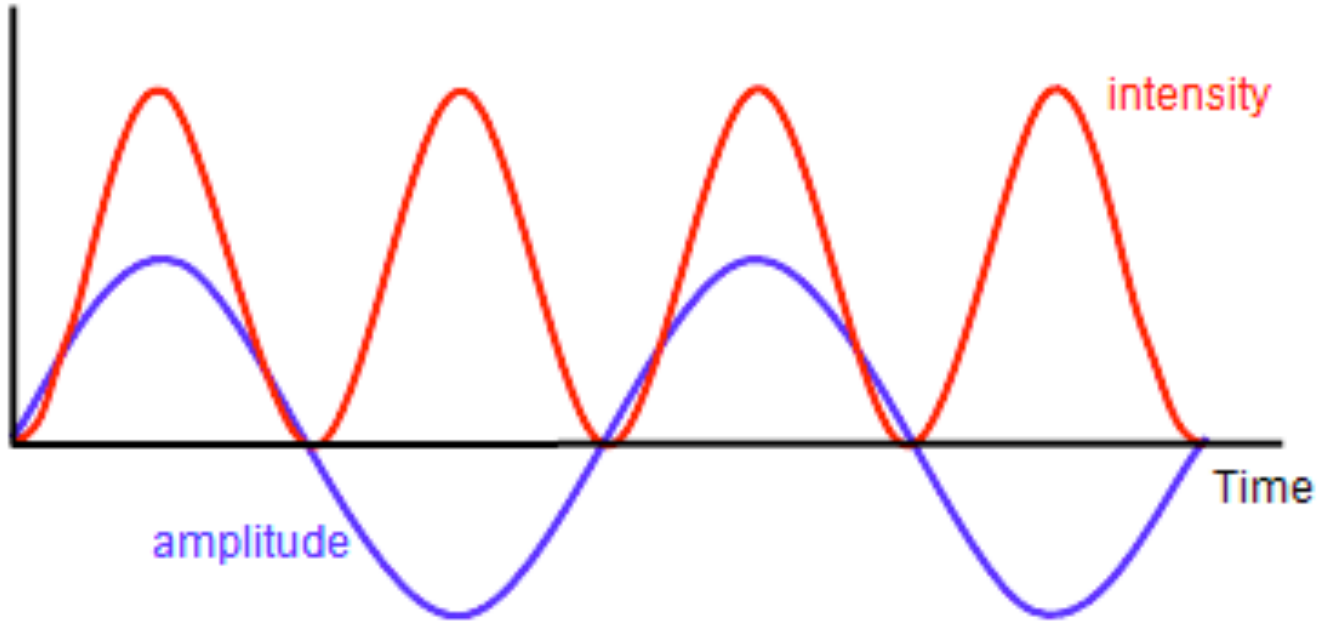
Experimental Phasing

Airlie McCoy

UNIVERSITY OF CAMBRIDGE

Interference

- X-ray detectors only detect X-ray intensities

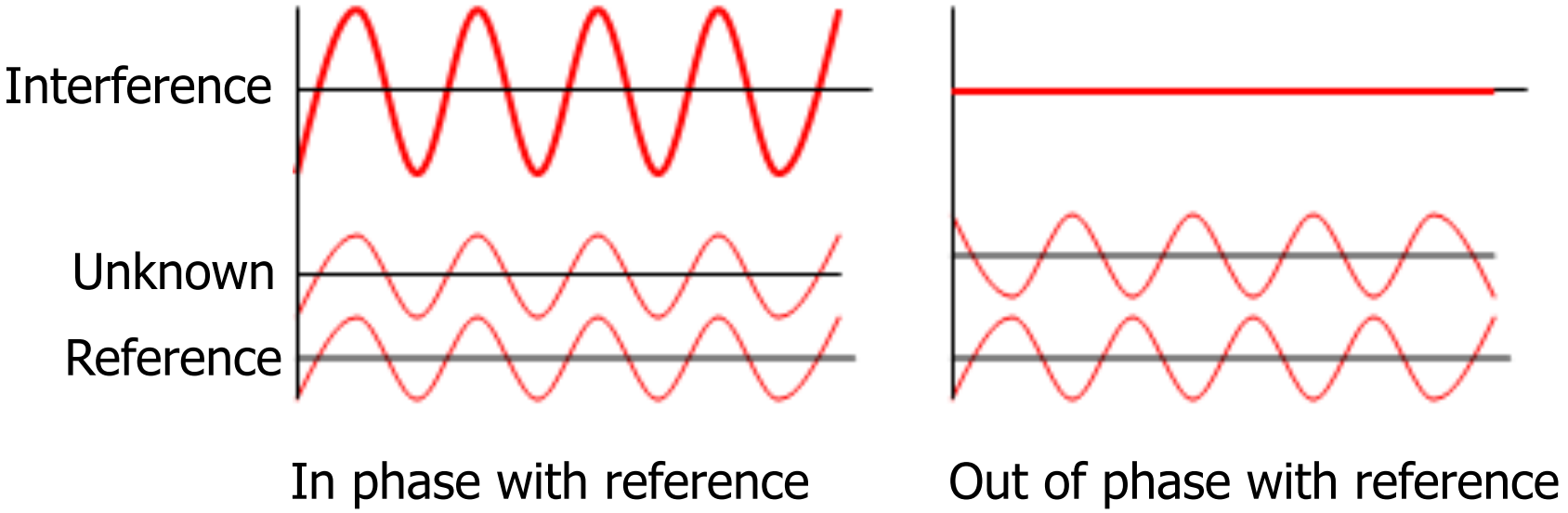


$$E \propto A^2$$

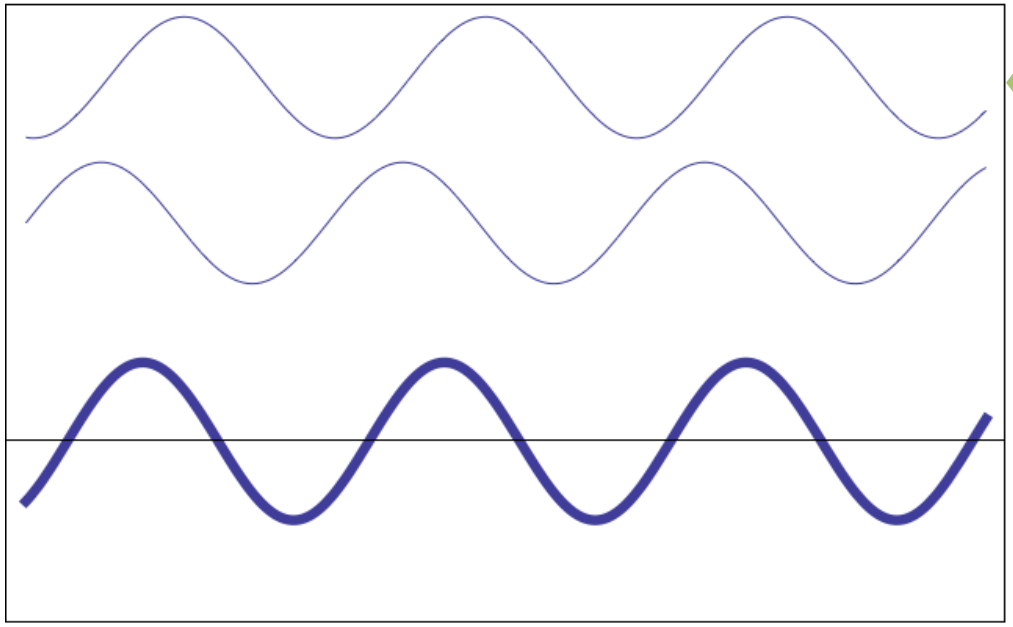
because they respond to the energy of the wave

Reference Wave

- To phase a wave of known amplitude by interference, look at the amplitude of the wave after interference with a reference wave of known phase and amplitude



Interference and Phase



The phase and amplitude of the reference wave are known from ***calculating*** *them from a* ***substructure*** *of atoms*

Substructure determination

- How do you determine the substructure?
- If you had the intensities of the substructure atoms floating in space you could solve the substructure
  - Patterson methods
  - Direct methods
  - Dual space methods

Problem

- You don't have these intensities!

Solution

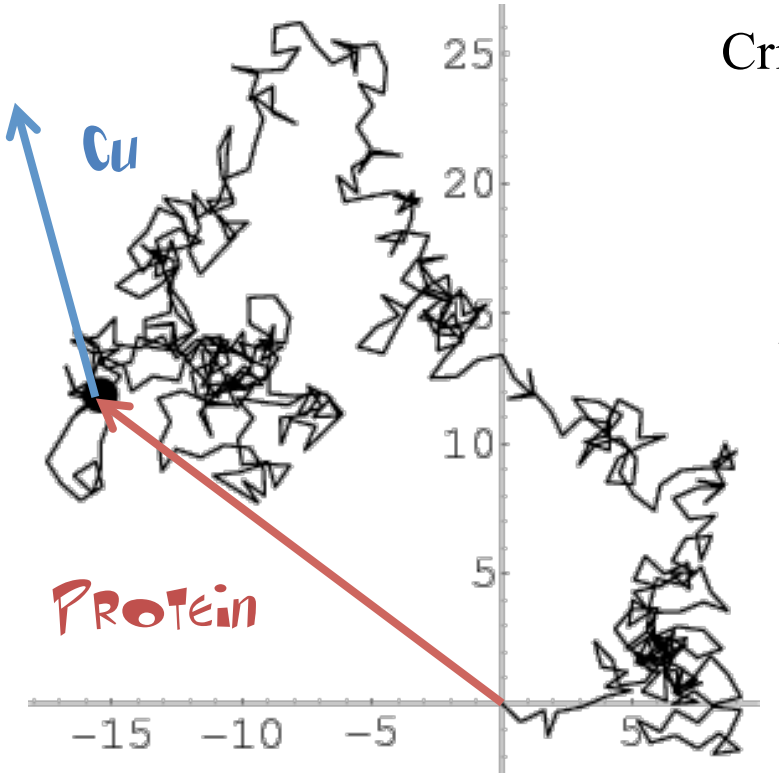
- Use approximations (e.g. Blundell & Johnson 1976)

Types of Interference Experiments

- Isomorphous replacement
- Anomalous scattering

Heavy Atom Scattering is Significant

- How can e.g. a single Cu atom in a 100 kDa protein make any difference to the intensities?
- Structure factors add up as a "random walk"

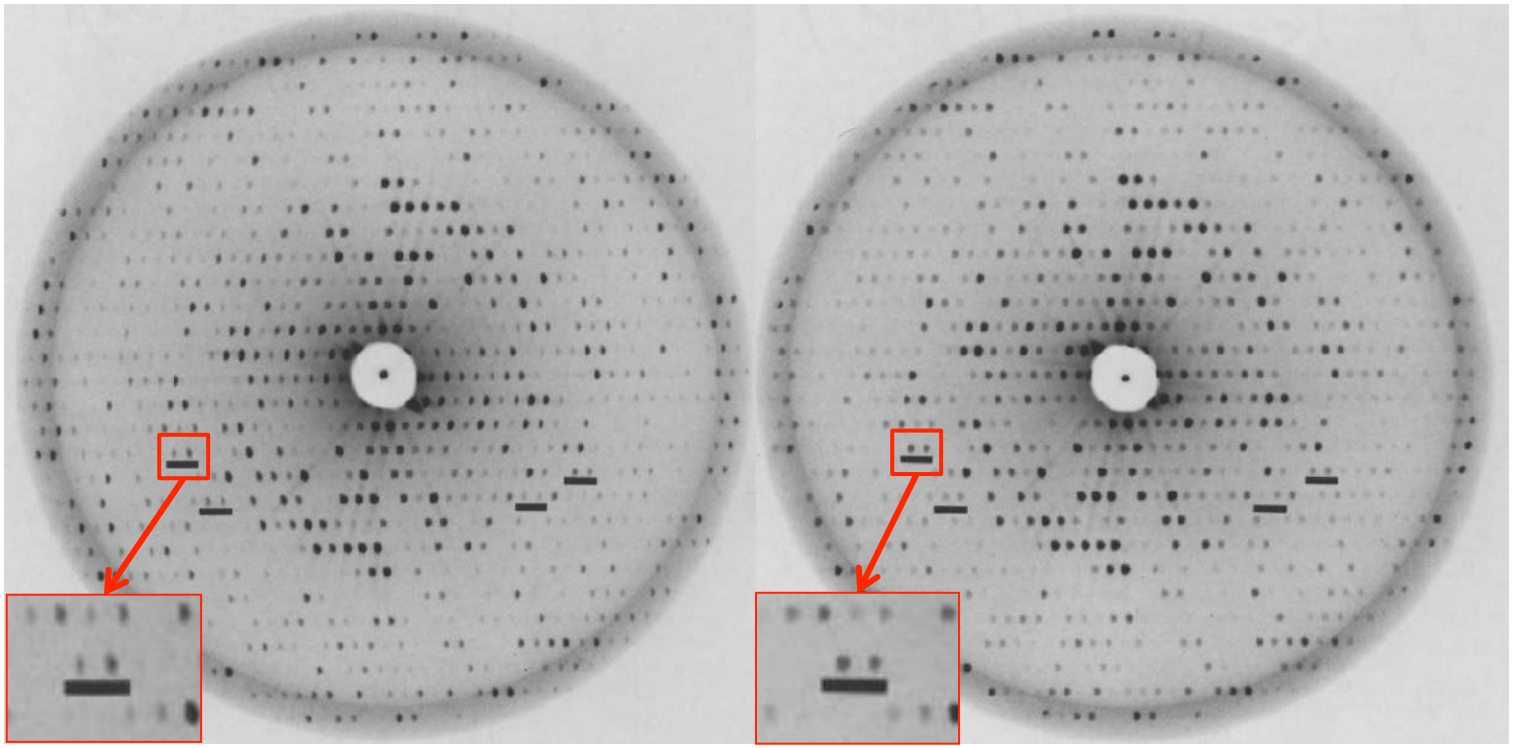


Crick and Magdoff (1956)

$$\frac{\langle \Delta F \rangle}{\langle |F| \rangle} = \frac{Z_H}{Z_{eff}} \sqrt{\frac{2N_H}{N_p}}$$

N_h, N_p are number of heavy, protein atoms
Z_h, Z_{eff} are atomic numbers of heavy atom and average of protein atoms (~6.7)

100 kDa protein with Copper (Z=28)
5.6%



Native

Heavy atom derivative

Intensity reversals (underlined)
indicate heavy atom has bound

Isomorphous Replacement

The clue is in the name…

- Native and derivative must be isomorphous
  - Same unit cell and space group
  - Same position and orientation of protein in unit cell
- Can require searching many different compounds to find one or two isomorphous ones

Anomalous Scattering

A·nom·a·lous

adj.

