MRC Laboratory of Molecular Biology

Data processing – integration & scaling

Harry Powell
MRC Laboratory of Molecular Biology
8th February 2010 – ESRF Grenoble

Overview – processing diffraction data

May be divided into stages:

- Integration
- Applying corrections
- Scaling and merging
  - merging partials to form complete reflections
  - merging symmetry equivalents
- Truncation (converting $|F|^2$ to $|F|$)
- Analysis

Preamble – rationale for the experiment

What are we doing, and why are we doing it?

Measuring intensities of spots to obtain structure factor *amplitudes*

$$|F_{hkl}| = \left[\frac{KI_{hkl}}{Lp}\right]^{1/2} \qquad (1)$$

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l F_{hkl}\, e^{-2\pi i(hx+ky+lz)} \qquad (2)$$

Careful data collection and careful measurement of intensities can be used to recover the phases which are otherwise lost

Optimization of Data Collection

Pre-process at least one image *before starting the full data collection* (preferably two at 90º to each other) to obtain:

- Cell parameters, crystal orientation and putative Laue group
- Estimate of mosaicity
- Effective resolution limit }
- Optimal crystal to detector distance } *e.g.* use *BEST*
- Exposure time }
- Strategy for data collection }

Remember! This is the last experimental stage - if you collect bad data now you are stuck with it. No data processing program can rescue the irredeemable!

Don't necessarily do what your PI or post-doc suggests – think! At Diamond or ESRF use DNA/Edna

First things first - load the images and look at them

Questions:

- are there any spots on the image?
- has the detector been used efficiently?
- do the spots look reasonable – split? large? above background?
- can you see separate lunes?
- is there a single lattice?
- should I throw the crystal away now and collect a dataset on another crystal instead?

Check two images at 90° to each other – some pathologies are not apparent from a single image.

Before starting to process

- Use the program tools to mask backstop, cryostream, other shadows.
- Set resolution limit to about 0.2Å higher than visible spots.
- Make sure beam position is more-or-less correct.
- Make sure other parameters (distance, wavelength, oscillation angle) are what you expect (do they correspond to what is in your notebook?).

Data Processing

We want to measure the intensities of the diffraction spots, so we need to know where they are (on each image and on which images) and how big they are; "what is spot and what is not".

Steps

- spot finding - what the spots on the current images are like
- **indexing** – roughly where they are on the image
- mosaic spread – how many images the spots are spread over
- **refinement** – accurately locate the spots
- **integration** – actually measure the intensities
- analyse the results – how well did we do the above?

Indexing

- Find spots on the image
- Convert 2D co-ordinates (image) to scattering vectors (corresponding to 3D RL co-ordinates)
- Index
- Cell reduction
- Apply Bravais lattice symmetry
- Pick a putative solution
- (Estimate mosaic spread)

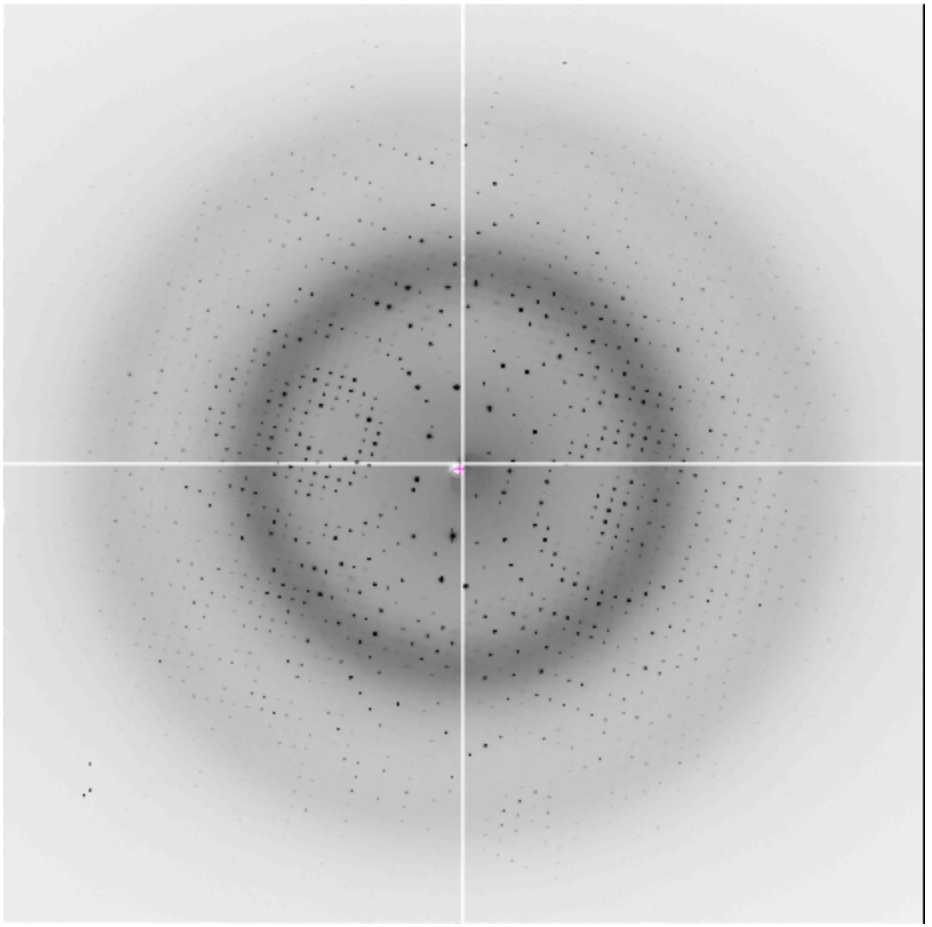
Note that indexing only gives an approximate solution; we *hope* it will be good enough to proceed.

Indexing

After locating the spots on an image, we can convert the 2D image co-ordinates to scattering vectors that correspond to lattice points in a (distorted) 3D reciprocal lattice by means of the relationships

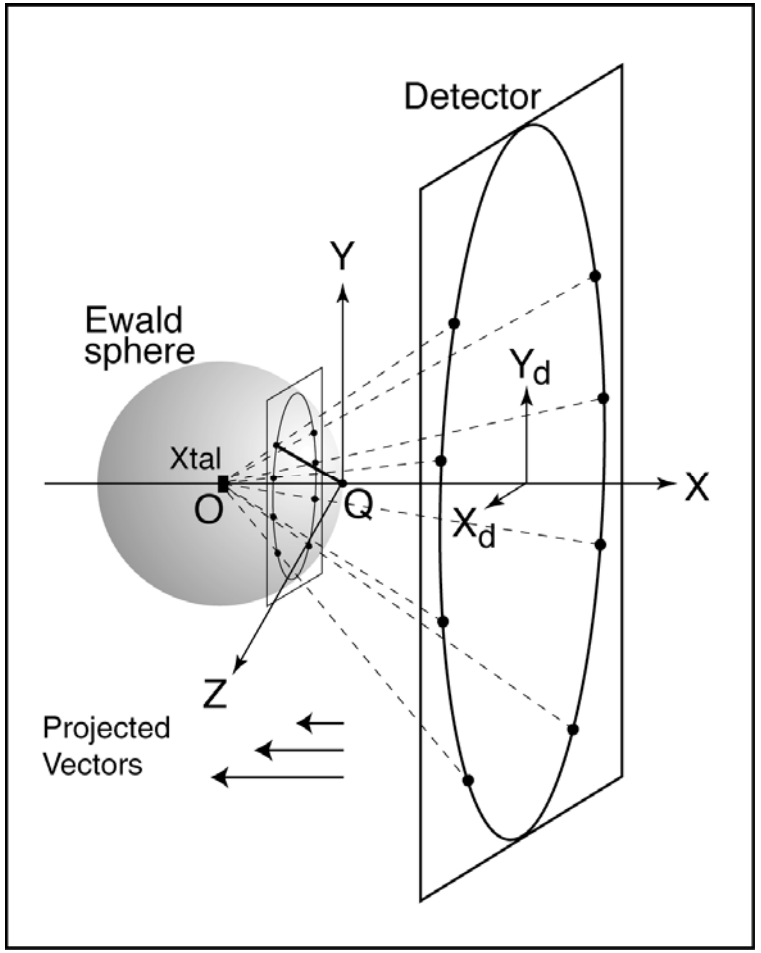
$$s = \begin{bmatrix} D/r - 1 \\ X_d/r \\ Y_d/r \end{bmatrix}$$

$$r = \sqrt{D^2 + X_d^2 + Y_d^2}$$



Indexing

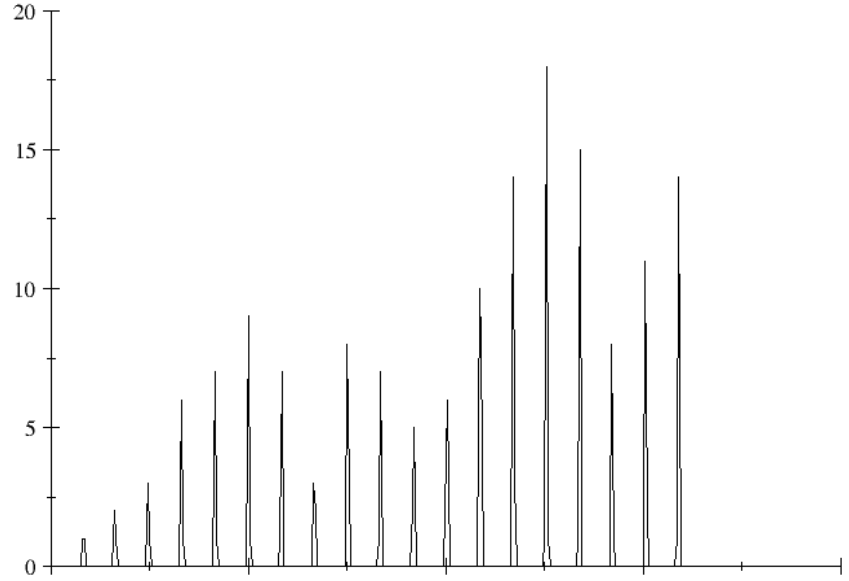
If the scattering vectors calculated from the 3D R.L. co-ordinates are projected along a real space axis direction (such as a, b or c) all the projected vectors for spots in the same reciprocal space plane will have the same length, as will all those spots in the next plane *etc*. This will give a large peak in the Fourier transform.



