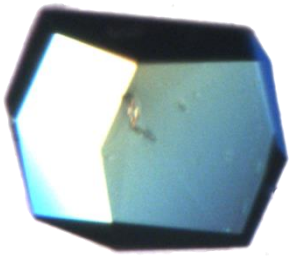
The Phase Problem

Dr. Ed Lowe –
edward.lowe@bioch.ox.ac.uk

Department of Biochemistry

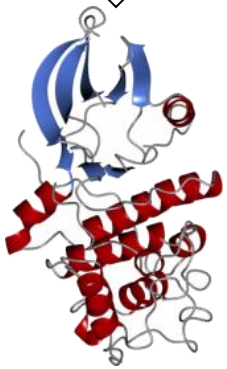
UNIVERSITY OF OXFORD

Phasing



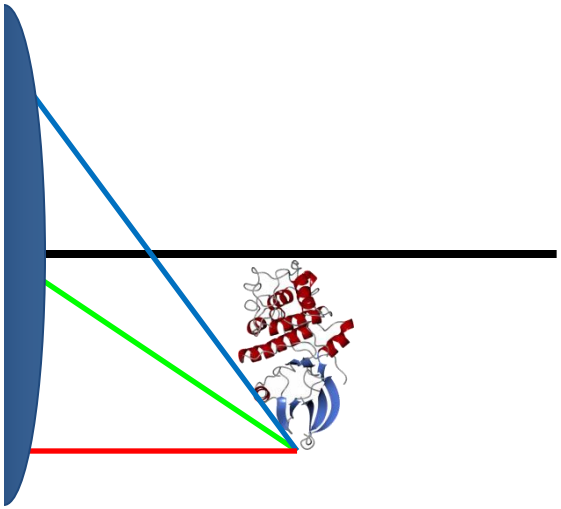
Grow Crystals

Magic happens…



Structure

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

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

Electron density at a point

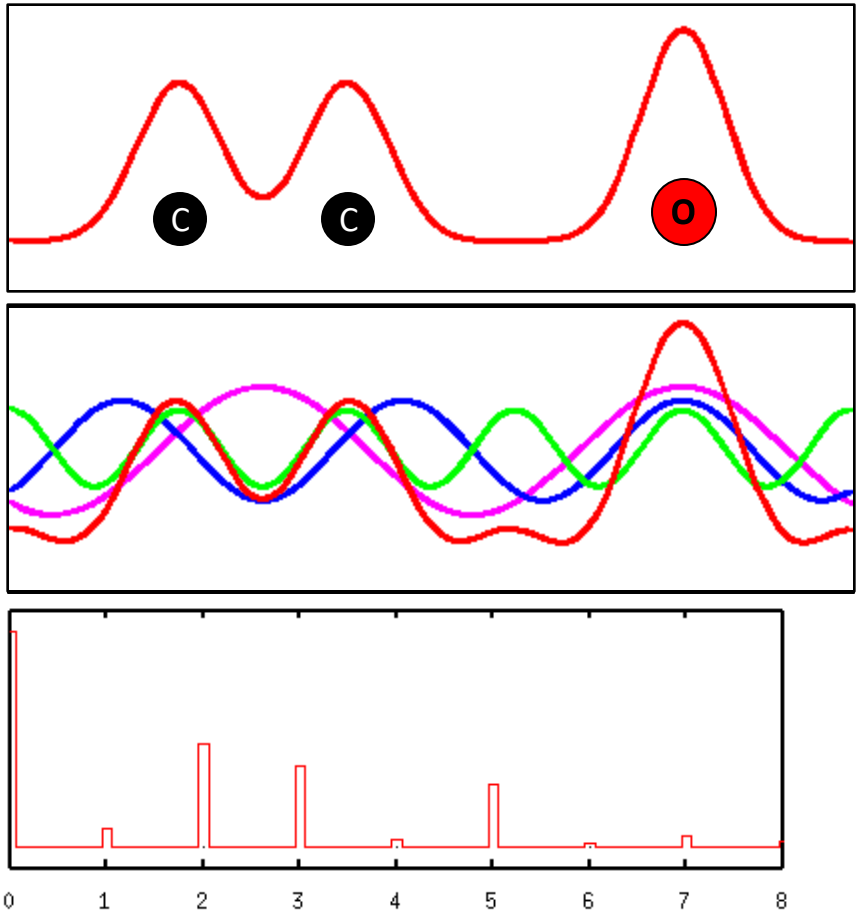
Amplitude

Phase

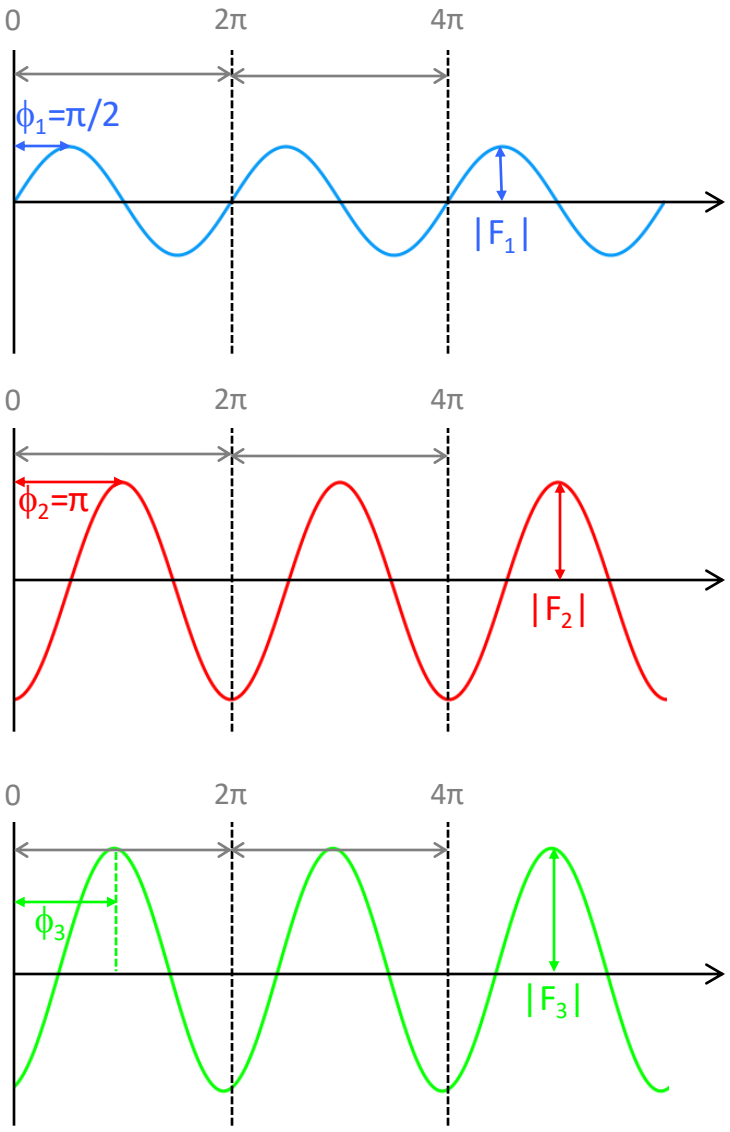
Fourier Syntheses and Transforms

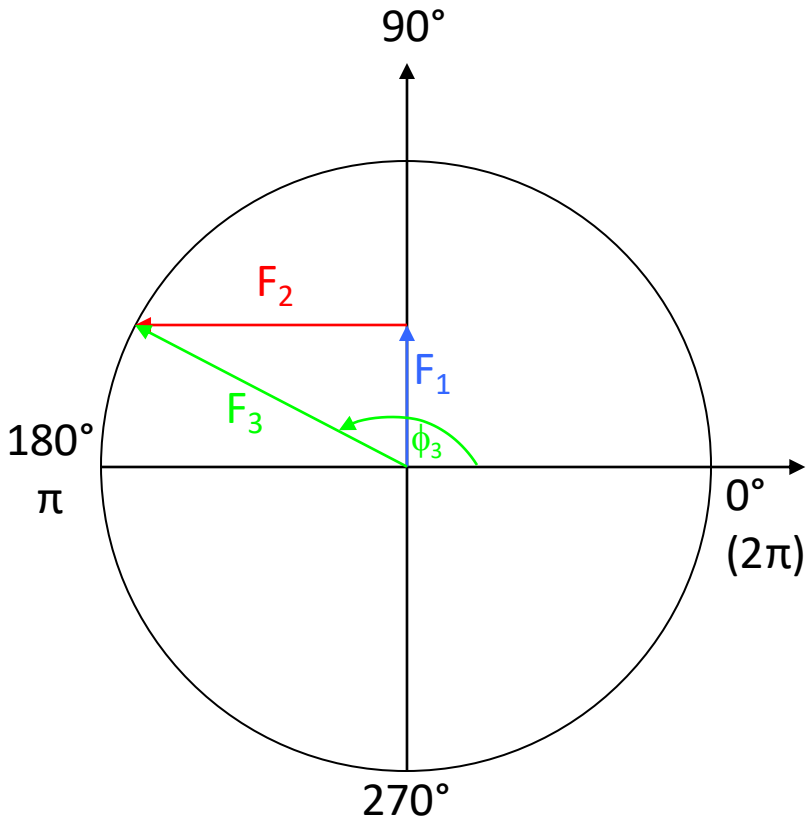
A Fourier synthesis is the representation of a function in terms of a set of sine waves

- A simple crystal with a unit cell containing one oxygen and two carbon atoms
- Trying to represent the unit cell in terms of sine waves, we start by adding a wave of frequency 2
- Note that one peak crosses the oxygen atom whilst the other spans the two carbons
- Now we add a wave of frequency 3. It has a different phase, meaning that we start at a different point in the wave. The amplitude is also different
- Finally we add a wave of frequency 5. Two of the peaks correspond to the carbon atoms
- The sum of the waves is a good approximation of the contents of the unit cell (given appropriate choices of frequency, amplitude and phase)
- The Fourier transform of the unit cell consists of a series of peaks, the largest of which are at 2, 3 & 5 on the x-axis
- These correspond exactly to the sine-wave frequencies used to reconstruct the unit cell. The peak height also corresponds to the sine-wave amplitudes
- For more Fourier theory and some excellent figures: http://www.ysbl.york.ac.uk/~cowtan/fourier/ftheory.html



