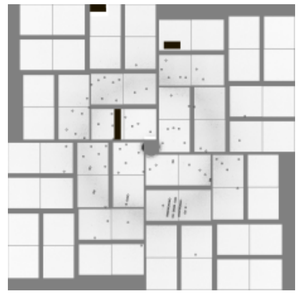
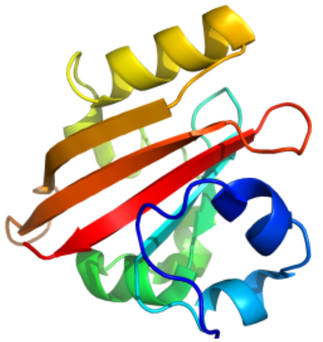
Data processing

Theory of data processing for serial crystallography





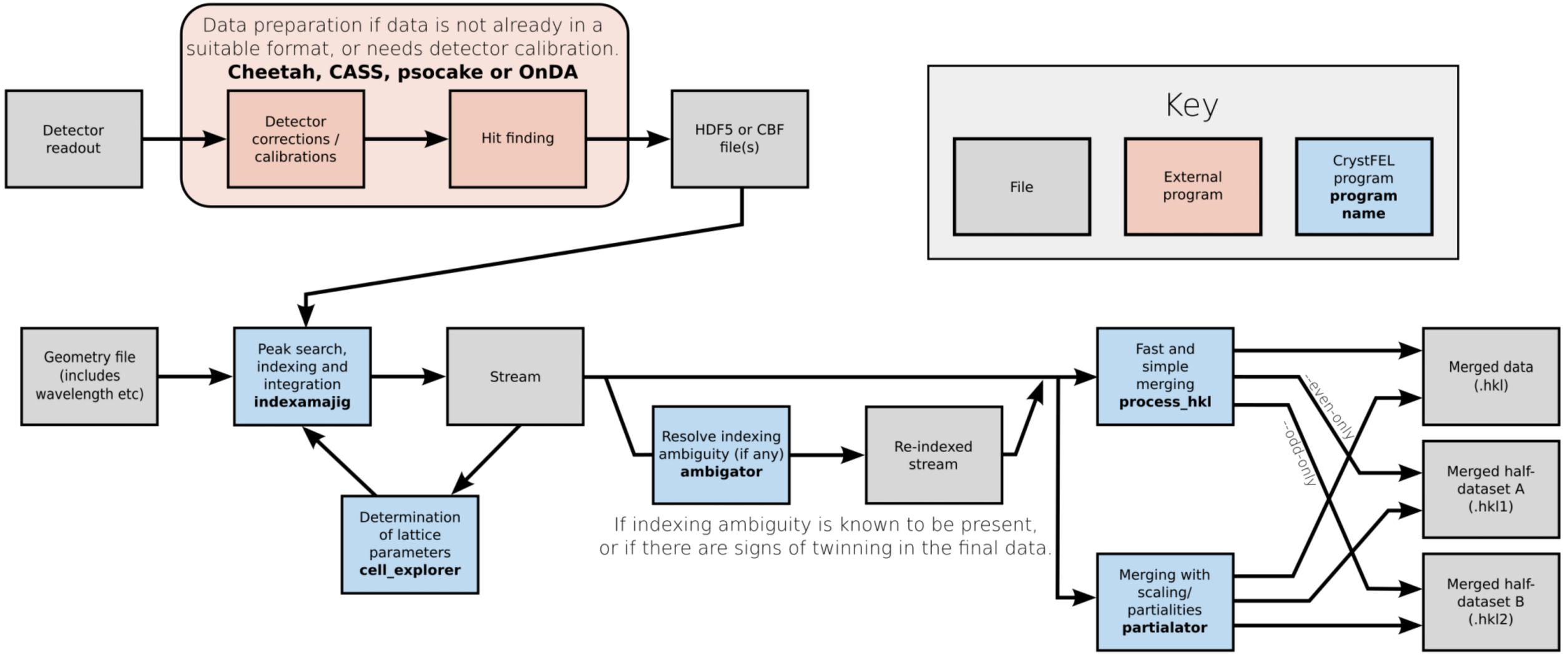
Thomas White
Gothenburg CrystFEL workshop
27. January 2020

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

CFEL SCIENCE

DESY.

Data processing pipeline

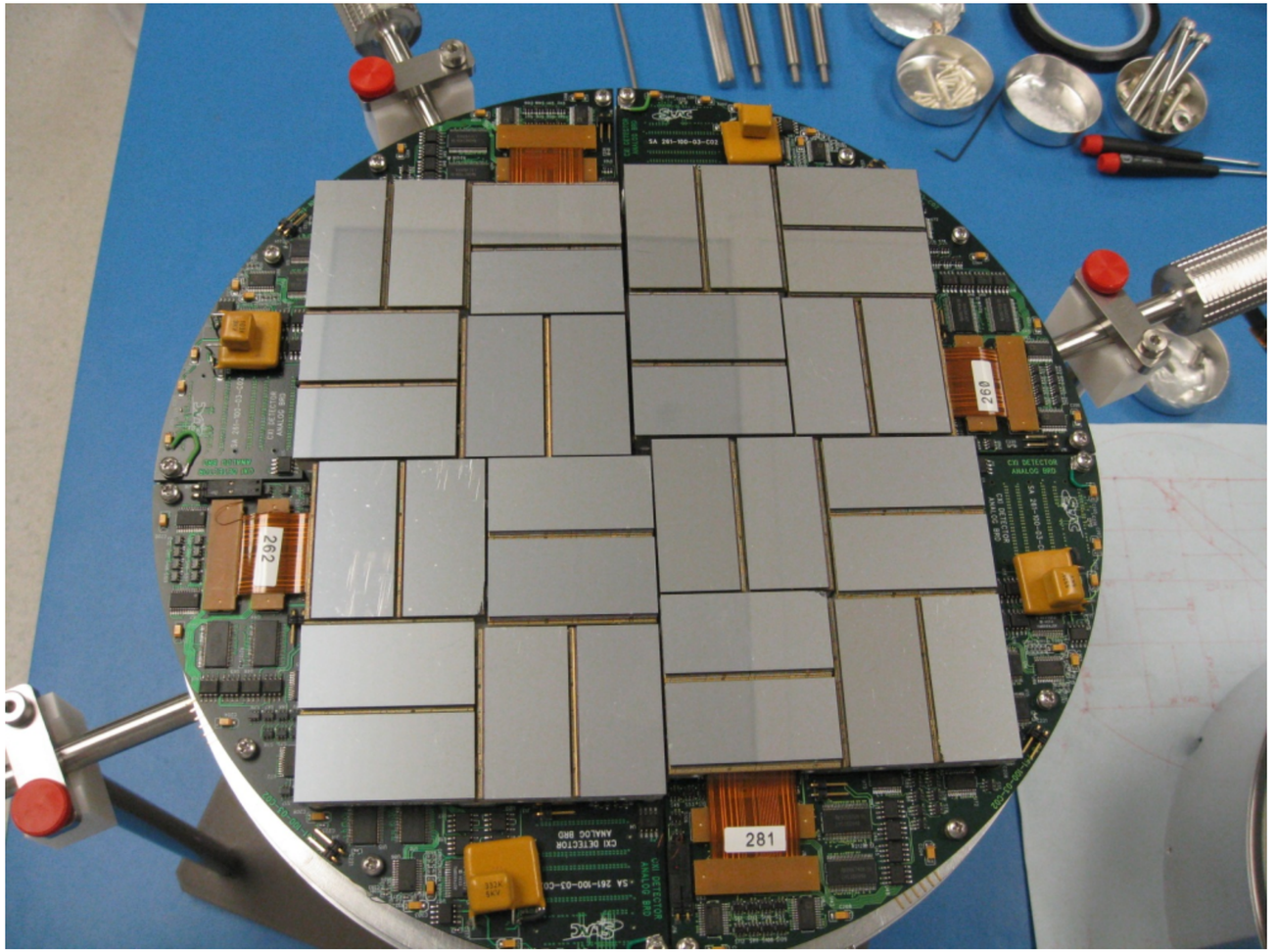


The past and future

The oscillation method (Arndt & Wonacott, 1977) is now generally accepted as the most efficient means for photographic data collection from crystals with unit-cell dimensions greater than about 100 Å. For very large unit cells (average dimensions greater than about 250 Å), the usual method must be modified to realize its full efficiency – in particular, to cope with situations where only one good photograph, instead of a series of contiguous ones, can be taken within the lifetime of a crystal. Essentially, the resulting 'one crystal–one photograph' condition requires a different treatment of partially recorded reflections. This paper describes general aspects of the new procedure and gives a detailed account of its application to data collection from crystalline tomato bushy stunt virus (TBSV: space group $I23$, $a = 383{\cdot}2$ Å) to 2·9 Å resolution.

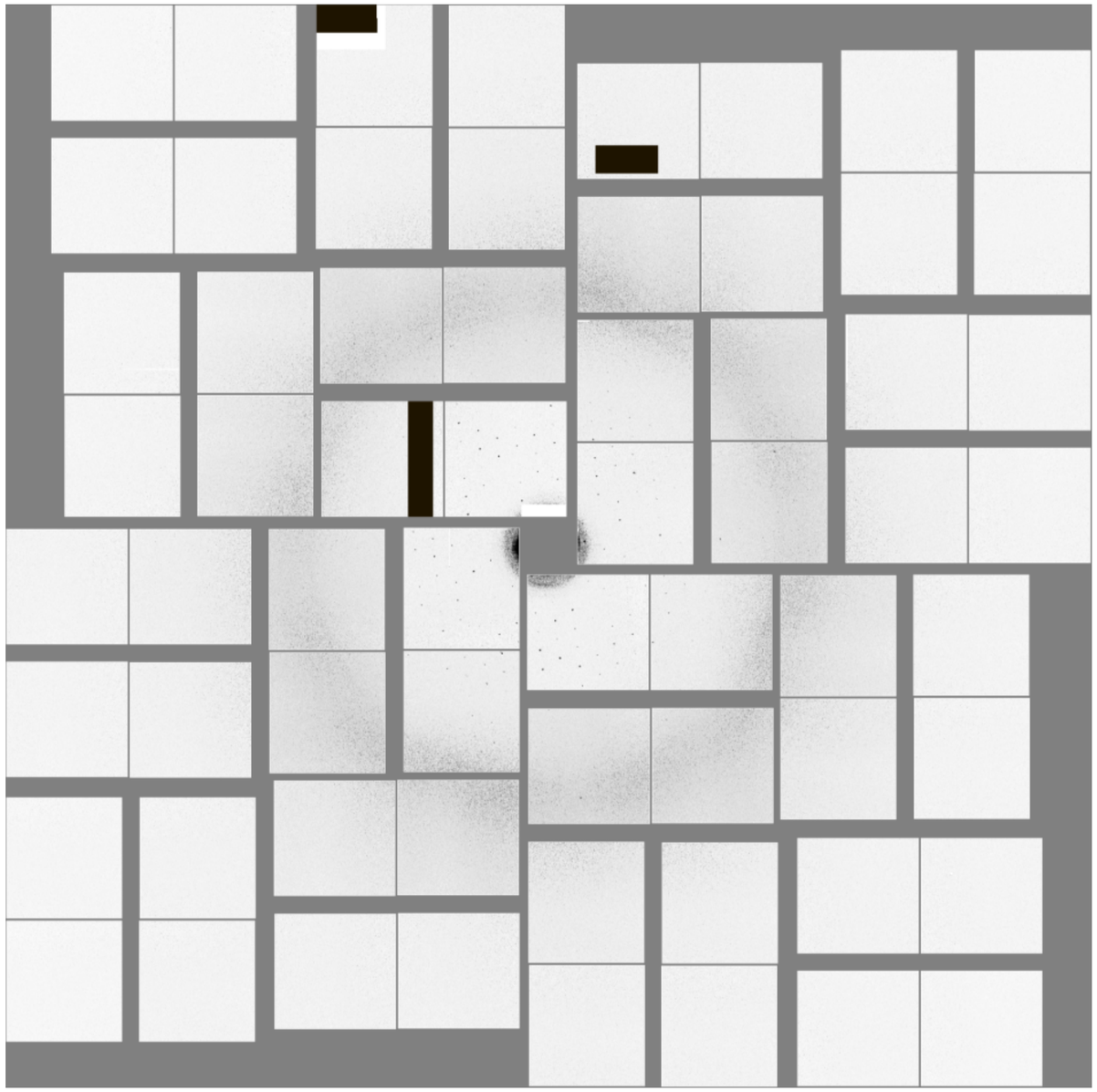
Winkler, Schutt and Harrison, Acta Cryst. A35 (**1979**), p901-911