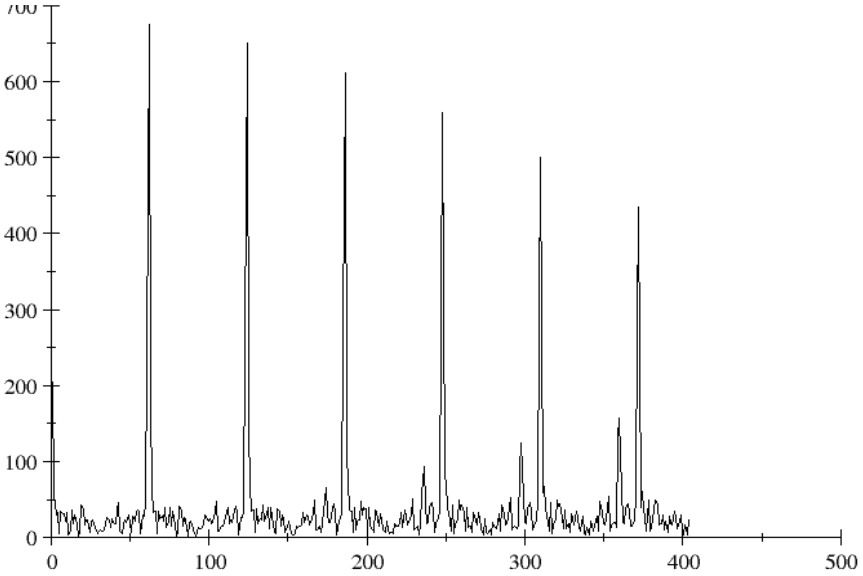
Indexing

The first large peak in the Fourier transform corresponds to a real space cell edge length. In this case, ~56Å.

Provided that a single image samples enough of reciprocal space, we can get information about all three crystal axes from one image.



scattering vectors projected along axis a



Fourier transform of scattering vectors

Bravais lattice identification - metric symmetry

Indexing gives us a basis solution that is triclinic.

Applying symmetry transformations (44 characteristic lattices) allows us to see how well this triclinic solution fits the cell edges and angles of lattices with higher symmetry, *e.g.* monoclinic, orthorhombic etc.

Mosflm and *XDS* give 44 solutions: each of these corresponds to one of the 14 Bravais Lattices (each of which may occur several times as a result of different transformations); *Denzo* and *HKL* only give the "best" 14 Bravais Lattice solutions.

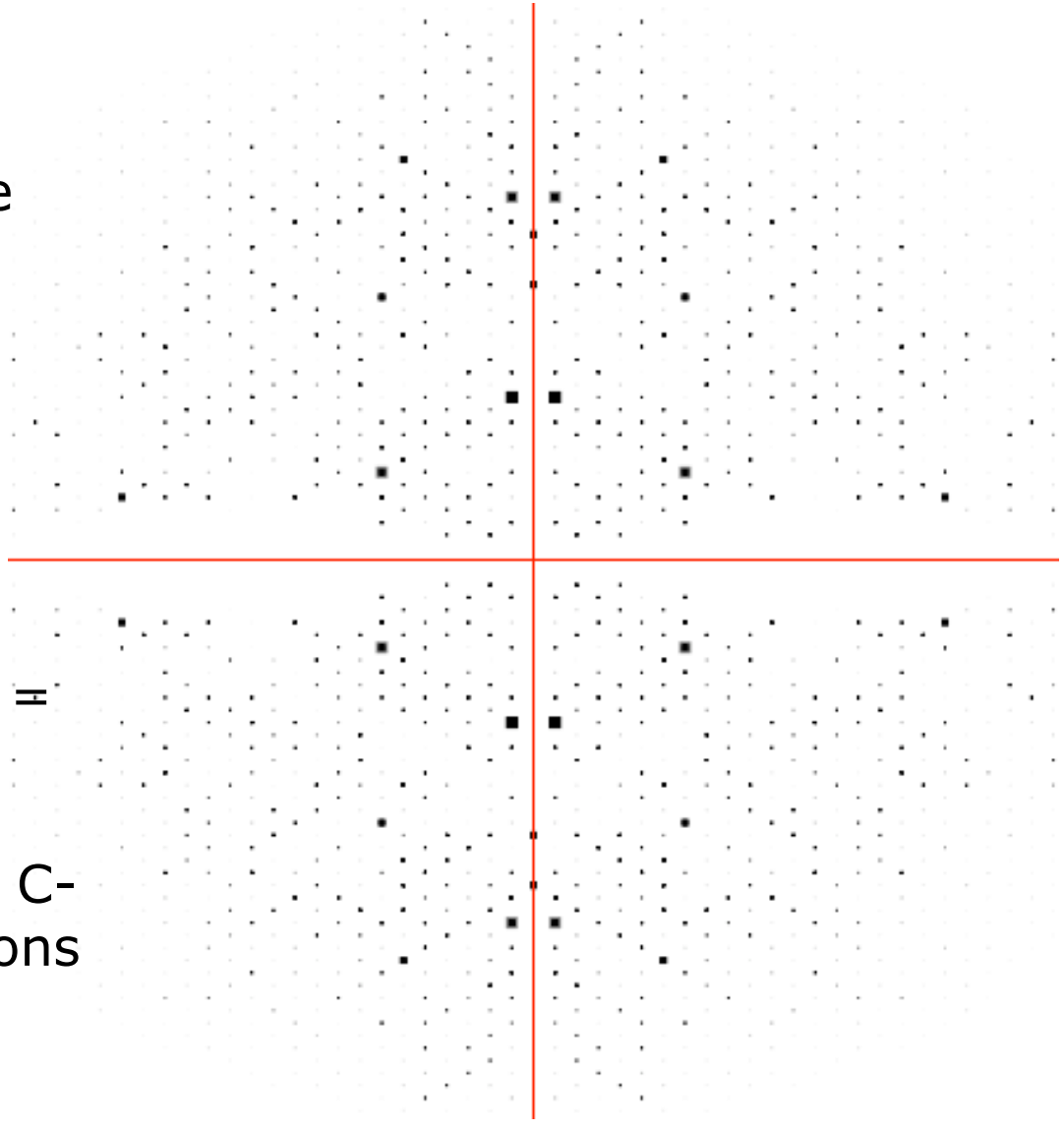
Metric symmetry may not be the correct crystal symmetry, but it usually is.

Bravais lattice identification – from intensities

The true Bravais Lattice symmetry can only be determined by analysing the intensities of symmetry equivalent reflections.

example of $C222_1$ with a = 74.7Å, b = 129.2Å, c = 184.3Å, which could be (incorrectly) indexed as hexagonal a = b = 74.7Å, c = 184.3Å

There are also two incorrect C-centred orthorhombic solutions



Refine the parameters -

Optimise the fit of observed to predicted spot positions, so that the measurement boxes can be placed accurately over the spots.

Specifically, improve estimates of:

- Crystal parameters
- Instrument parameters

Can be performed by either (or both):

- Positional refinement using spot co-ordinates on detector
- Post-refinement using intensity measurements of partial reflections.

Positional refinement and post-refinement

Positional refinement

- uses the spot positions on each image, so it can be done for each image without reference to the others. Both fully and partially recorded reflections can be used.

Post-refinement

- needs intensity measurements for spots which are partial across at least two images; we cannot use fully recorded reflections for this
- □ needs at least two adjacent images (and probably more for fine-phi slicing, where the mosaic spread is more than twice the oscillation angle)

Positional refinement

Minimises

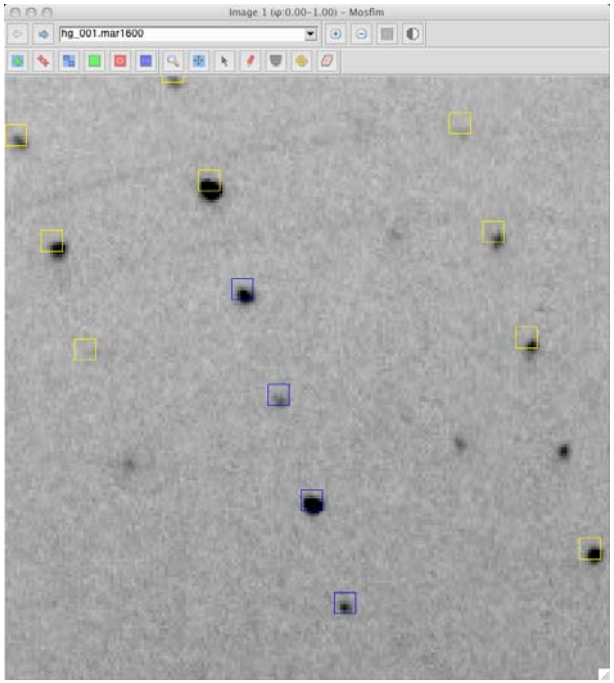
$$\Phi_1 = \sum_{i=1}^{n} w_{ix} \left[X_i^{calc} - X_i^{obs} \right]^2 + w_{iy} \left[Y_i^{calc} - Y_i^{obs} \right]^2$$

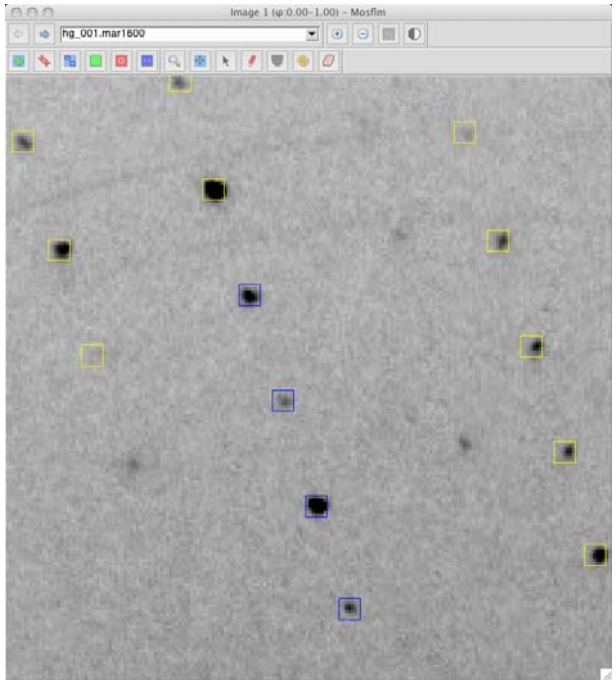
n.b. i) rotation of crystal about phi axis has no effect on this residual so it can't be refined.
ii) cell dimensions and other parameters (*e.g.* crystal to detector distance) may be strongly correlated.

