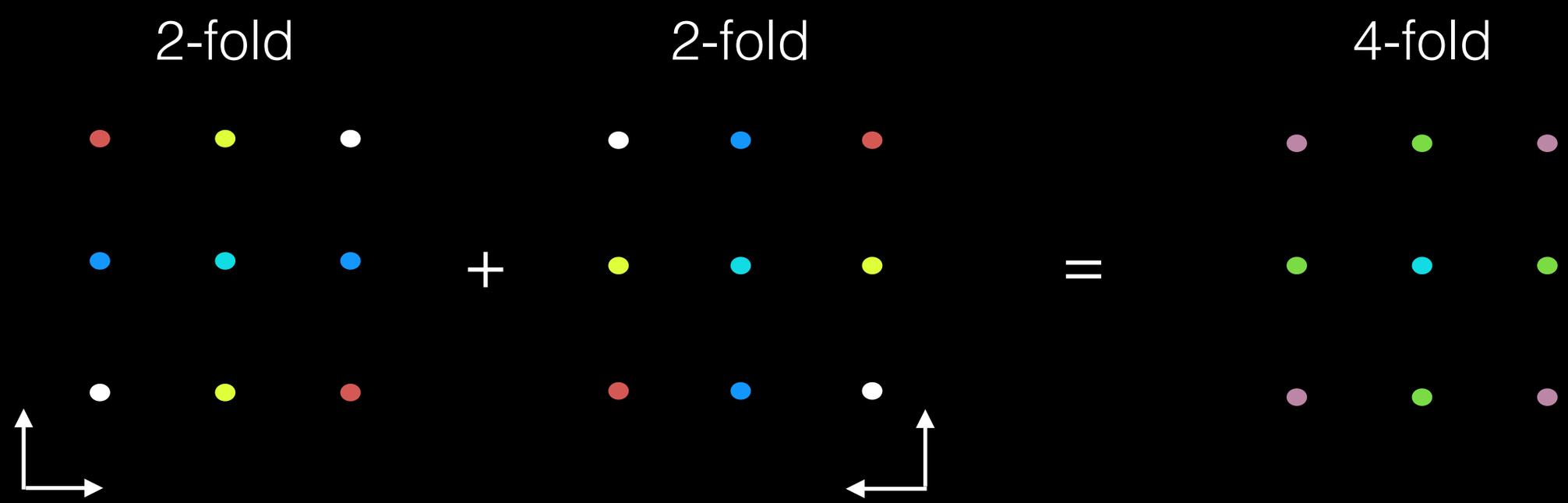
Scoring all possible sub-groups

Patterson group		Likelihood	NetZcc	Zcc+	Zcc-	CC	CC-	delta	Reindex operator
I m -3	***	1.000	4.70	9.74	5.04	0.97	0.50	0.0	-a,b,-c
I m m m		0.000	2.91	9.74	6.84	0.97	0.70	0.0	-a,b,-c
I m -3 m		0.000	7.62	7.62	0.00	0.74	0.00	0.0	-a,b,-c
I 4/m m m		0.000	0.78	8.07	7.29	0.74	0.74	0.0	-c,-a,b
I 4/m m m		0.000	0.78	8.07	7.29	0.74	0.74	0.0	-a,b,-c
I 4/m m m		0.000	0.78	8.07	7.29	0.74	0.74	0.0	b,-c,-a
R -3 :H		0.000	2.58	9.85	7.27	0.98	0.71	0.0	b-c,-a-b,-1/2*a+1
R -3 :H		0.000	2.57	9.85	7.27	0.98	0.71	0.0	-a+b,a-c,-1/2*a-1
I 1 2/m 1		0.000	2.56	9.84	7.28	0.98	0.72	0.0	a,-b,-c
I 1 2/m 1		0.000	2.51	9.80	7.28	0.98	0.72	0.0	-a,-c,-b
R -3 :H		0.000	2.50	9.79	7.28	0.98	0.71	0.0	-b-c,-a+c,-1/2*a+
R -3 :H		0.000	2.50	9.78	7.29	0.97	0.71	0.0	-a-c,b+c,1/2*a+1/
I 1 2/m 1		0.000	2.45	9.75	7.29	0.98	0.72	0.0	b,-a,c
I 4/m		0.000	1.13	8.54	7.41	0.75	0.74	0.0	-c,-a,b
I 4/m		0.000	1.09	8.51	7.42	0.75	0.74	0.0	-a,b,-c
I 4/m		0.000	1.04	8.47	7.43	0.74	0.74	0.0	b,-c,-a
P -1		0.000	2.44	9.90	7.46	0.99	0.73	0.0	x+y,x-z,y-z
F m m m		0.000	0.26	7.82	7.56	0.75	0.74	0.0	-b,-a+c,-a-c
F m m m		0.000	0.22	7.79	7.57	0.75	0.74	0.0	-c,-a-b,-a+b
F m m m		0.000	0.18	7.76	7.58	0.74	0.74	0.0	-a,b+c,b-c

Additionally, assess systematic absences to determine full space group symmetry (not needed for scaling).

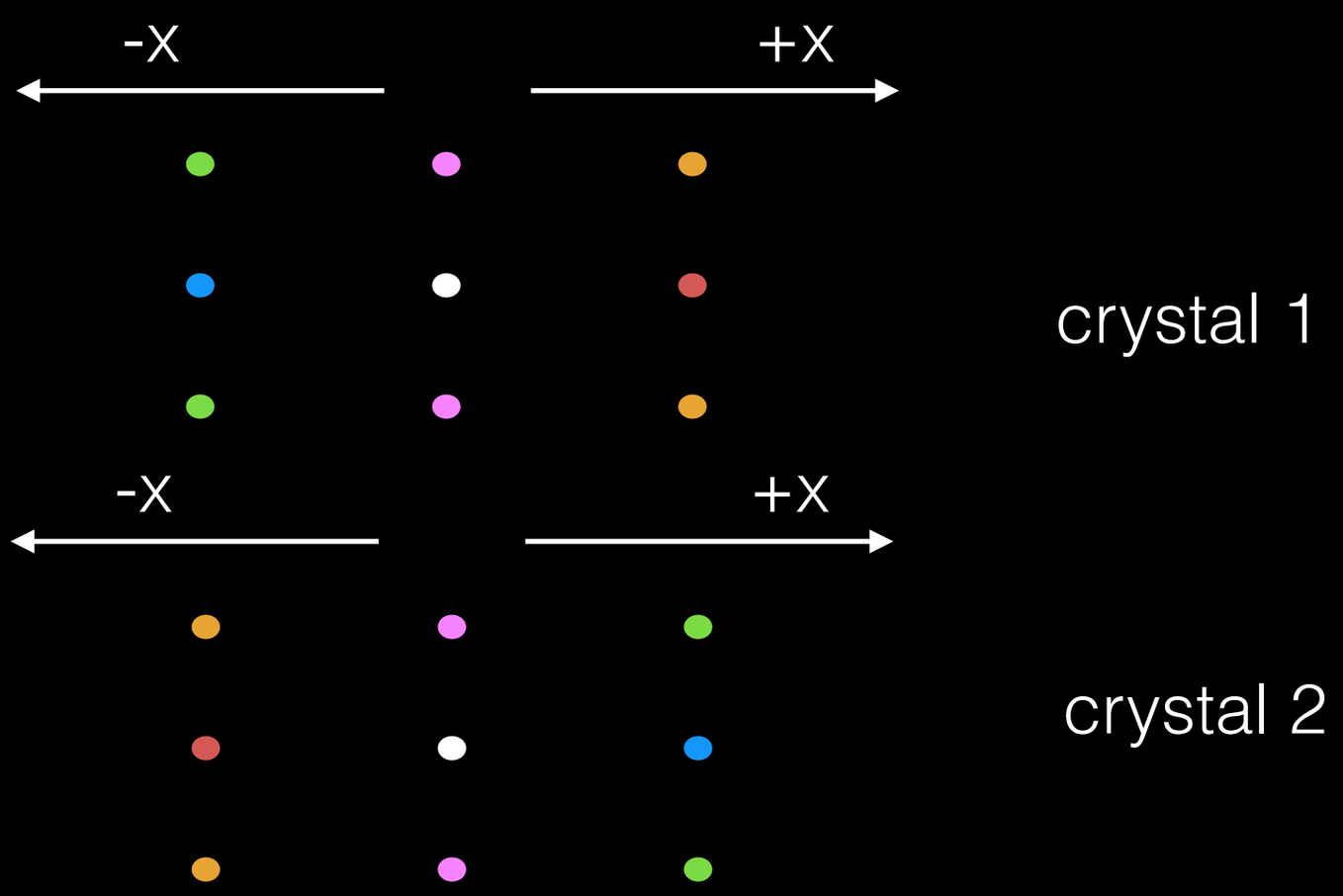
SYMMETRY ANALYSIS DIFFICULTIES

Effects of twinning & pseudo-symmetry, can give appearance of higher symmetry. 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, there are 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. Alternatives: xscale_isocluster, 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)

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 *P*1.

SCALING

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

Create a 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 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: dials.scale, XDS/XSCALE, Aimless, scalepack, …..

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

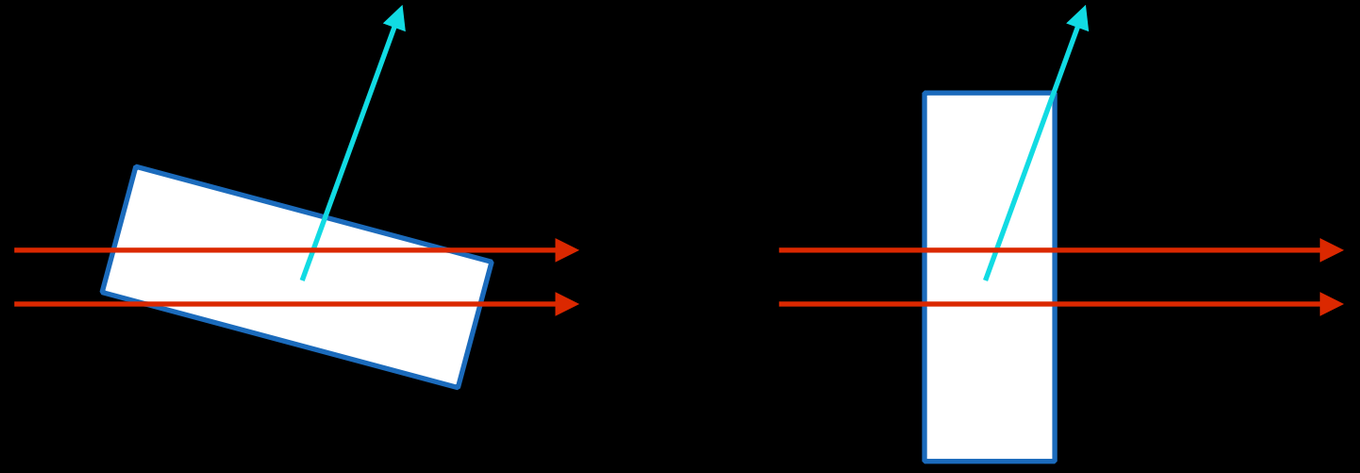
SCALING PROCESS COMPONENTS

- rounds of outlier rejection (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)

SCALING MODELS

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

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



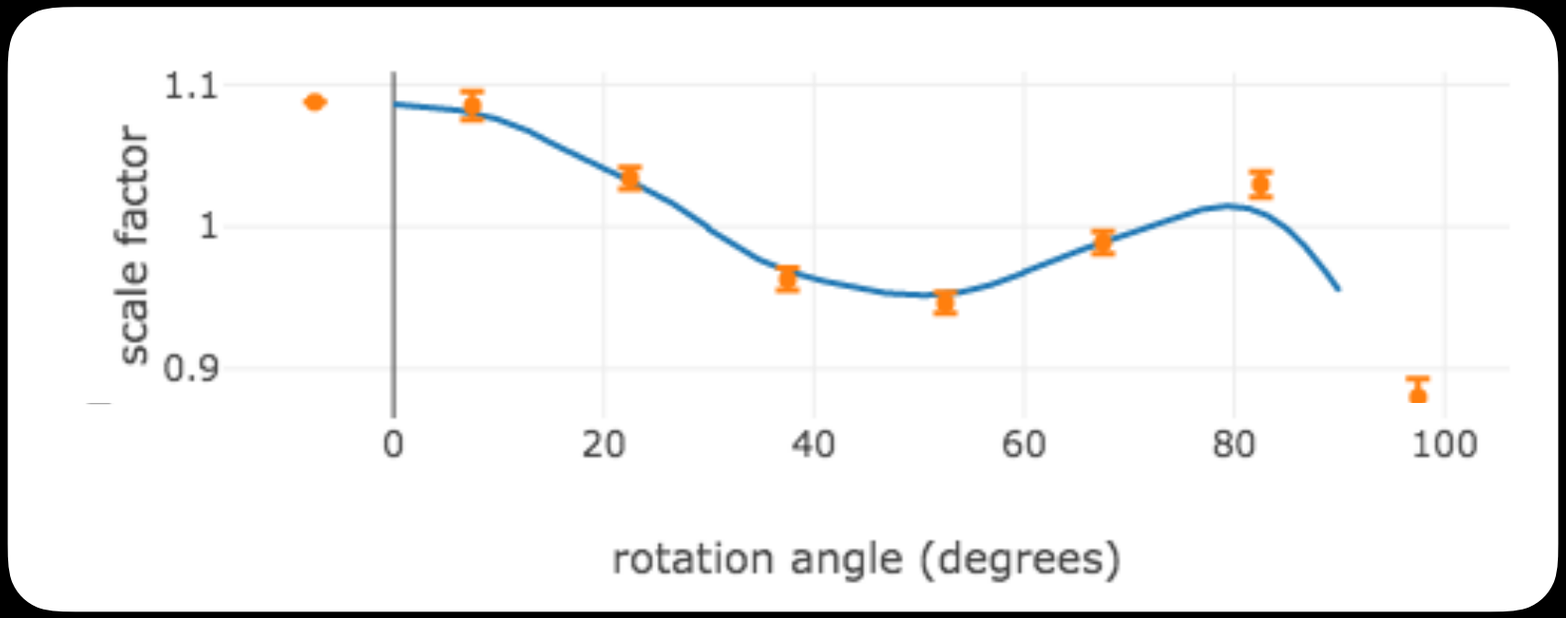
APPROACH 1: PHYSICAL SCALING MODEL

Three part model *per sweep/crystal*, approach used in Aimless & DIALS

Correction 1: Smoothly varying scale factor, dependent on crystal rotation angle.

Corrects for changes in illuminated volume, slow changes in beam intensity, primary absorption.

Gives the relative scale for multi-sweep/multi-crystal datasets.