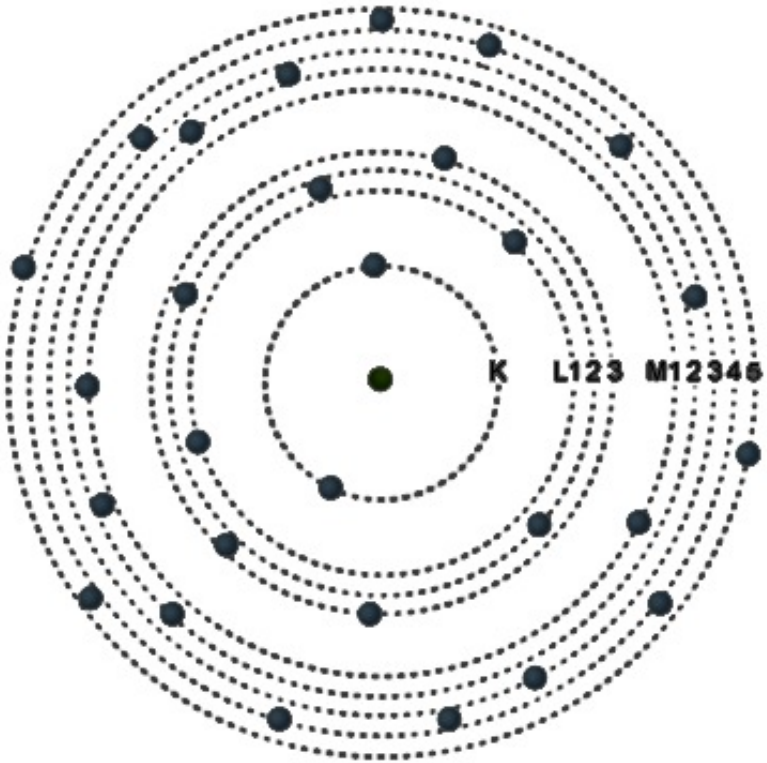
Deviating from the normal or common order, form or rule

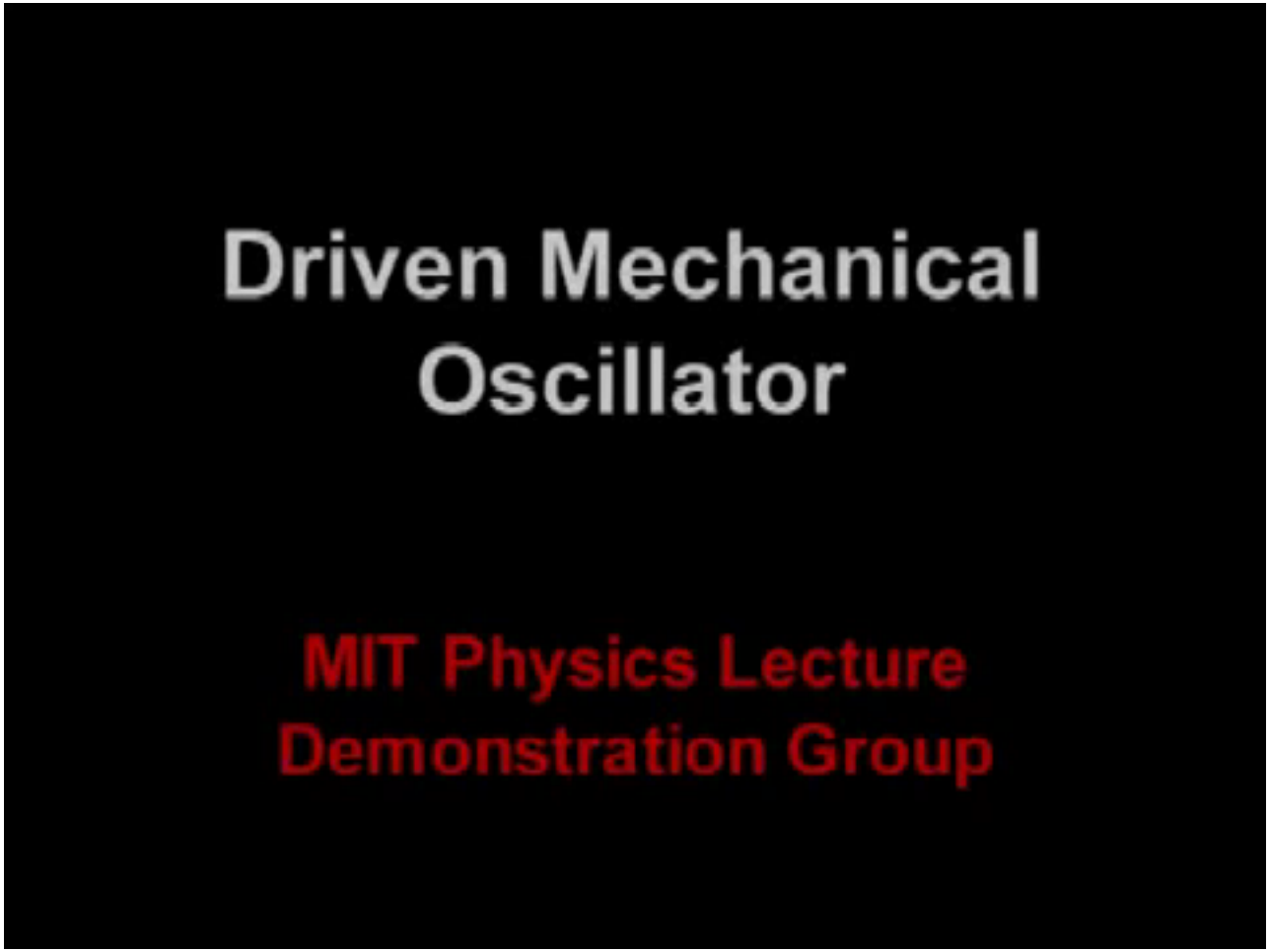
"**Anomalous** scattering" is **absolutely normal** while "**norma**l scattering" occurs only as an ideal, over simplified model, which can be used as a first approximation when studying scattering problems"

IUCR Pamphlet "Anomalous Dispersion of X-rays in Crystallography"
S. Caticha-Ellis (1998)

Anomalous Scattering

- 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





https://www.youtube.com/watch?v=aZNnwQ8HJHU

$$f = \frac{\omega^2}{\omega^2 - \omega_s^2 - ik\omega} = f^0 + \Delta f' + i\Delta f''$$

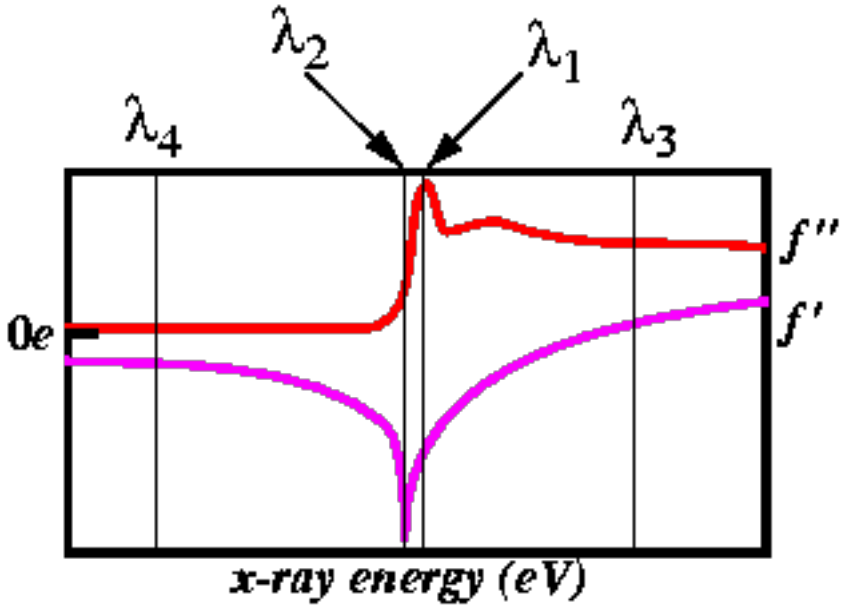
Scattering factor

Frequency of incident wave

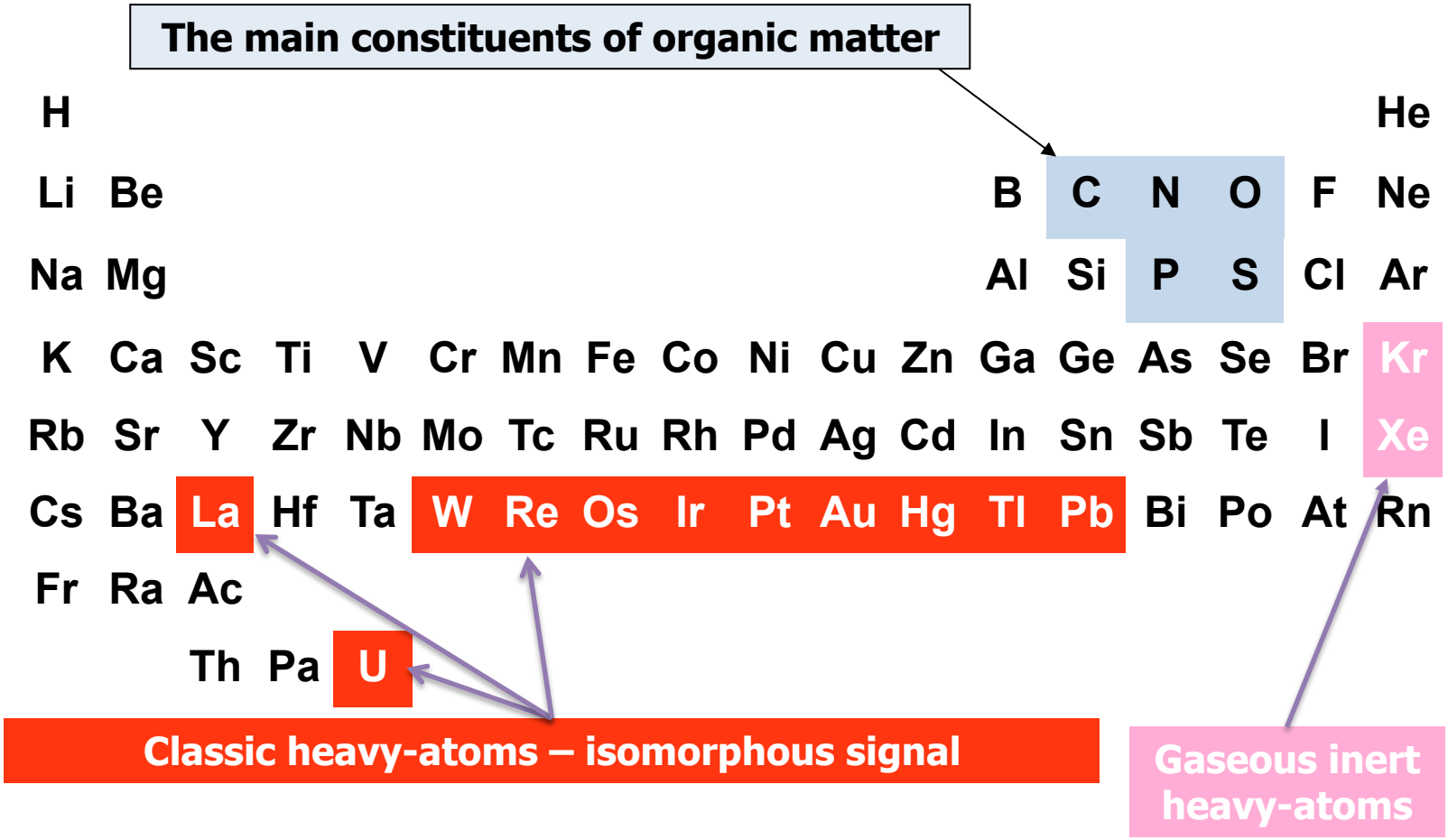
Resonant frequency

Damping factor

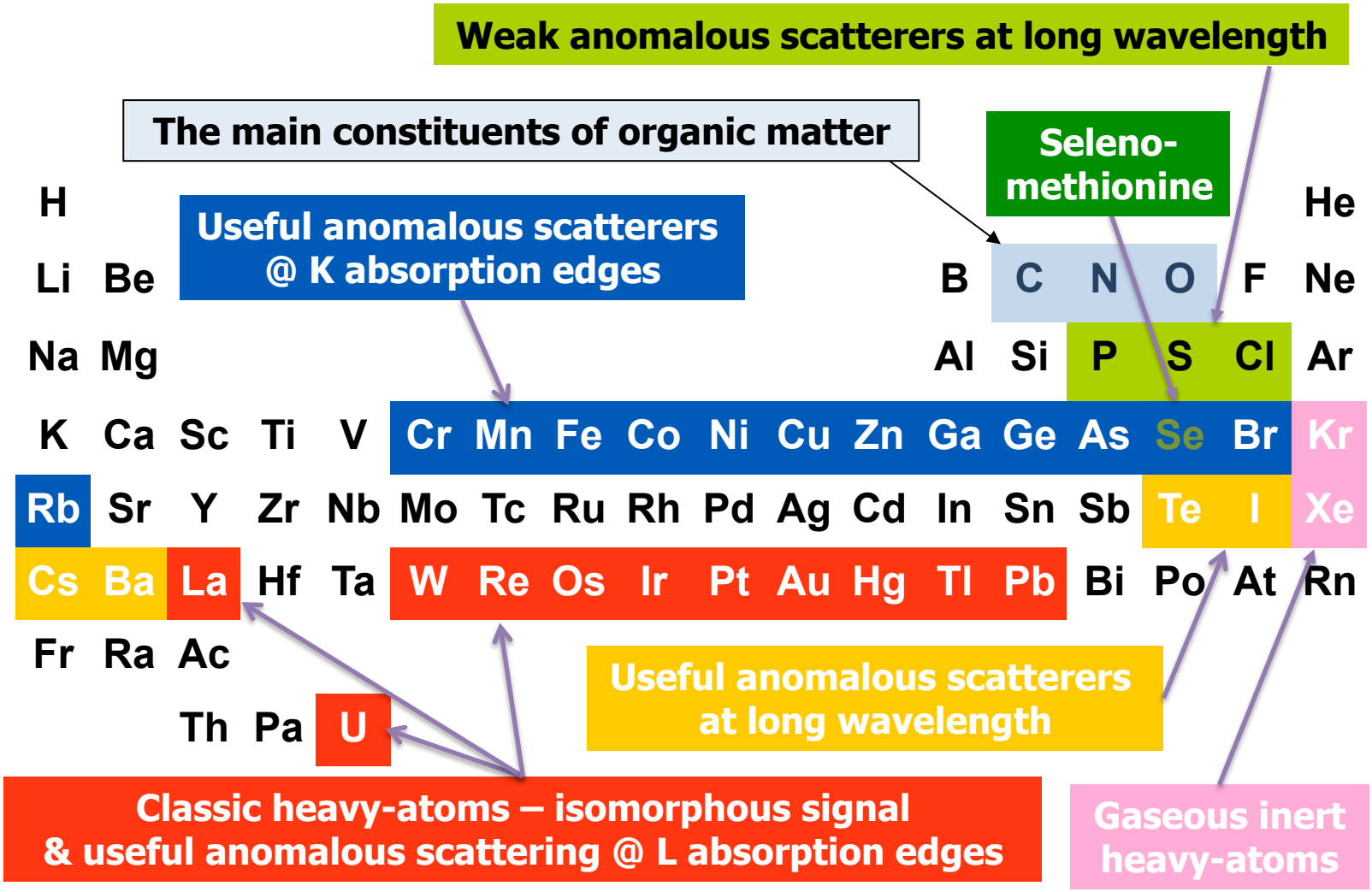
- Both $\Delta f'$ and $\Delta f''$ change with the frequency of the incident radiation
  - Large when the incident frequency is near resonant frequency
- Dispersive component
  $\Delta f'$ is in phase with f^0
- Anomalous component
  $\Delta f''$ is advanced by 90° by with respect to f^0



