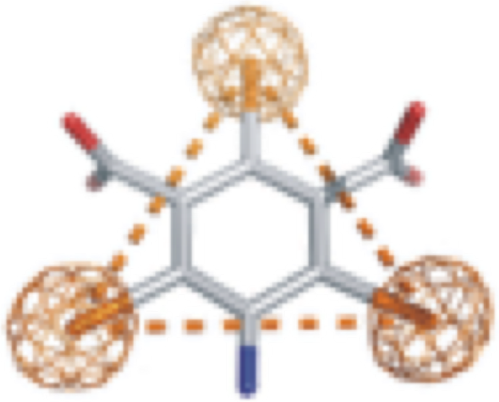
Experimental phasing with SHELX

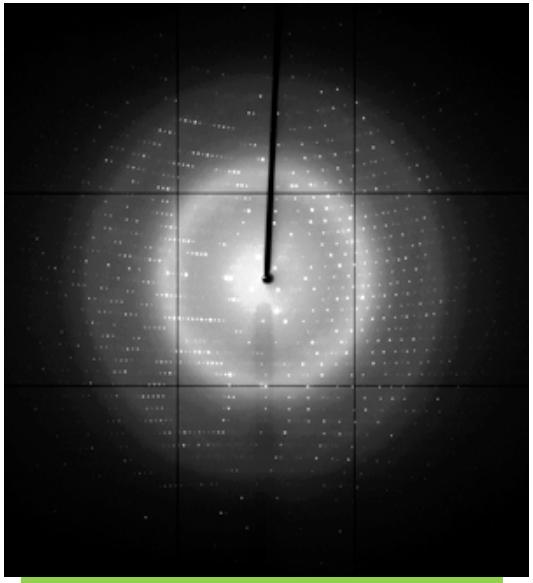


Isabel Usón, George M. Sheldrick

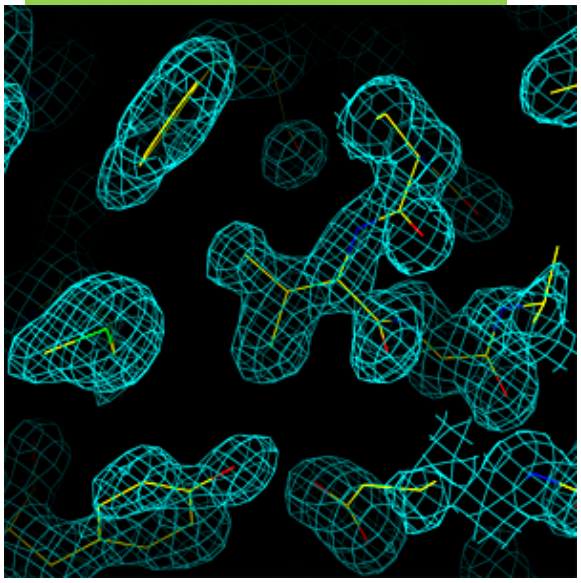
http://shelx.uni-goettingen.de

uson@ibmb.csic.es

SHELX Package (Macromolecules) George Sheldrick



The phase problem



Experimental Phasing

SHELXC
Data preparation

SHELXD
Substructure determination

SHELXE
Phasing, density modification, autotracing

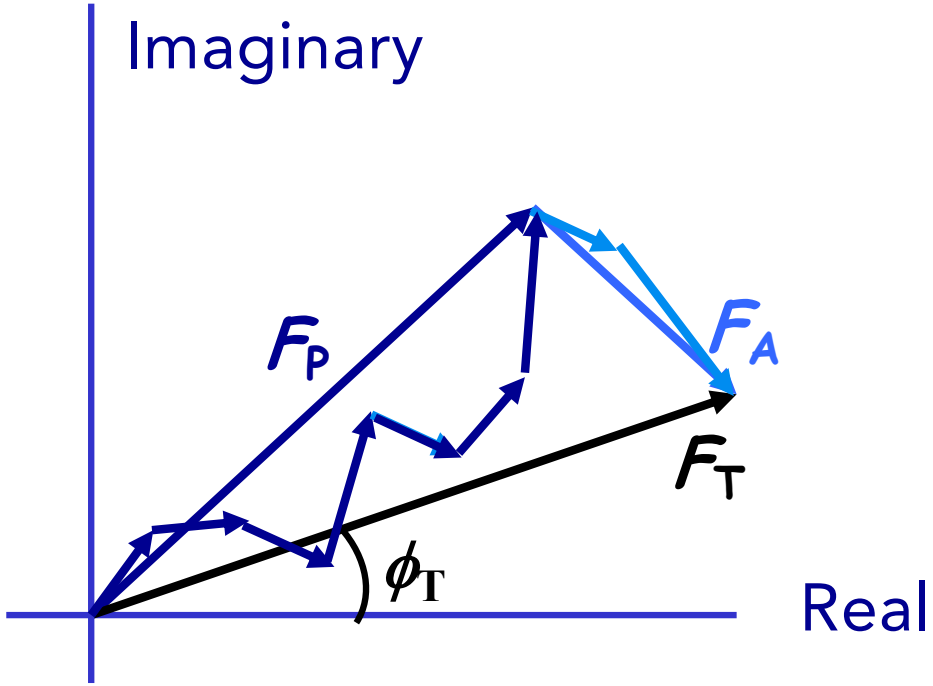
Interfaces
- HKL2MAP (Schneider, Pape)
- CCP4
- XDSGUI

SHELXL Refinement
- s.u. twins, disorder

Molecular Replacement
- Pipelines

ARCIMBOLDO

Experimental phasing



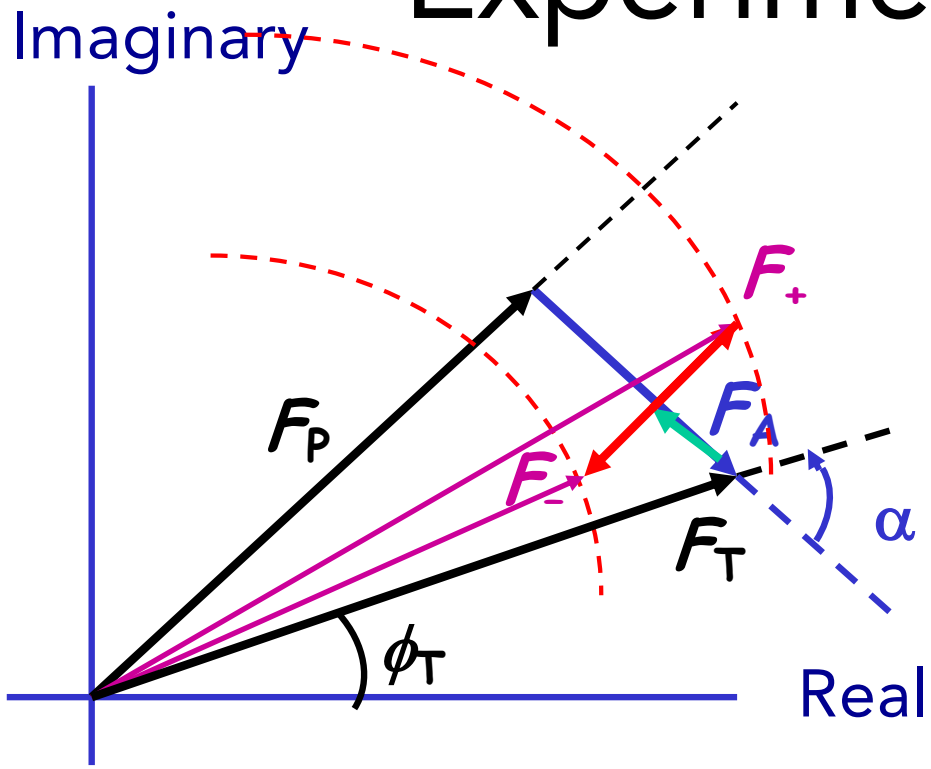
$$F_T = \Sigma F_i$$

$$F_T = F_P + F_A$$

Each structure factor is composed of the contributions of all atoms in the unit cell

Differentiate a small molecule in the macromolecule

Experimental phasing



$$F_T = F_P + F_A$$

$$f = f^0 + f' + i\,f''$$

- Differentiate a small molecule in the macromolecule and use as **marker atoms** to provide **reference phases**

Calculate phases ϕ_T of the full structure by: $\phi_T = \phi_A + \alpha$

ϕ_A is the calculated phase of the heavy atom substructure

α can be estimated from the experimental data

Experimental phasing with SHELXC/D/E

- In practice, phase determination requires the following stages:

  C: (Scale data), compute ΔF or F_A and estimate α

  D: Locate the heavy atoms

  (Refinement of heavy atoms) calculate ϕ_A

  E: Calculate start protein phases using

  $\phi_T = \phi_A + \alpha$

  Improve these phases by density modification

Experimental phasing methods

SAD (native S): $|F_T| \sim \frac{1}{2} (|F^+| + |F^-|)$ and $|F_A|\sin\alpha \sim \frac{1}{2} (|F^+| - |F^-|)$

SIR (RIP): $|F_T| = |F_{nat}|$ and $|F_A|\cos\alpha \sim |F_{deriv}| - |F_{nat}|$

SIRAS: By combining SAD and SIR, we obtain $|F_T|$, $|F_A|$ and α

MAD: Data from at least two wavelengths give $|F_T|$, $|F_A|$ and α

SHELXD uses $|F_A|$ (approximated as $|F_A|\sin\alpha$ or $|F_A|\cos\alpha$ if necessary) to find the heavy atom sites (substructure).

The resolution may be truncated to reduce noise

To calculate and refine the protein phases, $|F_T|$ and the shift angle α are needed

