

Molecular Replacement

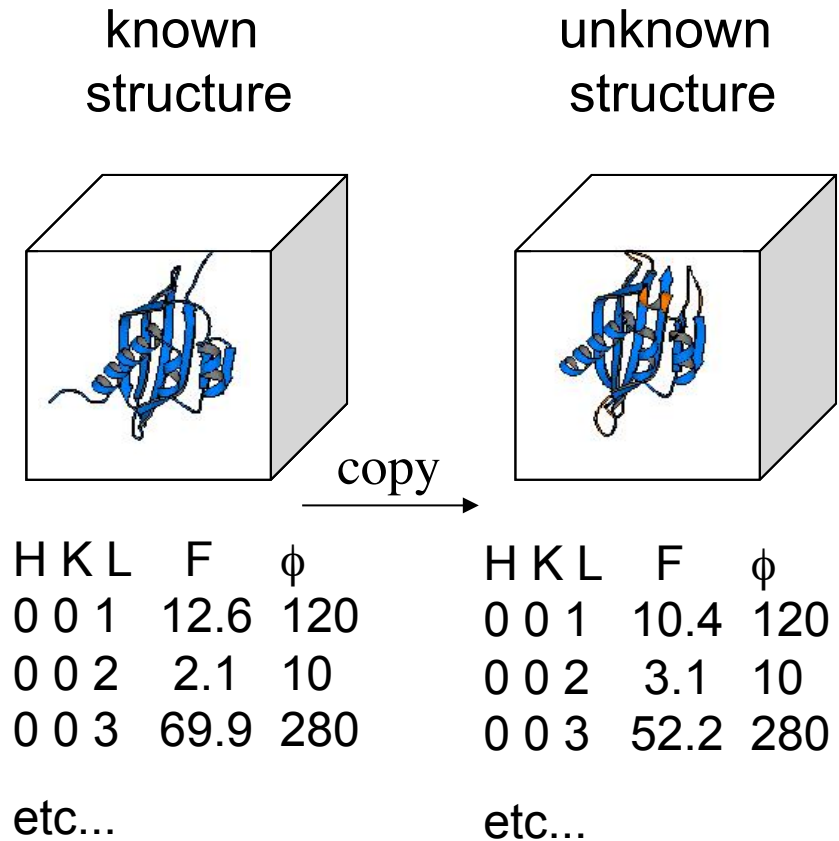
Airlie McCoy



UNIVERSITY OF
CAMBRIDGE

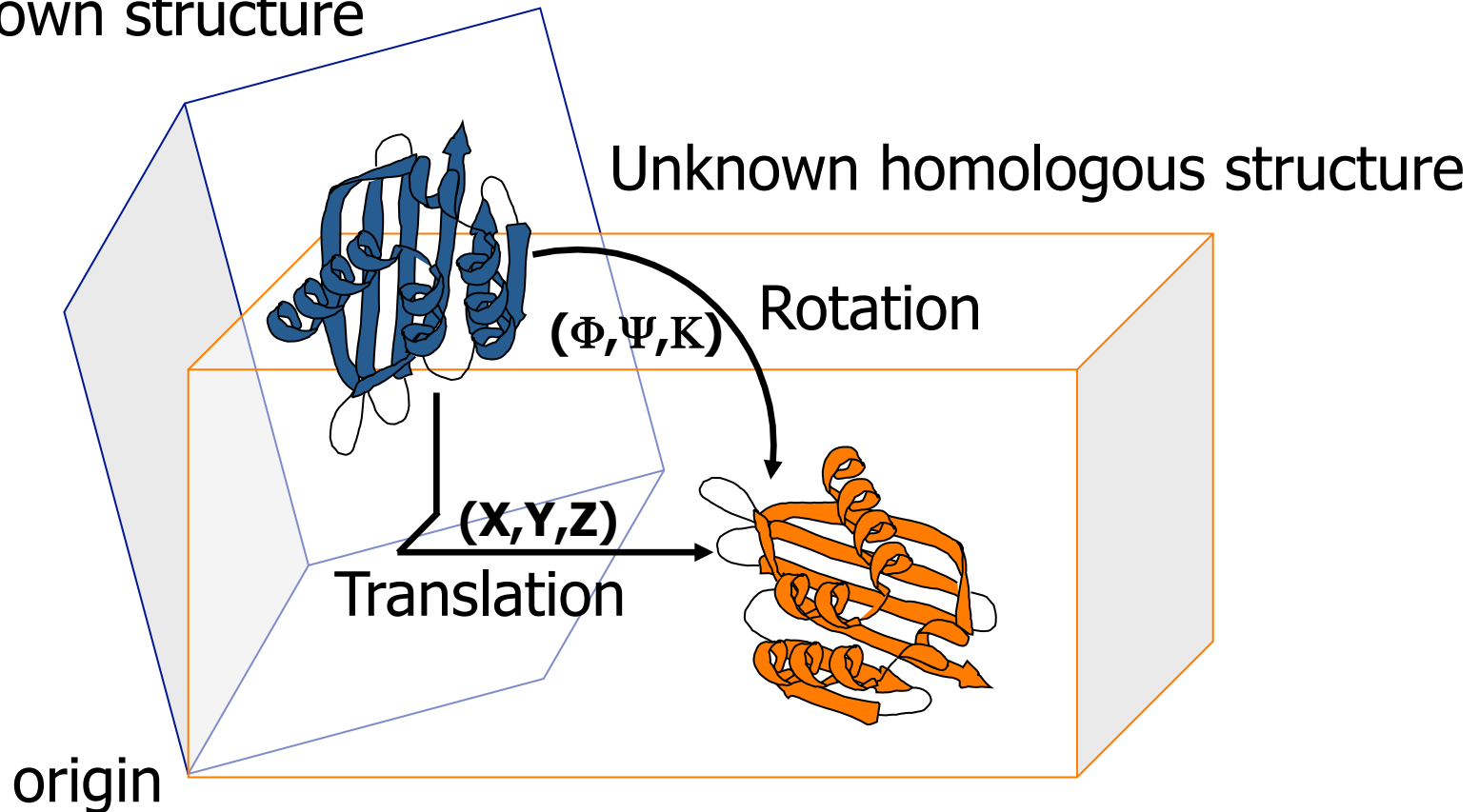
Molecular Replacement

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



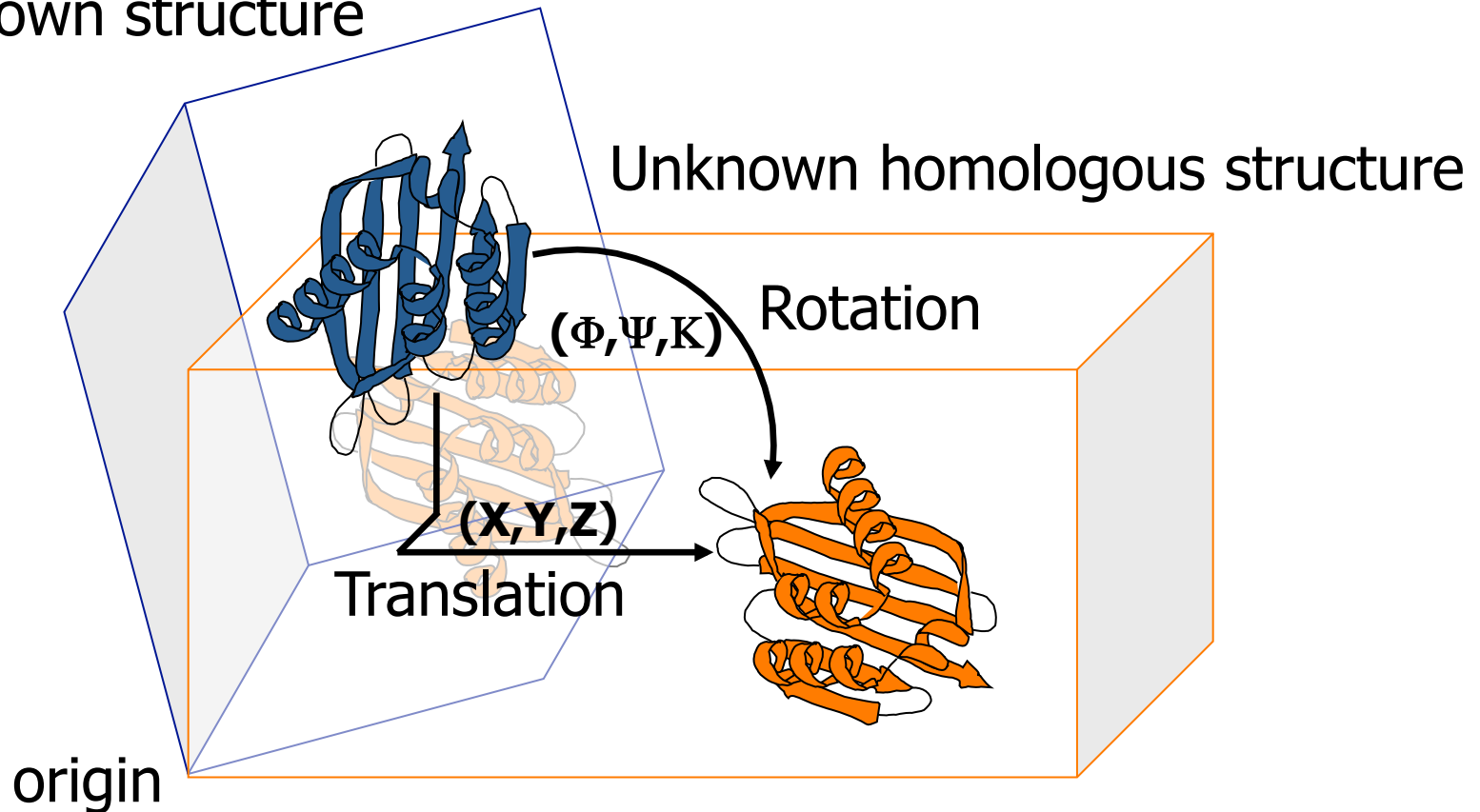
Molecular Replacement

Known structure

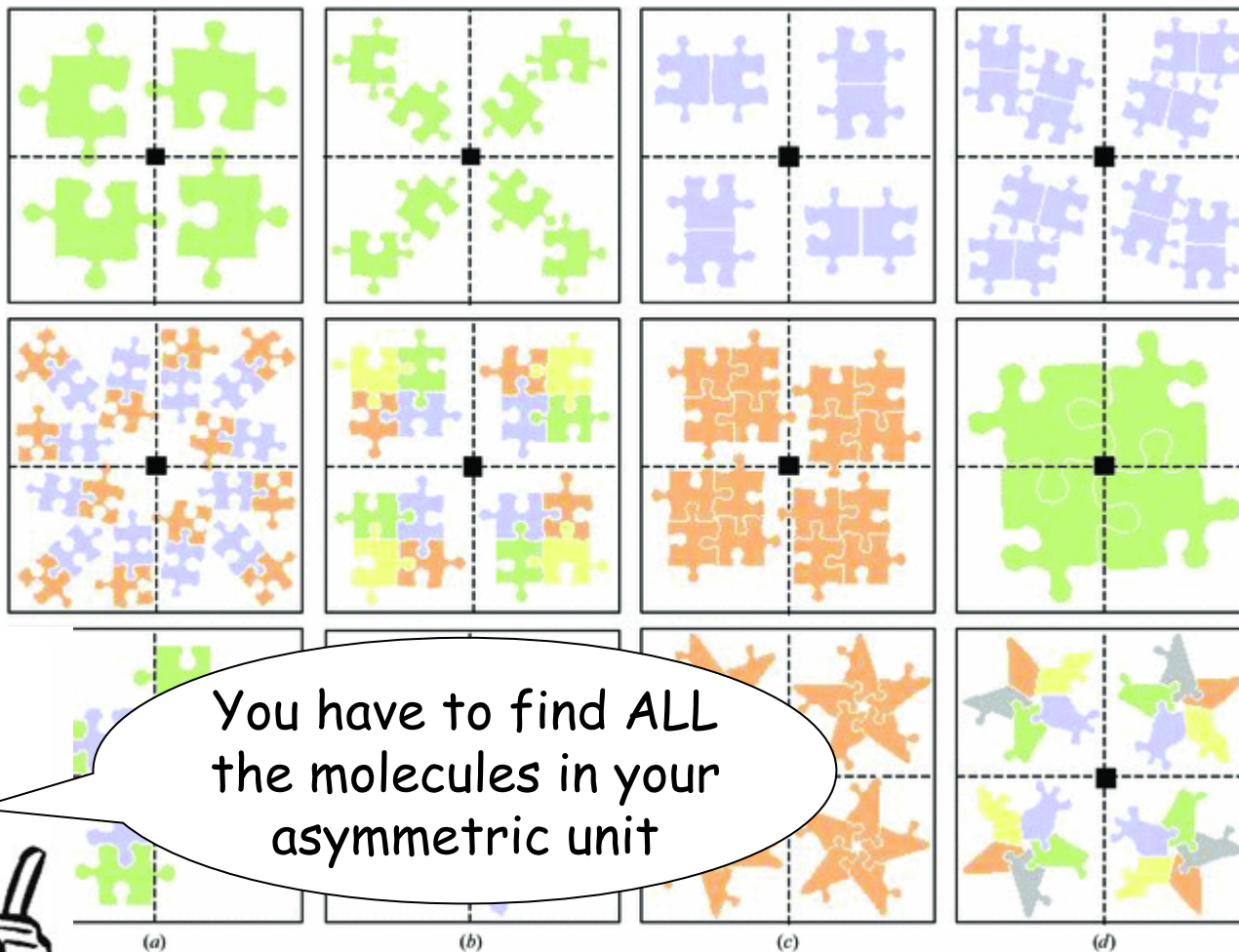


Molecular Replacement

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

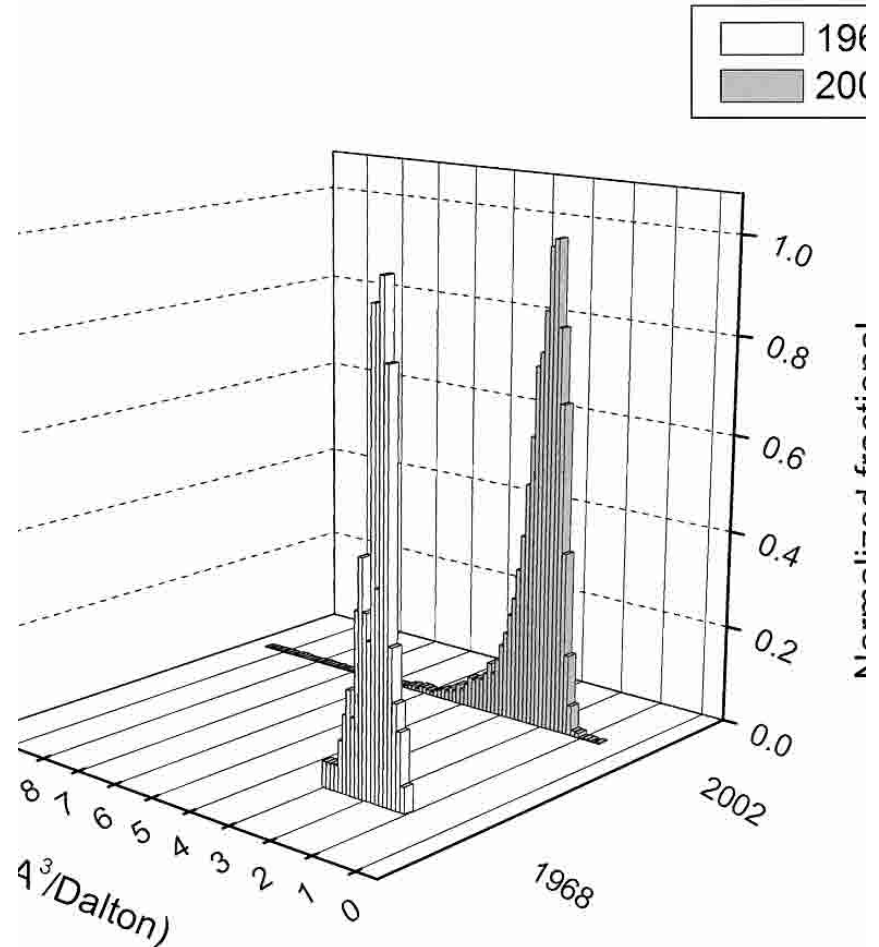
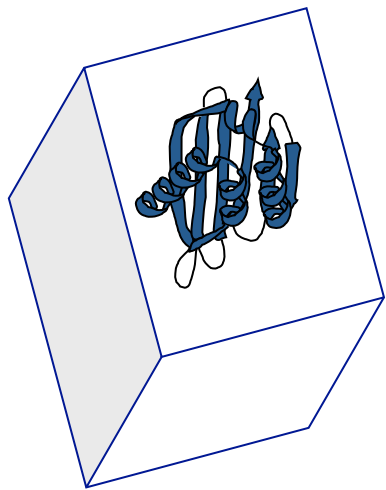


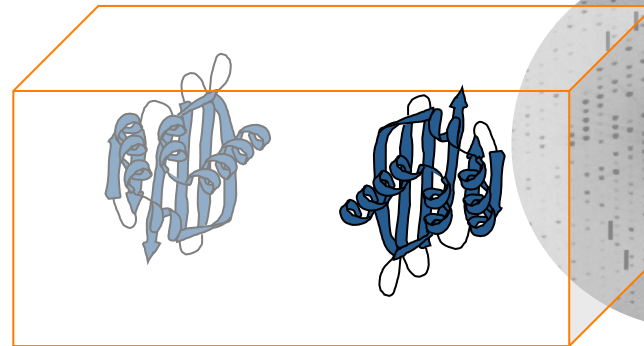
Figure 1: Kantardjieff and Rupp (2003)



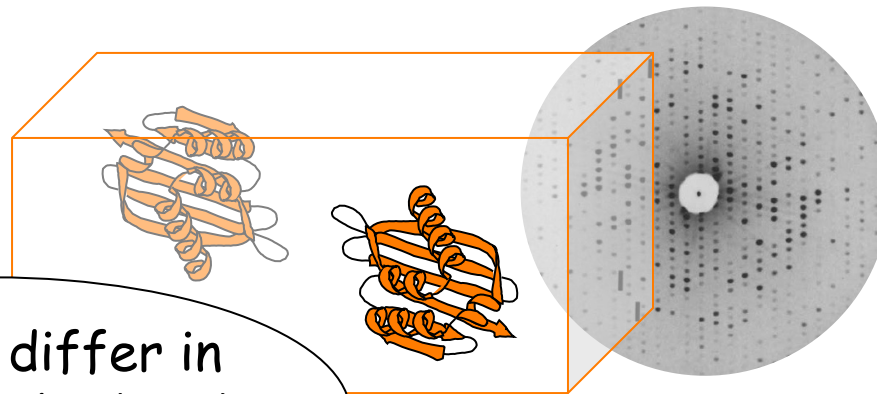
SEARCH

(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

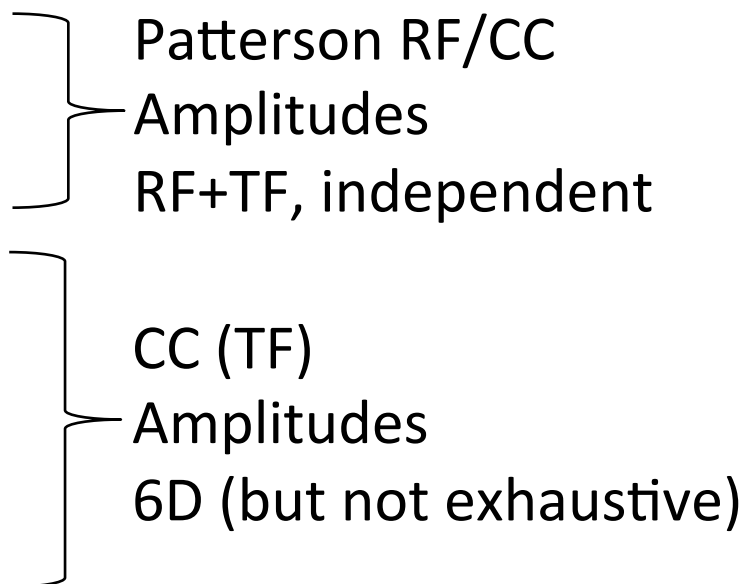
- Each molecule needs 6 parameters (6D)