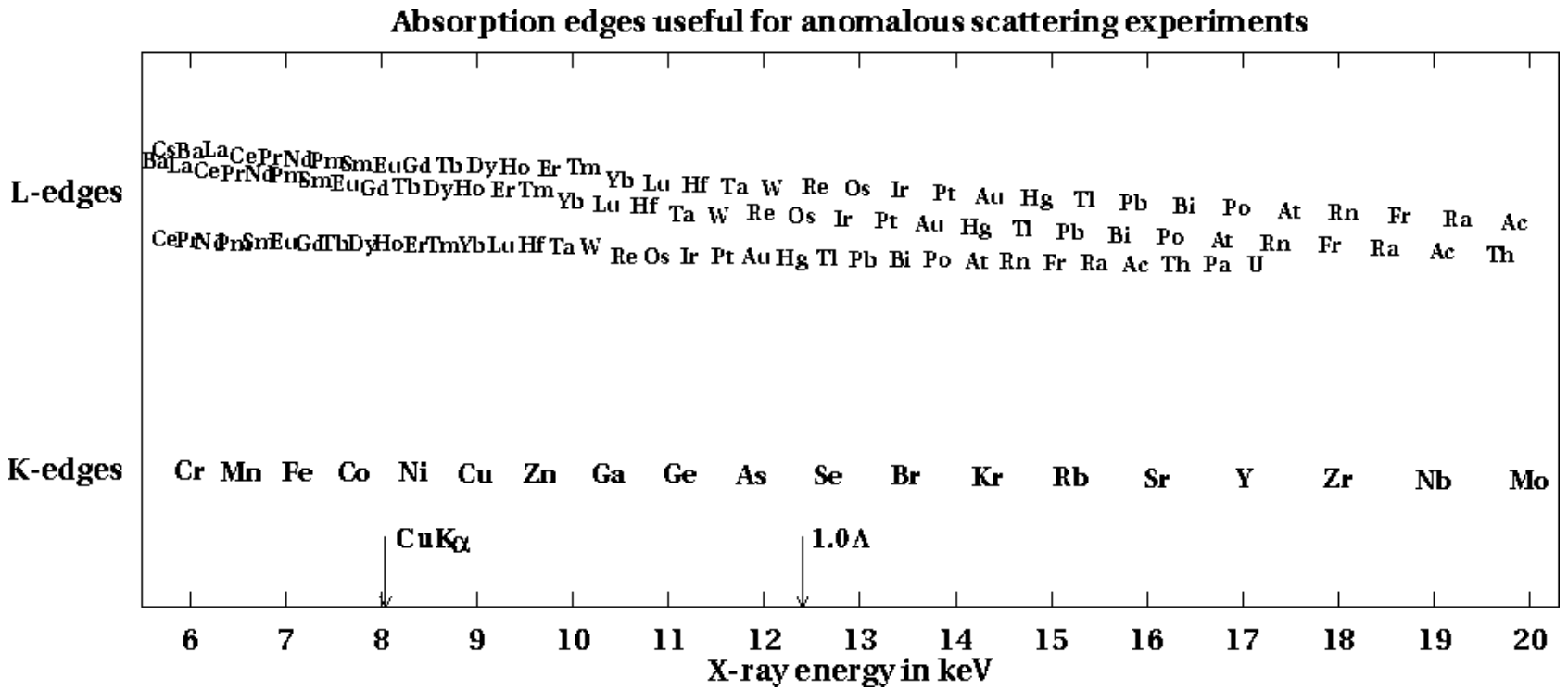
Isomorphous replacement



Anomalous Scatterering



Absorption edges




MIR
SAD
SIR
SIRAS
MIRAS
MAD
RIP

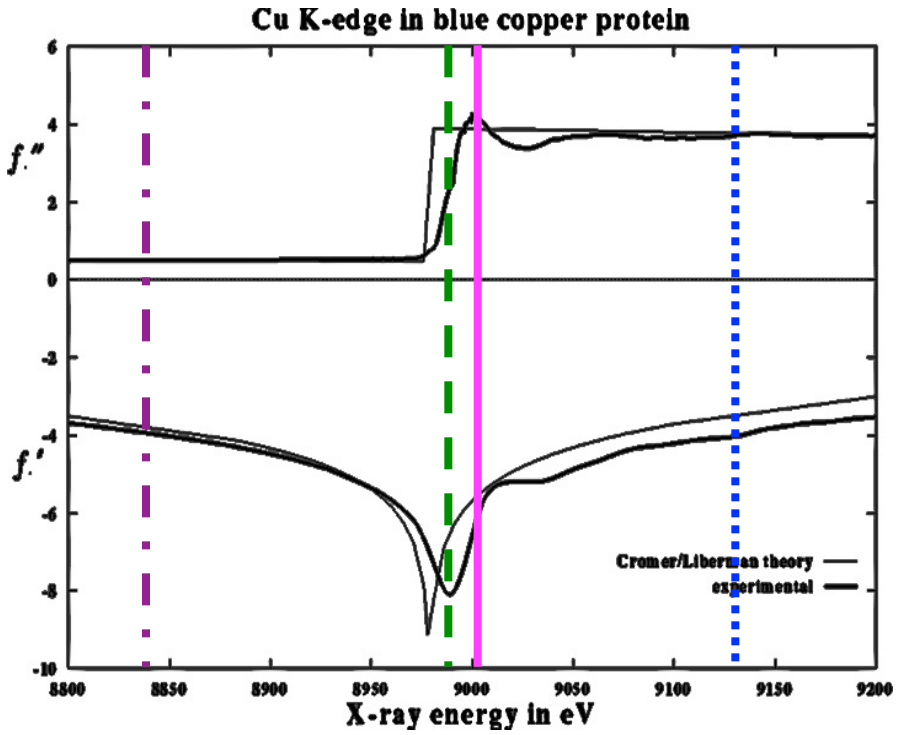
SAD

- The most popular way of solving structures by experimental phasing (over 70% and rising)
- Can be done with intrinsic S and CuKα X-rays
- SAD phasing theory is very good
- Easy to automate
- Can be very fast
  - Can be done from single dataset
- May need multiple crystals
  - And careful data processing

MAD

- Usually performed using seleno-methionines
- Can use intrinsically bound metal ions
  - e.g. Fe, Ni, Cu, Zn
- Or introduced heavy atoms
  - e.g. U, Pt, Au, Hg, Pb
- Requires a synchrotron
  - Tune wavelength to optimise f’ and f” values
- Theoretical problems in the phasing algorithms make MAD data not as useful as they could be

Wavelength Choice



PEAK: |f”| is large

INFLECTION: |f’| is large

REMOTE: low energy

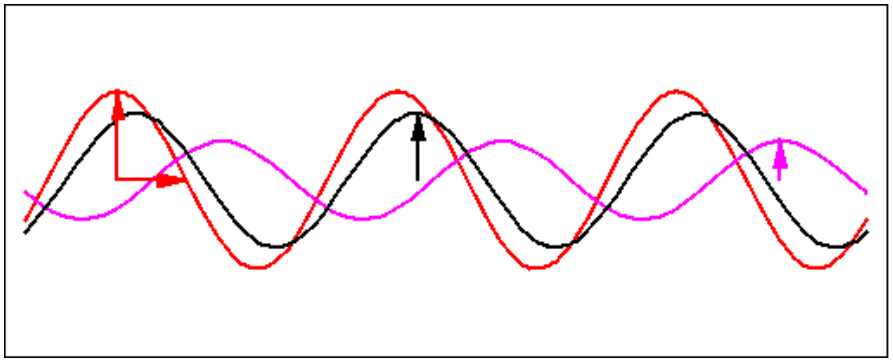
REMOTE: high energy

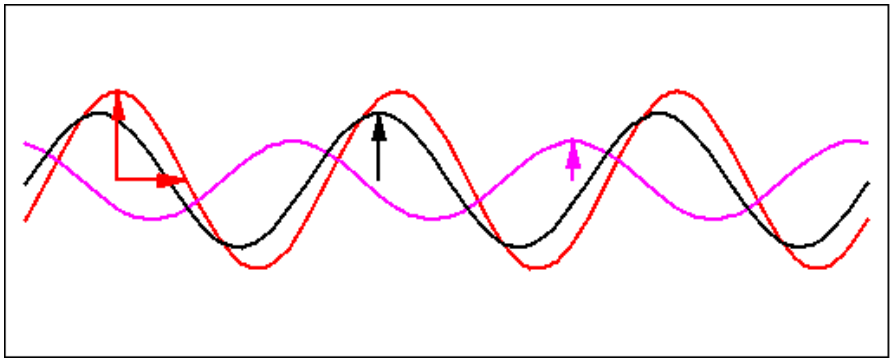
Warning
Radiation
Damage
INCREASES
at peak

Harker diagrams

Reference Wave

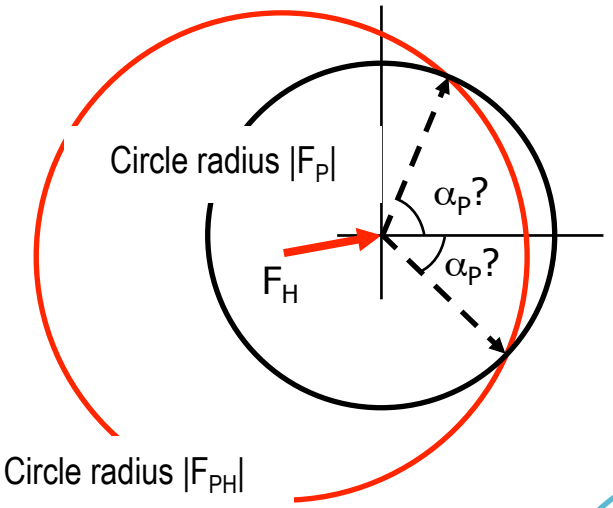
- Reference wave (red)
  - Amplitude and phase known
- Interference wave (black)
  - Amplitude known
- Unknown wave (magenta)
  - Amplitude known

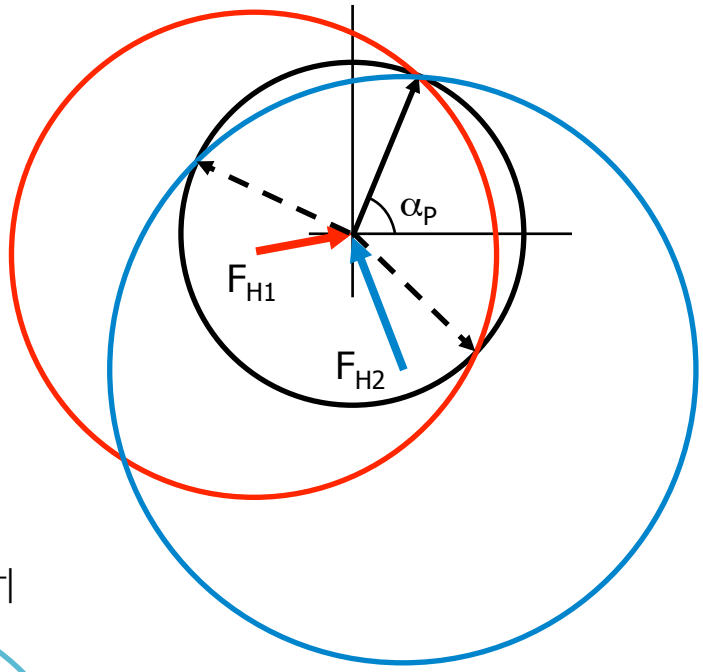


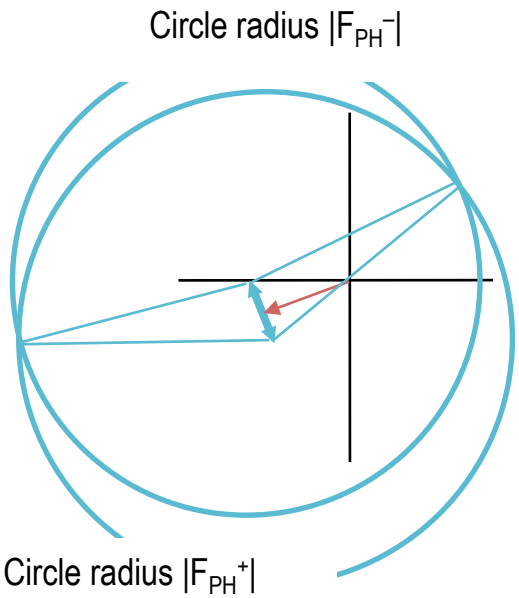


- Two solutions for phase of unknown wave
- *Also two solutions for phase of interference wave*