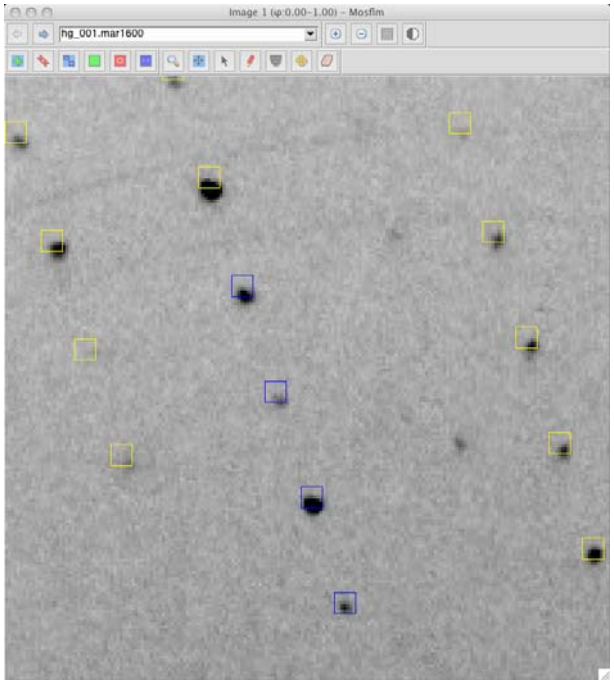
Can be used to refine unit cell dimensions, crystal to detector distance, Y scale, 2 of 3 crystal mis-setting angles, detector mis-setting angles & direct beam position.

Positional refinement

Can be used to refine crystal & detector parameters, but some are strongly correlated especially if using only low resolution data, *e.g.* cell dimensions and crystal to detector distance.







Post-refinement or the "phi-centroid" method

Minimises the *angular* residual δ (see later) *via*

$$\chi^2 = \sum_{i=1}^{n} w_i \left[\frac{R_i^{calc} - R_i^{obs}}{d_i} \right]^2$$

n.b. we need:

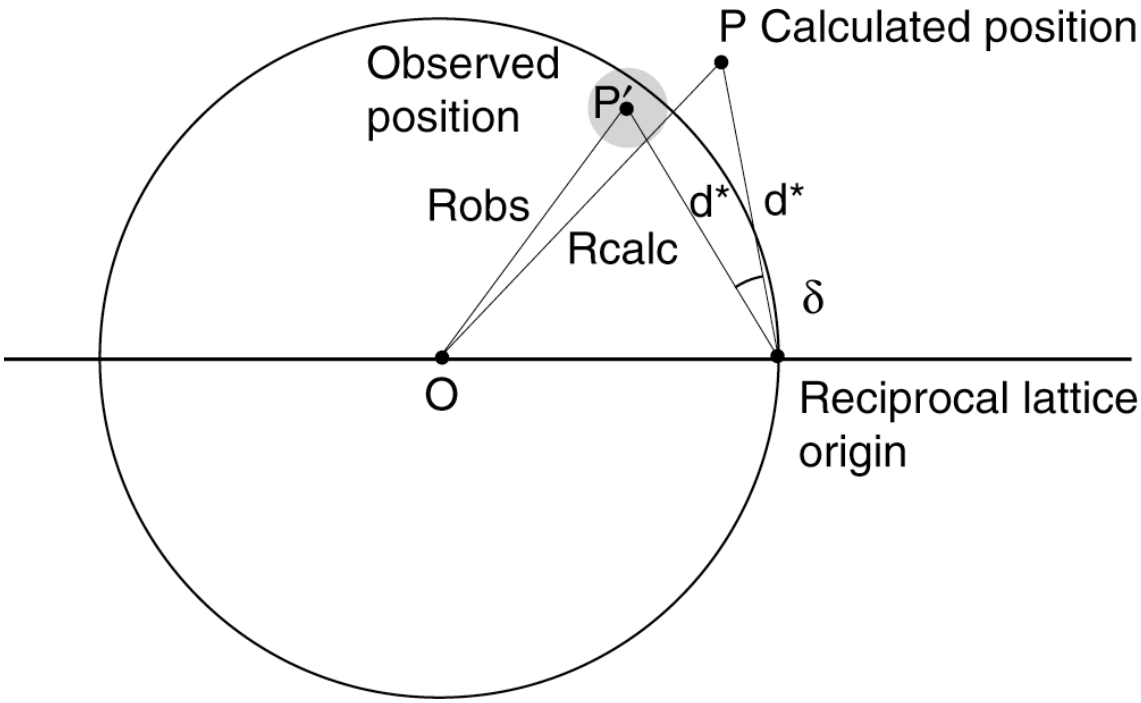
i) a reasonable knowledge of intensities for this, so it can only be done *after* integration - hence "*post-refinement*"

ii) a model for the "rocking curve"

We can refine the unit cell dimensions, 2 of 3 crystal mis-setting angles, and *either* the mosaicity *or* the beam divergence.

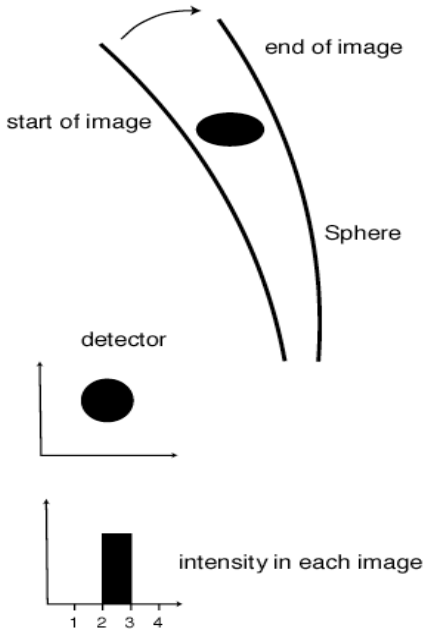
Post-refinement

We can visualise this in the Ewald sphere construction, minimising the angular residual δ. A suitable model for the rocking curve allows us to determine the "observed" position (P').

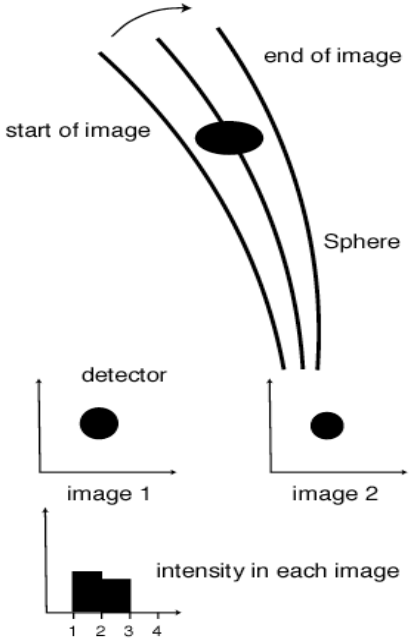


Ewald Sphere

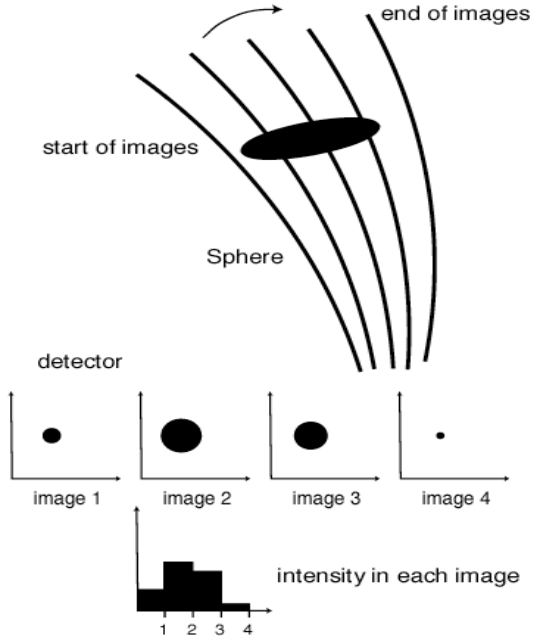
Fully recorded and partially recorded reflections



A fully-recorded spot is entirely recorded on one image



Partials are recorded on two or more images



"Fine-sliced" data has spots sampled in 3-dimensions

illustrations from Elspeth Garman

Integration itself

two basic ways -

(1) summation integration
simple, fast, okay for all except weak, overloaded or partially overlapping reflections

