Data Processing - opening the Black Box

Diamond-CCP4 Data Collection & Structure Solution Workshop 2025

Ana Gonzalez (MAX IV)

About this lecture

Data processing is done automatically after data collection, why is it useful to learn about it?:

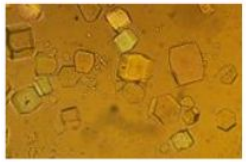
- Automation is not perfect. Sometimes it is possible to get better results by processing manually.
- Data processing provides feedback about the experiment.

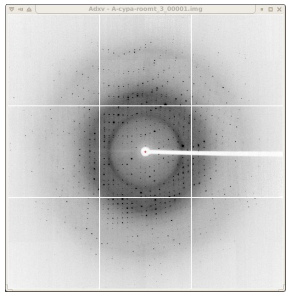
Opening the black box:

- Steps in data processing (general description, no details about implementation in different programs).
- Quality indicators. What can go wrong, when it is possible to do anything about it

Steps in structure determination

Data collection: From crystals to diffraction images

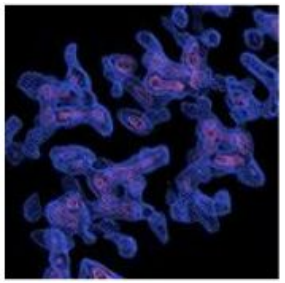




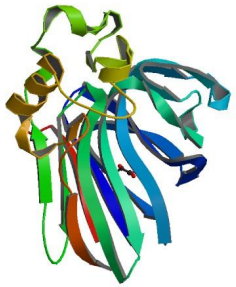
Data processing: From diffraction to structure factor intensities/amplitudes

F(hkl) or **I(hkl)** and uncertainties

Phasing: Structure factors/Electron density map



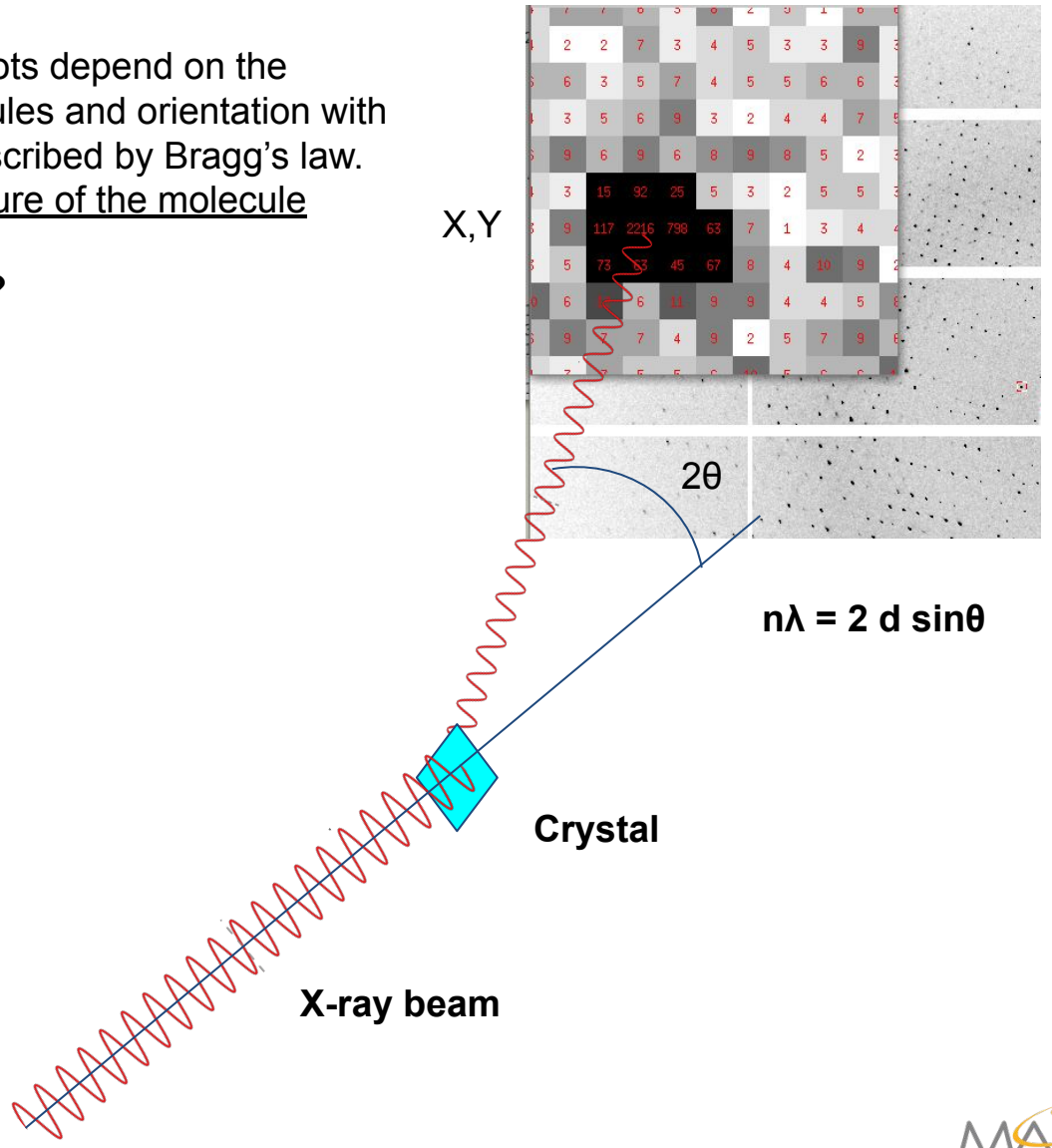
Model building and refinement



What can we get from the diffraction pattern?

The positions of the diffraction spots depend on the crystal arrangement of the molecules and orientation with respect to the X-ray beam, as described by Bragg's law.
<u>They do not depend on the structure of the molecule</u>

How about the spot intensities?



