

You can't get the staff - an electronic alternative...

(an introduction to xia2)

Graeme Winter

Diamond Light Source

CCP4 Study Weekend 2012



Overview

- Background
- What is xia2?
- What does it do and how do I use it?
- What decisions does it make?
- Conclusions / summary



Before we start...

- No MOSFLM¹, XDS², SCALA³, CCP4⁴
- → no xia2
- No LABELIT⁵, CCTBX⁶, POINTLESS⁷, etc.
- → harder to write xia2, less reliable

¹A.G.W. Leslie, Acta Cryst. (2006) D62, 48-57

²W. Kabsch, Acta Cryst. (2010) D66, 125-132

³P. Evans, Acta Cryst. (2006) D62, 72-82

⁴CCP4, Acta Cryst. (1994) D50, 760-763

⁵N.K. Sauter et al. J. Appl. Cryst. (2004) 37, 399-409

⁶R.W. Grosse-Kunstleve et al. J. Appl. Cryst. (2002) 35, 126-136

⁷P. Evans, Acta Cryst. (2006) D62, 72-82



Acknowledgements

- Andrew Leslie, Harry Powell, Phil Evans, Wolfgang Kabsch, Kay Diederichs, Nick Sauter, Ralf Grosse-Kunstleve
- Alun Ashton, Dave Stuart, Diamond beamline staff, Miroslav Papiz, Steve Prince, Colin Nave, xia2 users, providers of test data (esp. JCSG)
- Funding from Diamond Light Source, BBSRC e-Science e-HTPX project, BioXHit

Background (2005)

- Comprehensive, trusted software available
- Background of strong publications (esp. CCP4 study weekends)
- Massive advances in computing
- New synchrotron for UK



Background (2005)

- Comprehensive, trusted software available
- Background of strong publications (esp. CCP4 study weekends)
- Massive advances in computing
- New synchrotron for UK
- → a great time to develop automated data reduction



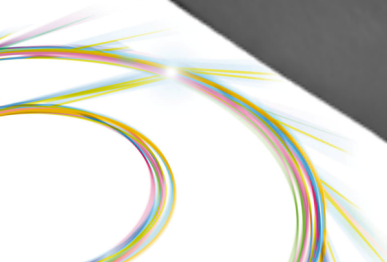
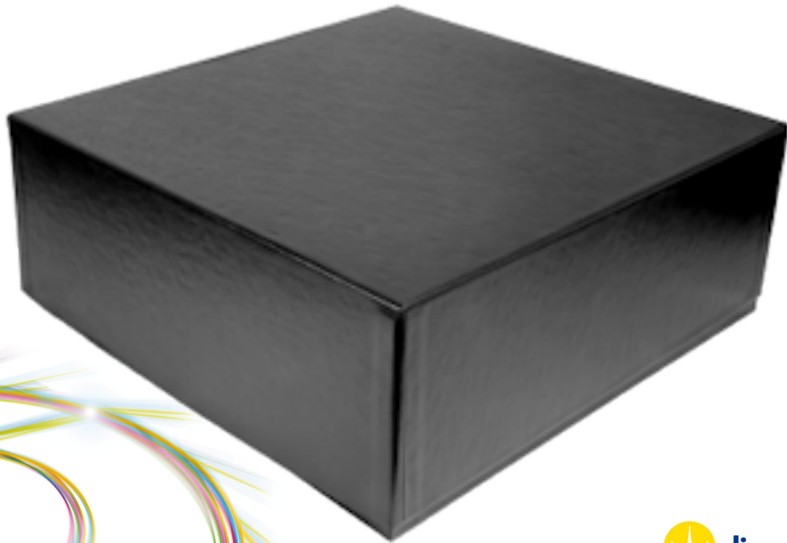
Background (2005)

- Comprehensive, trusted software available
- Background of strong publications (esp. CCP4 study weekends)
- Massive advances in computing
- New synchrotron for UK
- → a great time to develop automated data reduction
- *Also told that this is impossible and a waste of time*

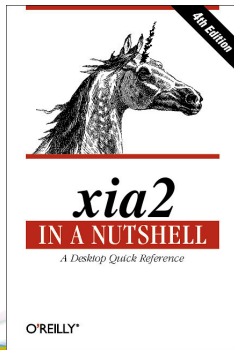
What is xia2?



