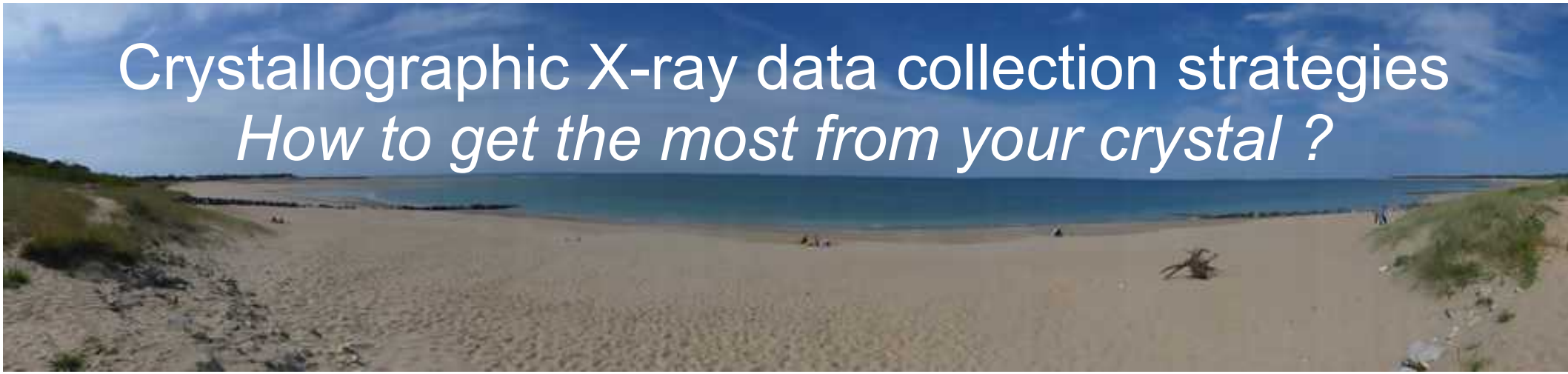


Crystallographic X-ray data collection strategies

How to get the most from your crystal ?

Dominique HOUSSET
E-mail : dominique.housset@ibs.fr

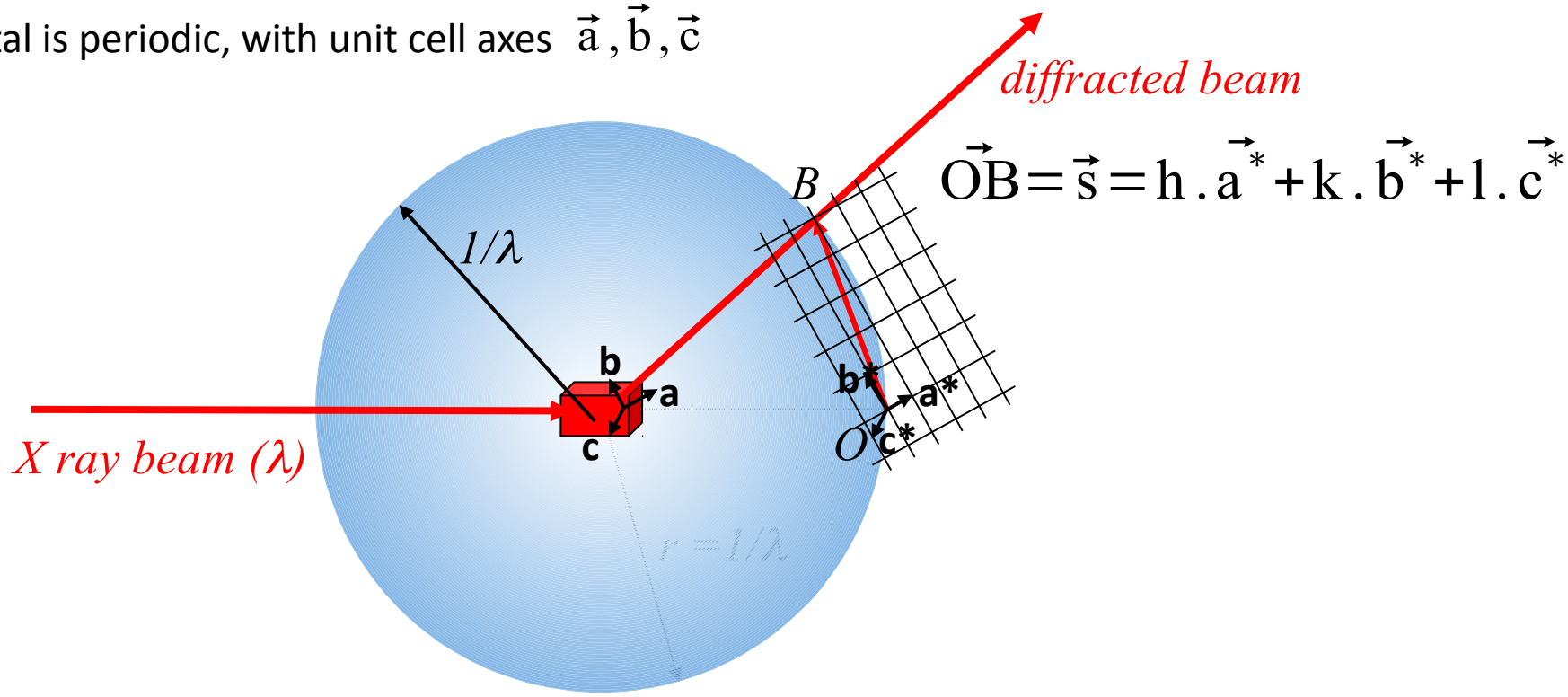
Starting point ...

- You just have placed your crystal on the head of the goniometer (on a inhouse diffractometer or at a synchrotron beamline)
  - What should you do obtain the best 3D structural information out of your crystal ?

Ewald sphere

- The Ewald sphere tells you which reflections you will be collecting
  - Which part of the reciprocal space you will have access to ?

The crystal is periodic, with unit cell axes $\vec{a}, \vec{b}, \vec{c}$



A wave is scattered when a node of the reciprocal lattice (indices h k l) touches the Ewald's sphere. The structure factor (amplitude F and phase ϕ) of the diffracted wave is :

$$F(\vec{s}) = N_{cell} \sum_J f_J . \exp[2i\pi(\vec{s}).\vec{r}_J]$$

Rotating the crystal

- To obtain complete 3D information, the crystal is rotated
- In theory …
  - The region of reciprocal space you have access to is the volume scanned by the Ewald sphere when it rotates around O
    - It is a tore, centrered in O
    - If you rotate the crystal by 360°, you will collect reflections corresponding to all the nodes included in tore
    - This technique of data collection is called the oscillation method or rotating crystal method
- In practice the resolution is also limited by :
  - The size and shape of your detector : θ_{max} due to the experimental setup
  - The sample that may not diffract strongly enough above a given resolution

Complete data

- Complete data at a given resolution in the reciprocal space
  - Reflections corresponding to nodes included in a sphere of radius 1/d, centred in O
  - ≠ from the tore
  - Even if you rotate your crystal by 360°, you miss a region of the reciprocal space :
    - The one located close to the rotation axis
- To have access to all the reciprocal space, one should collect rotation crystallographic data with at least 2 different rotation axes
- **But**, you may have symmetries in the reciprocal space !

Fridel's law

- **Fridel's law**
  - In the absence of anomalous dispersion:

$$F(\vec{s}) = \sum_J f_J . \exp[2i\pi(\vec{s}).\vec{r}_J] \text{ with } f_J \text{ real (no anomalous dispersion)}$$

$$F(-\vec{s}) = F^*(\vec{s})$$

$$I(\vec{s}) = F(\vec{s}).F^*(\vec{s}) = F^*(-\vec{s}).F(-\vec{s}) = I(-\vec{s})$$

- Since $\vec{s} = h.\vec{a^*} + k.\vec{b^*} + l.\vec{c^*}$ we often note: I(hkl) = I(-h-k-l)

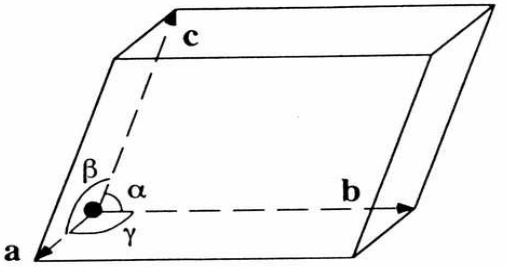
- ==> **Centre of symmetry in the reciprocal space**
- Only half of the sphere is required to get complete data

Symmetry in crystals => symmetry in reciprocal space

- Symmetries in crystals will induce symmetries in the diffraction pattern
  - A quick look at symmetries within crystals