 - An exhaustive search is big
 - All angles sampled at 2.5° ; $N_{\text{rot}} = 1.5 \times 10^6$
 - All positions sampled at $1\text{\AA}$ in a $100\text{\AA}$ cubic cell;
 $N_{\text{tra}} = 1.0 \times 10^6$
 - 6 dimensional search is $N_{\text{rot}} \times N_{\text{tra}} = 1.5 \times 10^{12}$ 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

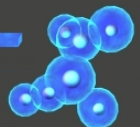
- MolRep
- AMoRe
- CNS
- EPMR
- COMO
- SOMoRe
- Queen of Spades



Maximum likelihood
Intensities
RF+TF, cumulative, amalgamation

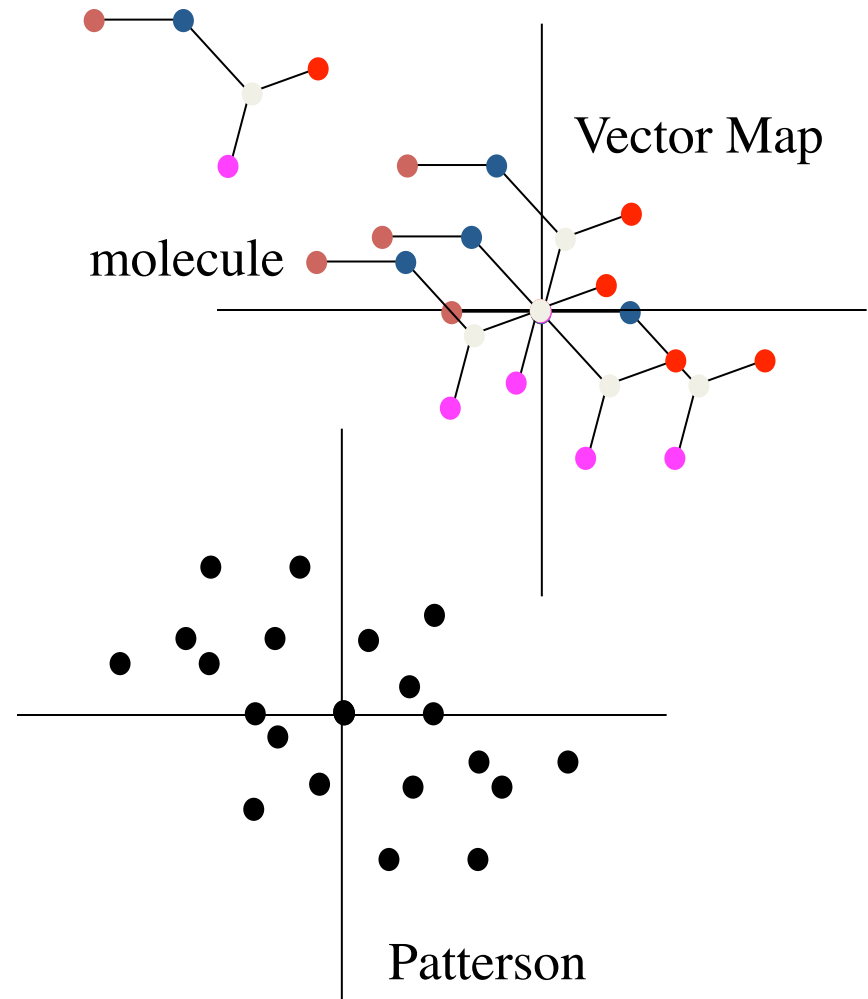
Some Important Features of Phaser

- 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

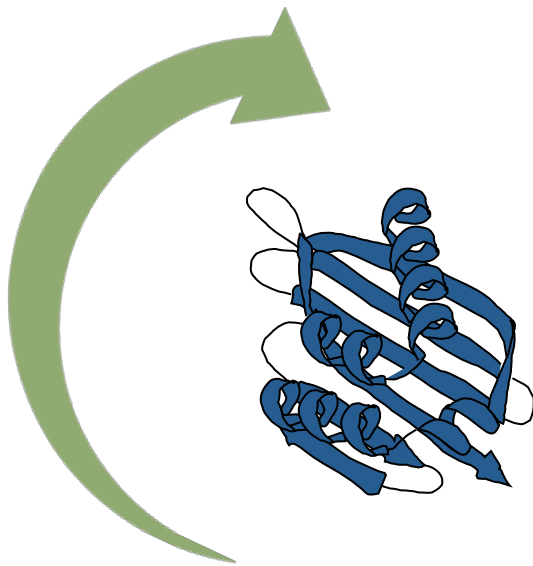
- Use probability
- Probabilities account for errors
 - Patterson methods cannot do this

