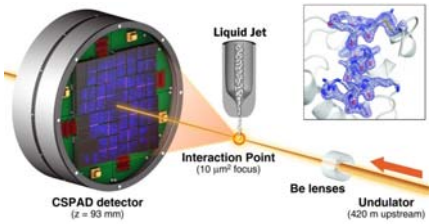
Experimental Phasing of SFX Data

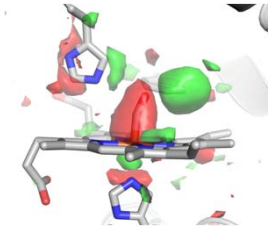
Thomas Barends
MPI for Medical Research, Heidelberg

Conclusions

X-ray Free-Electron Lasers are pushing back the boundaries of possibility in biological crystallography:

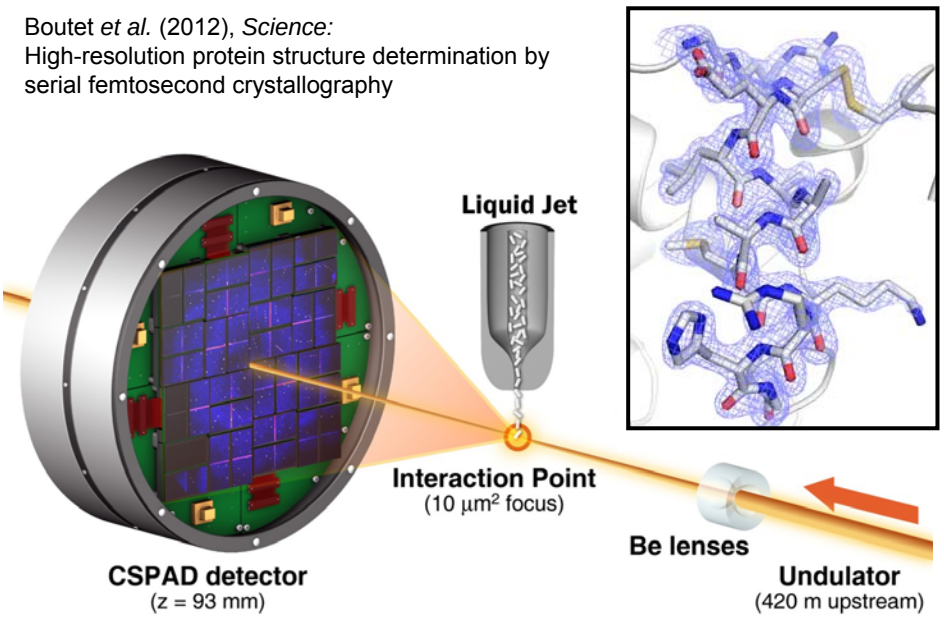
-Data collection from (sub)micrometer sized crystals
-Outrunning radiation damage (See Kern *et al.*, Suga *et al.*, Hirata *et al.*)
-Allowing subpicosecond time resolution





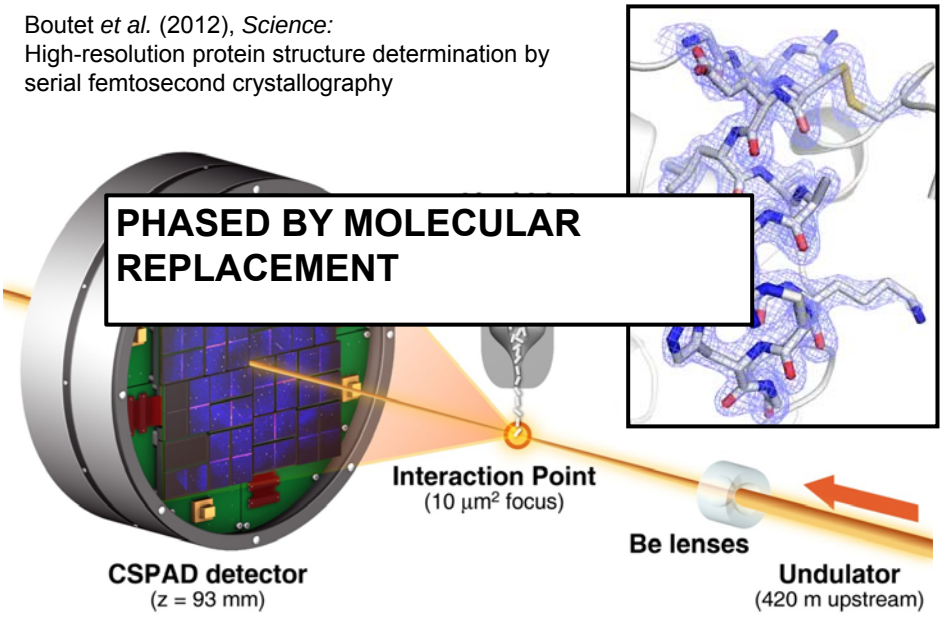
LYSOZYME DATA FROM CXI FEBRUARI 2011 – 1.9 Å resolution

Boutet *et al.* (2012), *Science:*
High-resolution protein structure determination by serial femtosecond crystallography

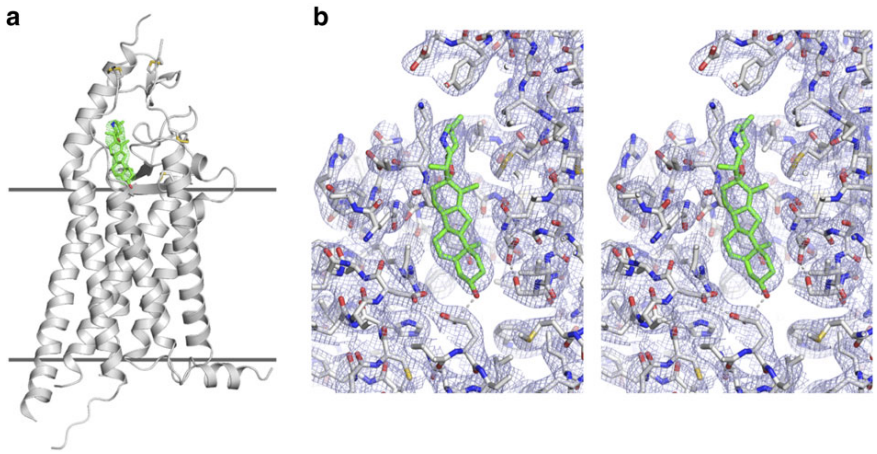


LYSOZYME DATA FROM CXI FEBRUARI 2011 – 1.9 Å resolution

Boutet *et al.* (2012), *Science:*
High-resolution protein structure determination by serial femtosecond crystallography

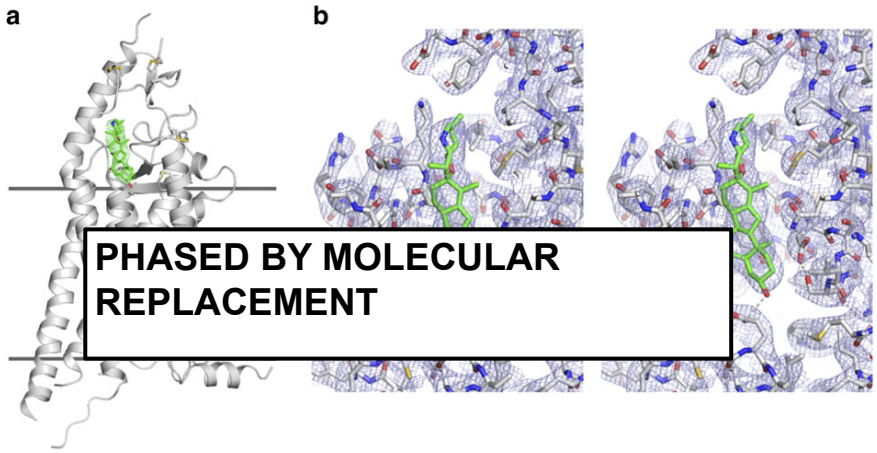


PHASED BY MOLECULAR REPLACEMENT



"Smoothened" GPCR complexed with cyclopamine
Structure determination of membrane proteins from less than 0.5 mg of protein!