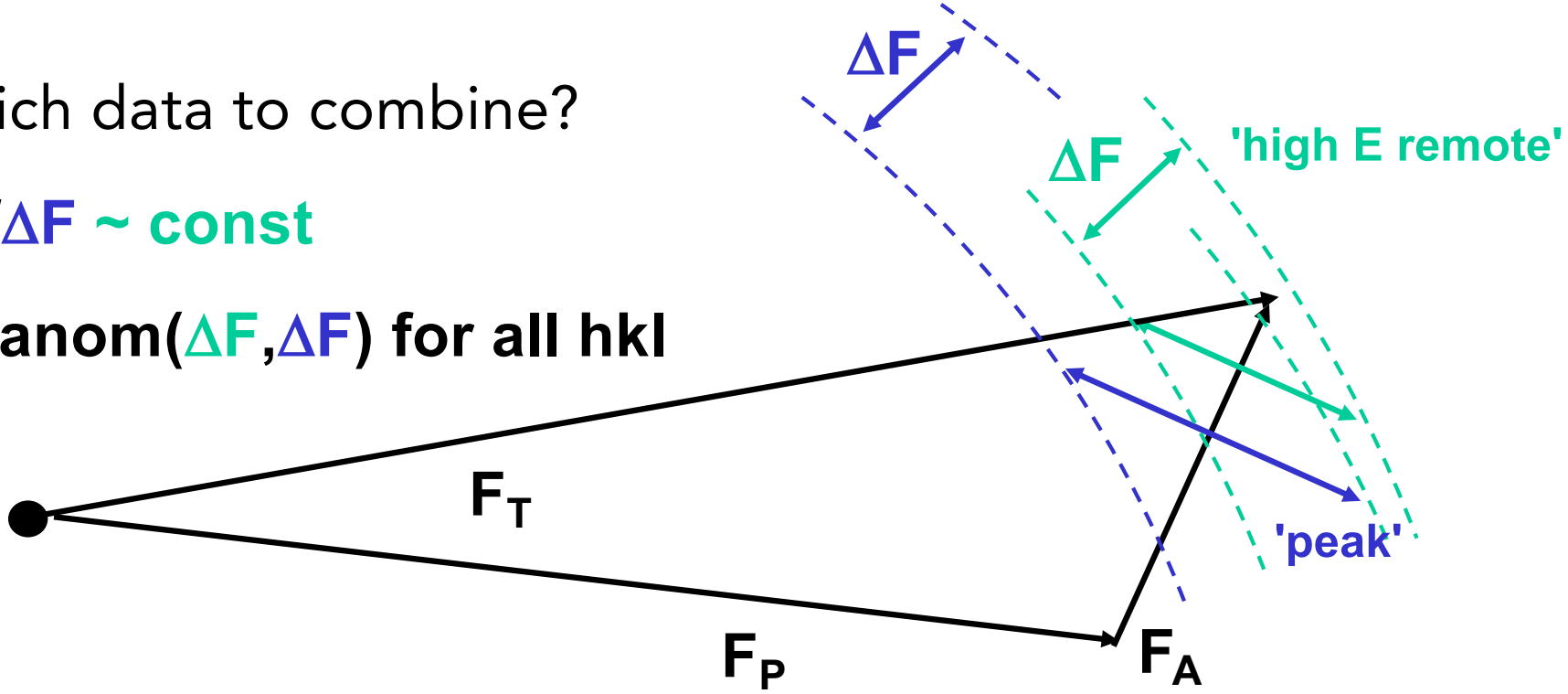
Multiple isomorphous replacement with anomalous scattering (MIRAS) not in SHELX

Data preparation in SHELXC

- Which data to combine?

→ **ΔF/ΔF ~ const**

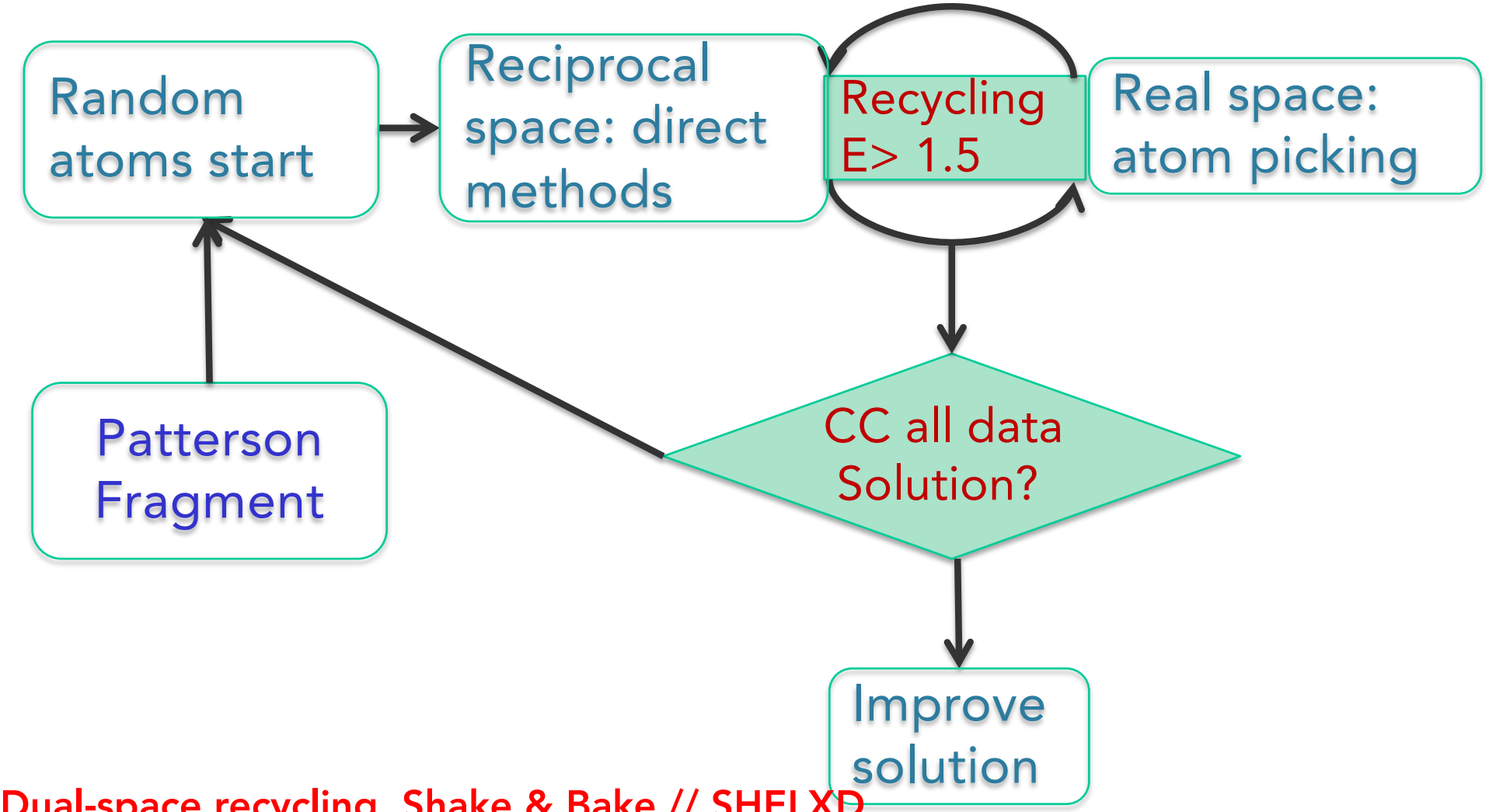
→ **CCanom(ΔF,ΔF) for all hkl**



- To what resolution limit?

  SHELXD and SHELXE use normalized amplitudes E, so high resolution will be upweighted

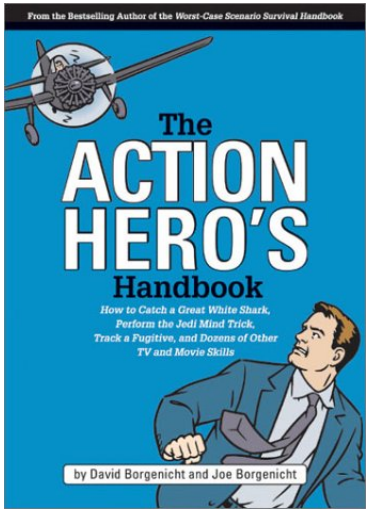
SHELXD: direct methods for macromolecules



Dual-space recycling, Shake & Bake // SHELXD

Sheldrick, Gilmore, Hauptman, Weeks, Miller & Usón (2012), *International Tables for Crystallography Vol. F*, eds. Arnold & Rossmann, pp. 333-351

The quiz problem 4



Why does dual space recycling work?

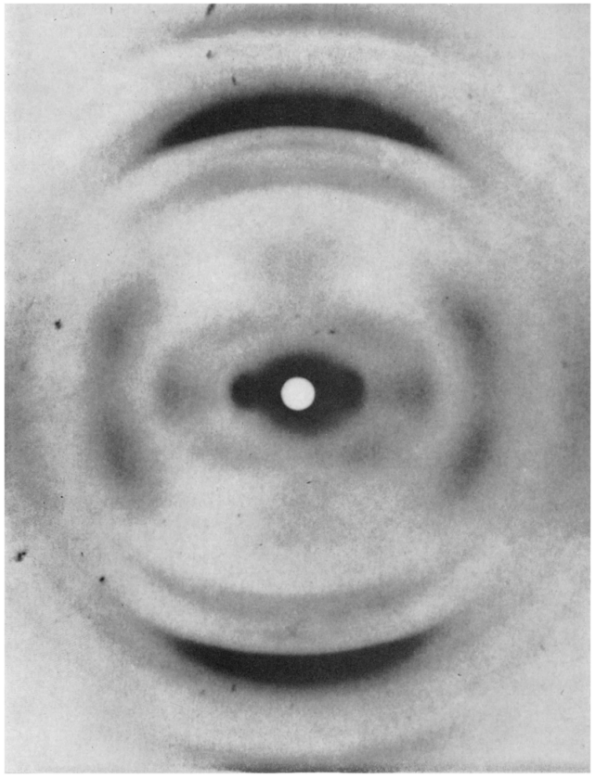
- E- You gain information by switching between real and reciprocal space
- O- you apply differently atomicity constraints in either space, thus modifying phases and maps

- Write down letter from correct answer!

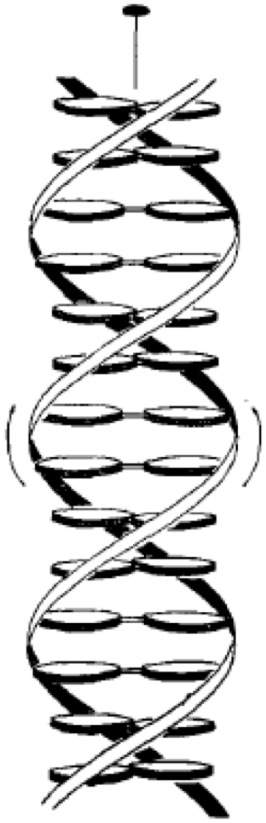
Structures solved *ab initio* by SHELXD

Compound	Spacegroup	N(+solv)	HA	dÅ
Actinomycin X2	P1	273(305)	-	0.9
Actinomycin Z3	$P2_12_12_1$	186(307)	2Cl	0.96
Vancomycin	$P4_32_12$	202(312)	6Cl	1.09
Actinomycin D	P1	270(314)	-	0.94
Ristocetin A	$P2_1$	294(420)	-	1.03
Hirustasin	$P4_32_12$	402(467)	10S	1.2
Cyclodextrin	$P2_1$	448(467)	-	0.88
Decaplanin	$P2_1$	448(635)	4Cl	1.00
PolyA RNA	$P4_12_12$	478(572)	22P	1.00
Cyclodextrin	P1	483(562)	-	1.00
Bucandin	C2	516(634)	10S	1.05
Amylose CA26	P1	572(719)	-	1.1
Viscotoxin B2	$P2_12_12_1$	722(812)	12S	1.05
Mersacidin	$P3_2$	750(826)	24S	1.04
Feglimycin	P65*	828(1026)	-	1.10
Tsuchimycin	P1	1069(1283)	24Ca	1.00
rc-WT CvHiPIP	$P2_12_12_1$	1264(1599)	8Fe	1.2
Cytochrome c3	$P3_1$	2024(2208)	8Fe	1.2