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

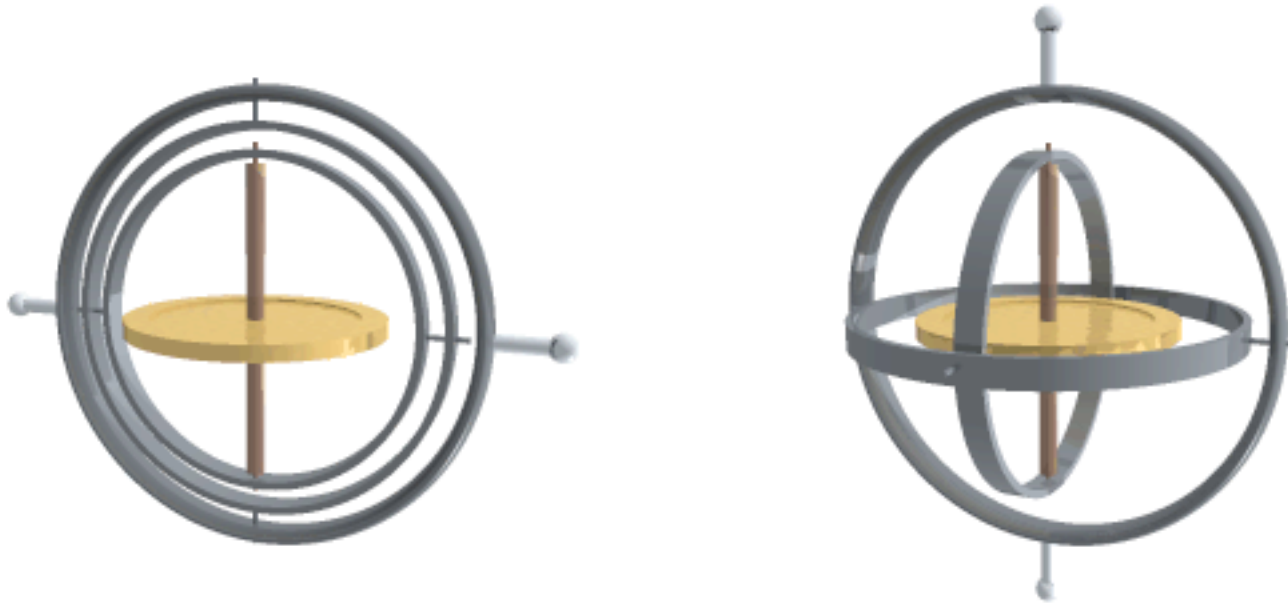
E_e and D_{obs} are defined as in (Read & McCoy, 2016); E_e is the effective E , representing information derived from E_{obs}^2 , and D_{obs} represents the reduction in correlation between observation and E_e arising from experimental error; $E_{obs}^2 = I_{obs}/(\varepsilon \Sigma_N)$ where ε and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

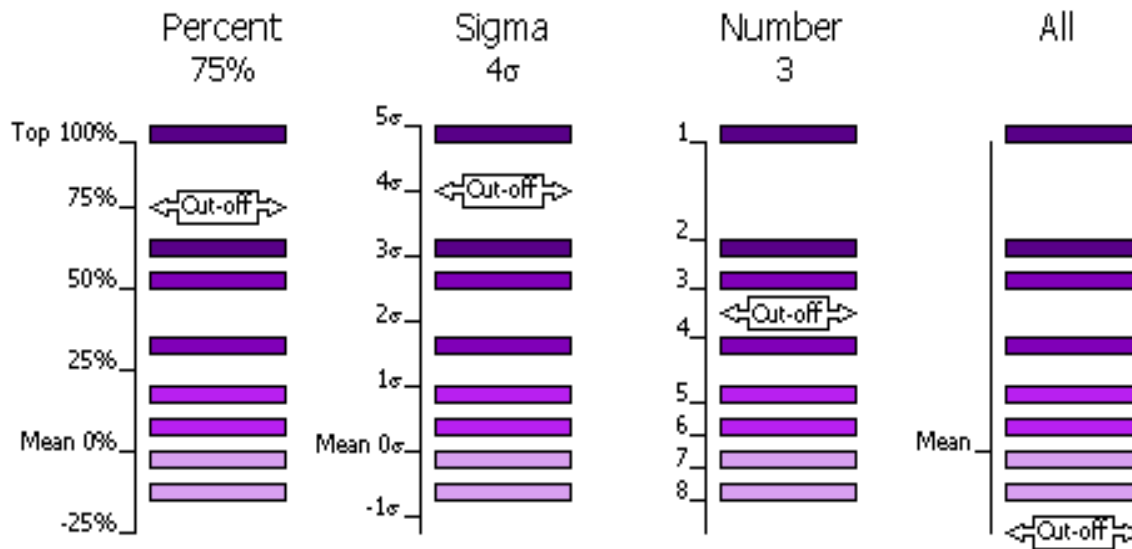


Euler Angles



Peak selection

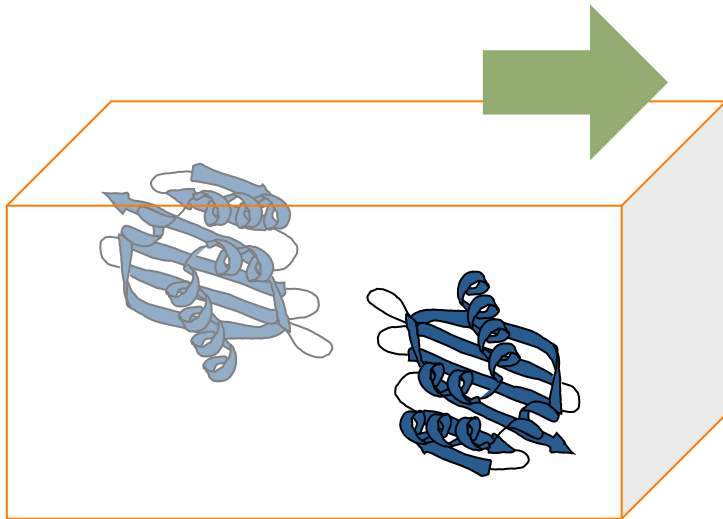
- Must choose a selection criteria to carry **potential** solutions through to the next step



