CCP4 School Chicago 2013

Data Processing with XDS

Tim Grüne
Georg-August-Universität
Institut für Strukturchemie

http://shelx.uni-ac.gwdg.de
tg@shelx.uni-ac.gwdg.de

X-ray Diffraction in a Nutshell

Model Building & Refinement — Ideal Crystal Data

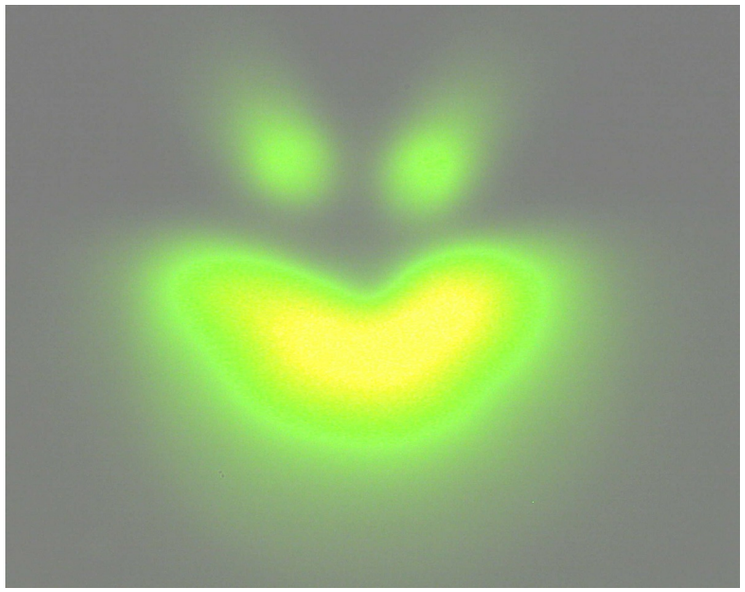
- Goal of crystal structure determination: Model of some molecule

- Refinement of model against $h, k, l, F(hkl), \sigma_F$ or $h, k, l, I(hkl), \sigma_I$

- Ideal: independent of machine, independent of sample

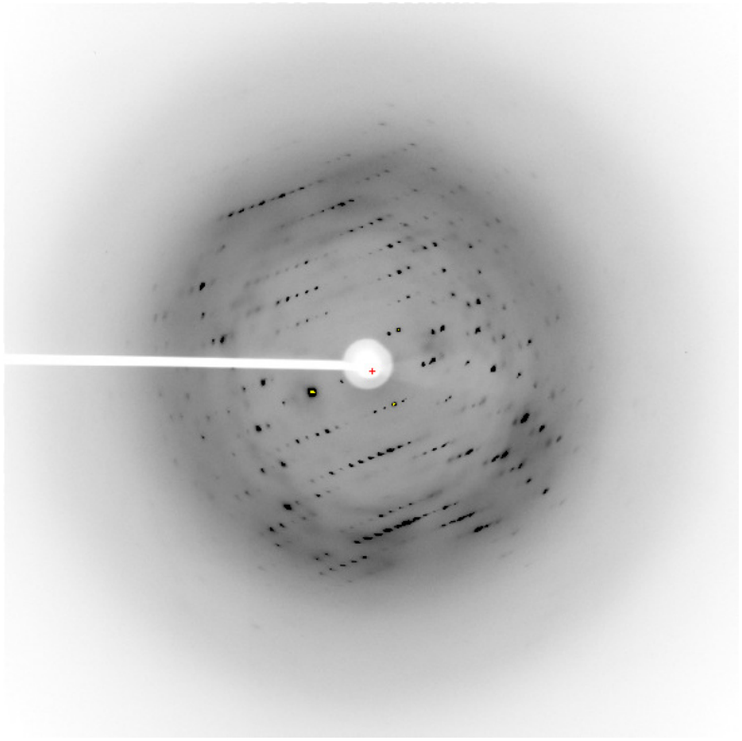
Experimental Errors

Misfocused beam



Courtesy N. Sanshvili, S. Corcoran; APS Chicago

Poorly diffracting crystal



Courtesy K. Pröpper, Uni Göttingen

Sources of Errors

- Beam/ beamline
- Crystal imperfections
- Overloads
- Detector Background Noise
- Settings inaccuracies
- . . .

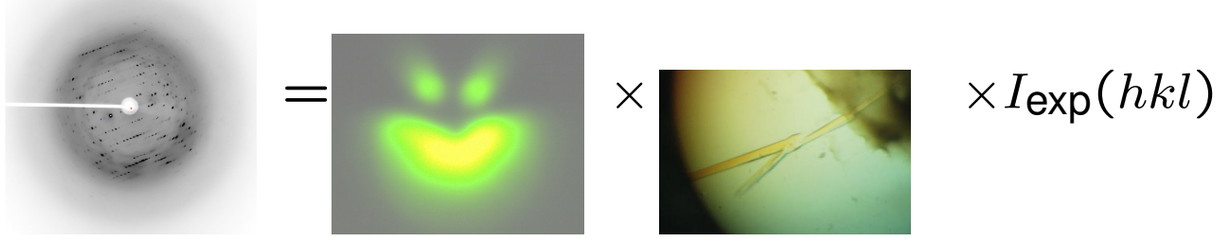
P.S.: Nukri's beam is NOT an experimental error, but simply the beam before it was focussed ;-)

Intensities and Amplitudes — Experiment *vs.* Theory

Intensities $I(hkl)$ are experimental quantities *measured* from the detector.

Amplitudes $|F(hkl)|$ are theoretical concepts necessary for refinement*

$$I_{\text{exp}}(hkl) = \frac{e^4}{{m_e}^2 c^4} \underbrace{\frac{\lambda^3 V_{\text{crystal}}}{V_{\text{u.c.}}^2} I_0 L P T E}_{\text{experimental factors}} |F_{\text{theor.}}(hkl)|^2 \qquad (1)$$

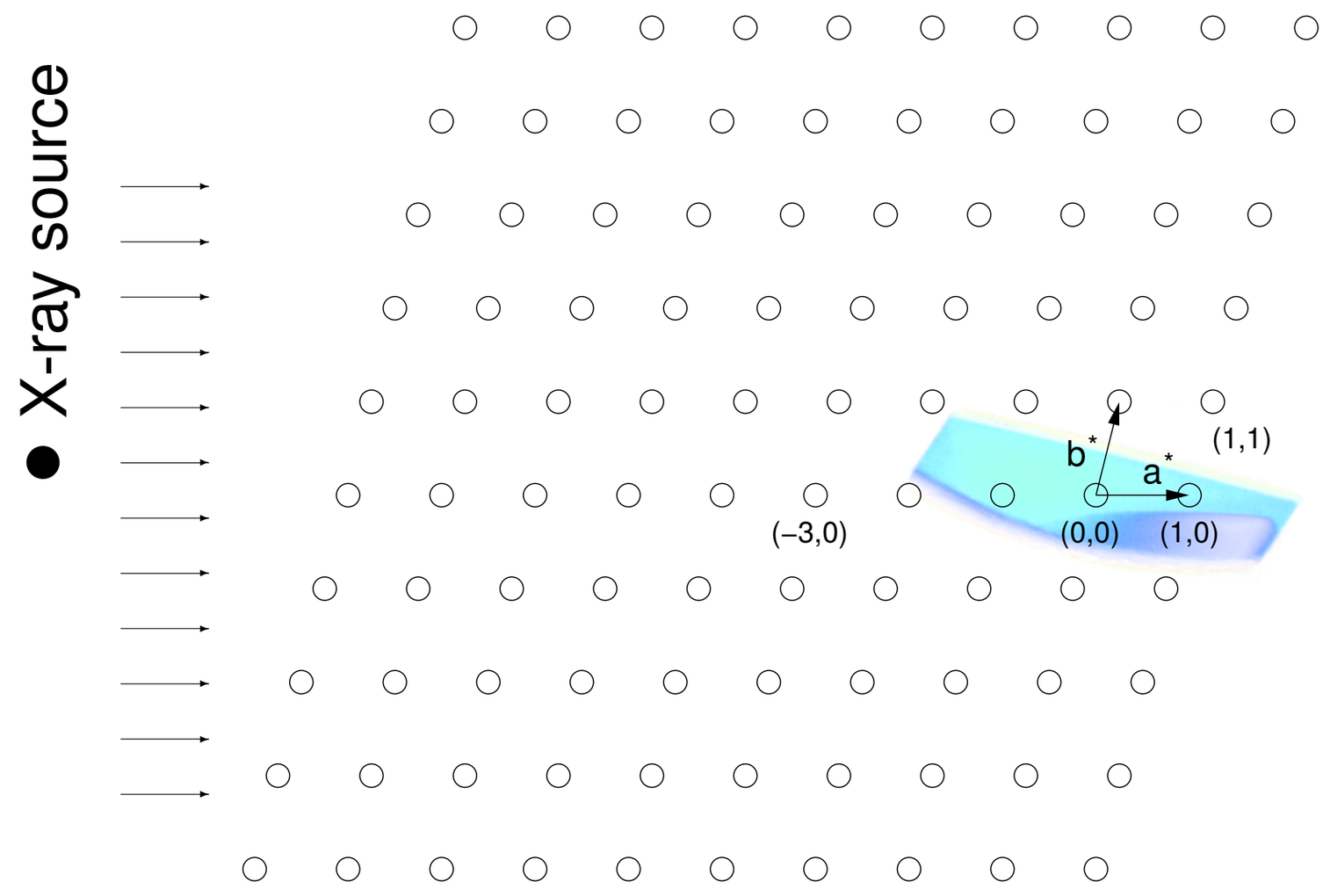


Data reduction $\approx$ Determination of intensities $I_{\text{exp}}(hkl)$ from frames

Scaling $\approx$ Determination of amplitudes $|F_{\text{theor.}}(hkl)|$ from $I_{\text{exp}}(hkl)$ and experimental settings

*Giacovazzo et el., "*Fundamentals of Crystallography*" (IUCr Texts on Crystallography), 1985, Kapitel "Diffraction Intensities"

Ewald Sphere Construction



Reciprocal Lattice:

$$\vec{a}^* = \frac{\vec{b}\times\vec{c}}{(\vec{a}\times\vec{b})\cdot\vec{c}}$$
$$\vec{b}^* = \frac{\vec{c}\times\vec{a}}{(\vec{a}\times\vec{b})\cdot\vec{c}}$$
$$\vec{c}^* = \frac{\vec{a}\times\vec{b}}{(\vec{a}\times\vec{b})\cdot\vec{c}}$$

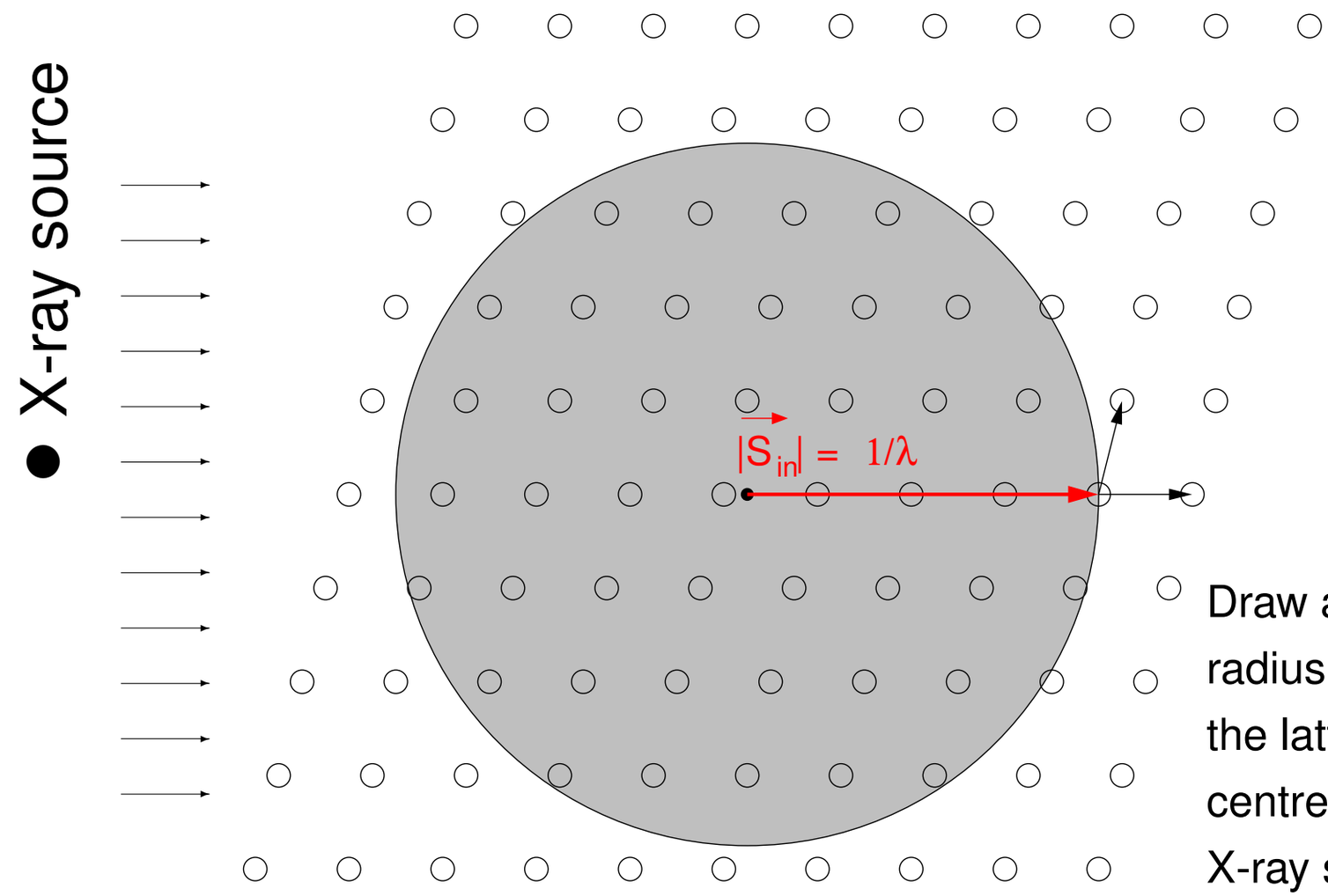
Lattice points at:

$h\vec{a}^* + k\vec{b}^*(+l\vec{c}^*)$

(hollow circles)

The crystal rotates about the origin of the reciprocal lattice.

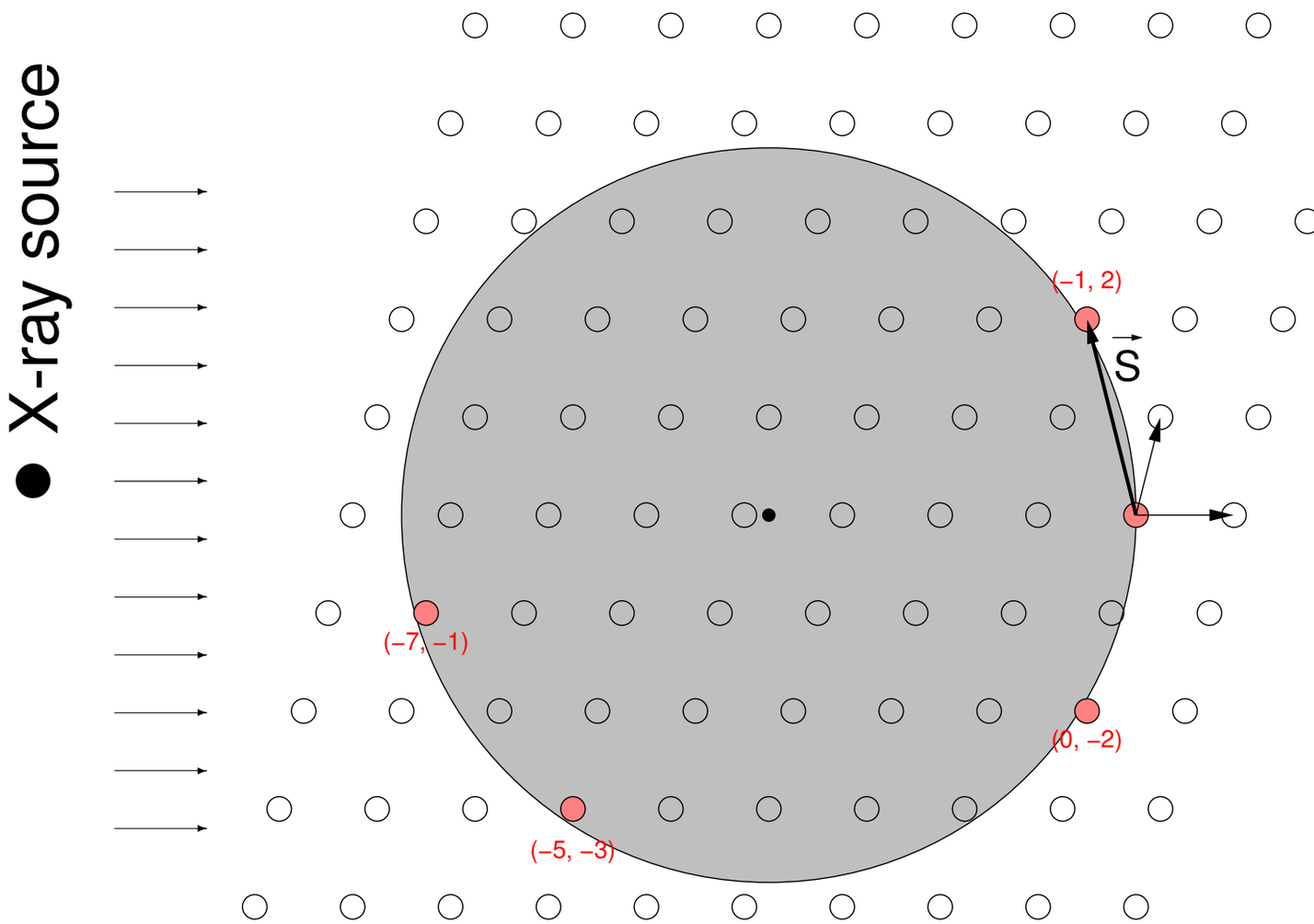
Ewald Sphere Construction



Draw a sphere with radius $1/\lambda$ that **touches** the lattice origin. The sphere centre lies aligned with the X-ray source.

This sphere is the **Ewald Sphere**.

Ewald Sphere Construction

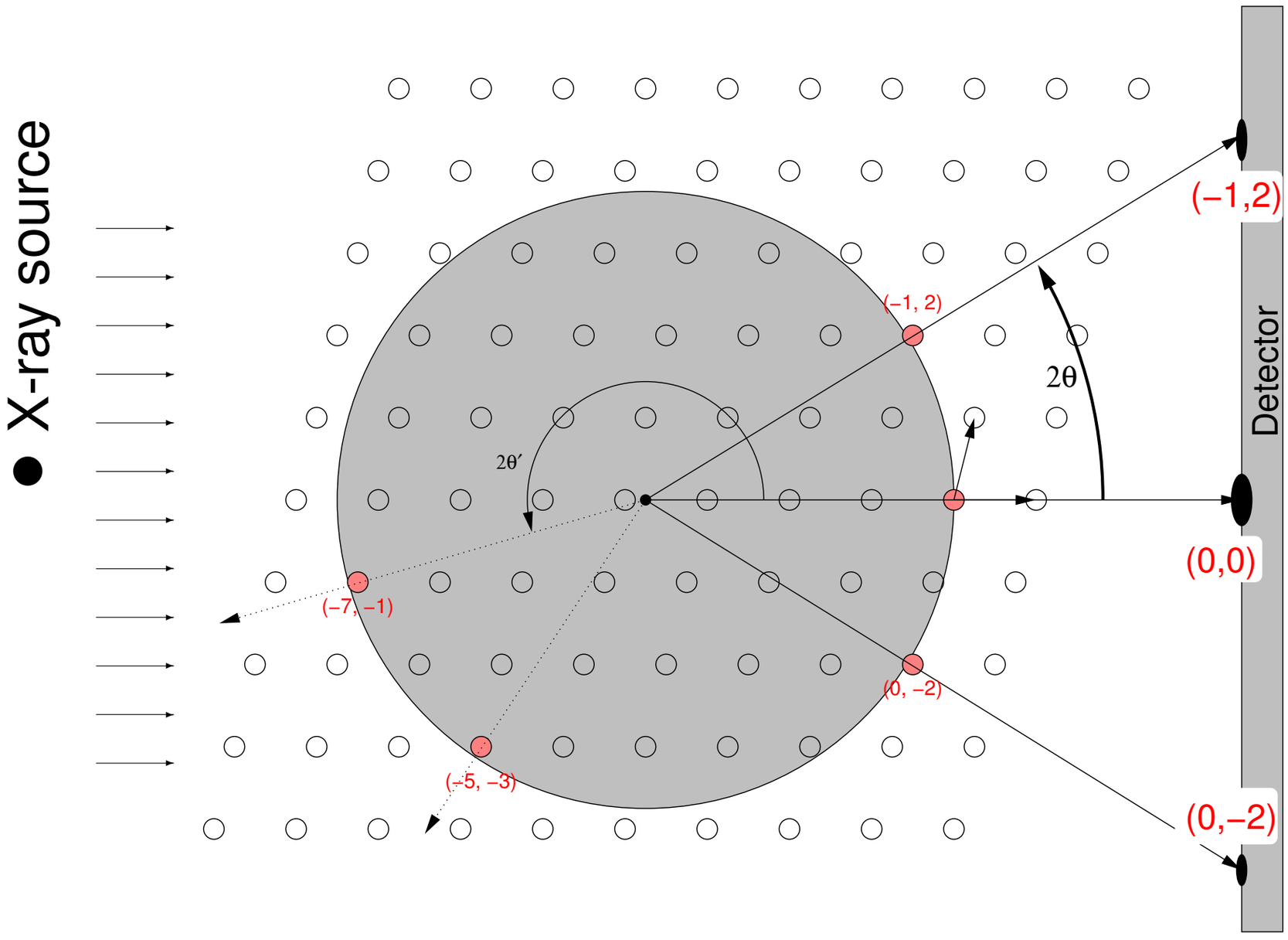


The **scattering vector** $\vec{S}$ points from the origin to the lattice point.

Exactly those lattice points on the surface of the Ewald sphere fulfil the Laue conditions.

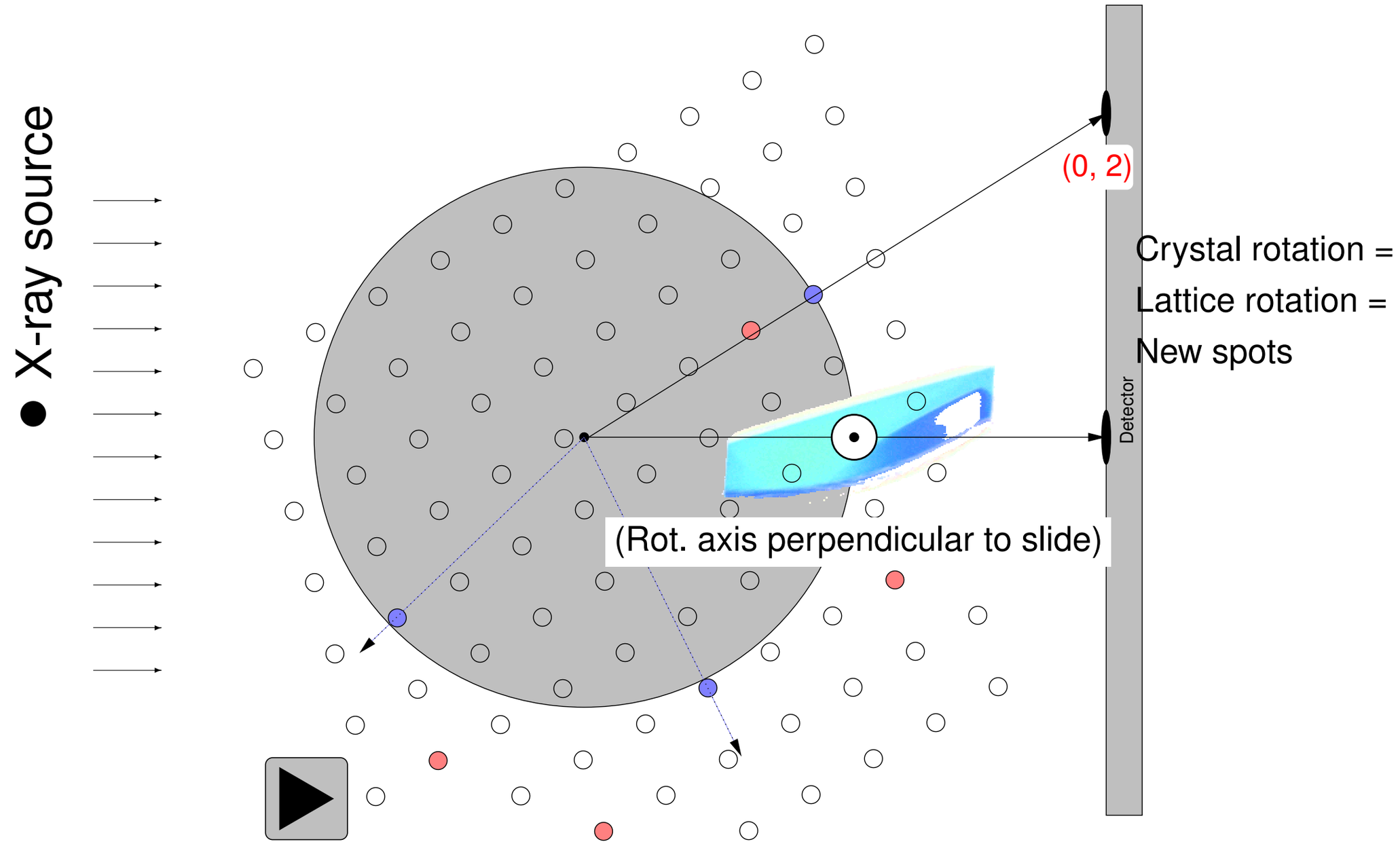
They are the **recordable** reflections.