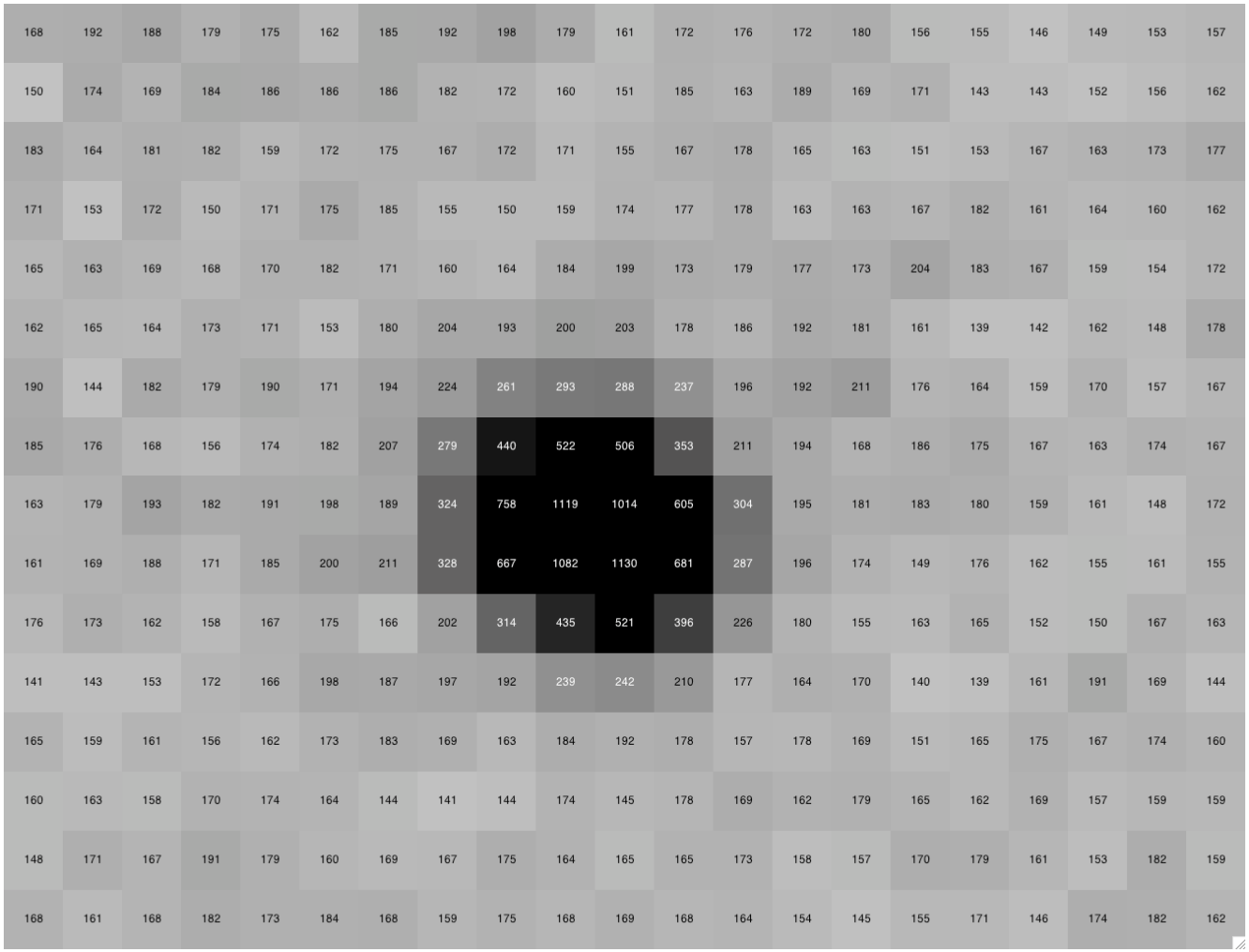
(2) profile fitting (only *intended* to improve weak spots)

can be sub-divided into

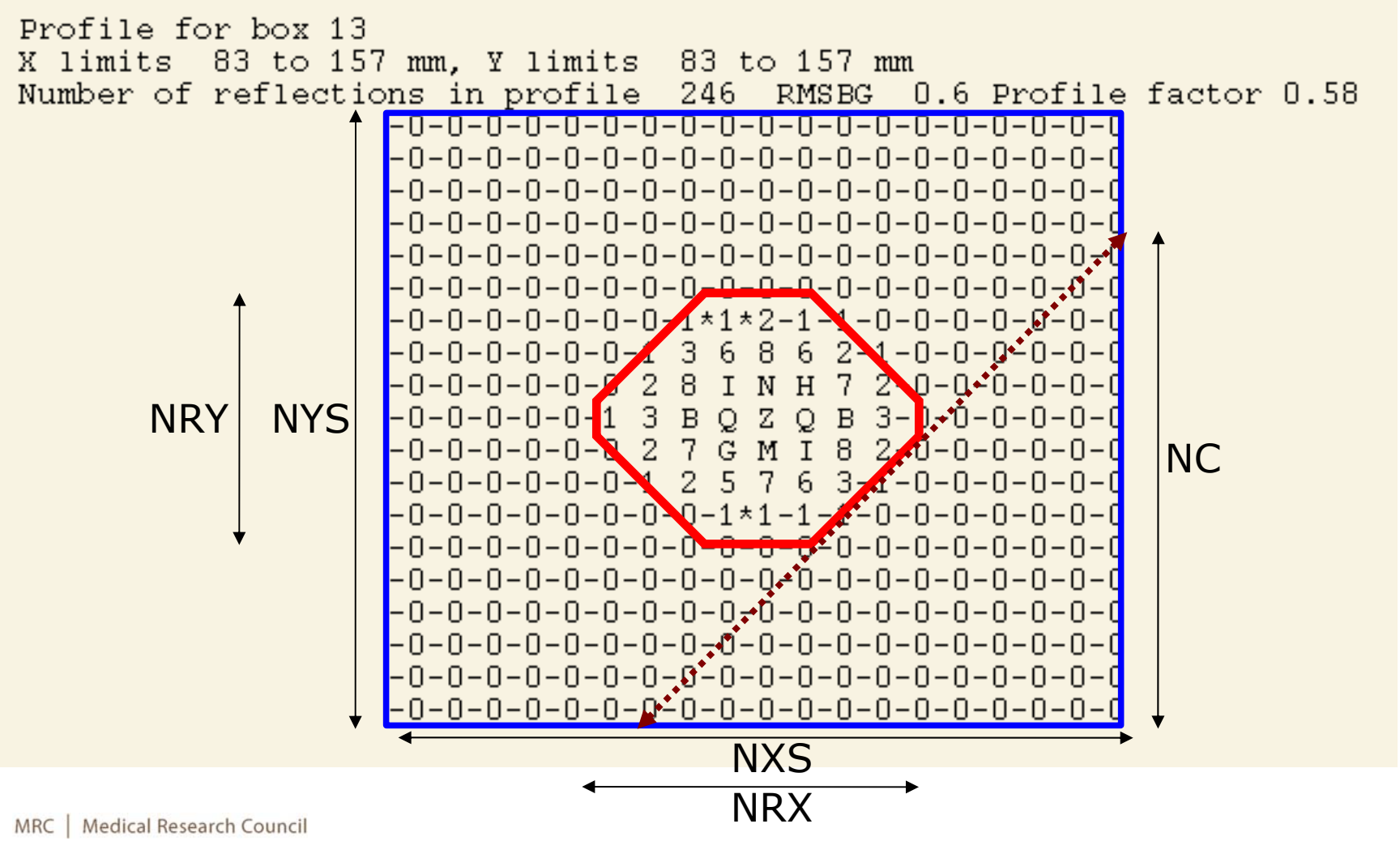
- two-dimensional (2D) – builds up profiles using information from single images (but we can use several images)
- three-dimensional (3D) – builds up profiles across several adjacent images

Measuring the intensity of a spot

Identify the background & spot regions, work out what the background level is around the spot, then *assume* it is the same under the spot.



Defining the spot and background region

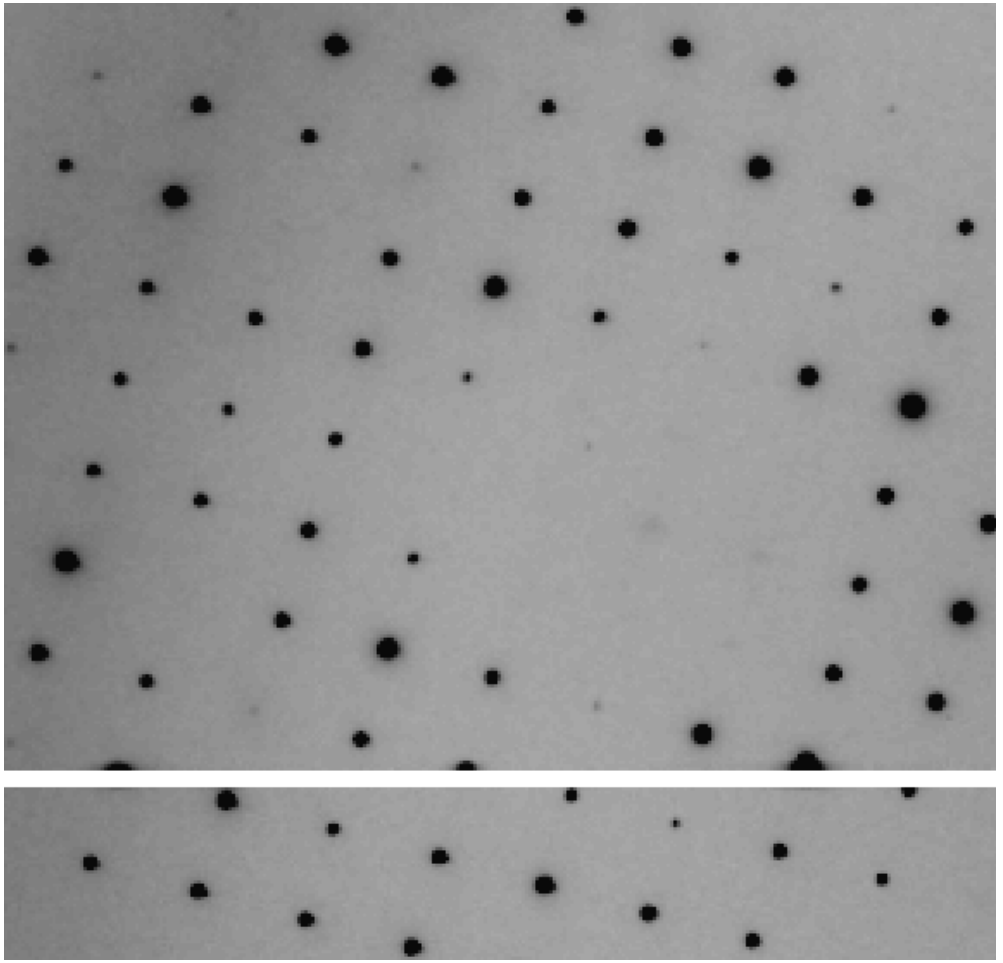


Summation integration

- In the absence of background, just add the pixel counts in the spot region together - but there is always background!
- Need to define spot and background regions - we cannot measure background directly under the spots, so we calculate a local background plane and slope from nearby non-spot pixels
- Use this to subtract the background under the spots
- Weak spots may have their shoulders under the background, so that their measurement is impaired.