Crystalline systems

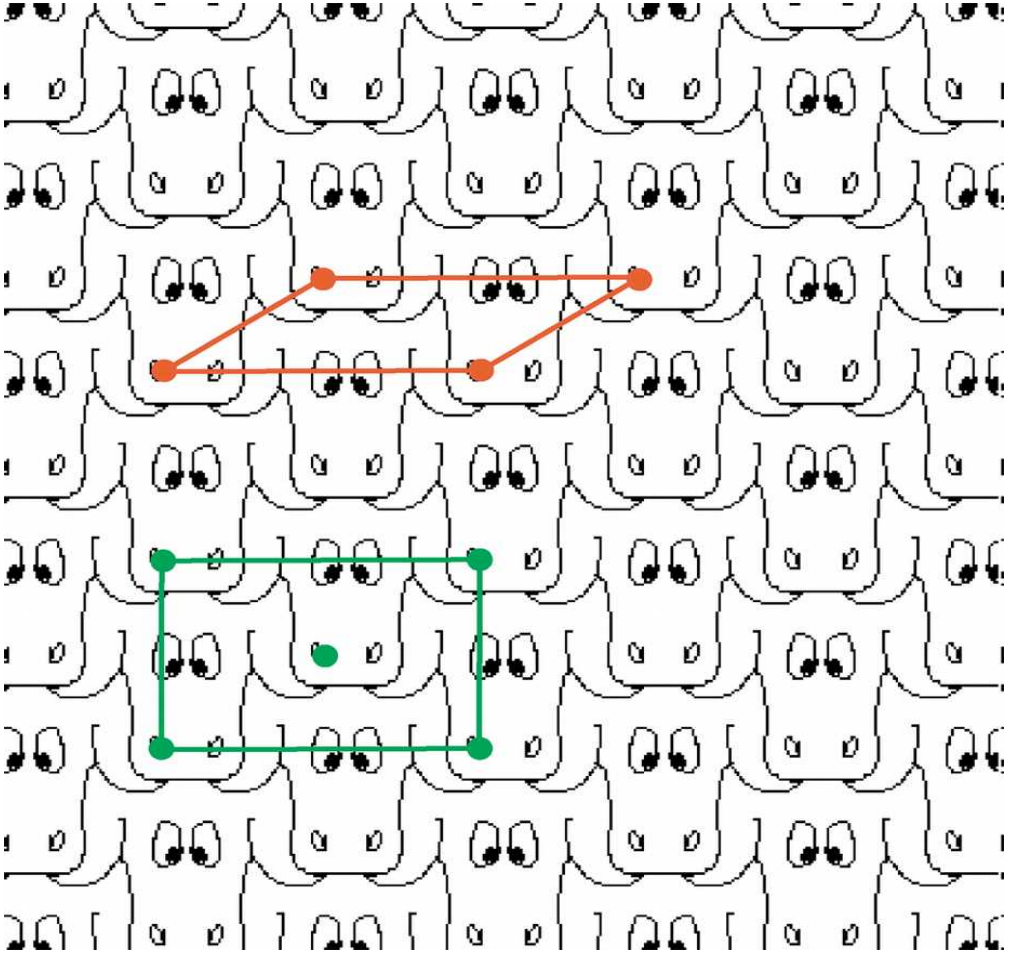
- There are only **seven** ways of paving the space in 3 dimension in a periodic manner:



Crystal system	Bravais lattice	Rotational symmetry	Cell parameters
Triclinic	P	none	a≠b≠c α≠β≠γ
Monoclinic	P, C	2-fold axis (//b)	a≠b≠c α=γ=90°≠β
Orthorhombic	P, C, I, F	3 2-fold (// a,b,c)	a≠b≠c α=β=γ=90°
Tetragonal	P, I	4-fold (//c)	a=b≠c α=β=γ=90°
Trigonal	P (R)	3-fold (//c)	a=b≠c α=β=90° γ=120°
Hexagonal	P	6-fold (//c)	a=b≠c α=β=90° γ=120°
Cubic	P, I, F	3-fold (//diag)	a=b=c α=β=γ=90°

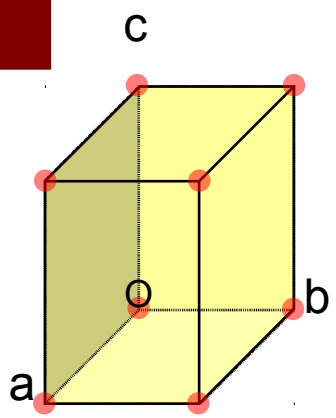
Primitive or multiple cells ?

- A primitive cell contains one lattice point per cell
- A multiple cell contains more than one lattice point per cell
- All crystals can be described by a primitive cell
- One use a multiple cell only when it has a higher symmetry than the primitive one

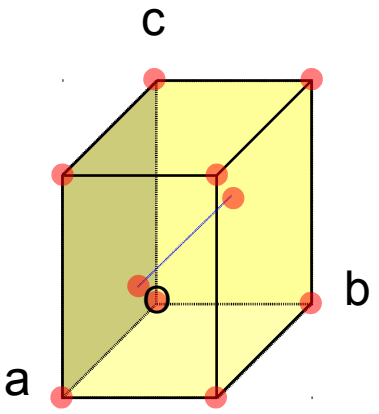


Three types of multiple cells

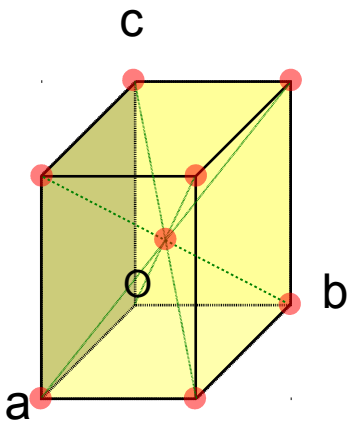
- Lattice P : Primitive
  - 1 node per unit cell: x, y, z



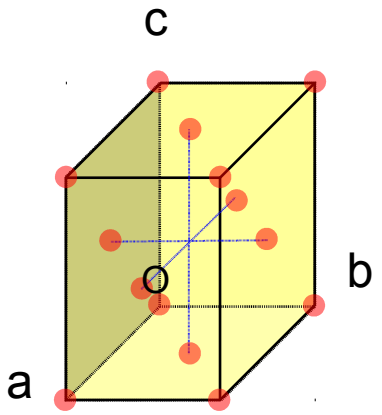
- Two faces centred, lattice C, B, A
  - 2 nodes on the (b, c) face for the A-type
  - x, y, z & x, y+1/2, z+1/2



- Boby centred, lattice I
  - 2 nodes: x, y, z & x+1/2, y+1/2, z+1/2



- All faces centred, lattice F
  - 4 nodes : x, y, z; x+1/2, y+1/2, z; x+1/2, y, z+1/2; x, y+1/2, z+1/2



14 Bravais lattices

- Triclinic P
- Monoclinic P,C
- Orthorhombic P,C,I,F
- Tetragonal P,I
- Rhombohedral R
- Hexagonal P
- Cubic P,I,F

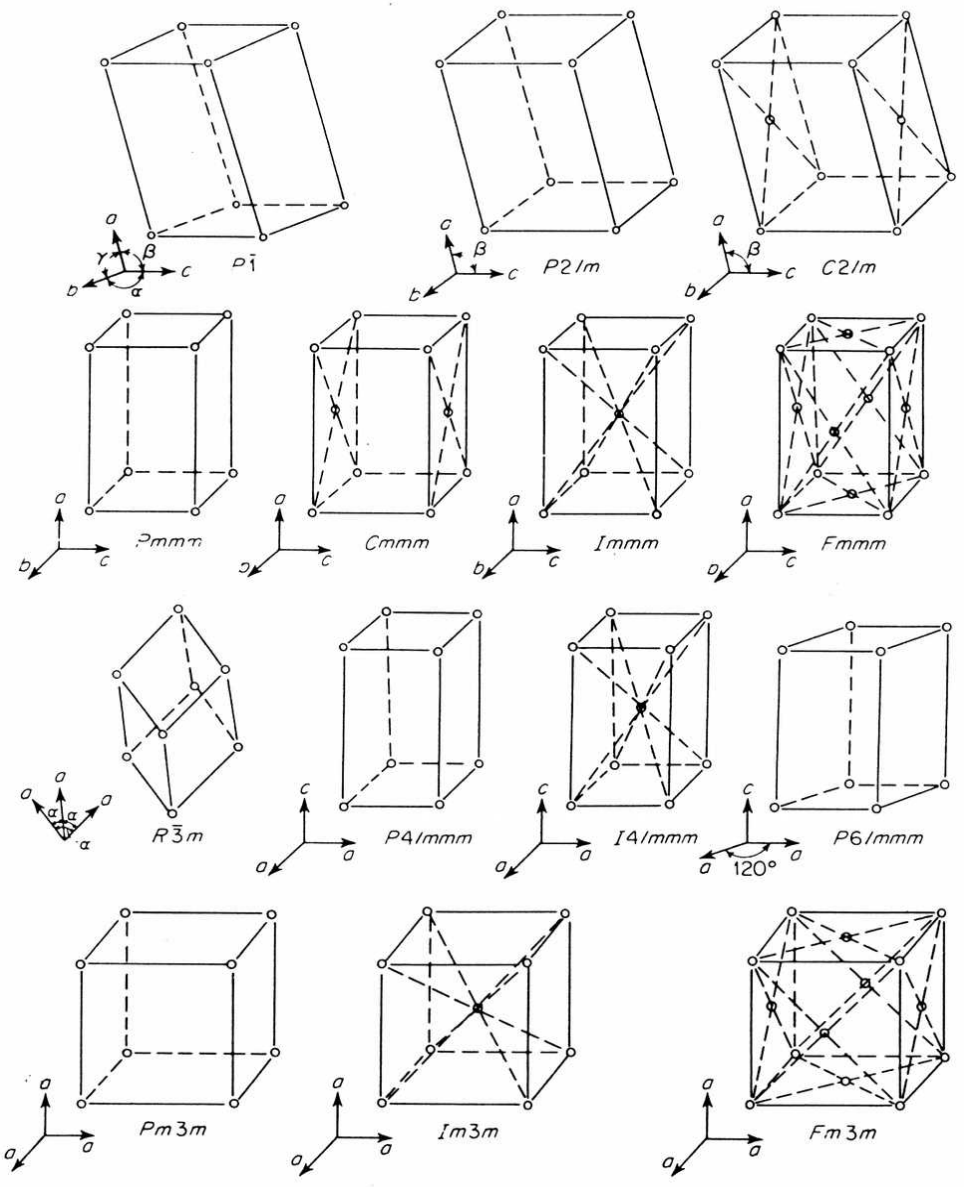
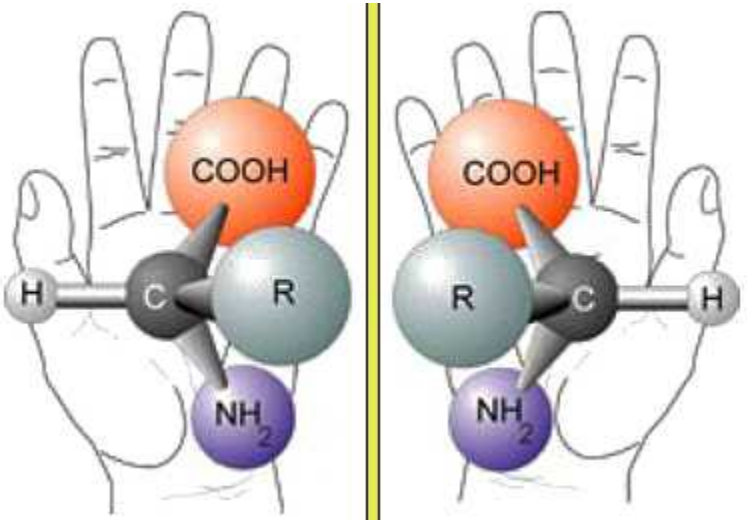


Figure 3.13. The 14 Bravais lattices.