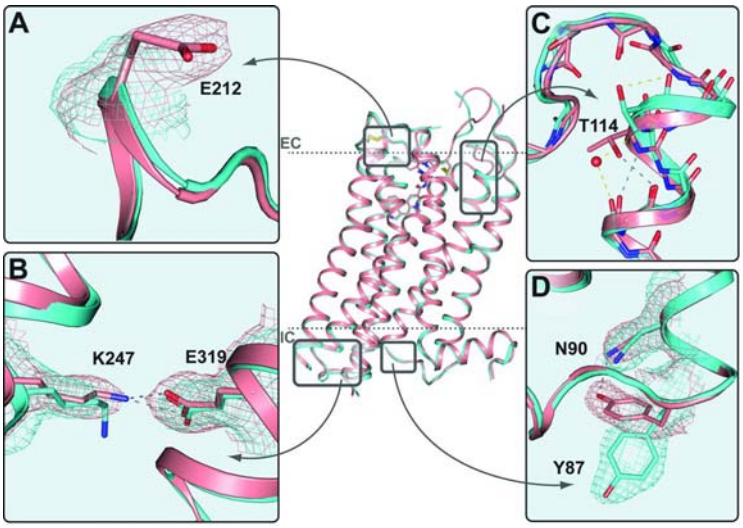
Weierstall *et al*, Nature Communications 5, Article number: 3309
doi:10.1038/ncomms4309



PHASED BY MOLECULAR REPLACEMENT

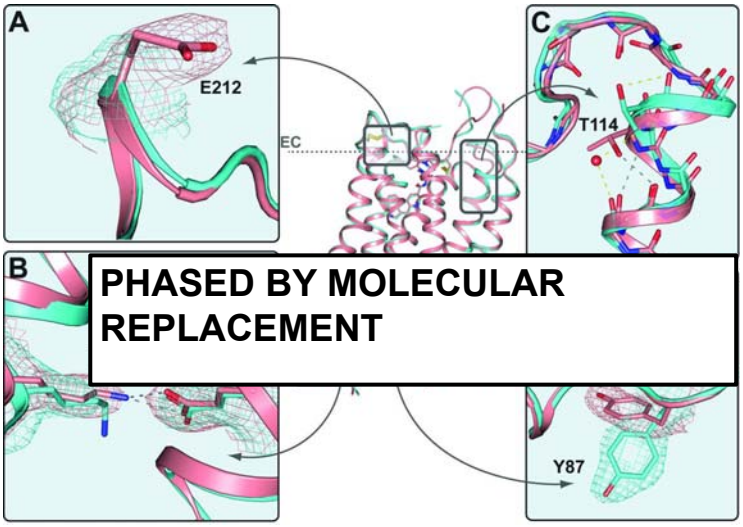
"Smoothened" GPCR complexed with cyclopamine
Structure determination of membrane proteins from less than 0.5 mg of protein!

Weierstall *et al*, Nature Communications 5, Article number: 3309
doi:10.1038/ncomms4309



Red: XFEL
Green: synchrotron

Liu et al., **Serial Femtosecond Crystallography of G Protein-Coupled Receptors**
Science 20 December 2013:
Vol. 342 no. 6165 pp. 1521-1524DOI:10.1126/science.1244142

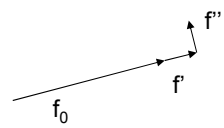


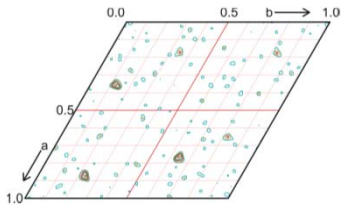
PHASED BY MOLECULAR REPLACEMENT

Red: XFEL
Green: synchrotron

Liu et al., **Serial Femtosecond Crystallography of G Protein-Coupled Receptors**
Science 20 December 2013:
Vol. 342 no. 6165 pp. 1521-1524DOI:10.1126/science.1244142

Phase determination for new structures?



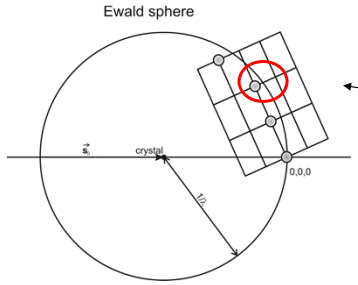


Introduce a heavy atom and measure
-the differences between I(hkl) and I(-h-k-l) (SAD)
-I(hkl) and I(-h-k-l) at different wavelengths (MAD)
-the differences between I(heavy atom) and I(no heavy atom)
(SIR(AS), MIR(AS))
-or introduce small changes by radiation damage (RIP(AS))

$$\rho(xyz) = \frac{1}{V} \sum_{hkl=-\infty}^{\infty} |F(hkl)| e^{-2\pi i(hx+ky+lz)+i\alpha(hkl)}$$

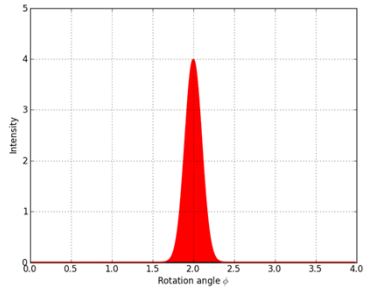
All of this relies on being able to accurately measure small differences between large values – requires a far greater accuracy than what is needed for your average difference Fourier map!

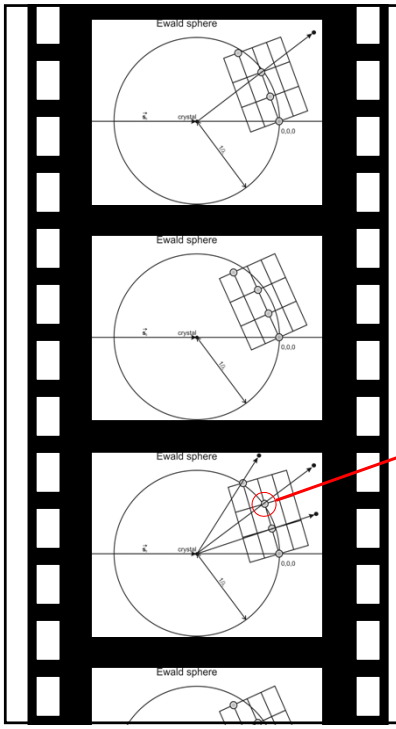
Indexing and integrating reflections: rotation/oscillation method



Accurate determination of profile → accurate I and σ(I)

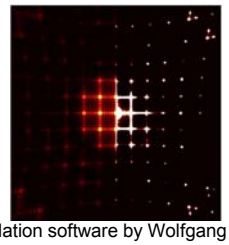
Rotation method
-rotate xtal over finite range
-calculate orientation matrix from observed spot positions
Can fully integrate whole reflections!