Ewald Sphere Construction

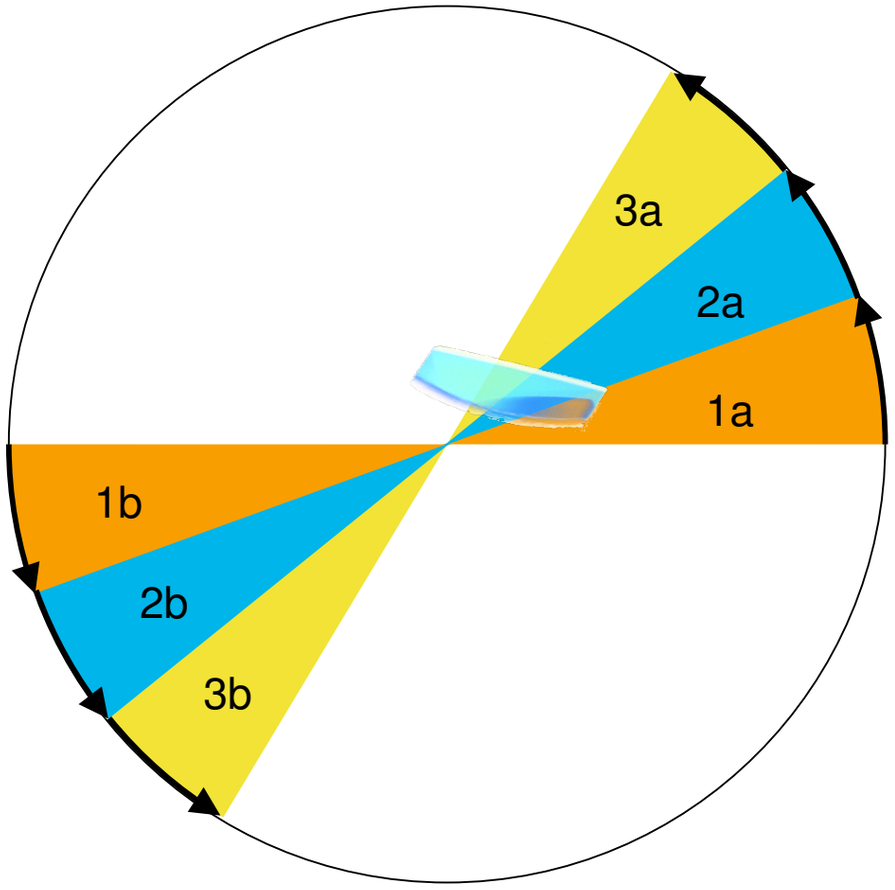


Some of these spots hit the detector.

Ewald Sphere Construction



Inverse Beam for Anomalous Phasing Strategies

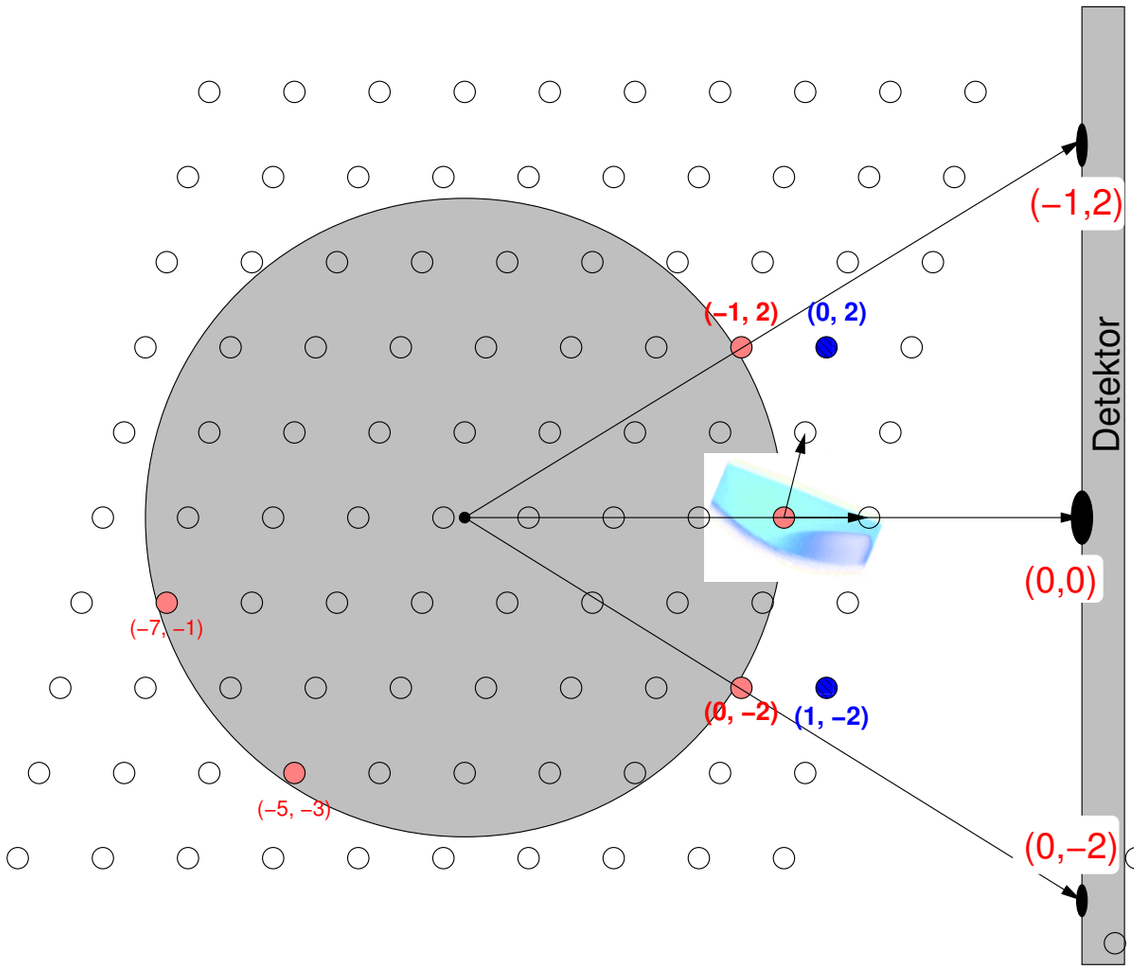


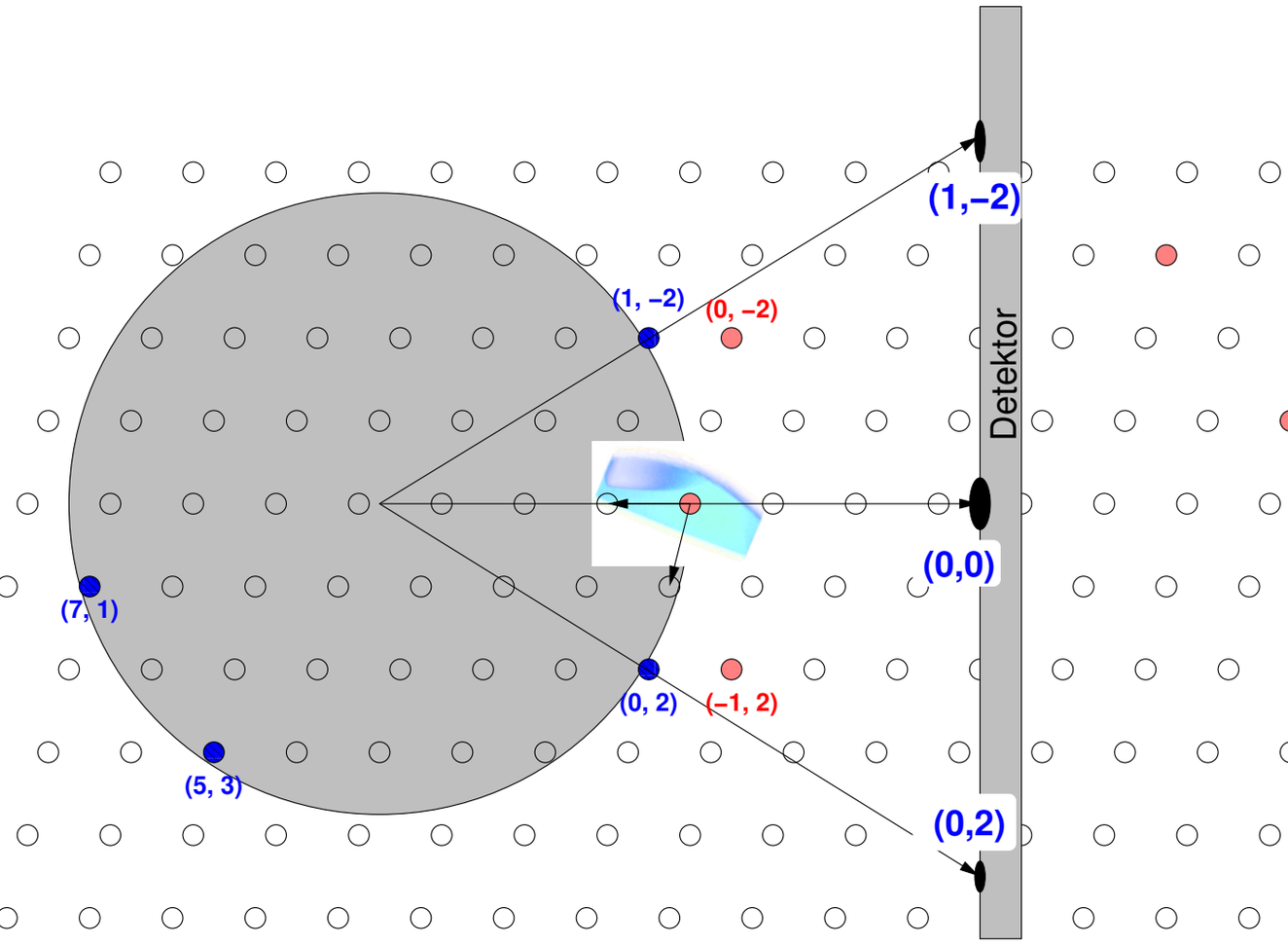
Wedges collected in the order

1. $0° - 5°$
2. $180° - 185°$
3. $5° - 10°$
4. $185° - 190°$
5. $10° - 15°$
6. $190° - 195°$
7. ...

Synchrotrons often offer 'inverse beam' collection strategies to accurately determine Bijvoet pairs for phasing based on anomalous dispersion.

Bijvoet Pairs and Inverse Beam





- Bijvoet are collected close together in space and time.

- Reduces random errors but may pronounce systematic errors.

