Data Processing with XDS

Kay Diederichs

Universität Konstanz

Outline

1) **Introduction to XDS**

2) **Data processing** is the crucial link between your experiment and your structure

   * garbage in – garbage out!
   * for important data, do not rely on automatic data processing alone
   * get to know your data!
   * manual checks are easily performed with XDSGUI
   * try to optimize data processing – this converts noise to signal, and may enable structure solution, and/or improve refinement

3) **Space group determination** may be anything between trivial and difficult.

   * not a good idea to just go to lower symmetry (P1) if structure solution "does not work"
   * it helps to understand a few basic relationships, and to know where to look things up
   * the space group is a hypothesis until proven by satisfactory structure solution/ refinement and electron density maps. So it is not determined by data processing alone!

The *XDS* program suite

- Original author: Wolfgang Kabsch (Max-Planck-Institute Heidelberg)
- Since ~1986
  I joined 2007

The *XDS*+ programs

- *XDS* - the main program: indexing, integrating, scaling, statistics
- *XSCALE*: - scale several *XDS* intensity data sets together; zero-dose extrapolation; statistics
- *XDSCONV*: convert to other programs' formats (e.g. MTZ file)

Programs independent from the *XDS* distribution:

- *XDS-Viewer* - inspect diagnostic images written by *XDS*, or (single) data frames (open source: sourceforge.net).Instead, *adxv* may be used
- *XDSSTAT* - additional statistics (not part of main distribution; download and use: see XDSwiki)
- *XDSGUI* – graphical user interface (open source: sourceforge.net) for *XDS* and *SHELX C/D/E and ARCIMBOLDO* (since latest version)
- *XDSCC12* – which data sets to re-index and merge?

Automatic processing: *XDS*

- beamline software (provides XDS.INP)
- scripts: *xia2* (CCP4), *autoPROC* (Globalphasing), *xdsme* (Soleil), *autoxds* (SSRL), *autoprocess* (CMCF), ... *generate_XDS.INP* (XDSwiki), *fast_dp* (Diamond)
- CCP4: *pointless, xdsconv* (type CCP4_I+F, or CCP4, or CCP4_I, or CCP4_F)
- SHELX: *shelxc* reads XDS_ASCII.HKL

Principle of *XDS* processing

- There is one JOB= line in XDS.INP which specifies a list of tasks:

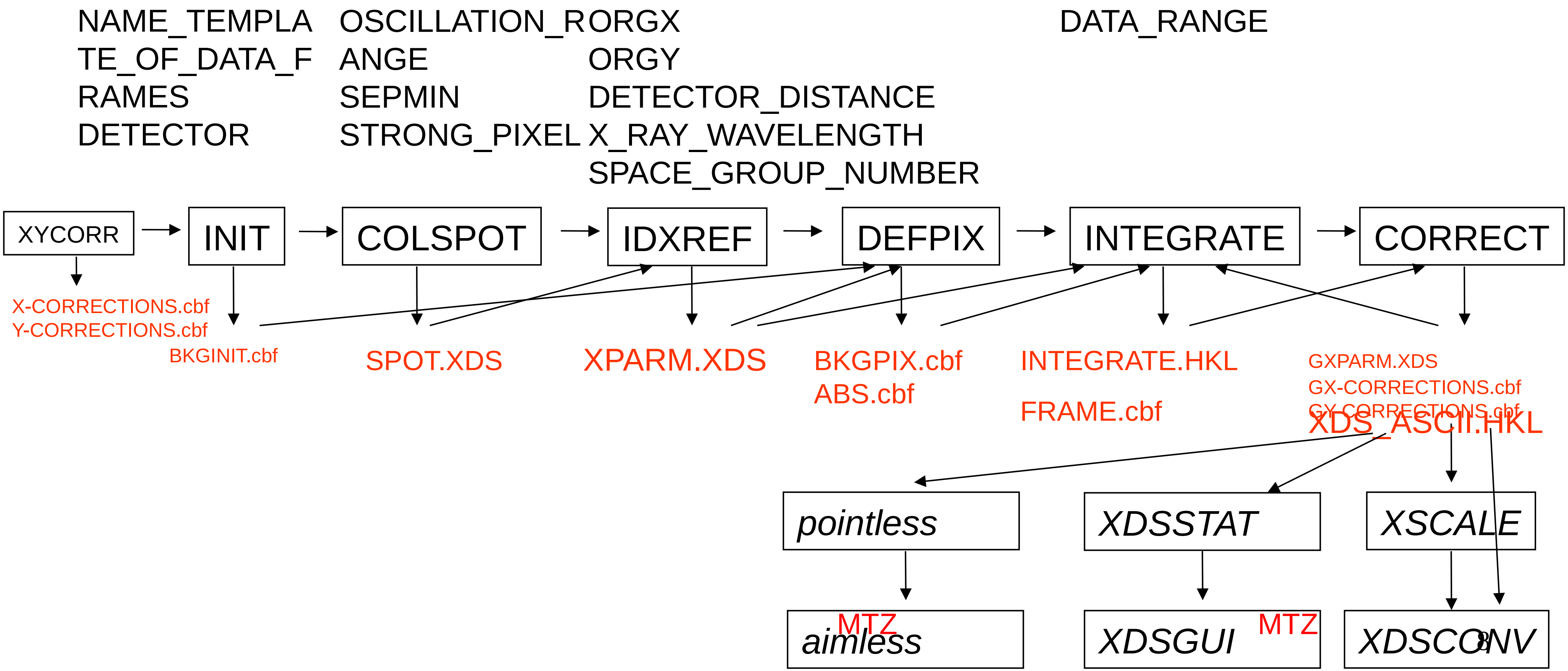
JOB= XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT

