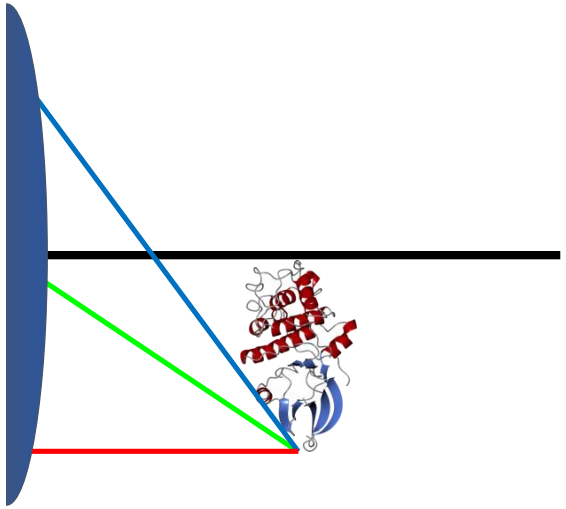
Experimental Phasing – The Phase Problem

Ed Lowe – University of Oxford

edward.lowe@bioch.ox.ac.uk

Electron Density Equation



- Normally a lens will
  - Bend the scattered light pattern
  - Apply a phase shift
- Each “pixel” of the image formed is therefore a sum of many scattered waves. These will add up in a way that depends upon their phase and on the position of the pixel.
- It is therefore not surprising that the electron density equation looks the way that it does.

All reflections

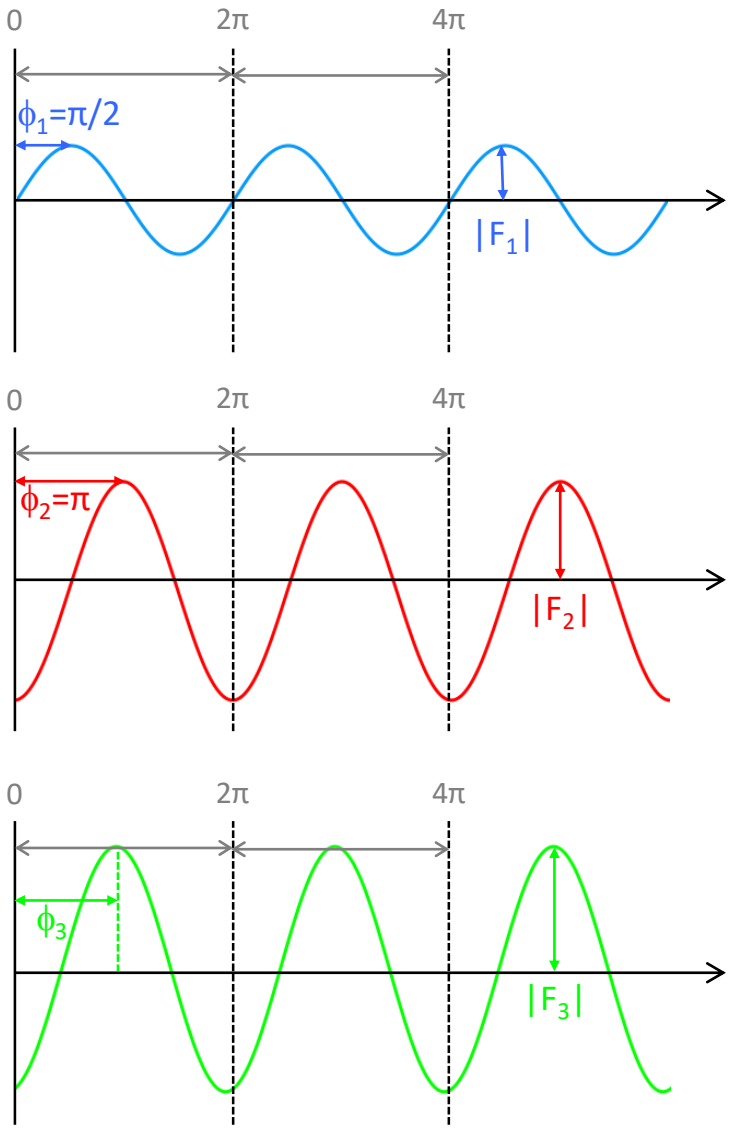
$$r(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l |F(h,k,l)| \exp\left\{2\pi i(hx+ky+lz) + i\alpha(h,k,l)\right\}$$

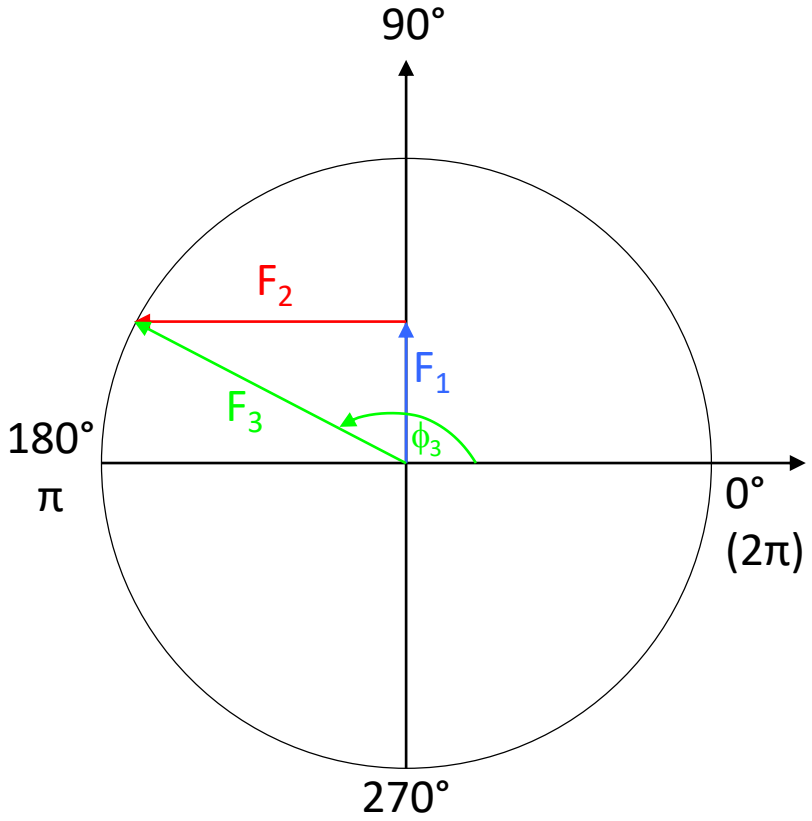
Electron density at a point

Amplitude

Phase

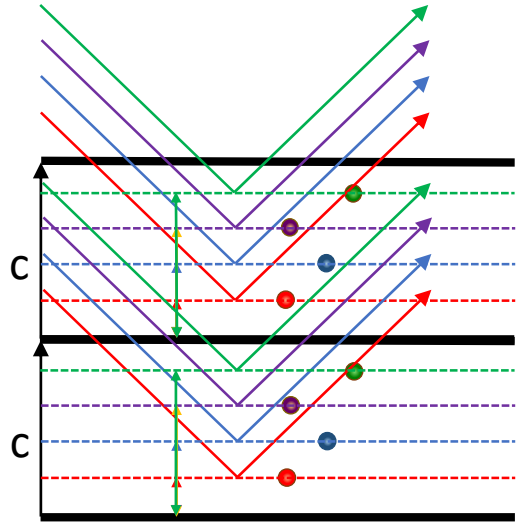
Addition of waves in the complex plane

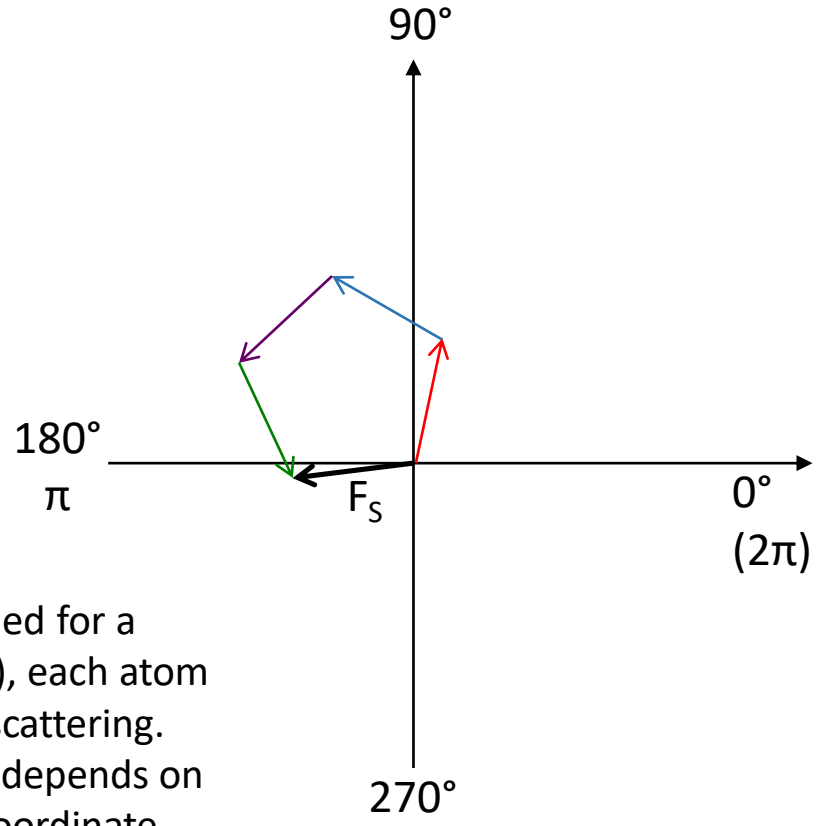




NB. Experiments are monochromatic, so all waves have the same wavelength

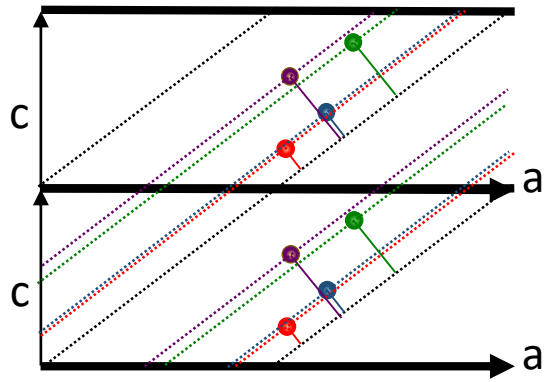
Addition of atomic scattering vectors to produce a structure factor



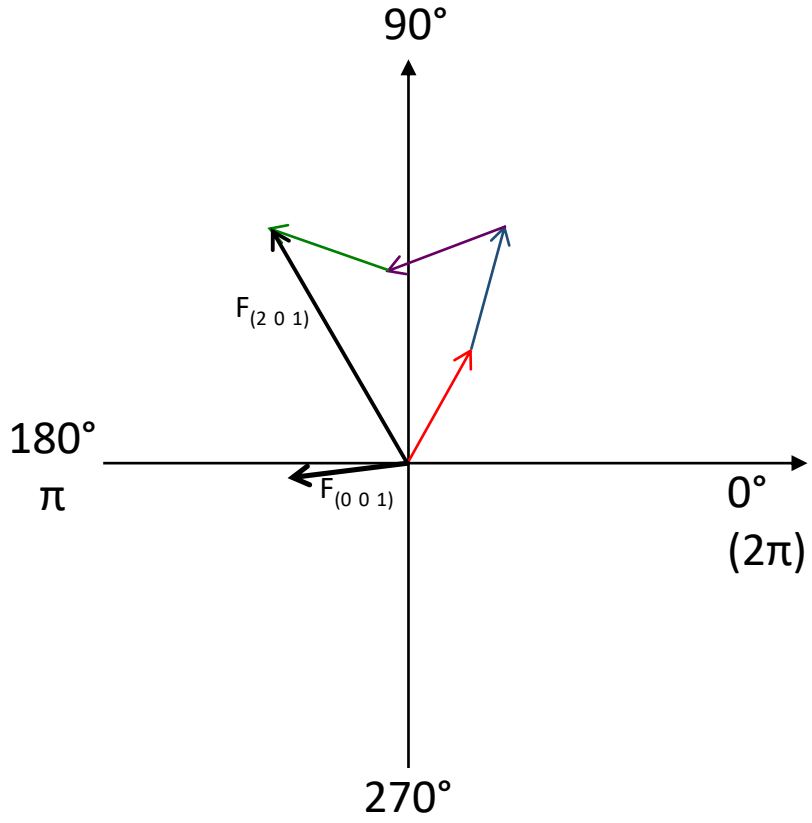


- When the diffraction condition is satisfied for a particular set of planes (drawn in black), each atom contributes some amount to the total scattering.
- The relative phase of that contribution depends on the position of the atom (fractional z-coordinate shown).
- The total scattering is the vector sum of the individual atomic scattering vectors.
- This is referred to as a structure factor because it is dependant on the arrangement (or structure) of atoms in the unit cell