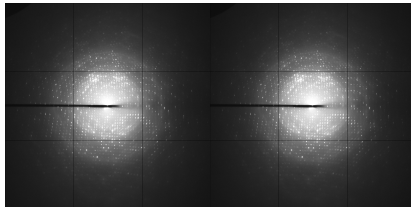
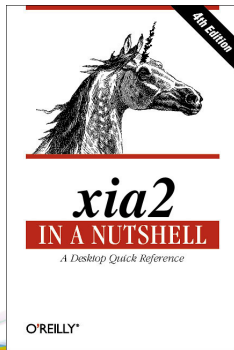
What is xia2?



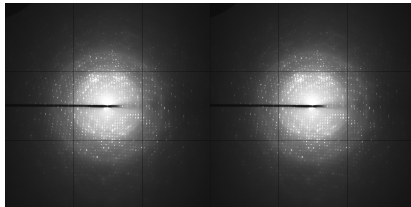
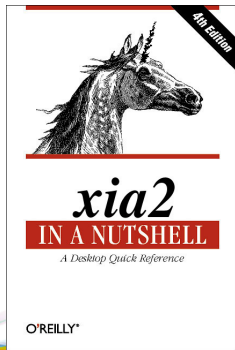
What is xia2?



What is xia2?



What is xia2?



→ $HKLI\sigma_I$

What is xia2?

- An *expert* system to perform diffraction data and processing on your behalf using your software



What is xia2?

- An *expert* system to perform diffraction data and processing on your behalf using your software
- A system which can correctly handle multi-pass, multi-wavelength data sets



What is xia2?

- An *expert* system to perform diffraction data and processing on your behalf using your software
- A system which can correctly handle multi-pass, multi-wavelength data sets
- *Not* a data processing package



Why “you can’t get the staff?”

- 12 datasets / hour possible
- Limited help
- Human endurance
- Intended xia2 as tool to delegate data processing to



Why “you can’t get the staff?”

- 12 datasets / hour possible
- Limited help
- Human endurance
- Intended xia2 as tool to delegate data processing to



Why is this useful?

- Second opinion
- Allows you to focus on problem cases
- Help busy / novice users
- Provides access to other tools
- Reproducible processing



Using xia2

xia2 -2d /here/are/my/data

- *or* -

xia2 -3d /here/are/my/data



Command line



Command line



```
File Edit View Bookmarks Settings Help
P 43 2 2
Unit cell:
42.180 42.180 39.240
90.000 90.000 90.000
mtz format:
scaled reflections: /home/gw56/3qrn/3dai/sub/DataFiles/AUTOMATIC_DEFAULT_free.at
sca format:
scaled reflections (SAD): /home/gw56/3qrn/3dai/sub/DataFiles/AUTOMATIC_DEFAULT_s
scaled.sca
sca unmerged format:
scaled reflections (SAD): /home/gw56/3qrn/3dai/sub/DataFiles/AUTOMATIC_DEFAULT_s
scaled.unmerged.sca
Processing took 00h 13m 02s
XIA2 used... ccp4 pointless xds xia2
Here are the appropriate citations (BIBTeX in xia-citations.bib.)
(1994) Acta Crystallogr. D 50: 760--763
Evans, Philip (2006) Acta Crystallographica Section D 62: 72--82
Kabsch, W. (1988) Journal of Applied Crystallography 21: 67--72
Kabsch, W. (1988) Journal of Applied Crystallography 21: 916--924
Kabsch, W. (1993) Journal of Applied Crystallography 26: 795--800
Winter, G. (2010) Journal of Applied Crystallography 43
Status: normal termination
gramee@iamr15191:~/processing/ccp4-study-weekend-2012/3qrn/3dai/sub>
```

sub : bash

Not GUI



Options (1)

- -atom X - tell xia2 to separate anomalous pairs
- -2d - tell xia2 to use MOSFLM and SCALA
- -3d - tell xia2 to use XDS and XSCALE
- -3dii - tell xia2 to use XDS and XSCALE, indexing with all images



How does it figure what to do?

- Read all of the image headers then



How does it figure what to do?

- Read all of the image headers then
- Organise these into sweeps then



How does it figure what to do?

- Read all of the image headers then
- Organise these into sweeps then
- Organise these into wavelengths then



How does it figure what to do?

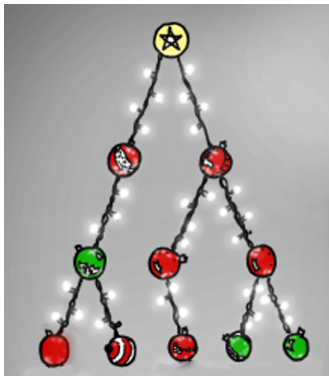
- Read all of the image headers then
- Organise these into sweeps then
- Organise these into wavelengths then
- Assign all of these wavelengths to a crystal



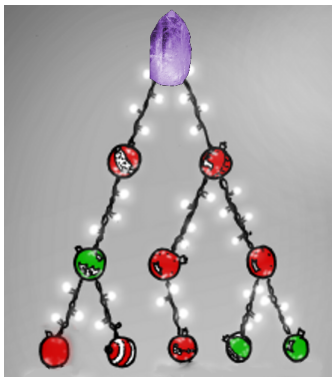
How does it figure what to do?