- data reduction is divided into tasks in a **modular** way
- information storage/exchange/flow between tasks by data files which may be inspected/analyzed
- each task needs the result from the previous tasks
- fine-tuning of a task does *not* require previous tasks to be repeated
- each task writes its output file <TASK>.LP

The tasks are ...

- XYCORR : write positional correction files (X-CORRECTIONS.cbf, Y-CORRECTIONS.cbf)
- INIT : find background pixels (defaults usually OK)
- COLSPOT: find reflection positions
- IDXREF : "index" reflections; user may supply/choose spacegroup
- XPLAN [not required] : strategy for data collection
- DEFPIX : mask shadows on detector (use XDSGUI!)
- INTEGRATE : evaluates intensities on all frames, writes INTEGRATE.HKL and FRAME.cbf
- CORRECT : scales, rejects outliers, statistics, writes scaled, unmerged XDS_ASCII.HKL (and other files)

Information flow



Example XDS.INP

```
JOB= XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT
ORGX=1546 ORGY=1552        !Detector origin (pixels); e.g. NX/2 NY/2
DETECTOR_DISTANCE=180      !(mm)
OSCILLATION_RANGE=0.50     !degrees (>0)
X-RAY_WAVELENGTH=0.980243 !Angstroem
NAME_TEMPLATE_OF_DATA_FRAMES=frms/wga2-27_1_???.img
DATA_RANGE=1 360           !Numbers of first and last data image collected
BACKGROUND_RANGE=1 10      !Numbers of first and last data image for background
SPACE_GROUP_NUMBER=  19    !0 for unknown crystals; cell constants are ignored.
UNIT_CELL_CONSTANTS= 44.4   86.4   104.5  90 90  90  ! not required if spgr=0
REFINE(IDXREF)=BEAM AXIS ORIENTATION CELL DISTANCE
REFINE(INTEGRATE)=DISTANCE BEAM ORIENTATION CELL ! AXIS
ROTATION_AXIS= 1.0 0.0 0.0
INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0
FRACTION_OF_POLARIZATION=0.99                    ! SLS X06SA
POLARIZATION_PLANE_NORMAL= 0.0 1.0 0.0
DETECTOR=CCDCHESS        MINIMUM_VALID_PIXEL_VALUE=1     OVERLOAD=65000
DIRECTION_OF_DETECTOR_X-AXIS= 1.0 0.0 0.0
DIRECTION_OF_DETECTOR_Y-AXIS= 0.0 1.0 0.0
VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= 7000 30000 !Used by DEFPIX
                                  !for excluding shaded parts of the detector.
INCLUDE_RESOLUTION_RANGE=50.0 1.3 !Angstroem; used by DEFPIX,INTEGRATE,CORRECT
```

Bold keyword/parameter pairs are required. Complete documentation at http://xds.mpimf-heidelberg.mpg.de/html_doc/xds_parameters.html