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

Brute translation function search

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

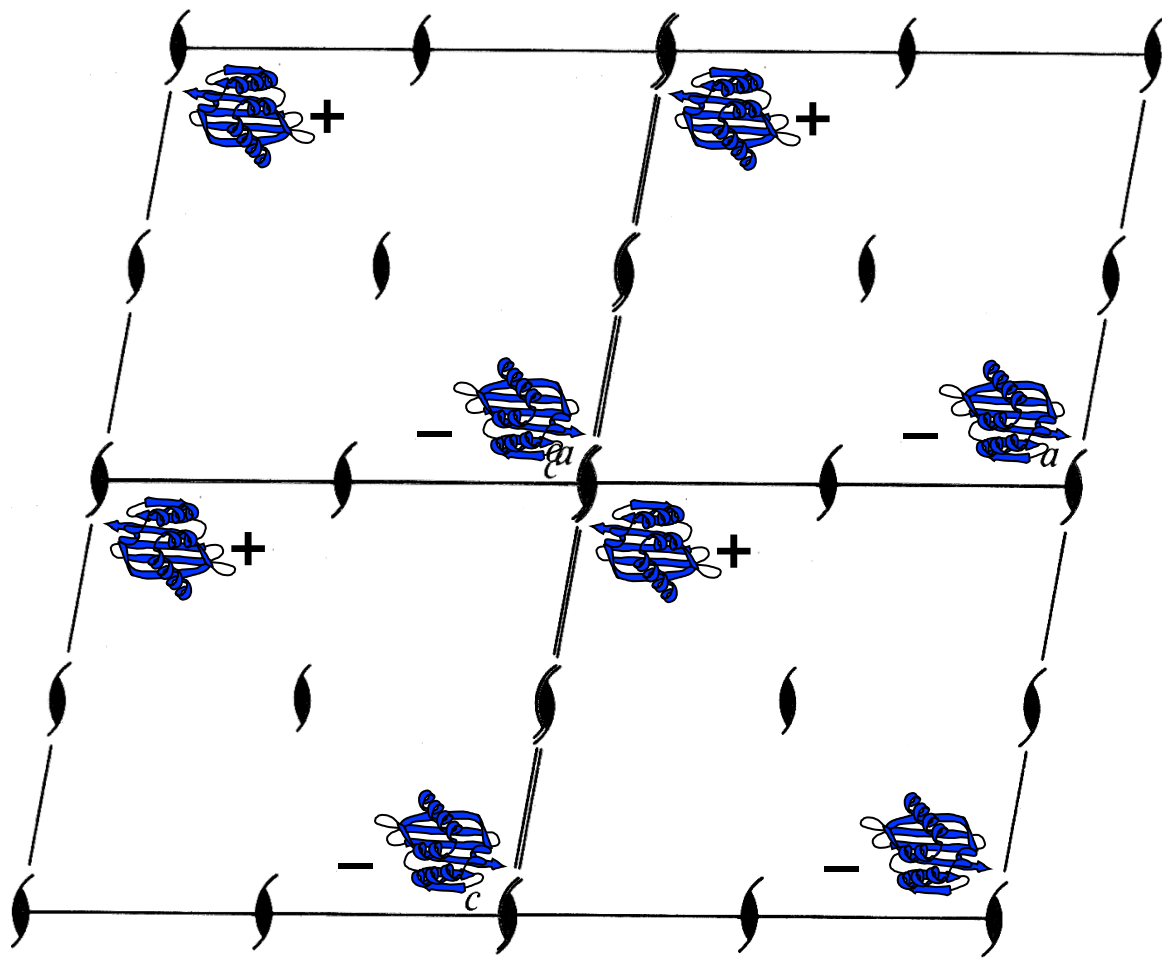


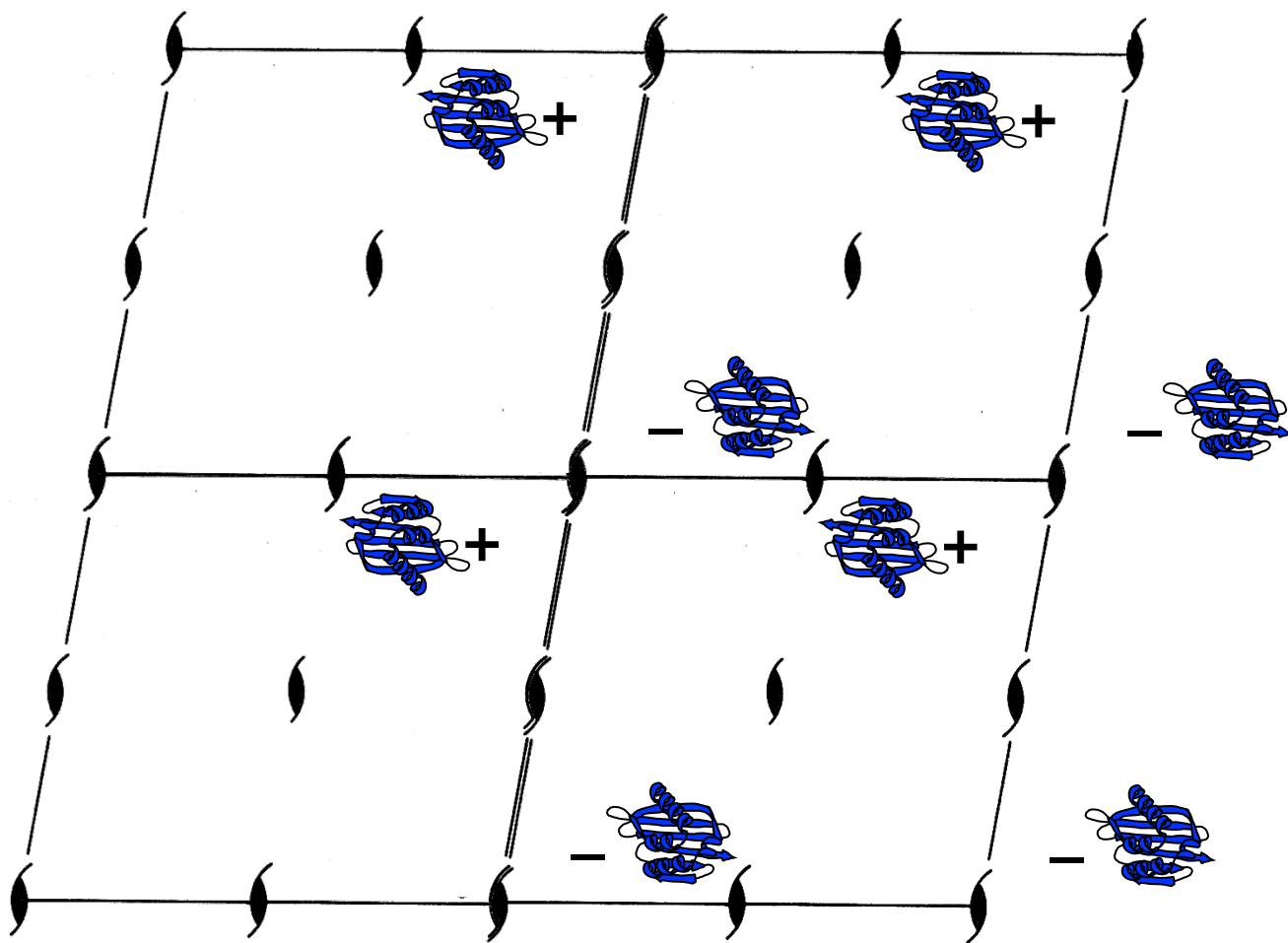
When is a model correctly placed?

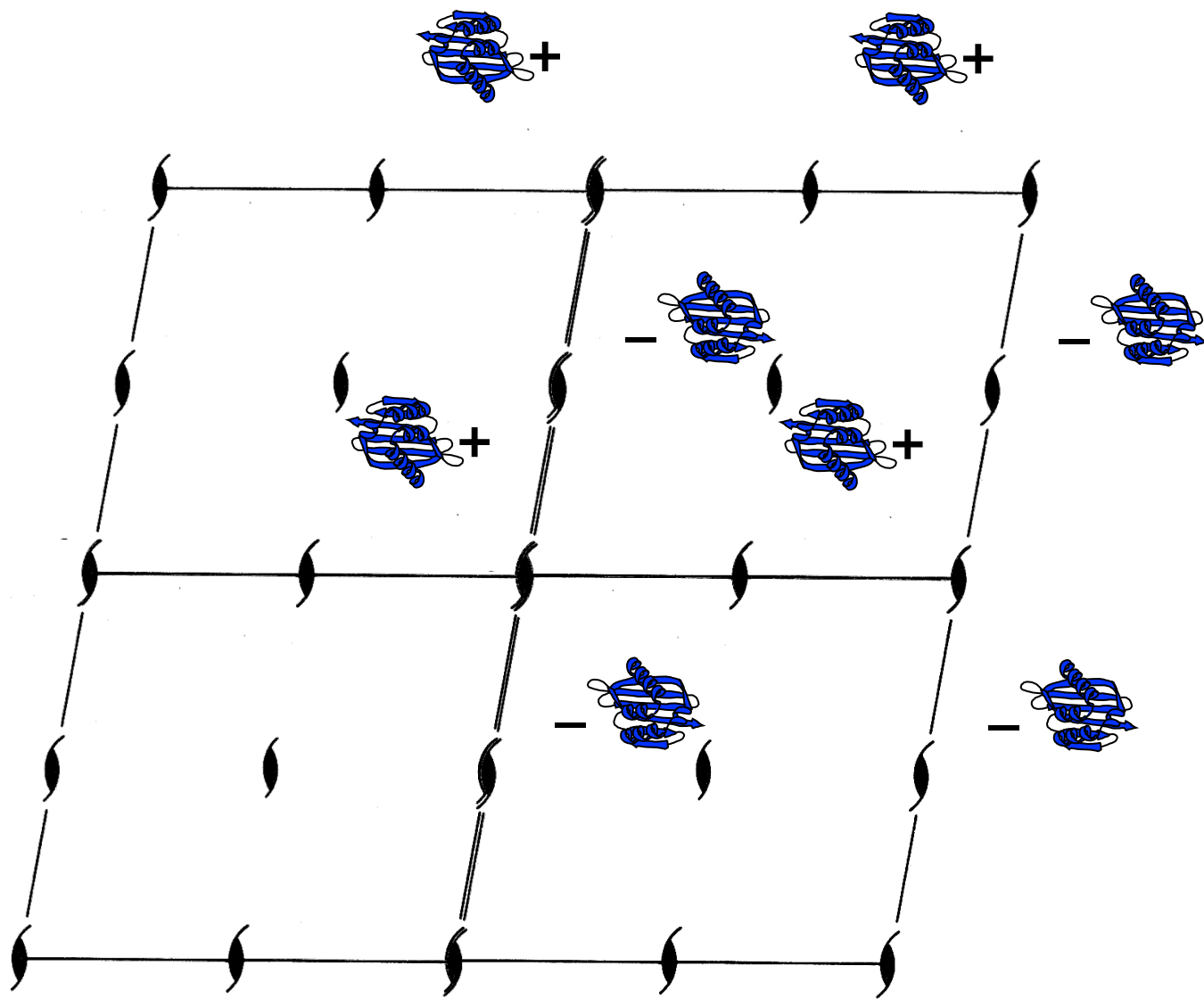
TF Z-score	Solved?
< 5	no
5 - 6	unlikely
6 - 7	possibly
7 - 8	probably
> 8	definitely