Serial femtosecond crystallography
- Each crystal is immediately destroyed
- Numerous shots of different crystals with possibly different sizes
- No *a priori* control over orientation
- Crystals effectively stand still during a 300 fs pulse
-Only part of reflection intersects Ewald sphere ("partials", no "fullies")
-Fringes rather than neat spots

6x6x6 unit cells



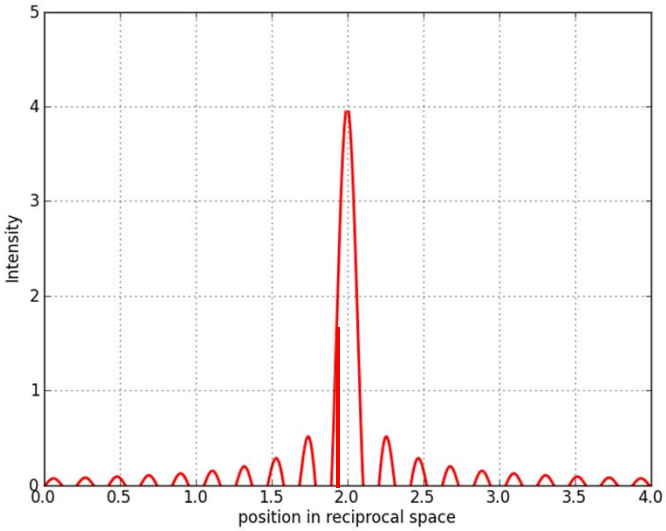
200x200x200 unit cells

(Simulation software by Wolfgang Kabsch)

- Crystals may be hit by center of the beam, or just be grazed by it
-Beam intensity and spectrum varies from shot to shot

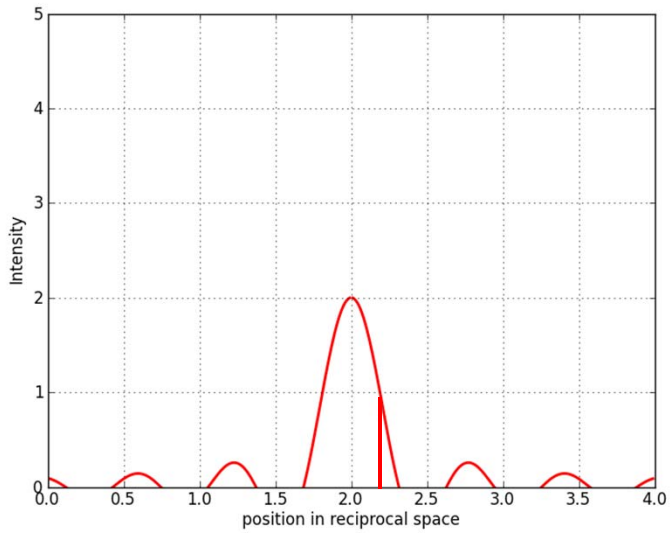
1st observation of a certain I(hkl)

Large crystal, hit full-on by the center of the direct beam



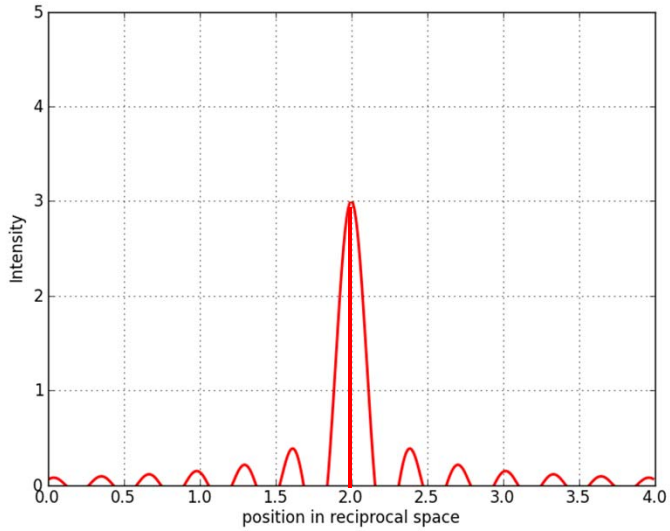
2nd observation of same I(hkl)

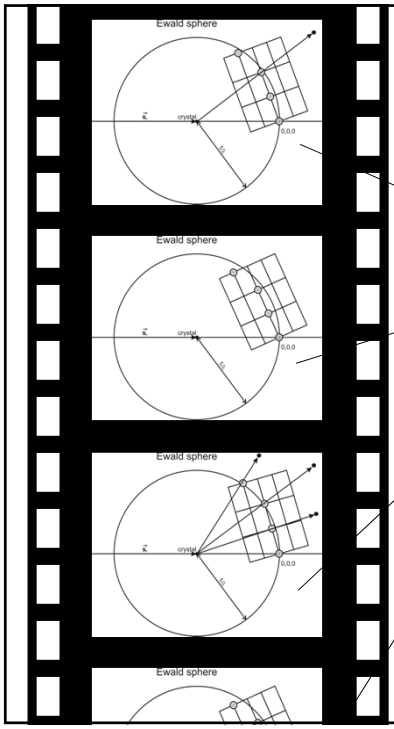
Much smaller crystal



3rd observation

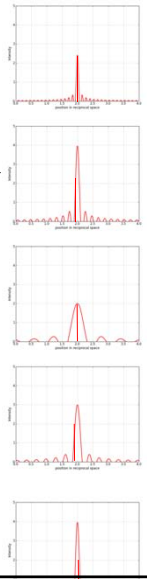
Bigger crystal than the last one…

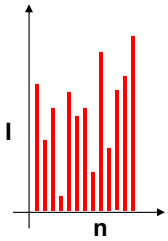




It is possible to do a *Monte Carlo* integration over multiple *indexed* femtosecond images and obtain a dataset of fully integrated reflections

Kirian *et al* (2010), *Optics Express*,**18**, 5713-5723:

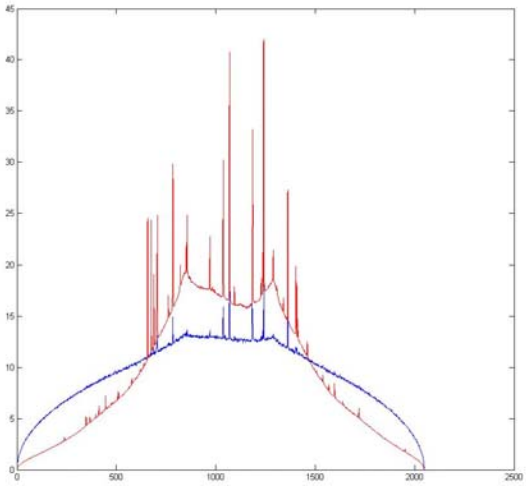


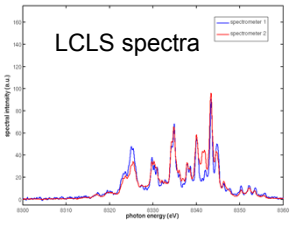


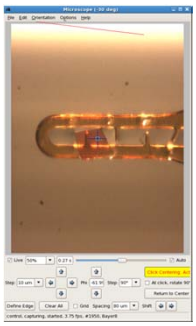
$$\langle I[hkl]\rangle = \Sigma I_n / n$$

…how good are these I? What are their sigma's?

Shot-to-shot beam differences







Projections of two images of the same macroscopic crystal taken shortly after each other

Test case for *de novo* phasing

- We want a really, really strong signal
- Idea: use a really good heavy atom derivative of something that scatters really, really well: Lysozyme:Gd

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Gd-HPDO3A, a complex to obtain high-phasing-power heavy-atom derivatives for SAD and MAD experiments: results with tetragonal hen egg-white lysozyme

Éric Girard, Laurent Chantalat, Jean Vicat and Richard Kahn*

Laboratoire de Cristallographie Macromoléculaire, Institut de Biologie Structurale (J.-P. Ebel CEA–CNRS–UJF, 41 Rue Jules Horowitz), 38027 Grenoble CEDEX 01, France

Correspondence e-mail: kahn@ibs.fr

A neutral gadolinium complex, Gd-HPDO3A, is shown to be a good candidate to use to obtain heavy-atom derivatives and solve macromolecular structures using anomalous dispersion. Tetragonal crystals of a gadolinium derivative of hen egg-white lysozyme were obtained by co-crystallization using different concentrations of the complex. Diffraction data from three derivative crystals (100, 50 and 10 m*M*) were collected to a resolution of 1.7 Å using Cu *K*α radiation from a rotating anode. Two strong binding sites of the gadolinium complex to the protein were located from the gadolinium anomalous signal in both the 100 and 50 m*M* derivatives. A single site is occupied in the 10 m*M* derivative. Phasing using the anomalous signal at a single wavelength (SAD method) leads to an electron-density map of high quality. The structure of the 100 m*M* derivative has been refined. Two molecules of the gadolinium complex are close together. Both molecules are located close to tryptophan residues. Four chloride ions were found. The exceptional quality of the SAD electron-density map, only enhanced by solvent flattening, suggests that single-wavelength anomalous scattering with the Gd-HPDO3A complex may be sufficient to solve protein structures of high

Received 18 July 2001
Accepted 3 October 2001

PDB Reference: Gd-derivative of lysozyme, 1h87.