Example XSCALE.INP

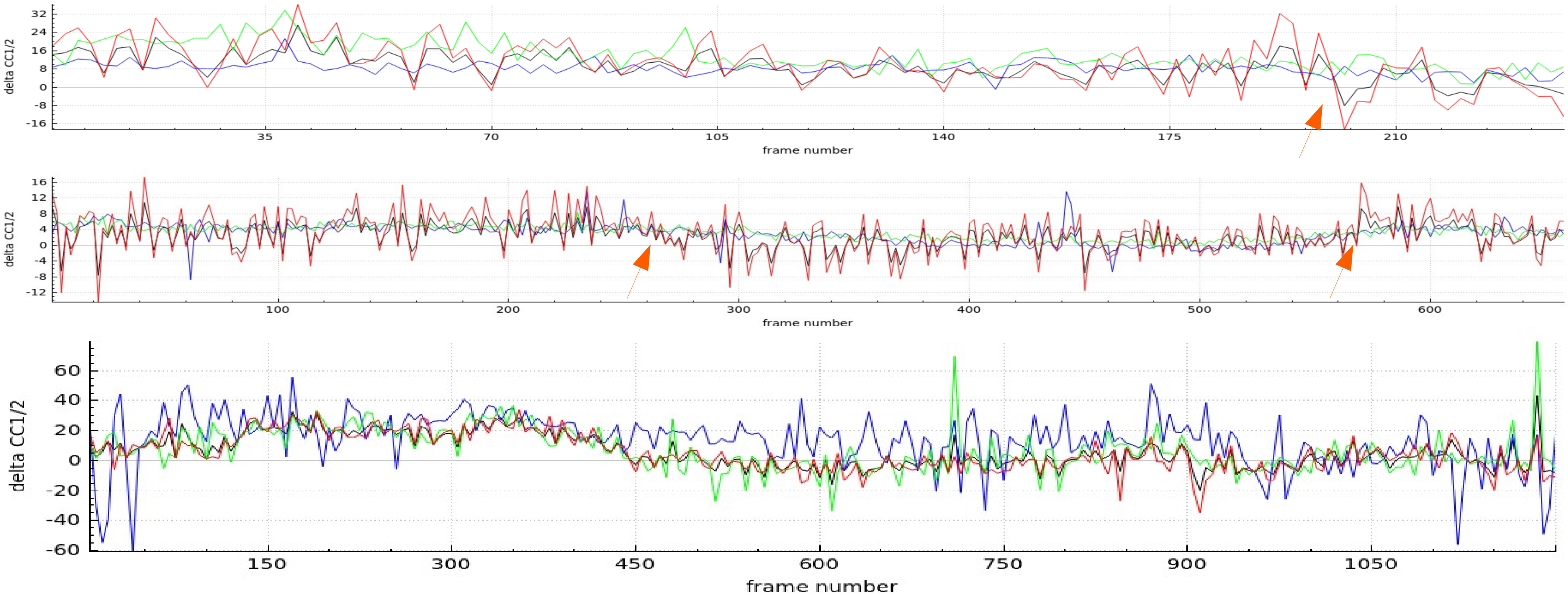
```
!============ EXAMPLE 3: specific reindexing of input data sets
!
!       Use of specific reindexing of input data sets for resolving
!       indexing ambiguities in the scaled output data set. This
!       happens if the crystal's space group symmetry is lower than
!       its lattice symmetry.
!
RESOLUTION_SHELLS= 100 10 6 4 3 2 1.9
SPACE_GROUP_NUMBER=78
UNIT_CELL_CONSTANTS=57.39 57.39 106.9   90 90 90
OUTPUT_FILE=scaf8_all_merged.hkl
MERGE=TRUE FRIEDEL'S_LAW=FALSE
STRICT_ABSORPTION_CORRECTION=TRUE
INPUT_FILE= ../xds-1_2/XDS_ASCII.HKL
REIDX_ISET= -1  0  0  0  0  1  0  0  0  0 -1  0
INPUT_FILE= ../xds-2_1/XDS_ASCII.HKL
INPUT_FILE= ../xds-3_1/XDS_ASCII.HKL
INPUT_FILE= ../xds-1_4/XDS_ASCII.HKL
INPUT_FILE=*../xds-5_1/XDS_ASCII.HKL
```

Bold keyword/parameter pairs are required. Complete documentation at http://xds.mpimf-heidelberg.mpg.de/html_doc/xscale_parameters.html