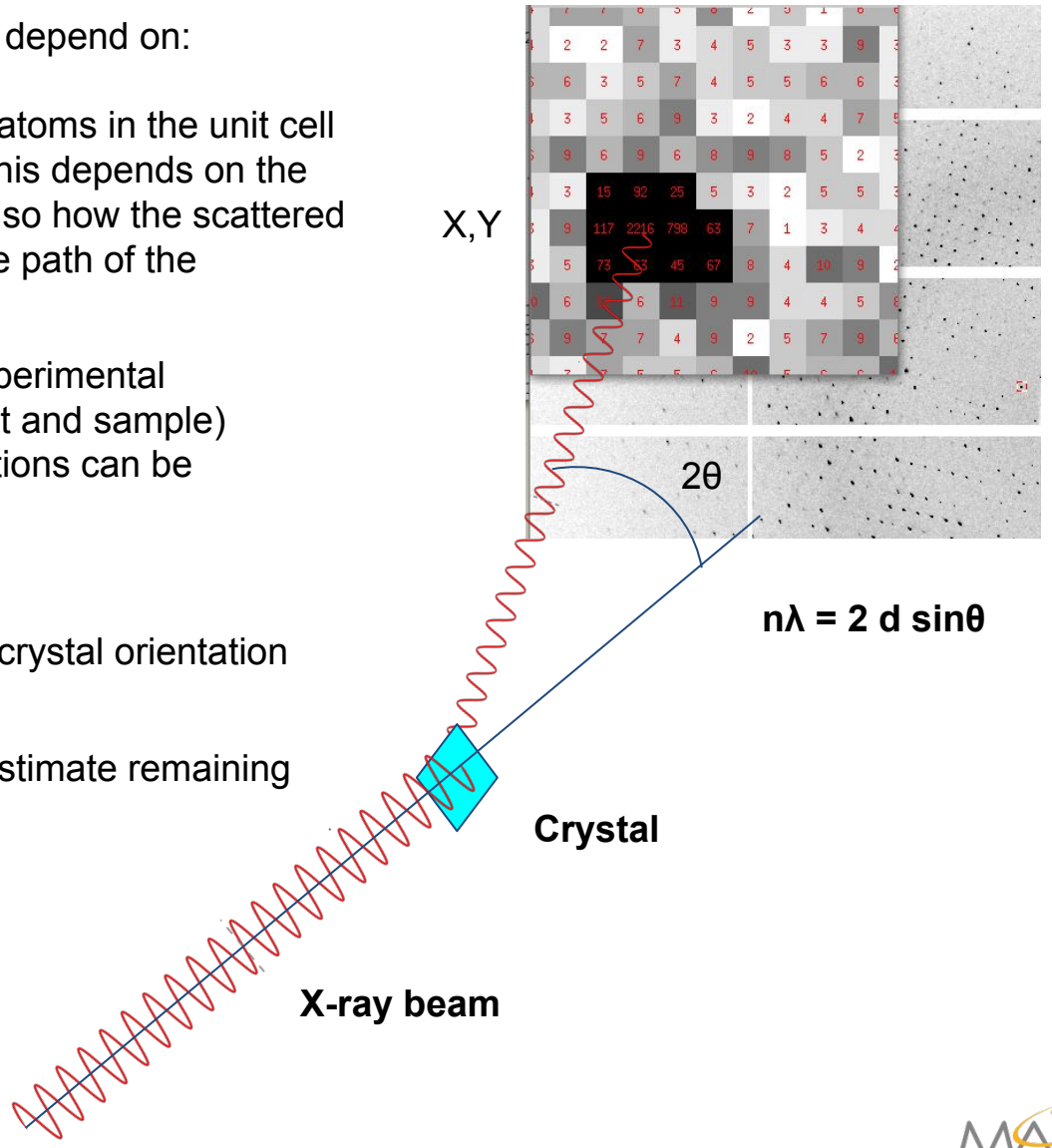
Data processing overview

The intensities of the diffraction spots depend on:

- The combined scatter of all the atoms in the unit cell (the **structure factor F(hkl)**). This depends on the atomic scattering factors, and also how the scattered X-rays phase changes along the path of the reflection.
- Non-crystal scattering, other experimental (physical/geometrical/instrument and sample) factors. Some of these contributions can be calculated or modelled.

During data processing:

- We sort the diffraction spots by crystal orientation and measure the intensities
- We correct the intensities and estimate remaining errors.



Main steps in data processing

1. **Indexing**: Determine the crystal lattice and crystal orientation from the observed diffraction spots.
2. **Integration**: Calculate the intensities I(hkl) of the diffraction spots.
3. **Scaling**: Place all intensities on the same scale.

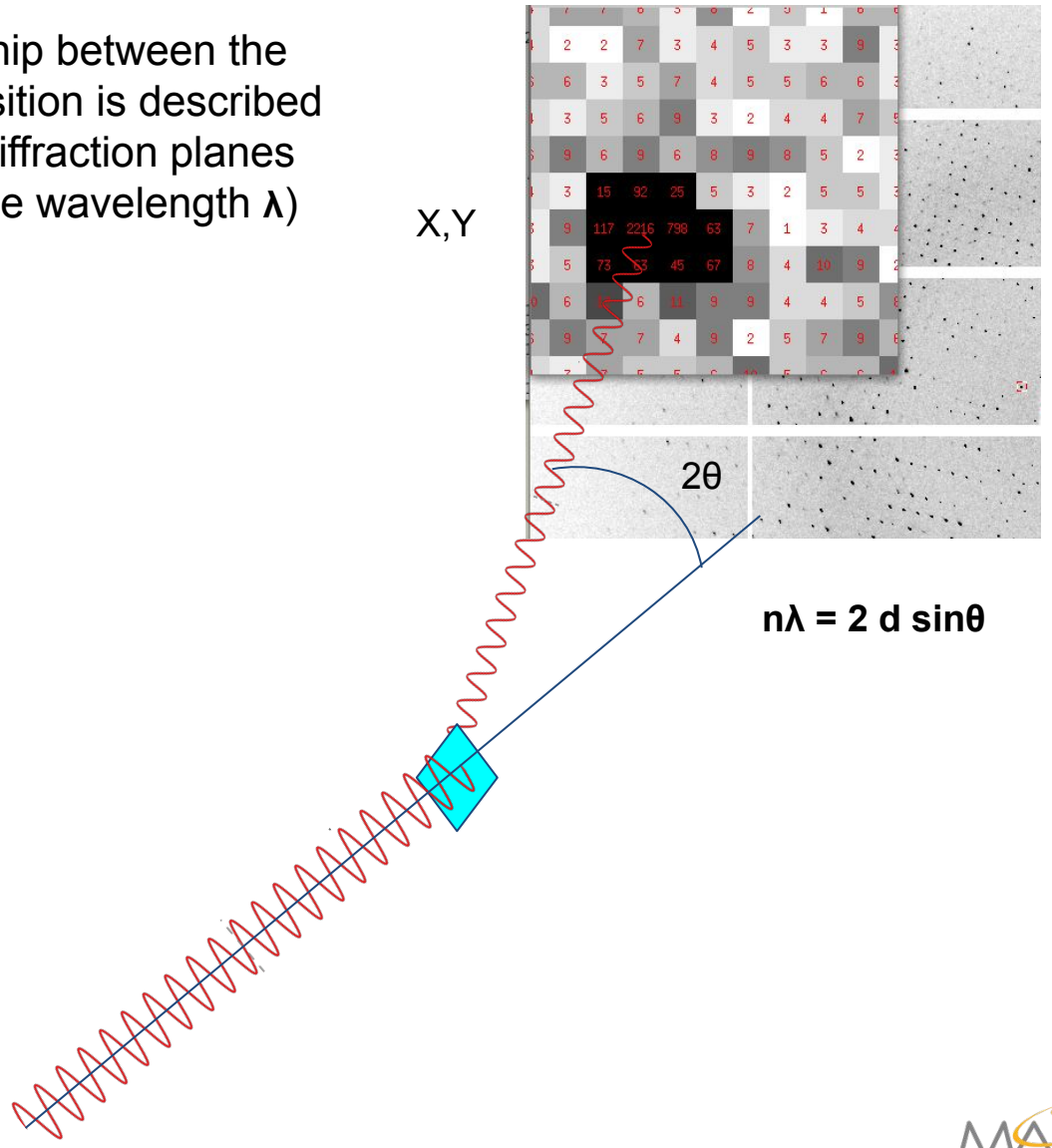
Other steps that are often done as part of processing:

4. **Merging**: Average all the measurements of the same reflection.
5. **Calculate structure factor amplitudes** |F(hkl)| from the intensities.

Indexing

Indexing overview

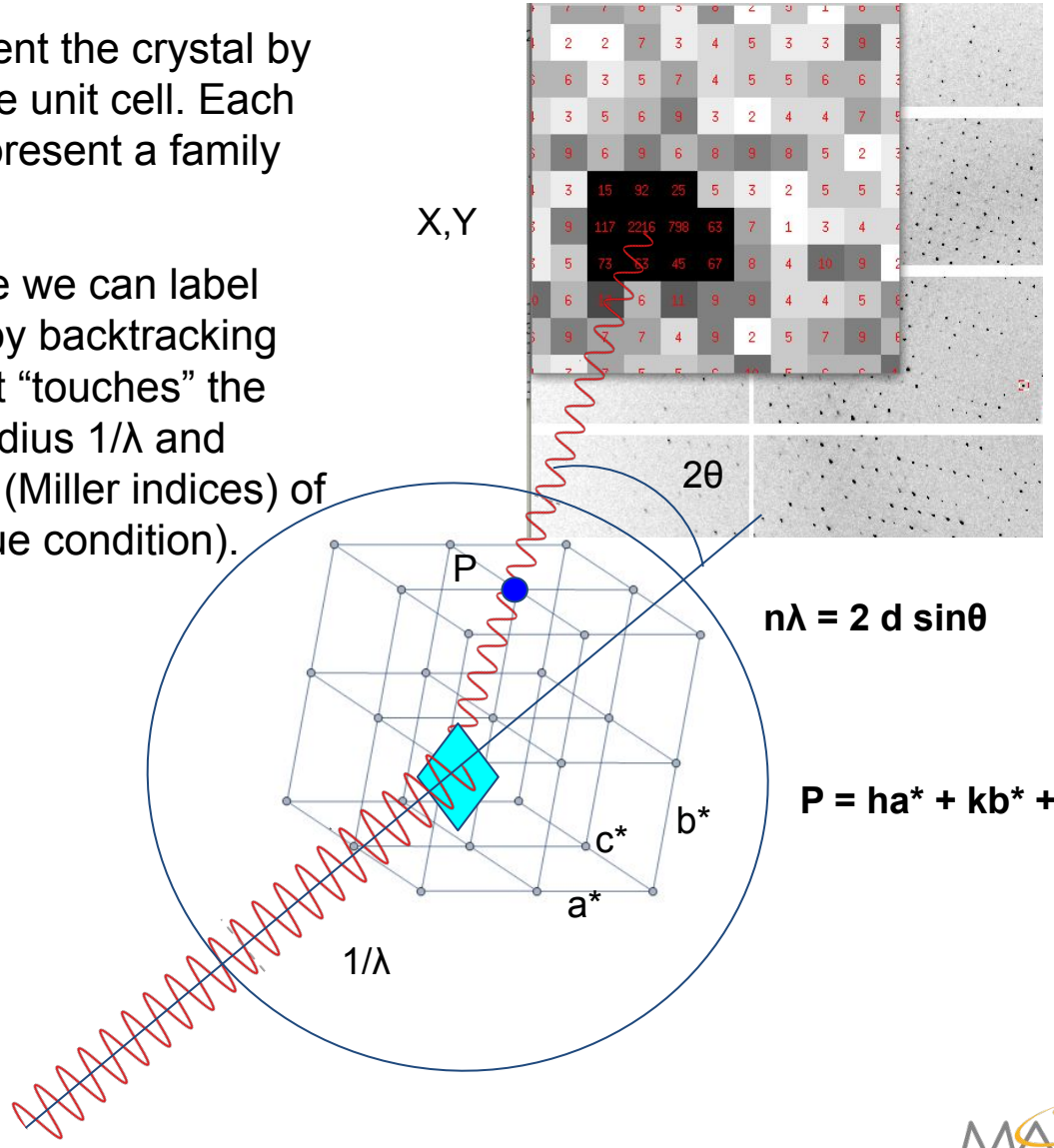
In the real space, the relationship between the crystal orientation and spot position is described by Bragg's law (d says which diffraction planes are reflecting at a given **θ** for the wavelength **λ**)



Indexing overview

It is more convenient to represent the crystal by a reciprocal lattice based on the unit cell. Each reciprocal lattice point h,k,l represent a family of parallel Bragg planes.