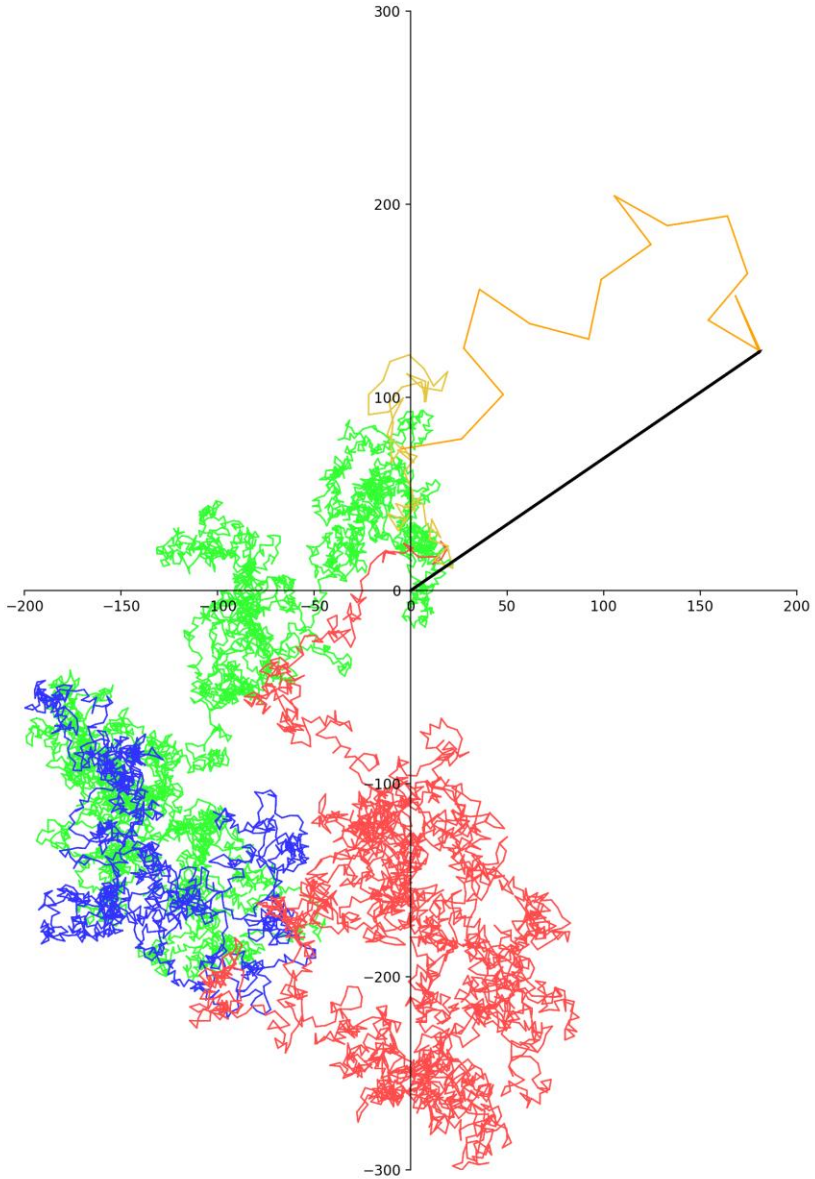
How does this look with a different set of Bragg planes?



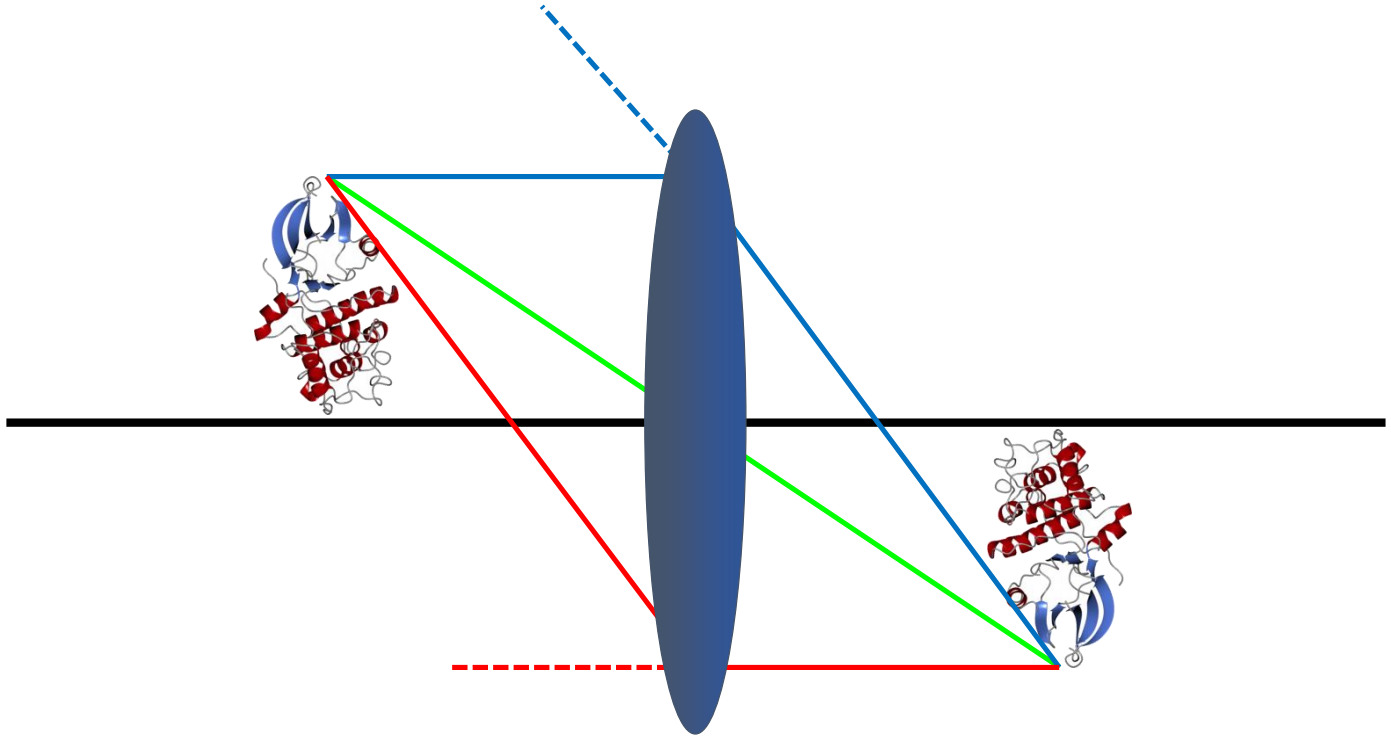
Indices (2 0 1)



A bit more complicated for a protein…

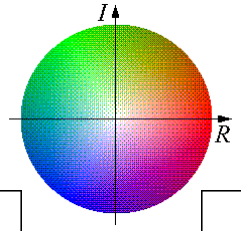


The Phase Problem

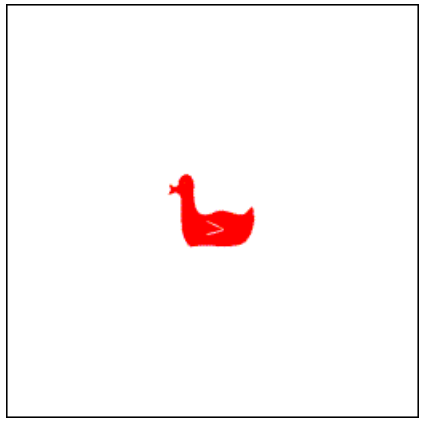


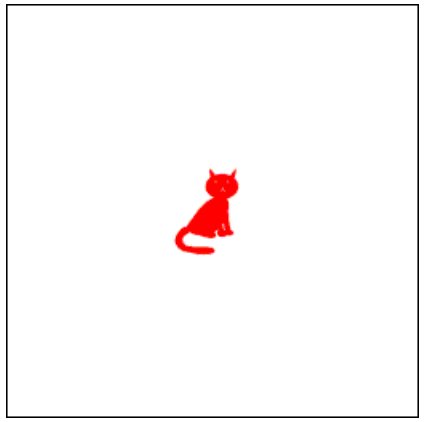
- It is straightforward to reconstruct an image from diffraction data when both the amplitudes and phases of the structure factors are known
- A diffraction image contains information on the amplitudes of structure factors but not their phases
- This is known as the phase problem

Our problem is how to achieve the recombination of scattered waves without having to invent an X-ray lens



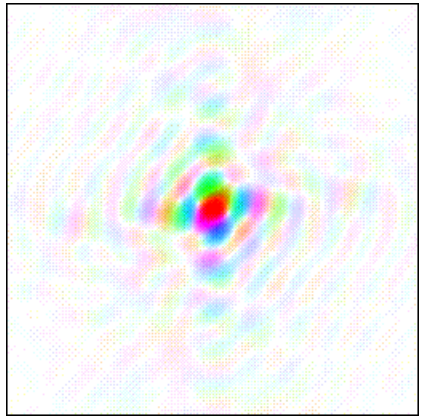
Representing **intens**

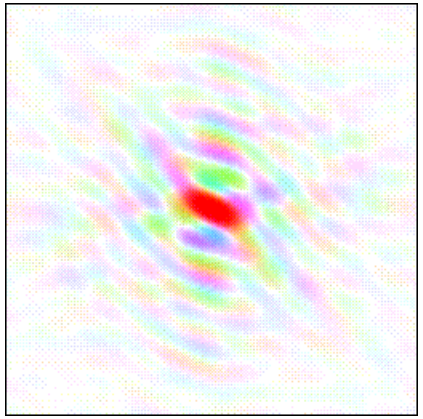


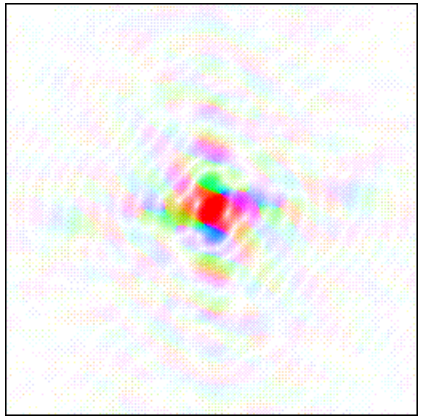


Fourier transform

Fourier transform







F(Duck),ϕ(Duck)

F(Cat),ϕ(Cat)

F(Duck),ϕ(Cat)

What will the reconstructed image most resemble?

A: A Duck

B: A Cat

C: Mixture (more duck)

D: Mixture (more cat)

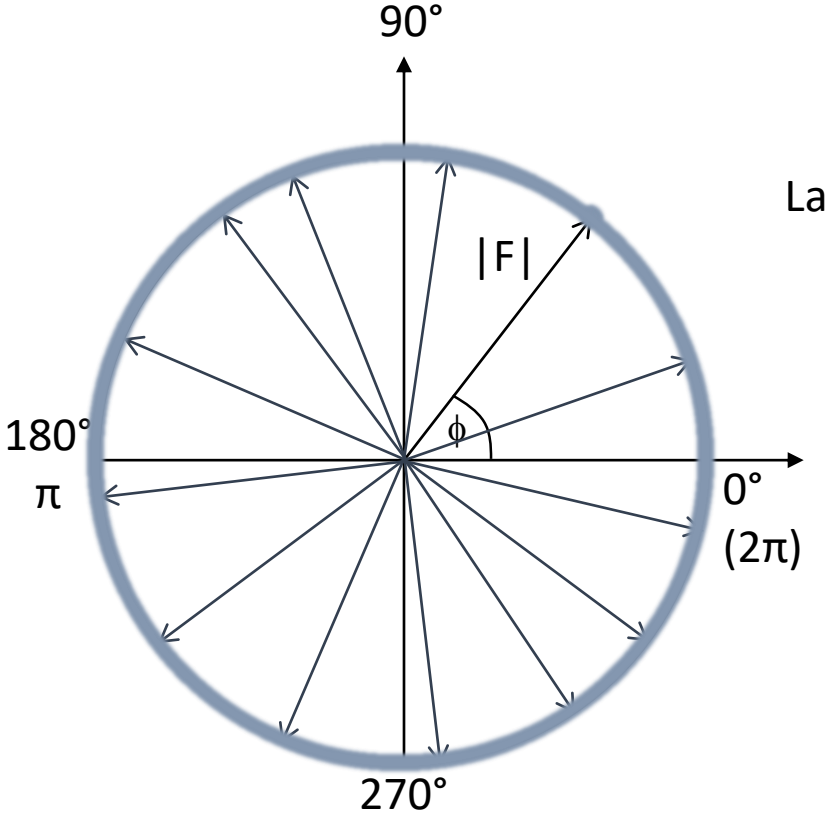
E: A rabbit

This sounds pretty hopeless – surely we can never trust anything we see in a crystal structure unless experimental phases were measured?