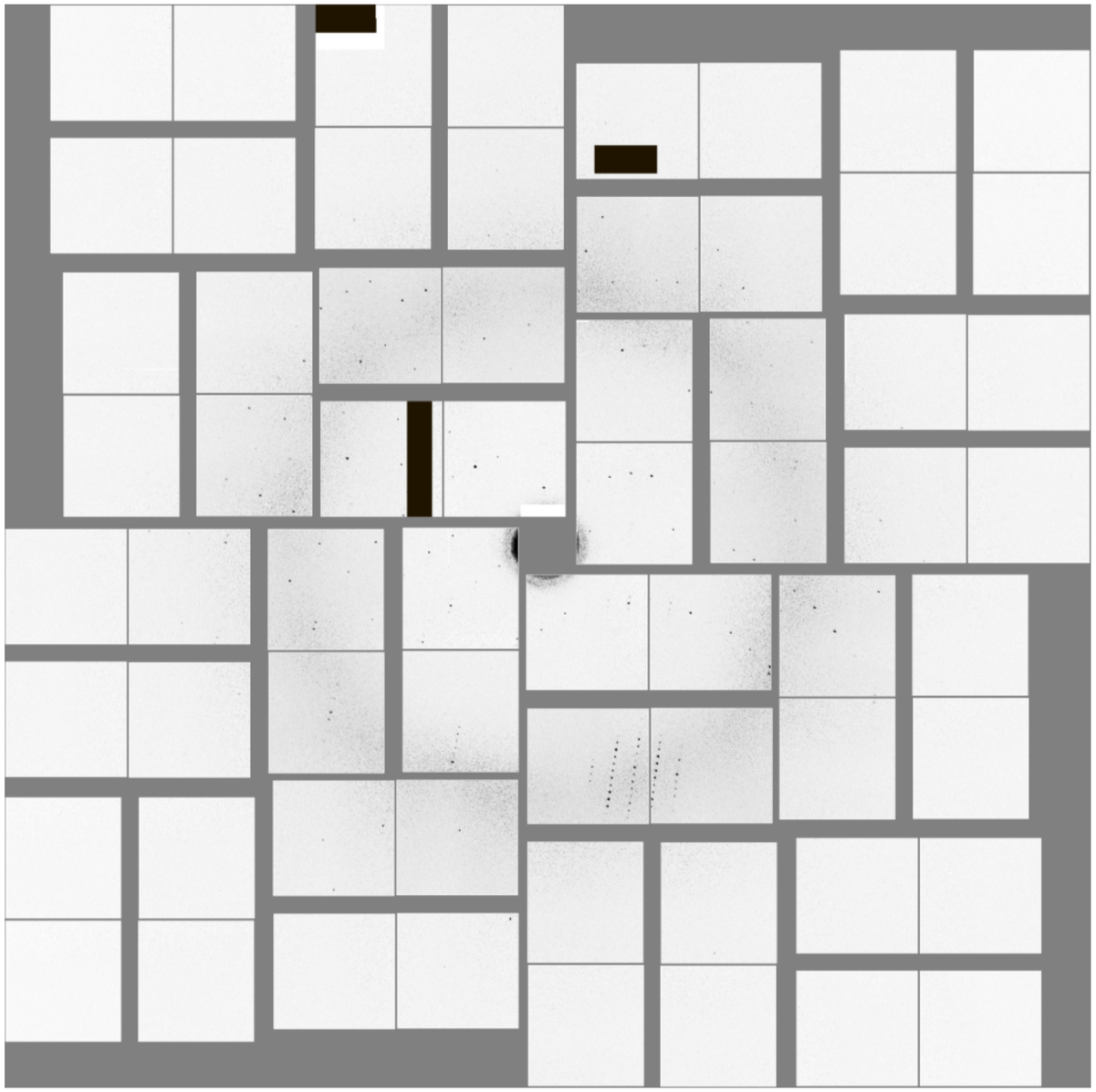
Multi-panel detector geometry



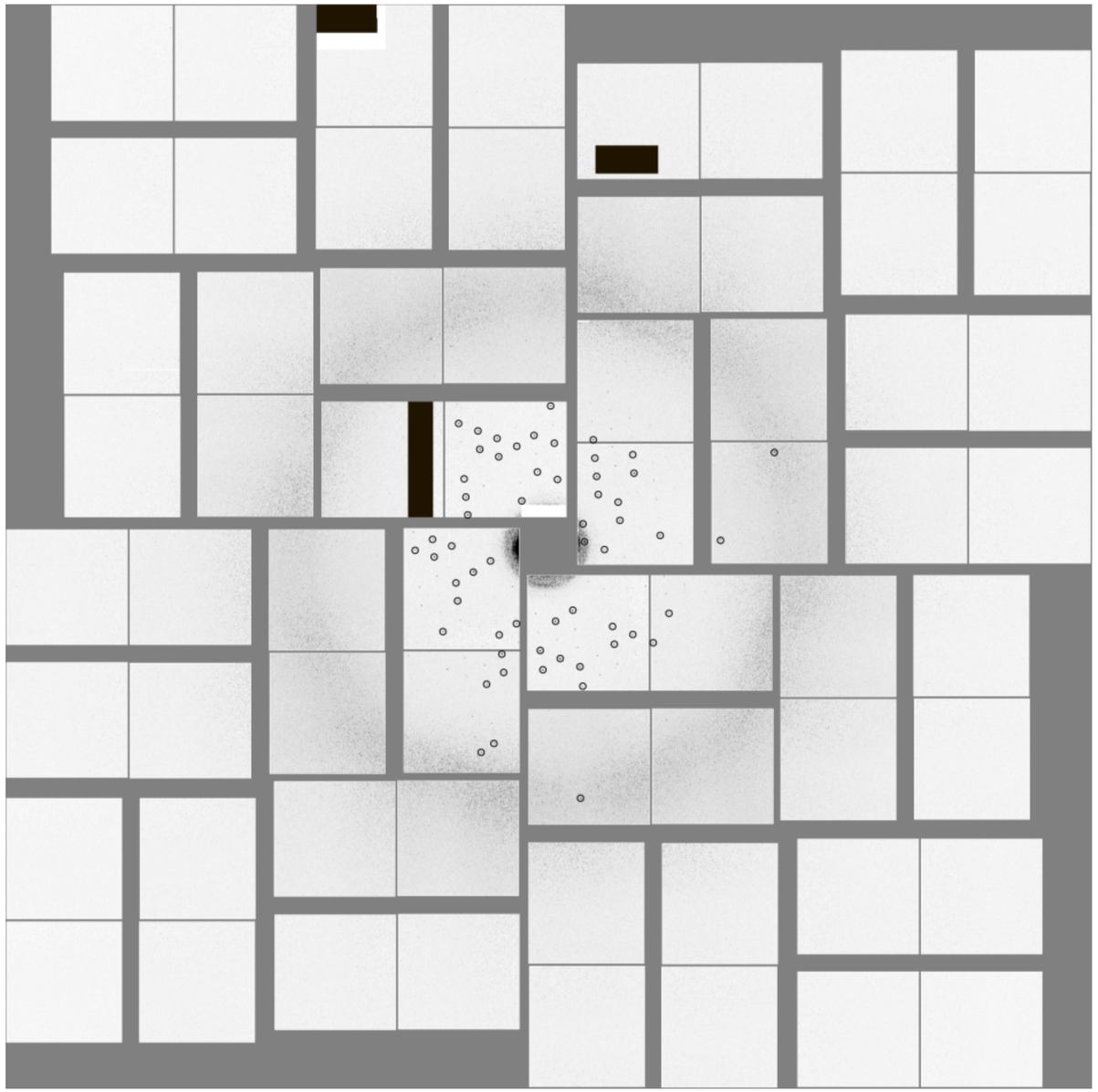
Cornell-SLAC Pixel Array Detector ("CSPAD")
Image: SLAC National Accelerator Laboratory