Poly rA: the other double helix

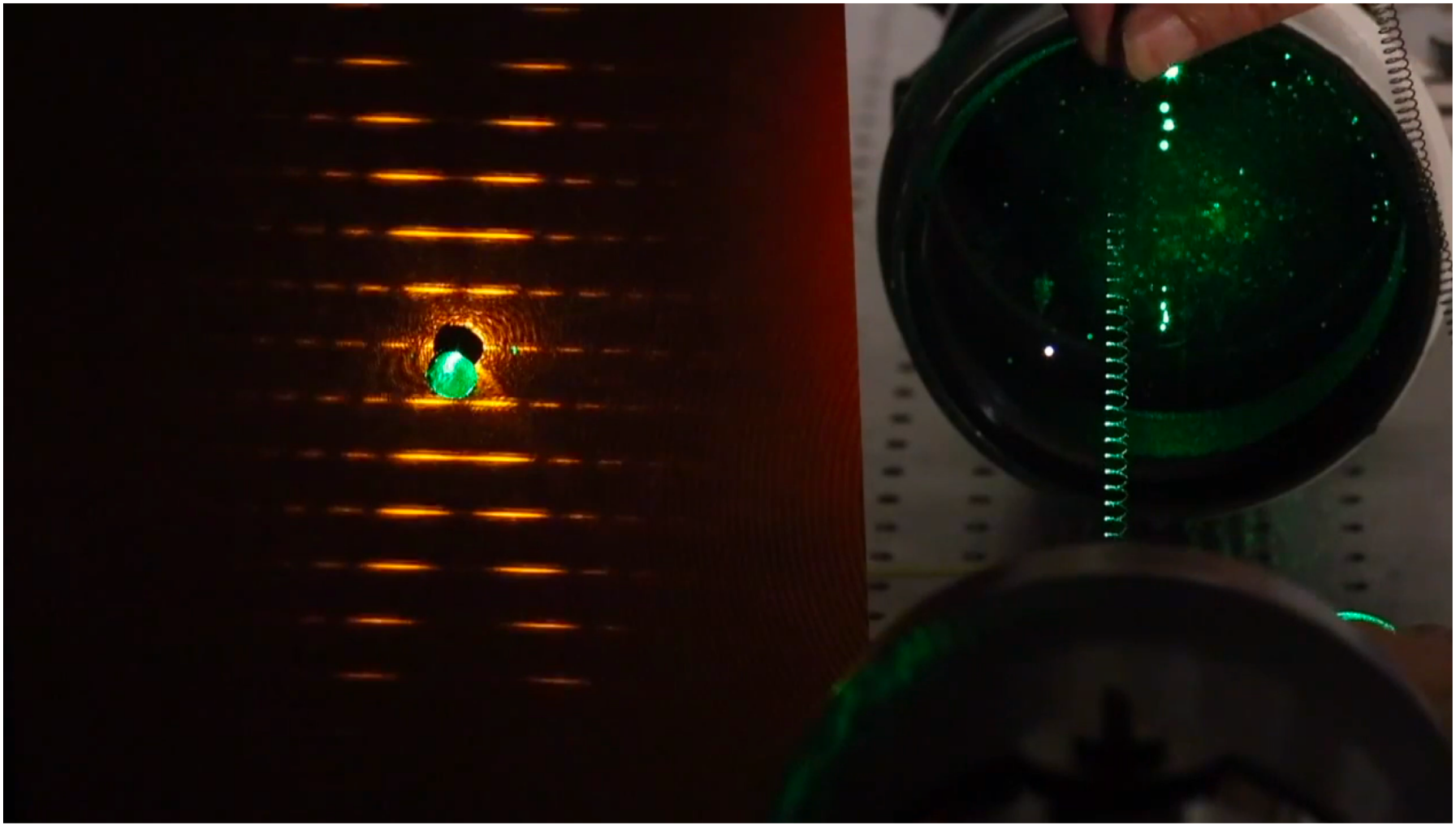


!?



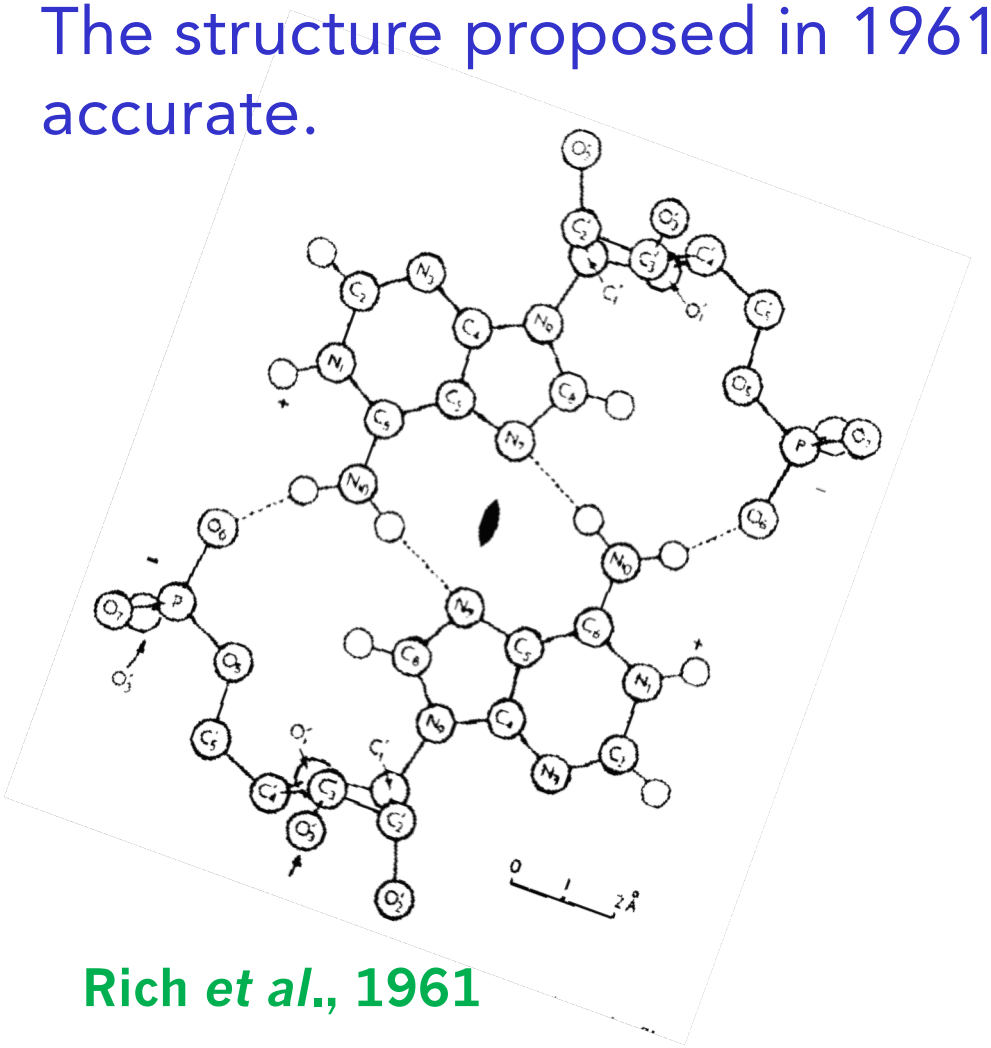
The solution of an RNA structure at atomic resolution illustrates SHELXD's original purpose. In 1961 Rich, Davies, Watson & Crick (*J. Mol. Biol.* **3** (1961) 71-86) proposed a *parallel double-helix* structure for polyadenylic acid, poly(rA), based on fibre diffraction photographs, using similar reasoning to that used for the *antiparallel double helix* of DNA

Poly rA: the other double helix

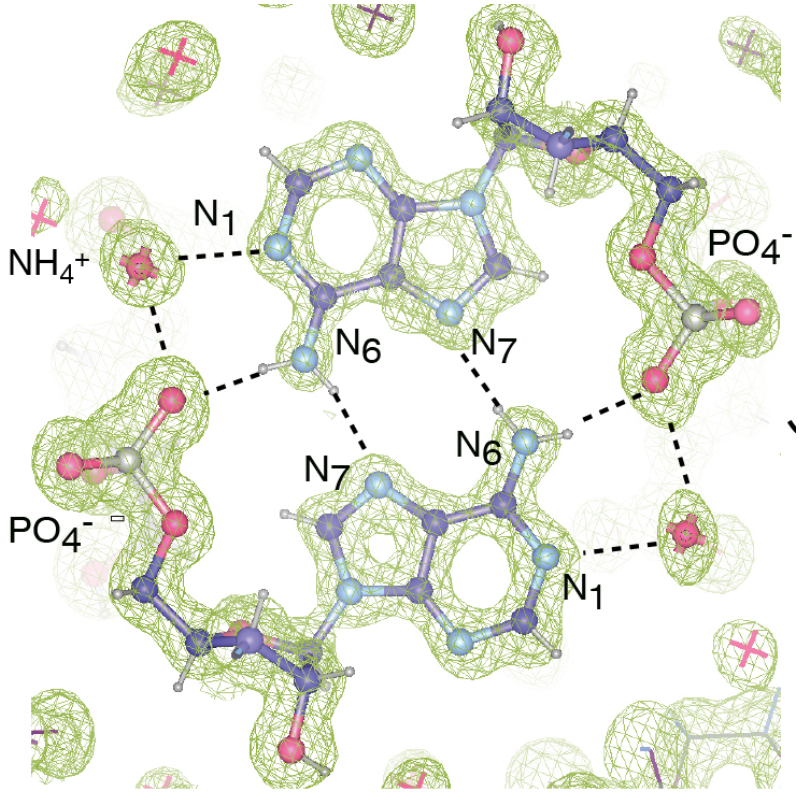


50 years later: the $(rA)_{11}$ structure (Nozhat Safaee, Kalle Gehring McGill)

- The structure proposed in 1961 was remarkably accurate.



Rich *et al*., 1961

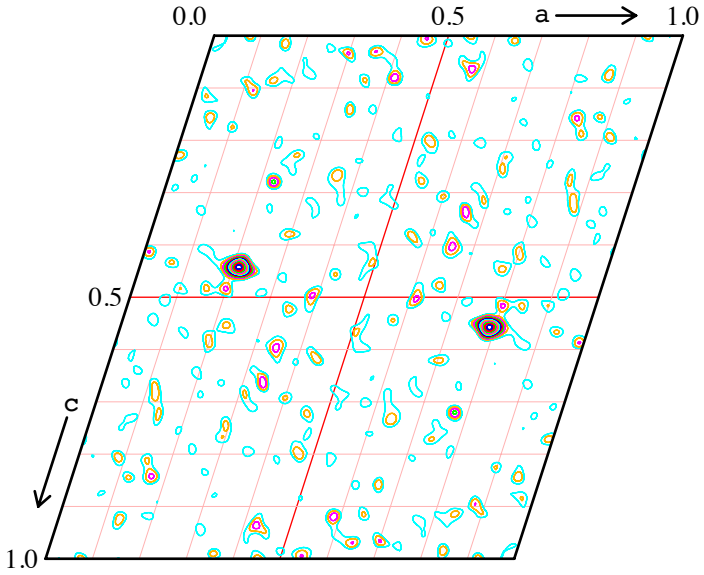


Safaee *et al*., 2013

Harker sections

Vectors between symmetry equivalents often have one or more constant coordinates

e.g. in C2 the atoms at x,y,z and –x, y,–z give rise to a Patterson peak at 2x,0,2z. This leads to an accumulation of peaks in the Harker section at y = 0