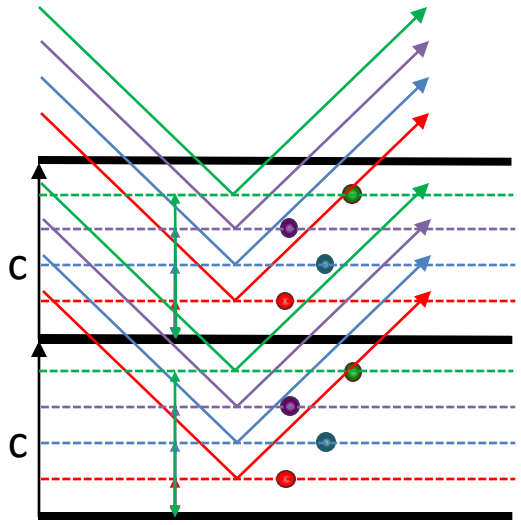
Addition of waves in the complex plane

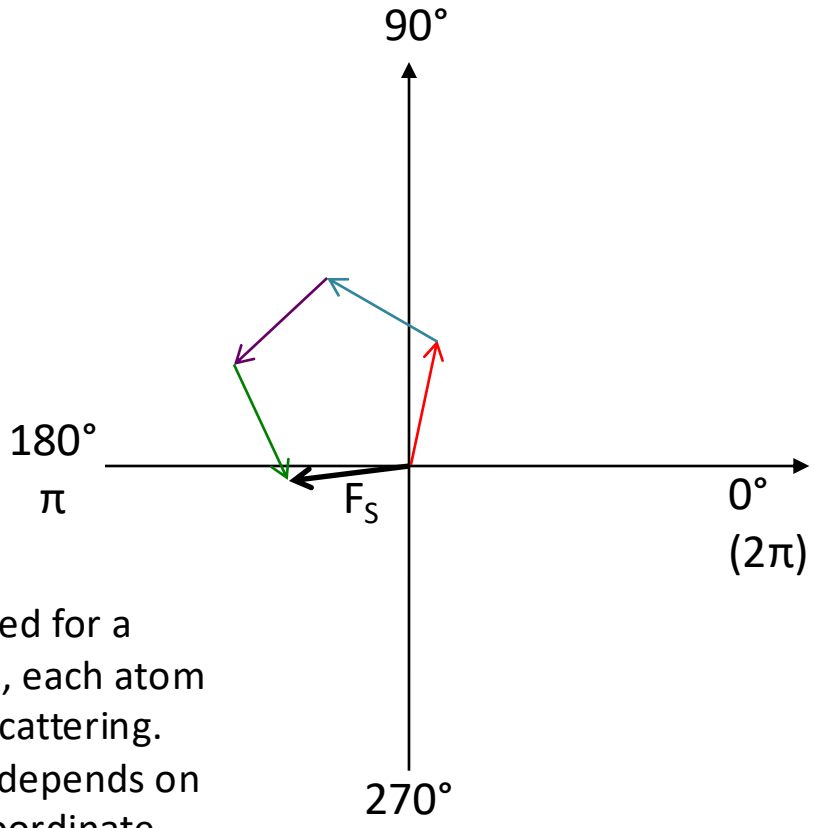




NB. Experiments are monochromatic, so all waves have the same wavelength

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 will the structure factor change if our atoms remain in the same positions but we consider a different set of Bragg planes?

A: The same phase but different amplitude

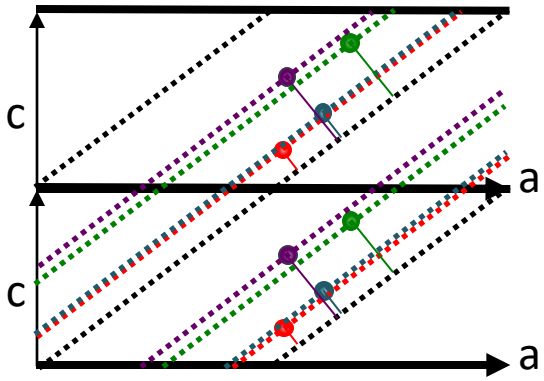
B: The same amplitude but different phase

C: The same phase and the same amplitude

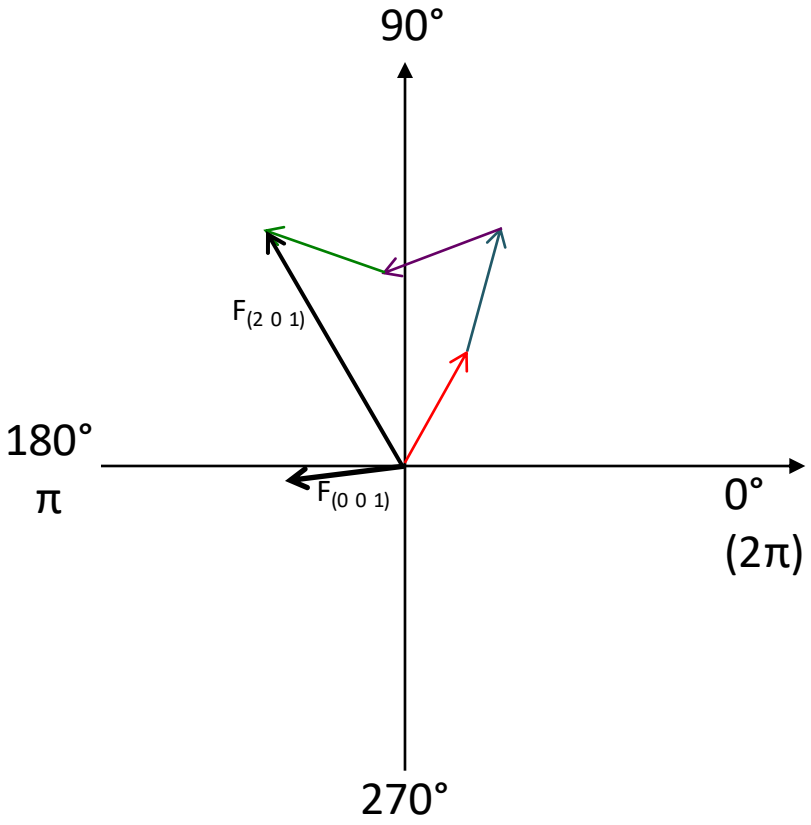
D: A different phase and a different amplitude

E: It will have a different name

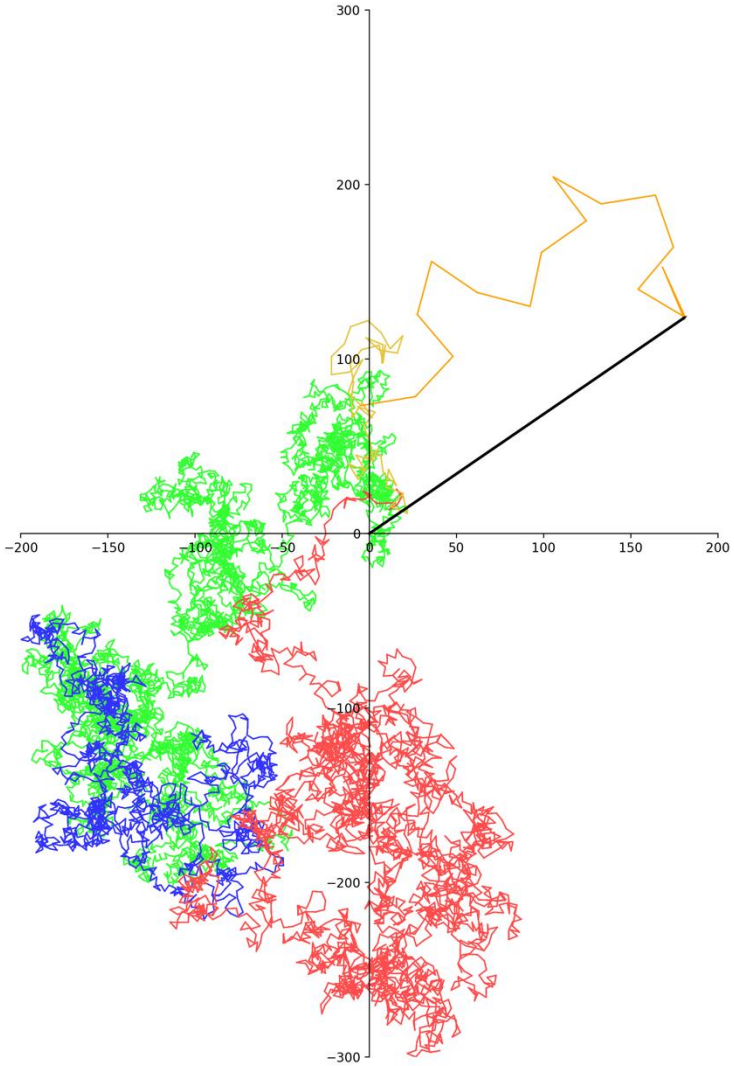
How does this look with a different set of Bragg planes?

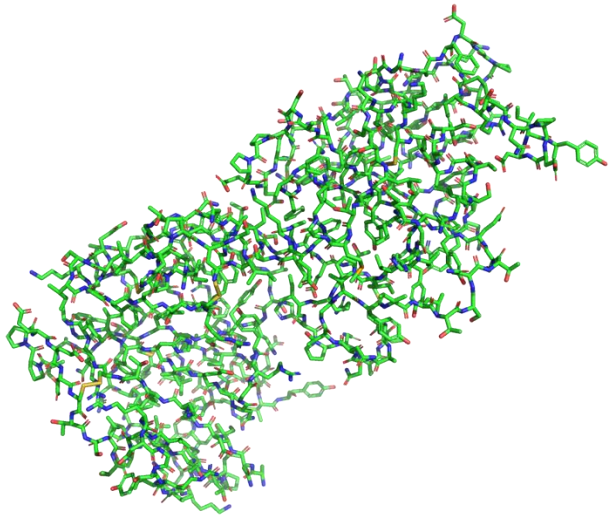


Indices (2 0 1)



A bit more complicated for a protein…





What can we achieve if phases are known?

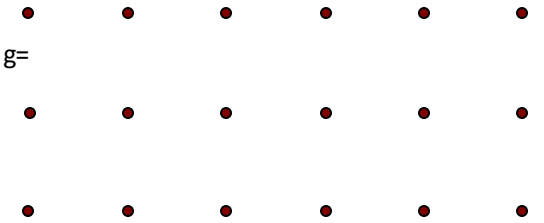
Convolution

- Convolution of two functions mixes them together. For functions f and g this can be expressed by the formula:

$$C(x) = \int_{\eta=0}^{1} f(\eta) g(x-\eta)\, d\eta$$

- It might be helpful to illustrate this:

If f= 

and g= 

- The convolution of f and g (written f*g) is:

- The convolution theorem states that the Fourier transform of the convolution of two functions is equal to the product of the Fourier transforms of the two separate functions.