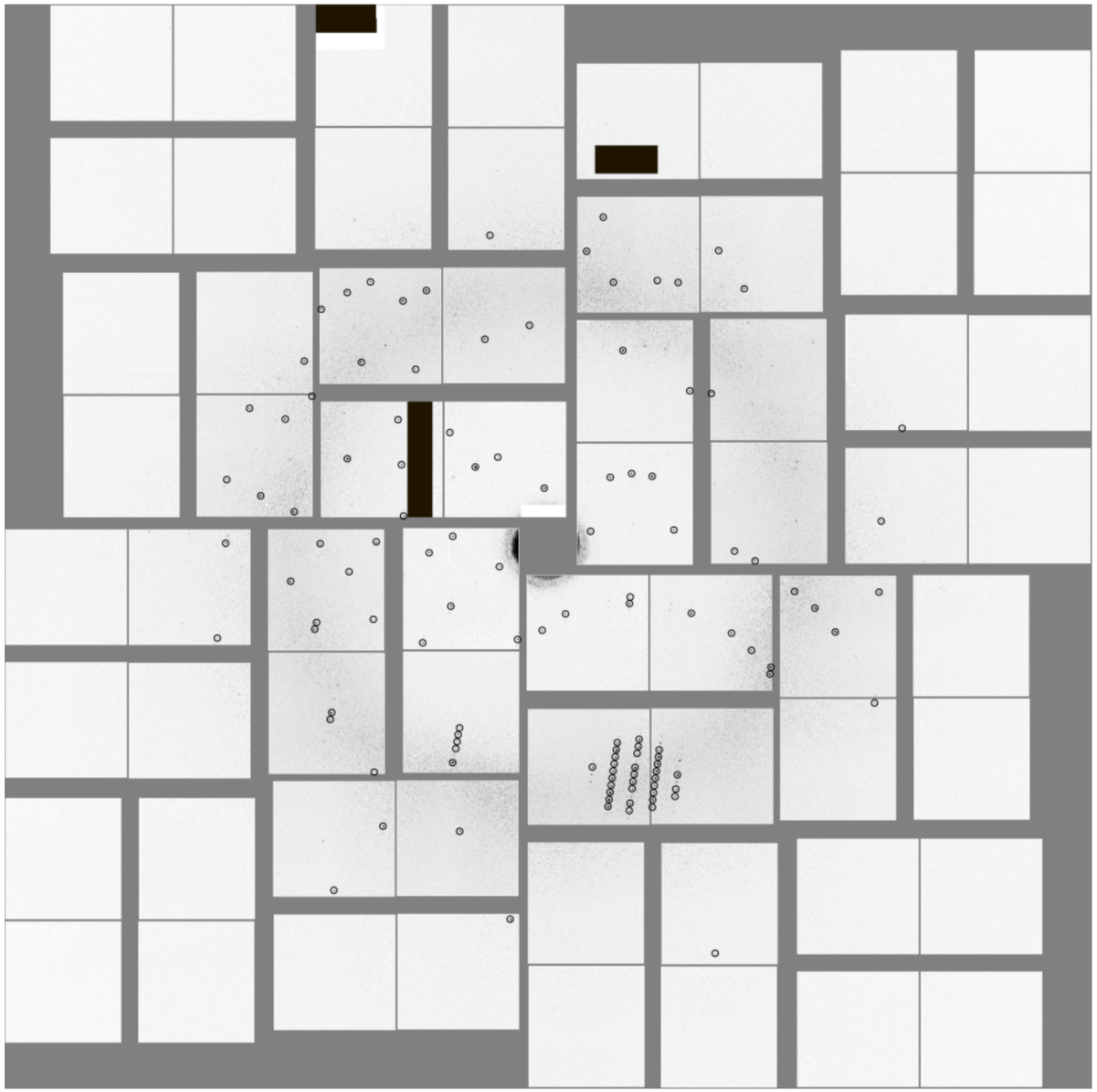
Diffraction patterns





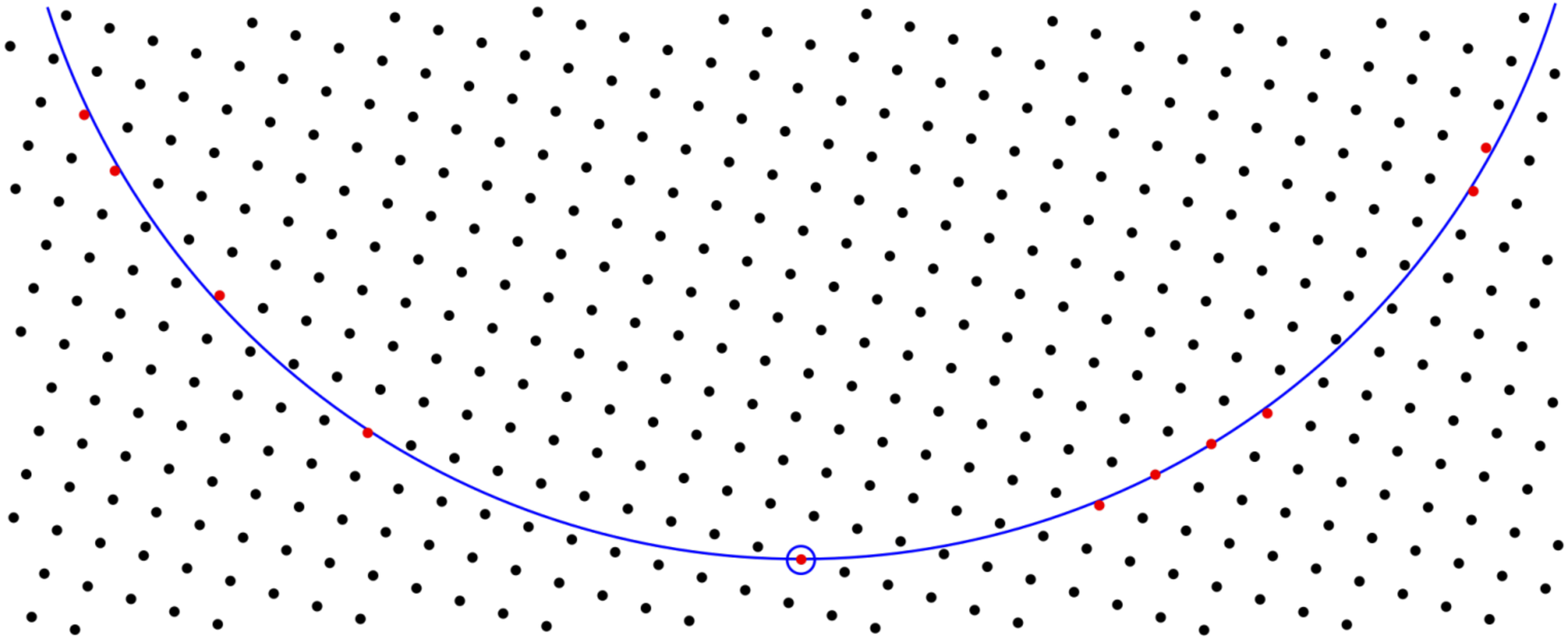
Diffraction patterns



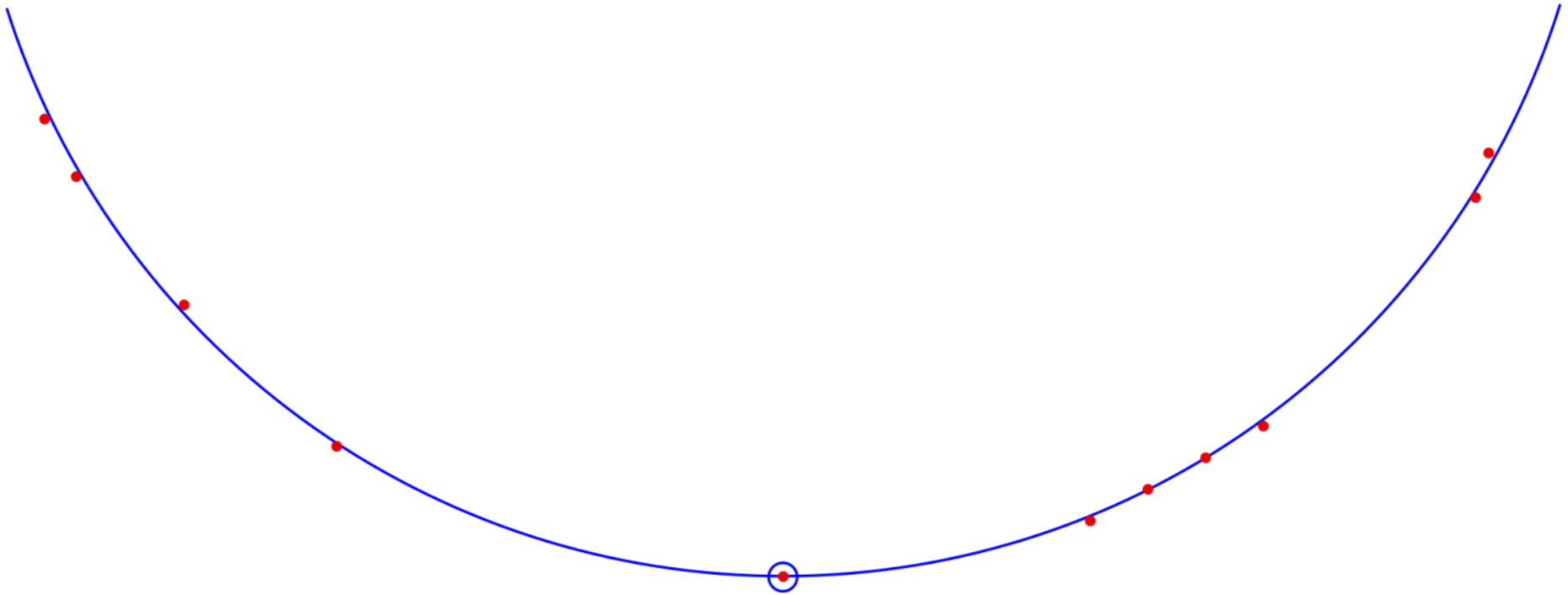


Indexing the diffraction pattern

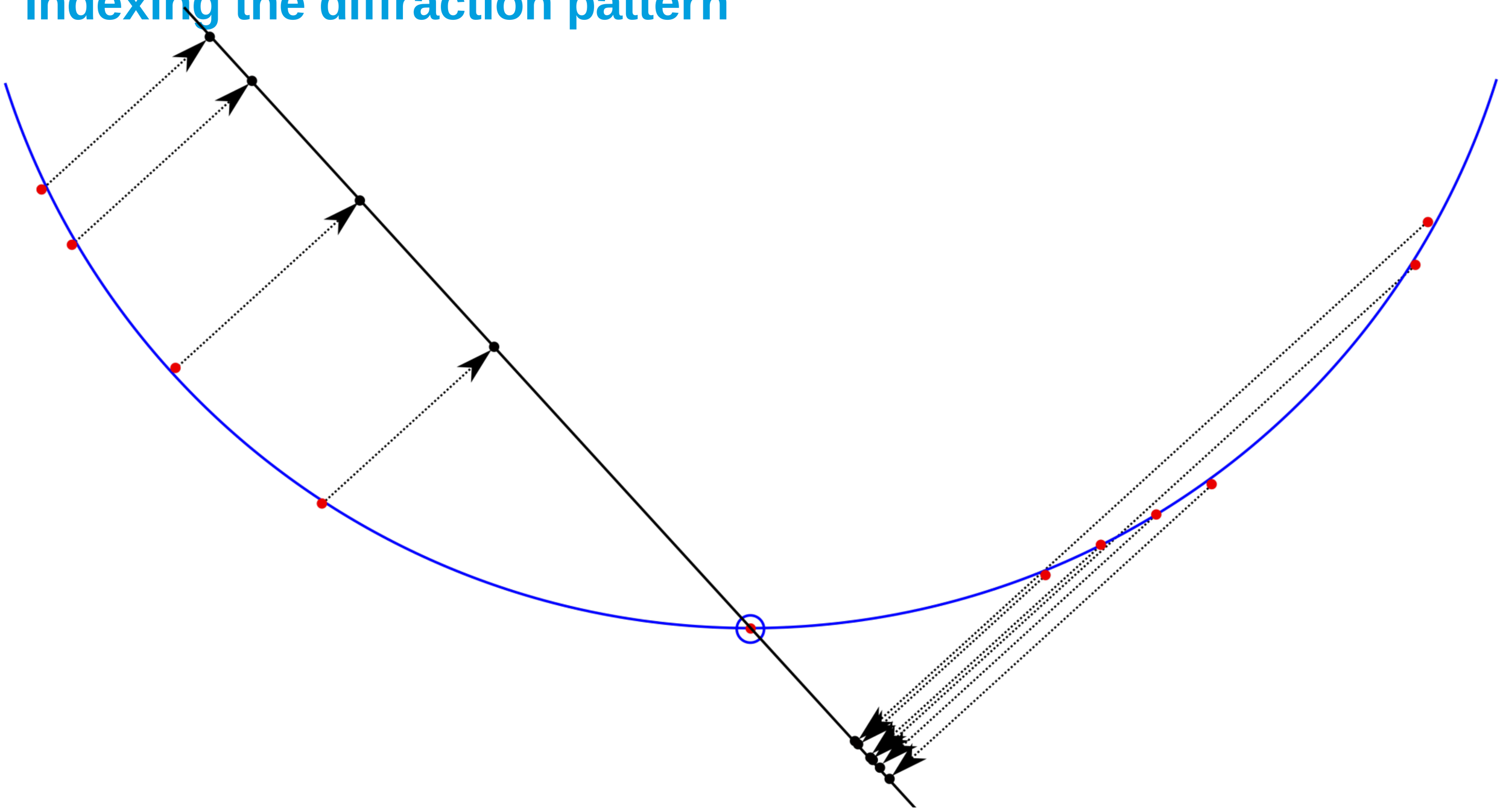
$$2d\sin\theta = \lambda$$



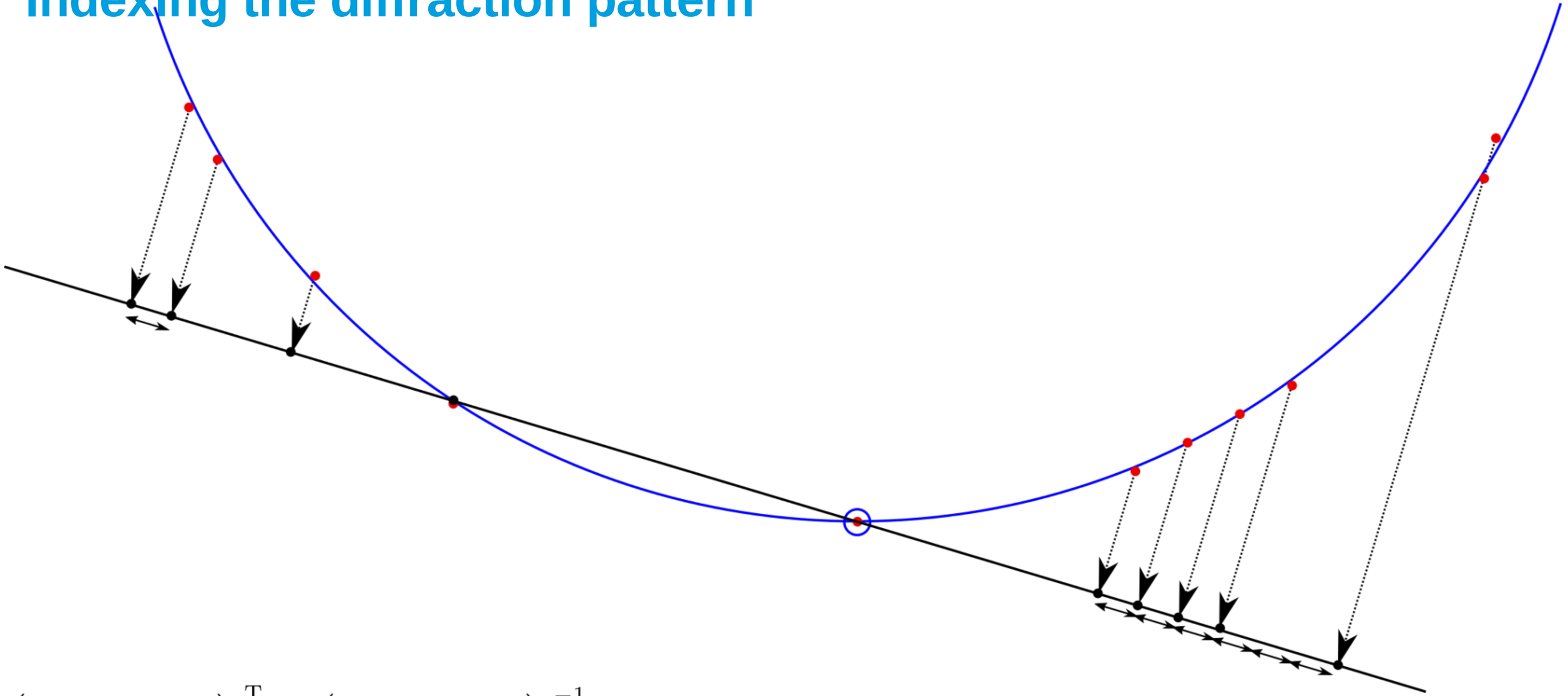
Indexing the diffraction pattern



Indexing the diffraction pattern

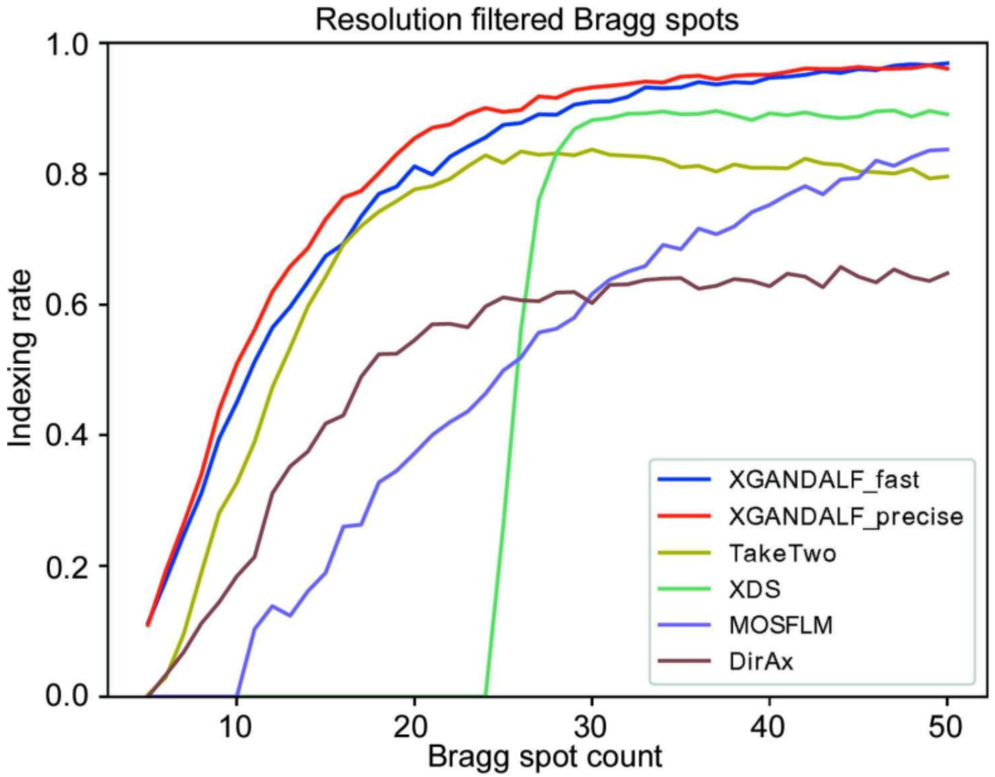


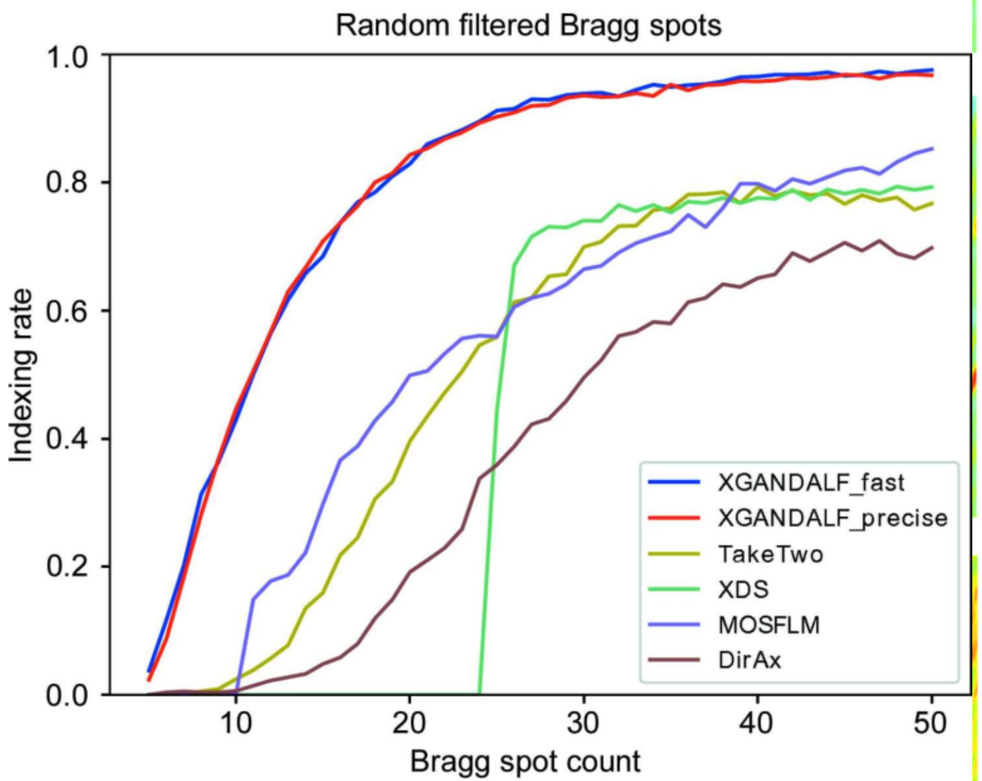
Indexing the diffraction pattern