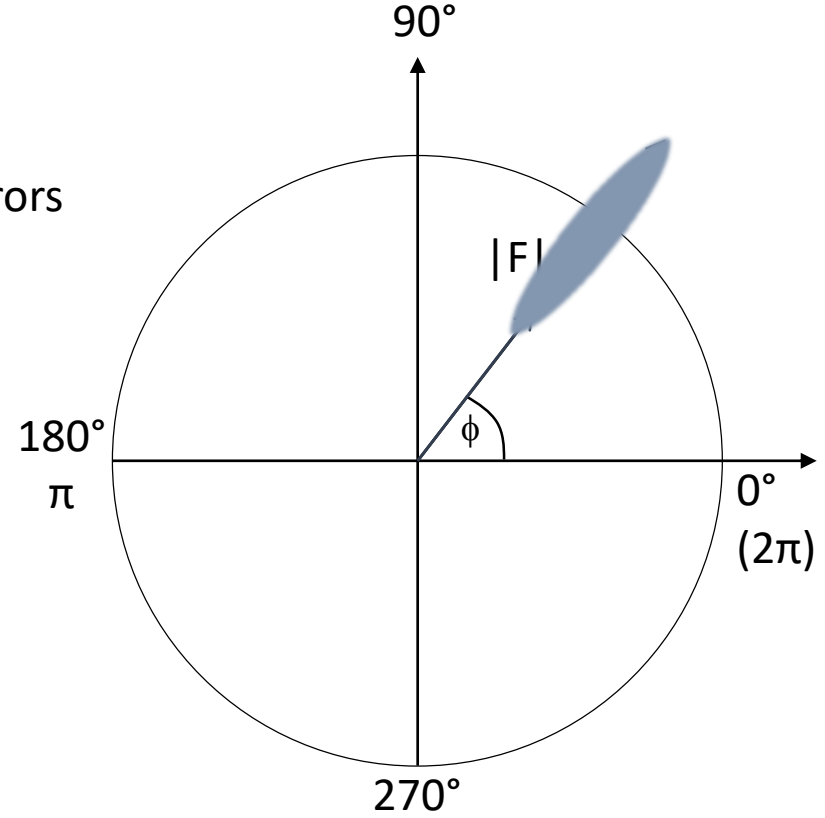
True or False?

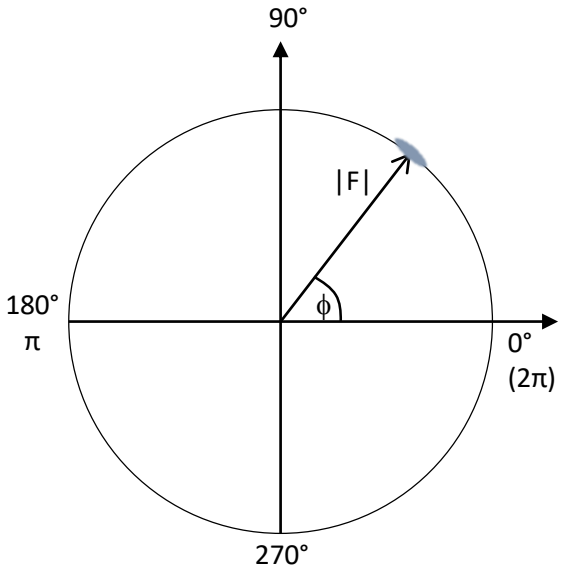
But… there is no need to panic just yet!

- This disastrous situation is greatly exaggerated when the two sets of phases are completely different.
- It relates to the fact that the vector you get by pointing a correct length vector in the wrong direction is (on average) further away from the true result the that obtained by pointing a vector of any random length in the right direction.
- Where the phase is close to the correct value, this ceases to be the case. This is why we can see features in difference maps (F_o-F_c), ϕ_{calc}
- Despite this, we would achieve a result much closer to the truth if we were able to always record phases with our amplitudes.

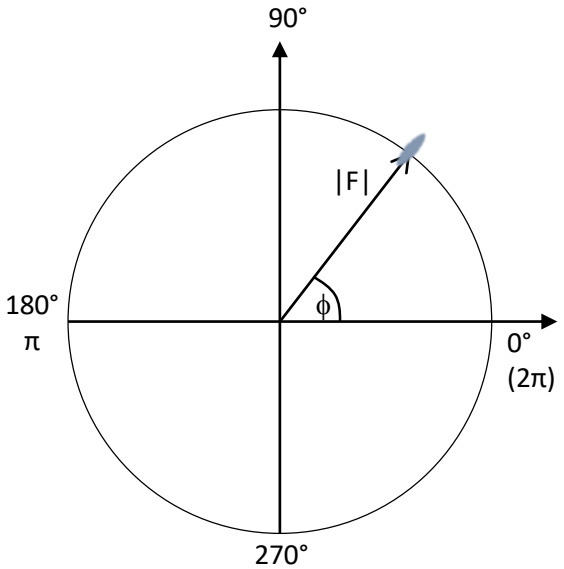


Large errors





Small errors



Direct methods

Structure factors are not independent of each other – they are related through the structure.
All atoms on a lattice plane scatter in phase, additionally those in parallel planes scatter in phase.

This leads to the derivation of the triplet phase relationship. This gives a reasonable estimation of phase for strong reflections.

$$j_{-h} + j_{k} + j_{h-k} \gg 0$$

- This works only if:
  - The magnitudes of the structure factor amplitudes are large – this relationship applies to strong reflections.
  - The atoms within the structure must be fully resolved from each other (for a protein structure this would require resolution of better than 1.0Å)
  - Relatively few atoms in the unit cell (approx. up to 1000)

Patterson function

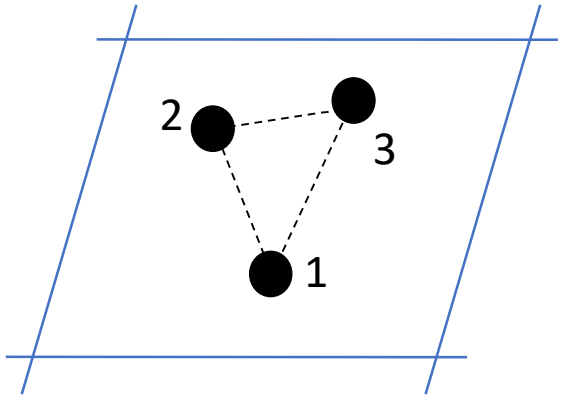
This is a Fourier summation of intensities (as opposed to Structure factors) with phase angles set to zero.

$$P(uvw) = \frac{1}{V} \sum_{hkl} |F(hkl)|^2 \cos[2\pi(hu + kv + lw)]$$

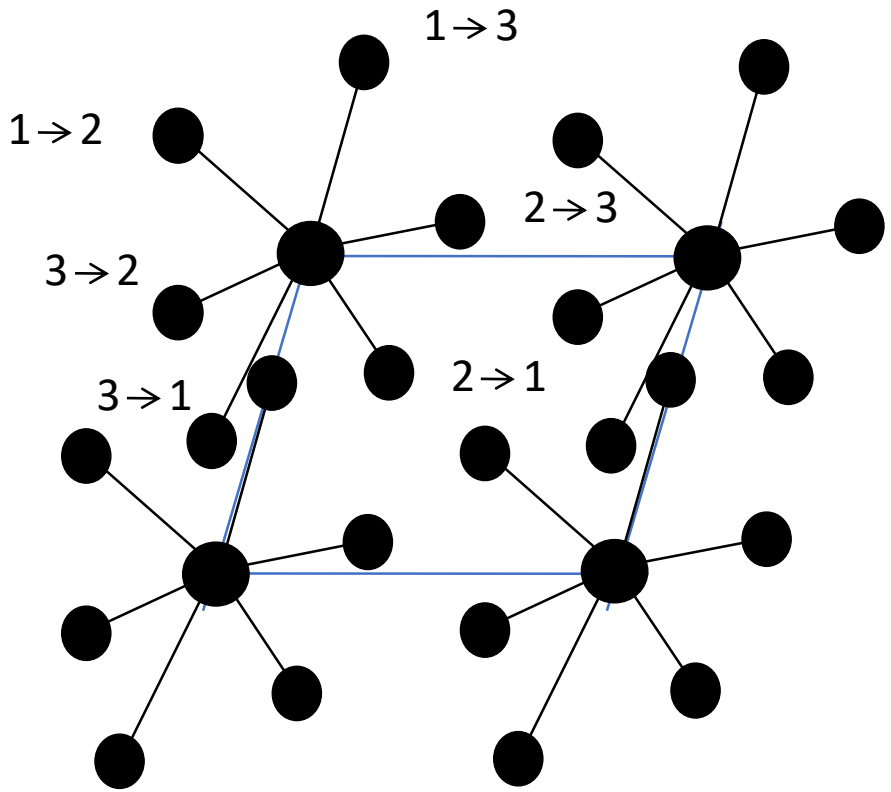
Where *u*, *v* and *w* are the coordinates in the Patterson cell to avoid confusion with x, y and z from the real unit cell (although both unit cells have the same dimensions).

By solving the Patterson equation for the measured intensities you can generate a vector map between the atoms.

Patterson Map



Unit cell containing 3 atoms



Patterson map for the 3 atoms. At the origin we see larger peaks as every atom can is placed at the origin in calculating the map. The peaks in the map are the vectors of the original 3 atoms

How many peaks will there be in the Patterson function of a small protein with 1000 atoms?

A: 9990

B: 99,900

No. of peaks = N^2-N

C: 999,000

D: 9,990,000

E: 42

So, what methods can we apply to a protein structure containing 5000 atoms at a resolution of 2Å?

A: Direct Methods

B: Patterson Methods

C: A combination of Direct and Patterson Methods

D: Neither, we need a simpler structure for either to work.

E: Throw all of our data at a computer and hope it works it out for us!

- It is possible to phase a small, simple structure directly.
- But we have a large, complex structure.

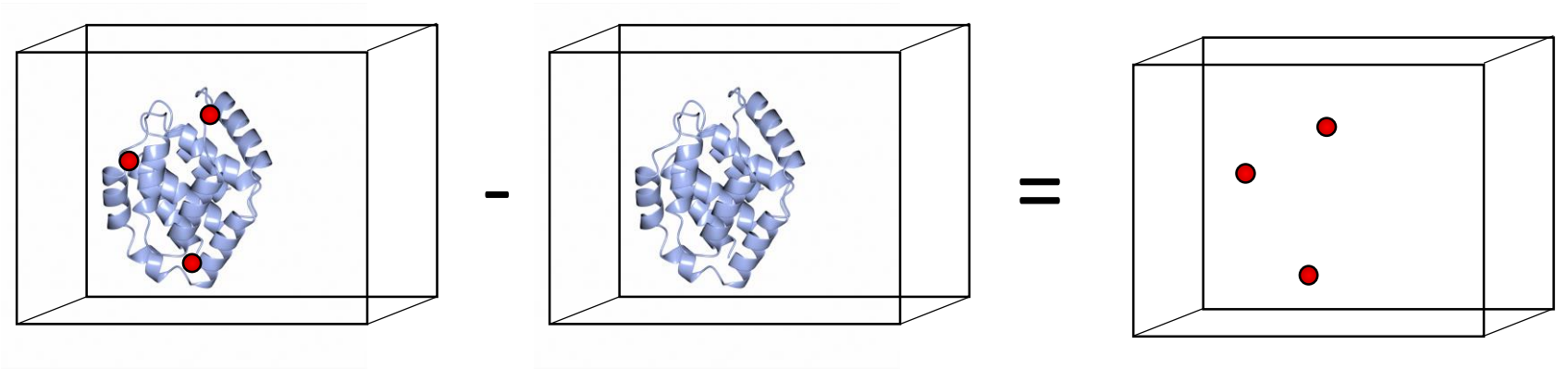
- We need a small structure that can act as a proxy for our large complex structure.
- This is our **substructure**
- Once we know the positions of a few atoms we can then bootstrap our way up.

