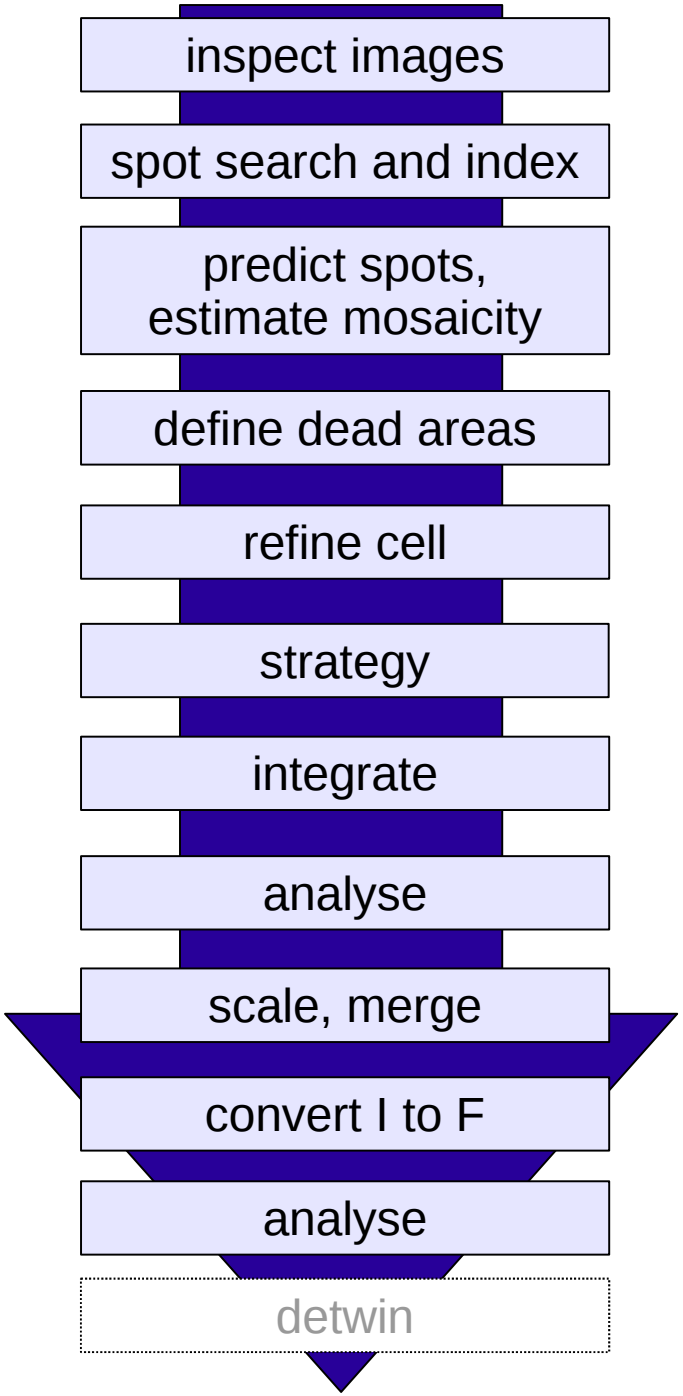
Data processing

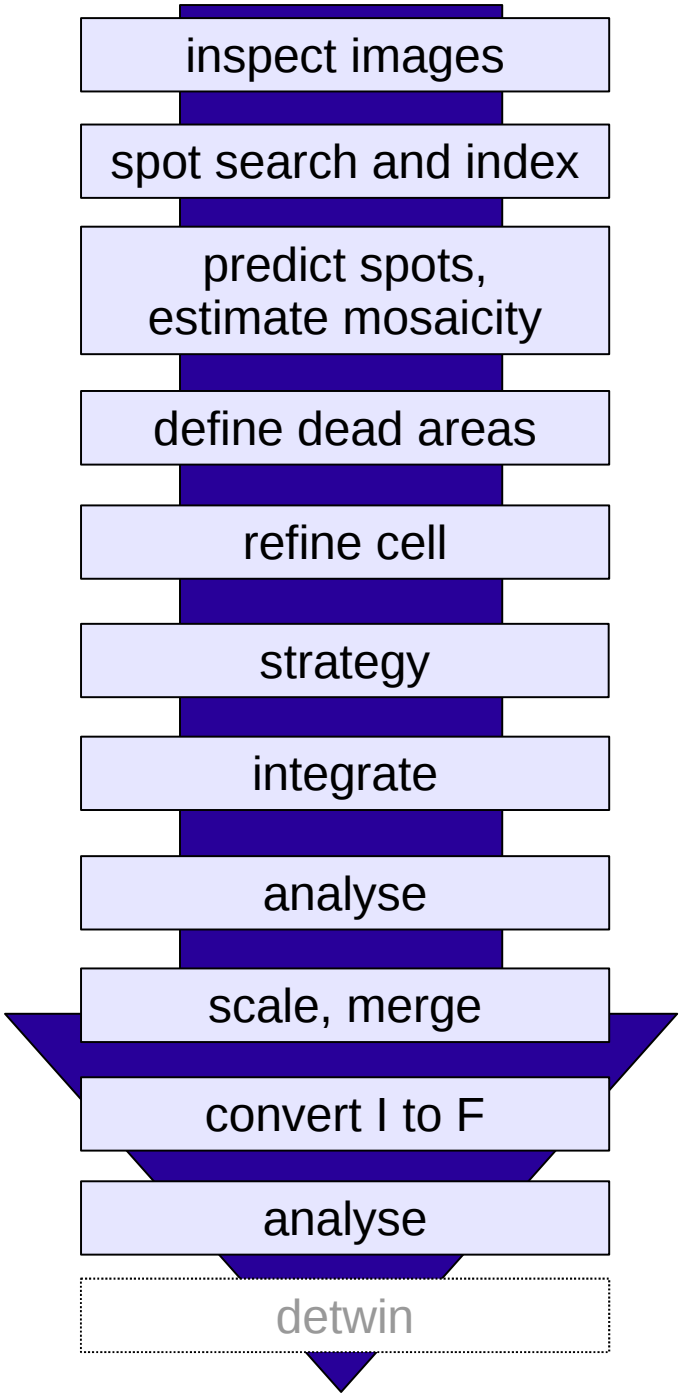
mosflm, imosflm, pointless, scala

Judit É Debreczeni

data processing flowchart

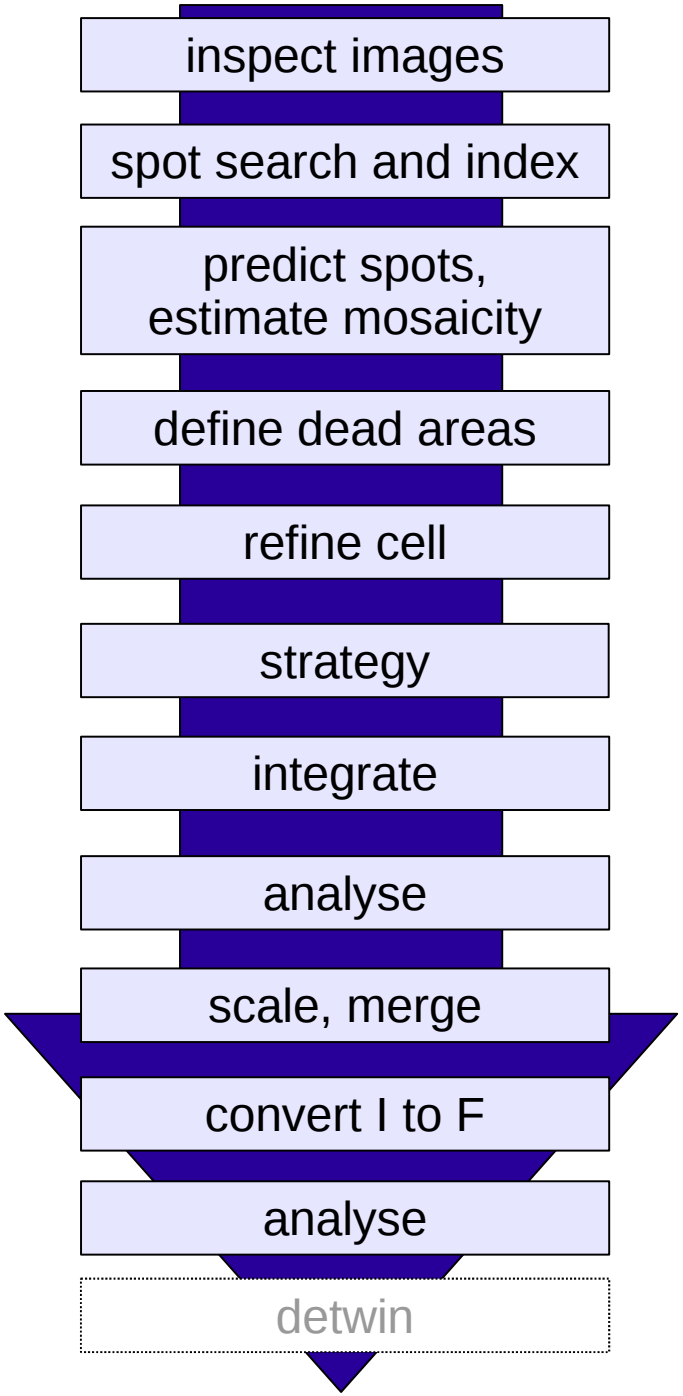


data processing flowchart



inspect images	is this my best crystal? Worth collecting/processing?
spot search and index	determine lattice
predict spots, estimate mosaicity	check solution
define dead areas	exclude shadows, icerings etc
refine cell	refine solution (multiple wedges)
strategy	get total rotation, oscillation, distance
integrate	determine intensities and σ
analyse	determine Laue group and space group
scale, merge	scale all data together and analyse
convert I to F	calculate structure factors
analyse	any bad parts to omit? etc
detwin	detwin if necessary

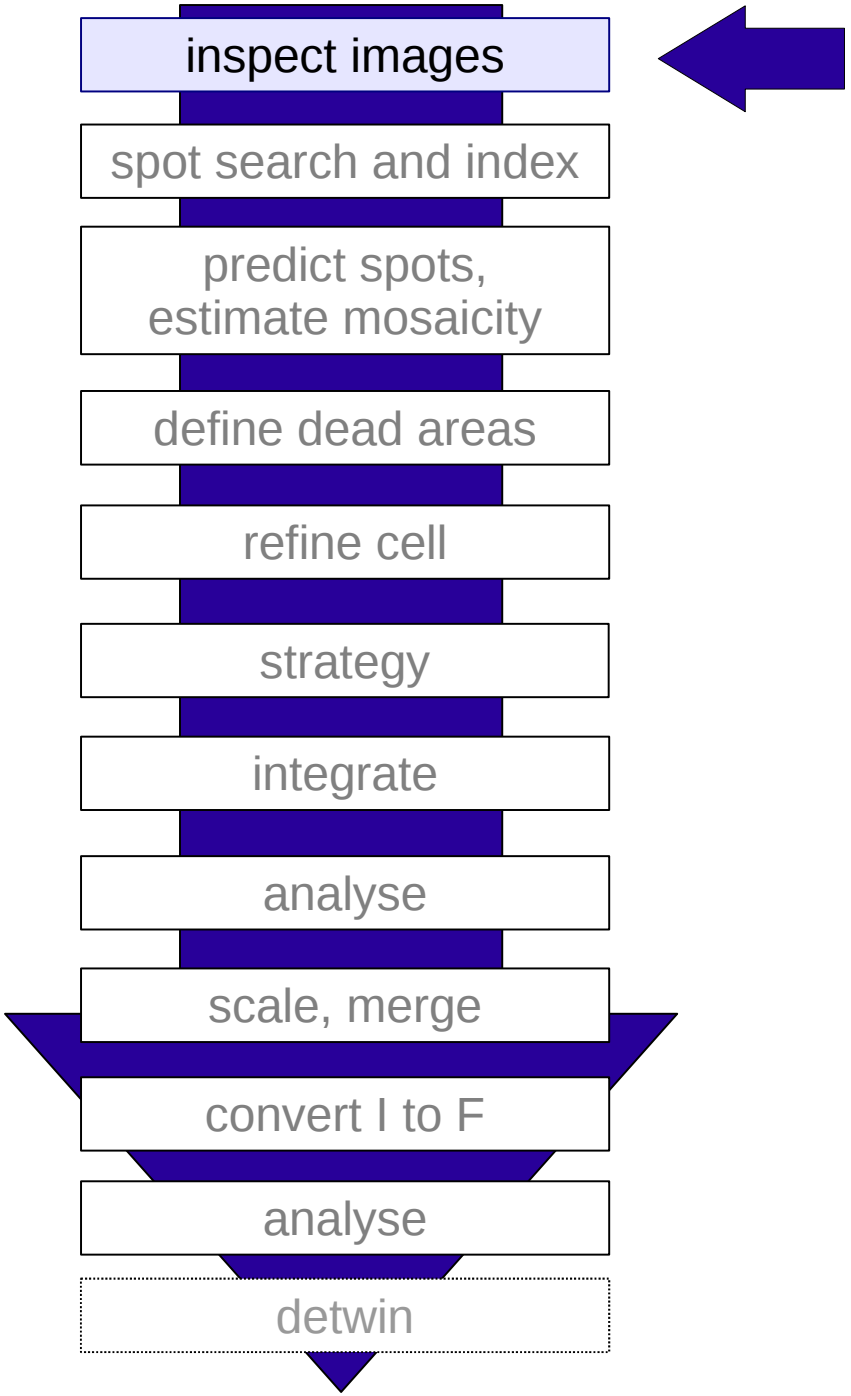
data processing flowchart



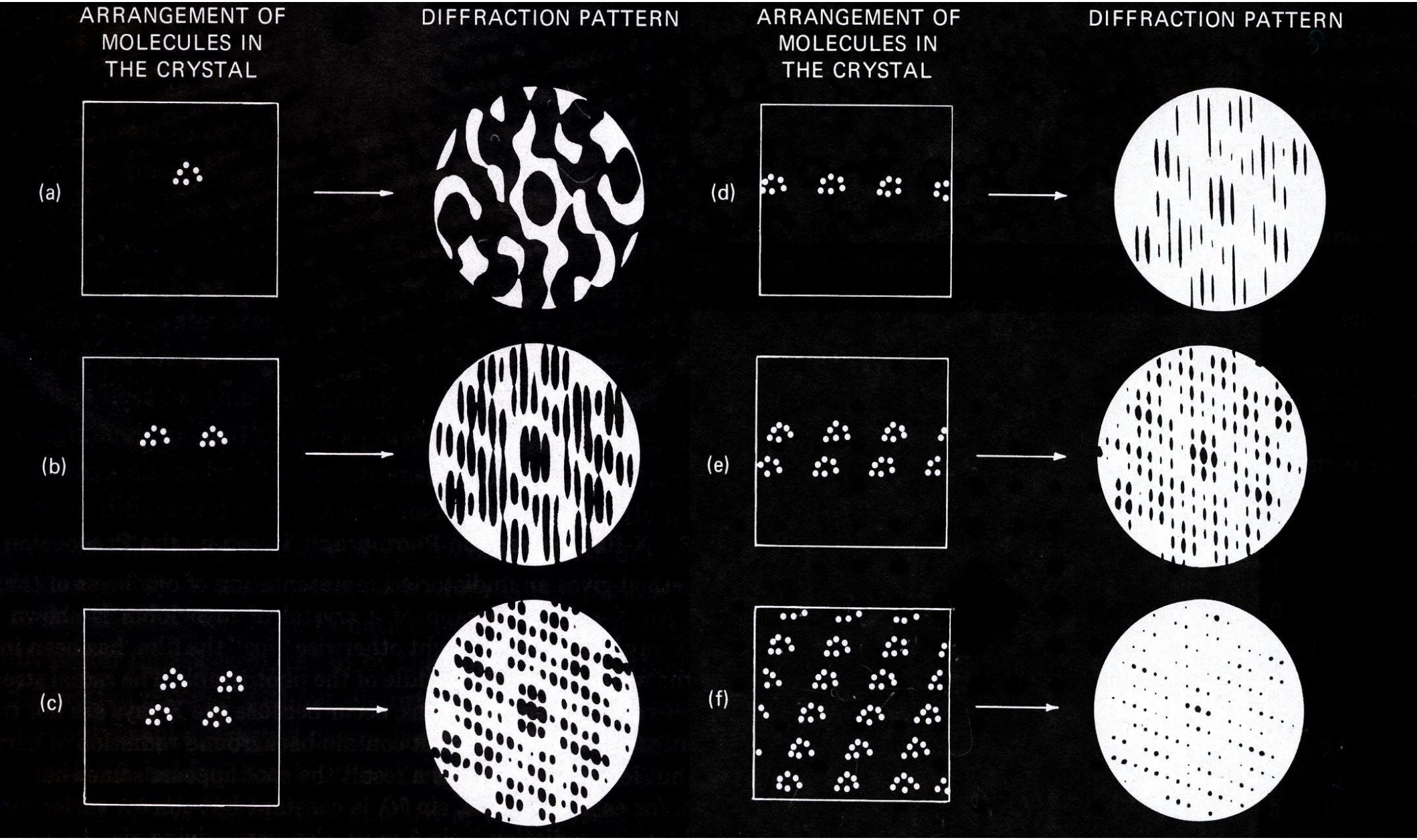
mosflm
imosflm

pointless
scala
truncate

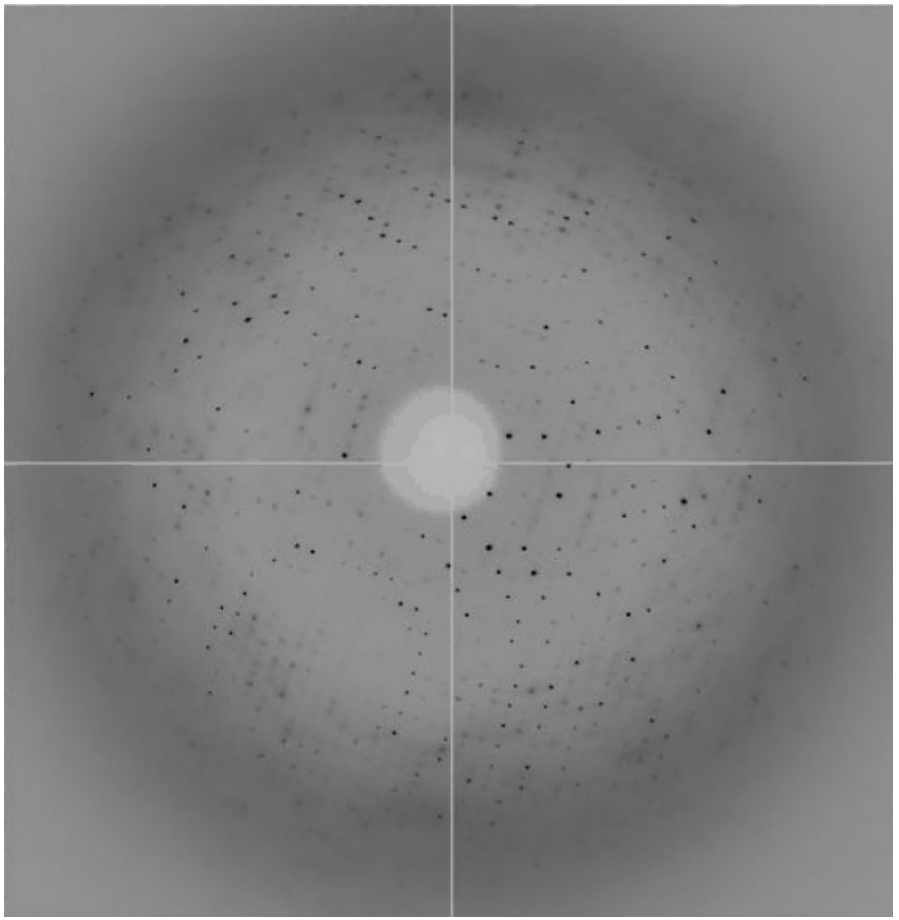
data processing flowchart



scattering and diffraction



scattering and diffraction

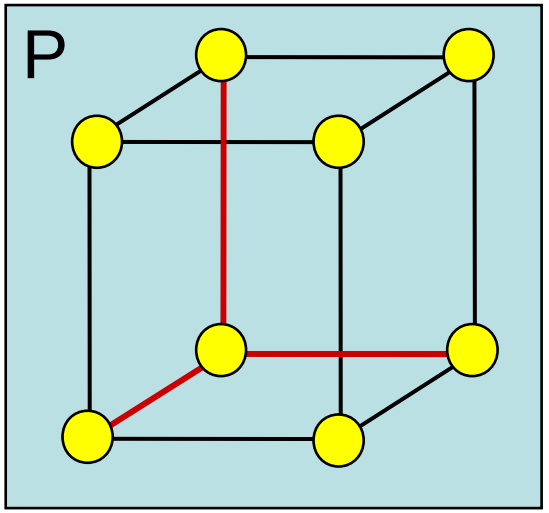


- diffraction images: not only diffraction, also scattering from the crystal's environment:
- solvent, air in the beam
- loop
- ice
- zingers, cosmic radiation

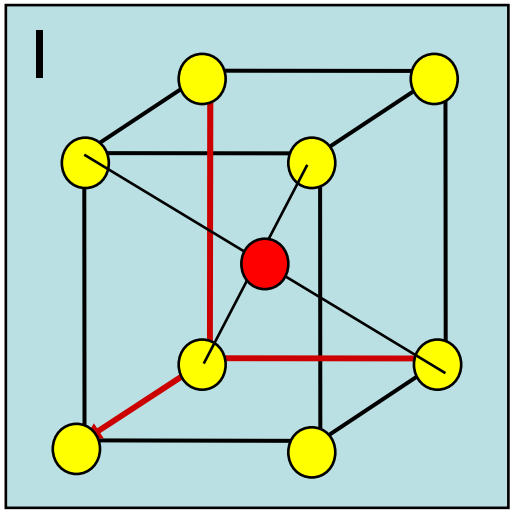
crystal systems

Crystal System	Minimum Symmetry	Bravais Lattices	Unit Cell Geometry
1. Triclinic	None	1. Primitive (*P*)	$a \neq b \neq c$; $\alpha \neq \beta \neq \gamma$
2. Monoclinic	One 2fold axis	2. Primitive (*P*) 3. Base-Centered (*C*)	$a \neq b \neq c$; $\alpha = \gamma = 90° \neq \beta$
3. Orthorhombic	Three orthogonal 2fold axes	4. Primitive (*P*) 5. Base-Centered (*C*) 6. Body-Centered (*I*) 7. Face-Centered (*F*)	$a \neq b \neq c$; $\alpha = \beta = \gamma = 90°$
4. Tetragonal	One 4fold axis	8. Primitive (*P*) 9. Body-Centered (*I*)	$a = b \neq c$; $\alpha = \beta = \gamma = 90°$
5. Trigonal	One 3fold axis	10. Primitive (*P*)	$a = b \neq c$; $\alpha = \beta = 90°, \gamma = 120°$
		11. Rhombohedral (*R*)	$a = b = c$; $\alpha = \beta = \gamma \neq 90°$
6. Hexagonal	One 6fold axis	10. Primitive (*P*)	$a = b \neq c$; $\alpha = \beta = 90°, \gamma = 120°$
7. Cubic	Four 3fold axes	12. Primitive (*P*) 13. Body-Centered (*I*) 14. Face-Centered (*F*)	$a = b = c$; $\alpha = \beta = \gamma = 90°$

