Automatic data processing at the ESRF: An overview of new and old features

Max Nanao
EMBL Grenoble

EMBL

Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

Currently at ESRF

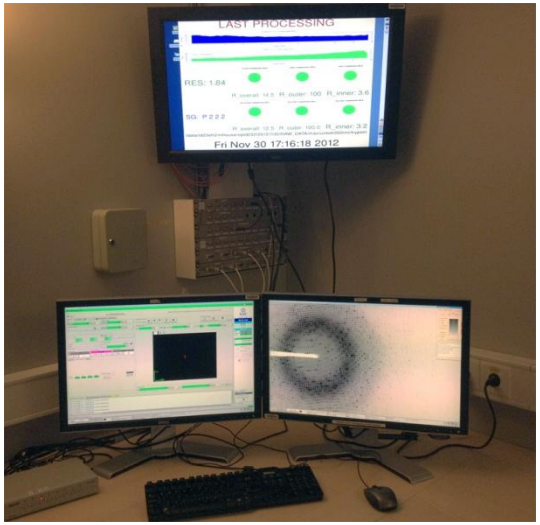
- A suite of data processing and analysis infrastructure for all datasets collected at MX undulator beamlines (and BM14)

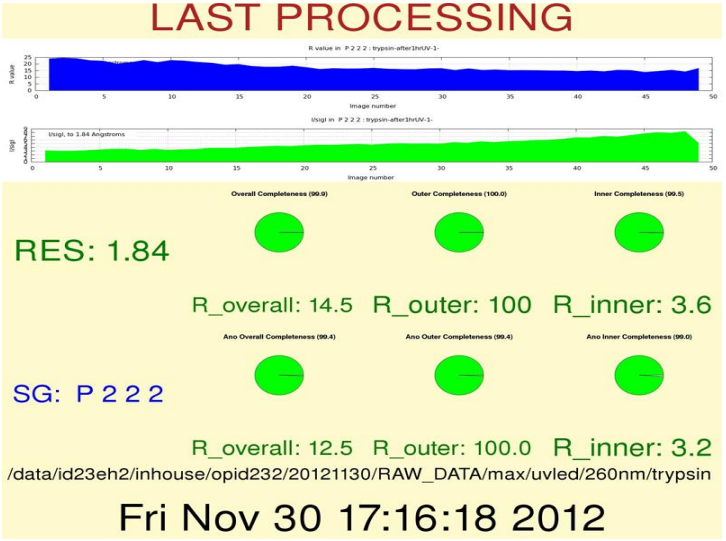
Some elements of the ESRF online data analysis portfolio

- “Heads up Display” of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB

Some elements of the ESRF online data analysis portfolio

- "Heads up Display" of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB

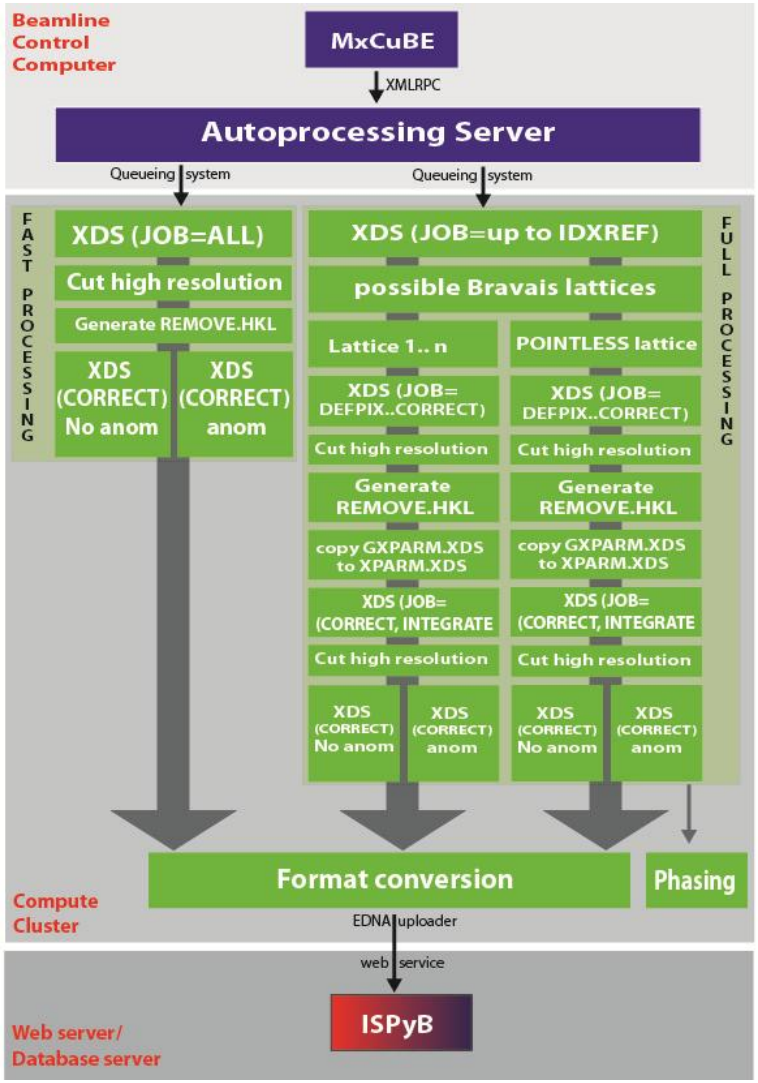




Monaco *et al.* J Appl Crystallogr. 2013 Jun 1;46(Pt 3):804-810.