Isomorphous replacement

To create a substructure the easiest way is to incorporate atoms heavier than the atoms normally found in proteins.

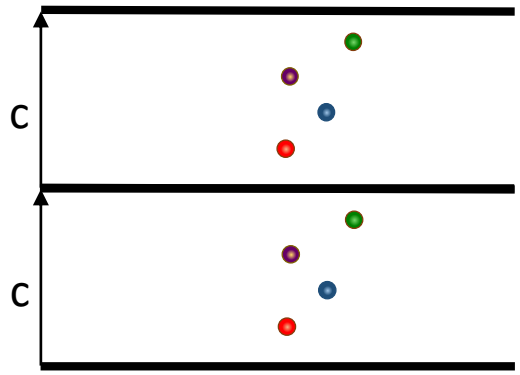
Isomorphous replacement compares the diffraction data from a native protein crystal (nothing bound) to one where a heavy atom has been bound.

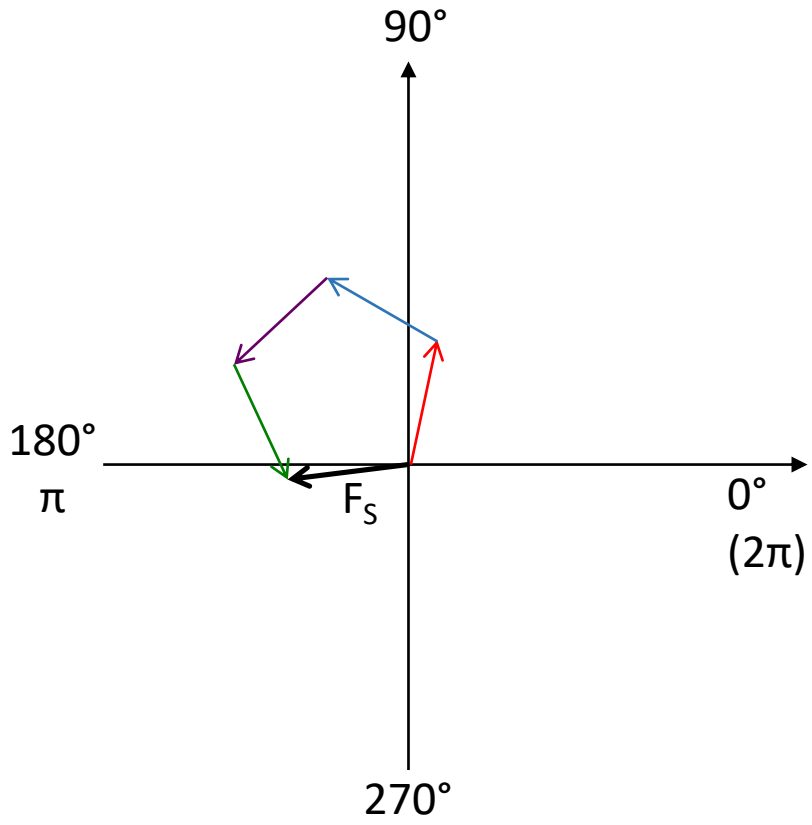
The Patterson function is then used to calculate the positions of the heavy atom(s).



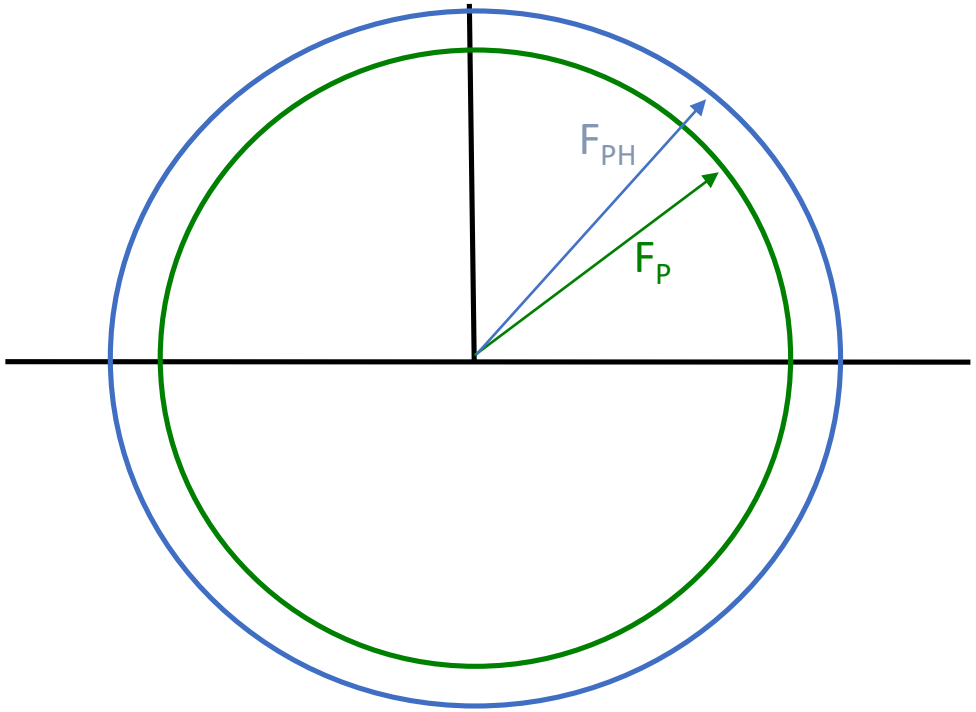
$$F_{PH} \quad - \quad F_P \quad = \quad F_H$$

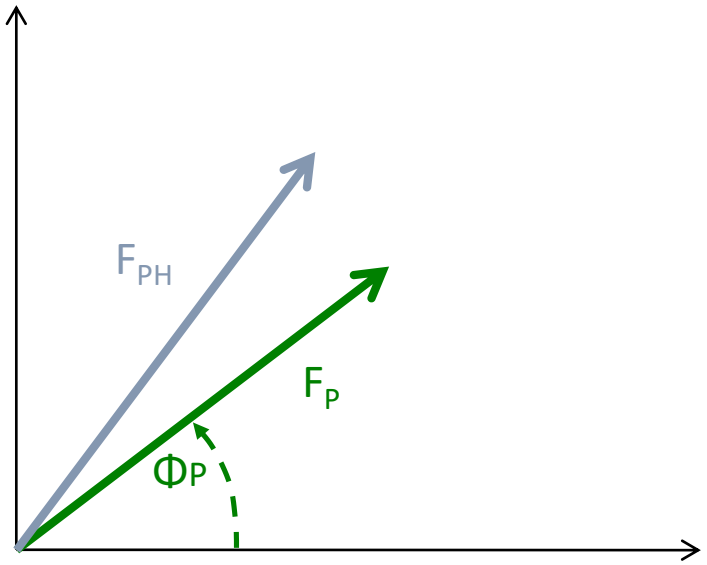
Remember – if we know the position of an atom we can calculate the phase of its contribution to scattering.





Solving the phases using Isomorphous Replacement

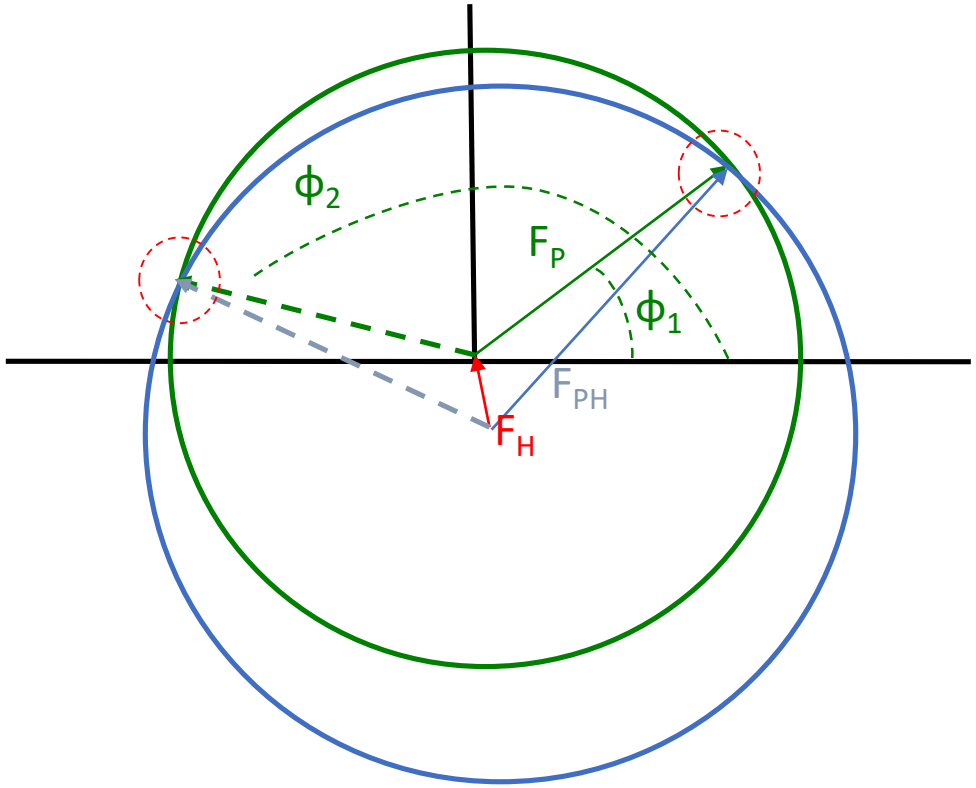




We collect Native (F_P) data

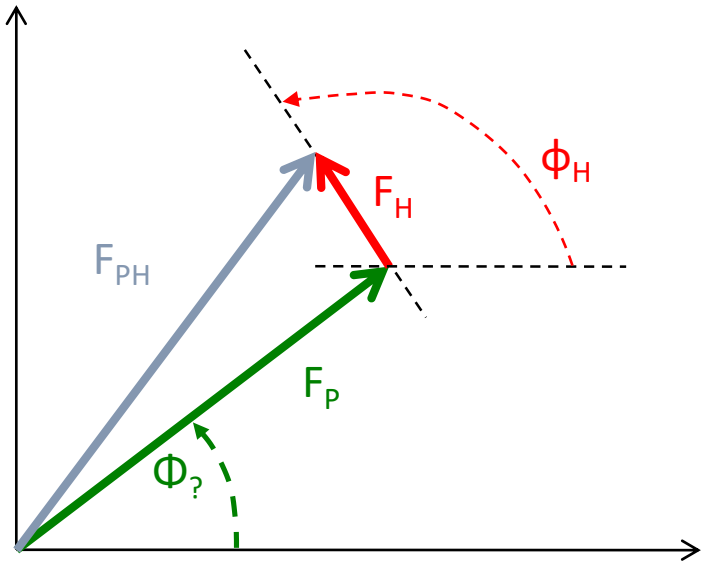
The phase is unknown

Next we collect the Derivative (F_{PH}) data



We offset the F_{PH} term from the origin by the value of the F_H term

The points where the two circles intersect are possible solutions for the phase. This is known as a Harker construction



$F_{PH} = F_P + F_H$

Using direct methods or the Patterson Function we can solve the position and phase of the heavy atom (F_H)

So, we have two possible solutions but only one can be right – how can we solve this ambiguity?