Manual processing: *XDSGUI*

- problems in phasing and refinement
  often due to bad / wrong data processing
- visually inspect frames; mask shadows
- optimize parameters, frame range,
  resolution cutoff ..
- presents tables as plots
- interfaces to *XDS* through its files, e.g. XDS.INP
- user – extensible / modifiable commands

XDSCC12: calculates $\Delta CC_{1/2,i} = CC_{1/2,with\_i} - CC_{1/2,without\_i}$



Examples of single data sets (plots from XDSGUI)

three resolution ranges (blue=low green=medium red=high) - i refers to batches of width 1°

- find bad frame ranges
- radiation damage

XDSCC12 for selection of data sets for *XSCALE*

Sources of information

- *XDS* main website: http://xds.mpimf-heidelberg.mpg.de - complete, accurate, up-to-date documentation; download
- XDSwiki: http://strucbio.biologie.uni-konstanz.de/xdswiki data sets; documentation; download; links to e.g. Matthew J. Whitley's excellent tutorial given at CSHL 2018
- CCP4 bulletin board
- SBGrid talk (May 2020) at https://www.youtube.com/watch?v=3WU9NrILECo
- Making a difference in multi-data-set crystallography : simple and deterministic data-scaling/selection methods. Assmann, G.M., Wang, M., Diederichs, K. (2020) Acta Cryst D76, 636 (serial crystallography, *XDSCC12*)

Garbage in – garbage out!

Lots of possibilities for mistakes in / problems due to the experiment:

- use a decent detector: Pilatus or Eiger
- wrong exposure → radiation damage: fix is to expose weakly, and collect 360°
- wrong oscillation range: fix is to collect with $\Delta\phi$=0.1°
- Badly centered crystal moves out of the beam during rotation: fix is …
- Ice: the fix is watching and listening to experts like E. Garman, I. Schlichting, and practicing. According to autoPROC, ice rings affect 5-10% of all reflections if present.

It is not the goal of the experiment to collect as many data sets as are technically possible, but rather to collect the best data technically possible, for a given project.

For important data, do not rely on automatic data processing

There are no clear-cut criteria for the quality of data set (“Table 1” is almost useless).

Synchrotrons typically run multiple pipelines, and the user has to choose ...

What can go wrong in automatic data processing?

- does not handle radiation damage (discard frames towards the end of the data set)
- does not handle shadowed areas of detector
- does not handle indexing problems (multiple lattices, ice, …) flexibly
- does not optimize processing

Automatic processing is ok for good data sets! GlobalPhasing’s *autoPROC* is mostly reliable. So is xia2 for DLS data. Difficult data sets benefit most from human insight.

Items that should be in "Table 1" (but are not)

- are the data anisotropic? how much (3 numbers)?
- is there anomalous signal? how much (resolution where it becomes insignificant, $CC_{1/2\text{-ano}}$ at low resolution)?
- Is there tNCS? how strong?
- Twinning? Estimate from L-test?
- Ice rings? How handled?
- Radiation damage? How measured? (*XDSCC12* and R_d plot)

Get to know your data: check the *pointless* logfile ...

```
Analysing rotational symmetry in lattice group I 4/m m m
----------------------------------------------


Scores for each symmetry element

Nelmt  Lklhd  Z-cc    CC        N  Rmeas    Symmetry & operator (in Lattice Cell)

  1   0.856   7.77   0.78    9368  0.224     identity
  2   0.139   2.98   0.30    6376  0.894     2-fold l ( 0 0 1) {-h,-k,l}
  3   0.842   8.02   0.80   18679  0.234 **  2-fold k ( 0 1 0) {-h,k,-l}
  4   0.148   3.10   0.31    5858  1.254     2-fold h ( 1 0 0) {h,-k,-l}
  5   0.073   1.25   0.13   12111  1.077     2-fold   ( 1-1 0) {-k,-h,-l}
  6   0.071   1.11   0.11   12293  1.152     2-fold   ( 1 1 0) {k,h,-l}
  7   0.068   0.83   0.08   24758  1.084     4-fold l ( 0 0 1) {-k,h,l}{k,-h,l}


Checking for possible twinning
      Estimated twin fraction alpha from cumulative N(|L|) plot 0.003 (+/-0.001)
   < |L| >:    0.496 (0.5 untwinned, 0.375 perfect twin)
      Estimated twin fraction alpha from < |L| >    0.000
   < L^2 >:    0.328 (0.333 untwinned, 0.2 perfect twin)
      Estimated twin fraction alpha from < L^2 >    0.000

The L-test suggests that the data are not twinned
```