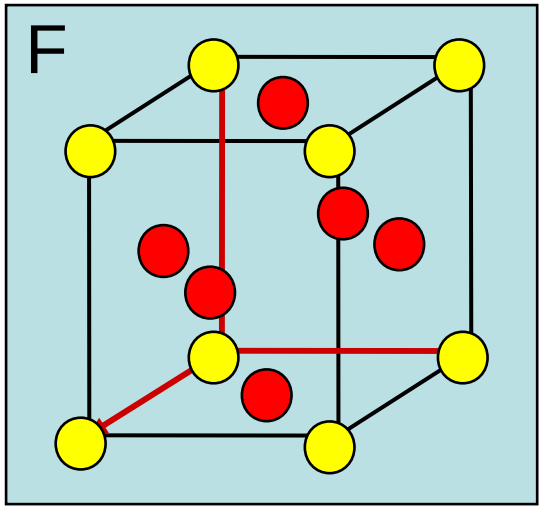
crystal systems



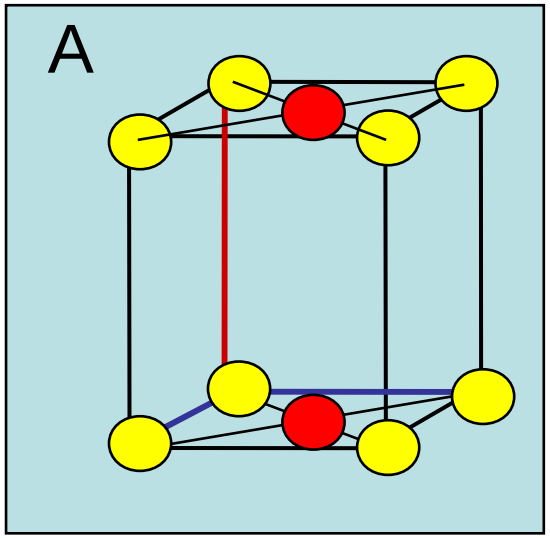
primitive

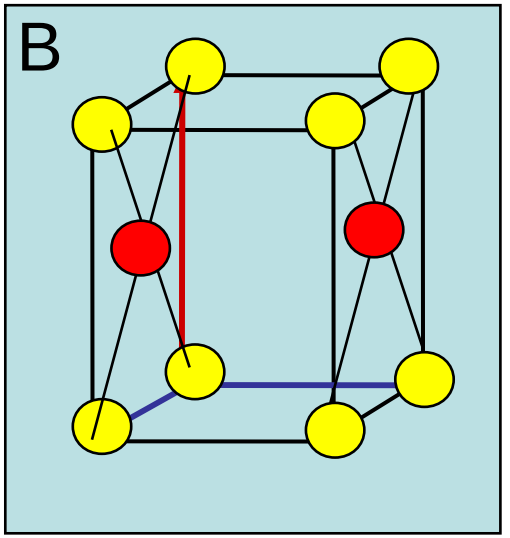


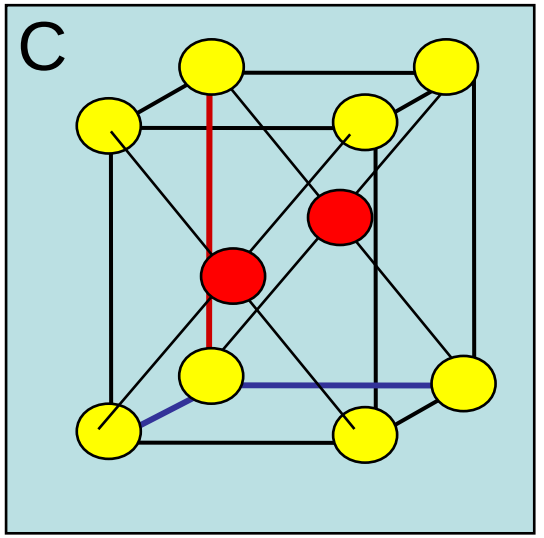
body centered



all face centered







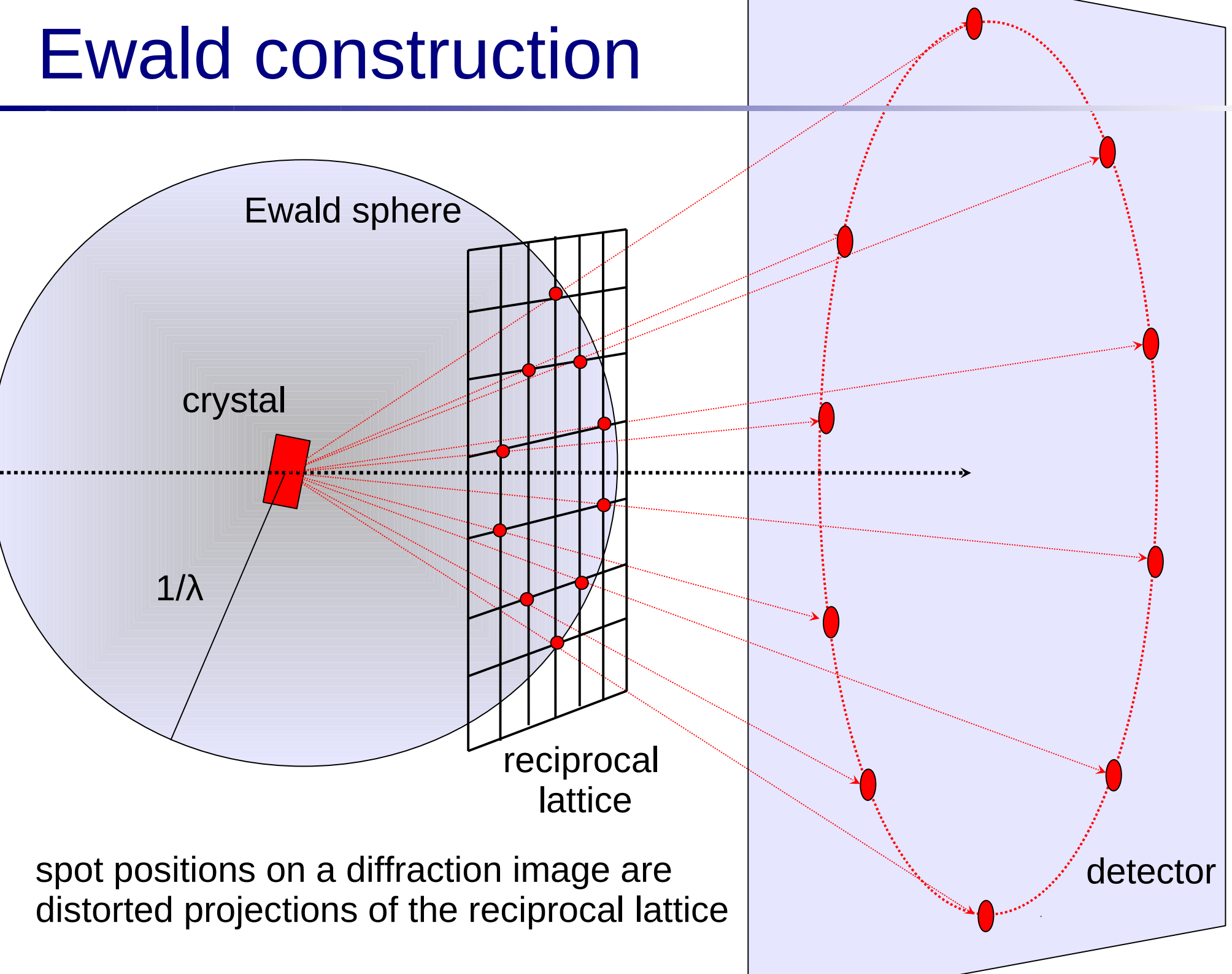
face centered

crystal systems

Crystal System	Minimum Symmetry	Bravais Lattices	Unit Cell Geometry
1. Triclinic	None	1. Primitive (*P*)	$a \neq b \neq c$; $\alpha \neq \beta \neq \gamma$
2. Monoclinic	One 2fold axis	2. Primitive (*P*) 3. Base-Centered (*C*)	$a \neq b \neq c$; $\alpha = \gamma = 90° \neq \beta$
3. Orthorhombic	Three orthogonal 2fold axes	4. Primitive (*P*) 5. Base-Centered (*C*) 6. Body-Centered (*I*) 7. Face-Centered (*F*)	$a \neq b \neq c$; $\alpha = \beta = \gamma = 90°$
4. Tetragonal	One 4fold axis	8. Primitive (*P*) 9. Body-Centered (*I*)	$a = b \neq c$; $\alpha = \beta = \gamma = 90°$
5. Trigonal	One 3fold axis	10. Primitive (*P*)	$a = b \neq c$; $\alpha = \beta = 90°, \gamma = 120°$
		11. Rhombohedral (*R*)	$a = b = c$; $\alpha = \beta = \gamma \neq 90°$
6. Hexagonal	One 6fold axis	10. Primitive (*P*)	$a = b \neq c$; $\alpha = \beta = 90°, \gamma = 120°$
7. Cubic	Four 3fold axes	12. Primitive (*P*) 13. Body-Centered (*I*) 14. Face-Centered (*F*)	$a = b = c$; $\alpha = \beta = \gamma = 90°$

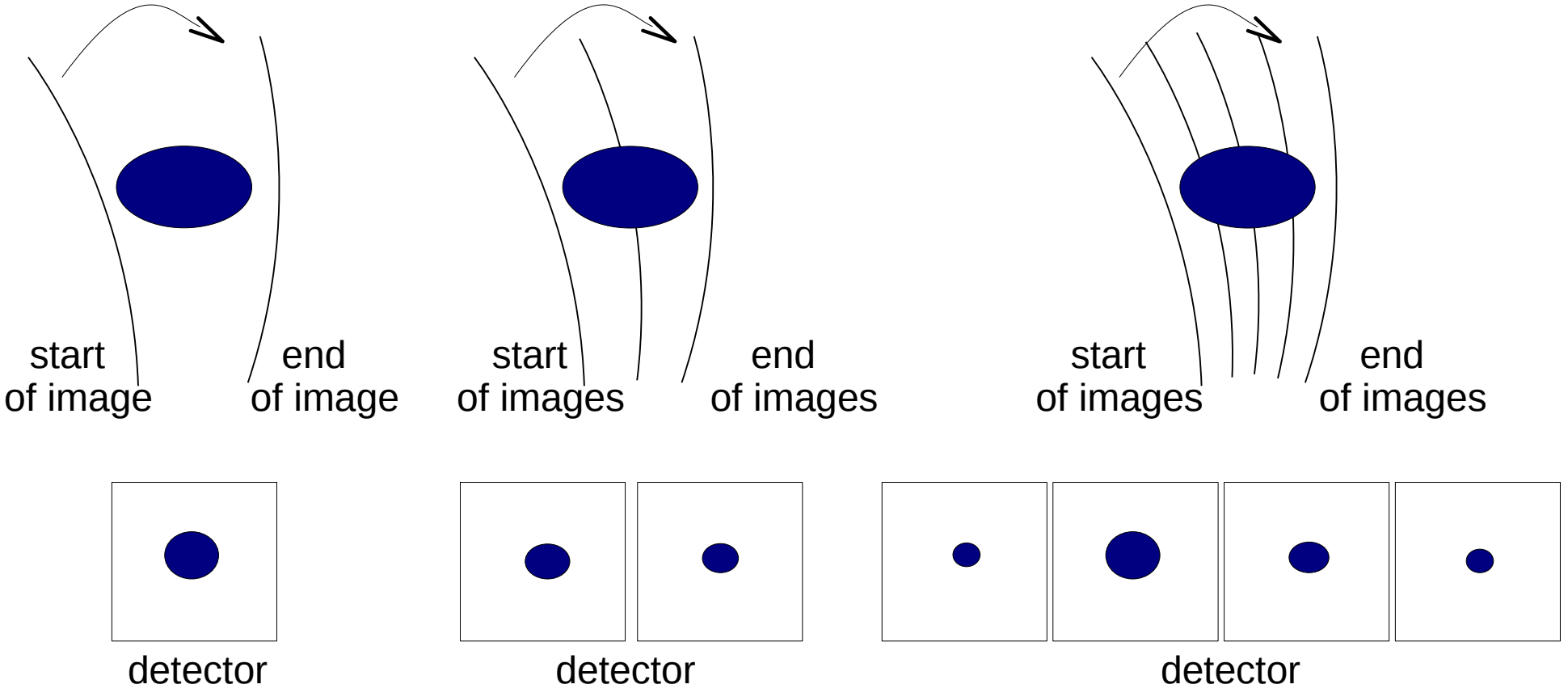
- 7 crystal systems + P, I, F, A, B, C -> 14 Bravais lattices
- many combinations reduce to one another

Ewald construction



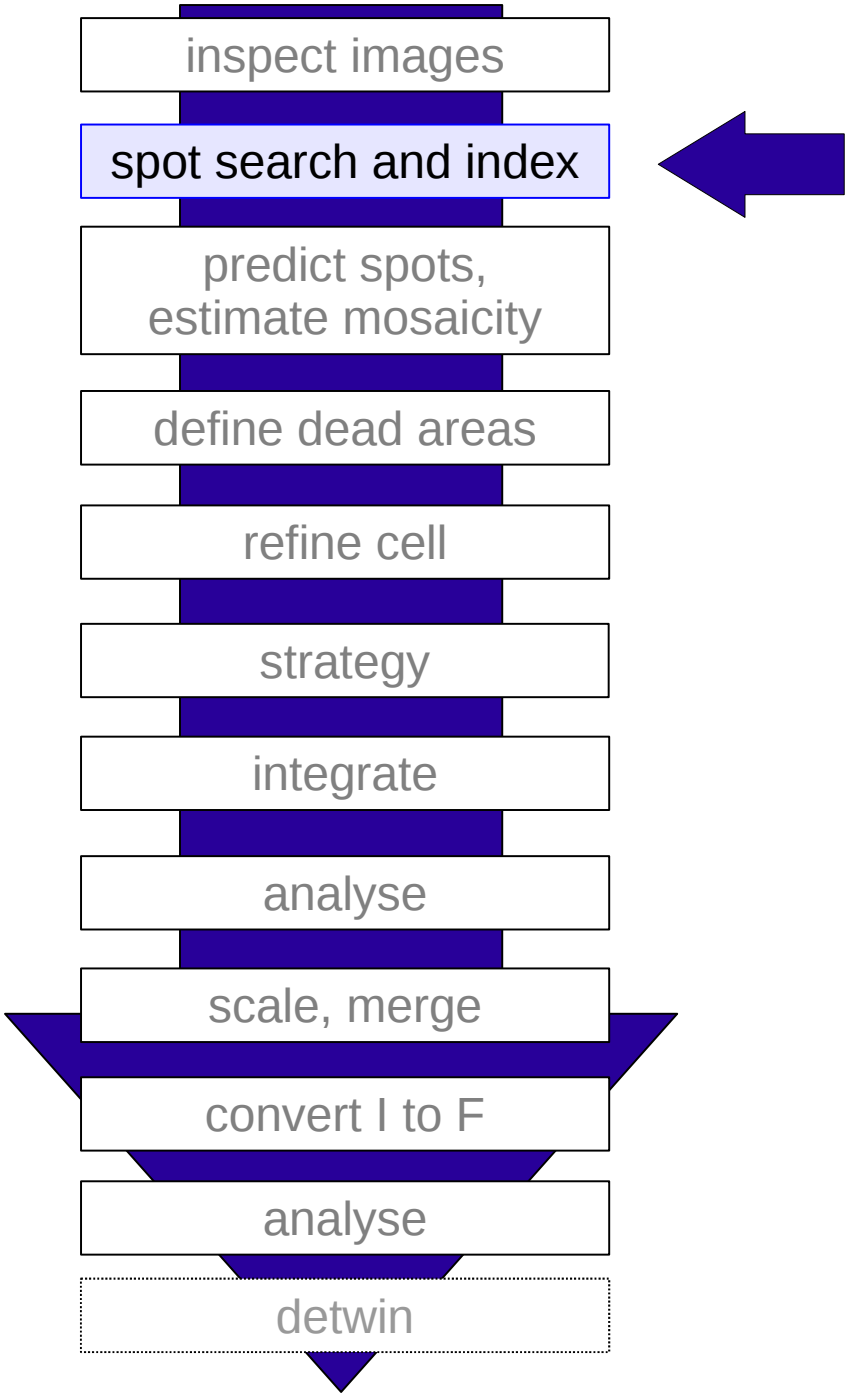
spot positions on a diffraction image are distorted projections of the reciprocal lattice

fulls and partials



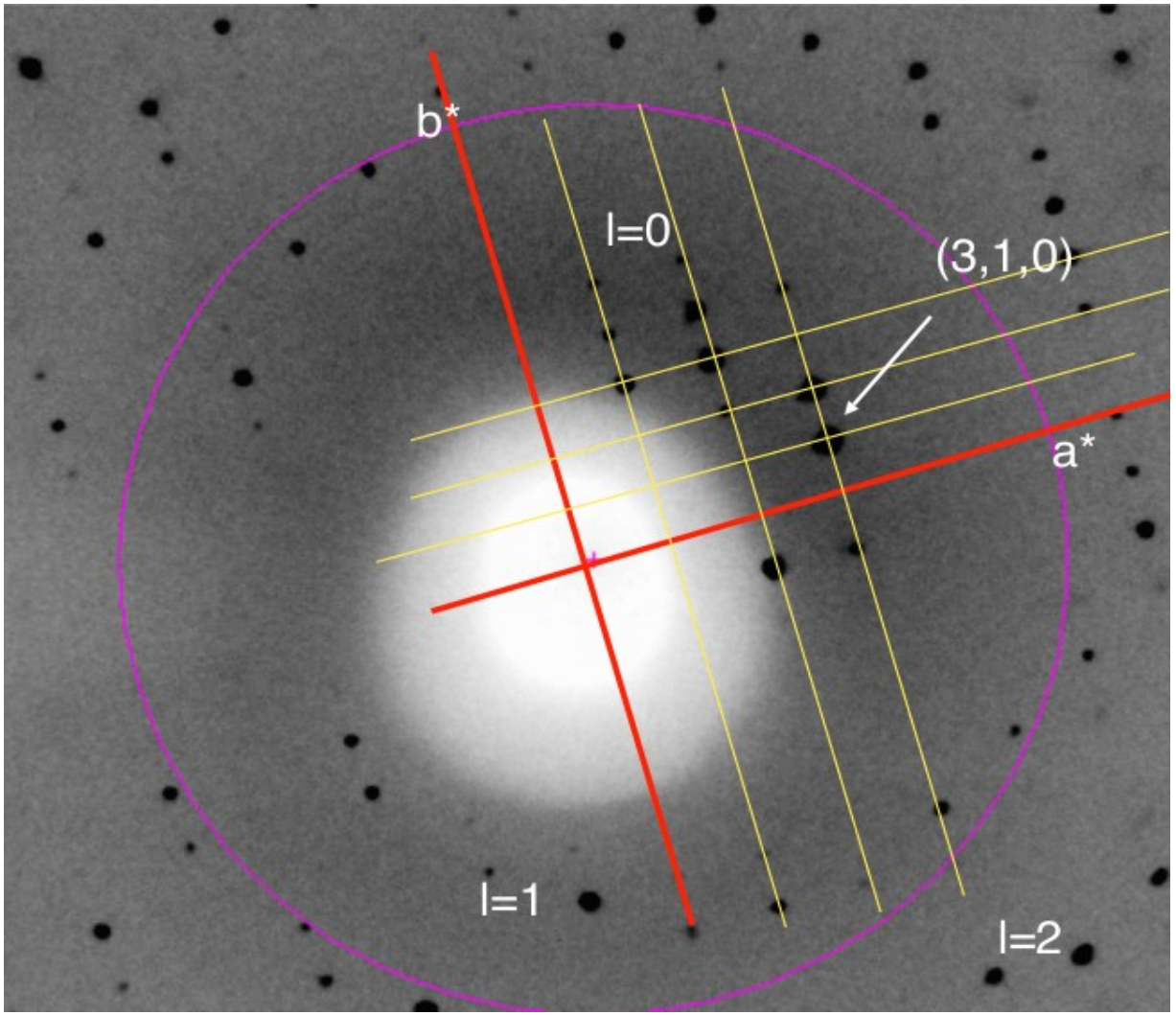
- fully recorded spot spans one image
- fine sliced data with spots sampled in 3D

data processing flowchart

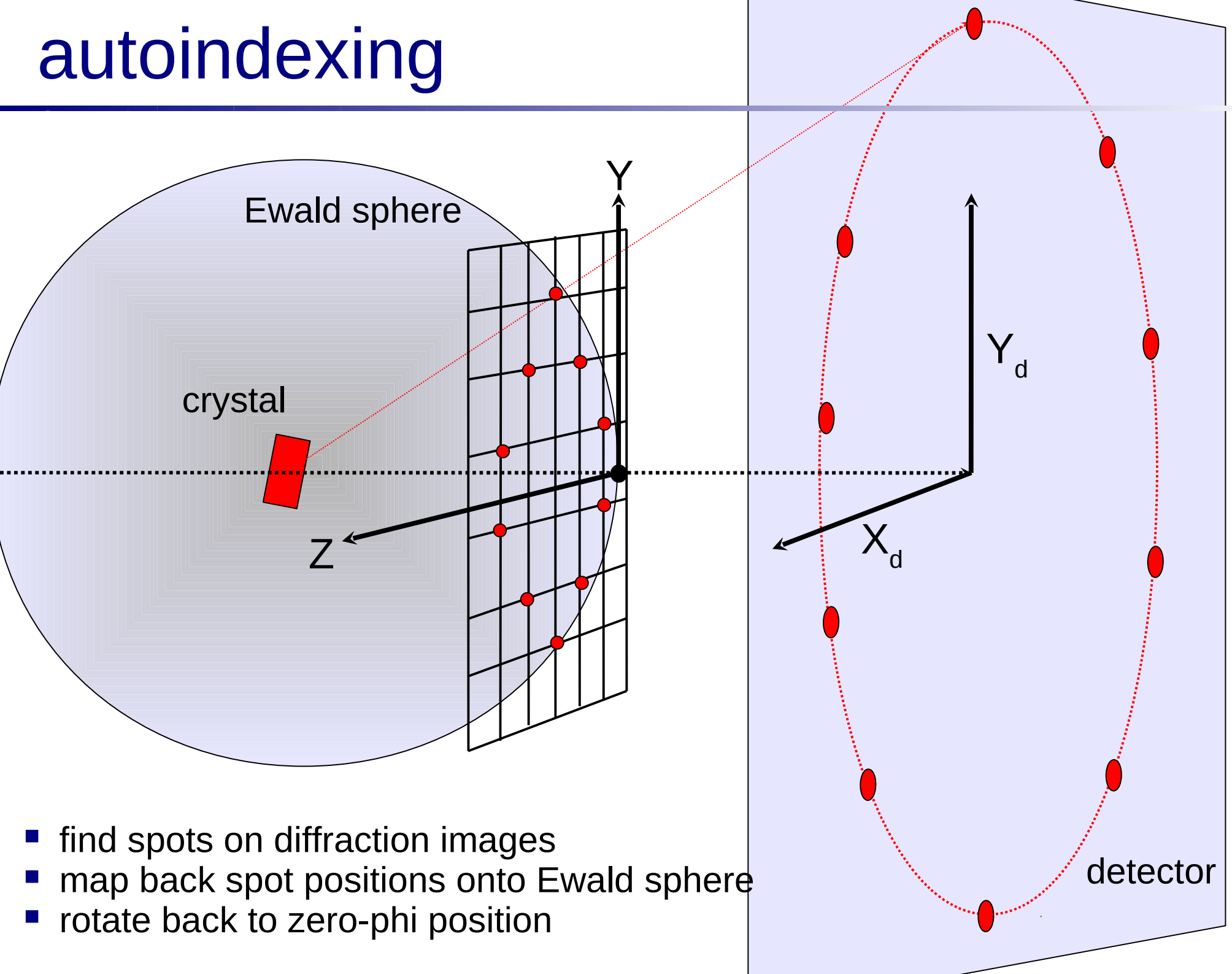


indexing

- if we know the main beam position on the image,
  we can count spots from the origin
- tedious, as we have to work in 3D and crystal is not aligned with rotation axis, detector etc

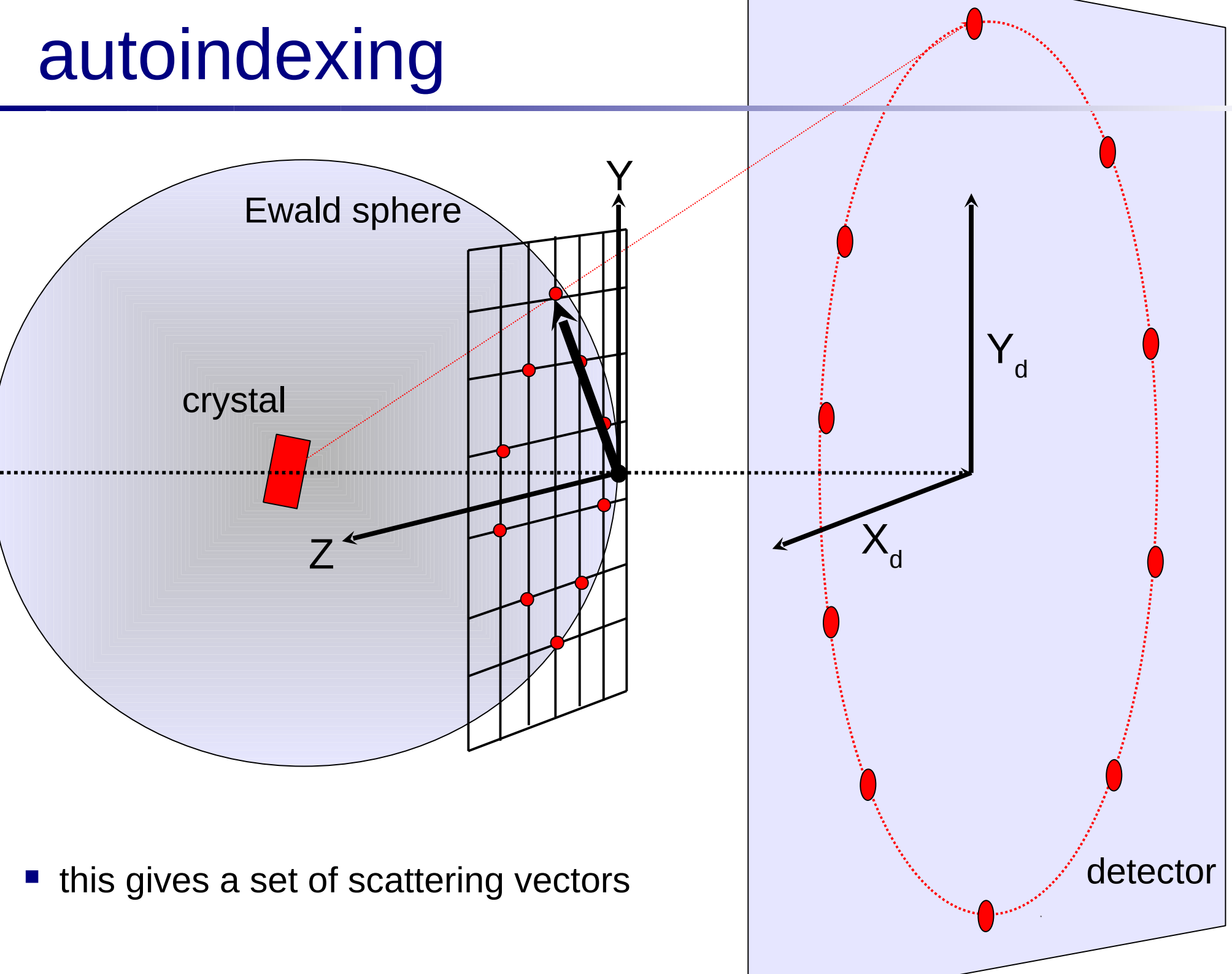


autoindexing



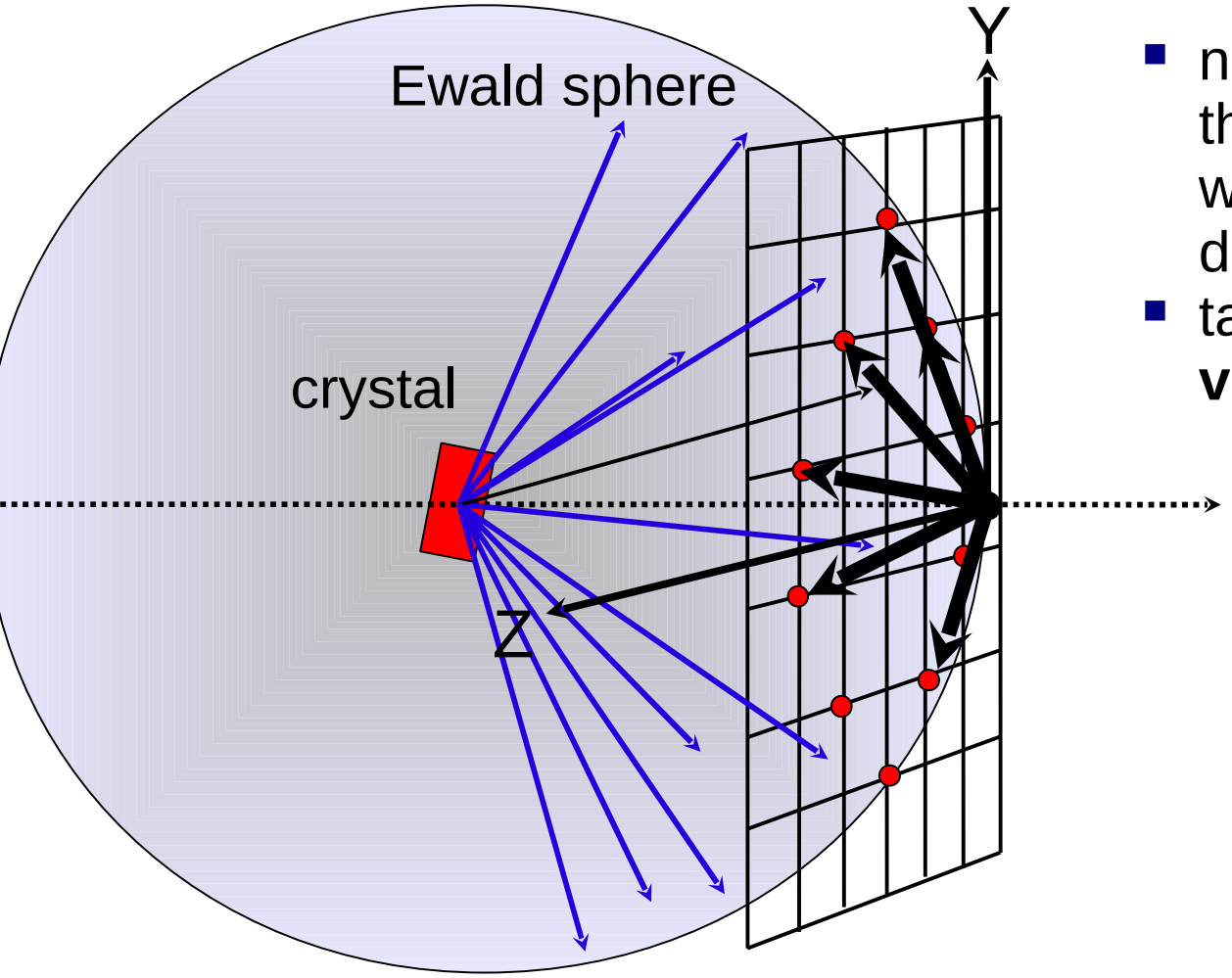
- find spots on diffraction images
- map back spot positions onto Ewald sphere
- rotate back to zero-phi position