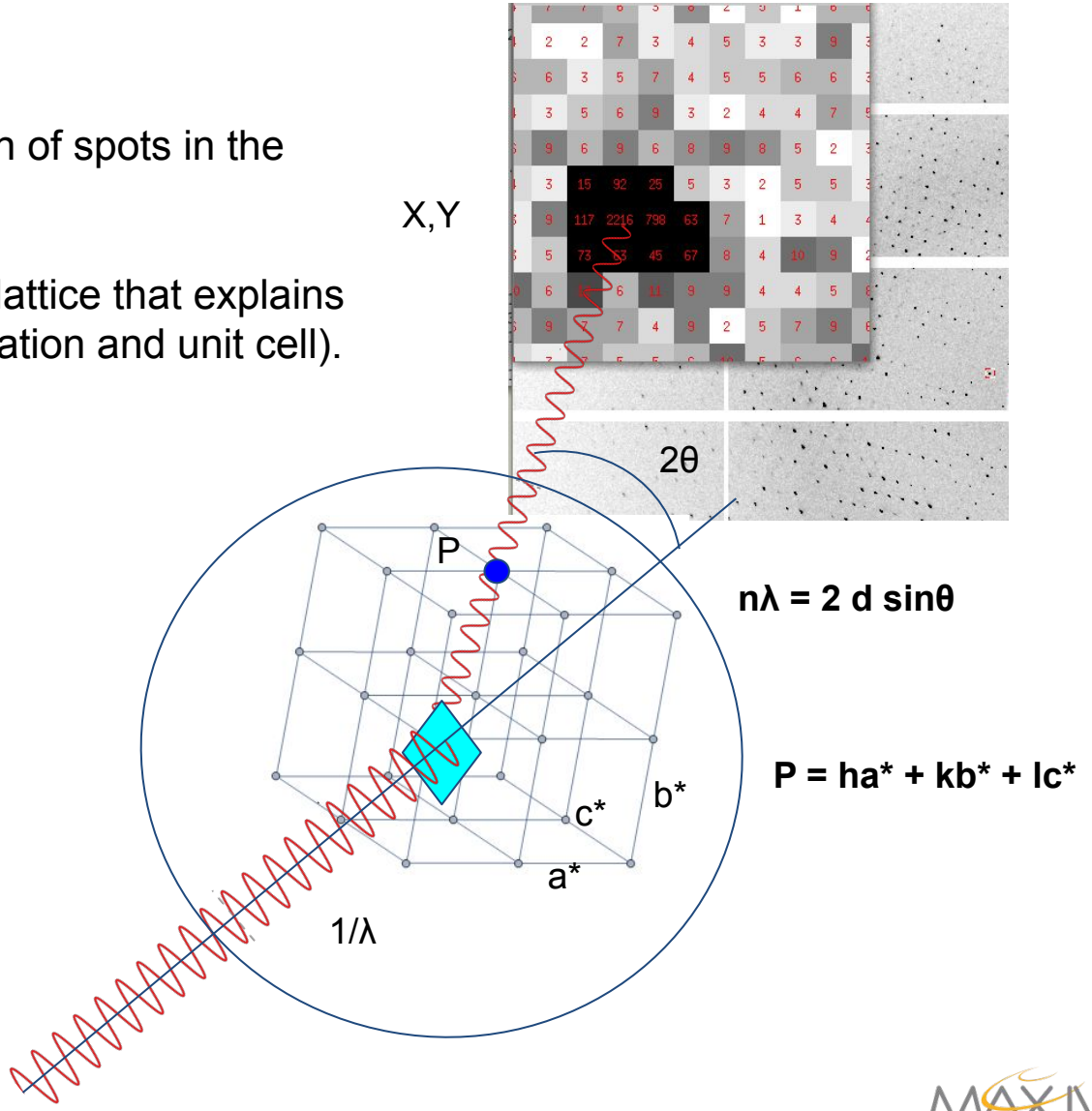
If we know the reciprocal lattice we can label (“index”) each diffraction spot by backtracking along the reflected X-ray until it “touches” the the surface Ewald sphere of radius 1/λ and assigning the h k l coordinates (Miller indices) of the point P at that location (Laue condition).



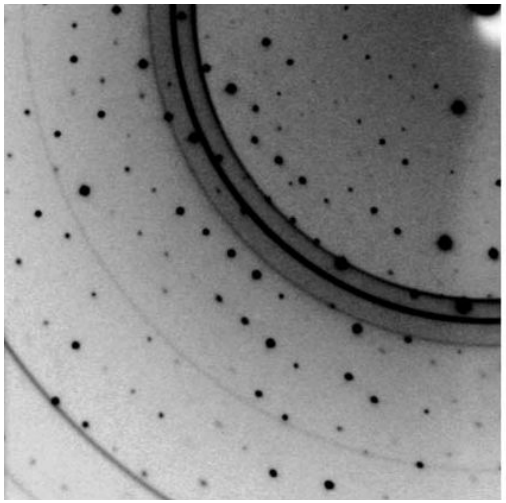
Indexing overview

Steps:

- Determine the X,Y position of spots in the image (spot finding).
- Figure out the reciprocal lattice that explains the spots positions (orientation and unit cell).



Spot finding



Can you find diffraction spots in this image?

Things that you (and the software) are looking for:

- Similarly shaped and sized spots lying at evenly spaced intervals.
- Spots clearly distinguishable from the background.

Only a fraction of the spots from one, several or all images are needed.

Mask shadows, bad detector areas, powder scatter rings.

Indexing

General procedure:

Project the detector spot coordinates onto the Ewald sphere to obtain the Reciprocal Lattice point P and the distance from the origin OP = $1/d_{hkl}$. The 1/d values cluster at even intervals along the principal axes of the crystal.

Search for a triclinic unit cell that describes the diffraction pattern and index it. This is done in different ways in different programs. See *X-ray data processing* by Harry Powell https://doi.org/10.1042/BSR20170227 .

Transform the initial cell into a cell in one of the characteristic lattices. Typically the highest symmetry lattice into which the initial cell fits relatively undistorted is chosen as the most probable. **This choice needs to be revised after the intensities of the reflections are known.**

Output from XDS, 44 possible solutions are listed (characteristic lattice, Bravais lattice, fit, unit cell)

```
 *   44     aP    0.0   35.5    54.0    57.1   90.0 103.5   90.1
 *   31     aP    0.5   35.5    54.0    57.1   90.0  76.5   89.9
 *   33     mP    1.6   35.5    54.0    57.1   90.0 103.5   90.1
     37     mC   64.0  111.5    35.5    54.0   90.1  90.1   85.4
     28     mC   64.4   35.5   111.5    54.0   90.1  90.1   85.4
     36     oC   64.5   35.5   111.5    54.0   89.9  90.1   94.6
     34     mP  140.7   35.5    54.0    57.1   90.0 103.5   90.1
     35     mP  188.3   54.0    35.5    57.1  103.5  90.0   90.1
```

etc.

Indexing problems?

Symptoms: The predicted pattern does not match the observed one or the scaling statistics are very poor

Is the diffraction good enough? **Look for another crystal.**

Is spot finding locating good spots? Are there many weak reflections being missed? Is there more than one lattice? **Play with spot determination parameters**.