Harker Diagrams

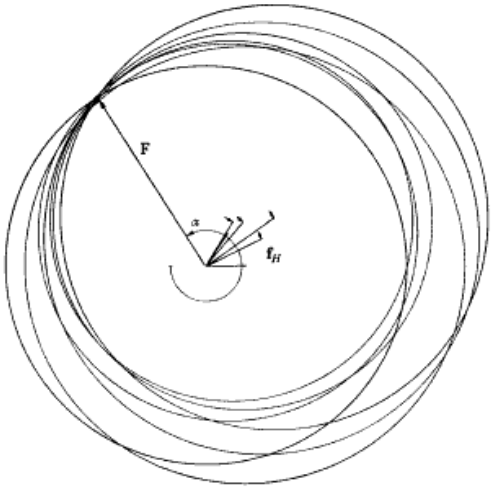


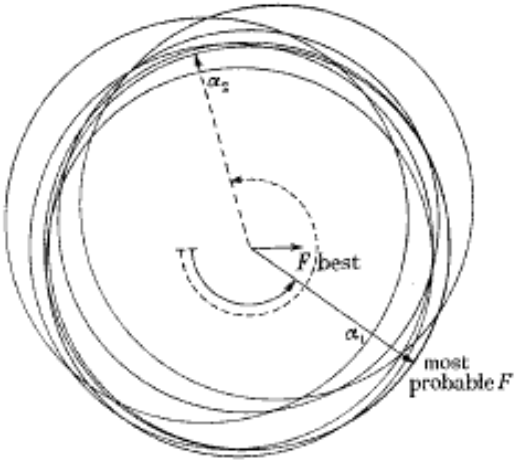


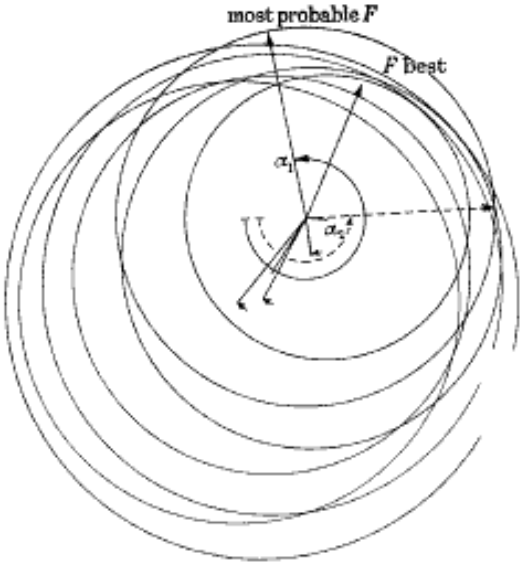


Reality…

- Some real Harker diagrams from the phasing of haemoglobin with 6 derivatives



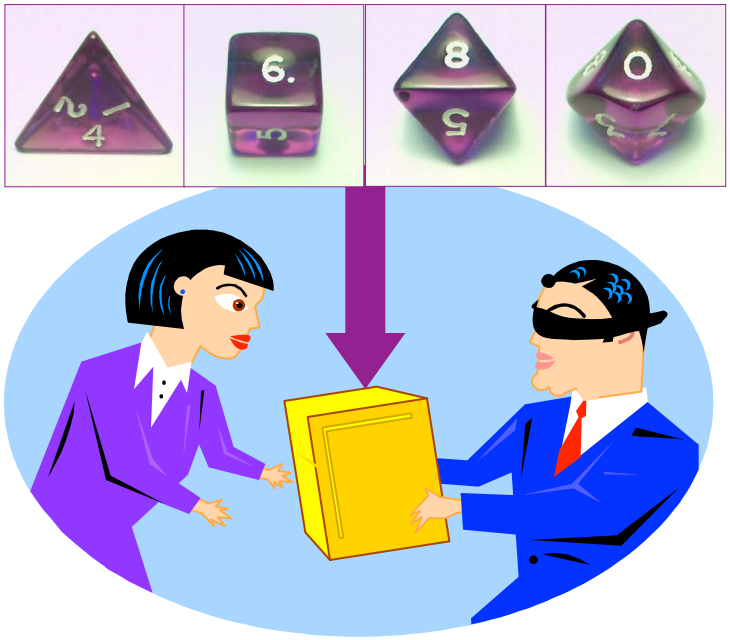


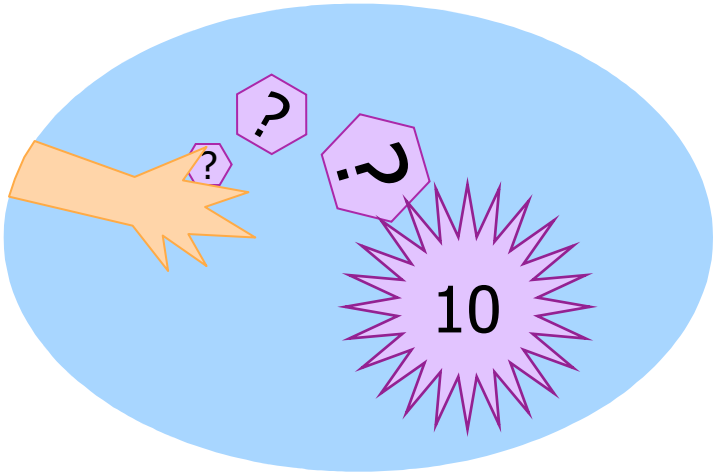


- Phase circles rarely cross exactly
- Need a **probabilistic** approach to determining the phase

Maximum Likelihood

A game of dice

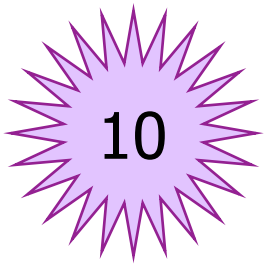




- Put four unbiased dice in a box
- I select a die at random
- I roll the die and tell you the result of the roll
- Which die did I most likely select?

Roll a 10

- The die obviously must have been the 10 sided die
- What does "must" mean in probabilities?



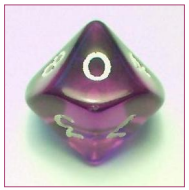
$$P(10;\boxed{10}) = \tfrac{1}{10}$$

$$P(10;\boxed{8}) = 0$$

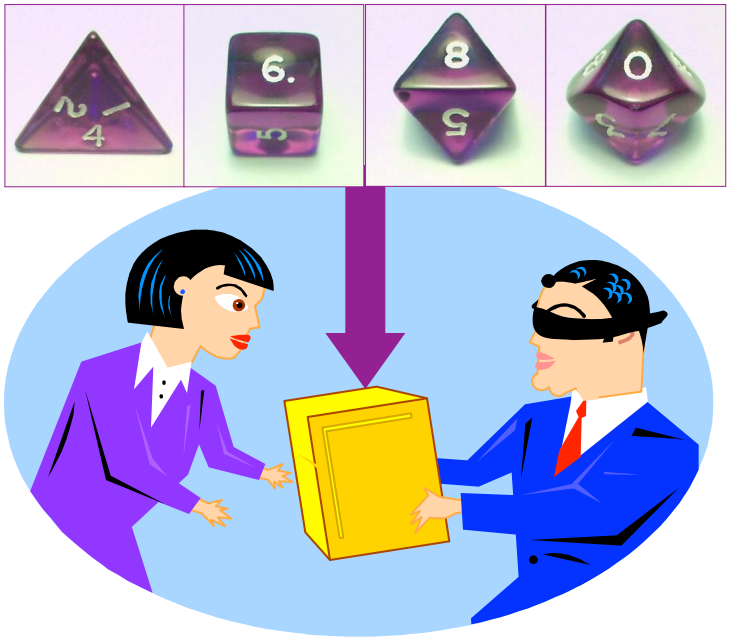
$$P(10;\boxed{6}) = 0$$

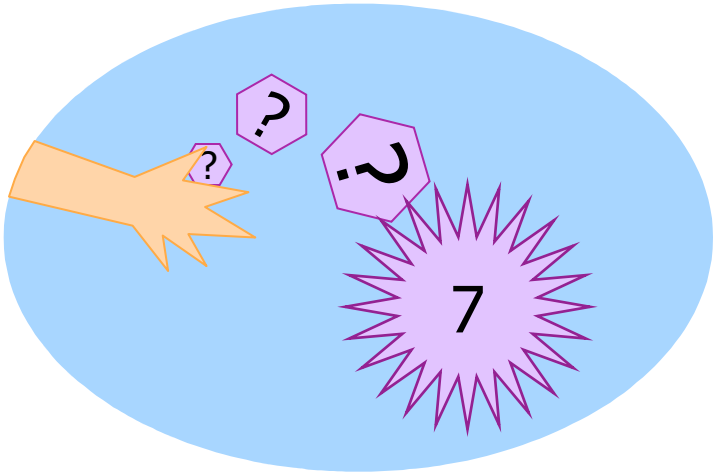
$$P(10;\boxed{4}) = 0$$

most likely



A game of dice with data

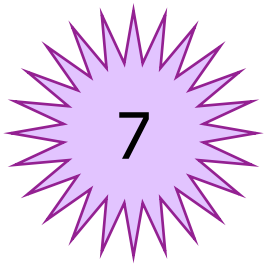




- Put four unbiased dice in a box
- I select a die at random
- I roll the die and tell you the result of the roll
- Which die did I most likely select?

Roll a 7

- The die could have been the 10 sided or the 8 sided die
- Which die is most likely?



$$P(7;\boxed{10}) = \frac{1}{10}$$

$$P(7;\boxed{8}) = \frac{1}{8}$$

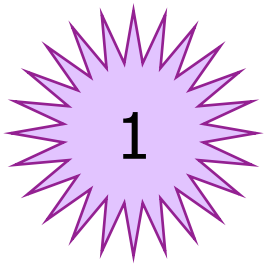
$$P(7;\boxed{6}) = 0$$

$$P(7;\boxed{4}) = 0$$

most likely

Roll a 1

- Could have been rolled by any of the dice
- The most likely die is the one with the highest probability of generating the data



$$P(1;\boxed{10}) = \frac{1}{10}$$

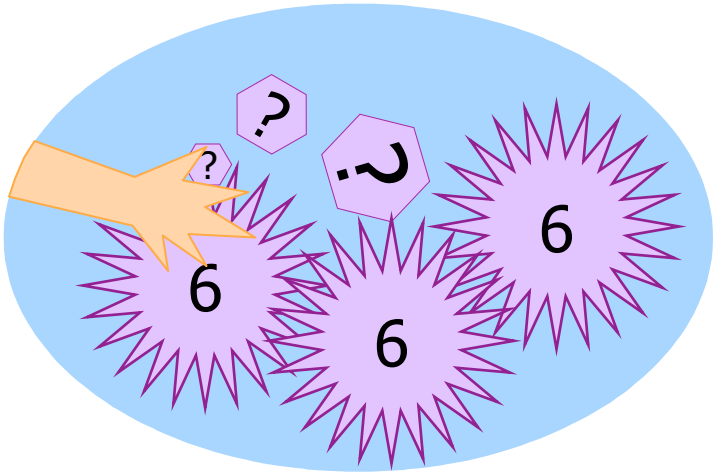
$$P(1;\boxed{8}) = \frac{1}{8}$$

$$P(1;\boxed{6}) = \frac{1}{6}$$

$$P(1;\boxed{4}) = \frac{1}{4}$$

most likely

A game of dice with more data

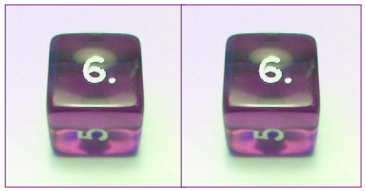


- Put four unbiased dice in a box
- I select a die at random
- I roll that die three times and tell you the results
- Which die did I most likely select?

Multiplying probabilities

- When probabilities are independent they multiply

$$P(6;\boxed{6}) = \tfrac{1}{6} = 0.16666667$$



$$P(6,6;\boxed{6}) = \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{36} = 0.0277778$$

$$P(6,6,6;\boxed{6}) = \tfrac{1}{6} \times \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{216} = 0.0046296$$

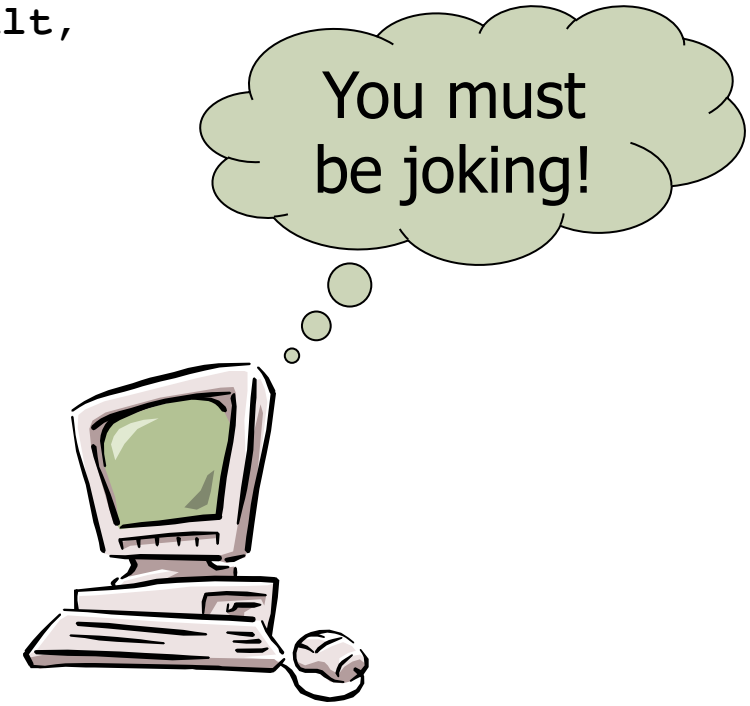
100 times

$$P(6\text{K} \times 100;\boxed{6}) = 6^{-100} = 1.53064 \times 10^{-78}$$

Computers and small numbers

"Oh great one, what is the probability of throwing a 6 from a six sided die one billion times?"

```
> SYSTEM-F-FLTOVF_F, arithmetic fault,
floating overflow at PC=00006244,
PSL=03C0 0020 %TRACE-F-TRACEBACK,
symbolic stack dump follows
module name     routine name    line
OVERF           OVERF           104
DPARA$MAIN      DPARA$MAIN      276
```



Computers can not store numbers very close to zero

Computers and log(small numbers)

"Oh great one, what is the logarithm of the probability of throwing a 6 from a six sided die one billion times?"