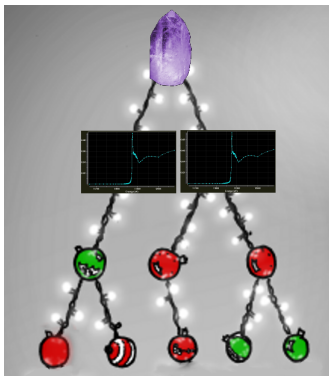
How does it figure what to do?



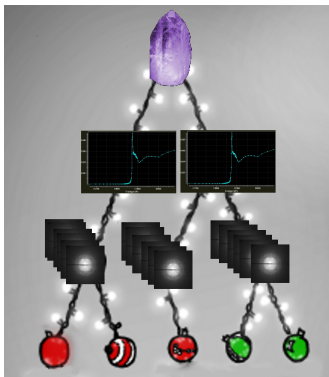
How does it figure what to do?



How does it figure what to do?



How does it figure what to do?



What the program sees (automatic.xinfo)

```
BEGIN PROJECT AUTOMATIC
BEGIN CRYSTAL DEFAULT
BEGIN HA_INFO
ATOM Ba
END HA_INFO
BEGIN WAVELENGTH SAD
WAVELENGTH 0.979500
END WAVELENGTH SAD
BEGIN SWEEP SWEEP1
WAVELENGTH SAD
DIRECTORY /dls/i02/data/2011/mx1234-5
IMAGE K5_M1S3_3_001.img
START_END 1 450
END SWEEP SWEEP1
END CRYSTAL DEFAULT
END PROJECT AUTOMATIC
```

Understanding the experiment

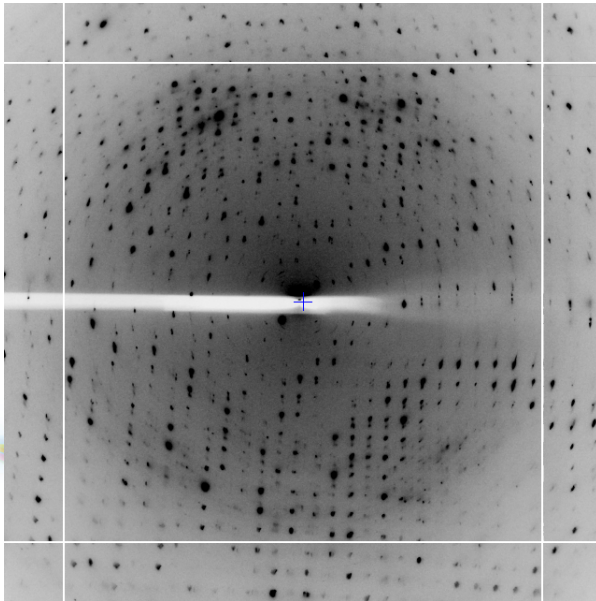
- SWEEP: one “scan” - basic unit of indexing / integration
- WAVELENGTH: container of SWEEPs
- WAVELENGTH: all H K L observations merged
- WAVELENGTH: CCP4 MTZ dataset
- CRYSTAL: contains WAVELENGTHs
- CRYSTAL: all data - basic unit of scaling

Example: 3QRN⁸

- Data recorded at Diamond I02
- DNA / ligand complex
- Demonstrates:
 - Radiation damage
 - Heavy atom
 - Resolution limits
- Better sample used for deposition

⁸J.P. Hall et al., Proc. Natl. Acad. Sci. USA 2011 108 (43) 17573-17574

Example: data



Example command line

```
xia2 -3d -atom Ba /dls/i02/data/...
```



Output

- xia2.txt: everything you should read - including program citations
- xia2-debug.txt: everything you probably shouldn't
- LogFiles: you should look at these
- DataFiles: MTZ + ersatz scalepack