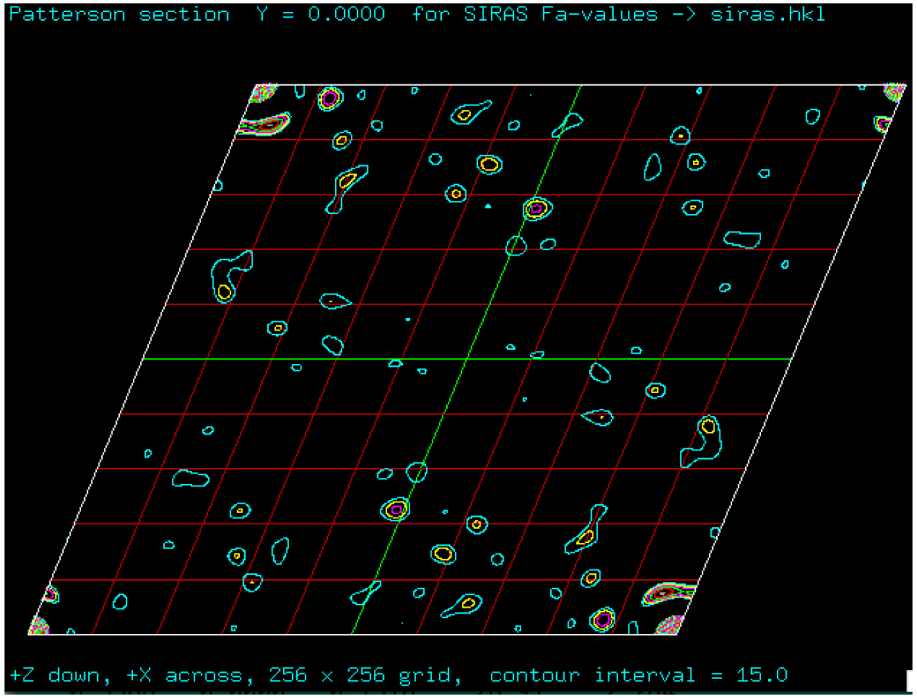
Harker sections can aid space group determination, (systematic absences)

Promising MAD FA or SAD or SIR ΔF when the Harker sections contain clear peaks of significant height (say $>8\sigma$). A good sign is consistent peaks in SIR and SAD Harker sections

Harker sections

Vectors between symmetry equivalents often have one or more constant coordinates

e.g. in C2 the atoms at x,y,z and –x, y,–z give rise to a Patterson peak at 2x,0,2z. This leads to an accumulation of peaks in the Harker section at y = 0



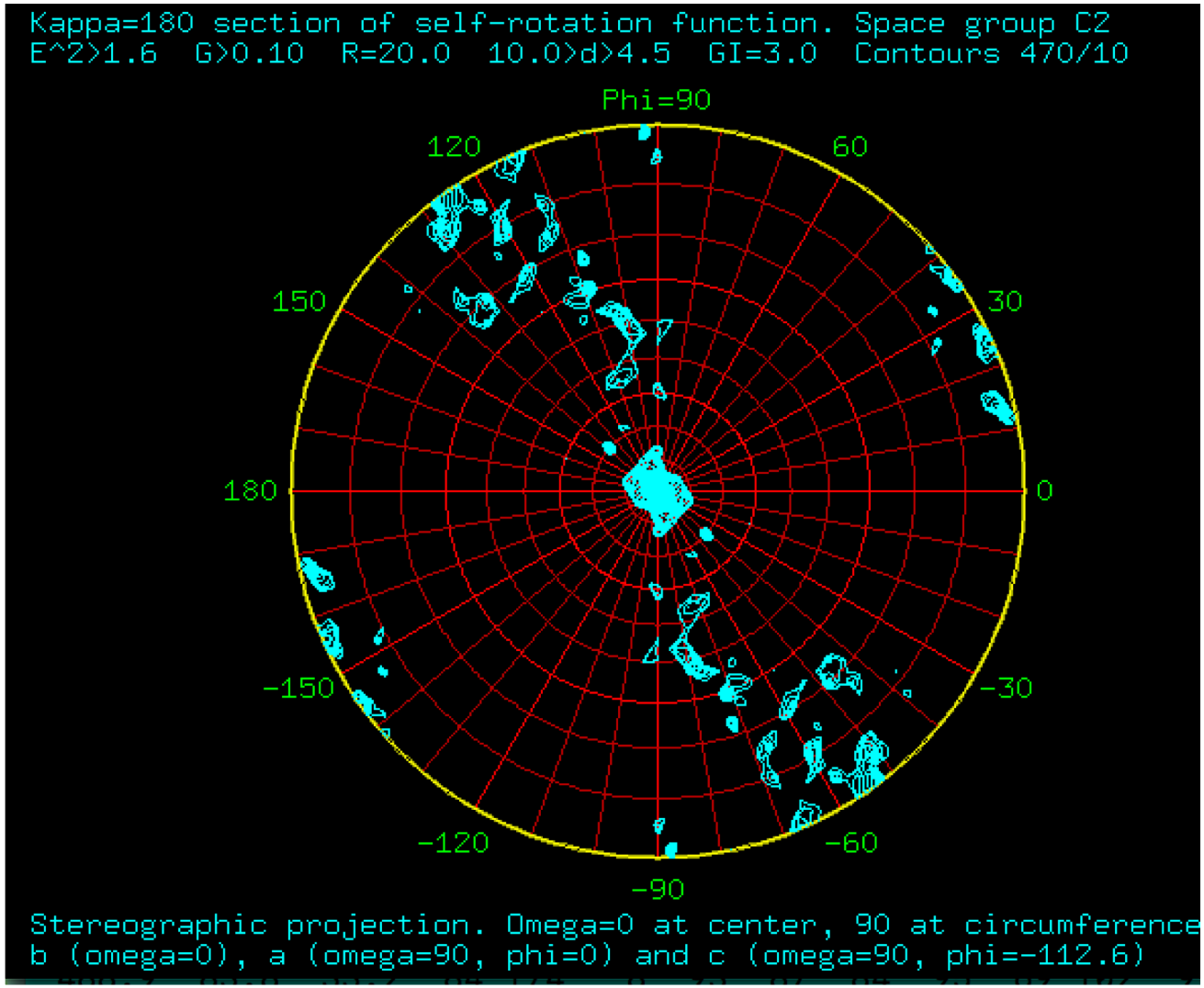
Harker sections can aid space group determination, (systematic absences)

Promising MAD FA or SAD or SIR ΔF when the Harker sections contain clear peaks of significant height (say $>8\sigma$). A good sign is consistent peaks in SIR and SAD Harker sections

Self-rotation

Same structure: a dimer with two domains/ monomer

Structure of Windbeutel, Qingjun Ma, Institute of Oceanology, CAS



Hg substructure in AtsK, Ilka Müller

Peak	Self	Cross-vectors (minimum distance/PSMF)							
1	43.5								
	17.1	Hg1							
2	42.9	21.4	Hg2#						
	14.3	14.2							
3	81.8	30.3	35.1						
	0.0	10.9	14.6						
4	44.0	26.2	13.2	32.7					
	17.4	1.7	18.3	13.4					
5	42.2	21.2	24.8	48.1	21.2				
	12.4	13.8	13.5	7.6	7.3				
6	83.2	47.7	32.5	37.5	31.9	29.9			
	10.2	3.6	10.4	12.4	6.3	0.0			
7	78.4	32.0	38.4	33.9	29.3	35.2	41.6		
	4.4	3.9	23.2	5.7	9.0	3.9	7.7		
8	78.3	35.1	29.3	41.6	37.9	31.7	33.9	34.7	
	19.5	16.8	10.0	3.5	11.2	2.4	5.5	13.0	
9	73.6	43.9	46.0	26.3	46.0	43.5	26.2	27.2	27.4
	36.4	23.4	11.0	11.3	13.3	13.8	19.2	8.1	6.2
10	77.5	46.3	43.8	27.6	43.7	46.0	27.5	26.0	26.3
	15.8	10.2	25.1	15.9	15.1	8.4	20.8	7.7	3.8