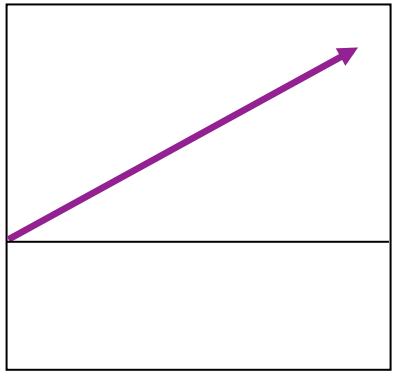
```
> -778151250.4
```

log(likelihood) is not close to zero

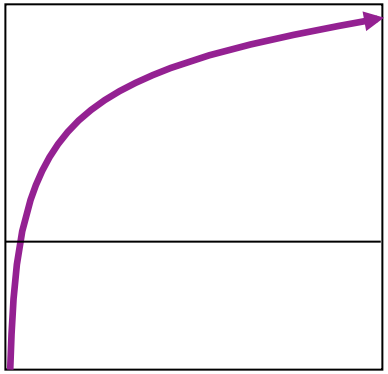
- So the log(likelihood) solves the small number problem
- But can we just switch to using the log(likelihood)?

Optimisation and logarithms

- Logarithmic functions are "monotonic" functions
  - *i.e.* they "preserve the given order"
  - If $y_1 < y_2$ for all $x_1 < x_2$ then $\log(x_1) < \log(x_2)$
- The parameter values obtained optimising log(likelihood) are the same as those obtained optimising likelihood
  - **Optimising log(likelihood) ≡ Optimising likelihood**



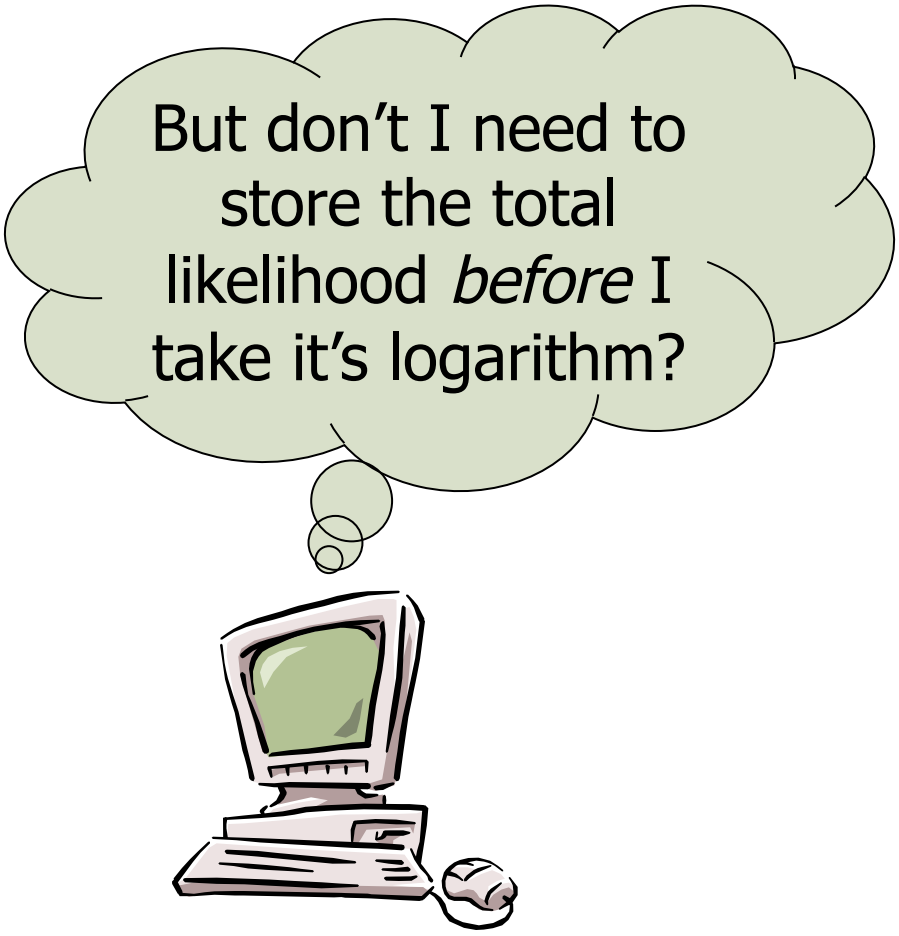
$y = x$



$y = \log(x)$

Logarithms, products and sums

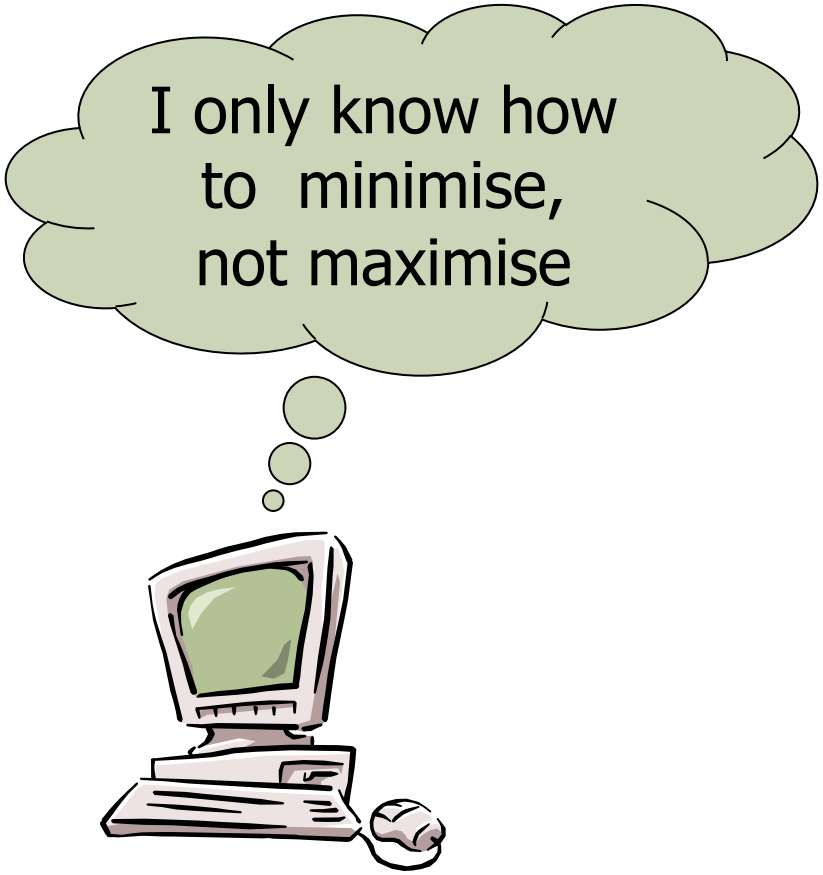
- No, there is a shortcut to the log(total likelihood) when total likelihood is a product of likelihoods
- If log(total likelihood) equals log(Π likelihoods)
  - product
- Then log(total likelihood) also equals Σlog(likelihoods)
  - sum



Minimising

- Computer algorithms are designed to minimise
- Therefore we optimise our parameters by minimising the

  -log(likelihood)



Logarithms and independence

$$\log(\prod \text{likelihoods}) = \sum \log(\text{likelihoods})$$

$$\log\left(P\left(3,3;\boxed{6}\right)\right)$$

$$= \log\left(P\left(3;\boxed{6}\right) \times P\left(3;\boxed{6}\right)\right)$$

$$= \log\left(\tfrac{1}{6} \times \tfrac{1}{6}\right)$$

$$= \log(0.0277)$$

$$= -1.556$$

$$\log\left(P\left(3,3;\boxed{6}\right)\right)$$

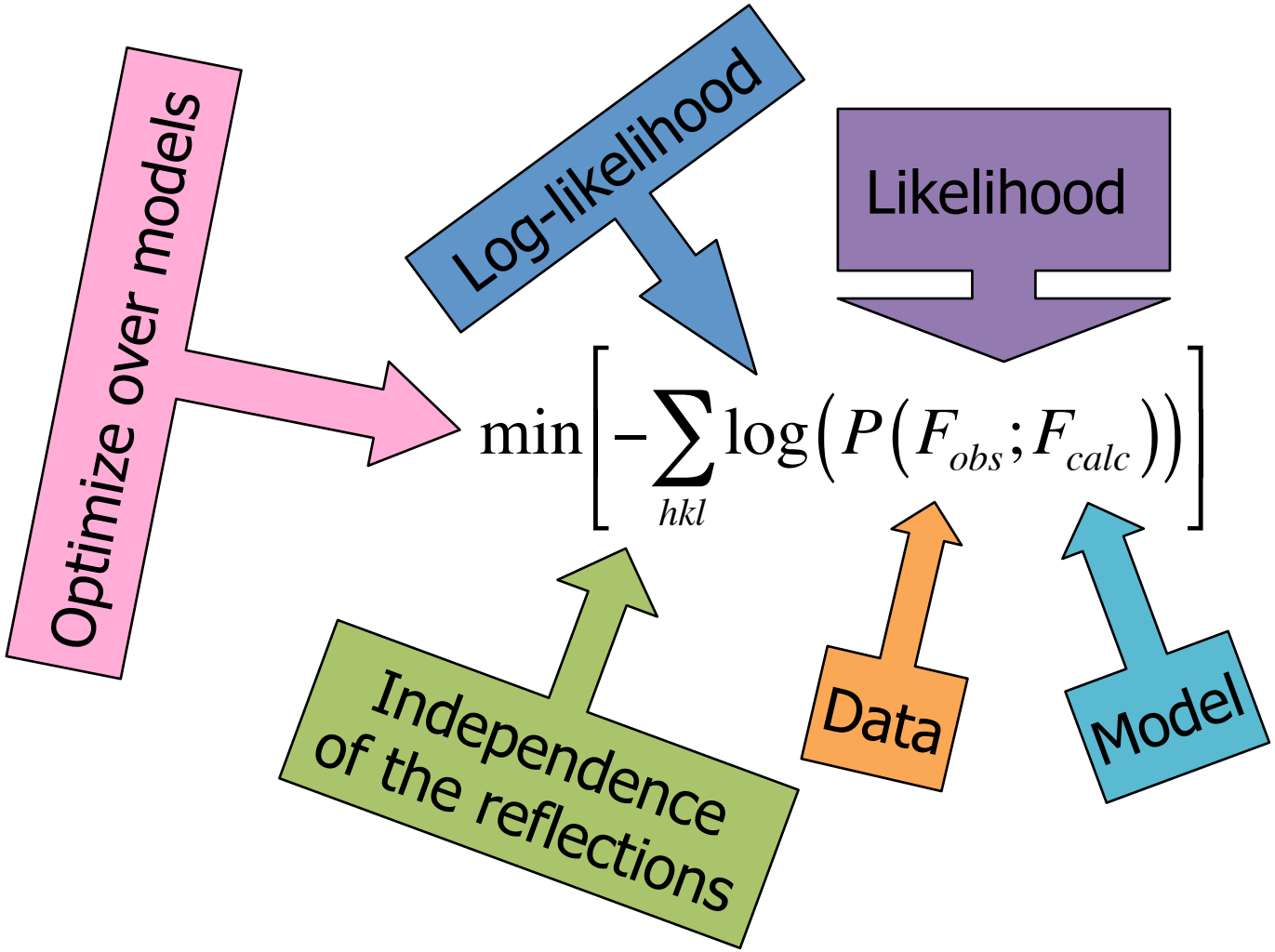
$$= \log\left(P\left(3;\boxed{6}\right)\right) + \log\left(P\left(3;\boxed{6}\right)\right)$$