Example results

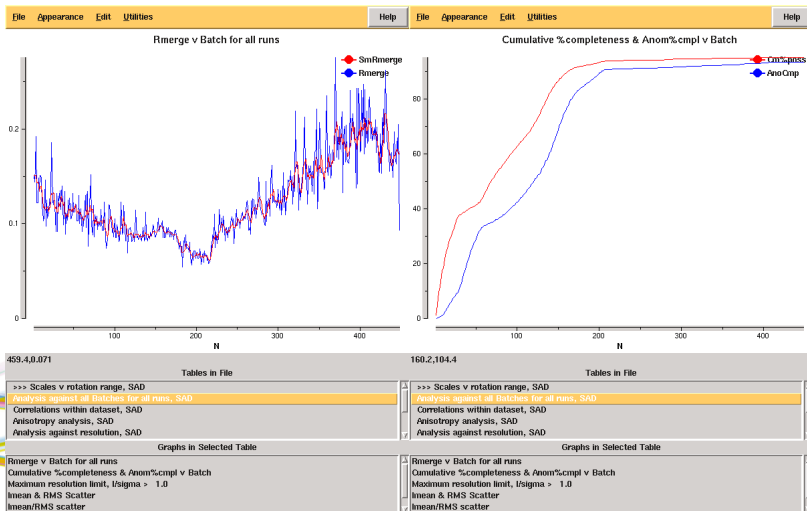
High resolution limit	1.25	6.45	1.25
Low resolution limit	18.85	18.85	1.27
Completeness	95.2	60.1	70.2
Multiplicity	12.2	8.4	4.8
I/sigma	12.3	18.5	2.6
Rmerge	0.113	0.096	0.564
Rmeas(I)	0.129	0.118	0.633
Rmeas(I+/-)	0.121	0.105	0.679
Rpim(I)	0.034	0.038	0.267
Rpim(I+/-)	0.043	0.041	0.368
Wilson B factor	12.131		
Anomalous completeness	93.3	52.6	58.0
Anomalous multiplicity	6.4	5.0	2.0
Anomalous correlation	0.544	0.791	-0.297
Anomalous slope	1.085	0.000	0.000
Total observations	118588	529	1634
Total unique	9749	63	337

Example command line (I cheated, I like AIMLESS)

xia2 -3da ...



LogFiles/*aimless.log



What to do next?

- Edit automatic.xinfo
- Only process first 200 frames



Modify automatic.xinfo → modified.xinfo

```
BEGIN PROJECT AUTOMATIC
BEGIN CRYSTAL DEFAULT
BEGIN HA_INFO
ATOM Ba
END HA_INFO
BEGIN WAVELENGTH SAD
WAVELENGTH 0.979500
END WAVELENGTH SAD
BEGIN SWEEP SWEEP1
WAVELENGTH SAD
DIRECTORY /dls/i02/data/2011/mx1234-5
IMAGE K5_M1S3_3_001.img
START_END 1 200 ! THIS WAS 450
END SWEEP SWEEP1
END CRYSTAL DEFAULT
END PROJECT AUTOMATIC
```

Running again

```
xia2 -3d -xinfo modified.xinfo
```



Example results II

High resolution limit	1.22	6.34	1.22
Low resolution limit	19.62	19.62	1.24
Completeness	86.9	49.1	37.8
Multiplicity	5.3	4.9	1.7
I/sigma	20.1	37.0	2.3
Rmerge	0.036	0.020	0.355
Rmeas(I)	0.060	0.038	0.448
Rmeas(I+/-)	0.043	0.023	0.491
Rpim(I)	0.023	0.014	0.297
Rpim(I+/-)	0.022	0.011	0.339
Wilson B factor	10.70		
Anomalous completeness	77.7	41.0	18.3
Anomalous multiplicity	2.7	3.5	0.5
Anomalous correlation	0.779	0.931	0.000
Anomalous slope	1.553	0.000	0.000
Total observations	50875	272	342
Total unique	9552	55	199

Resolution

- Data incomplete at high resolution
- Add RESOLUTION to xinfo file (in either SWEEP or WAVELENGTH
- Add -resolution to the command line



Options (2)

- -xinfo modified.xinfo - use specific input file
- -image /path/to/an/image.img - process specific scan
- -spacegroup spacegroup_name - set the spacegroup, e.g. P21
- -cell a,b,c, α , β , γ - set the cell constants
- -small_molecule - don't run things like TRUNCATE



What did it do? and why?