Get to know your data: Check the *phenix.xtriage* log

... for anisotropy information:

```
----------Maximum likelihood anisotropic Wilson scaling----------

ML estimate of overall B_cart value:
  33.97,  0.00,  4.46
          72.36,  0.00
                 62.21
```

→ **consider *StarAniso* or similar (Phaser)**

... for pseudosymmetry (tNCS) information:

```
 ----------Patterson analyses----------

 Largest Patterson peak with length larger than 15 Angstrom:
 Frac. coord.              :   -0.273    0.500   -0.298
 Distance to origin        :   60.425
 Height relative to origin :   11.106 %
 p_value(height)           :    1.076e-01

Explanation ...
```

Manual checks are easily performed with *XDSGUI*

- before running *XDS*: check frames (*adxv*, *Albula*, *XDSGUI* ...)
- **Frame** tab: FRAME.cbf for match obs/calc, and masked areas
- **IDXREF** tab: for ice rings / indexed *vs* un-indexed reflections
- look at indexing in reciprocal space: **tools** tab/Further analyses/spot2pdb-coot command; see XDSwiki:XDSGUI#tools
- **statistics** tab: *XDSCC12* and *XDSSTAT*

Try to optimize data processing

- XDSwiki:Optimisation#Re-INTEGRATEing_with_the_correct_spacegroup.2C_refined_geometry_and_fine-slicing_of_profiles

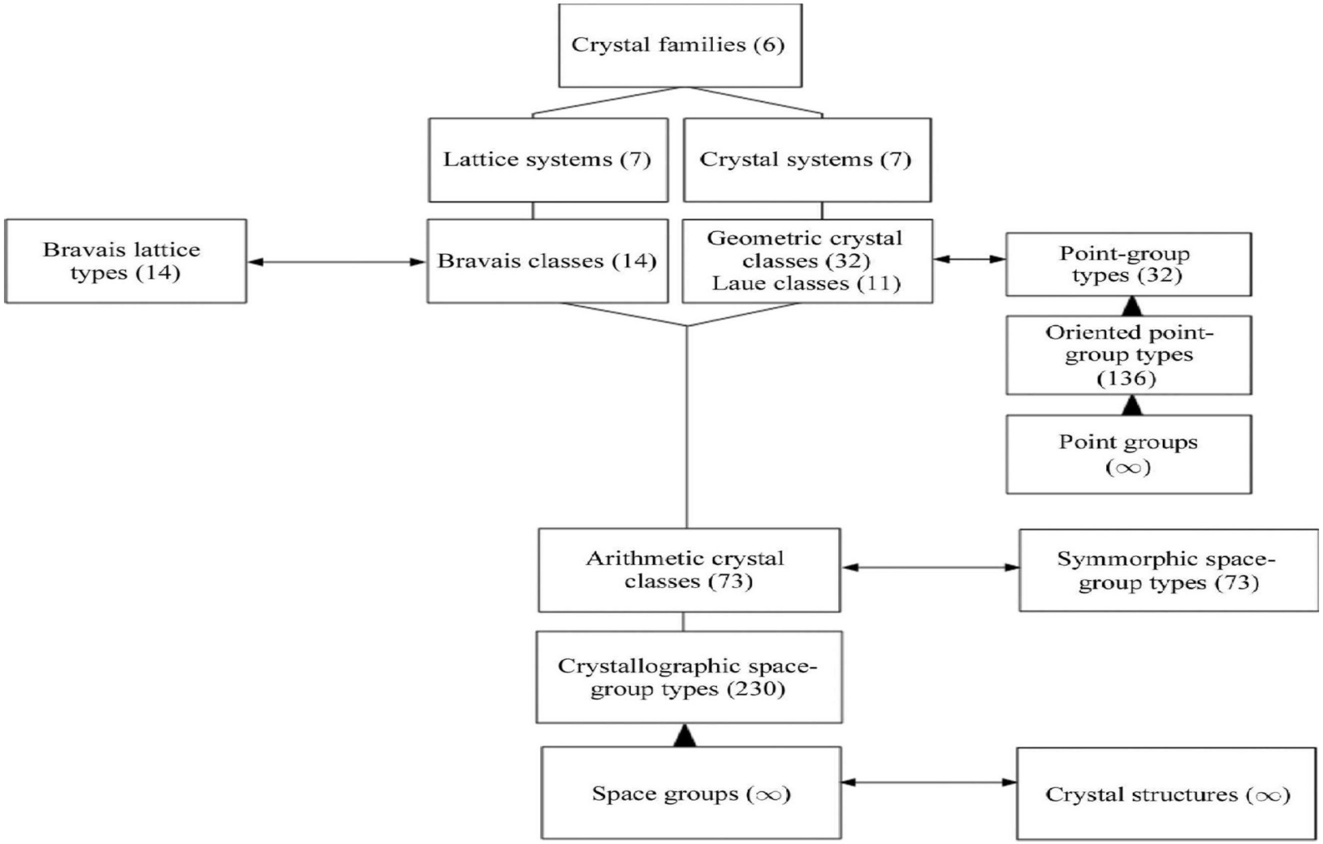
  XDSwiki:Optimisation#using_the_refined_values_for_beam_divergence_and_mosaicity_for_re-integration
- **tools** tab: "Saving and comparing good results" and "Optimizing data quality". After changing parameters, run "JOB=DEFPIX INTEGRATE CORRECT", compare and save if better/restore old if worse.
- consider use of *StarAniso* if anisotropy