autoindexing



- this gives a set of scattering vectors

autoindexing

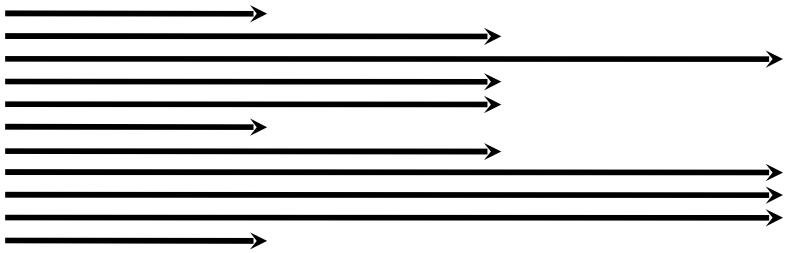


- now consider every direction in the Ewald sphere (sample the whole hemishpere with ~8000 directions)
- take the projection of **scattering vectors** onto these directions

set of projections

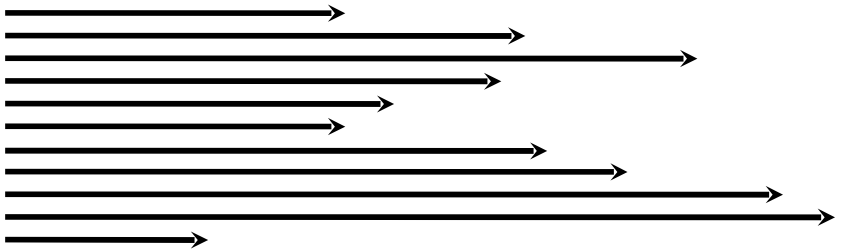
autoindexing

- if the direction in question is a real space cell axis (perpendicular to a reciprocal space lattice plane), the projections will cluster at lengths which are multiples of lattice spacing



clustering of lengths

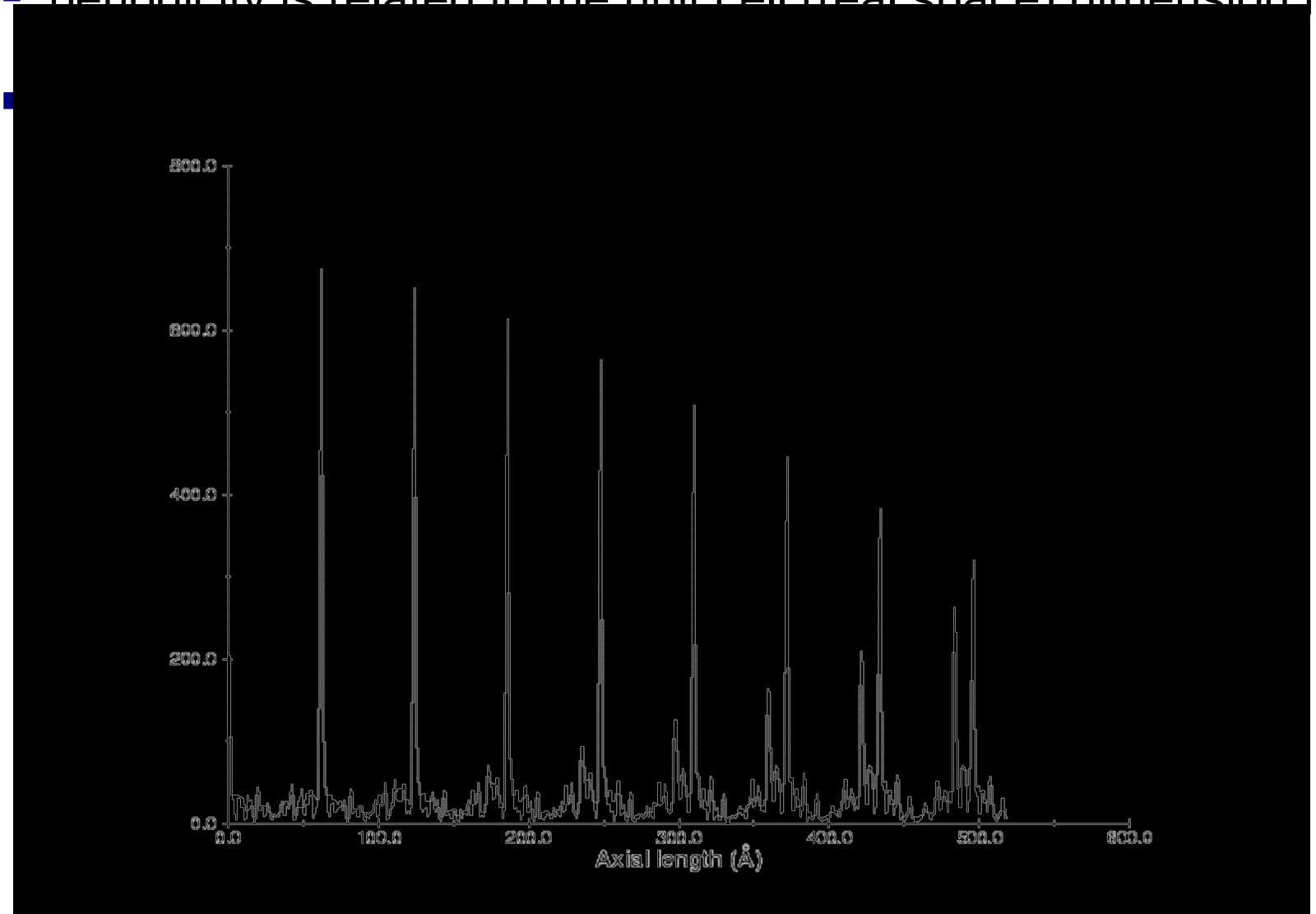
- otherwise lengths of projections will be random, no clustering



random lengths

autoindexing

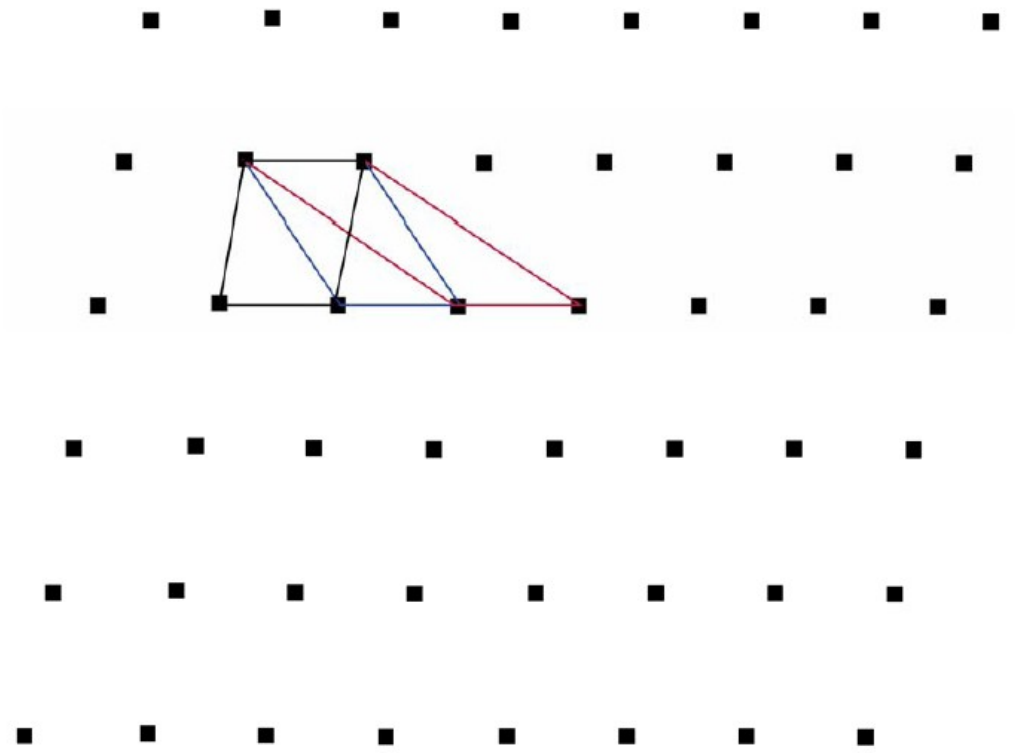
- Evenness of spacing is evaluated by a 1D Fourier transform of the projection
- if peaks are evenly spaced, a few low frequency terms will be large
- periodicity is related to the unit cell (real space) dimension in that



autoindexing

- pick three non-coplanar directions with large peaks in the Fourier
- the lattice found this way is not necessarily the simplest cell:

  black: reduced cell
  red, blue: possible autoindexing solutions



autoindexing - in practice

- multiple solutions with penalties: how well does the cell obey the constraints of that lattice type
- sharp drop between solutions and non-solutions
- solution's penalty < 10 (20 okay for dodgy crystal)
- pick the highest symmetry from the low-penalty bunch

autoindexing - in practice

Solutions:

Solution	Lat.	Pen.	a	b	c	α	β	γ	σ(x,y)	σ(φ)	δ beam
1 (ref)	aP	0	67.1	67.4	67.5	90.1	90.1	90.0	0.14	0.49	0.34 (0.1)
2 (ref)	mP	0	67.1	67.4	67.5	90.0	90.0	90.0	0.14	0.49	0.34 (0.1)
3 (ref)	mP	0	67.1	67.5	67.4	90.0	90.0	90.0	0.14	0.49	0.34 (0.1)
4 (ref)	mP	0	67.4	67.1	67.5	90.0	90.0	90.0	0.14	0.49	0.34 (0.1)
5 (ref)	aP	0	67.1	67.4	67.5	89.9	89.9	90.0	0.14	0.49	0.34 (0.1)
6 (ref)	mC	1	95.3	95.4	67.1	90.0	90.0	90.0	0.14	0.59	0.34 (0.1)
7 (ref)	oP	1	67.1	67.4	67.5	90.0	90.0	90.0	0.14	0.49	0.34 (0.1)
8 (ref)	tP	1	67.4	67.4	67.1	90.0	90.0	90.0	0.14	0.49	0.34 (0.1)
9 (ref)	oC	1	95.4	95.4	67.1	90.0	90.0	90.0	0.14	0.59	0.34 (0.1)
10 (ref)	mC	1	95.3	95.4	67.1	90.0	90.0	90.0	0.14	0.59	0.34 (0.1)
11 (ref)	mC	2	95.1	95.0	67.5	90.0	90.3	90.0	0.16	0.66	0.34 (0.1)
12 (ref)	oC	2	95.1	95.0	67.5	90.0	90.0	90.0	0.19	0.65	0.33 (0.1)
13 (ref)	mC	2	95.1	95.0	67.5	90.0	90.3	90.0	0.16	0.66	0.34 (0.1)
14 (ref)	tP	3	67.2	67.2	67.5	90.0	90.0	90.0	0.18	0.50	0.34 (0.1)
15 (ref)	hR	3	95.1	95.1	116.9	90.0	90.0	120.0	0.17	0.58	0.35 (0.2)
16 (ref)	cP	3	67.3	67.3	67.3	90.0	90.0	90.0	0.21	0.55	0.34 (0.1)
17 (ref)	hR	3	95.2	95.2	116.5	90.0	90.0	120.0	0.20	0.46	0.34 (0.2)
18 (reg)	oC	249	67.1	150.7	67.4	90.0	90.0	90.0	-	-	-
19 (reg)	mC	249	150.7	67.1	67.4	90.0	90.1	90.0	-	-	-
20 (reg)	mC	250	67.1	150.6	67.5	90.0	90.0	90.0	-	-	-
21 (reg)	mC	250	67.1	150.7	67.4	90.0	90.0	90.0	-	-	-
22 (reg)	mC	250	150.5	67.1	67.5	90.0	90.1	90.0	-	-	-
23 (reg)	oC	250	67.1	150.5	67.5	90.0	90.0	90.0	-	-	-
24 (reg)	mC	251	150.8	67.4	67.1	90.0	90.0	90.0	-	-	-
25 (reg)	oC	251	67.4	150.8	67.1	90.0	90.0	90.0	-	-	-
26 (reg)	mC	252	67.4	150.8	67.1	90.0	90.0	90.0	-	-	-
27 (reg)	hP	252	67.2	67.2	67.5	90.0	90.0	120.0	-	-	-
28 (reg)	hP	252	67.4	67.4	67.1	90.0	90.0	120.0	-	-	-
29 (reg)	mC	499	150.5	67.1	95.4	90.0	129.3	90.0	-	-	-
30 (reg)	oI	499	67.1	95.4	116.5	90.0	90.0	90.0	-	-	-
31 (reg)	oI	500	67.1	67.4	165.0	90.0	90.0	90.0	-	-	-
32 (reg)	tI	500	67.2	67.2	165.0	90.0	90.0	90.0	-	-	-
33 (reg)	mC	502	95.1	95.1	95.1	90.0	119.9	90.0	-	-	-
34 (reg)	oF	502	95.1	95.1	165.0	90.0	90.0	90.0	-	-	-

Spacegroup: h3

autoindexing - why it goes wrong

- weak diffraction: enough spots? -- needs 50+
- experimental parameters:
  - **beam position (never trust the header!)**
  - wavelength
  - detector distance

  these are almost always indicated by no separation between solution and noise (in penalties)
- multiple lattices
- overlapping spots
- true symmetry lower than metric symmetry
- indexing can be misleading for low symmetry if only from one image

autoindexing - is it OK?

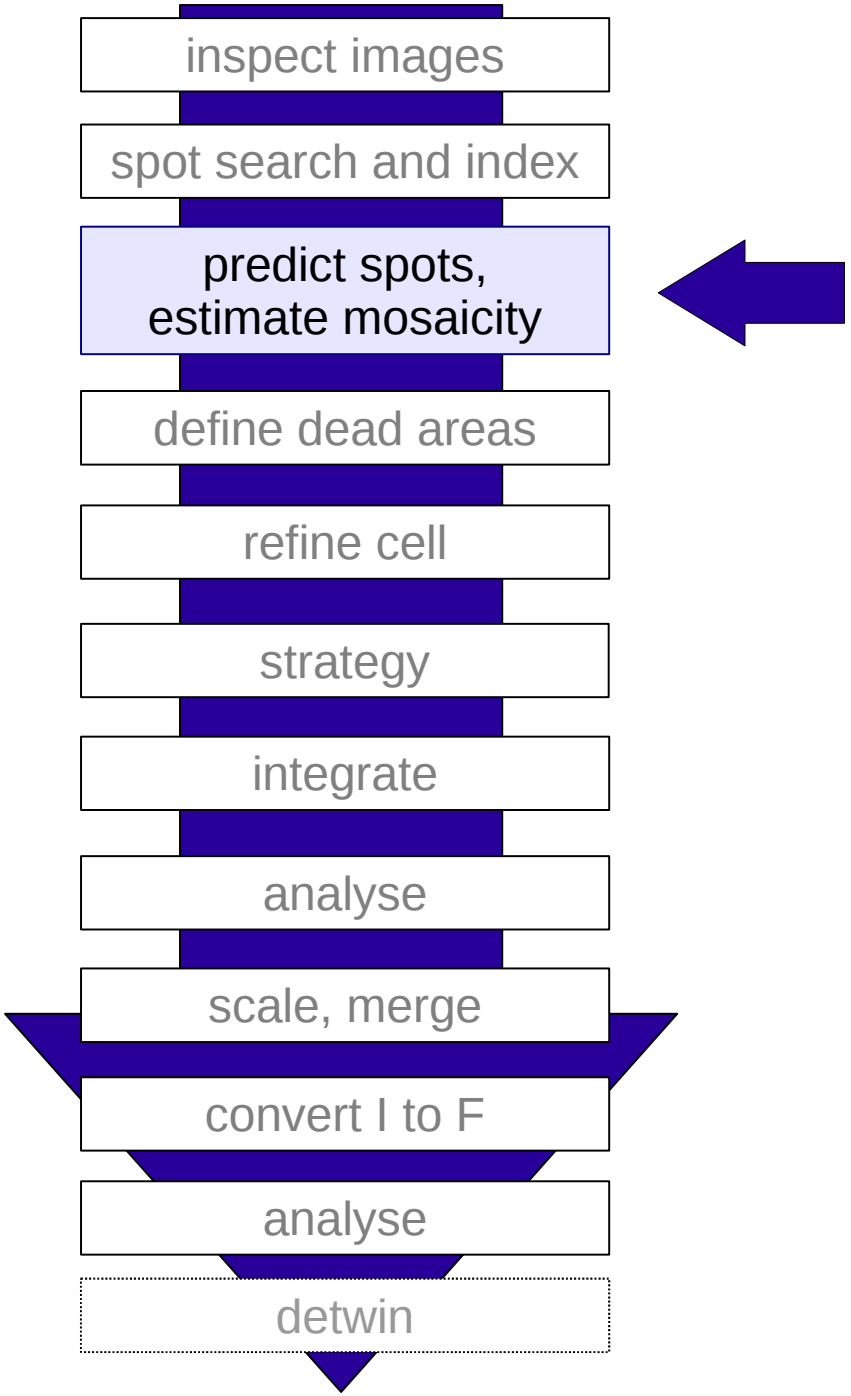
check predicted pattern at different angles:

- does it match?
- are all reflections predicted?

- unpredicted spots:
  - incorrect mosaicity
  - multiple lattices
  - pseudosymmetry

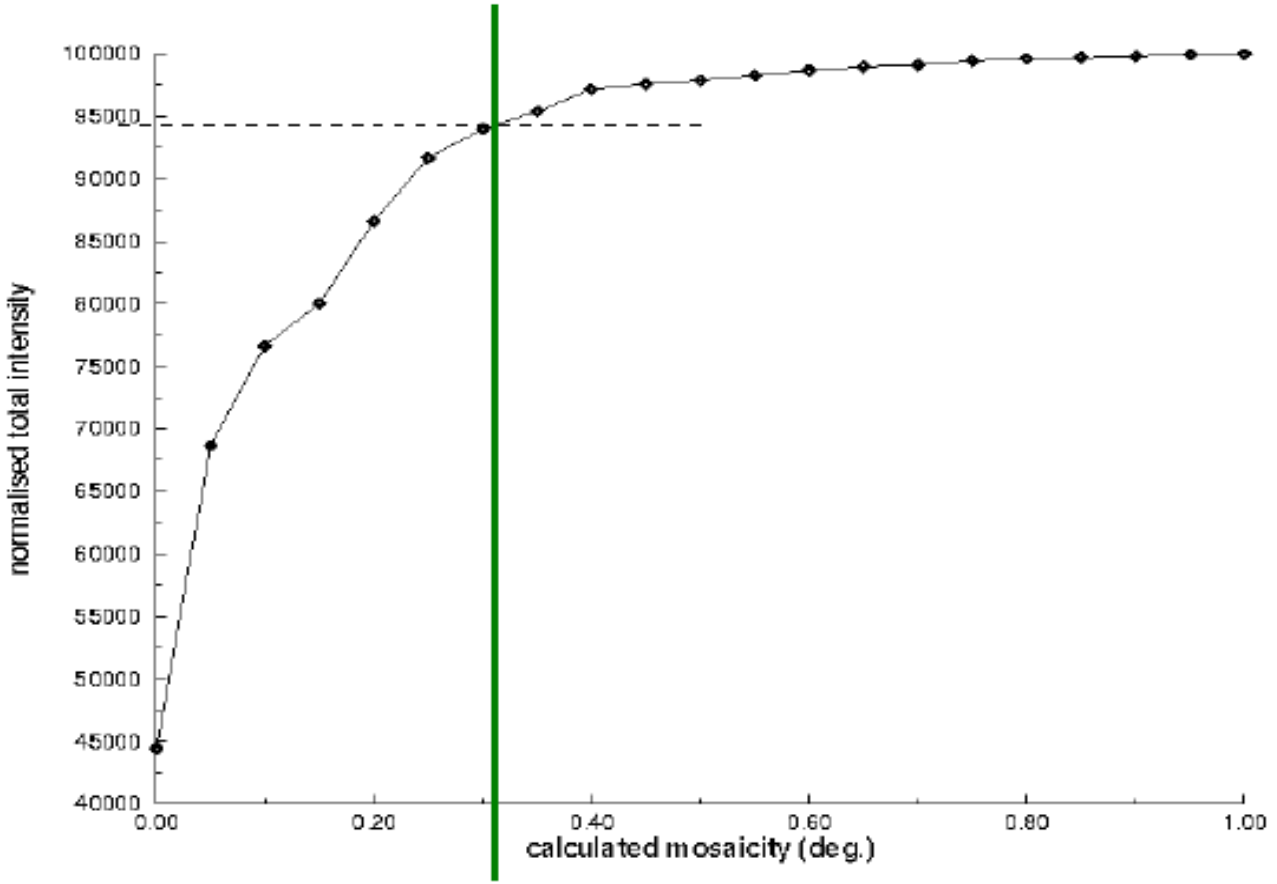
- solutions given by mosflm are the same, but imposing different lattice symmetries.
  In other words, if the P1 solution is wrong, all the others will be wrong too...

data processing flowchart

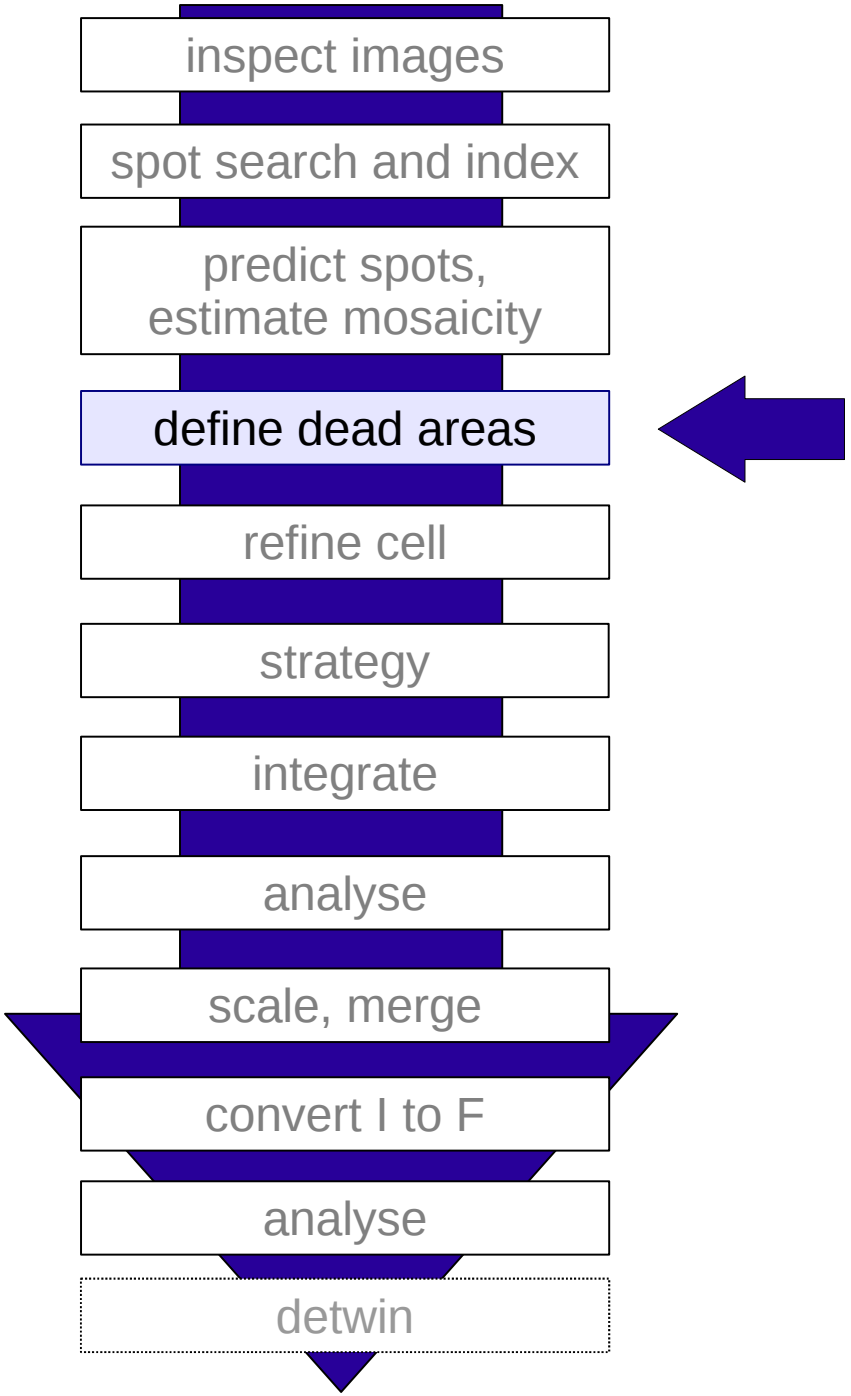


estimating mosaicity

- relevant directly after autoindexing, improve estimate later
- for a series of mosaicity values (0.0, 0.05, 0.1 etc) integrate the total intensity under all predicted spots
- surprisingly accurate (don't be mislead by tails of strong spots)