Space groups and chirality

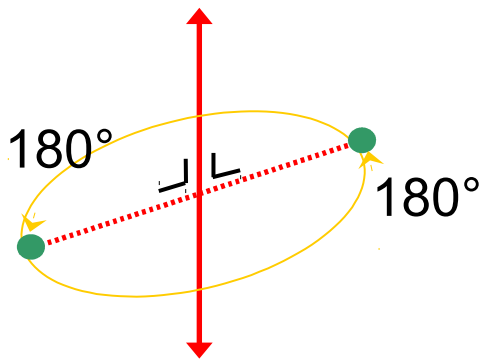
- There are **230 space groups** corresponding to the 230 manner to associate these symmetry elements with the different unit cells described by the **14 Bravais lattices.**
- Macromolecules are chiral (L-aminoacids in protein): it cannot be superposed on its image by a mirror:

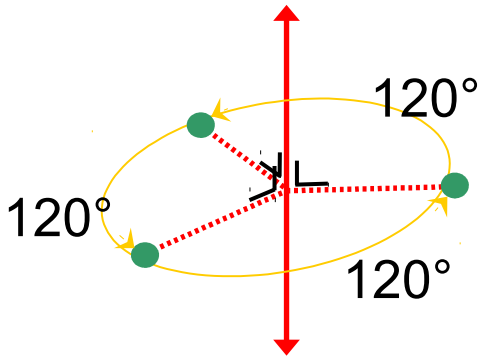


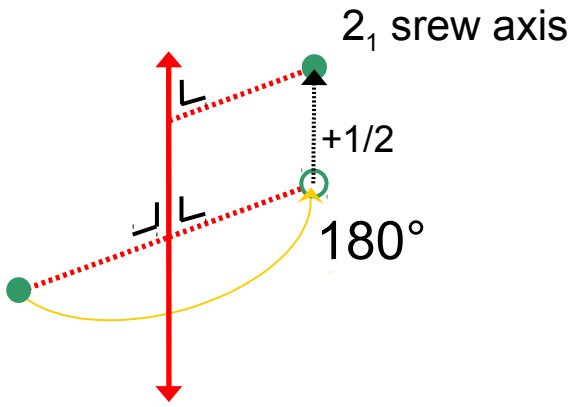
- mirror and center of symmetry are not allowed.
- Only **65 space groups** for biological macromolecules.

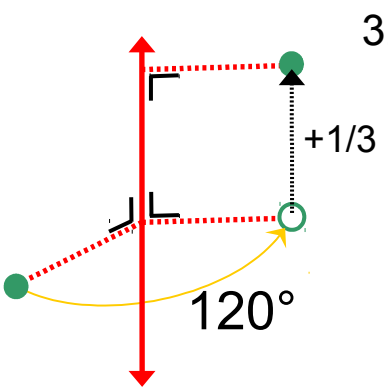
Symmetries and space group for protein crystals

- Symmetry operators are limited:
  - Axes 2 or 2_1 (axe 2 plus translation of 1/2 along axis 2)
  - Axes 3, 3_1 or 3_2 (axe 3 plus translation of 1/3 or 2/3 along axis 3)
  - Axes 4, 4_1, 4_2 or 4_3 (axe 4 plus translation of 1/4, 1/2 or 3/4 along axe 4)
  - Axes 6, 6_1, 6_2, 6_3, 6_4 or 6_5 (axe 6 plus translation of 1/6, 1/3, 1/2, 2/3 or 5/6 along axe 6)









What are the consequence in the reciprocal space of symmetries in the real space ?

- A rotation or screw rotation axis of order n, parallel to a given direction in the real space will induce a rotation axis of same order, parallel to the equivalent direction in the reciprocal space.
  - In space group ***P***2_1, there is a screw rotation axis of order 2, along **b**
  - => rotation axis of order 2 along **b*** in the reciprocal space
  - In space group ***P***$4_3 2_1 2$, there is a screw rotation axis of order 4, along **c**
  - => rotation axis of order 4 along **c*** in the reciprocal space
  -
- In ***P***2_1 space group, what are the reflections equivalent (same intensity) by symmetry?

11 Laue classes in reciprocal space

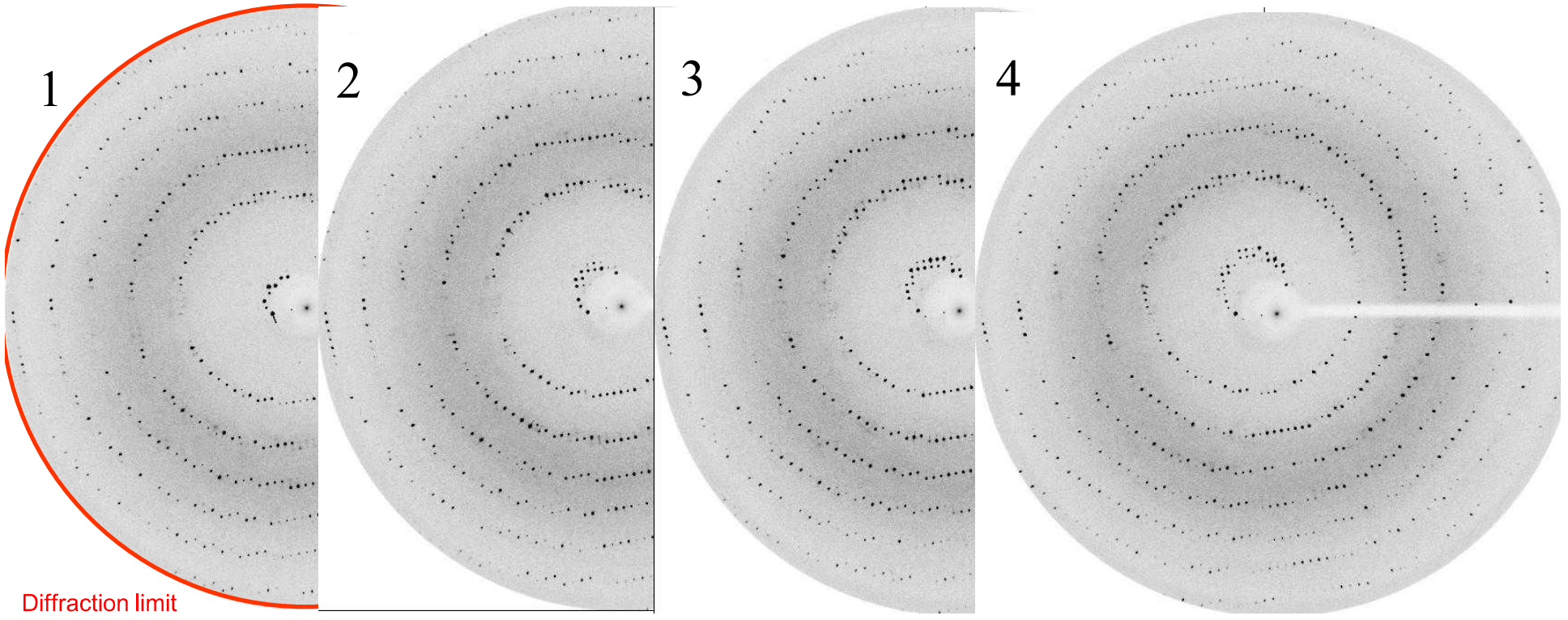
- If the Fridel's law is valid (always, except for non-centrosymmetric crystal in the presence of significant anomalous dispersion)

Crystal System	Laue Class	Space Group	Nb of eq. Refl.
Triclinic	-1	***P***1	2
Monoclinic	2/m	***P***2, ***P***2_1, ***C***2	4
Orthorhombic	mmm	***P***222, ***P***222_1, ***P***2_12_12, ***P***$2_12_12_1$, ***C***222_1, ***C***222, ***F***222, ***I***222, ***I***$2_12_12_1$	8
Tetragonal	4/m	***P***4, ***P***4_1, ***P***4_2, ***P***4_3, ***I***4, ***I***4_1,	8
	4/mmm	***P***422, ***P***42_12, ***P***4_122, ***P***4_12_12, ***P***4_222, ***P***4_22_12, ***P***4_322, ***P***4_32_12, ***I***422, ***I***4_122	16
Trigonal	-3	***P***3, ***P***3_1, ***P***3_2, ***R***3,	6
	-3m	***P***312, ***P***321, ***P***$3_1$12, ***P***$3_1$21, ***P***$3_2$12, ***P***$3_2$21, ***R***32	12
Hexagonal	6/m	***P***6, ***P***6_1, ***P***6_5, ***P***6_2, ***P***6_4, ***P***6_3,	12
	6/mmm	***P***622, ***P***6_122, ***P***6_522, ***P***622, ***P***6_422, ***P***6_322	24
Cubic	m3	***P***23, ***F***23, ***I***23, ***P***2_13, ***I***2_13,	24
	m3m	***P***432, ***P***4_232, ***F***432, ***F***4_132, ***I***432, ***P***4_332, ***P***4_132, ***I***4_132	48

Impact of multiple cells and helical rotation axes?

- **Systematic extinction**:
  - Multiple unit cells
    - C : h + k = 2n
    - I : h + k + l = 2n
    - F : h, k, l of same parity, i.e. h + k = 2n, k + l = 2n et h + l = 2n
  - Helicoidal rotation axes
    - Axe 2_1 / a (h00) => h = 2n
    - Axe 6_1 / c (00l) => l = 6n
    - Axe 6_2 / c (00l) => l = 3n
    - Axe 6_3 / c (00l) => l = 2n
    - Axe 6_4 / c (00l) => l = 3n
    - Axe 6_5 / c (00l) => l = 6n
- In most cases, the space group can be determined form symmetry in the reciprocal space
- However, we can't distinguish between ***P***6_5 and ***P***6_1, ...