30Å distance, Patterson FOM

Hg3

Hg4

Hg5#

Hg6#

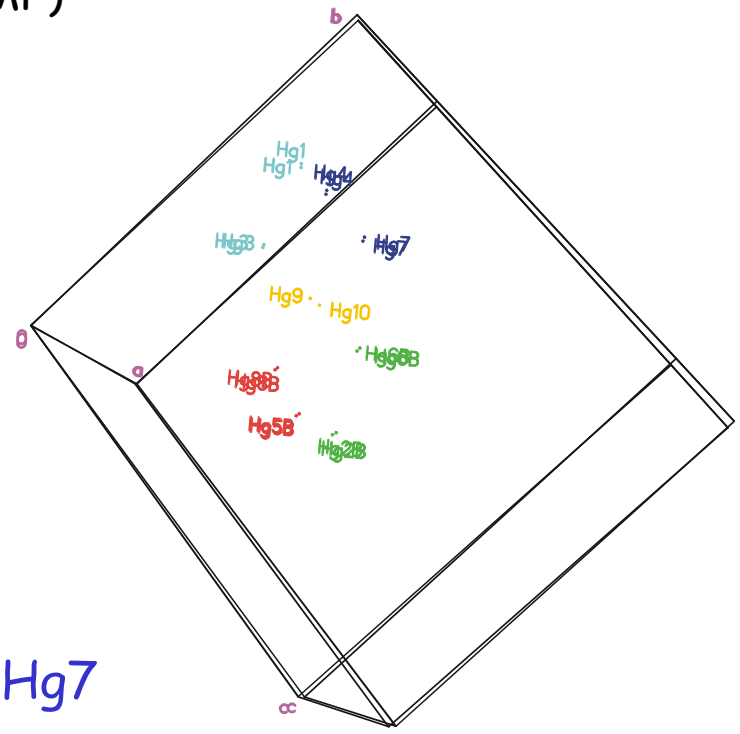
Hg7

Hg8#

Hg9

7.9
0.0

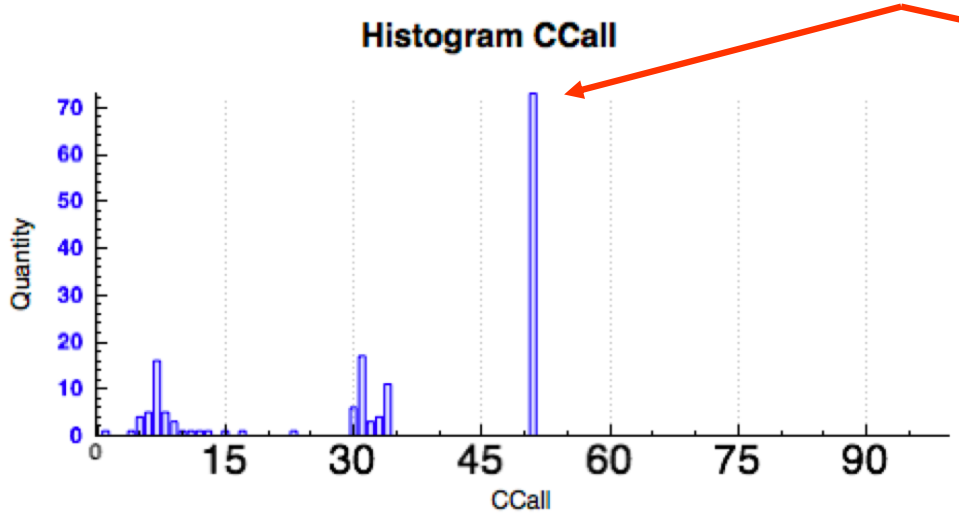
Hg10#

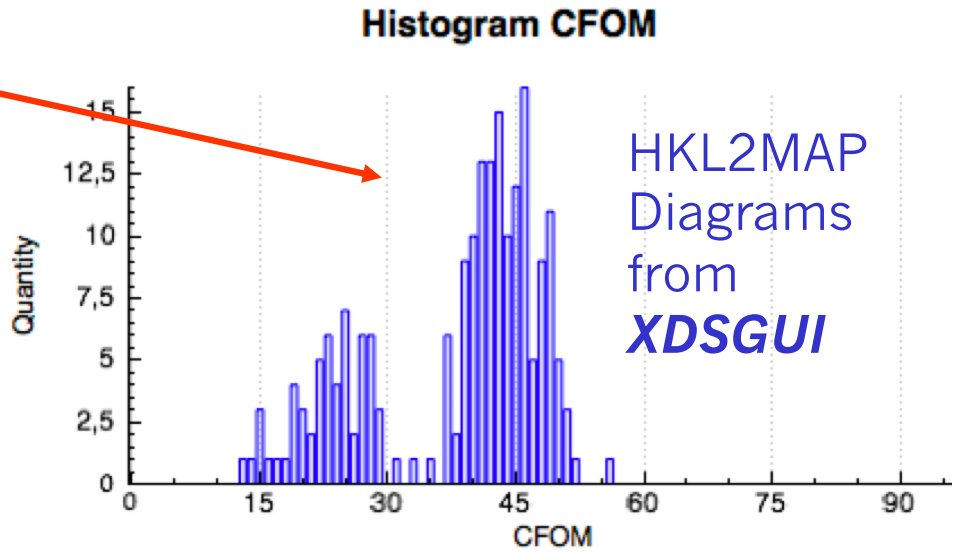


-x, 0.5+y, 0.5-z

SHELXD histograms and occupancies for two structures: Se and Br^- derivative

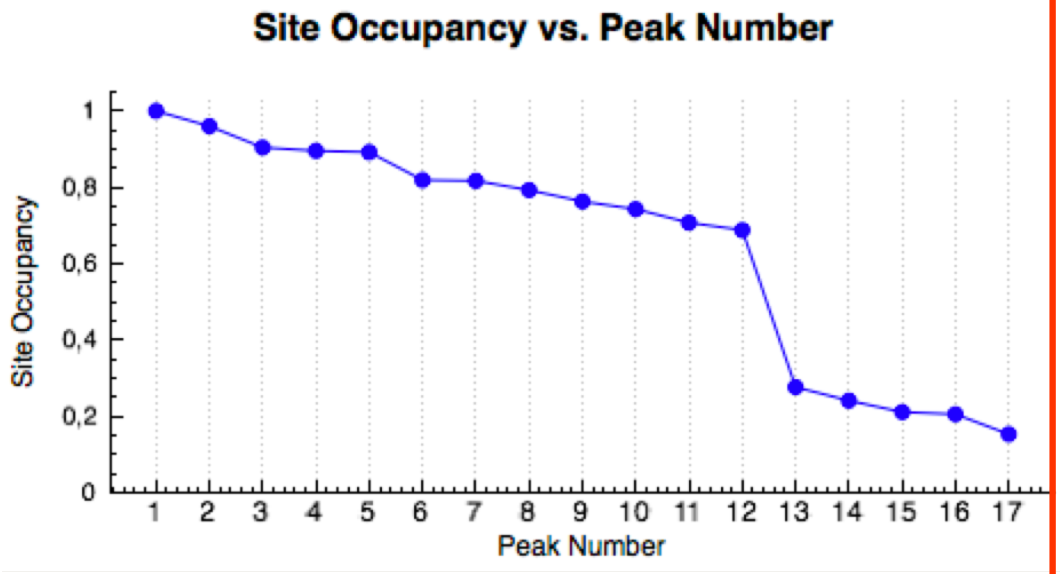
Solutions

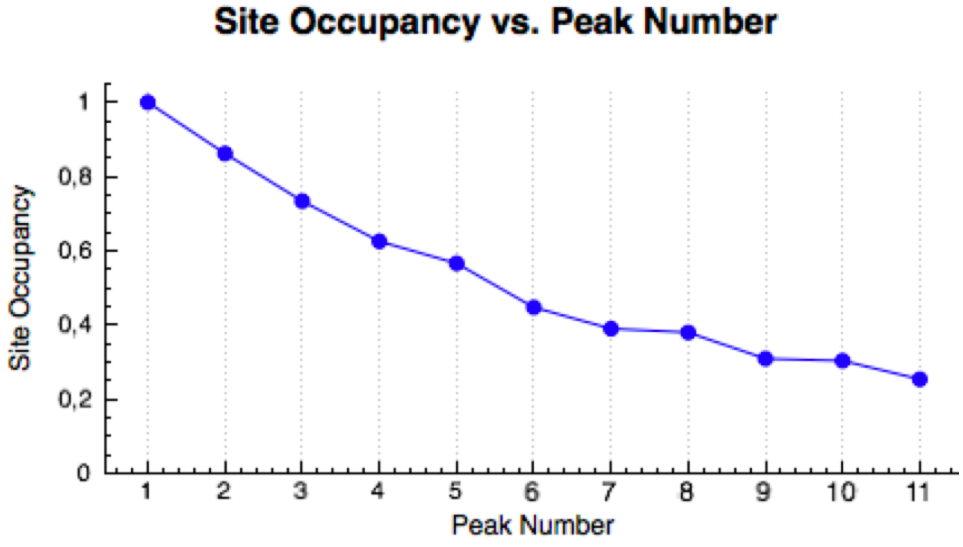




HKL2MAP Diagrams from ***XDSGUI***

Which is Se-MAD and which Br-SAD?

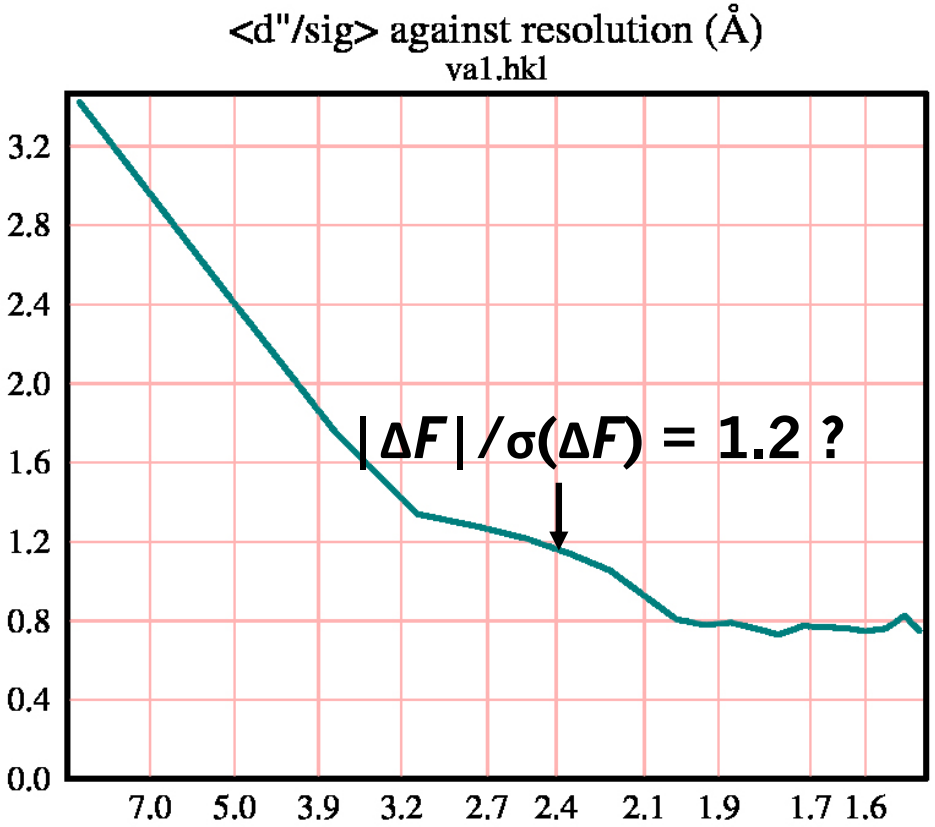


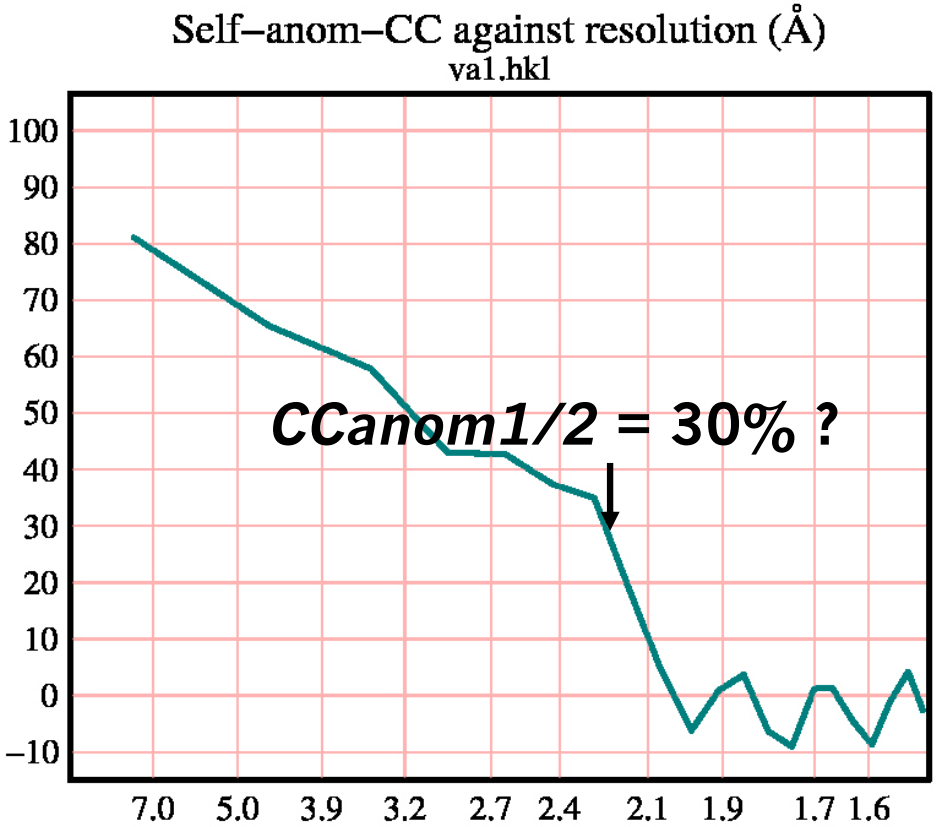


The heavy-atom enantiomorph problem

- The location of the heavy atoms from the $|F_A|$-values does not define the enantiomorph of the heavy-atom substructure; there is exactly a 50% chance of getting the enantiomorph right. In general, original and inverted substructures are tried and only one of the maps will show protein features.
- If the space group is one of an enantiomorphic pair (e.g. $P4_12_12$ and $P4_32_12$) the space group and the atom coordinates must be inverted. For $I4_1$, $I4_122$ and $F4_132$ inversion not at the origin!