A: Solve for both solutions and see which one is correct

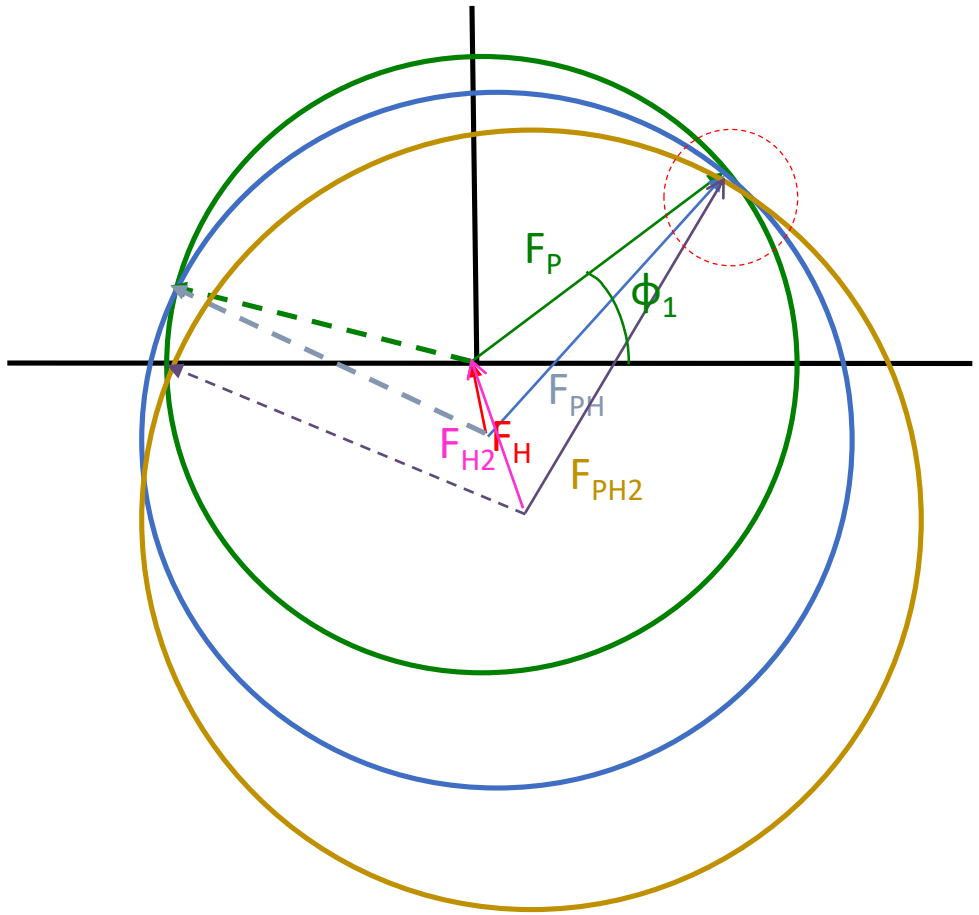
B: Take a guess – one of them has to be right

C: Repeat the whole experiment with a second derivative and look for agreement with one of the first derivative solutions

D: Select the solution that generates a map compatible with a protein composed of L-amino acids

E: Throw all of our data at a computer and hope it works it out for us!

Multiple Isomorphous Replacement



However, there are two possible solutions for ϕ_P, so we need more information.

By using a second derivative binding in a different site on the protein we can potentially solve our problem

In this case the second derivative suggests that ϕ_1 is correct.

Anomalous Scattering

A•nom•a•lous

adj.

Deviating from the normal or common order, form or rule

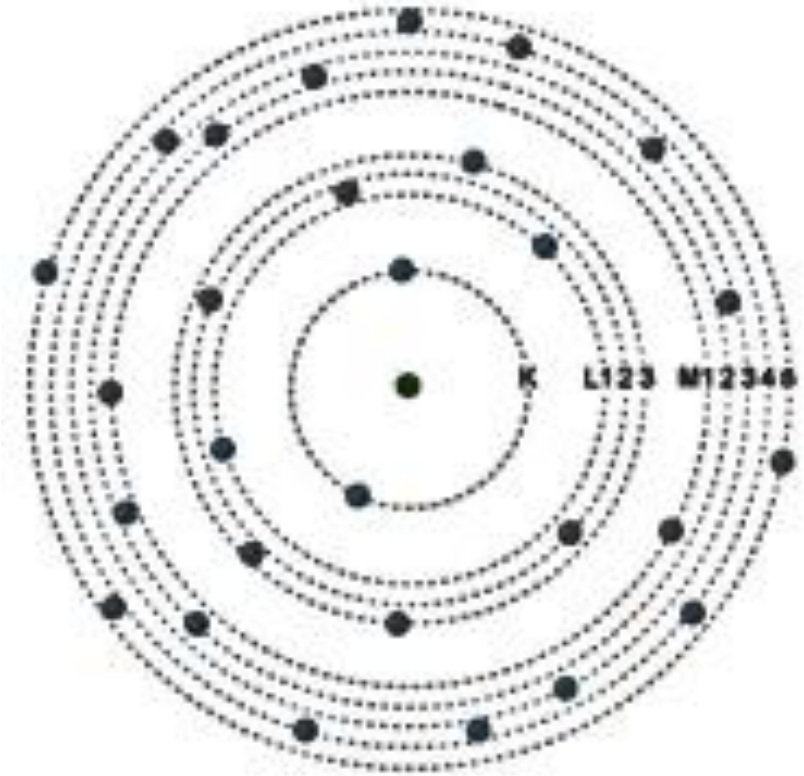
"**Anomalous** scattering" is **absolutely normal** while "**norma**l scattering" occurs only as an ideal, over simplified model, which can be used as a first approximation when studying scattering problems"

IUCR Pamphlet "Anomalous Dispersion of X-rays in Crystallography"

S. Caticha-Ellis (1998)

i.e. All atoms are anomalous scatterers – but not all are <u>significant</u> anomalous scatterers

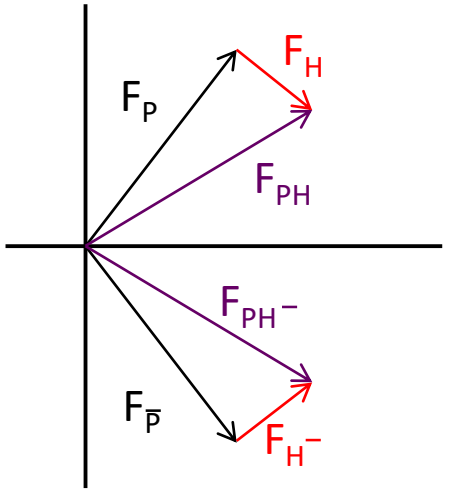
- Anomalous scattering is due to the electrons being tightly bound (particularly in K & L shells)
- In classical terms, the electrons scatter as though they have resonant frequencies

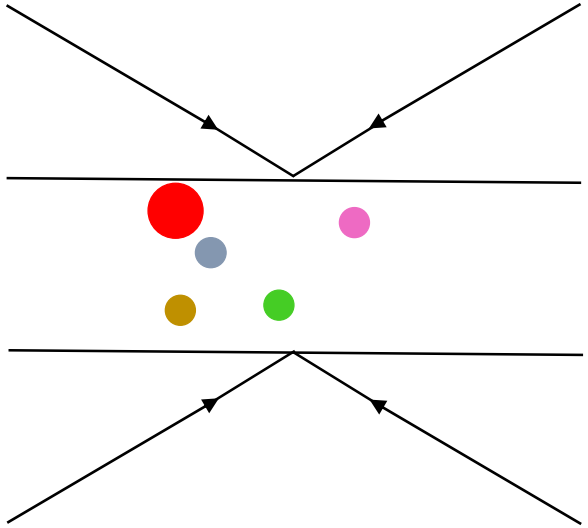


Friedel's Law in normal scattering conditions

Friedel pairs are Bragg reflections related by inversion through the origin

Friedel's Law – A Friedel pair have equal amplitude and opposite phase

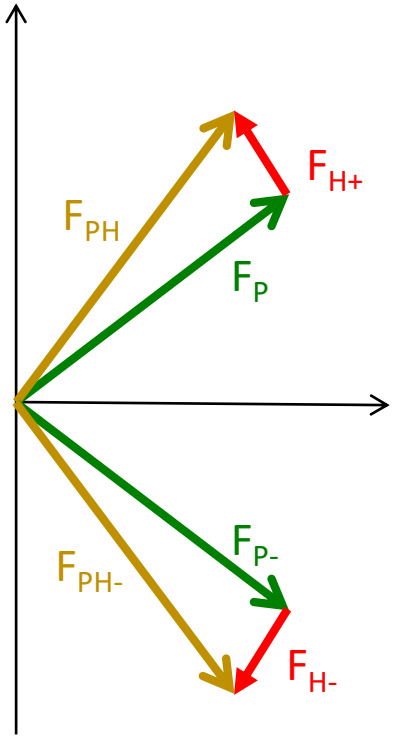




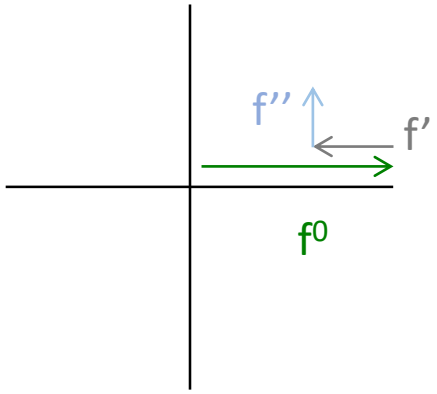
Normal scattering conditions

$$|F_{hkl}| = |F_{\overline{hkl}}| \quad j_{hkl} = -j_{\overline{hkl}}$$

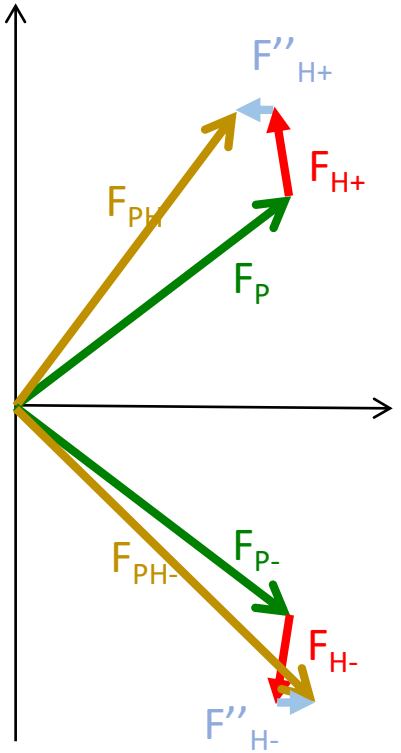
Breaking Friedel's Law



Normal scattering conditions



Under anomalous scattering conditions, the f'' component of atom A lags the phase component of the f^0+f' by 90°. Its phase is always 90° different.



Friedel's Law is broken.

$$|F_{PH}| \neq |F_{\overline{PH}}|$$

How can this help solve the phase problem?

- The atoms normally found in proteins (carbon, nitrogen, oxygen) from do not scatter anomalously at the X-ray wavelengths (energies) we routinely use.

- But heavy atoms do. So we can create a heavy atom substructure again and collect anomalous data.

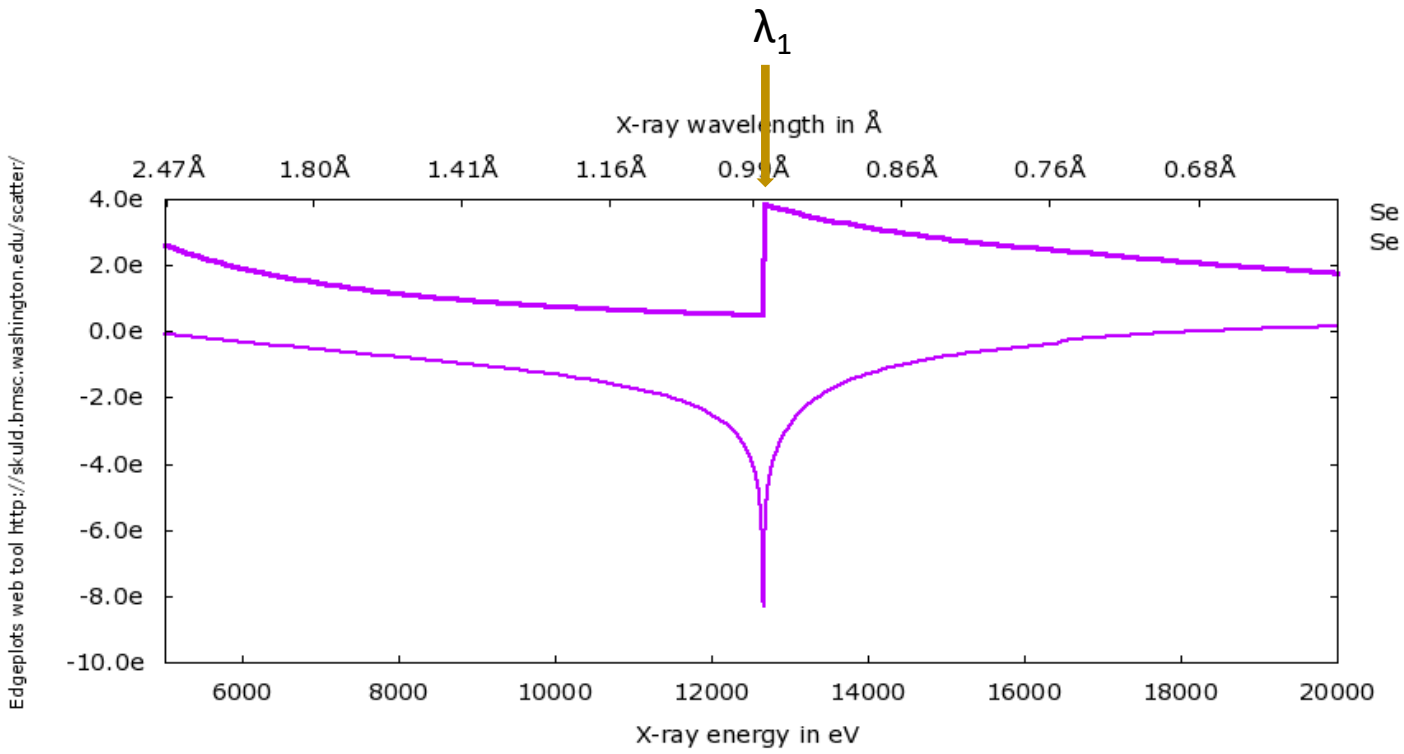
- An additional method of heavy atom incorporation can be used here by incorporating selenomethionine into the protein in place of methionine.