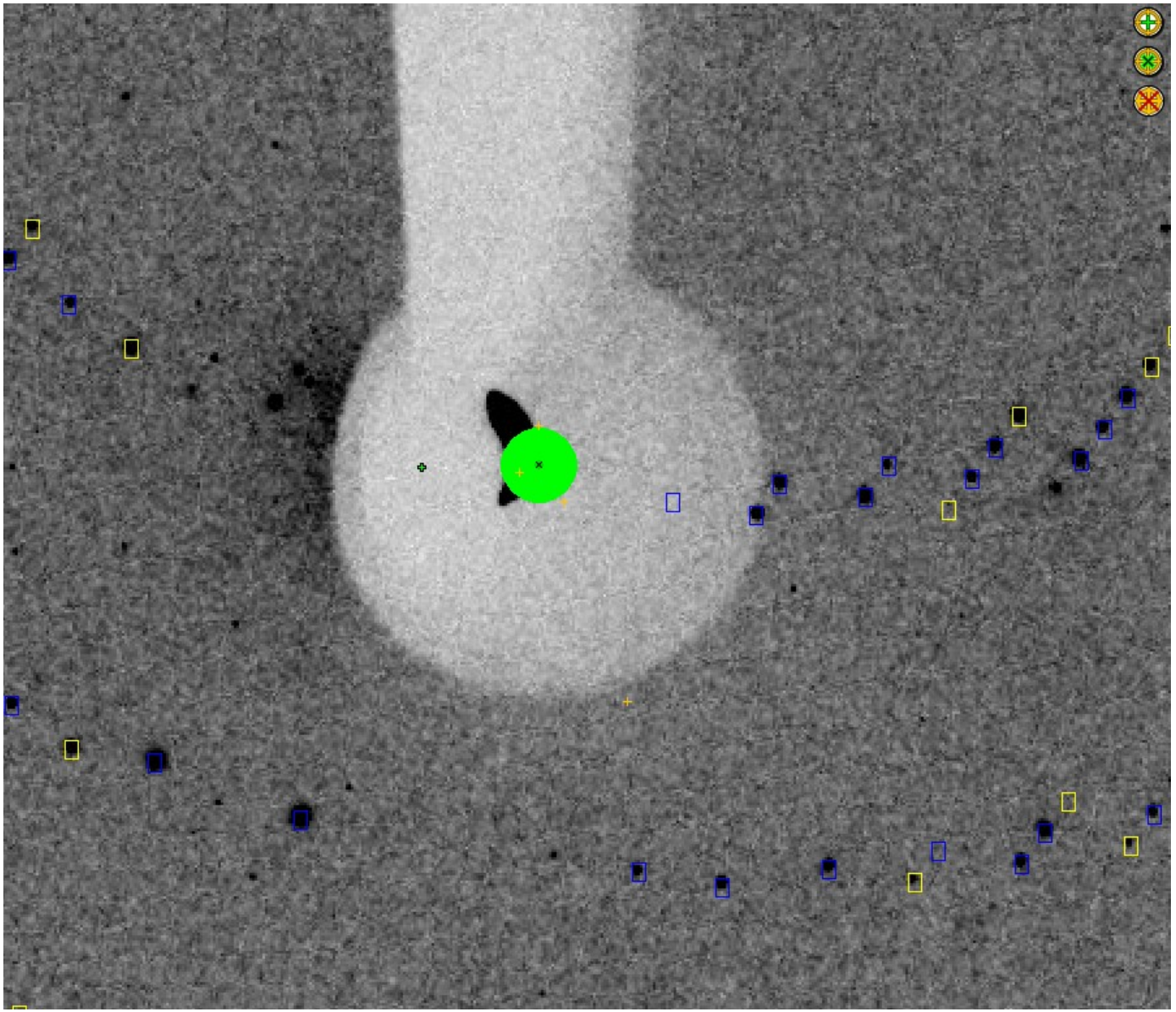
data processing flowchart



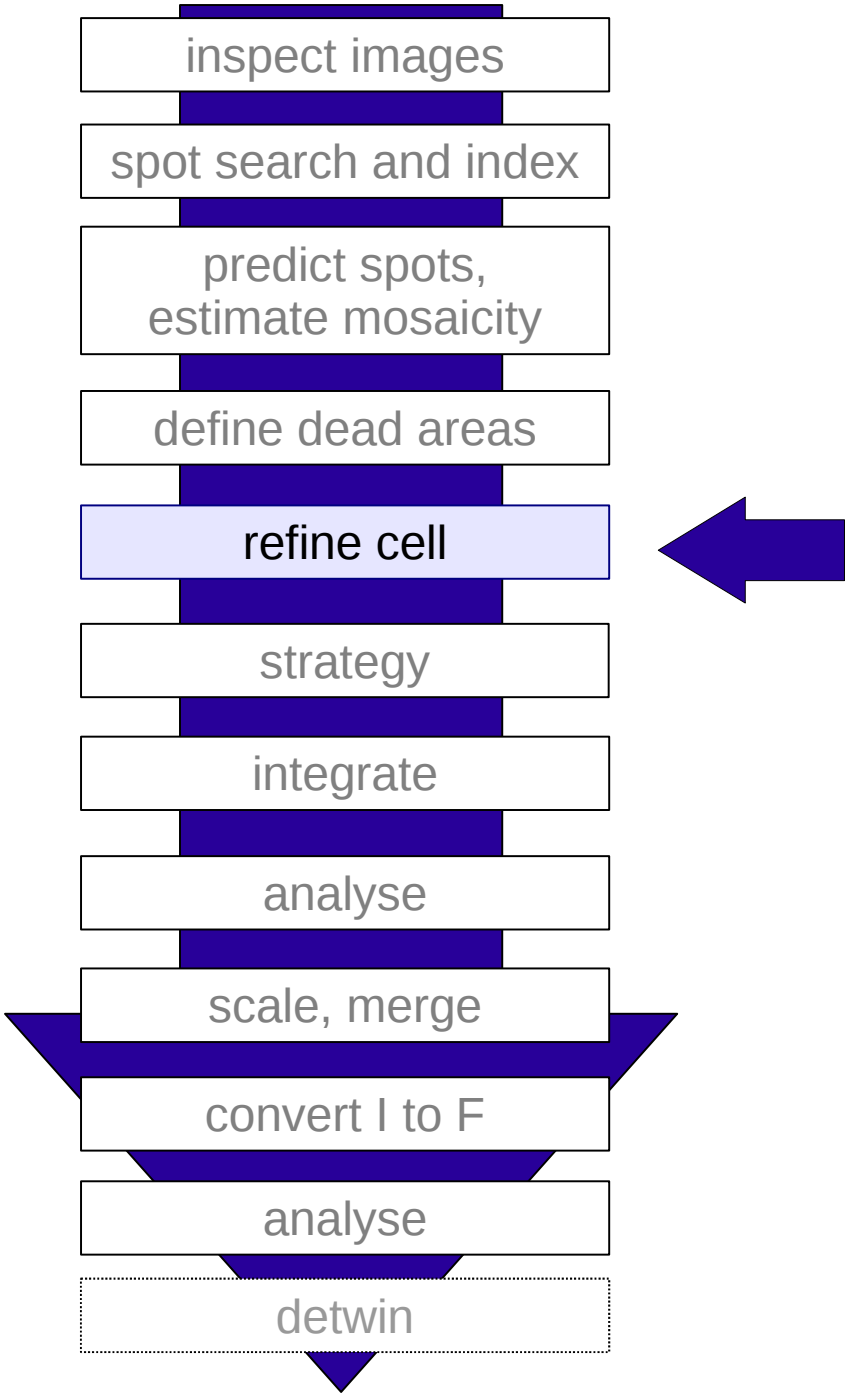
mask dead areas

- otherwise shadowed areas will be processed as active areas!

  - beamstop
  - cryo nozzle



data processing flowchart



cell refinement

- idea:
  - first obtain a good estimate of cell parameters
  - then integrate with fixed cell
- cell refinement
  - on well separated segments
  - each segment should ideally span > 2 x mosaicity
    e.g. 1-5, 91-95
  - collect these before start of data collection for better strategy
- monitor spot profile, residuals

  two kinds of residuals for independent refinement of quantities
  - based on spot coordinates
  - based on spot position in phi

- other programs have different philosophy:
  - HKL2000: postrefinement with all spots
  - XDS: refine cell as we go in integration

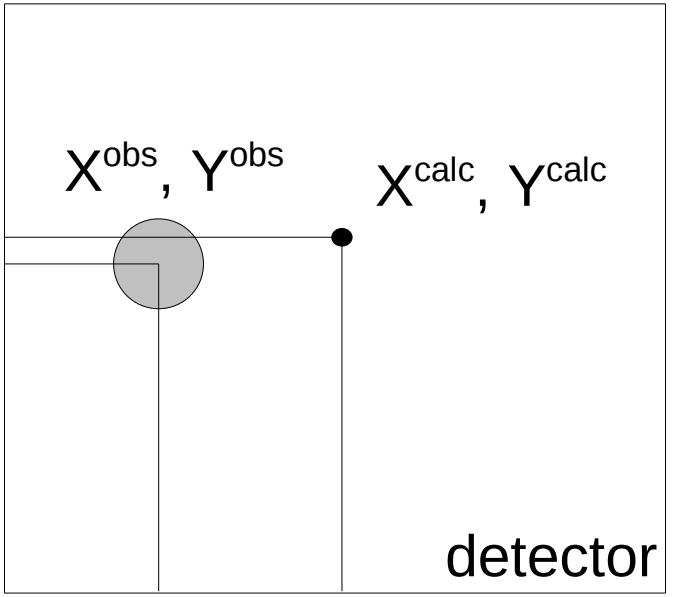
cell refinement

Parameters to refine:

- Crystal parameters
  - cell dimensions
  - orientation
  - mosaic spread
- Detector parameters
  - position
  - orientation
  - distortion parameters
- Beam parameters
  - orientation
  - beam divergence

cell refinement: spot coordinates

- Refinement against spot positions:
  - crystal to detector distance
  - beam centre
  - detector tilts and other miscalibrations
  - X-vsY fudge factor

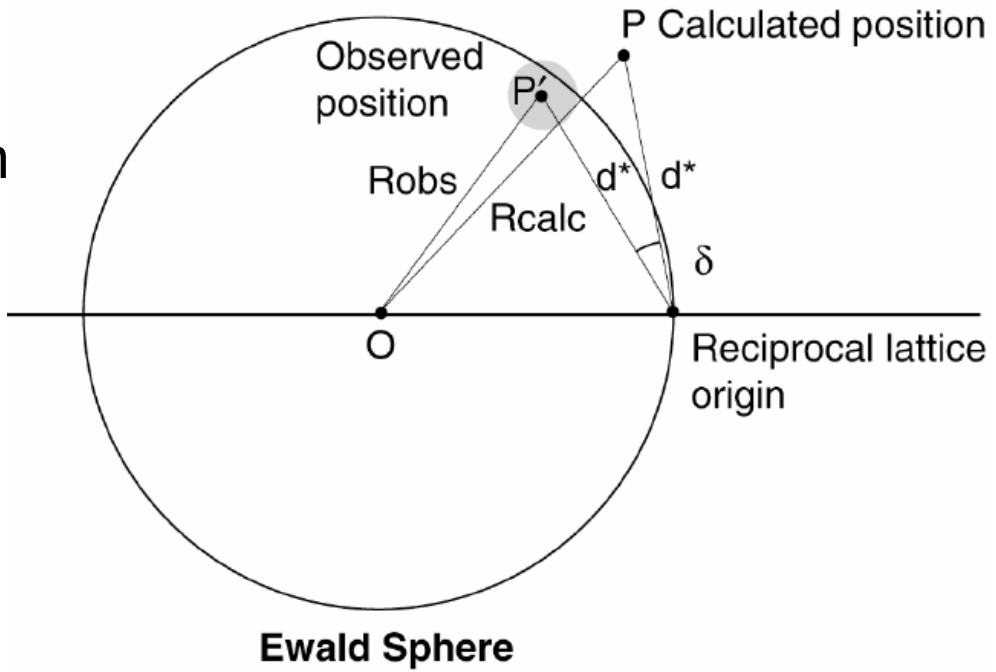


- what to expect:
  < 0.03 : excellent
  ~ 0.08 : still good
  ~ 0.12 : okay if high mosaicity explains it

- residual: $\Omega_1 = \sum_i \omega_{ix}(X_i^{calc} - X_i^{obs})^2 + \omega_{iy}(Y_i^{calc} - Y_i^{obs})^2$

cell refinement: spot *phi* position

- Refinement against spot position
  - cell
  - crystal orientation
  - mosaicity



Ewald Sphere

- residual: $\Omega = \sum \left(\frac{Robs - Rcalc}{d^*} \right)^2$
- partial reflections:
  e.g. spreading over two images with intensity on the two images: I_1 and I_2

  To determine the observed position on the first image from the fraction of the total intensity on that image: $F=I_1/(I_1 + I_2)$

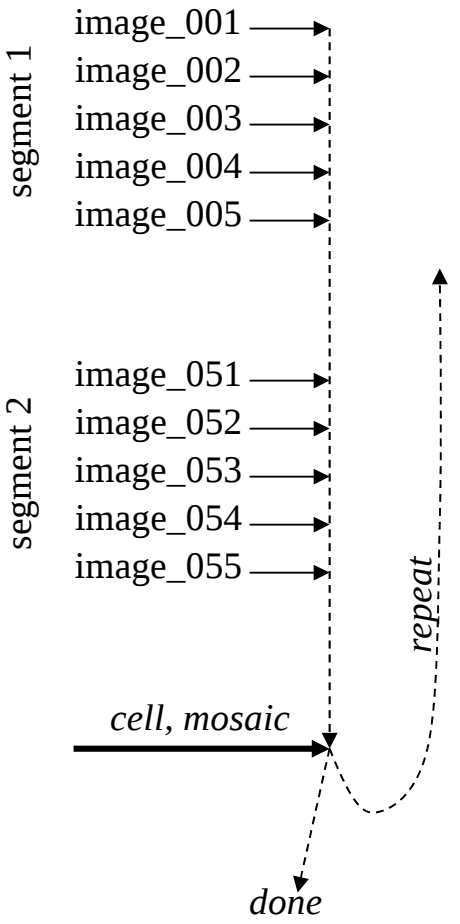
  requires a model for the rocking curve of the spot

$$F = \frac{1}{2}\left(1 \pm Sin\left[\frac{\pi}{2}\frac{\Delta R}{\varepsilon}\right]\right)$$

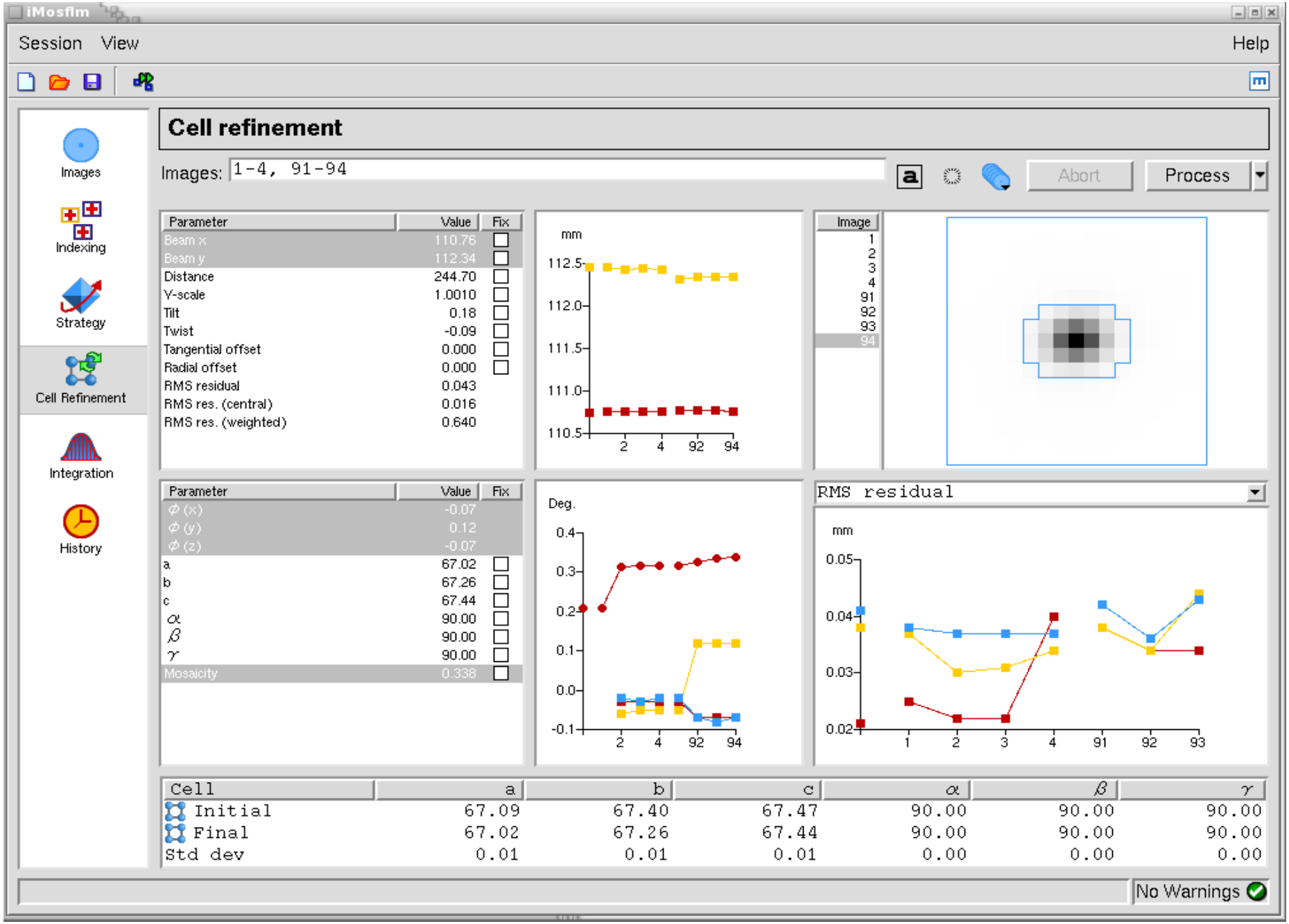
  (epsilon: radius of a reciprocal lattice point)

cell refinement flow

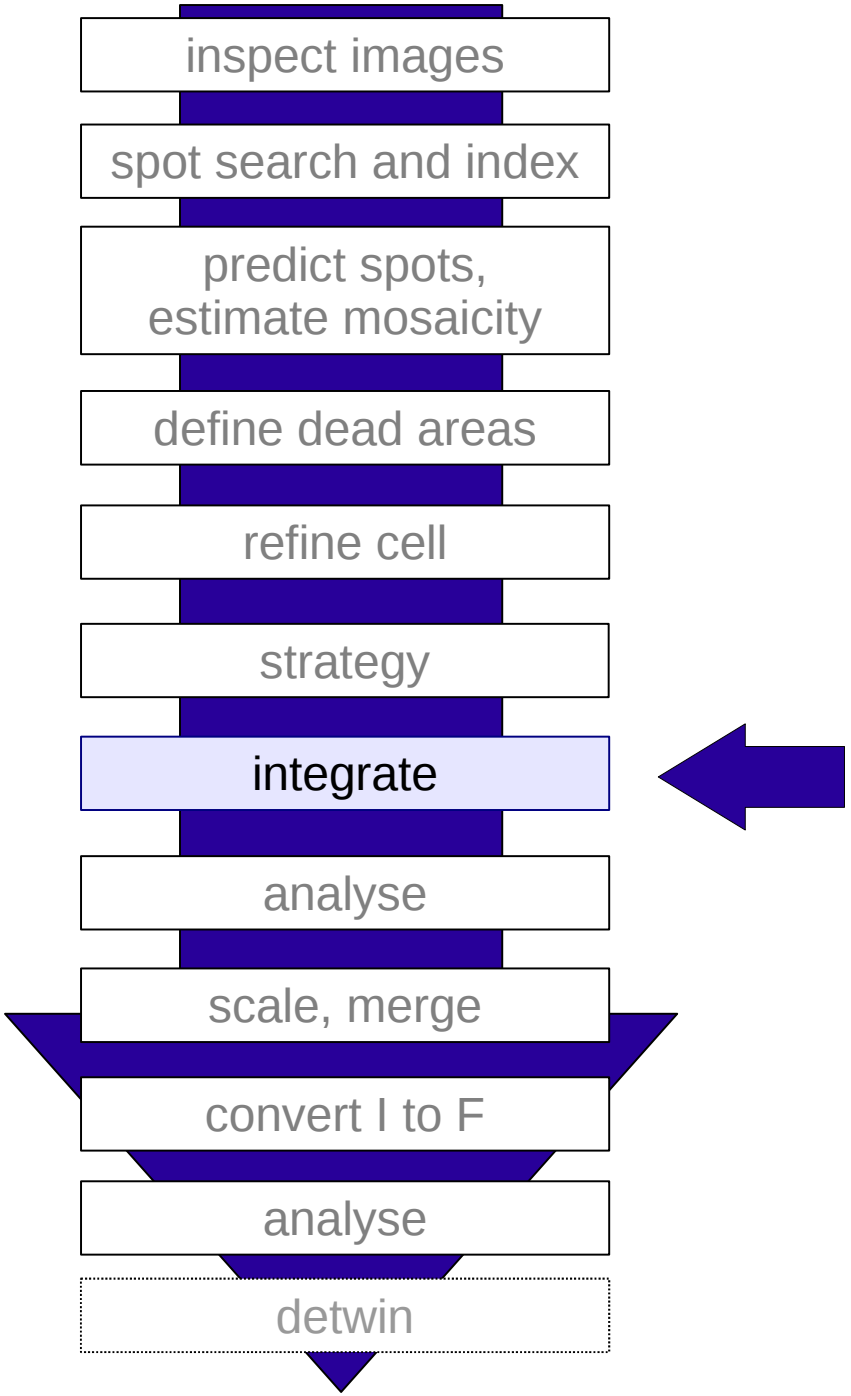
- first orientations are refined for all specified images
  - angular and positional residuals printed for each image
  - spot profile printed
- finally, cell parameters and mosaicity are refined from all images
- if parameters not converged, whole cycle is repeated
- how to judge refinement:
  - values drifted a lot?
  - negative mosaicity? (misindexed?)
  - predictions, profiles



cell refinement flow



data processing flowchart



integration

Two steps:

- Predict the position in the digitised image of each Bragg reflection
- Estimate intensity (substract X-ray background) and determine an error estimate oft the intensity

1. Predicting reflection positions:

- Accuracy in prediction is crucial. Ideally, cell parameters should be known to better than 0.1%.
- Errors in prediction will introduce systematic errors in profile fitting.
- Refined during integration:
  - detector parameters
  - crystal orientation
  - mosaicity
- Cell parameters not refined.