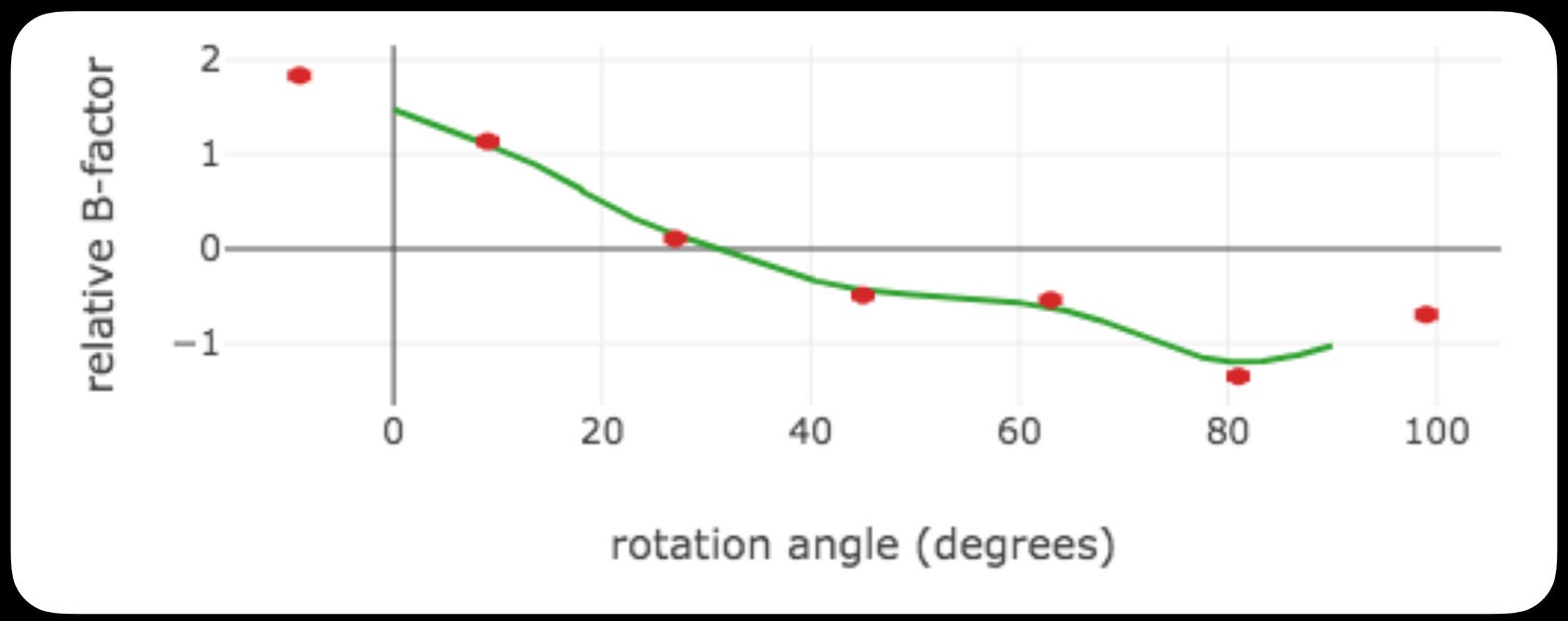
APPROACH 1: PHYSICAL SCALING MODEL

Three part model, approach used in Aimless & DIALS

Correction 2: Smoothly varying isotropic B-factor, dependent on crystal rotation angle (i.e. accumulated exposure time).

$$e(B(t)/2d^2)$$

Corrects for global/average radiation damage (*not* site specific).



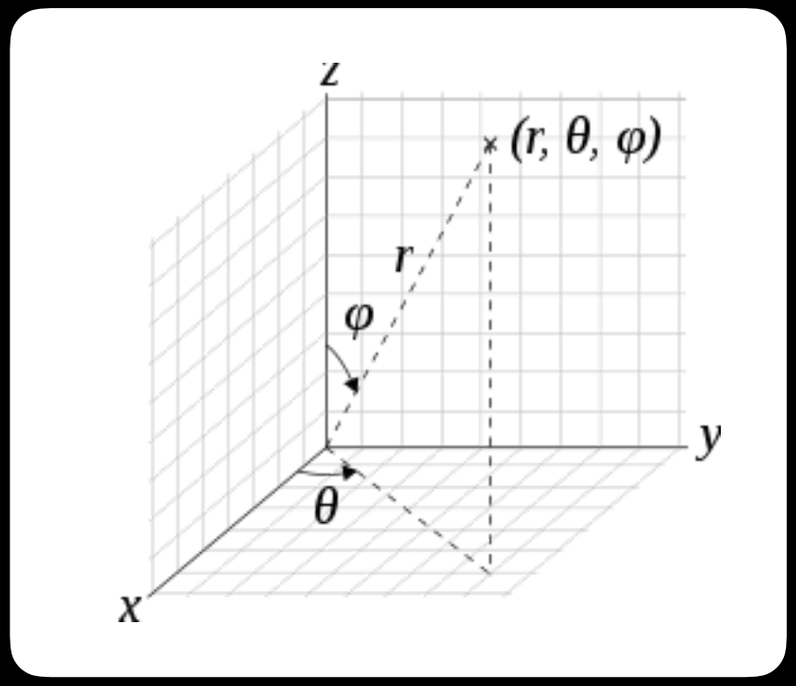
APPROACH 1: PHYSICAL SCALING MODEL

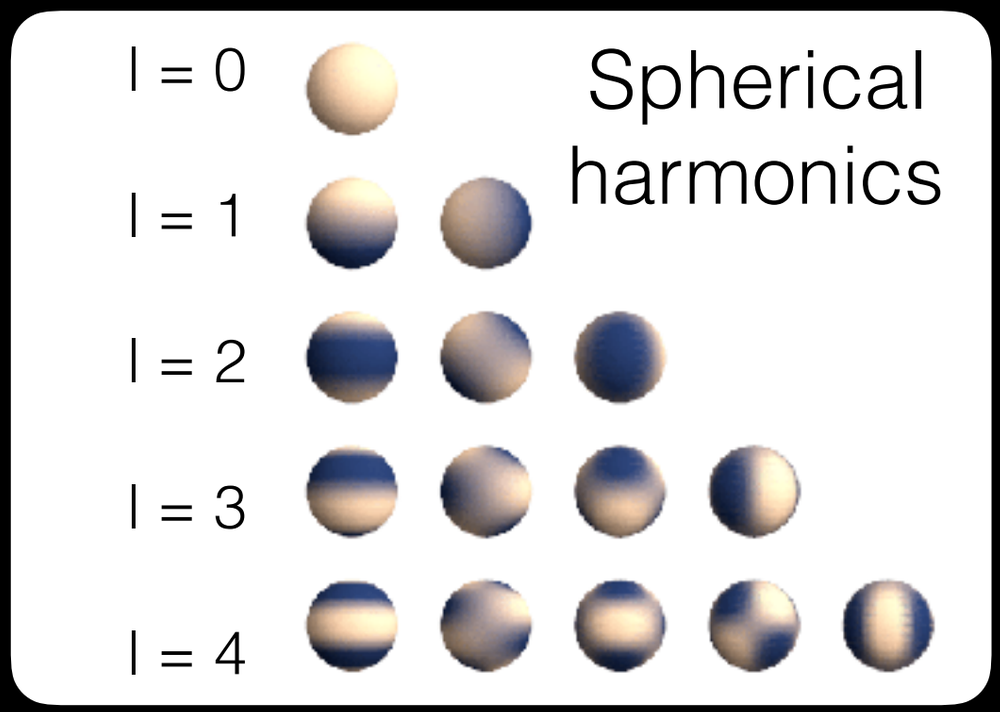
Three part model, approach used in Aimless & DIALS

Correction 3: relative absorption correction for different directions in space, based on scattering vectors.

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

Typically small for MX (~1%), unless at long wavelength/absorption edge.

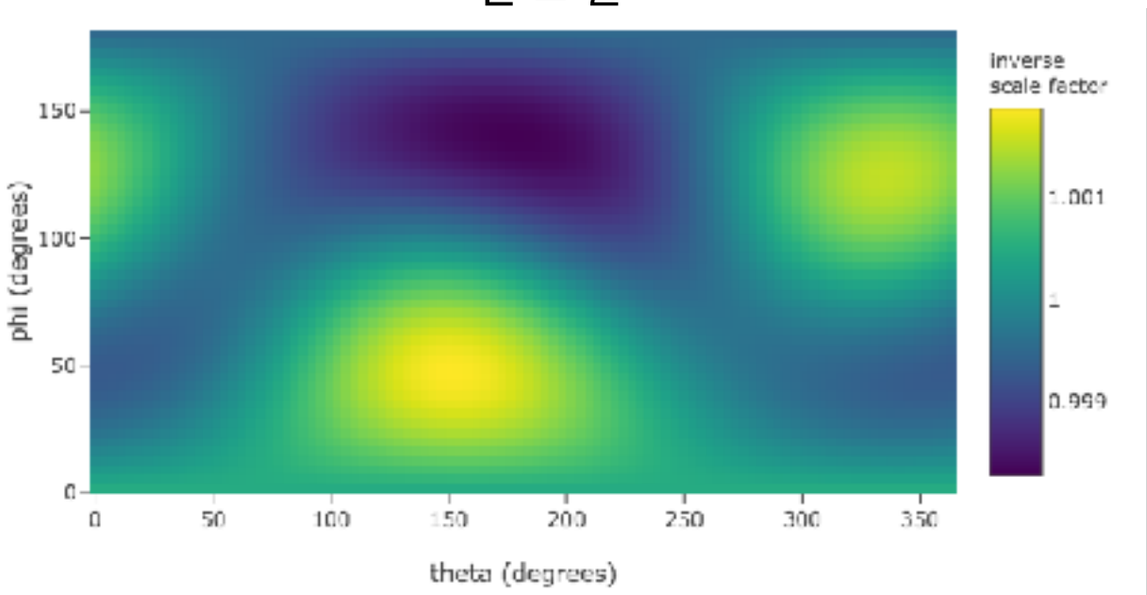


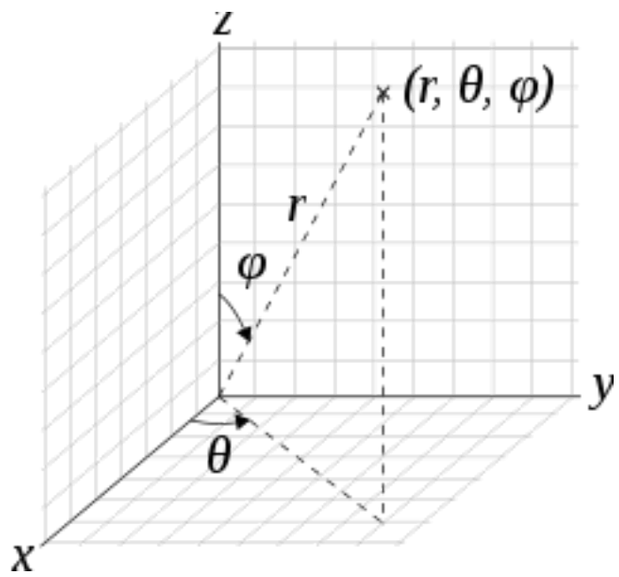


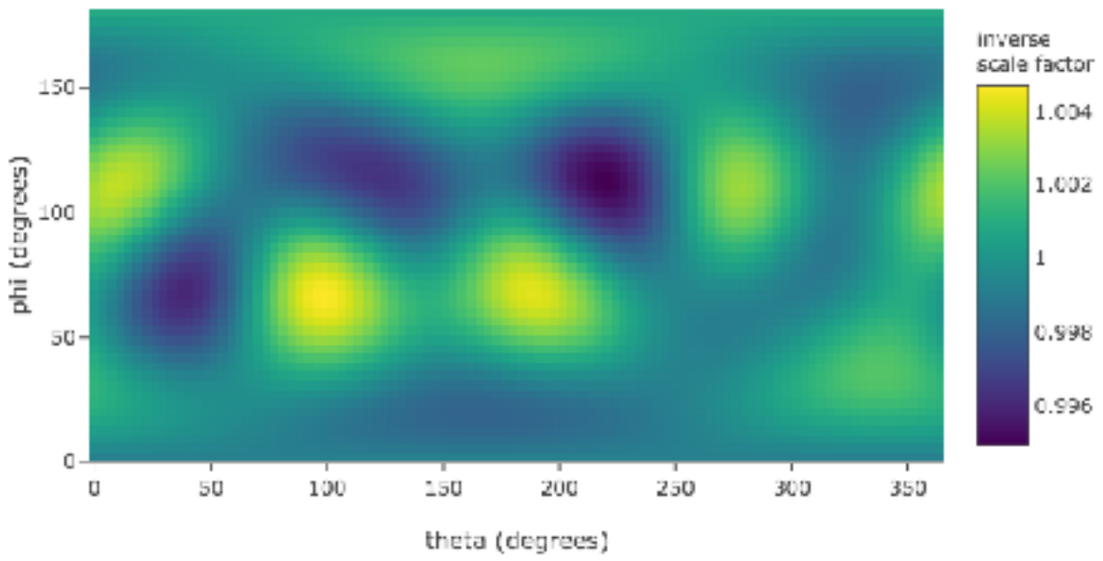
APPROACH 1: PHYSICAL SCALING MODEL

Effect of the number of spherical harmonics on the absorption surface

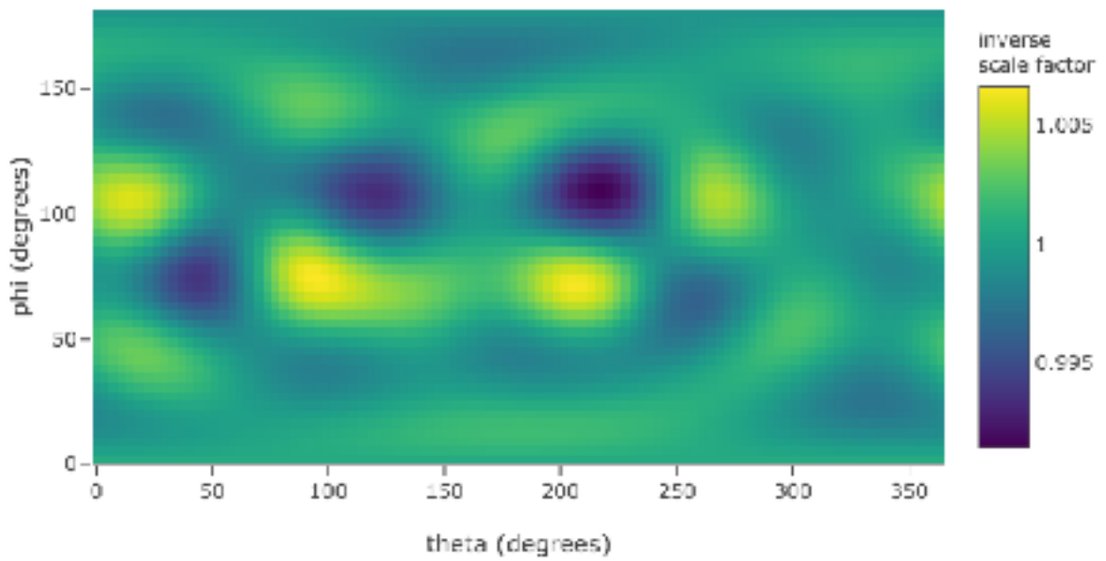
L = 2







L = 4



L = 6

APPROACH 2: 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.

Components are arrays/grids of parameters e.g. 2D grid of parameters equally spaces in resolution and rotation angle.

Differences to physical scaling model: Less restrained by assumptions on form of corrections, more general/adaptable. Physical interpretation of the corrections may be less clear. Array model implementation typically requires many more parameters than physical model (100s vs ~60).

APPROACH 2: ARRAY SCALING MODEL

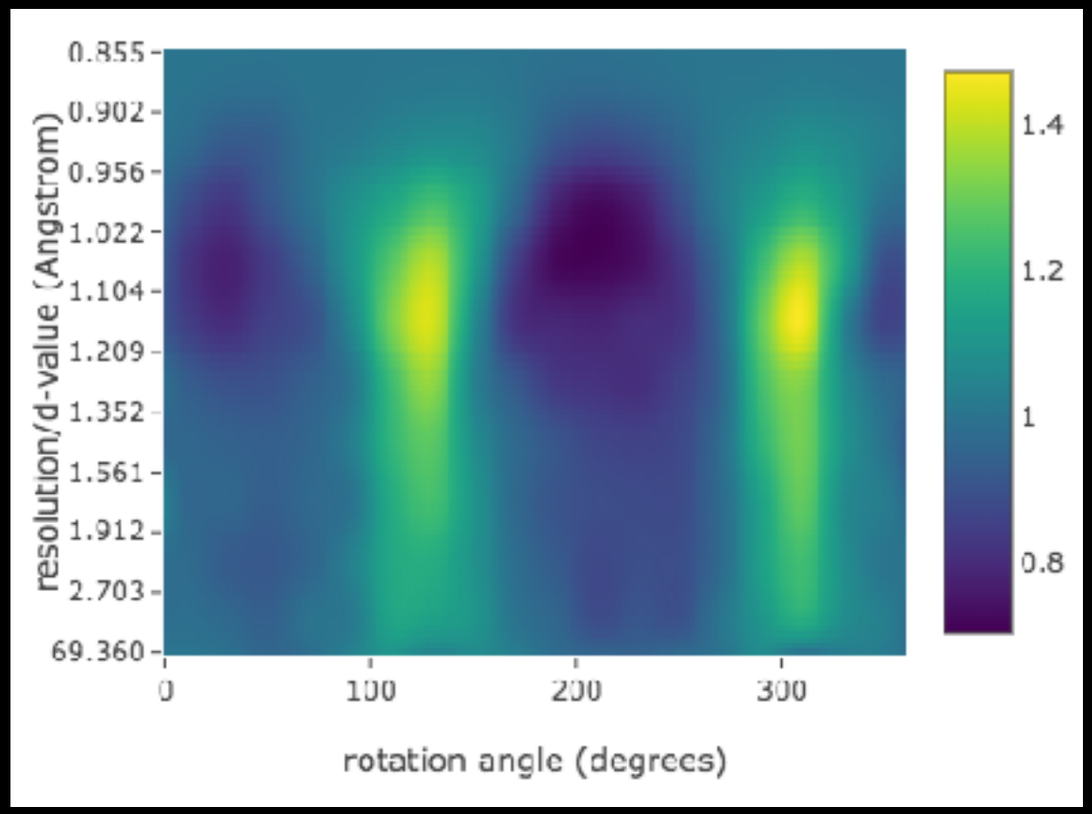
Have an implementation of this approach as an option within dials.scale.
Same principles as behind the model from XDS/Xscale.

model=array (default model=physical)

Smoothly weight a grid of parameters to get correction for a reflection.