Inversion

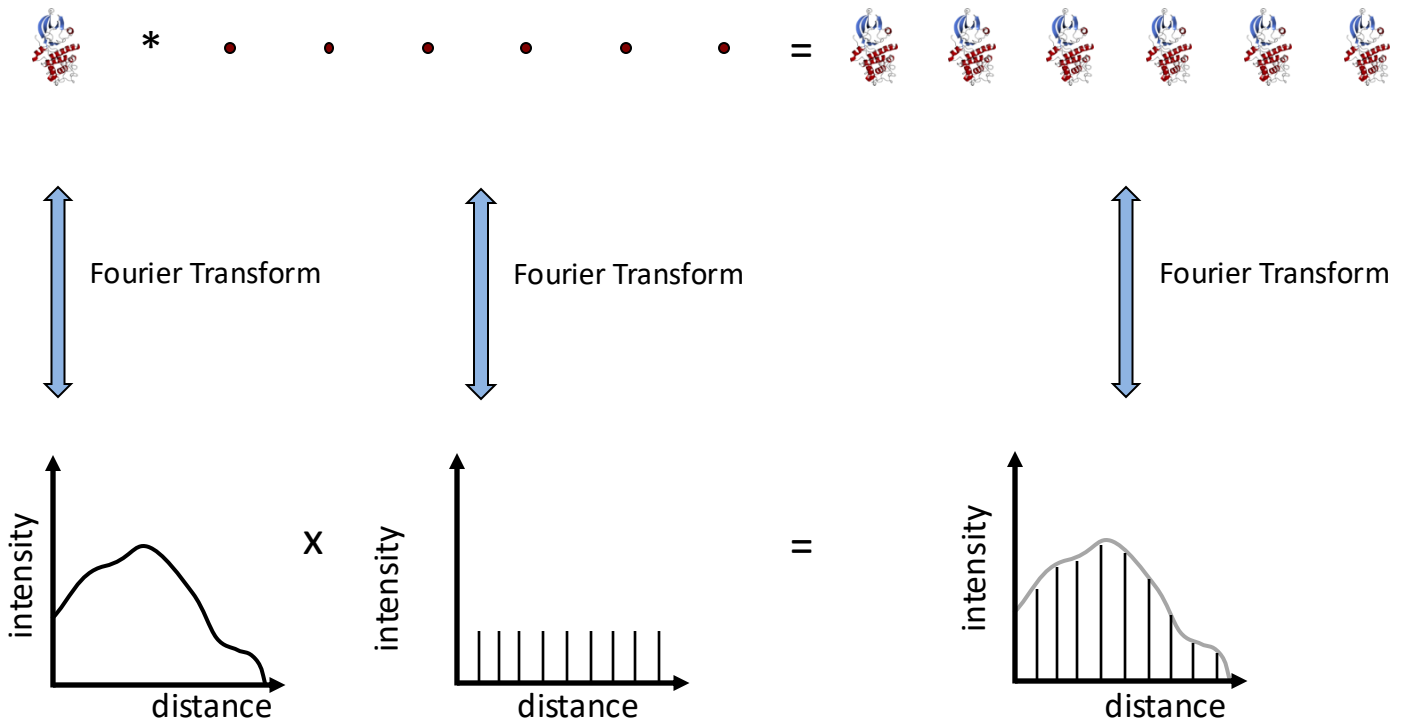
- The inversion theorem states that:

  If function F is the Fourier transform of function G, then function G is the Fourier transform of function F

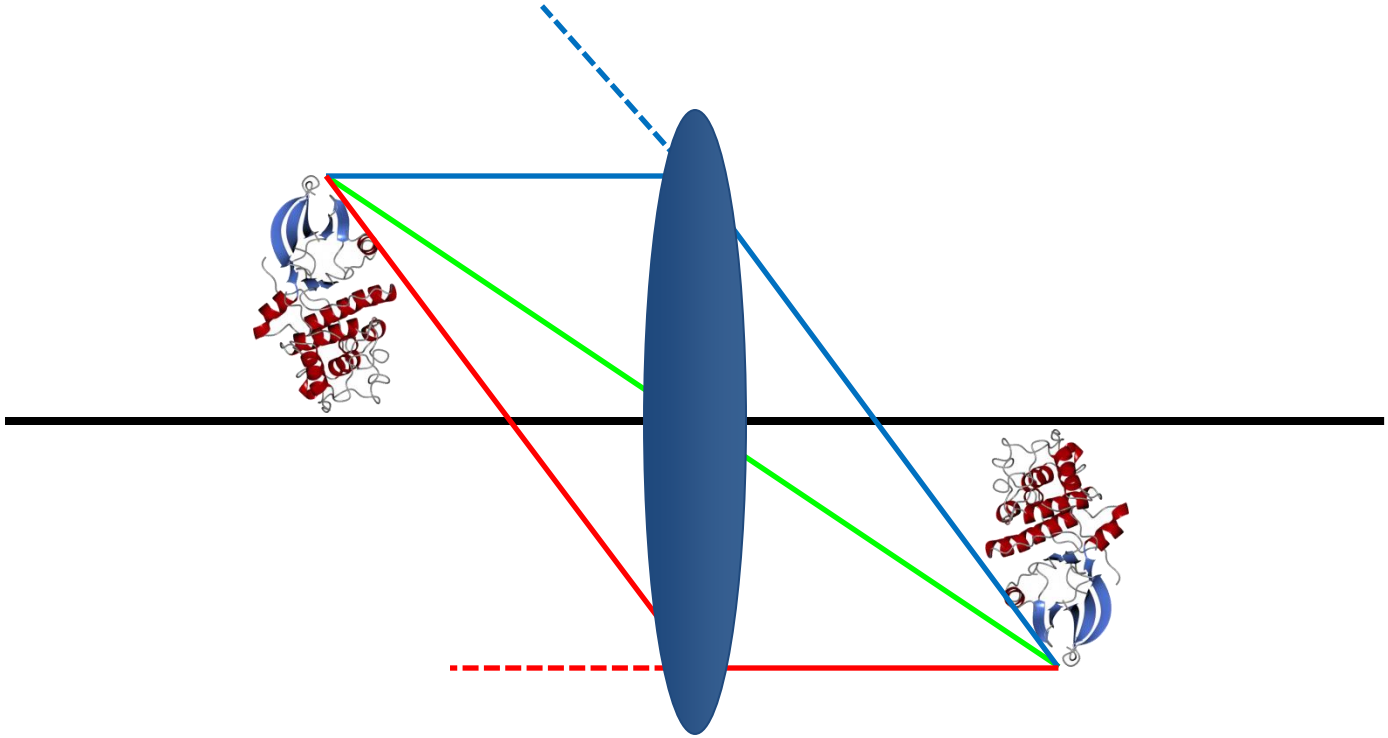
- This is important because:

  In the same way as we can use Fourier transforms to calculate the diffraction pattern of a protein crystal, we can use Fourier transforms to calculate the electron density distribution in a crystal from an observed diffraction pattern

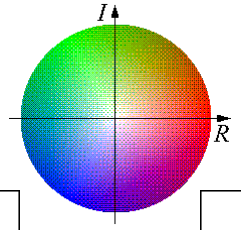
Putting all of this together:



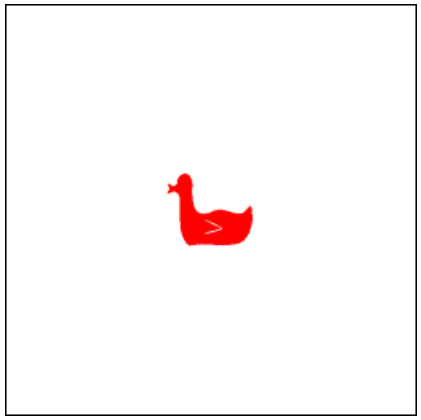
The Phase Problem

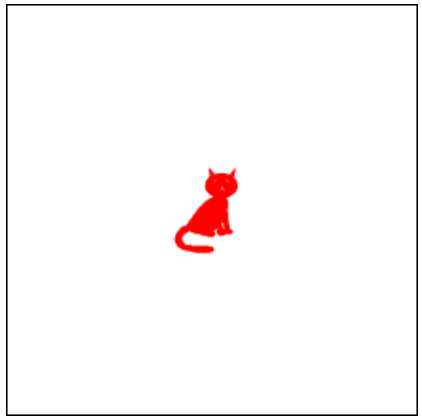


Our problem is how to achieve the recombination of scattered waves without having to invent an X-ray lens



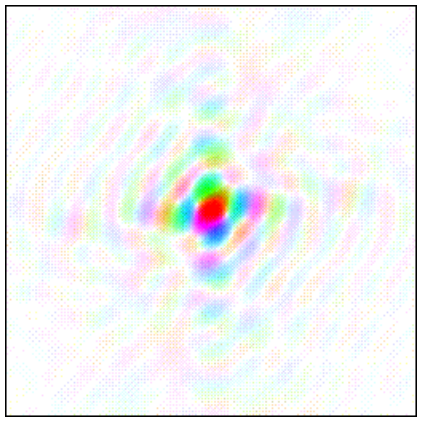
Representing **intens**

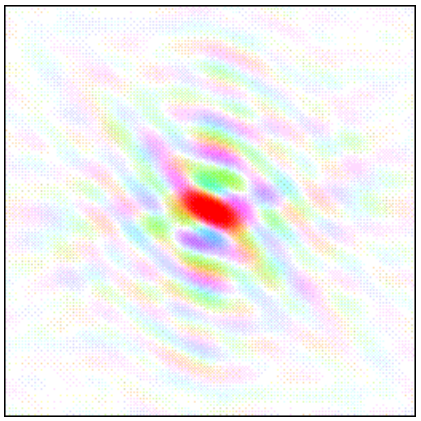


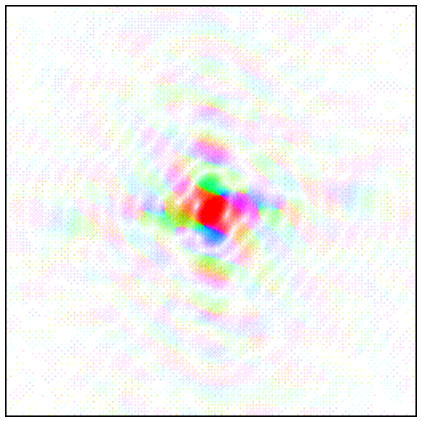


Fourier transform

Fourier transform







F(Duck),ϕ(Duck)

F(Cat),ϕ(Cat)

F(Duck),ϕ(Cat)

What will the reconstructed image most resemble?

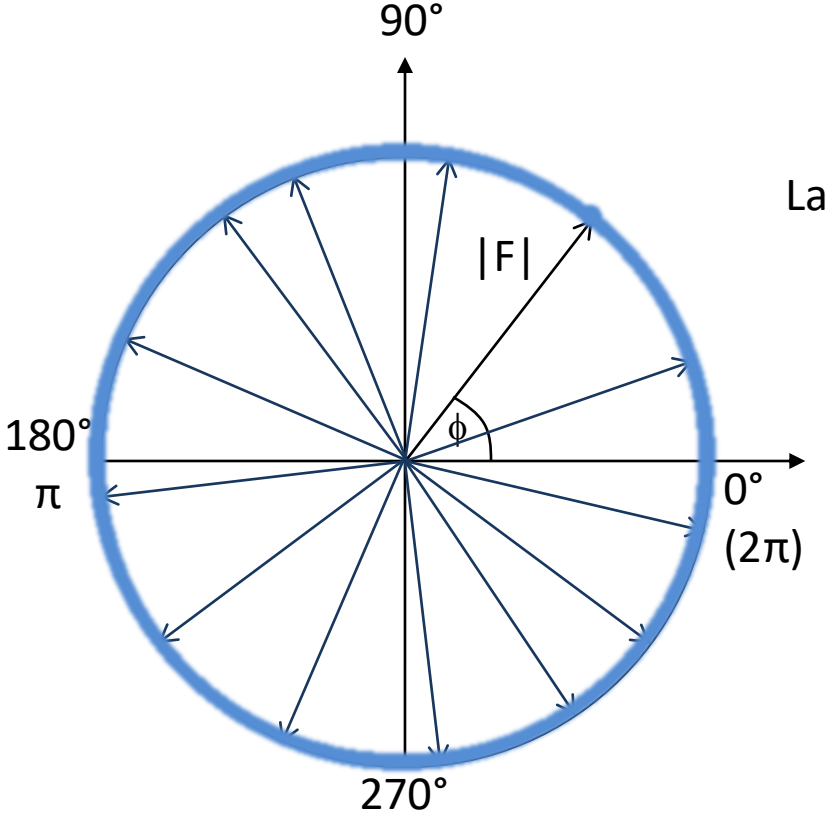
- A: A Duck
- B: A Cat
- C: Mixture (more duck)
- D: Mixture (more cat)
- E: A Tin of Sardines