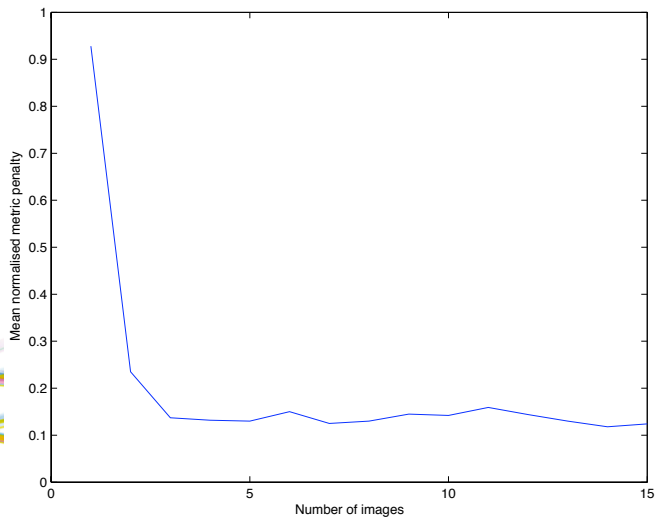


Indexing

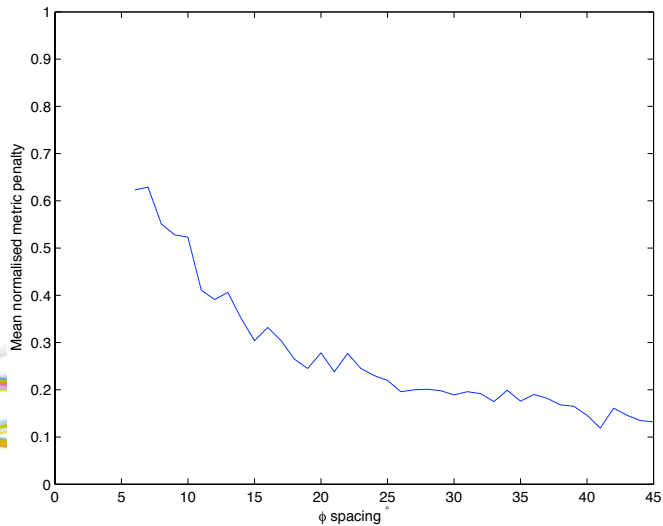
- Initial indexing with LABELIT from 3 images⁹
- Refine results with XDS indexing
- Use data based on general analysis @ 0, 45, 90 degrees