Do you know the following:

- the wavelength
- the direct beam coordinates on the detector
- the crystal to detector distance

Use beamline provided input files

Do you have the correct geometry description? **Use beamline provided input files**

Very rarely there something wrong with the detector, diffractometer, etc. **Contact the beamline**

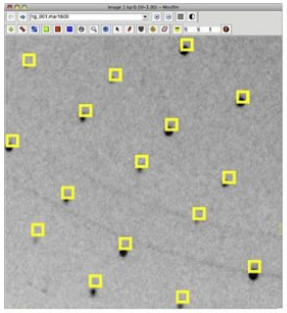
Integration

Integration overview

Add upp (integrate) all the pixel counts in the predicted spot positions.

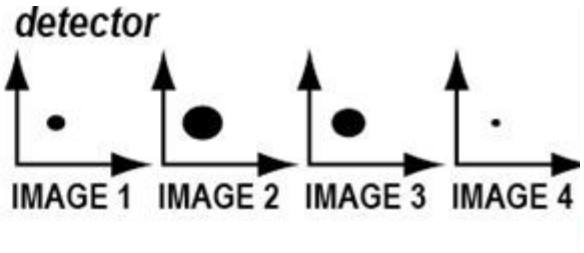
Steps:

- Positional refinement of the crystal (cell, orientation and mosaicity) and experiment parameters (detector position and orientation; beam divergence) to minimize the difference between predicted and observed spots.
- Integration of all the pixels in the reflection, estimation of standard deviations.
- Post-refinement after integration to determine accurately the center of partial reflections and get better crystal parameter values (this step is sometimes done during scaling or skipped)
- Correction of intensities for well determined factors (Lorentz, polarization, air and detector absorption)



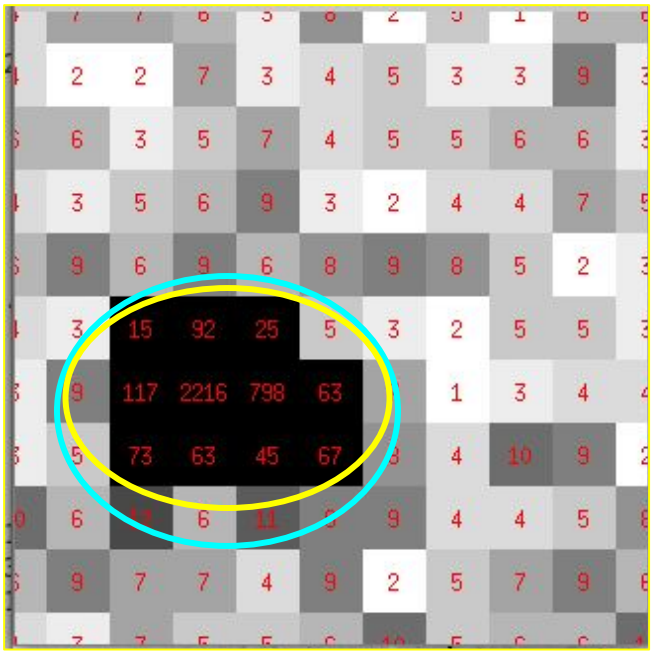
Ref: Harry Powell

Difference between prediction after indexing and actual spots



Reflection recorded partially (in different images).

Integration



How do we integrate this reflection?

We **add** the counts in the pixels that belong to a reflection and **subtract** the estimated background counts in these pixels.

It is important to have a good “flat” area around the spots. Moving the detector back can help.

Instead of simple summation, it is possible to weight the contribution of different pixels based on the profile (shape, size) of strong reflections in different parts of the detector. This improves the statistics for weak spots.

Standard deviations can be calculated as $N^{1/2}$ (N is the actual number of photons and requires knowledge of the detector conversion rate).

See Leslie, https://doi.org/10.1107/S0907444905039107

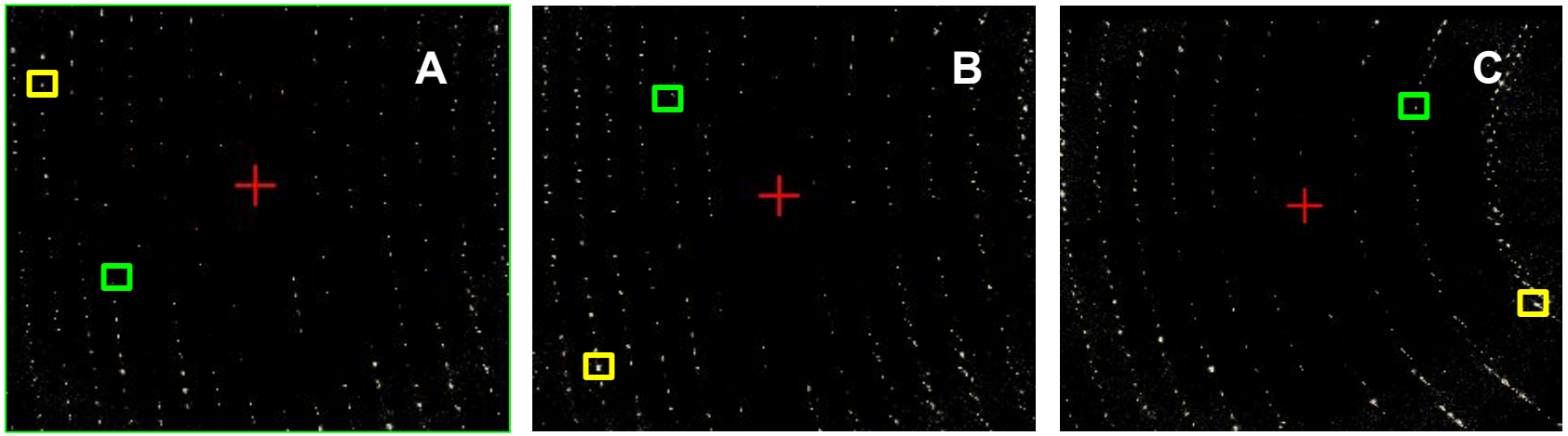
Scaling

Scaling overview

	hkl_1	I_1	hkl_2	I_2
Image A	1 2 3	100	-5 2 9	15
Image B	1 -2 -3	110	-5 -2 9	30
Image C	-1 -2 3	90	5 2 9	32

Example of simple scaling between images:

Which scale between images A, B and C minimizes the differences between reflection hkl_1 and hkl_2?



Scaling overview

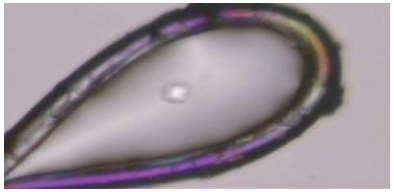
Scaling consists in the comparison of symmetry-related reflections (expected to have the exact same intensity, in theory) and calculation of corrections to put them in the same scale. Scaling can be done within a data set or between data sets from the same or different crystals.