  … may make the difference between interpretable and non-interpretable map, substructure solution, ...

Space group determination

- seems to frighten (every year more) new crystallographers
- part of this is because the terminology is rich, and often inconsistent: “crystal class” or “crystal system”, “Laue class” or “Laue point group”, ...? The differences are often subtle, and are only appreciated by experts.
- another part of this is because there are many rules that come into play: re-indexing, alternate indexing, quantitative vs qualitative aspects …

Hierarchy of the classification of space groups and point groups into classes, systems and families. Horizontal double-ended arrows indicate a 1:1 correspondence. Black triangles indicate that, in ...



JAC Appl Cryst
JOURNAL OF APPLIED CRYSTALLOGRAPHY

Nespolo *et al.*

Volume 51 | **Part 5** | October 2018 | Pages 1481–1491 | 10.1107/S1600576718012724

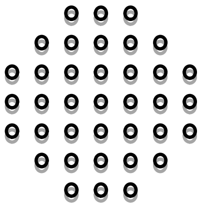
How do we deduce the Space Group in practice?

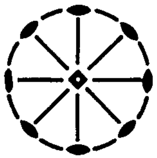
- We start in reciprocal space (point group)
- We go all way back from symmetry in reciprocal space to crystal space group
  - Data processing gives values of the unit cell parameters
  - Lattice symmetry is derived from the unit cell parameters
  - Comparison of related intensities gives crystal point group
  - Systematic absences allow to reduce the number of possible space groups.
  - Space group is only a hypothesis until structure is complete

(A. Lebedev)

Space group assignment (e.g. *pointless*)

Reciprocal space lattice (positions of reflections)

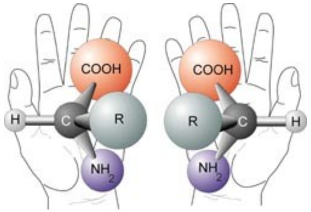


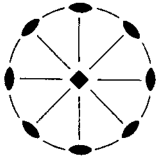


4/mmm

(Approximate) lattice point group

Mirror symmetry is <u>not</u> allowed in biological macromolecules

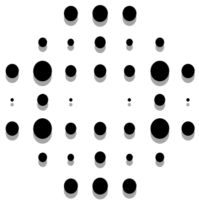


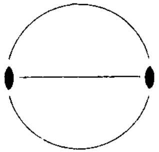


4 2 2

Highest possible crystal point group

Intensities of reflections

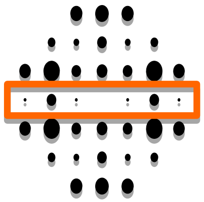


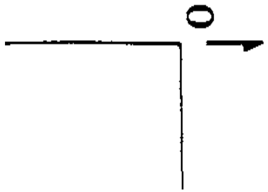


2

Probable crystal point group

Intensities of axial reflections





$P\ 2_1$

Probable crystal space group

User: decision making, structure solution, final space group assignment

(A. Lebedev)