$$= \log\left(\tfrac{1}{6}\right) + \log\left(\tfrac{1}{6}\right)$$

$$= -0.778 - 0.778$$

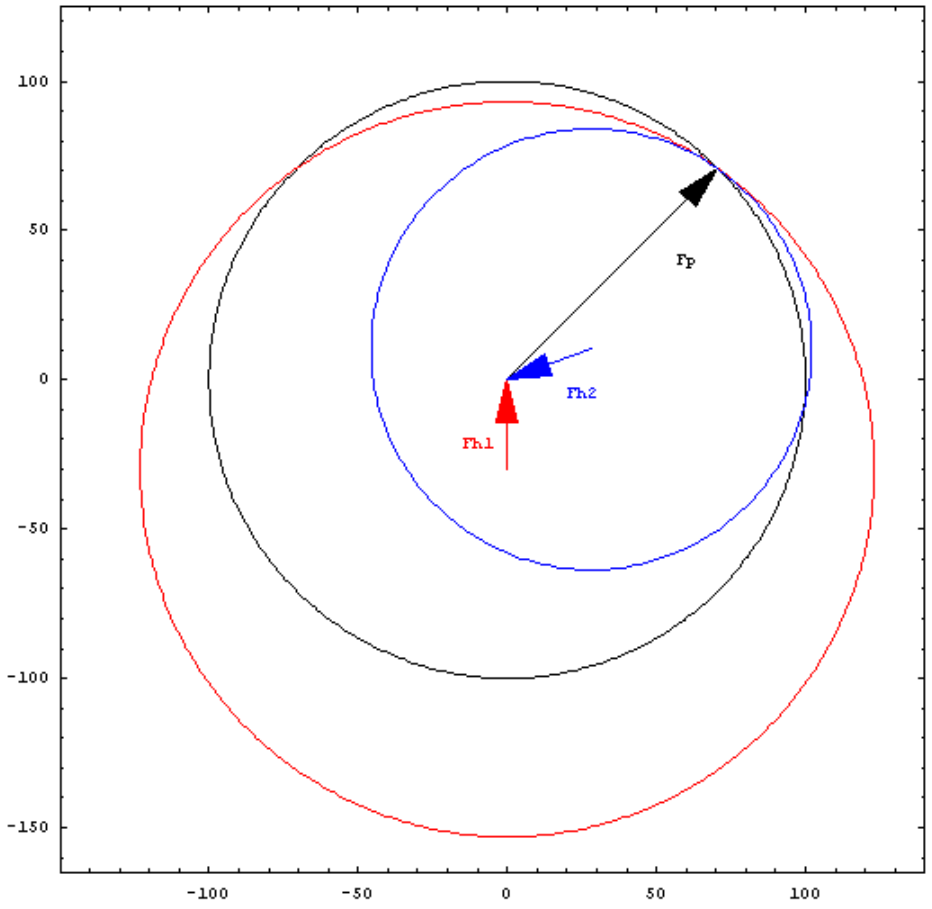
$$= -1.556$$

Likelihood Functions



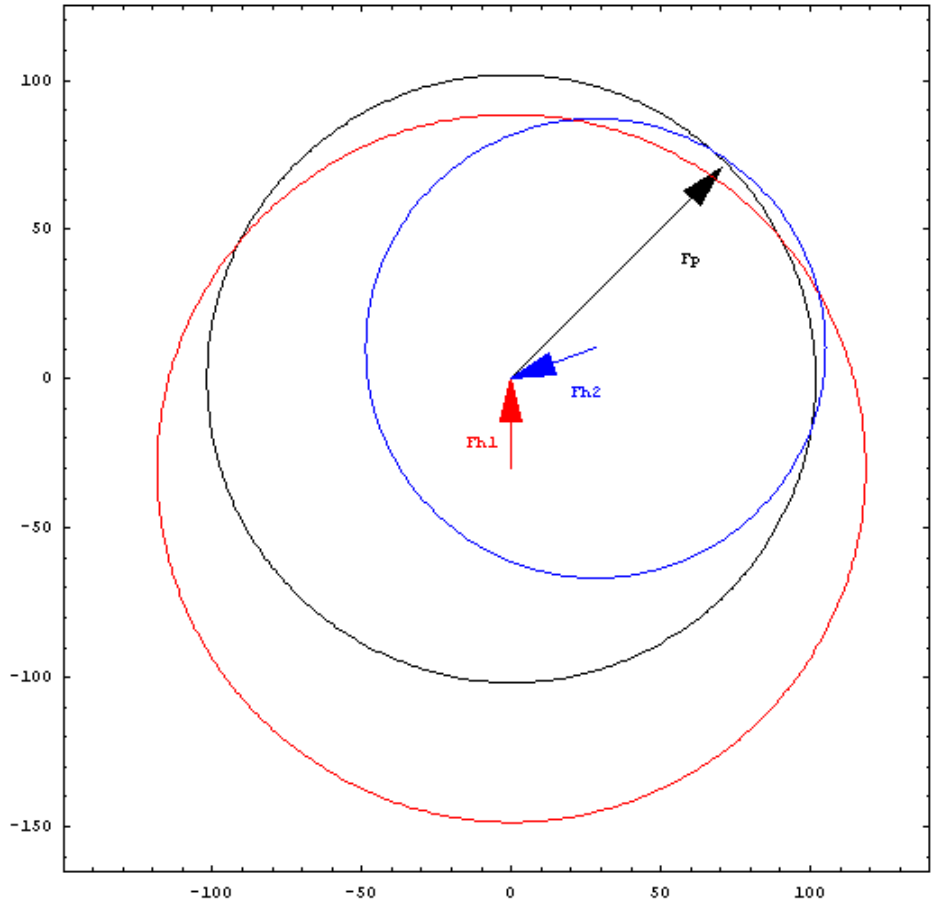
Harker construction

- Phasing of one refection using two derivatives with no errors
- Phase determined with very high probability



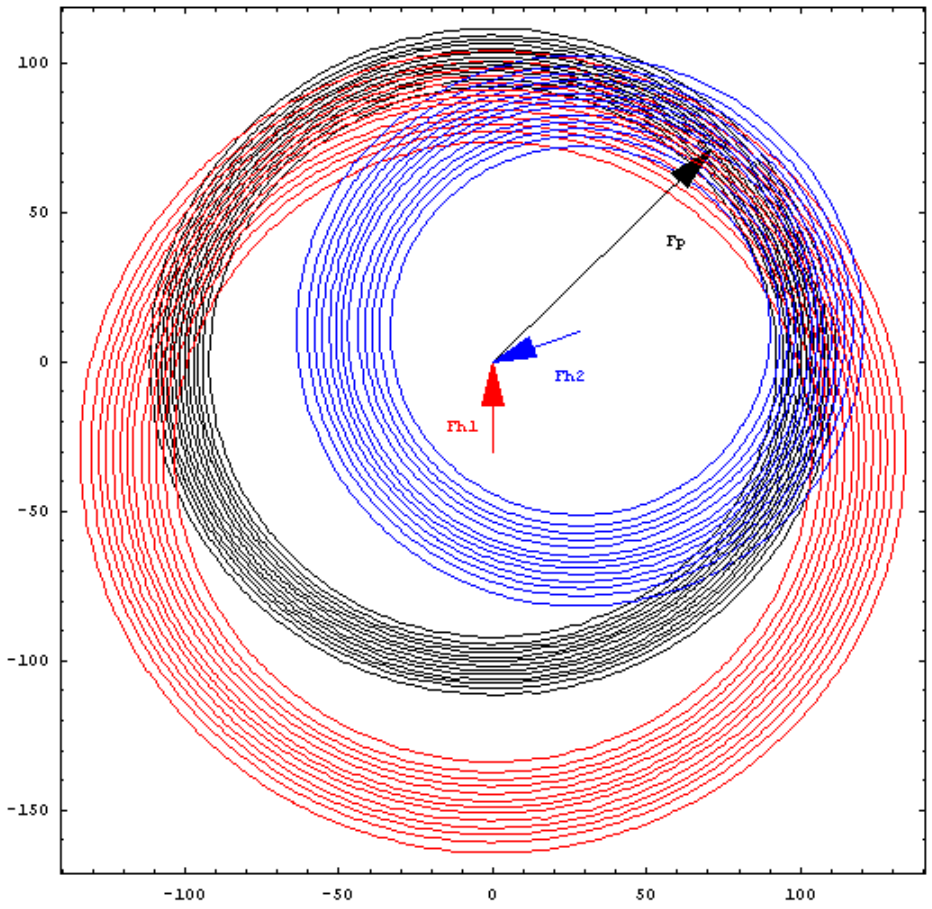
Harker construction

- There are many sources of error in the experiments
  - Mainly model errors
  - Also data errors
- The errors are large
- We are looking for the best phase
- We therefore need a probability function



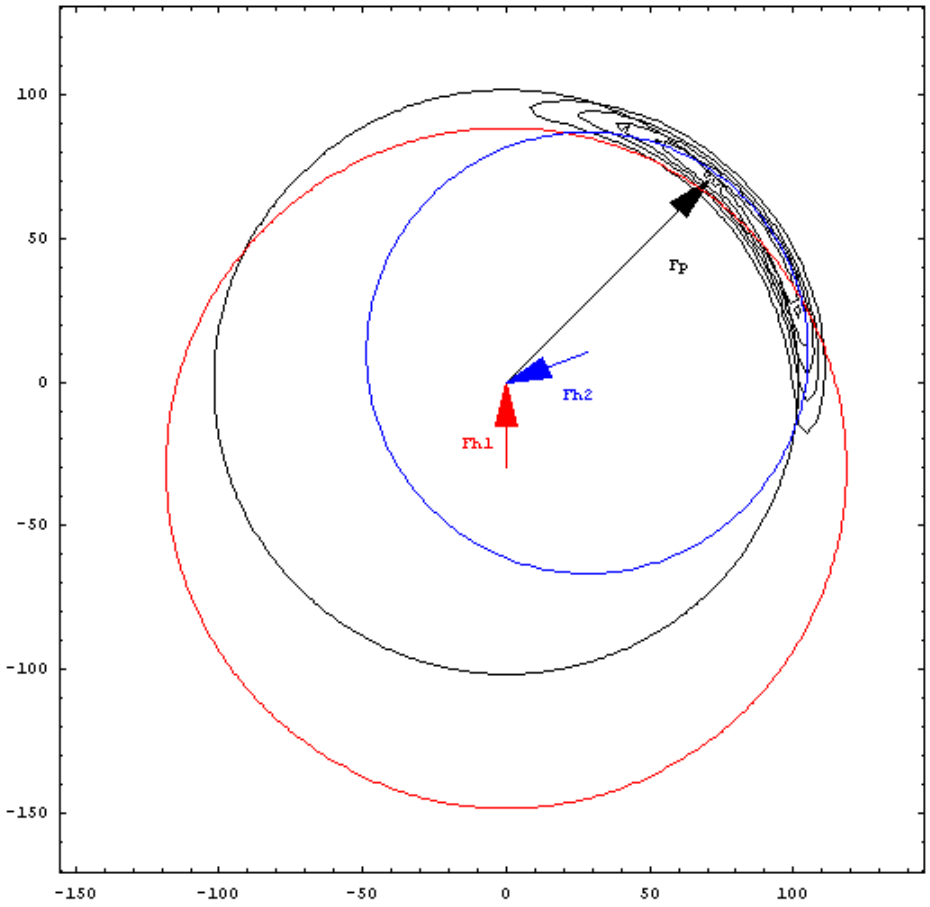
Probabilistic Harker Diagram

- Each circle has an error associated with it to give a distribution
- The total likelihood is the volume under the curve of the product of the distributions



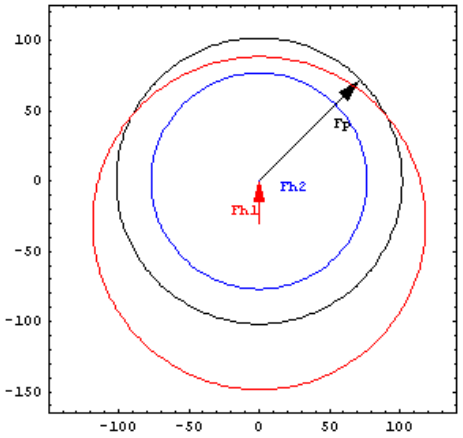
Probabilistic Harker Diagram

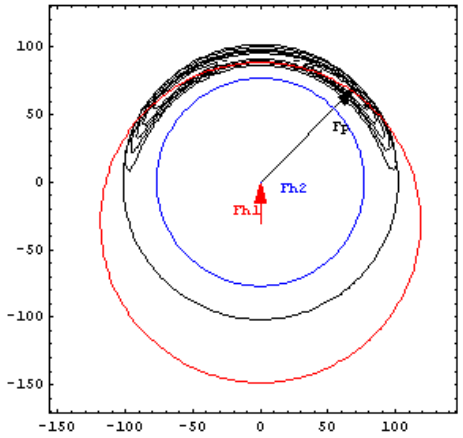
- The final distribution is high only where all three circles overlap



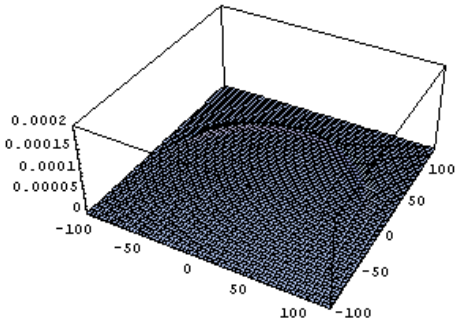
Refining Occupancy

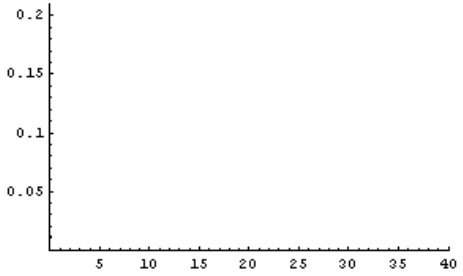
To refine the occupancy of a heavy atom, maximise the likelihood (area under the curve)





Final refined value is the optimum for ALL reflections (movie shows ONE reflection)

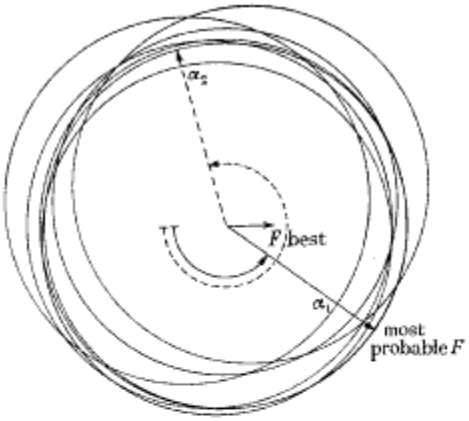




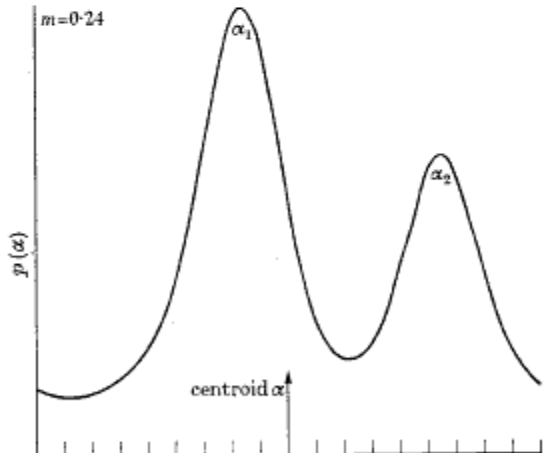
Density modification

Phase probability

- Each reflection really has a phase probability density function (PDF) rather than a single phase
- This is a complicated mathematical function
  - Requires lots of memory
- Four Hendrickson-Lattman coefficients (A,B,C,D) are used to store this PDF in a compact form



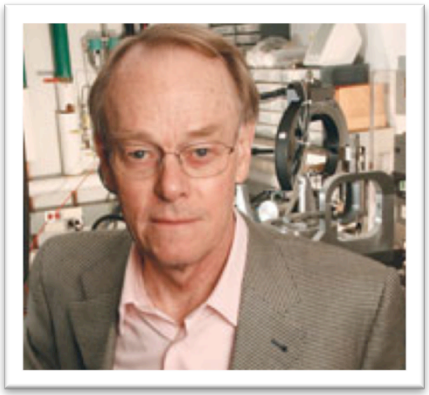
(b)