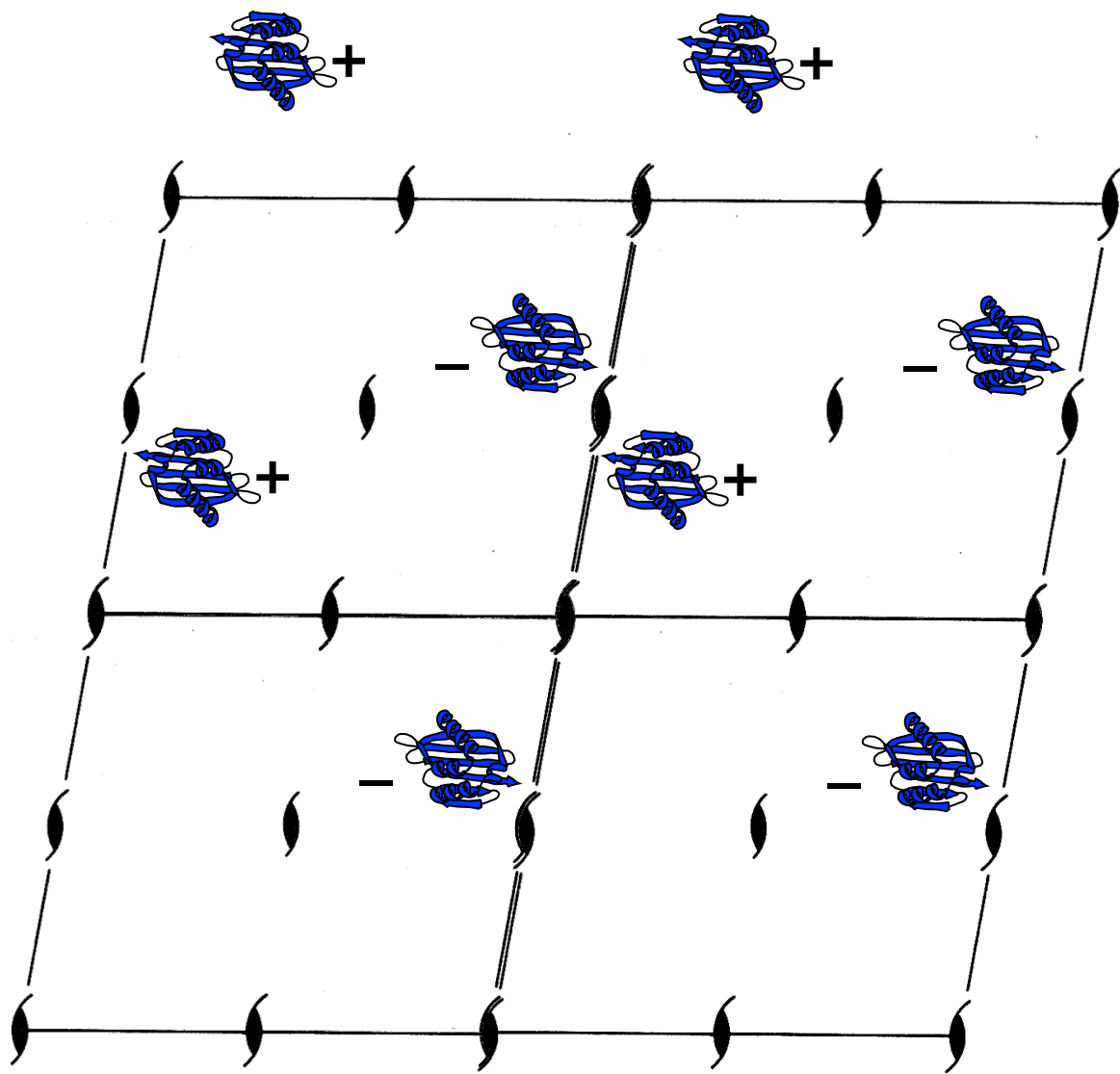
Origins

- 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
-



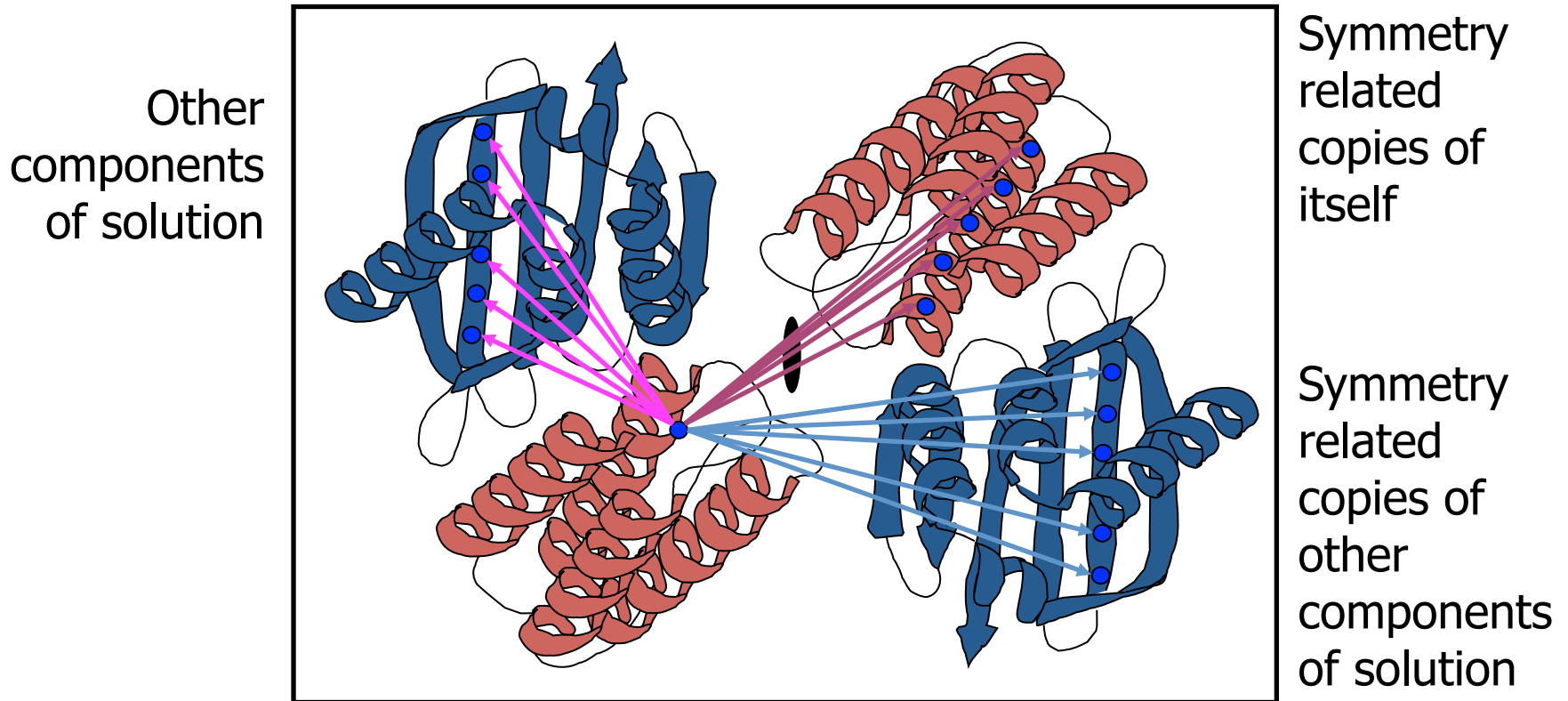






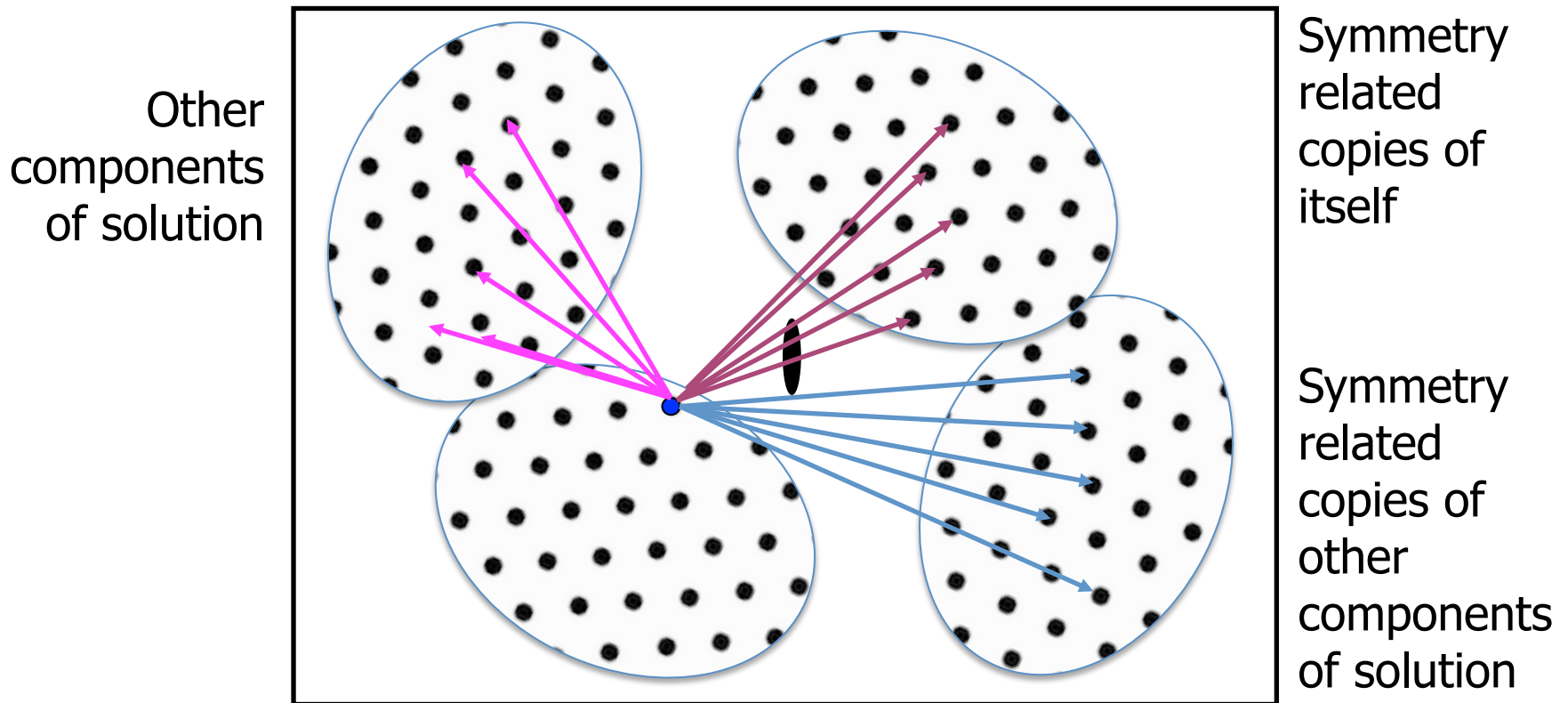
Packing Function

- $\text{C}\alpha$ clash test



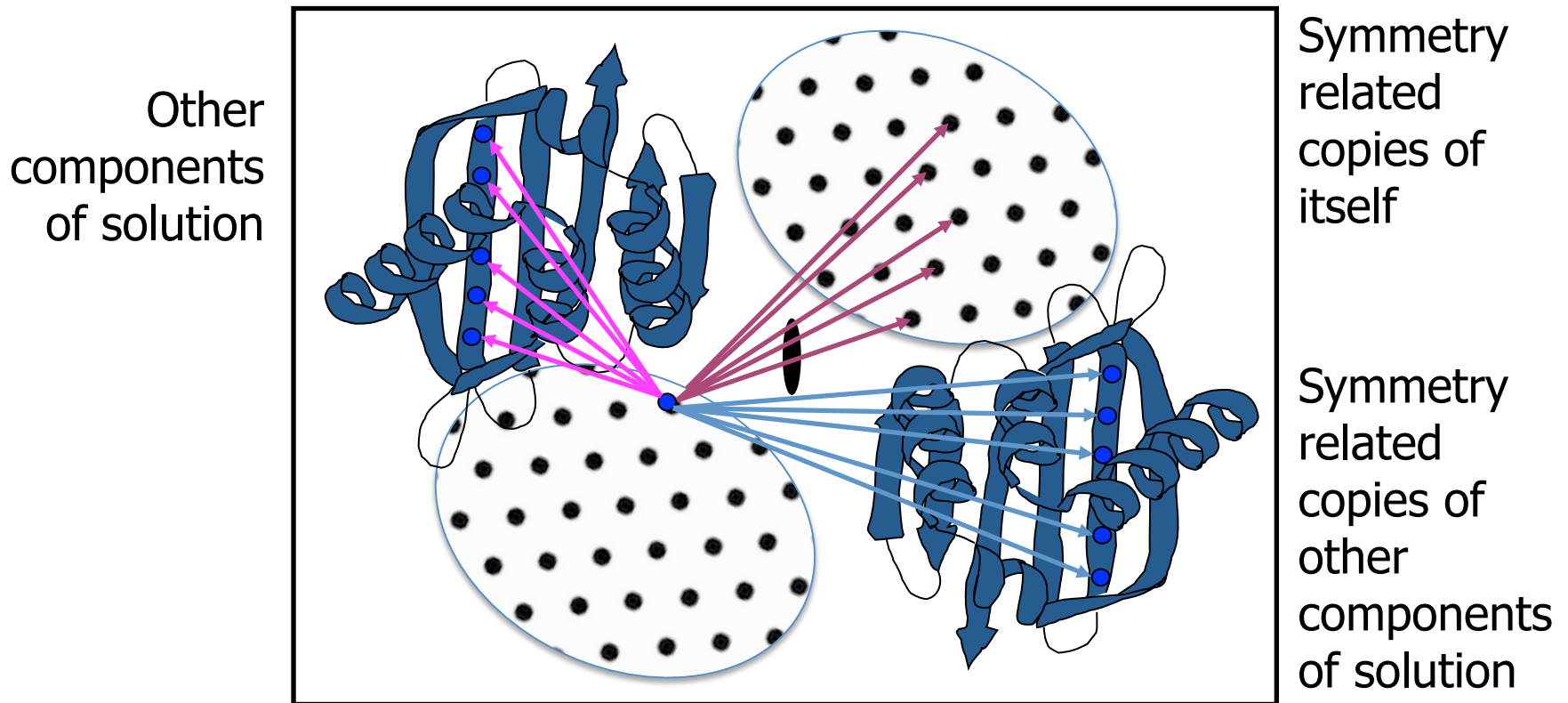
Packing Function

- Hexagonal Grid clash test



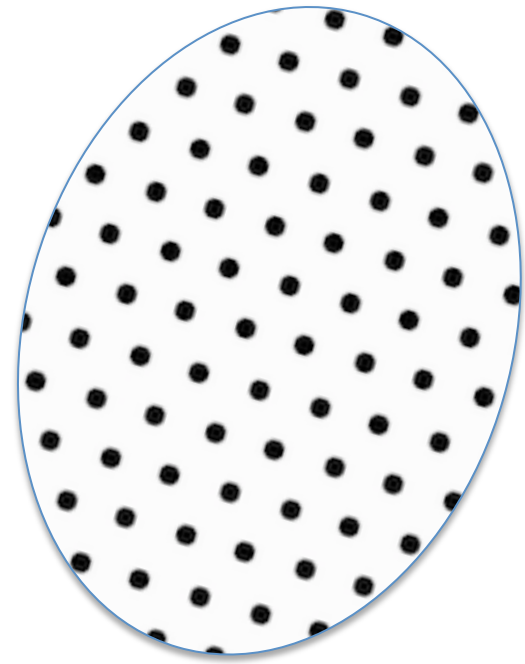
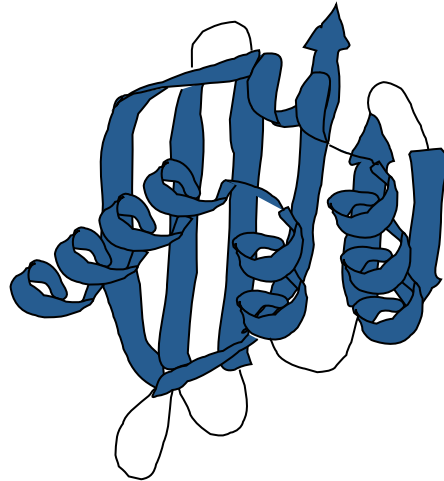
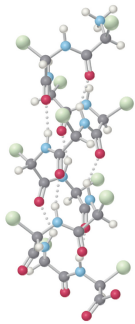
Packing Function

- Mixed Hexagonal Grid and C α clash test



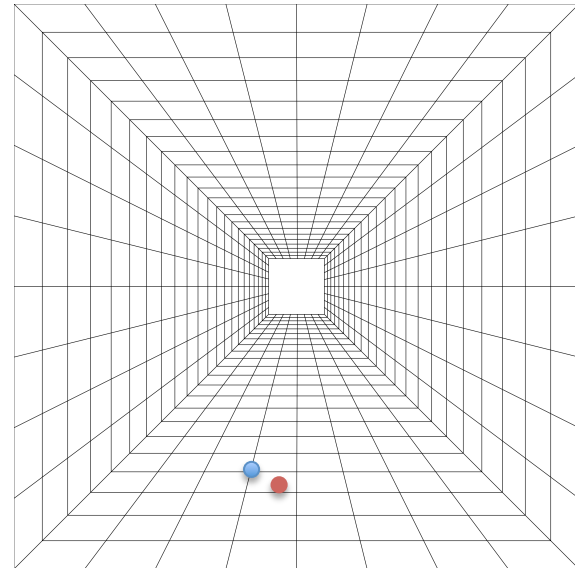
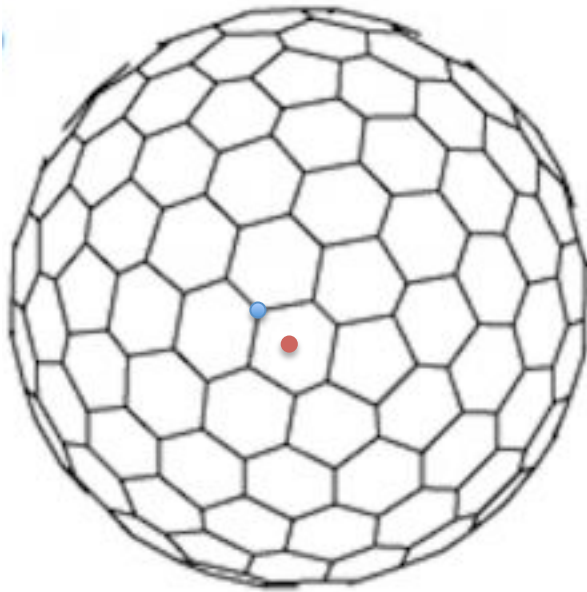
Packing Function