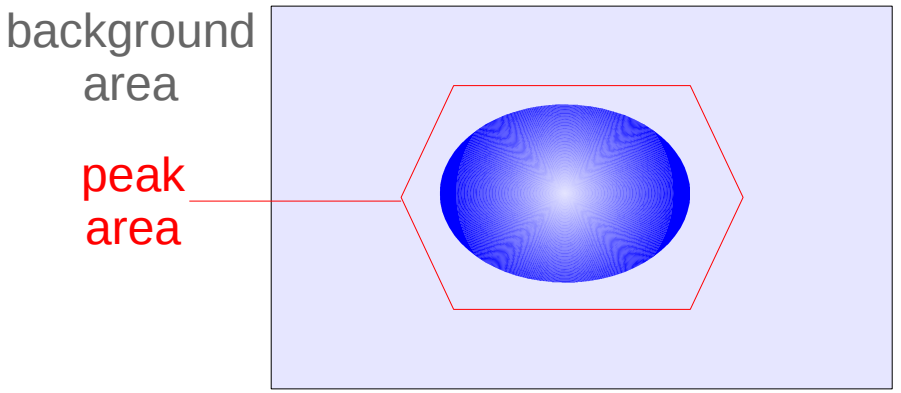
2. Estimating intensity and error:

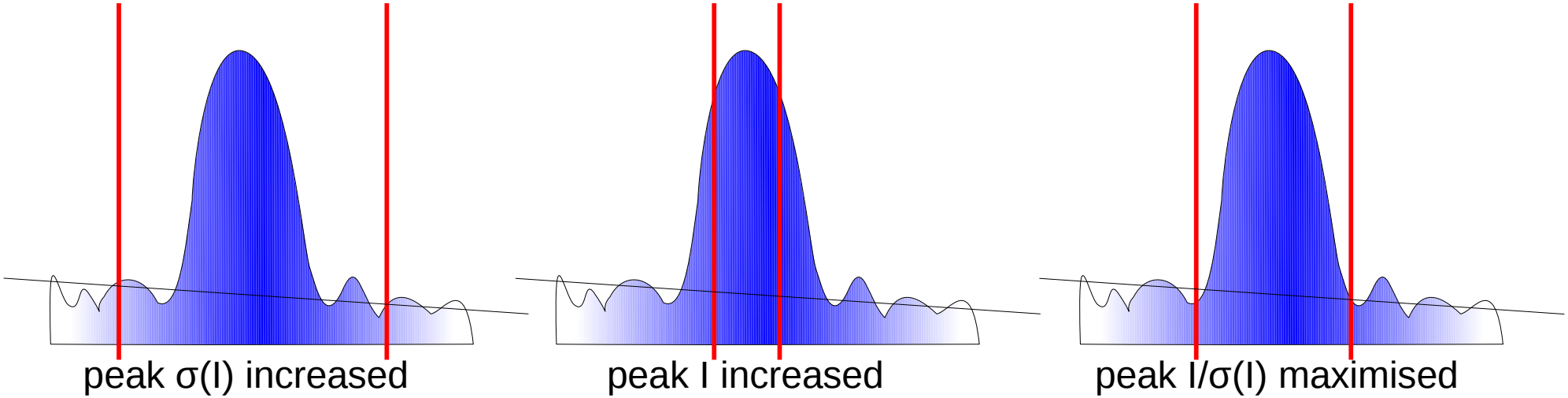
- profile fitting
- summation integration

integration - summation

Summation integration:

- Sum the pixel values of all pixels in the peak area of the mask.
- Subtract the sum of the background values calculated from the background plane for the same pixels.
-

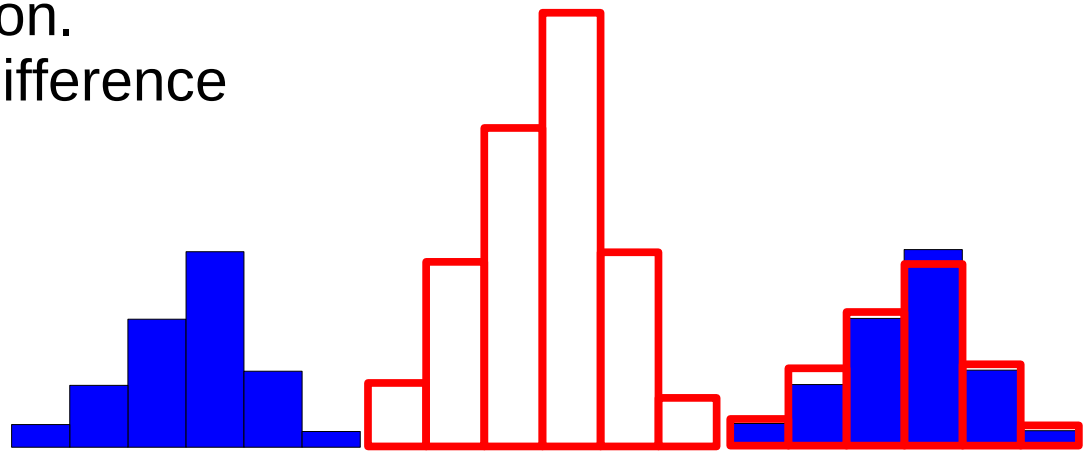




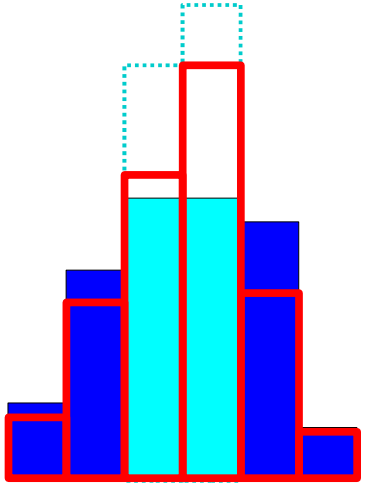
integration - profile fitting

Profile fitting:

- Assume that the shape or profile (2D or 3D) of the spots is known.
- Assume that spot shape does not depend on intensity
- Determine the scale factor, which, when applied to the standard spot profile, gives the best fit to the observed spot profile.
- This scale factor is then proportional to the profile fitted intensity fot the reflection.
- Minimise difference



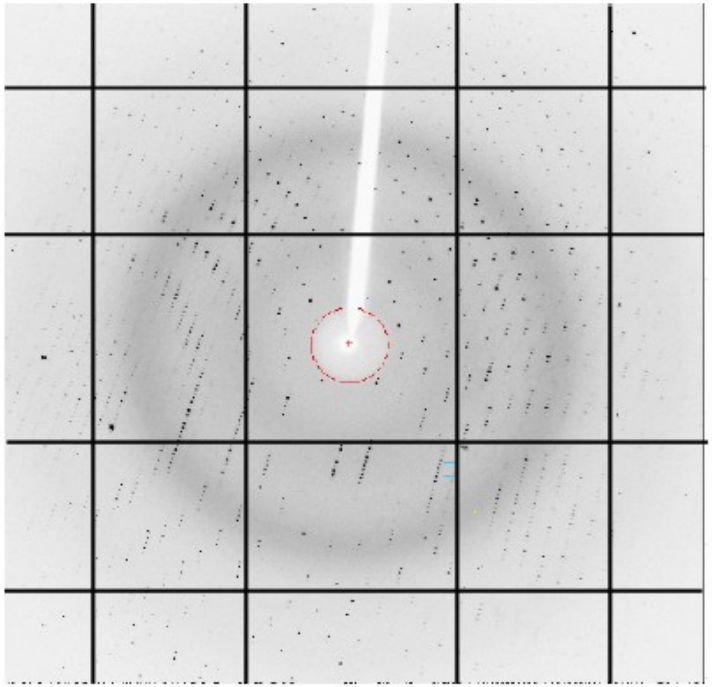
- More reliable intensity estimate of weak reflections (down-weighting peripheral peak pixels)
- Assumption #2 not true for saturated spots.

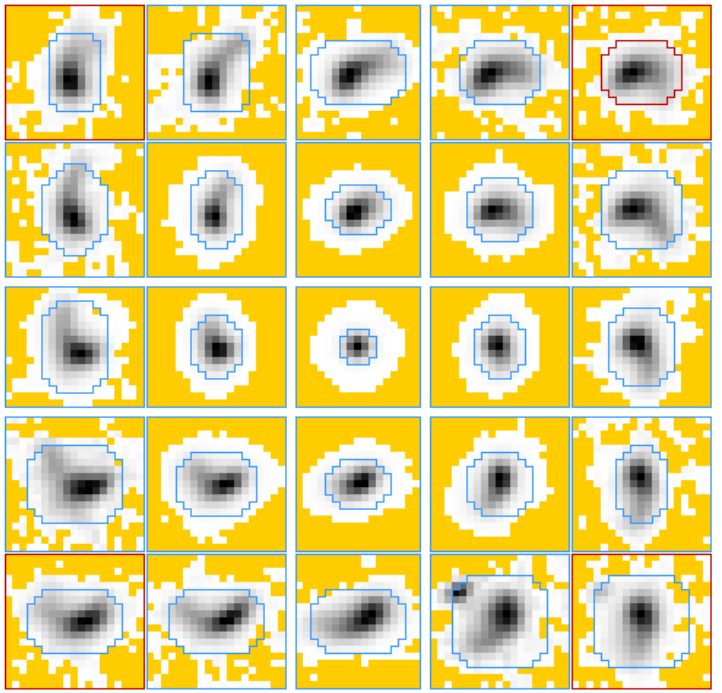


integration - profile fitting

How do we know what the standard profile looks like?

- Split the detector surface into tiles
- Average strong reflections within tile (i.e. different detector areas will have different profiles -- other programs work differently)





- For each reflection a new profile is calculated as a weighted mean of profiles in adjacent regions (weight: scales with distance)
- Used for both full and partial reflections, although not really valid for partials, but in practice just works fine.

integration - standard deviations

Estimation of standard deviations:

- For summation integration or profile fitted partials:
  based on Poison statistics.
- For profile fitted intensities:
  goodness of fit of the scaled standard profile to the true reflection profile

- BUT: these underestimate the true errors, and should be modified accordingly at the merging step so that reflect the actual differences between multiple measurements.

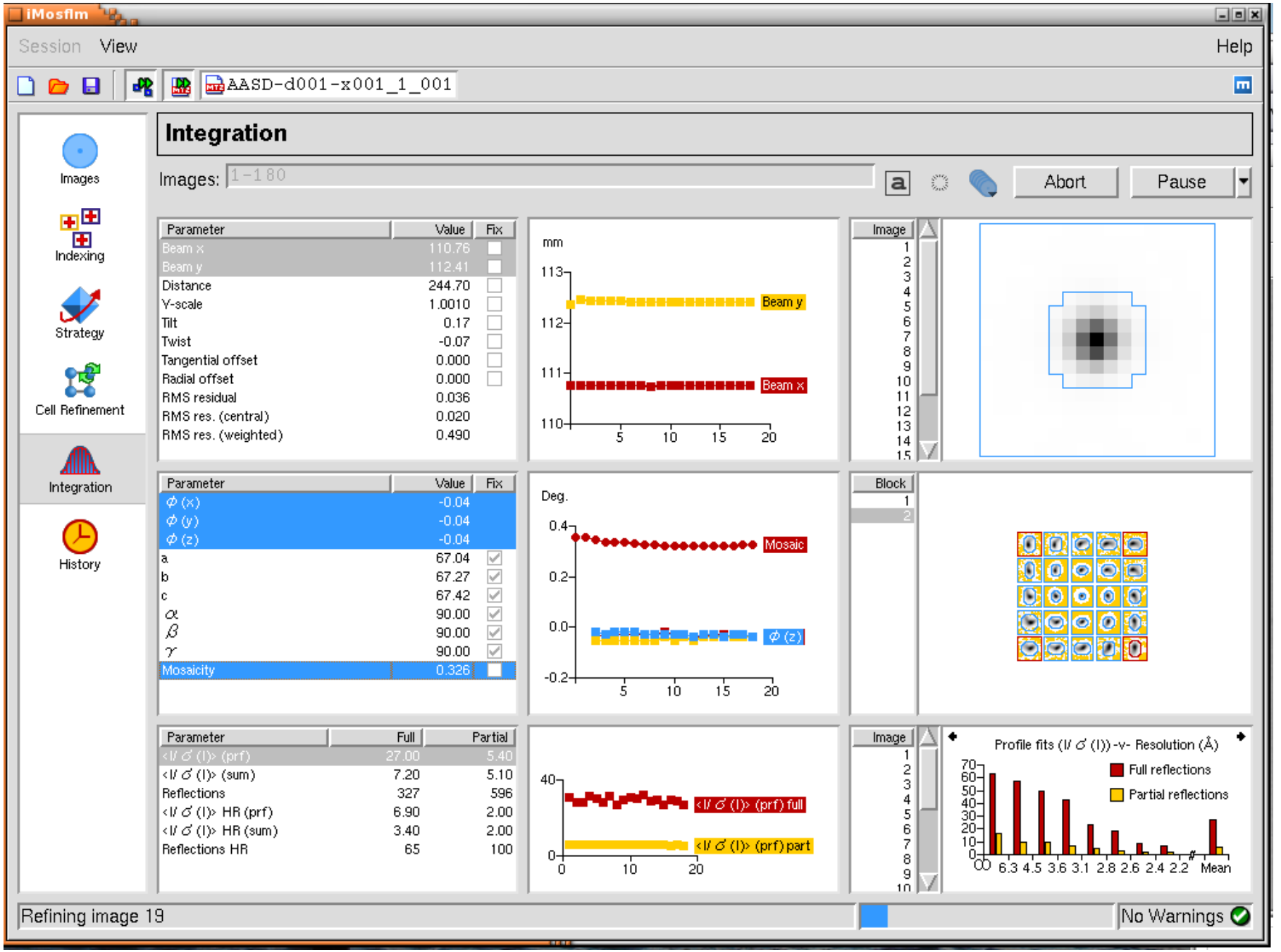
integration in practice

- images integrated in blocks (typically 10 images/block)
- first all images are postrefined with fixed cell
  then images in block are integrated

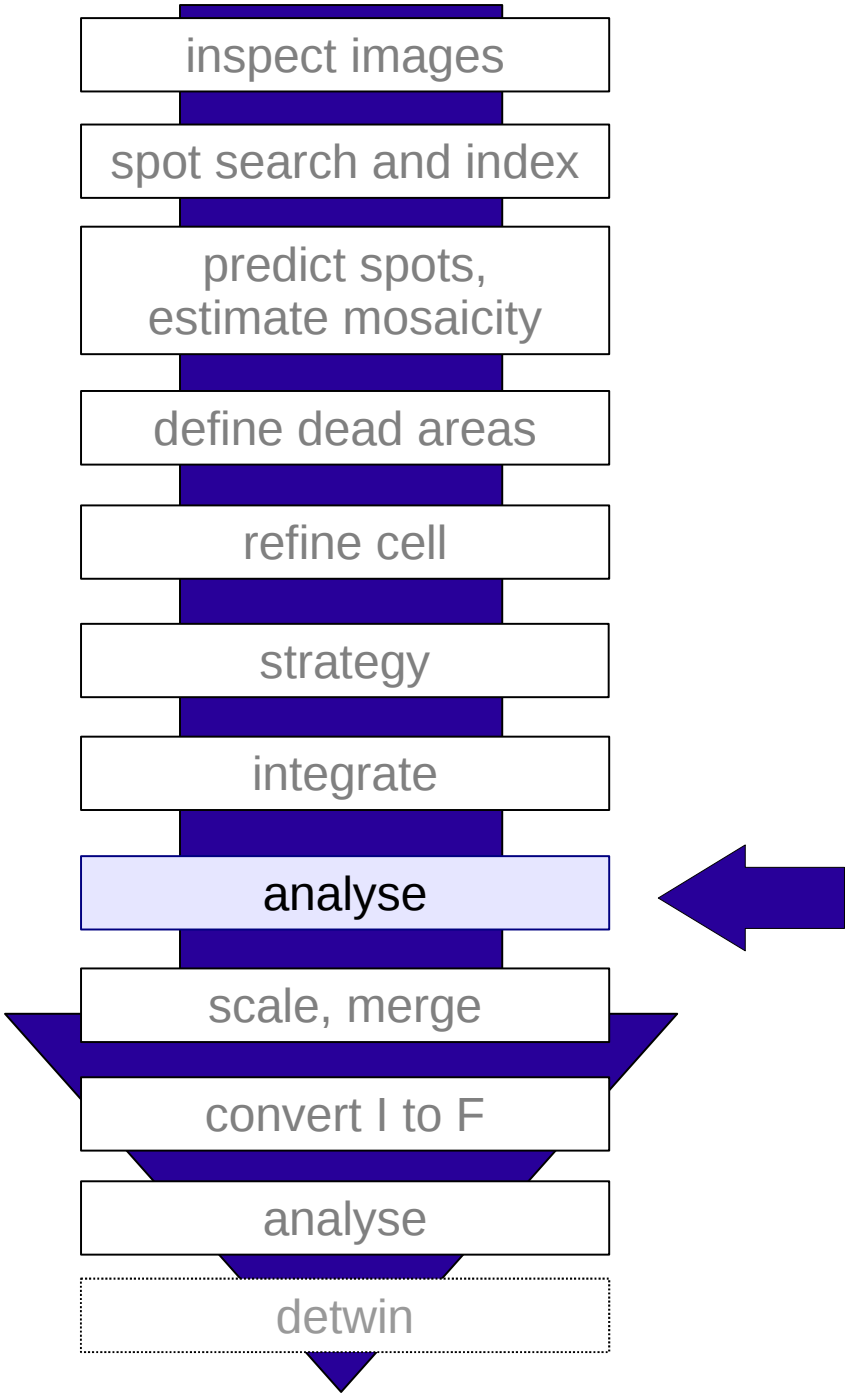
- look at:
  - standard profiles
  - I/sigma per resolution bin
  - warning messages at end of run
  - mosaicity refinement (not on per default)

- most likely warning messages:
  - gain must be adjusted (detector-dependent)
  - YSCALE -- sensitive, can ignore
  - poor profiles, averaging
    -- can be genuine if reflections are weak at edge
  - crystal slippage --

integration in practice



data processing flowchart



space group determination

- The space group is only a hypothesis until the structure is solved!
  - It is hard to distinguish between true (crystallographic) and approximate (NCS) symmetry
  - Metric symmetry can be higher than symmetry of the space group

- It would be great, however, to know the symmetry early to determine the data collection strategy... hmm.

space group determination

- Stages in space group determination

1. Lattice symmetry (symmetry of crystal class)
   - Crystal class imposes restrictions on cell dimensions (spot positions), and this is needed for accurate prediction.
   - Determined at indexing
   - BUT: pseudo-symmetry (metric symmetry higher than true symmetry). E.g. a=b=c α=β=γ=90 is necessarily cubic.

2. Laue group symmetry
   - determined by the symmetry of the diffraction pattern (spot positions and intensities)
   - corresponds to the space group without translations with an added centre of symmetry (from Friedel's law)

space group determination

- Stages in space group determination (cont.)

3. Point group symmetry
   - For chiral space groups (like macromolecular crystals), there is only one point group corresponding to each Laue group.
   - It is the space group without any translations.

4. Space group symmetry
   - point group + translations (screw axes etc)
   - translations are only visible as systematic absences, usually along axes, which are not very reliable indicators -- there are only a few axial reflections and chances are that we don't record them. Also: accidental absences.

space group determination

Systematic absences:

- For a P_q axis along e.g. c (index l), axial reflections are only present if
  l = n(p/q) where n is an integer

- examples:

2_1	2n	2, 4, 6, ...
3_1	3n	3, 6, 9, ...
4_1, 4_3	4n	4, 8, 12, ...
4_2	2n	2, 4, 6, ...
6_1, 6_5	6n	6, 12, 18, ...
6_2, 6_4	3n	3, 6, 9, ...
6_3	2n	2, 4, 6, ...

But we may not record many axial reflections!

space group determination

Systematic absences (cont.)

- centering causes systematic absences too, but these conditions apply to all h, k, l (i.e. not to axial reflections only)

A centered	$k + l = 2n$
B centered	$h + l = 2n$
C centered	$h + k = 2n$
F centered	$k + l = 2n$, $h + l = 2n$, $h + k = 2n$
I centered	$h + k + l = 2n$
R (obverse)	$-h + k + l = 3n$
R (reverse)	$h - k + l = 3n$

pointless

Program to determine Laue group from unmerged intensities.
Strategy used in pointless:

- From unit cell dimensions, find the highest compatible lattice symmetry
- Score each symmetry element (rotation) belonging to lattice symmetry using all pairs of observations related by that element,
  Score based on correlation coefficient, which is relatively independent of unknown scales.
- Score combinations of symmetry elements of all possible sub-groups (Laue-groups) of lattice symmetry group.
- Score possible space groups from axial systematic absences.

pointless example

Example: a false diamond

- data integrated in P1
- cell: 67.01 67.31 67.43 90.11 89.96 90.05
  whoa, looks like cubic....
  except that integration is difficult and scaling fails
- ... and indeed, pointless says:

  `Lattice point group: P 4 3 2`

- OK, let's score the symmetry elements!

pointless example (cont.)

- Scoring each symmetry element for P432:

likelihood — correlation-coefficient on E^2

Z-score — R-factor

Nelmt	Lklhd	Z-cc	CC	N	Rmeas		Symmetry & operator (in Lattice Cell)		
1	0.933	8.93	0.89	23366	0.148		identity		
2	0.044	0.77	0.08	45821	0.943		2-fold (1 0 1)	{+l,-k,+h}	
3	0.044	1.00	0.10	45091	0.906		2-fold (1 0-1)	{-l,-k,-h}	
4	0.055	-0.08	-0.01	45251	0.998		2-fold (0 1-1)	{-h,-l,-k}	
5	0.050	0.11	0.01	45053	0.979		2-fold (0 1 1)	{-h,+l,+k}	
6	0.045	0.53	0.05	45050	0.943		2-fold (1-1 0)	{-k,-h,-l}	
7	**0.898**	**8.62**	**0.86**	**44891**	**0.248**	**	**2-fold k (0 1 0)**	**{-h,+k,-l}**	
8	0.045	0.47	0.05	45125	0.959		2-fold (1 1 0)	{+k,+h,-l}	
9	**0.866**	**8.45**	**0.84**	**44834**	**0.260**	**	**2-fold h (1 0 0)**	**{+h,-k,-l}**	
10	**0.883**	**8.53**	**0.85**	**44874**	**0.273**	**	**2-fold l (0 0 1)**	**{-h,-k,+l}**	
11	0.056	-0.12	-0.01	91238	1.568		3-fold (1 1 1)	{+l,+h,+k}	{+k,+l,+h}
12	0.058	-0.21	-0.02	91364	1.614		3-fold (1-1-1)	{-l,-h,+k}	{-k,+l,-h}
13	0.053	-0.01	-0.00	91229	1.556		3-fold (1-1 1)	{+l,-h,-k}	{-k,-l,+h}
14	0.056	-0.14	-0.01	91231	1.565		3-fold (1 1-1)	{-l,+h,-k}	{+k,-l,-h}
15	0.055	-0.10	-0.01	91251	1.132		4-fold h (1 0 0)	{+h,-l,+k}	{+h,+l,-k}
16	0.044	0.65	0.07	91270	1.066		4-fold k (0 1 0)	{+l,+k,-h}	{-l,+k,+h}
17	0.046	0.37	0.04	91221	1.093		4-fold l (0 0 1)	{-k,+h,+l}	{+k,-h,+l}

Ooops, only the orthorhombic symmetry operators are present!

pointless example (cont.)

- Laue group determination:

likelihood

correlation-coefficient on E^2

Z-score

R-factor

	Laue Group		Lklhd	NetZc	Zc+	Zc-	CC	CC-	Rmeas	R-	Delta	ReindexOperator
> 1	**P m m m**	*******	**0.944**	**8.38**	**8.62**	**0.24**	**0.86**	**0.02**	**0.23**	**1.14**	**0.1**	**[h,k,l]**
2	P 1 2/m 1		0.019	7.46	8.78	1.31	0.88	0.13	0.19	0.96	0.1	[-h,-k,l]
3	P 1 2/m 1		0.017	7.40	8.73	1.33	0.87	0.13	0.20	0.96	0.1	[-h,-l,-k]
4	P 1 2/m 1		0.014	7.36	8.68	1.33	0.87	0.13	0.20	0.96	0.1	[-k,-h,-l]
- 5	P -1		0.002	7.17	8.93	1.77	0.89	0.18	0.15	0.89	0.0	[-h,k,-l]
6	P 4/m		0.001	4.66	6.03	1.36	0.60	0.14	0.36	0.95	0.4	[l,h,k]
7	P 4/m		0.001	4.27	5.70	1.43	0.57	0.14	0.38	0.95	0.1	[k,l,h]
8	P 4/m		0.001	4.46	5.87	1.40	0.59	0.14	0.38	0.95	0.3	[h,k,l]
9	P 4/m m m		0.000	4.56	4.88	0.32	0.49	0.03	0.47	1.18	0.1	[k,l,h]
10	H -3		0.000	2.41	4.31	1.90	0.43	0.19	0.53	0.86	0.4	[-k+l,h+k,-h+k+l]
11	H -3		0.000	2.44	4.34	1.90	0.43	0.19	0.52	0.86	0.4	[-h+k,h+l,h+k-l]
12	H -3		0.000	2.46	4.36	1.89	0.44	0.19	0.52	0.86	0.4	[k-l,-h+l,h+k+l]
13	C 1 2/m 1		0.000	2.53	4.42	1.89	0.44	0.19	0.42	0.88	0.1	[k+l,-k+l,h]
14	H -3		0.000	2.52	4.41	1.89	0.44	0.19	0.52	0.86	0.4	[h-l,k+l,h-k+l]

Clear preference for primitive orthorhombic.

(Likelihood allows for the possibility of pseudo-symmetry.)

pointless example (cont.)

- Translational symmetry: from systematic absences

```
        Zone                 Number  PeakHeight  SD  Probability
  ReflectionCondition

1   screw axis 2(1) [a]        40   1.081   0.124   *** 0.968   h00: h=2n
2   screw axis 2(1) [b]        41   1.050   0.089   *** 0.984   0k0: k=2n
3   screw axis 2(1) [c]        37   1.086   0.143   *** 0.960   00l: l=2n
```

- There's only one primitive orthorhombic space group with these elements:

```
 Spacegroup          TotProb SysAbsProb Reindex   Conditions

<P 21 21 21> ( 19)    0.863  0.914         h00: h=2n, 0k0: k=2n, 00l: l=2n (zones 1,2,3)
  ..........
 <P 21 21 2> ( 18)    0.036  0.038         h00: h=2n, 0k0: k=2n (zones 1,2)
  ..........
 <P 2 21 21> ( 18)    0.029  0.030         0k0: k=2n, 00l: l=2n (zones 2,3)
  ..........
 <P 21 2 21> ( 18)    0.014  0.015         h00: h=2n, 00l: l=2n (zones 1,3)
```

alternative indexing

- Alternative, valid but non-equivalent indexing schemes are possible in space groups where the point group symmetry is lower than that of the lattice group.
  These are related by symmetry operators present in the crystal system but not in the point group.