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

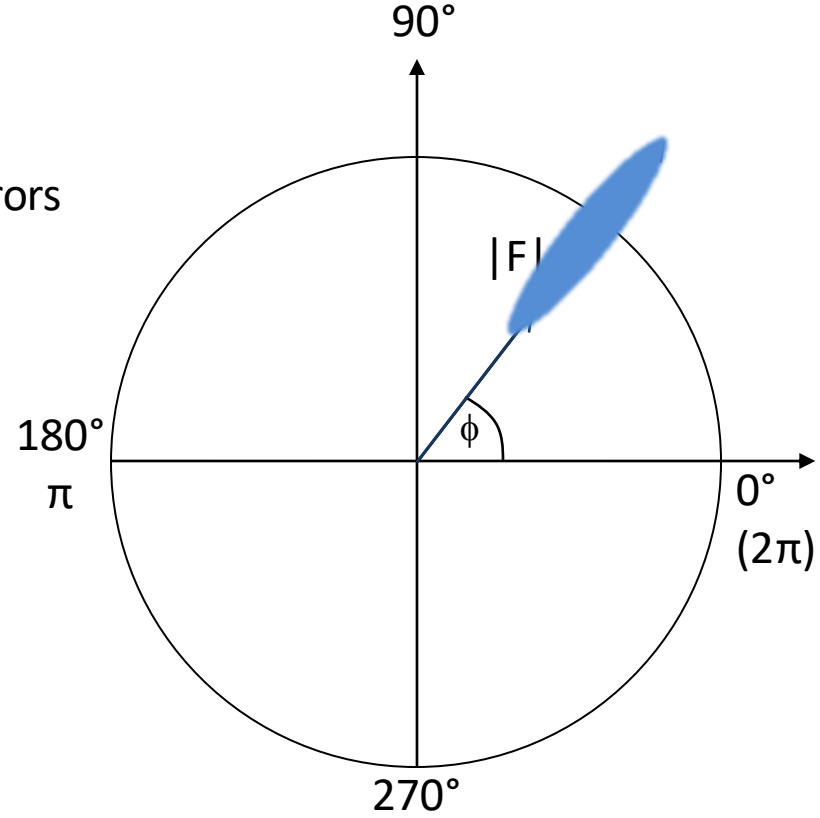
True or False?

But… there is no need to panic just yet!

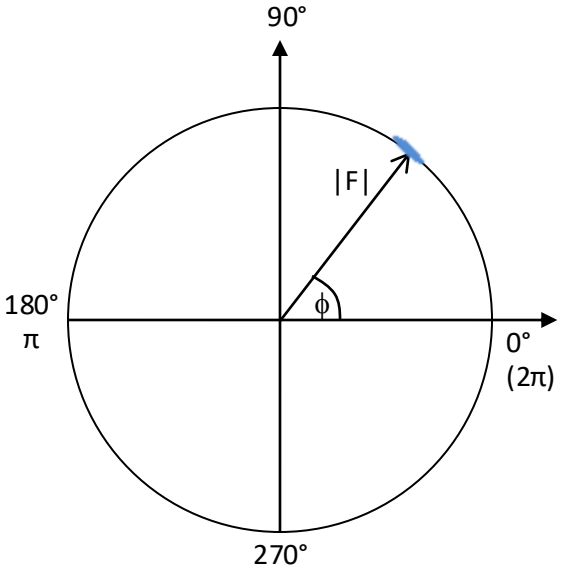
- This disastrous situation is greatly exaggerated when the two sets of phases are completely different.
- It relates to the fact that the vector you get by pointing a correct length vector in the wrong direction is (on average) further away from the true result the that obtained by pointing a vector 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.

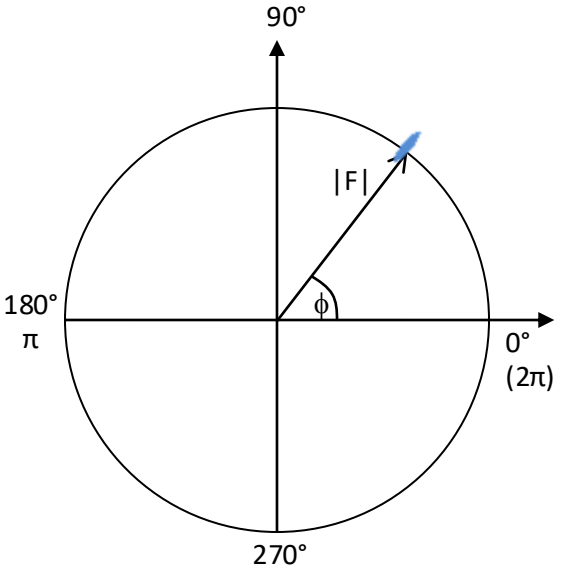


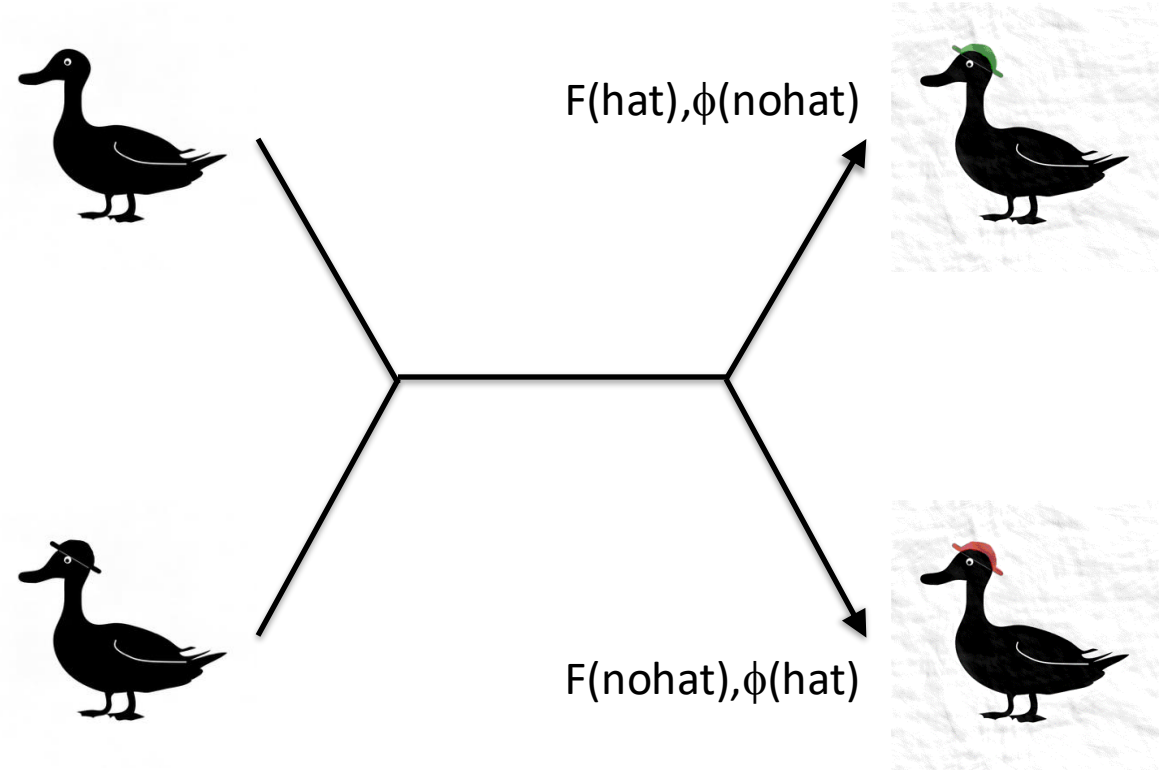
Large errors



Small errors






F(hat),ϕ(nohat)
F(nohat),ϕ(hat)

Phasing

When we talk about solving protein structures we mean we want to calculate the electron density.

The structure factors are calculated from the intensities of the measured reflections, these in turn can be described as a function of electron density.

Refers to the volume of the unit cell

As a result of discrete spots

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

Electron density

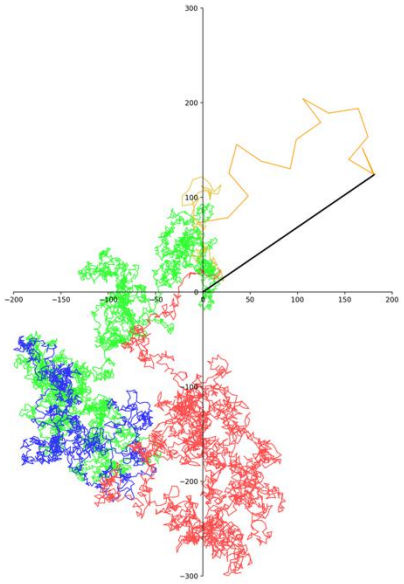
Sum of Structure Factors for all hkl

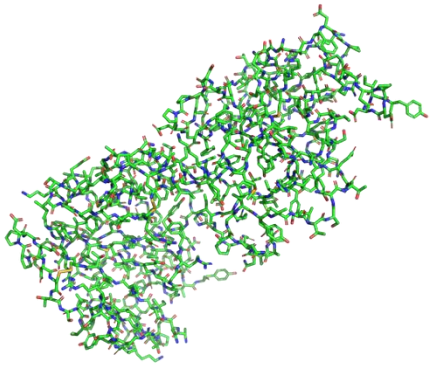
Phase contribution

Solutions to the Phase Problem

- "Copy" the phase contributions from a sufficiently similar known structure
- Calculate the phase contributions from experimental data

Think back to this Argand diagram...





- If the constellation of atoms is sufficiently similar.
- And if it can be placed in an appropriate position and orientation in the unit cell
- Then the structure factors calculated from these atoms will be similar to those derived from our experimental structure
- Meaning that the phase terms can be borrowed to allow the calculation of an initial electron density map

This is Molecular Replacement

Thanks to Alphafold and Rosettafold we now have access to a suitable model structure in most cases

What can we achieve if phases are not known?

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.

$$\varphi_{-h} + \varphi_{k} + \varphi_{h-k} \square 0$$

- 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)

- This method can be useful in solving the heavy atom substructures.

- There are fewer heavy atoms in the unit cell and they are further apart, so high resolution data is not needed.