Profile fitting integration - background

Based on the assumption that spots corresponding to fully recorded reflections in the same region of the detector (and on images nearby in phi) have similar profiles.



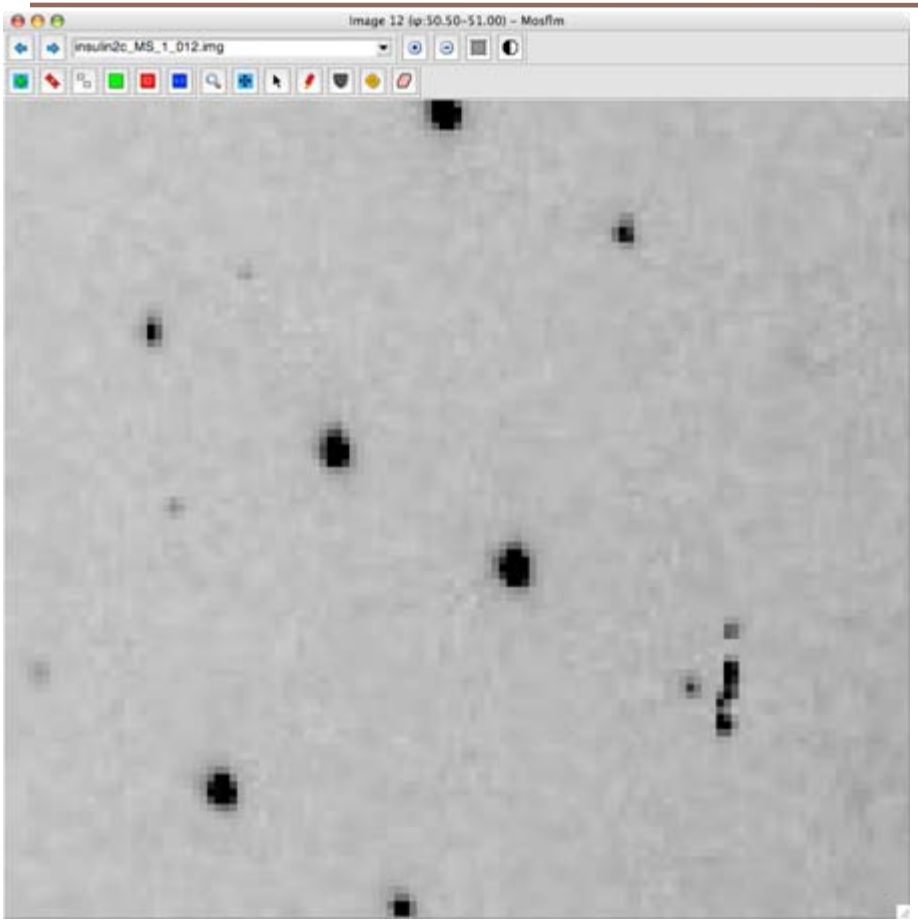
Profile fitting integration – standard profiles

use a profile determined empirically from well-measured reflections to measure the intensity of weak reflections (whose shoulders disappear below the background), simply by minimising *R to* optimise a scale factor (*K*):

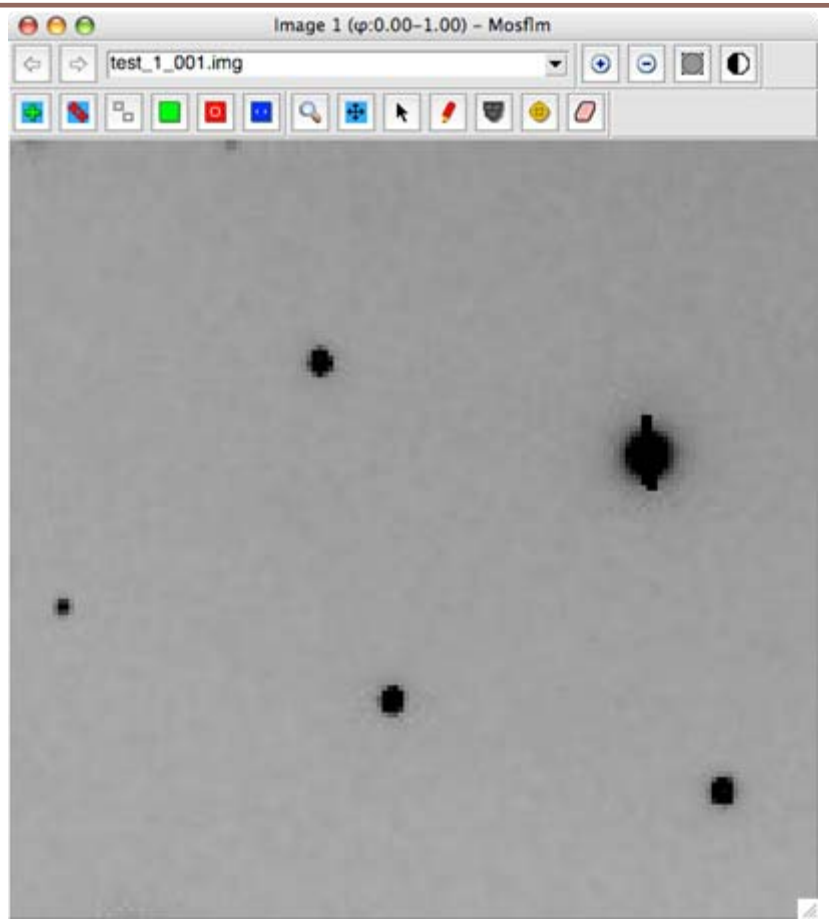
$$R = \sum_{peak\ pixels} w \left(X_i - K P_i \right)^2$$

- requires accurate (sub-pixel) placement of the profile
- improves variance estimates for weak reflections
- should reduce random error (weak reflections)
- may increase systematic error (strong reflections)

Other improvements offered by profile fitting



identify zingers



measure overloads

Analysing the results of integration

- Check graphs - they should vary smoothly without obvious discontinuities.

- Large changes in parameters may indicate problems with the crystal or instrument.

- Look at any images corresponding to discontinuities in the graphs.

- $I/\sigma(I)$ at (high resolution limit-0.2Å) should be >1

- Check any warnings issued by the program; it may be best to re-process after following the advice given (all warnings given by *Mosflm* are accompanied by suggestions on how to improve the processing).

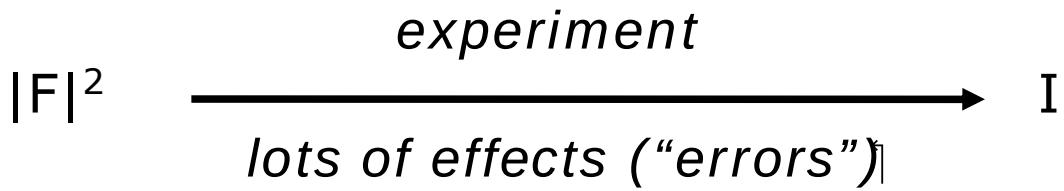
Scaling and merging (a)

Scaling and merging the data is the next step following integration. It is important because:

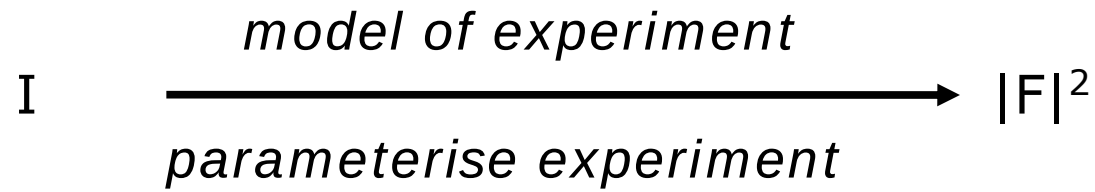
- it attempts to put all observations on a common scale
- it provides the main diagnostics of data quality and whether the data collection is satisfactory

Because of this diagnostic role, it is important that data are scaled as soon as possible after collection, or during collection, preferably while the crystal is still on the camera.

Scaling and merging (b)



Our job is to invert the experiment



Why are reflections on different scales?

Various physical factors lead to observed intensities being on different scales:

(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 the incident beam and camera

(a) incident beam intensity

(b) illuminated volume of crystal

(c) absorption of primary beam by crystal

(d) variations in rotation speed and shutter synchronisation.
If you suspect this, *tell the beamline staff immediately*.

Factors related to the crystal and diffracted beam

(e) Absorption in secondary beam - serious at long wavelength (including CuKα), worth correcting for MAD data

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

Maybe extrapolate back to zero time?

The relative B-factor is largely a correction for 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.

- The useful area of the detector should be calibrated or told to the integration program

- Calibration should flag defective pixels and dead regions *e.g.* 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.*)

Checking the output of *SCALA*

Check these files/plots:

- *SCALA* log file
- ROGUES file
- Normal probability plot(s)
- Correlation plot
- Surface plot
- *loggraph* output

Checking the output of *SCALA*

Check these files/plots:

- *SCALA* log file
- ROGUES file
- Normal probability plot(s)
- Correlation plot
- Surface plot
- *loggraph* output

Questions about the data

- What is the overall quality of the dataset?
  - How does it compare to other datasets for this project?
- What is the real resolution?
  - Should you cut the high-resolution data?
- Are there bad batches
  - individual duff batches or ranges of batches?
- Was the radiation damage such that you should exclude the later parts?
- Is the outlier detection working well?
- Is there any apparent anomalous signal?