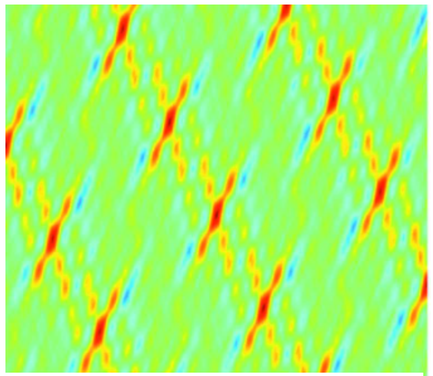


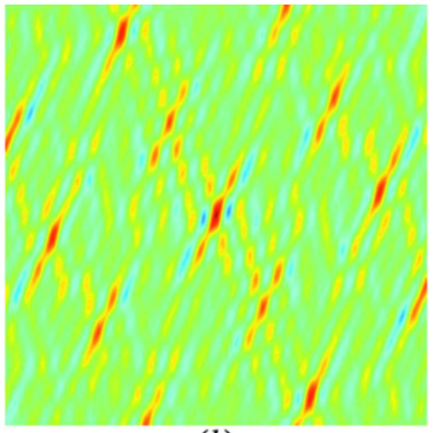
$$\begin{pmatrix} a_x^* & a_y^* & a_z^* \\ b_x^* & b_y^* & b_z^* \\ c_x^* & c_y^* & c_z^* \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{pmatrix}^{-1}$$

XGANDALF indexer for snapshots

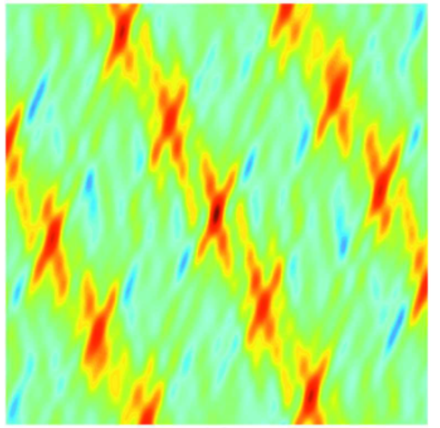




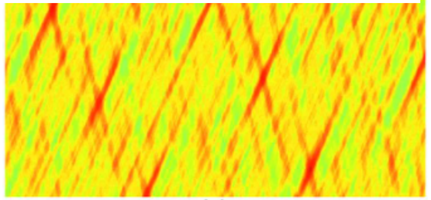




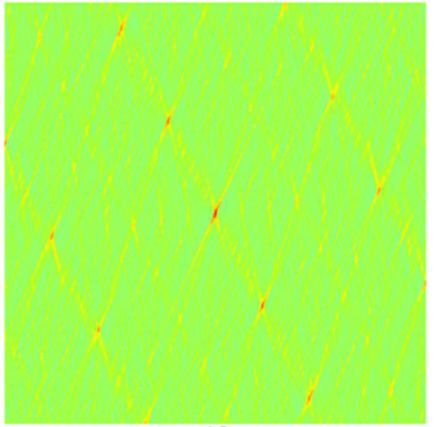
(b)



(d)



(e)

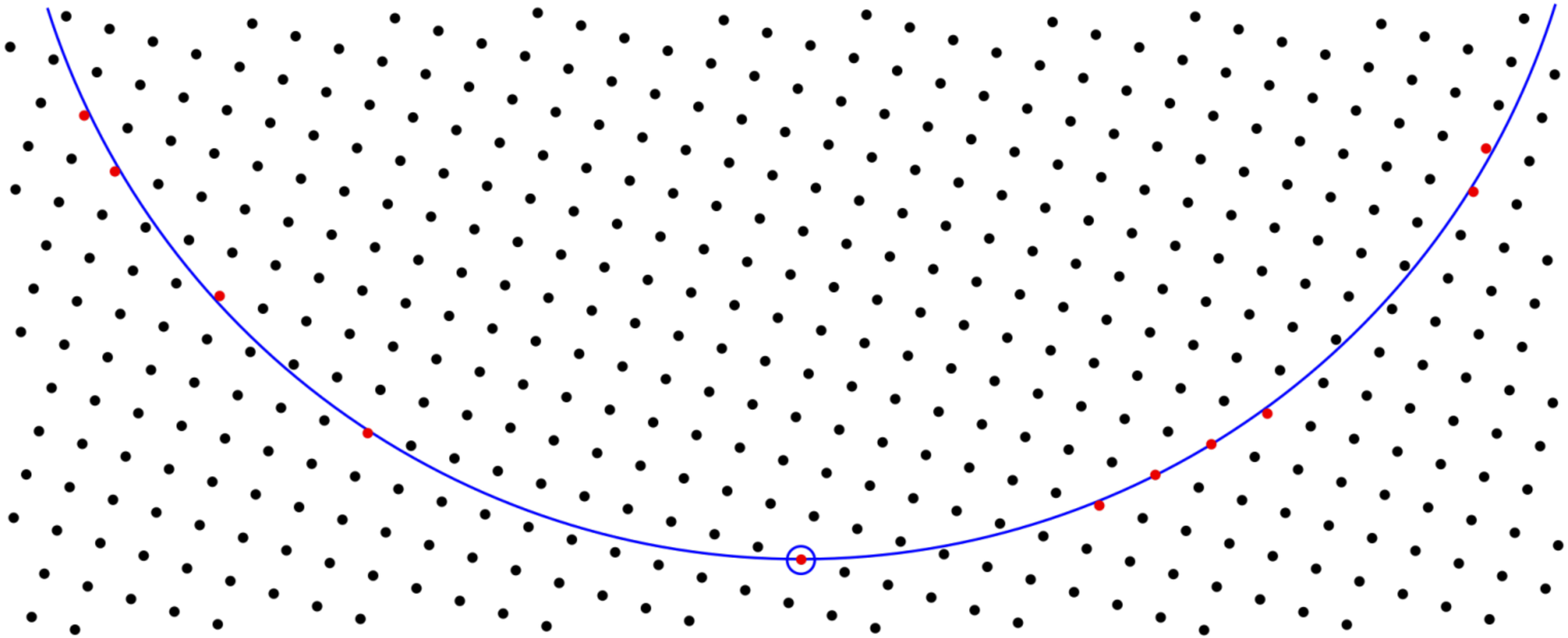


(f)

