

Molecular Replacement

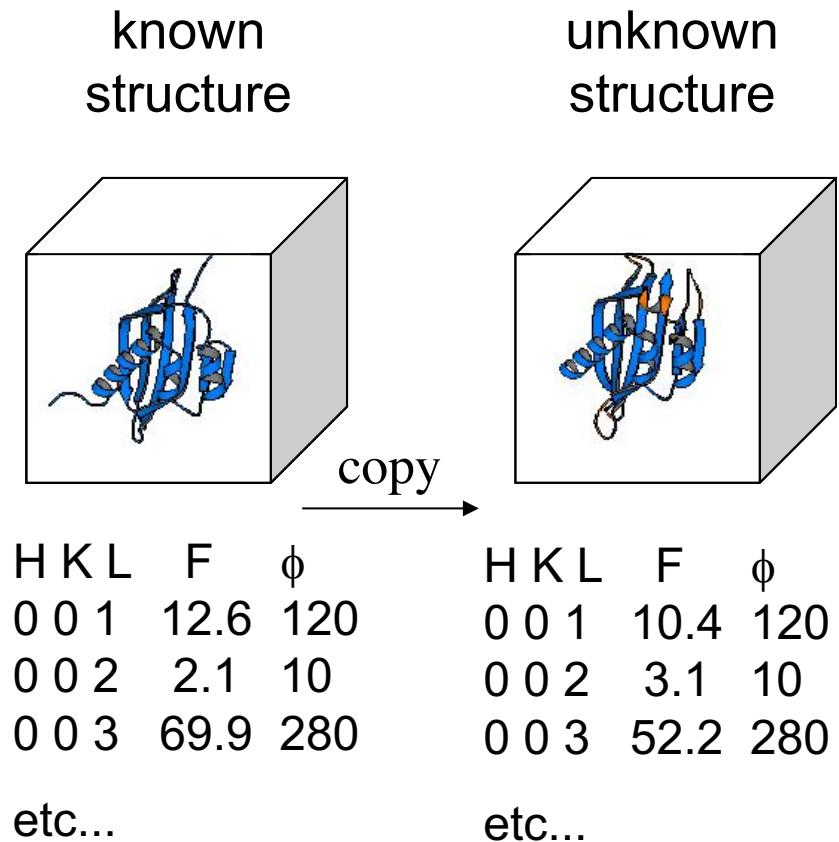
Airlie McCoy



UNIVERSITY OF
CAMBRIDGE

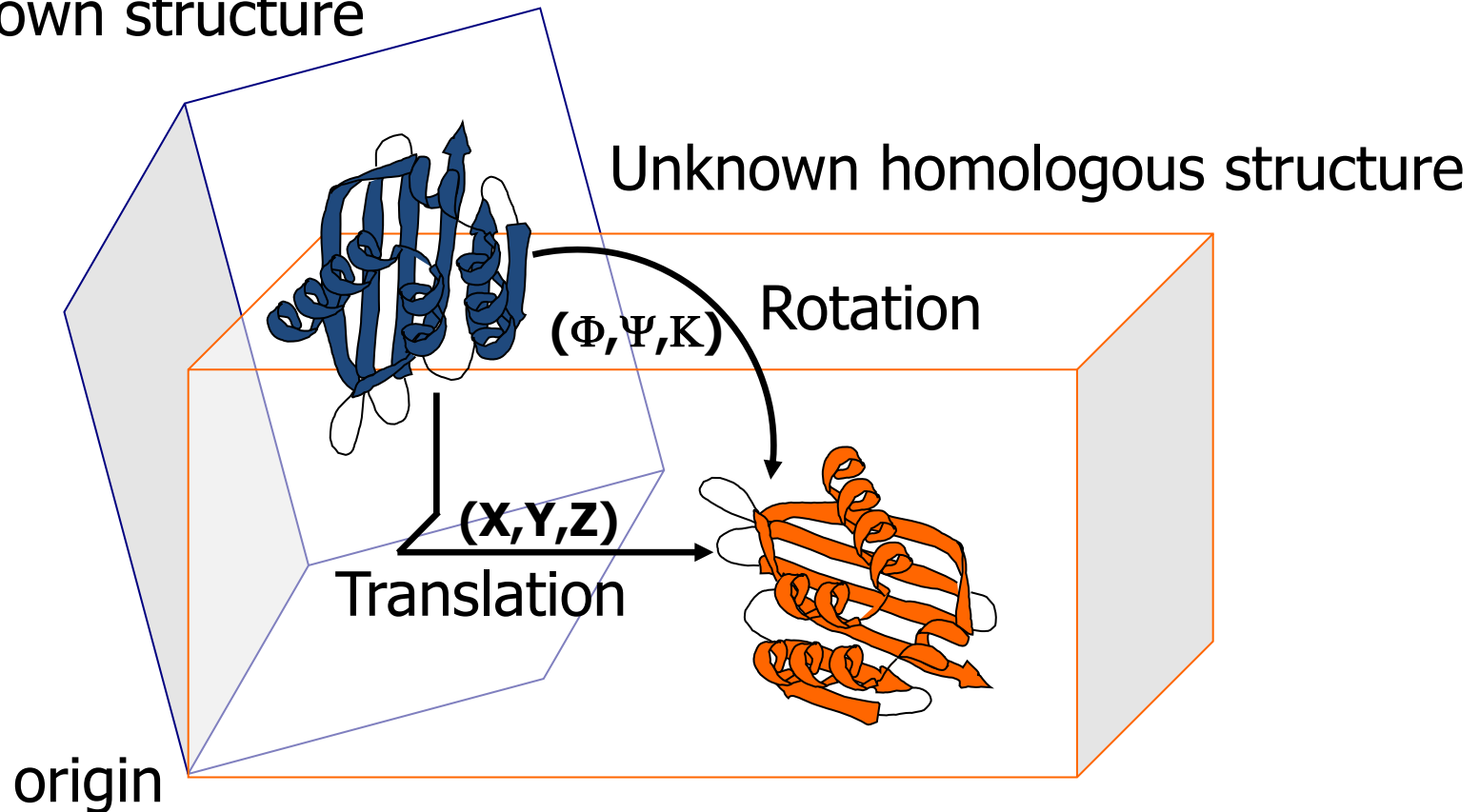
Molecular Replacement

- Find orientation and position where model overlies the target structure
- Borrow the phases
- Then it becomes a refinement problem – the phases change



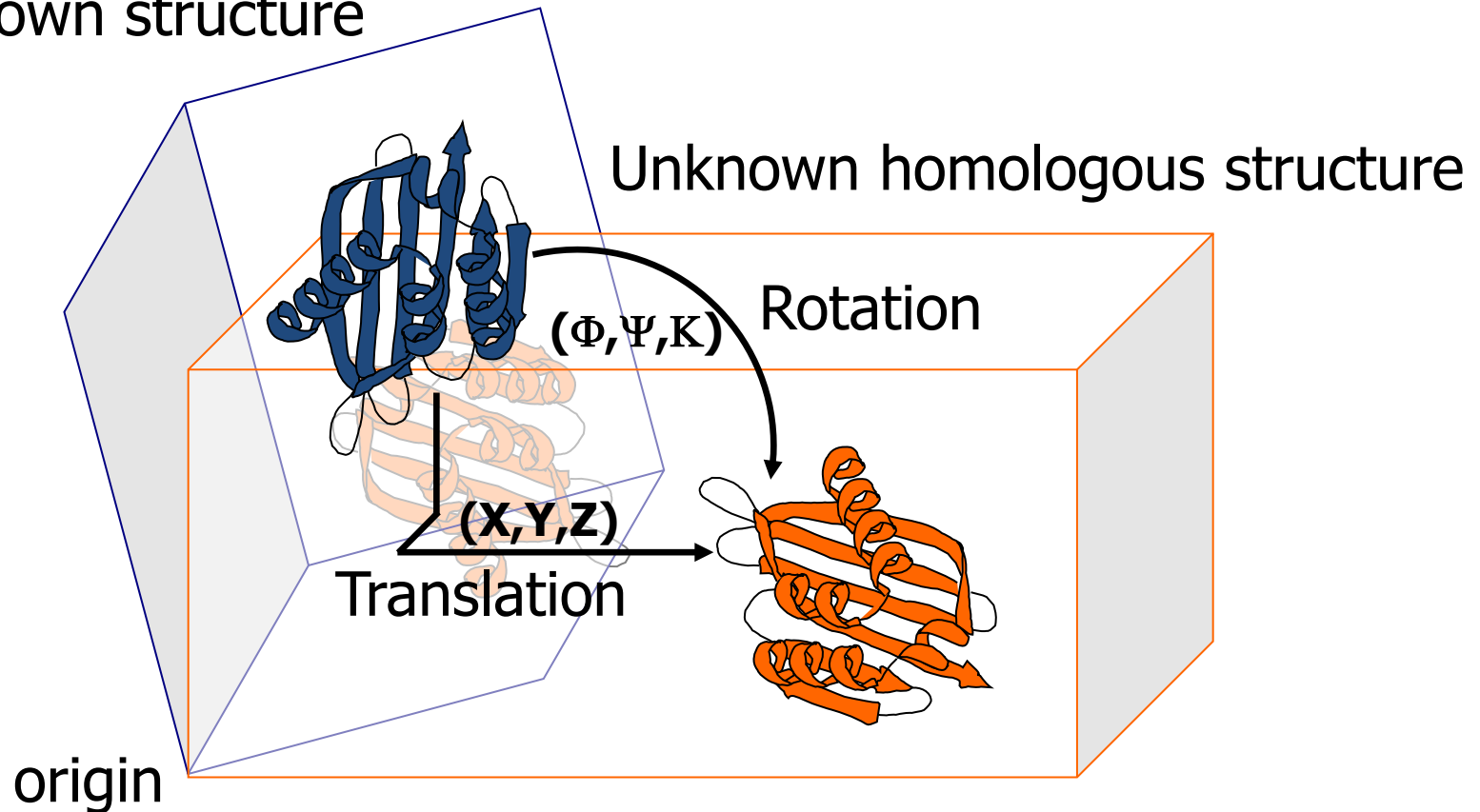
Molecular Replacement

Known structure

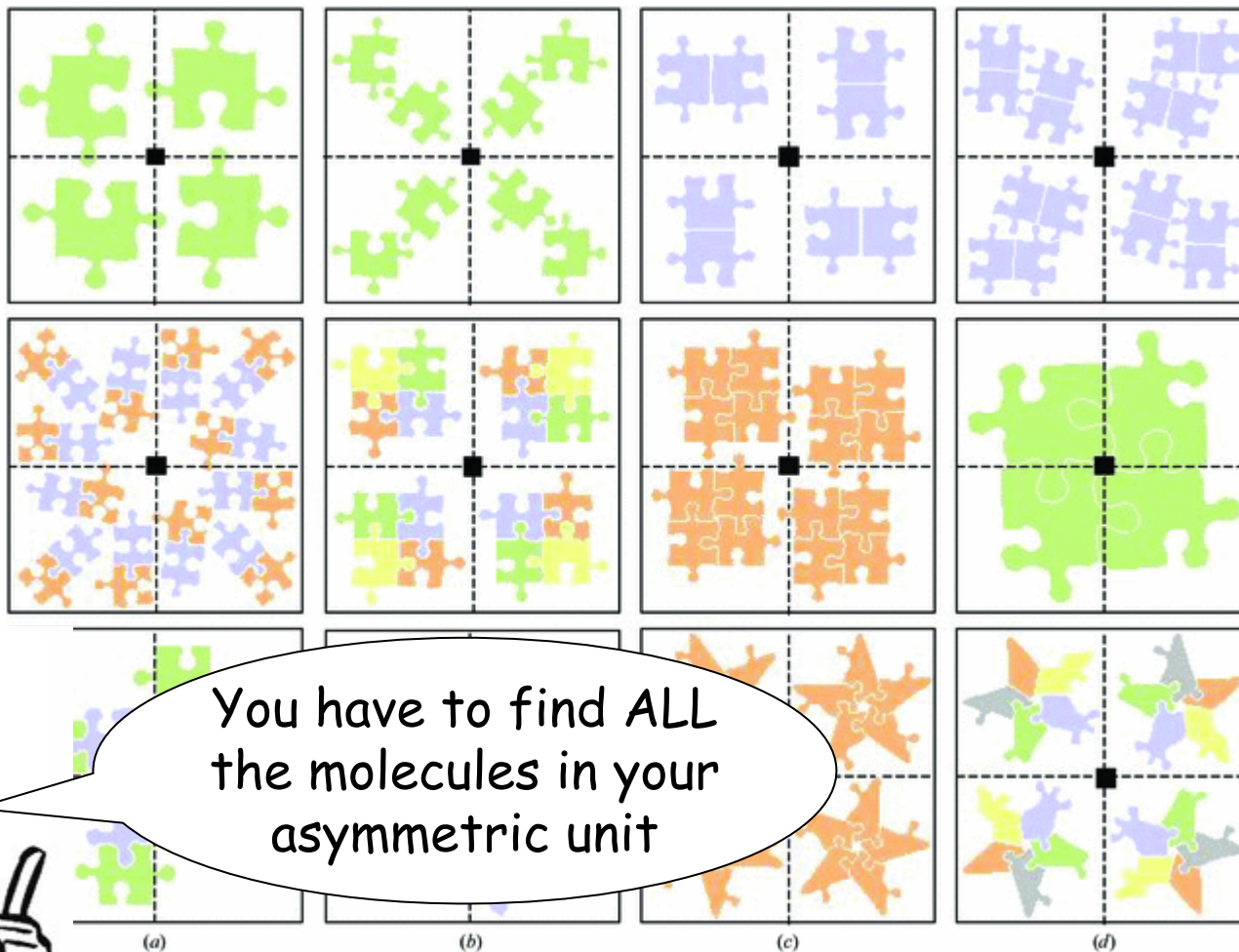


Molecular Replacement

Known structure



Contents of the Asymmetric Unit



You have to find ALL
the molecules in your
asymmetric unit

Matthew's coefficient

- First calculated by Brian Matthews in 1968 (over 3500 citations)
- Most crystals are 50% protein by volume
- Can be used to estimate the contents of the asymmetric unit
- *Self Rotation Function*
- *TNCS Order*

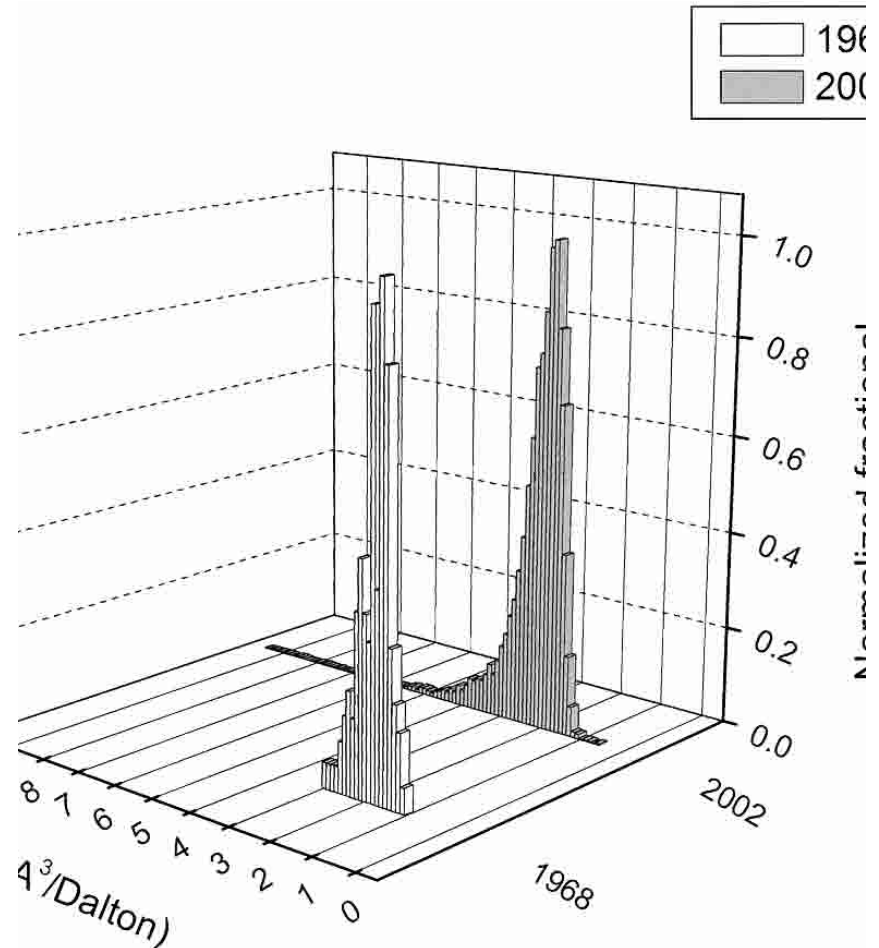
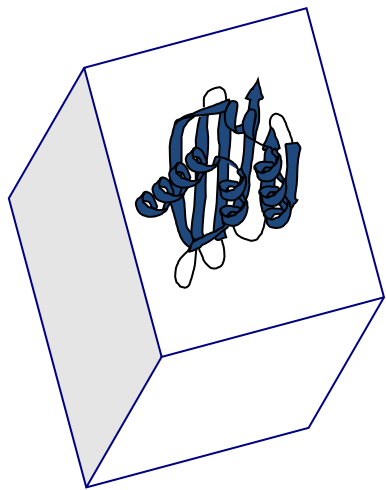


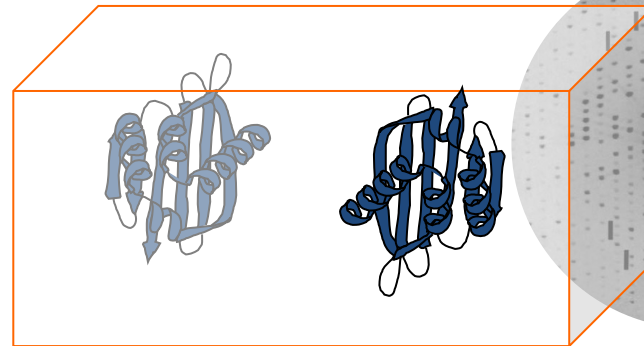
Figure 1: Kantardjieff and Rupp (2003)



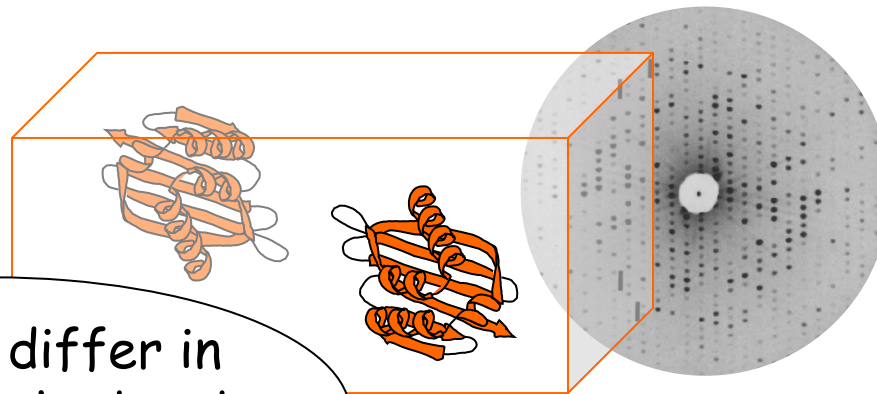
SEARCH

(Φ, Ψ, χ)

(x, y, z)



SCORE



Programs differ in
search method and
scoring function

Molecular Replacement

- Issues
 1. How to **score** each orientation and position so as to find when the model best fits the target structure
 2. How to **search** for solutions: strategies for exploring rotations and translations
 - MR can fail due to suboptimal choices in either
 - **Choice of search method and scoring function are not independent**
 - Different scoring functions allow different search methods
-

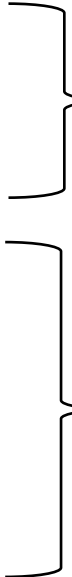
Search strategy

- Each molecule needs 6 parameters (6D)
 - An exhaustive search is big
 - All angles sampled at 2.5° ; $N_{\text{rot}} = 1.5 \times 10^6$
 - All positions sampled at $1\text{\AA}$ in a $100\text{\AA}$ cubic cell;
 $N_{\text{tra}} = 1.0 \times 10^6$
 - 6 dimensional search is $N_{\text{rot}} \times N_{\text{tra}} = 1.5 \times 10^{12}$ points
 - This is only ONE component of asymmetric unit
 - MR search strategies can be divided into rotation and translation separately (2x3D)
 - $N_{\text{rot}} + N_{\text{tra}} = 2.5 \times 10^6$ points
-

Scoring

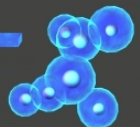
- What is the “best match” between the observed and calculated structure factors
 - Can use
 - Patterson methods/Correlation Coefficient
 - Maximum Likelihood methods
 - Can use
 - Structure factor amplitudes
 - Structure factor intensities
-

Software

- MolRep
 - AMoRe
 - CNS
 - EPMR
 - COMO
 - SOMoRe
 - Queen of Spades
- 
- Diagram illustrating software groupings:
- Group 1 (Red text):
 - Patterson RF/CC TF
 - Amplitudes
 - Independent rotation functions
 - Group 2 (Orange text):
 - CC (TF)
 - Amplitudes
 - 6D (but not exhaustive)
- Phaser
 - Maximum likelihood
 - Intensities
 - Additive rotation functions
 - Amalgamation of TF solutions
-

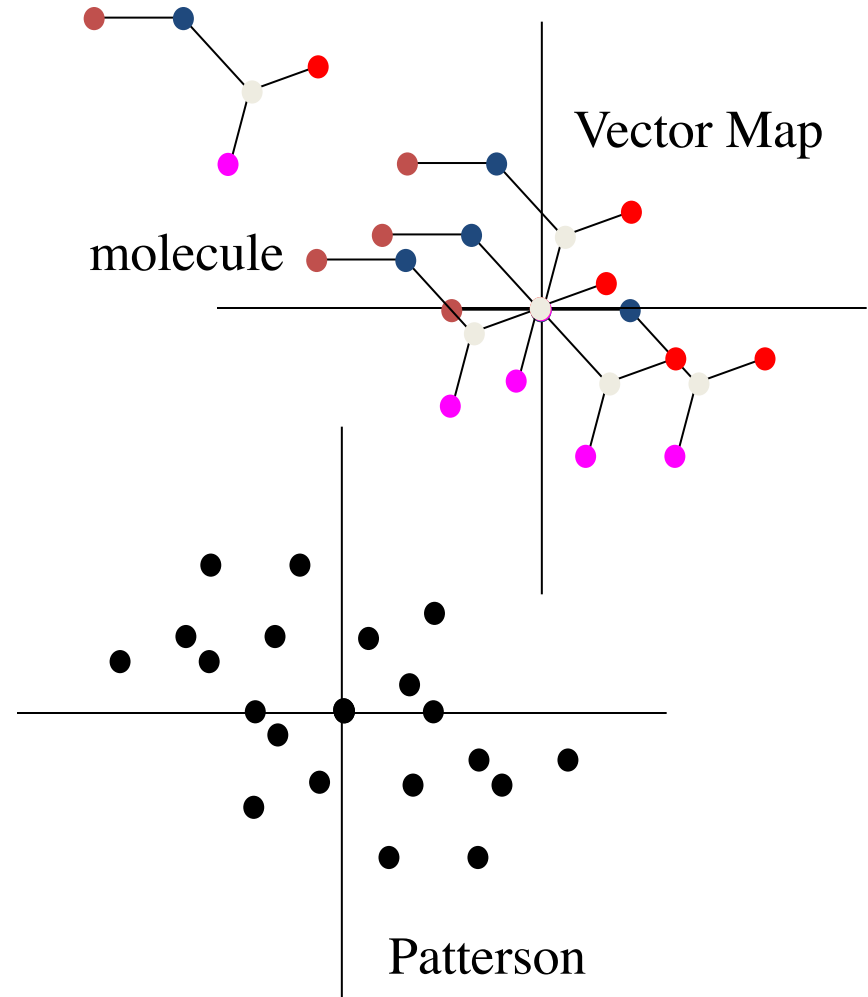
Some Important Features of Phaser

- ML can take account of errors
 - Model errors can be very large
 - Data errors can be very large for weak reflections
- ML allows solutions to be built up by addition
- Expected LLG allows artificial intelligence to be built into resolution selection, search order, and termination criteria
 - Bias free structure pruning
- Translational NCS correction allows new classes of structures to be solved by MR



Patterson Scoring

- Patterson is the FT of the amplitude² and the phases set to zero
- Can be calculated from the intensities
- This is the vector map of the atoms
 - Can be deconvoluted if the structure is small



Maximum Likelihood Scoring

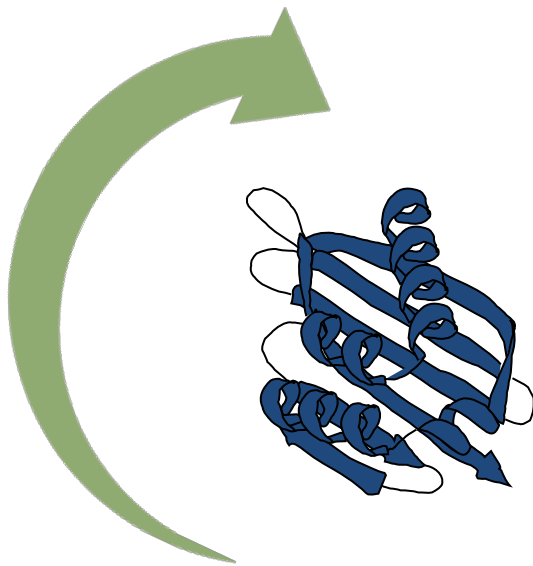
- Use probability
- Probabilities account for errors
 - Patterson methods cannot do this

$$LLGI = \sum_{\mathbf{h}} \log \left(\frac{2E_e}{1 - D_{obs}^2 \sigma_A^2} \exp \left(-\frac{E_e^2 + D_{obs}^2 \sigma_A^2 E_C^2}{1 - D_{obs}^2 \sigma_A^2} \right) I_0 \left(\frac{2E_e D_{obs} \sigma_A E_C}{1 - D_{obs}^2 \sigma_A^2} \right) \right)$$

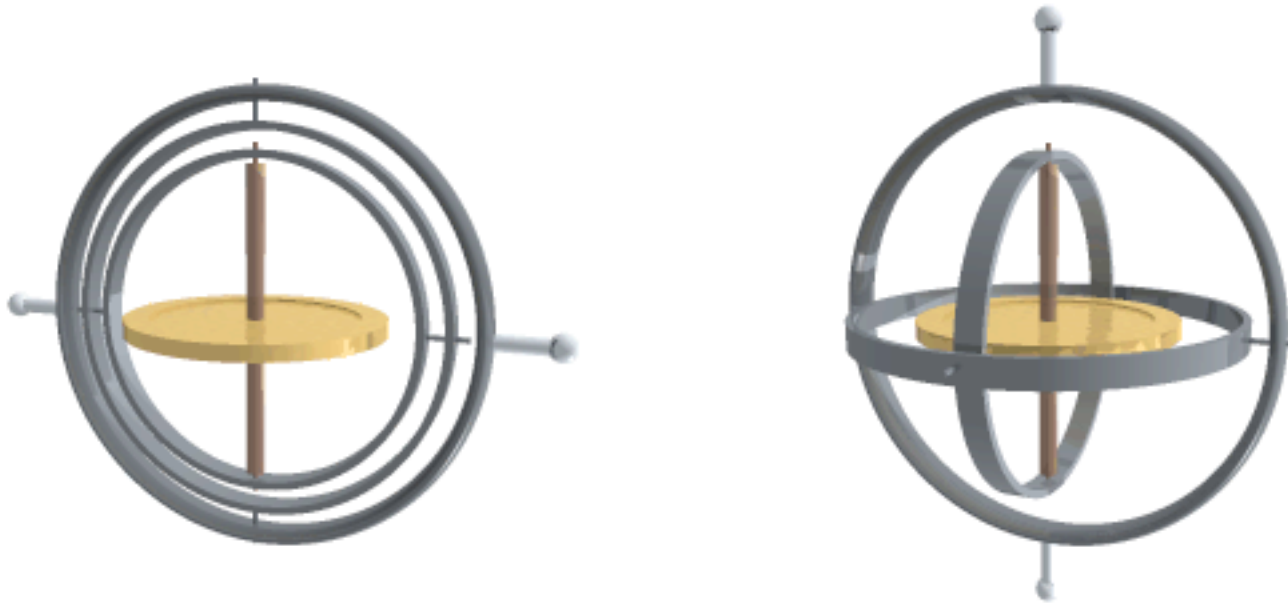
E_e and D_{obs} are defined as in (Read & McCoy, 2016); E_e is the effective E , representing information derived from E_{obs}^2 , and D_{obs} represents the reduction in correlation between observation and E_e arising from experimental error; $E_{obs}^2 = I_{obs}/(\varepsilon \Sigma_N)$ where ε and Σ_N includes correction terms for anisotropy and tNCS modulations

Brute rotation function search

- Place model at orientations and calculate probability of each orientation being correct

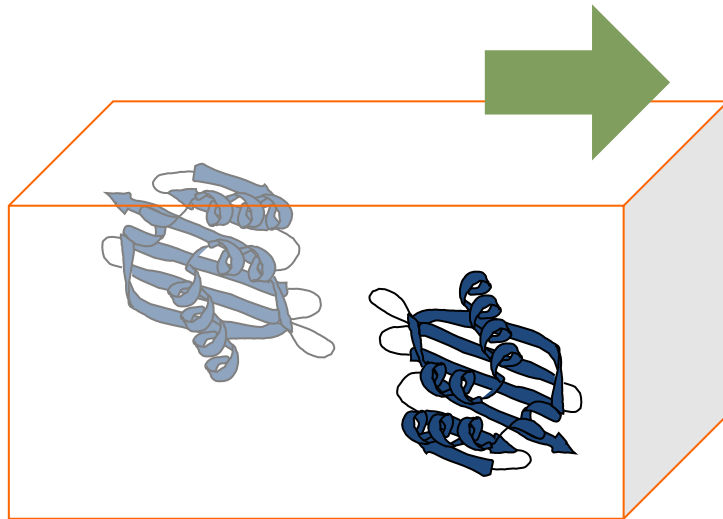


Euler Angles



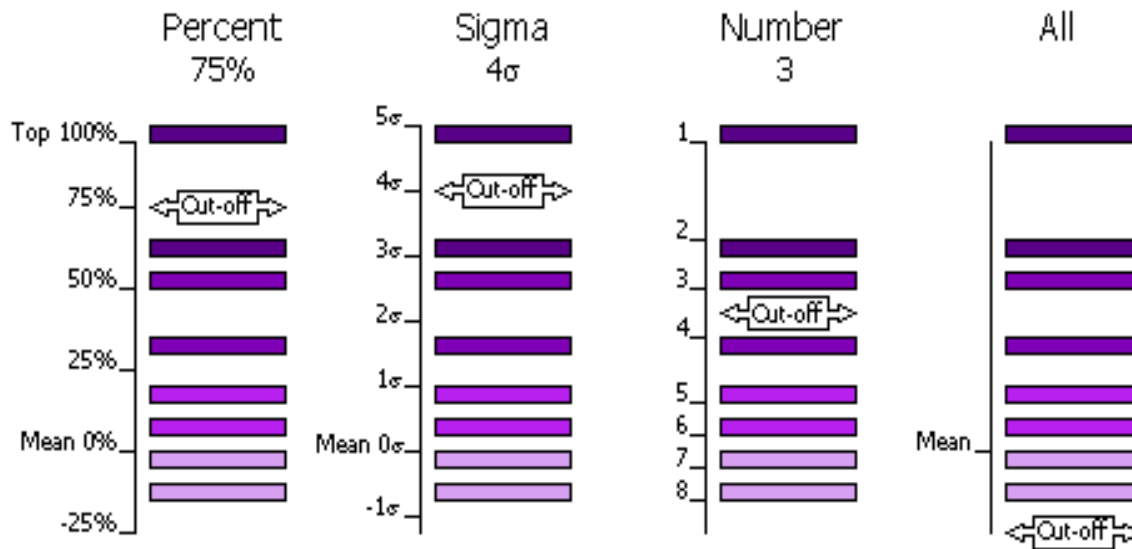
Brute translation function search

- Place model at points in unit cell and calculate probability that it is in each position



Peak selection

- Must choose a selection criteria to carry **potential** solutions through to the next step