Agreement between equivalent reflections

R-factors: traditional overall measures of quality

(a) R_{merge} (R_{sym}) = Σ | I_{hl} - $\langle I_h \rangle$ | / Σ | $\langle I_h \rangle$ |

Avoid referring to this!

(b) R_{meas} = $R_{r.i.m.}$ = $\Sigma \sqrt{(n/n-1)}$ | I_{hl} - $\langle I_h \rangle$ | / Σ | $\langle I_h \rangle$ |
multiplicity-weighted R-factor – good

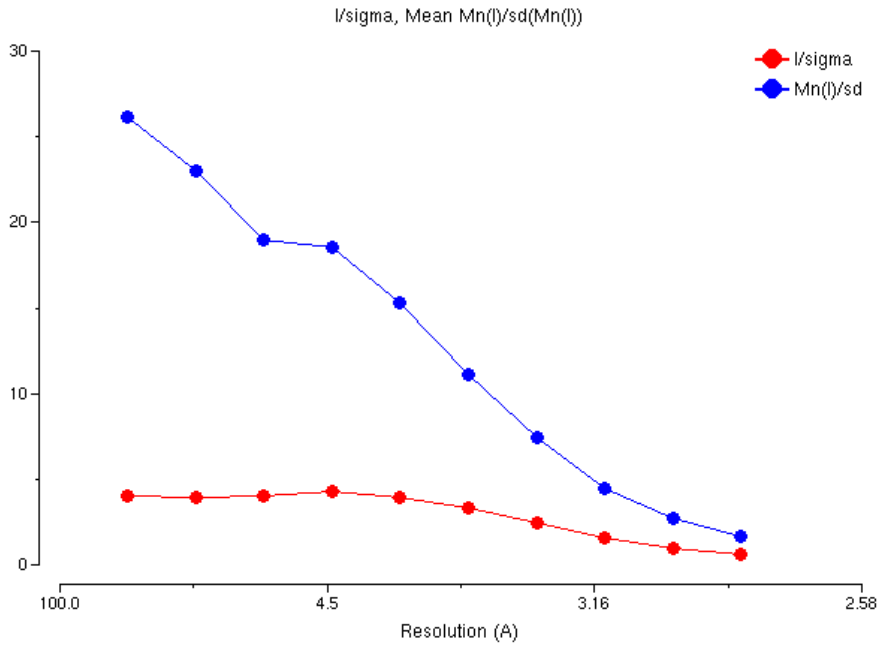
(c) $R_{p.i.m.}$ = $\Sigma \sqrt{(1/n-1)}$ | I_{hl} - $\langle I_h \rangle$ | / Σ | $\langle I_h \rangle$ |

"Precision-indicating R-factor" gets better (smaller) with increasing multiplicity

What is the real resolution (a)?

Look at intensities and standard deviations:

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

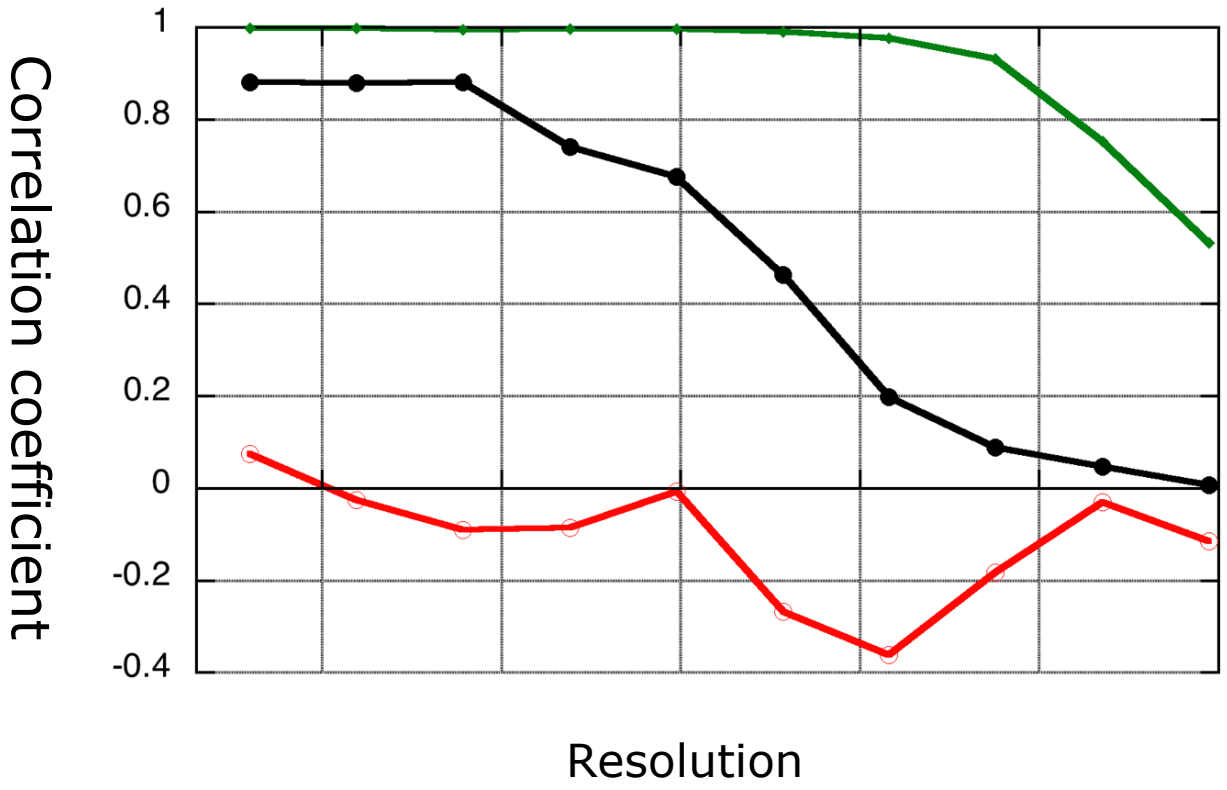


$\langle \langle I \rangle / \sigma(\langle I \rangle) \rangle$ greater than ~ 2 (or so)

Maybe lower for anisotropic data, 1.5 to 1.0

What is the real resolution (b)?

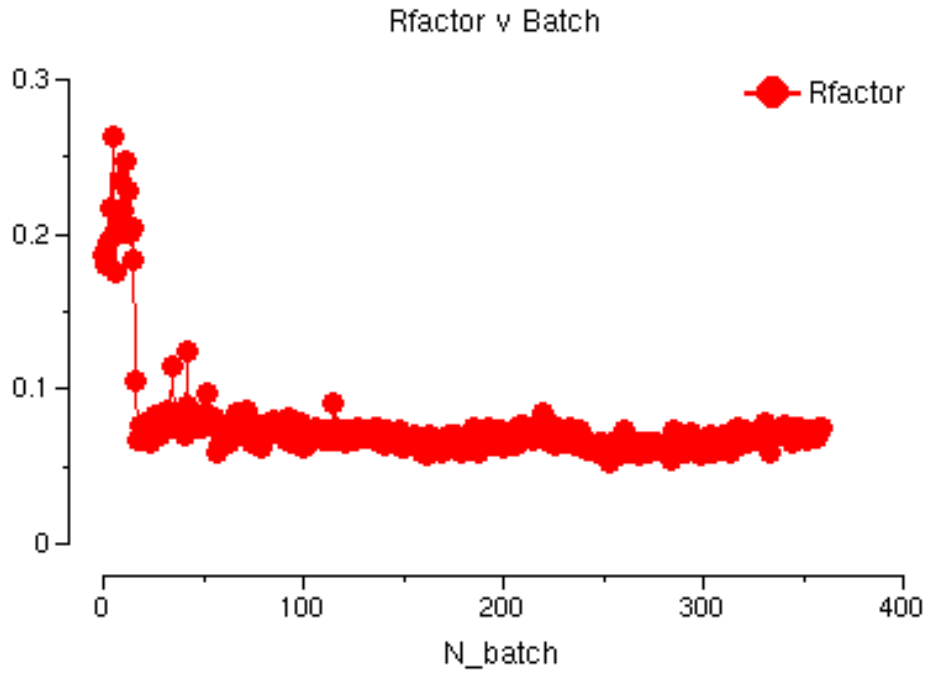
Correlation between half datasets (random halves)



Correlation of <I> indicating a resolution limit

Are some parts of the data bad?

Analysis of R_{merge} against batch number gives a very clear indication of problems local to some regions of the data. Perhaps something has gone wrong with the integration step, or there are some bad images



Here the beginning of the dataset is wrong due to problems in integration (Mosflm)

Outliers

Detection 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

Scala also has facilities for omitting regions of the detector (rectangles and arcs of circles)

Inspect the ROGUES file to see what is being rejected (at least occasionally)