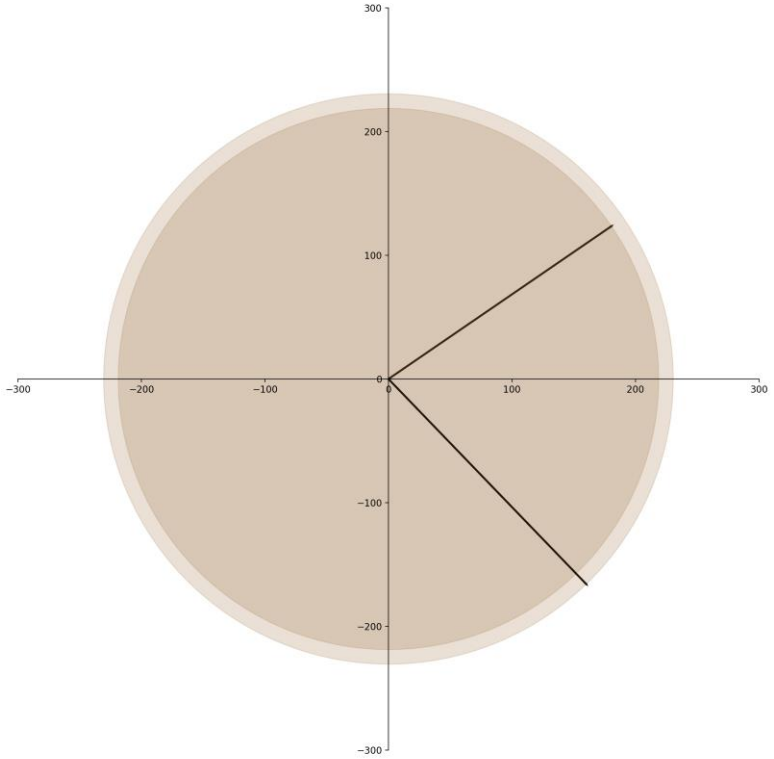
- We need to collect data at a synchrotron as we can select the wavelength and cause our substructure atoms to scatter anomalously.

There are two ways of phasing using anomalous scattering

SAD – Single anomalous diffraction – where we collect a single dataset with the maximum anomalous signal.
MAD – multiwavelength anomalous dispersion – where we collect several datasets with various levels of anomalous scatter and make use of the dispersive differences between wavelengths.
Also SIRAS and MIRAS if we have isomorphous native data.

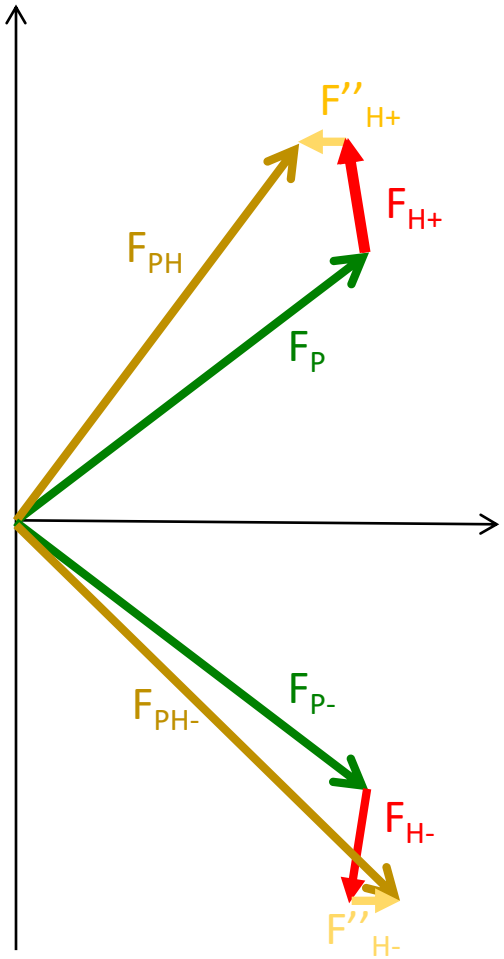
Solving the phases using Single Anomalous Diffraction (SAD)

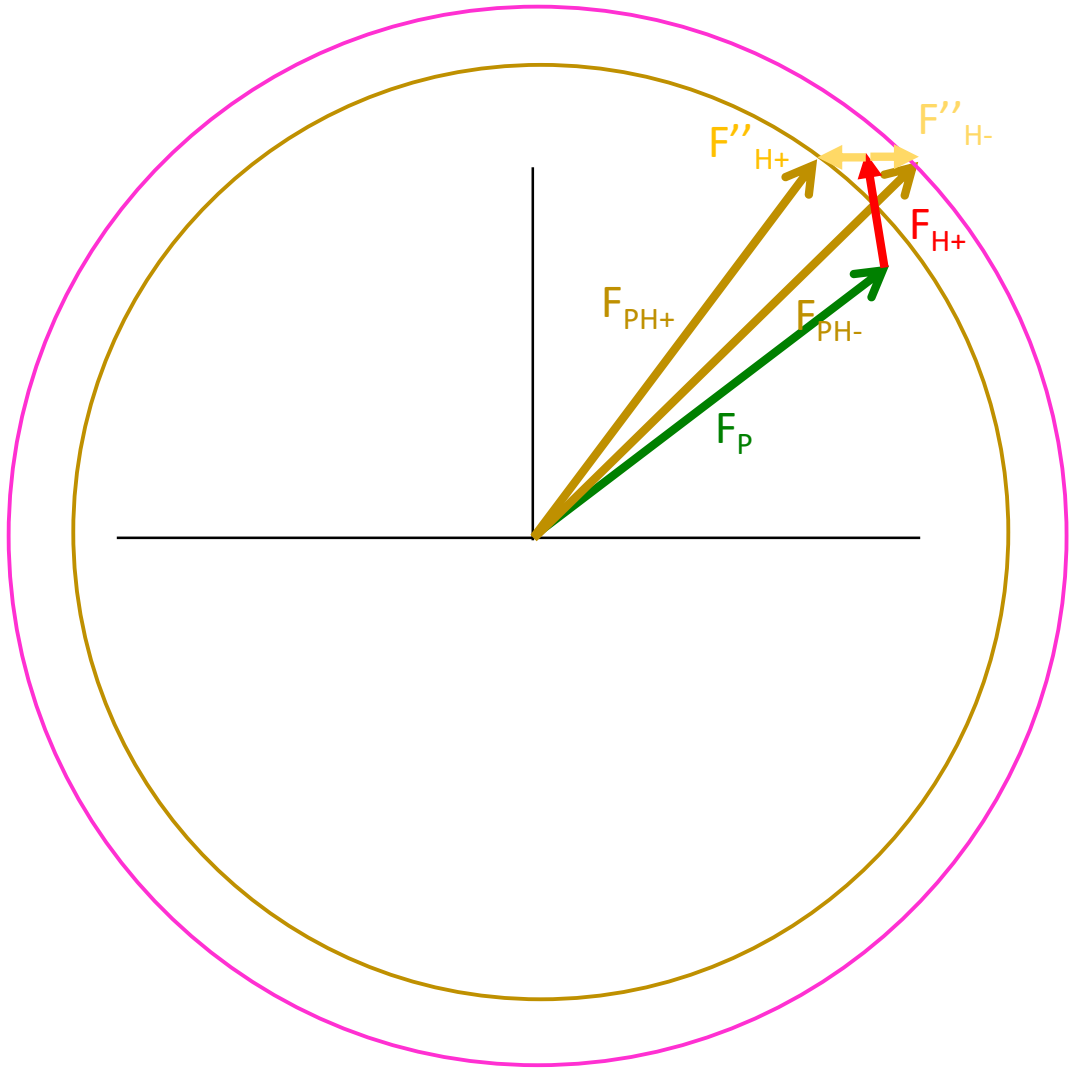




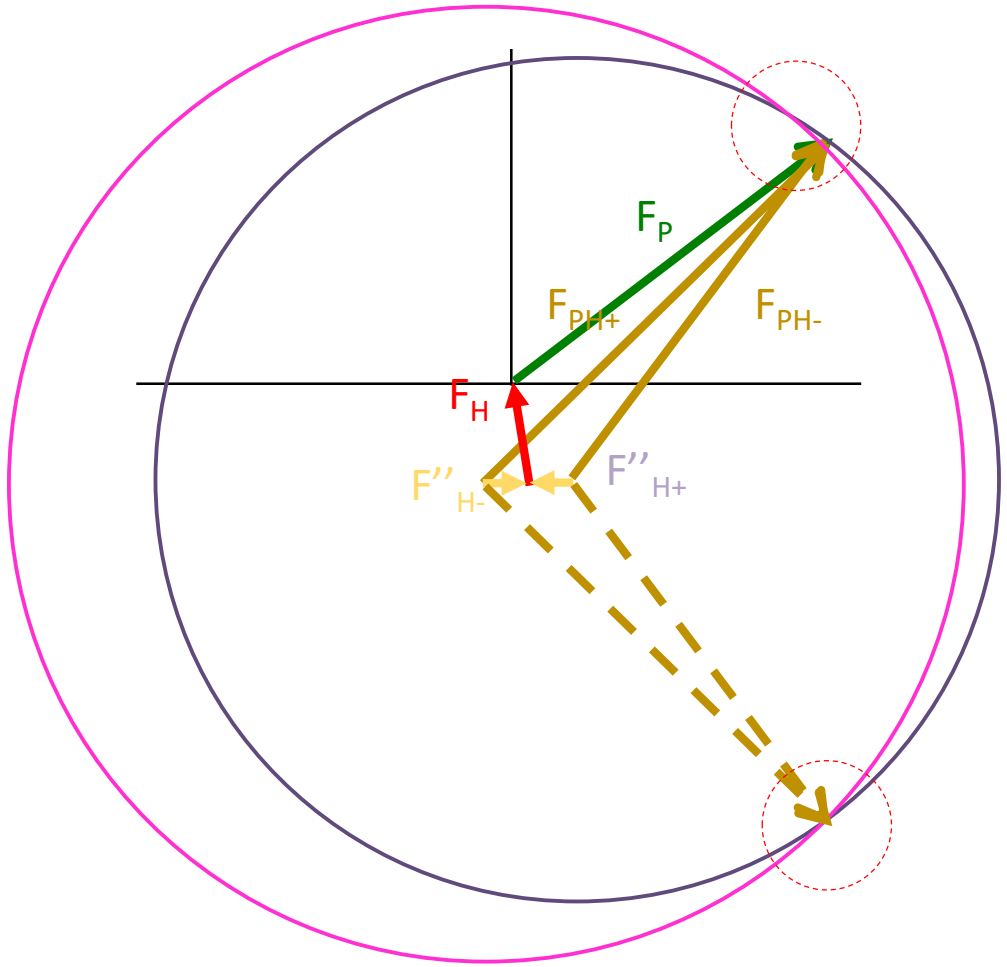
λ_1 = peak = maximum anomalous (f'')

Solving the phases using Single Anomalous Diffraction (SAD)