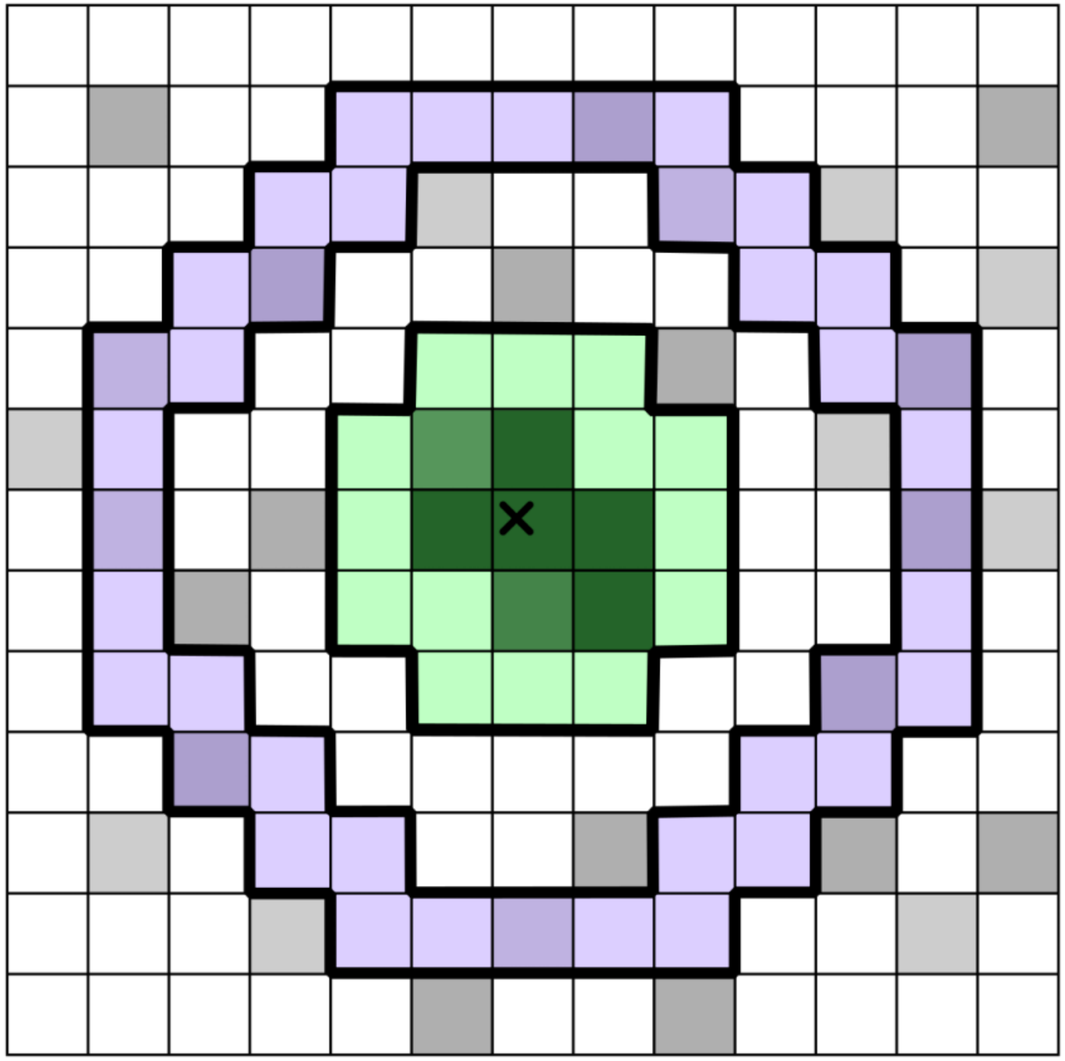
Y. Gevorkov et al, Acta Crystallographica A75 (2019) p694-704

Reflection "prediction"

Important difference from before: all lattice points get projected, not just the ones with peaks



Reflection integration

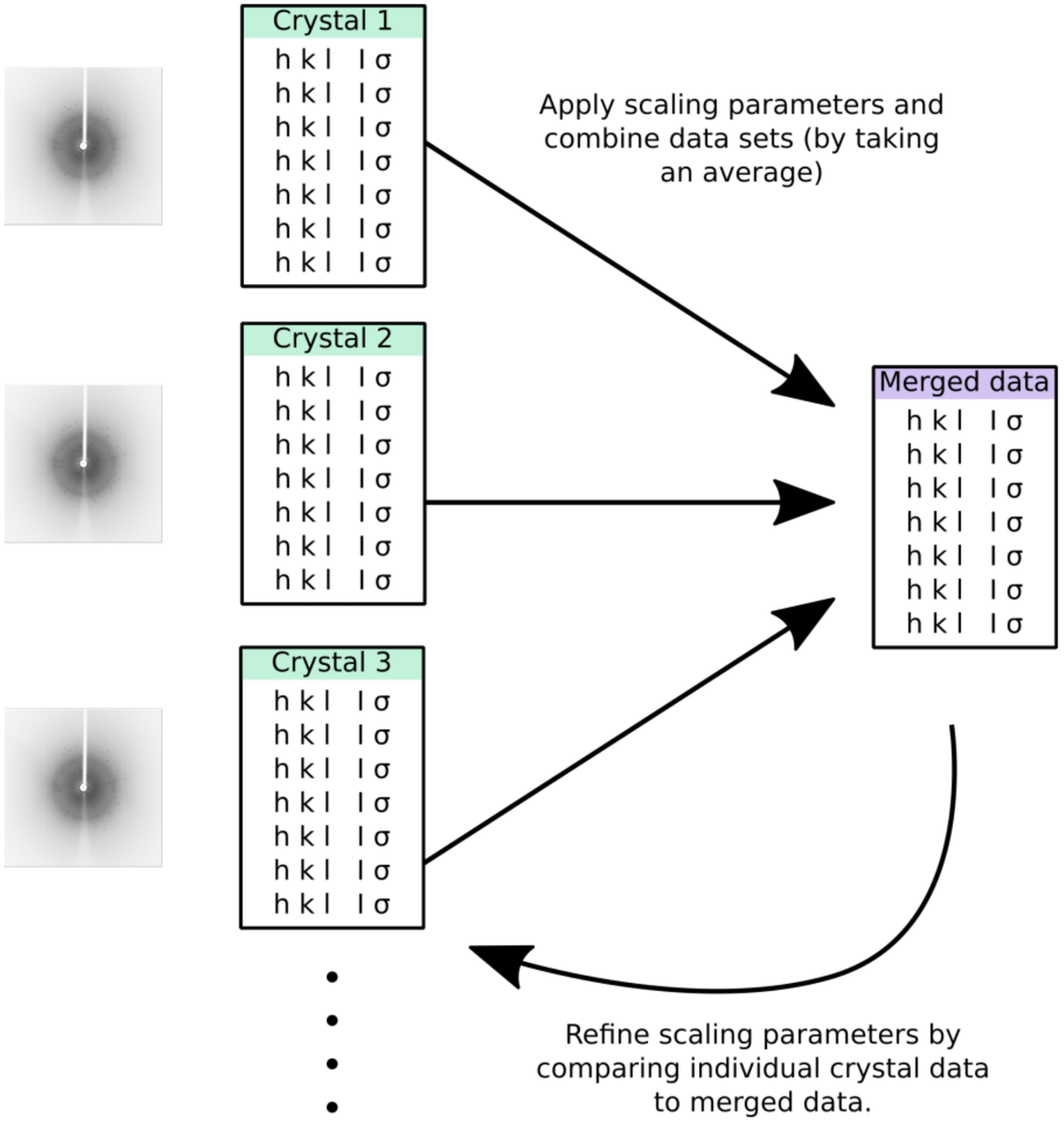


Integration fiducial

Peak region

Background region

Merging the measurements



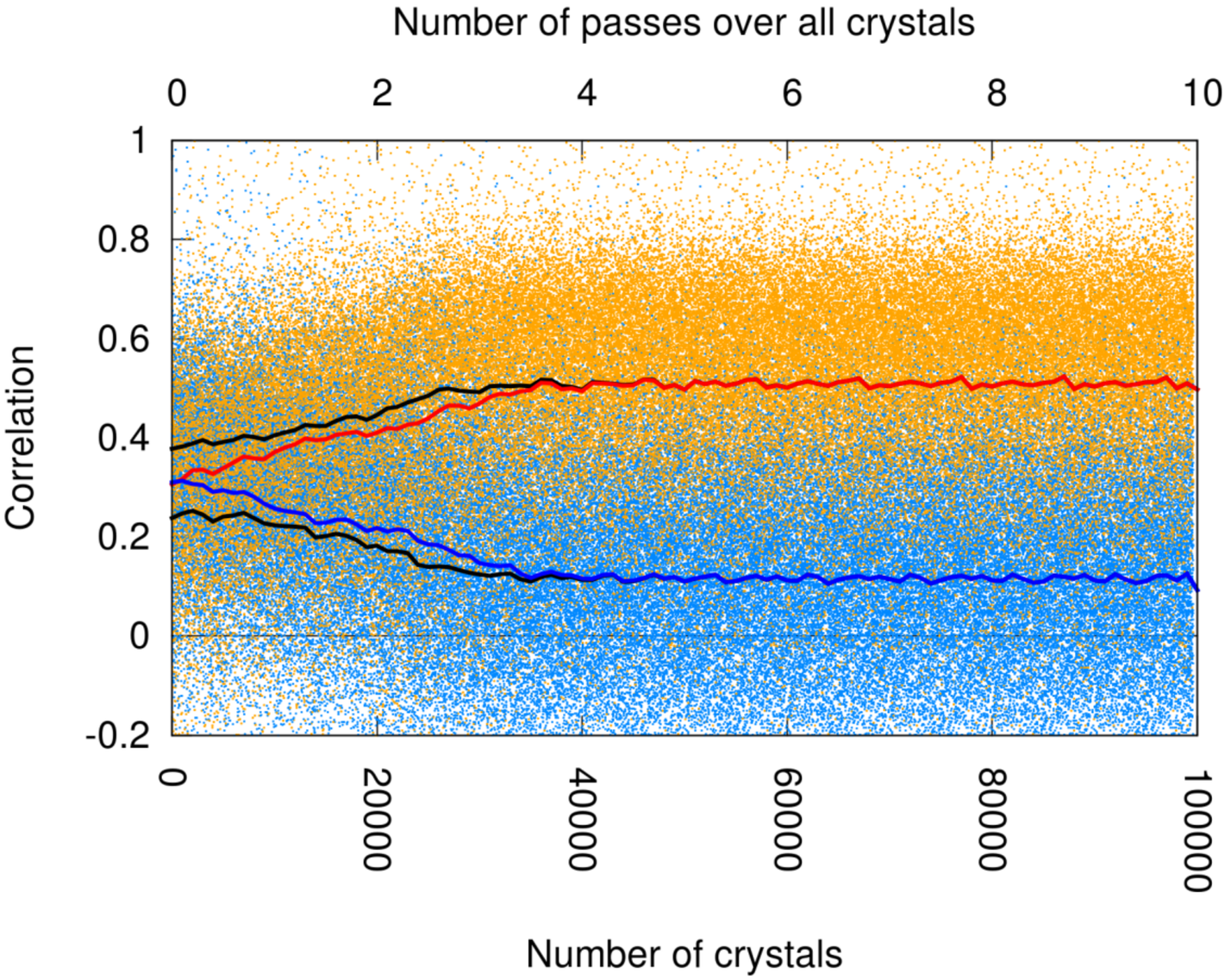
Indexing ambiguities

Clustering algorithm:

Brehm and Diederichs
Acta Cryst. D70 (2014) p101-109

Simplified (1D) version:

White, ... Brehm, ... Diederichs, ...
J. Appl. Cryst. 49 (2016) p680
(see section 4)



Figures of merit

From merged data:

Average I/sigma(I)
Number of measurements per merged reflection

Split pattern into two halves, merge each half separately:

R_split
CC½
CC* ..... all measure pretty much the same thing

Go back and look at individual patterns again:

ΔCC½ ("delta CC half") - see Assmann et al. J. Appl. Cryst. 49 (2016) p1021

What is CrystFEL?

SFX data processing pipeline
- → **indexamajig** - Indexing and integration
- → **process_hkl** - Monte Carlo merging, simple scaling
- → **partialator** - Scaling, partialities, post-refinement
- → **ambigator** - Resolve indexing ambiguities

Figures of merit
- → **check_hkl** - Completeness, multiplicity, SNR
- → **compare_hkl** - Compare reflection data

Simulation tools
- → **partial_sim** - Calculate partial reflections
- → **pattern_sim** - Fourier simulation

Visualisation tools
- → **hdfsee** - Display diffraction patterns
- → **render_hkl** - Visualise reflection data
- → **cell_explorer** - Exampline unit cell distributions

Utilities / other
- → **get_hkl** - Twin, example, add noise
- → **list_events** - Create event lists
- → **geoptimiser** - Refine detector geometry
- → **whirligig** - Find rotation series
- → **cell_tool** - Unit cell manipulations

Lots of supporting scripts (mostly in Python)

Documentation (manual pages, --help, example files)

Lots more on website (tutorial, demonstration data, citation lists, best practice guidelines, ...)