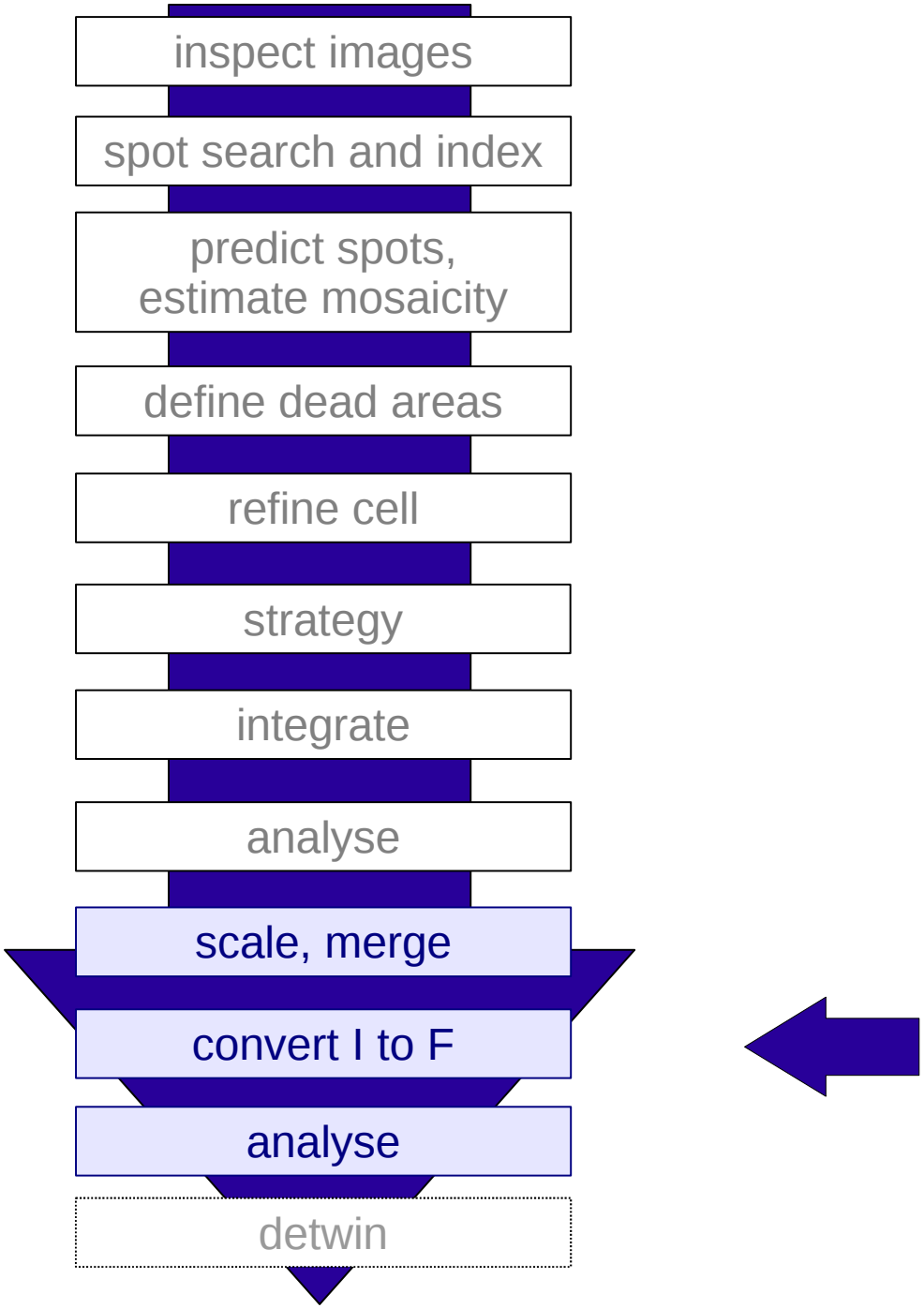
- tetragonal, trigonal, hexagonal, cubic

  e.g. in P3 there are four different ways to index:
  (h, k, l)
  (-h, -k, -l)
  (k, h, -l)
  (-k, -h, -l)

  For the first dataste, you can choose any of these.

  For subsequent datasets, autoindexing might pick an alternative one and you have to fix it in order to scale them together.
  Pointless will do this for you by comparing the unmerged test data to a merged reference dataset.

data processing flowchart



scaling, merging - why?

- Various physical factors lead to observed intensities being on different scales.
  - Some corrections are know (e.g. Lorentz and polarisation corrections.
  - Others can only be determined from the data.

- Scaling parameters should match the experiment

- Factors influencing scaling:
  1. Incident beam and camera
  2. Crystal and diffracted beam
  3. Detector

- Merging: all observations of each unique reflection to a single I_{mean}
  - reject outliers
  - estimate standard deviations
  - calculate statistics

scaling - 1. incident beam

- beam intensity:
  - variable on synchrotrons (beam decay, esp short half life beams)
  - warming-up effects of in-house optics

- illuminated volume might change with phi
  if beam smaller than crystal (synchrotrons)

- absorption of primary beam by crystal

- rotation speed variations
  - difficult to detect, therefore we assume that rotation speed is constant.
  - can be disastrous for data quality

- shutter synchronisation errors

scaling - 2. crystal and diffraction

- absorbtion in secondary beam
  - significant at long wavelengths (CuKα or long wavelength SAD/MAD)

- radiation damage
  - synchrotron, or even high intensity in-house sources
  - not easily correctable, as structural changes are involved.
  - approaches: zero-dose extrapolation
    low-dose plus high redundancy data collection

scaling - 3. detector

- detector errors are difficult to detect

- detectors should be properly calibrated:
  - spacial distortion
  - sensitivity of response (floodfield)

- Dead area should be excluded at integration stage
  - including bad pixels, ideally
  - shadows of beamstop, cryo nozzle

scaling - scaling function

- scales are determined by comparison of symmetry-related reflections.
  - gives internal consistency of intensities.
  - However, we do no know the true intensities and hence internally consistent data does not mean that it is correct.
  - This might lead to systematic errors.

- Scaling function:

  g = g (rotation/image number) g (time) g (s) ...and others

  primary beam B-factor absorbtion e.g. tails

  scale is smooth function of spindle rotation

  A discontinuous function of is usually less appropriate

  time dependent radiation damage correction

scaling - scaling models

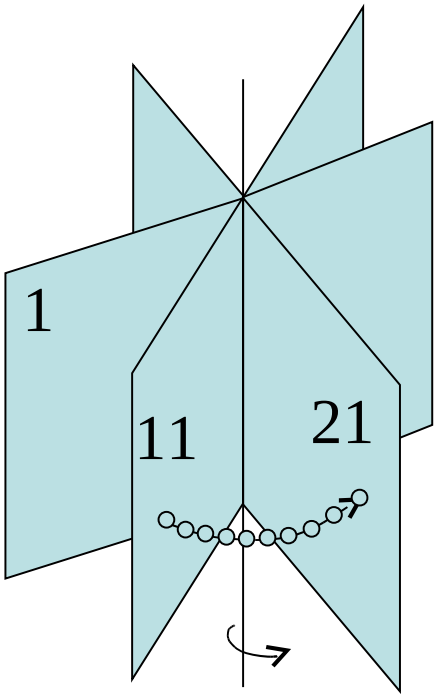
- Batch scaling
  - each frame gets separate scale
  - slow for many images
- Smooth scaling
  - extrapolate between images
- Local scaling
  - if reference dataset exists
  - higher reliability of outlier rejection
  - e.g. scale all wavelengths of MAD experiment together helps to extract weak signal
  - approximation of f'' and relative Δf'

Single B-factor

- Smooth B-factor
  - equivalent of smooth scales
- Anisotropic B-factor

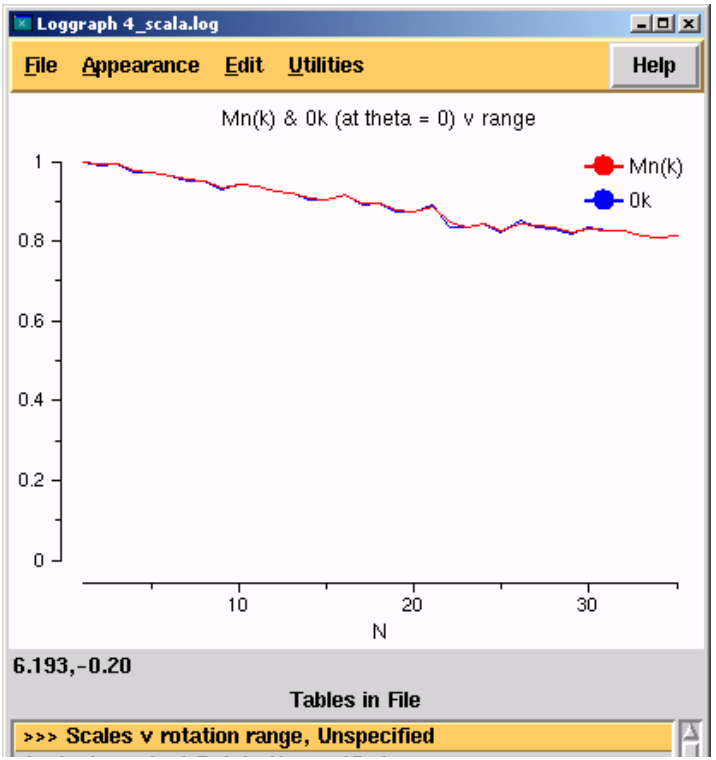
- Secondary beam correlation
  - in spherical harmonics
- Absorption correction

scaling - smooth scaling

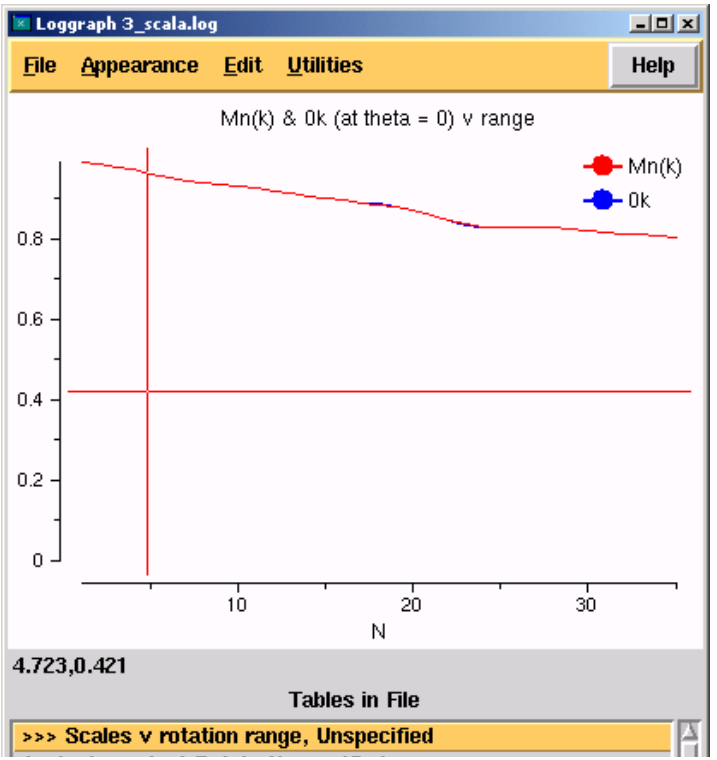


- determine scale factor for every n^{th} image
- interpolate between images
- fast (few scale factors to compute)

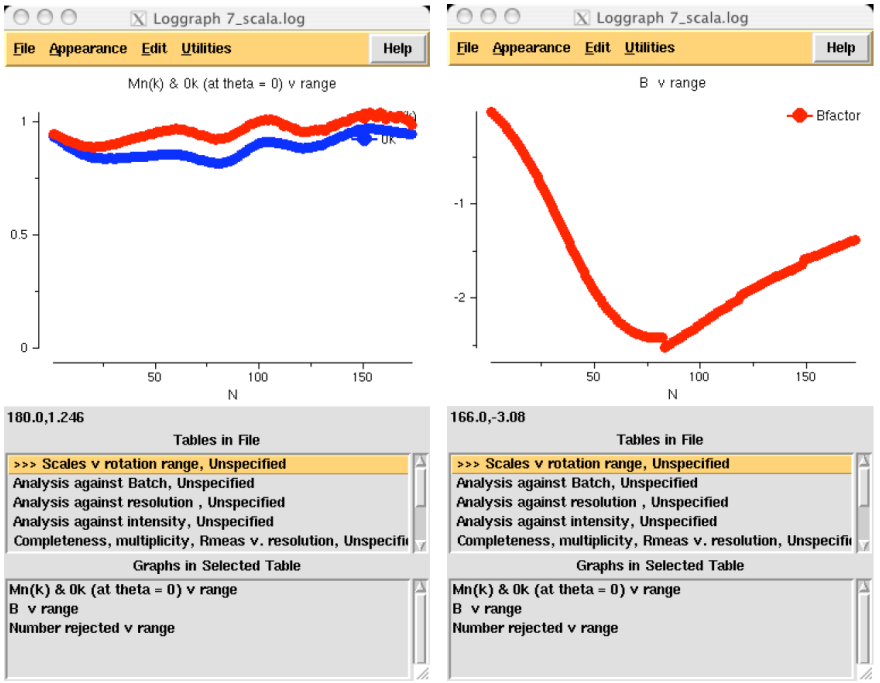
batch scales



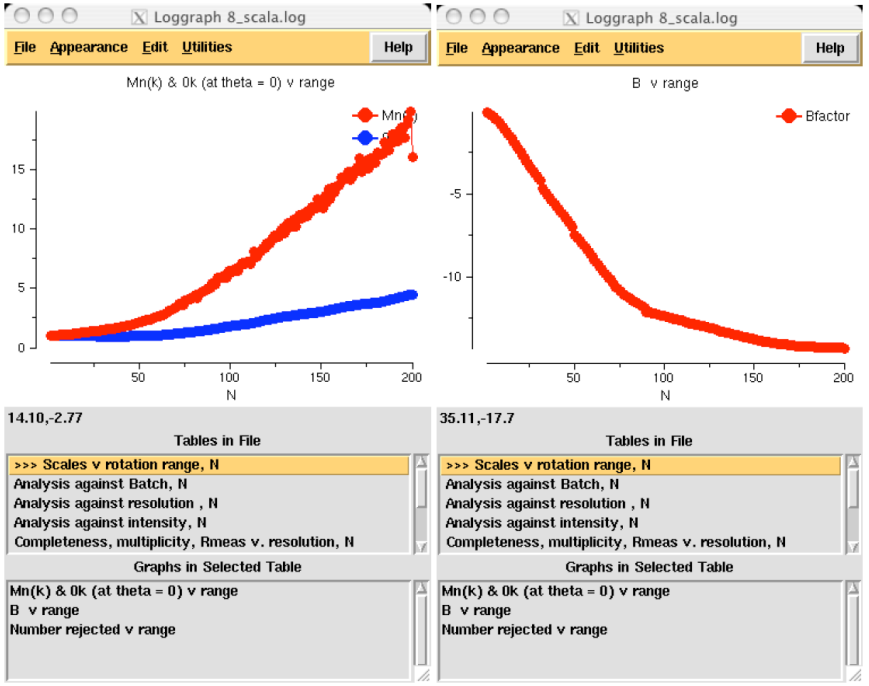
smooth scales



scaling - B-factors



- dataset with little radiation decay
- small B-factor variation

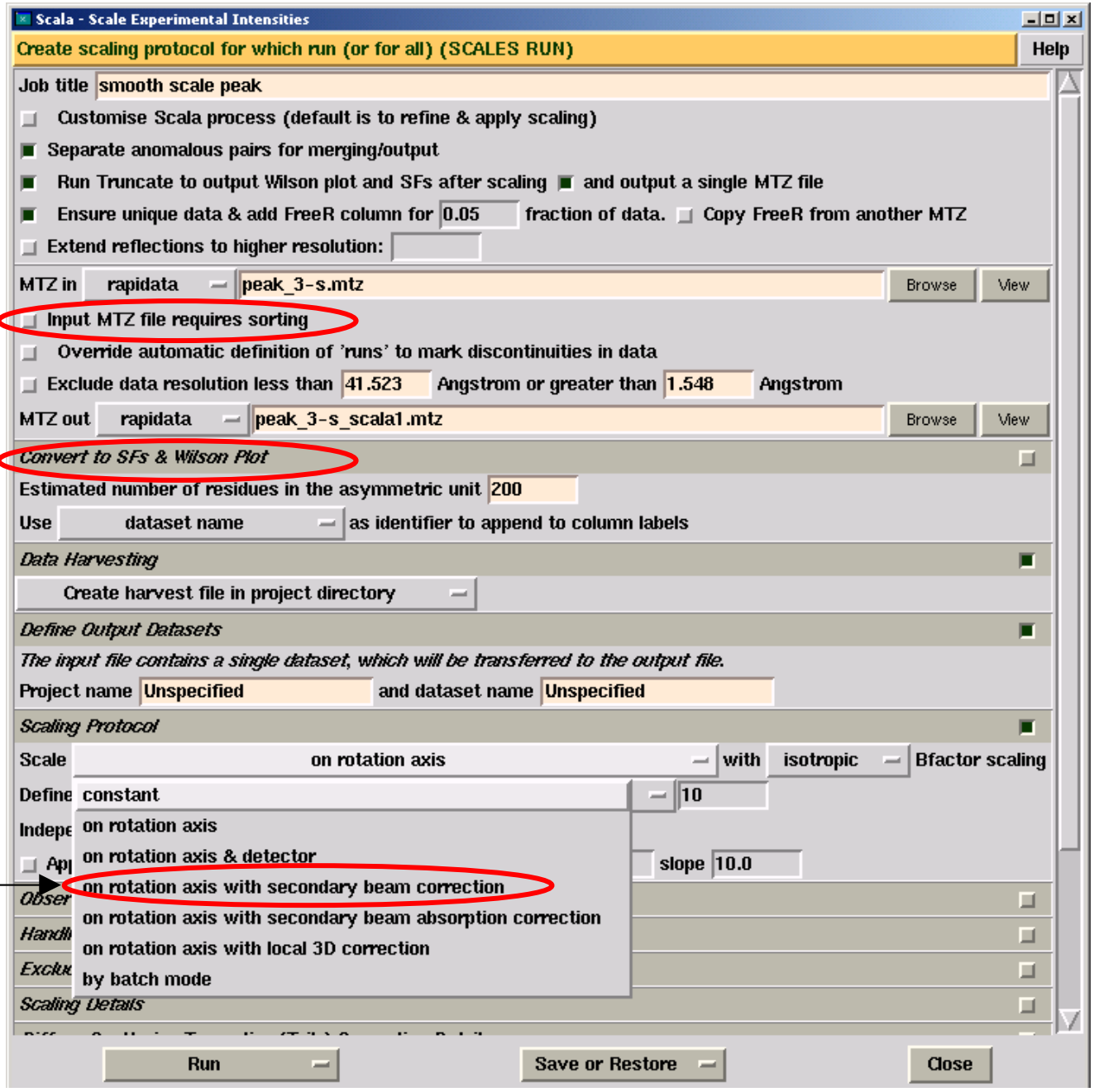


- dataset with severe radiation decay
- large B-factor drop (larger than 10Å^2)

scaling - in practice

For single input mtz

Smooth scaling



scaling - quality indicators

- Determination of scaling parameters depends on symmetry related reflections having different scales. If all observations of a reflection have the same value of the scale component, then there is no information about that component and it remains as a systematic error in the merged data (e.g. absorbtion)

- Thus, to get intensities with the lowest absolute error, the symmetry-related observations should be measured in as different way as possible (e.g. rotation about multiple axes, different slicing). This will increase the merging R, but improve the esimate of I.

- However, to measure the most accurate difference for phasing (e.g. anomalous experiments), observations should be measured in as similar way as possible...

scaling - quality indicators

Questions you might want to ask:

- Overall quality -- How does this dataset compare to others in the project?
- What is the resolution?
- Are there bad batches that should be excluded?
- Is there significant radiation damage?
- Is outlier rejection working well? Are outliers arranged in any systematic way?
- Is there anomalous signal?

scaling - R-factors

R-factors: traditional measures of overall quality

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

The traditional measure of agreement. But: it increases with higher multiplicity even though data are better

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

The multiplicity-weight R-factor allows for the improvement of data with higher multiplicity. Useful when comparing different point-pointgroups.

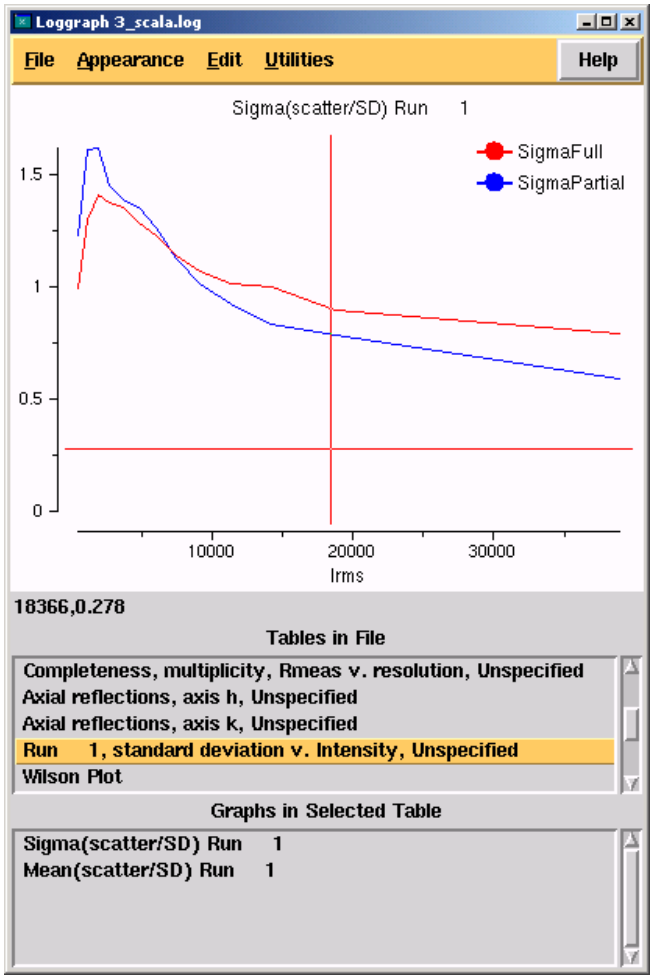
- $R_{p.i.m.} = \Sigma \sqrt{(1/n-1)} \mid I_{hl} - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$

Precision-indicating R-factor gets better with increasing multiplicity, estimates the precision of the merged <I>.

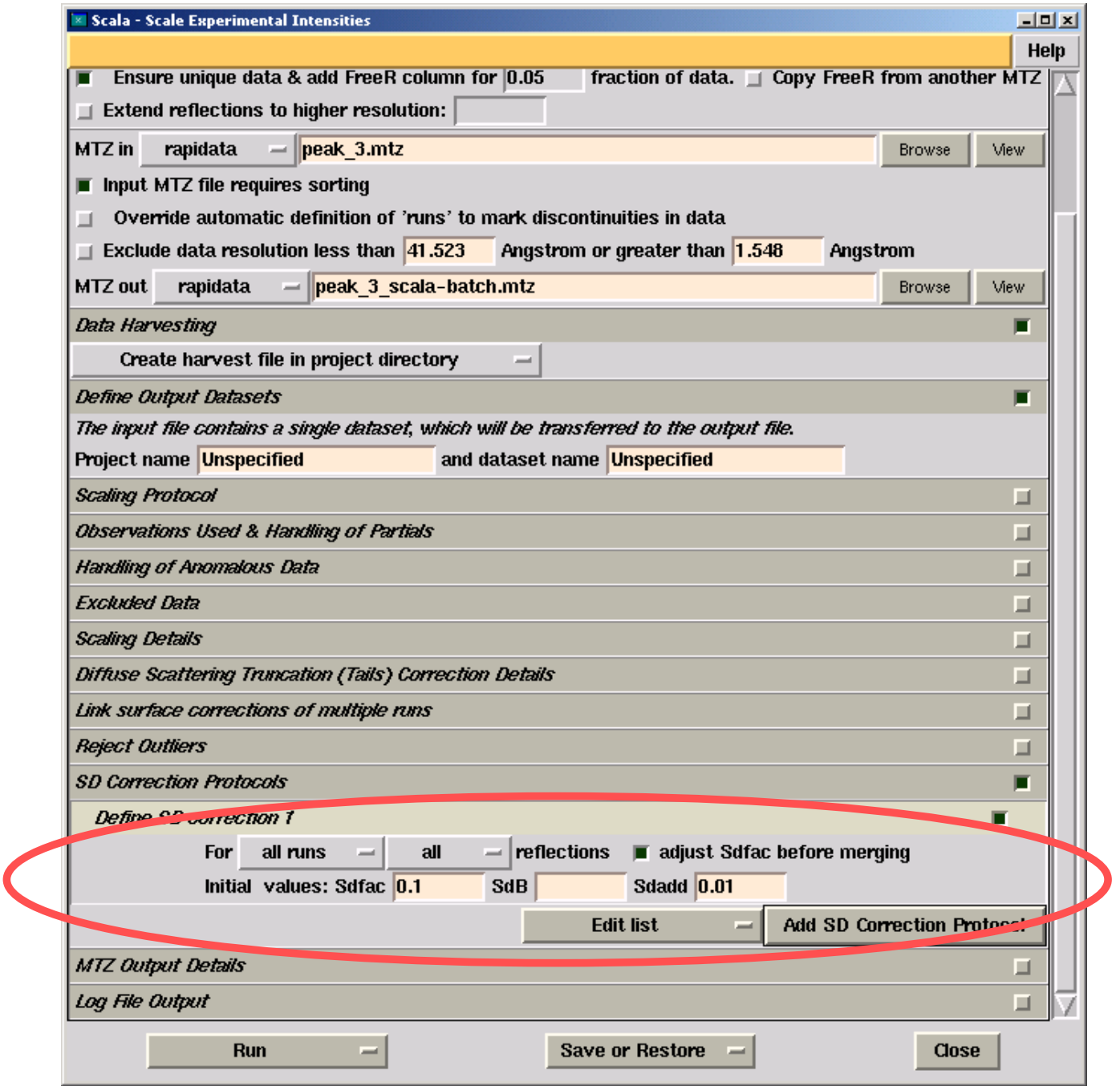
scaling - error model

- Calculate standard deviation based on sigma from profile fitting for each reflection
  - needs calibration against scatter in observations of that reflection
  - error model adjusts sigma vs scatter
- Standard deviation vs. intensity plots:
  $\delta = (I - \langle I \rangle) / \sigma$
  - should be around 1
  - similar to chi-squared test in HKL2000 or d*trek, but plotted by intensity
- If there is a significant variation from 1, one should adjust the error model:
  - SDFAC automatically modified (typically 1.2-1.3)
  - increase SDADD in steps of 0.01

  $\sigma(I) = SDFAC \times \sqrt{\{\sigma^2(I) + (SDADD \times I)^2\}}$
- Should not have too many rejections per image.
  - mild outliers mean large standard deviations, not rejection