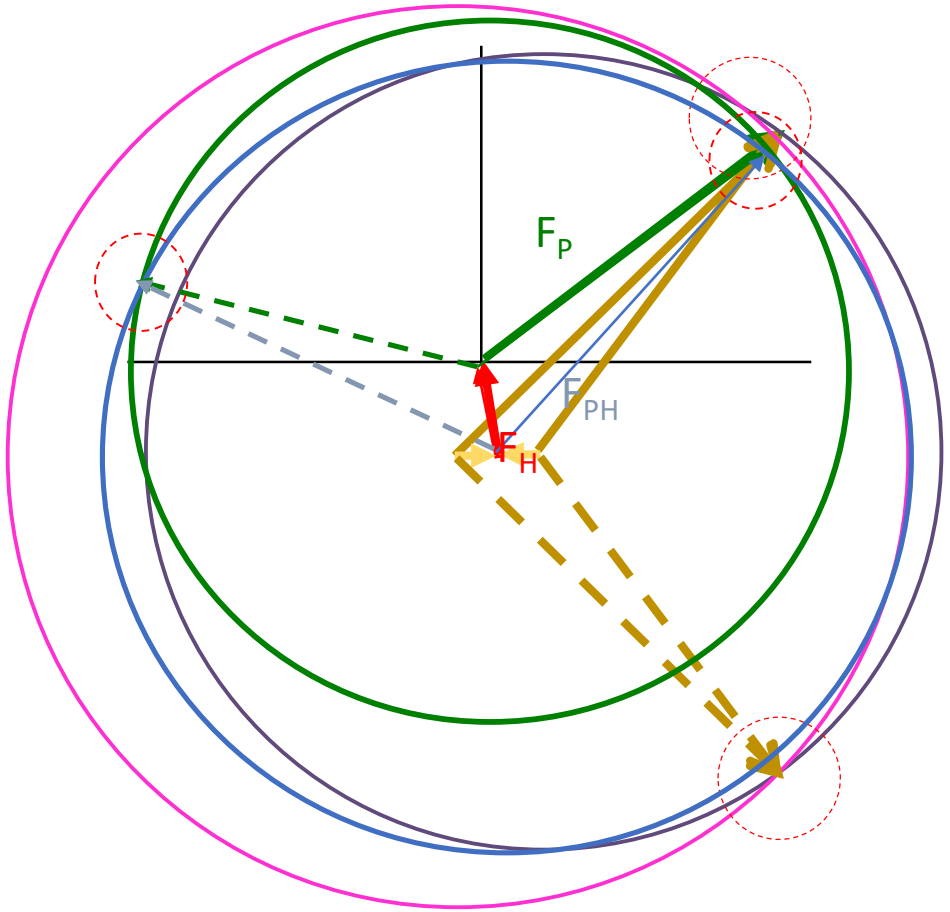


Solving the phases using Single Anomalous Diffraction



- Again we have two phase solutions
- If we had collected data for a MAD experiment we would add the addition wavelengths on to the construction in the same way we added an addition derivative in isomorphous replacement.
- We can also use **density modification** techniques on the electron density maps calculated from both phase solutions.
- In most cases it would be possible to tell which was the correct solution by the fact only one map would look like protein electron density.

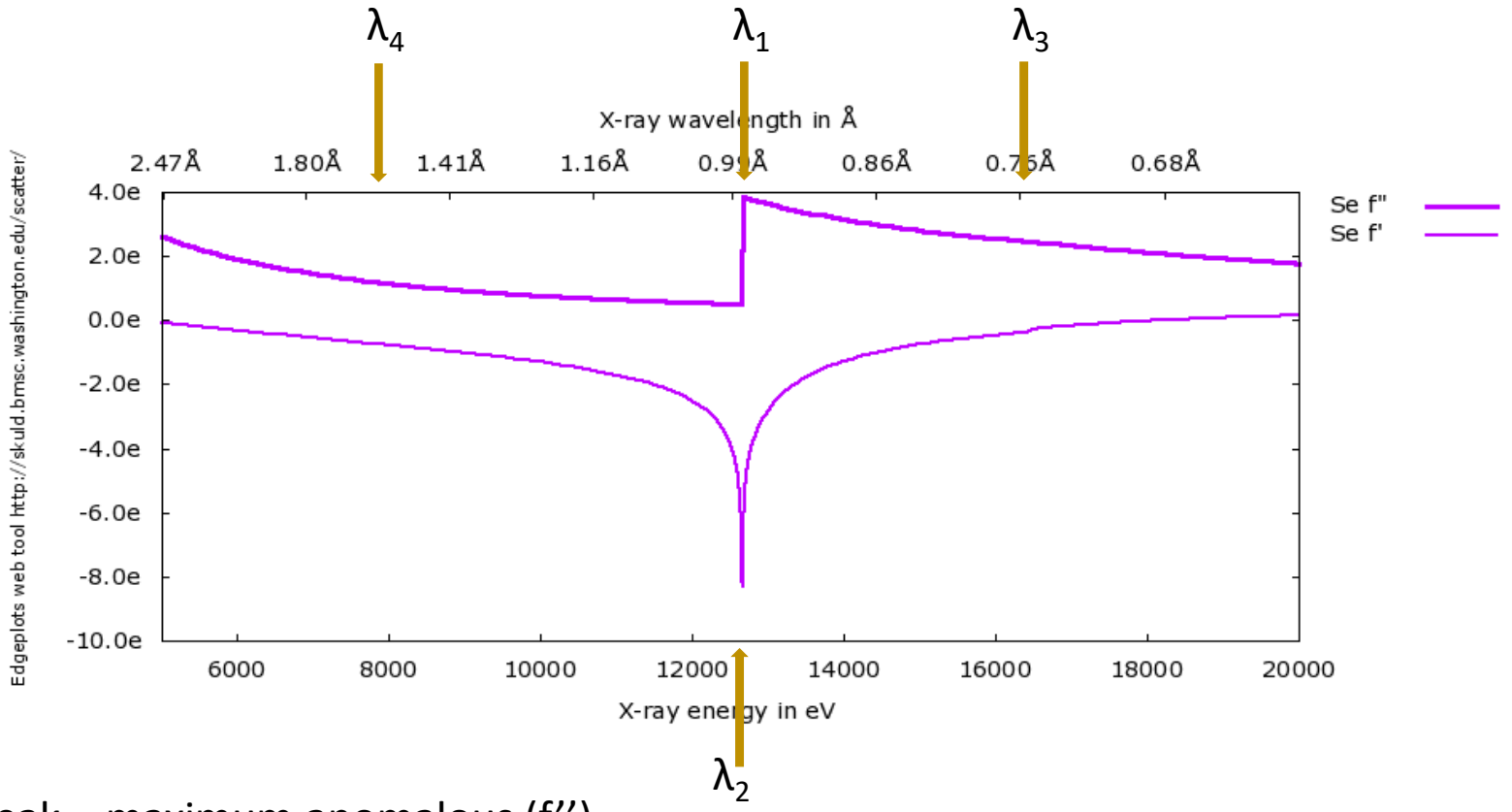
Solving the phases using Single Anomalous Diffraction



However, we can also combine the phasing experiment we did for the isomorphous with the SAD phasing experiment.

This is called SIRAS – Single isomorphous replacement anomalous dispersion.

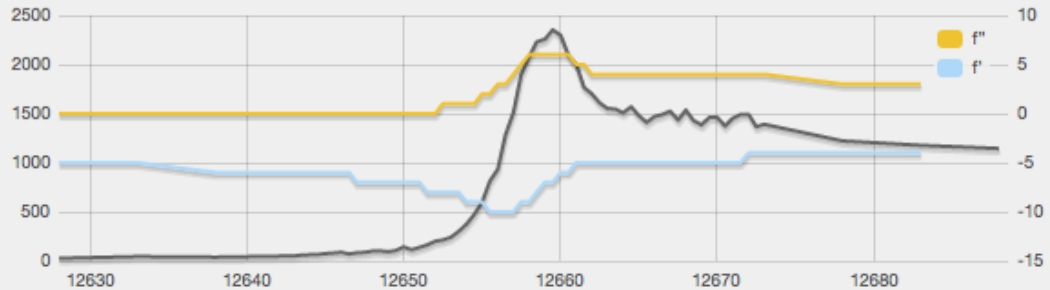
Solving the phases using Multiwavelength Anomalous Diffraction

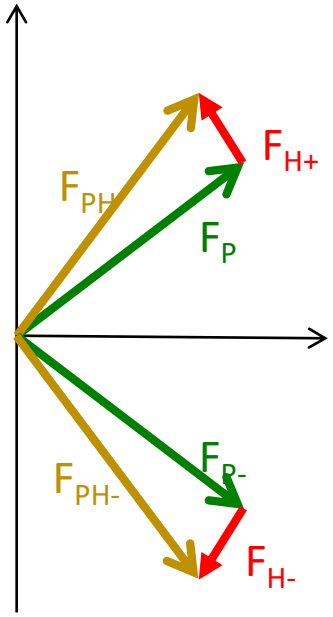


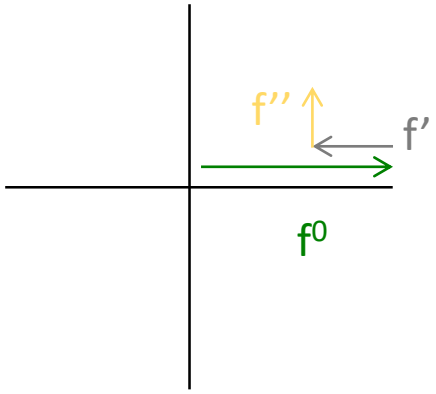
λ_1 = peak = maximum anomalous (f'')
λ_2 = inflection = minimum f'
λ_3 = high energy remote
λ_4 = low energy remote

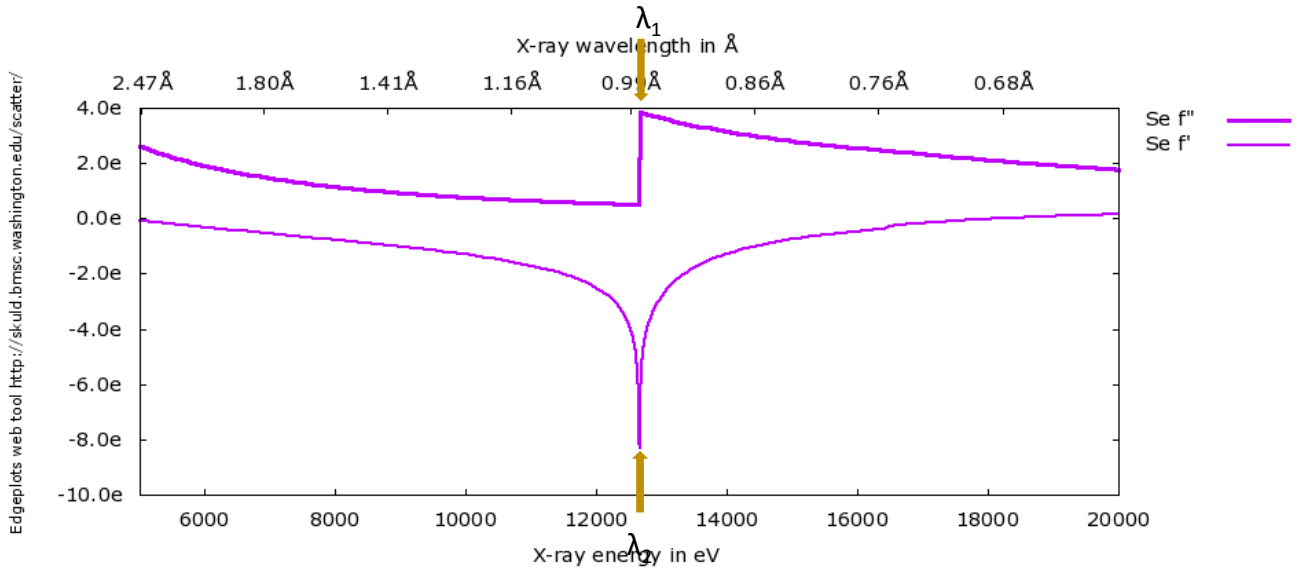
12-05-2016 18:22:26 - Se Edge Scan

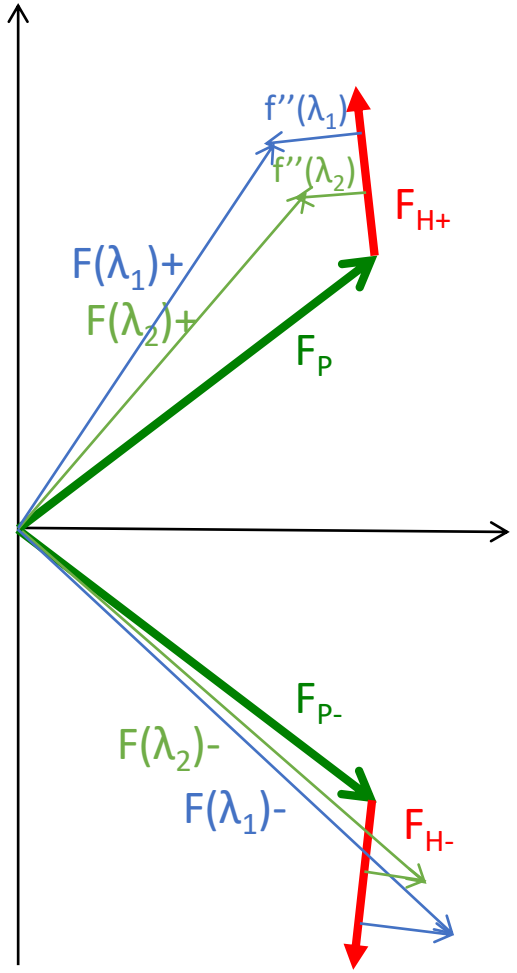
Sample: CV_PACT_OPT_C7_A	Scan File: CV_1.fluo
E(Peak): 12659.5eV (0.9794Å)	f'': 6.67 / f': -7.22e
E(Inf): 12656.490234375eV (0.9796Å)	f'': 3.7699999809265 / f': -10.289999961853e
Exposure: 1s	Transmission: 0.40%
Beamsize: 59.46993637085x20µm	
Comment: Click to edit	