CrystFEL 0.9.0

→ New cell_tool
→ Interface to PinkIndexer
→ Reflection prediction for electrons and wide bandwidth X-rays
→ Arbitrary spectrum API
→ ZMQ/MsgPack interface
→ New cell comparison algorithm
→ Faster deltaCChalf calculation
→ Arbitrary polarisation/degree of polarisation (needed for LCLS-II)
→ Improvements to multi-lattice indexing
→ Handle CBF files natively

Advanced topics

Indexing ambiguities

... how to know there's an ambiguity, and how to fix it

Partiality modelling and post-refinement

Special considerations for TR-SX, MAD and MIR

Recognise when there's an ambiguity

Warning signs:

Point group 3, 4, 6, 312, 321 or 23
(432, 622, 32, 422, 222, 2 and 1 are OK)
.... but you'd have to know the structure already

Twin warnings from Phenix, CCP4 etc, or check_hkl --ltest :

```
$ check_hkl test.hkl -p 4ET8.pdb --ltest --ignore-negs
Using symmetry from reflection file: 4/mmm
Discarded 0 reflections (out of 3033) with I/sigma(I) < -inf
Discarded 672 reflections because they had negative intensities.
2631 pairs
<|L|> = 0.512 (ideal untwinned 0.500, twinned 0.375)
<L^2> = 0.344 (ideal untwinned 0.333, twinned 0.200)
$
```

Twin warnings mean...

1. There's an indexing ambiguity which you need to resolve

2. The crystals are really (physically) twinned

3. There's some weird data artifact (e.g. bad detector calibration)

.... any combination of the above.

Symmetry classification for serial crystallography

Triclinic lattice

$\bar{1}$	**1**	P$\bar{1}$	**P1**

Monoclinic lattice

	m		Pm, Pc, Cm, Cc
2	2/m	**P2, P2$_1$, C2**	P2/m, P2$_1$/m, C2/m, P2/c, P2$_1$/c, C2/c

Orthorhombic lattice

	mm2		Pmm2, Pmc2$_1$, Pcc2, Pma2, Pca2$_1$, Pnc2, Pmn2$_1$, Pba2, Pna2$_1$, Pnn2, Cmm2, Cmc2$_1$, Ccc2, Amm2, Aem2, Ama2, Aea2, Fmm2, Fdd2, Imm2, Iba2, Ima2
222	mmm	**P222, P222$_1$, P2$_1$2$_1$2, P2$_1$2$_1$2$_1$, C222$_1$, C222, F222, I222, I2$_1$2$_1$2$_1$**	Pmmm, Pnnn, Pccm, Pban, Pmma, Pnna, Pmna, Pcca, Pbam, Pccn, Pbcm, Pnnm, Pmmn, Pbcn, Pbca, Pnma, Cmcm, Cmce, Cmmm, Cccm, Cmme, Ccce, Fmmm, Fddd, Immm, Ibam, Ibca, Imma

Tetragonal lattice

	$\bar{4}$						
4	$\bar{4}$2m	$\bar{4}$m2	4/m	4mm	**P4, P4$_1$, P4$_2$, P4$_3$, I4, I4$_1$**	P$\bar{4}$, I$\bar{4}$	P4mm, P4bm, P4$_2$cm, P4$_2$nm, P4cc, P4nc, P4$_2$mc, P4$_2$bc, I4mm, I4cm, I4$_1$md, I4$_1$cd
						P$\bar{4}$2m, P$\bar{4}$2c, P$\bar{4}$2$_1$m, P$\bar{4}$2$_1$c, I$\bar{4}$2m, I$\bar{4}$2d	P$\bar{4}$m2, P$\bar{4}$c2, P$\bar{4}$b2, P$\bar{4}$n2, I$\bar{4}$m2, I$\bar{4}$c2	P4/m, P4$_2$/m, P4/n, P4$_2$/n, I4/m, I4$_1$/a
422	4/mmm				**P422, P42$_1$2, P4$_1$22, P4$_1$2$_1$2, P4$_2$22, P4$_2$2$_1$2, P4$_3$22, P4$_3$2$_1$2, I422, I4$_1$22**	P4/mmm, P4/mcc, P4/nbm, P4/nnc, P4/mbm, P4/mnc, P4/nmm, P4/ncc, P4$_2$/mmc, P4$_2$/mcm, P4$_2$/nbc, P4$_2$/nnm, P4$_2$/mbc, P4$_2$/mnm, P4$_2$/nmc, P4$_2$/ncm, I4/mmm, I4/mcm, I4$_1$/amd, I4$_1$/acd	

Rhombohedral lattice

3	$\bar{3}$	3m	**R3 (H3)**	R$\bar{3}$ (H$\bar{3}$)
32	$\bar{3}$m		**R32 (H32)**	R$\bar{3}$m (H$\bar{3}$m

Hexagonal lattice

3	$\bar{3}$					**P3, P3$_1$, P3$_2$**	P$\bar{3}$					
6	**312**	**321**	3m1	$\bar{6}$	31m	6mm	**P6, P6$_1$, P6$_5$, P6$_2$, P6$_4$, P6$_3$**	**P312, P3$_1$12, P3$_2$12**	**P321, P3$_1$21, P3$_2$21**	P3m1, P3c1	P$\bar{6}$	P31m, P31c	P6/m, P6$_3$/m	P6mm, P6cc, P6$_3$cm, P6$_3$mc
			$\bar{3}$m1	$\bar{6}$m2	$\bar{6}$2m	$\bar{3}$1m	6/m			P$\bar{3}$m1, P$\bar{3}$c1	P$\bar{6}$m2, P$\bar{6}$c2	P$\bar{6}$2m, P$\bar{6}$2c	P$\bar{3}$1m, P$\bar{3}$1c	
622			6/mmm				**P622, P6$_1$22, P6$_5$22, P6$_2$22, P6$_4$22, P6$_3$22**			P6/mmm, P6/mcc, P6$_3$/mcm, P6$_3$/mmc				

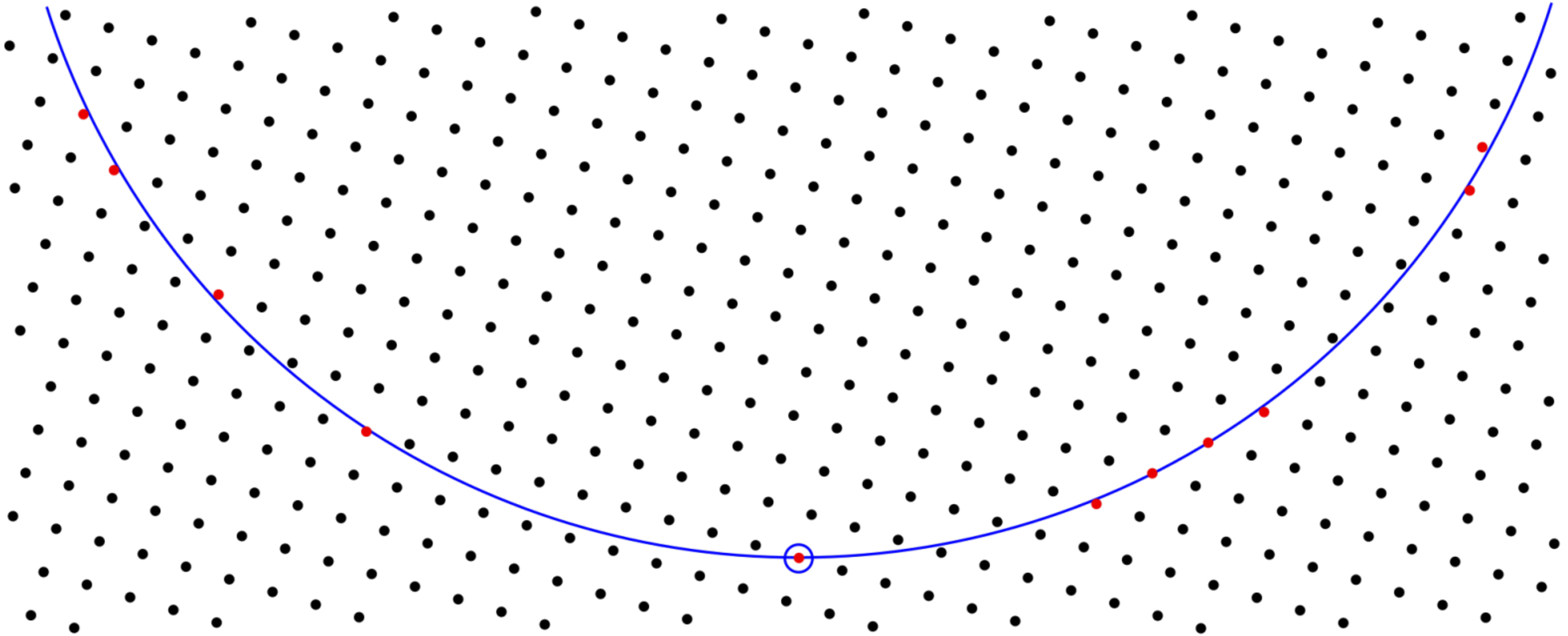
Cubic lattice

23	$\bar{4}$3m	m$\bar{3}$	**P23, F23, I23, P2$_1$3, I2$_1$3**	P$\bar{4}$3m, F$\bar{4}$3m, I$\bar{4}$3m, P$\bar{4}$3n, F$\bar{4}$3c, I$\bar{4}$3d	Pm$\bar{3}$, Pn$\bar{3}$, Fm$\bar{3}$, Fd$\bar{3}$, Im$\bar{3}$, Pa$\bar{3}$, Ia$\bar{3}$
432	m$\bar{3}$m		**P432, P4$_2$32, F432, F4$_1$32, I432, P4$_3$32, P4$_1$32, I4$_1$32**	Pm$\bar{3}$m, Pn$\bar{3}$n, Pm$\bar{3}$n, Pn$\bar{3}$m, Fm$\bar{3}$m, Fm$\bar{3}$c, Fd$\bar{3}$m, Fd$\bar{3}$c, Im$\bar{3}$m, Ia$\bar{3}$d	