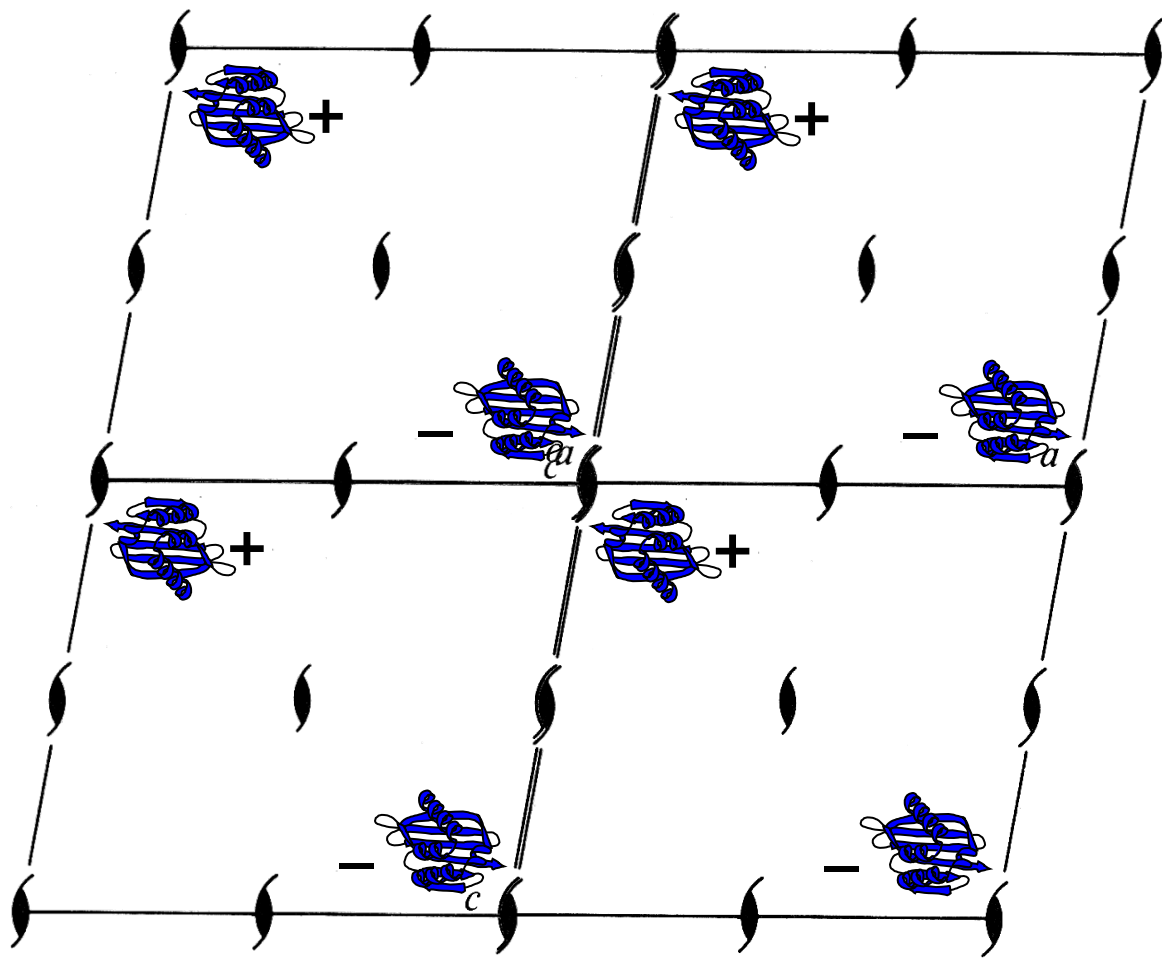
- By default, solutions over 75% of the difference between the top peak and the mean are selected

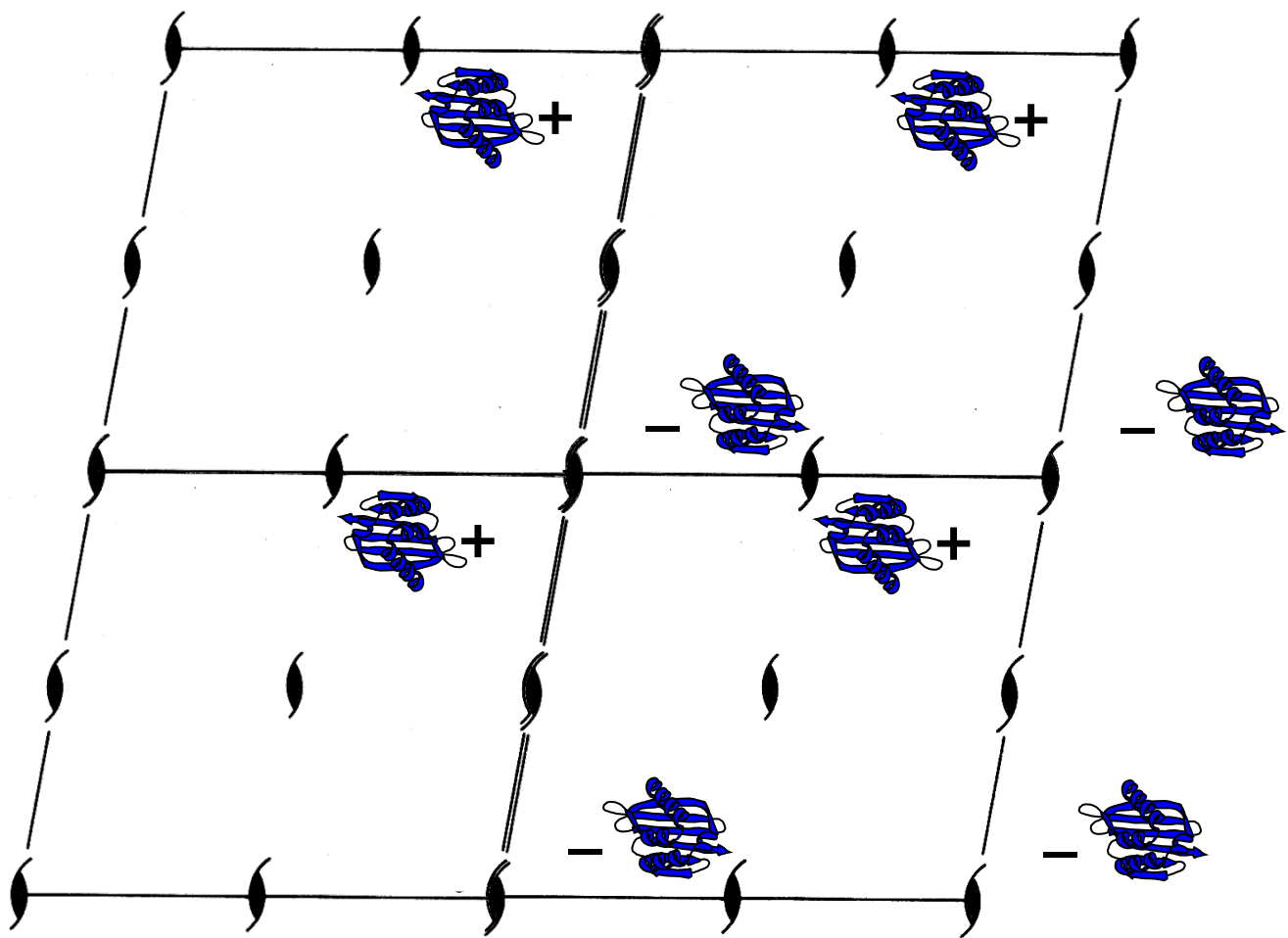
When is a model correctly placed?

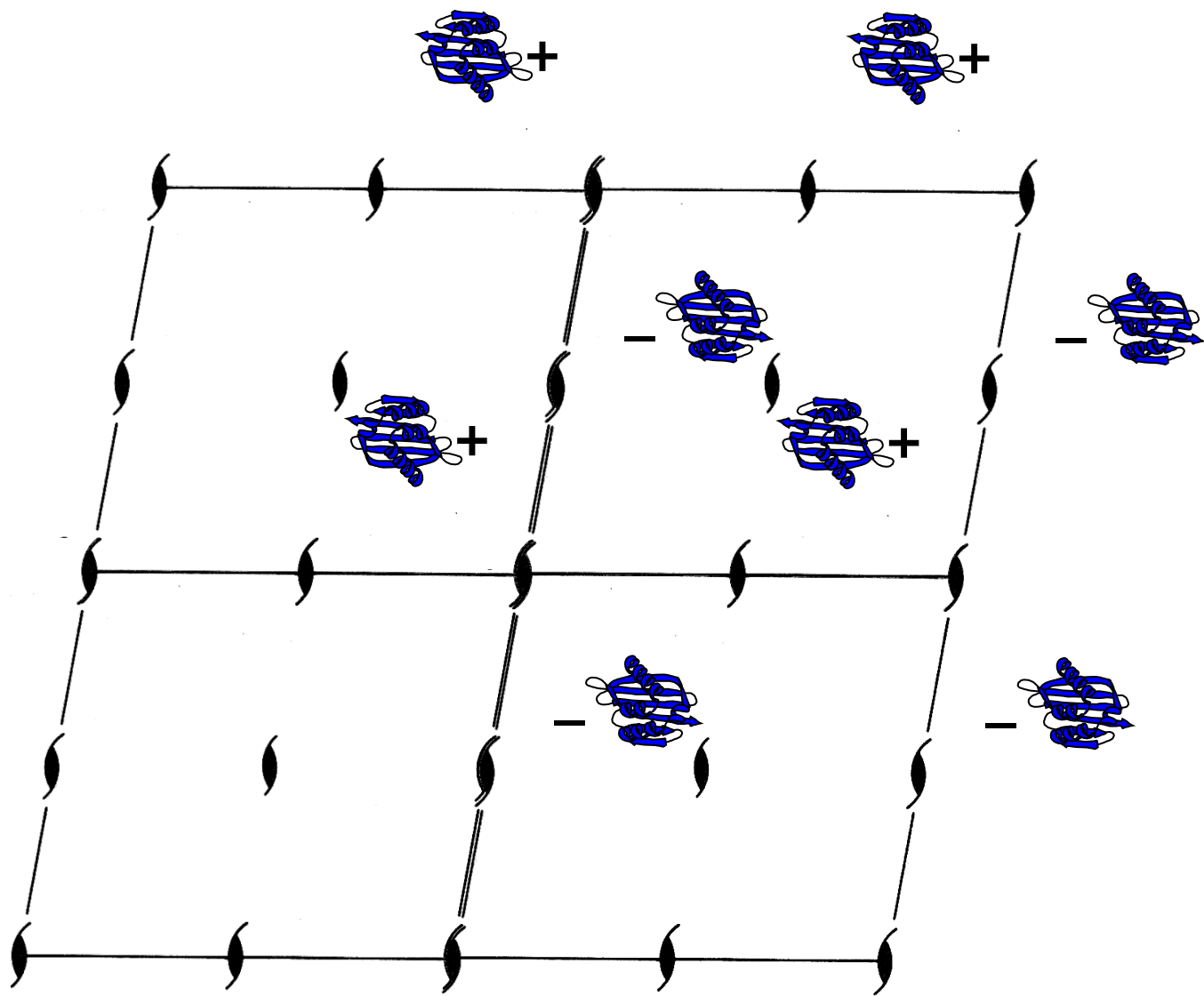
TF Z-score	Solved?
< 5	no
5 - 6	unlikely
6 - 7	possibly
7 - 8	probably
> 8	definitely

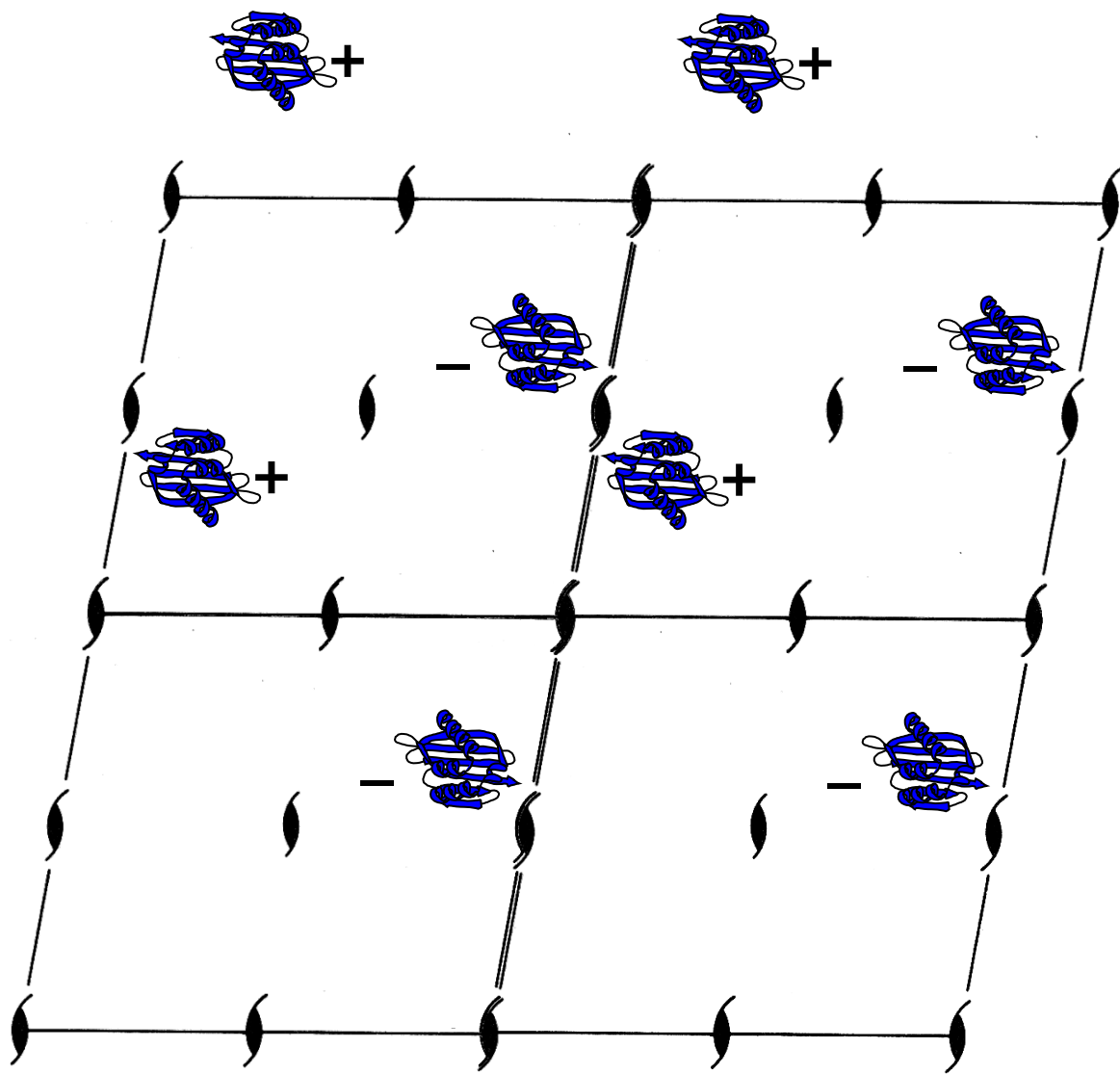
Origins

- Can only find the translation perpendicular to a rotation axis
 - no rotation symmetry, no translation to find!
 - If there are multiple symmetry axes of the same order of rotation in a plane then the translation can be defined with respect to any one of these
 - These are equivalent to different choices of origin
 - Different MR solutions may be on “different origins” and look different when they are really the same
 - PHASER SOLUTION TEMPLATE
-



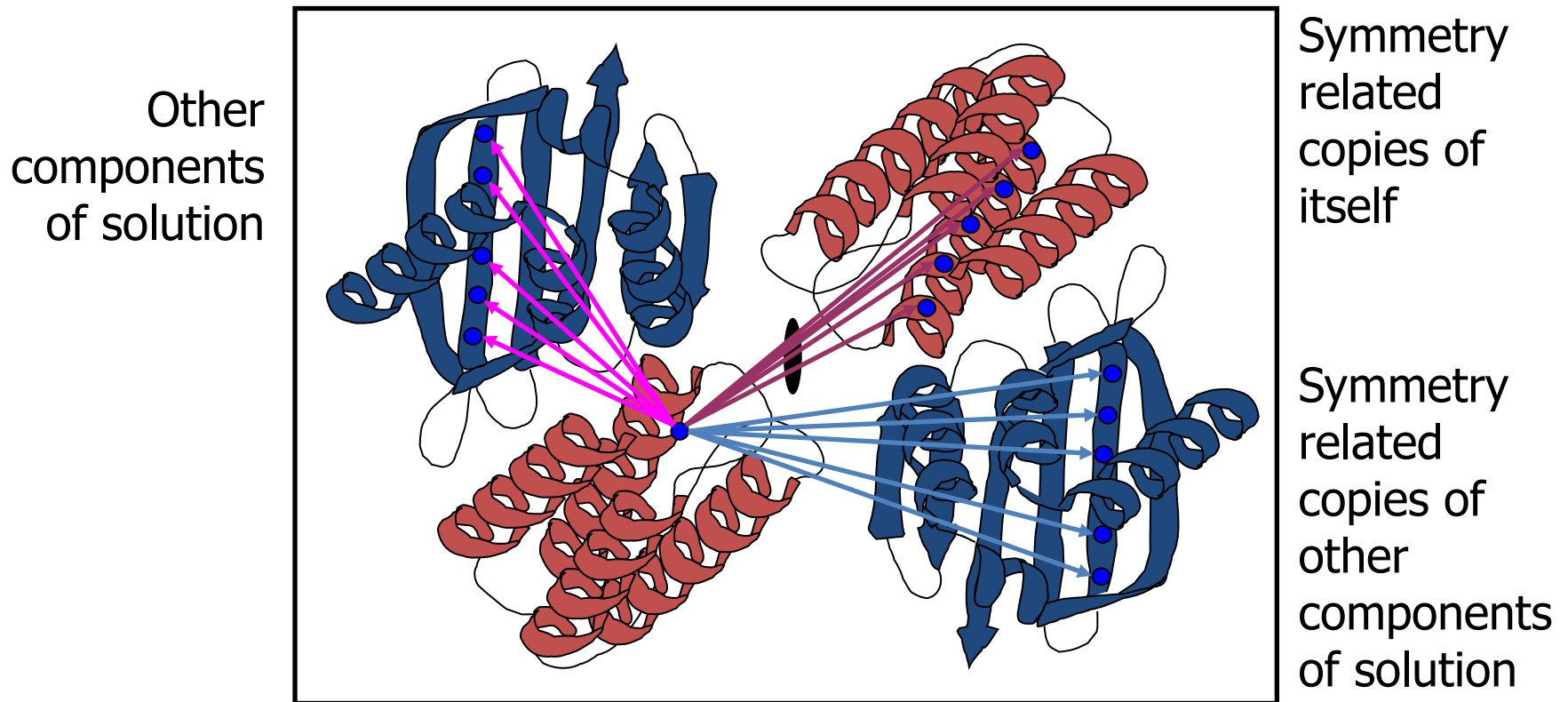






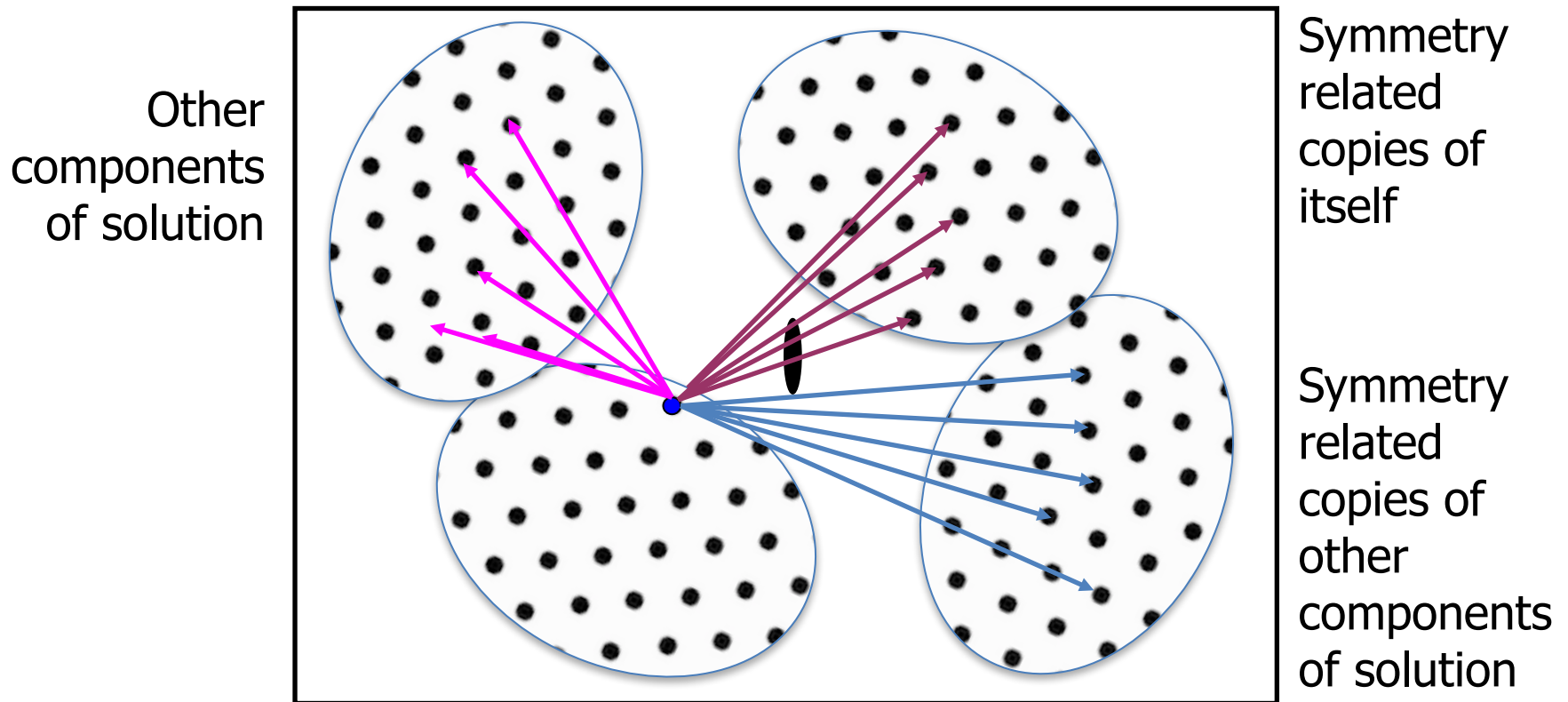
Packing Function

- α clash test



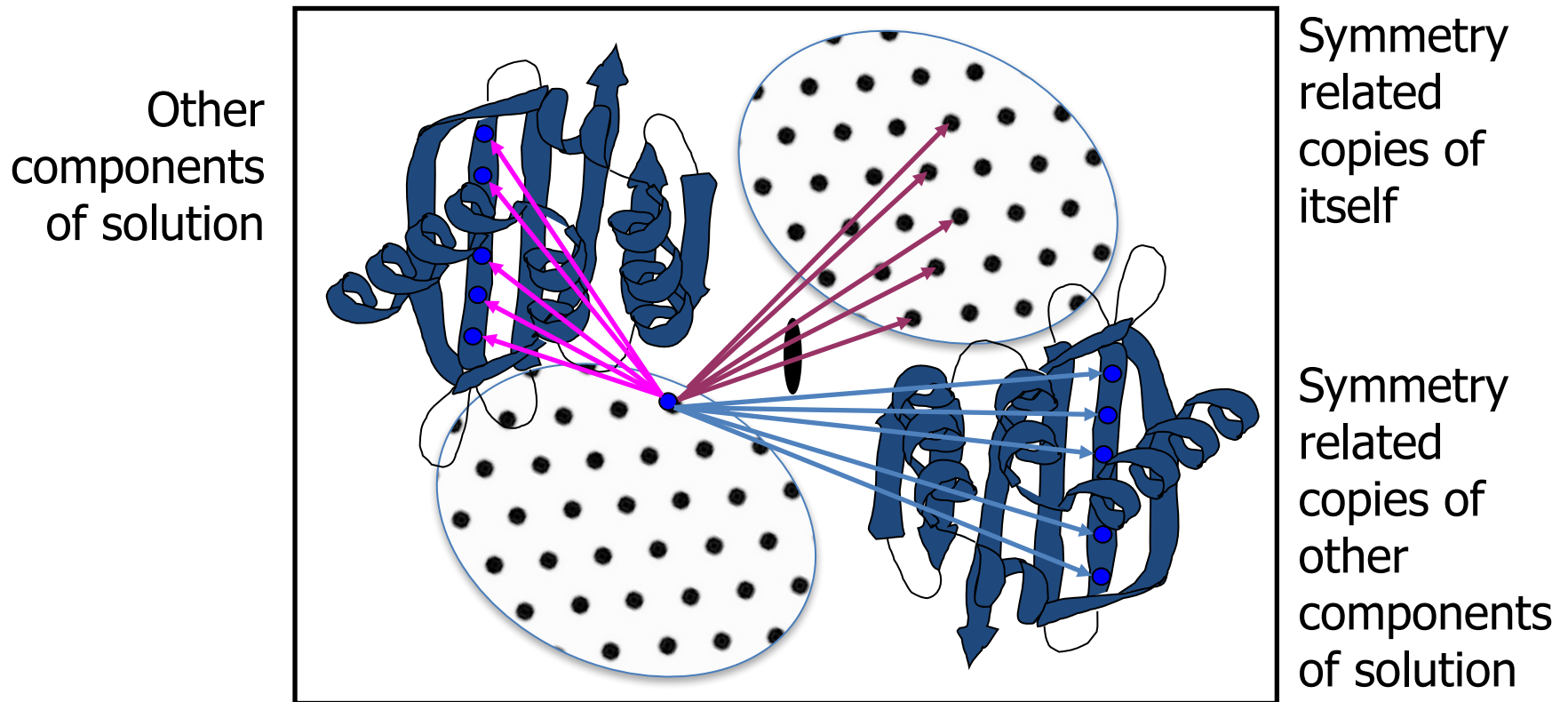
Packing Function

- Hexagonal Grid clash test



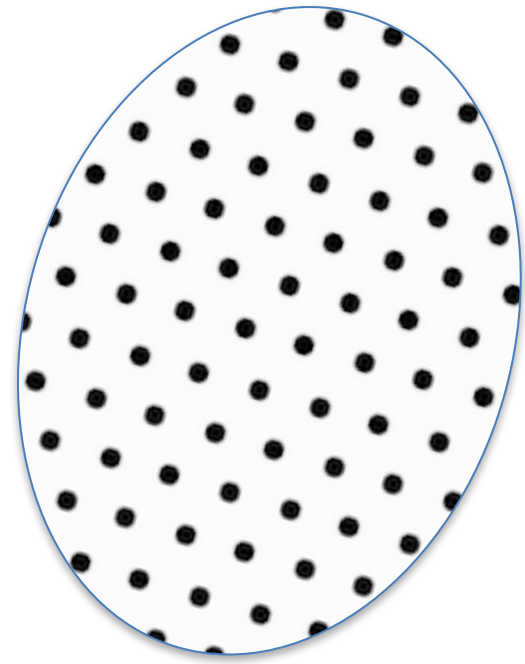
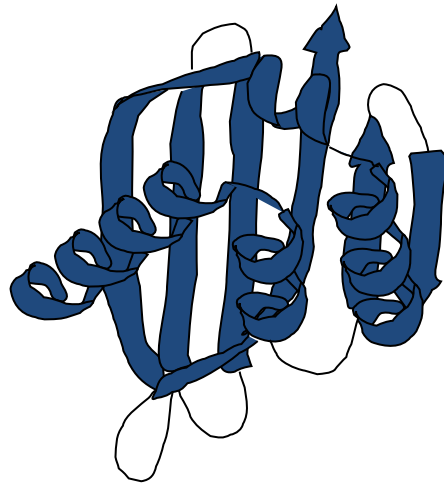
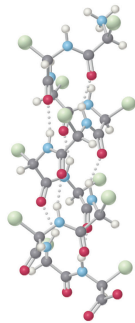
Packing Function

- Mixed Hexagonal Grid and C α clash test



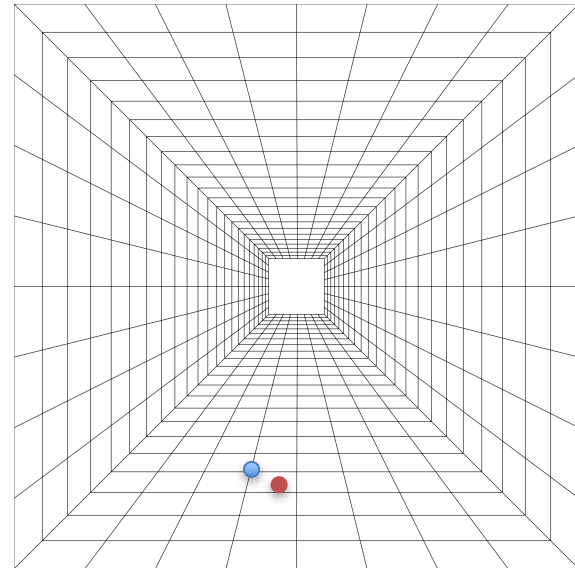
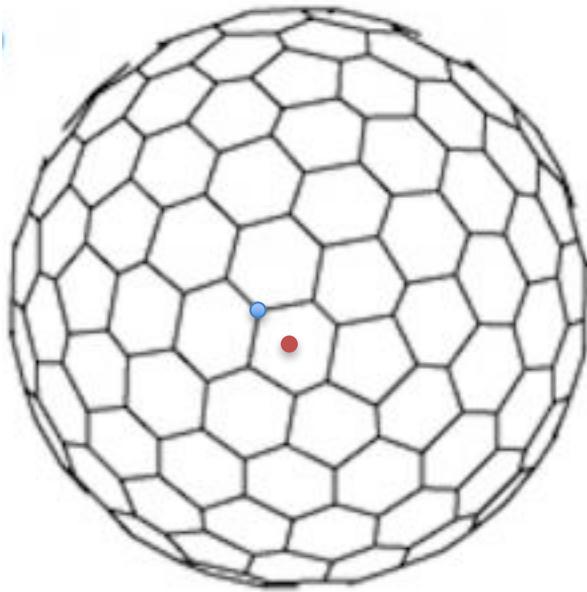
Packing Function