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



Refinement

- Rotation and Translation searches are scored on a grid



1. Rotation search

choose top peaks (eg > 75% of best score)

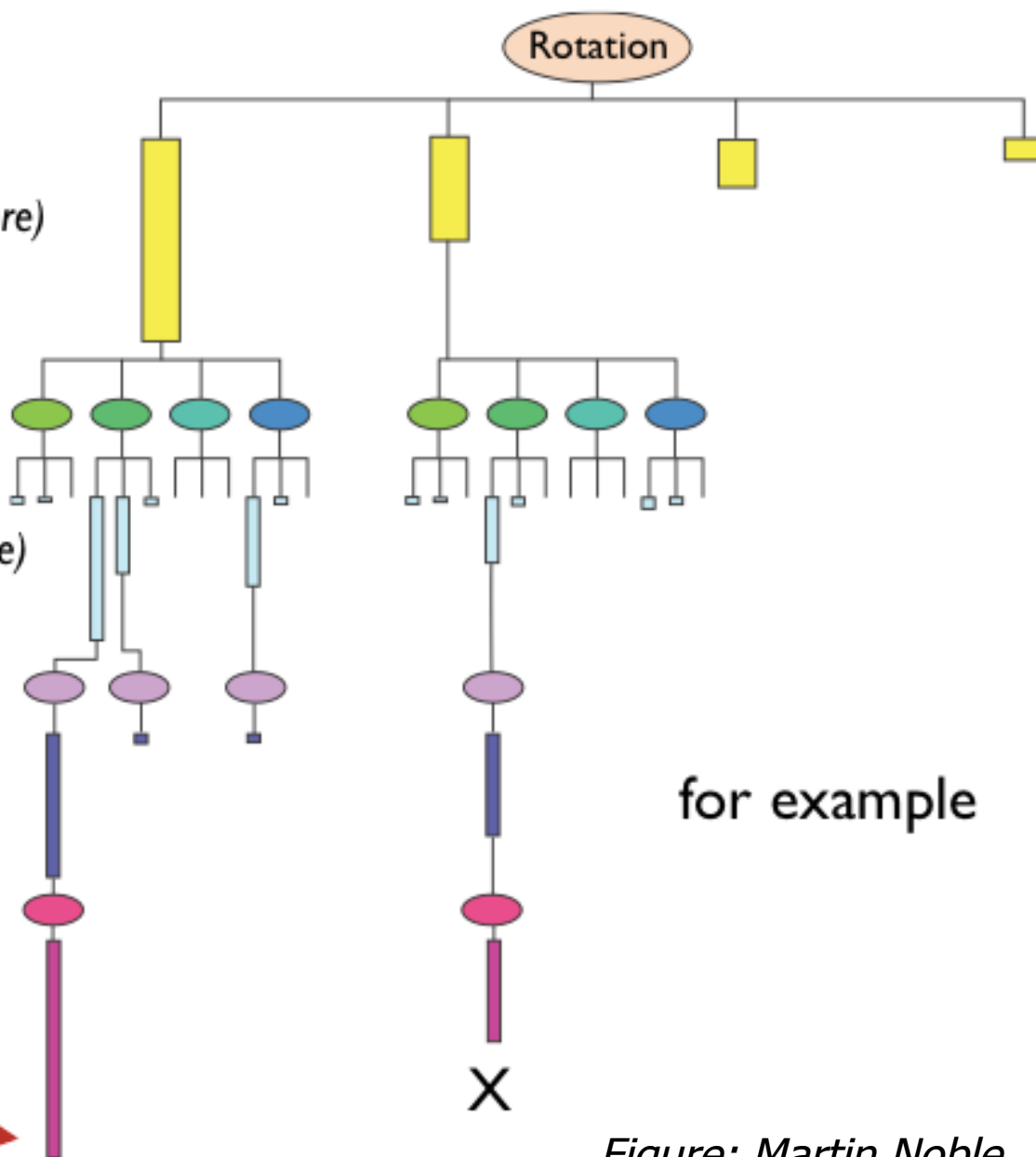
2. Translation search in all possible space groups

choose top peaks (eg > 75% of best score)

3. Packing check

4. Rigid body refinement to optimise score

This one



for example

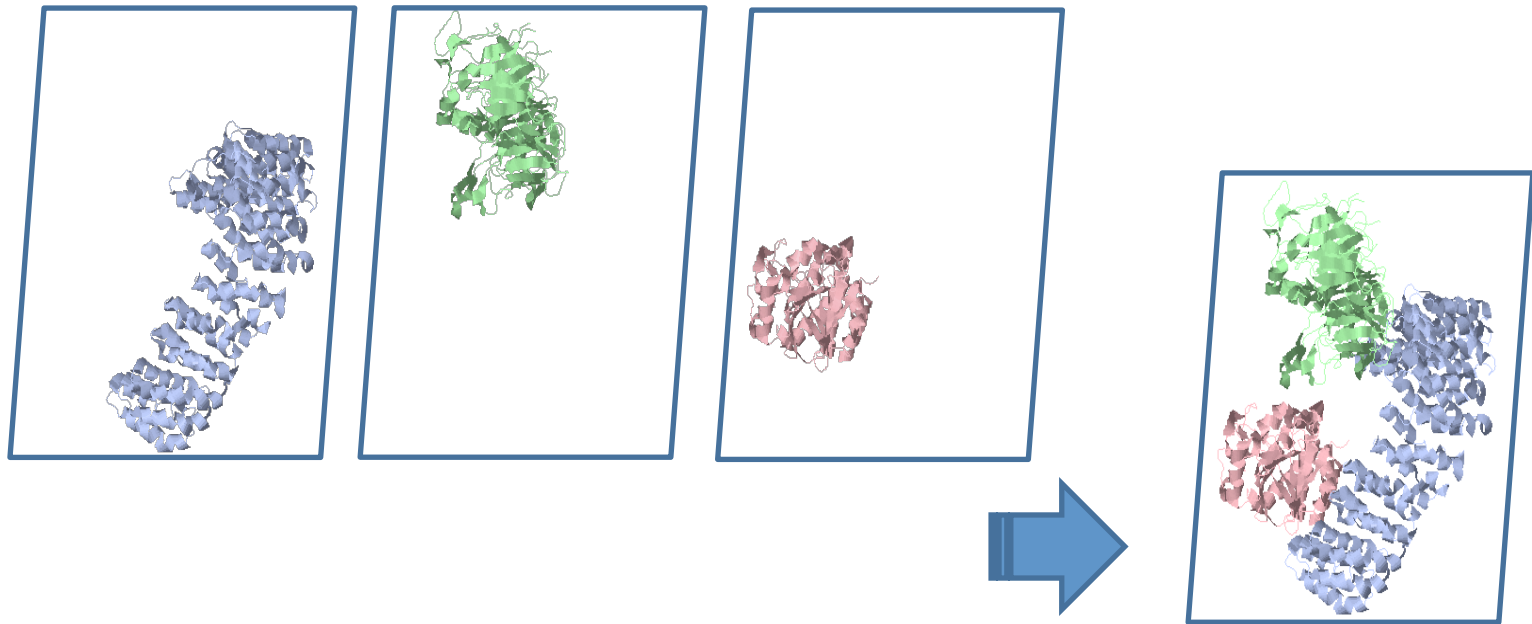
Figure: Martin Noble

Using Partial Structure

- The MLRF and MLTF can use models that have already been placed in the asymmetric unit
 - Patterson RF cannot account for placed models
 - The Patterson Correlation Coefficient 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
-

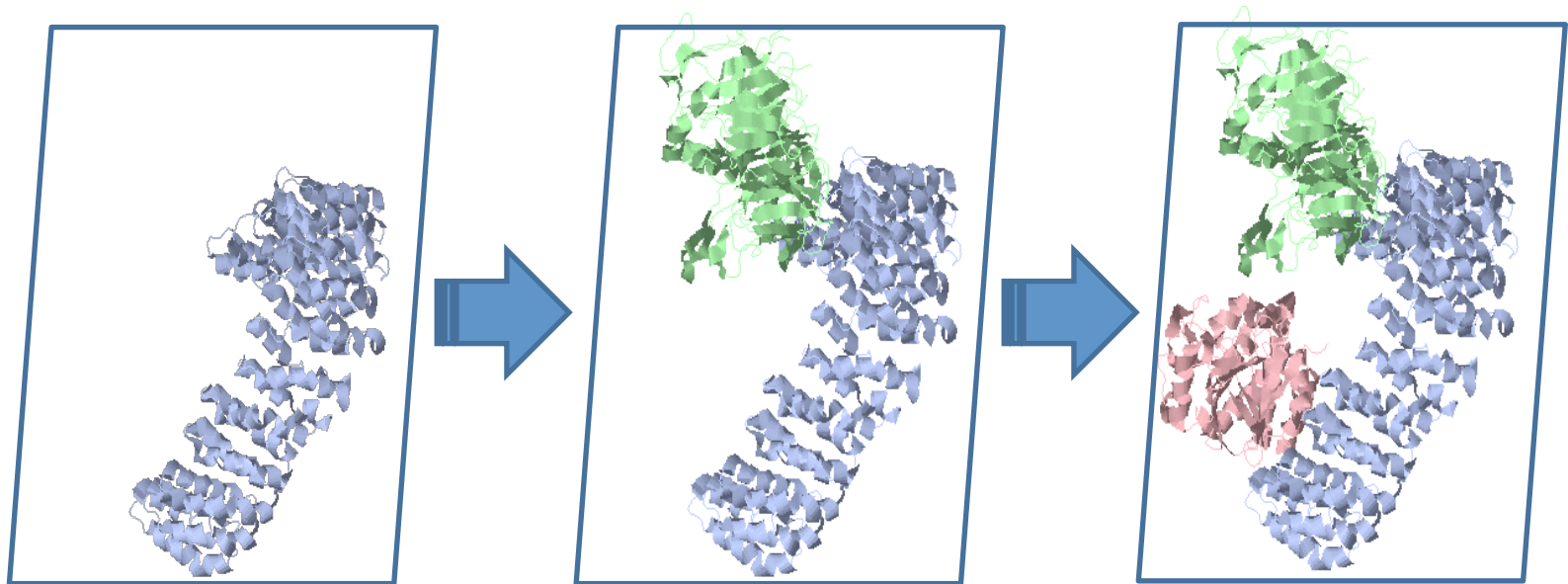
Searches for multiple components

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

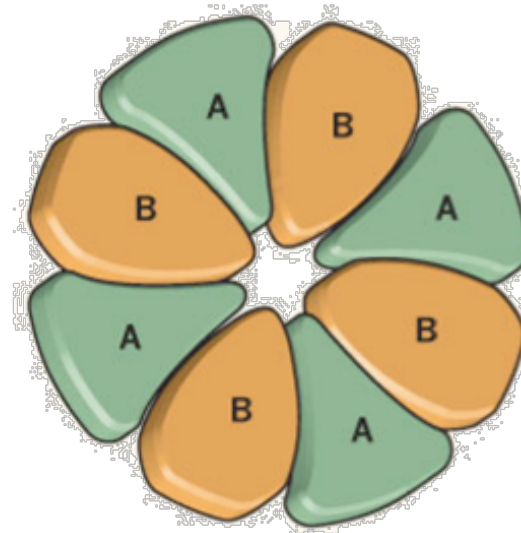
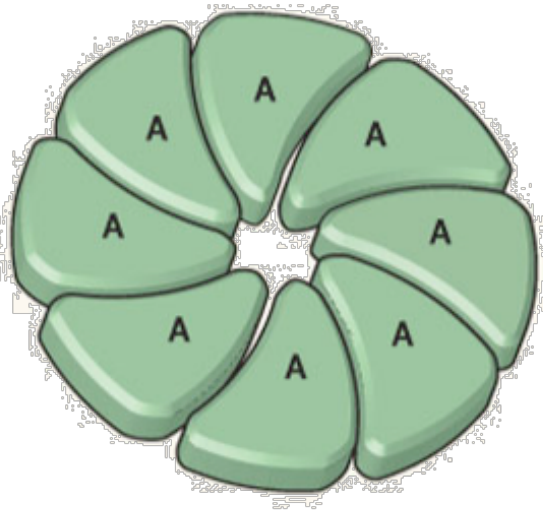