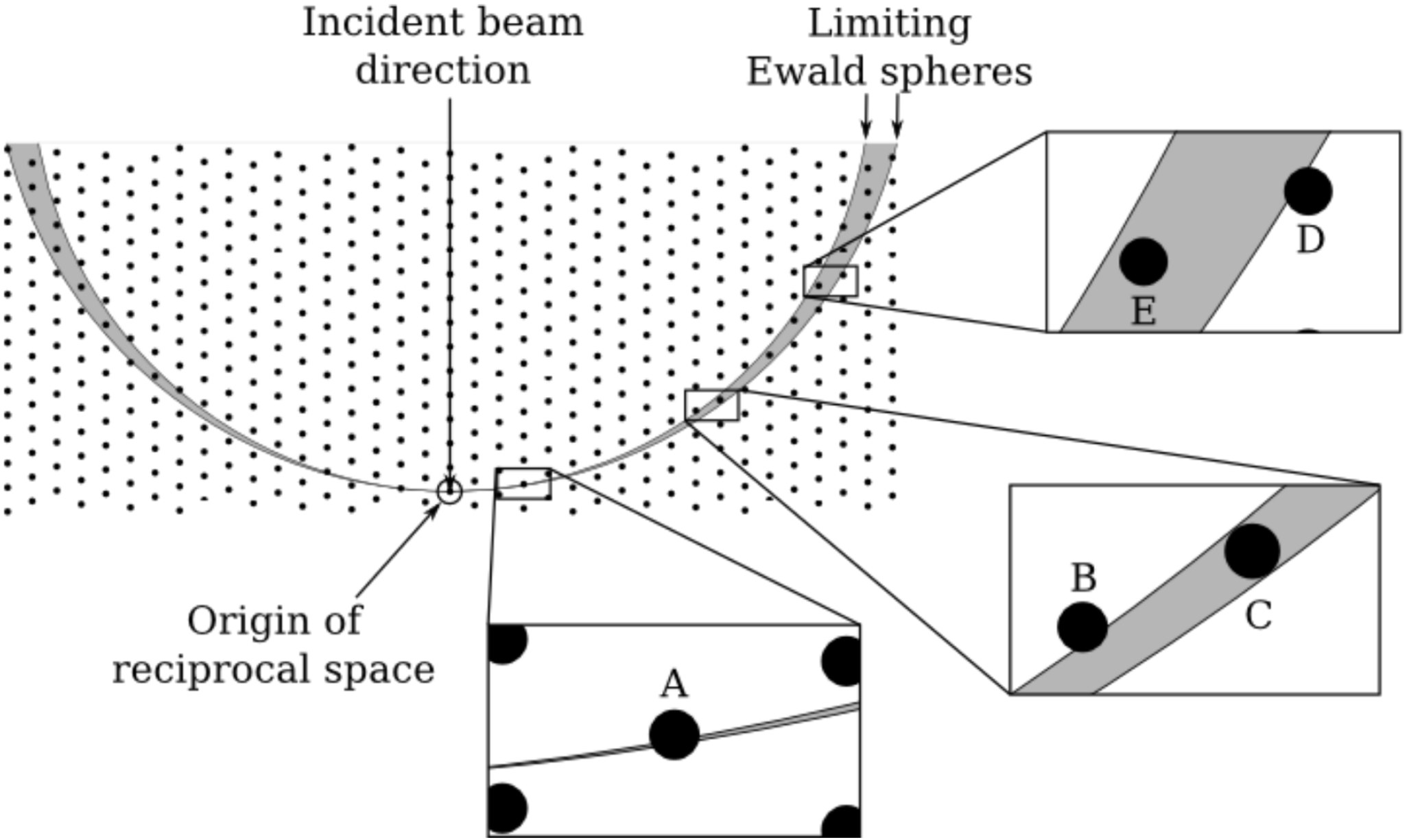
Spotting "accidental" ambiguities

```
$ cell_tool 4ET8.pdb --find-ambi -y 422
Input unit cell: 4ET8.pdb
------------------> The input unit cell:
tetragonal P, unique axis c, right handed.
a      b      c            alpha   beta  gamma
 79.00  79.00  38.00 A     90.00  90.00  90.00 deg
Looking for ambiguities up to 3x each lattice length.
This will take about 30 seconds.  Please wait...

[....]

Observed symmetry operations:
      Observed : hkl       -h,-k,l    -h,k,-l    -k,-h,-l   k,-h,l     -k,h,l     k,h,-l     h,-k,-l
Ambiguity operations:
Observed -> 422 :
```

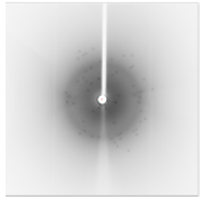
Reflection partiality

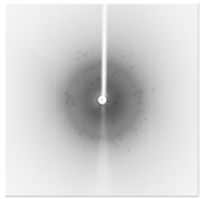


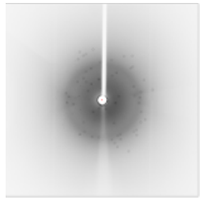
Reflection partiality

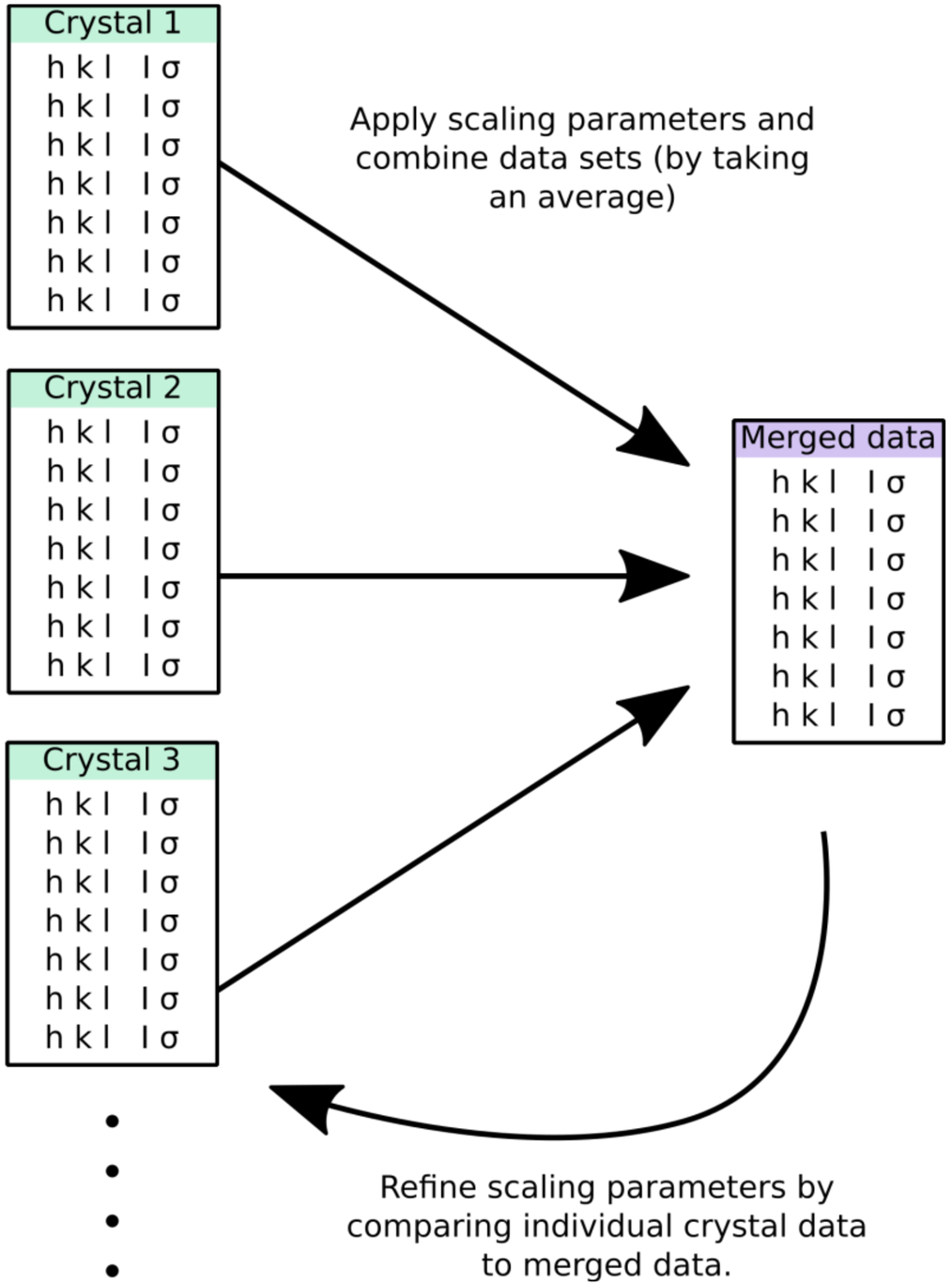


Post-refinement









Partiality models available in CrystFEL (v0.9.0)

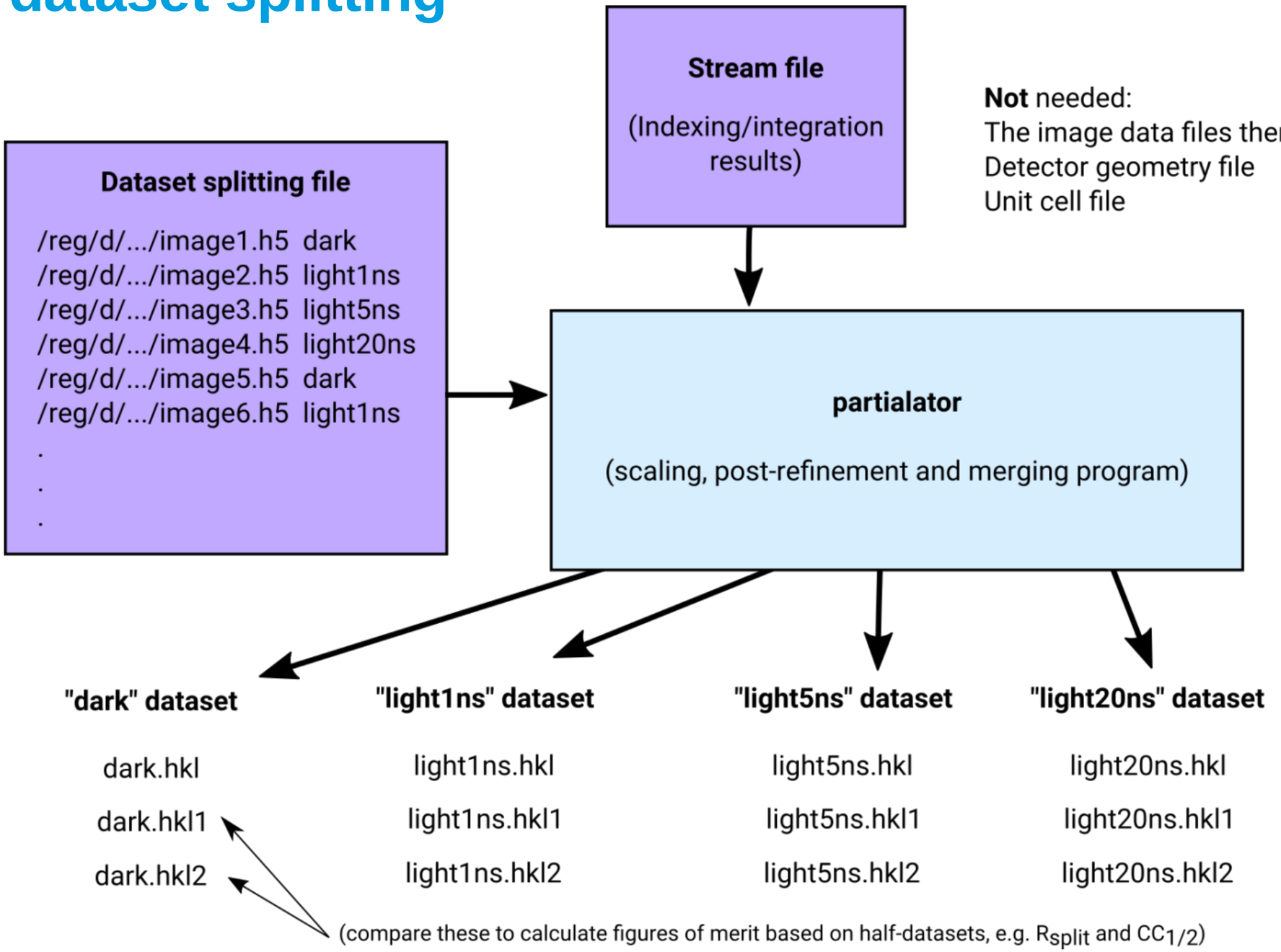
partialator --model=unity
Set all partialities to 1 (no modelling)

partialator --model=xsphere
Numerical integration across arbitrary spectrum
See Ginn et al, Acta Cryst. D71, p1400 (2015)

partialator --model=offset
Thin Ewald sphere, like Kabsch, Acta Cryst. D70, p2204 (2014)

partialator --model=ggpm
Analytically-defined model based on Gaussian overlap integrals
(Needs X-ray spectrum to be an analytical function)