Collecting data: the oscillation method



Essential steps of data processing

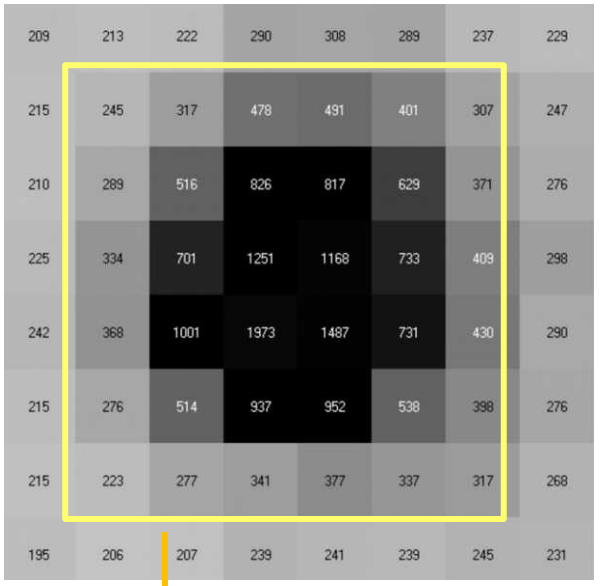
- **Several steps**:
  - 1) *Indexation*
  - 2) *Integration*
  - 3) *scaling & merging*
  - 4) *providing data statistics*

Indexation in theory

- Localisation of bright, well define spots on several images
  - Determination of x,y coordinates and rotation angle φ for each spot; need for 100 – 1000 spots
- Calculation of coordinates in reciprocal space
  - Calculation of $\vec{s}_i$ for spot i
  - Each $\vec{s}_i$ vector should be a vector of the reciprocal lattice: $\vec{s}_i = h_i . \vec{a}^* + k_i . \vec{b}^* + l_i . \vec{c}^*$
- Determination of difference vectors between reciprocal lattice points
  - $\vec{s}_{ij} = \vec{s}_j - \vec{s}_i$ Advantage: the same vector is found many times: better precision
- Fitting of cell and orientation parameters on this sets of vectors
  - Determining a unit cell and an orientation matrix that can describe all these vectors by:
  - $\vec{s}_{ij} = h_{ij} . \vec{a}^* + k_{ij} . \vec{b}^* + l_{ij} . \vec{c}^*$
- Cell parameters, orientation matrix, quality of fit between predicted and observed spots provided.
  - All spots on all images can be associated with its three indices h,k,l that define its position in the reciprocal space

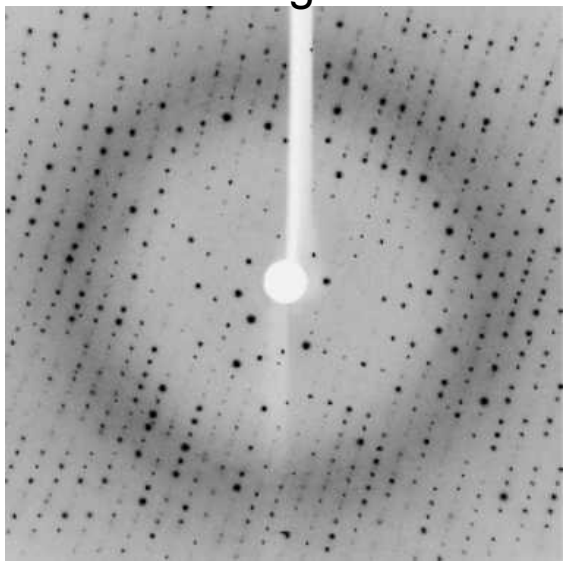
Indexation in practice

- Unit cell axes and orientation matrix are determined
- Indices h,k,l are associated to each spots

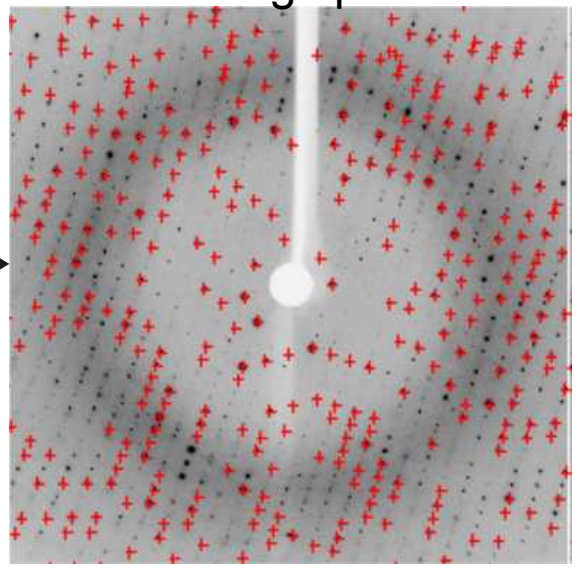


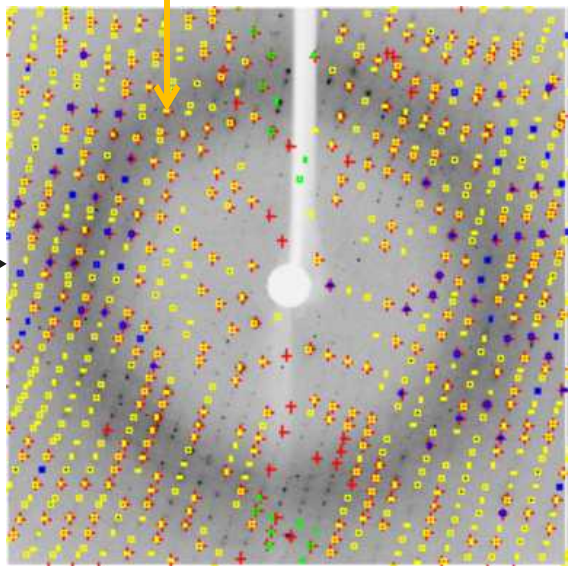
Autoindexing
Cell parameters a, b,c α,β,γ