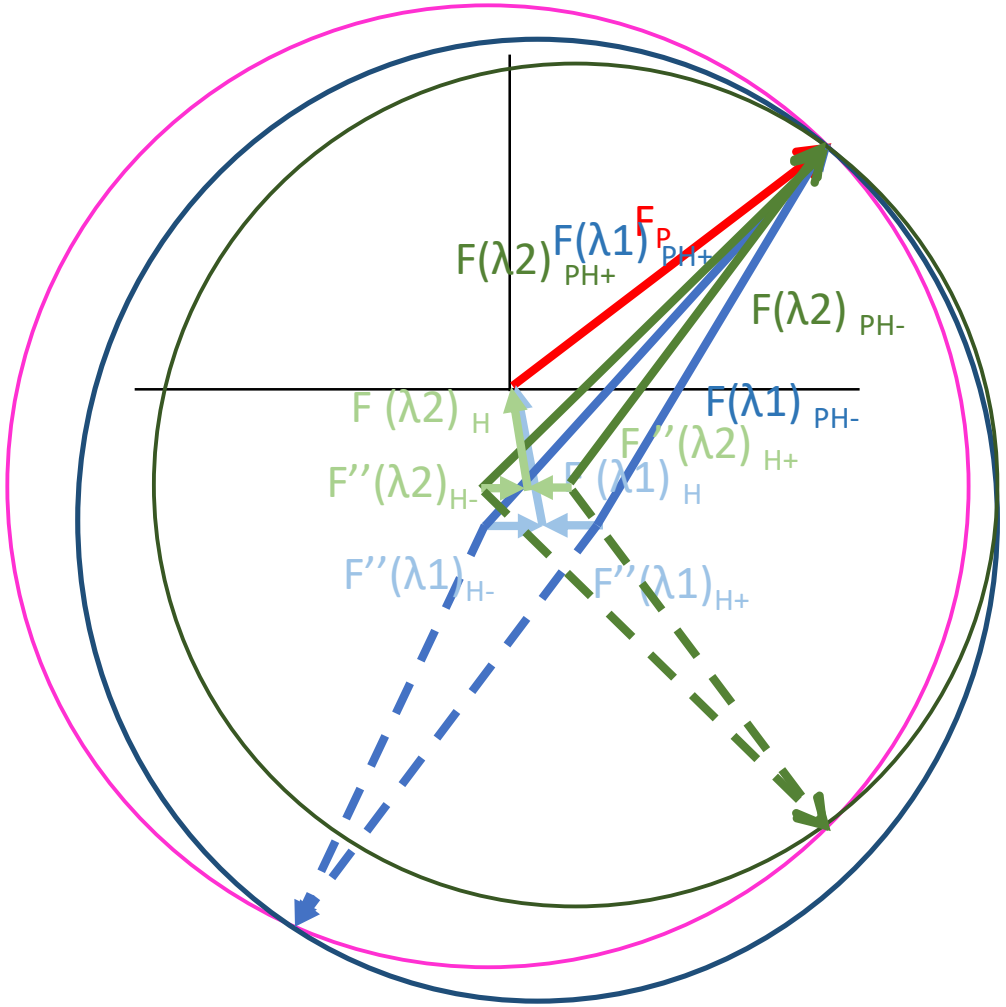

F_{H+}
F_{PH}
F_P
F_{P-}
F_{PH-}
F_{H-}


f''
f'
f^0

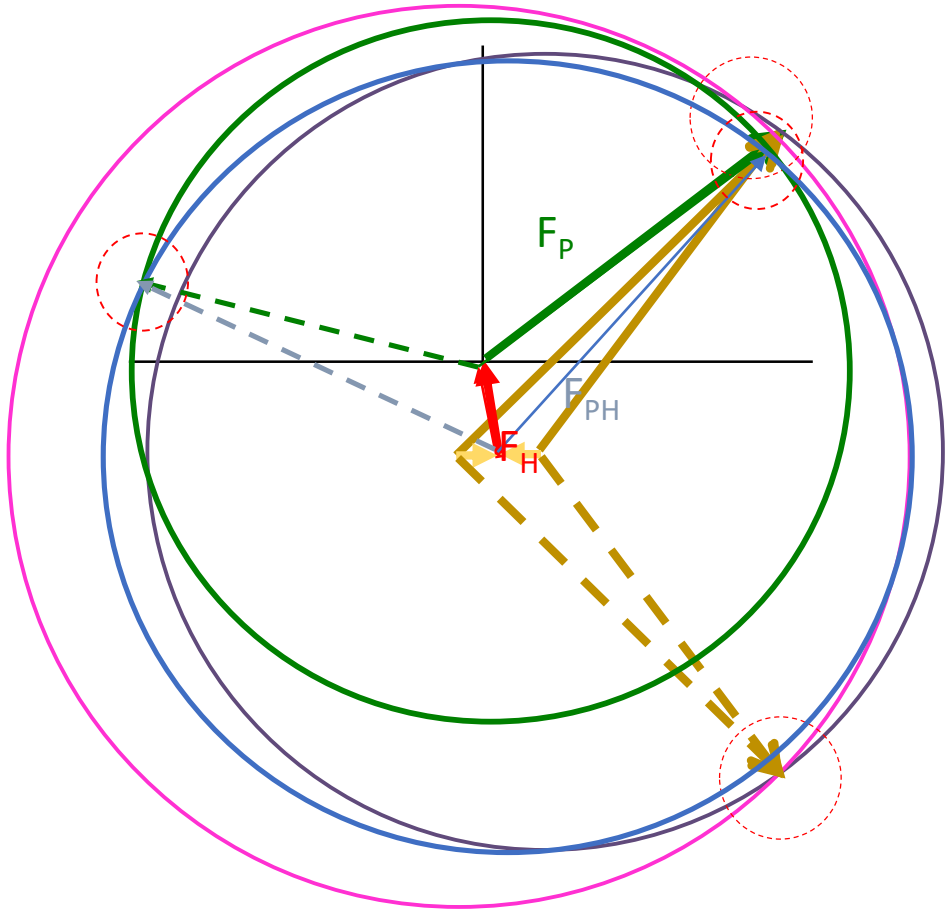

λ_1
X-ray wavelength in Å
2.47Å
1.80Å
1.41Å
1.16Å
0.86Å
0.76Å
0.68Å
4.0e
2.0e
0.0e
-2.0e
-4.0e
-6.0e
-8.0e
-10.0e
6000
8000
10000
12000
14000
16000
18000
20000
λ_2
X-ray energy in eV
Se f''
Se f'
Edgeplots web tool http://skuld.bmsc.washington.edu/scatter/


$f''(\lambda_1)$
$f''(\lambda_2)$
F_{H+}
$F(\lambda_1)+$
$F(\lambda_2)+$
F_P
F_{P-}
$F(\lambda_2)-$
$F(\lambda_1)-$
F_{H-}

Solving the phases using Multiwavelength Anomalous Diffraction



Phase Error



- You may have noticed that the rings in the Harker constructs do not overlap perfectly at a point. This is not (just) sloppy draftsmanship on my part!
- In reality, experimental errors in the measurement of structure factors result in there being a range of possible phase values normally described as a phase probability distribution.
- Modern software normally uses Maximum Likelihood methods to derive the phase probability distributions.

Density Modification (Improvement)

- Initial maps phased purely from substructure atoms are rarely sufficiently interpretable to build a model
- Density modification encompasses a range of techniques to take advantage of prior knowledge in order to improve our electron density maps, generally by improving our estimates of the phases.
- Many of these calculations are carried out in real space.

- Solvent Flattening
  - Solvent Flipping
- Histogram Matching
- Averaging

Solvent Flattening (and phase extension)

- Disordered region of electron density $\rho(x)$ have a constant flat value of around 0.33 $e^-/Å^3$ for pure water – protein has a higher average electron density of around 0.44 $e^-/Å^3$
- A mask is constructed to describe contiguous solvent regions of the initial map (if this is not possible, density modification will not work)
- Within the solvent region the electron density is set to the average value and new structure factors are calculated.
- This procedure is most effective when solvent content is relatively high.

Solvent Flipping

- There is a risk of bias in the procedure described so far. The new information (the flattened solvent region) and the original information (the protein region) will tend to correlate which can bias towards the original map.
- Solvent flipping does not set the solvent region to an average density value but rather inverts the sign of voxels in the solvent region.
- This allows the protein and solvent regions to be treated independently and reduces bias.
- Solvent flipping is implemented in SOLOMON and used in autoSHARP. SHELXE also uses solvent flipping to flip voxels considered unlikely to be protein.

Sphere of Influence

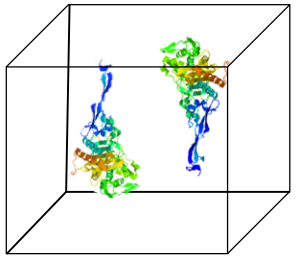
- The program SHELXE uses a somewhat different implementation of solvent flipping, relying on calculating the electron density at the surface of a sphere of 2.42Å radius.
- The 5% highest values of electron density within an initial map are treated as protein within this method and two further definitions are allowed, a solvent region and a “fuzzy” region .
- Note that this method does not rely on the initial definition of a solvent mask region and will self correct over many cycles if the definitions of protein and solvent are not initially perfect.

Histogram Matching

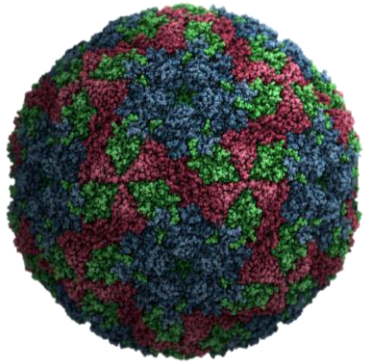
- Histogram matching is a complementary method to solvent flattening.
- It is based on methods used for image enhancement.
- Images can be dramatically improved by matching the distribution of voxel intensities to that of a high quality image of a similar object.
- Electron density histogram density matching takes advantage of the observation that the distribution of electron density values in a map is almost identical at a given resolution for any protein structure, regardless of the structural details.

Noncrystallographic symmetry (NCS)

Proteins often have more than one molecule in the asymmetric unit. They could be true representations of the biological unit or it could be coincidental and purely a result of the way the protein molecules pack.



- By calculating the relationship between the molecules in the ASU, it is possible to average the electron density of the molecules.
- This creates a better signal to noise ratio and therefore more defined electron density for the individual molecules.



A great example of using NCS is in solving virus structures where there are large numbers of identical structure forming the capsid. The use of NCS makes it possible to generate interpretable maps.

Cross-crystal Averaging

- Sometimes a protein will crystallise in more than one crystal form, but none of them provide both the resolution and phase information needed to make map interpretation straightforward.
- It is possible to take a region of electron density from one crystal and relate it to that in a second crystal via rotation and translation transformations.
- These regions of electron density may then be average in a similar way to NCS related regions within the same crystal.
- This is a very powerful method, since the protein structure will be the same (or very similar) in both crystals but errors from noise will be different and therefore lost in averaging.