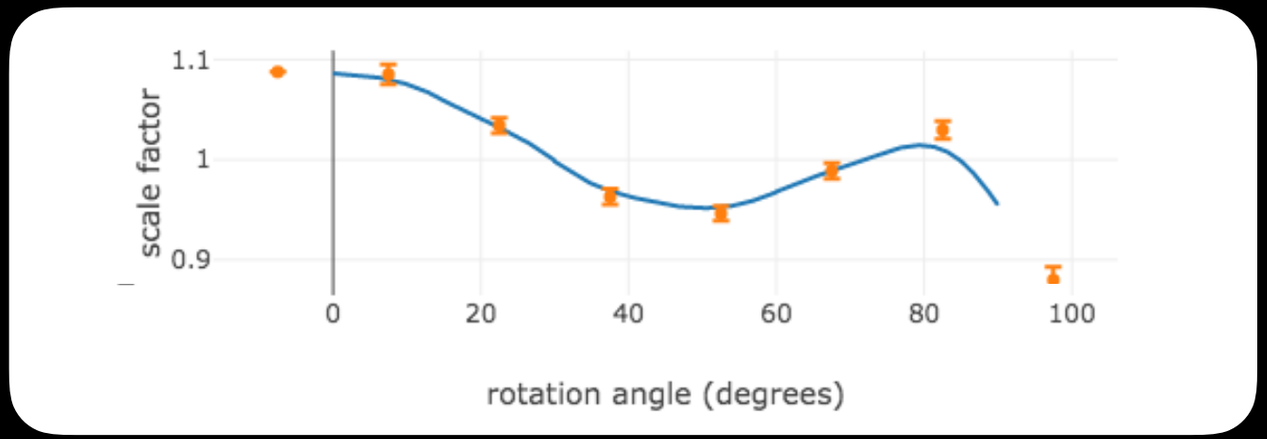
Decay correction - 2D grid of parameters; d-values & rotation angle.

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

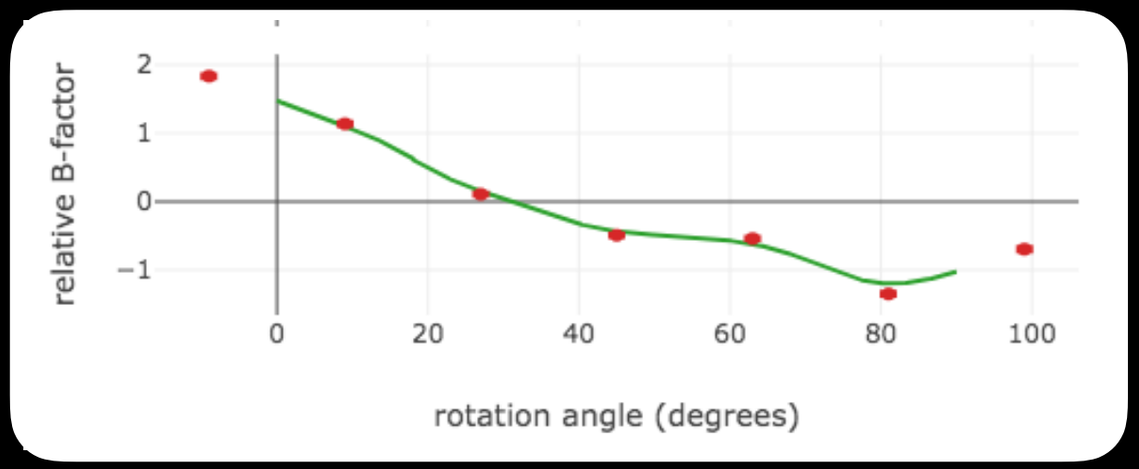


RECAP: DIALS SCALING MODEL

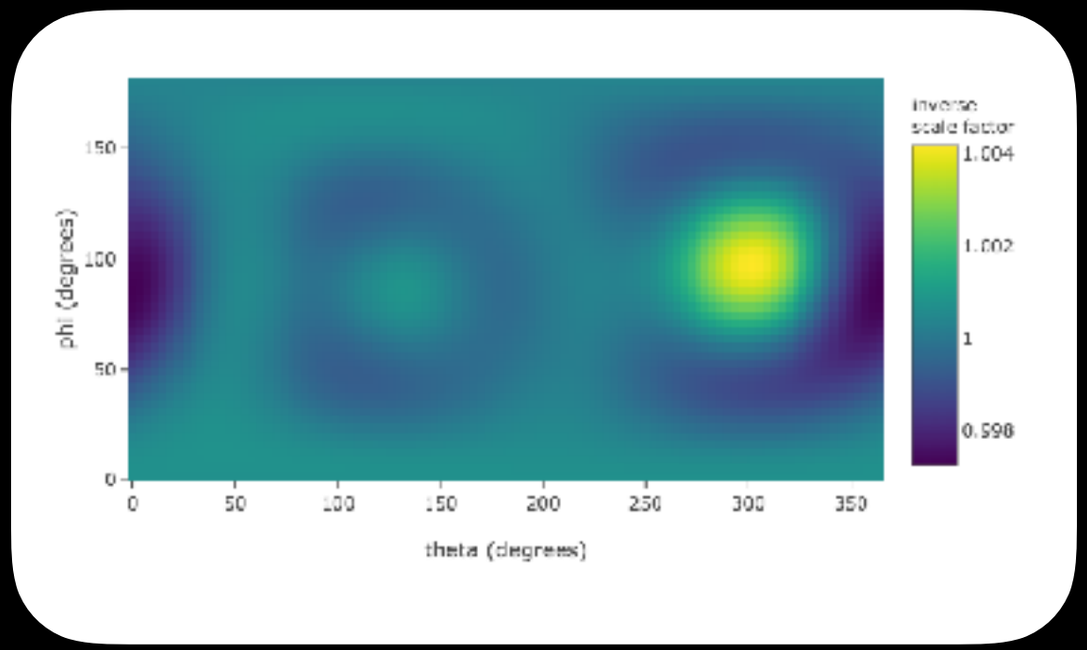
Scaling correction =



x



X



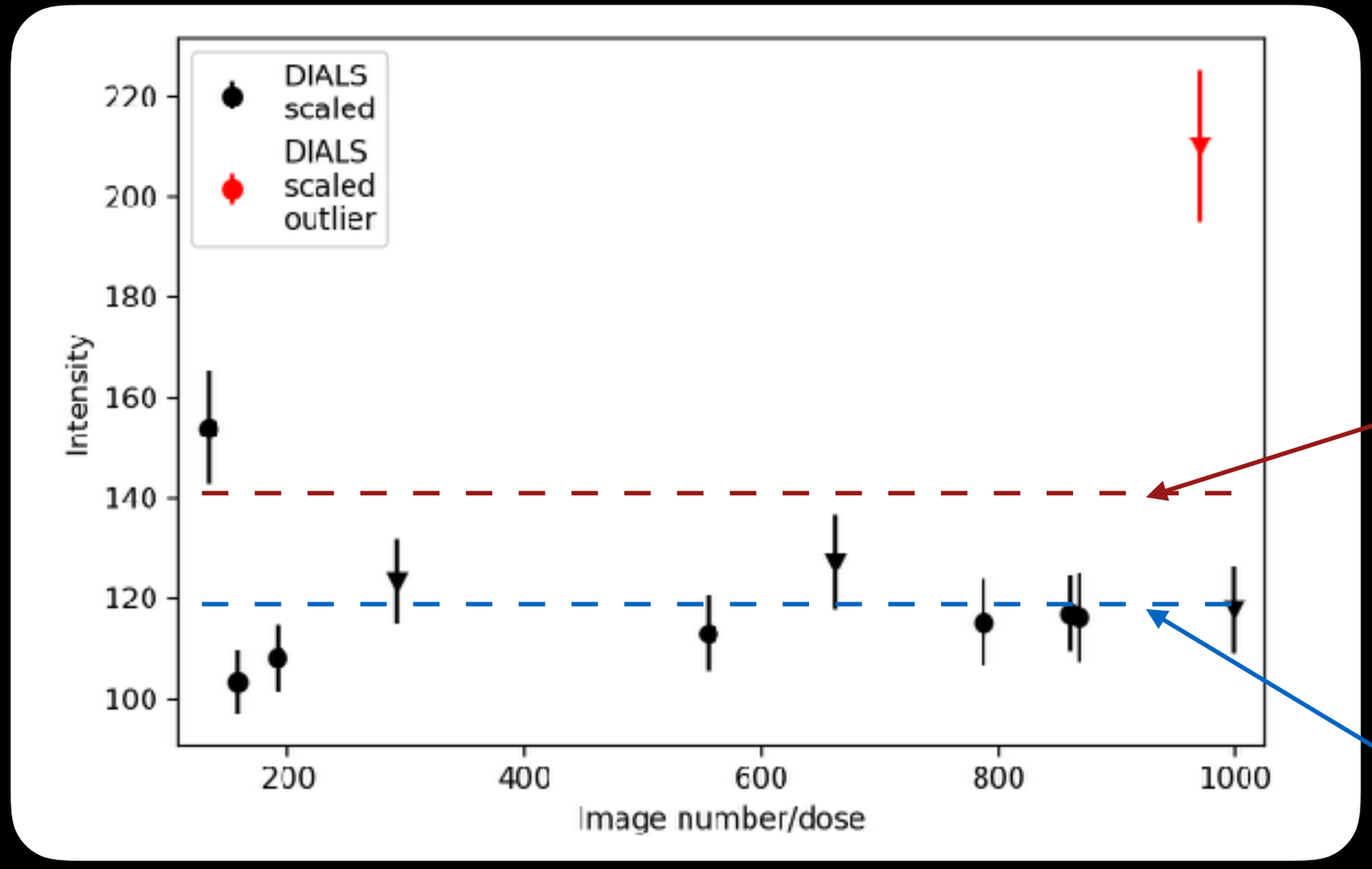
OUTLIER REJECTION

Outliers have a disproportionate effect - largest RMSD.

Test the normalised deviation from the symmetry group, reject those above a threshold.