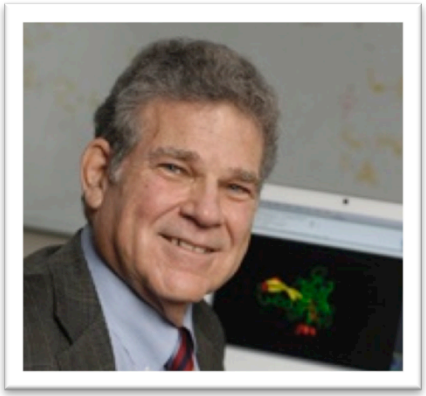
Hendrickson-Lattman Coefficients

$$P(\alpha) \propto \exp\left[A\cos(\alpha) + B\sin(\alpha) + C\cos(2\alpha) + D\sin(2\alpha)\right]$$

- HL coefficients allow for easy combination of phase information from multiple sources
  - the combined PDF is formed simply by adding the A,B,C, and D from the two distributions



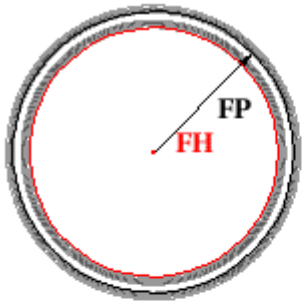
Wayne Hendrickson

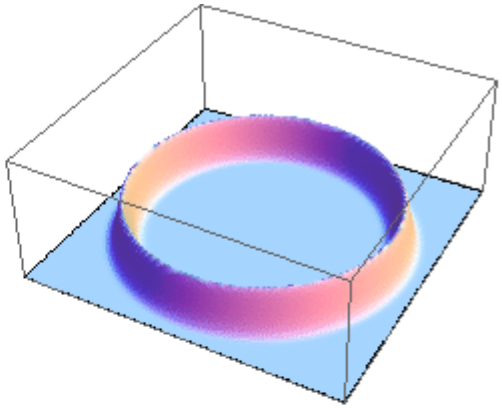


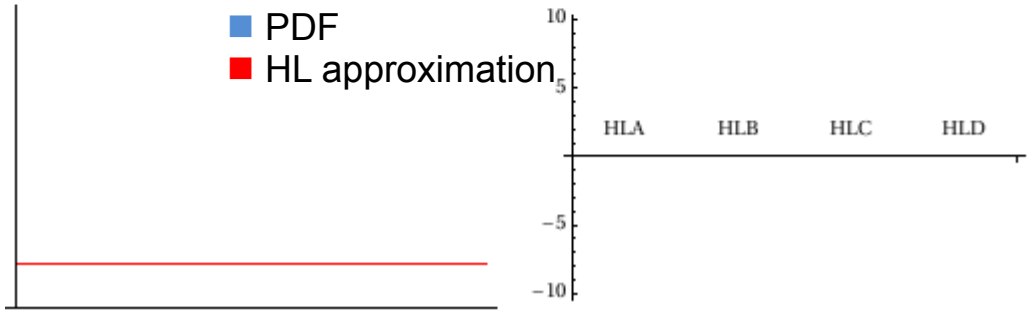
Eaton Lattman

Hendrickson-Lattman Coefficients

HL coefficients as a function of F_H occupancy



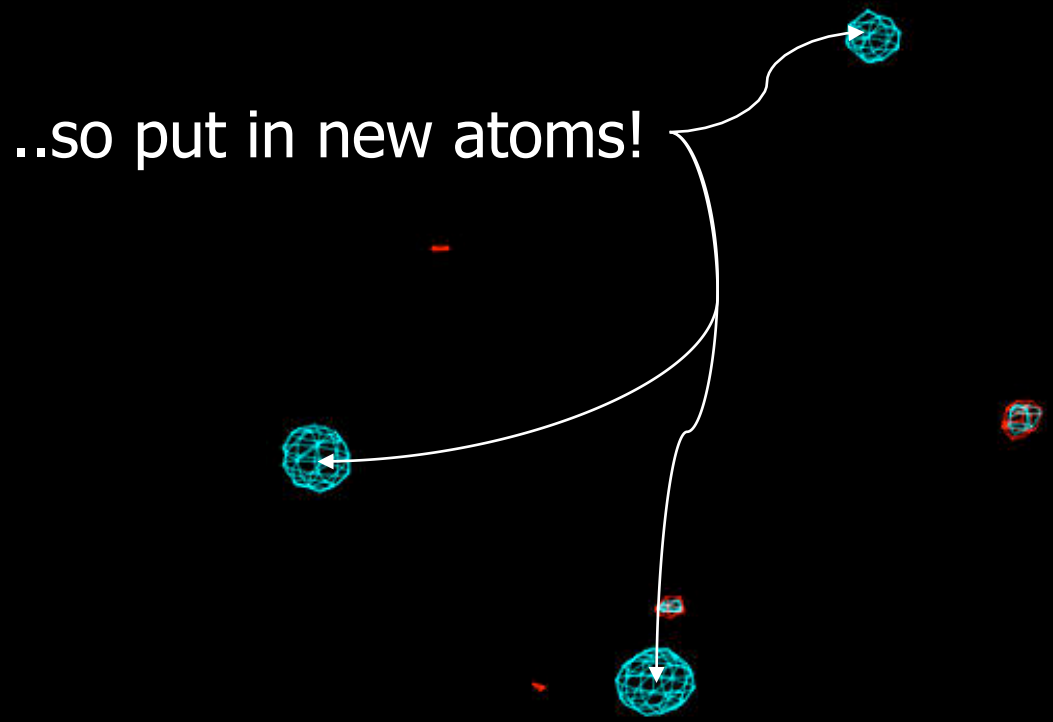




Log-likelihood gradient maps

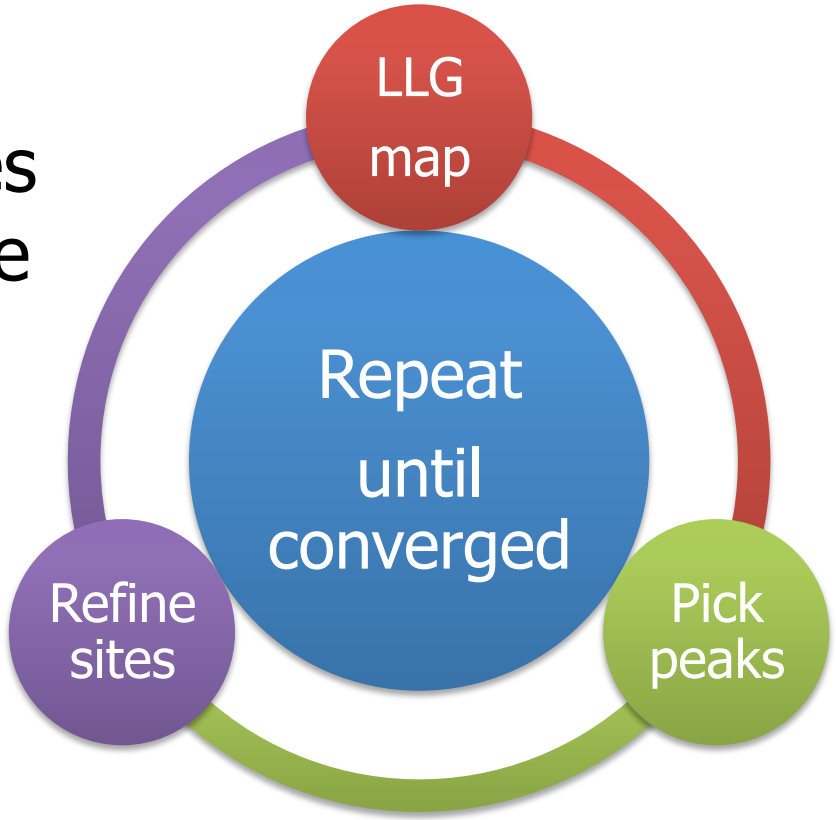
Log-likelihood gradient maps

Gradient of likelihood plotted with respect to coordinates shows where the likelihood would increase if there were atoms present…



Completion of sub-structure

- LLG maps are very sensitive
- Inclusion of minor sites **greatly improves** the phases
- Could include all intrinsic sulphurs
- Also finds bound halides



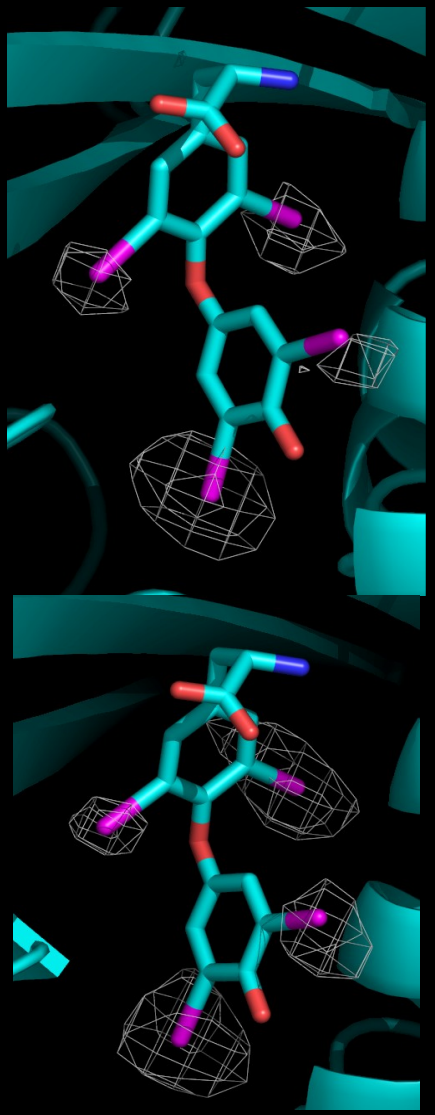
Thyroxine binding globulin

Where does thyroxine bind?

Thyroxine contains 4 iodine atoms

Two molecules in asymmetric unit
d_{min}=2.8Å
λ=0.979Å
f'' ≈ $3e^-$

Phaser LLG map @5.5σ



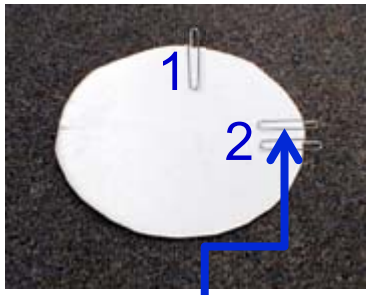
Zhou *et al*. (2006). *PNAS* **103**: 13321

Model building

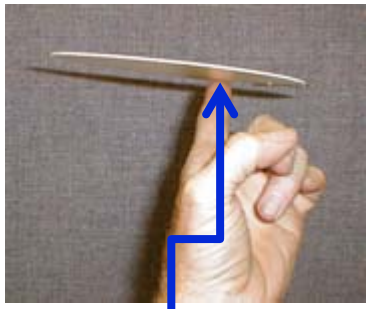
Calculating Electron Density

- ML function is good for refining the parameters, but what phase should be used in the electron density equation?
  - Have to pick one phase
- **We want the phase that gives the electron density with the lowest rms error**
  - Parseval's theorem relates the rms error in real space to the rms error in reciprocal space and vice versa
- This phase (the "**best phase**") is the probability-weighted average of all the phases
  - It is not the "most probable phase"

- Cut the centre out of a polystyrene foam plate
- Balance the disk on your finger
  - The centre of mass is at the centre
- Now put 3 paperclips on the edge of the disc
  - 2 together
  - 1 a distance away
- The balancing point is **between** the 3 paperclips
  - Not on the 2 paperclips