- Small – Medium – Large
- all atoms – C α atoms – Hexagonal Grid



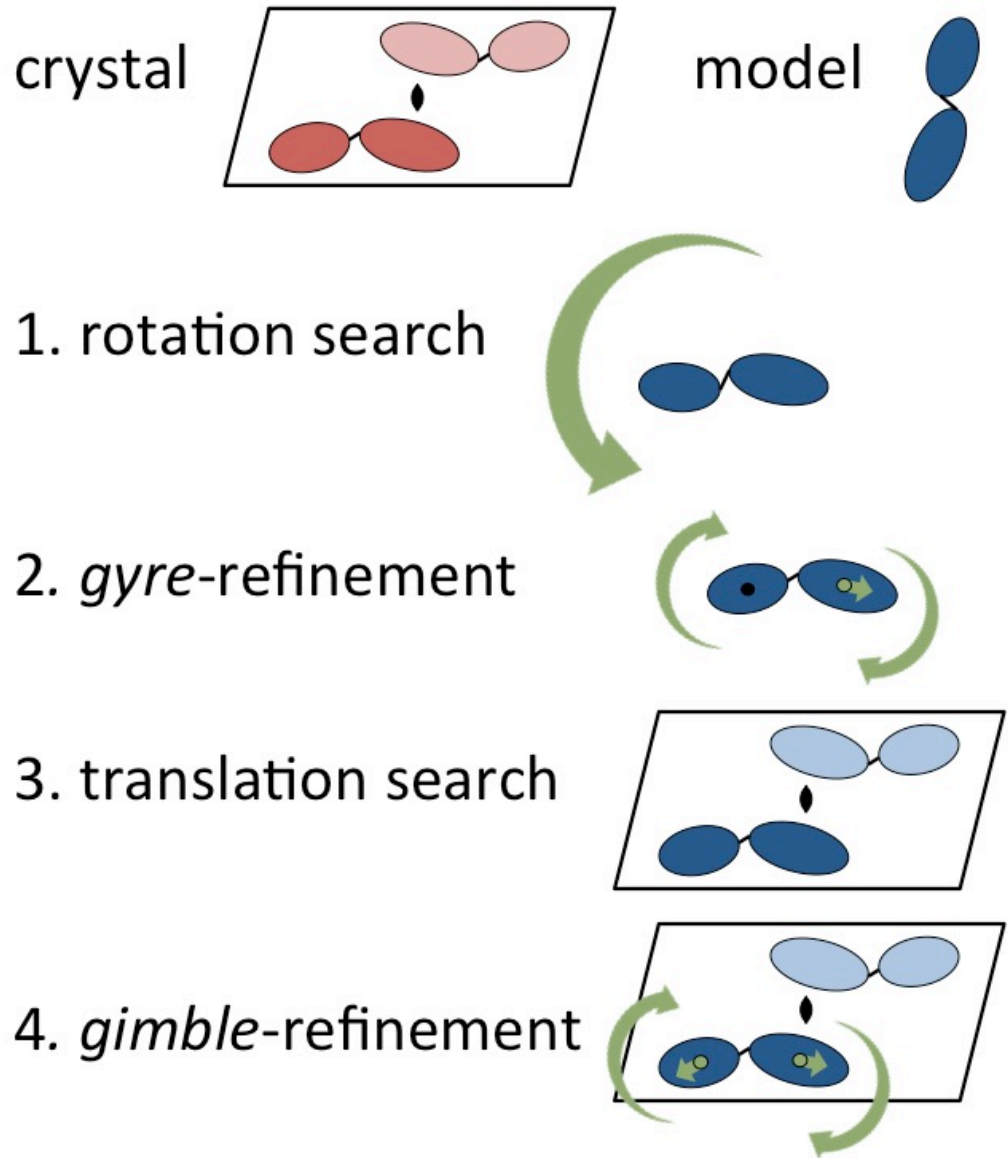
Refinement

- Rotation and Translation searches are scored on a grid



Gyre and Gimble

ML replacement for PC- refinement



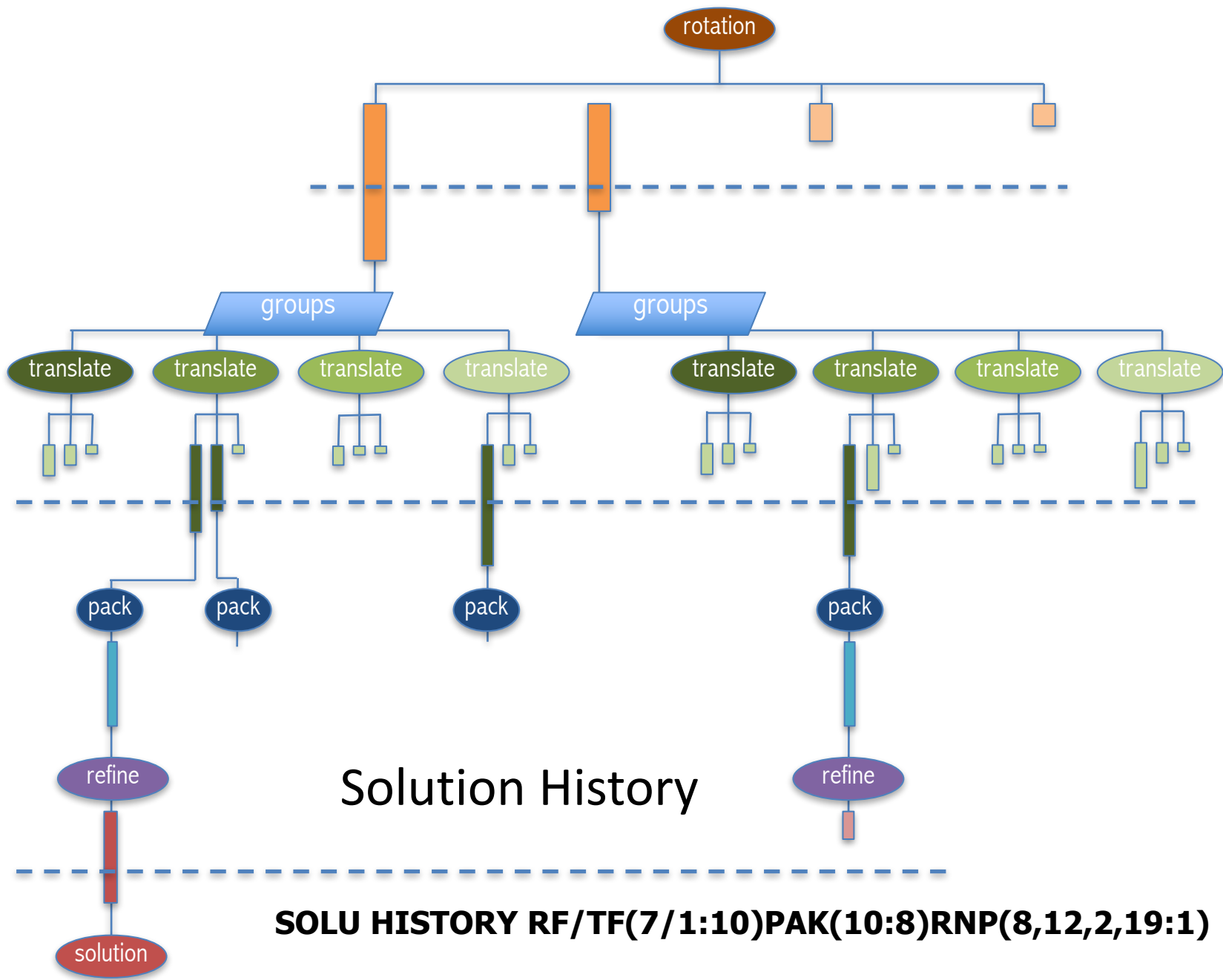


Figure: Martin Noble

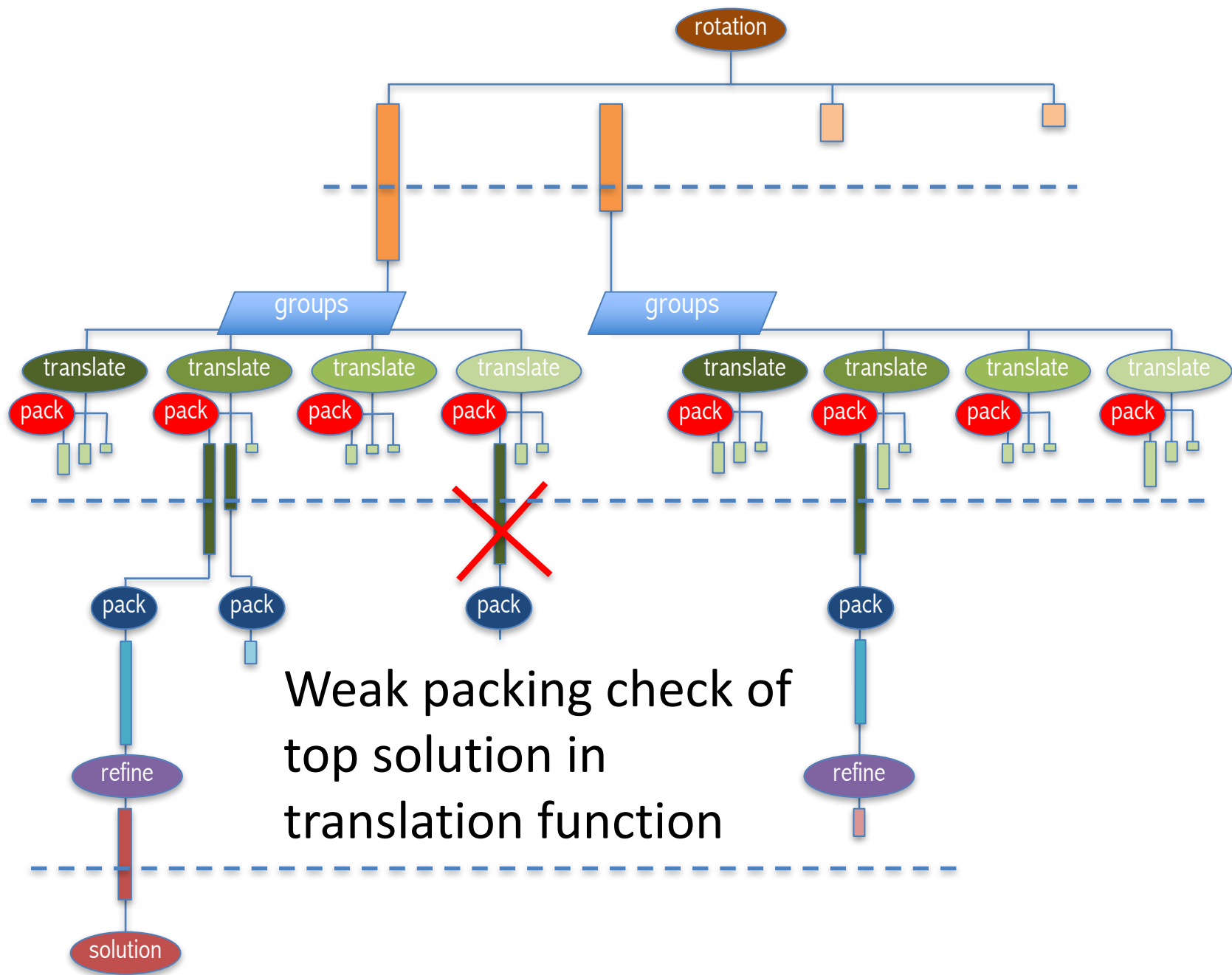


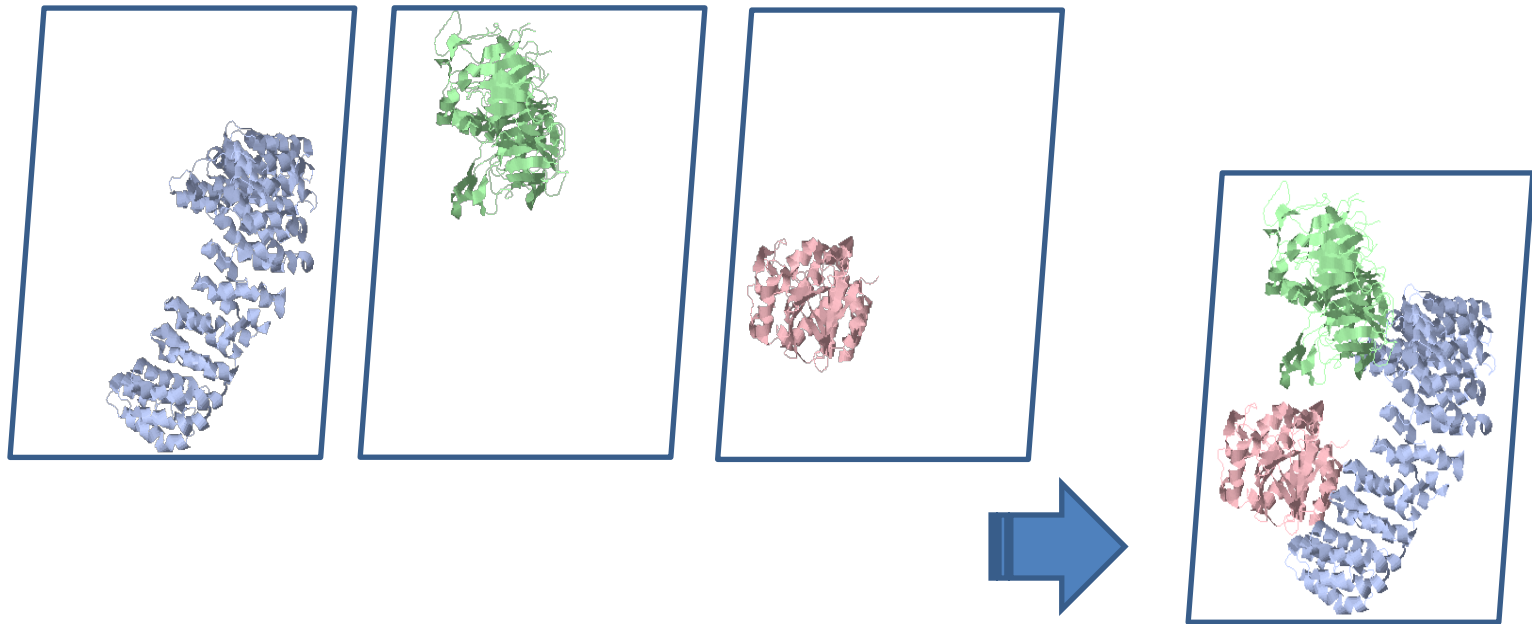
Figure: Martin Noble

Using Partial Structure

- The MLRF and MLTF can use models that have already been placed in the asymmetric unit
 - Patterson Rotation Function cannot account for placed models
 - The Patterson Correlation Coefficient Translation Function can account for known positions...
 - But most of the difficulty in MR is the rotation function, because the signal to noise ratio is much lower
-

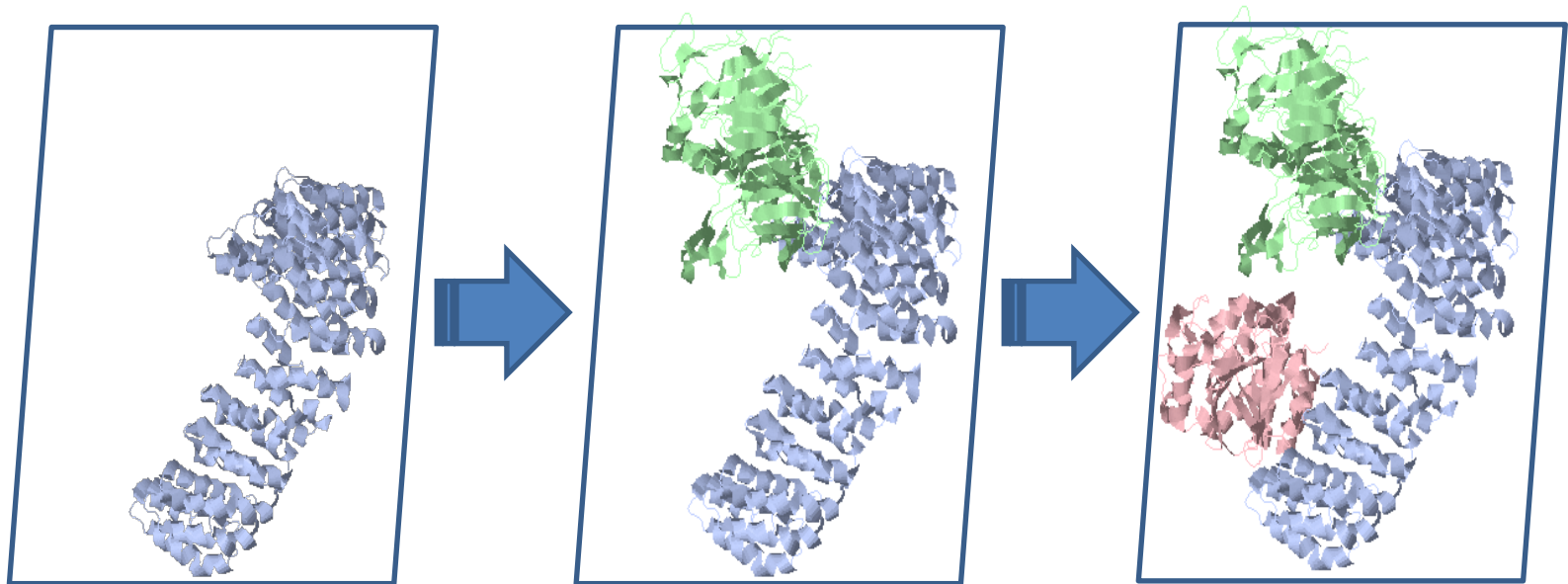
Searches for multiple components

- It may not be possible to find each component in a separate search

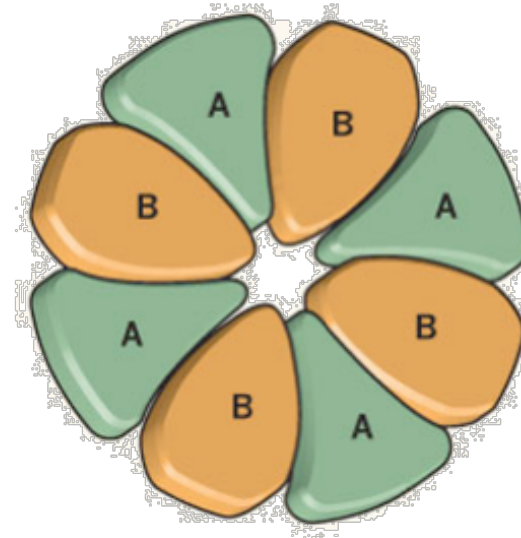
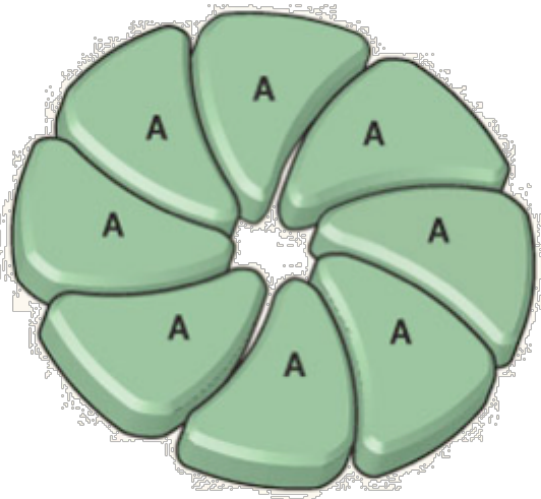


Searches for multiple components

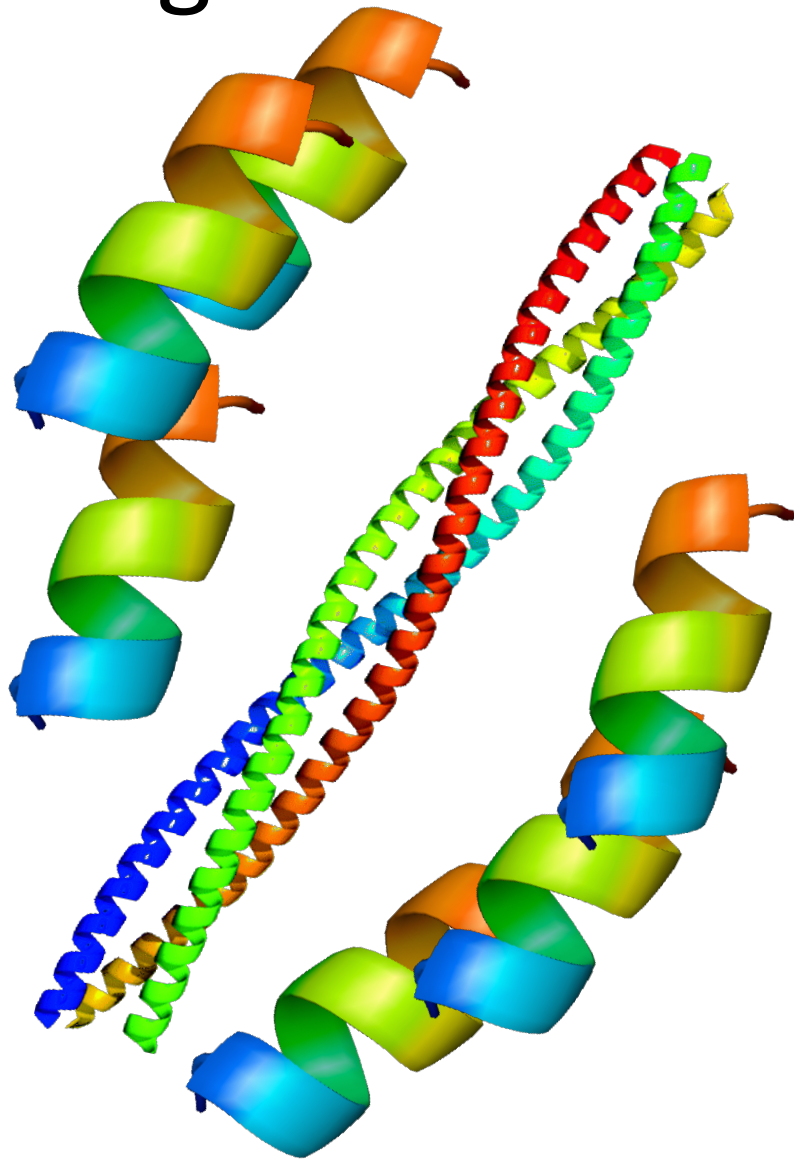
- ML includes partial structure information from previous placements
- Structure builds up by addition



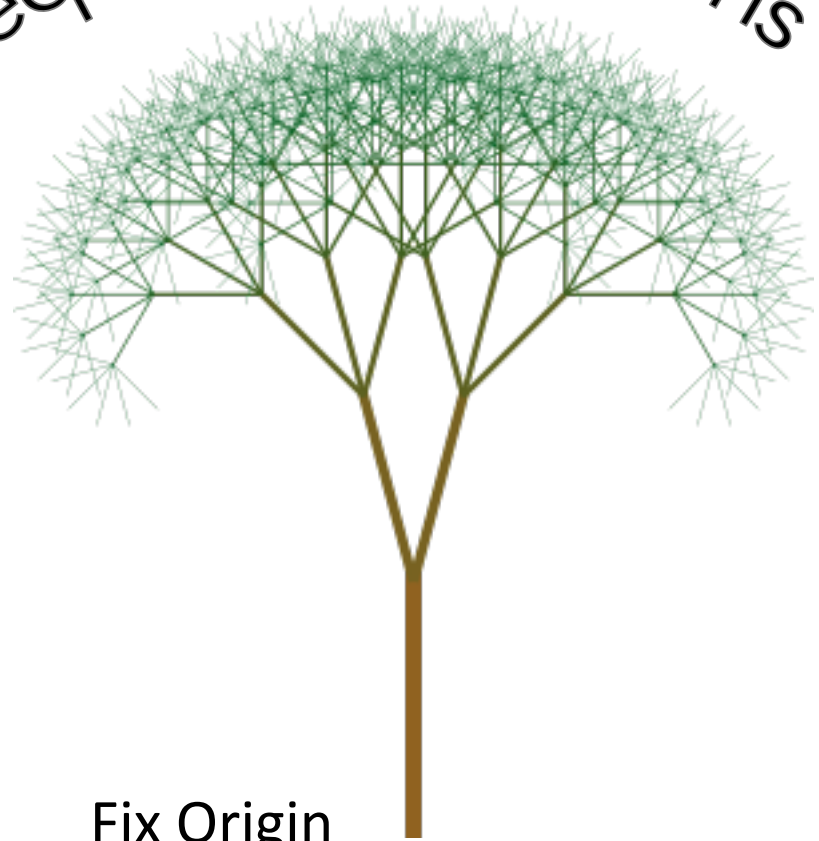
Multi-copy searches



Fragments and Tree Search

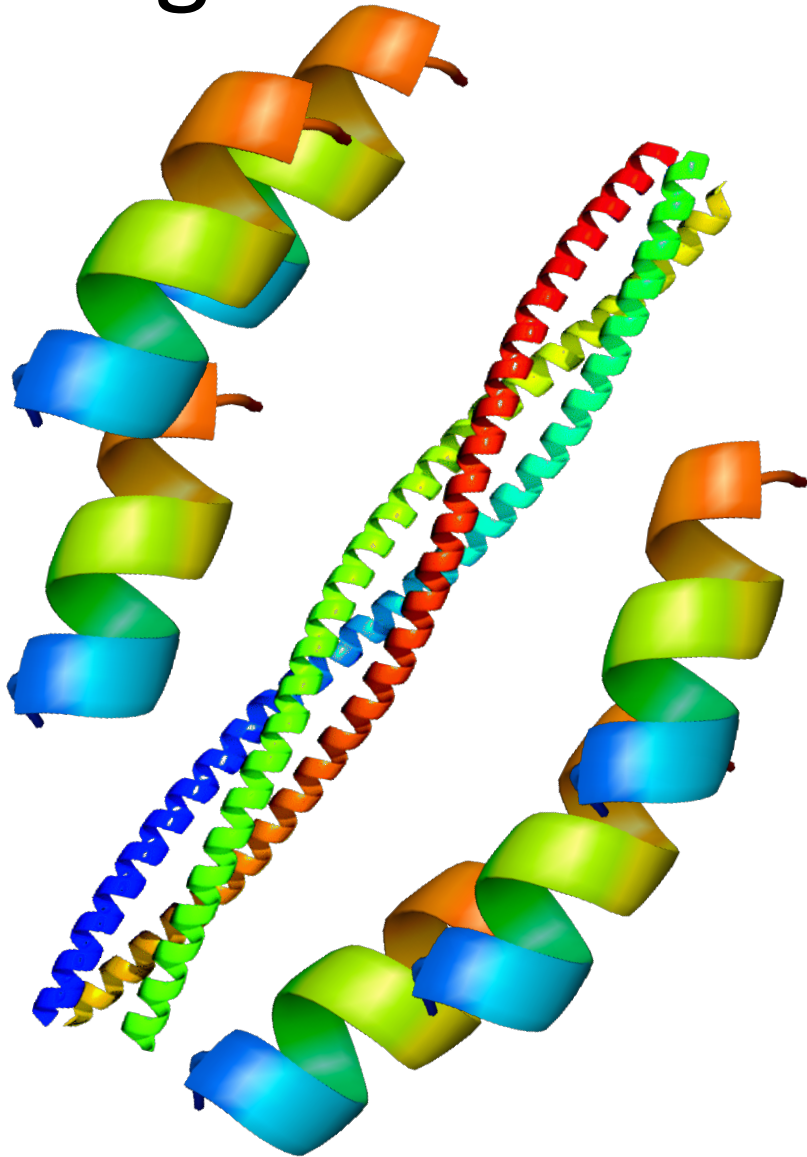


equivalent solutions



Fix Origin

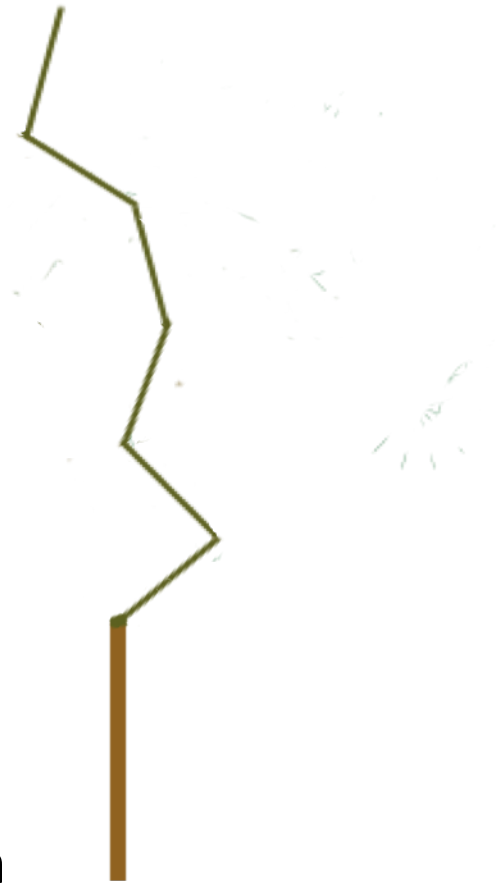
Fragments and Amalgamation



Amalgamation is possible because LLGI shows which intermediate structures are correct

one solution

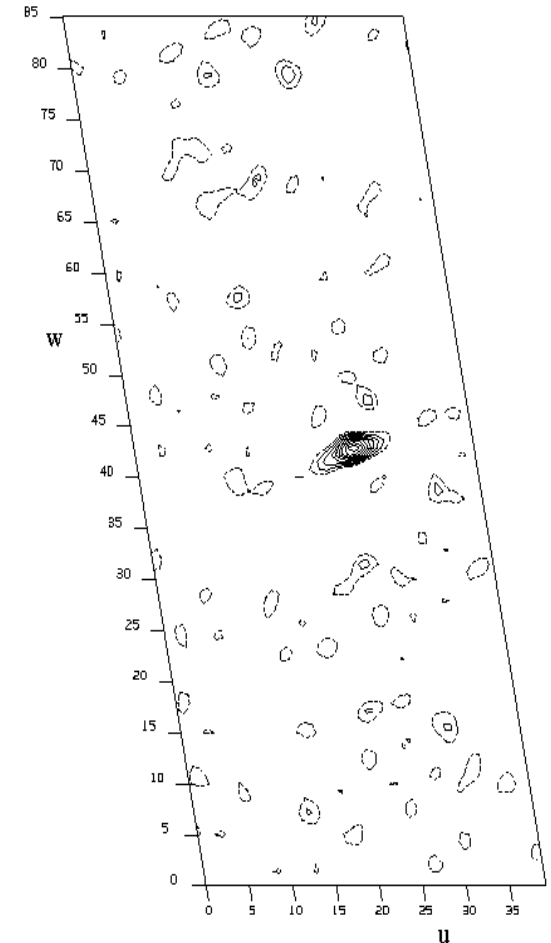
Fix Origin



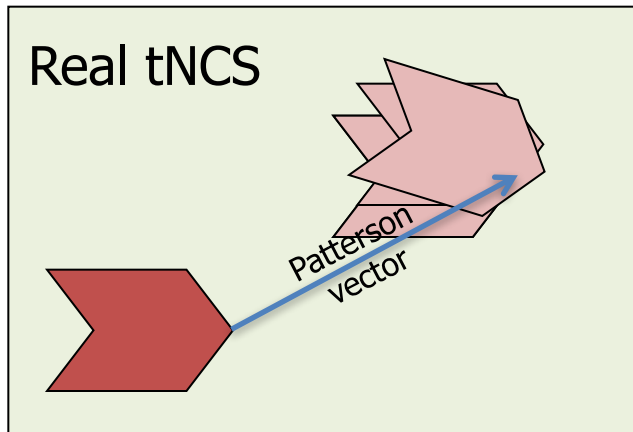
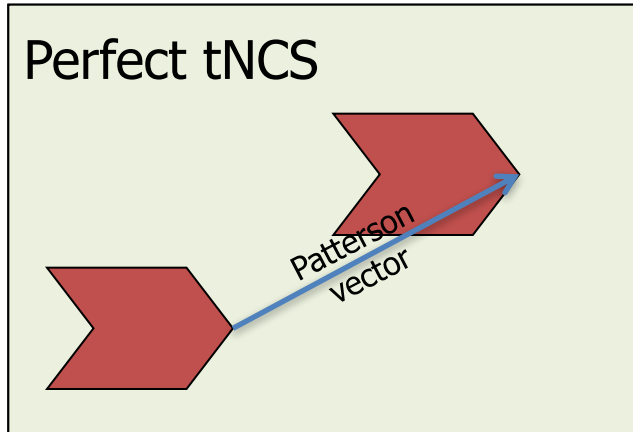
Translational NCS

- If tNCS is not accounted for then $TFZ > 8$ does not indicate a correct placement
 - TFZ values are always higher
 - $TFZ > 12$ can be wrong
- When TFZ is accounted for the TFZ values are those expected of data without tNCS

Native Patterson of mouse renin.