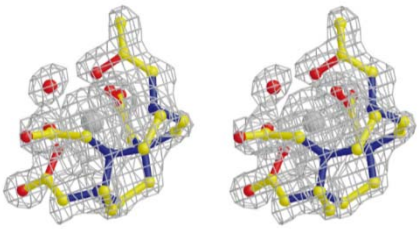
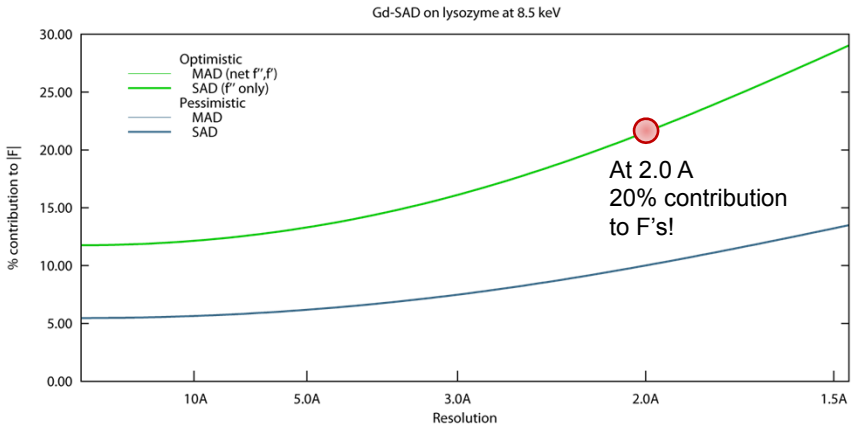


Figure 5
Stereoview of the final $2|F_o| - |F_c|$ electron-density map around Gd-HPDO3A (site Gd 1) and its ninth ligand, water molecule Wat152. The map is contoured at the 1σ level.

Acta Cryst. (2002). D**58**, 1–9

Girard *et al.*, *Acta Cryst.* **D**58, 1-9

Anomalous differences from Gd

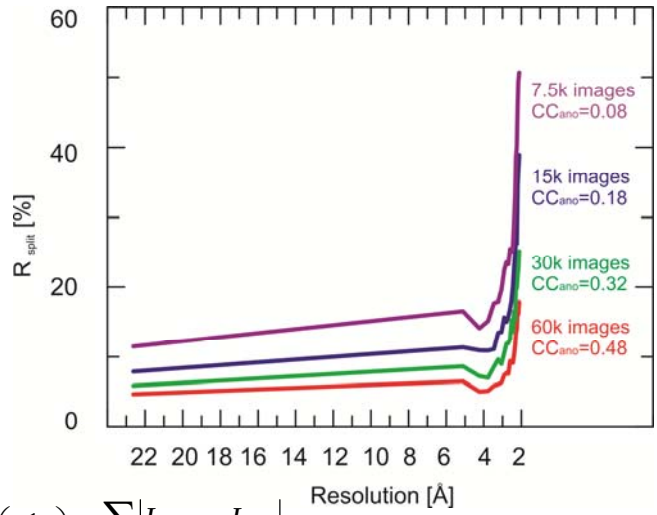


Can we get a strong enough anomalous signal from this to
-find the heavy atom positions?
-get the phases for the protein?

	LCLS SFX data	Rotating anode data
Data collection		
Space group	$P4_32_12$	$P4_32_12$
Cell dimensions [a]		
a, *b*, *c* (Å)	79, 79, 39	78.1, 78.1, 39.2
a, *b*, *g* (º)	90, 90, 90	90, 90, 90
Wavelength (Å)	1.45	1.5418
Pulse energy/fluence at sample	60 µJ/ $8 \cdot 10^{23}$ photons/s [b]	10^7 photons/s
Dose (MGy)	22 MGy/crystal	1 kGy/dataset
Number of collected diffraction patterns	2,402,199	360 (high-res run) 140 (low-res run)
Number of crystal hits	191,060	500
Number of indexed images patterns	59,667	500
Resolution (Å) [c]	40-2.1 (2.14-2.10)	40-1.8 (1.83-1.8)
$R_{split} = \frac{1}{\sqrt{2}} \cdot \frac{\sum_{hkl} \left\| I_{hkl}^{even\_images} - I_{hkl}^{odd\_images} \right\|}{\sum_{hkl} \frac{1}{2} \left(I_{hkl}^{even\_images} + I_{hkl}^{odd\_images} \right)}$	0.061 (0.179)	n.a.
R_{merge}	n.a.[d]	0.030 (0.068)
CC*	1.0 (0.99)	1.0 (0.99)
$CC_{1/2}$	0.99 (0.96)	1.0 (0.95)
CC_{ano}	0.48	0.92
I/s*I*	11.9 (4.7)	39.7(7.3)
Completeness (%)	100 (100)	96.6 (93.9)
SFX multiplicity of observations[e]	1383.5 (1310.0)	n.a.
Redundancy[e]	n.a.	16.1 (11.7)

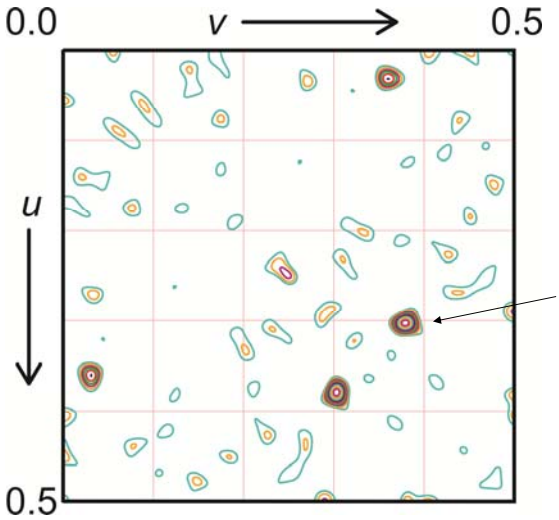
Refinement		
Resolution (Å)	40-2.1	
No. reflections	7287	
R_{work}/ R_{free}	0.230/0.259	
No. atoms		
Protein	992	
Ligand/ion	58 (2 gadoteridol)	
Water	70	
B-factors		
Protein	29.9	
Ligand/ion	39.5	
Water	43.7	
R.m.s deviations		
Bond lengths (Å)	0.008	
Bond angles (º)	1.17	

Experimental phasing (SAD on lysozyme:gadolinium)



$$R_{split} = \left(\frac{1}{\sqrt{2}}\right)\frac{\sum\left|I_{h,odd} - I_{even}\right|}{\frac{1}{2}\sum\left|I_{h,odd} + I_{h,even}\right|}$$

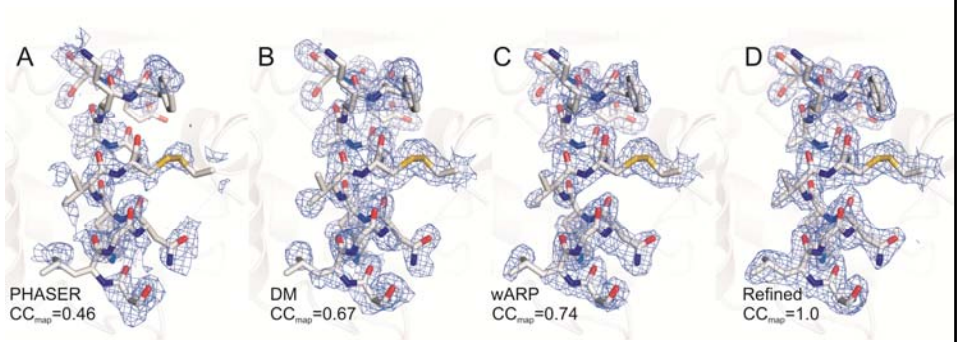
Experimental phasing (SAD on lysozyme:gadolinium)



w=1/4 section of anomalous difference Patterson using 60,000 images

No. Images	Peak height
7,500	3.1
15,000	4.0
30,000	5.7
60,000	6.7

Experimental phasing (SAD on lysozyme:gadolinium)



→First demonstration of *de novo* phasing from FEL data
→60,000 images needed

Perhaps I was a bit strict by insisting on automatic building though...