$$\delta = (I - g\langle I_h \rangle)/\sigma'$$

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



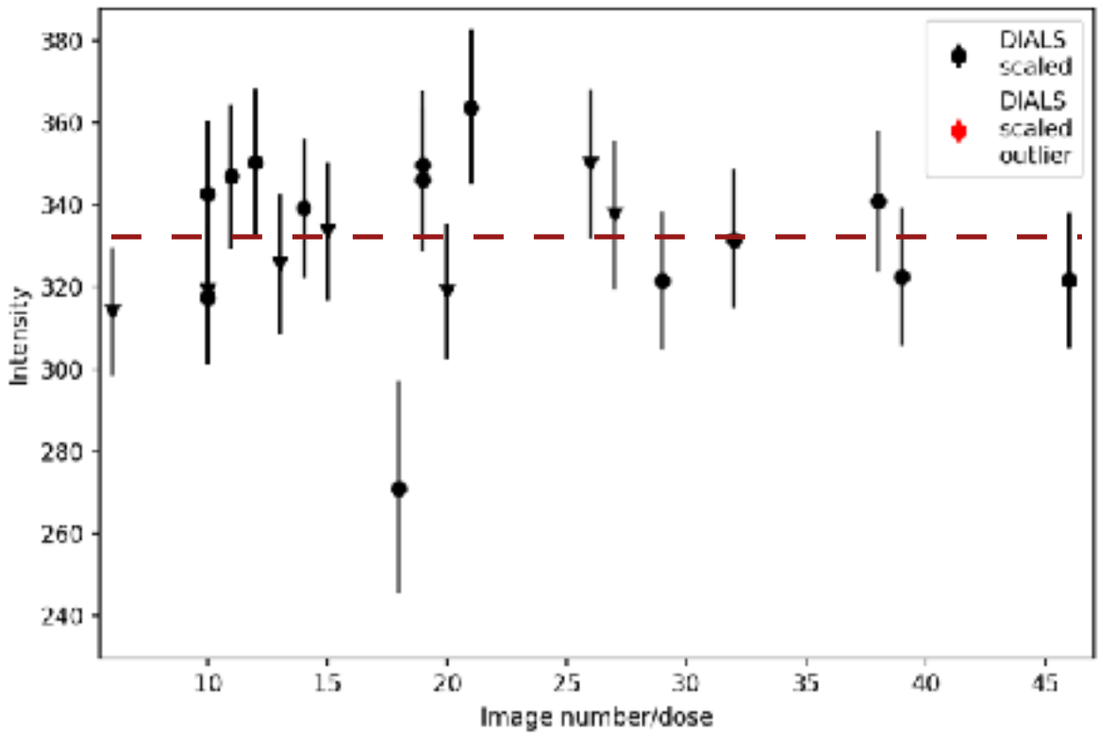
Average with outlier

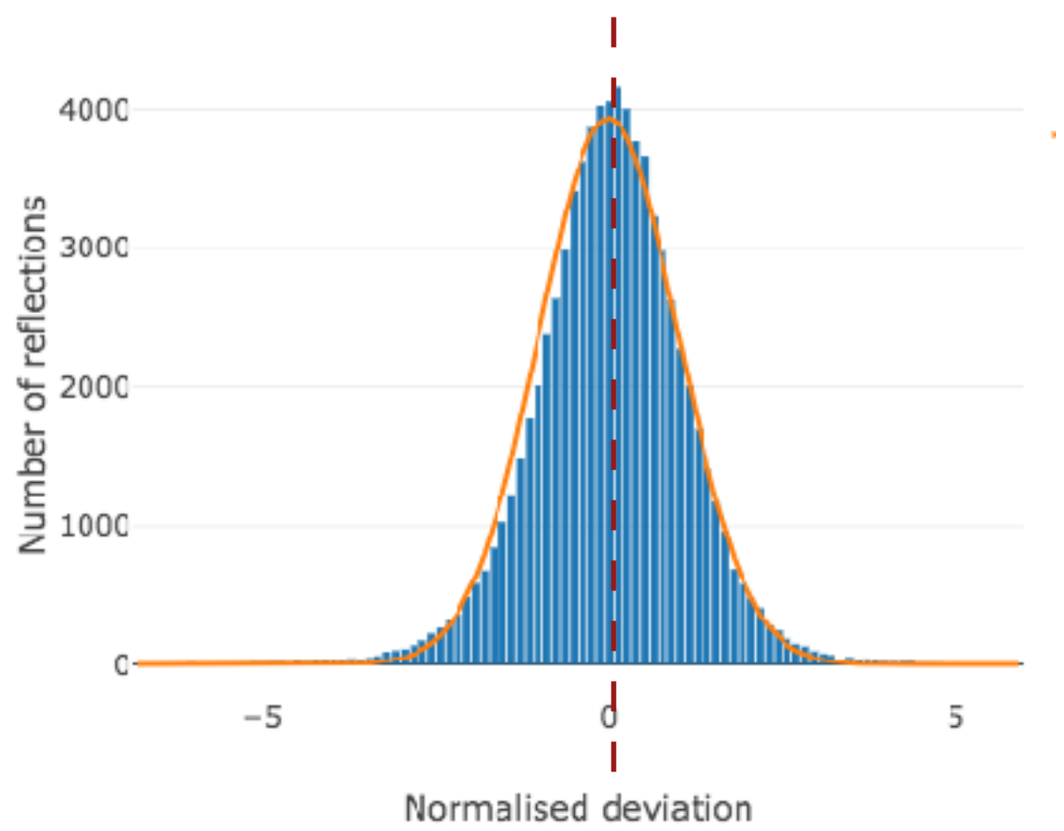
Average excluding outlier

ERROR (SIGMA) DISTRIBUTION

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

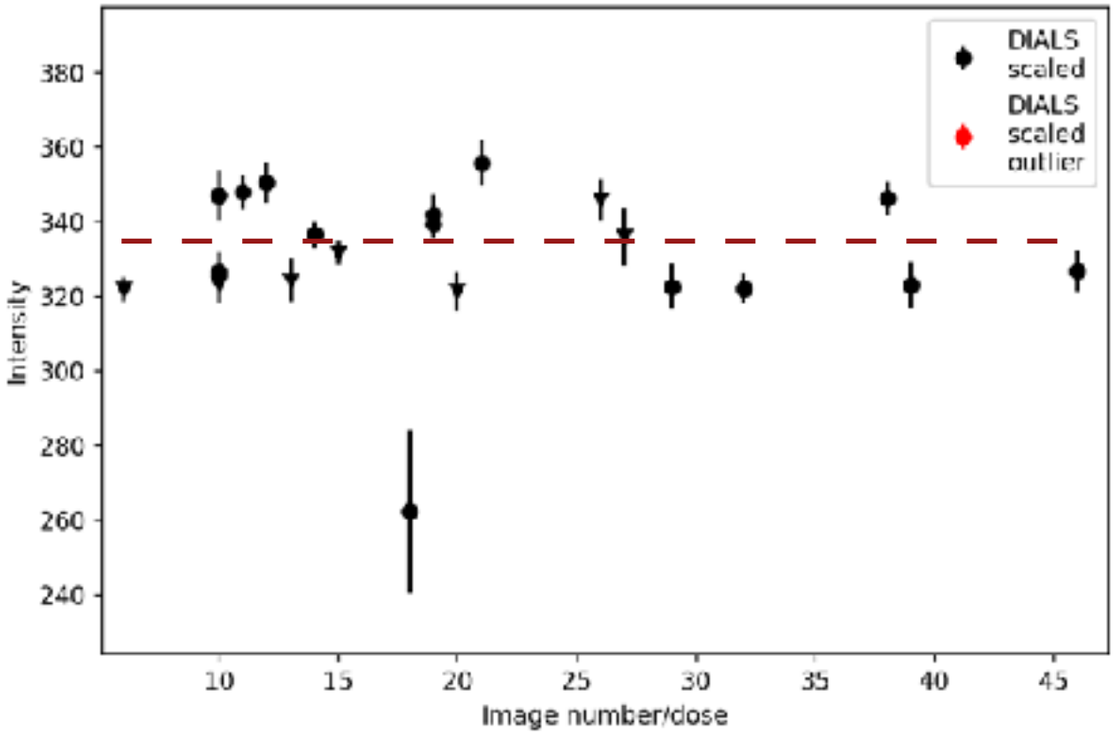
Principle: for realistic errors, the overall distribution of intensities is consistent with the errors on individual reflections.



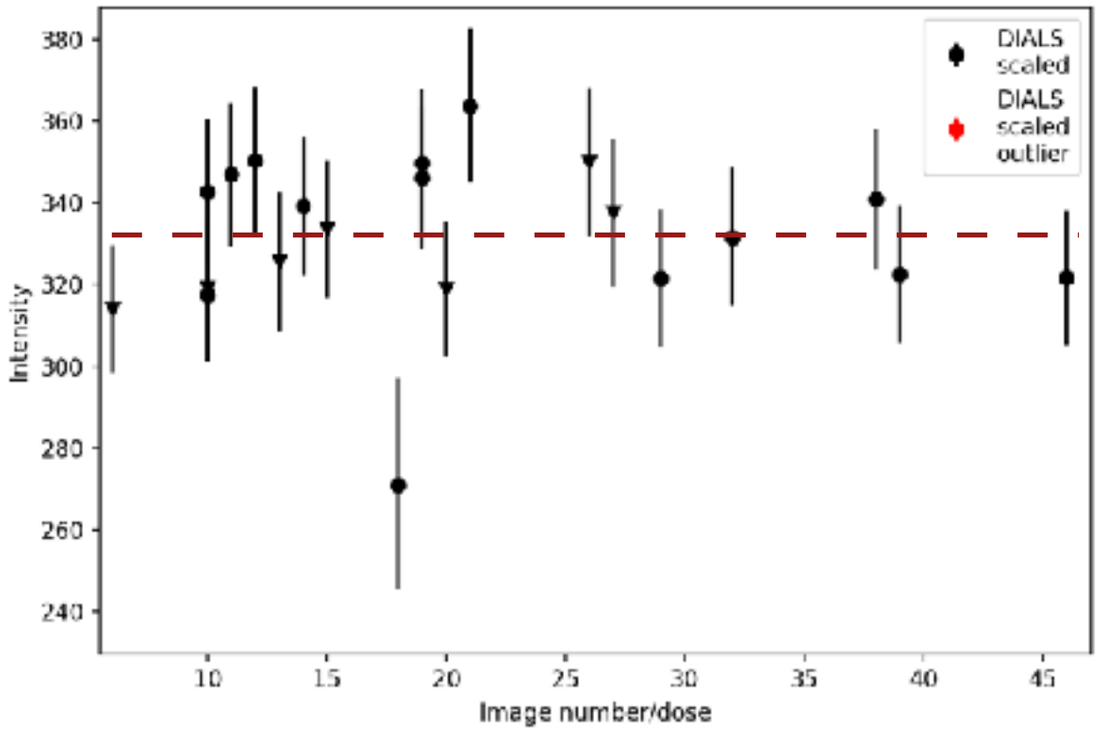


ERROR (SIGMA) DISTRIBUTION

In reality, the intensities deviate from the average by more than expected based upon the errors after integration.
The true error has been *underestimated,* and the effect is most pronounced for the most intense reflections.



Scaled, no error adjustment



Scaled, errors adjusted

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

ERROR (SIGMA) ADJUSTMENT

General approach is to use two parameters to describe the effect of systematic errors and optimise these by assessing the distribution of intensities and uncertainties.

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

First parameter *a* - scales all sigmas by a constant factor - represents the effect of systematic errors that affect all reflections.

Second term *b* - intensity dependent factor - represents the effect of systematic errors that scale with intensity.

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

a x b is measure of fraction of systematic error (2% very good)

ERROR (SIGMA) ADJUSTMENT

Similarities in form across different data processing programs

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

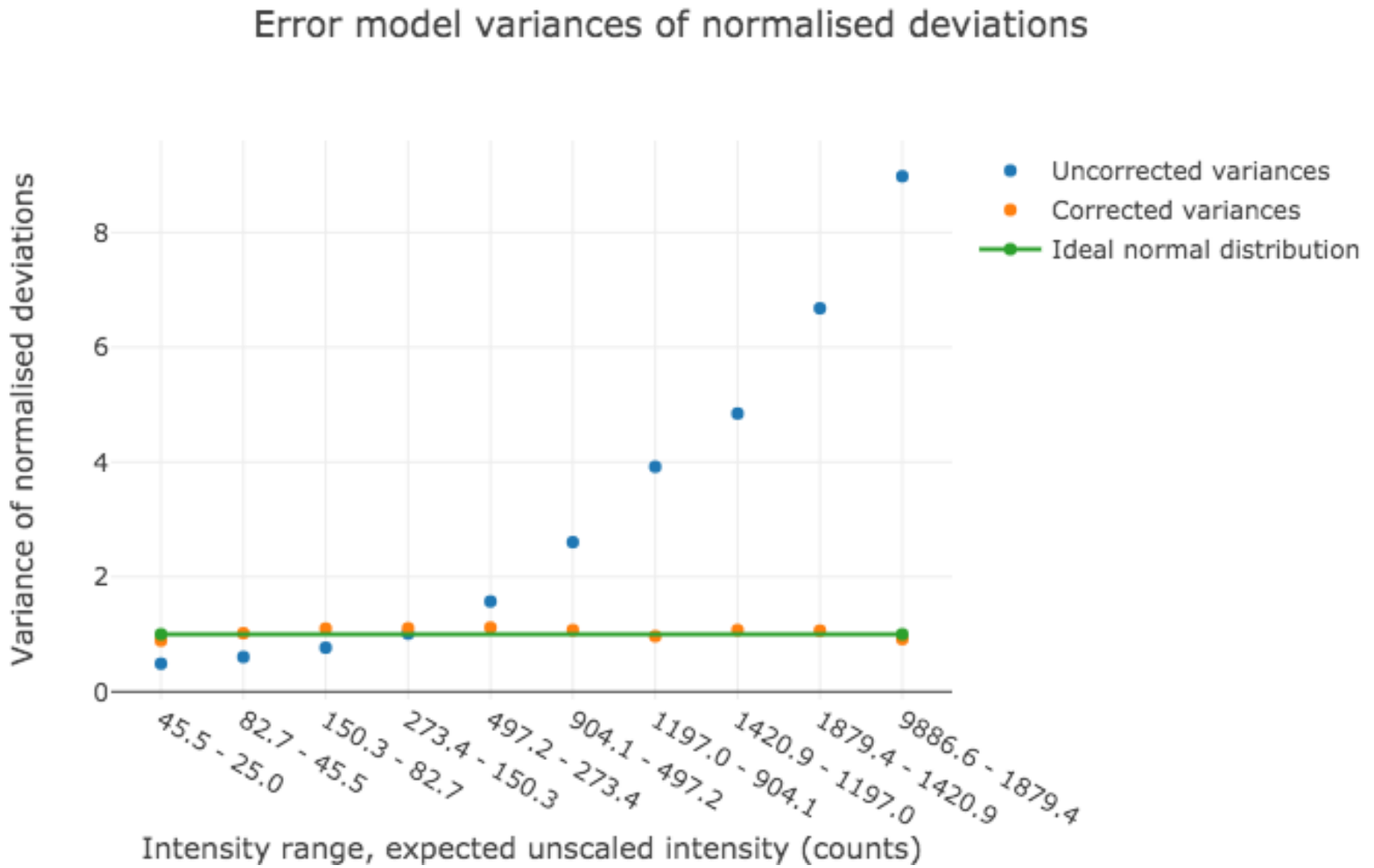
Aimless $$\sigma(I)_{\text{corr}} = SdFac * [\sigma(I)^2 + (SdB * I) + (SdAdd * I)^2]^{\frac{1}{2}}$$

XDS $$\sigma(I)^2 = a * (\sigma(I)_0^2 + b * I^2)$$

Approaches to parameter fitting differ between programs, see papers for technical details.

DIALS ERROR (SIGMA) ADJUSTMENT

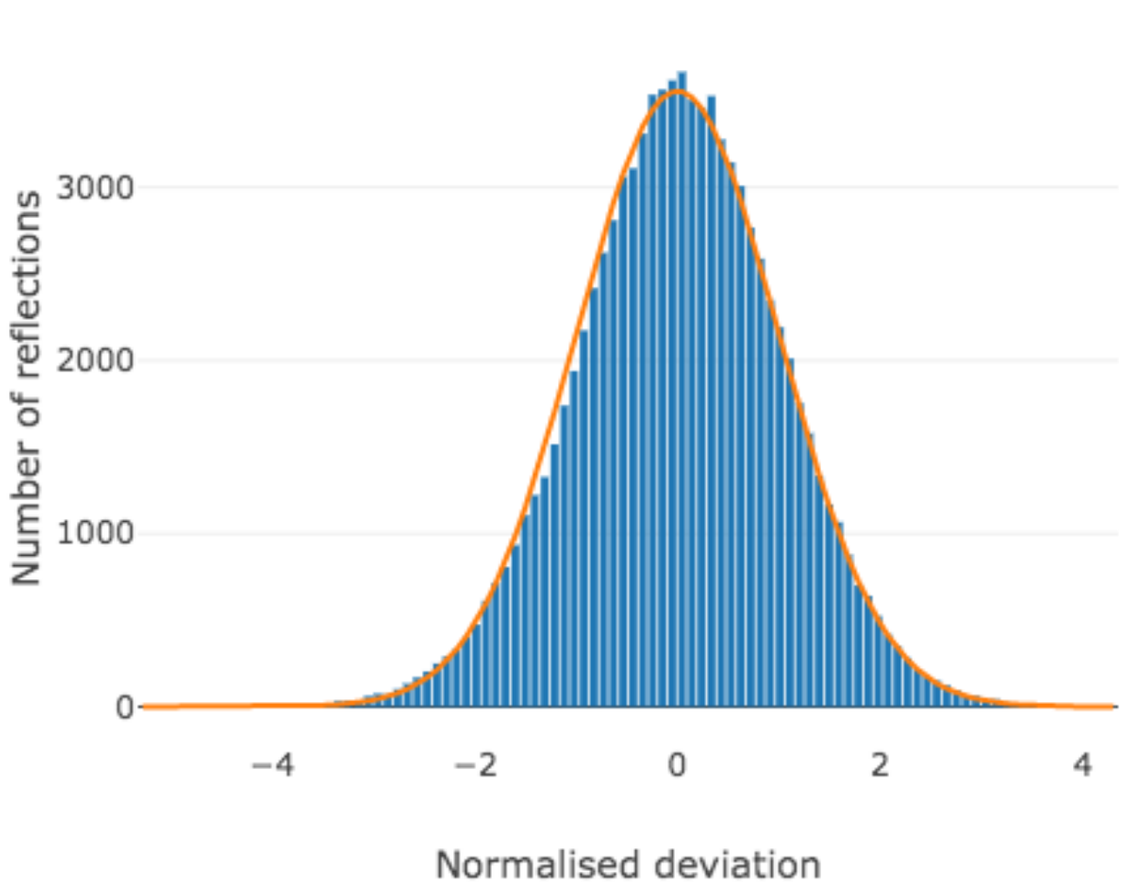
$$\sigma'^2 = a^2(\sigma^2 + (bI)^2)$$

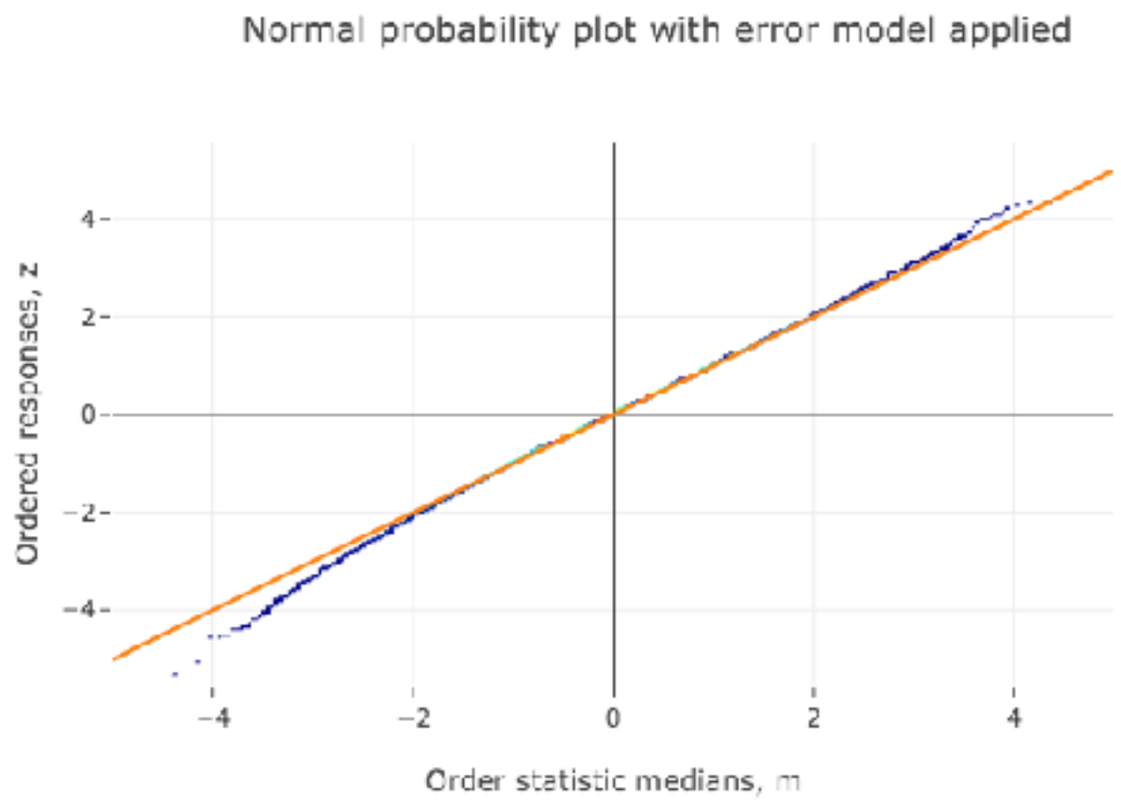


dials.scale.html, ‘Error distribution plots’