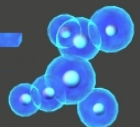
Translational NCS



- Three tNCS parameters refined from data alone
 1. Rotation
 2. Translation
 3. RMSD between copies
- tNCS correction factors used in MR and SAD
- Two classes of tNCS cases accounted for
 - These cover the majority of cases

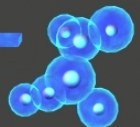
Translational NCS

- Pairs of molecules related by one vector
 - One peak in Patterson
 - Molecules in pairs
 - There can be any number of pairs of molecules related by the same tNCS vector
- Molecules related by multiples of one vector
 - Peaks in Patterson are multiples of same vector
 - Molecules in sets related by same vector
 - There can be any number of sets of molecules related by the same tNCS vector



Twin Detection

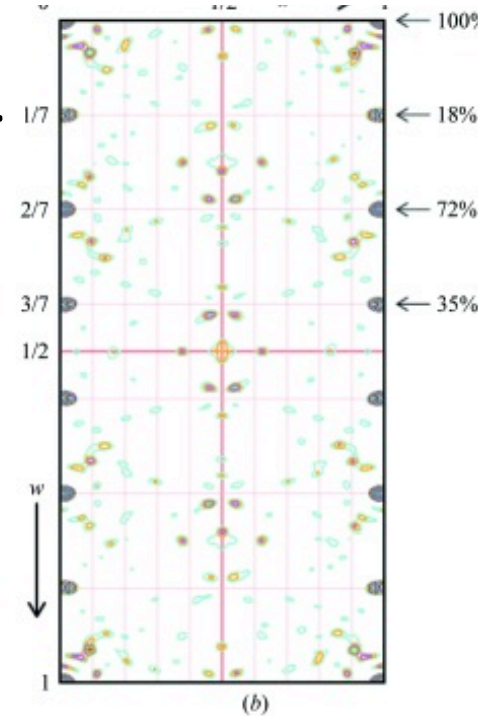
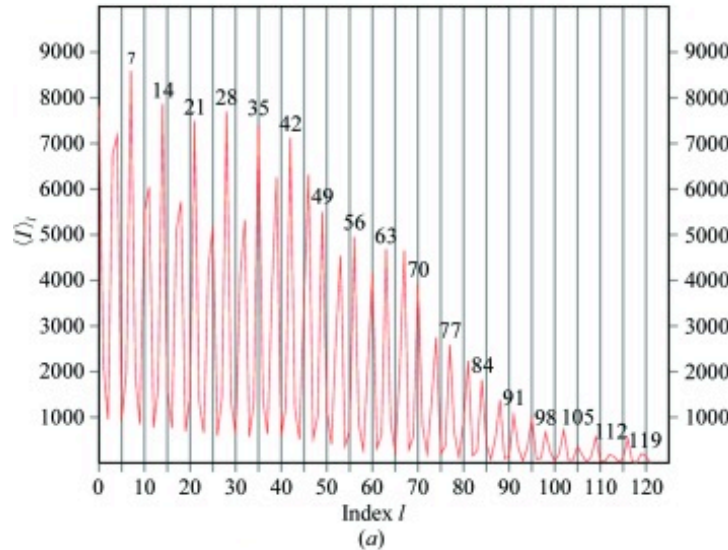
- Translational NCS masks twinning
 - Has opposite effect on intensity statistics
- Correcting the data for tNCS unmask twinning
- Phaser generates cumulative intensity plots for centric and acentric reflections after correction for tNCS and anisotropy
- Phaser gives a P-value for there being twinning in the presence of tNCS



Likelihood-based molecular-replacement solution for a highly pathological crystal with tetartohedral twinning and sevenfold tNCS

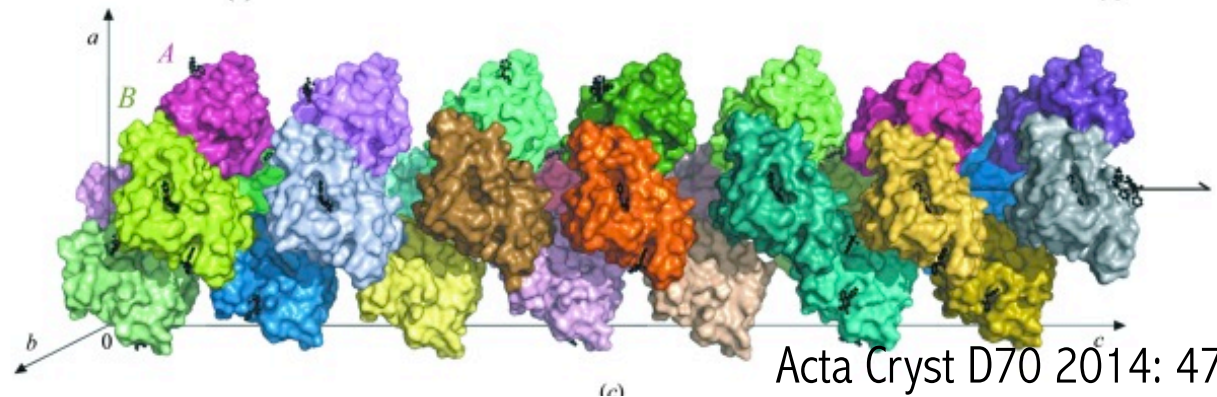
Sliwiak J, Jaskolski M, Dauter Z, McCoy AJ, Read RJ.

Repeats $7/2$
along c^*



Solved finding 56 copies
of the monomer in P1

28 copies in C2
(true space group)



Acta Cryst D70 2014: 471–480.

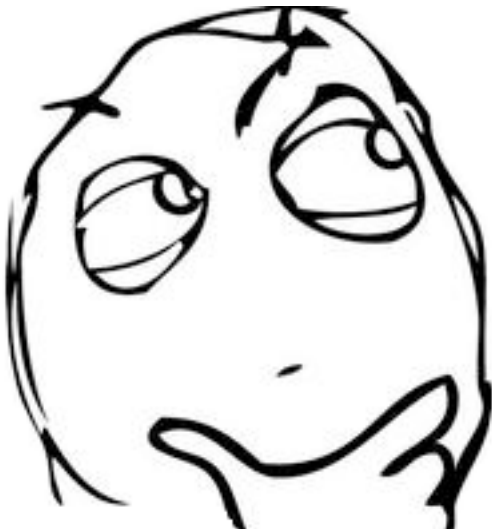


Will MR work
for me?

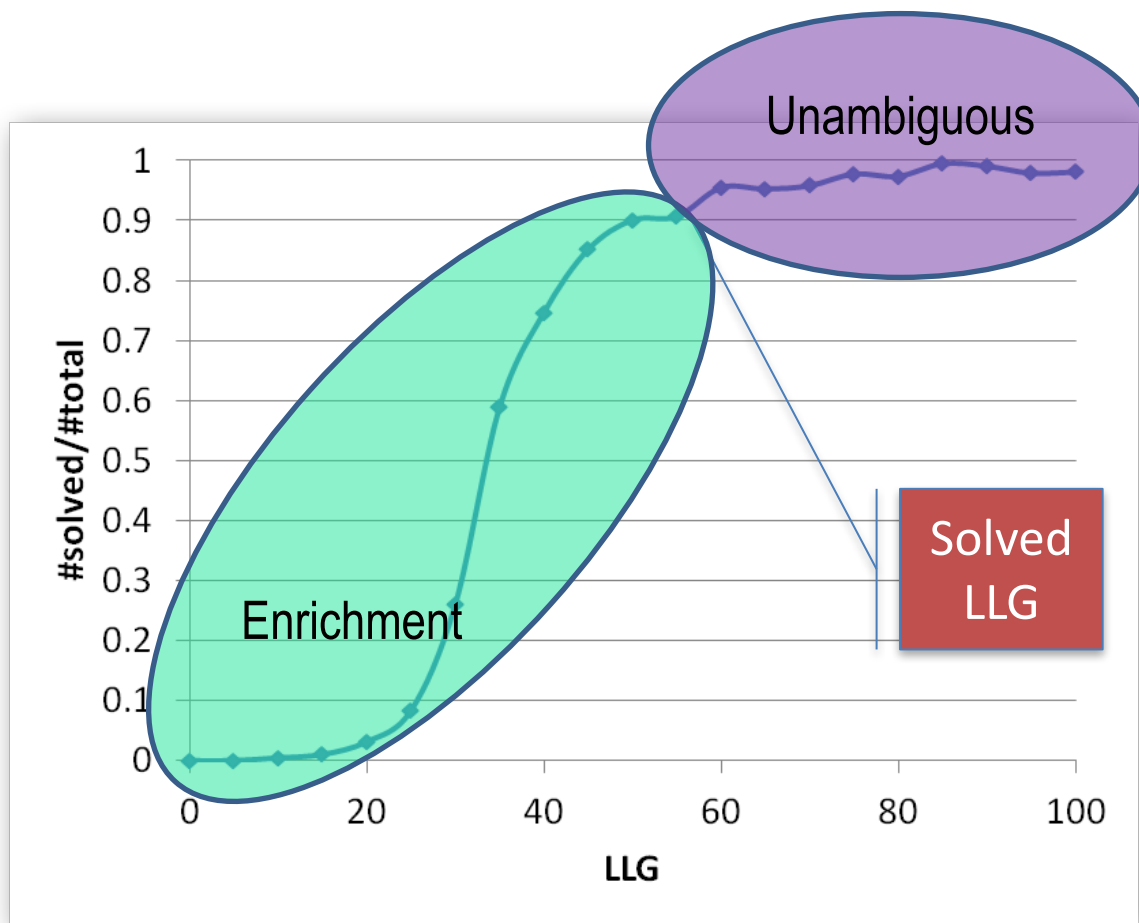
- Is my model good enough?
- Is my data good enough?
- Do I need to place multiple models simultaneously to get a signal?
- Will fragment-based MR work?
- Will α -helices work?
- How big does my helix/fragment have to be?
- Will single-atom MR work?



First, find the
criteria for deciding
that MR has
worked!



Final LLG for MR solutions



Database of
over 23000
MR problems

Plot of LLG versus success in structure solution

R.D. Oeffner

When is a model correctly placed?

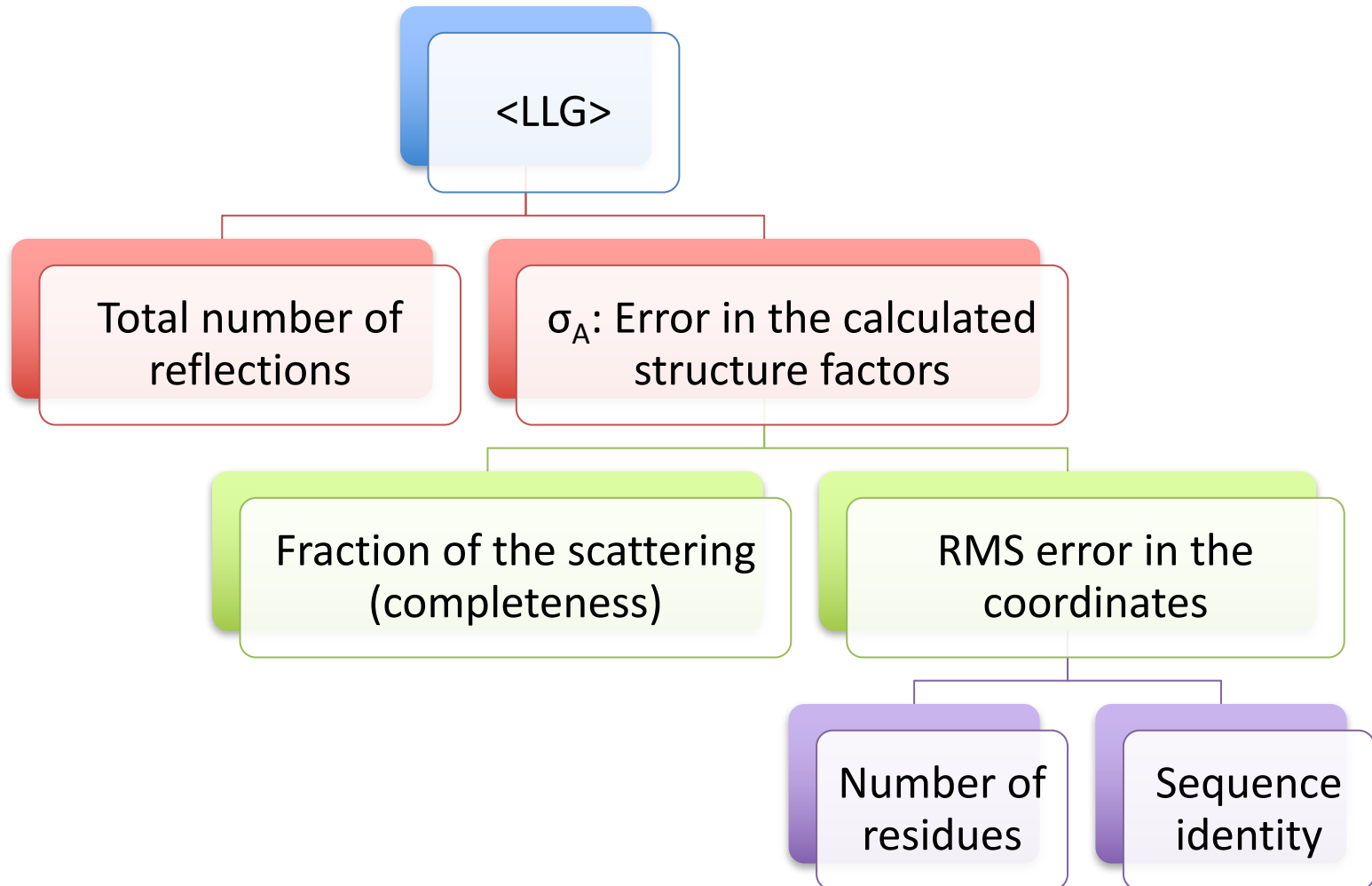
TF Z-score	LLG score	Solved?
< 5	< 25	no
5 - 6	25 - 36	unlikely
6 - 7	36 - 49	possibly
7 - 8	49 - 64	probably
> 8	> 64	definitely

Predicting LLG of solution

- So if you can predict the LLG...
 - You know how easy/difficult will be MR
 - You can prioritize structure solution strategies
- Removes uncertainty in MR
 - Knowing when to start/stop has always been the problem with MR

You can minimize the time to structure solution

Expected LLG



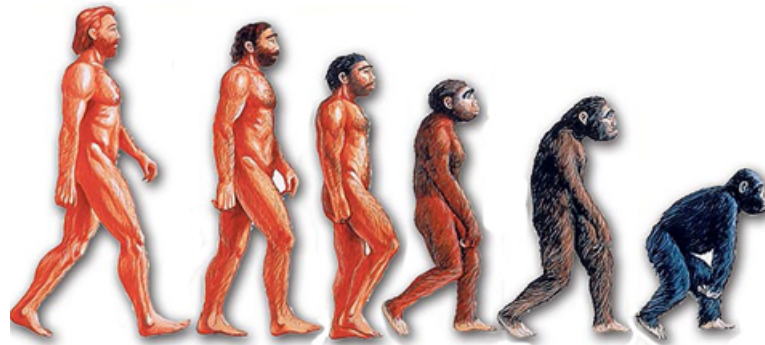
MR with Fragment/Atom

- Fragments or even atoms are just models with low completeness
 - Low σ_A
- Success of fragment/atom based MR relies on other contributions to the $\langle \text{LLG} \rangle$ being favourable
 - Lots of reflections
 - Low RMS

Success does NOT depend on

- The resolution (strictly)
- The type of fragment

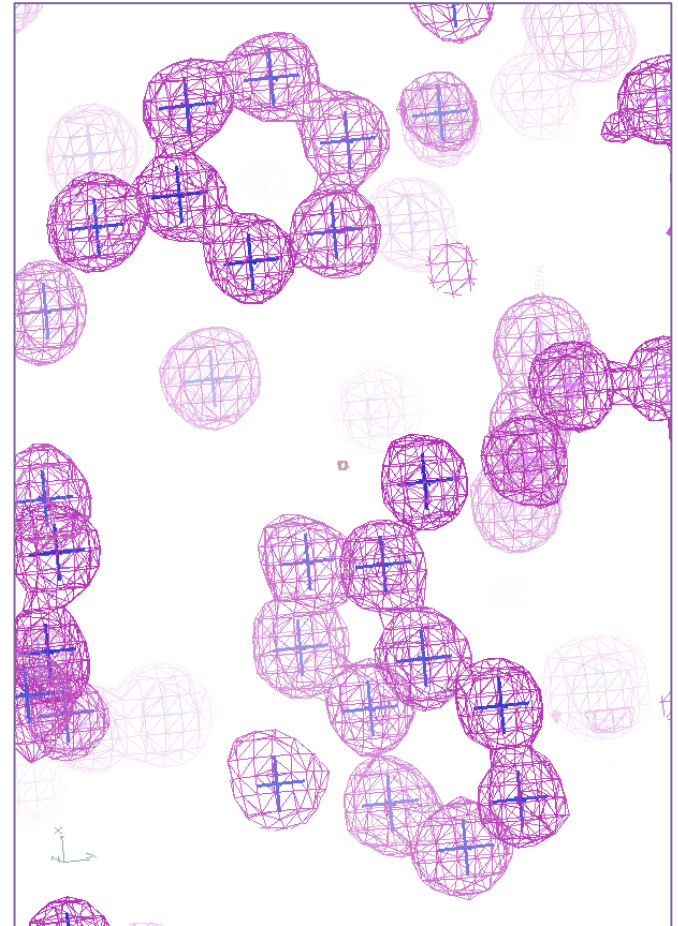
Devolution



- ↓ Homologous structures
 - ↓ Domains
 - ↓ Ab Initio models
 - ↓ Fragments
 - ↓ Helices
 - Atoms
-

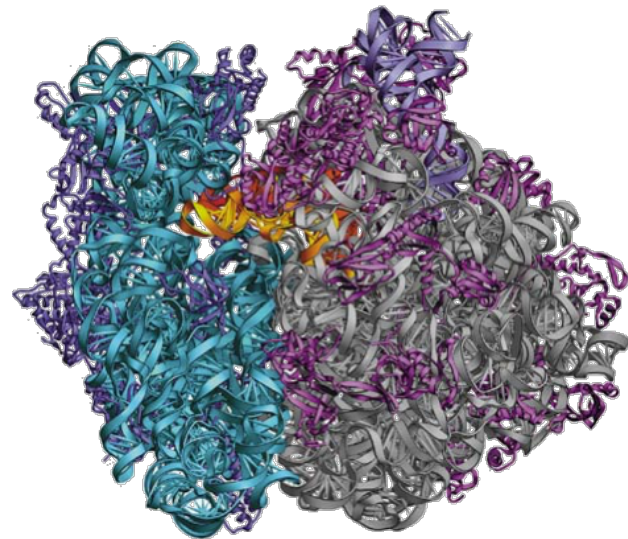
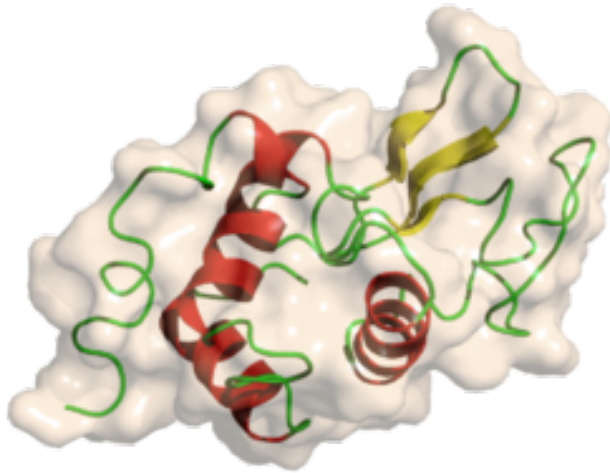
Single Atom MR

- Aldose Reductase
- 36 kDa, 0.78Å resolution
- 2525 non-H atoms in structure
- $\langle \text{LLG} \rangle$ for first N ≈ 0.1
 - For first S ≈ 4
 - For first S ≈ 16 if B-factor is $2 \text{ \AA}^2 < \text{Wilson B}$



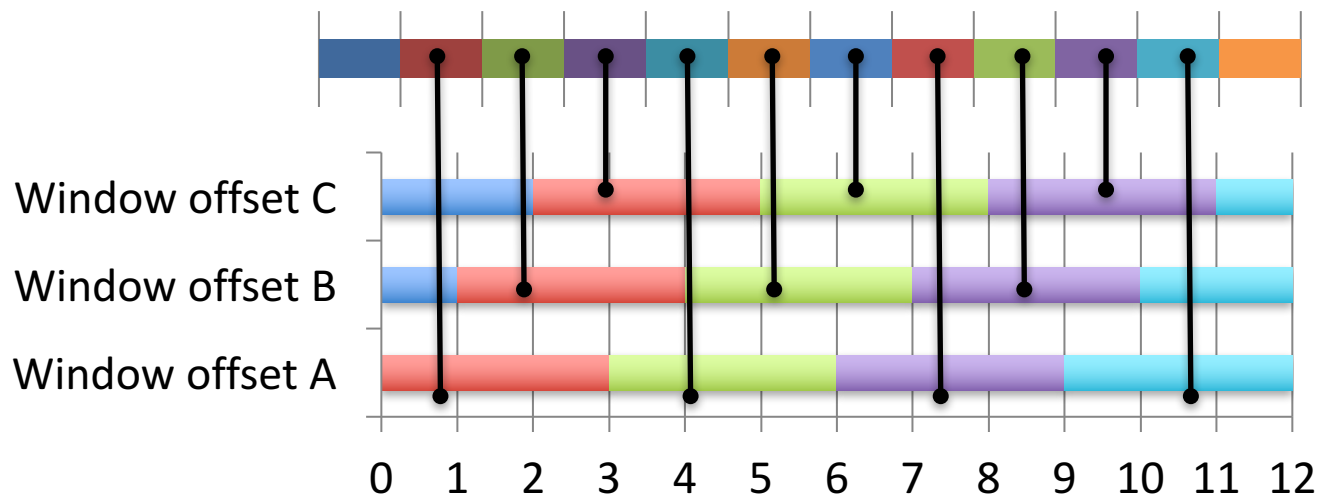
<LLG> and Resolution

Example	Data	<LLG> target	resolution
HEWL	1.9 Å	120	5.6 Å
Ribosome	3.6 Å	120	10.8 Å



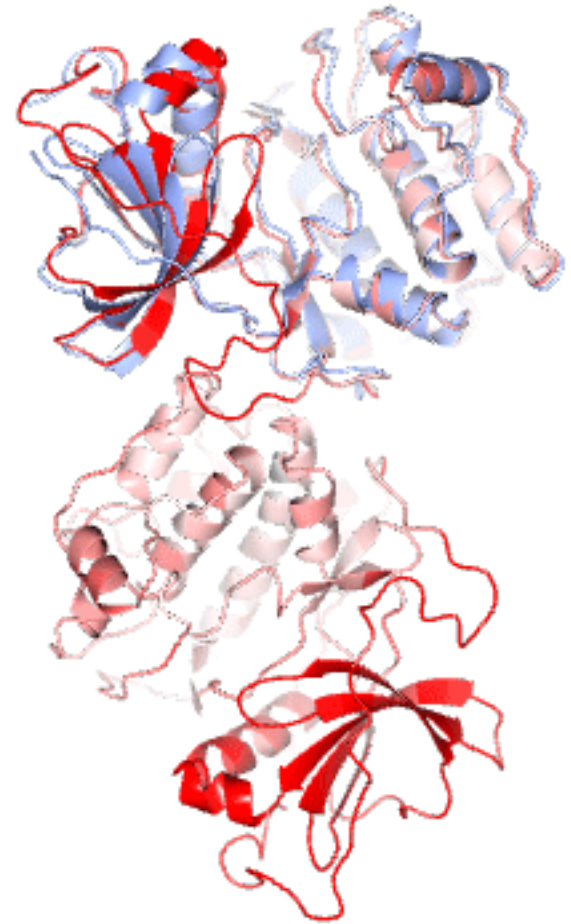
<LLG> Pruning

- Occupy refinement would normally over-parameterize a model
- eLLG can be used to determine number of atoms for significant change in LLG



<LLG> Pruning

- C-Amp-Protein K
(1ctp solved with 1atp)
- Clash due to
conformational change
- eLLG guided pruning
removes sections of
smaller domain
- Pruned structure passes
packing tests



After MR



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)

Find best
model

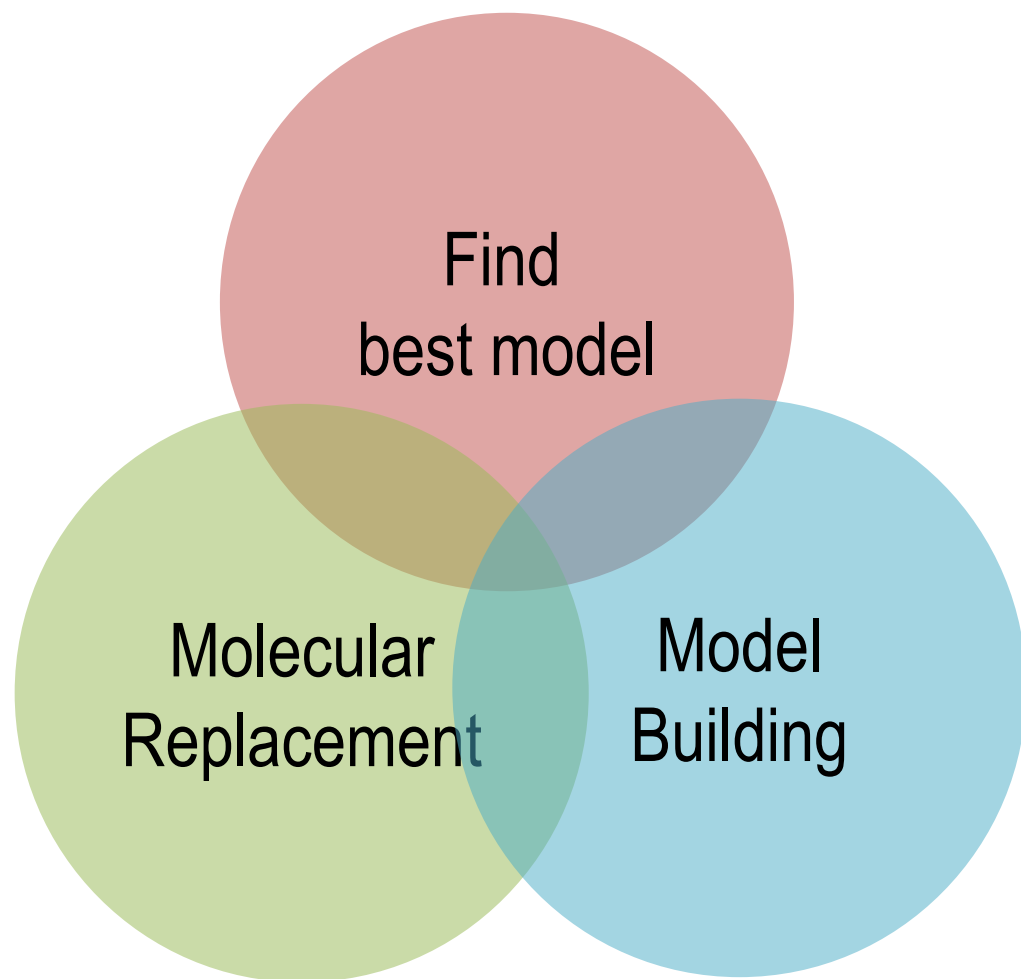
Molecular
replacement

Model
building



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



phenix.mr_rosetta

- Find MR solutions with Phaser, rebuild them with ROSETTA using techniques from ab initio modelling (ROSETTA energy term) to bring the structures within the radius of convergence of standard rebuilding/refinement in phenix.autobuild
- Correct solution must be in list passed to ROSETTA
 - phenix.mr_rosetta takes top 5 by default, regardless
 - relies on enrichment
- Rebuilding in ROSETTA includes map information through a density term in the ROSETTA energy

Increasing the Radius of Convergence of Molecular Replacement by
Density and Energy Guided Protein Structure Optimization
DiMaio et al (2011) Nature, 473, 540-543.

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



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

Cambridge University



Duke University

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



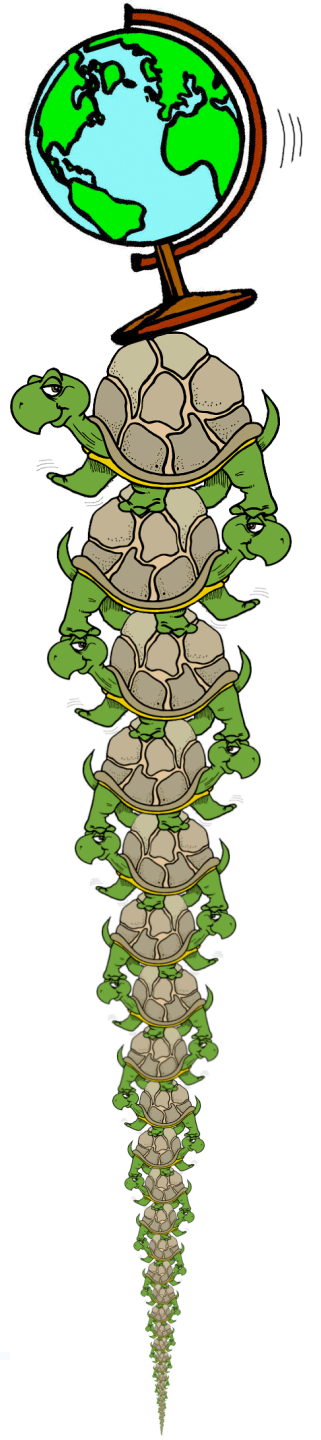
***An NIH/NIGMS funded
Program Project***