Searches for multiple components

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

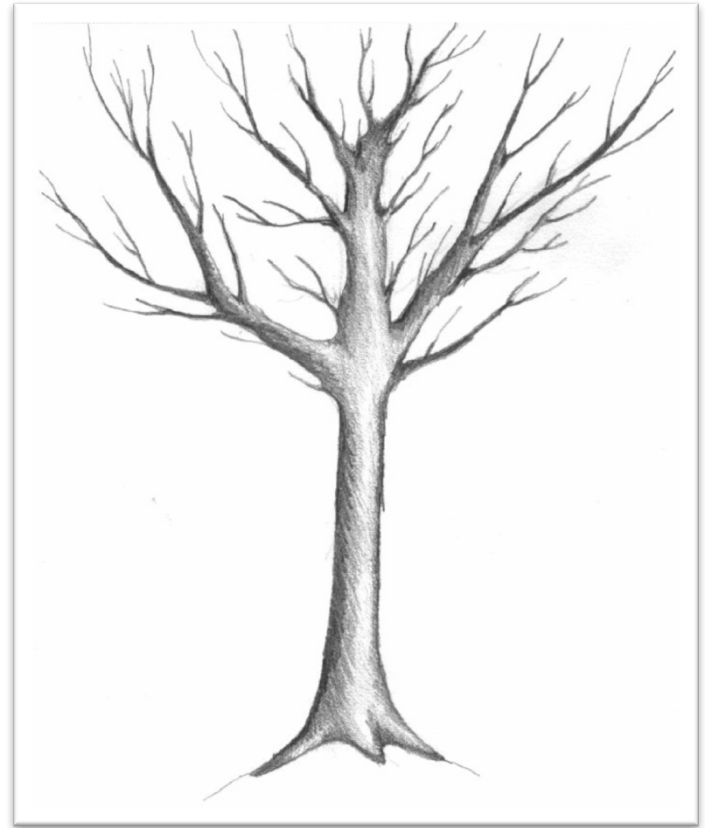


Multi-copy searches



Multi-copy searches

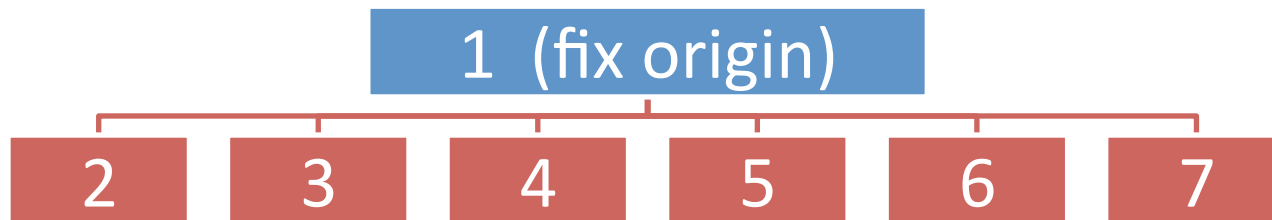
- One RF finds all orientations
- One TF for each orientation finds component
- Tree search generates a heavily branched search
- All solutions equivalent after sequential RF/TF searches



Naive search does more
work than necessary

Fast Search Algorithm

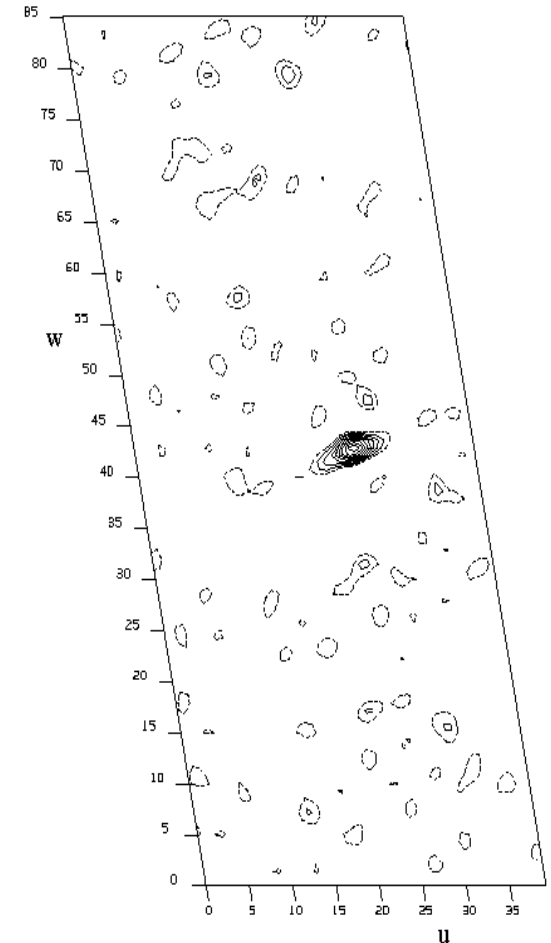
- Phaser has a search algorithm that amalgamates more than one solution per RF/TF pair
- Fixes origin with first, then reuses RF peaks to find other placements



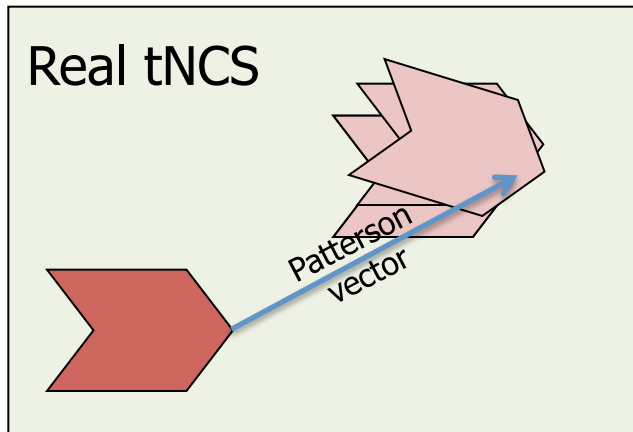
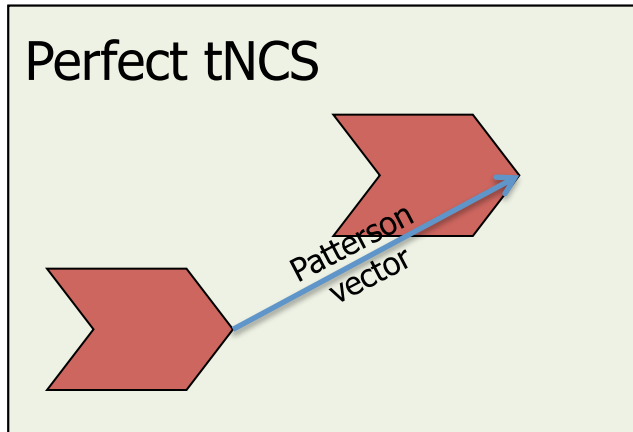
Translational NCS

- 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.



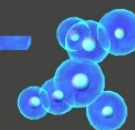
Translational NCS



- 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

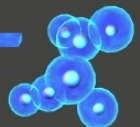
Translational NCS

- 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

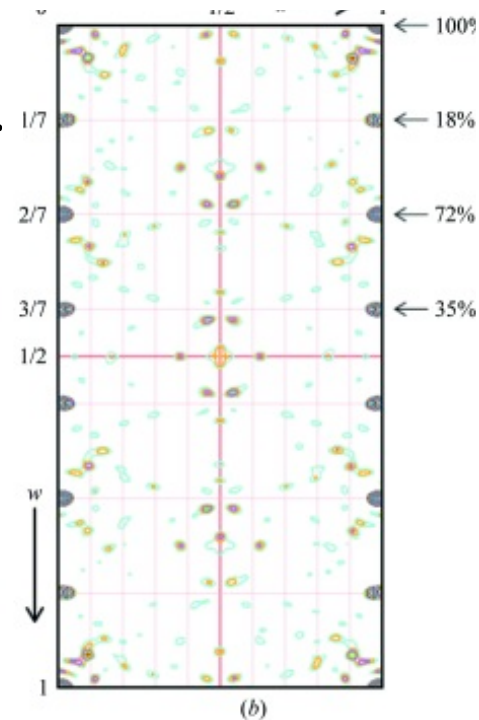
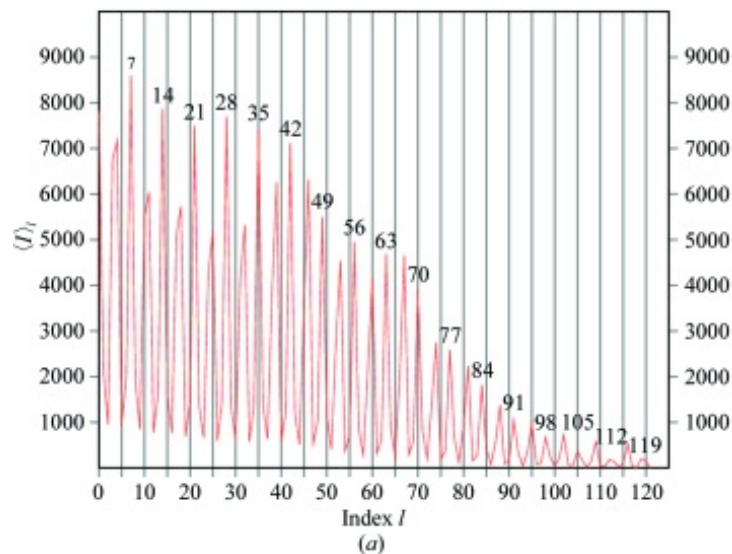
- Translational NCS masks twinning
 - Has opposite effect on intensity statistics
- Correcting the data for tNCS unmask twinning
- Phaser generates cumulative intensity plots for centric and acentric reflections after correction for tNCS and anisotropy
- Phaser gives a P-value for there being twinning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



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

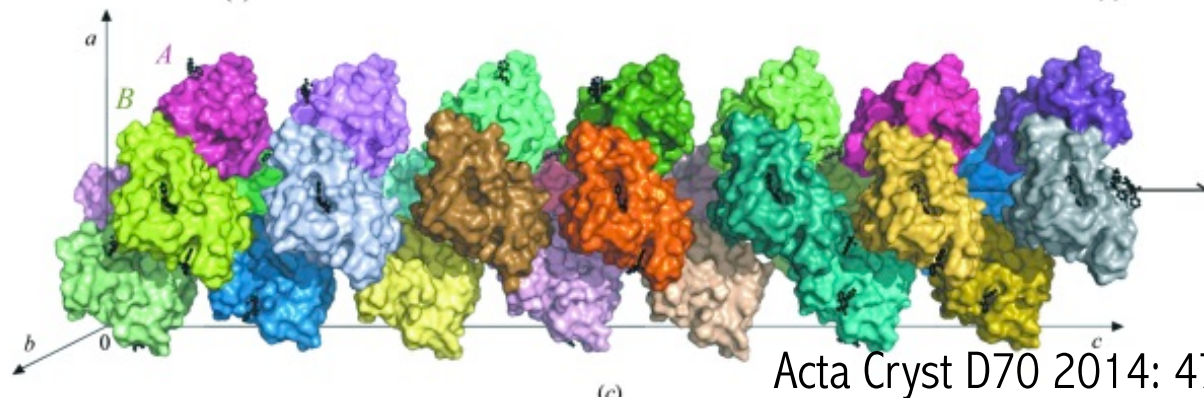
Sliwiak J, Jaskolski M, Dauter Z, McCoy AJ, Read RJ.