DIALS ERROR (SIGMA) ADJUSTMENT

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





dials.scale.html, ‘Error distribution plots’

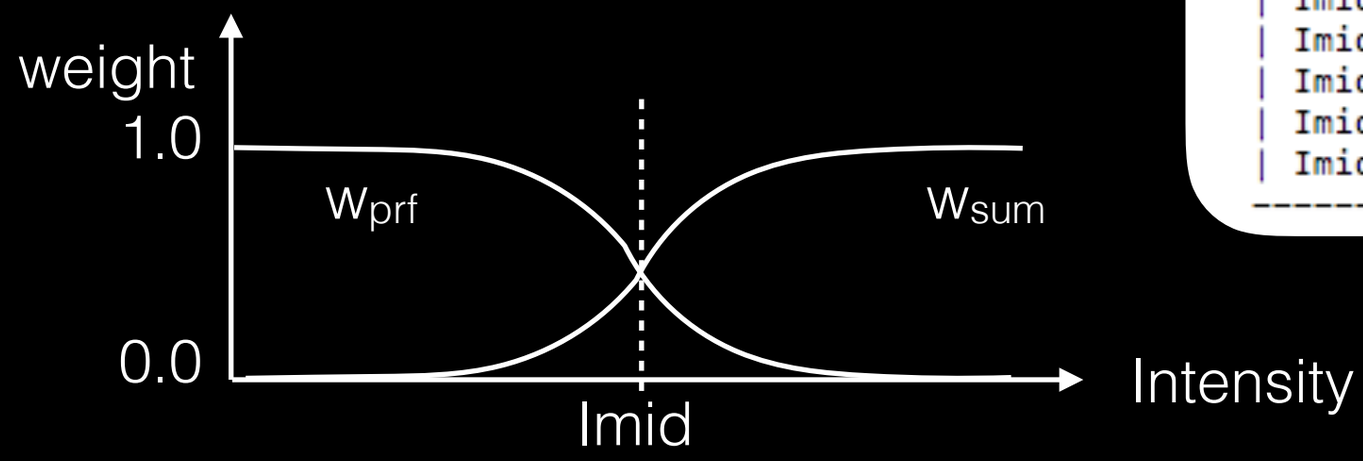
COMBINING SUMMATION/PROFILE INTENSITIES.

Integration programs give summation and profile fitted intensities. Profile intensity estimates are usually best, but summation estimates can be more accurate for the most intense observations.

Try to find best combination - smooth cross over between profile and summation, optimise crossover point. Score based on R_{meas} in dials.scale.

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

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




```
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 = 2003.81 | 0.99959 | 0.21789 |
| Imid = 200.38  | 0.9996  | 0.21773 |
| Imid = 20.04   | 0.99956 | 0.22256 |
| Imid = 2.0     | 0.99946 | 0.25967 |
---------------------------------------
```

RECAP: DATA REDUCTION STEP 2

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

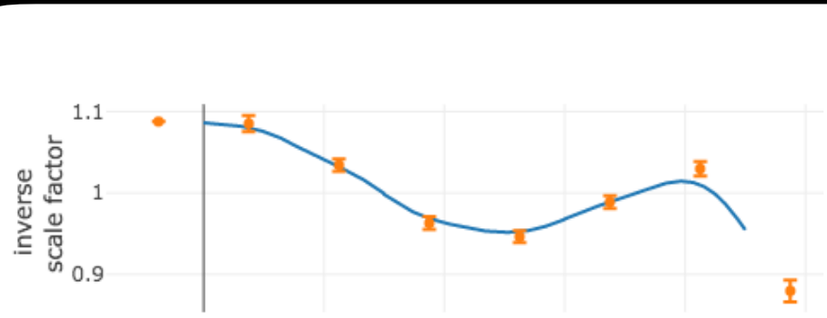
Integrated reflections

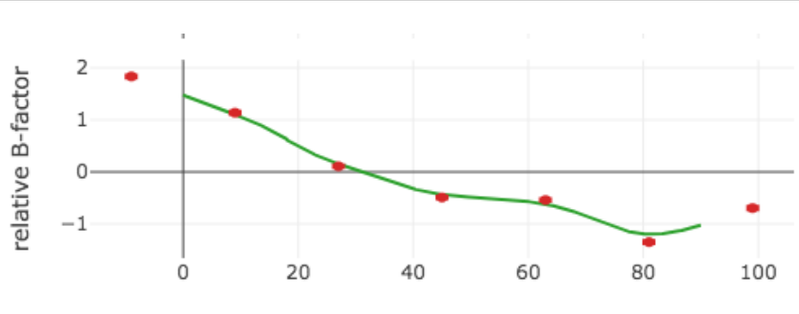
↓

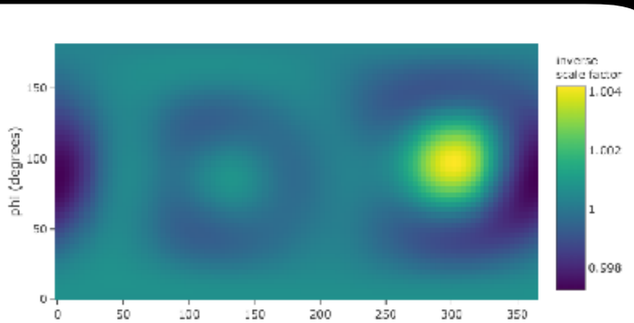
Integrated reflections, in correct point group

↓

Scaled (corrected) intensities and sigmas







ASSESSING SCALING RESULTS

Summary statistics tables

----------Merging statistics by resolution bin----------

d_max	d_min	#obs	#uniq	mult.	%comp	<I>	<I/sI>	r_mrg	r_meas	r_pim	cc1/2	cc_ano
69.31	3.29	22601	3392	6.66	98.63	532.6	69.2	0.037	0.040	0.015	0.999*	0.297*
3.29	2.61	22267	3319	6.71	97.53	193.3	47.8	0.050	0.054	0.021	0.998*	0.338*
2.61	2.28	22519	3258	6.91	96.76	99.2	35.7	0.065	0.070	0.027	0.997*	0.362*
2.28	2.07	22169	3243	6.84	95.95	67.5	27.3	0.078	0.085	0.032	0.996*	0.249*
2.07	1.93	21502	3167	6.79	95.19	44.0	20.5	0.099	0.107	0.041	0.994*	0.204*
1.93	1.81	21655	3187	6.79	94.37	25.0	14.1	0.136	0.147	0.056	0.991*	0.188*
1.81	1.72	21246	3151	6.74	93.89	15.6	10.0	0.180	0.195	0.074	0.986*	0.146*
1.72	1.65	20887	3104	6.73	93.07	10.4	7.2	0.232	0.252	0.096	0.978*	0.148*
1.65	1.58	21639	3123	6.93	92.64	8.3	6.1	0.275	0.297	0.112	0.973*	0.055*
1.58	1.53	20674	3082	6.71	92.00	6.3	4.7	0.327	0.354	0.136	0.955*	0.063*
1.53	1.48	20145	3066	6.57	91.58	5.0	3.8	0.382	0.415	0.161	0.939*	0.021
1.48	1.44	20801	3029	6.87	90.74	3.7	3.0	0.473	0.512	0.194	0.922*	0.055*
1.44	1.40	18330	2982	6.15	90.31	3.2	2.5	0.523	0.571	0.227	0.891*	-0.002
1.40	1.37	13881	2480	5.60	72.83	2.7	2.0	0.585	0.646	0.268	0.811*	-0.001
1.37	1.33	9578	1749	5.48	52.71	2.4	1.8	0.638	0.704	0.292	0.748*	-0.019
1.33	1.31	6726	1287	5.23	38.73	2.3	1.6	0.654	0.725	0.305	0.737*	0.003
1.31	1.28	4850	1001	4.85	29.69	2.0	1.3	0.745	0.833	0.362	0.603*	-0.018
1.28	1.26	2923	700	4.18	21.03	1.8	1.1	0.854	0.973	0.452	0.406*	0.029
1.26	1.23	1214	407	2.98	12.37	1.5	0.8	0.895	1.067	0.567	0.284*	-0.156
1.23	1.21	305	191	1.60	5.64	1.6	0.6	0.746	0.987	0.640	0.489*	-0.357
69.19	1.21	315912	48918	6.46	72.87	69.3	17.1	0.064	0.070	0.027	0.999*	0.358*

ASSESSING SCALING RESULTS

Summary statistics tables

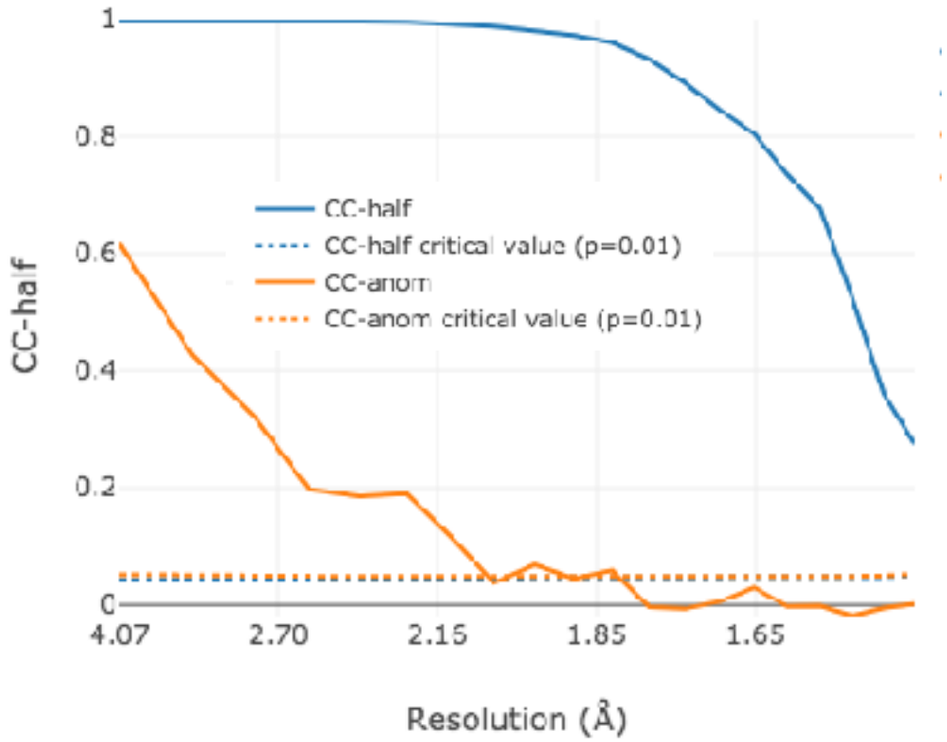
----------Summary of merging statistics----------

	Overall	Low	High
High resolution limit	1.21	3.29	1.21
Low resolution limit	69.19	69.31	1.23
Completeness	72.9	98.6	5.6
Multiplicity	6.5	6.7	1.6
I/sigma	17.1	69.2	0.6
Rmerge(I)	0.064	0.037	0.746
Rmerge(I+/-)	0.055	0.030	0.697
Rmeas(I)	0.070	0.040	0.987
Rmeas(I+/-)	0.065	0.036	0.986
Rpim(I)	0.027	0.015	0.640
Rpim(I+/-)	0.035	0.019	0.697
CC half	0.999	0.999	0.489
Anomalous completeness	71.5	98.8	1.4
Anomalous multiplicity	3.3	3.5	1.3
Anomalous correlation	0.358	0.297	-0.357
Anomalous slope	1.097		
dF/F	0.070		
dI/s(dI)	1.265		
Total observations	315912	22601	305
Total unique	48918	3392	191

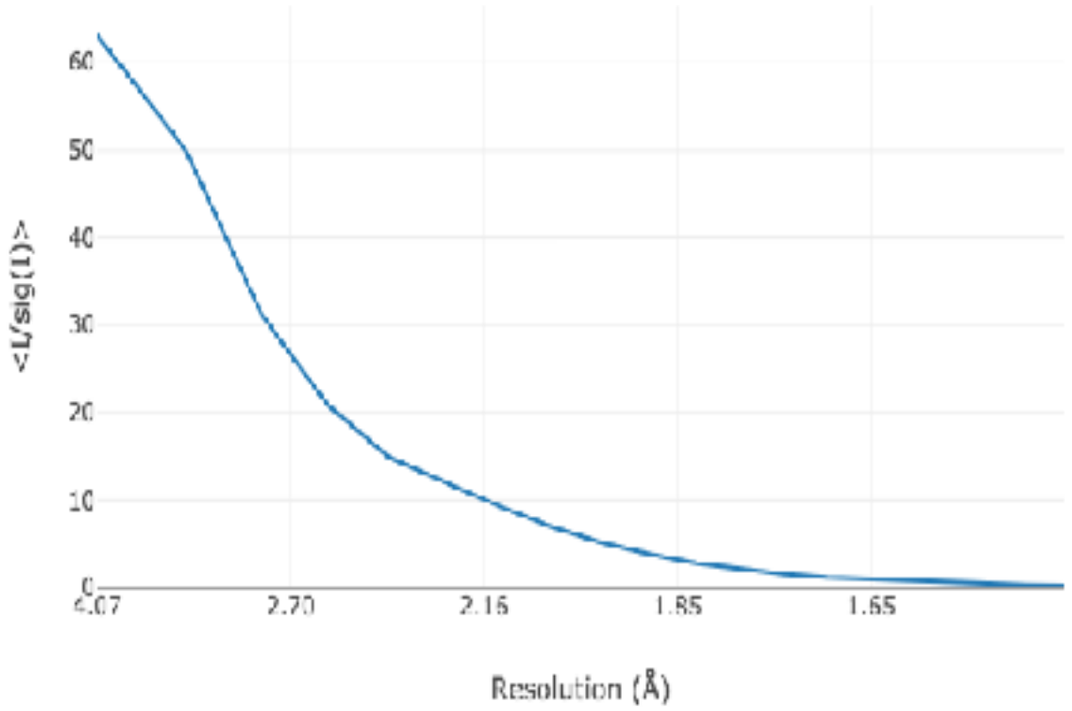
ASSESSING SCALING RESULTS

Resolution dependent plots

Normal and anomalous $CC_{1/2}$



Average I/σ(I)

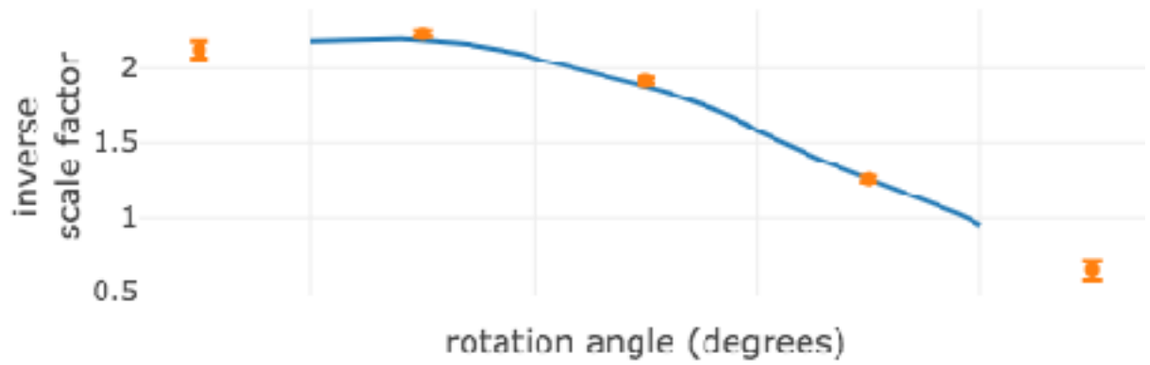


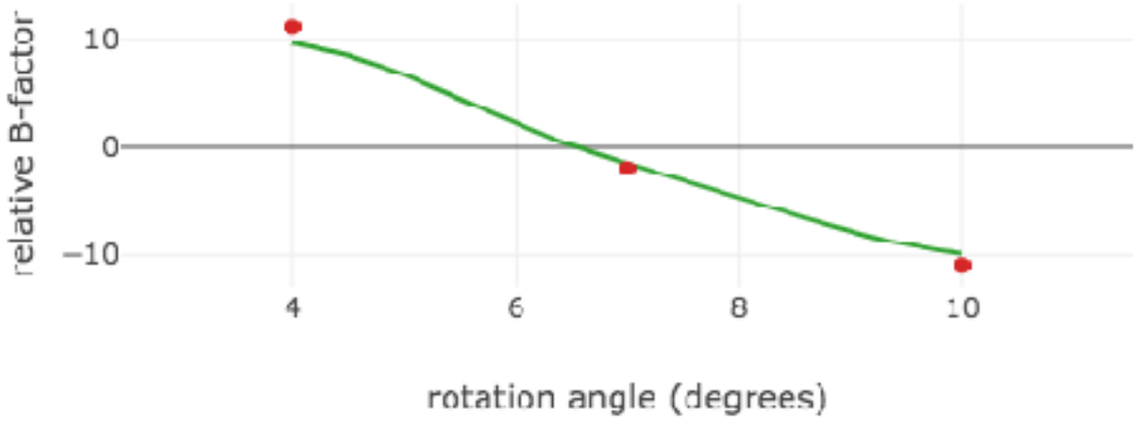
$CC_{1/2}$ is best metric for assessing resolution limits [1]. Clear interpretation of values & known statistical properties. Not dependent on σ(I), which may be inaccurate.