Detectors, software etc. all still under development, so this will all become easier in the forseeable future

Data quality

I didn't think of this – a referee kindly pointed this out – thank you!

R_{split} is very similar to $R_{merged\text{-}I}$:

$$R_{split} = \left(\frac{1}{\sqrt{2}}\right)\frac{\sum\left|I_{h,odd} - I_{even}\right|}{\frac{1}{2}\sum\left|I_{h,odd} + I_{h,even}\right|} \quad R_{merged-I} = \frac{\sum\left|I_{h,p} - I_{h,q}\right|}{\frac{1}{2}\sum\left|I_{h,p} + I_{h,q}\right|}$$

Which in turn is closely related to R_{pim} : (an analytical form of $R_{merged\text{-}I}$)

$$R_{pim} = \frac{\sum_h \sqrt{\frac{1}{N-1}} \sum_i \left|I_{h,i} - \overline{I_h}\right|}{\sum_h \sum_i I_{h,i}}$$

Diederichs, K. & Karplus, P. A. *Nat. Struct. Biol.* 4, 269–275 (1997)
Weiss, M. S. *J. Appl. Cryst.* **34**, 130-135 (2001)
Weiss, M. S., Sicker, T. & Hilgenfeld, R. *Structure* **9**, 771-777, (2001)
Panjikar, S. & Tucker, P. A. *J. Appl. Cryst.* **35**, 261-266, (2002)

SAD: Prediction of solvability

$$R_{pim} = \frac{\sum_h \sqrt{\frac{1}{N-1}} \sum_i \left| I_{h,i} - \overline{I_h} \right|}{\sum_h \sum_i I_{h,i}}$$

-good indicator of noise

$$R_{ano} = \frac{\sum \left| I_h - I_{-h} \right|}{\frac{1}{2} \sum \left| I_h + I_{-h} \right|}$$

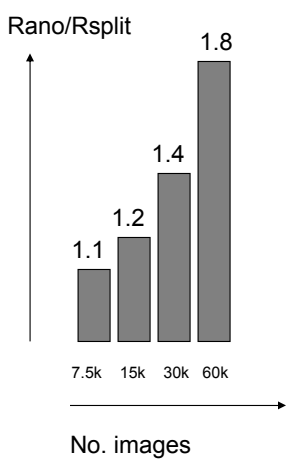
-contains signal as well as noise

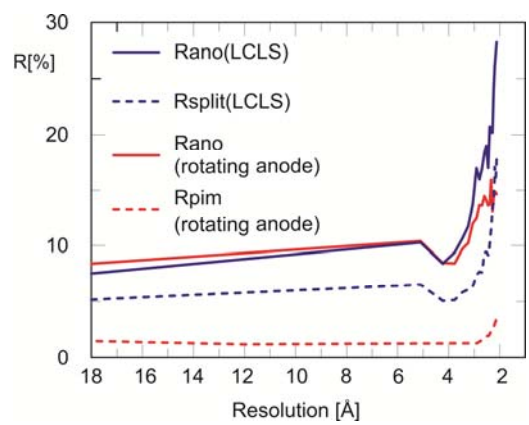
When R_{ano} exceeds R_{pim}, substructure solutions start to appear in trials
At higher R_{ano}/R_{pim} ratios, phasing becomes possible

Tip from referee: does R_{ano}/R_{split} have such a predictive value, too?

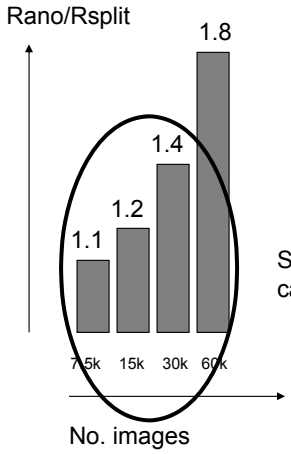
Diederichs, K. & Karplus, P. A. *Nat. Struct. Biol.* 4, 269–275 (1997)
Weiss, M. S. *J. Appl. Cryst.* **34**, 130-135 (2001)
Weiss, M. S., Sicker, T. & Hilgenfeld, R. *Structure* **9**, 771-777, (2001)
Panjikar, S. & Tucker, P. A. *J. Appl. Cryst.* **35**, 261-266, (2002)

Predictive value of Rano/Rsplit?

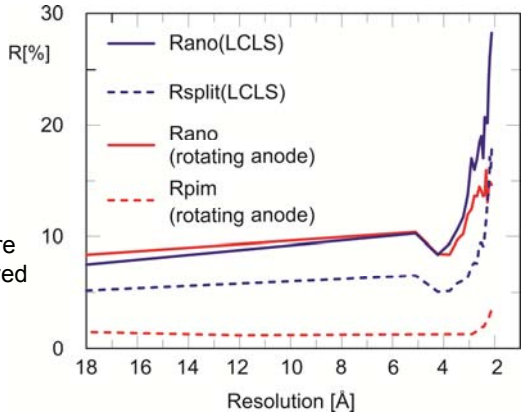




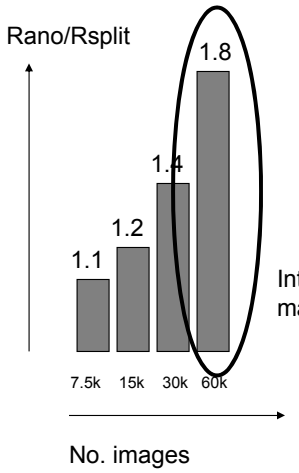
Predictive value of R_{ano}/R_{split}?



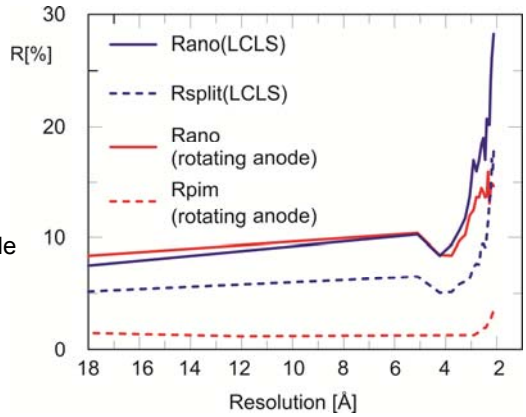
Substructure can be solved



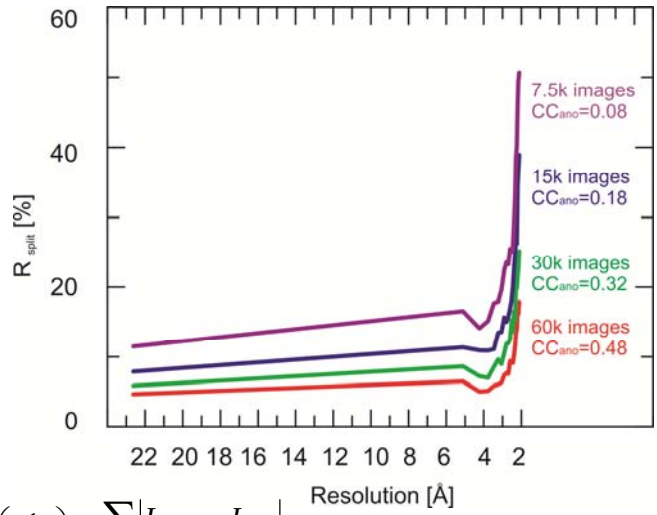
Predictive value of R_{ano}/R_{split}?