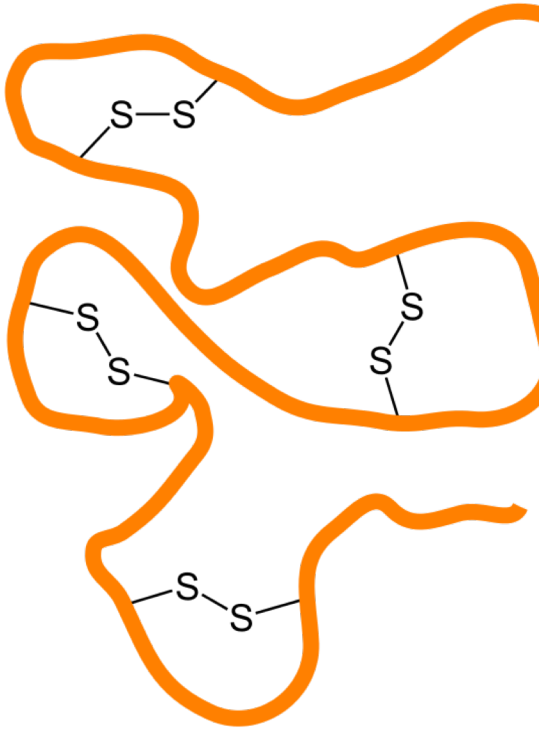
Criteria for data truncation





- Anomalous differences are usually truncated for substructure solution. These XPREP diagrams for the viscotoxin-A1 data show suitable criteria. **|Δ*F*|/σ(Δ*F*)** requires good estimates of the intensity esds, but if it extrapolates at high resolution to the pure noise value of **√(2/π) = 0.798,** the esds must be OK. The **CC** between two random Δ*F* subsets (**CCanom1/2**) is independent of the esds, but requires unmerged data.

Disulfide bond resolution



- When the anomalous signal does not extend to sufficient resolution to resolve disulfides, it has been standard practice to search for *super-sulfur* atoms.
- Modify the peaksearch to locate the best positions for S-S units in the slightly elongated electron density maxima. These *resolved disulfides* not only improve the performance of the substructure solution, they also give a much better phase extension to higher resolution and better final map correlation coefficients.

Critical parameters for SHELXD

The Patterson-seeded dual-space recycling in SHELXD is effective at finding the heavy atoms, however attention needs to be paid to:

- The *resolution* at which the data are truncated, e.g. where the internal CC (CC1/2) between the signed anomalous differences of two randomly chosen reflection subsets falls below 40%
- The *number of sites* to find should be within about 20% of the true value
- Are sites on *special positions* possible? Switch off or not accordingly
- For S-SAD, resolving disulfides (DSUL) can be very useful
- In difficult cases, run more trials (say 50000)

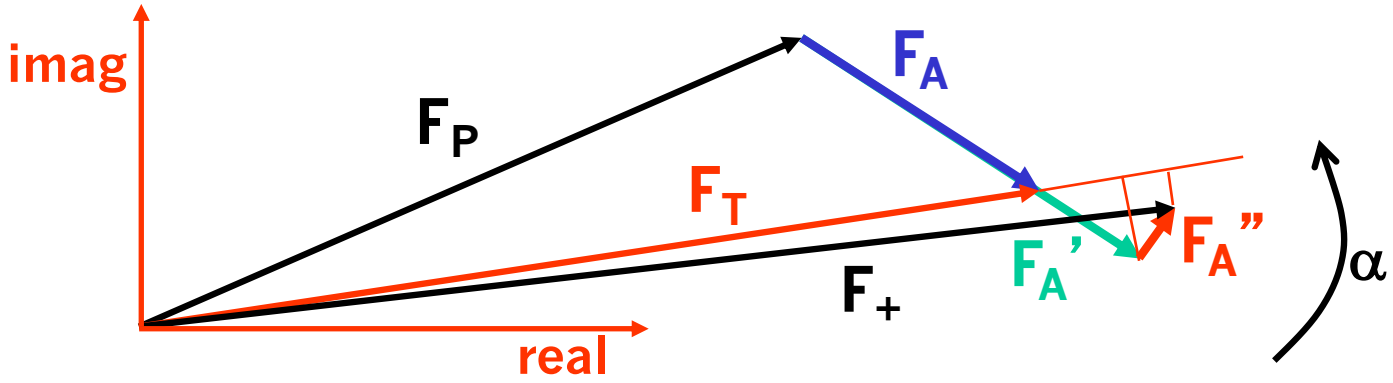
SHELXD is highly parallel and scales up well on a multiple-CPU machine (about 29 times faster on a 32-CPU machine)

The analysis of MAD data

Assume that all the anomalous atoms are of the same element:

$\mathbf{F_A},\ \mathbf{F_A}' = (f'/f_0)\mathbf{F_A}$ **and** $\mathbf{F_A}'' = i(f''/f_0)\mathbf{F_A}$.

[note: f ' is normally negative, so F_A' is parallel to $-F_A$ rather than $+F_A$ as shown here]



We can find F_+ by projecting F_A' and F_A'' onto F_T followed by Pythagoras:

$$|F_+|^2 = [F_T + (f'/f_0)|F_A|\cos\alpha + (f''/f_0)|F_A|\sin\alpha]^2 + [(f'/f_0)|F_A|\sin\alpha - (f''/f_0)|F_A|\cos\alpha]^2$$

which can be rearranged and combined with the $|F_-|^2$ equation to give:

$$|F_\pm|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\alpha \pm c|F_T||F_A|\sin\alpha$$

where: $a = ((f'')^2 + (f')^2)/f_0^2$, $b = 2f'/f_0$, $c = 2f''/f_0$ and $\alpha = \phi_T - \phi_A$.

Unknowns: $|F_T|$ (native F ignoring f' and f'' contributions), $|F_A|$ (heavy atom) and α (phase shift) for each reflection.

Why does SAD phasing work?

We would like to be able to use F_A, the heavy atom structure factor, to solve the substructure, but instead we have to make do with:

$$|F_+| - |F_-| = c\,|F_A| \sin\alpha$$

Which can be derived from the MAD equations by assuming that the anomalous differences $||F_+| - |F_-||$ are small.

We first normalize $||F_+| - |F_-||$ to get E-values. This gets rid of c and its resolution dependence! But we still have one item of data for two unknowns, $|F_A|$ and α, We need $|F_A|$ to solve the substructure and α to get starting values for the native phases using $\phi_T = \phi_A + \alpha$.

The trick is to use only the largest $||F_+|-|F_-||$ and to assume that they correspond to $\alpha \approx 90°$ (when $I_+ >> I_-$) and 270° ($I_- >> I_+$). Direct methods only uses the larger E-values anyway.

Density modification in SHELXE

Starting phases from MR or Experimental phasing are often not accurate enough to make the solution evident

Once we have a starting solution, we can constrain it to be more protein-like by making a chemically sensible change

If we back-transform the modified map and combine with the original phases, they should improve

Sphere of influence: attempt to bring in more stereochemical knowledge

Hoppe & Gassmann (1968) Phase correction method. Acta Cryst. B24, 97
Main, Bricogne, Wang, Kleywegt, Abrahams, Read, Veilleux, Terwilliger, Zhang, Cowtan…

The sphere of influence algorithm

- For each voxel in the map, calculate the variance *V* of the density on a spherical surface of radius 2.42 Å.

Atomic position should fall on voxels with high *V*: sharpen or set to 0!

if $\rho > 0$, $\rho_{mod} = [\rho^4/(\nu^2\sigma^2(\rho)+\rho^2)]^{1/2}$ ($\nu = 0.5$)

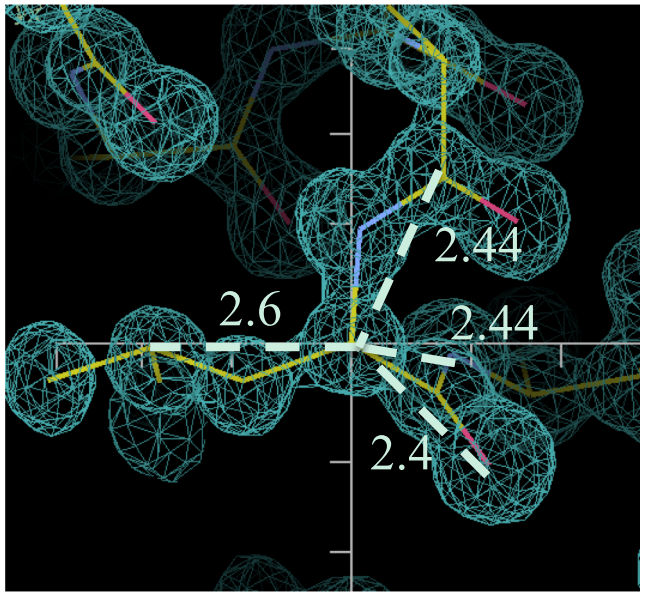
if $\rho < 0$, $\rho_{mod} = 0$

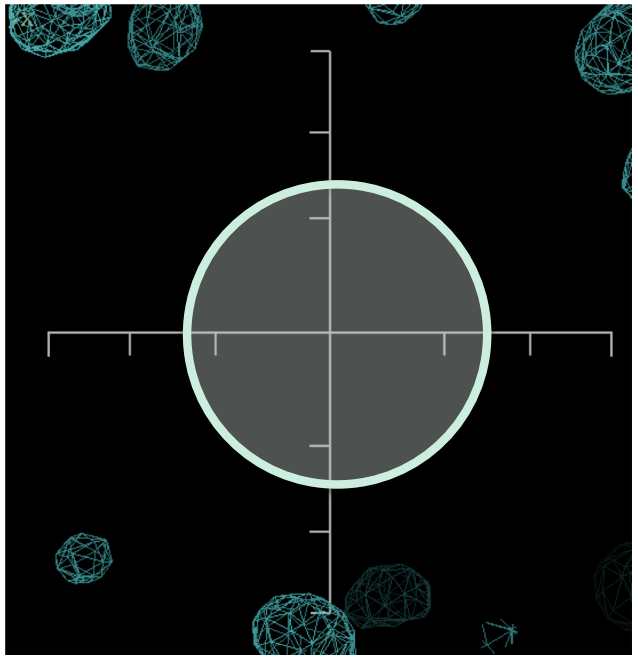
Similar to ACORN.

Solvent should fall on voxels seeing low *V*: flip!

$\rho_S' = -\gamma\rho$ where γ is usually set to 1.0).