XDS

The XDS Resources

originally written by W. Kabsch
co-authored by K. Diederichs

availability:

XDS http://xds.mpimf-heidelberg.mpg.de/ main program suite

Wiki and auxiliary programs http://strucbio.biologie.uni-konstanz.de/xdswiki/

GUIs

- xdsGUI http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/XDSgui
- XDSi http://cc.oulu.fi/~pkursula/xdsi.html
- Xdsapp http://www.helmholtz-berlin.de/forschung/funkma/soft-matter/forschung/bessy-mx/xdsapp

XDS Characteristics

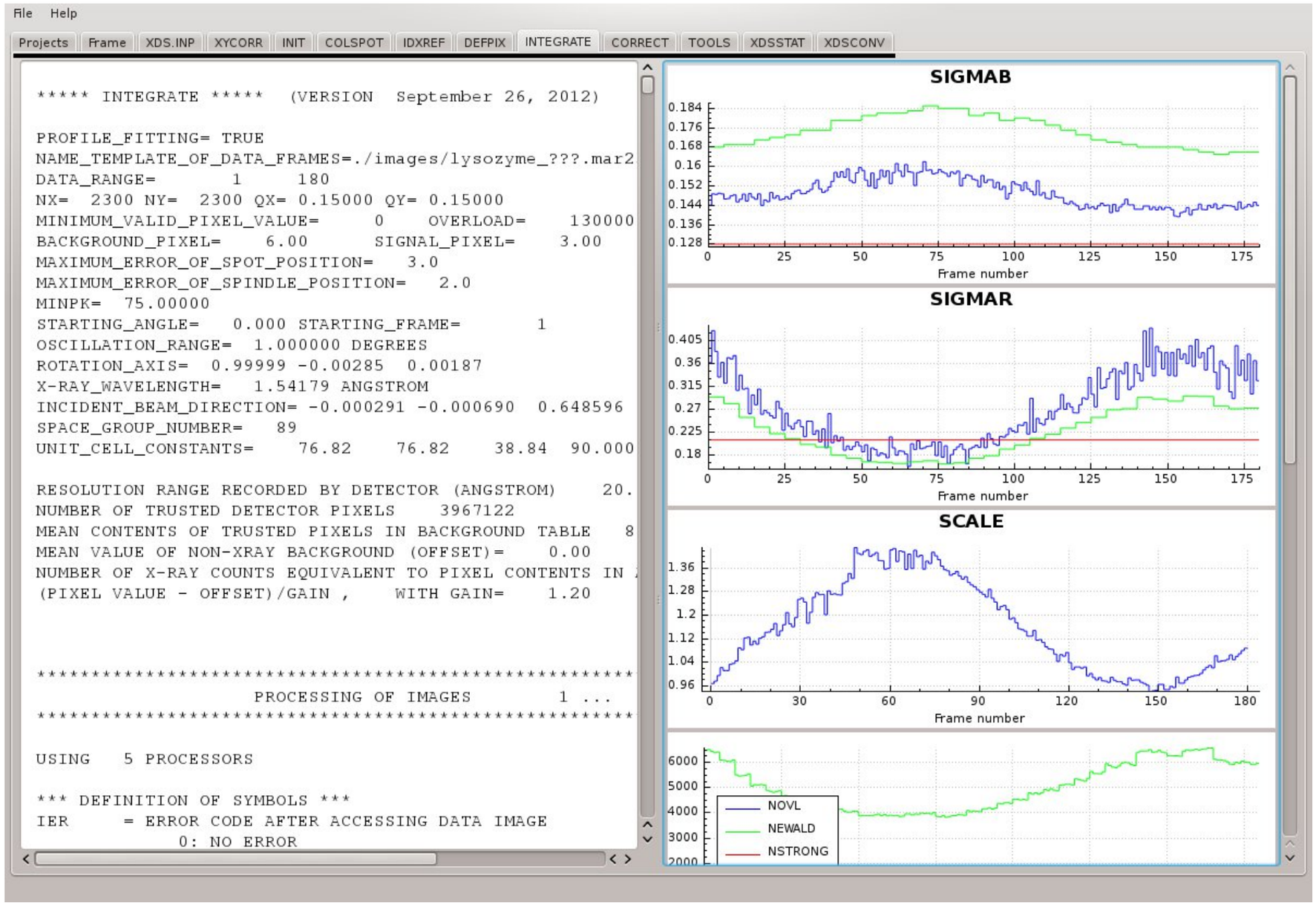
Some of the special features of XDS:

- Arbitrary Geometry (axes orientation, etc.)
- 3-dimensional spot integration
- Correction for Radiation Damage available
- Optimised for new Pilatus Detector
- Command-line program (pro's and con's. . .)
- Parallelised: Fast!
- Clear documentation

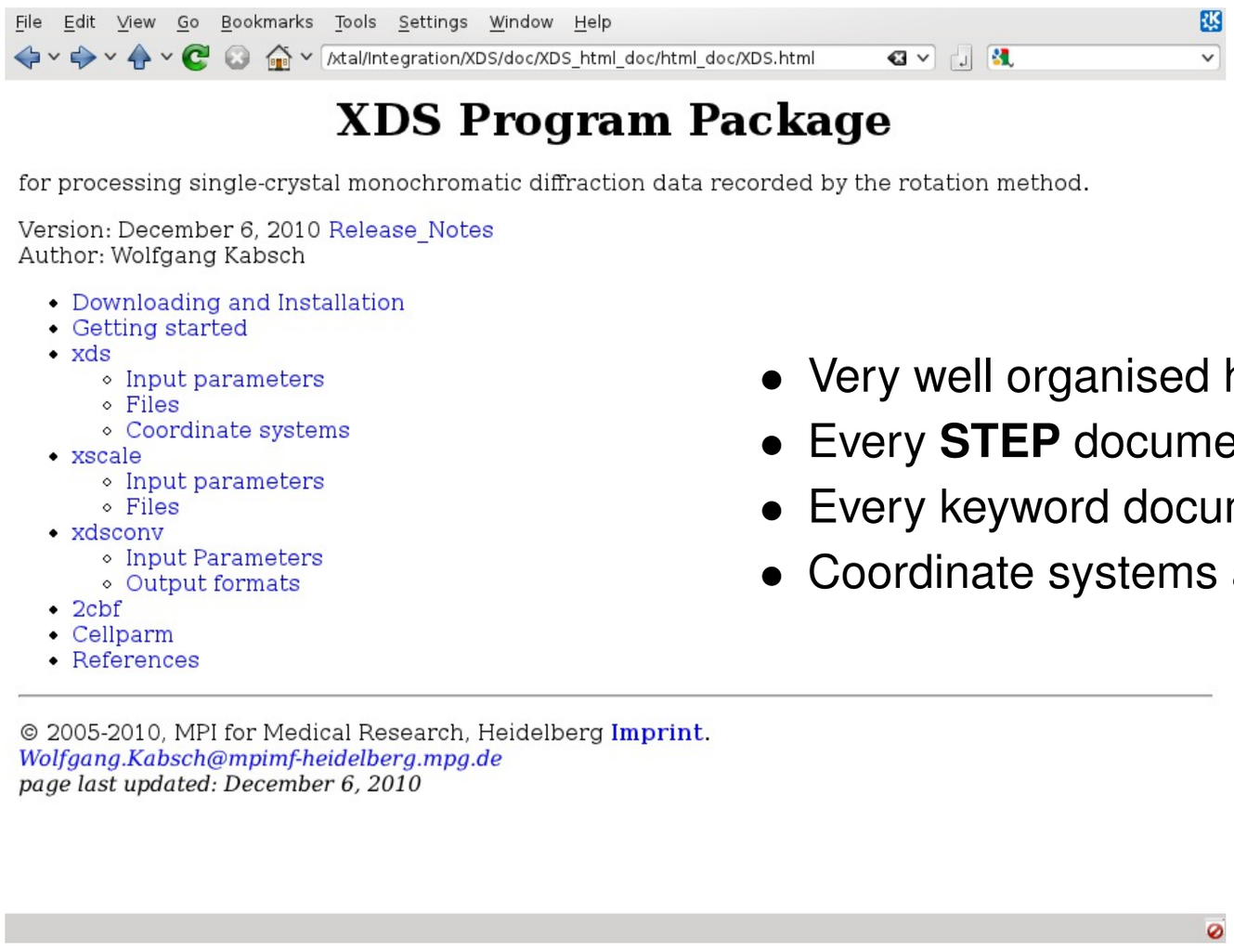
xdsGUI

File
Help
Projects
Frame
XDS.INP
XYCORR
INIT
COLSPOT
IDXREF
DEFPIX
INTEGRATE
CORRECT
TOOLS
XDSSTAT
XDSCONV
Project Folders
Choose or create new folder
Load recent project
/home/tg/uni/datasets/lehre/ggnb/xds

xdsGUI


File
Help
Projects
Frame
XDS.INP
XYCORR
INIT
COLSPOT
IDXREF
DEFPIX
INTEGRATE
CORRECT
TOOLS
XDSSTAT
XDSCONV
***** INTEGRATE ***** (VERSION September 26, 2012)
PROFILE_FITTING= TRUE
NAME_TEMPLATE_OF_DATA_FRAMES=./images/lysozyme_???.mar2
DATA_RANGE= 1 180
NX= 2300 NY= 2300 QX= 0.15000 QY= 0.15000
MINIMUM_VALID_PIXEL_VALUE= 0 OVERLOAD= 130000
BACKGROUND_PIXEL= 6.00 SIGNAL_PIXEL= 3.00
MAXIMUM_ERROR_OF_SPOT_POSITION= 3.0
MAXIMUM_ERROR_OF_SPINDLE_POSITION= 2.0
MINPK= 75.00000
STARTING_ANGLE= 0.000 STARTING_FRAME= 1
OSCILLATION_RANGE= 1.000000 DEGREES
ROTATION_AXIS= 0.99999 -0.00285 0.00187
X-RAY_WAVELENGTH= 1.54179 ANGSTROM
INCIDENT_BEAM_DIRECTION= -0.000291 -0.000690 0.648596
SPACE_GROUP_NUMBER= 89
UNIT_CELL_CONSTANTS= 76.82 76.82 38.84 90.000
RESOLUTION RANGE RECORDED BY DETECTOR (ANGSTROM) 20.
NUMBER OF TRUSTED DETECTOR PIXELS 3967122
MEAN CONTENTS OF TRUSTED PIXELS IN BACKGROUND TABLE 8
MEAN VALUE OF NON-XRAY BACKGROUND (OFFSET)= 0.00
NUMBER OF X-RAY COUNTS EQUIVALENT TO PIXEL CONTENTS IN
(PIXEL VALUE - OFFSET)/GAIN , WITH GAIN= 1.20
PROCESSING OF IMAGES 1 ...
USING 5 PROCESSORS
*** DEFINITION OF SYMBOLS ***
IER = ERROR CODE AFTER ACCESSING DATA IMAGE
0: NO ERROR
SIGMAB
Frame number
SIGMAR
Frame number
SCALE
Frame number
NOVL
NEWALD
NSTRONG

Documentation



- Very well organised html-documentation
- Every **STEP** documented
- Every keyword documented
- Coordinate systems are explained