Some elements of the ESRF online data analysis portfolio

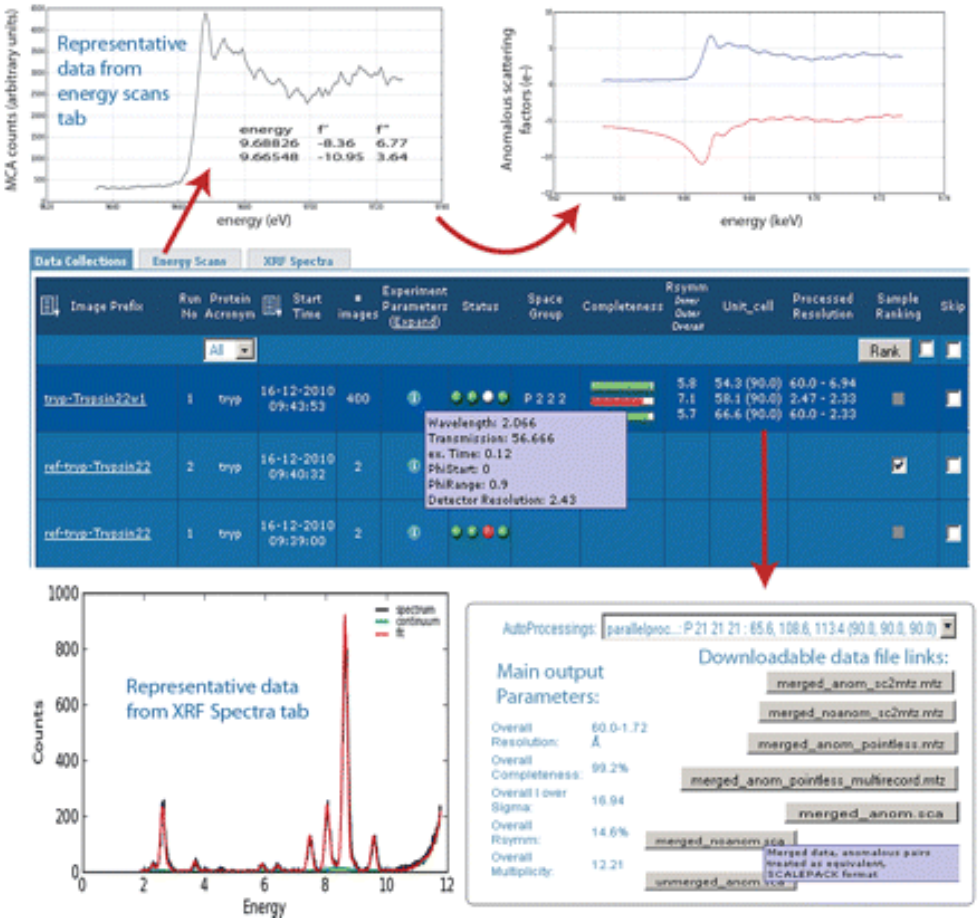
- "Heads up Display" of data from the last crystal
- Automatic indexing and integration
- Presentation of results in ISPyB



Monaco *et al.* J Appl Crystallogr. 2013 Jun 1;46(Pt 3):804-810.

Some elements of the ESRF online data analysis portfolio

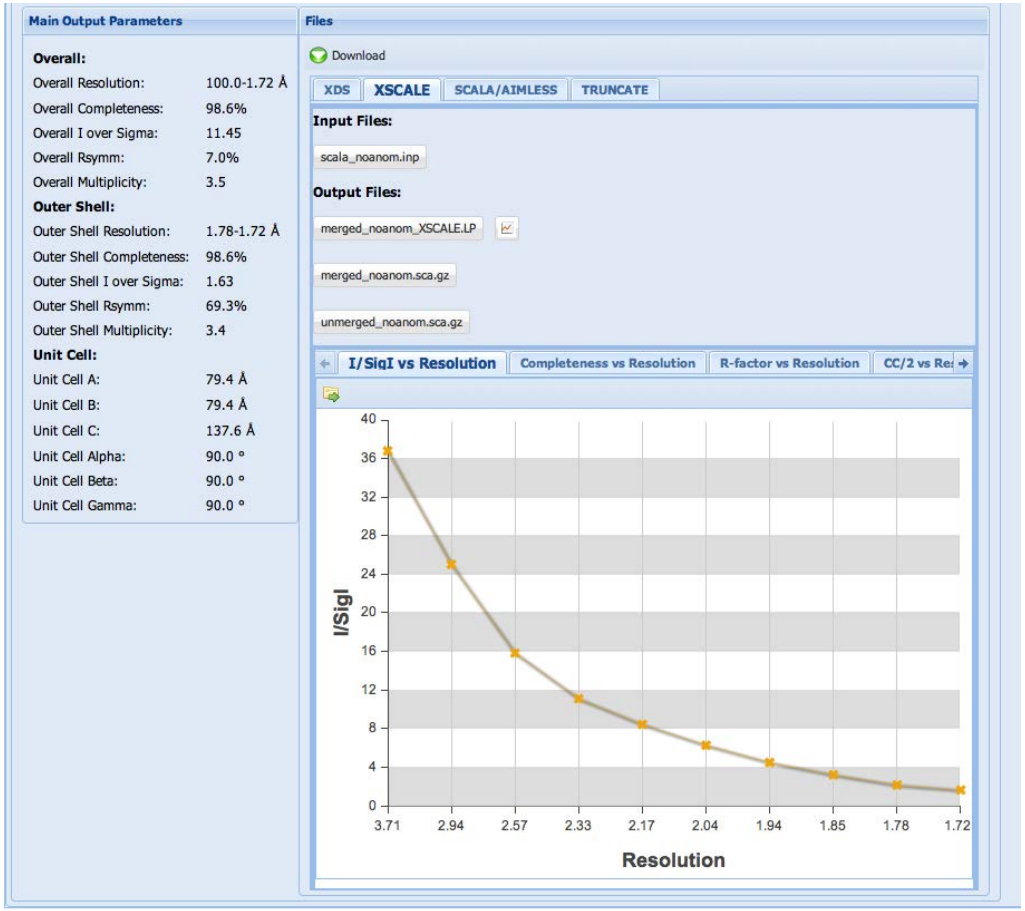
- "Heads up Display" of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB



Delagenière *et al*. Bioinformatics. 2011 Nov 15;27(22):3186-92

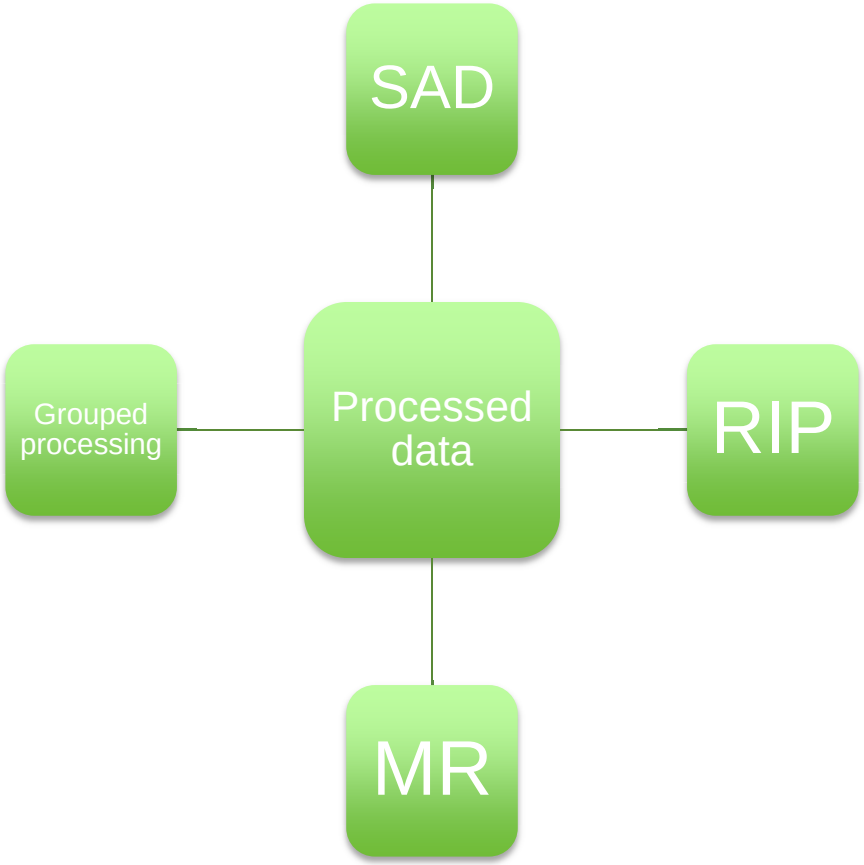
Some elements of the ESRF online data analysis portfolio