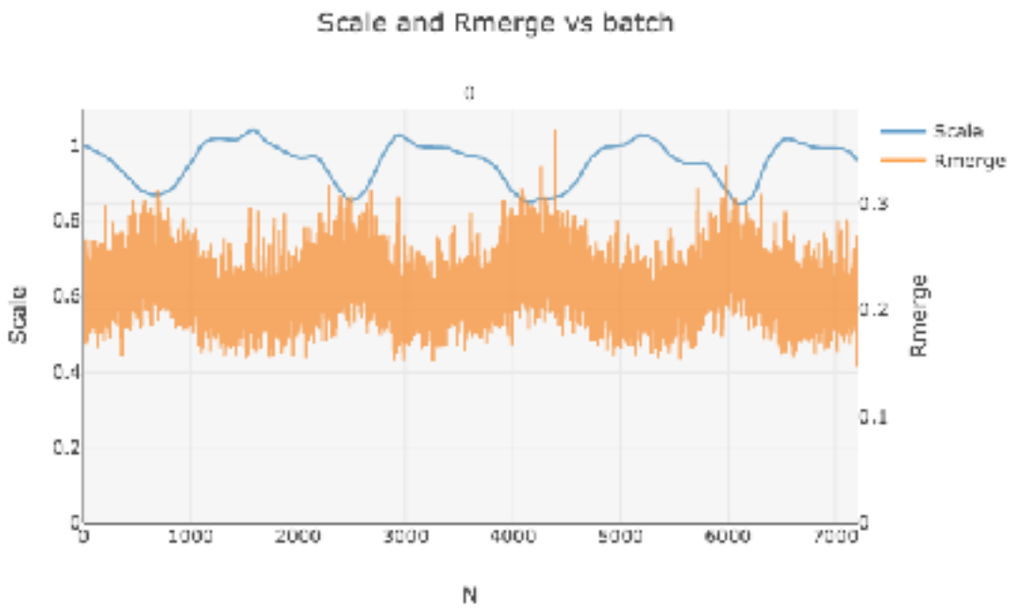
[1] Assessing and maximising data quality in macromolecular crystallography, Karplus & Diederichs, 2015.

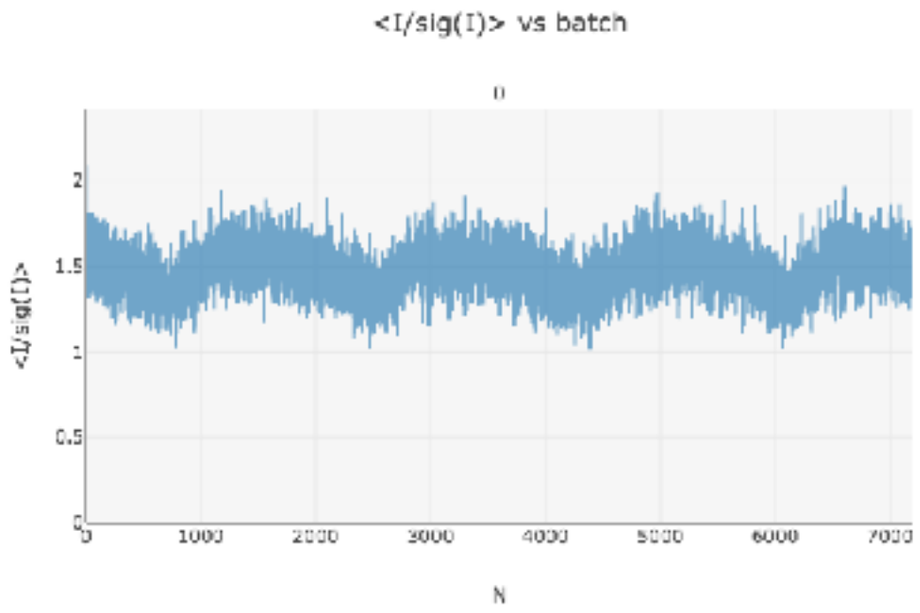
ASSESSING SCALING RESULTS

Plots as a function of image number/rotation



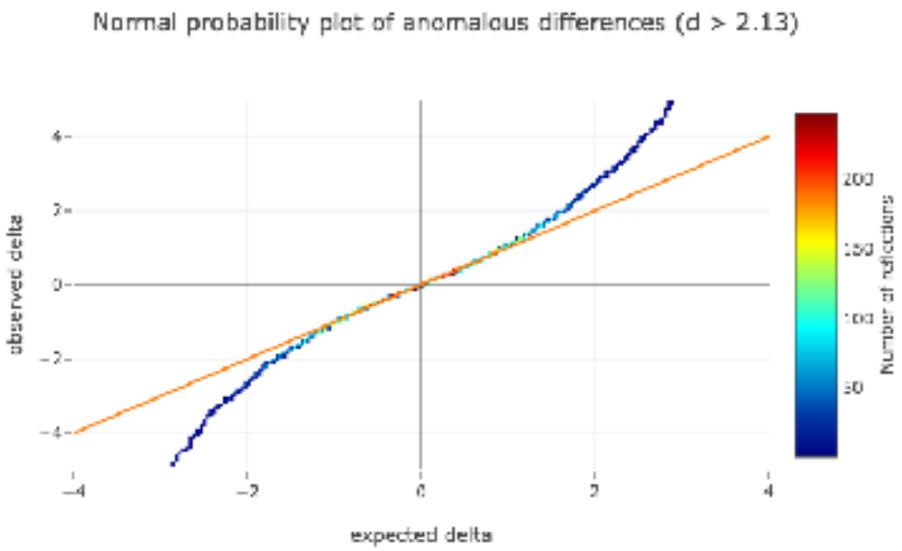


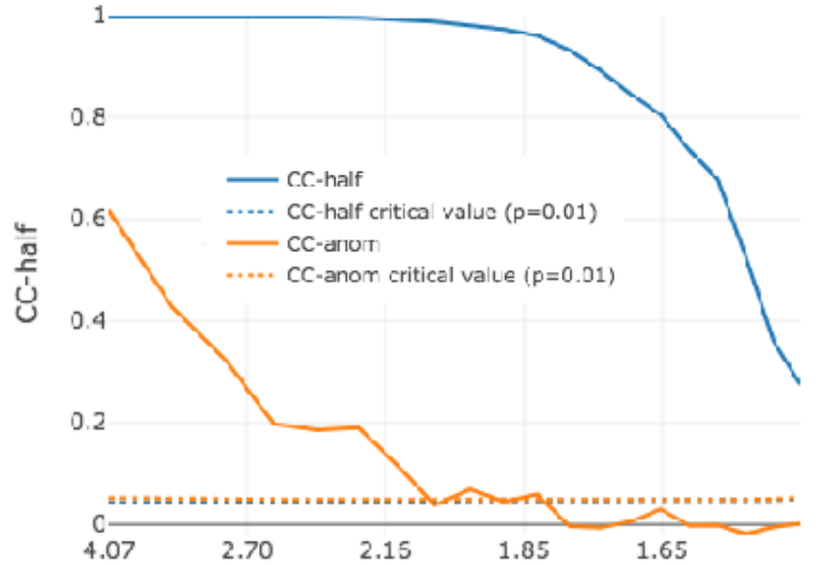




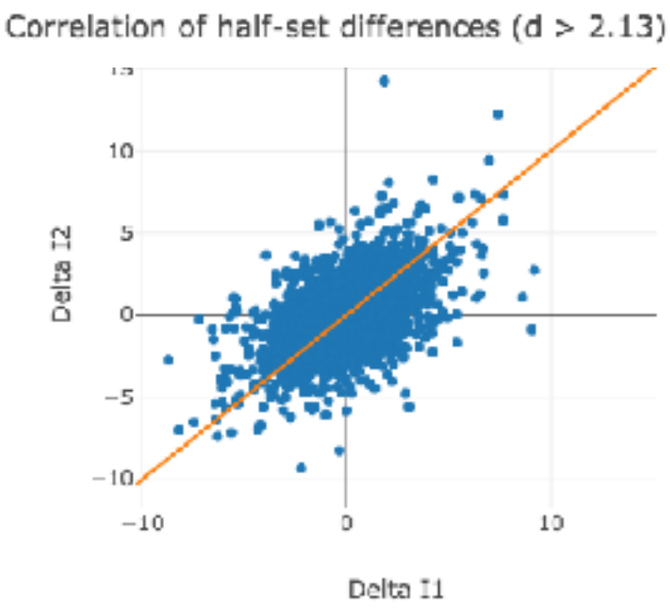
ASSESSING SCALING RESULTS

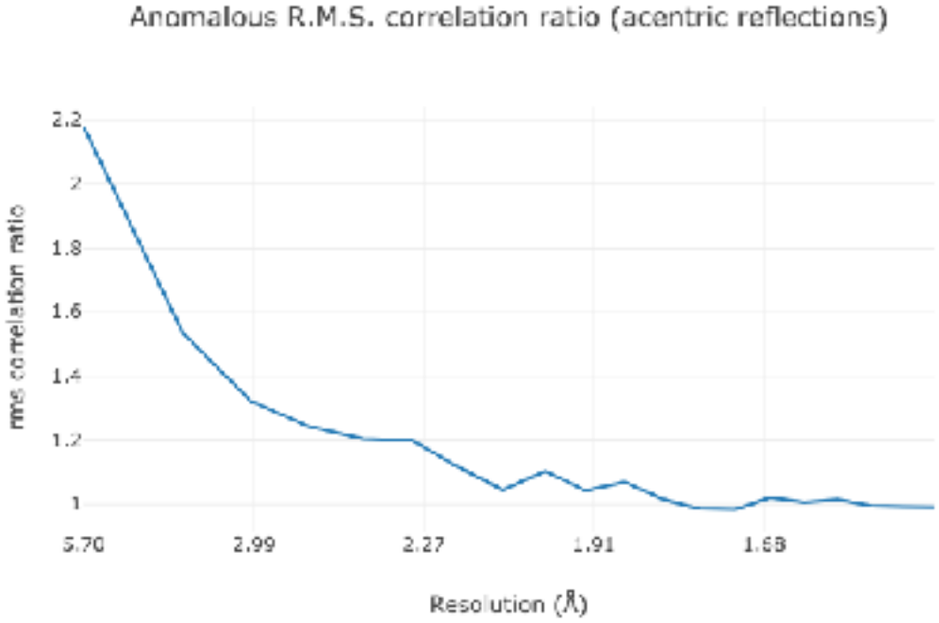
Anomalous plots





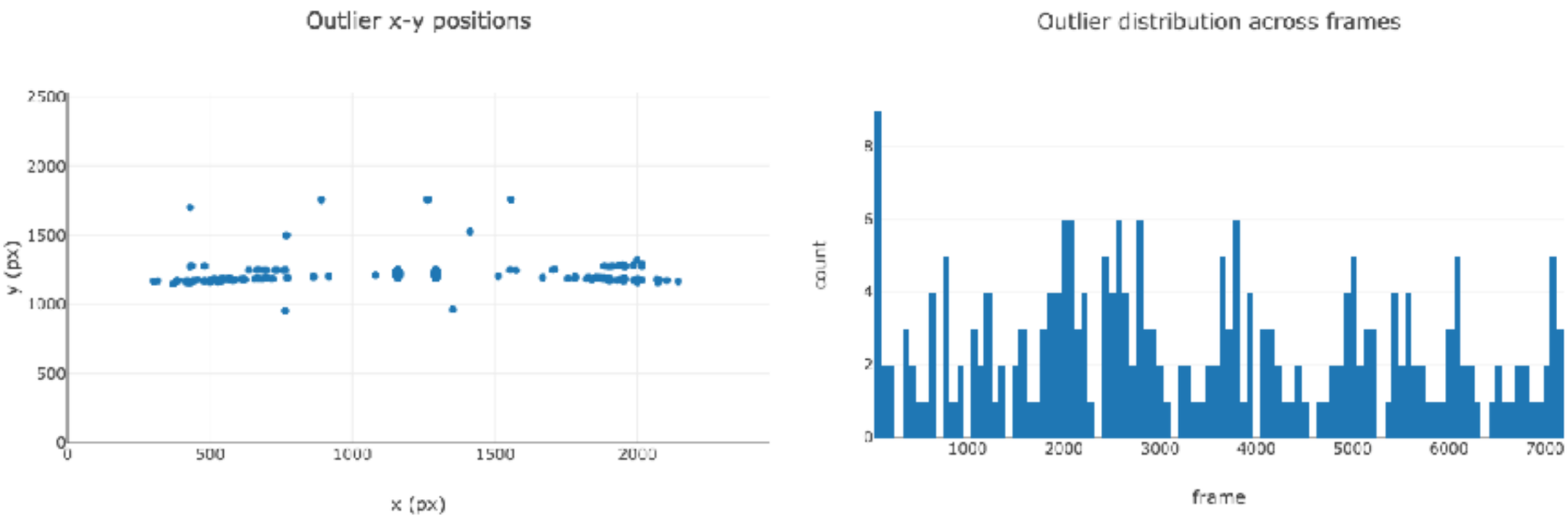
$$\delta_{\mathrm{anom}} = (I^{+} - I^{-})/(\sigma^{2}(I^{+}) + \sigma^{2}(I^{-}))^{\frac{1}{2}}$$





ASSESSING SCALING RESULTS

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 < ~0.05?)
- How many outliers were rejected, too many?

```
Performing a round of error model refinement.

Error model details:
  Type: basic
  Current parameters: a = 0.91327, b = 0.03893
  Error model formula: σ'² = a²(σ² + (bI)²)
  estimated I/sigma asymptotic limit: 28.125
```