XDS Programs

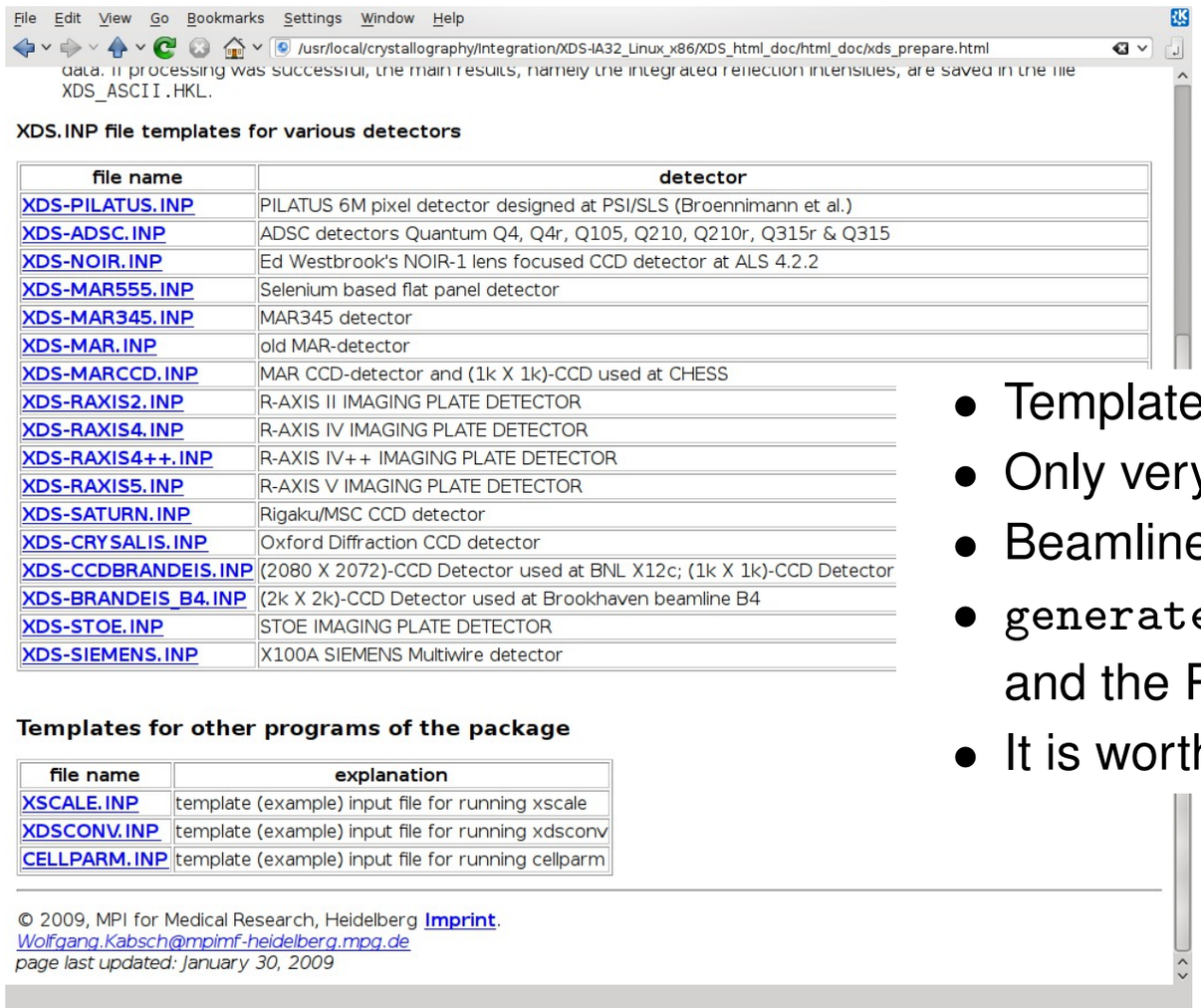
xds_par Main program for data integration

xscale scaling program for multiple datasets. **N.B.: Do not scale data that have already been scaled! Otherwise: corruption of standard deviations (σ_I). Can hamper phasing**

xdsconv Conversion to various other data formats

cellparam Weighted average unit cell parameters from several runs

Templates



- Templates of input scripts for all supported detector formats
- Only very few adjustments necessary to get started
- Beamlines often generate appropriate input scripts
- `generate_XDS.INP` from XDSwiki for MARCCD, ADSC, and the Pilatus 6M
- It is worth learning how to set it up from scratch!

http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Generate_XDS.INP

XDS.INP

XDS is controlled by one single input file: `XDS.INP`.

- Name cannot be changed
- **Each data set must be run in separate directory to avoid overwriting of files.**
- Contains about 100 Keywords [a] of the form

  `KEYWORD=VALUE`
- Only about 10 Keywords must be modified for most data sets (*e.g.* image names, detector distance, number of images, etc.)
- Most important one:

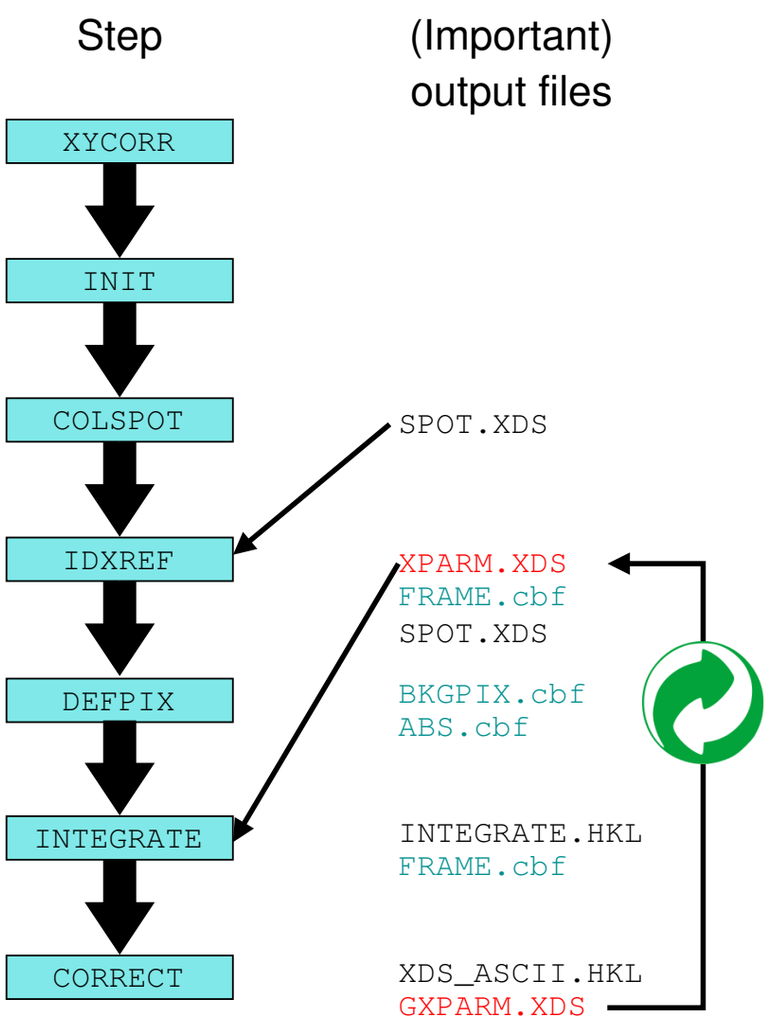
  `JOB= XYCORR INIT COLSPOT IDXREF DEFPIX XPLAN INTEGRATE CORRECT`

  Each name stands for one of the steps XDS carries out during data integration.

  (`XPLAN` is optional and corresponds to the BEST [2] or STRATEGY [1] output to report optimal data collection range(s))

[a] also called “cards” for historical reasons

Steps of Data Processing



red: Text files with parameters or data

cyan: control images (use xds-viewer)

black: data files for further processing

The Steps

`XYINIT` writes files for positional corrections of the detector plane. Most modern detectors provide already corrected images so that these to files are normally flat.

`INIT` determines initial detector background

`IDXREF` indexing: unit cell dimensions and crystal orientation

`DEFPIX` set active dectector area (exclude resolution cut-off, beam stop shadow, ...)

`XPLAN` (optional)generate “strategy” tables with data completeness

`INTEGRATE` extract reflection intensities from frames

`CORRECT` applies corrections (polarisation, Lorentz-correction, ...), scales reflections, reports data statistics

Program Flow

- Each step must be passed at least once - the subsequent steps depend on files produced by the previous steps.
- Each step creates a log-file (`XYINIT.LP`, `INIT.LP`,...).
- `IDXREF` is the main hurdle - once unit cell and crystal orientation are determined, integration usually runs smoothly.
- `CORRECT` summarises the quality of the data.
- Mostly `IDXREF.LP` and `CORRECT.LP` should be inspected.

IDXREF

- Indexing step: Find cell parameters and cell orientation.
- First refinement of experimental parameters (Detector distance, ...)
- Writes solution to `XPARM.XDS`

```
XPARM.XDS     VERSION March 30, 2013
    1         0.0000     0.1000  0.001263 -0.999998 -0.001591
    0.976260        0.000701       -0.000942        1.024317
 1     122.8045     140.1587     211.8464 109.309   99.583   90.012  <--- Cell
   -4.603119       87.195305      -86.352295
   -8.289775       98.209831       99.652267
  201.774994      -60.712154      -21.904343
      1       2463       2527     0.172000     0.172000
 1231.734253    1265.753540      620.964600                          <--- Distance
    1.000000        0.000000        0.000000
    0.000000        1.000000        0.000000
    0.000000        0.000000        1.000000
      1          1       2463          1       2527
 0.00     0.00     0.00   1.00000   0.00000   0.00000   0.00000   1.00000   0.00000
```

SPOT.XDS

- COLSPOT: Detector coordinates and Intensity of **strong** spots to be used for indexing:

```
X(pixel)   Y(pixel)    #image     counts
1056.11    1529.51      15.35      2544.
1895.52    1525.49       9.19      2481.
1913.43    1547.90       2.63      1999.
```

- IDXREF: Miller-Indices according to XPARM.XDS

```
   X(pixel)   Y(pixel)    #image     counts         H         K         L
   1056.11    1529.51      15.35      2544.        14        13        18
   1895.52    1525.49       9.19      2481.        -7       -13       -14
   1913.43    1547.90       2.63      1999.         0         0         0   <---
```

- 0 0 0: not indexed with current cell

"!!! ERROR !!! SOLUTION IS INACCURATE"

Correct indexing is crucial for data integration. If XDS cannot index more than 70 % of all reflections listed in `SPOT.XDS`, it stops with the above error message.

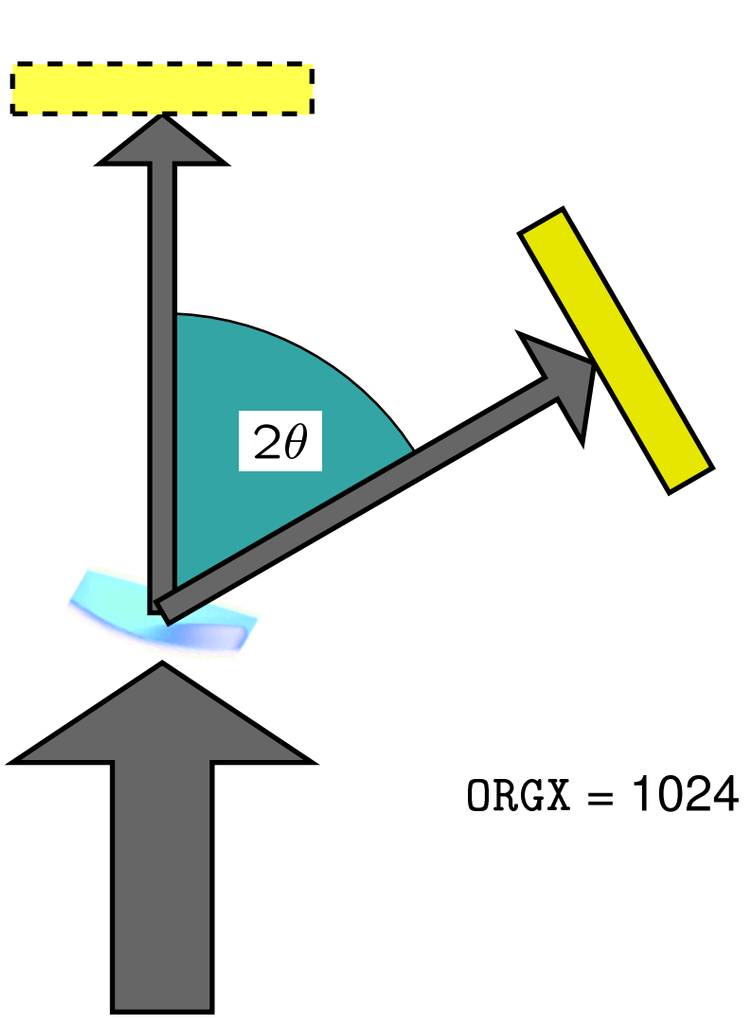
Most common reasons:

1. Wrong `ORGX`, `ORGY`
2. Poor data quality

ORGX, ORGY

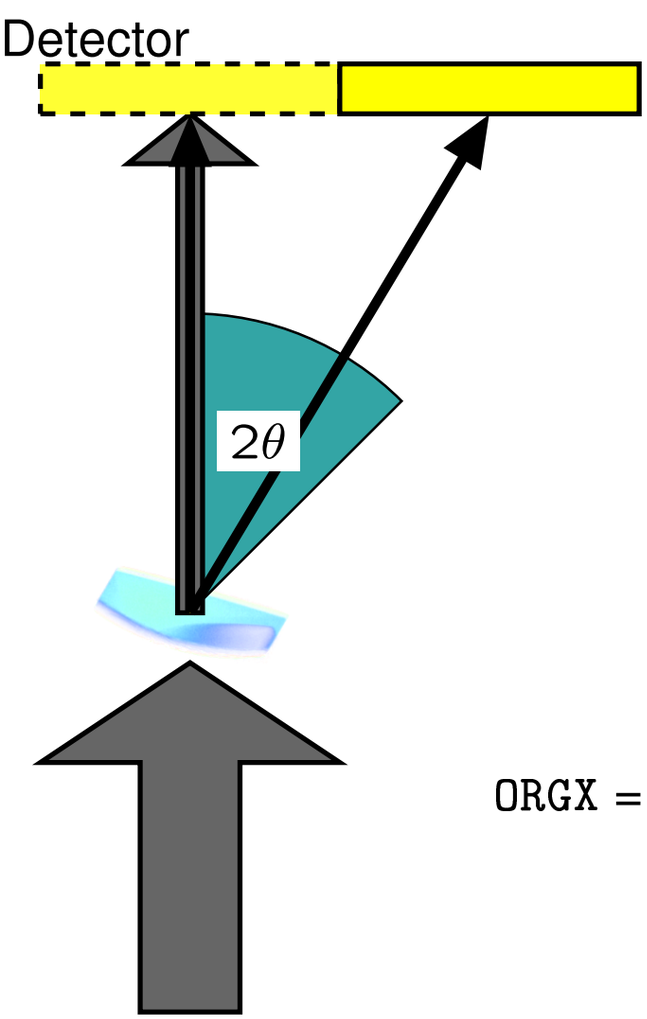
- many programs take the direct beam position

- XDS takes the detector origin `ORGX`, `ORGY`: point normal to **detector plane** through crystal



`ORGX` = 1024, `ORGY` = 1024

2048x2048 Detector



`ORGX` = -2048, `ORGY` = 1024

Wrong ORGX, ORGY

The "detector origin" ORGX, ORGY in XDS.INP is provided in pixels. Typically it is in the centre of the detector, so a good start is ORGX=NX/2, ORGY=NY/2, *i.e.* half the detector dimensions.