- "Heads up Display" of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB



Delagenière *et al*. Bioinformatics. 2011 Nov 15;27(22):3186-92

Automatic processing is the starting point for more advanced analysis

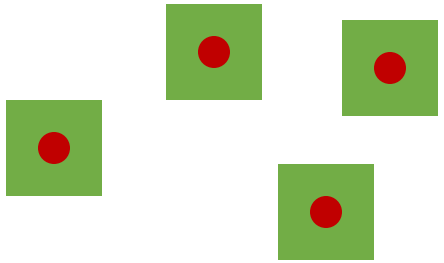


Today:

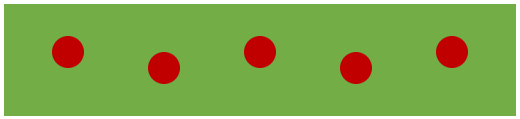
- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

Grouped data processing…

- Multi position datasets becoming more and more common (and easier in MxCube 2.0!)
- The palette of data collection schemes is beginning to outpace the “back end”: what can be automatically processed
- → Need for a simple, rapid and robust system to automatically process data coming from these more complex schemes



Multiple crystals from one or many sample holders



Multiple discrete positions from one xtal

Multiple sweeps from one position

A few challenges

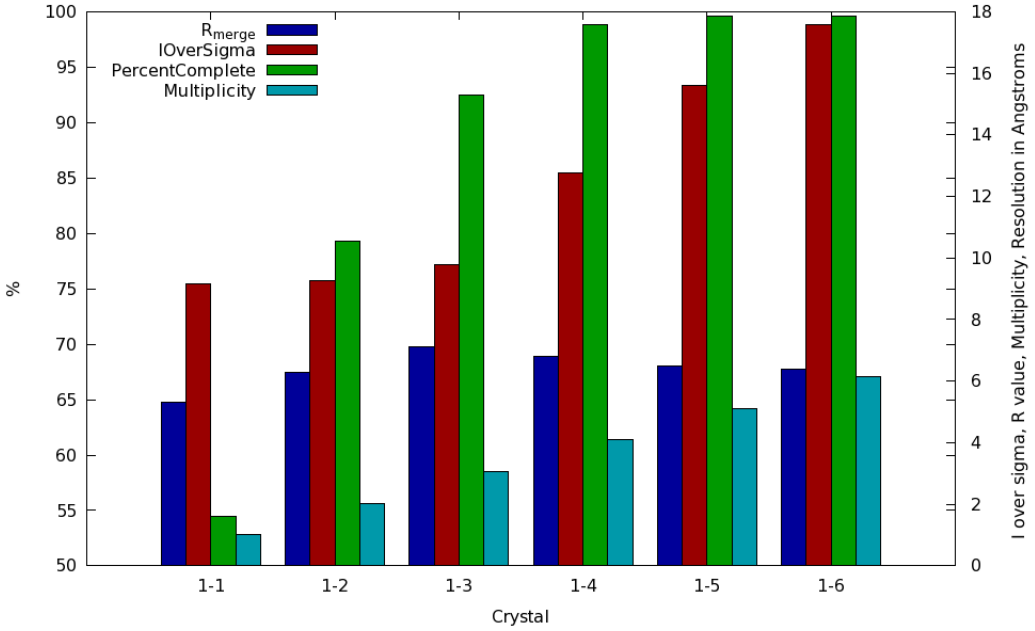
- Consistent indexing between wedges
- Reliable indexing
- Timely results
- Minimal user intervention
- Multiple elements require coordinated modification
  - MXCuBE
  - autoprocessing server
  - ISPyB data model + display

Internals

- Pointless used on all data to determine the SG+Bravais lattice
- Individual wedges are re-indexed as required
- XSCALE used to scale wedges together

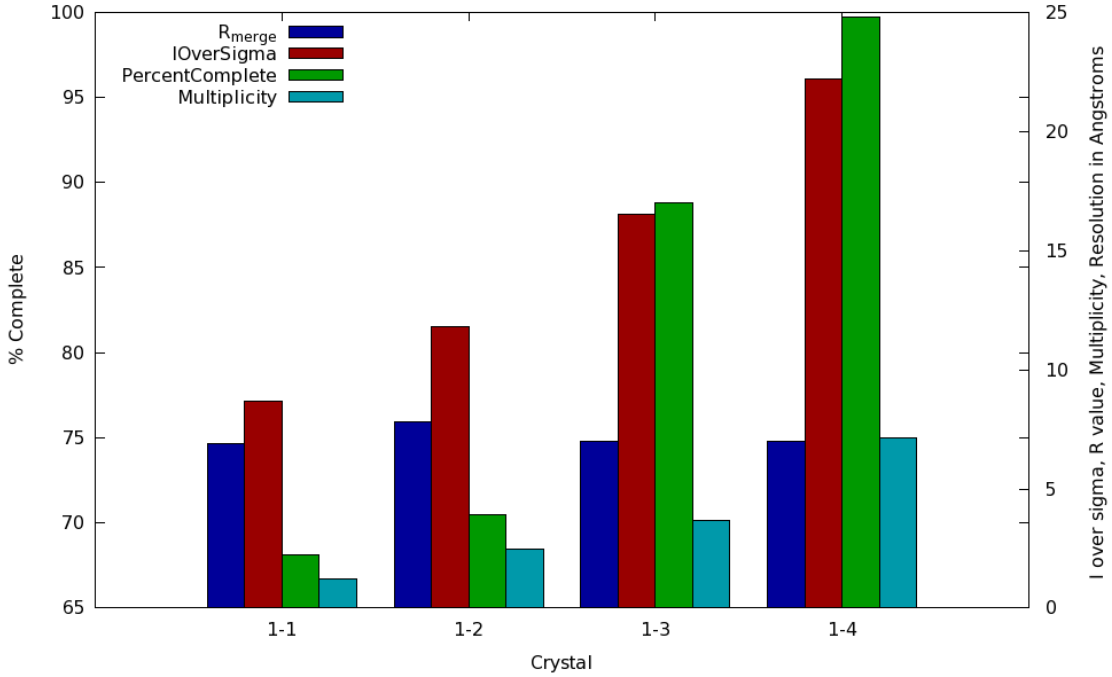
A few examples:

- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature



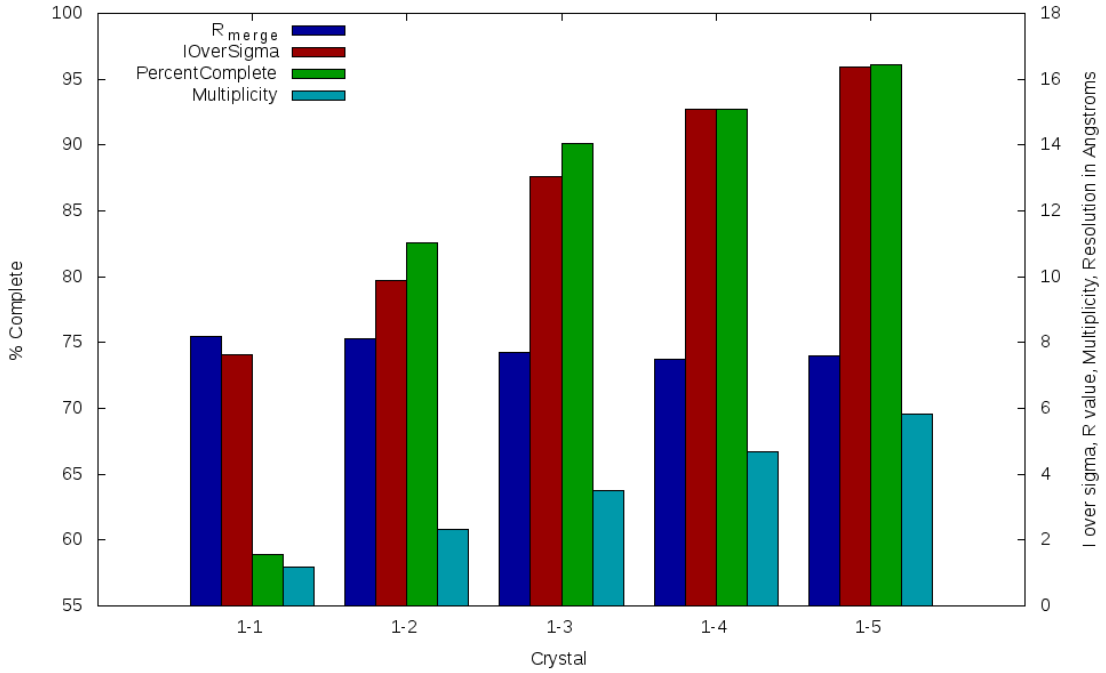
A few examples:

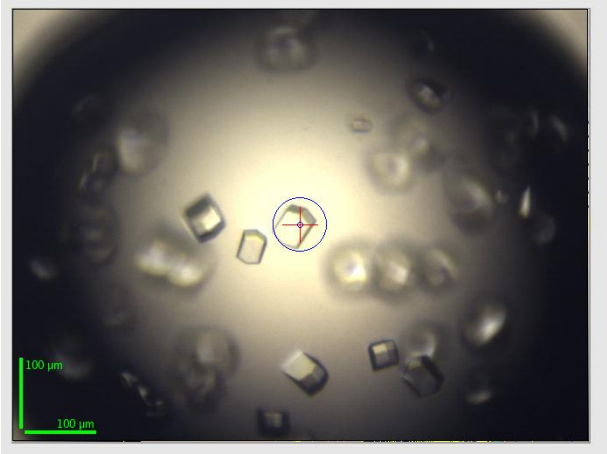
- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature



A few examples:

- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature

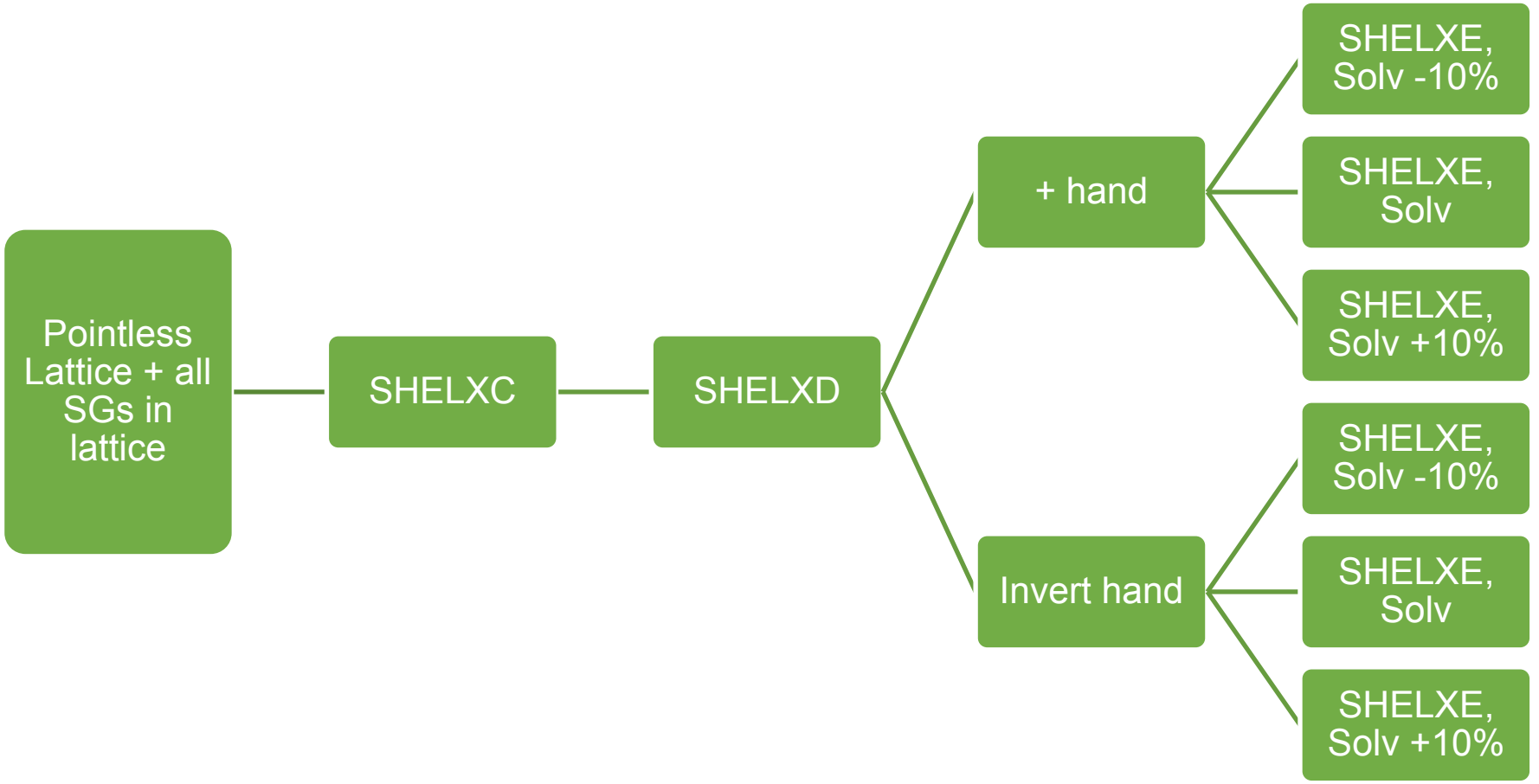




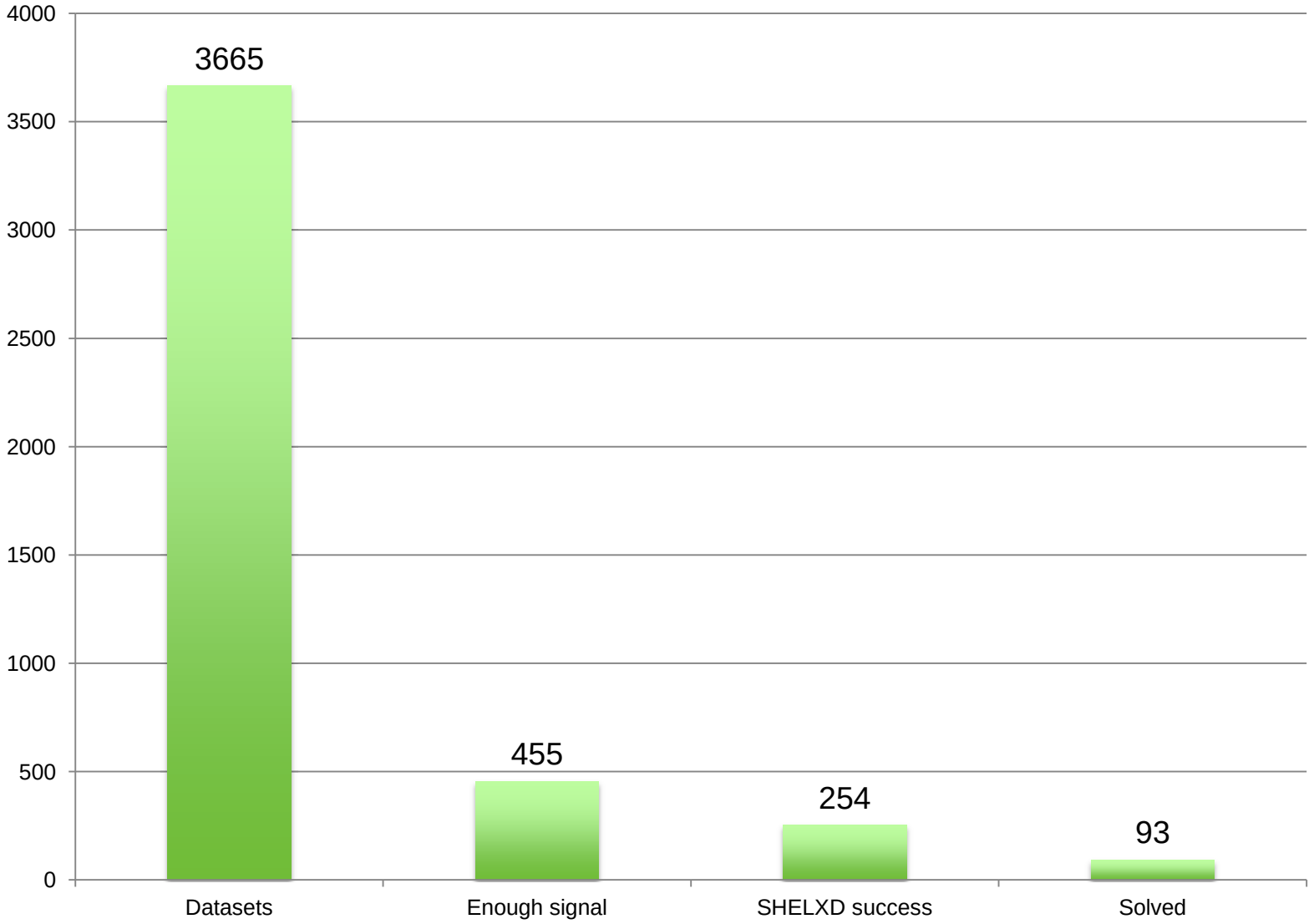
Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