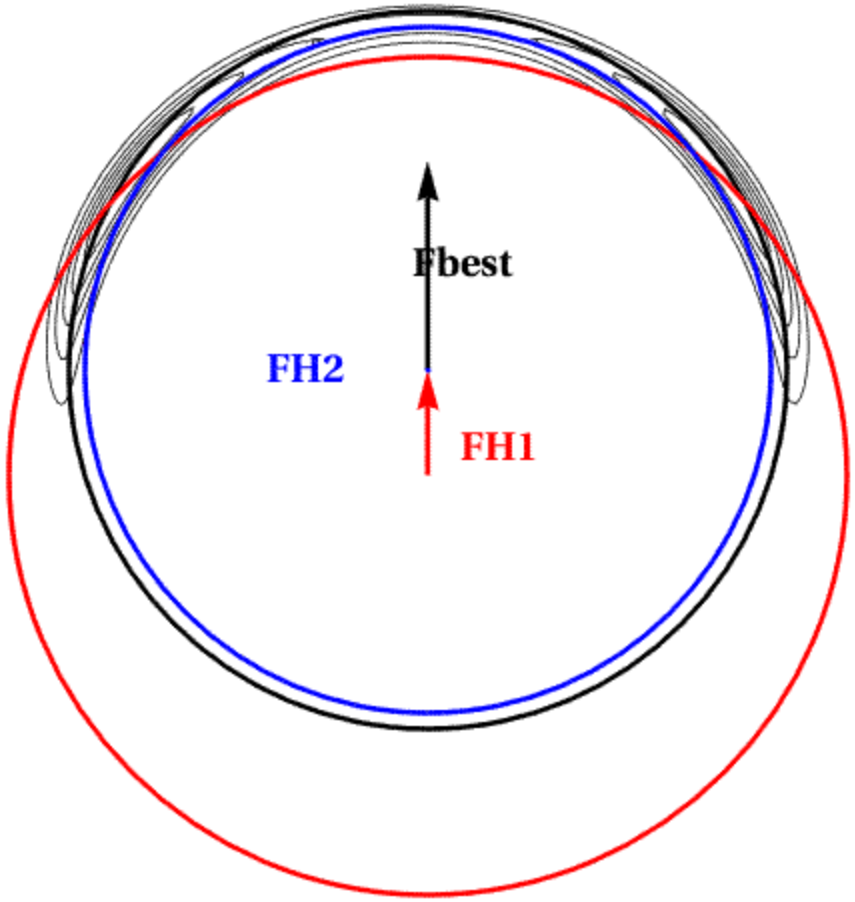


Most Probable Structure Factor



Best Structure Factor

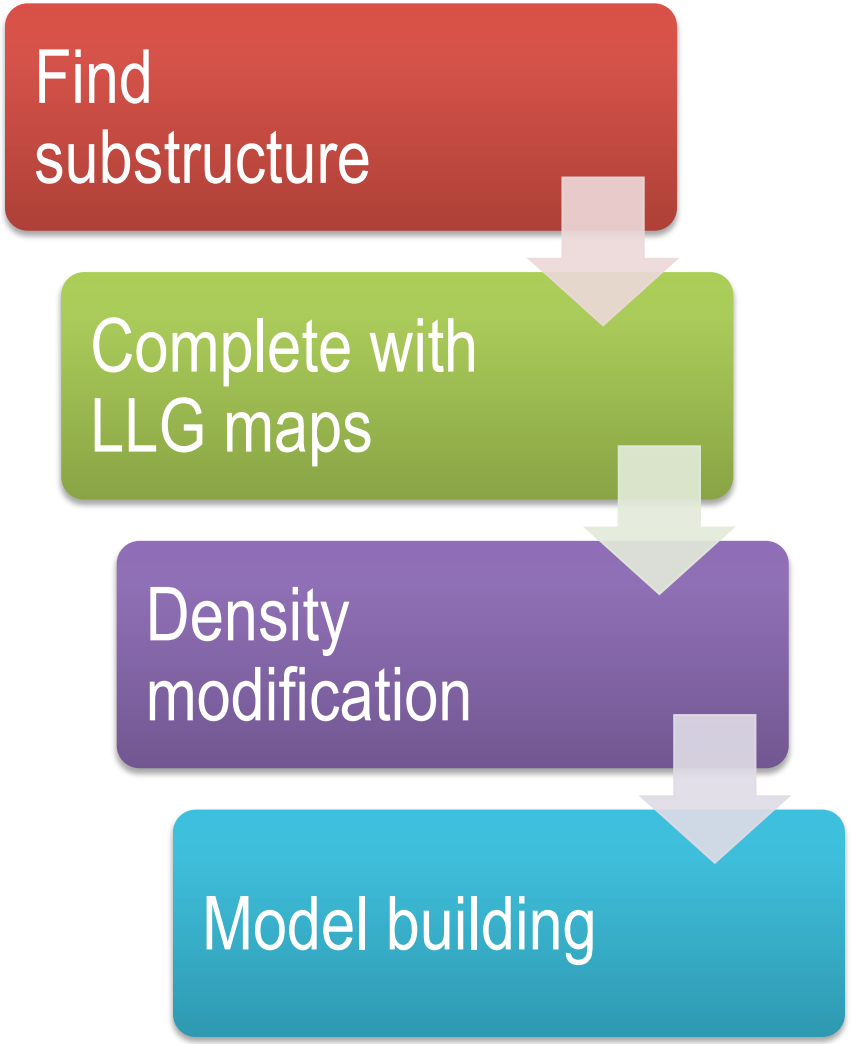
- F_{best} has a lower |F| amplitude than F_{obs}
- The reduction in F_{obs} to give F_{best} is expressed as the "figure of merit" (m)
  - **0<m <1:** F_{best} lies inside the F_{obs} circle
  - **m=1 :** Perfect phase information
  - **m=0:** No phase information
  - The higher the average value of the figure of merit, the better



New approaches

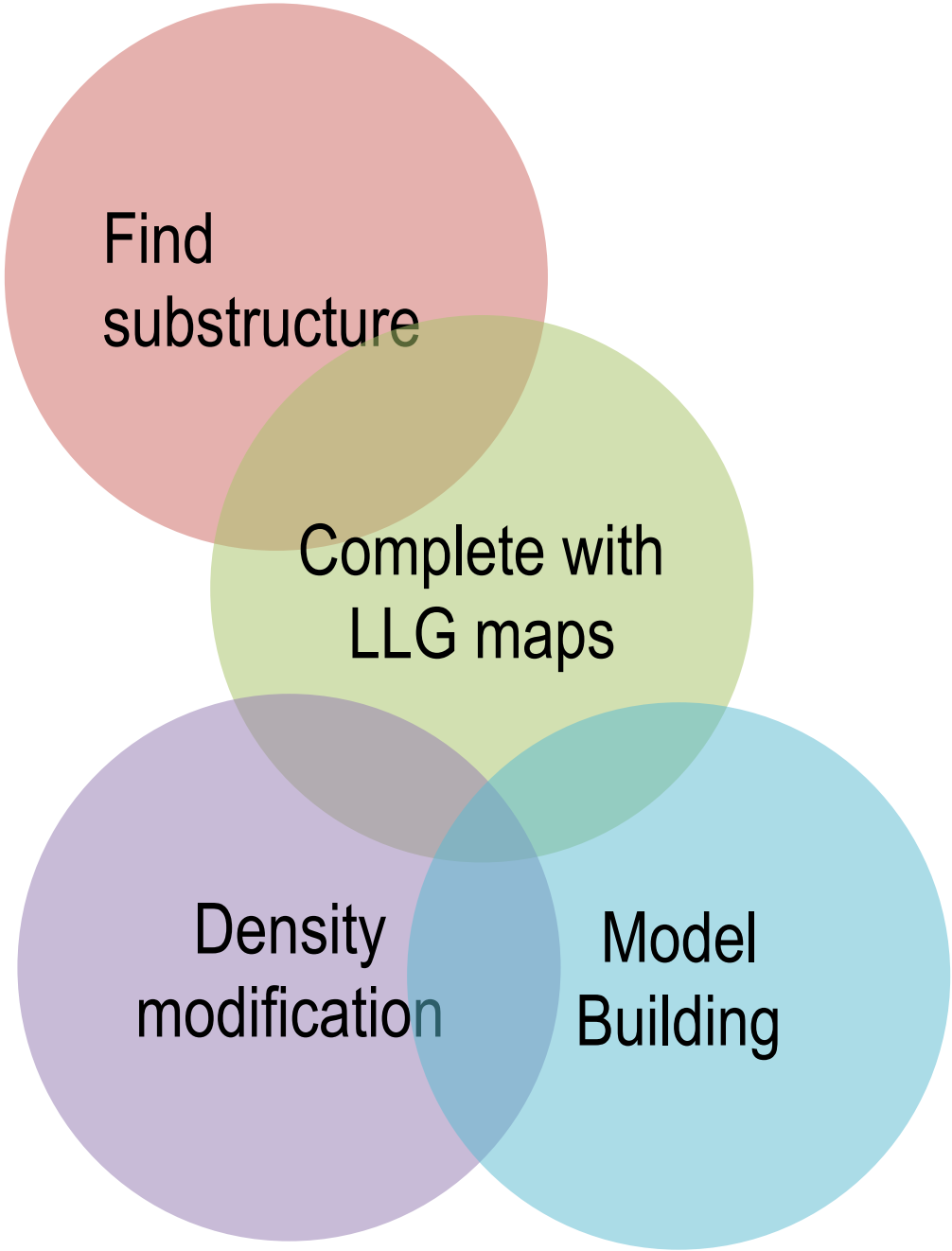
The pathway of structure solution

- Historically, there has been a linear progression through structure solution
- You had to be sure each step is correct before progressing to the next
- When signal is low you cannot be sure (of anything)

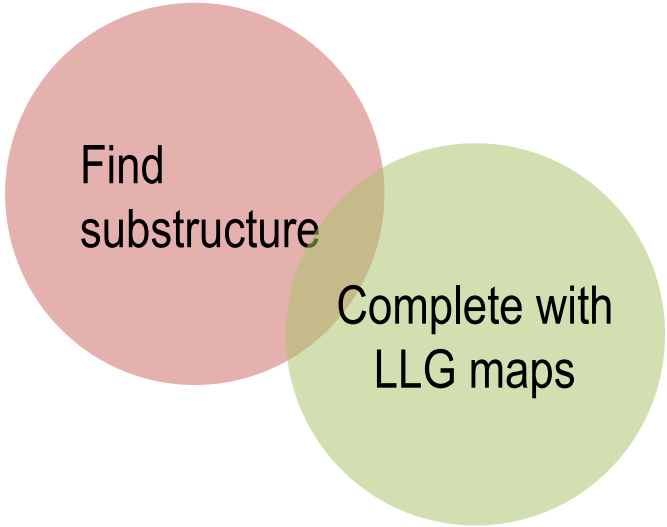


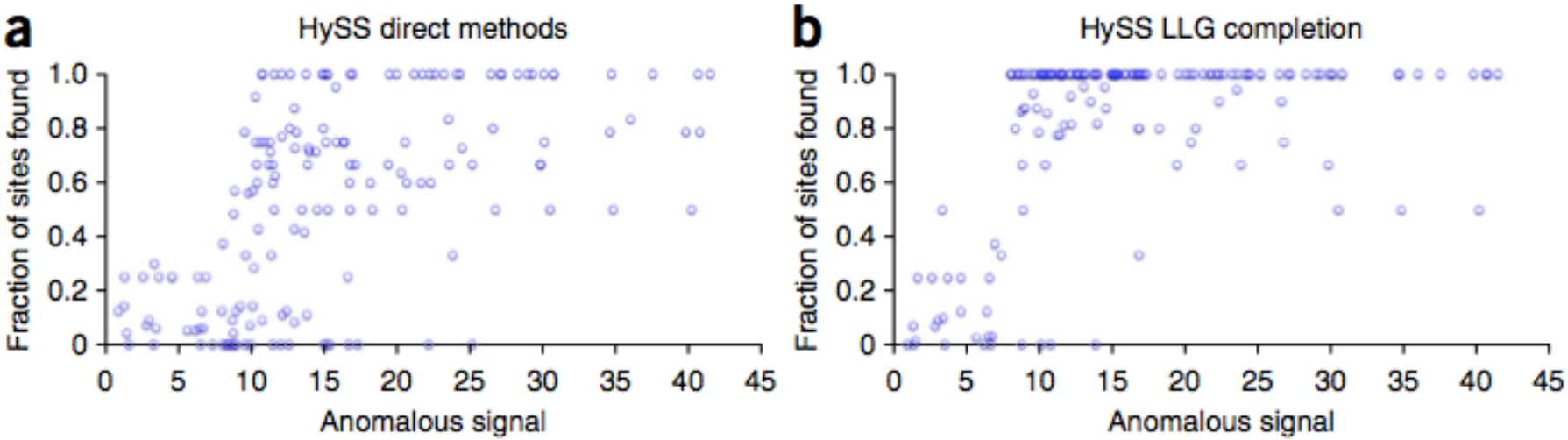
New approaches

- Take multiple possibilities for each step and uses subsequent steps to distinguish correct from incorrect solutions
- Enables structure solution when signal is low

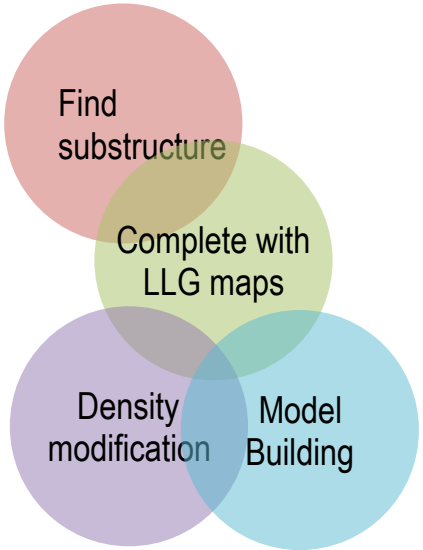


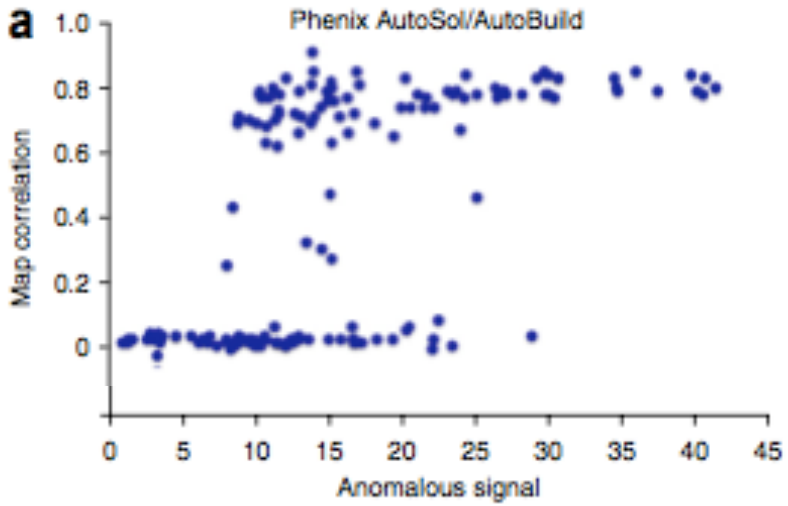
Find phasing substructure

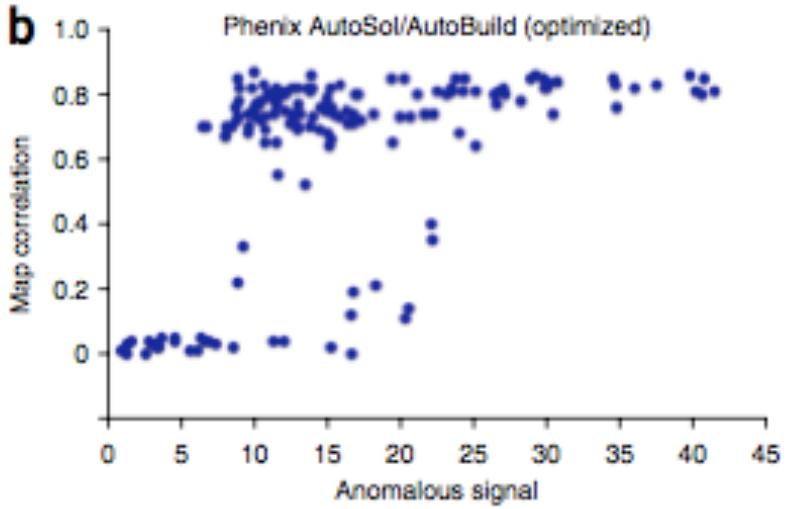





Find substructure
Complete with LLG maps
Density modification
Model building


Find substructure
Complete with LLG maps
Density modification
Model Building


a
Phenix AutoSol/AutoBuild
Map correlation
1.0
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35 40 45
Anomalous signal


b
Phenix AutoSol/AutoBuild (optimized)
Map correlation
1.0
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35 40 45
Anomalous signal

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Youval Dar, Nat Echols, Nigel Moriarty, Nader Morshed, Ian Rees, Oleg Sobolev

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner, Richard Mifsud

Duke University

Jane & David Richardson, Chris Williams, Bryan Arendall, Bradley Hintze

An NIH/NIGMS funded Program Project