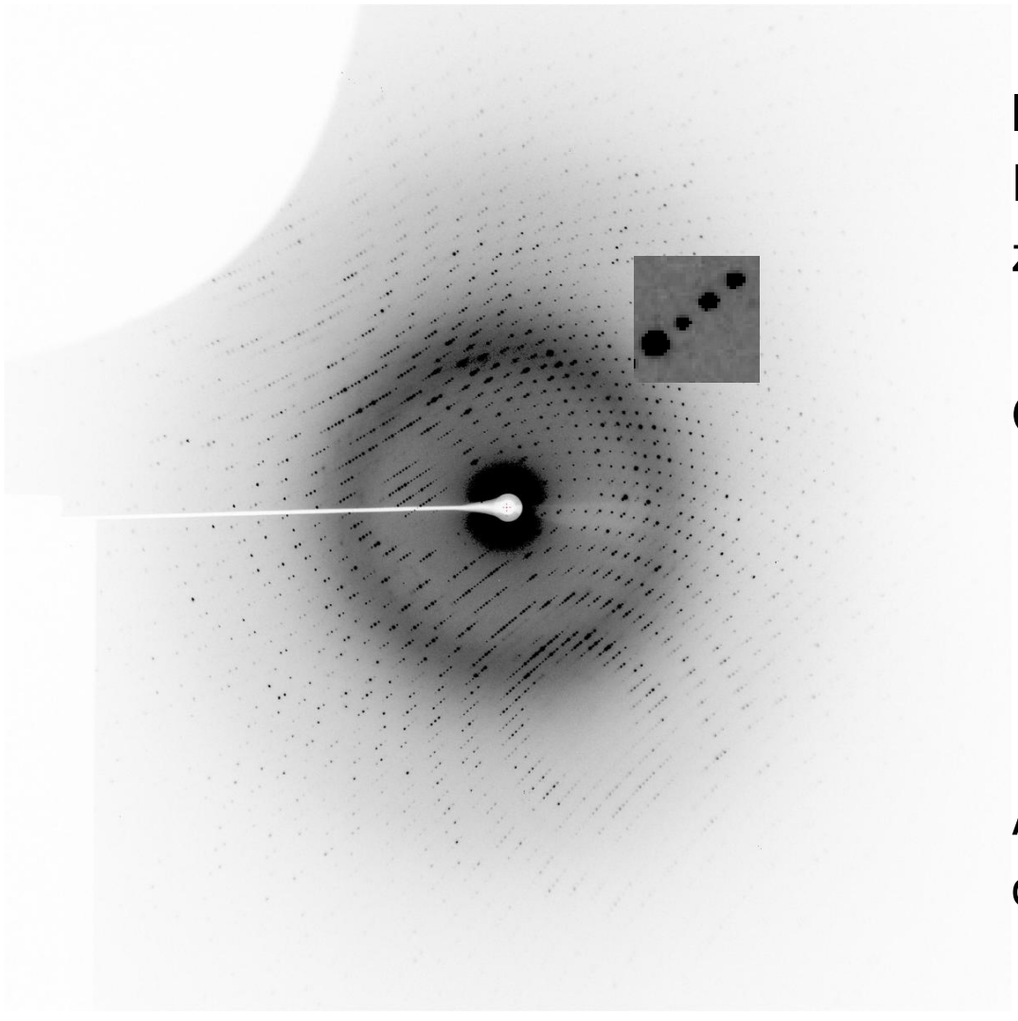


Illustration of the sensitivity to ORGX, ORGY
Diffraction image from Thaumatin with a medium size cell 58 Å, 58 Å, 150 Å, 90°, 90°, 90°.

Coordinates of the four magnified spots:

Coordinates		Difference		Difference(mm)	
1505	1800				
1514	1793	9	7	0.66	0.51
1522	1788	8	6	0.59	0.44
1530	1780	8	7	0.59	0.51

An offset of the origin of about 0.5 mm means a shift of 1 in Miller-Indices

mis-indexing

Complaints about ORGX, ORGY

Indicators of a wrongly set origin:

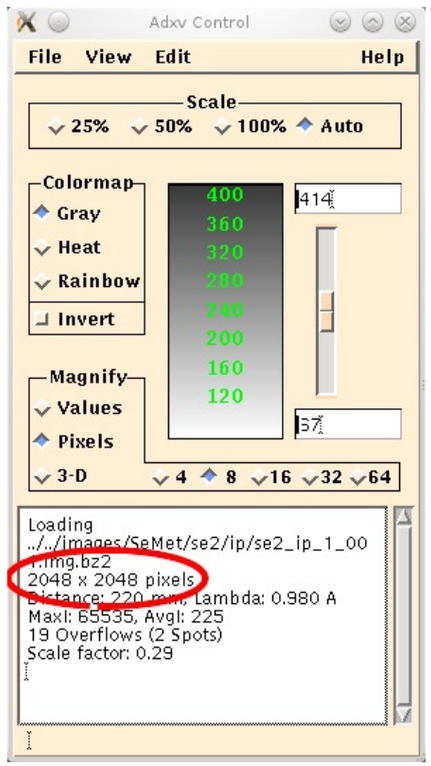
1. refined detector distance far from real value
2. ORGX, ORGY drift far away from initial value

```
IDXREF.LP:
 ***** DIFFRACTION PARAMETERS USED AT START OF INTEGRATION *****
                      [...]
 DETECTOR ORIGIN (PIXELS) AT         1580.99   1506.71   (XDS.INP: 1450 1450)
 CRYSTAL TO DETECTOR DISTANCE (mm)       211.85          (XDS.INP: 200)
```

The detector distance reported during data collection (image header) should be correct within fractions of mm at a proper beamline and not drift by 11 mm.

Correcting ORGX, ORGY

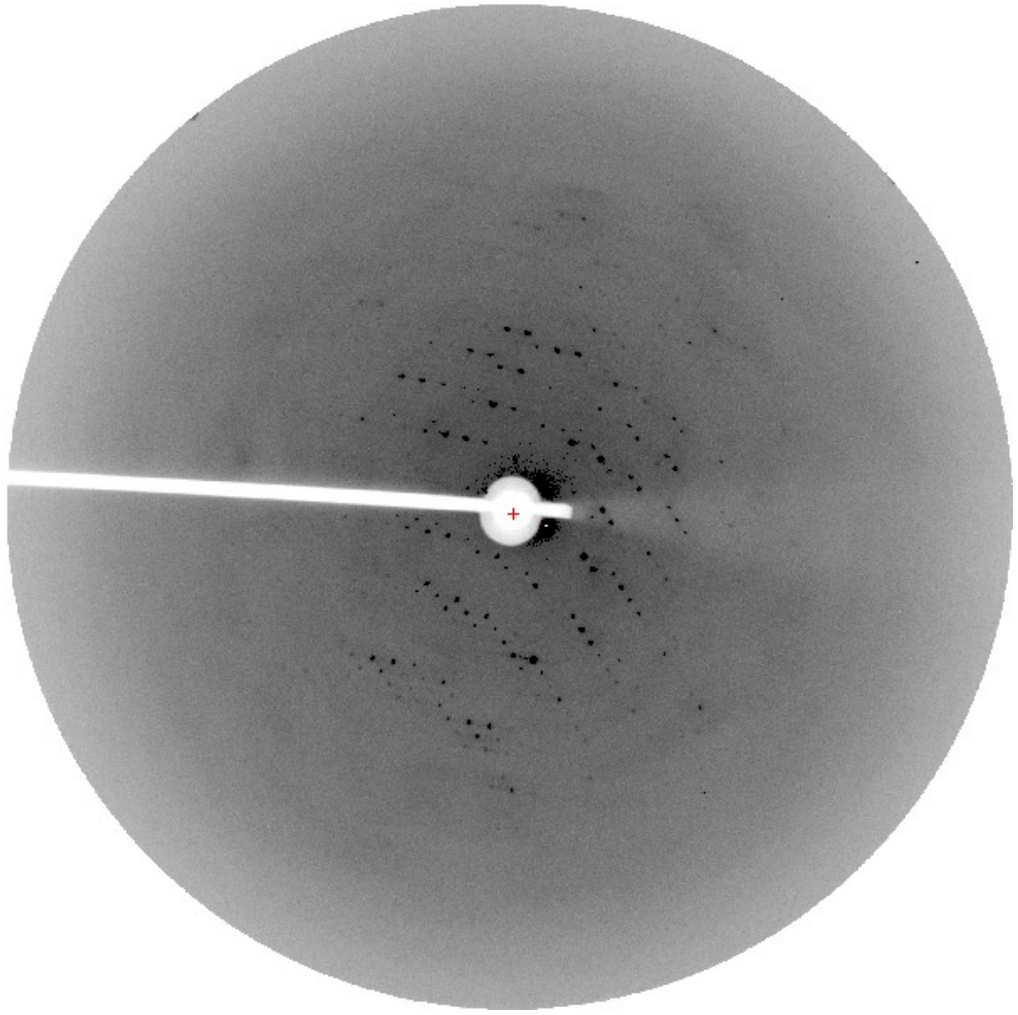
1. Start with ORGX=NX/2, ORGY=NY/2 (use adxv [3] to read out header information). Often XDS refines these values reasonably well.



Correcting ORGX, ORGY

2. Check the image for the beam centre. At $\theta = 0^\circ$ this correpsonds well to ORGX, ORGY. Ice rings are particularly useful, they are centred at about the beam centre (use circle fitting option in mosflm).

In case you **expect** poorly diffracting crystals (3 Å and worse), consider recording one image with an empty loop filled with a drop of water to get an image with ice rings!!!



Carry on regardless

The XDS error message `SOLUTION IS INACCURATE` does not necessarily mean that something is seriously wrong. The step IDXREF still refines the experimental parameters and writes them into `XPARM.XDS`. This is all we need to continue.

Check `IDXREF.LP` for

- Number of indexed reflections

```
***** INDEXING OF OBSERVED SPOTS IN SPACE GROUP #    1 *****
 1909 OUT OF    2506 SPOTS INDEXED.
    0 REJECTED REFLECTIONS (REASON: OVERLAP)
  597 REJECTED REFLECTIONS (REASON: TOO FAR FROM IDEAL POSITION)
```

  With $\approx$ 30 % of indexed reflections the parameters in `XPARM.XDS` are probably good enough to converged to the correct values during `INTEGRATE` and `CORRECT`.
- refined detector distance and detector origin do not shift too much

Set `JOB = DEFPIX INTEGRATE CORRECT` and integrate your data.

If something **did** seriously go wrong, the file XPARM.XDS would not been written.