⁹This is *not* good for small molecule data

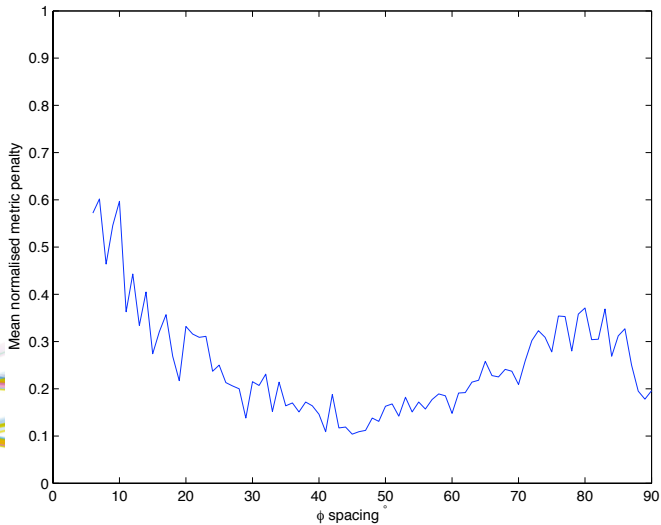
Indexing



Indexing

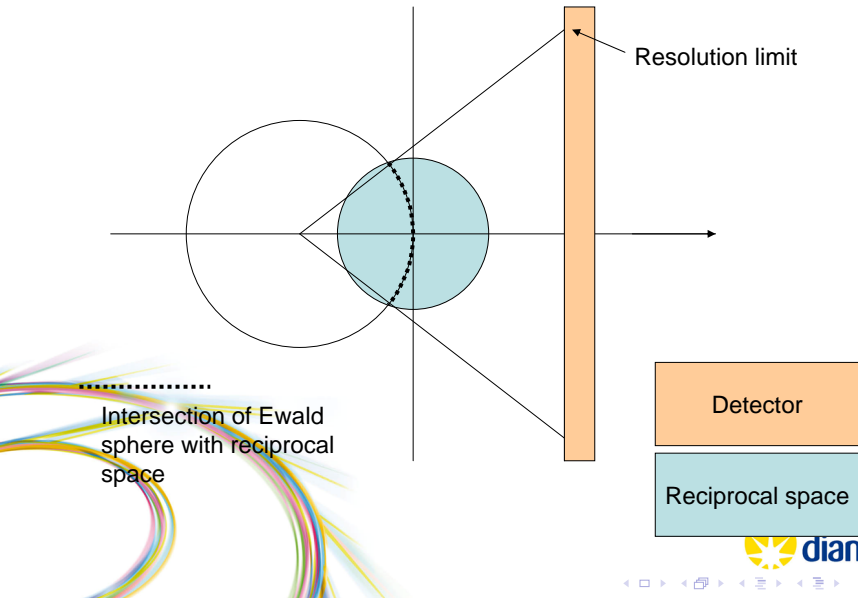


Indexing

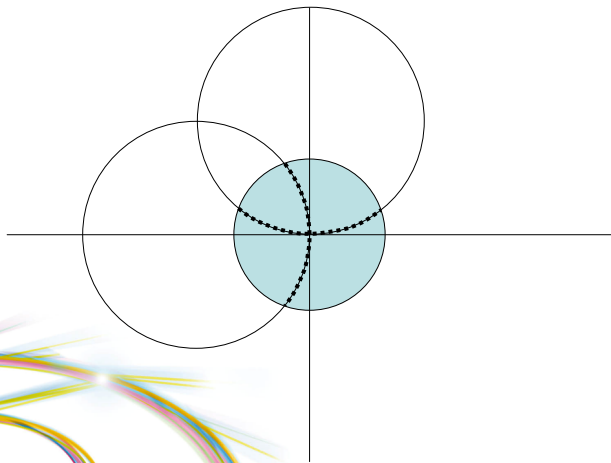


diamond

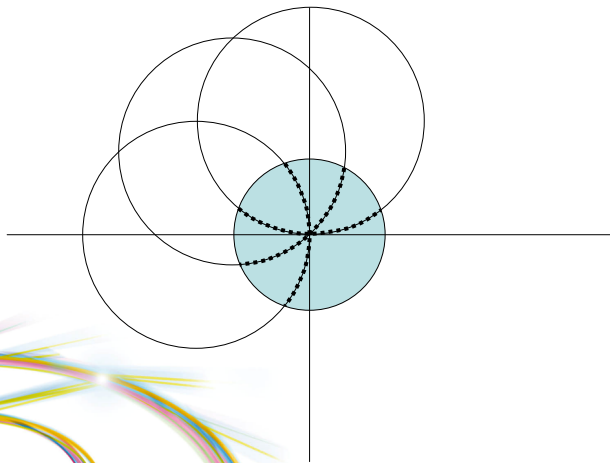
Indexing



Indexing



Indexing



Integration

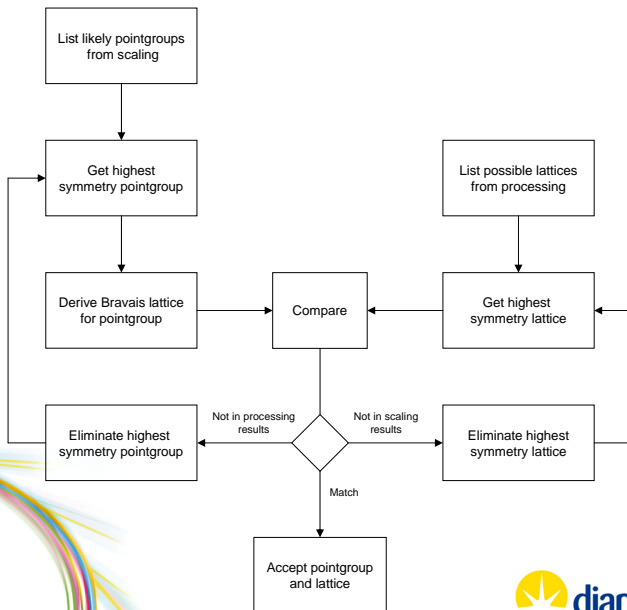
- Integrate with lattice constraints applied
- Integrate to corners of detector
- If good reason, repeat integration e.g. with results of postrefinement
- Perform postrefinement in P1, assumed lattice - may reject lattice, feed back to indexing
- At the end of this we have LATTICE
- If XDS, includes iterative elimination of outliers in CORRECT step