MR-SAD Phasing

Airlie McCoy

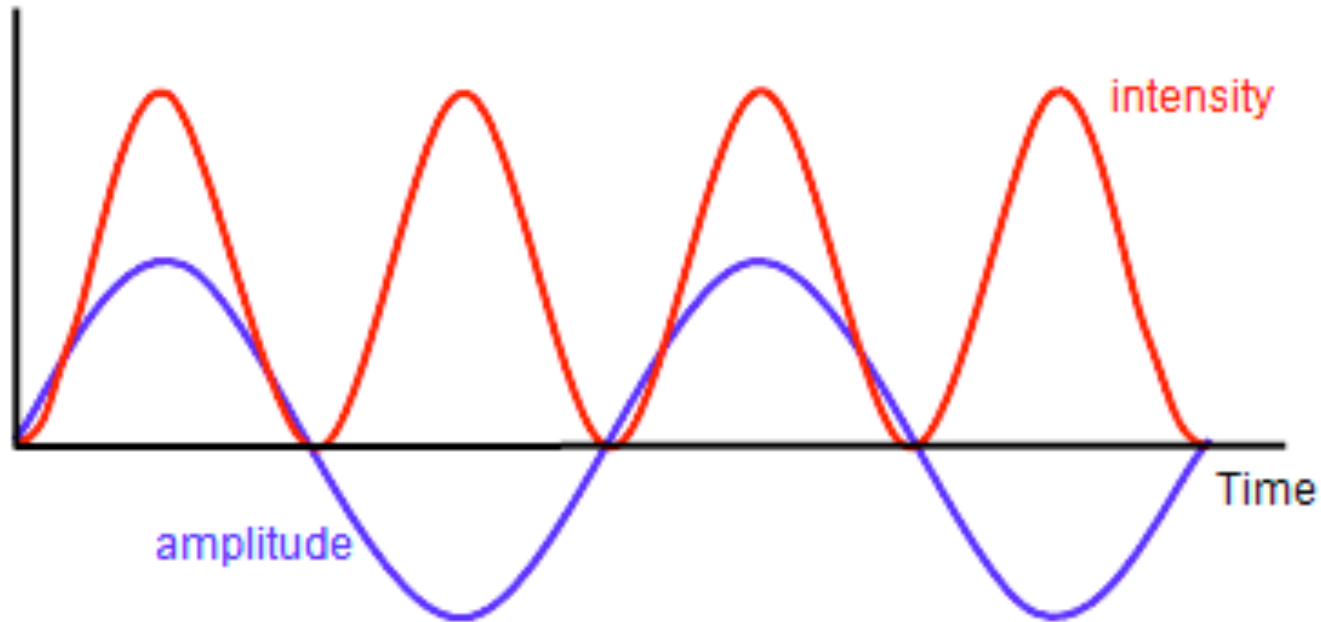


UNIVERSITY OF
CAMBRIDGE



Interference

- X-ray detectors only detect X-ray intensities

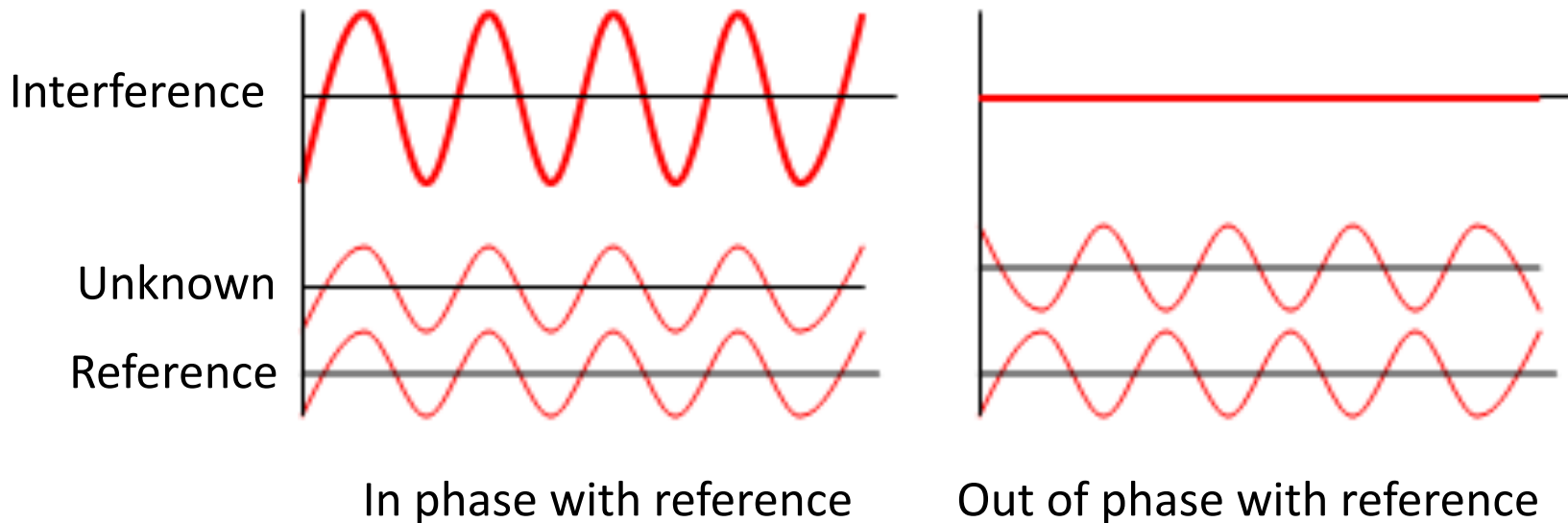


$$E \propto A^2$$

because they respond to the energy of the wave

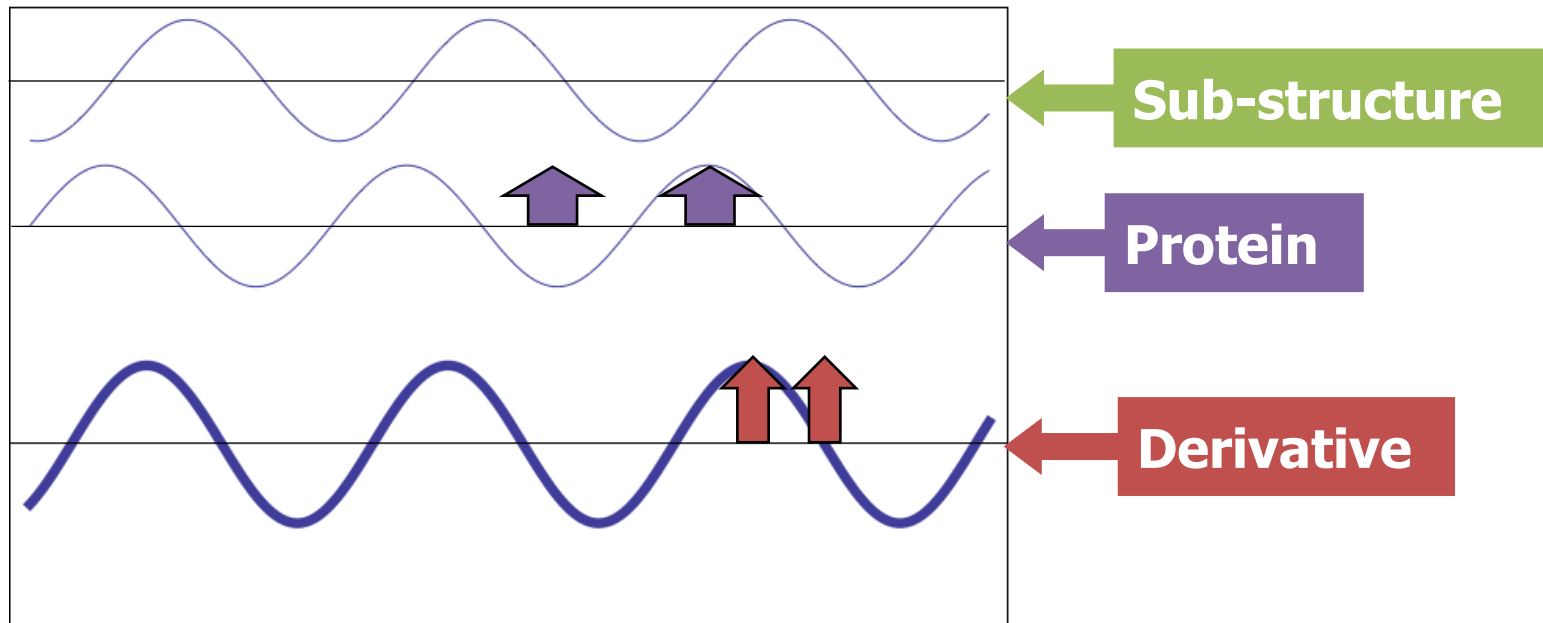
Reference Wave

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

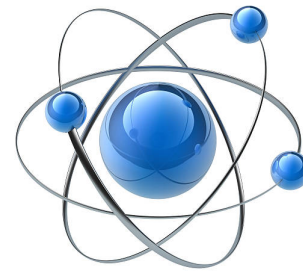


Interference and Phase

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



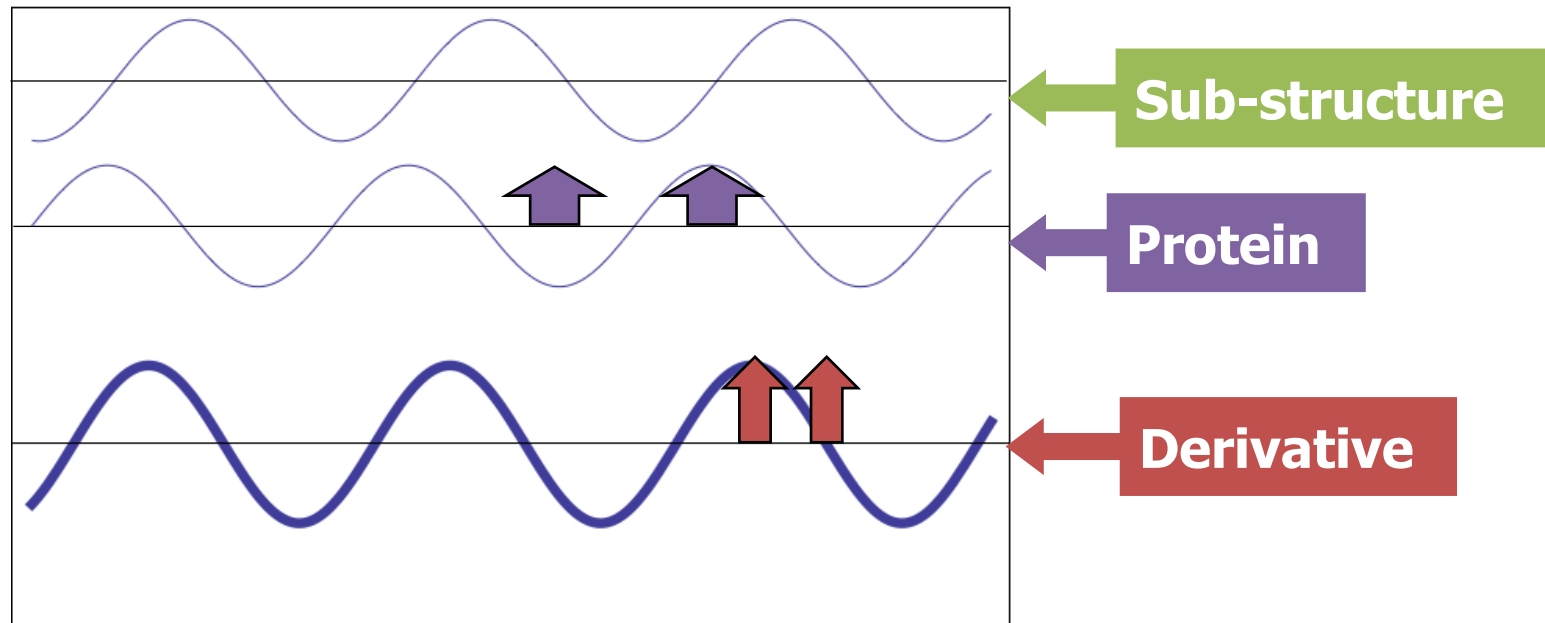
Phassade



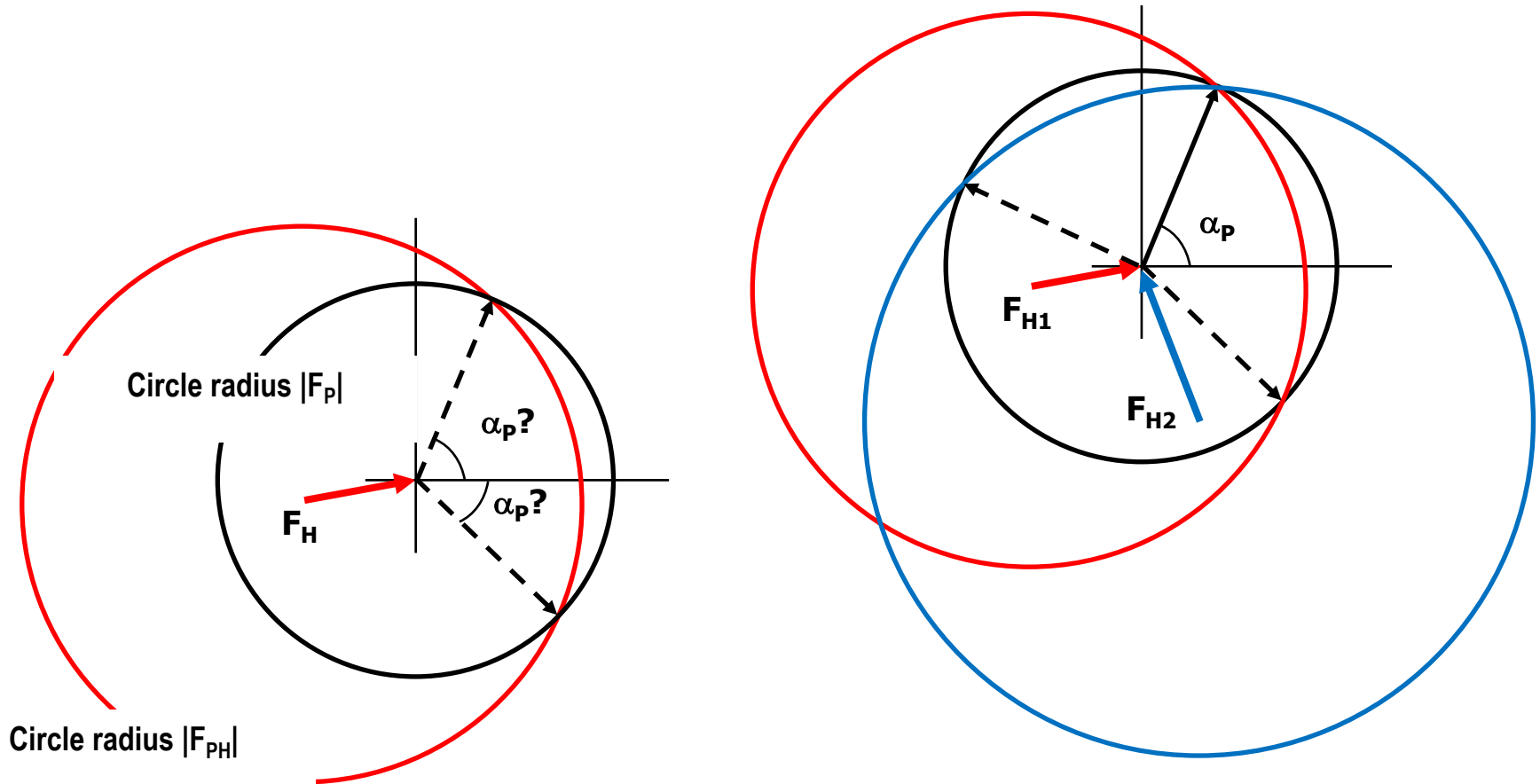
- Single Atom MR has inspired a new way to find the SAD substructure
- Phassade is an FFT-substructure search
 - based on the SAD likelihood target
 - Very fast
- Phassade avoids any requirement to estimate F_A
 - automatically takes account of the effects of both real and imaginary scattering in the substructure
- Particularly good for substructures in which the
 - signal comes from mainly from real scattering
 - anomalous scattering contributions are very large

Interference and Phase

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

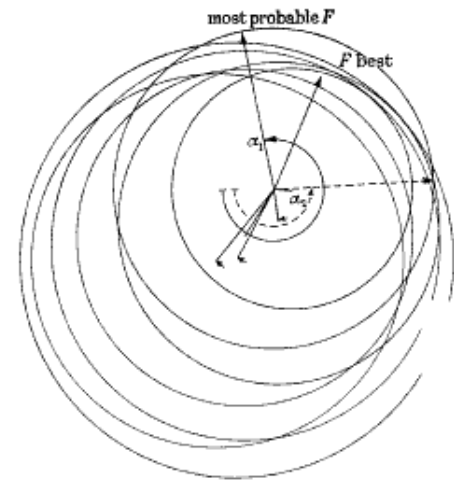
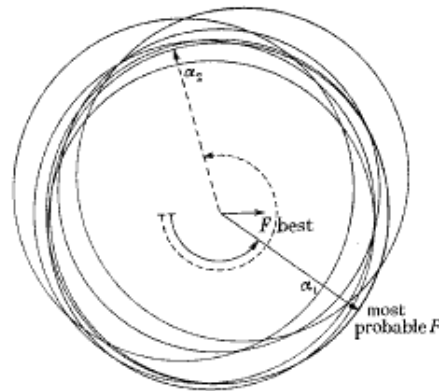
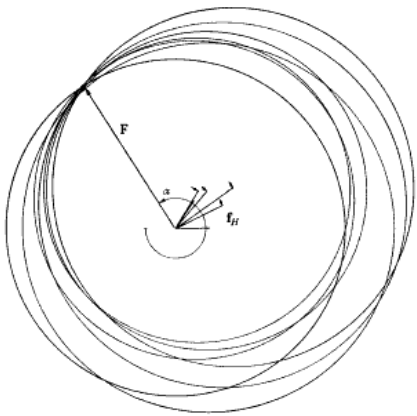


Harker Diagrams



Reality...

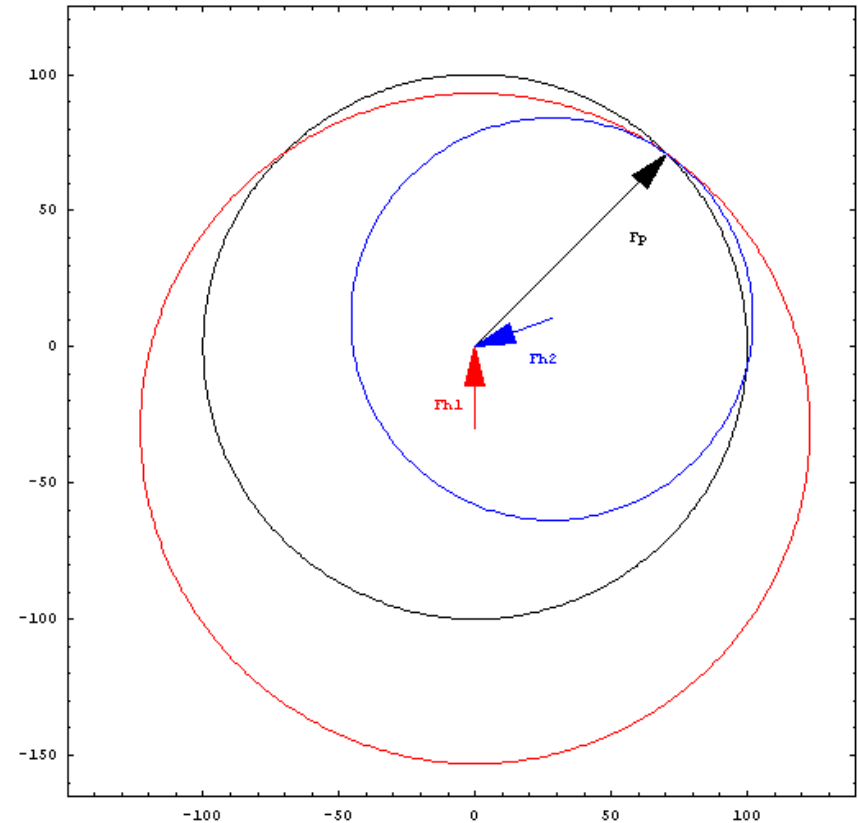
- Some real Harker diagrams from the phasing of haemoglobin with 6 derivatives



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

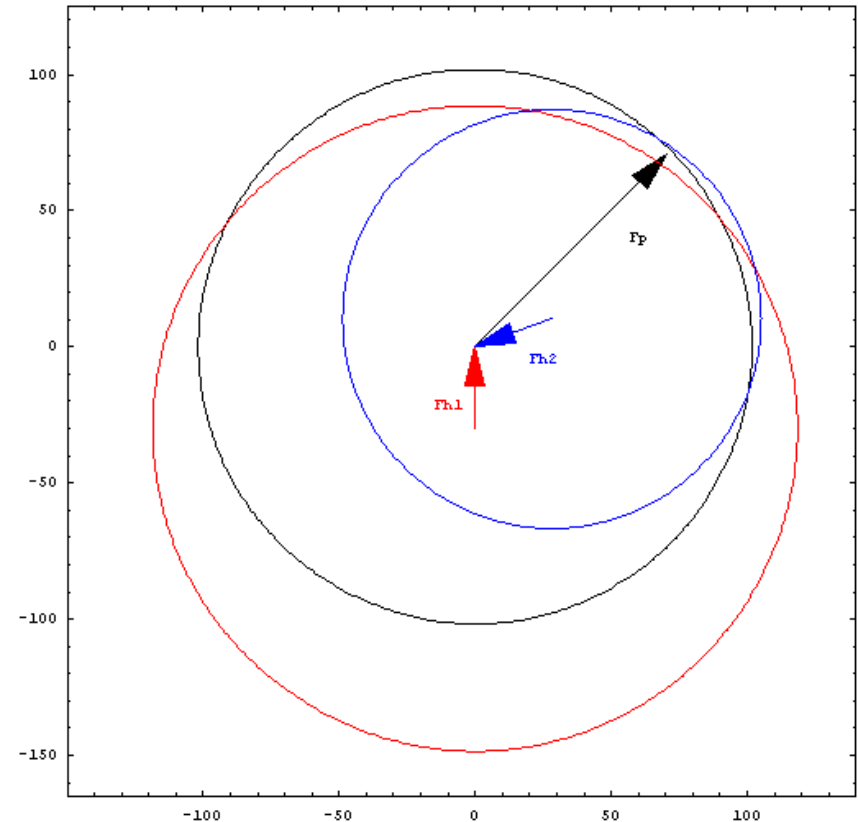
Harker construction

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



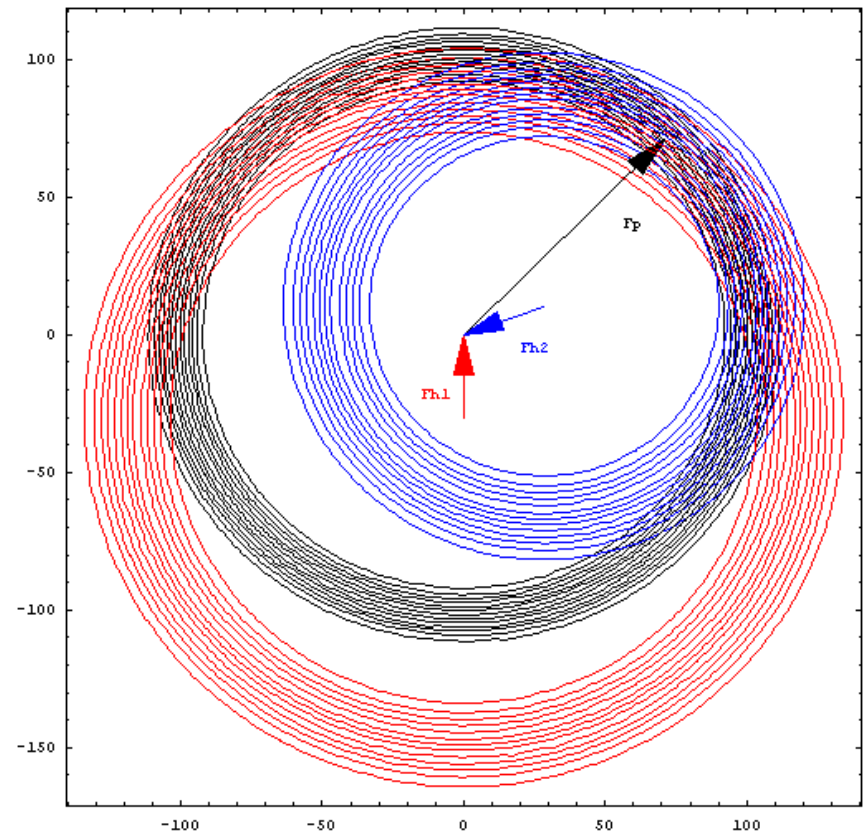
Harker construction

- 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



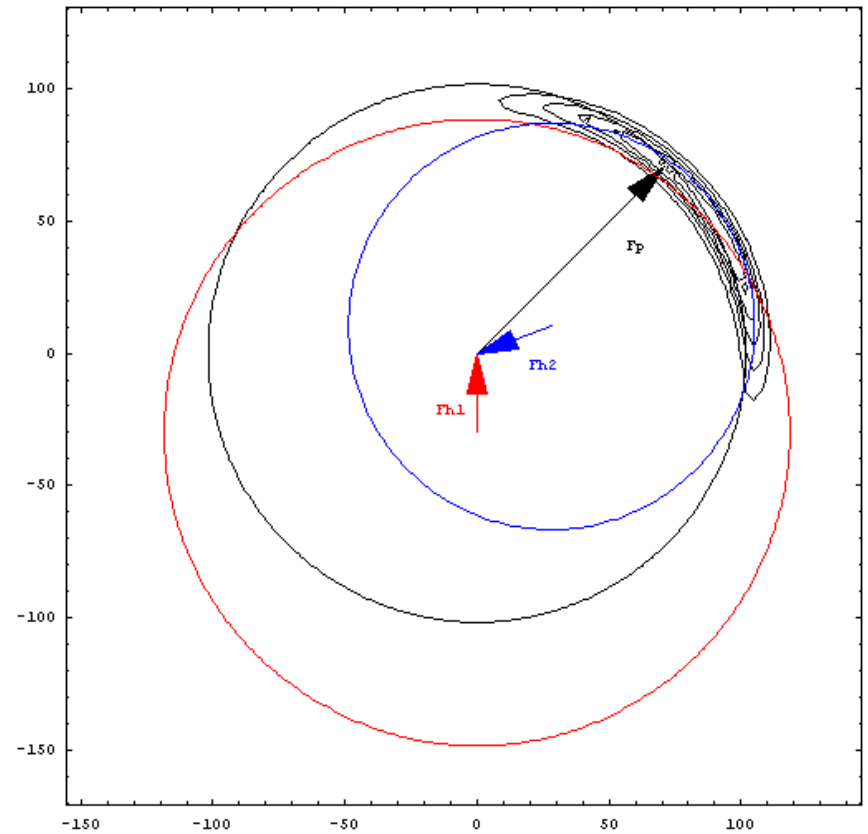
Probabilistic Harker Diagram

- 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



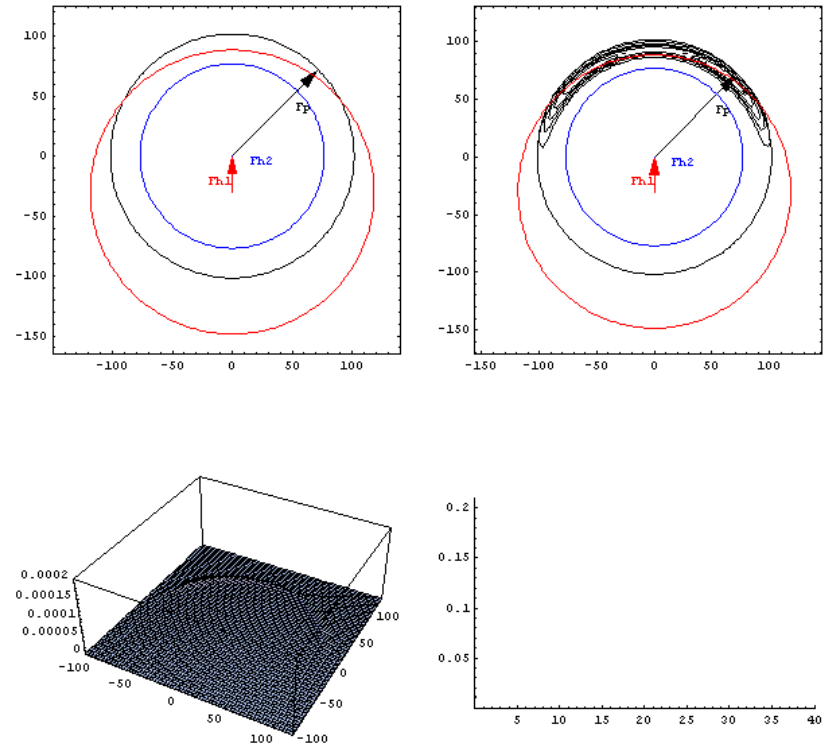
Probabilistic Harker Diagram

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



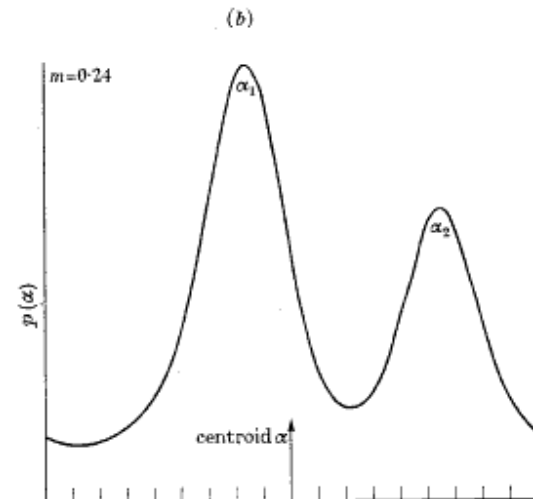
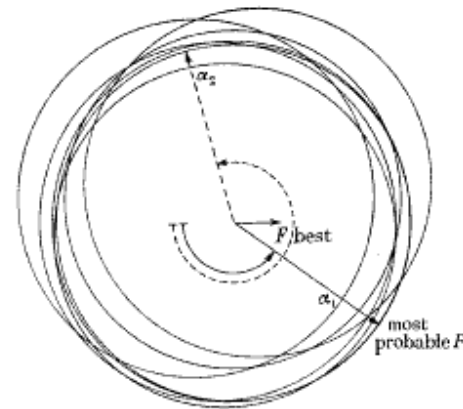
Refining Occupancy

- 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)



Phase probability

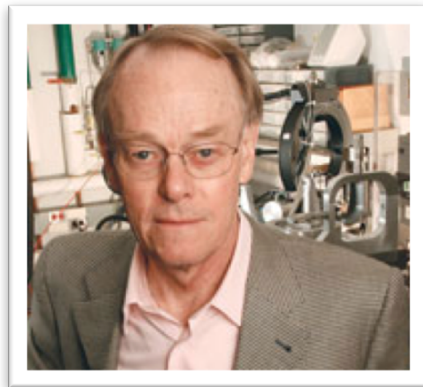
- Each reflection has a phase probability density function (PDF) rather than a single phase
- This is a complicated mathematical function
 - Requires lots of memory
- Hendrickson-Lattman coefficients are used to store the PDF in a compact form



Hendrickson-Lattman Coefficients

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

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



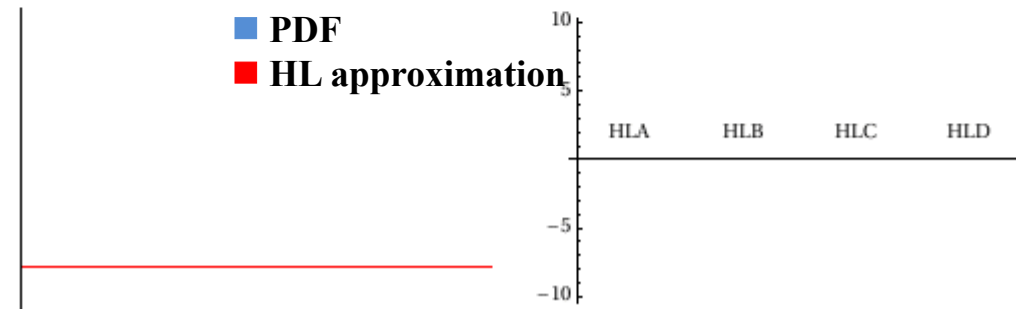
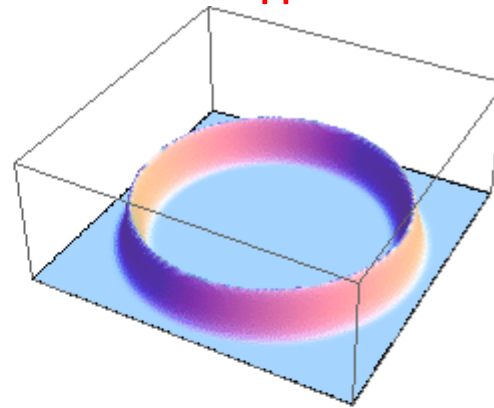
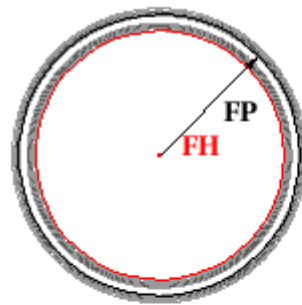
Wayne Hendrickson



Eaton Lattman

Hendrickson-Lattman Coefficients

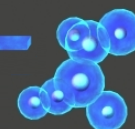
HL coefficients as a function of F_H occupancy





SAD

- The most popular way of solving structures by experimental phasing (over 70%)
- Can be done with intrinsic S and CuK α 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



Anomalous Scattering

A·nom·a·lous

adj.