DEFPIX: active detector mask

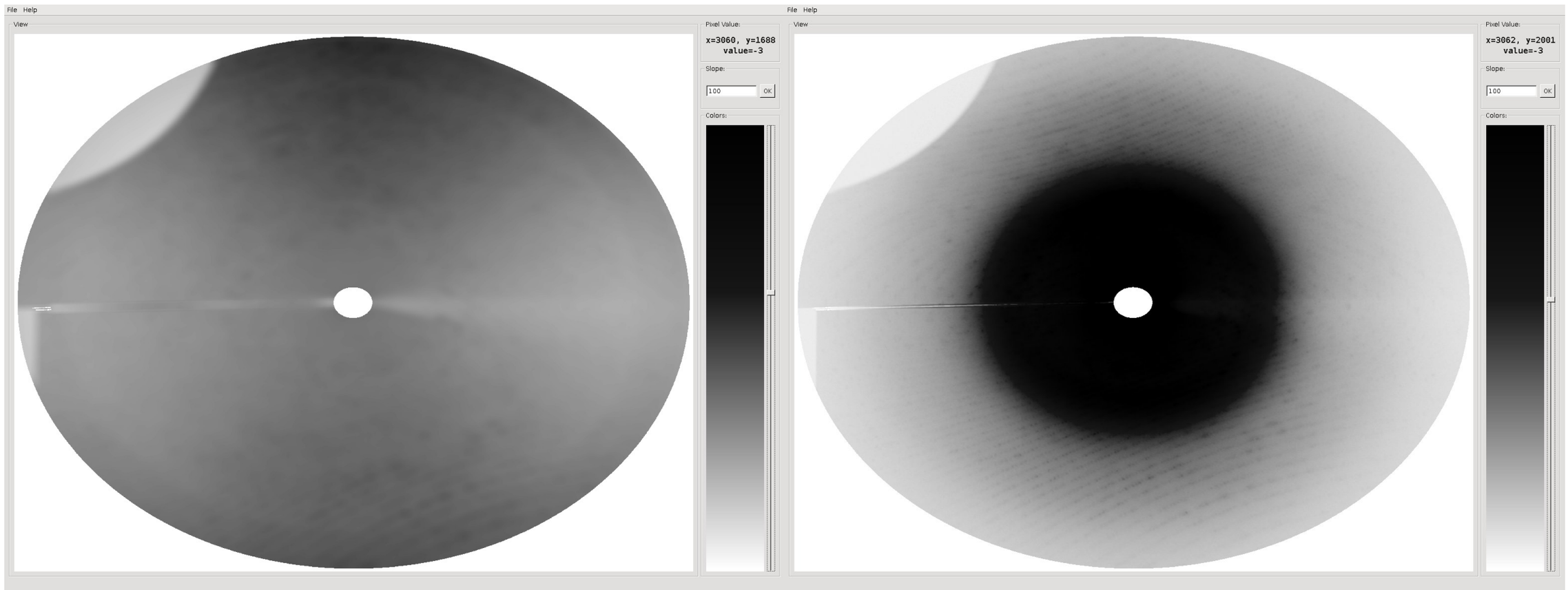
DEFPIX sets the area of the detector which is taken into account during integration. It takes into account:

1. INCLUDE_RESOLUTION_RANGE default: 20 Å to detector edge.
2. VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS exclude shadowed regions, *e.g.* beamstop, cryo stream nozzle
3. UNTRUSTED_RECTANGLE exclude gaps of CCD chips *e.g.* Pilatus detector
4. EXCLUDE_RESOLUTION_RANGE exclude ice rings from data integration

VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS

`ABS.cbf`

`BKGPIX.cbf`, all included

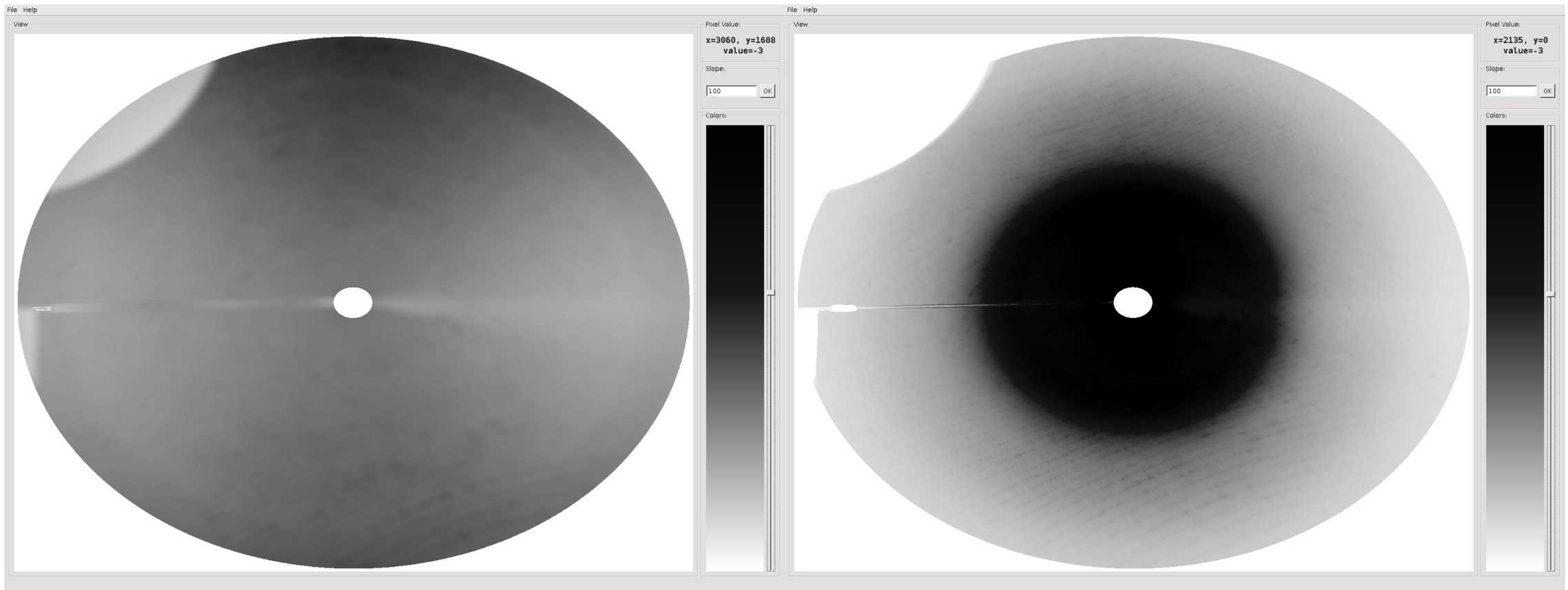



```
VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= 1000 30000
```

VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS

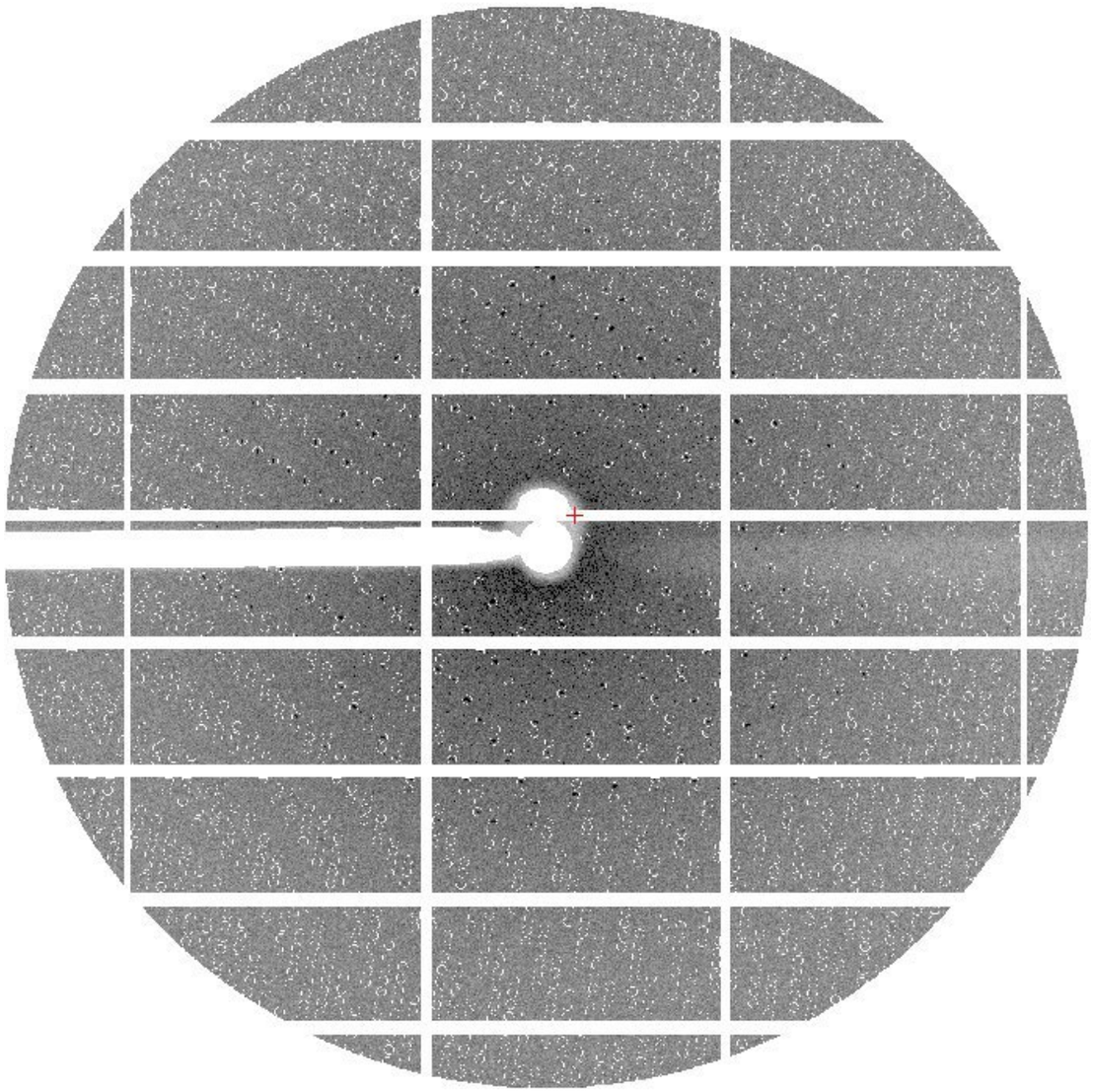
`ABS.cbf`

`BKGPIX.cbf`, shadows removed




```
VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= 6100 30000
```

Pilatus and other gaps: UNTRUSTED_RECTANGLE



- Beamline provides rectangles for their detector
- `UNTRUSTED_RECTANGLE= 981 989 0 2528`
  $X_{\text{from}} \rightarrow X_{\text{to}}$ and $Y_{\text{from}} \rightarrow Y_{\text{to}}$: thin vertical strip
- Phasing: be generous to avoid shading!

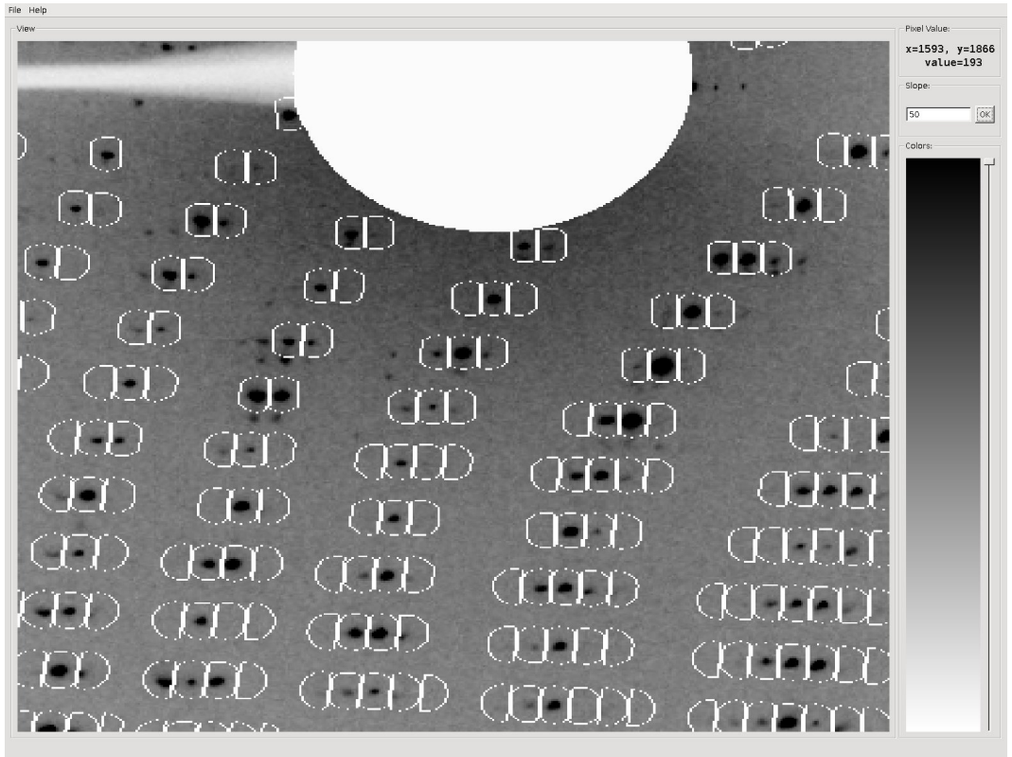
INTEGRATE and CORRECT: Intensity determination and fine tuning

INTEGRATE determine spot intensities based on parameters (saved in XPARM.XDS)

CORRECT experimental corrections (*e.g. Lorentz- and Polarisation-correction*), refine parameters and determine space group (saved to GXPARM.XDS)

CORRECT write FRAME.cbf:

- predicted spot positions encircled
- check correctness of predictions



Recycling

- Parameters (in `GXPARM.XDS`) depend on measured intensities
- Intensities (including corrections) depend on Parameters

  ⇒ rename `GXPARM.XDS` to `XPARM.XDS` and rerun XDS (`JOB = DEFPIX INTEGRATE CORRECT`) to improve results.