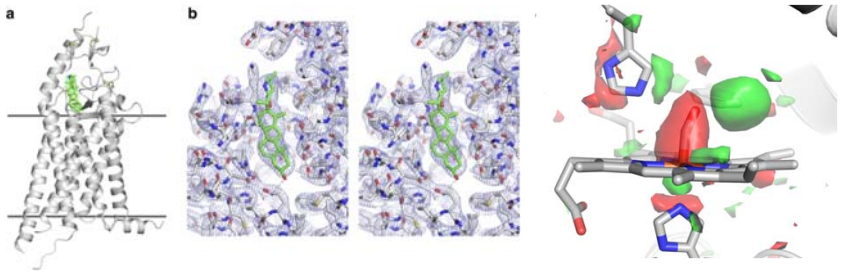
Interpretable map



Is a high Rsplit so bad?
Many SFX structures have high Rsplit...



$$R_{split} = \left(\frac{1}{\sqrt{2}}\right)\frac{\sum\left|I_{h,odd} - I_{even}\right|}{\frac{1}{2}\sum\left|I_{h,odd} + I_{h,even}\right|}$$



Difference Density maps are amazing – you can get away with very low precision and still see real things!

See the paper by Henderson and Moffat (1971): the error level in a difference map is much lower than in e.g. an experimental map

The features in a map scale with $(fom)^2$ – the phases are all-important!

Peter Atkins' textbook "Physical Chemistry" has a nice exercise in which a one-dimensional electron density is calculated from a set of structure factor amplitudes. We can rewrite this exercise to simulate a difference Fourier synthesis.

We prepare two sets of "true"structure factor amplitudes $\left|F_{hkl}^{h}\right|$ and $|F_{hkl}|$. To stay in the right range we choose these amplitudes such that the difference map coefficients will be about 10% of the measured amplitudes.

To these amplitudes we add random errors δ^h and δ to obtain the "observed" amplitudes:

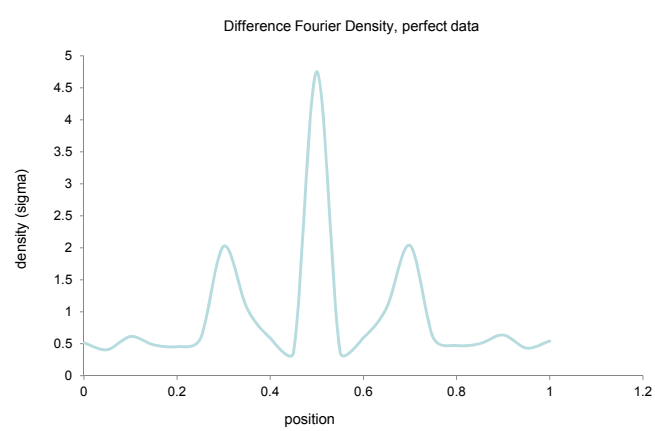
$$\left|F_{hkl}^{h,obs}\right| = \left|F_{hkl}^{h}\right| + \delta^h$$
$$\left|F_{hkl}^{obs}\right| = |F_{hkl}| + \delta$$

We then calculate the observed difference map amplitudes ΔF_{hkl}^{obs}:

$$\Delta F_{hkl}^{obs} = \left|F_{hkl}^{h,obs}\right| - \left|F_{hkl}^{obs}\right|$$

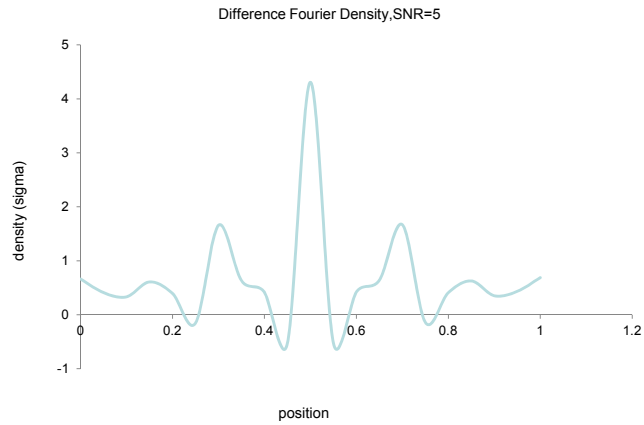
and look at the features in the one-dimensional map as a function of the root-mean-square errors in the observed amplitudes $\sqrt{\langle(\delta^h)^2\rangle}$ and $\sqrt{\langle\delta^2\rangle}$.

Without errors, i.e. when $\sqrt{\langle(\delta^h)^2\rangle} = \sqrt{\langle\delta^2\rangle} = 0$, the map contains one major peak of over 4.5σ at x=0.5 :

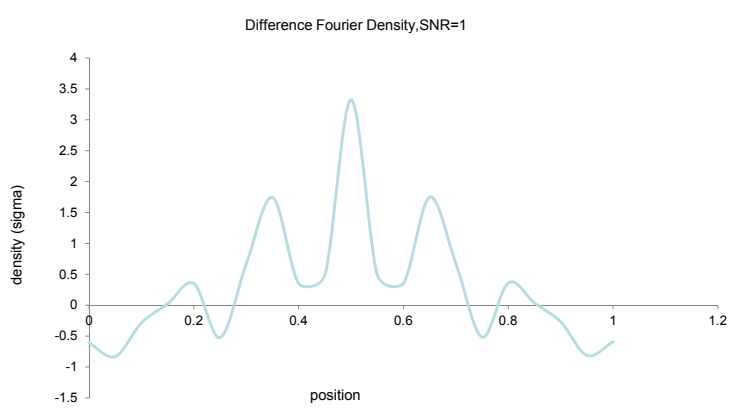


We can now set the errors to the measured structure factors to such values that the root-mean-square error is about 1/5 of the root-mean-square value of the true values of ΔF_{hkl} , i.e. a signal-to-noise ratio of 5:

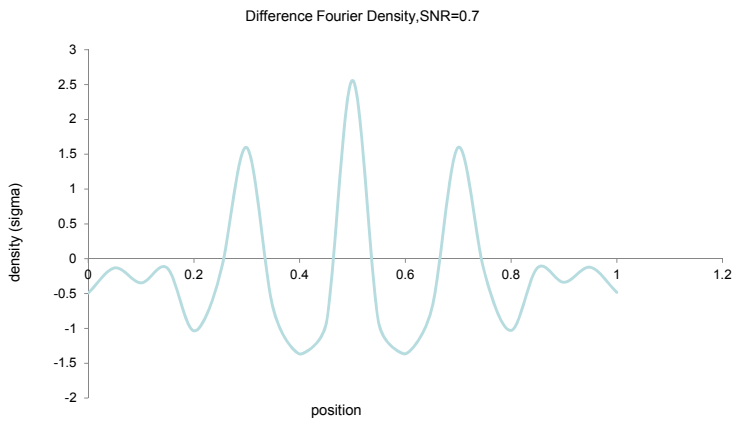
$$\sqrt{\langle(\delta^h)^2\rangle} = \sqrt{\langle\delta^2\rangle} = 0.2\sqrt{\langle\Delta F_{hkl}^2\rangle}$$

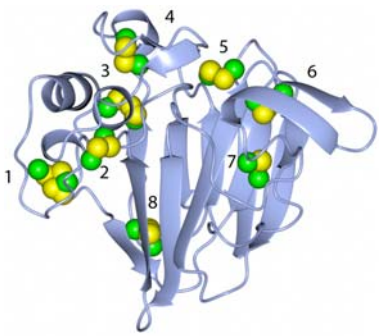


How far can we push this? Let us now put the signal-to-noise ratio at 1 and look at the map:

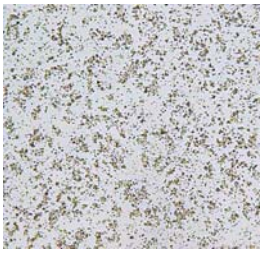


Even at a signal-to-noise ratio of 0.7, the magic of the Fourier transform still reveals the main features of the density:

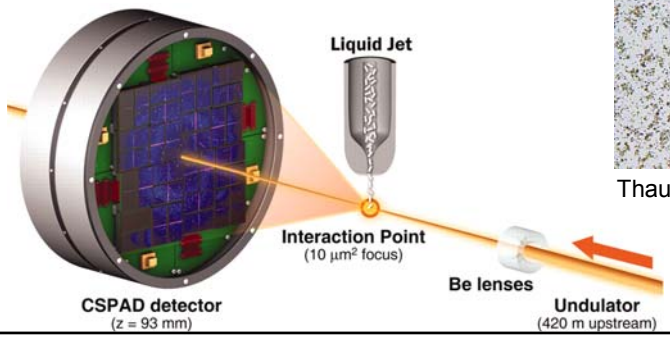




Thaumatin – an S-SAD benchmark
200 residues, 17 sulfur atoms, high res.
Signal now only ~ 1% but:
SOLVING THAUMATIN USING S-SAD IS A SIMPLE EXERCISE IN CONVENTIONAL CRYSTALLOGRAPHY

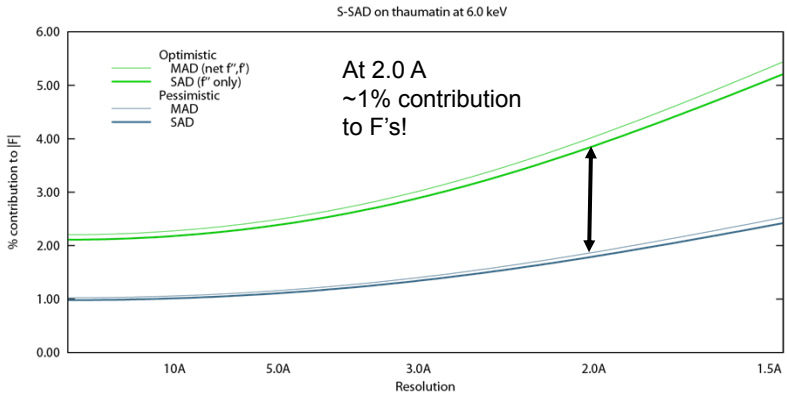


Thaumatin microxtals

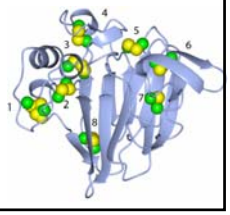


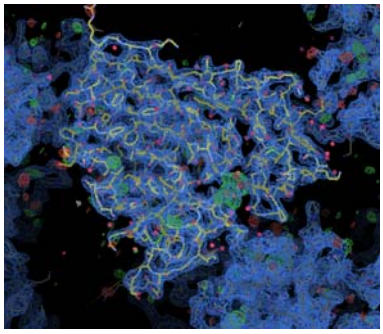
Karol Nass

Anomalous differences from S



This is a much weaker signal than the Gd anomalous signal



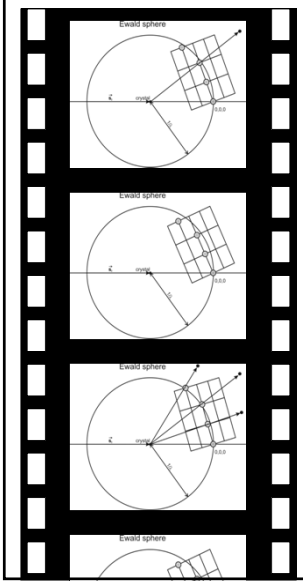


Successful S-SAD phasing from SFX data
…but it needed 900k images…
Why do we need so many?
-is it the partiality of the measurements?
(*remember I am spoiled by programs such as XDS)*
-is it the shot-to-shot variations in the beam?
(remember I am spoiled by synchrotron beams..)
-is it the experimental detectors?
(remember I am spoiled by highly developed, almost perfect detectors such as the Pilatus etc.)

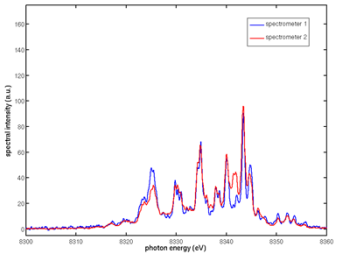
Data set statistics	
Data collection time	15.4 hours
Number of frames collected	6,662,315
Hit rate	22.6%
Number of frames indexed	871,859
Indexing rate	59.9%
Resolution range	36.7 – 2.2 A (2.3-2.2 A)
R_{split}	5.1% (11.9%)
CC_{ano}	0.225 (0.116)

Karol Nass

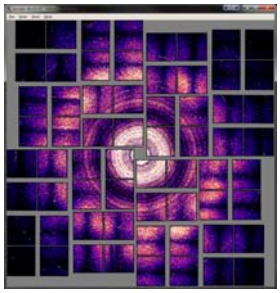
Why do we need so many images?
Is it one or more of these three factors?
Is it something else?



LCLS spectra



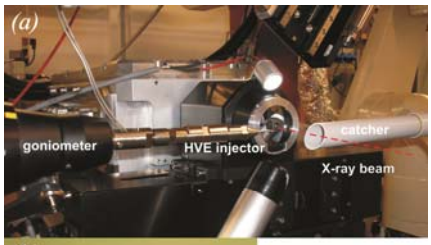
(early) CSPAD detector



I fear I have accused all of these factors (and others) of being the cause for the need for thousands of images at one point or another…

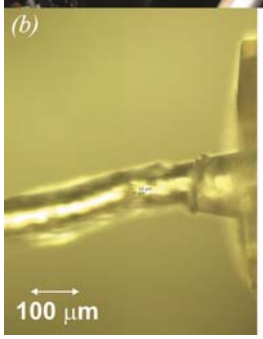
A high viscosity extrusion injector that works in vacuum as well as in air
for use at FELs and at synchrotrons

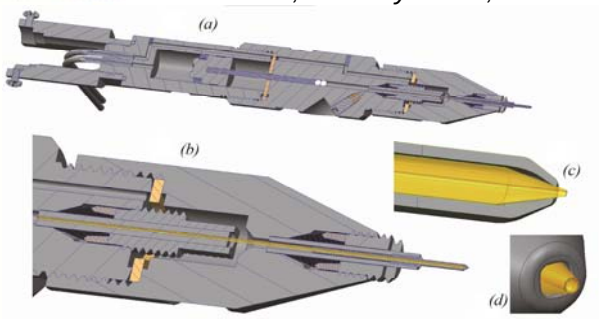
(Sabine Botha, Bruce Doak)



-Serial data collection from small xtals
-Can use the whole acceptable dose/xtal for a single shot
-Low sample consumption
-Soluble as well as membrane proteins
-Can work with LCP, grease, polymers, etc.