Space group determination

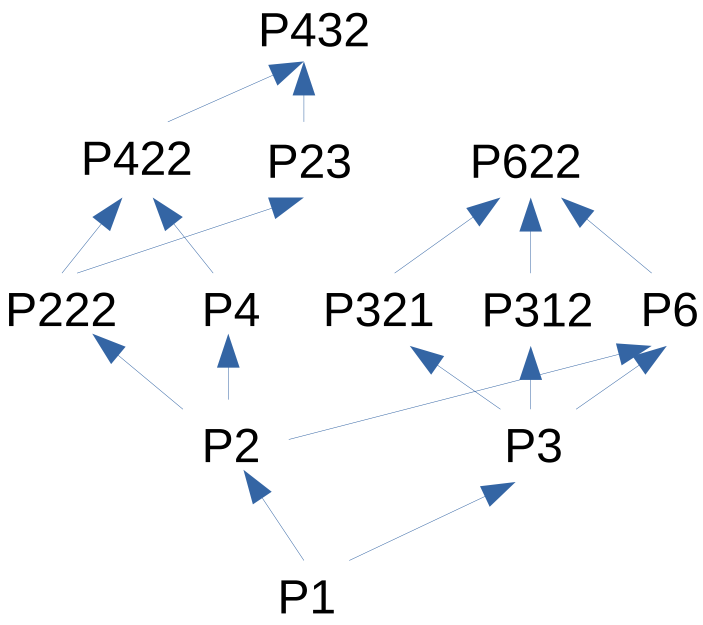
- seems to frighten (every year more) new crystallographers
- part of this is because the terminology is often inconsistent: “crystal class” or “crystal system”, “Laue class” or “Laue point group”, ...? The differences are often subtle, and only for experts.
- another part of this is because there are many rules that come into play: re-indexing, alternate indexing, quantitative vs qualitative aspects…
- It helps to think of space groups being built up hierarchically by adding symmetry elements. Adding a symmetry element means going to the supergroup of a space group. If this happens along a new direction, it also means an additional constraint on the cell parameters. Removing a symmetry element means going to a subgroup.
- The tables at https://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Space_group_determination should make the understanding easier.

https://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Space_group_determination

spacegroup number	maximum *translationengleiche* subgroup	minimum *translationengleiche* supergroup	spacegroup name
1	-	3, 4, 5, 143, 144, 145, 146	P 1
3	1	16, 17, 18, 21, 75, 77, 168, 171, 172	P 2
4	1	17, 18, 19, 20, 76, 78, 169, 170, 173	P 21
5	1	20, 21, 22, 23, 24, 79, 80, 149, 150, 151, 152, 153, 154, 155	C2
16	3	89, 93, 195	P 2 2 2
17	3, 4	91, 95	P 2 2 21
18	3, 4	90, 94	P 21 21 2
19	4	92, 96, 198	P 21 21 21
20	4, 5	91, 92, 95, 96, 178, 179, 182	C 2 2 21
21	3, 5	89, 90, 93, 94, 177, 180, 181	C 2 2 2
22	5	97, 98, 196	F 2 2 2
23	5	97, 197	I 2 2 2
24	5	98, 199	I 21 21 21
75	3	89, 90	P 4
76	4	91, 92	P 41
77	3	93, 94	P 42
78	4	95, 96	P 43
79	5	97	I 4
80	5	98	I 41
89	16, 21, 75	207	P 4 2 2
90	18, 21, 75	-	P 4 21 2
91	17, 20, 76	-	P 41 2 2
92	19, 20, 76	213	P 41 21 2
93	16, 21, 77	208	P 42 2 2
94	18, 21, 77	93, 97	P 42 21 2
95	17, 20, 78	-	P 43 2 2
96	19, 20, 78	212	P 43 21 2
97	22, 23, 79	209, 211	I 4 2 2
98	22, 24, 80	210, 214	I 41 2 2
143	1	149, 150, 168, 173	P 3
144	1	151, 152, 169, 172	P 31
145	1	153, 154, 170, 171	P 32
146	1	155, 195, 196, 197, 198, 199	R 3
149	5, 143	177, 182	P 3 1 2
150	5, 143	177, 182	P 3 2 1
151	5, 144	178, 181	P 31 1 2
152	5, 144	178, 181	P 31 2 1
153	5, 145	179, 180	P 32 1 2
154	5, 145	179, 180	P 32 2 1
155	5, 146	207, 208, 209, 210, 211, 212, 213, 214	R 3 2
168	3, 143	177	P 6
169	4, 144	178	P 61
170	4, 145	179	P 65
171	3, 145	180	P 62
172	3, 144	181	P 64
173	4, 143	182	P 63
177	21, 149, 150, 168	-	P 6 2 2
178	20, 151, 152, 169	-	P 61 2 2
179	20, 153, 154, 170	-	P 65 2 2
180	21, 153, 154, 171	-	P 62 2 2