Patterson function

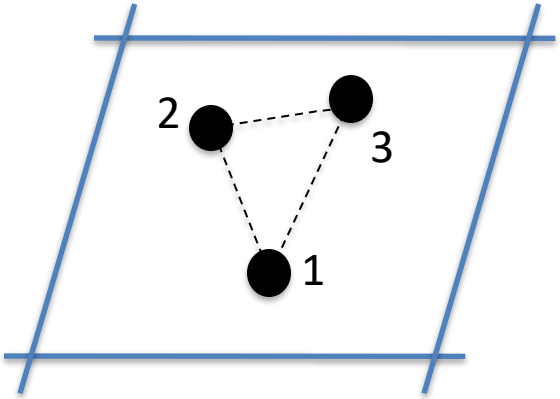
This is a Fourier summation of intensities (as opposed to Structure factors) with phase angles set to zero.

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

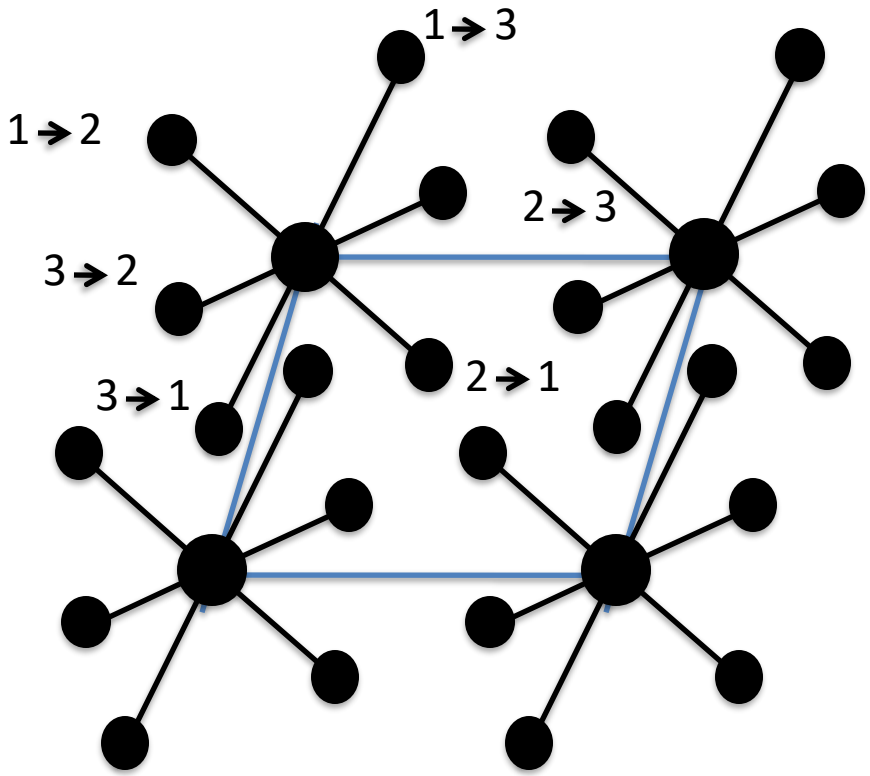
Where *u*, *v* and *w* are the coordinates in the Patterson cell to avoid confusion with x, y and z from the real unit cell (although both unit cells have the same dimensions).

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

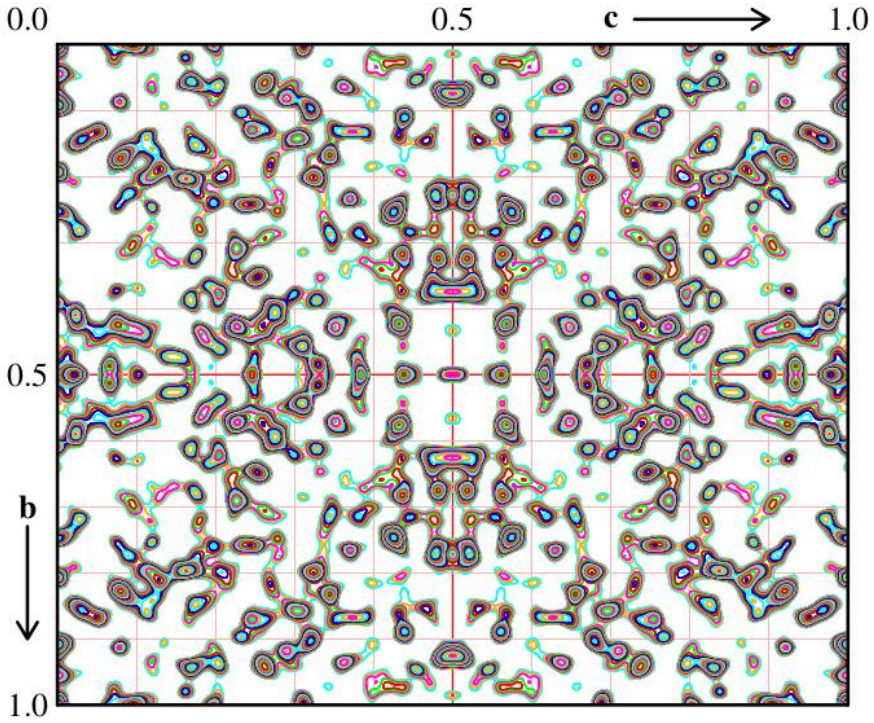
B: 99,900

C: 999,000

D: 9,990,000

E: 42

Patterson map



A protein contains 1000's of atoms that all contribute to it's diffraction. The Patterson becomes too complex and you can't solve the atomic positions. A unit cell containing N atoms will generate a Patterson map containing N^2-N non-origin peaks.

- Is there a way we can utilise Patterson functions to help us phase the protein diffraction data?

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!

A simpler structure…

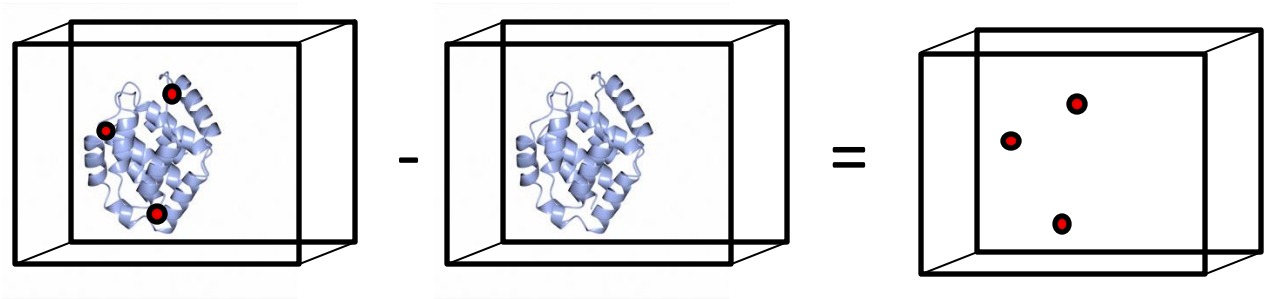
- We need some way by which we can identify the positions of a small number of atoms – from this we can estimate the phases for all reflections and bootstrap our way up.
  - Atoms placed at random positions
  - Atoms with much higher atomic number than the C, O, N and S making up most of protein molecules
  - Atoms that scatter ‘anomalously’ in a wavelength dependent manner

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.

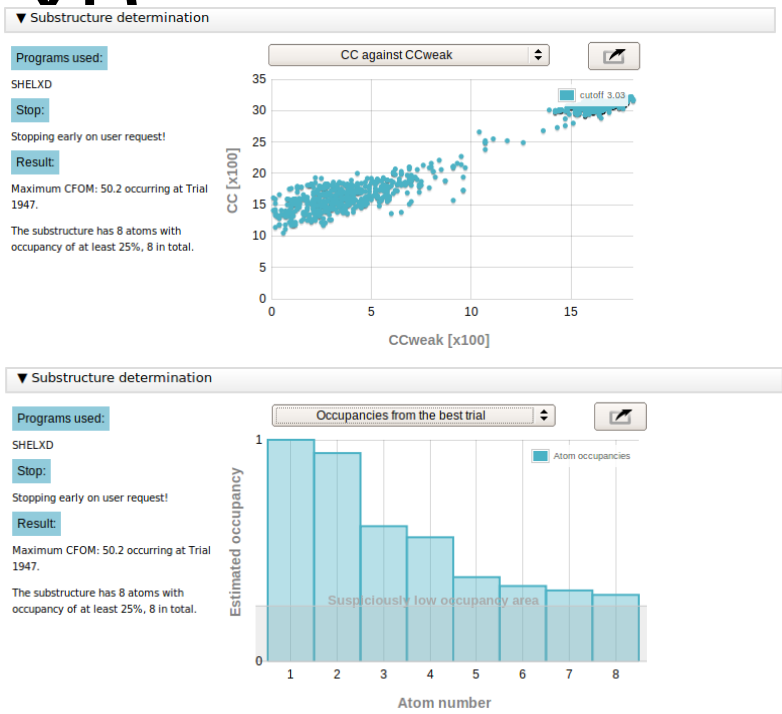
The Patterson function is then used to calculate the positions of the heavy atom(s).



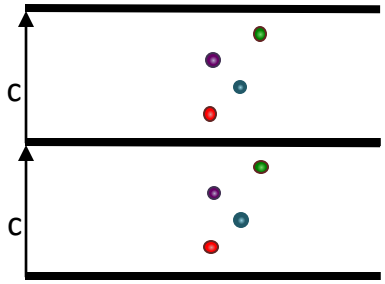
$$F_{PH} \quad - \quad F_P \quad = \quad F_H$$

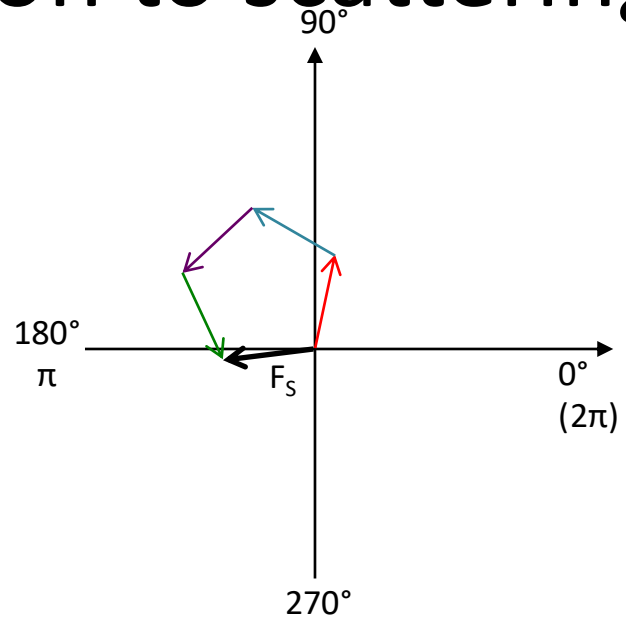
SHELXD

- Very powerful (and flexible) for finding sites
- Assumes all sites are the same element and compensates for differences in scattering via occupancy
- Can seed with Patterson vector rather than random or Patterson derived atomic coordinates
- Can search for a single site at low resolution and split at higher resolution (SS bonds in SSAD)