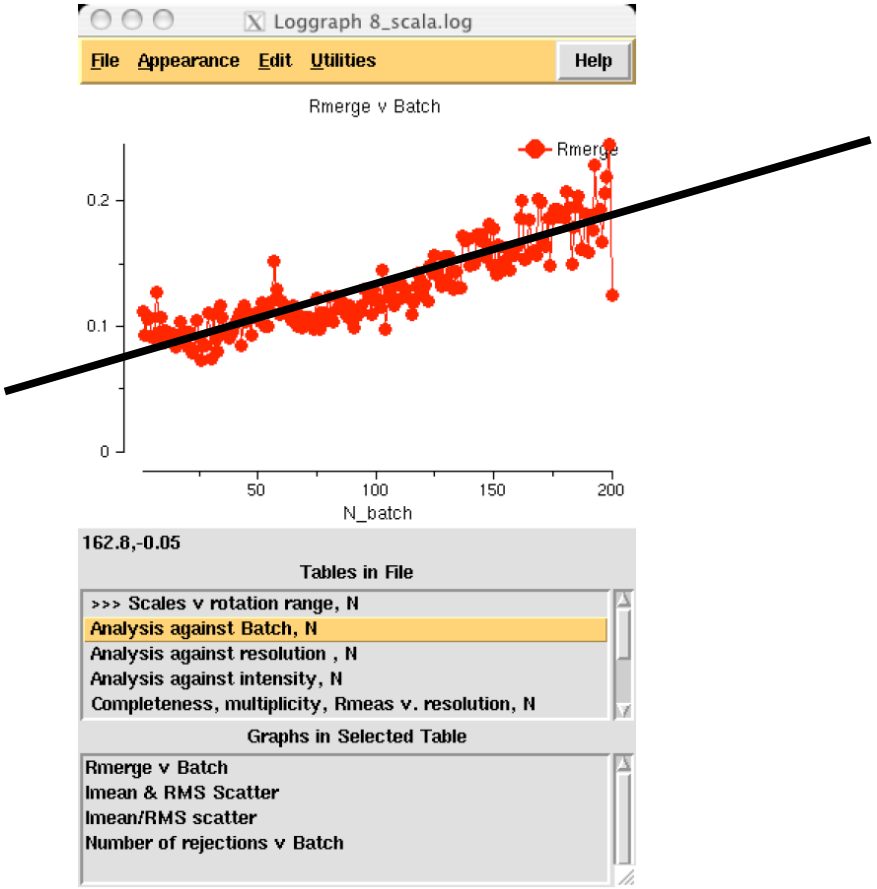


scaling - error model

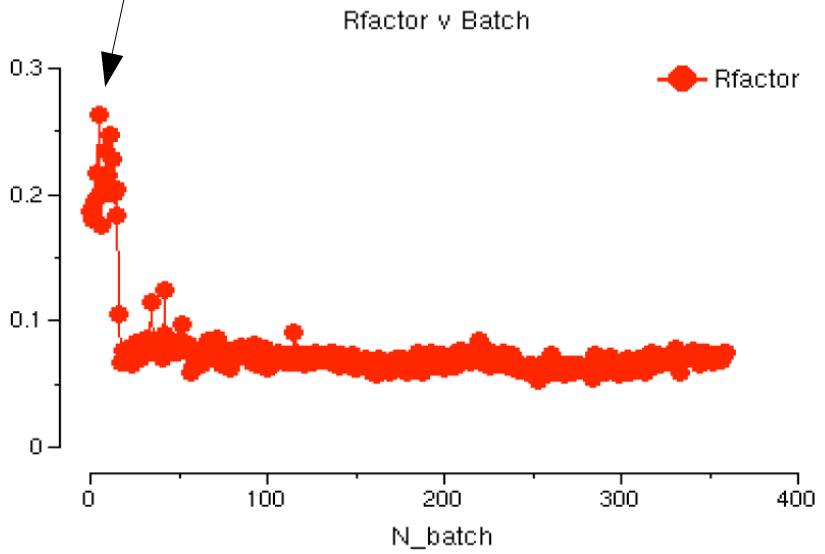


scaling - Rmerge plot

- Radiation decay: steady increase in Rmerge

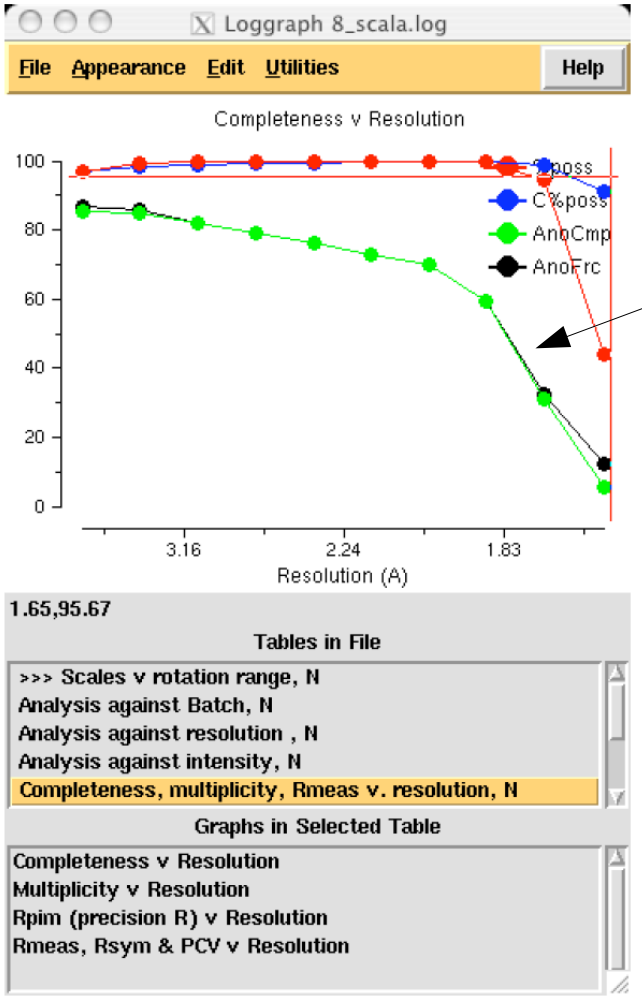


- Suboptimal initial batches: can be fixed by reindexing and reprocessing



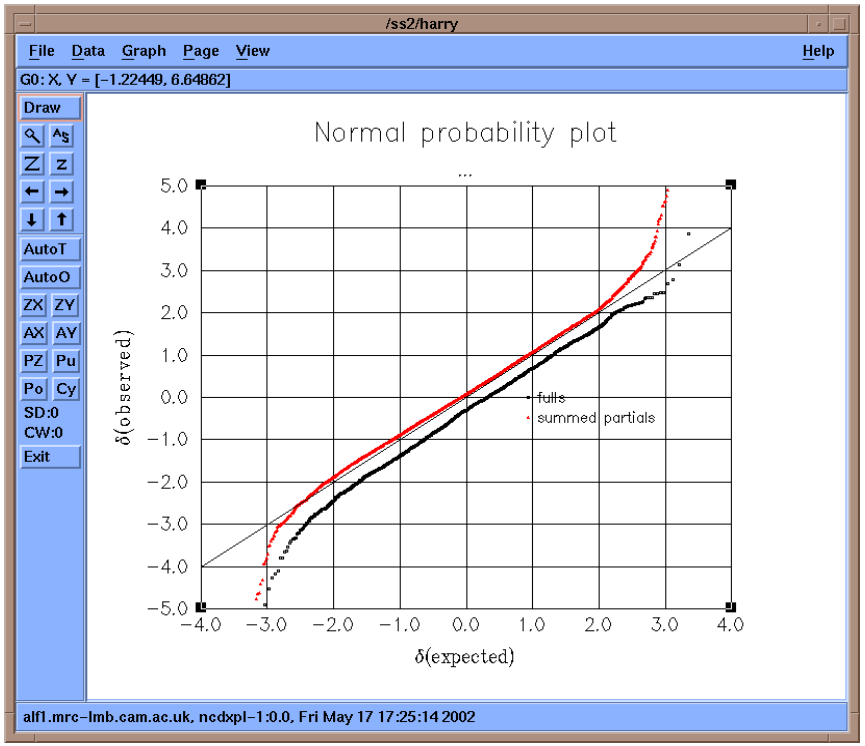
scaling - completeness

- data completeness vs resolution

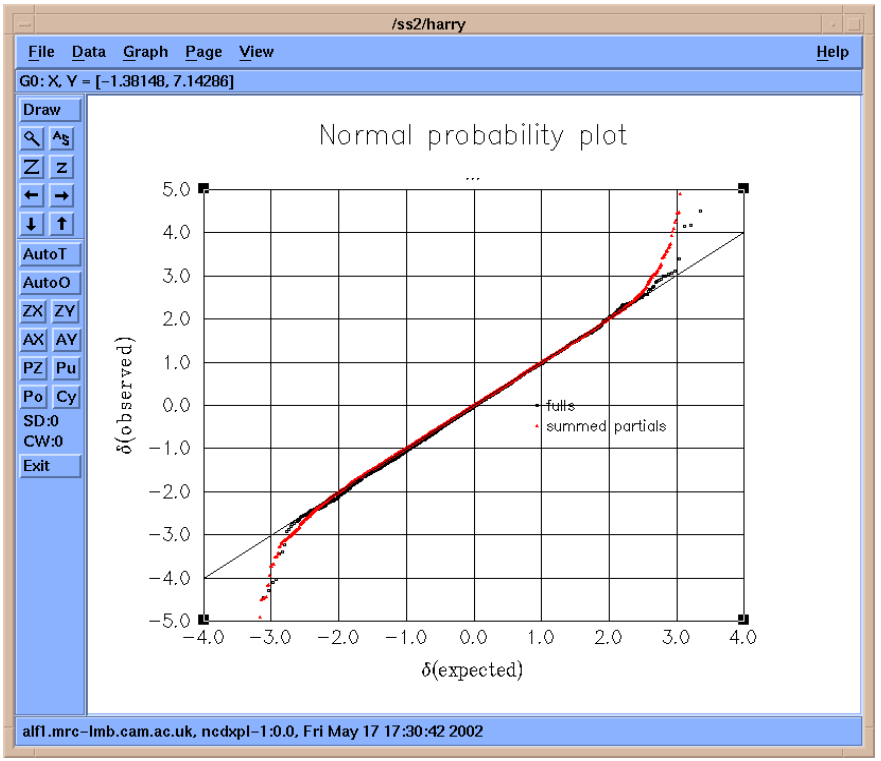


anomalous completeness!

scaling - the tails correction



- It there is significant Thermal Diffuse Scatter (TDS, spots have halos), full and summed partial reflections may not behave the same in a normal probability plot.

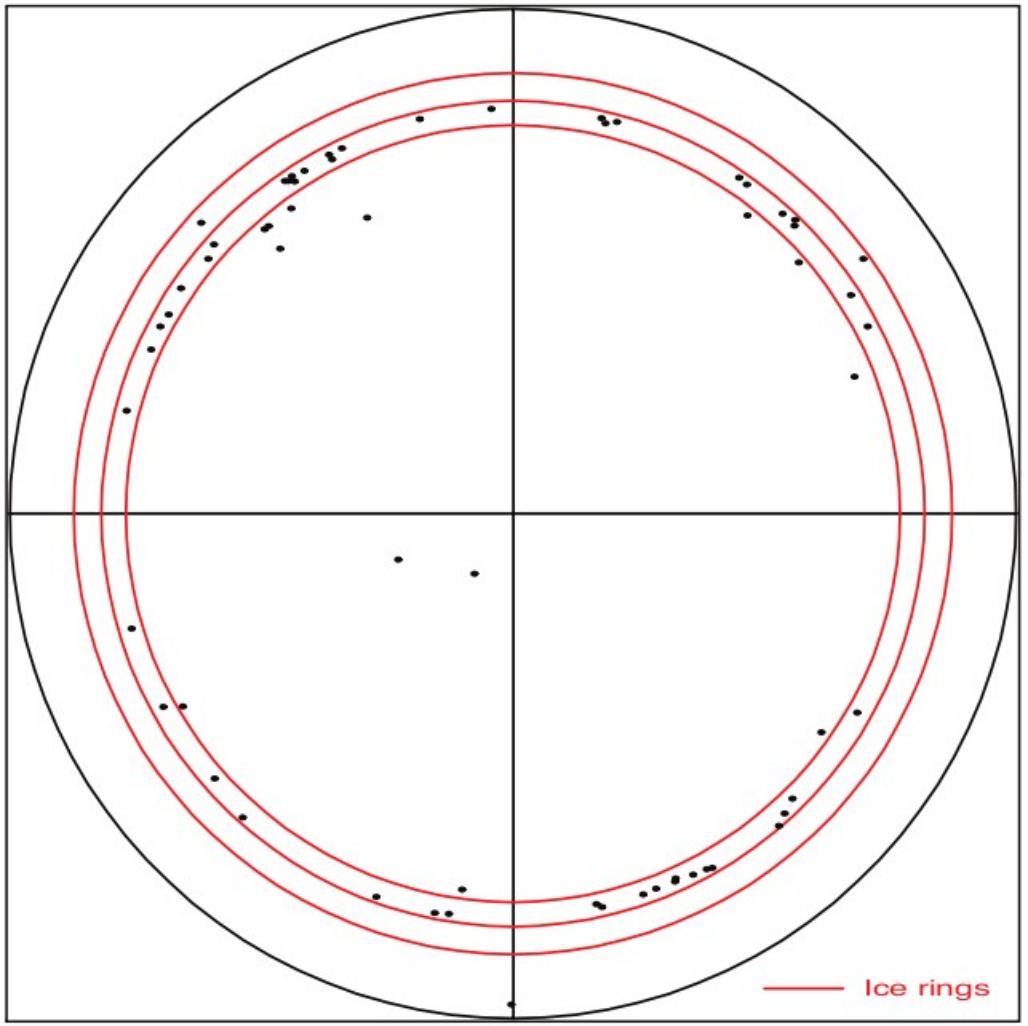


- Tails correction: correct this discrepancy.

outlier rejection

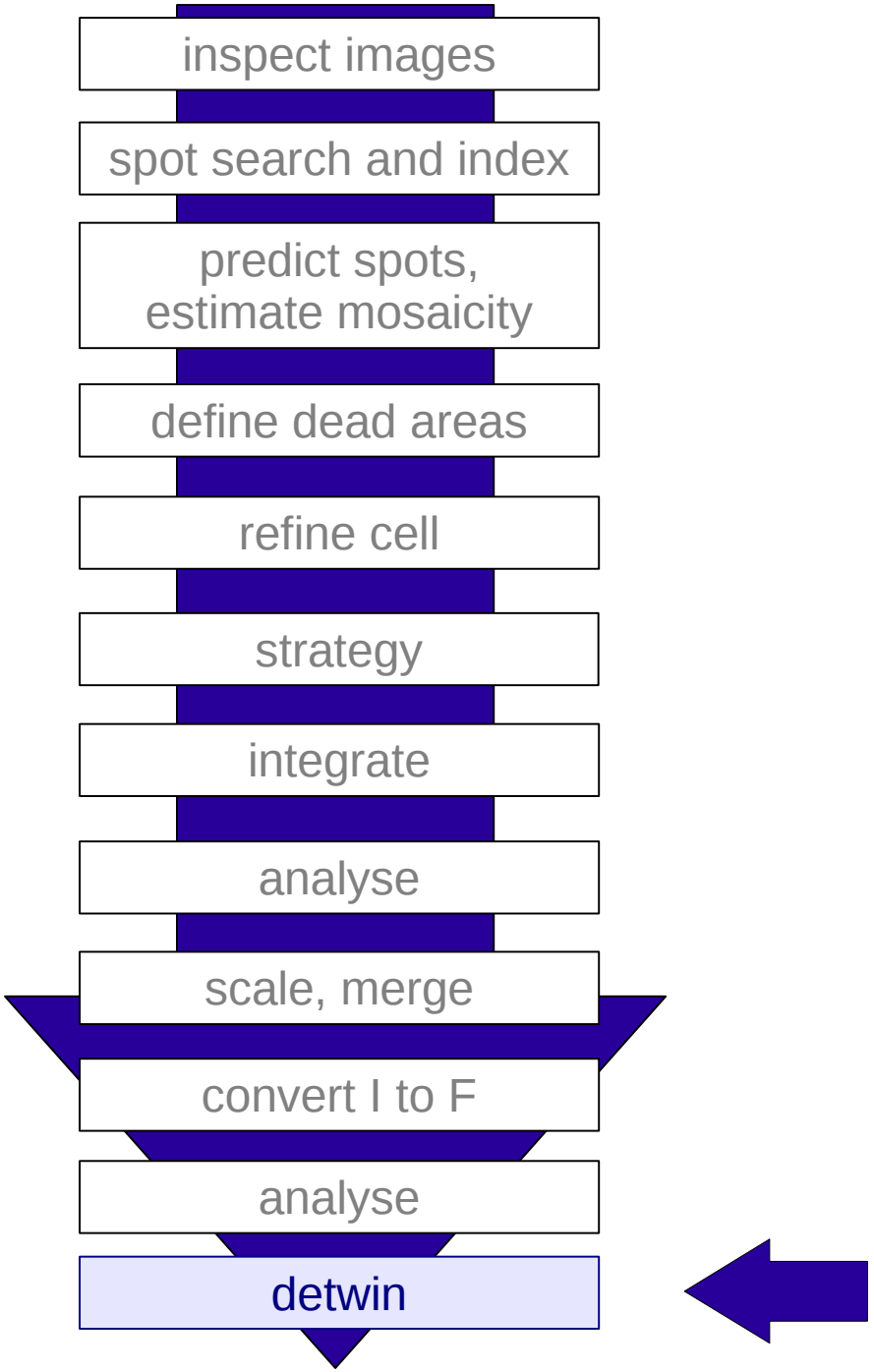
- reliable outlier rejection requires high multiplicity
- remember the backstop shadow, ideally at integration stage, but scala has mechanisms to exclude regions too.
- inspect the ROGUEPLOT

- what are outliers?
  - spots on dead areas
  - icerings
  - zingers
  - bad prediction (spot not there)
  - spot overlap
    - high mosaicity
    - multiple lattices



- position of outliers mapped on detector surface: ice rings...

data processing flowchart



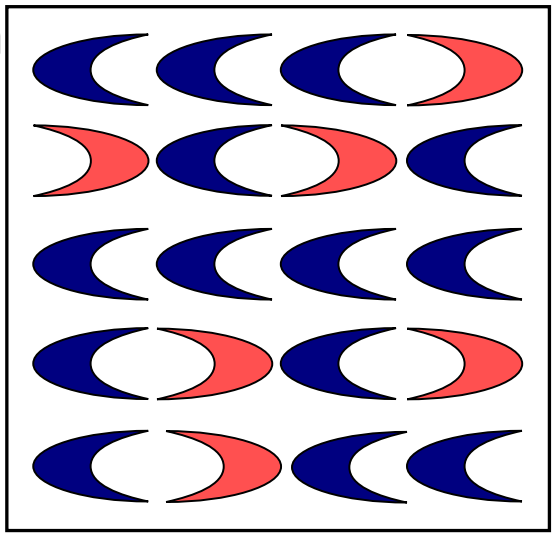
twins

1. classification of twins
2. twin detection
3. detwinning
4. structure solution and refinement (examples)

what is a twin?

-two (or more) crystals of the same kind
-joined together in a well defined orientation

2D twin crystal:



twin operator (twin law):

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

twin fraction:

α_1 = 14/20

α_2 = 6/20

perfect twin: $\alpha = 0.5$

partial twin: $\alpha \neq 0.5$

Diffration intensities: $I_{twin} = \alpha_1 I + \alpha_2 I$

Detwinning: deconvolution of twin components

classification

twin	twin operator
merohedral twin	belongs to the crystal system but not the point group
pseudo-merohedral twin	belongs to a higher crystal system
reticular merohedral twin	obverse-reverse twinning in rhombohedral
non-merohedral twin	arbitrary operator

epitaxial twinning: non-merohedrals, split crystals
lattice overlap not perfect
hemihedral twins: 2 domains are present

merohedral twins

twin operator:
belongs to the crystal system but not the Laue group

- exact overlap of lattices
- all reflections affected by twinning
- determination of the space group and Laue group difficult
- low $|E^2-1|$ value
- possible in the following crystal systems:
  - tetragonal
  - trigonal/hexagonal
  - cubic

pseudo-merohedral twins

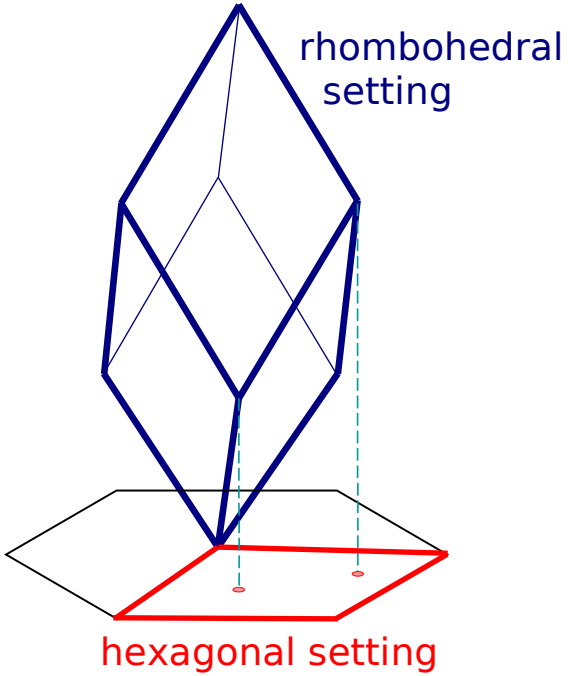
twin operator:
belongs to a higher crystal system

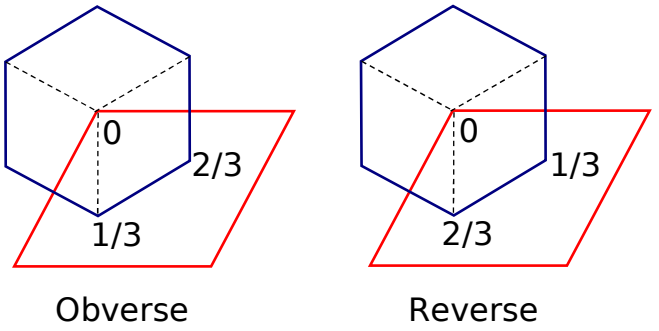
- if the metric symmetry is higher than the symmetry of the crystal
  typical examples:
  - monoclinic with β≈90º mimics orthorhombic
  - monoclinic with a≈c "
  - monoclinic with β≈120º and 3 domains mimics 6/m
- overlap of lattices not exact,
  depending on how well the higher metric symmetry is fulfilled
- all reflections affected by twinning

obverse-reverse twins

twin operator:
1. twofold axis || to the threefold
2. twofold axis || to **a-b**

- exact overlap of lattices
- lattice centering not easy to detect because
  - h+k+l=3n (obverse)
    h-k+l=3n (reverse) are overlapping
- reciprocal lattice plot:
- l=3n layers: every third reflection observed
  l≠3n layers: every third reflection absent
- high ratio of non-overlapping reflections → easy to solve/refine





non-merohedral twins

twin operator:
arbitrary operator,
e.g. a twofold rotation about a cell edge or diagonal

- overlap of lattices not exact; some spots split, some sharp
- twinning is obvious from the diffraction pattern
- indexing, cell refinement and integration problematic
  (non-protein or commercial software)
- majority of reflections not affected
- refinement problems
  (HKLF 5 format, free R-set selection - twin pairing error)

detection of twins

1. Indexing

- split reflections (pseudo- and non-merohedral)
- metric symmetry higher than Laue symmetry (pseudo-merohedral)

2. Data reduction, data statistics

- systematic absences not consistent with any space groups
- R_{int} for the higher symmetry Laue group is just a bit higher than for the lower one
- $|E^2-1|<<0.736$
- the second moment of $\langle I \rangle$ $(\langle I^2 \rangle / \langle I \rangle^2)$ for acentric reflections

(perfect twins)

twin: 1.5 no twinning 2.0

- H-test: $H=(I_{1,twin}-I_{2,twin})/(I_{1,twin}+I_{2,twin})$ (partial twins)

$\alpha = \frac{1}{2}(1-2\langle H \rangle)$

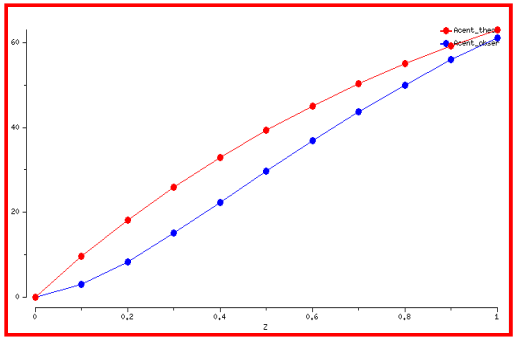
- intensity distribution tests affected by anisotropy and pseudo-symmetry!
- estimation of α with ± 5% error

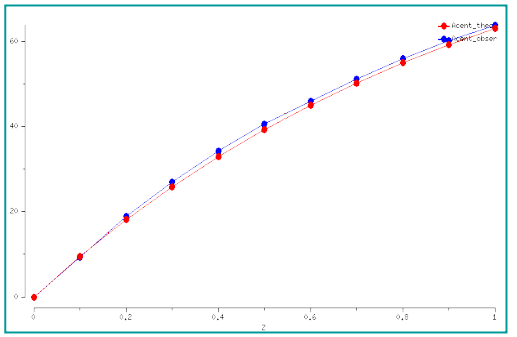
detection of twins

- L-test: (not affected by anisotropy): $L=(I_{1,twin}-I_{2,twin})/(I_{1,twin}+I_{2,twin})$
  (on reflection close in reciprocal space)

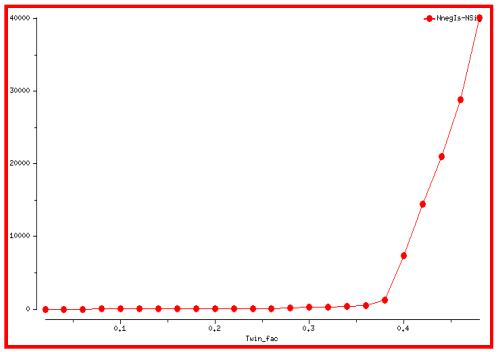
	<|L|>	<L^2>
no twinning, acentric	1/2	1/3
no twinning, centric	$2/\pi$	1/2
perfect twin, acentric	3/8	1/5

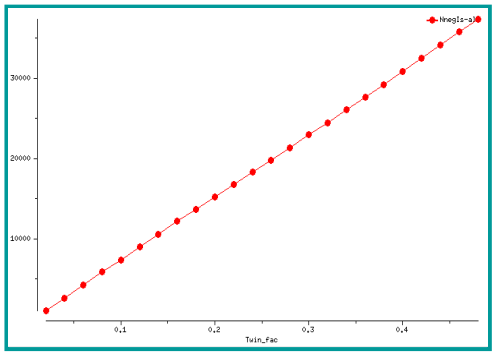
- cumulative intensity plot





- Britton plot (# of **–** intensities *vs* **α**)