Image 0-1°



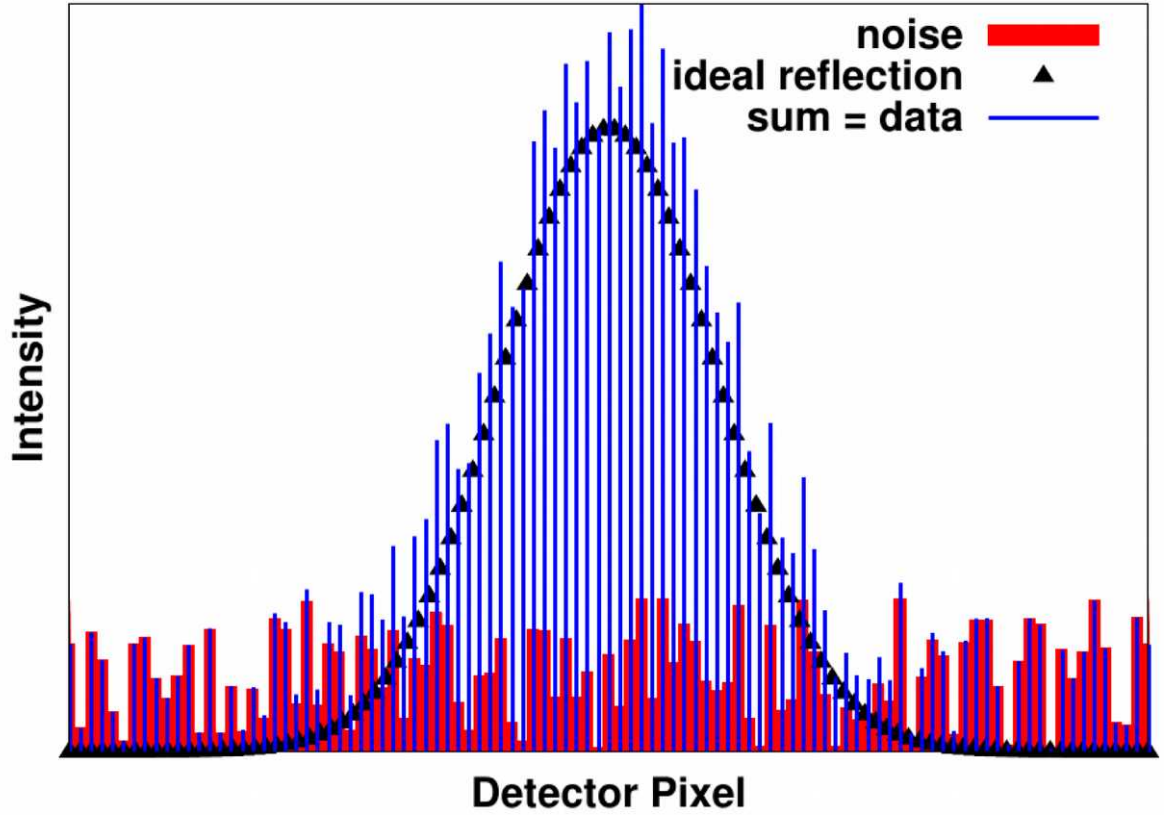
Finding spots





Integration

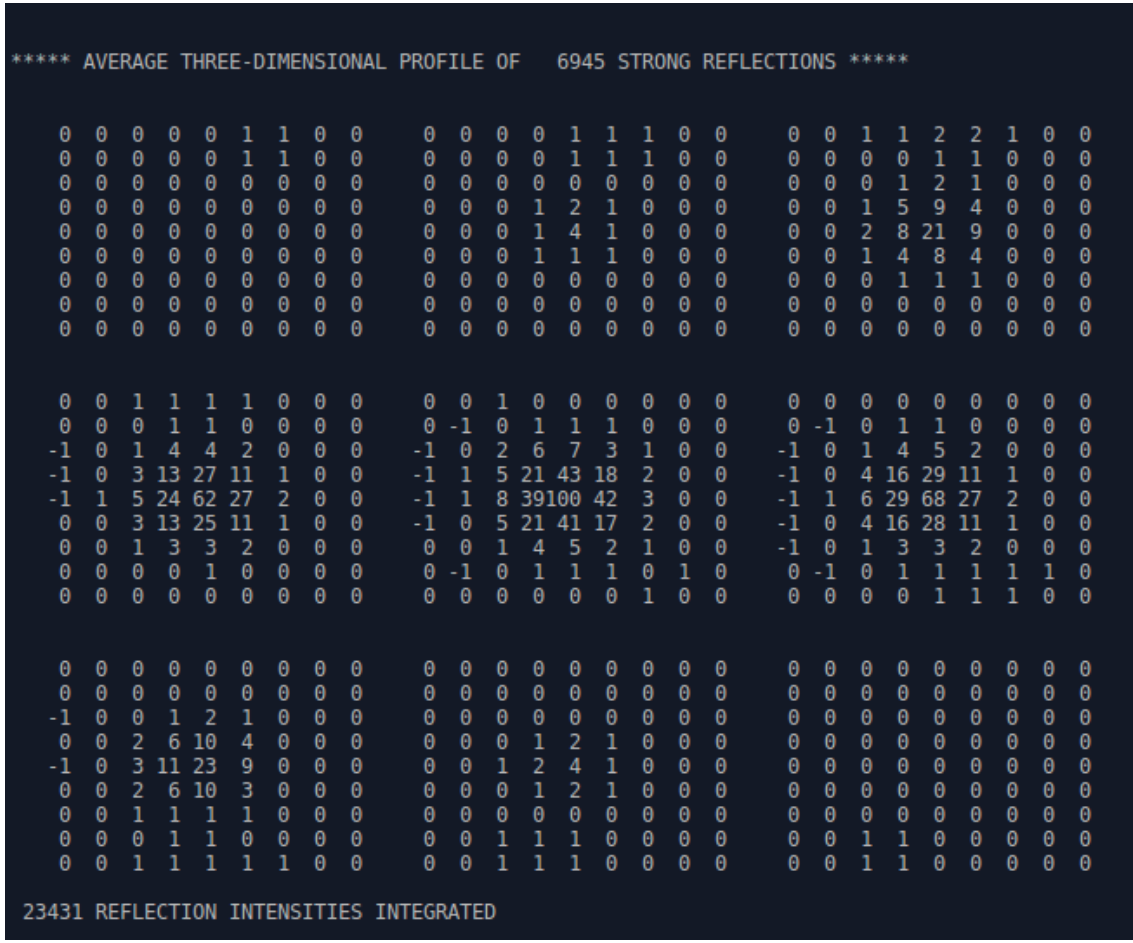
- Since the position of each spots on each image is now known, the intensity is measured for all of them.
  - Background needs to be removed
  - Reflections splitted on more than on image are summed
  - 3D profiles of reflections are calculated



3D Profile fitting

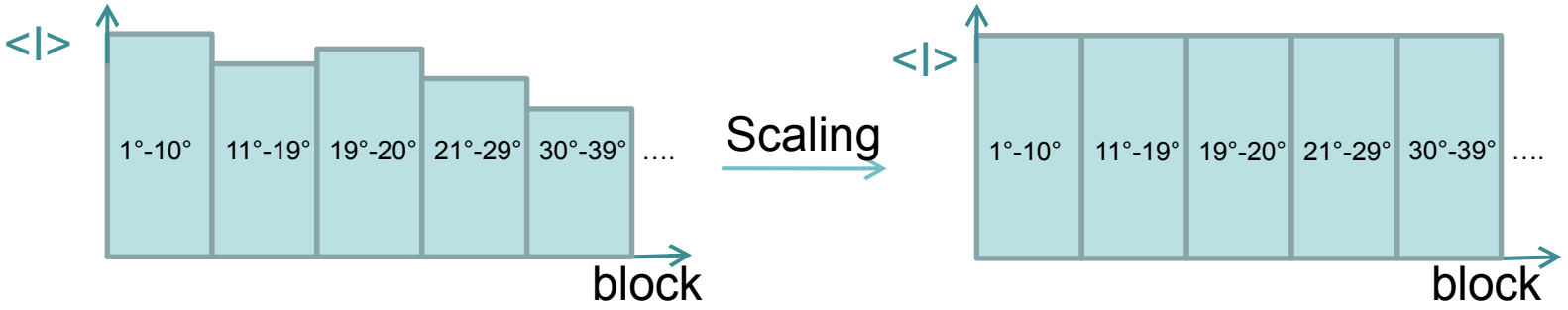
- A diffraction spot can present on several successive images
  - Several parameters impact the number of images on which a given diffraction spot is present:
    - Crystal mosaicity
    - Beam divergence
    - Distance from rotation axis
    - Wavelength dispersion
    - Oscillation range chosen for the data collection (0.05° to 1°)
  - If most spots are on more than one image, 3D (x,y,φ) is preferable

3 dimensionnal profiles with XDS



Scaling & Merging

- 3) *scaling & merging*
  - Determination of the symmetry in the reciprocal space
  - Scaling of all images based on background and then on symmetry related reflections



  - Guess about the symmetry of the crystal
  - Calculation of unique reflections from multiple measurements of symmetry related reflections

- 4) *provide a file containing merged reflections and* ***data statistics***

Reflection file

- Amplitude of structure factor are calculated from intensities : $I = F^2$

h	k	l	F	SIGF	DANO	SIGDANO	F(+)	SIGF(+)	F(-)	SIGF(-)
0	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0	2	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	3	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	4	101.12	6.29	0.00	0.00	100.92	9.00	100.05	9.11
0	0	5	5087.18	868.91	5087.18	868.91	5087.18	868.91	5004.75	871.44
0	0	6	-1.00	868.91	-1.00	868.91	-1.00	868.91	5004.75	871.44
0	0	7	-1.00	868.91	-1.00	868.91	-1.00	868.91	5004.75	871.44
0	0	8	712.77	26.26	0.00	0.00	713.90	35.18	706.38	40.04
0	0	9	251303.12	24365.59	251303.12	24365.59	251303.12	24365.59	246856.75	27390.66
0	0	10	-1.00	24365.59	-1.00	24365.59	-1.00	24365.59	246856.75	27390.66
0	0	11	-1.00	24365.59	-1.00	24365.59	-1.00	24365.59	246856.75	27390.66
0	0	12	374.42	11.63	0.00	0.00	377.39	14.45	367.19	19.85
.....										
.....										
36	20	1	239.06	4.01	-32.37	8.15	221.41	6.19	253.78	5.30

Highest resolution

Statistical indicators and data quality

- I / σ(I) overall (as high as possible)
  - in the last resolution shell: I / σ(I) > 2

- The agreement between reflections equivalent by symmetry is assessed by the R_{sym} (or R_{merge}):

$$R_{sym} = \frac{\sum_{hkl} \sum_{i=1}^{n_{hkl}} \left| I_i(hkl) - \langle I(hkl) \rangle \right|}{\sum_{hkl} \sum_{i=1}^{n_{hkl}} I_i(hkl)}$$

  - n_{hkl} is the number of equivalent measurements for a reflection hkl, taking crystal symmetry and Fridel's law, typically, $2 < n_{hkl}$ 12
  - If the symmetry is correct, usually $R_{sym} < 0.1$ ($R_{sym} < 0.05$ for very good data)

Statistical indicators and data quality

- R_{sym} (or R_{merge}) is a poor, but still widely used, statistical indicator
  - Precision on intensities increases with the number of measurements
  - R_{sym} increases also with the number of measurements
    - Scientists were pushed to avoid high redundancy and to eliminate outliers to obtain better R_{sym}, at the price of a reduced precision on intensities
- Two other R-factors were introduced, to take into account the $\sqrt{n}$ error propagations:

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

$$R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n_{hkl}-1}} \sum_{i=1}^{n_{hkl}} |I_i(hkl) - \langle I(hkl) \rangle|}{\sum_{hkl} \sum_{i=1}^{n_{hkl}} I_i(hkl)}$$

Statistical indicators and data quality

- $CC_{1/2}$ (Pearson correlation coefficient) is good indicator of the significance of the signal (information content)
  - For each unique reflection hkl, the set of measurements is randomly divided into two halves, subset 1 & 2.
  - The average intensity of each unique reflection hkl is calculated for each subset
  - The correlation between averaged intensities of subset 1 and subset 2 is calculated:

$$CC_{1/2} = \frac{\sum_{hkl} \left(I_1 - \langle I_1 \rangle\right)\left(I_2 - \langle I_2 \rangle\right)}{\sqrt{\sum_{hkl} \left(I_1 - \langle I_1 \rangle\right)^2} \sqrt{\sum_{hkl} \left(I_2 - \langle I_2 \rangle\right)^2}}$$

  - The significance of $CC_{1/2}$ can be assessed by the Student's test
    - For a significance of p = 0.01:
      - $CC_{1/2} > 0.3$ for n = 100
      - $CC_{1/2} > 0.08$ for n = 1000

Karplus, P.A. and Diederichs, K. (2015) Assessing and maximizing data quality in macromolecular crystallography. Current Opinion in Struct.Biol. 34, 60-68
Karplus, P.A. and K. Diederichs, K. (2012) Linking Crystallographic Data with Model Quality. Science 336, 1030-1033

Parameters to be obtimized

- How many images to collect ?
  - Crystal symmetry, phasing method
- Which oscillation angle ?
  - Cell parameters, type of detector, type of processing
- Which crystal to detector distance ?
  - Resolution limit of the crystal, cell parameters
- Which exposure time ?
  - Type of detector, no saturated spots