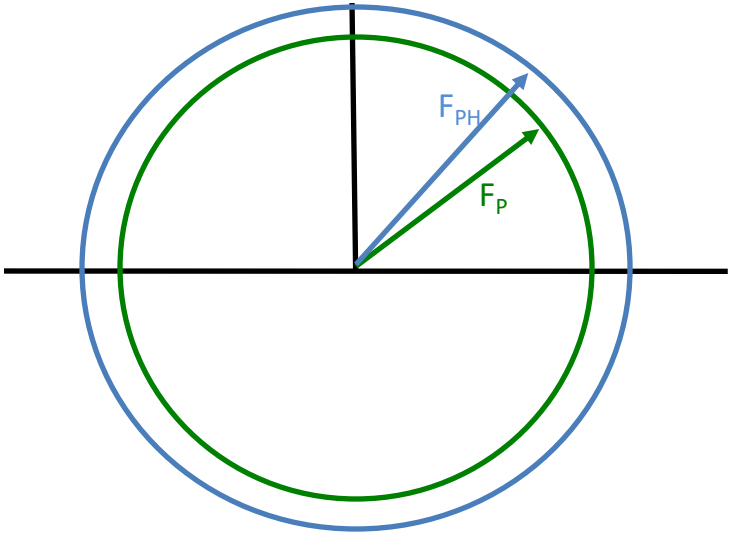


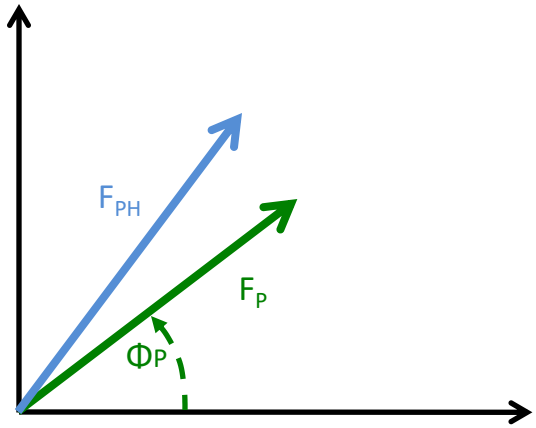
Remember – if we know the position of an atom we can calculate the phase of its contribution to scattering.





Solving the phases using Isomorphous Replacement

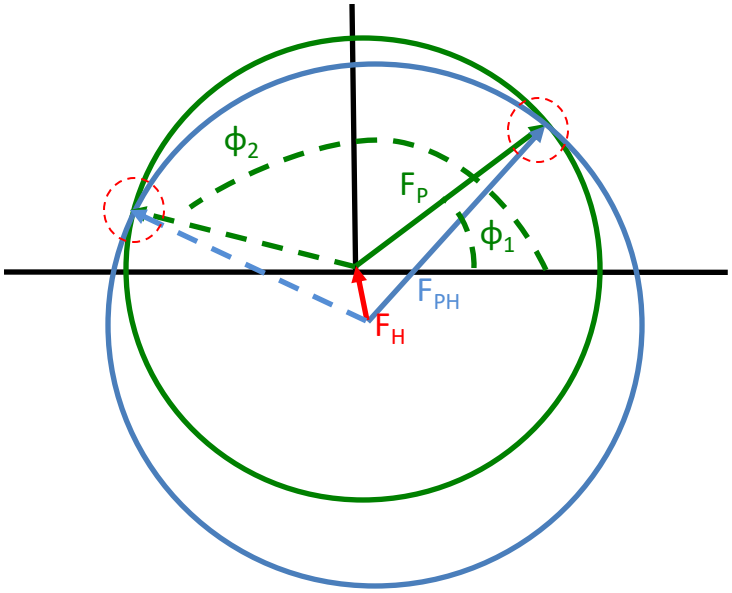




We collect Native (F_P) data

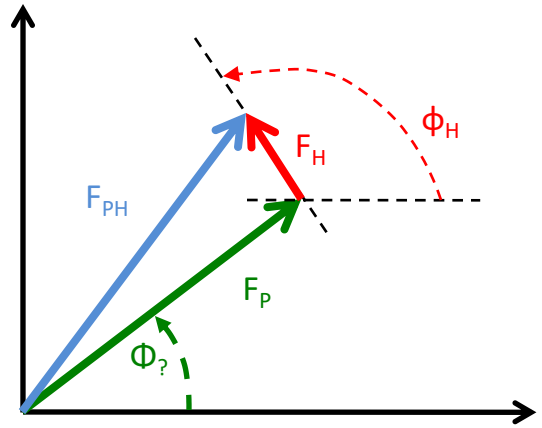
The phase is unknown

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

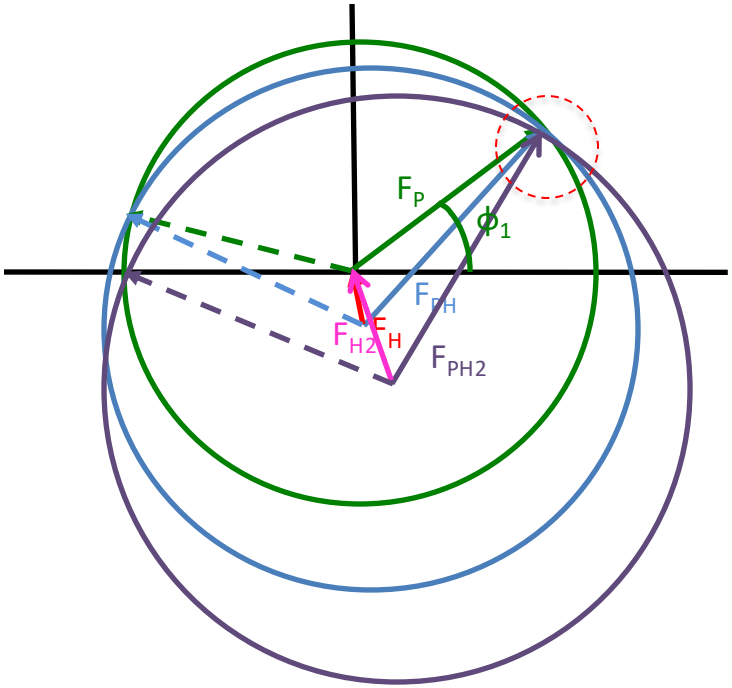
B: Take a guess – one of them has to be right

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

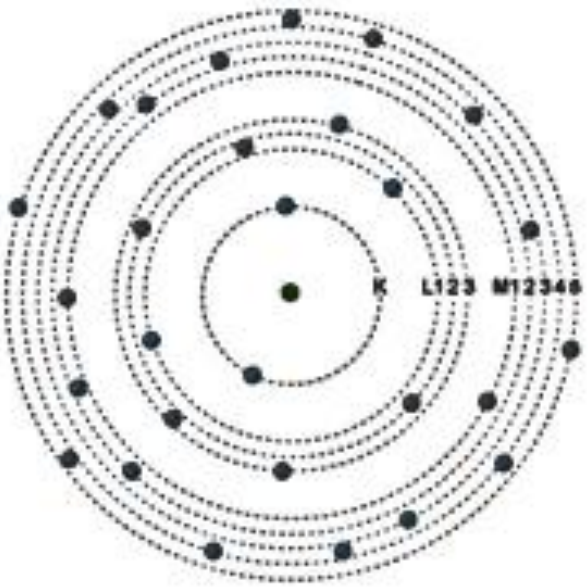
"**Anomalous** scattering" is **absolutely normal** while "**norma**l scattering" occurs only as an ideal, over simplified model, which can be used as a first approximation when studying scattering problems"

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

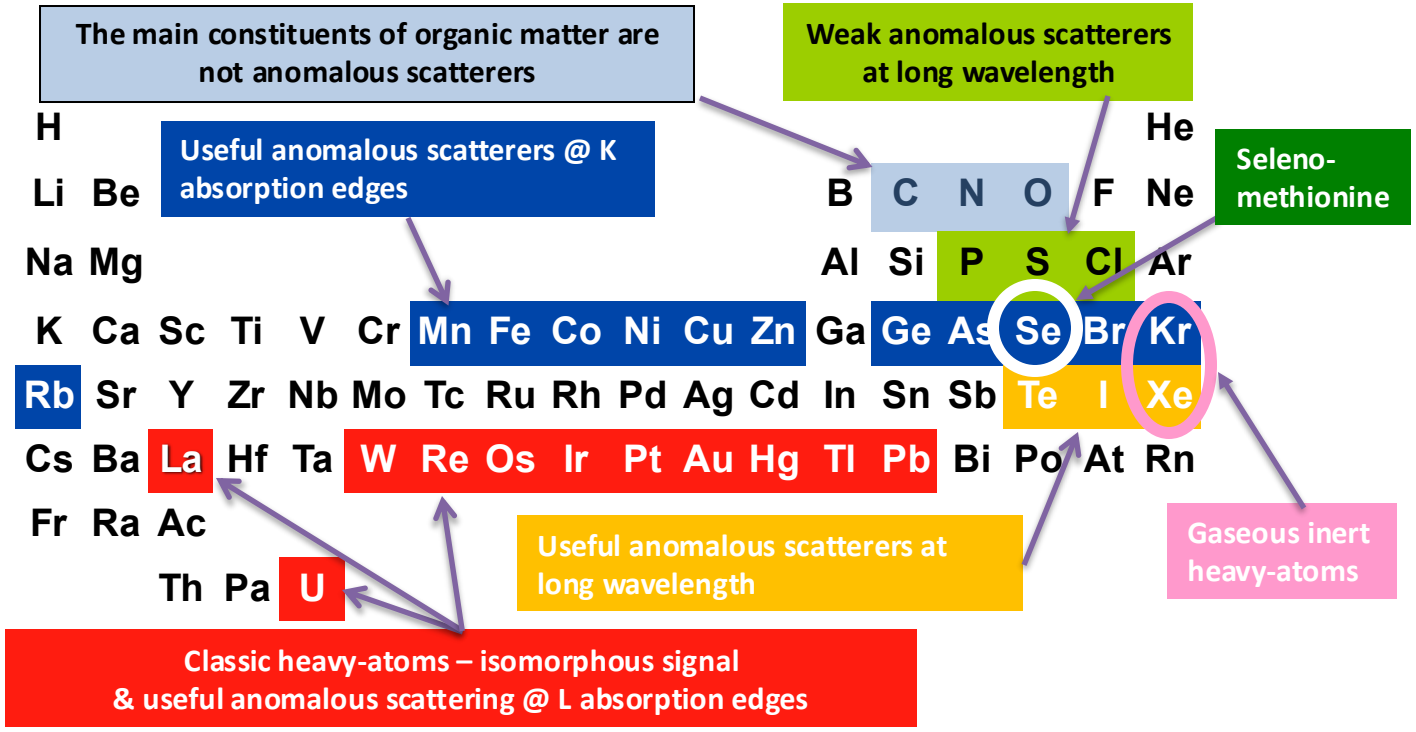
S. Caticha-Ellis (1998)