Deviating from the normal or common order, form or rule

“**Anomalous** scattering” is **absolutely normal** while “**normal** 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”

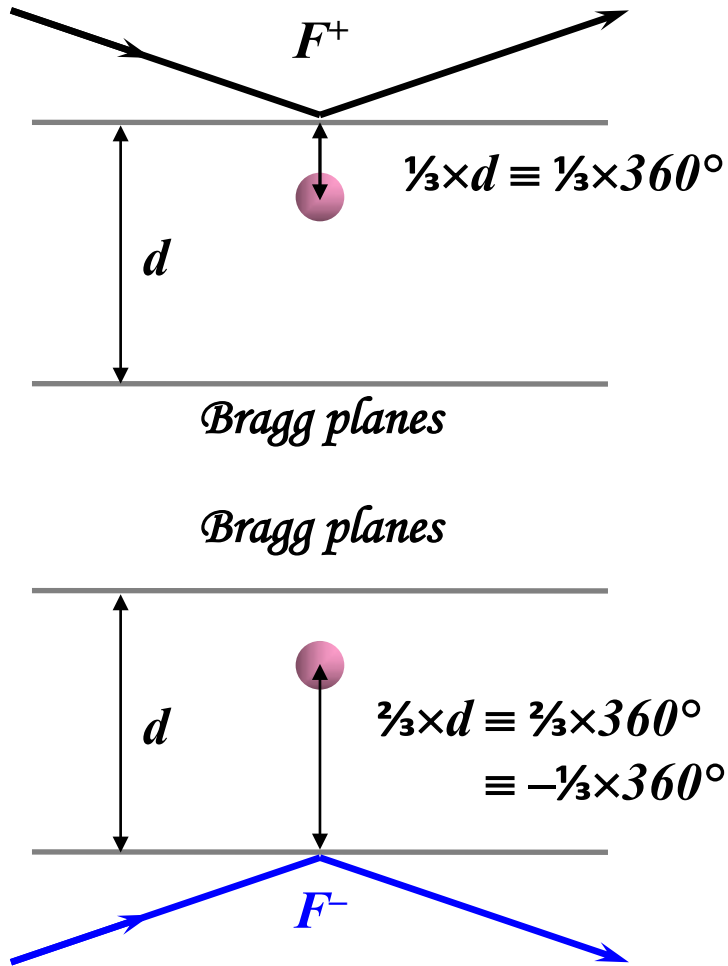
S. Caticha-Ellis (1998)

Driven Mechanical Oscillator

MIT Physics Lecture
Demonstration Group

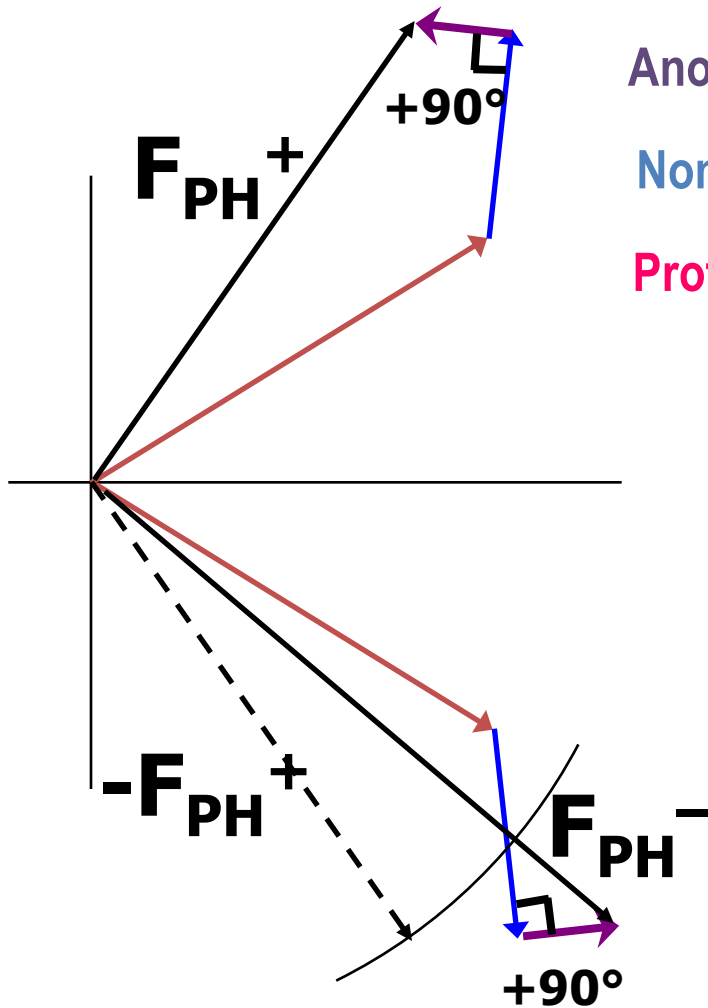
<https://www.youtube.com/watch?v=aZNnwQ8HJHU>

SAD Harker Diagram



- Phase of non-anomalous component depends on the atom's location between Bragg planes
- Anomalous scattering component is always **advanced by 90°**
 - irrespective of position of atom between Bragg planes

SAD Harker Diagram



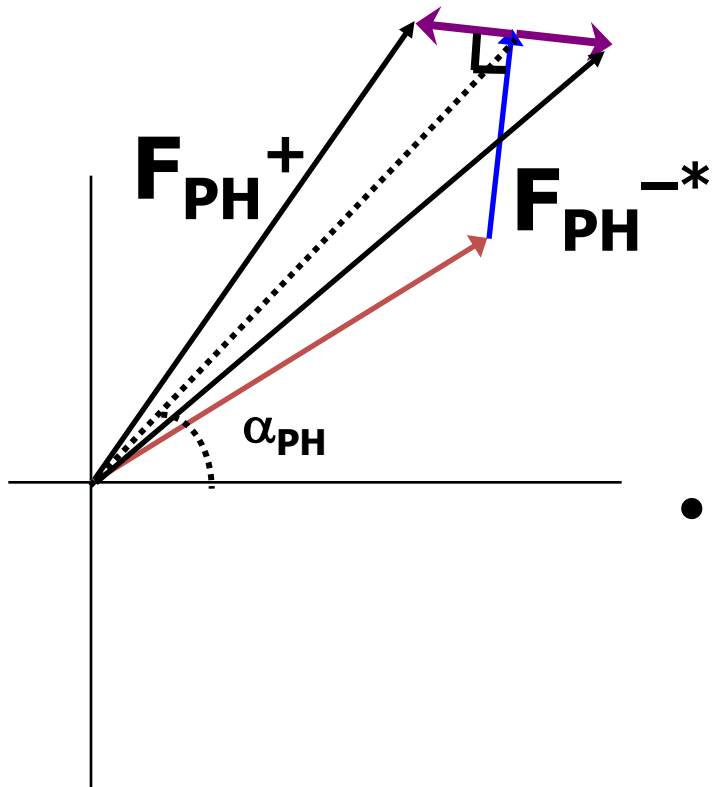
Anomalous part of heavy atom

Non-anomalous part of heavy atom

Protein (non-anomalous scattering)

- Non-anomalous and anomalous scattering components sum differently in the Friedel mates F_{PH}^+ and F_{PH}^-
 - F_{PH}^+ and F_{PH}^- have different intensities

SAD Harker Diagram



Anomalous part of heavy atom

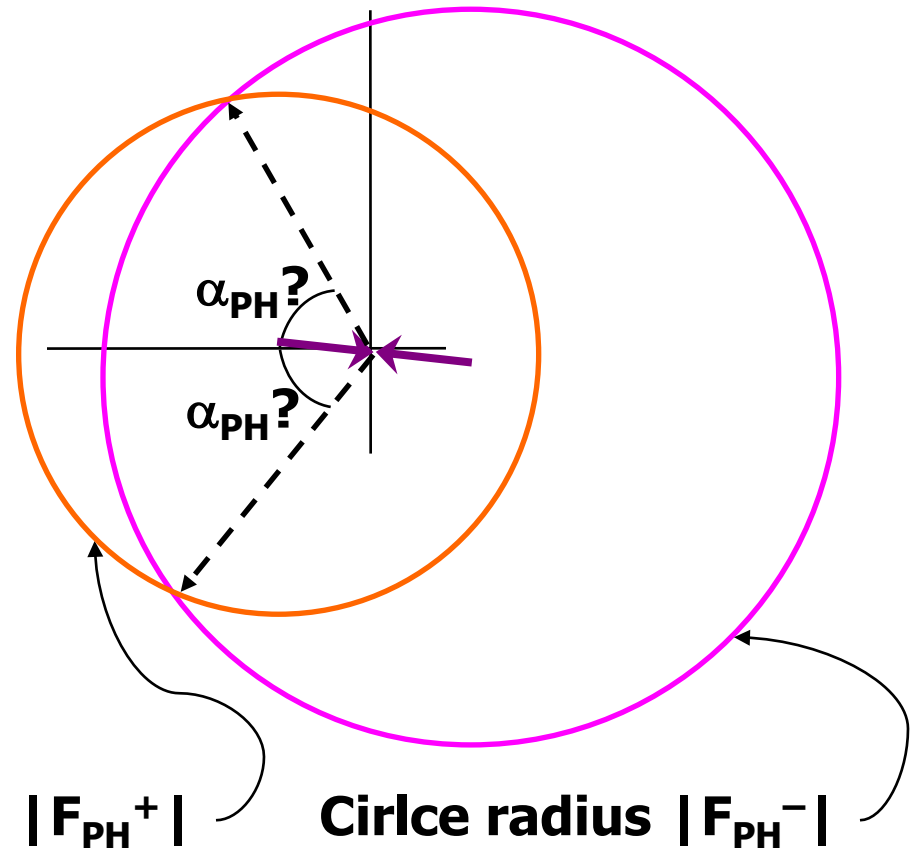
Non-anomalous part of heavy atom

Protein (non-anomalous scattering)

- It is easier/usual to use the complex conjugate of all Friedel F^-
 - i.e. all structure factors for $(-h, -k, -l)$
 - Phase $F_{PH}(\alpha_{PH})$

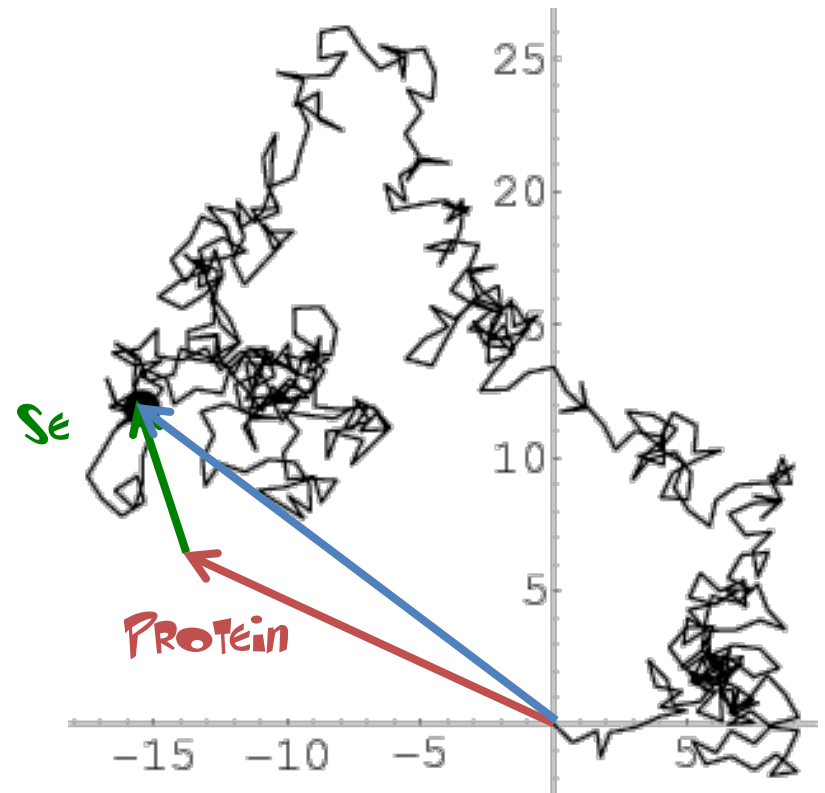
SAD Harker Diagram

- There is a phase ambiguity for F_{PH} given by the two places where the two circles with centres displaced by the anomalous scattering vectors overlap



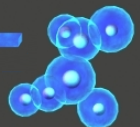
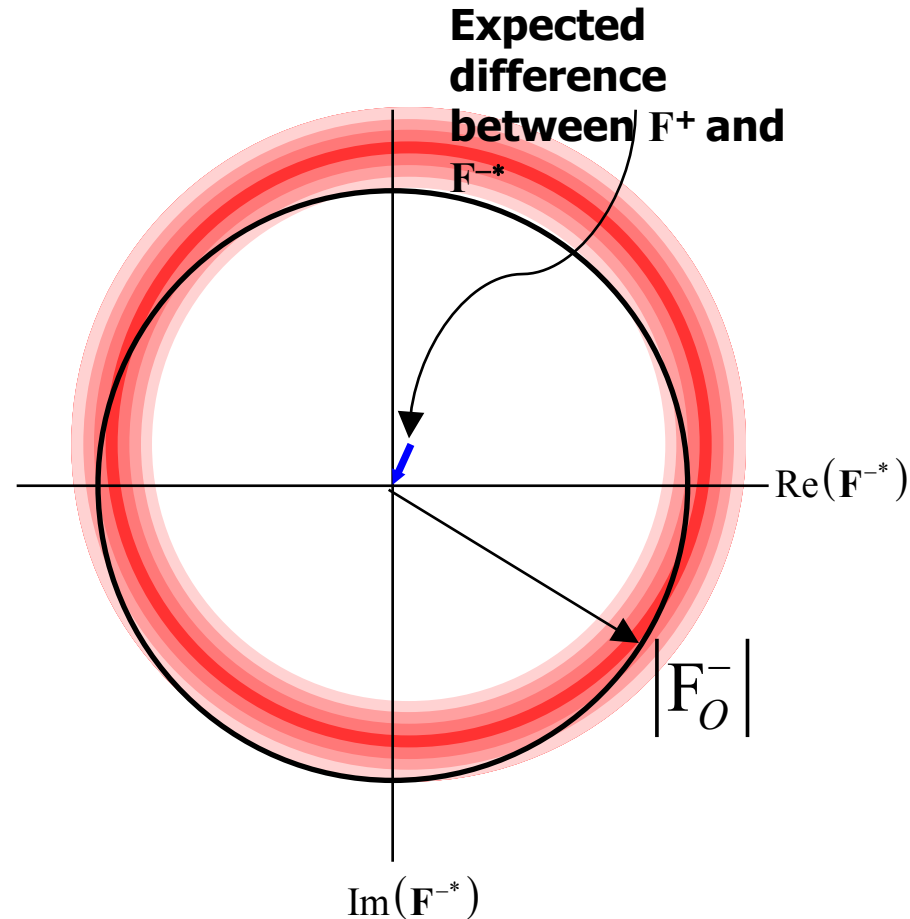
Partial Structure

- But there is more phase information!
- The anomalous scatterer is also a part of the structure
 - Gives additional phase information
 - Begins to break phase ambiguity



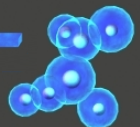
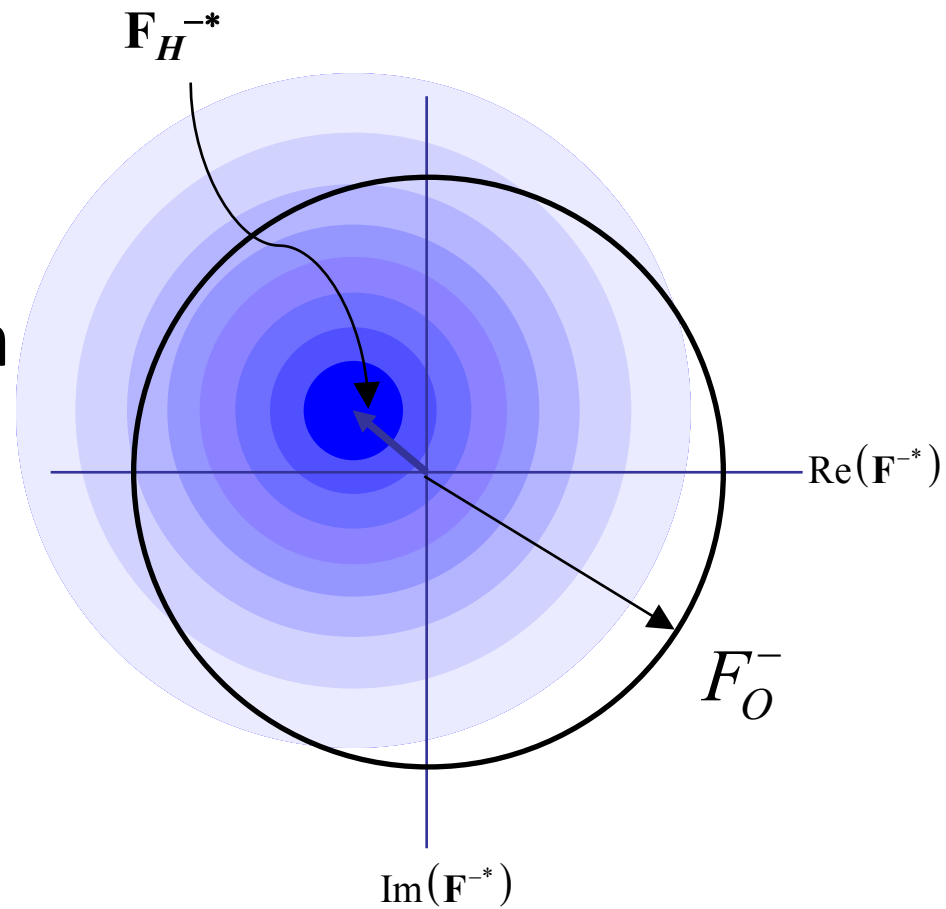
SAD Phasing

- Primarily anomalous scattering
- Gives a narrow probability distribution



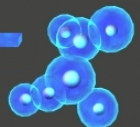
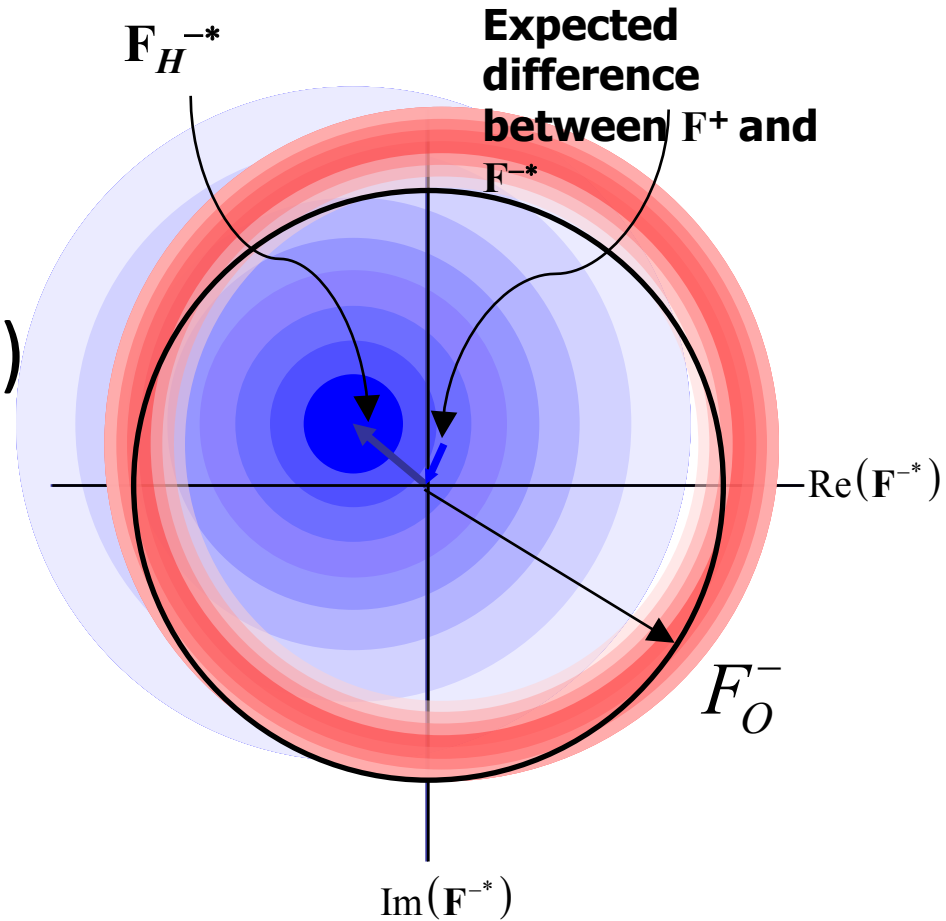
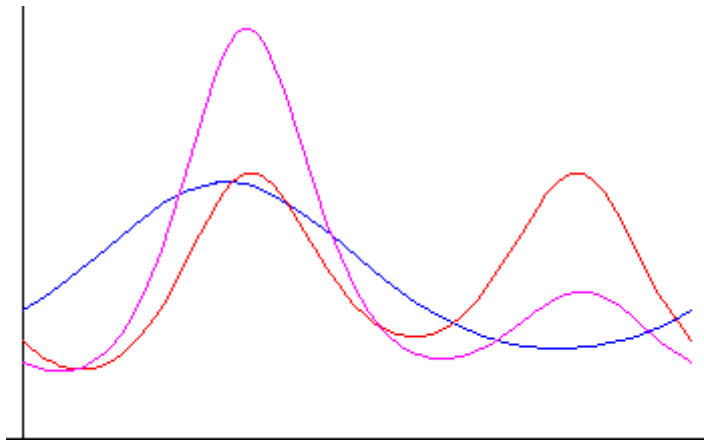
SAD Phasing

- Primarily Normal scattering
- Gives a broad probability distribution



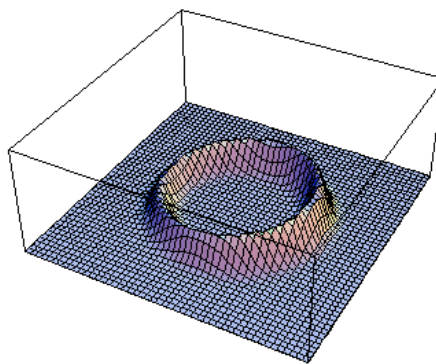
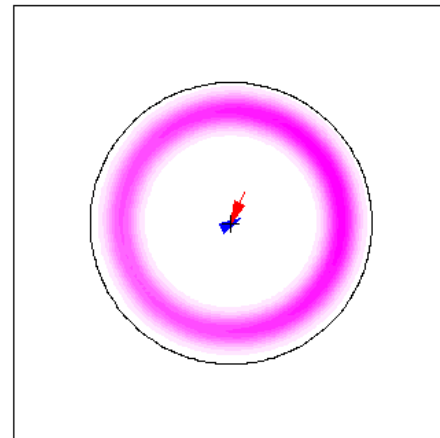
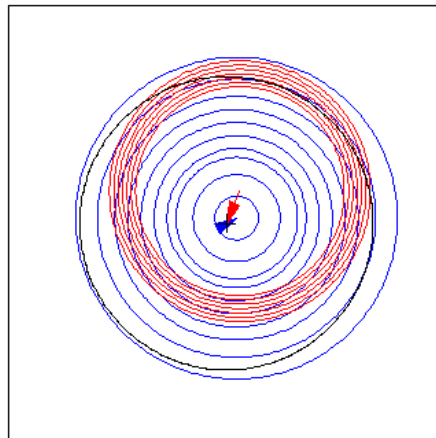
SAD Phasing

- Likelihood is proportional to the product of the two distributions (magenta) under the black circle

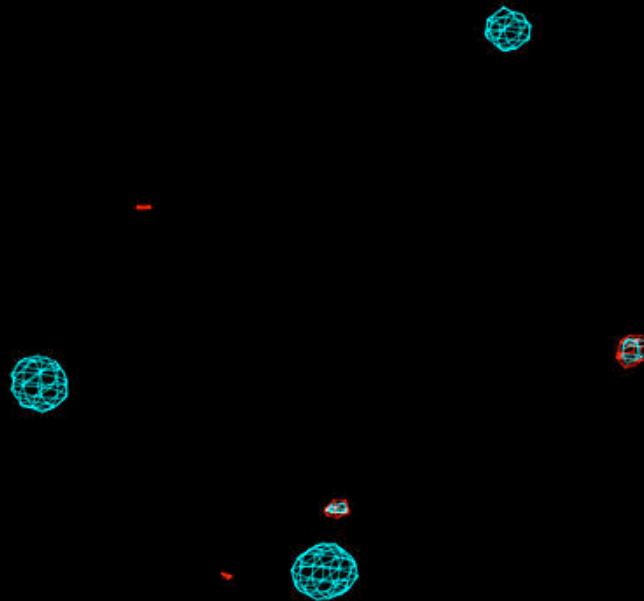


Refining Occupancy

- To refine the occupancy of a heavy atom, maximise the SAD likelihood
- Final refined value is the optimum for ALL reflections (movie shows ONE reflection)

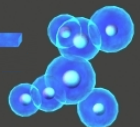
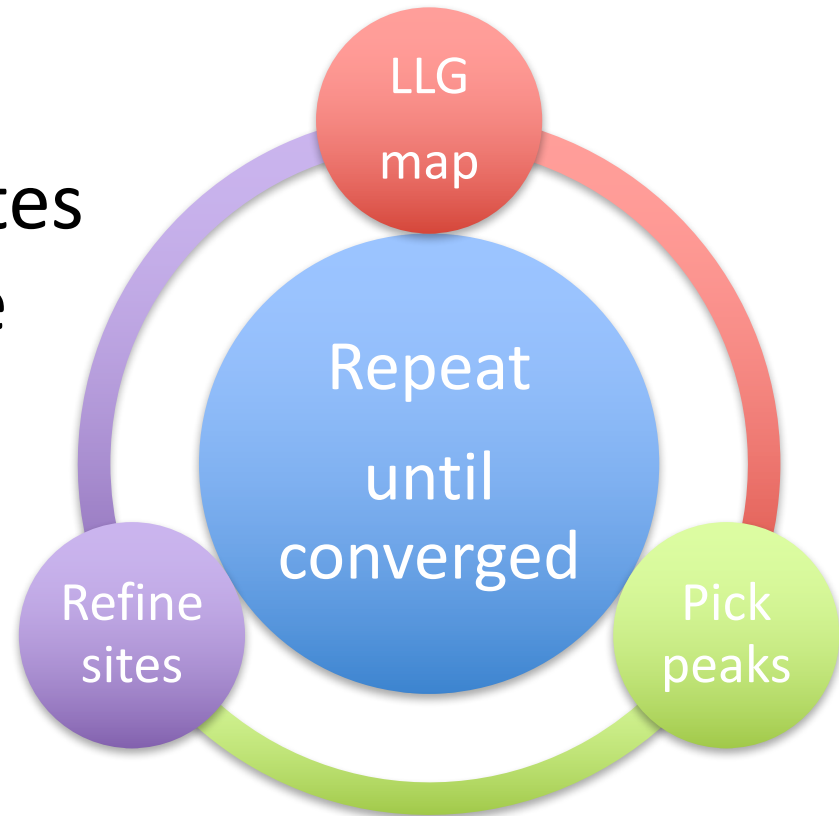


Log-likelihood gradient maps



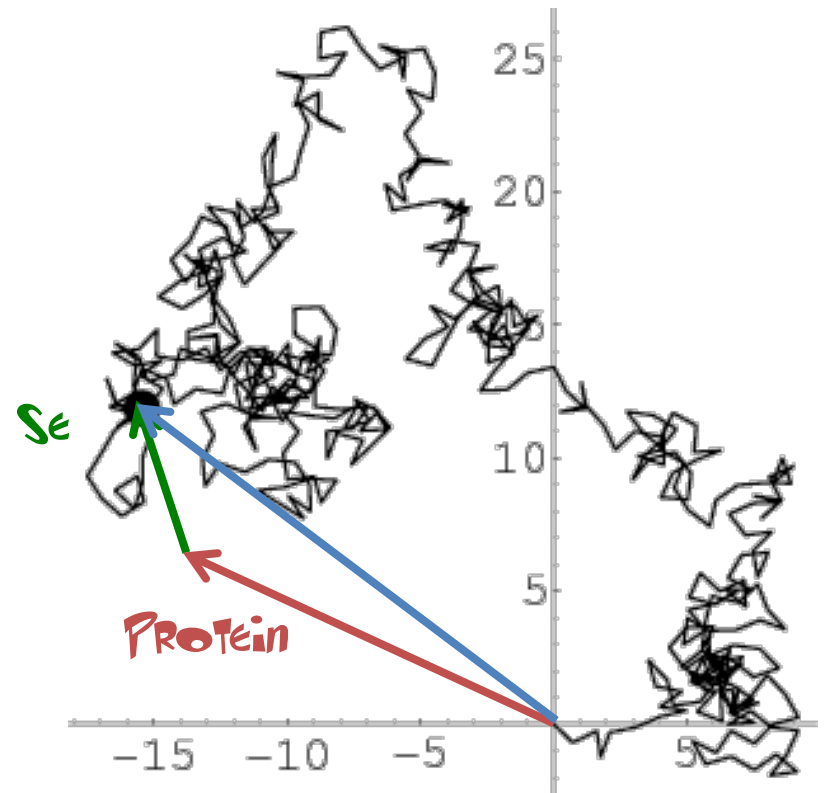
Completion of sub-structure

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



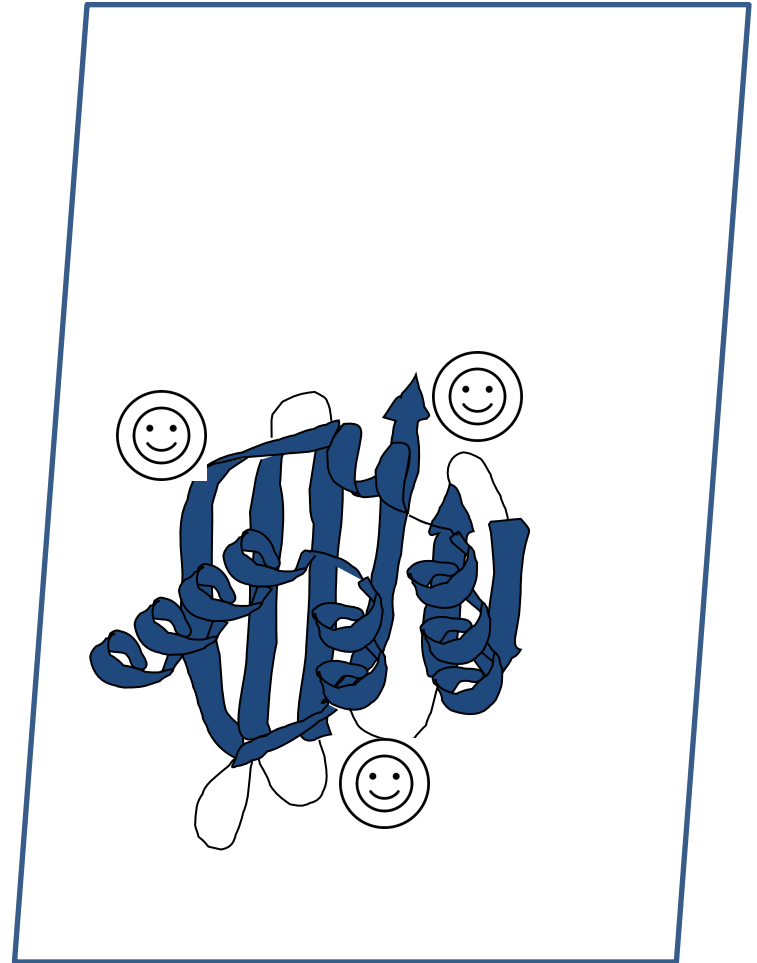
Partial Structure

- But there is more phase information!
- The anomalous scatterer is also a part of the structure
 - Gives additional phase information
 - Begins to break phase ambiguity



Locating Sites

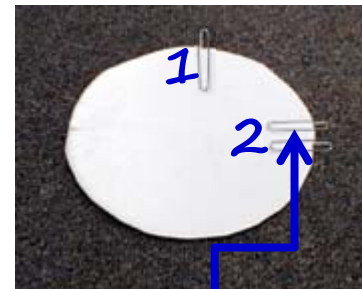
- Log-likelihood
Gradient maps
- Can find very weak sites
- Helps to remove model bias when the model is very poor



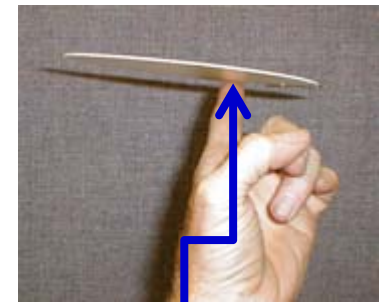
Calculating Electron Density

- 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 edge of the disc
 - 2 together
 - 1 a distance away
- The balancing point is between the 3 paperclips
 - Not on the 2 paperclips