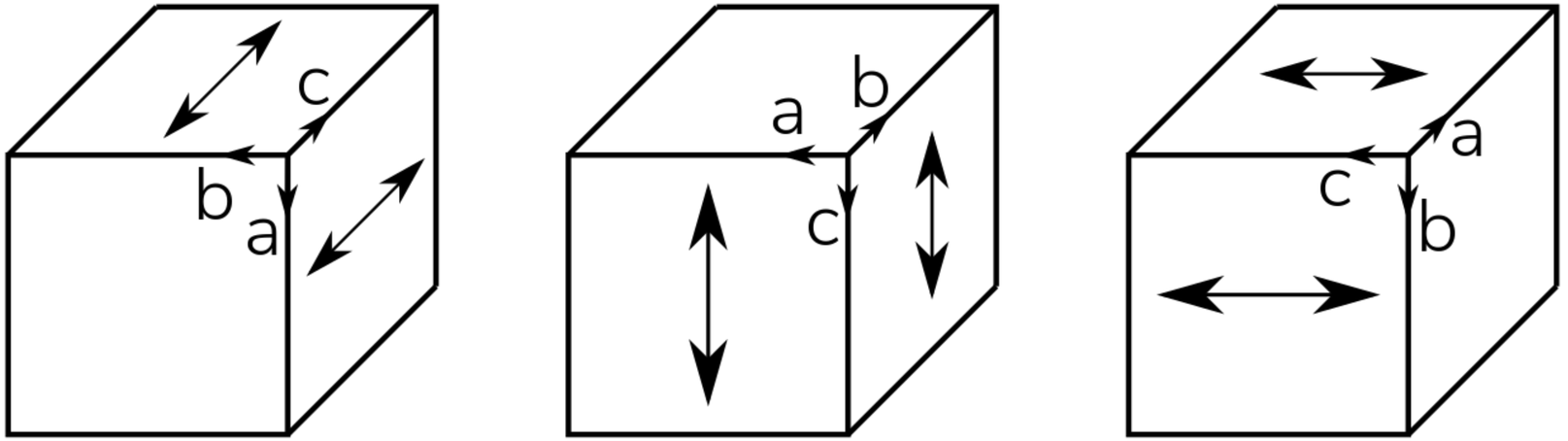
Custom dataset splitting



Three-way ambiguity in human melatonin receptor

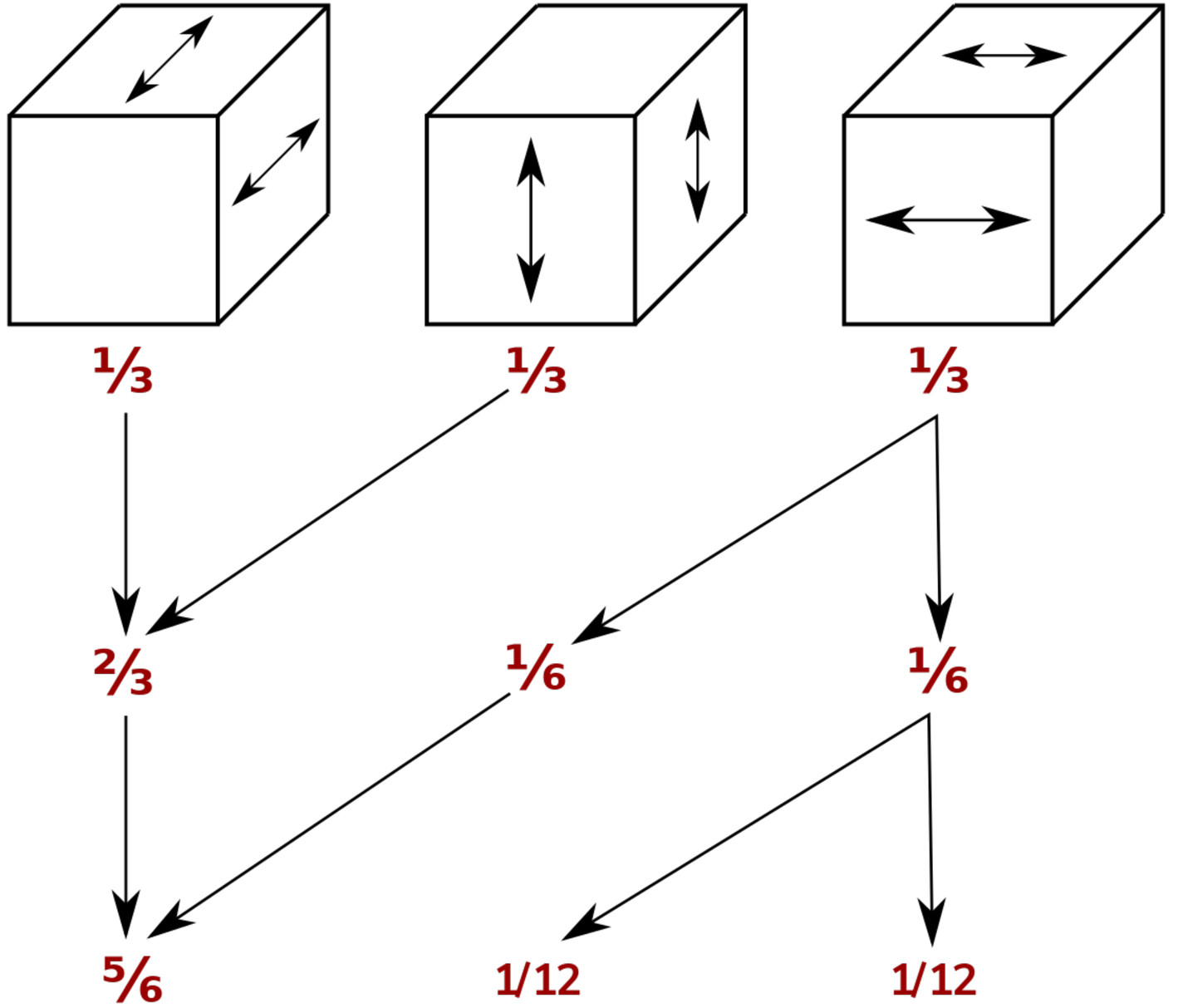
Tetragonal P
a = 122.3 Å
b = 122.3 Å
c = 122.8 Å
alpha = 90°
beta = 90°
gamma = 90°

Indexing ambiguity: hkl → klh → lhk → hkl → ...
One indexing ambiguity, three possibilities

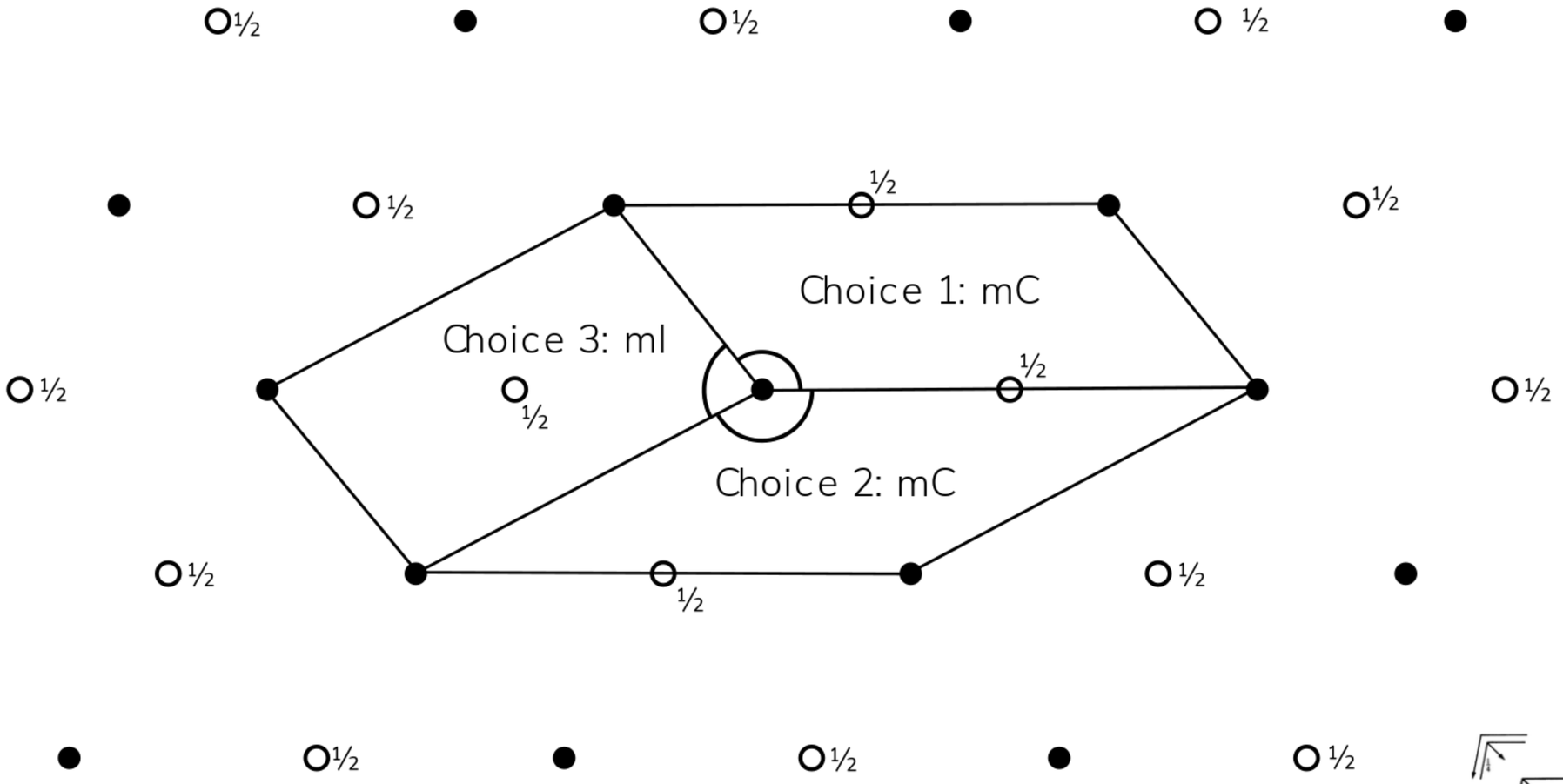


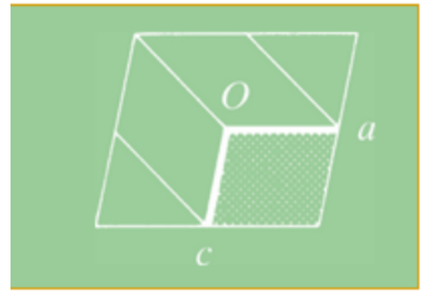
B. Stauch, L. C. Johannson, J. D. McCorvy, N. Patel, G. Won Han, X.-P. Huang, C. Gati, A. Batyuk, S. T. Slocum, A. Ishchenk, W. Brehm, T. A. White, N. Michaelian. C. Madsen, L. Zhu, T. D. Grant, J. M. Grandner, A. Shiriaeva, R. H. J. Olsen, A. R. Tribo, S. Yous, R. C. Stevens, U. Weierstall, V. Katritch, B. L. Roth, W. Liu, V. Cherezov.
Nature 569 (2019) p284)

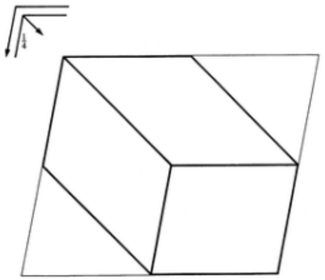
Resolving a three-way cyclic ambiguity

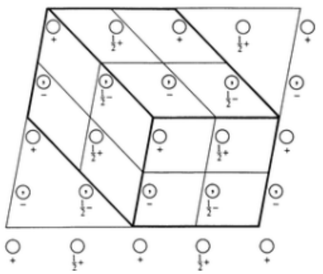


Cell choices for monoclinic C-centered structures









CrystFEL unit cell tool

Find possible indexing ambiguities:

```
$ cell_tool -p mystructure.cell --find-ambi
```

Calculate radii of all powder rings (useful for calibration):

```
$ cell_tool -p mystructure.cell --rings
```

Calculate all unit cell choices for monoclinic C cell:

```
$ cell_tool -p mystructure.cell --cell-choices
```

Transform a unit cell:

```
$ cell_tool -p mystructure.cell --transform=b,a-2b,c
```

Calculate primitive unit cell from centered cell:

```
$ cell_tool -p mystructure.cell --uncenter
```

Find transformation relating two cells:

```
$ cell_tool -p mystructure.cell --compare other.cell
```

Acknowledgements

Richard Kirian (Arizona State U.)
Kenneth Beyerlein
Andrew Aquila
Andrew Martin
Lorenzo Galli
Chun Hong Yoon
Karol Nass
Nadia Zatsepin (Arizona State U.)
Anton Barty
Thomas Grant (Hauptman-Woodward Inst., Buffalo)
Cornelius Gati
Aleksandra Tolstikova
Wolfgang Brehm (U. Konstanz, now CFEL)
Valerio Mariani
Parker de Waal (Van Andel Inst.)
Takanori Nakane (U. Tokyo)
Keitaro Yamashita (SPring-8)
Oleksandr Yefanov
Helen Ginn (U. Oxford)
Steve Aplin
Nicolas Riebesel (Technical U. Hamburg-Harburg)
Yaroslav Gevorkov

... and many more for feedback, ideas and guidance!

Funding: Helmholtz Association (Program-Oriented Funding); PIER Helmholtz Graduate School; European Union's 2020 Research and Innovation Programme (Marie Skłodowska-Curie grant agreement 637295 "X-Probe"); X-ray Free Electron Laser Priority Strategy Programme (MEXT, Japan); SACLA HPC and Mini-K; BMBF German-Russian Cooperation "SyncFELMed" (grant 05K14CHA); BioStruct-X (Seventh Framework Programme of the European Commission)