How many degrees should be collected

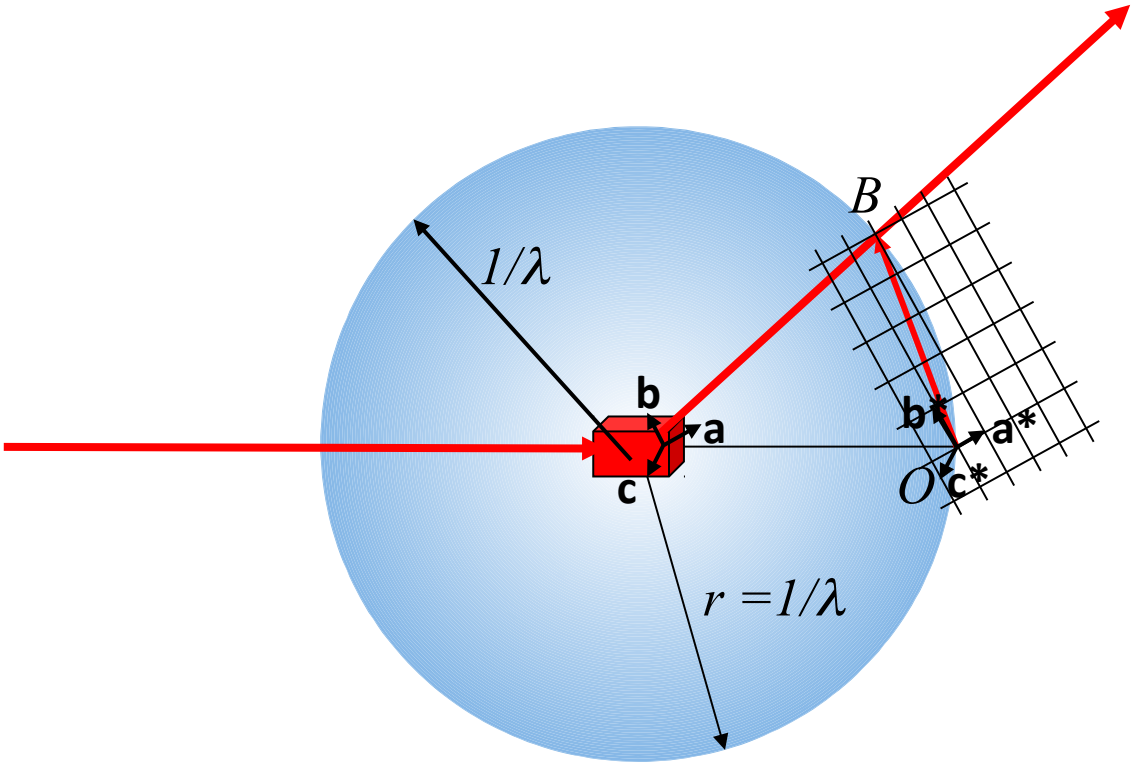
- The goal is to get 100 % complete data !
- The total rotation that should be collected depends on :
  - The symmetry of the crystal
    - The higher the symmetry, the smaller the range
  - Whether you expect anomalous signal
  - On the starting orientation of the crystal
- For a maximum completeness and a redundancy between 3 and 4, the minimum range is (if anomalous scattering is present, multiply by 2)
  - If optimal starting oritentation is used
    - ***P***1 : 360°, ***P***2_1 : 180°, ***P***222 : 90°, ***P***422 : 45°, ***P***6 : 60°, ***P***622 : 30°, …
- **But …**
  - A higher redundancy will always improve the data
  - Data collection time is no longer a limiting factor
    - The more you collect the better it is !

Optimal φ slicing ?

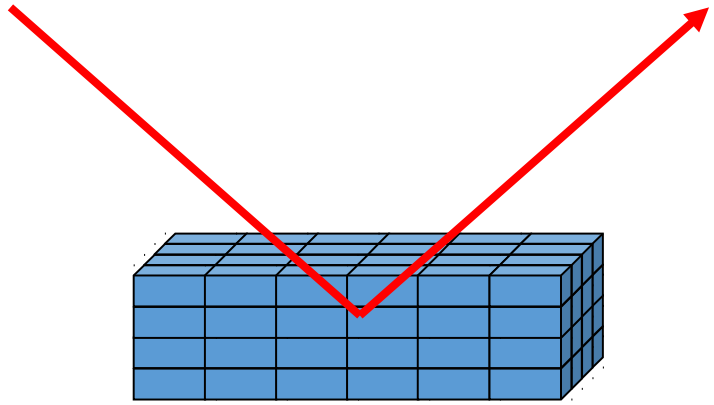
- Depends on :
  - Your crystal
  - Your beam
  - The detector type
  - The cell parameters

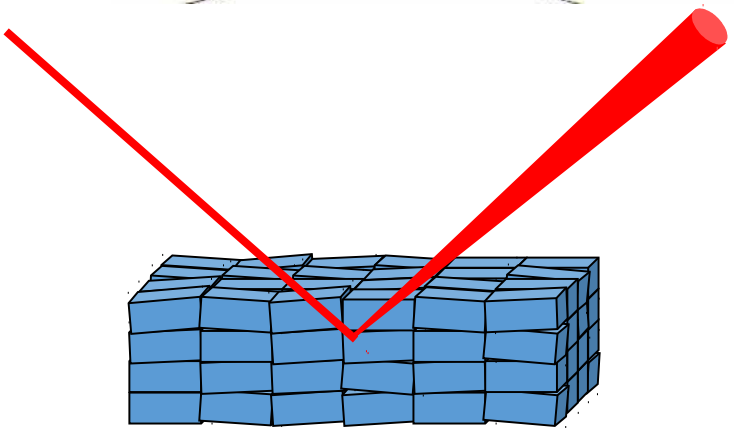
An unperfect world …

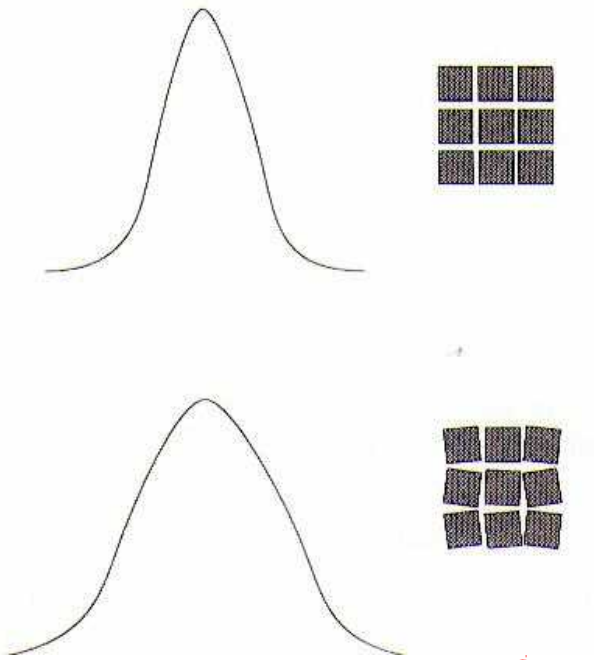
- In theory :
  - perfect crystal: reciprocal lattice is built of points
  - perfect beam (no wavelength dispersion, no divergence…) : infinitely thin Ewald sphere