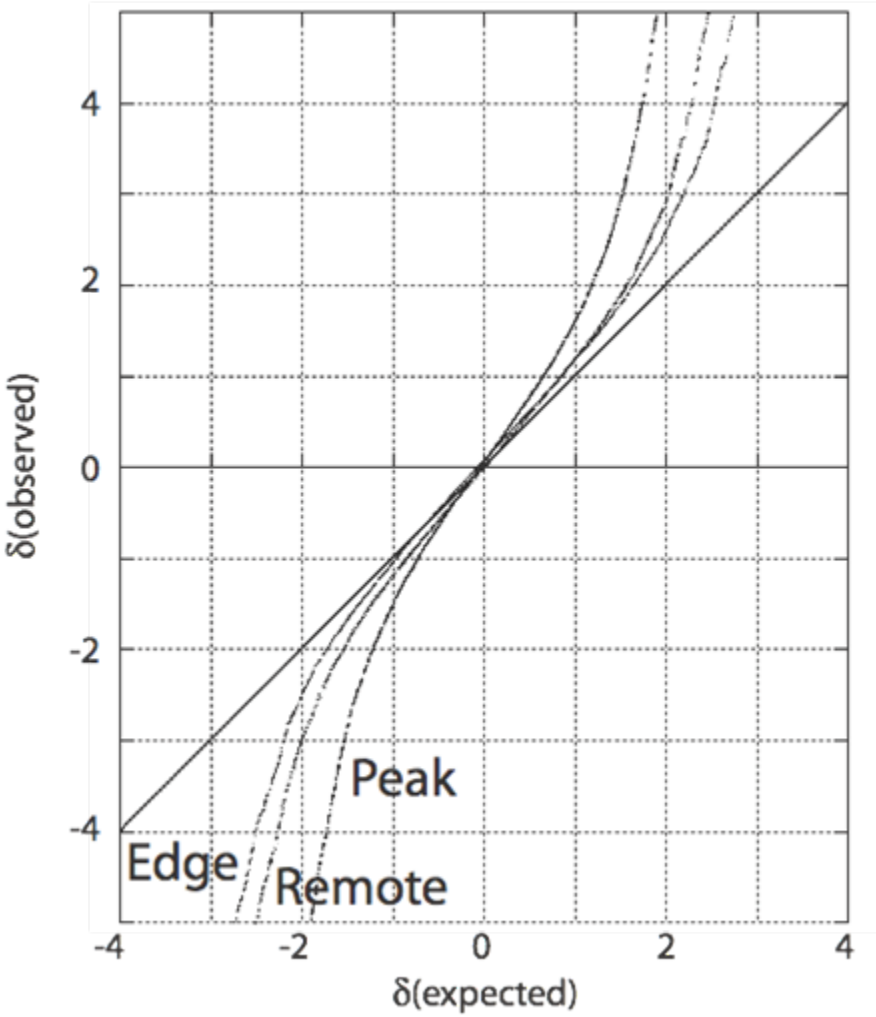
h	k	l	h	k	l	Batch	I	sigI	E	TotFrc	Flag	Scale
(measured)			(unique)									
-2	-2	0	2	2	0	1220	24941	2756	1.03	0.95p	I-	2.434
-4	2	0	2	2	0	1146	9400	2101	0.63	0.99p	*I+	3.017
4	-2	0	2	2	0	1148	27521	2972	1.08	1.09p	I-	2.882
2	-4	0	2	2	0	1075	29967	2865	1.13	0.92p	I+	2.706
						Weighted mean	27407					

Detection of anomalous signal (a)

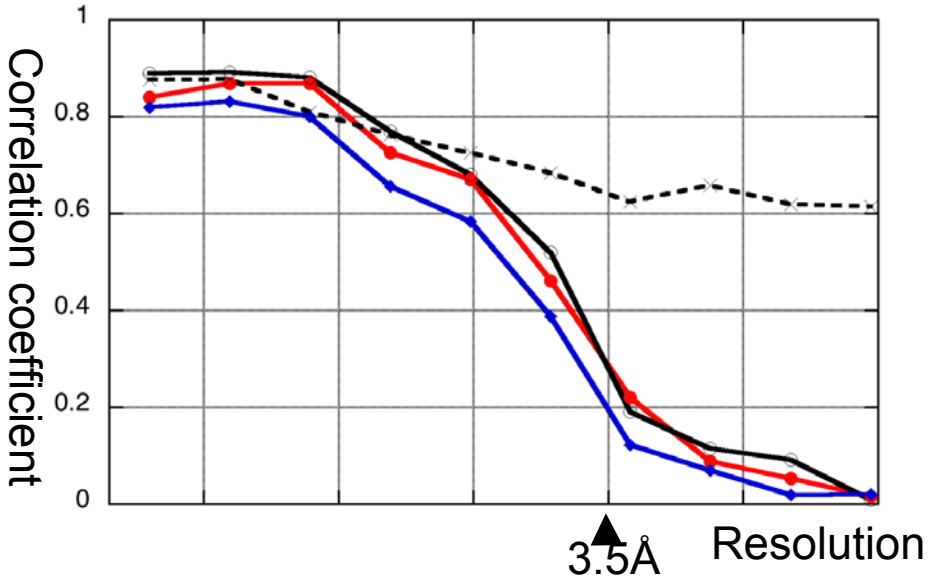
Are the differences greater than would be expected from the errors?

Test using a Normal Probability Plot: a slope > 1.0 means a significant difference

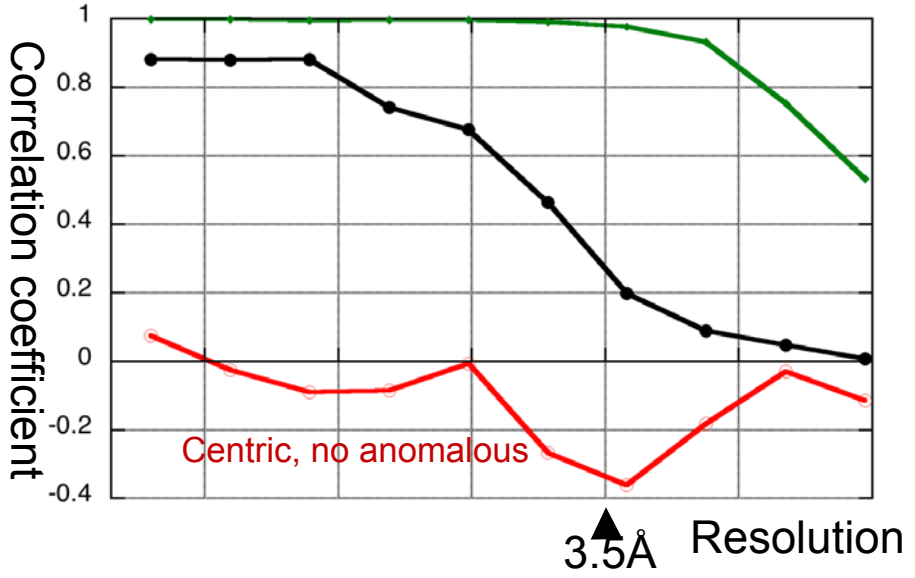
Differences are largest at the peak wavelength



Detection of anomalous signal (b)



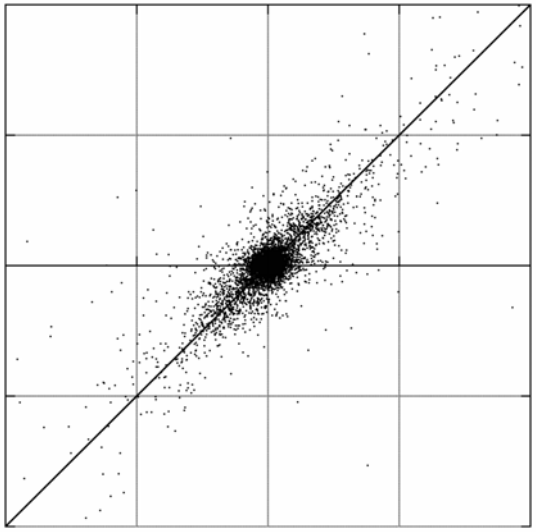
Correlation between wavelengths (MAD)



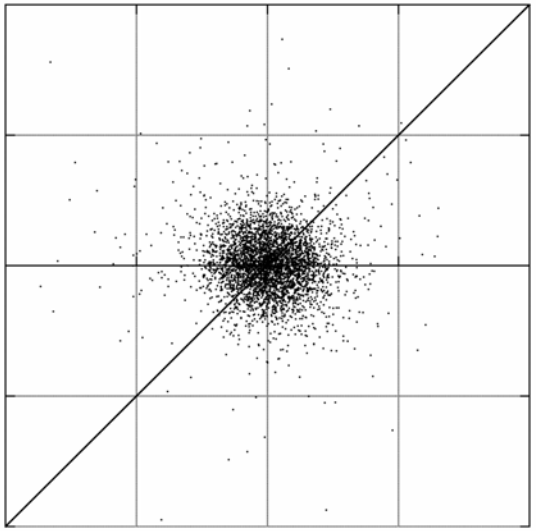
Correlation between half datasets at peak wavelength