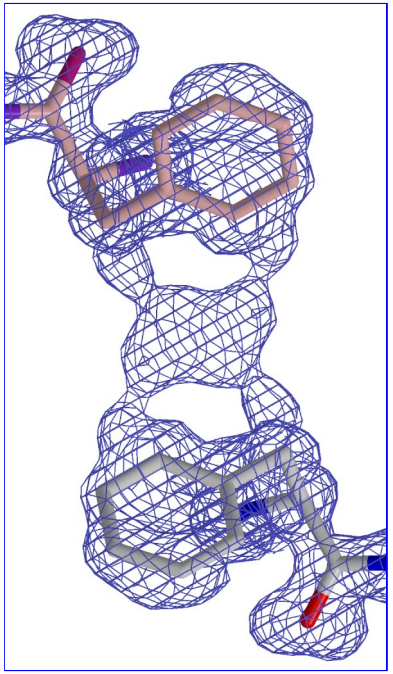
detection of twins

3. Structure solution

- Patterson map physically impossible (for heavy atoms)
- molecular replacement problematic even if the search model highly homologous
- twin ghosts, messy maps (unexplained blobs in density)
- disorders that cannot be explained or modeled reasonably
- high R-factors

detection of twins

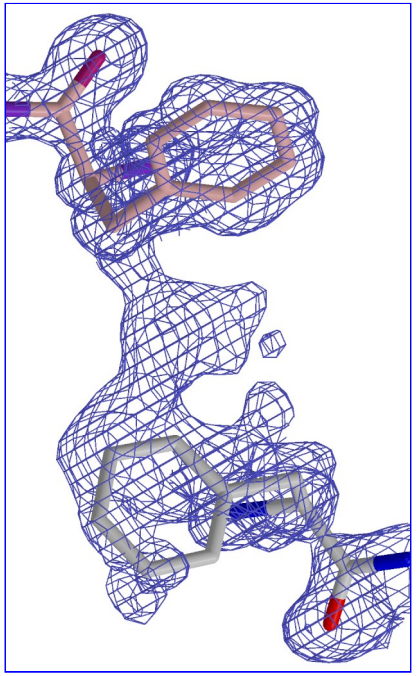
alternate conformations *vs* different species



$P6_522$

1 mol / ASU

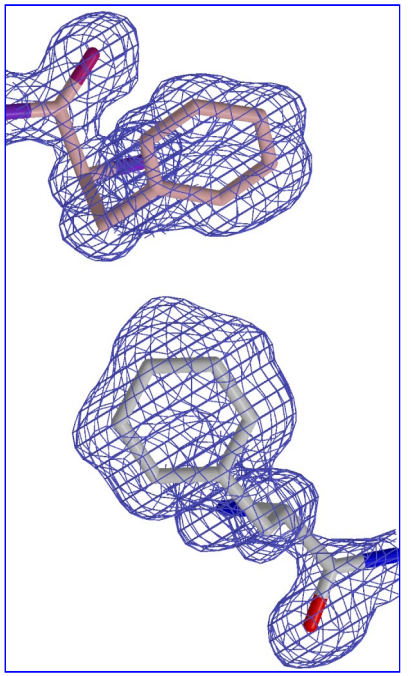
alternate conformations of symmetry equivalents clashing



$P6_5$

2 mol / ASU

after 1 cycle of twin refinement



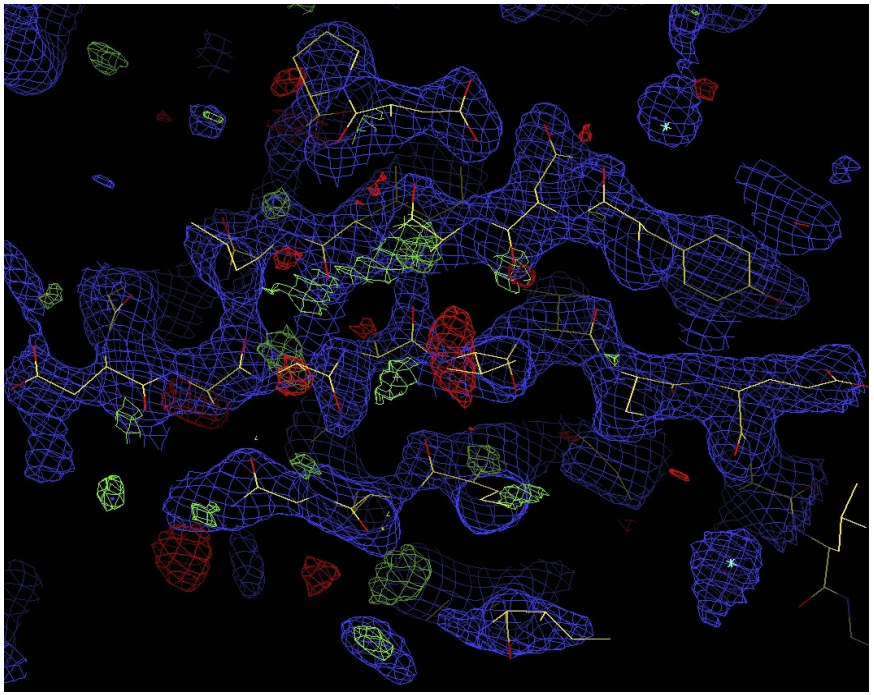
$P6_5$

2 mol / ASU

final refined structure

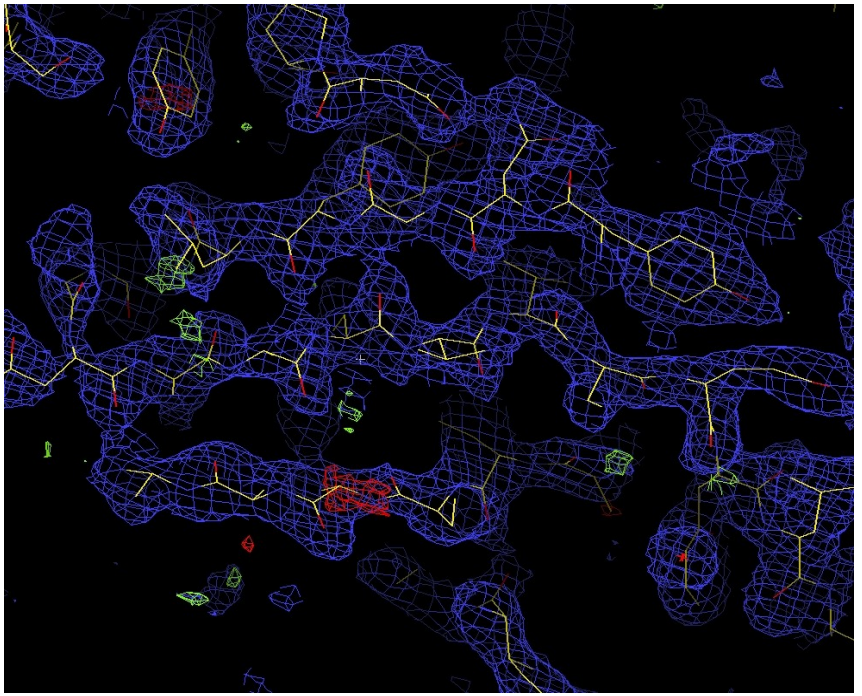
detection of twins

map quality



$P3_221$

unexplained blobs,
R>30

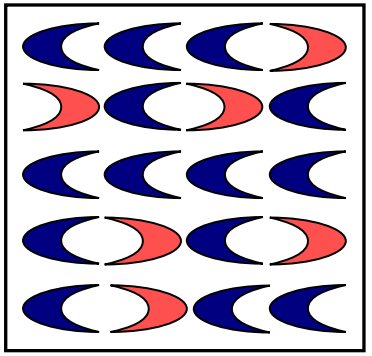


$P3_2$

`TWIN 0 1 0 1 0 0 0 0 -1`

better maps,
R~20

detwinning



$\alpha_1 = 14/20$

$\alpha_2 = 6/20$

$\alpha_1 + \alpha_2 = 1$

$I_{twin} = \alpha I + (1-\alpha) I$

detwinning: deconvolution of intensities of overlapping reflections

Diffraction intensities of two overlapping reflections related by the twin operation but not by cystallographic symmetry:

$$I_{1,twin} = (1-\alpha) I_1 + \alpha I_2$$
$$I_{2,twin} = \alpha I_1 + (1-\alpha) I_2$$

} measured intensities

$$I_1 = \left((1-\alpha) I_{1,twin} - \alpha I_{2,twin}\right)/(1-2\alpha)$$
$$I_2 = \left((1-\alpha) I_{2,twin} - \alpha I_{1,twin}\right)/(1-2\alpha)$$

} intensities after deconvolution

$$v(I_1) = \left((1-\alpha)/(1-2\alpha)\right)^2 (\sigma I_{1,twin})^2 + \left(\alpha/(1-2\alpha)\right)^2 (\sigma I_{2,twin})^2$$
$$v(I_2) = \left((1-\alpha)/(1-2\alpha)\right)^2 (\sigma I_{2,twin})^2 + \left(\alpha/(1-2\alpha)\right)^2 (\sigma I_{1,twin})^2$$

} variance of I_1 and I_2

unreliable results if $\alpha \approx 0.5$

example: merohedral

Human Hydroxyacid Oxidase 1

- cell:

```
97.300 97.300 80.540 90 90 90
```

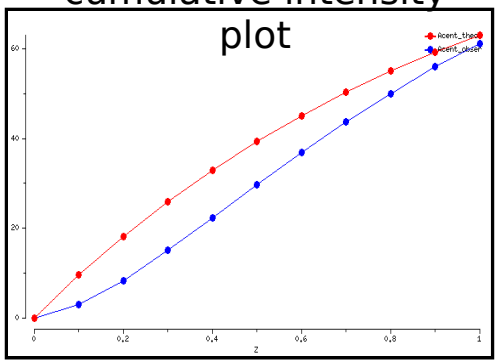
I4 or **I422**

Laue group	4/mmm	4/m
R_{int}	10.1%	7.5%
Matthews	1.36	2.72
Solvent c.	9.7%	54.8%

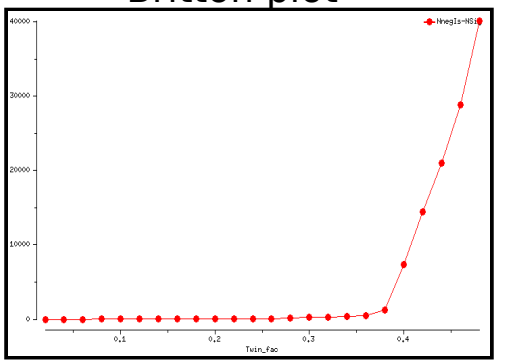
- xprep:

```
Mean |E*E-1| = 0.586
TWIN  0 1 0  1 0 0  0 0 -1
BASF  0.363 [NC]
```

cumulative intensity plot

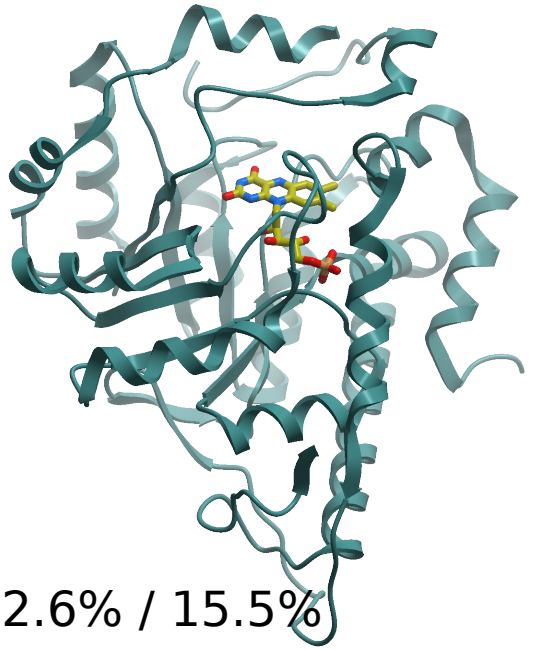


Britton plot



SHELXL in I4

twin law: 0
1 0
1 0 0
0 0 -1



R/R_{free} = 12.6% / 15.5%

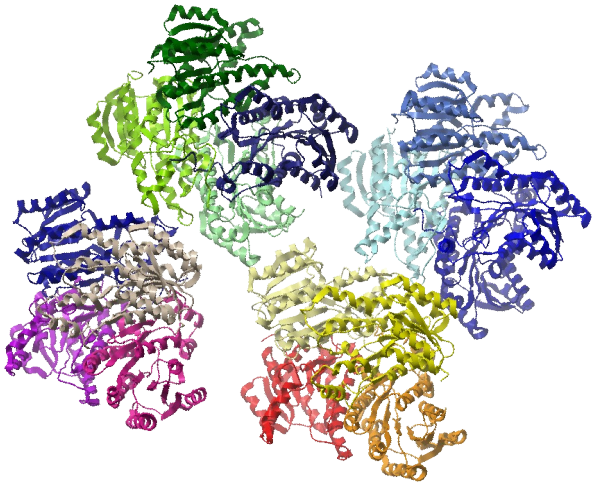
example: pseudo-merohedral

Human Retinal Short-Chain Dehydrogenase/Reductase 3

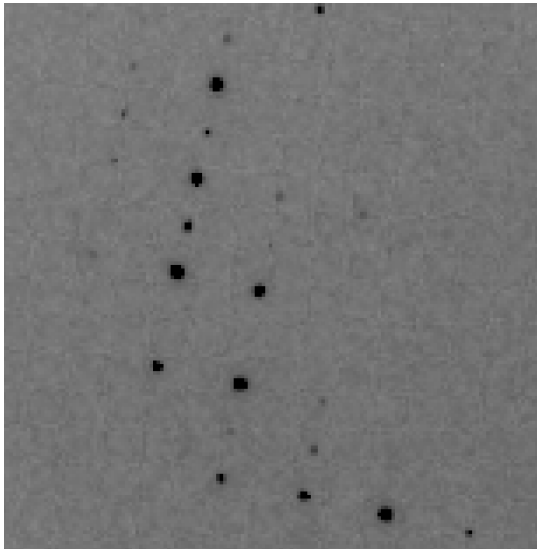
- cell: **167.115** 98.823 **167.463**
  90.00 115.87 90.00
  $P2_1$ (Rmerge=0.08)

  or: 178.666 284.619 98.990
  90.00 90.00 90.90
  $C222_1$ (Rmerge=0.17)

$P2_1$
16 mol / ASU
R/R_{free} = 18.2% / 22.9%



- overlapping reflections



$P2_1$
TWIN 0 0 -1 0 -1 0 -1 0 0
twin refinement
not pursued as
no significant improvement

example: obverse-reverse

Human Transplantation Specific Transplantation Antigen 3

- cell: 91.640 91.630 429.980 90.00 90.00 120.00

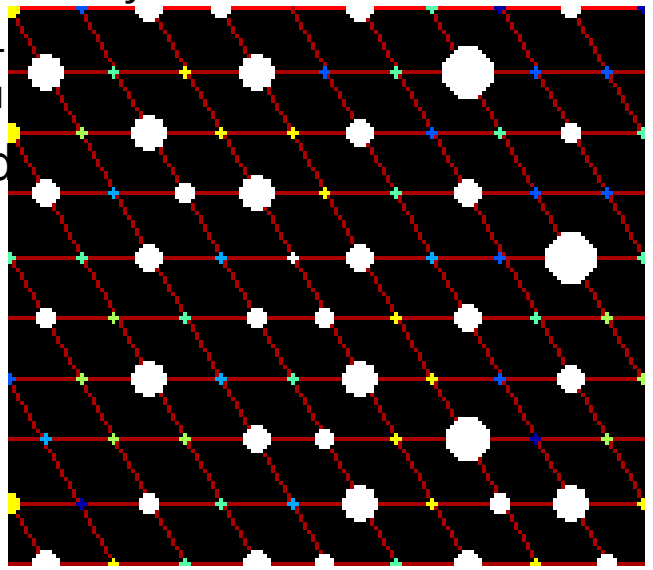
P6 or R3

- obverse/reverse test: Mean I for
  - obv only 81.4
  - rev only 78.0
  - nei

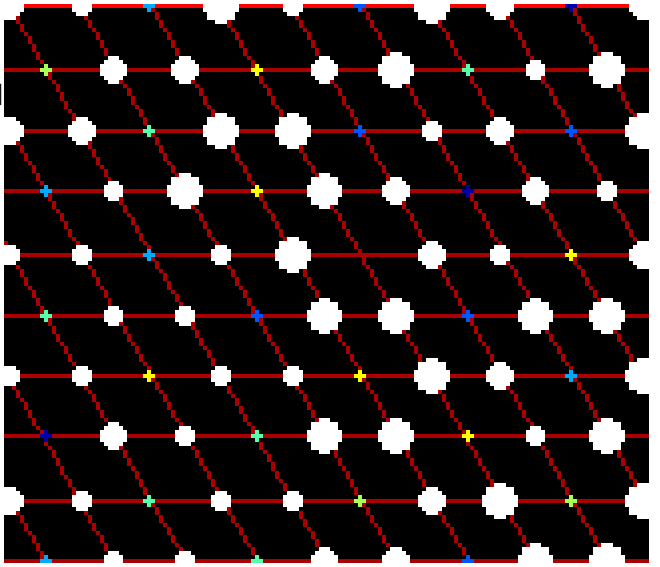
TWIN -1 0 0 0 -1 0 0 0 1

BASF 0.489

I=3 every 3rd observed

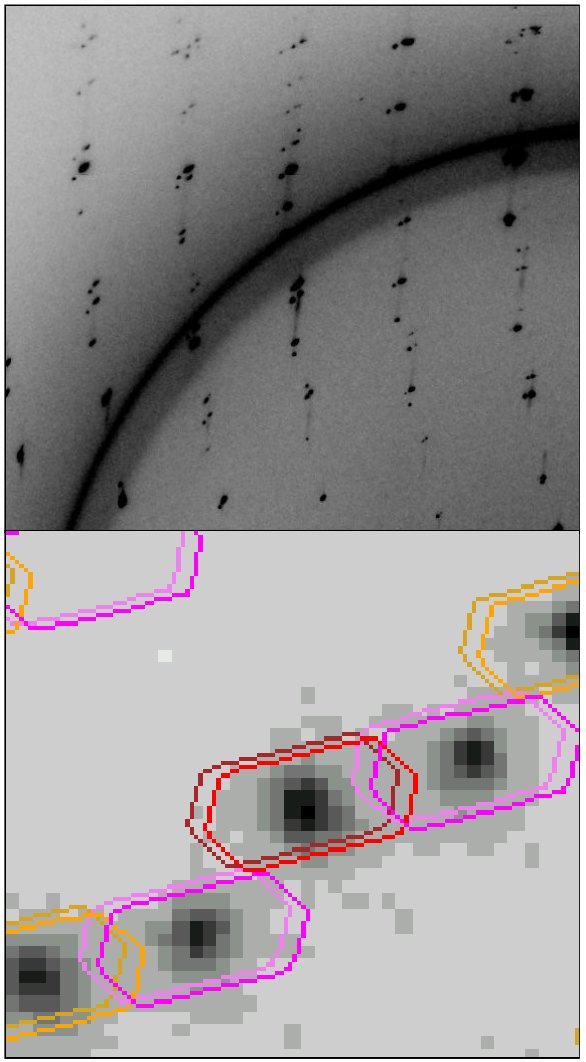


I=2 every 3rd missing



example: non-merohedral

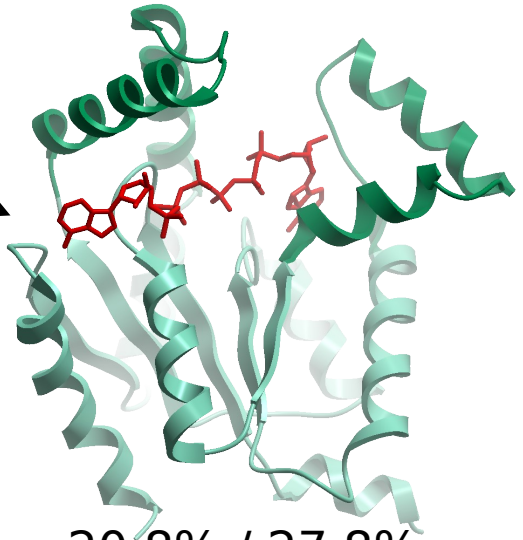
Human Adenylate Kinase 1



- cell: 35.184 43.919 64.742
  90.80 95.61 106.26
  P1
- twin operation:
  twofold (180°) rotation around cell diagonal
- partially overlapping reflections

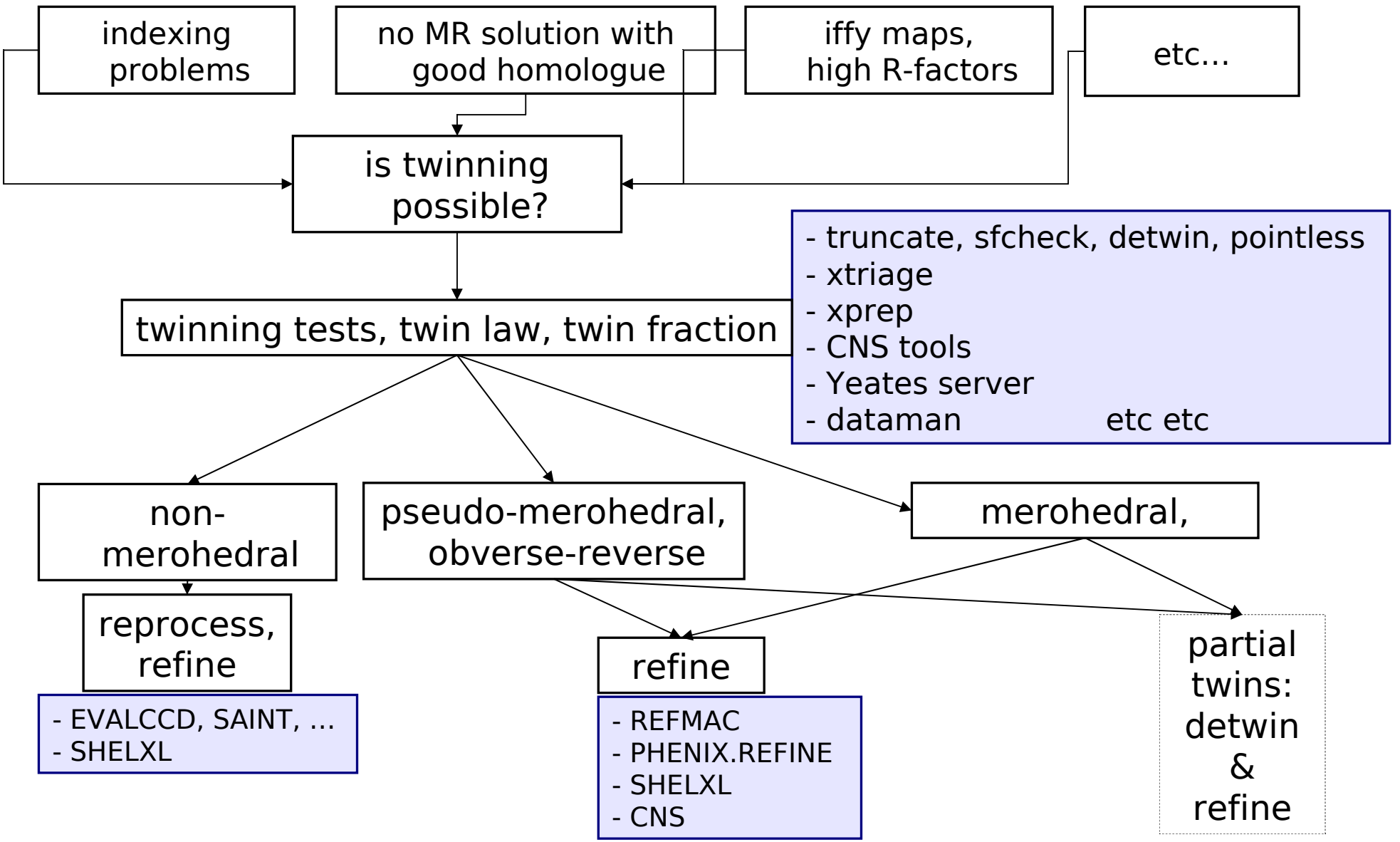
Data reduction:
EVALCCD
Refinement:
SHELXL, HKLF 5

free R set!



R/R_{free} = 20.8% / 27.8%

structure solution



reading

- Rees, C.D. (1980), The influence of twinning by merohedry on intensity statistics. *Acta Cryst* **A36**, 578-581
- Redinbo, M.R. & Yeates, T.O. (1993) Structure determination of plastocyanin with a hemihedral twinning fraction of one-half. *Acta Cryst* **D49**, 375-380
- Gomis-Rüth et al. (1995), Determination of Hemihedral Twinning and Initial Structural Analysis of Crystals of the Procarboxypeptidase. A Ternary Complex. *Acta Cryst* **D51**, 819-823
- Yeates, T.O. (1997), Detecting and overcoming crystal twinning. *Methods in Enzymology* **276**, 344-358
- Herbst-Irmer, R. & Sheldrick, G.M. (1998), Refinement of twinned structures with SHELXL97. *Acta Cryst* **B54**, 443-449
- Herbst-Irmer, R. & Sheldrick, G.M. (2002), Refinement of obverse/reverse twins. *Acta Cryst* **B58**, 477-481
- Padilla, J.E. & Yeates, T.O. (2003) A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Cryst.* **D59**, 1124-1130
- Lebedev, A.A., Vagin, A.A. & Murshudov, G.N. (2006) Intensity statistics in twinned crystals with examples from the PDB. *Acta Cryst.* **D62**, 83-95.
- Yeates twin server and tutorial: www.doe-mbi.ucla.edu/Services/Twinning
- CCP4 about twinning: www.ccp4.ac.uk/dist/html/twinning.html
- Regine's twin site: http://shelx.uni-ac.gwdg.de/~rherbst/twin.html