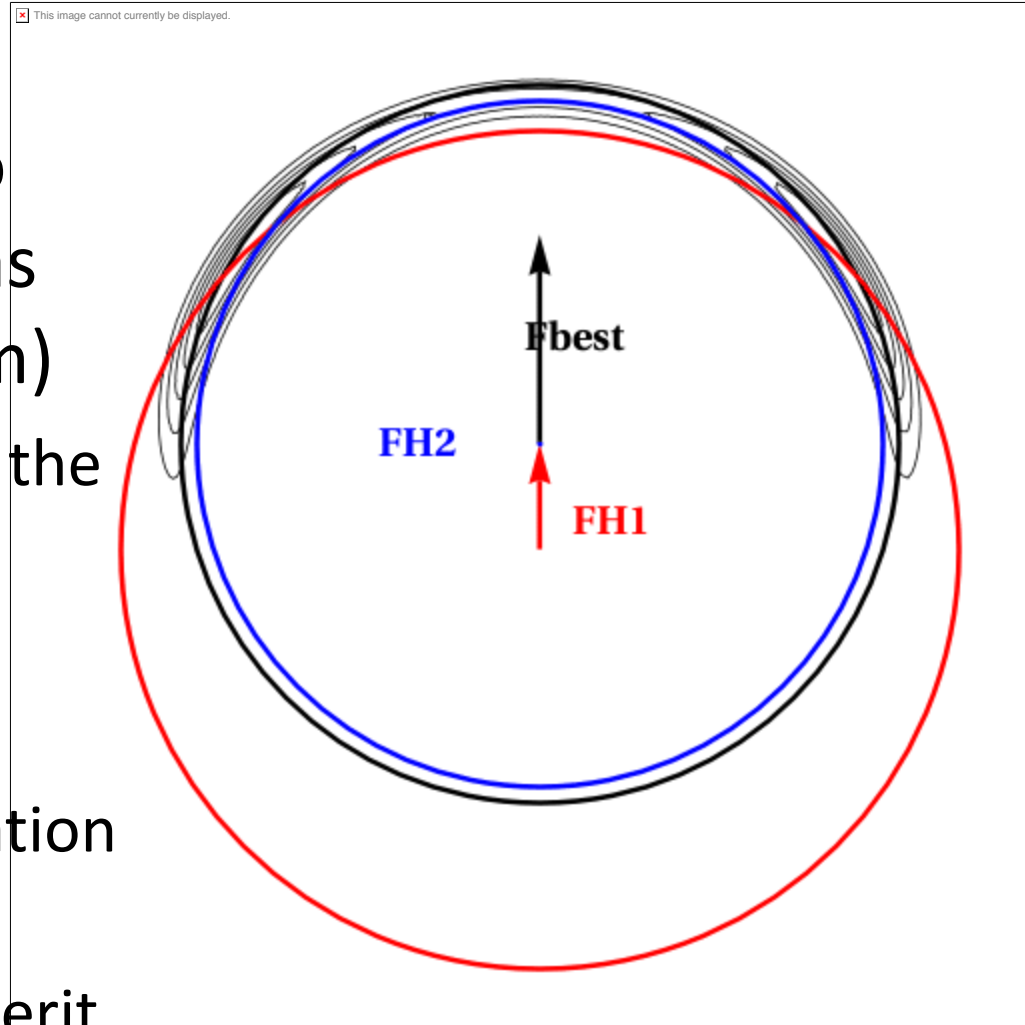


*Most
Probable
Structure
Factor*



*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

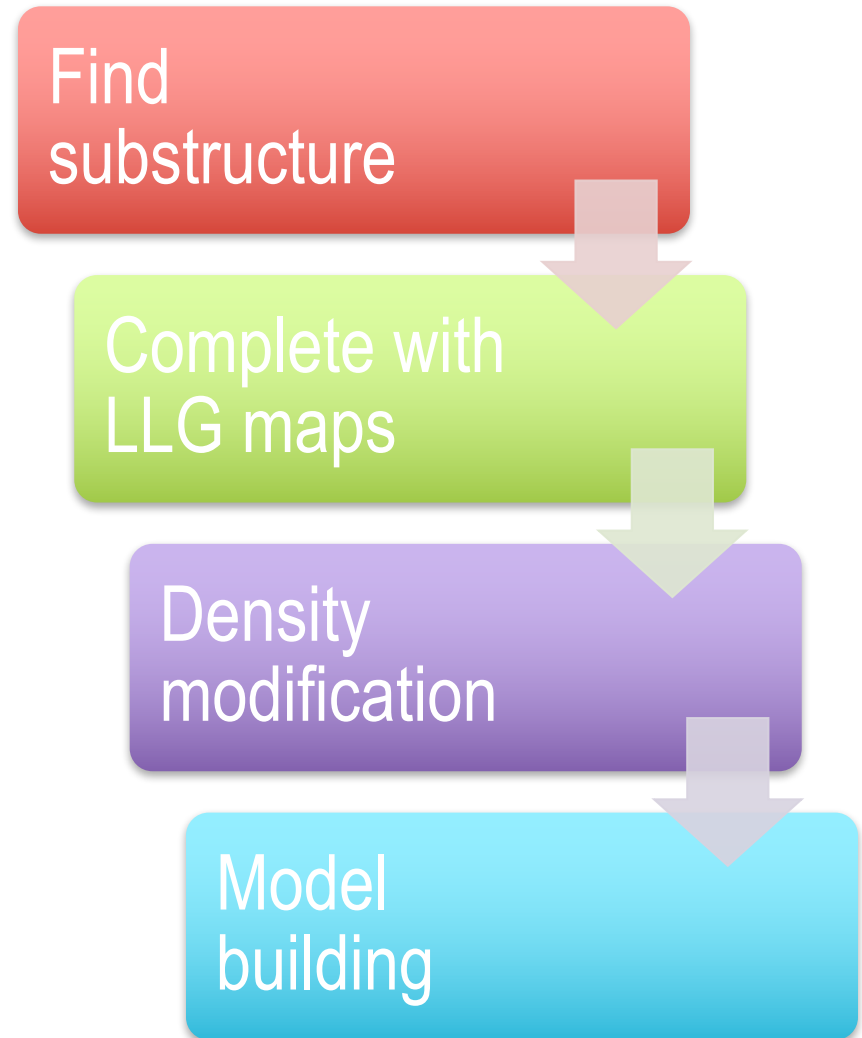


New approaches



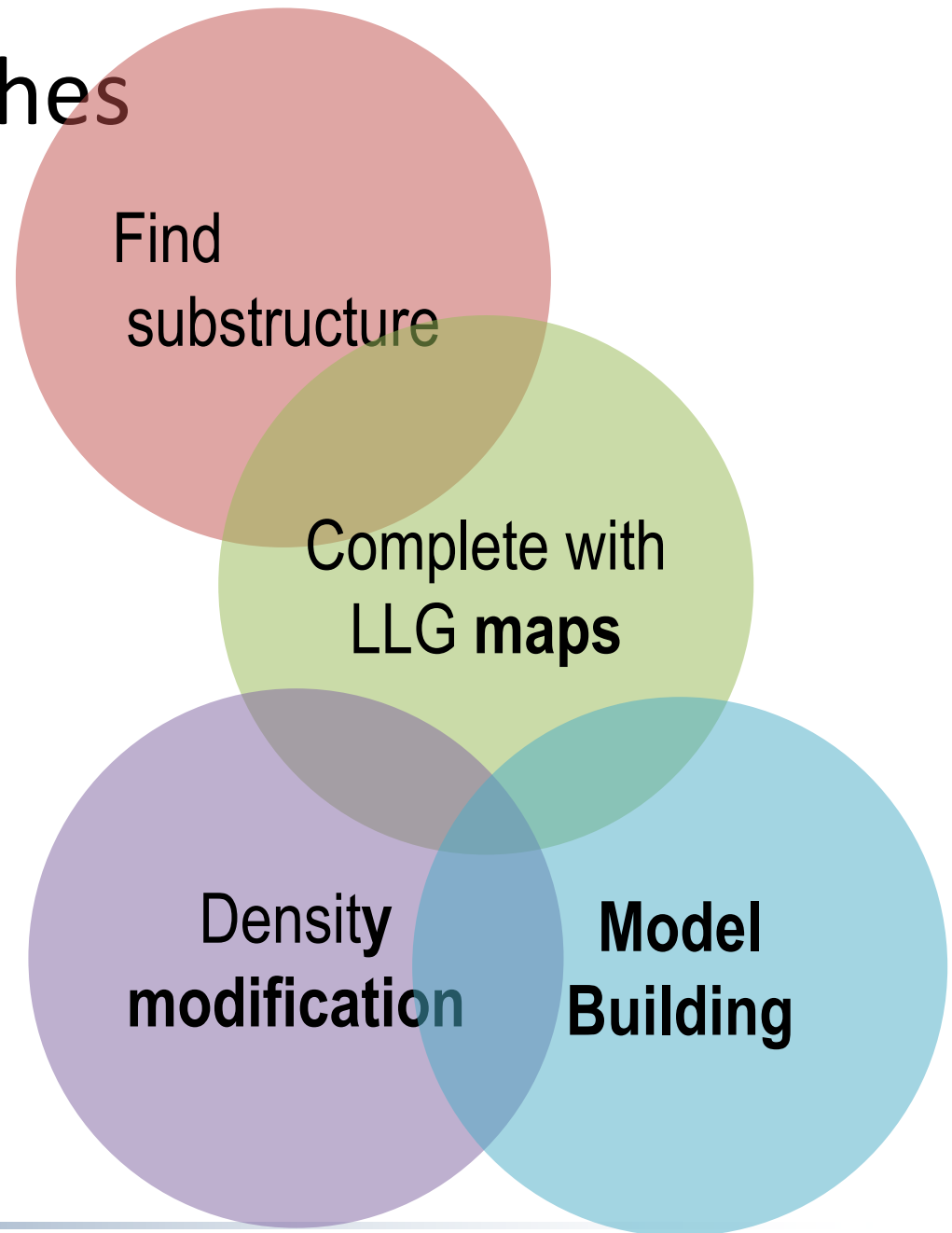
The pathway of structure solution

- 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

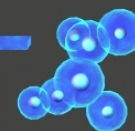
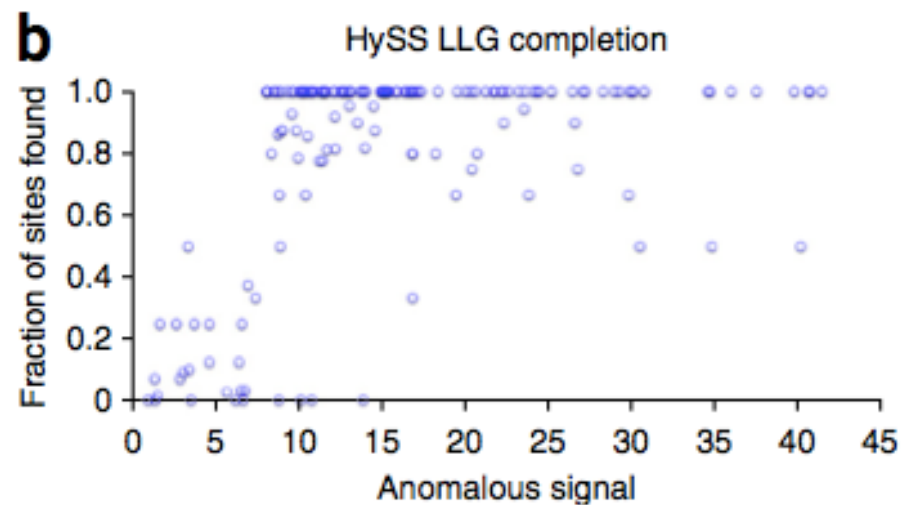
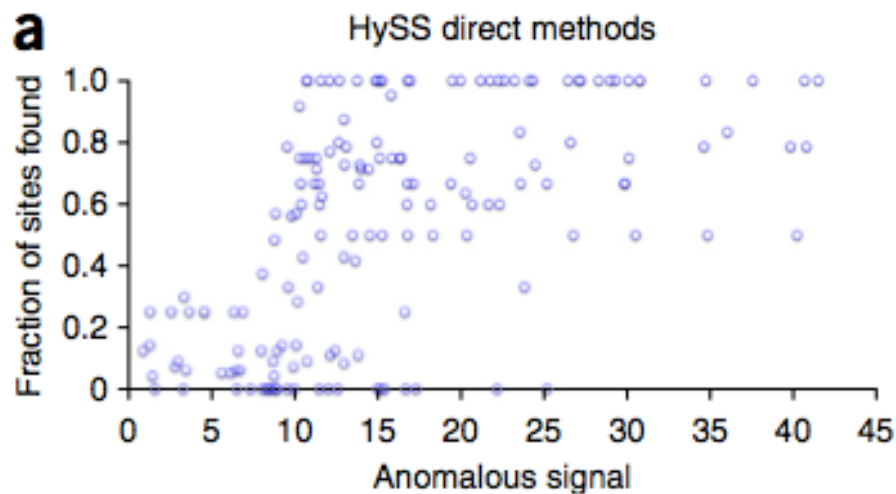
- **Take multiple possibilities for each step and uses subsequent steps to distinguish correct from incorrect solutions**
- **Enables structure solution when signal is low**

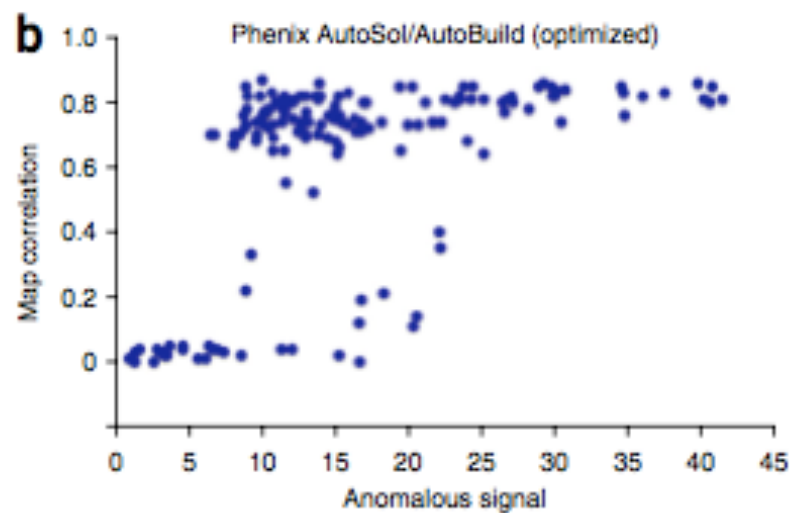
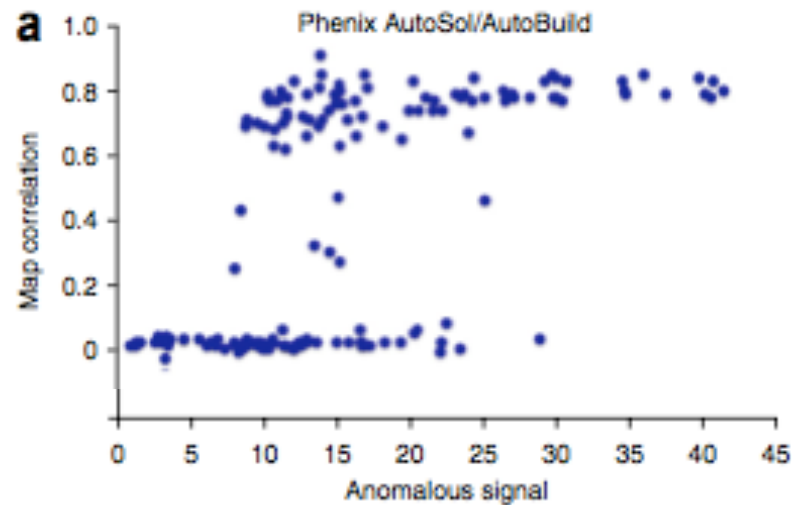
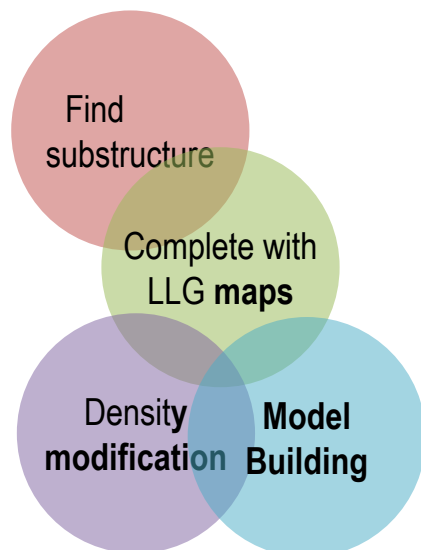


Find phasing
substructure

Find
substructure

Complete with
LLG maps



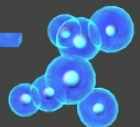
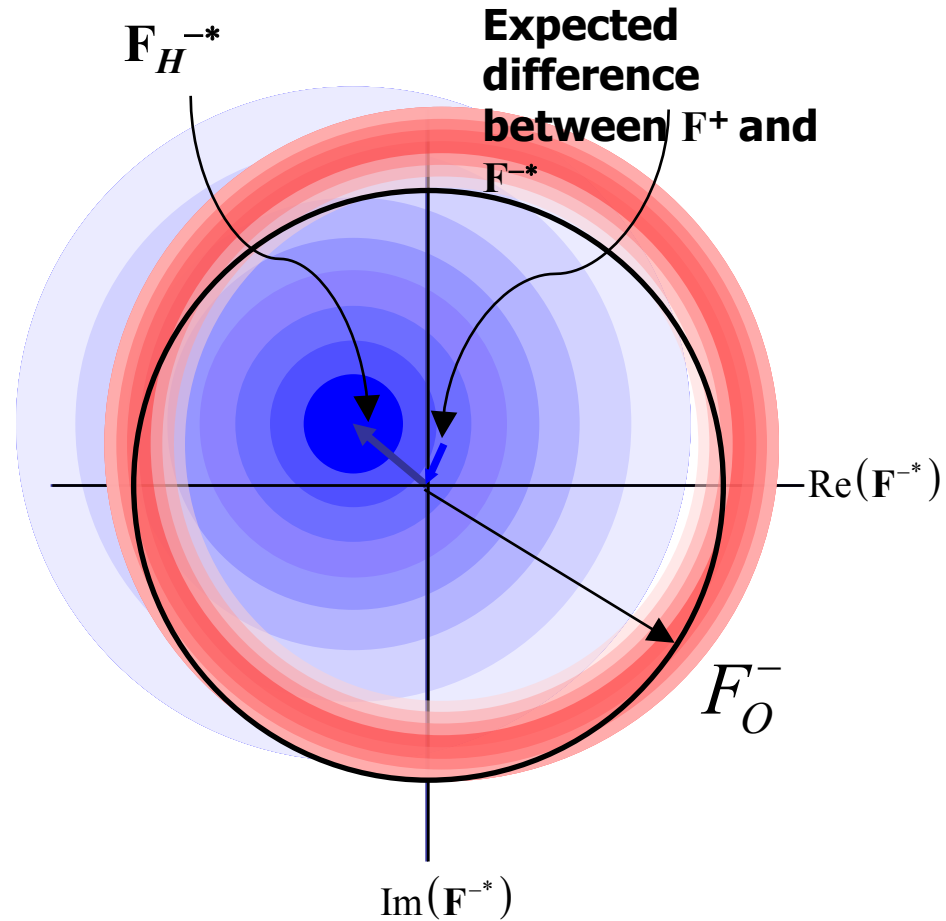
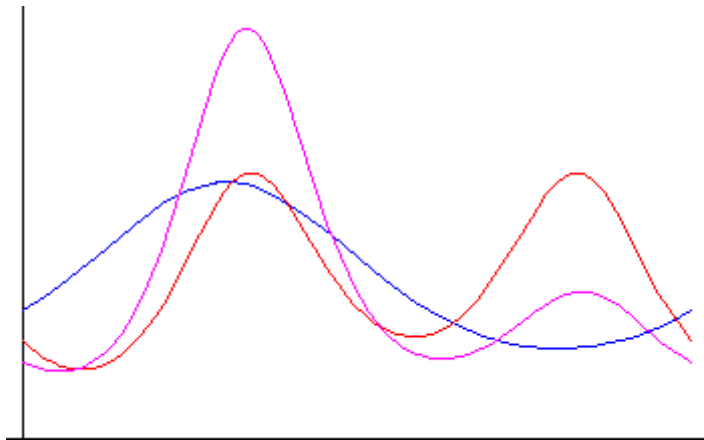


New approaches



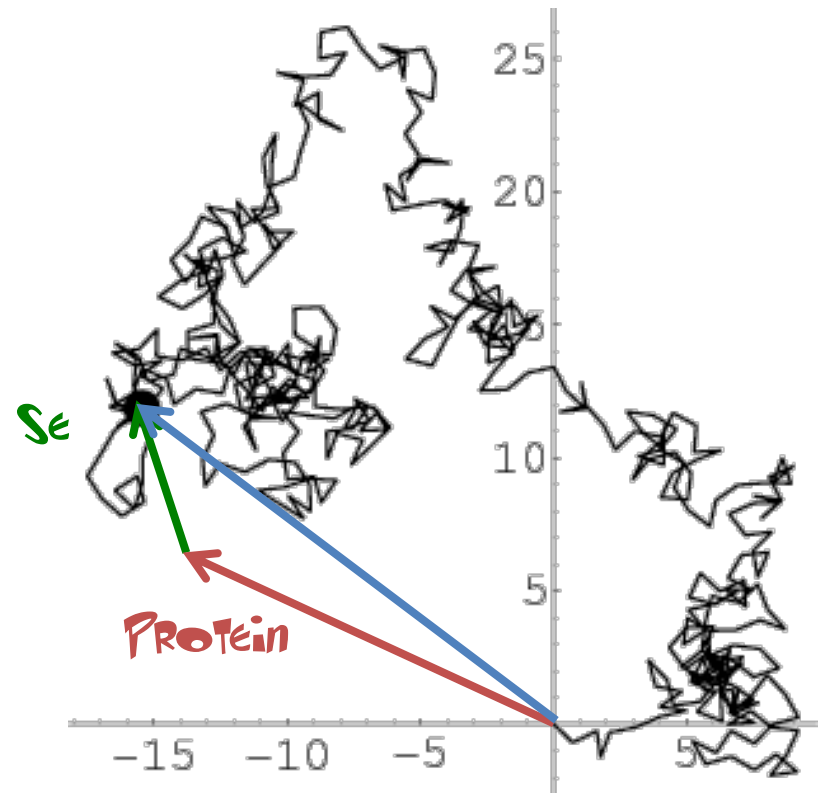
SAD Phasing

- Reminder!
 - SAD likelihood contains information from the anomalous AND REAL scattering



Partial Structure

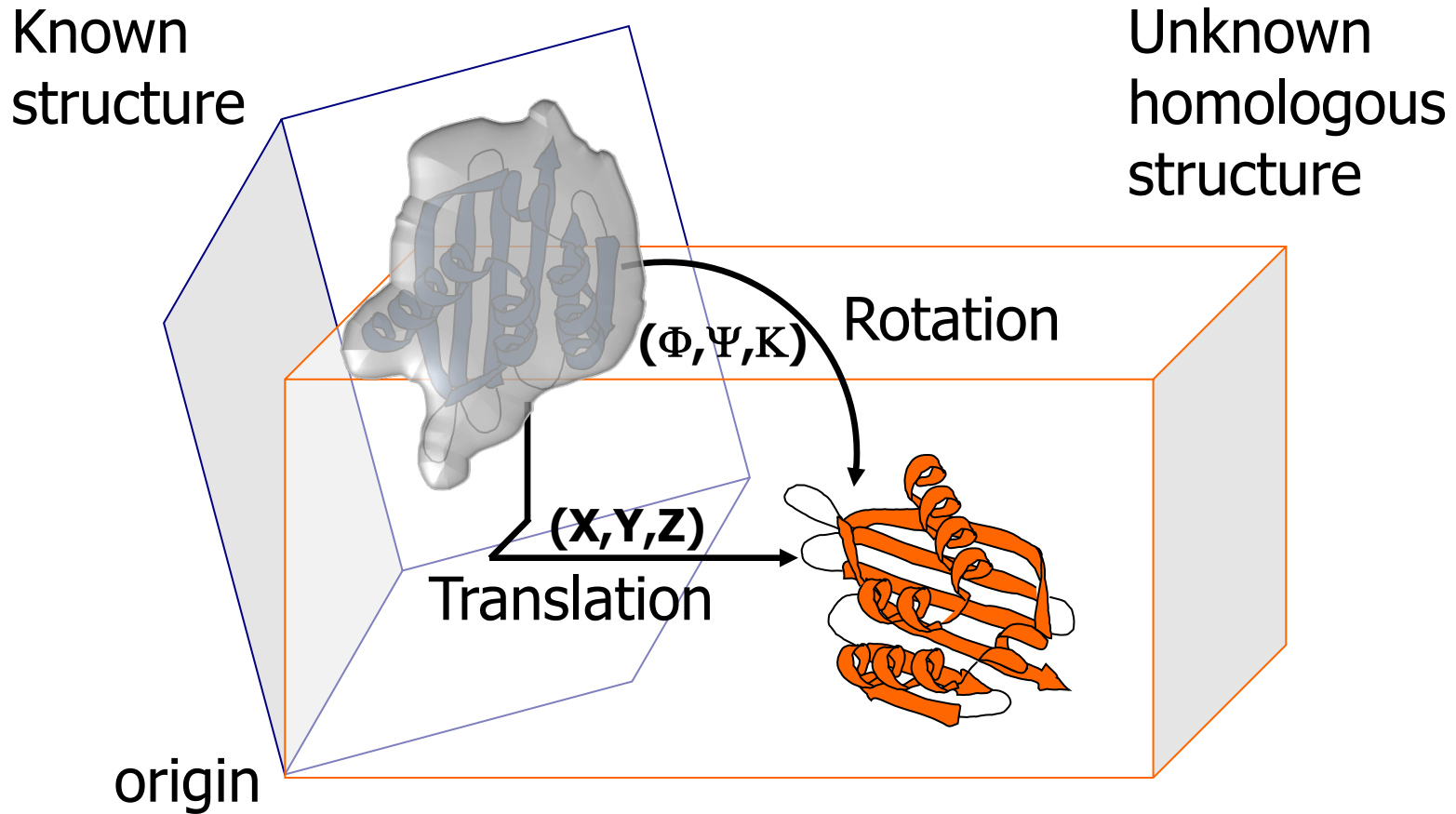
- A partial real scatterer can be used to find the anomalous sites
- This is MR-SAD
- Partial structure does not need to be coordinates



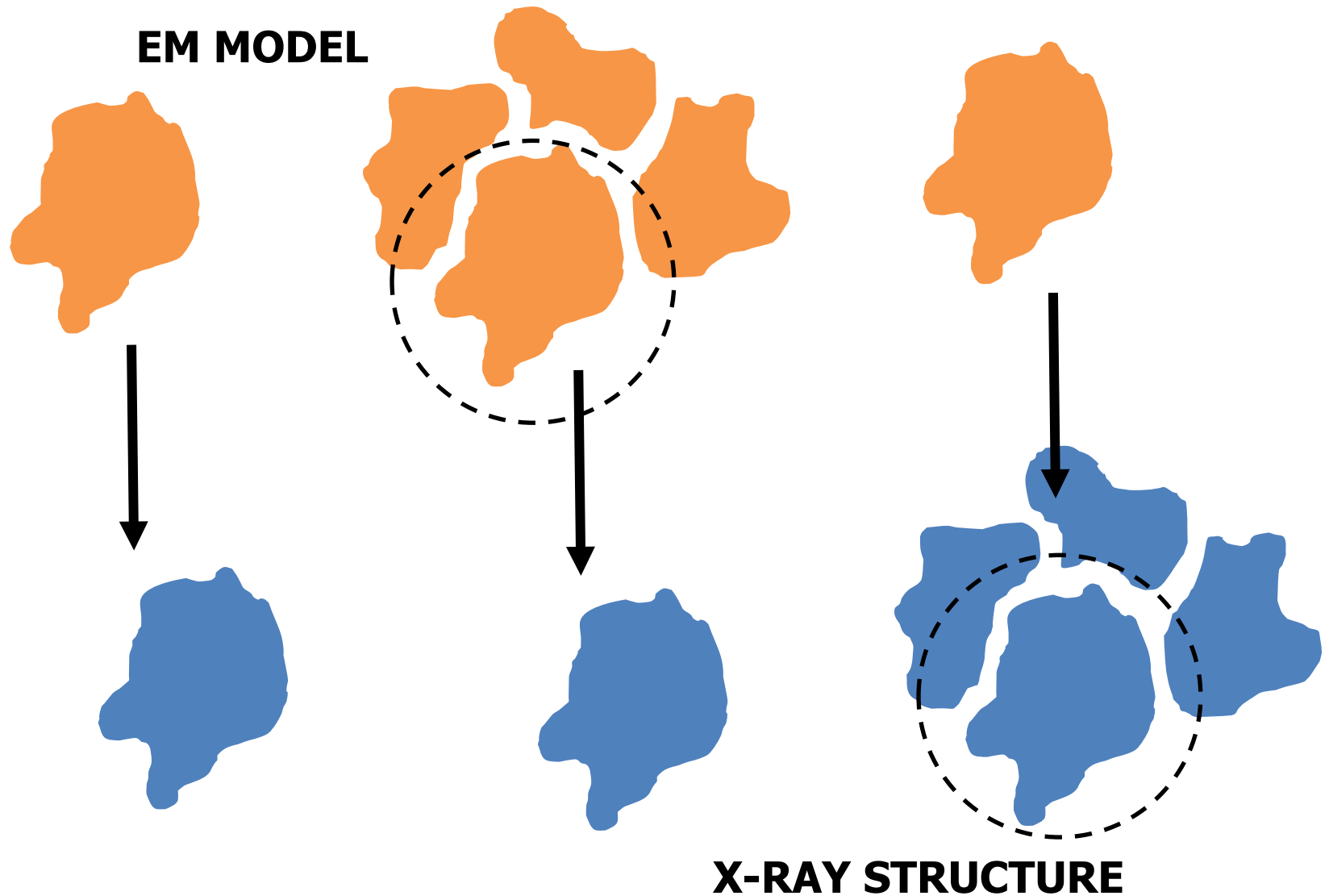
MR with MAPS



Molecular Replacement



Molecular Replacement



EM and X-ray crystallography

- Early examples were hybrid studies of icosahedral viruses
 - Both negative stain and cryo-EM images can be used
 - We are interested in cases where EM images are lower resolution than X-ray data
 - E.g 8- $\rightarrow$ 3, 5- $\rightarrow$ 1.8
 - phase extension is required after molecular replacement
 - Technically similar to cross-crystal phasing methods in X-ray crystallography
-

Problems

Using electron-microscopy images as a model for molecular replacement

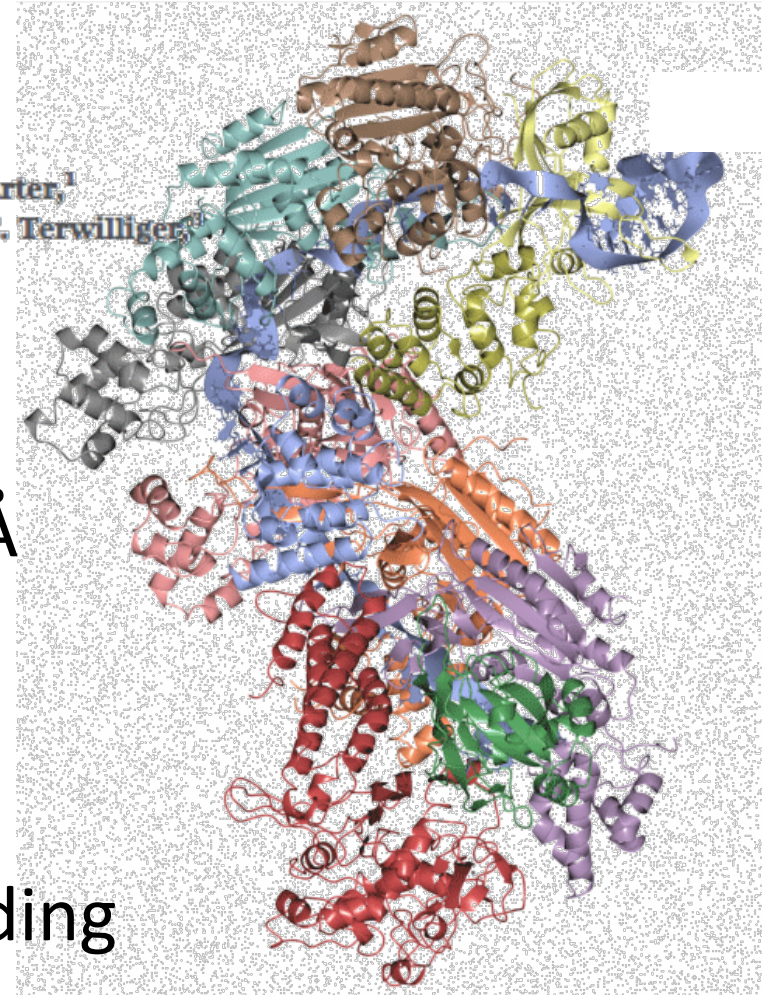
Eleanor J. Dodson, *Acta Cryst.* (2001). D57 1405-1409

- i. Problem of precise scaling of the EM model
 - ii. Problem of cryo-EM density contrast
 - iii. ~~Problem of reliable phase contrast function correction, and the possibility of inverted phases~~
 - iv. Problem of how to scale (together/absolute) the structure factors at low resolution
 - v. High internal symmetry of many target structures
-

Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*

Ryan N. Jackson,¹ Sarah M. Golden,¹ Paul B. G. van Erp,¹ Joshua Carter,¹
Edze R. Westra,^{2*} Stan J. J. Brouns,² John van der Oost,² Thomas C. Terwilliger,³
Randy J. Read,⁴ Blake Wiedenheft^{1†}

- Cryo-EM reconstruction at 8Å
- Crystals diffracted to 3Å
- Average over 2-fold non-crystallographic and internal symmetry before model building



Jackson *et al.*, Science (2014)

Protocol

Published in final edited form as:

Nat Protoc. 2015 September ; 10(9): 1275–1284. doi:10.1038/nprot.2015.069.

X-ray structure determination using low-resolution electron microscopy maps for molecular replacement

Ryan N. Jackson¹, Airlie J. McCoy³, Thomas C. Terwilliger², Randy J. Read³, and Blake Wiedenheft^{1,*}

¹Department of Microbiology and Immunology, Montana State University, Bozeman MT 59717, USA

²Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545

³Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Cambridge CB2 0XY, U.K.