Steps:

- Symmetry determination (but the space group is still a hypothesis until the structure has been determined!)
- Determination of scale factors for the intensities and recalculation of the intensity uncertainties.
- Rejection of outliers (intensities which are too low or too high).
- Statistical analysis.

Sources of differences in intensities

- Random errors (counting statistics).
- Non-isomorphism (scaling data from different crystals, but also from radiation damage).
- Intensity fluctuations at the beamline, other instrumentation issues.
- Bad samples (disorder, ice). Bad data processing.
- Radiation damage.
- Primary and secondary absorption of X-rays (by the crystal, buffer, crystal support).
- Changes in exposed volume (miscentering, crystal larger than the beam).



Scaling can serve as empirical correction for some suboptimal experiment setup (not completely, not always). **To the extent that is possible, optimize sample preparation and data collection to minimize differences arising from experimental issues.**

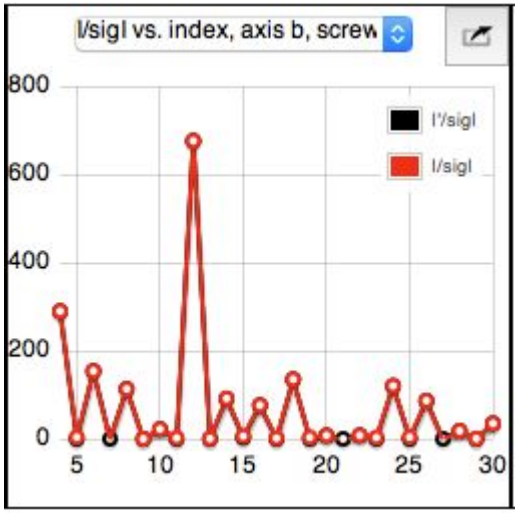
Symmetry determination

Lattice group name P 6 2 2

Reindex operator from input to lattice: [-1/2h-1/2k,-1/2h+1/2k,-l]

Likelihood	CC	R		Symmetry
0.927	0.91	0.176		identity
0.913	0.89	0.191	***	2-fold l (0 0 1) {-h,-k,l}
0.050	0.03	0.686		2-fold k (0 1 0) {-h,h+k,-l}
0.942	0.94	0.146	***	2-fold h (1 0 0) {h+k,-k,-l}
0.056	0.00	0.631		2-fold (1-1 0) {-k,-h,-l}
0.048	0.05	0.697		2-fold (2-1 0) {h,-h-k,-l}
0.939	0.93	0.138	***	2-fold (-1 2 0) {-h-k,k,-l}
0.054	0.01	0.669		2-fold (1 1 0) {k,h,-l}
0.058	-0.00	0.731		3-fold l (0 0 1) {k,-h-k,l}{-h-k,h,l}
0.058	-0.00	0.691		6-fold l (0 0 1) {h+k,-h,l}{-k,h+k,l}

Pointless documentation in ccp4i2.gitlab.io



We determine the Laue space groups by comparing the intensities for separate symmetry operations.

Screw axes can be detected by analysing systematic absences along an axis of the reciprocal cell (*eg,* in the space group $P2_1$ the reflections 0k0 are only present when k is even.

The analysis of the intensities can also detect twinning, pseudo-translation, etc.

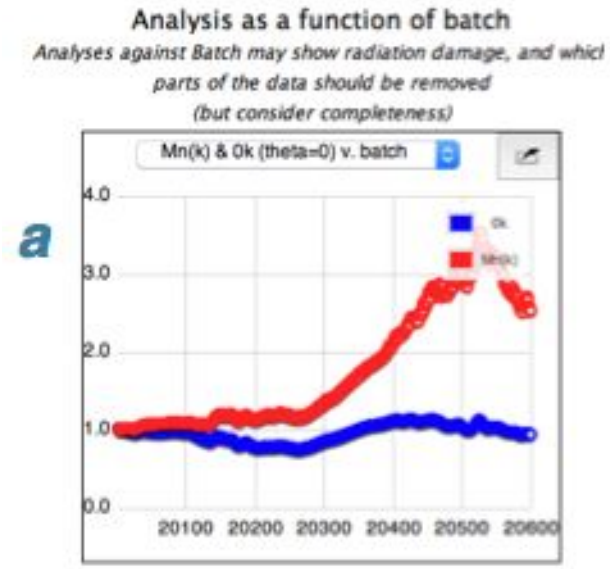
However, it cannot distinguish between enantiomorphic space groups.

Spacegroup	ITnumber	TotalProb	SysAbsProb	Reindex	Conditions
P 62 2 2	180	0.637	0.637	[h,k,l]	00l: l=3n
P 64 2 2	181	0.637	0.637	[h,k,l]	00l: l=3n

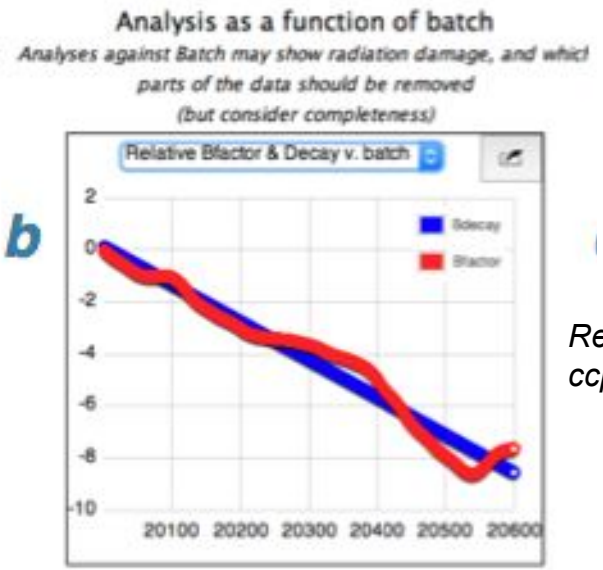
Scaling factors

Scale factors are determined such that they minimize the differences between intensities of symmetry-related reflections collected on different images or on different parts of the detector. A temperature factor-like scale, and a zero-dose correction can also be applied.

Always look at the scaling factors. Scaling as a function of image should vary smoothly and in a consistent way for reflections at different resolutions.