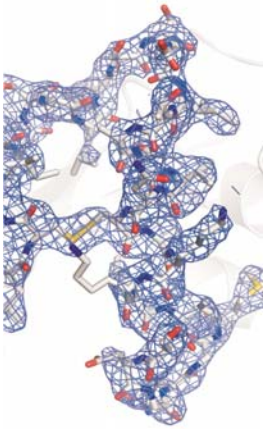
Botha et al., *Acta Cryst D***71**, 387-397





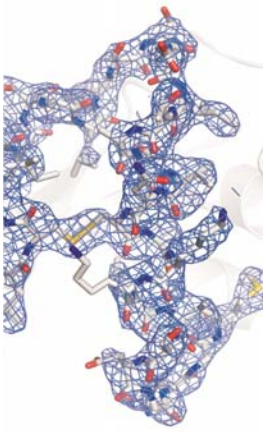
Synchrotron Serial Crystallography on Lysozyme in LCP
MIRAS phasing using an iodine and a gold derivative (example: I-derivative)
Hit rate around 100%, indexing rate around 25%
Reasonable map but again required >40,000 images
With a Pilatus detector and a synchrotron beam…

	CrystFEL	
Res. range	15-2.5 (2.6-2.5) A	
No. indexed	42,115	
Completeness	100% (100%)	
Multiplicity	1107 (816)	
I/σI	12.4 (4.4)	
R_{split}/R_{meas}	0.259 (0.335)	
CC*	0.923 (0.905)	
CC_{ano}	0.391 (0.083)	
FOM	0.49	
PP iso (cen/acen)	1.748 / 1.318	
PP ano	1.039	
% auto-built	~50%	



Synchrotron Serial Crystallography on Lysozyme in LCP
MIRAS phasing using an iodine and a gold derivative
nXDS: postrefinement, profile fitting → partiality taken into account explicitly
Kabsch, *Acta Cryst* D**70**,2204-2216
→Compare pure Monte Carlo with the nXDS approach

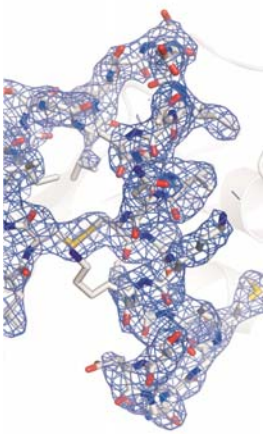
	CrystFEL	nXDS
Res. range	15-2.5 (2.6-2.5) A	19-2.7 (2.9-2.7)
No. indexed	42,115	9,098
Completeness	100% (100%)	99% (99%)
Multiplicity	1107 (816)	119 (92)
I/σI	12.4 (4.4)	32.2 (15.5)
R_{split}/R_{meas}	0.259 (0.335)	0.264 (0.581)
CC*	0.923 (0.905)	0.996 (0.973)
CC_{ano}	0.391 (0.083)	0.743 (0.479)
FOM	0.49	0.44
PP iso (cen/acen)	1.748 / 1.318	1.438 / 1.043
PP ano	1.039	1.077
% auto-built	~50%	>90%



Synchrotron Serial Crystallography on Lysozyme in LCP
MIRAS phasing using an iodine and a gold derivative

-nXDS data has much stronger signal with fewer images than CrystFEL

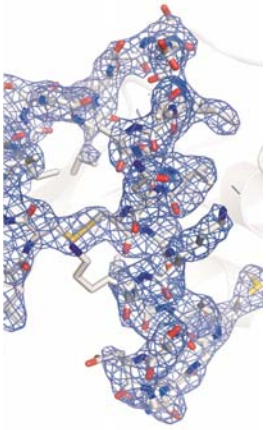
	CrystFEL	nXDS
Res. range	15-2.5 (2.6-2.5) A	19-2.7 (2.9-2.7)
No. indexed	**42,115**	**9,098**
Completeness	100% (100%)	99% (99%)
Multiplicity	1107 (816)	119 (92)
I/σI	12.4 (4.4)	32.2 (15.5)
R_{split}/R_{meas}	0.259 (0.335)	0.264 (0.581)
CC*	0.923 (0.905)	0.996 (0.973)
CC_{ano}	**0.391 (0.083)**	**0.743 (0.479)**
FOM	0.49	0.44
PP iso (cen/acen)	1.748 / 1.318	1.438 / 1.043
PP ano	1.039	1.077
% auto-built	~50%	>90%



Synchrotron Serial Crystallography on Lysozyme in LCP
MIRAS phasing using an iodine and a gold derivative

-Phases from CrystFEL better than from nXDS (more images?)

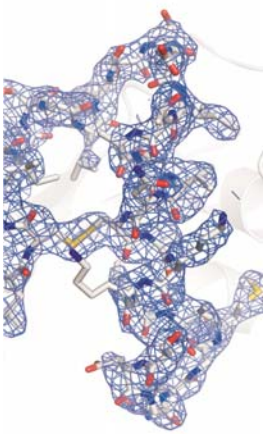
	CrystFEL	nXDS
Res. range	15-2.5 (2.6-2.5) A	19-2.7 (2.9-2.7)
No. indexed	42,115	9,098
Completeness	100% (100%)	99% (99%)
Multiplicity	1107 (816)	119 (92)
I/σI	12.4 (4.4)	32.2 (15.5)
R_{split}/R_{meas}	0.259 (0.335)	0.264 (0.581)
CC*	0.923 (0.905)	0.996 (0.973)
CC_{ano}	0.391 (0.083)	0.743 (0.479)
FOM	**0.49**	**0.44**
PP iso(cen/acen)	**1.748 / 1.318**	**1.438 / 1.043**
PP ano	**1.039**	**1.077**
% auto-built	~50%	>90%



Synchrotron Serial Crystallography on Lysozyme in LCP
MIRAS phasing using an iodine and a gold derivative

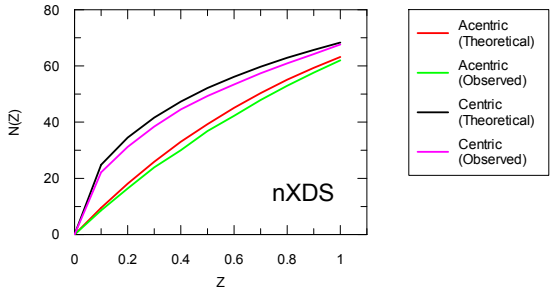
-Autobuilding with ARP/wARP easier from nXDS intensities (I-distributions?)

	CrystFEL	nXDS
Res. range	15-2.5 (2.6-2.5) A	19-2.7 (2.9-2.7)
No. indexed	42,115	9,098
Completeness	100% (100%)	99% (99%)
Multiplicity	1107 (816)	119 (92)
I/σI	12.4 (4.4)	32.2 (15.5)
R_{split}/R_{meas}	0.259 (0.335)	0.264 (0.581)
CC*	0.923 (0.905)	0.996 (0.973)
CC_{ano}	0.391 (0.083)	0.743 (0.479)
FOM	0.49	0.44
PP iso (cen/acen)	1.748 / 1.318	1.438 / 1.043
PP ano	1.039	1.077
% auto-built	**~50%**	**>90%**



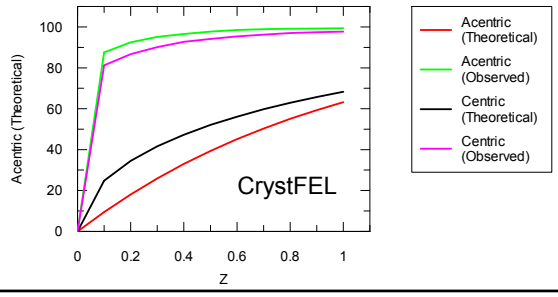
More "conventional" intensity distributions from nXDS data – reason for easier autobuilding?

a)



B-factor 31.3 Å^2

b)



B-factor 42.0 Å^2

Conclusions

-SFX data can be phased from anomalous differences **even from the very weak signal provided by endogenous sulfur with many images**

-**Serial Synchrotron Crystallograph**y data can be phased also **but again many images needed** → might suggest neither FEL beam, nor CSPAD detectors cause this need

-**nXDS approach appears to greatly reduce the need for many images**

→ looking forward to using postrefinement in the new CrystFEL 0.6.0

→ if nXDS can be made to accept/index more images even bigger gains possible than nXDS already yields

People

LCLS

Sebastien Boutet
Marc Messerschmidt
Garth Williams
Mengning Liang
Andy Aquila
Jason Koglin

SLS

Meitian Wang
Florian Dworkowski
Ezequiel Panepucci

MPI Heidelberg

Lutz Foucar
Sabine Botha
Karol Nass
Beatrice Latz
Robert Shoeman
Wolfgang Kabsch
Bruce Doak
Ilme Schlichting
MPImF Mechanical Workshop