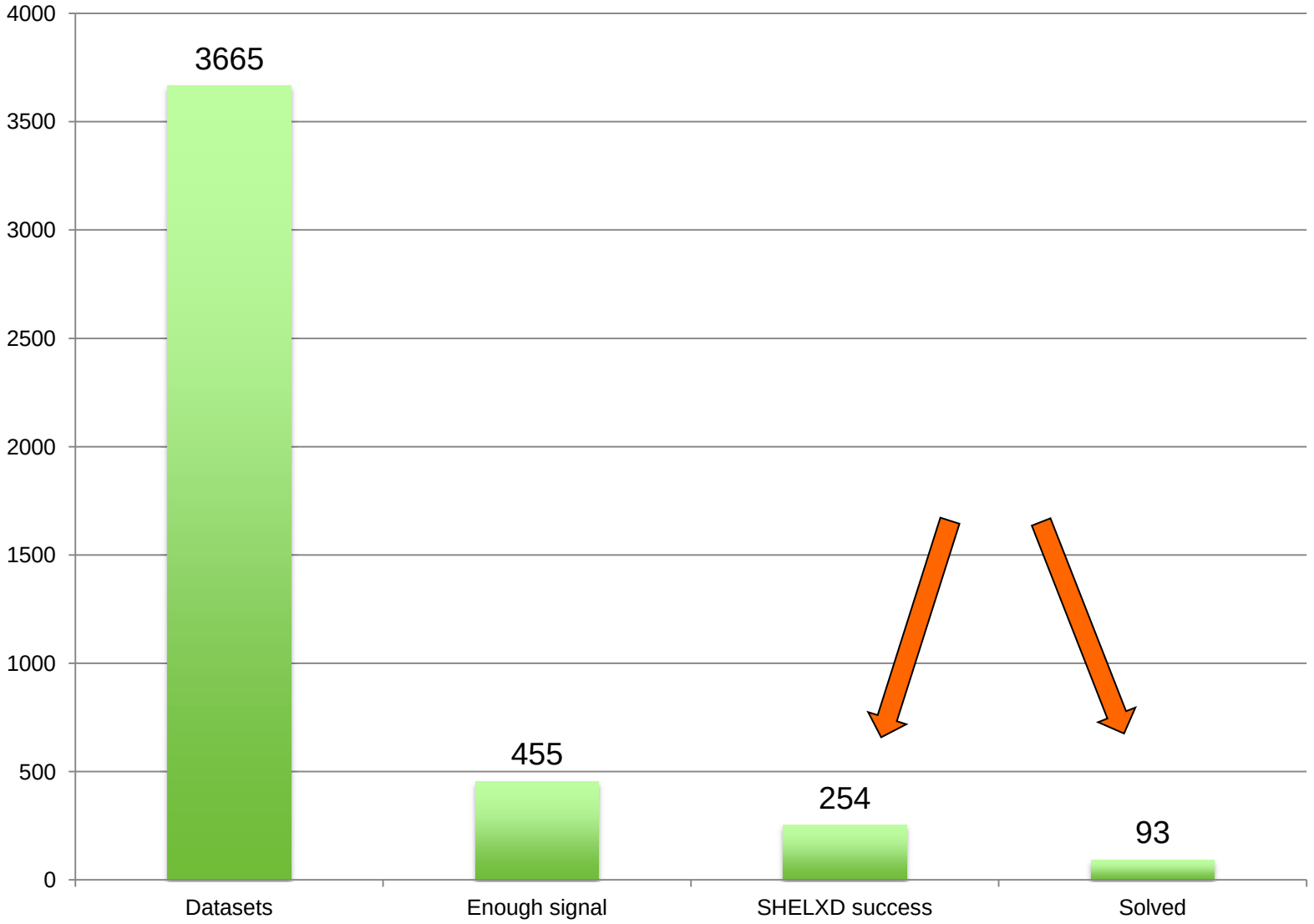
Automatic Structure Solution



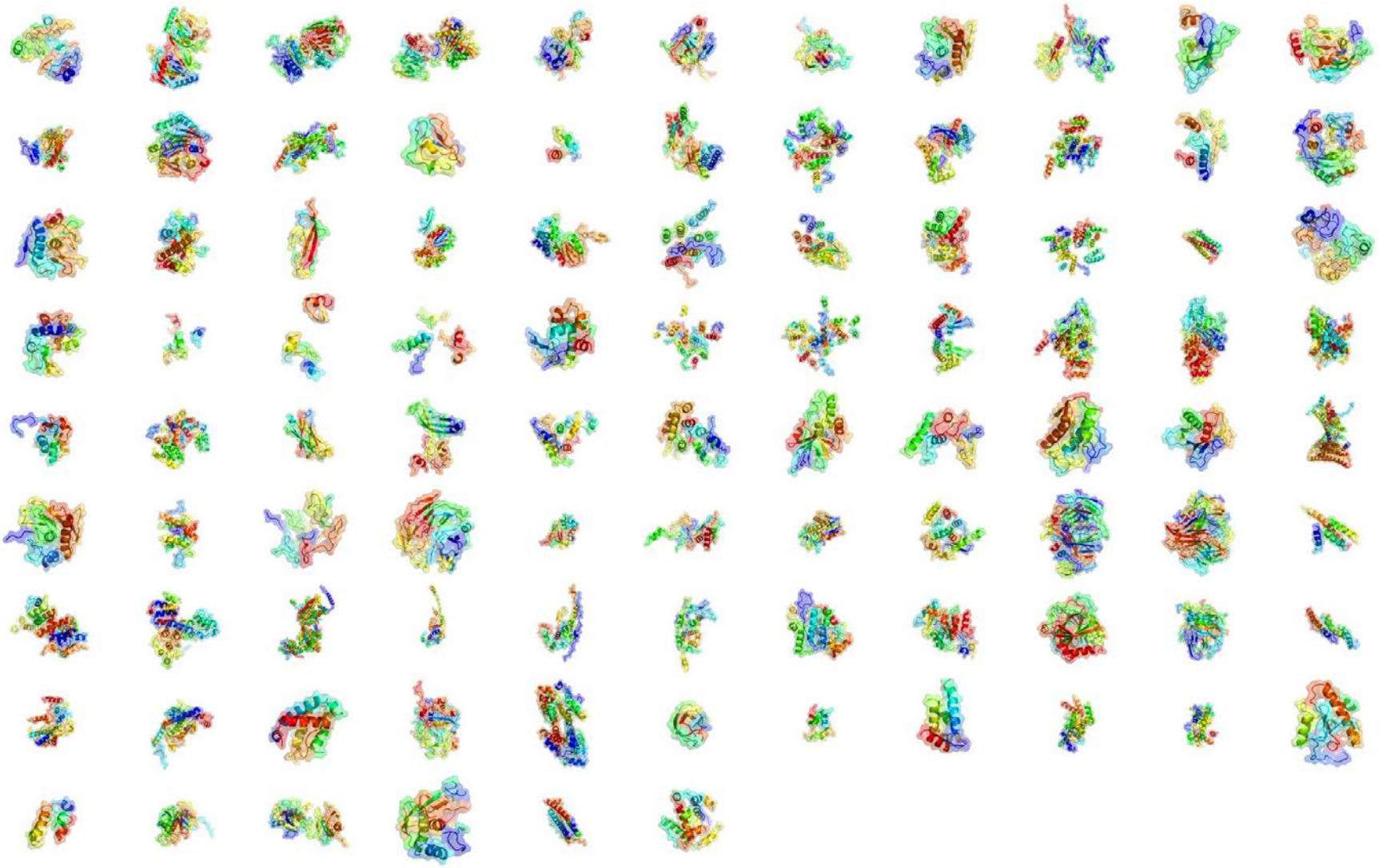
Some statistics from visitor data



Some statistics from visitor data



Some structures solved automatically



Monaco *et al.* J Appl Crystallogr. 2013 Jun 1;46(Pt 3):804-810.