Scaling

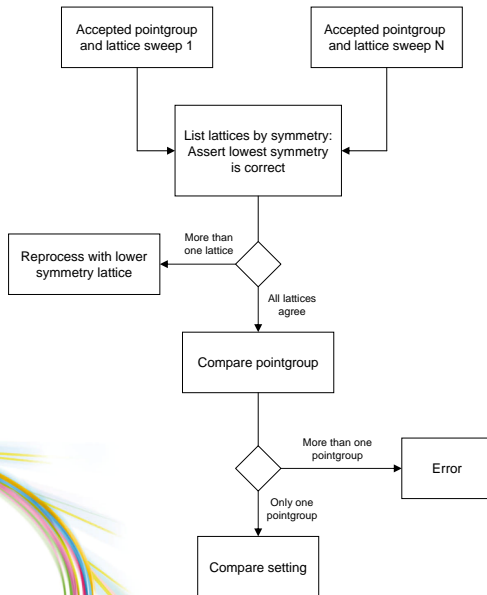
- Compare results of pointless with remaining allowed lattices:
 - If agree, proceed
 - If lattice not allowed, consider next solution
 - If solution lower symmetry than lattice, reject and return to indexing
- Ensure conclusions consistent
- Now have correct LAUE GROUP
- Ensure consistent setting / origin choice
- Place data into data collection order
- Analyse absences to decide likely SPACE GROUPs
- Decide scaling model¹⁰

¹⁰For XDS use not corrections in CORRECT, apply all corrections in XSCALE ↻ 🔍 🔗

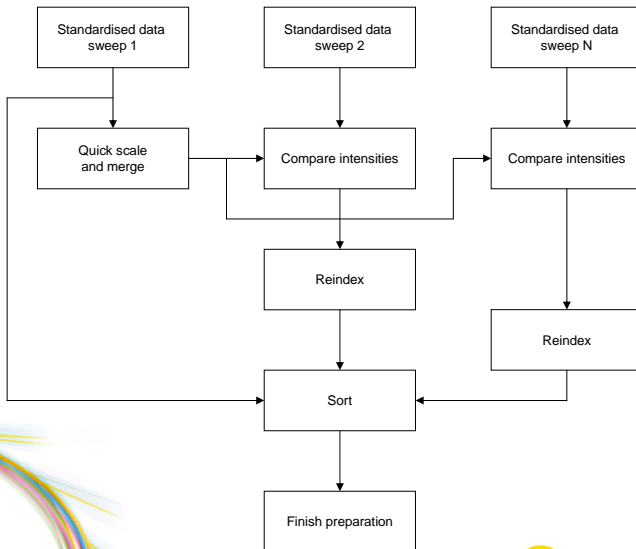
Scaling



Scaling



Scaling



Merging and analysis

- If using XDS for integration and XSCALE for scaling, data still merged with SCALA / AIMLESS
- Resolution limits calculated from the intensities, not program output
- “Downstream” analysis (e.g. TRUNCATE and SFCHECK) identical



Merging and analysis

- If using XDS for integration and XSCALE for scaling, data still merged with SCALA / AIMLESS
- Resolution limits calculated from the intensities, not program output
- “Downstream” analysis (e.g. TRUNCATE and SFCHECK) identical
- Working on scaling data direct from XDS with AIMLESS

Decision making



Indexing - LABELIT

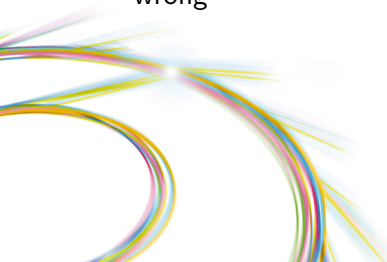
Solution	Metric fit	rmsd	#spots	crystal_system	unit_cell
:) 9	0.2097 dg	0.327	533	tetragonal tP	42.32 42.3
:) 8	0.2097 dg	0.364	541	orthorhombic oP	39.29 42.2
:) 7	0.2097 dg	0.300	519	monoclinic mP	39.26 42.3
:) 6	0.1950 dg	0.299	523	monoclinic mP	39.26 42.3
:) 5	0.1307 dg	0.411	523	orthorhombic oC	59.71 59.9
:) 4	0.1307 dg	0.412	524	monoclinic mC	59.91 59.7
:) 3	0.0937 dg	0.429	524	monoclinic mC	59.71 59.9
:) 2	0.1010 dg	0.298	512	monoclinic mP	42.27 39.3
:) 1	0.0000 dg	0.291	509	triclinic aP	39.31 42.2

Indexing - IDXREF

*	31	aP	0.0	39.1	42.1	42.1	90.0	90.
*	44	aP	0.1	39.1	42.1	42.1	90.0	90.
*	34	mP	0.7	39.1	42.1	42.1	90.0	90.
*	20	mC	0.7	59.6	59.6	39.1	90.1	90.
*	33	mP	0.8	39.1	42.1	42.1	90.0	90.
*	25	mC	0.9	59.6	59.6	39.1	89.9	90.
*	35	mP	1.7	42.1	39.1	42.1	90.0	90.
*	23	oC	1.7	59.6	59.6	39.1	89.9	90.
*	32	oP	1.8	39.1	42.1	42.1	90.0	90.
*	21	tP	1.9	42.1	42.1	39.1	90.0	90.
	10	mC	79.5	57.5	57.4	42.1	90.0	90.
	13	oC	79.9	57.4	57.5	42.1	90.0	90.
	14	mC	79.9	57.4	57.5	42.1	90.0	90.