This way one should also set the correct high- and low- resolution cut-offs

Resolution Cut-Off

The default resolution range in XDS is 20 Å to the detector edge

```
INCLUDE_RESOLUTION_RANGE=20.0 0.0
```

- Medium to low resolution data: increase 20 Å to 30 Å or even 50 Å (check `BKGPIX.cbf`)
- After **second** round of integration: determine high-resolution cut-off. My favourite: xprep - higher number of resolution shells than *e.g.* listed in `CORRECT.LP`

Why after **second** round?

- Correct space group rather than $P1$

$\Rightarrow$ more symmetry related reflections

$\Rightarrow$ more reliable data statistics, especially I/σ_I

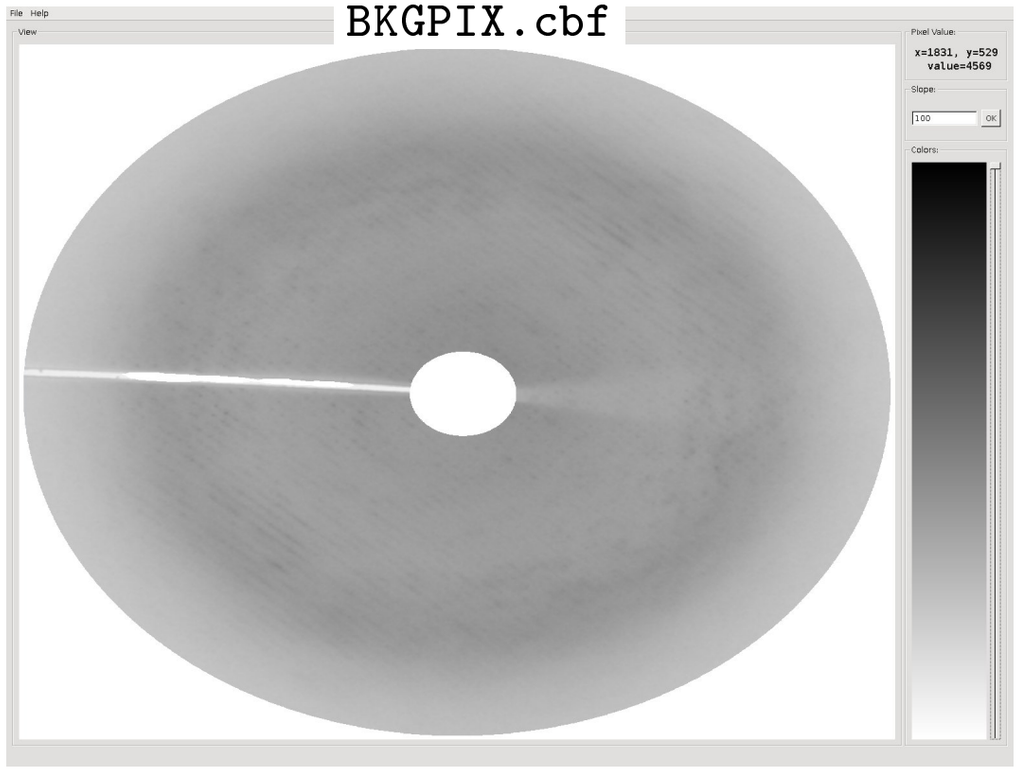
High Resolution Cut-Off

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

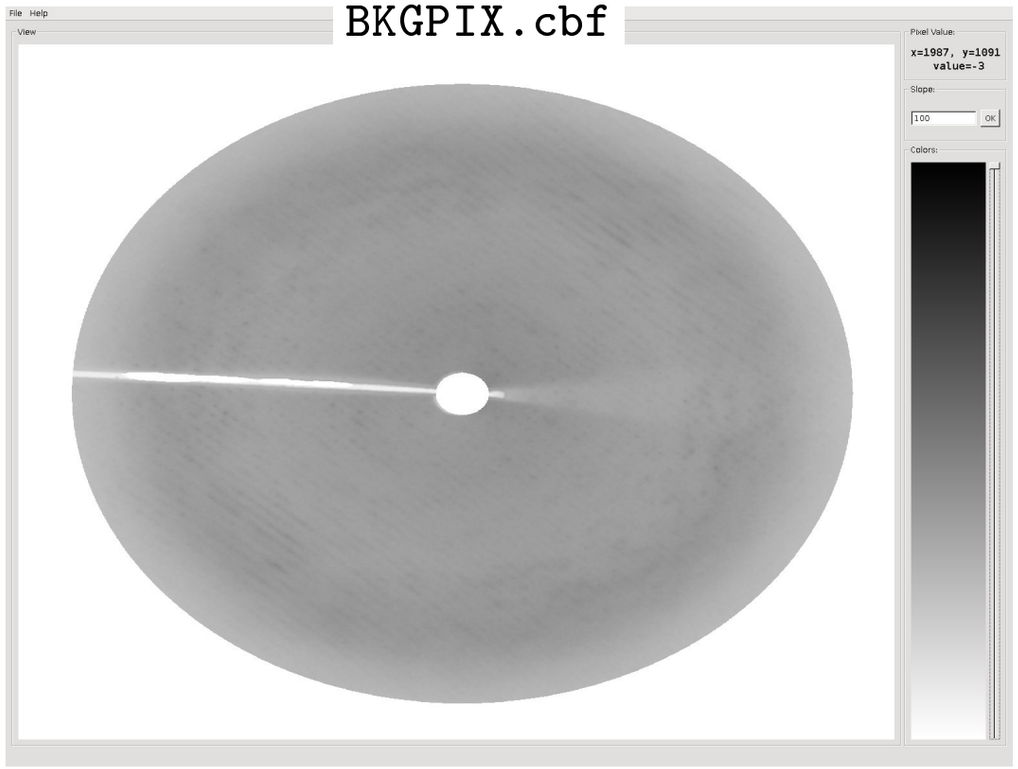
	RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
	13.63	9181	1459	1511	96.6%	2.5%	2.9%	9180	55.25	2.7%	100.0*	10	0.901	1206
	9.64	17093	2592	2672	97.0%	2.7%	3.1%	17091	53.02	3.0%	99.9*	4	0.858	2332
	7.87	23196	3381	3433	98.5%	3.2%	3.4%	23195	46.27	3.5%	99.9*	-1	0.848	3115
	6.81	27489	3965	4049	97.9%	4.5%	4.4%	27483	35.88	4.8%	99.9*	-2	0.851	3705
	6.09	31672	4498	4565	98.5%	5.8%	5.7%	31671	28.72	6.3%	99.8*	-3	0.834	4258
	5.56	35456	4969	5040	98.6%	6.8%	6.6%	35451	25.07	7.3%	99.7*	-5	0.816	4752
	5.15	37852	5387	5465	98.6%	7.2%	7.1%	37847	23.06	7.8%	99.7*	-3	0.818	5168
	4.82	40555	5829	5894	98.9%	7.2%	7.1%	40546	22.85	7.8%	99.7*	-5	0.795	5549
	4.54	41126	6125	6213	98.6%	7.8%	7.7%	41105	20.88	8.4%	99.7*	-2	0.807	5788
	4.31	45588	6568	6647	98.8%	9.5%	9.5%	45578	17.93	10.3%	99.5*	-2	0.804	6302
	4.11	47633	6842	6900	99.2%	12.4%	12.4%	47622	14.05	13.4%	99.2*	-1	0.805	6574
	3.93	50103	7192	7284	98.7%	16.8%	16.8%	50091	10.89	18.1%	98.7*	-1	0.800	6940
	3.78	51899	7414	7519	98.6%	23.0%	22.9%	51889	8.35	24.9%	97.7*	-1	0.796	7172
	3.64	54983	7766	7842	99.0%	31.6%	31.8%	54975	6.30	34.1%	96.0*	-1	0.781	7533
	3.52	56772	8033	8114	99.0%	43.3%	43.5%	56759	4.72	46.7%	93.2*	0	0.774	7785
	3.41	58909	8313	8412	98.8%	59.1%	59.8%	58897	3.45	63.8%	87.7*	-1	0.748	8065
-->	3.31	60300	8607	8702	98.9%	83.0%	84.5%	60285	2.42	89.7%	79.6*	0	0.728	8345
									^^^^^^					
	3.21	62086	8804	8904	98.9%	116.7%	119.8%	62073	1.72	126.0%	64.9*	1	0.705	8563
	3.13	64144	9057	9159	98.9%	156.4%	160.9%	64134	1.28	168.9%	49.9*	-2	0.666	8820
	3.05	61751	9009	9395	95.9%	206.2%	212.5%	61650	0.92	223.0%	34.5*	-2	0.643	8497
	total	877788	125810	127720	98.5%	10.5%	10.7%	877522	13.37	11.4%	99.9*	-1	0.769	120469

Logfile `XSCALE.LP`

Low Resolution Cut-Off



Default resolution range (20 Å - edge)
- includes noise at edge
- loose low resolution reflections



Adjusted resolution range (40 Å - 2.85 Å)
- $I/\sigma_I \approx 2$ in outer shell
- low resolution reflection important for “shape” of molecule (MR, refinement)

Format Conversion

The final integrated data are written to the file `XDS_ASCII.HKL`.

Data are best scaled with `XSCALE`.

1. for phasing with shelx c/d/e

2. for refinement (`mtz`-file)

HKL → hkl

How to:

1. Phasing with shelx c/d/e requires no conversion

2. `shelxc` and `xprep` both read XDS and XSCALE output (HKL–) files

3. format detected automatically

HKL $\rightarrow$ mtz

How to (suggested by Clemens Vonrhein, Globalphasing:

1. Conversion to `mtz`, `hklref`: indexing correction if necessary (e.g. space group $C2$)

```
pointless -copy xdsin xscale.HKL hklref ${freeRefMtz} hklout pointless.mtz
```

2. Merging of equivalent indices:

```
aimless hklin pointless.mtz hklout aimless.mtz << eoaimless
ONLYMERGE
eoaimless
```

3. Intensities to amplitudes ($I \rightarrow |F|$):

```
truncate hklin aimless.mtz hklout truncate.mtz << eotruncate | tee truncate.log
eotruncate
```

HKL → mtz (continued)

4. Copying R_{free} flags if necessary

```
cad hklin1 truncate.mtz hklin2 ${freeRefMtz} hklout cad.mtz << eocad | tee cad.log
LABIN FILE 1 ALL
LABIN FILE 2 E1=FreeR_flag
eocad
```

5. Correction for reported resolution, e.g. by `mtzdmp`

```
sftools << eosftools | tee sftools.log
read cad.mtz
SELECT ONLY COLUMN IMEAN PRESENT
write sftools.mtz
eosftools
```

6. Extension of R_{free} reflections to full resolution

```
freerflag hklin sftools.mtz hklout freerflag.mtz << eofreerflag | tee freerflag.log
COMPLETE FREE=FreeR_flag
eofreerflag
```

References

1. Leslie, A.G.W., *Recent changes to the MOSFLM package for processing film and image plate data*, Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography (1992), No. 26
2. A.N. Popov and Bourenkov , *Choice of data-collection parameters based on statistic modeling* Acta Crystallogr. (2003). **D**59, 1145-1153
3. Andrew Arvai, http://www.scripps.edu/~arvai/adxv.html