Space group relationships

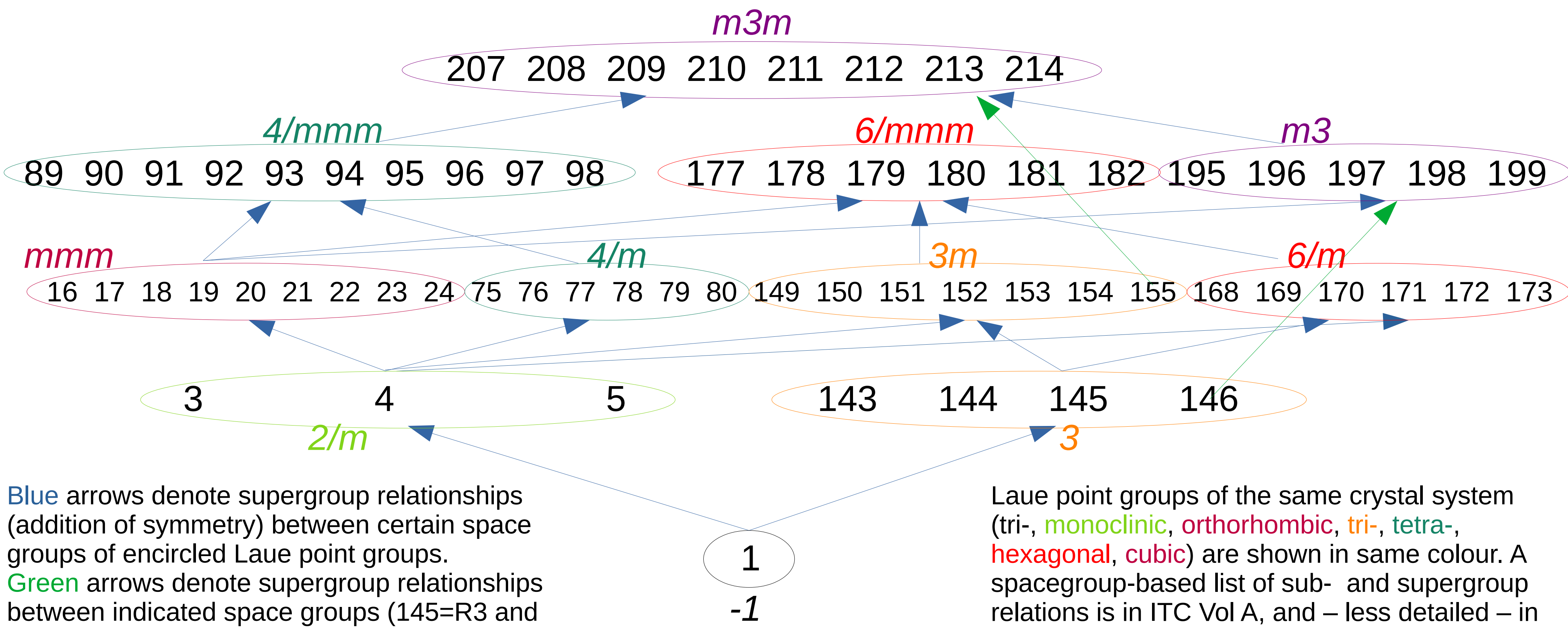
Example: relations of subgroups and supergroups of primitive space groups without screw axes



This is relevant also for twinning

(the figure could be expanded to encompass all 65 spacegroups)

Crystal systems, Laue point groups, the 65 Sohncke space groups, and their sub- und supergroups



Blue arrows denote supergroup relationships (addition of symmetry) between certain space groups of encircled Laue point groups.
Green arrows denote supergroup relationships between indicated space groups (145=R3 and 155=R32) and a higher-symmetry Laue point group – only these two “jump“ a layer.

Laue point groups of the same crystal system (tri-, monoclinic, orthorhombic, tri-, tetra-, hexagonal, cubic) are shown in same colour. A spacegroup-based list of sub- and supergroup relations is in ITC Vol A, and – less detailed – in the `Space_group_determination` article in XDSwiki.

Thank you!