- Were profile/summation intensities combined?

```
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 = 2003.81 | 0.99959 | 0.21789 |
| Imid = 200.38  | 0.9996  | 0.21773 |
| Imid = 20.04   | 0.99956 | 0.22256 |
| Imid = 2.0     | 0.99946 | 0.25967 |
-------------------------------------
Combined intensities with Imid = 200.380803531 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).

Some structure determination programs require structure factor amplitudes:

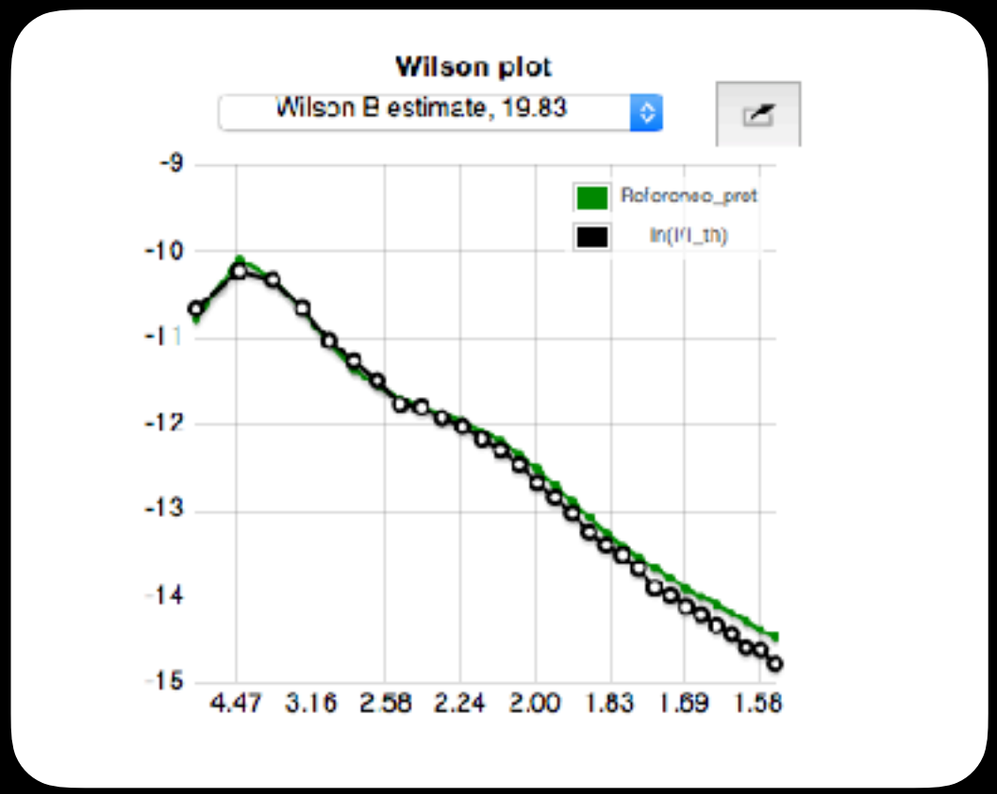
$$|F| = \sqrt{I}$$

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 σ(I)).

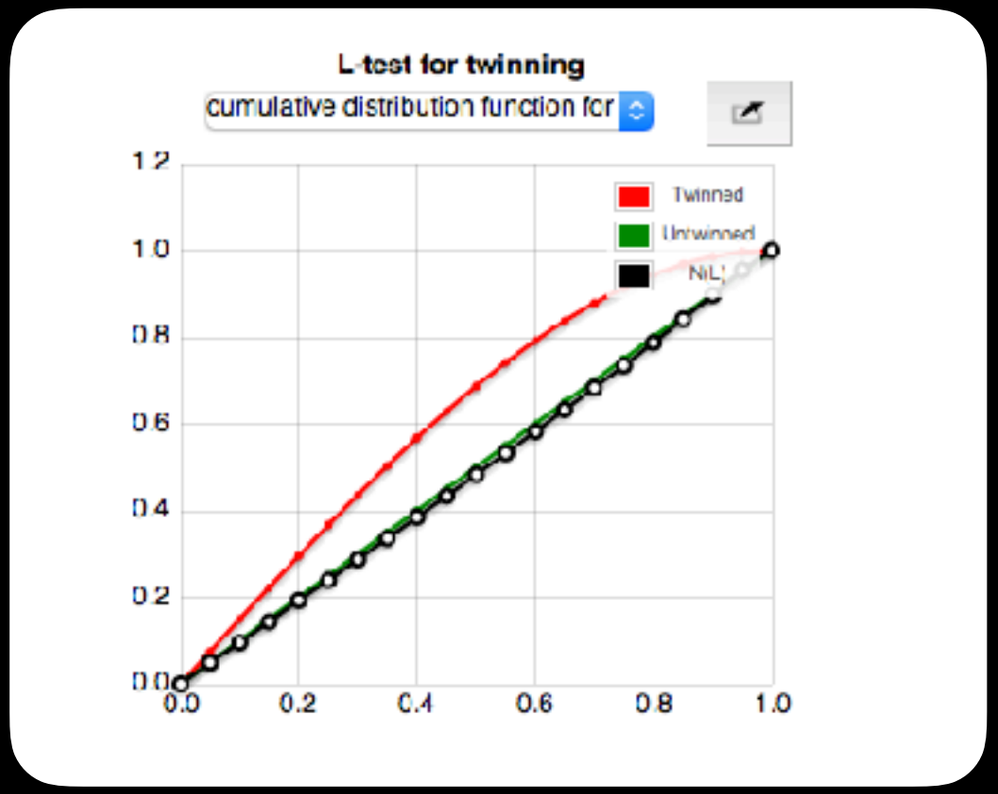
Program can use: dials.merge, ctruncate (ccp4),

INTENSITY STATISTICS



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

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 Fs*

Integrated reflections

↓

Integrated reflections, in correct point group

↓

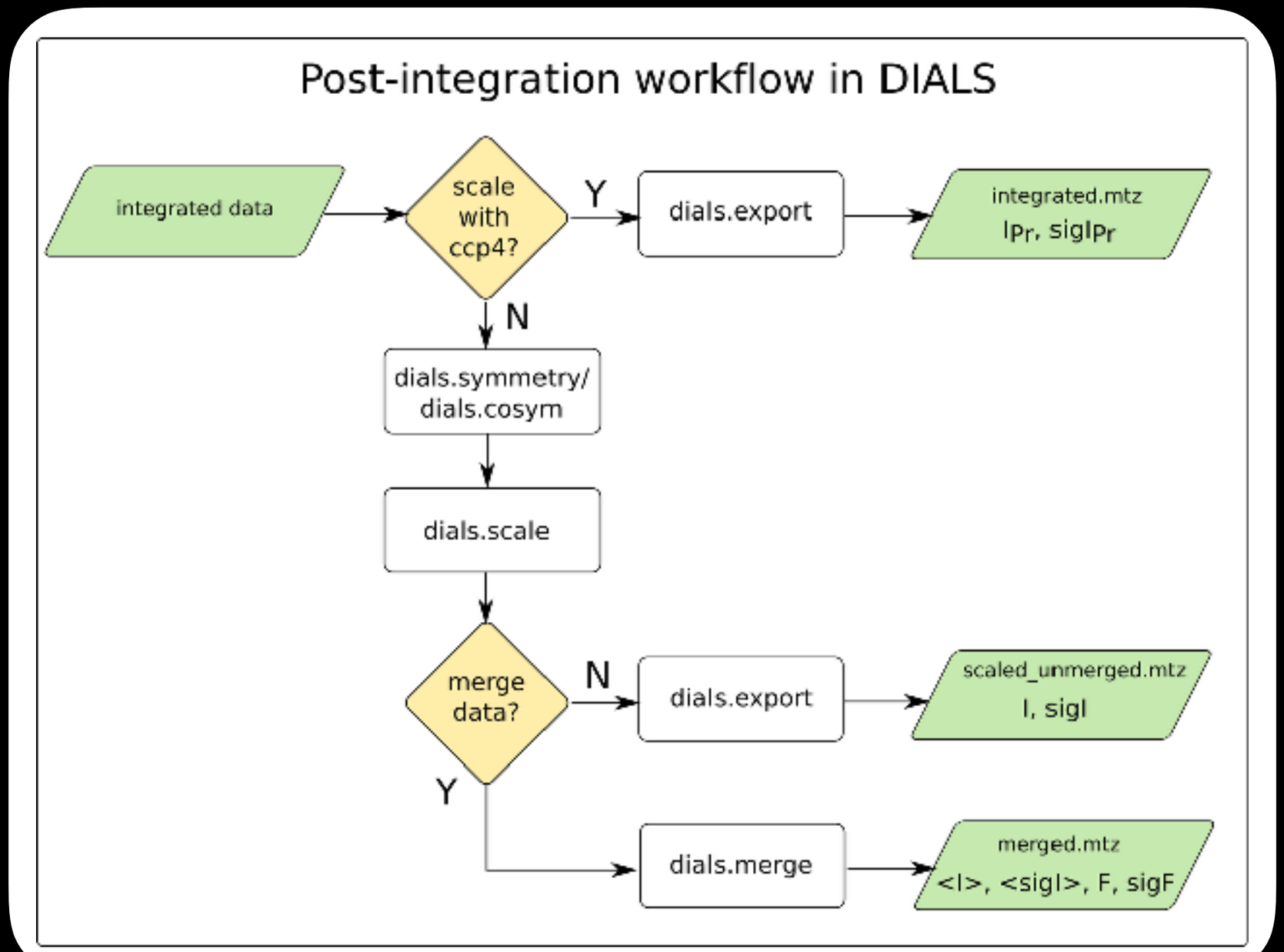
Scaled (corrected) intensities and sigmas

↓

Merged (symmetry-unique) set of I/F

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

DIALS DATA REDUCTION WORKFLOW



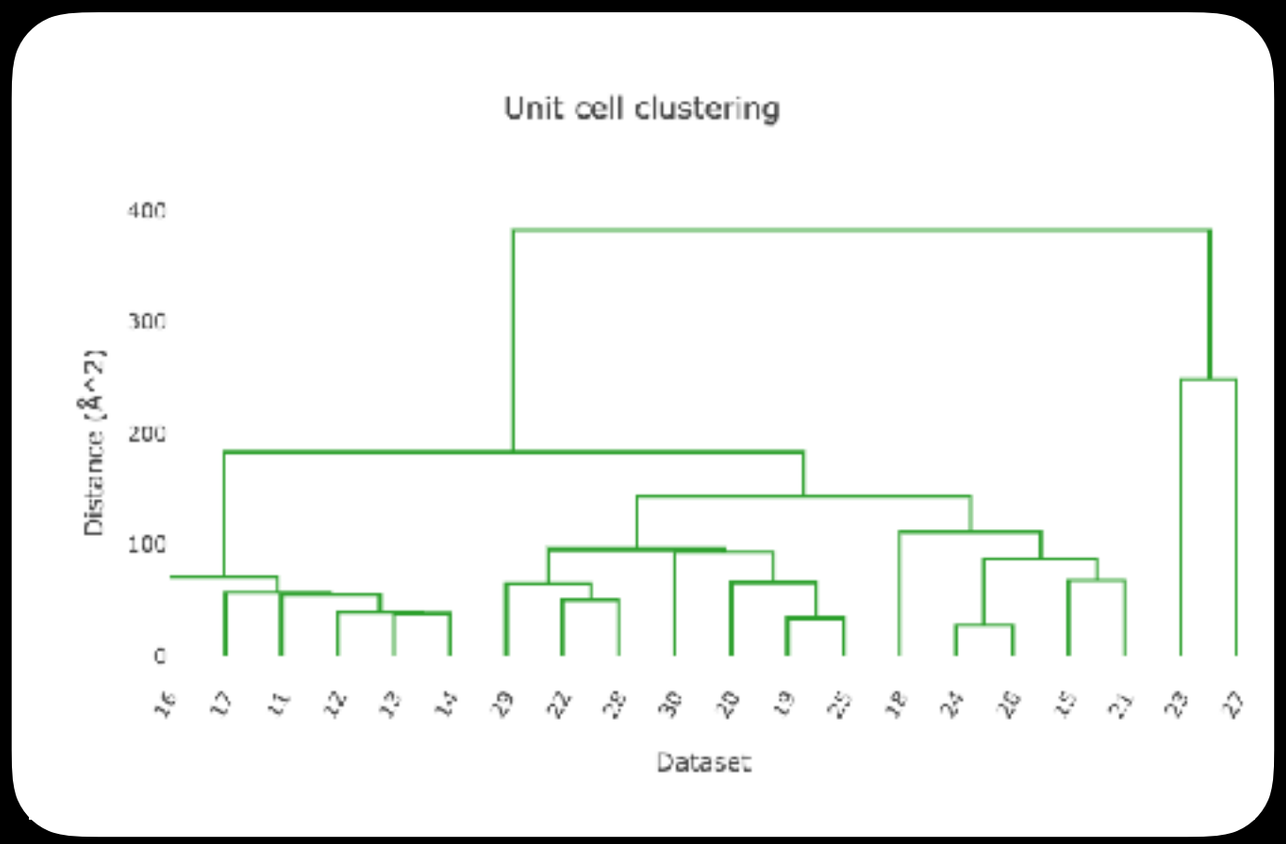
CCP4 equivalent: Pointless ⟶ Aimless ⟶ Ctruncate

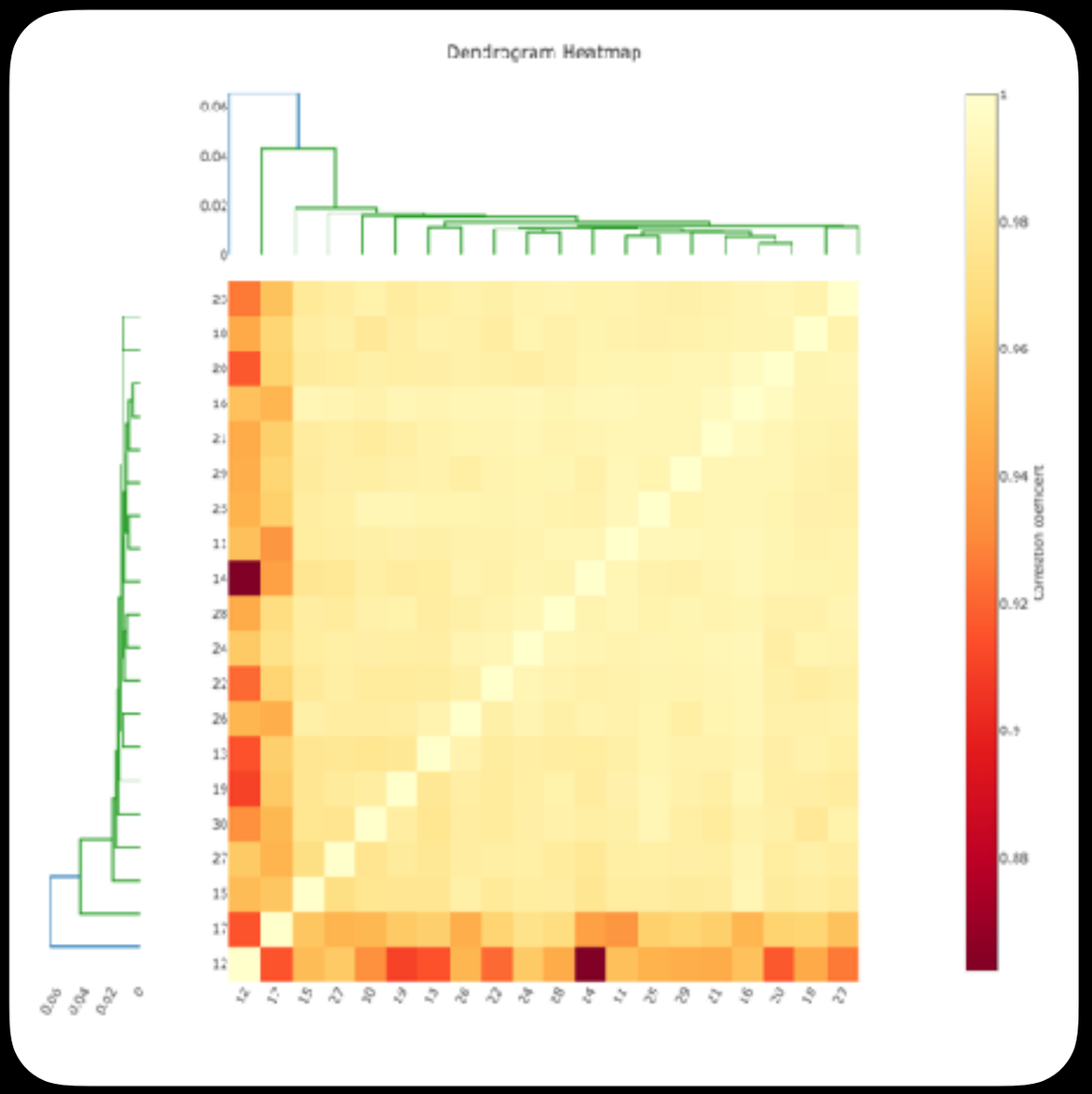
MULTI-CRYSTAL DATA REDUCTION

When scaling & merging multi-crystal data (a large number of small-rotation measurements), need to decide which subset of the data to keep due to nonisomorphism. In addition, must resolve indexing ambiguities.

Clustering/dataset rejection:

- unit cell clustering
- delta cc1/2 clustering

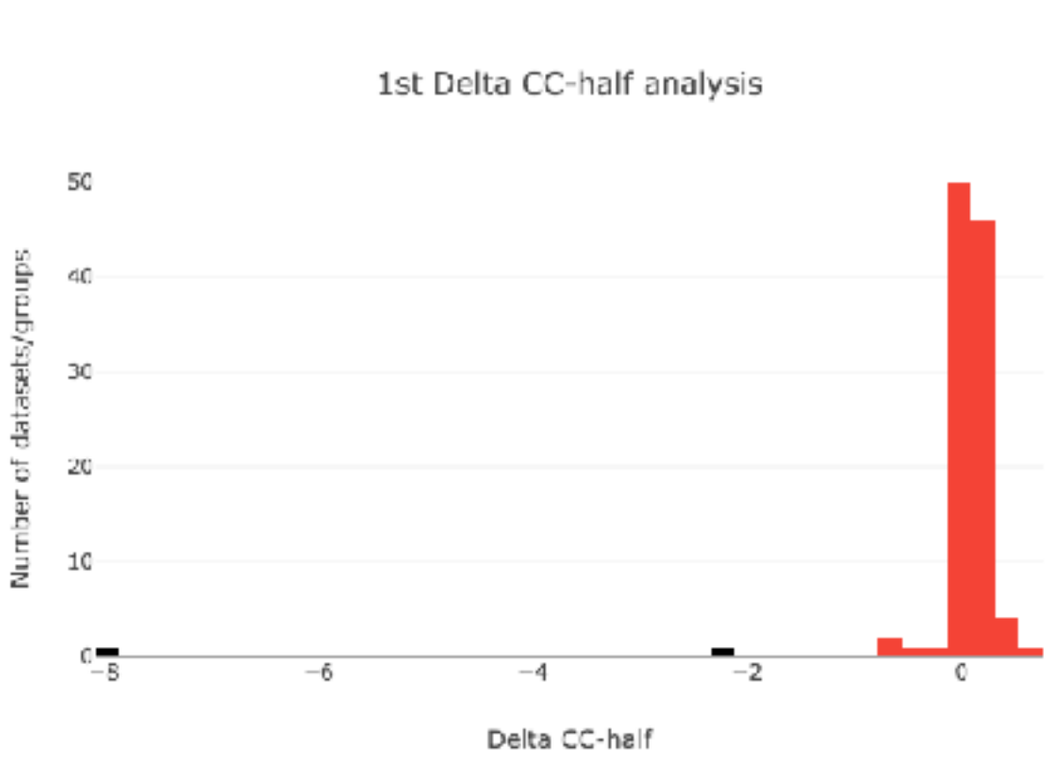


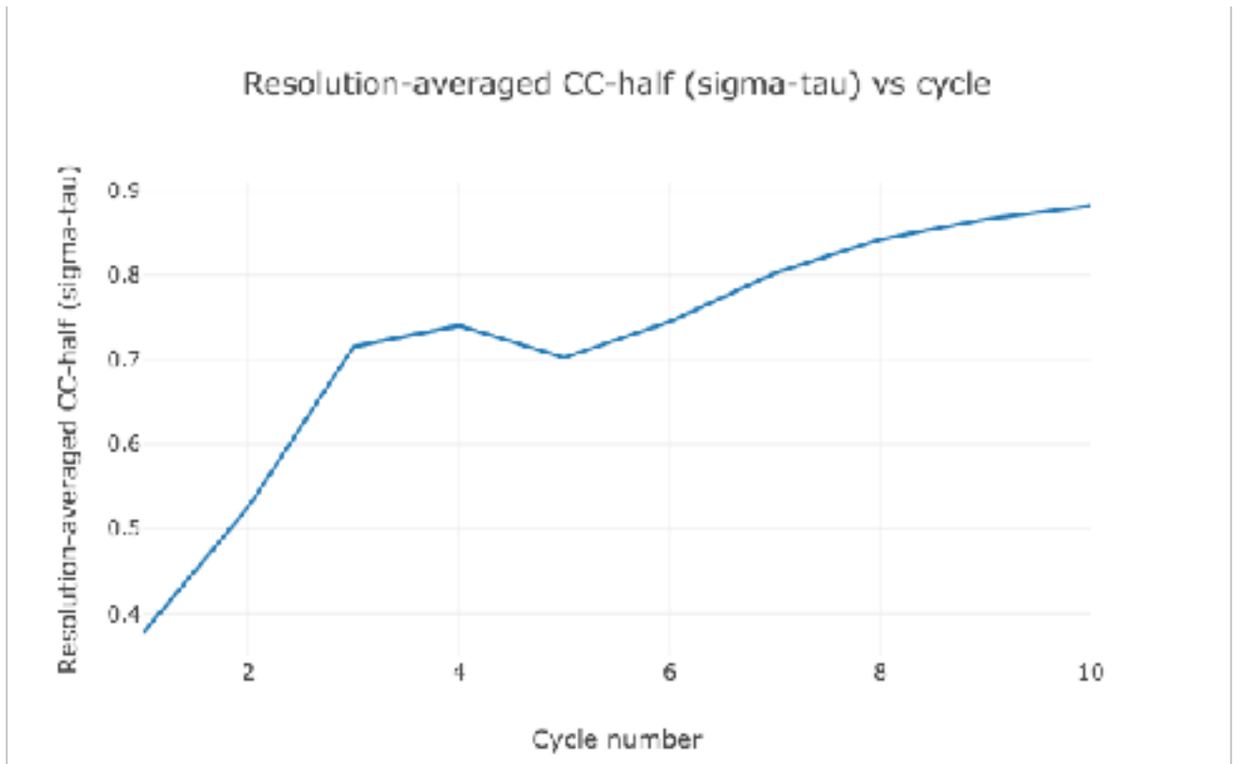


MULTI-CRYSTAL DATA REDUCTION

Within DIALS: dials.cosym, dials.cluster_unit_cell, dials.compute_delta_cchalf