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



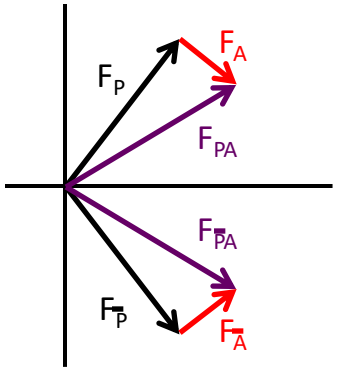
Significant Anomalous Scatterers



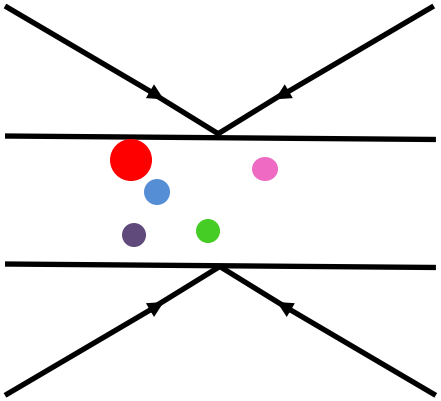
Friedel's Law in normal scattering conditions

Friedel pairs are Bragg reflections related by inversion through the origin

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

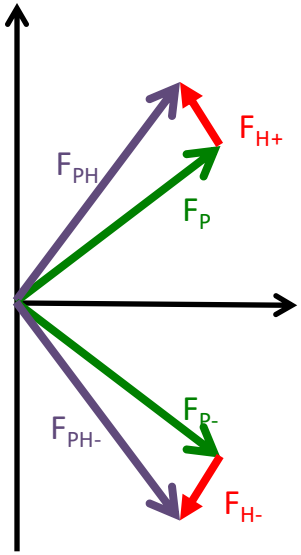


Normal scattering conditions

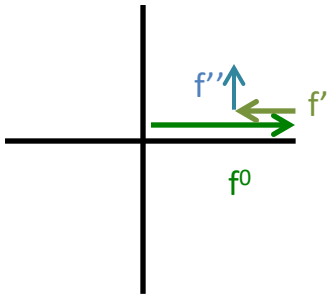


$$|F_{hkl}| = |F_{\overline{hkl}}| \qquad \varphi_{hkl} = -\varphi_{\overline{hkl}}$$

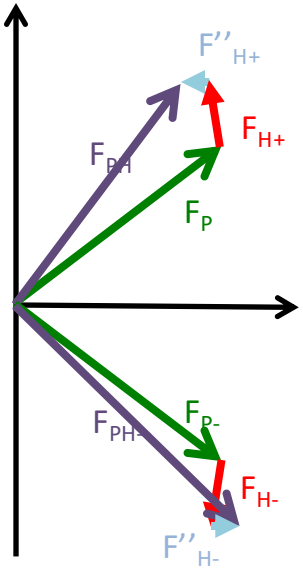
Breaking Friedel's Law



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.

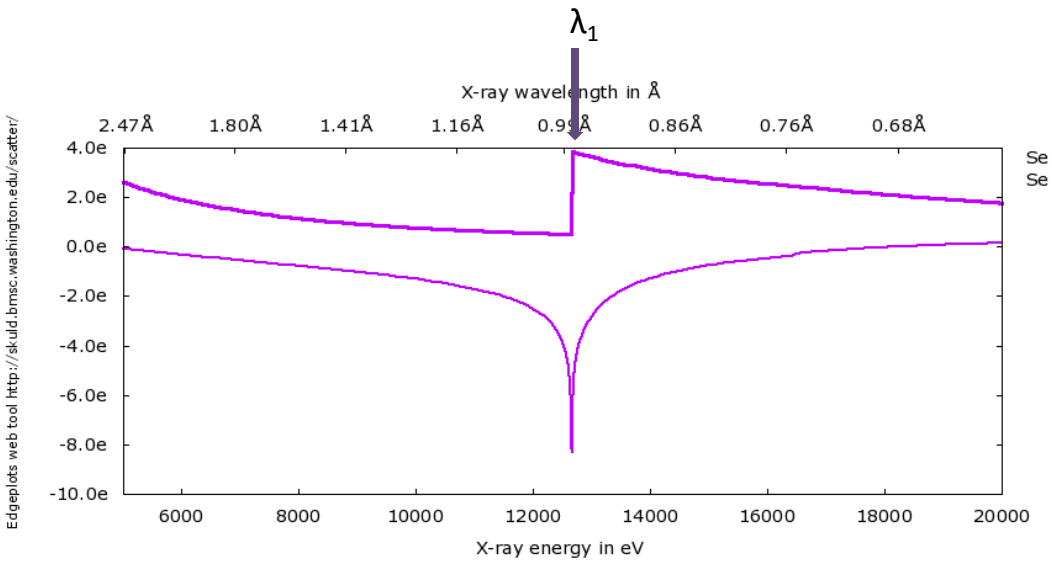
- An additional method of heavy atom incorporation can be used here by incorporating selenomethionine into the protein in place of methionine.

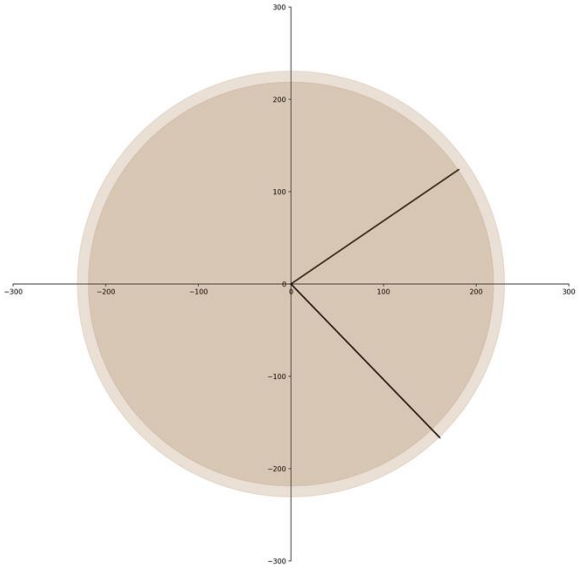
- We need to collect data at a synchrotron as we can select the wavelength and cause our substructure atoms to scatter anomalously.

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.

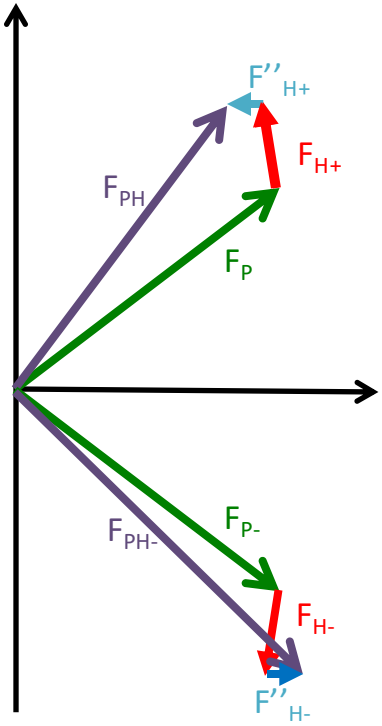
Solving the phases using Single Anomalous Diffraction (SAD)

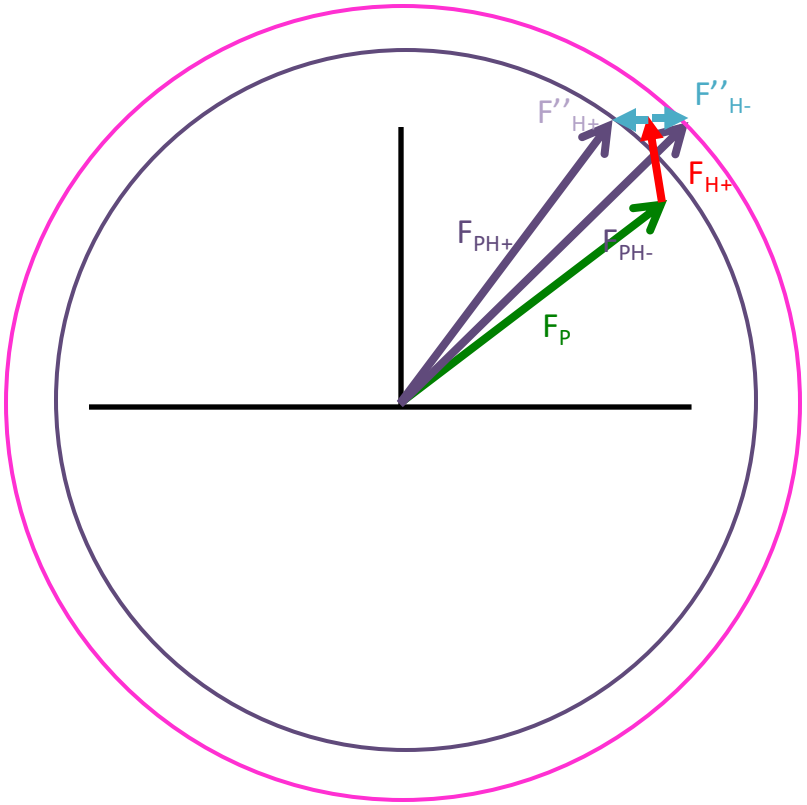




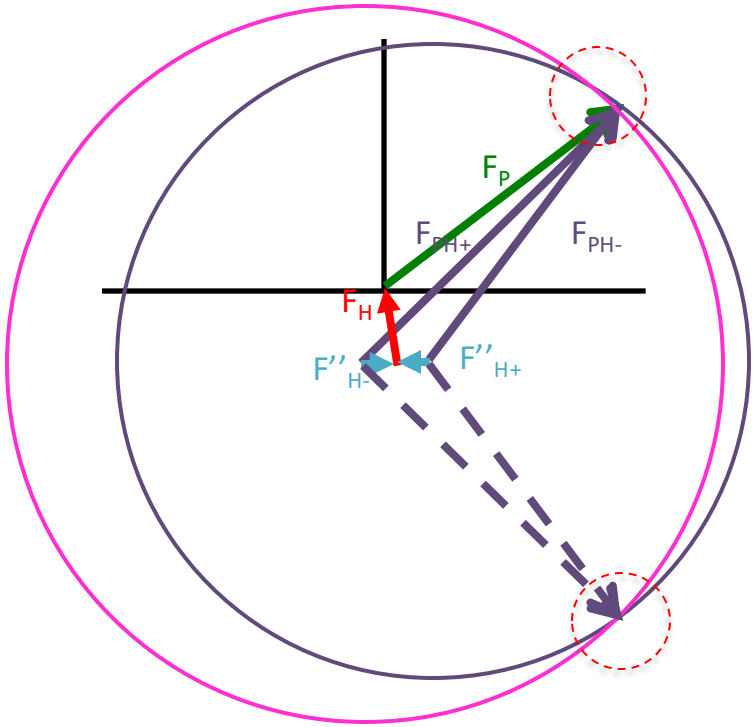
λ_1 = peak = maximum anomalous (f'')

Solving the phases using Single Anomalous Diffraction (SAD)



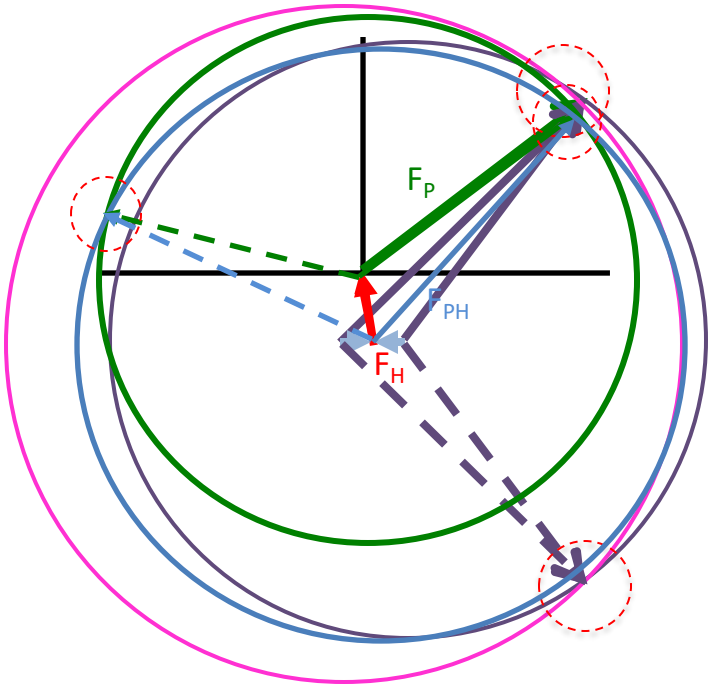


Solving the phases using Single Anomalous Diffraction



- 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.