Testing lattice choice

- Perform postrefinement (MOSFLM and XDS) in P1 and putative lattice
- Compare R.M.S. deviation of observed / predicted centres
- Results comparable → lattice probably OK
- Results worse with lattice constraints → lattice probably wrong



Testing lattice choice 1

REFINED PARAMETERS: DISTANCE BEAM ORIENTATION CELL AXIS
USING 27389 INDEXED SPOTS
STANDARD DEVIATION OF SPOT POSITION (PIXELS) 1.28
STANDARD DEVIATION OF SPINDLE POSITION (DEGREES) 0.23

...

UNIT CELL PARAMETERS	42.180	42.183	39.236	90.002	89
E.S.D. OF CELL PARAMETERS	1.8E-02	4.3E-02	1.5E-02	1.4E-02	1.0E
SPACE GROUP NUMBER	1				

Testing lattice choice 2

REFINED PARAMETERS: DISTANCE BEAM ORIENTATION CELL AXIS
USING 27378 INDEXED SPOTS
STANDARD DEVIATION OF SPOT POSITION (PIXELS) 1.29
STANDARD DEVIATION OF SPINDLE POSITION (DEGREES) 0.23

...

UNIT CELL PARAMETERS	42.187	42.187	39.242	90.000	90.000
E.S.D. OF CELL PARAMETERS	1.6E-02	1.6E-02	1.2E-02	0.0E+00	0.0E+00
SPACE GROUP NUMBER	75				

Lattice observations

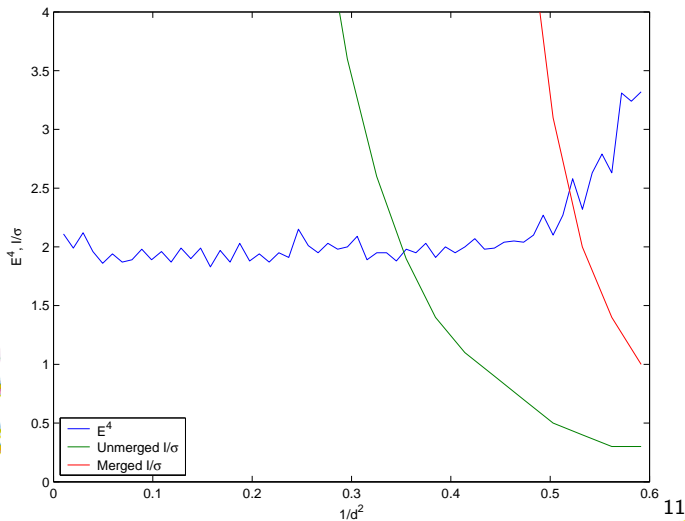
- Selecting lattice from indexing safe, as tested and challenged
- However strong argument for performing all processing in P1:
 - Processing only performed once
 - Incorrect constraints cannot break things
 - Results generally comparable
- This is on the to-do list...

Resolution limits - default criteria

- Merged $\frac{I}{\sigma_I} > 2$
- Unmerged $\frac{I}{\sigma_I} > 1$
- Control with -misigma, -isigma



Resolution limits



11



diamond

Which options work best?

- It depends ...
- ... try for yourself!



Which options work best?

- It depends ...
- ... try for yourself!
- Sometimes -2d (MOSFLM / SCALA) works better, sometimes -3d (XDS etc.)



Which options work best?

- It depends ...
- ... try for yourself!
- Sometimes -2d (MOSFLM / SCALA) works better, sometimes -3d (XDS etc.)
- Run both - compare results, make up your own mind



Which options work best?

- It depends ...
- ... try for yourself!
- Sometimes -2d (MOSFLM / SCALA) works better, sometimes -3d (XDS etc.)
- Run both - compare results, make up your own mind
- Hint for small molecule: -3dii -small_molecule

Conclusions

- System available which can reduce your data on your behalf
- Relies on your software: MOSFLM / LABELIT / CCP4 / XDS
- Handles complex strategies so use them
- Works on Windows / OS X / Linux / laptop / workstation / cluster



Conclusions

- Best way to learn data reduction is to teach it
- Computer is very dim but diligent pupil
- Have a go yourself, or feel free to contribute to xia2



Getting xia2

- Blog: xia2.blogspot.com
- Code: xia2.sf.net
- List: xia2-list@lists.sourceforge.net