For *intermediate V*, use a weighted mixture of both treatments





MODES in SHELXE: shelxe…

Start phases	name.fcf
	name.hlx
Fragment or map	name.pda (pdb format)
	name (.ins with xtal coord)
	name.phi (phs format)
Experimental	name name_fa
Fragment + Experimental	name.phi name_fa
map + Experimental	name.pda name_fa
find or input substructure	-z (n / 0 /-)

Starting phase information

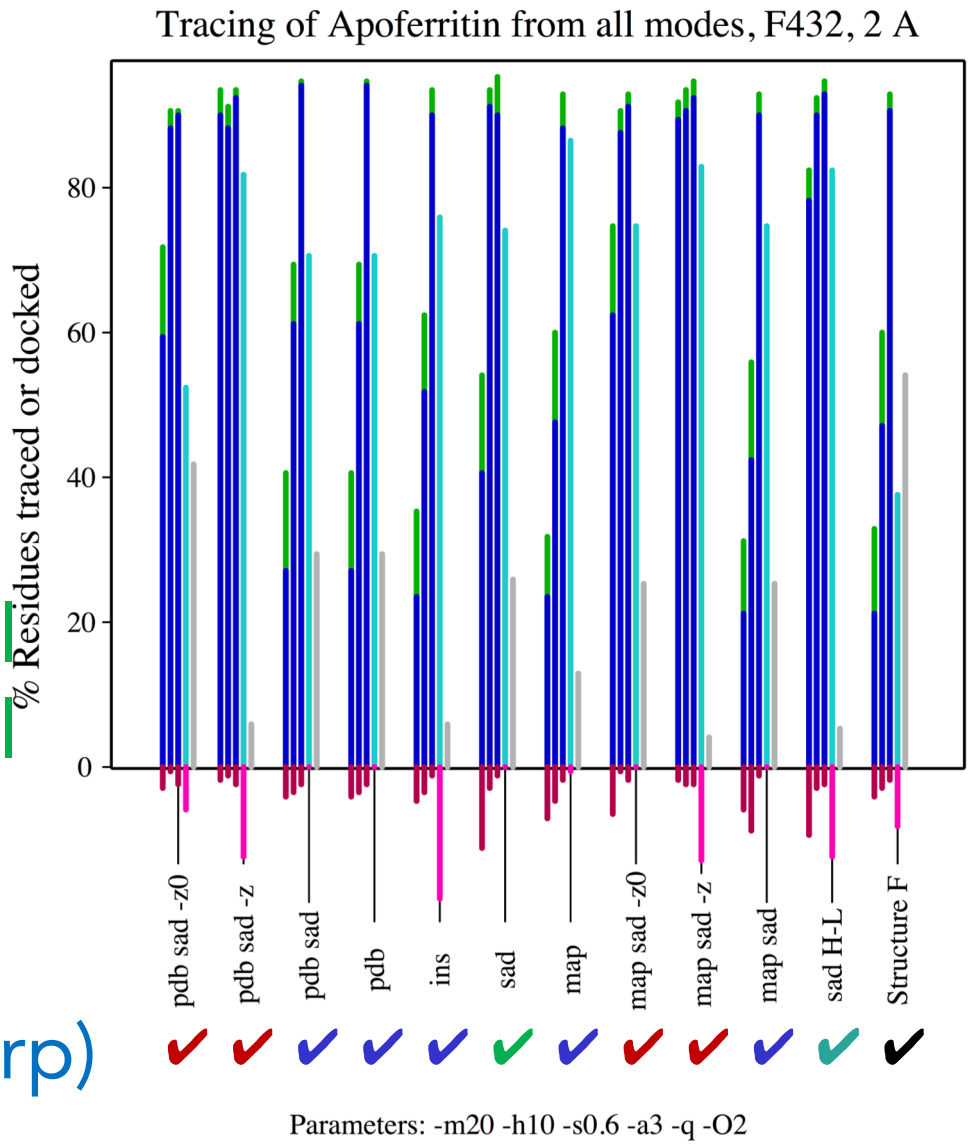
Start phases

Fragment or map

Experimental

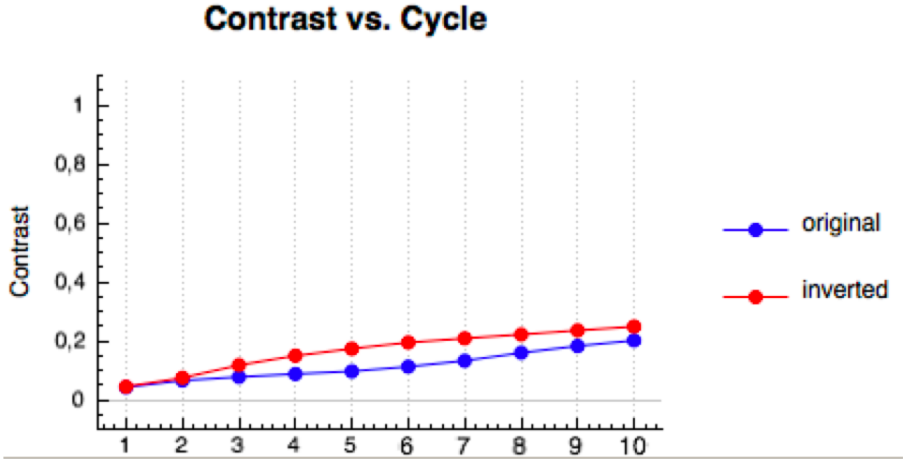
Fragment + Experimental
map + Experimental
find or input substructure

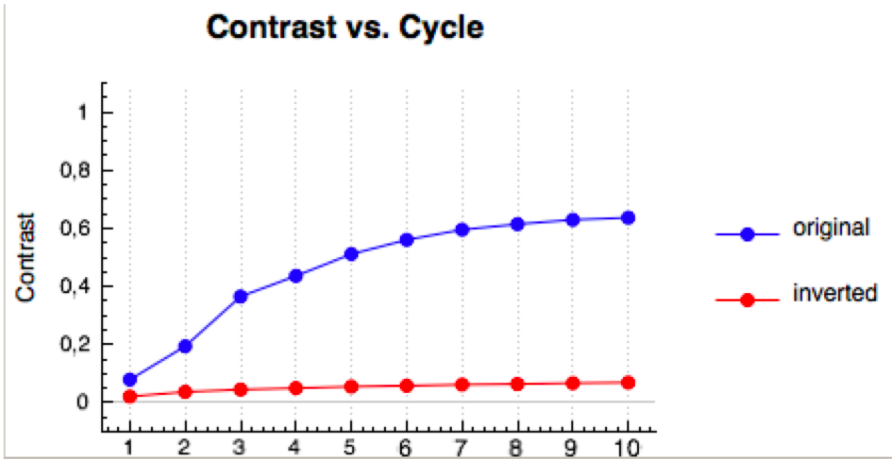
Hendrickson-Lattman (Sharp)



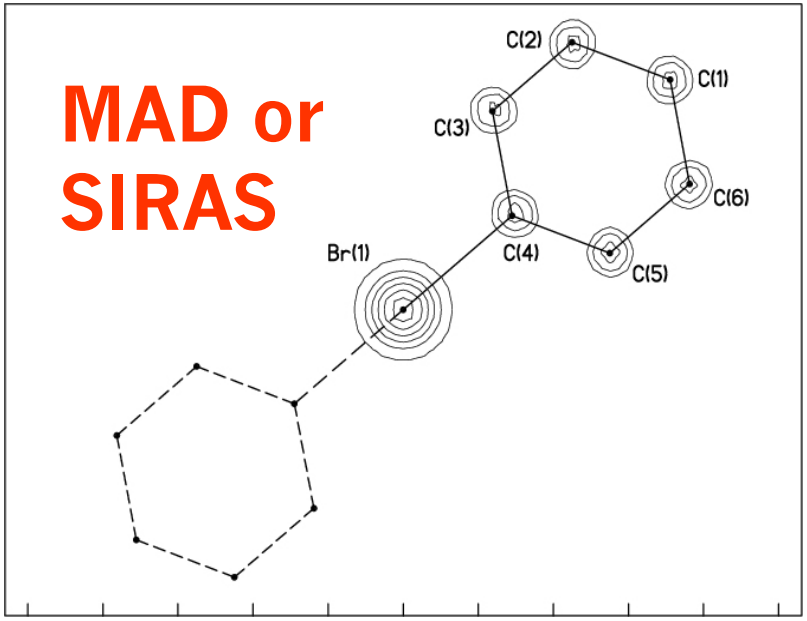
HA enantiomorph discrimination

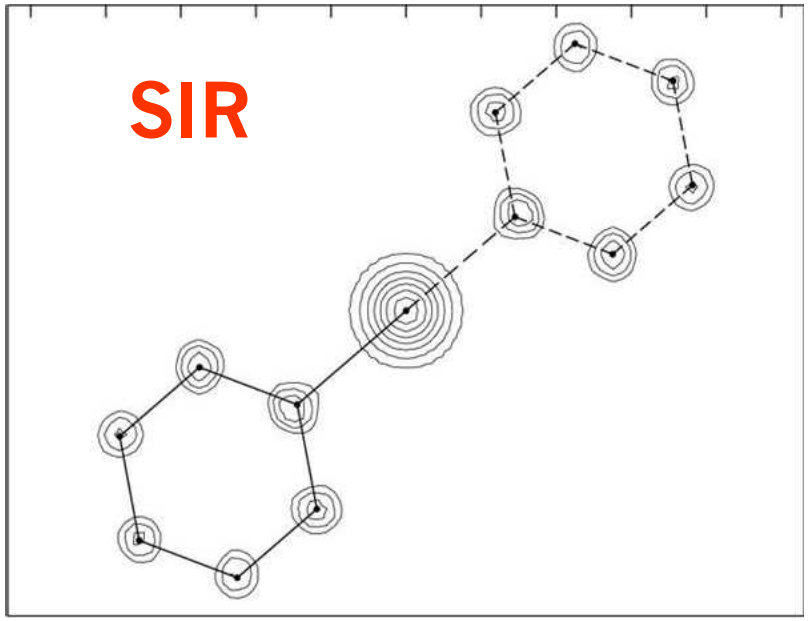
- The ***contrast*** is the variance of V (the variance of the density on a spherical surface of radius 2.42Å from a given voxel) calculated over all voxels. If there are regions of low V (solvent) and high V (protein) the contrast is high. This is a good indication of the correct heavy atom enantiomorph. If the contrast differs strongly for the two hands, the structure is solved!