Scale factors: zero θ (blue) and average over resolution (red)



Relative B-factor showing radiation damage

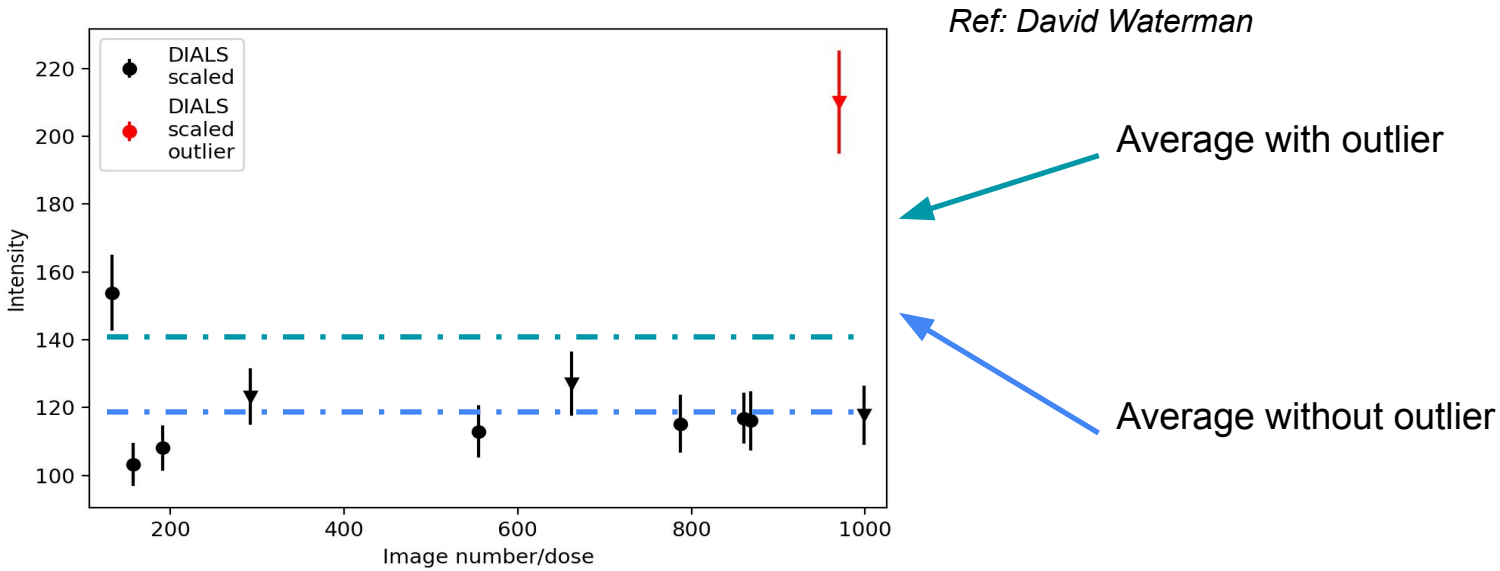
Ref: Scaling and merging statistics in ccp4i2.gitlab.io

Beware of outliers!

Outliers can throw off the estimate of the reflection intensity. Thus programs calculate the normalised deviation from the symmetry equivalents, and reject measurements above a threshold.

Sometimes outliers indicate bad areas of the detector that need to be masked. **Look at the diffraction images and redo spot finding if needed.**

Note: Anomalous pairs (h k l and -h -k -l) can very often be treated as symmetry equivalents **but**…if the anomalous signal is very strong they could end up rejected as outliers. Make sure that the software is treating them correctly.



Ref: David Waterman

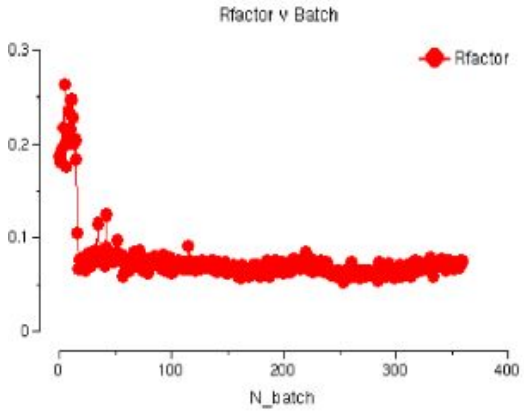
Quality indicators

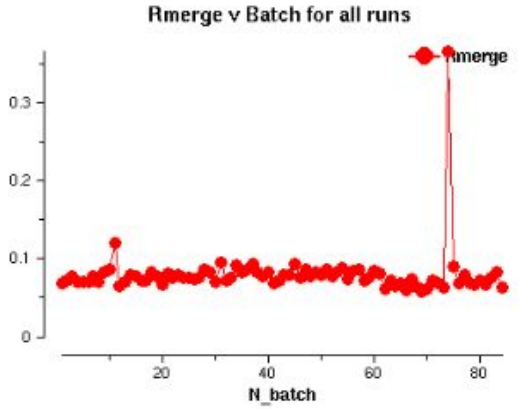
Most programs output several statistics to evaluate the agreement of intensities (R-values, correlation coefficients, and I/σ(I)). *See Karplus & Diederichs, https://doi.org/10.1016/j.sbi.2015.07.003*

Quality indicators

R factors measure the ratio between the sum of deviations from the average value of the intensity and the sum of all the measurements. The smaller the better.

Useful for comparisons across images and stronger reflections at lower resolution. Values much higher than 0.01 at low resolution indicate bad data or problems with indexing. R factor are useless to estimate the data quality for weaker reflections.





- Rmeas is preferred to Rmerge/Rsym because it is less sensitive to multiplicity. Rpim decreases with multiplicity. R-split (used for SX data) is similar to Rpim.
- Rd and Rcum are designed to detect radiation damage.
- Ranom can be used to detect anomalous differences to low resolution.

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

Quality indicators

Correlation coefficients (like $CC_{1/2}$) have better statistical properties than other quality indicators because it is easier to estimate their significance and do not depend on estimating σ(I) correctly. They tend to be more reliable to estimate data quality at high resolution and to detect the presence of small signals (eg CC_{anom}) .