dials.scale filtering.method=deltacchalf. Successive cycles of scaling and filtering based on delta cchalf

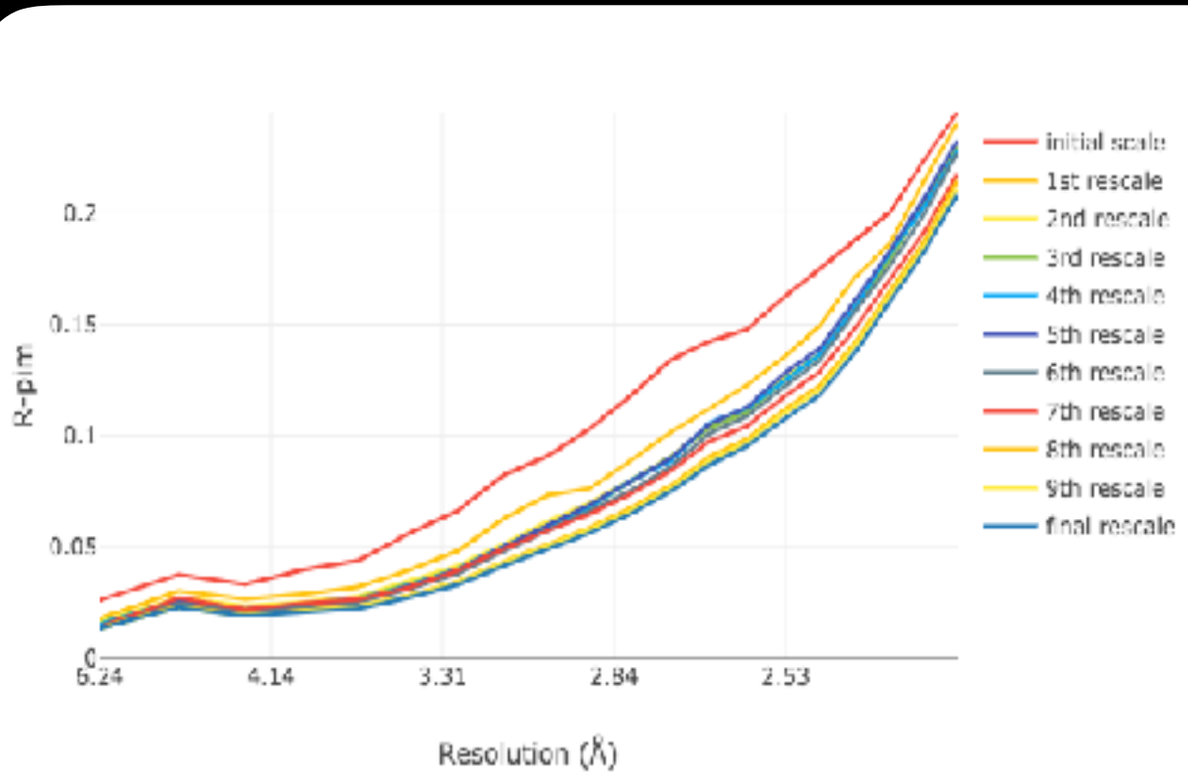


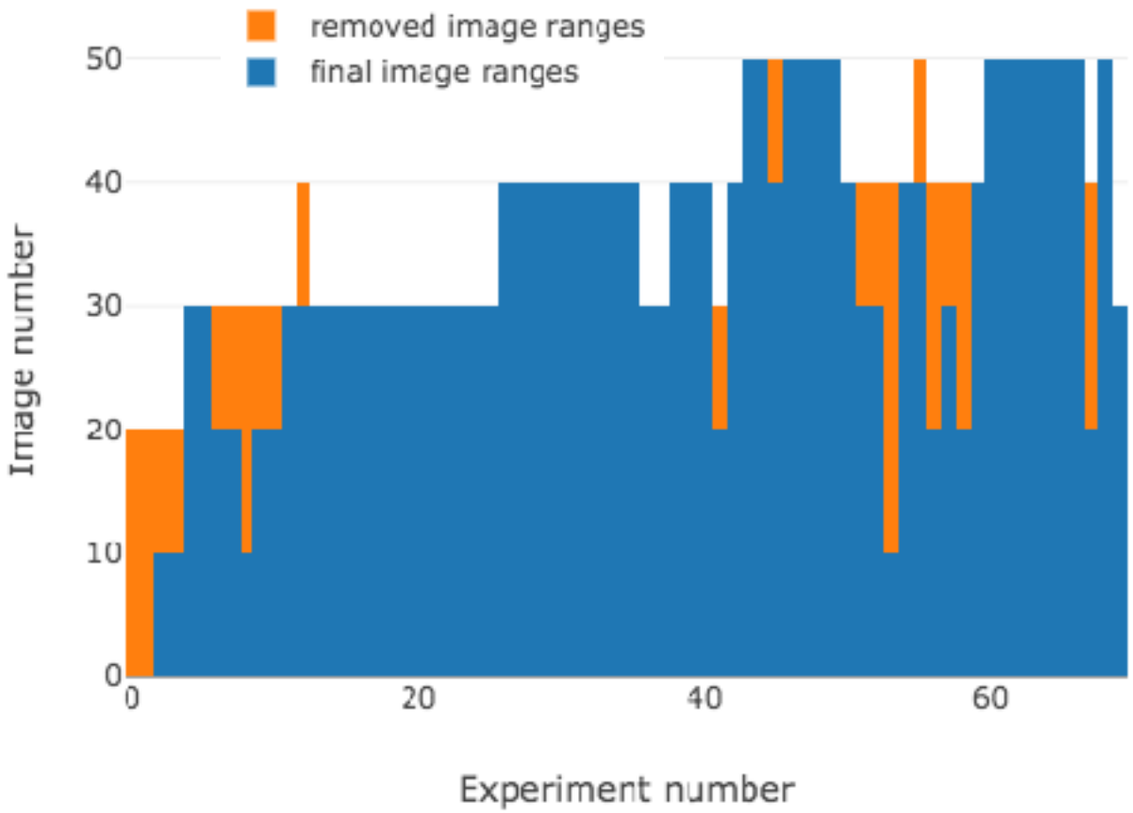


dials.scale.html, ‘Scaling and filtering plots’

MULTI-CRYSTAL DATA REDUCTION

dials.scale filtering.method=deltacchalf. Successive cycles of scaling and filtering based on delta cchalf





dials.scale.html, ‘Scaling and filtering plots’

MULTI-CRYSTAL DATA REDUCTION

xia2.multiplex

Combines DIALS data reduction tools into automatic processing workflow for multi-crystal data, starting with dials-integrated data.

Runs dials.cosym, dials.scale, clustering analysis, on all data.
Can trigger additional scaling and filtering to give improved subsets of the data.
Produces extensive html report.

Run as part of autoprocessing at Diamond for appropriate datasets.

Alternative programs as part of XDS package implementing similar algorithms:
XDSCC12, XSCALE_ISOCLUSTER

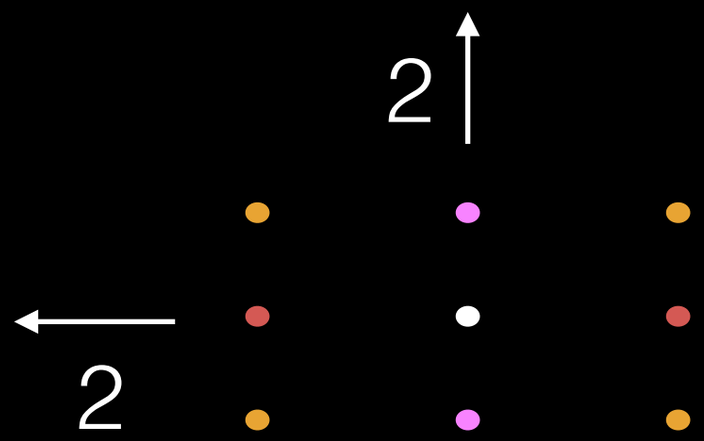
SUMMARY

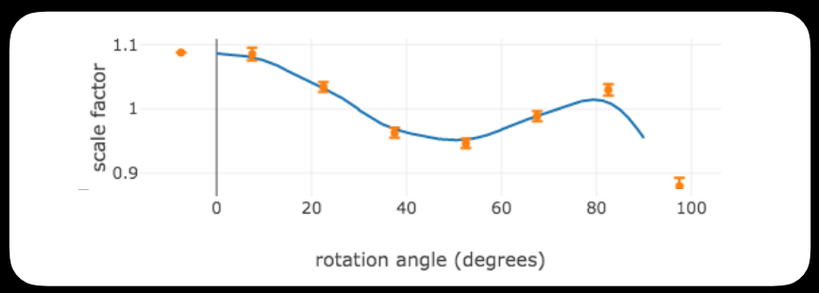
- Data reduction process of scaling and merging integrated intensities.

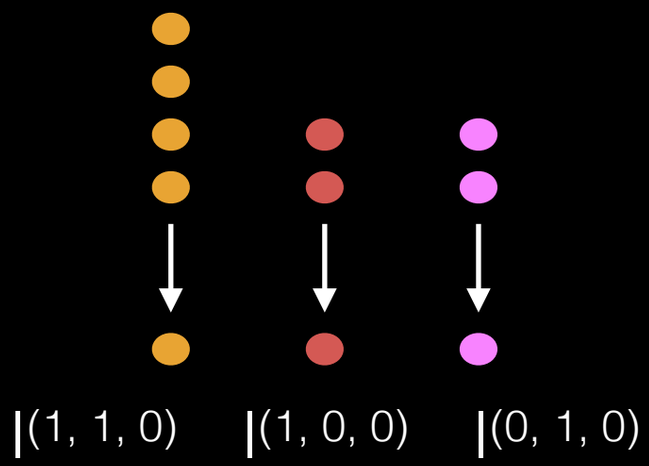
$$I_1^{(1,1,1)}, \sigma_1^{(1,1,1)}, I_2^{(1,1,1)}, \sigma_2^{(1,1,1)}, \ldots \rightarrow \mathbf{I}^{(1,1,1)}, \boldsymbol{\sigma}^{(1,1,1)}$$

- 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?







Acknowledgements

Dials project: Gwyndaf Evans, Graeme Winter, David Waterman. Richard Gildea, Nicholas Devenish, Luis Fuentes-Montero, Markus Gerstel, Ben Williams, James Parkhurst, Melanie Vollmar.

Useful discussions: Phil Evans, Kay Diederichs

Many DLS staff for datasets, feedback.

Thank you for your attention!