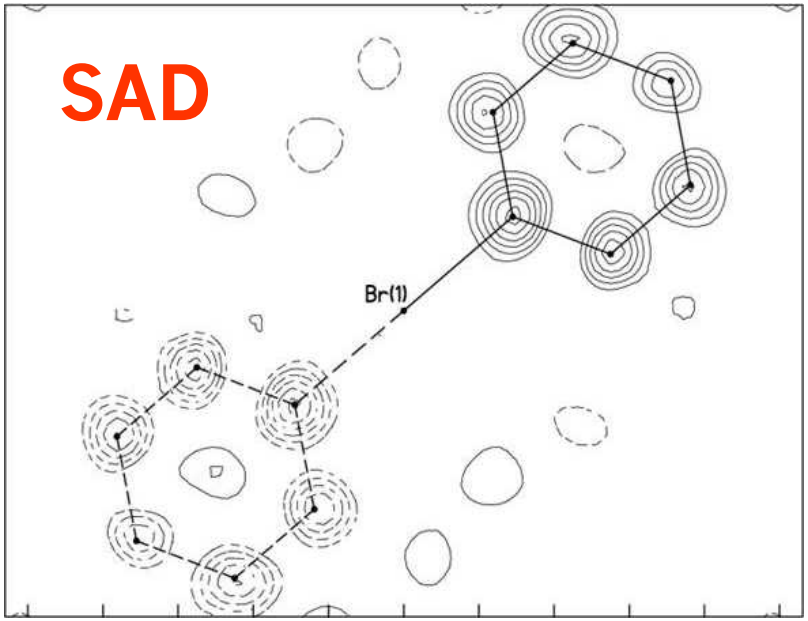




Simulations with one heavy atom in P1





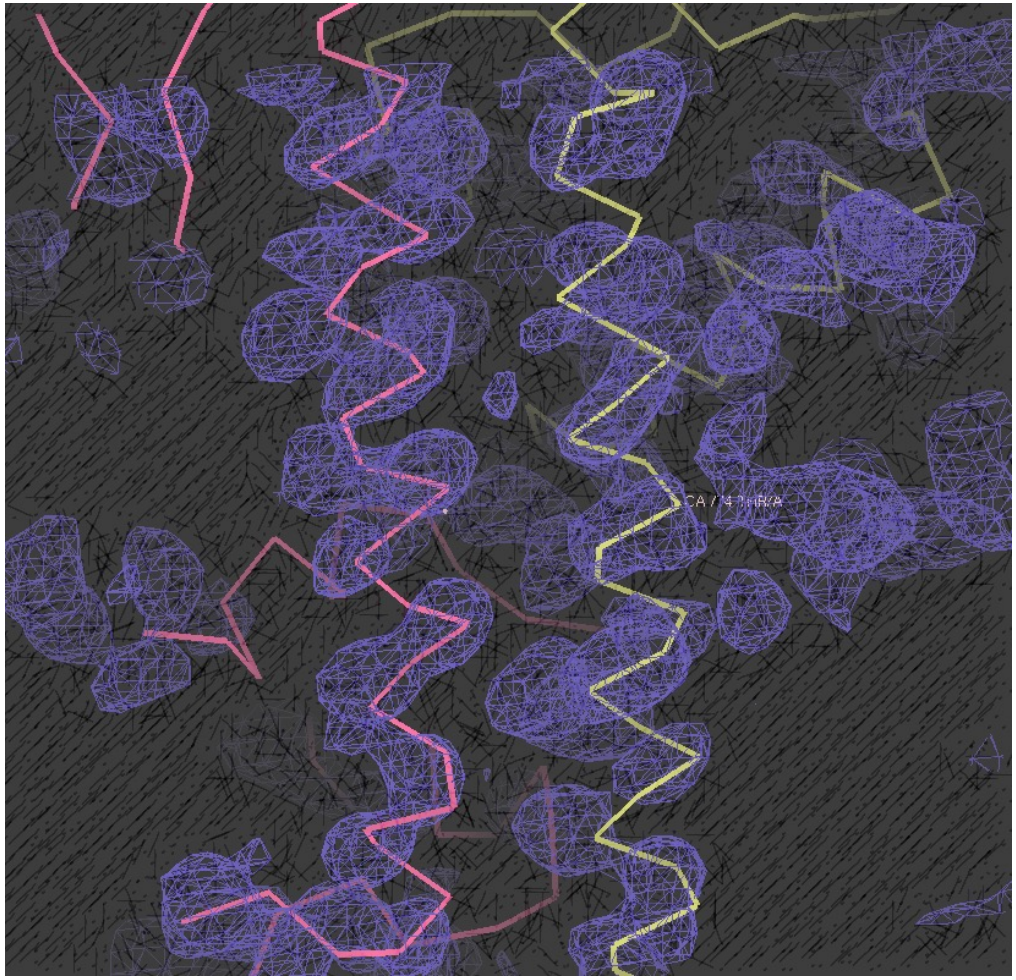


A perfect MAD or SIRAS experiment should give perfect phases! A centrosymmetric array of heavy atoms is fatal for SIR, but one heavy atom is enough for SAD even in space group P1, because it is easy to remove the negative image by setting negative density to zero.

How accurate need phases be...?

Mean phase errors as large as 78 degrees at 3Å still let you recognize features

If you can interpret the map, you can improve it: we know what the density for a helix should be



Map calculated at 3Å, wMPE=78°

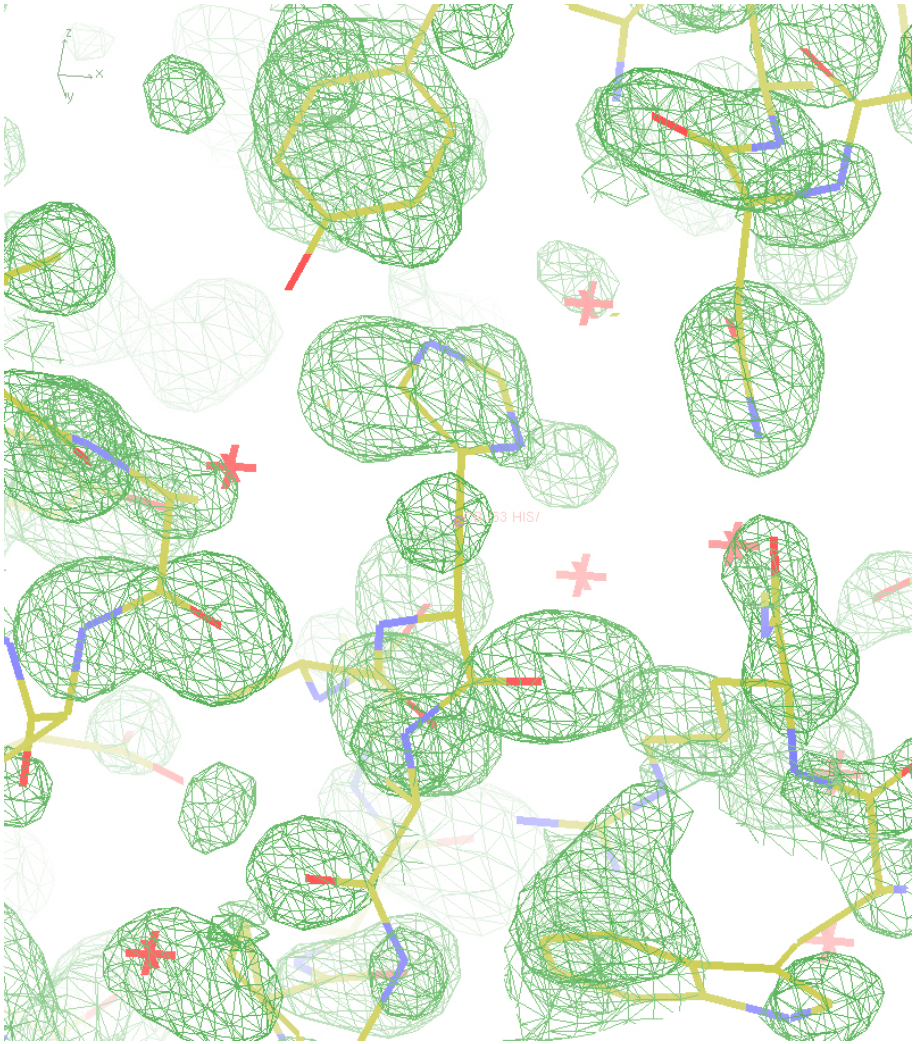
The free lunch algorithm

Extend the resolution of the data by including inner missing reflections and reflections at higher resolution than have been measured

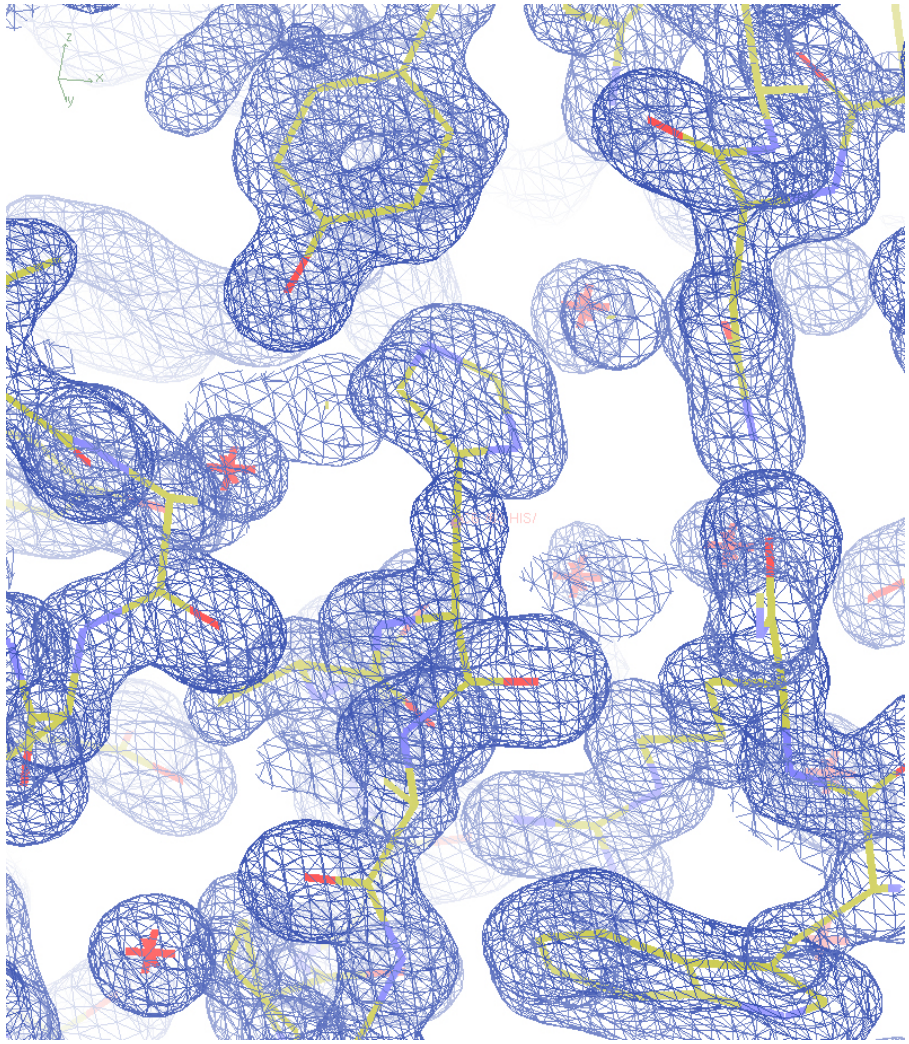
Introduced by Giacovazzo (*Acta Cryst*. D61 (2005) 556-565 and 1080-1087), and also implemented in ACORN (*Acta Cryst*. D61 (2005) 1465-1475).

Surprisingly, if these phases are now used to recalculate the density, using very rough guesses for the (unmeasured) amplitudes, the density actually improves!

Maps before and after a free lunch



Best experimental phases after density modification (MapCC 0.57)



After expansion to 1.0 Å with virtual data (MapCC 0.94)

Usón *et al.*, *Acta Cryst.* D63 (2007) 1069-1074.

Why do we get a free lunch?

- It is not immediately obvious why inventing extra data improves the maps. Possible explanations are:

1. The algorithm corrects Fourier truncation errors that may have had a more serious effect on the maps than we had realised.
2. Phases are more important than amplitudes, so as long as the extrapolated phases are OK any amplitudes will do.
3. Zero is a very poor estimate of the amplitude of a reflection that we did not measure.

Iterative generation of a poly-alanine backbone trace turns out to be an even more effective way of improving the phases, CC for the trace is a good indication of correctness