let's face reality



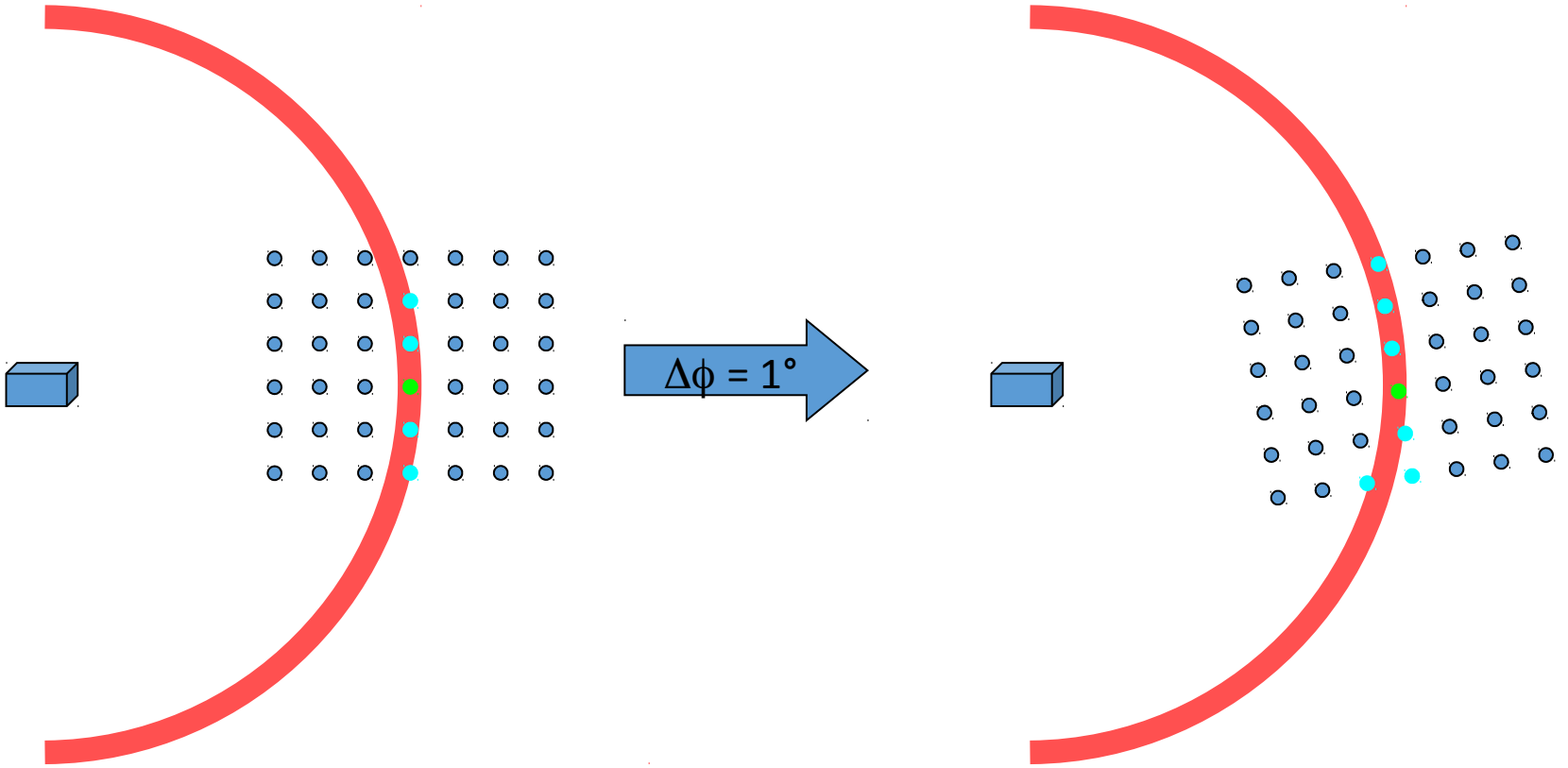




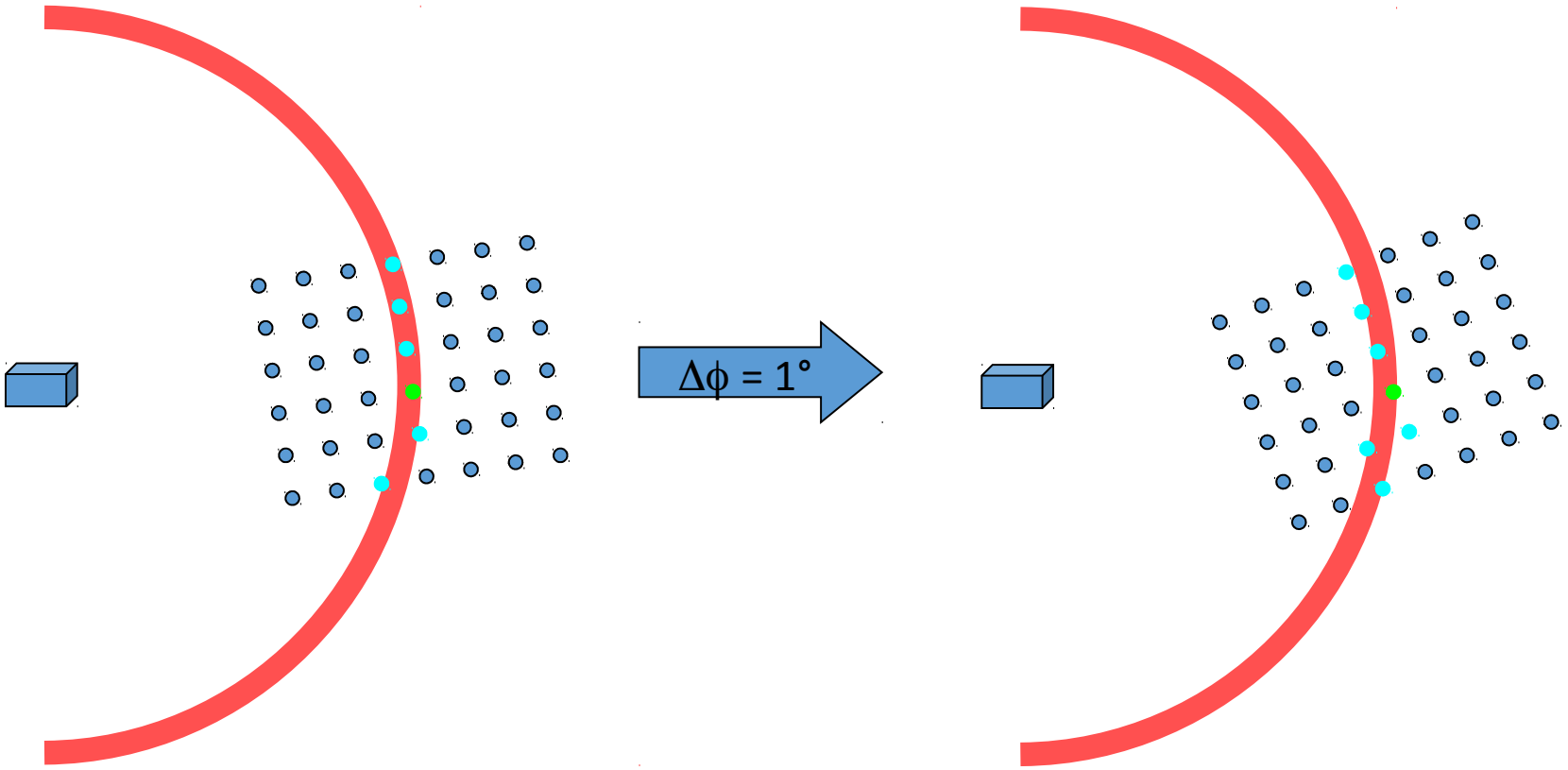
- perfect crystal: reciprocal lattice is built of points
- perfect beam (no wavelength dispersion, no divergence…)

- real life:
- mosaic crystal
- real beam (wavelength dispersion, divergence…)