Today:

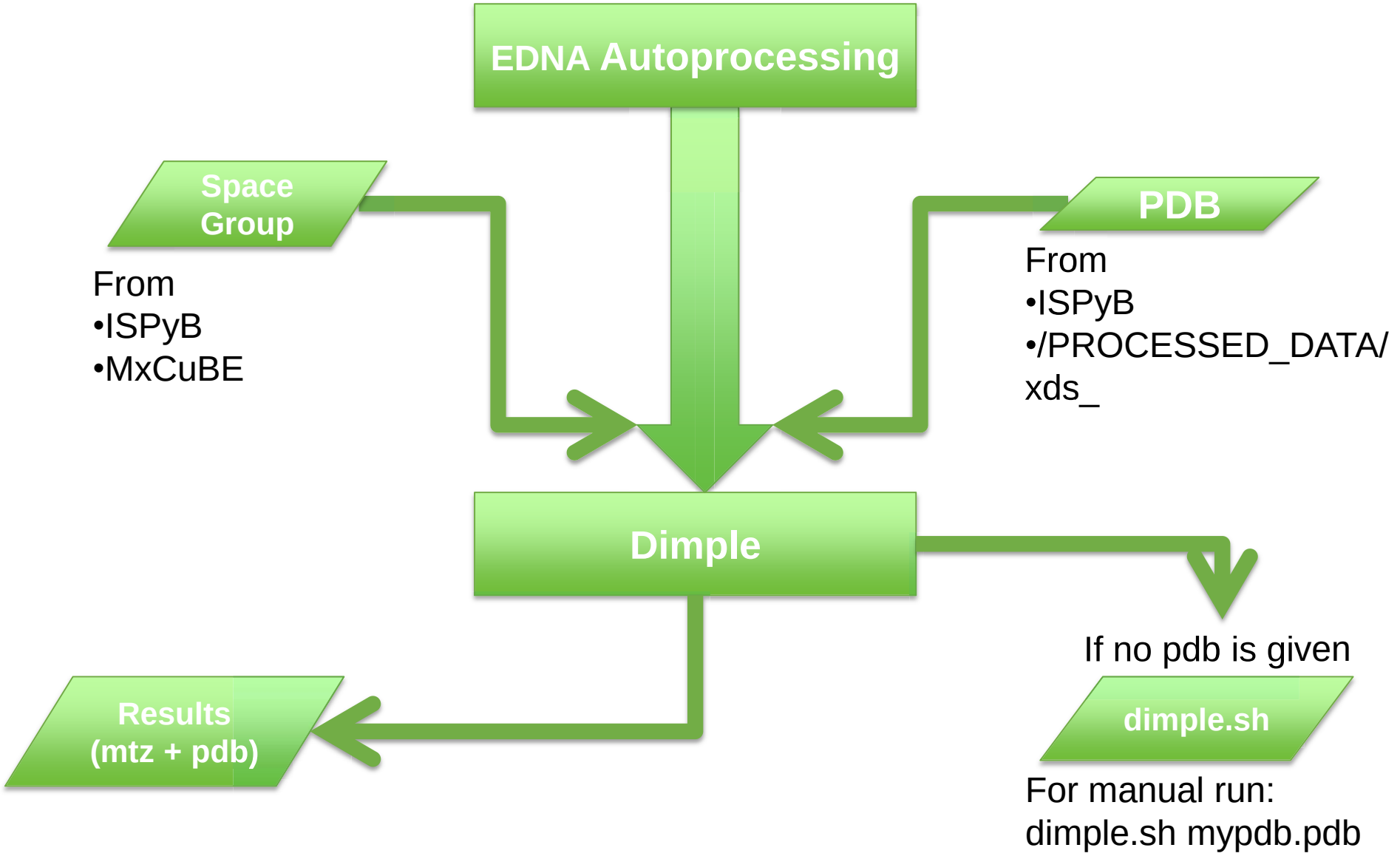
- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

Automatic MR

- MR and difference map peak identification possible through DIMPLE if the user specifies the PDB in MXCube

***DIMPLE* - a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands**
M. Wojdyr, R. Keegan, G. Winter and A. Ashton
*Acta Cryst A. (2013). A**69**, s299*

Dimple MR and difference peak identification



Automatic MR

- If the user does not specify the PDB?
  - The DIMPLE pipeline makes a script for running DIMPLE later
  - **Search PDB for similar unit cell parameters**

Macro-to-micro structural proteomics: native source proteins for high-throughput crystallization.
Totir M, Echols N, Nanao M, Gee CL, Moskaleva A, Gradia S, Iavarone AT, Berger JM, May AP, Zubieta C, Alber T.
PLoS One. 2012;7(2):e32498.

Automatic MR by unit cell

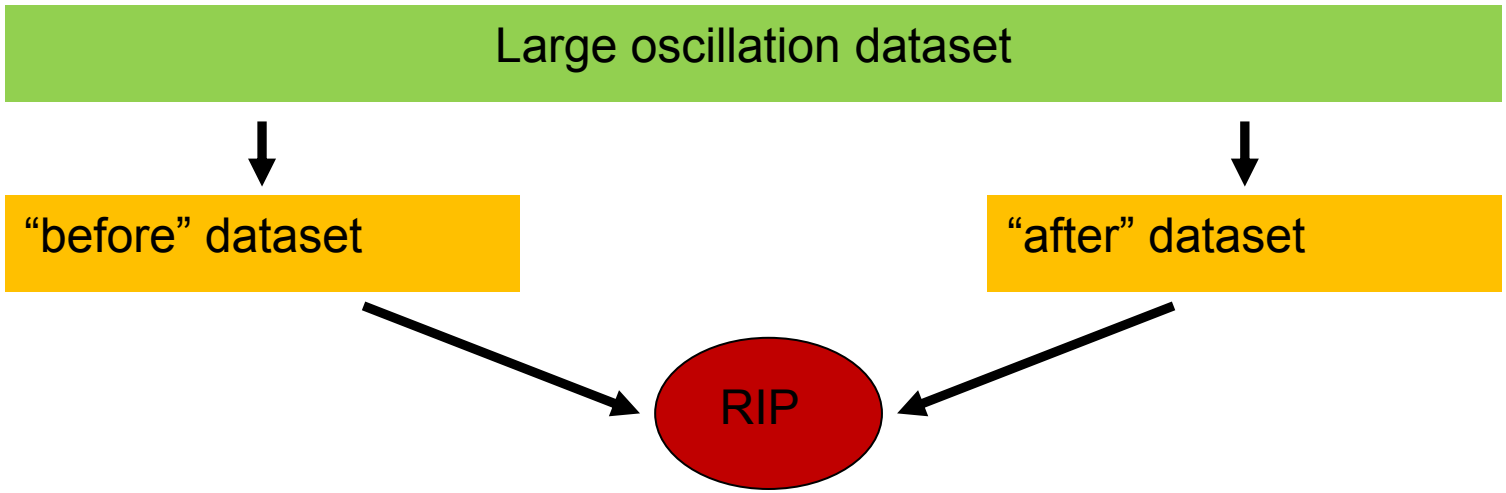
- Fast processing cell parsed
- Unit cell dimensions submitted to Nearest-Cell
- List of PDBs in different families returned
- Top *n* downloaded from EBI
- Search model cleaned up (First protein chain, no waters)
- PHASER
- Evaluation by Z-scores. Failed directories are deleted
- Refmac restrained refinement
- Water updating in Phenix.refine

Nearest-cell: a fast and easy tool for locating crystal matches in the PDB.
Ramraj V, Evans G, Diprose JM, Esnouf RM.
Acta Crystallogr D Biol Crystallogr. 2012 Dec;68(Pt 12):1697-700.

Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP

Automatic segmented Radiation Damage Induced Phasing



- Large oscillation dataset split into sub datasets for RIP
- Very well adapted to large datasets (i.e. PILATUS datasets)

Segmenting data sets for RIP. de Sanctis D and Nanao MH. Acta Crystallogr D Biol Crystallogr. 2012 Sep;68(Pt 9):1152-62.

Feature grid

Feature	Status
Automatic SAD phasing	Functional and solving structures Needs IspyB integration
Automatic Grouped data processing	Re-coded for new MxCube, still in testing but deployed Needs IspyB integration
Auto-MR	Installed January 2014, in testing
Dimple MR	Functional but needs MxCube 2.0 modification, Needs IspyB integration
Auto-RIP	Beta testing

People:

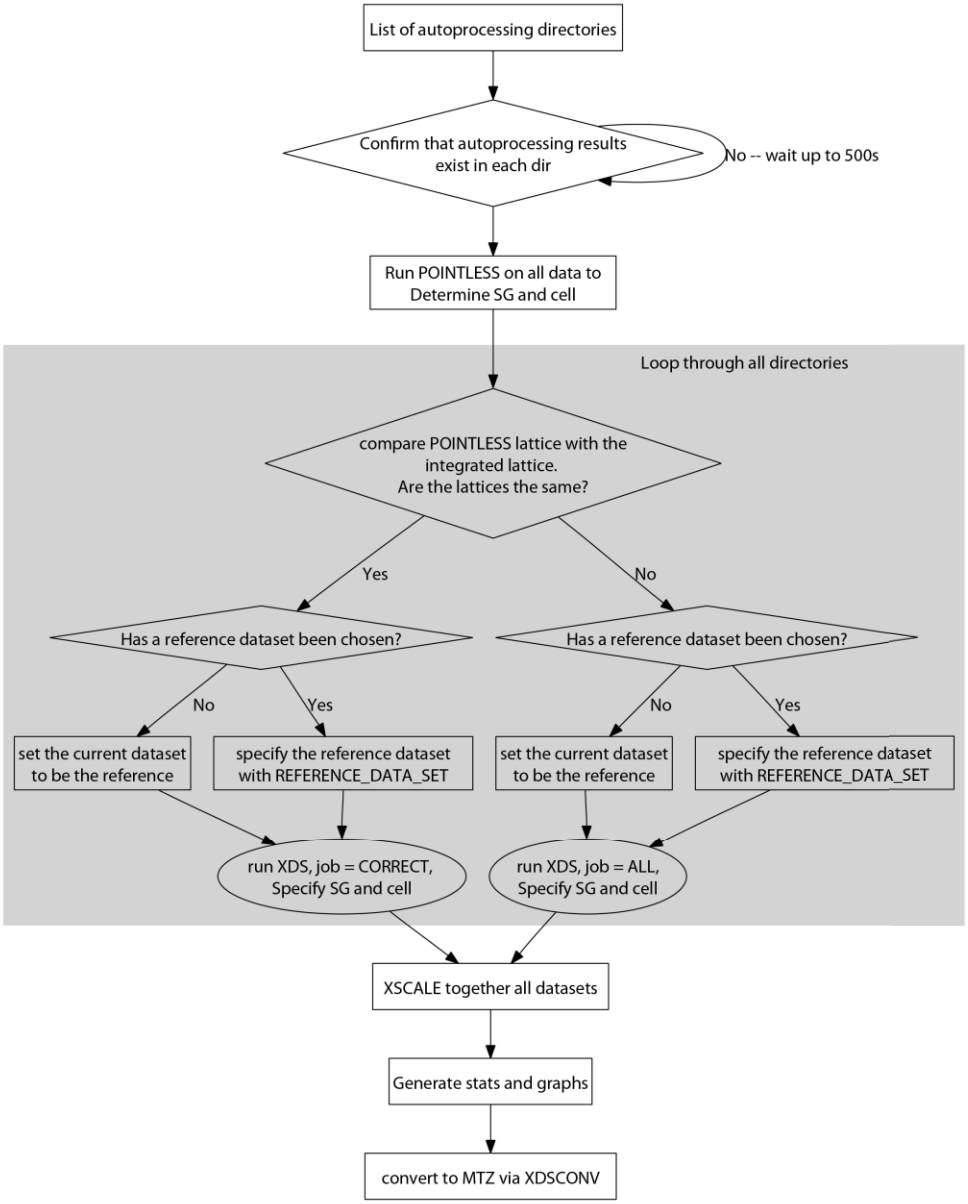
ESRF

- Matias Guijarro
- Sasha Popov
- Marjolaine Bodin
- Solange Delageniere
- Stephanie Monaco
- Daniele de Sanctis
- Thomas Boeglin
- Gordon Leonard
- Olof Svensson
- Marcus Oscarsson

EMBL

- Hassan Belrahili
- Florent Cipriani
- Josan Marquez
- Ulrich Zander
- Andrew McCarthy

Scheme


List of autoprocessing directories
Confirm that autoprocessing results exist in each dir
No -- wait up to 500s
Run POINTLESS on all data to Determine SG and cell
Loop through all directories
compare POINTLESS lattice with the integrated lattice. Are the lattices the same?
Yes
No
Has a reference dataset been chosen?
Has a reference dataset been chosen?
No
Yes
No
Yes
set the current dataset to be the reference
specify the reference dataset with REFERENCE_DATA_SET
set the current dataset to be the reference
specify the reference dataset with REFERENCE_DATA_SET
run XDS, job = CORRECT, Specify SG and cell
run XDS, job = ALL, Specify SG and cell
XSCALE together all datasets
Generate stats and graphs
convert to MTZ via XDSCONV