This can be used to set the useful resolution for finding anomalous scatterers

Another way of looking at correlations: scatter plot of Δanom1 v. Δanom2

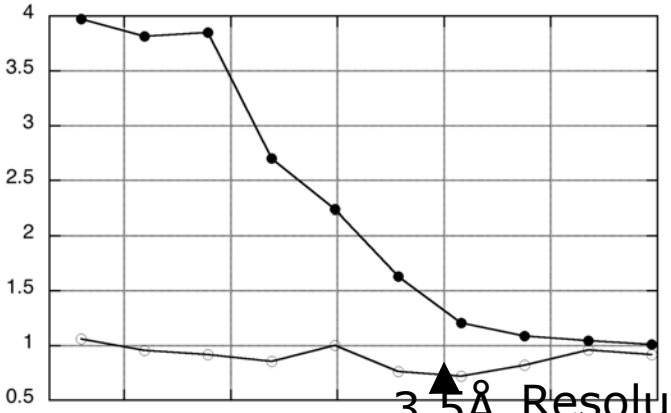


Correlated differences



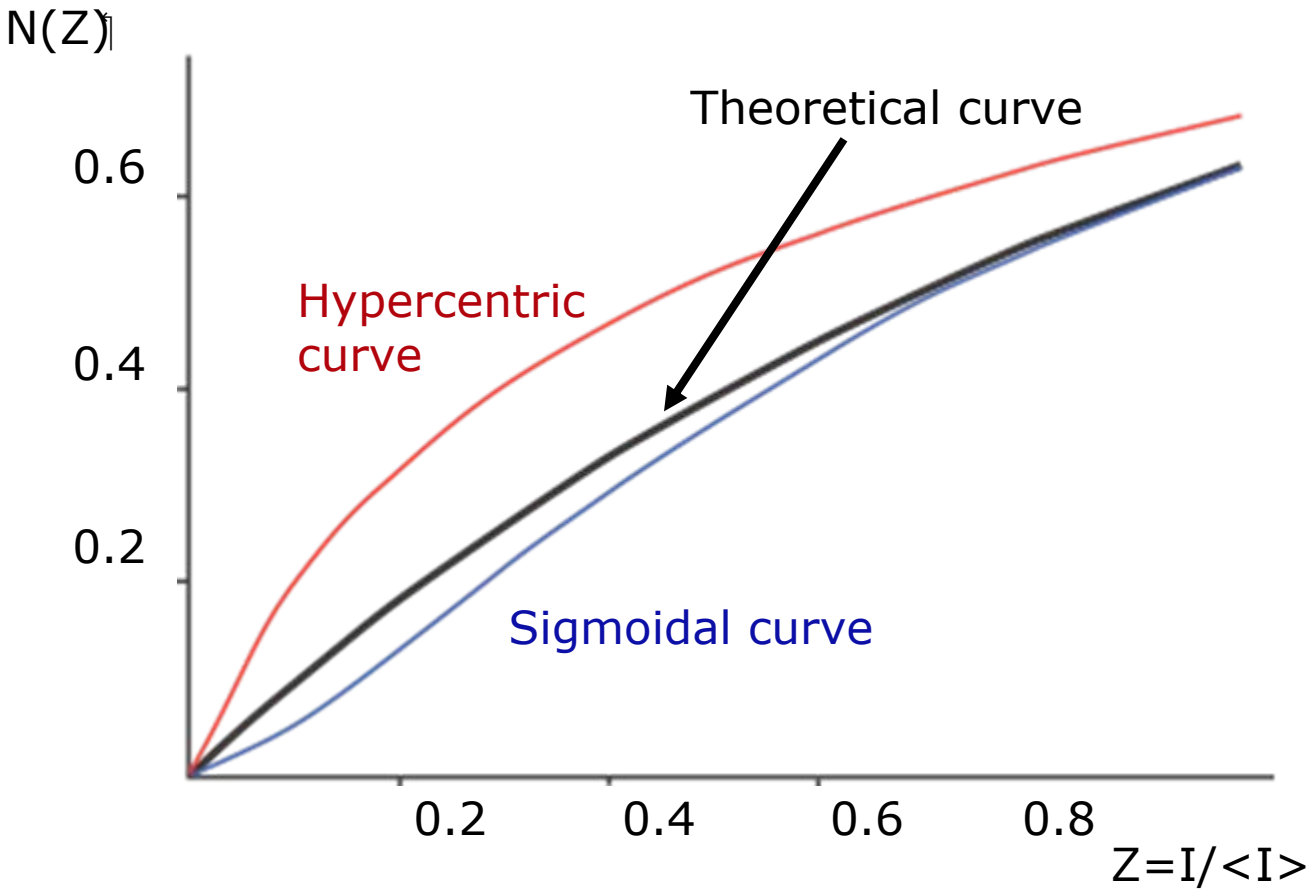
Uncorrelated native

Ratio of distribution width along to width across diagonal ~= signal/noise



Output from *Truncate* or *Ctruncate*

Intensity distributions and their pathologies



Identifiable pathologies

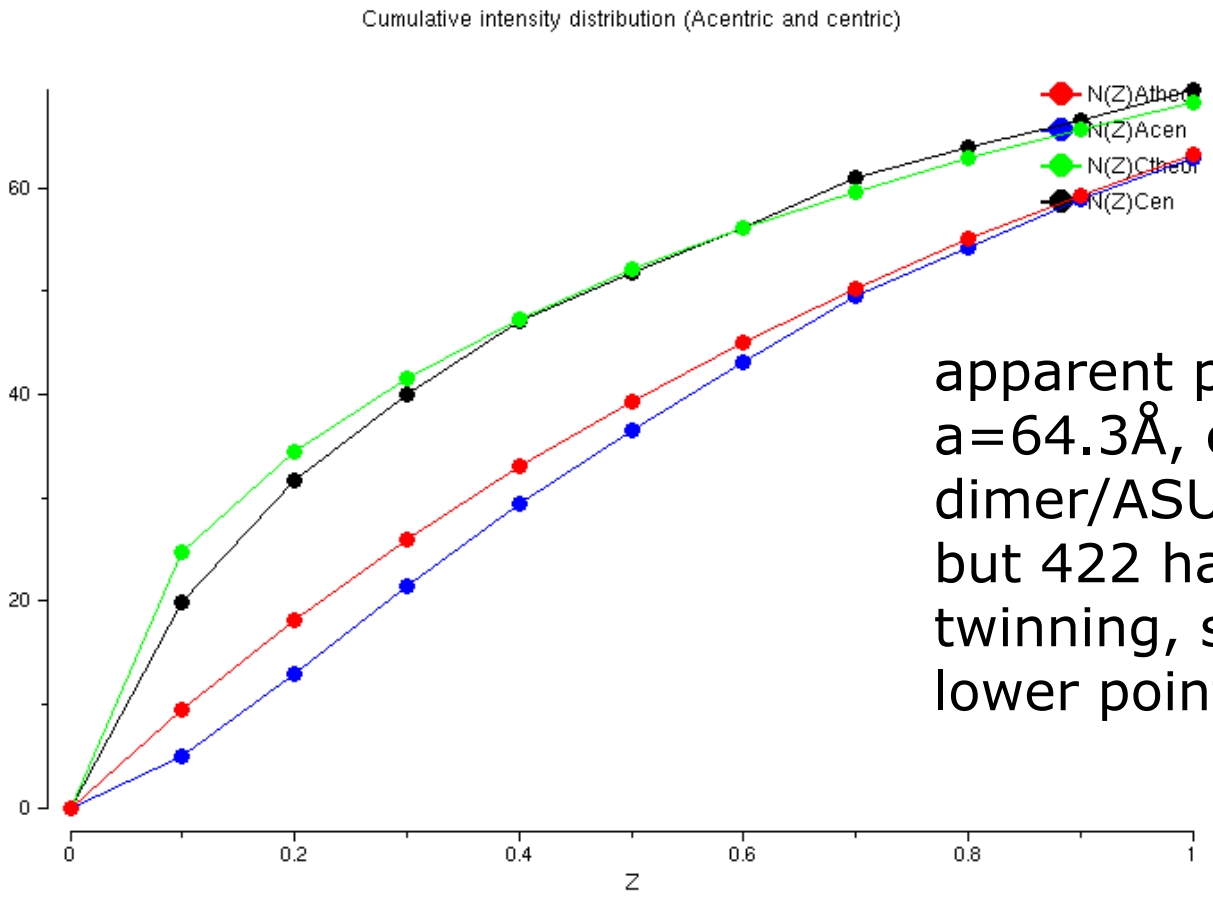
Arising from the intensities themselves - too few weak reflections due to:

- Twinning
- Overlapped reflections
- Systematic underestimation of background – from an underestimate of the detector gain

Arising from the intensity averages - too many (usually) weak reflections due to:

- Anisotropic diffraction
- Translational NCS

Merohedral twinning (exact overlap of lattices)

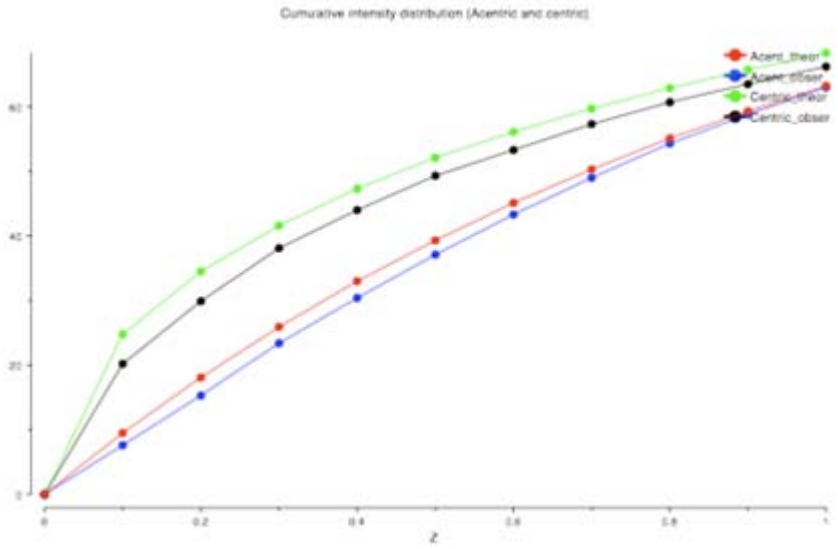


apparent point group 422
a=64.3Å, c=198.8Å,
dimer/ASU, 35kDa, 2.0Å –
but 422 has no possibility of twinning, so this must be lower point group (4).

Pseudo-merohedral twinning

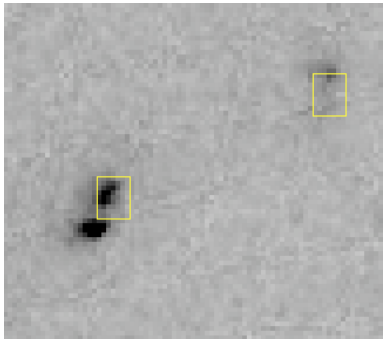
An example of an approximate overlap of lattices:

Apparently point group 422 (twin operator k,h,-l) - 79.2 = a≈b = 81.3Å
True space group: $P2_12_12_1$, 79.2, 81.3 81.2Å



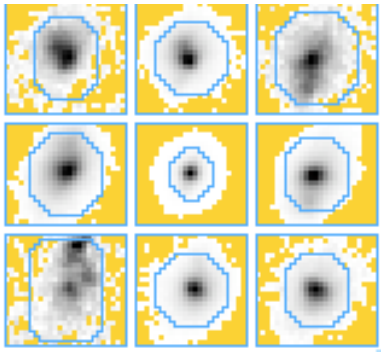
Sigmoidal cumulative intensity plot

on image



Split spots due to non-overlapping lattices

in average profiles



Finally...

Remember:

- Don't expect software to correct for a badly designed (or badly performed) experiment
- Scaling & merging provide the best statistics on the quality of your data
- Good luck!

Acknowledgements

MRC Laboratory of Molecular Biology


Andrew Leslie


Phil Evans


Geoff Battye


Luke Kontogiannis

CCP4
MRC
St Andrews University