Consequences on the Ewald's construction



Consequences on the Ewald's construction



- Width in φ of the spots and width of the Ewald sphere influence optimal Δφ

Direct detectors

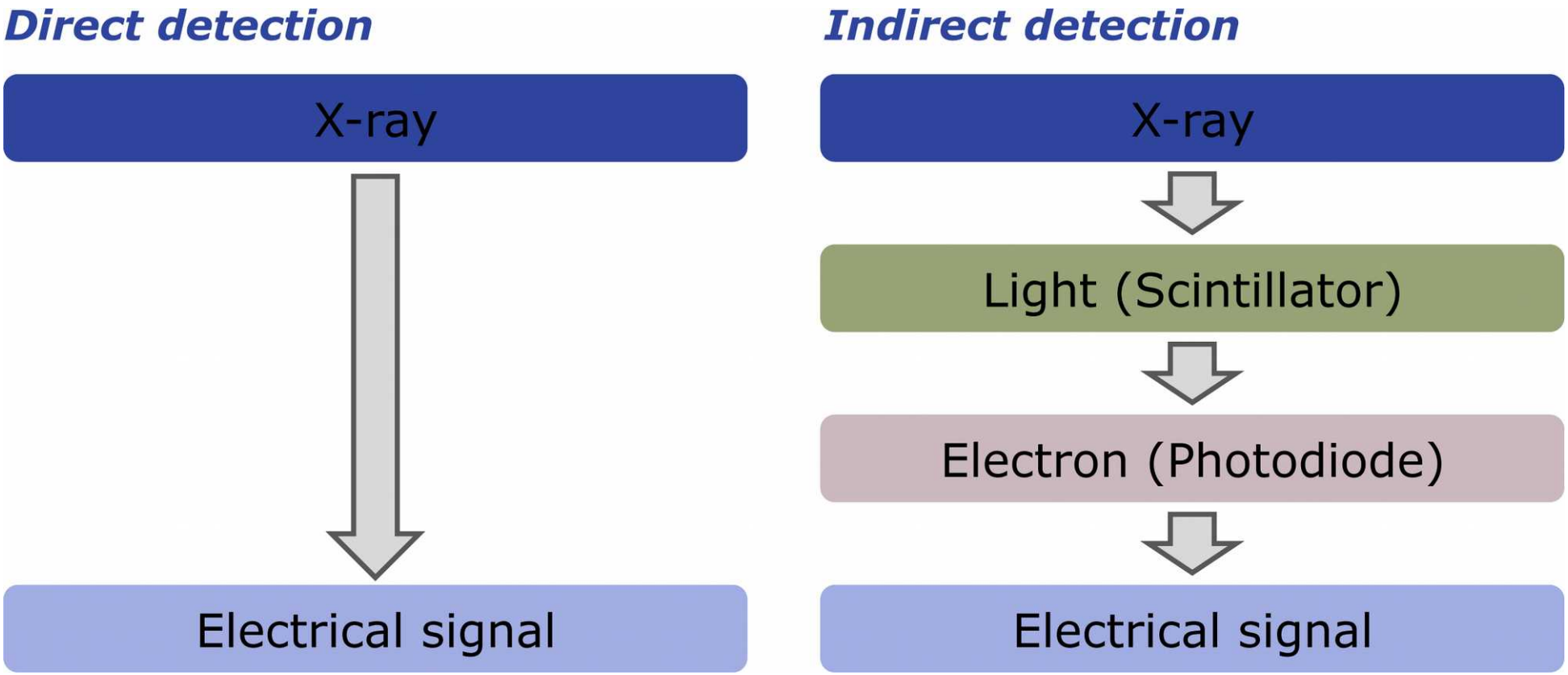


Illustration taken from Andreas Förster presentation, Oléron 2018

- No readout noise
- Lower backgroud

Direct detector

Noise-free detection

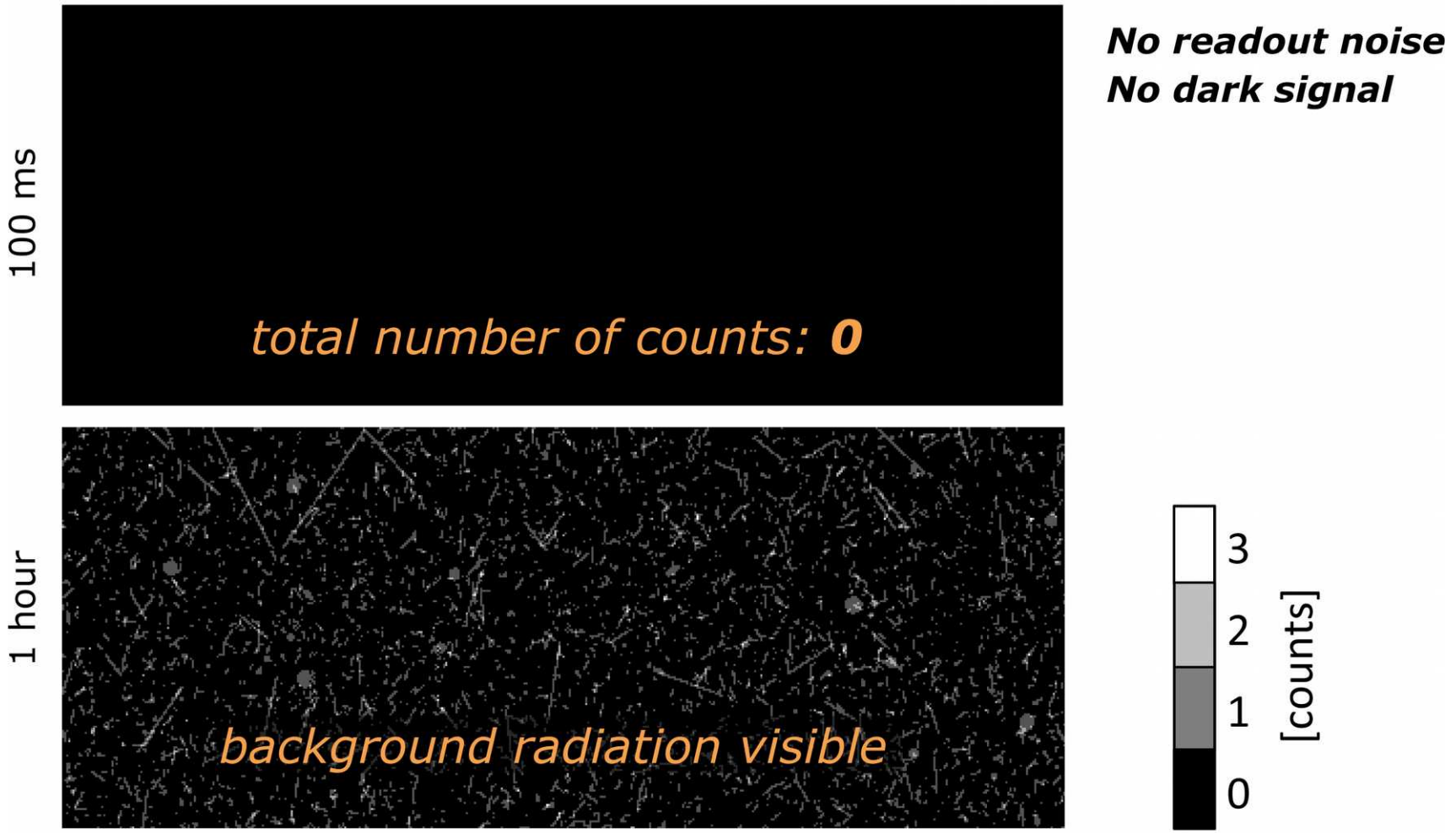
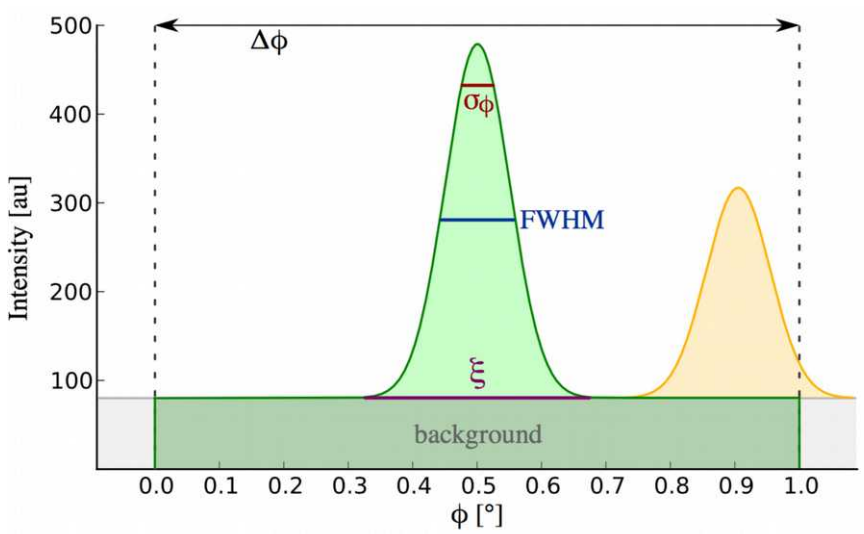
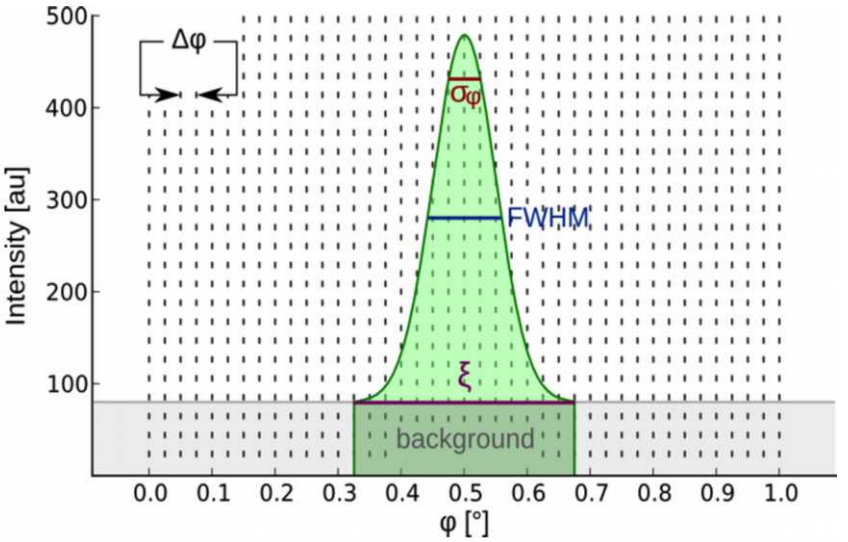


Illustration taken from Andreas Förster presentation, Oléron 2018

φ slicing ?





Wide φ-slicing

- Δφ > ξ
- Lots of background
- Few images

Fine φ-slicing

- Δφ ≪ ξ
- Minimal background
- Many images

Illustration taken from Andreas Förster presentation, Oléron 2018

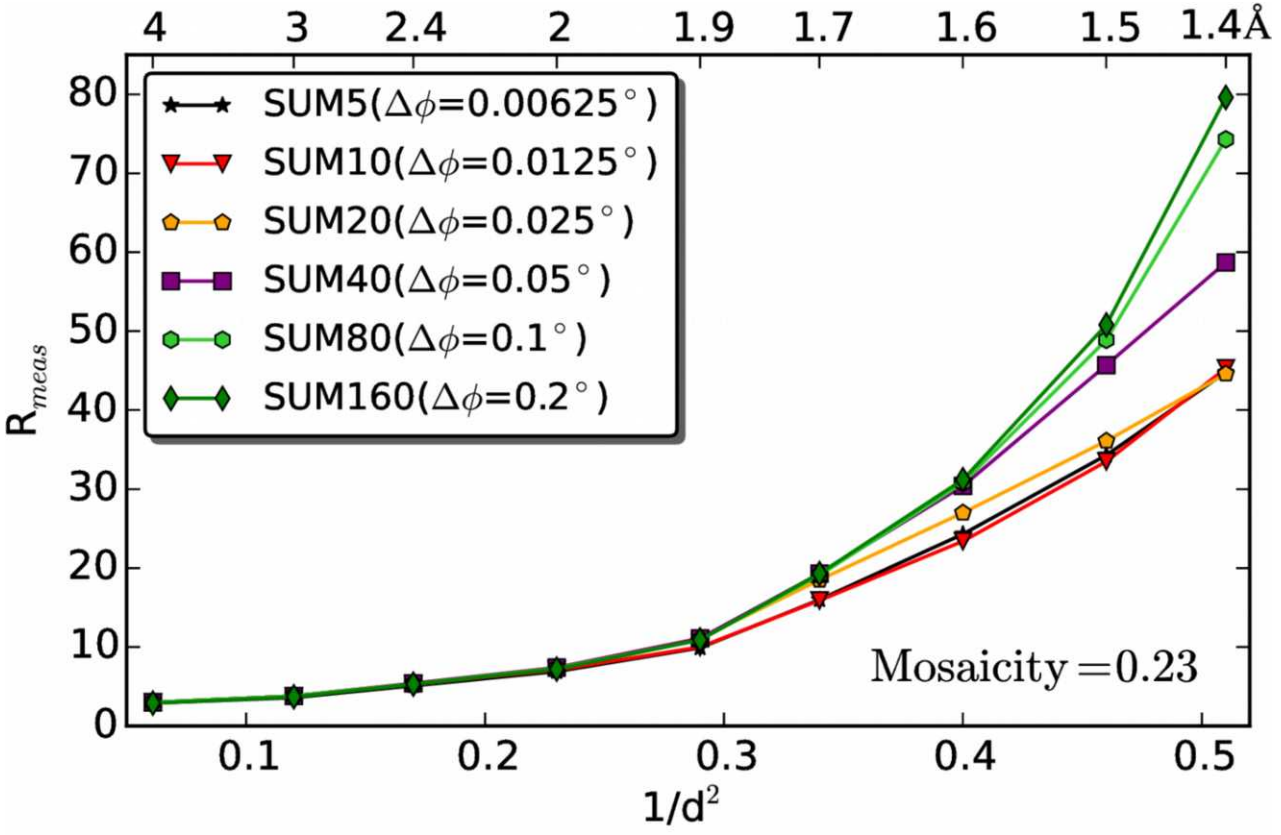
Mueller et al. (2012)

- Fine slicing is preferable (less background and less spot overlap)
- Signal should be strong enough for spot detection
- Dealing with large number of image (2000 – 8000) per data sets is no longer an issue

Fine slicing improves R_{meas}

- $\Delta\varphi \approx 1/10$ mosaicity optimal for
  - Overall statistics
  - High resolution shell statistics
  - Anomalous signal
  - Number of overlaps

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



Casañas et al. (2016)

Other parameters

- Time of exposure
  - Direct detectors have high dynamic and low noise
    - Low intensity measurements do not suffer from noise
    - Saturation occurs at very high count rate
  - Prefer shorter exposure
    - Less radiation dammage
    - Allow to collect more images and check when radiation damage becomes significant
- Crystal-detector distance
  - Spots should not overlap
    - Depend on the wavelength and the unit cell dimension
  - Depend on how your crystal diffracts
    - Which is the highest resolution you wish to collect

In conclusion

- Always collect the maximum amount of data
  - High redundancy lead to more accurate data
  - Accuracy of structure factor is crucial for anomalous signal
  - Some time the symmetry is lower than expected
  - Do not hesitate to go for 360° or more
    - No need to find the optimal starting point
- Prefer shorter exposure time
  - Avoid radiation damage before obtaining complete data
- Prefere fine φ slicing
  - Less background superposed to your signal
- Set the proper crystal-detector distance
  - Avoid spot overlap
  - Obtain the highest possible resolution

Thanks to Patrice Gouet and Laurent Maveyraud for sharing slides !