Direct use of EM map

ENSEMBLE EM MAPIN embd_513.map &

RMS 3.5 PROTEIN MW 390000 NUCLEIC MW 0



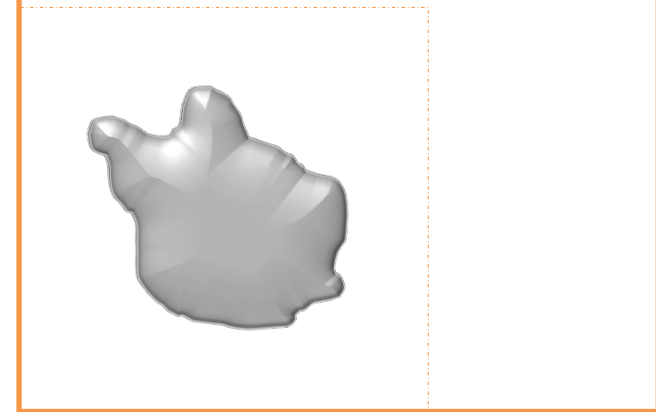
Direct use of EM map



**Volume
cutout
determined
by
molecular
weight of
model**

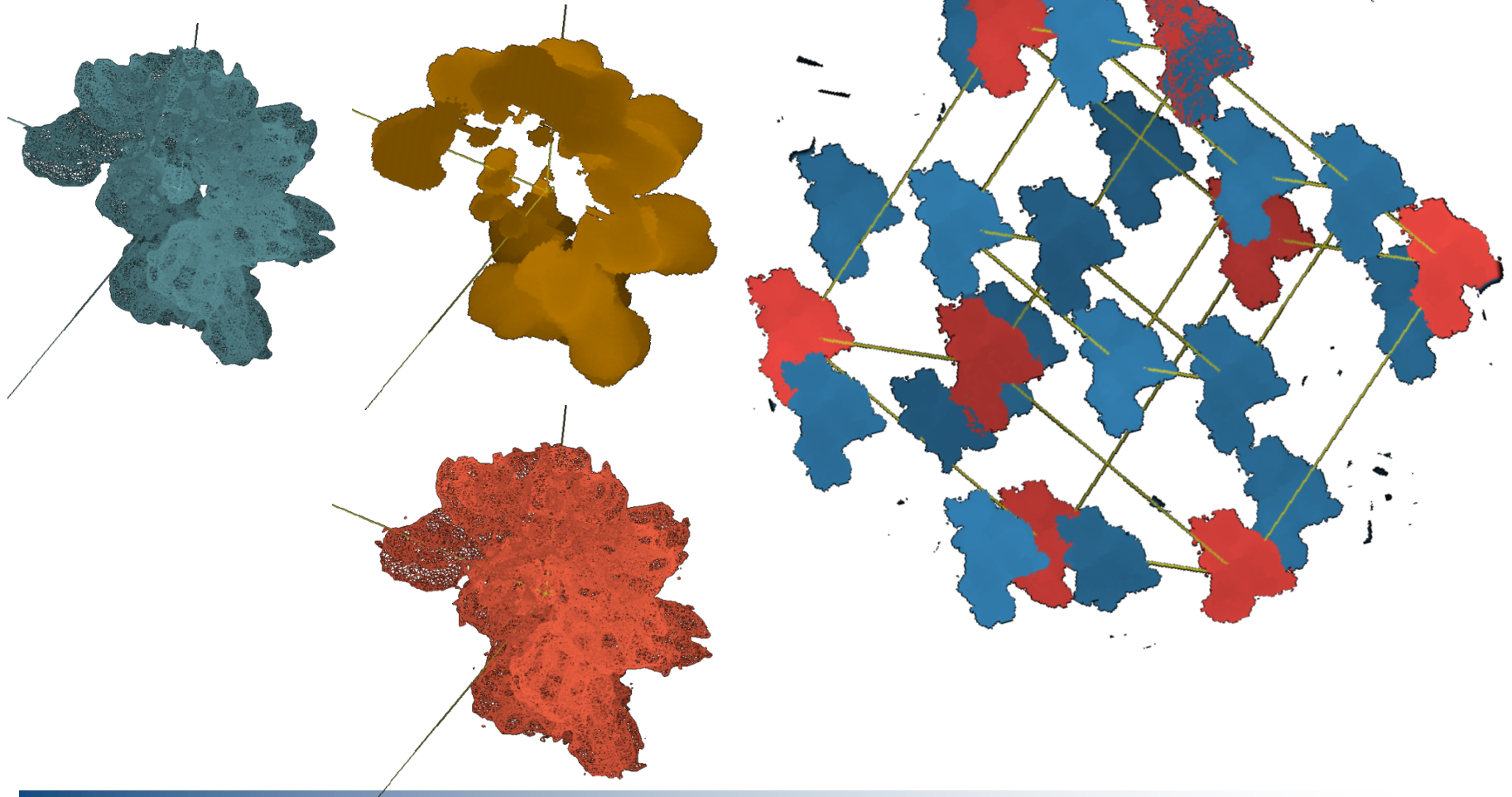


**Model in large P1
unit cell**



**No shift from input position
*c.f. Nature Methods Protocol***

Example: Cascade

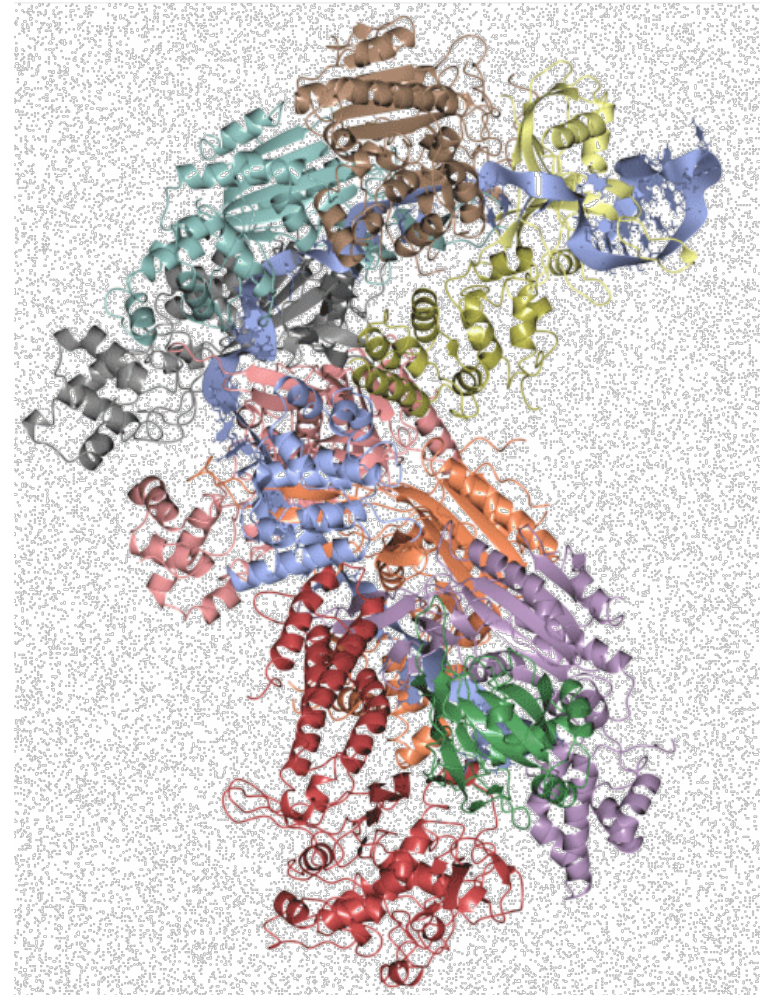


EM Scale Factor

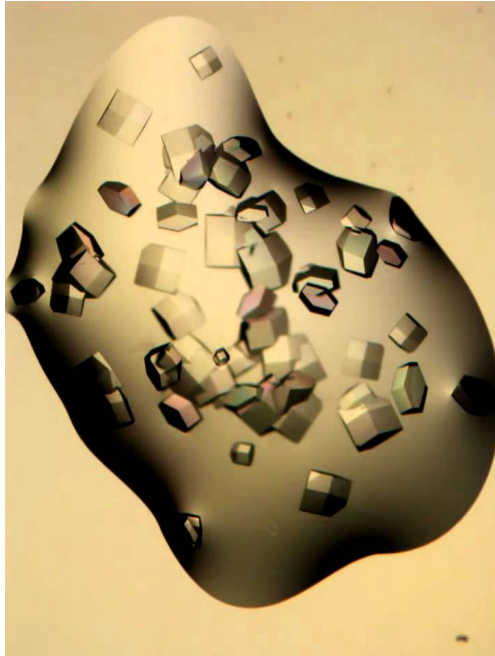
- The scale factor of maps can be refined against the likelihood function
 - Cell refinement assumes that the solution is present in the list of putative solutions, possibly with low signal-to-noise
 - Refinement will increase signal-to-noise
 - A grid search is required if the scale factor is so large that a solution is not found
-

Example: Cascade

- Essential to refine the magnification factor to identify the correct solution
- Scale factor refined to 0.949
 - changes length of reconstruction from 190Å to 180Å



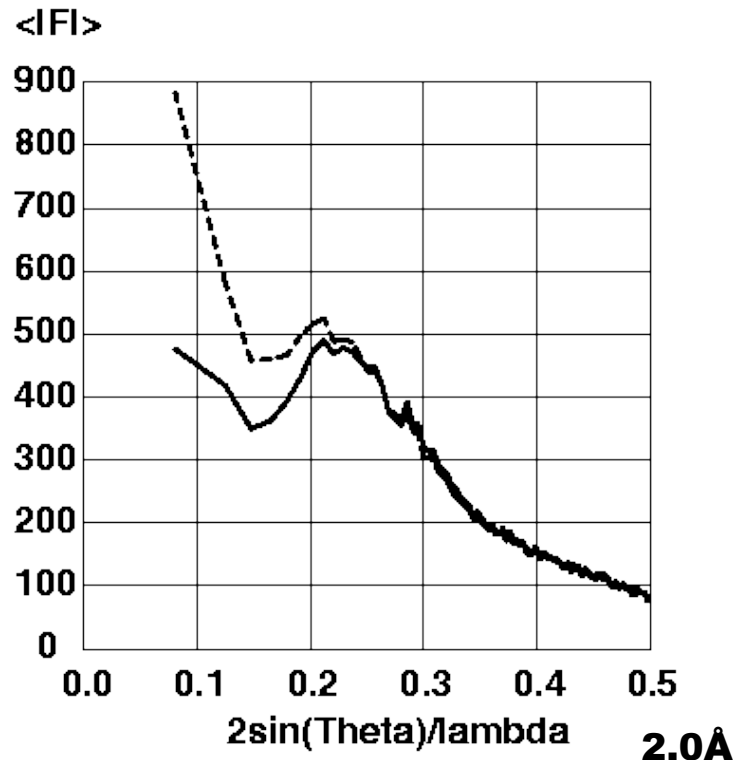
Solvent in Crystals



**Protein crystals are
30-70% solvent**

- The solvent contribution to the scattering is significant
- It affects data below 5Å and by 10Å is very significant
- Particularly relevant for Cryo-EM MR

Bulk Solvent and Structure Factors



Magnitude of observed structure factor amplitudes (solid line) and calculated structure factor amplitudes without a bulk solvent correction (dashed line) on an arbitrary scale against resolution

Dirk Kostrewa (1997). *CCP4 Newsletter* 34, 9

Solvent

Constant

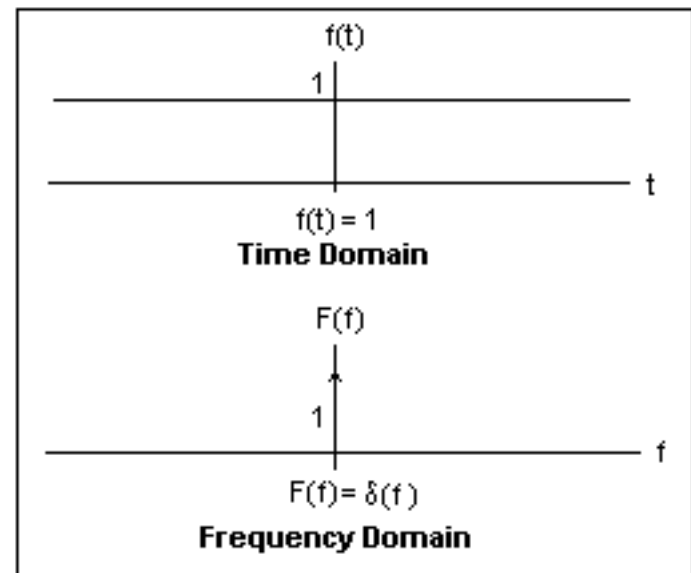


Contrast



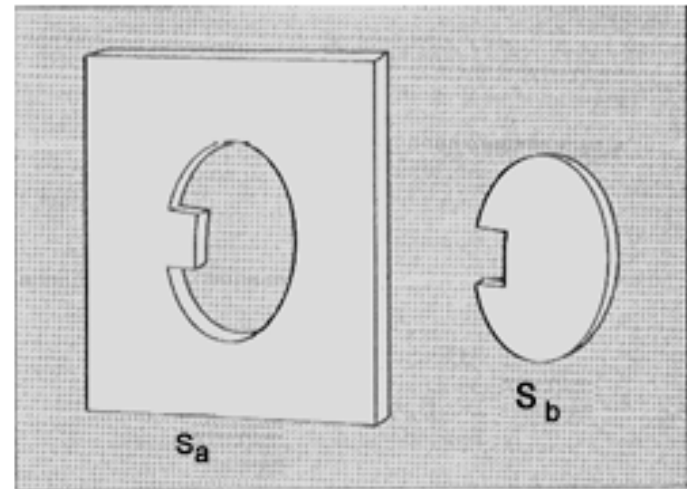
Solvent

- Fourier Transform of a constant in real space appears as the F000 term in reciprocal space
- F000 term cannot be measured in the diffraction experiment (direct beam)



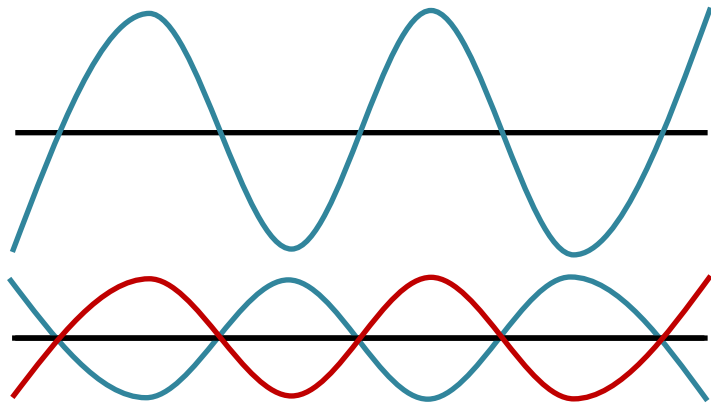
Babinet Principle

- Bulk solvent modeled as uniform electron density complementary to the protein
- Applying the Babinet correction results in lower protein amplitudes and no shift in protein phases
- Screens S_a and S_b produce the same diffraction patterns – except F_{000}



c Hayden-McNeil Publishing, Inc.

Babinet Solvent Correction



Protein contribution

**Babinet solvent
correction**

$$|F_{sol}| = -k_{sol} \exp\left(-B_{sol} \frac{\sin^2 \theta}{\lambda^2}\right) |F_P|$$

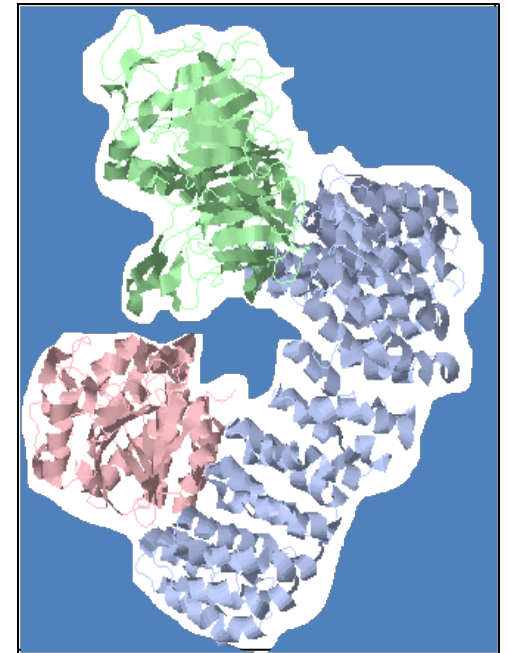
k_{sol} bulk solvent density height

B_{sol} B-factor: limits the correction to low resolution

Animation: Andrea Thorn

Mask Model of Bulk Solvent

- Fourier transformation of binary density mask
 - F_{mask} phase shifts the protein structure factors
- Like Babinet correction, a k_{mask} and B_{mask} are used to limit the correction to low resolution



**Jiang & Bruenger
(1994).**

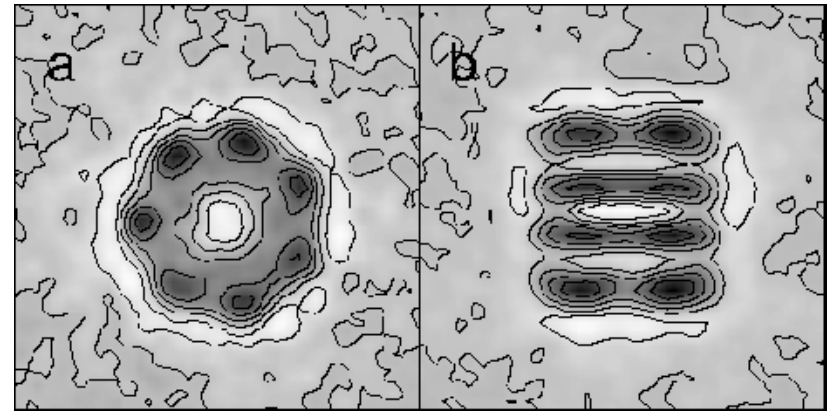
***J. Mol. Biol.* 243,**

100-115.

$$F_{calc} = k_{mask} \exp\left(\frac{-B_{mask}}{4d^2}\right) F_{mask}^{EM}$$

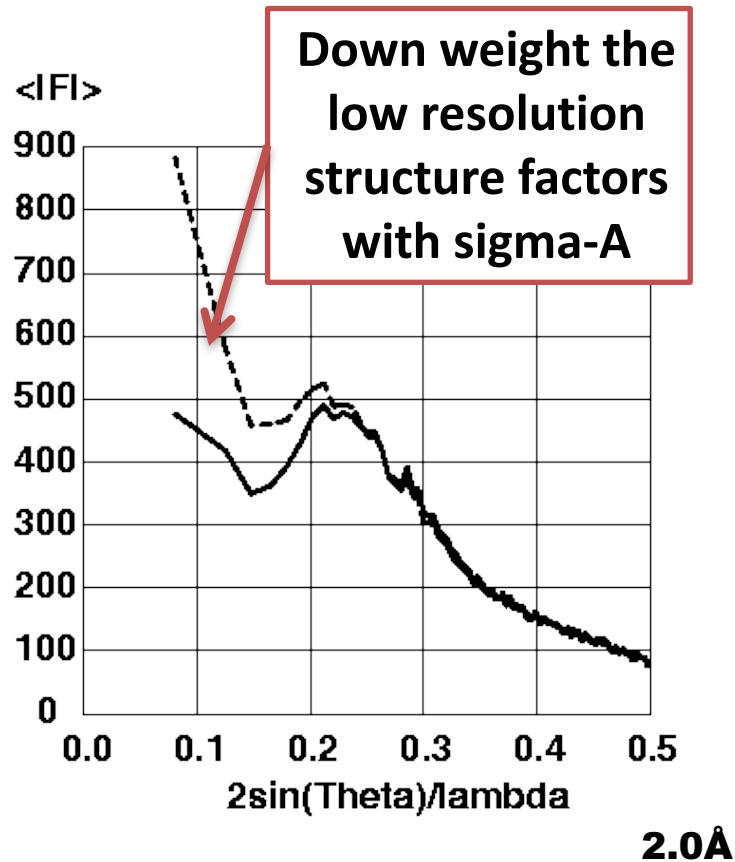
Cryo-EM contrast

- Density is averaged within a mask
- Outside the mask the density can be set to any value (zero)
- Density inside mask may be shifted by absolute value
- Contrast correction



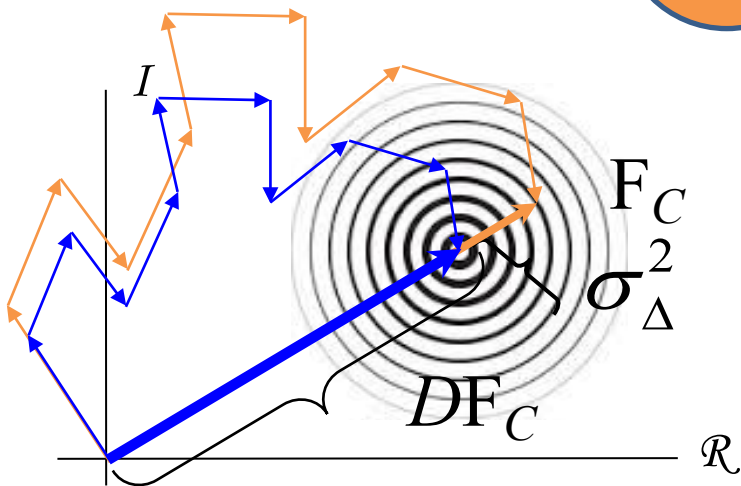
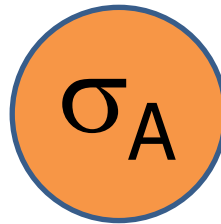
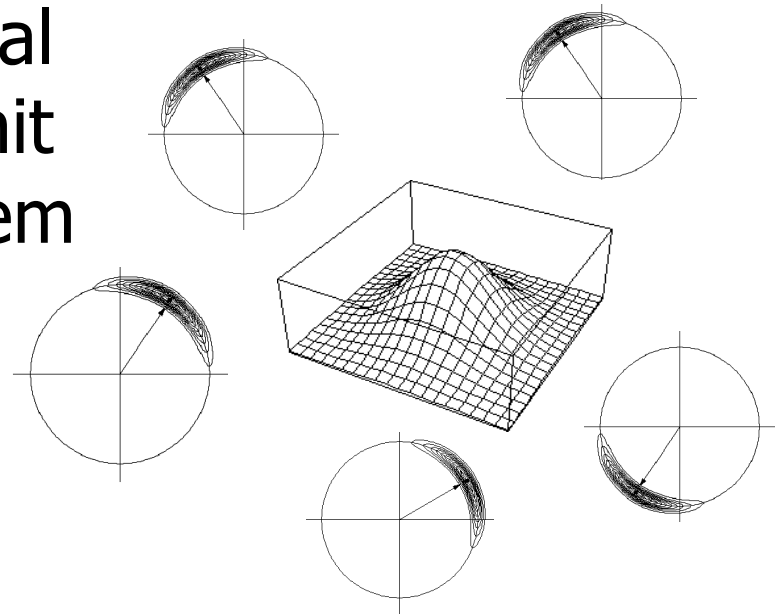
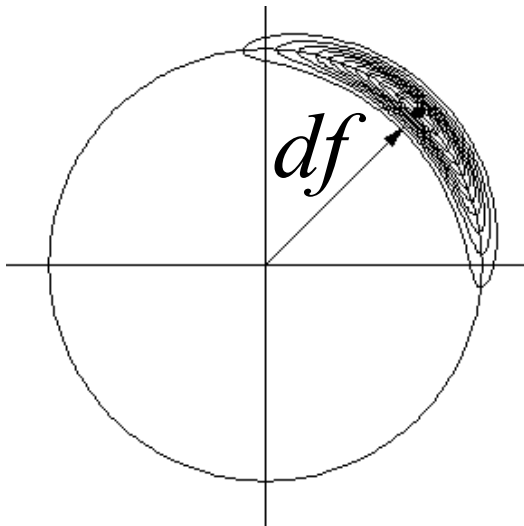
Cryo-EM averaged end and side views of GroEL (a,b), from Chen *et al.* (1994).

Bulk Solvent and Structure Factors

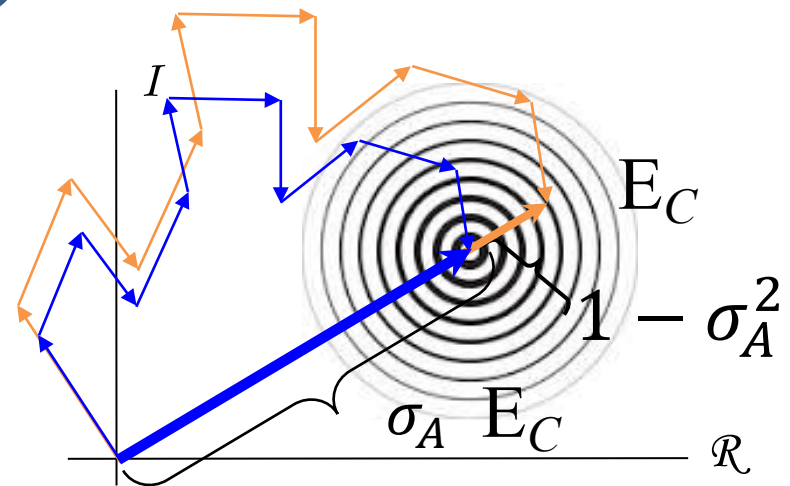


- Likelihood is optimized when the model and the estimates of error in the model are optimized
- Errors are estimated with σ_A

Central Limit Theorem



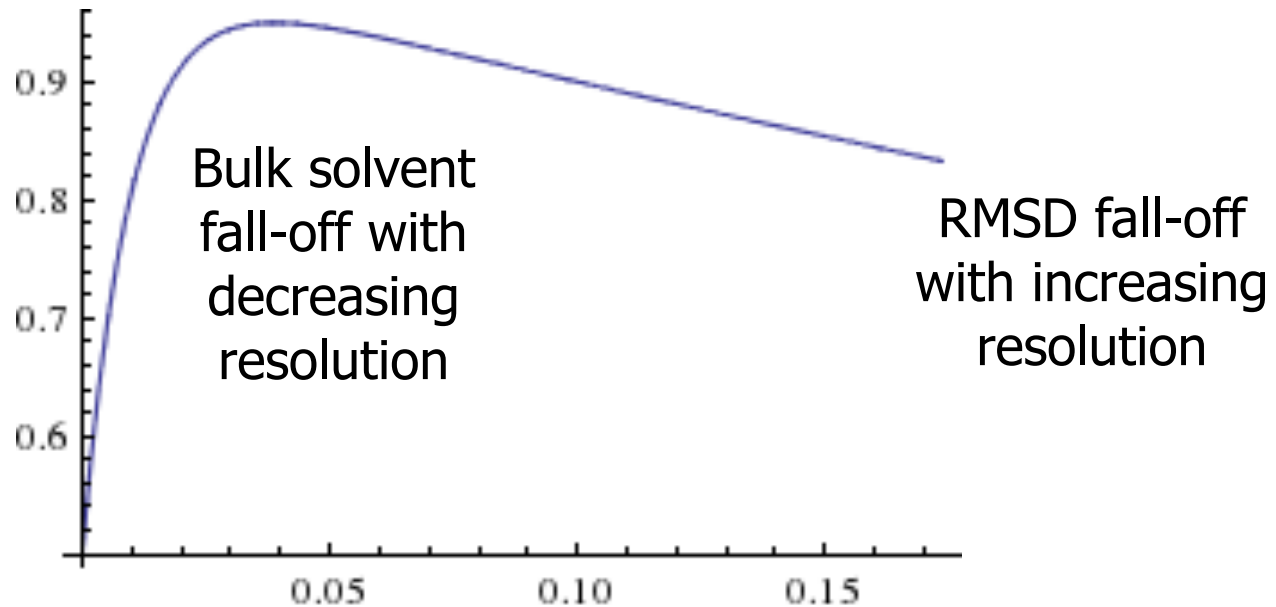
Structure Factor (F, D, σ_Δ)



Normalized Structure Factor (E, σ_A)

Bulk Solvent σ_A

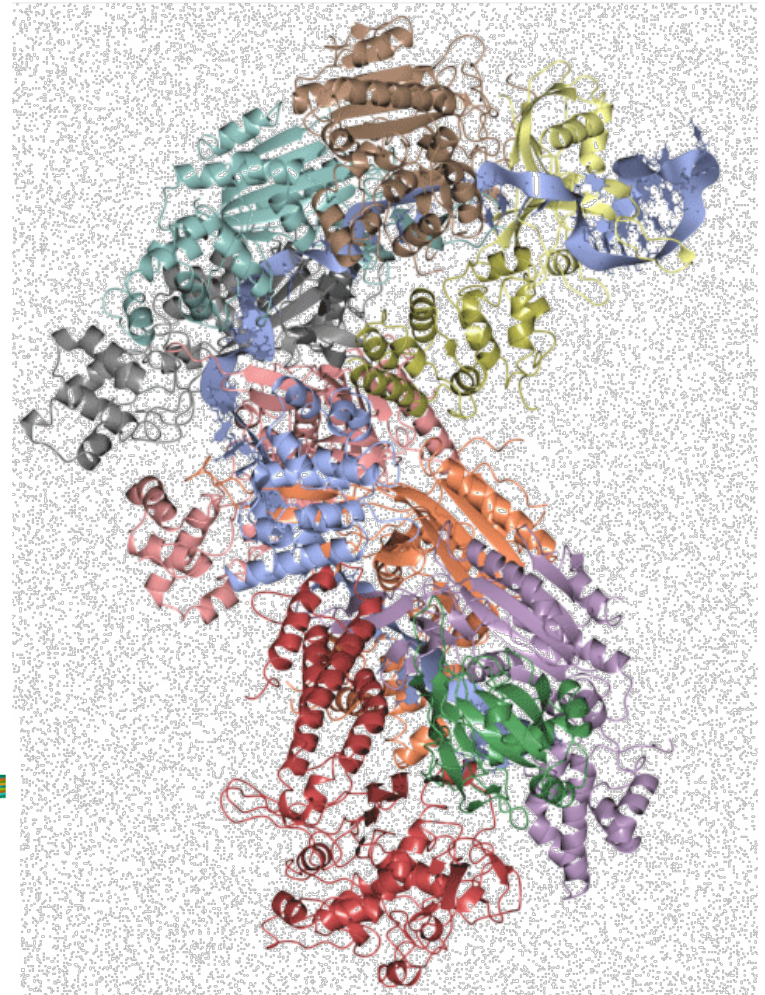
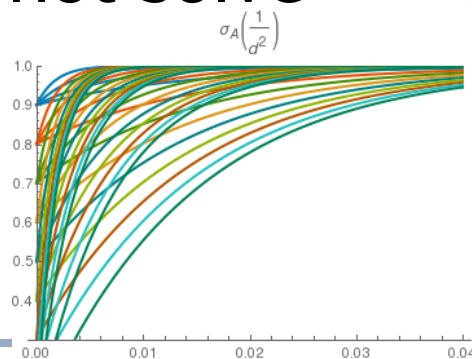
$$\sigma_A = f_p \left[1 - k_{sol} \exp\left(-B_{sol} \frac{\sin^2 \theta}{\lambda^2}\right) \right] \exp\left(\frac{8}{3} \pi^2 RMSD^2 \frac{\sin^2 \theta}{\lambda^2}\right)$$



Cascade

ksol Bsol	0.1	0.3	0.5	0.7	0.9
70	118	118	118	77	22
140	118	118	118	33	32
282	118	120	27	33	32
562	118	26	26	33	33

Default values of $k_{sol}=1.05$,
 $B_{sol}=502$ did not solve



Database Study

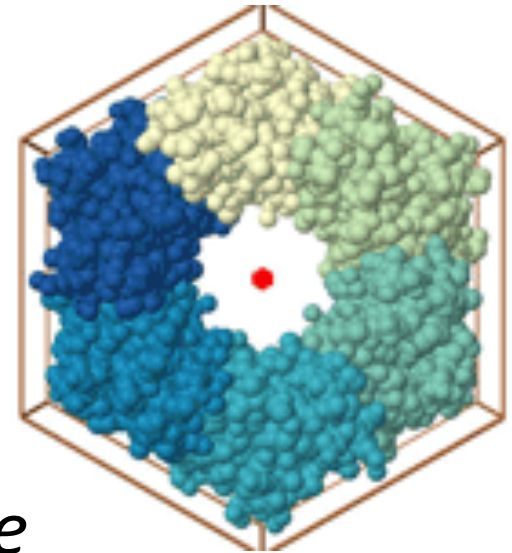
ksol Bsol	0.1	0.3	0.5	0.9	1.3
70	41	51	45	24	14
140	39	47	47	29	21
282	40	45	46	37	30
562	40	47	46	37	37
Default values of $k_{sol}=1.05$, $B_{sol}=502$					

33

- Eleven “good” EM maps taken from EMDB
- EM maps are better than coordinates for predicting structure factors at low resolution
- The σ_A needs to be increased for MR with EM

Non-crystallographic symmetry

- Phaser detects point group symmetry in the model when it is entered as coordinates
 - Algorithm by Gabor Bunkoczi
 - This symmetry used to filter the list for unique solutions
- *Phenix routines will now generate matrices from recorded symmetry*
 - *Plans to use this to filter solutions*



The Phenix Project

Lawrence Berkeley Laboratory

**Paul Adams, Pavel Afonine,
Dorothee Leibschnher, Nigel
Moriarty, Billy Poon, Oleg
Sobolev**



Los Alamos National Laboratory

**Tom Terwilliger,
Li-Wei Hung**



**Randy Read, Airlie McCoy,
Gábor Bunkóczi, Rob
Oeffner, Tristan Croll**

Cambridge University



Duke University

**Jane & David Richardson,
Bradley Hintze,
Christopher Williams,
Vincent Chen,
Michael Prisant**



**In NIH/NIGMS funded
Program Project**