Calculation of $CC_{1/2}$ (*Ref: Kay Diederichs*)

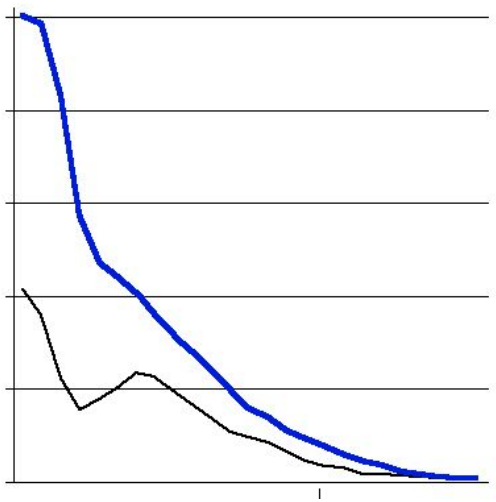
H,K,L	I_i in order of measurement	Assignment to half-dataset	Average I of X	Y
1,2,3	100 110 120 90 80 100	X, X, Y, X, Y, Y	100	100
1,2,4	50 60 45 60	Y X Y X	60	47.5
1,2,5	1000 1050 1100 1200	X Y Y X	1100	1075
...				

$$CC = \frac{\sum_j (x_j - \langle x \rangle)(y_j - \langle y \rangle)}{\sqrt{\sum_j (x_j - \langle x \rangle)^2 \sum_j (y_j - \langle y \rangle)^2}}$$

Quality indicators

<I/σ(I)> (NOT $<I_i/\sigma_i>$ for individual reflections) is also used to provide an estimate of the diffraction limit of the crystal. σ(I) needs to be calculated properly:

- Correction factors are applied to make the variance of the intensity deviations around 1 for all intensity ranges. Eg, a, v0 and b in XDS, Sdfac, Sdadd and SdB in AIMLESS.
- ISa = $1/ (ab)^{1/2}$ (from XDS) is an estimate of the maximum I/σ(I) (ie, for an infinite intensity) possible for the data. ISa is limited by systematic errors.



$<I/\sigma(I)>$ and $<I_i/\sigma_i>$ as a function of resolution

Ref: David Waterman

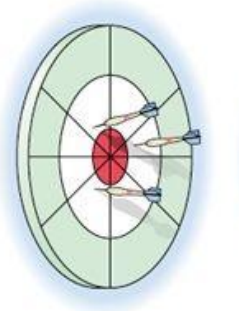
Advice: Do not remove any data based just on poor values of quality indicators at processing.
Test first if these data can improve the structure model (e.g, doing "paired refinement", *Karplus and Diederichs* https://www.science.org/doi/10.1126/science.1218231)

Other steps

To merge or not to merge

Options:

1. The symmetry equivalent reflections are averaged after scaling (keep separate records of the intensities for the anomalous pairs I+ and I- , they may carry unbiased structure information).
2. Each scaled intensity is written out with its original index.



Merging is often a good practice since averaging increases the accuracy of a measurement.

In some cases it can be useful to preserve the unmerged measurements::

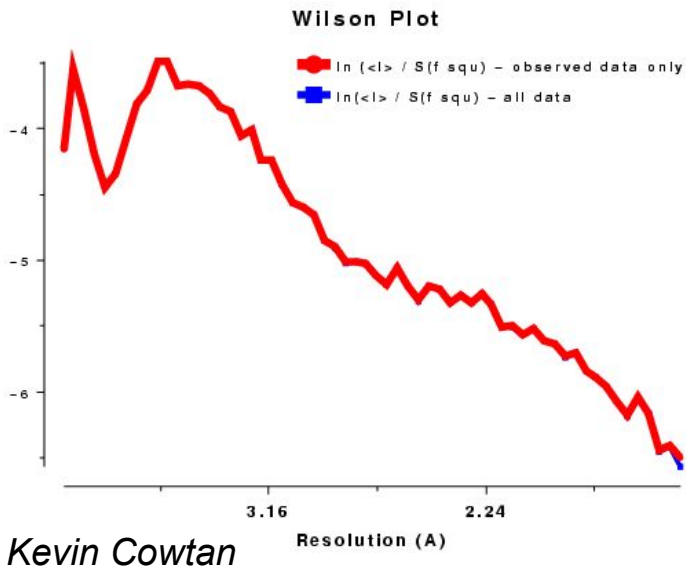
- Unmerged data can be reanalyzed with other software (eg, symmetry analysis with CCP4 pointless or Phenix xtriage).
- Phasing programs (eg. shelx, autosol, autosharp) may also perform better with unmerged data, as they can use their own local scaling routines for improved estimates of anomalous differences.

Structure factor amplitudes

I(hkl) is the energy that the diffracted X-ray dumps on the detector, but the amount that really describes the diffraction by the structure is the structure factor F(hkl)

Theoretically $|F^2| = I$, but we need an estimate of contents of the cell to put F in an absolute scale.

Negative intensities can be transformed ("truncated") into positive amplitudes using the Wilson distribution. *Ref: French and Wilson* https://doi.org/10.1107/S0567739478001114



Ref: Kevin Cowtan

Wilson plot: $\ln(\langle I \rangle / \Sigma f^2)$ as a function of resolution, named after Arthur Wilson. Deviations from a straight line at high resolution can be an indication of problems with the data or during processing

Thanks for your attention!

Main programs used for data processing

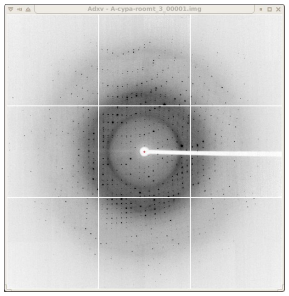
- **HKL2000, HKL3000** (denzo, scalepack) *https://hkl-xray.com/about-hkl-research*
- **XDS** (indexing to scaling) **XSCALE** (scaling of one or different data sets) **XDSCONV** (I to |F|) . GUIs: XDSGUI, XDSAPP, etc. *https://xds.mr.mpg.de/*
- **DIALS** (indexing, integration) GUI: DUI. *https://dials.github.io/index.html*
- **Mosflm** (indexing, integration). GUI: imosflm CCP4 software (**pointless**, **aimless**, **ctruncate**) is used to determine the symmetry, scale, merge and calculate |F|. *https://www.mrc-lmb.cam.ac.uk/mosflm/mosflm/*

There is no program that always outperforms the others. If you have a difficult case, try several.

Most synchrotron beamlines have automated processing pipelines that run one or more of these programs. Usually you can get the XDS (CORRECT.LP) or aimless output in a couple of minutes. **Become very familiar with at least one of those or both to be able to figure out quickly if you need to modify your data collection**.

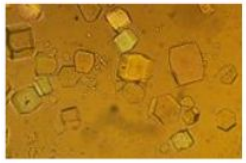
Final words

Collect the best data you can

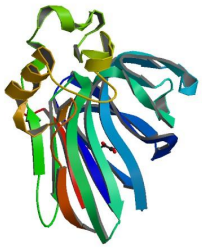


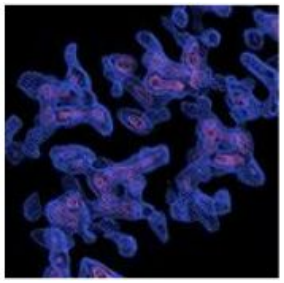
Data processing will be easier

$$|F_o(hkl)|$$



Prepare your sample properly





So will all the subsequent steps