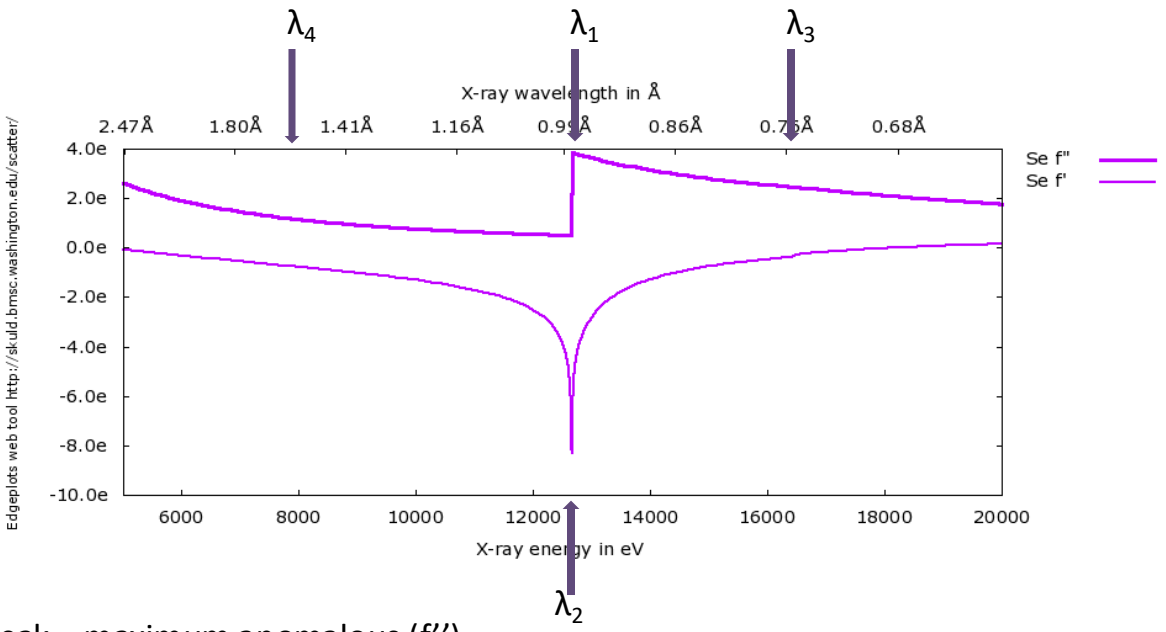
Solving the phases using Single Anomalous Diffraction



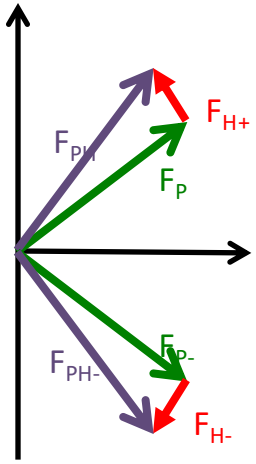
However, we can also combine the phasing experiment we did for the isomorphous with the SAD phasing experiment.

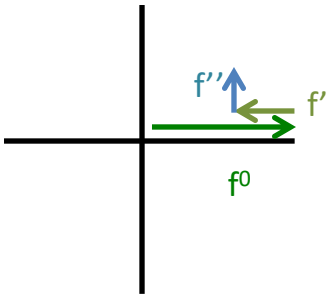
This is called SIRAS – Single isomorphous replacement anomalous dispersion.

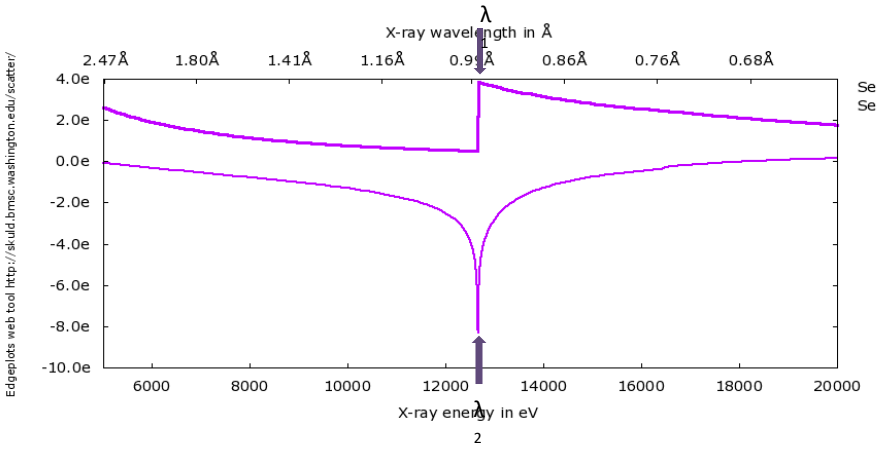
Solving the phases using Multiwavelength Anomalous Diffraction

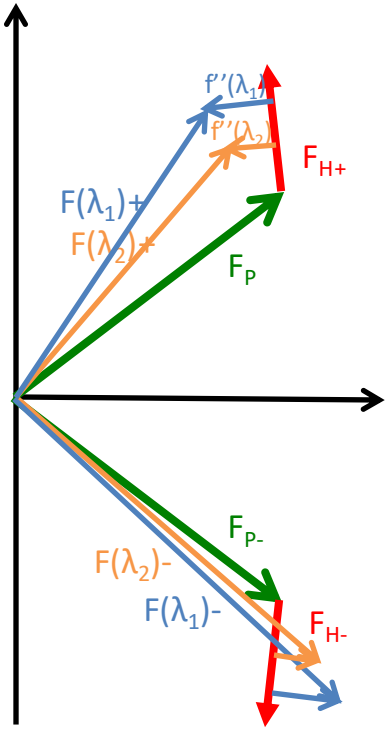


λ_1 = peak = maximum anomalous (f'')
λ_2 = inflection = maximum f'
λ_3 = high energy remote
λ_4 = low energy remote

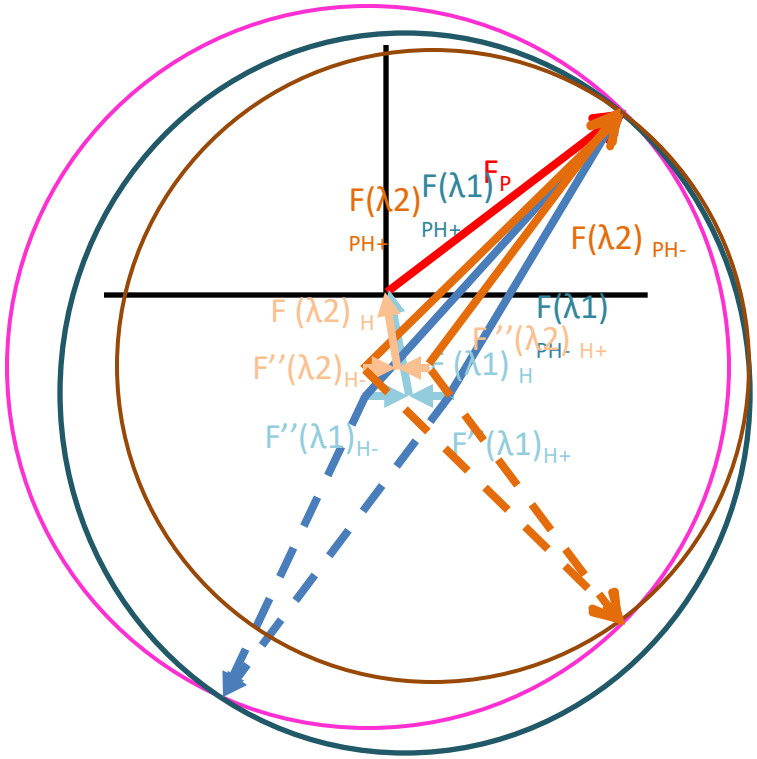

F_{PH+}
F_{H+}
F_P
F_{P-}
F_{PH-}
F_{H-}


f''
f'
f^0

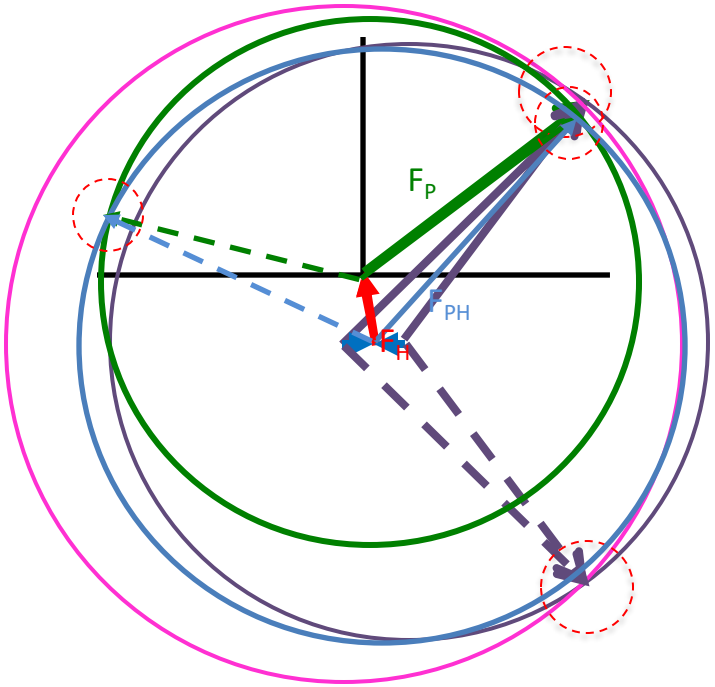

λ
X-ray wavelength in Å
2.47Å
1.80Å
1.41Å
1.16Å
0.91Å
0.86Å
0.76Å
0.68Å
4.0e
2.0e
0.0e
-2.0e
-4.0e
-6.0e
-8.0e
-10.0e
6000
8000
10000
12000
14000
16000
18000
20000
X-ray energy in eV
Se f''
Se f'
Edgeplots web tool Http://skuld.bmsc.washington.edu/scatter/


$f''(\lambda_1)$
$f''(\lambda_2)$
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.

Questions?