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.

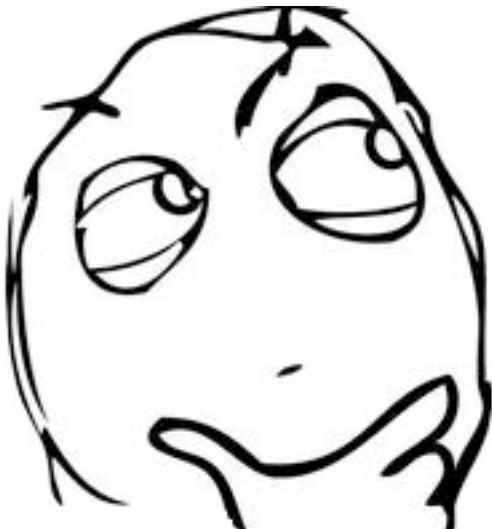


Will MR work
for me?

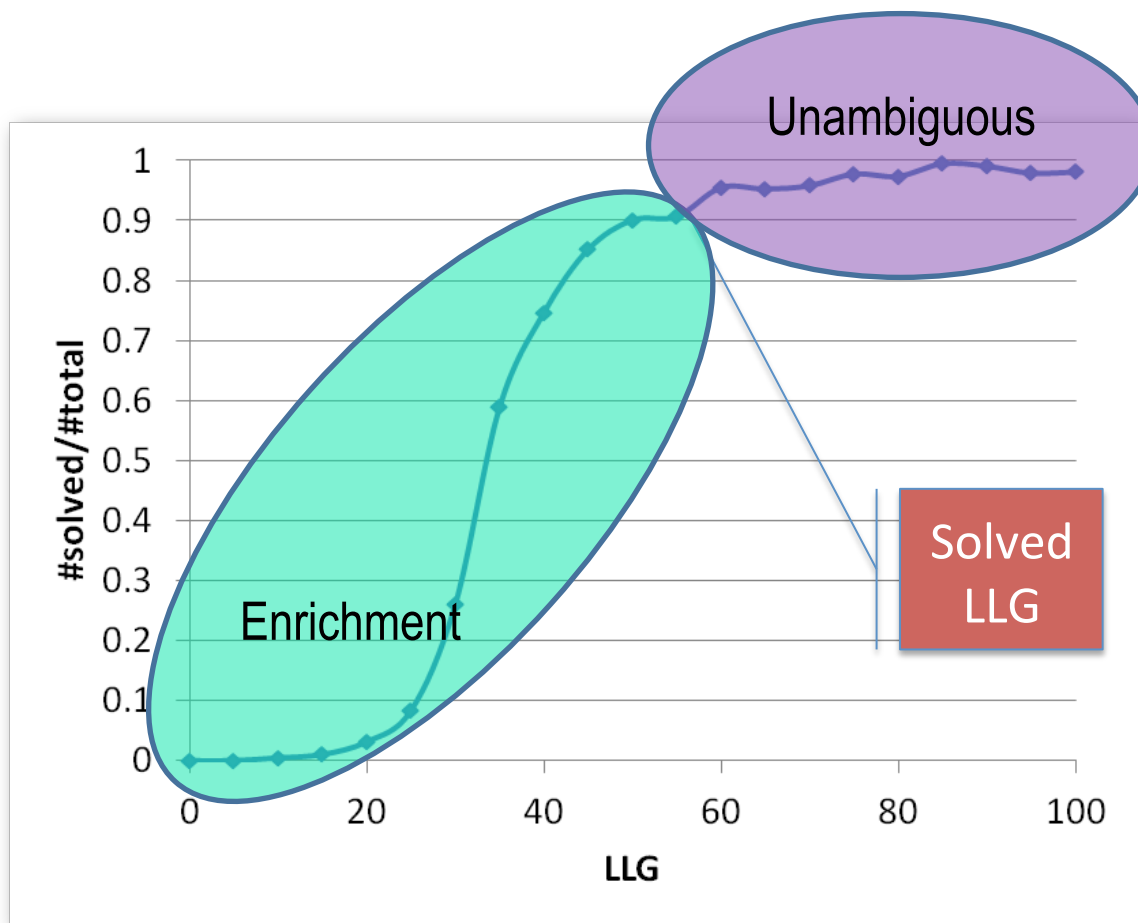
- Is my model good enough?
- Is my data good enough?
- Do I need to place multiple models simultaneously to get a signal?
- Will fragment-based MR work?
- Will α -helices work?
- How big does my helix/fragment have to be?
- Will single-atom MR work?



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?

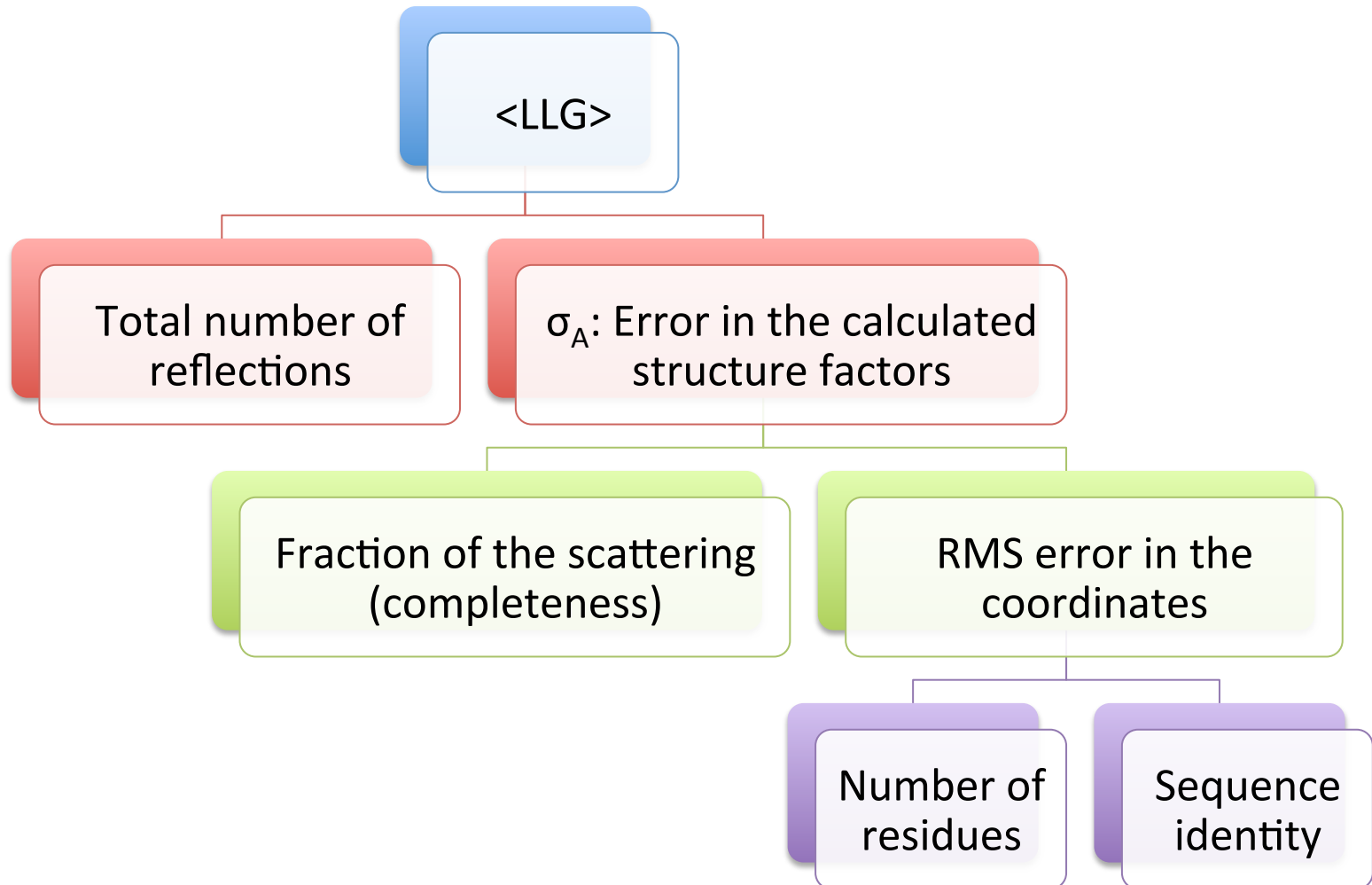
TF Z-score	LLG score	Solved?
< 5	< 25	no
5 - 6	25 - 36	unlikely
6 - 7	36 - 49	possibly
7 - 8	49 - 64	probably
> 8	> 64	definitely

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



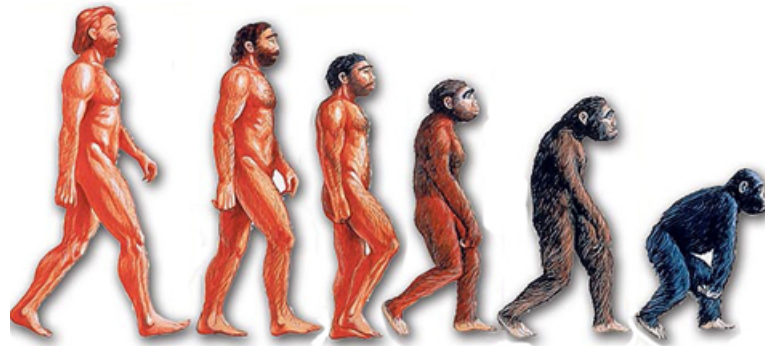
MR with Fragment/Atom

- 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 $\langle LLG \rangle$ being favourable
 - Lots of reflections
 - Low RMS

Success does NOT depend on

- The resolution (strictly)
- The type of fragment

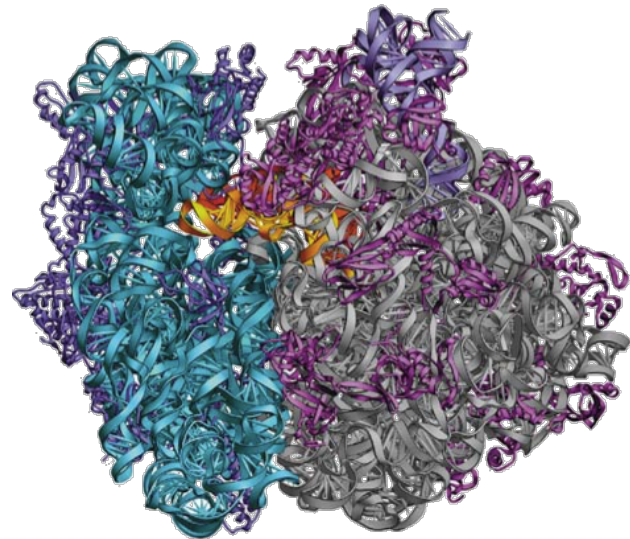
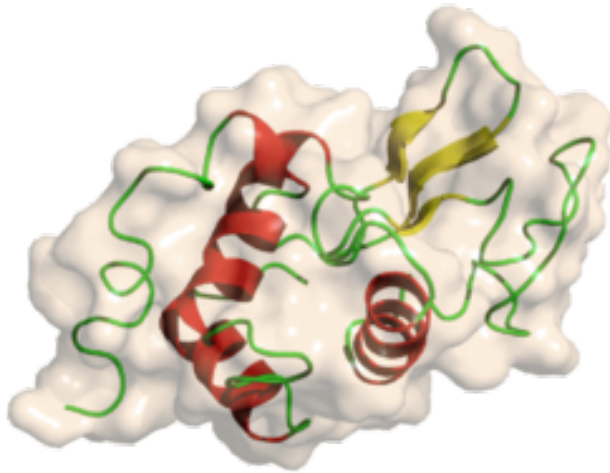
Devolution



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

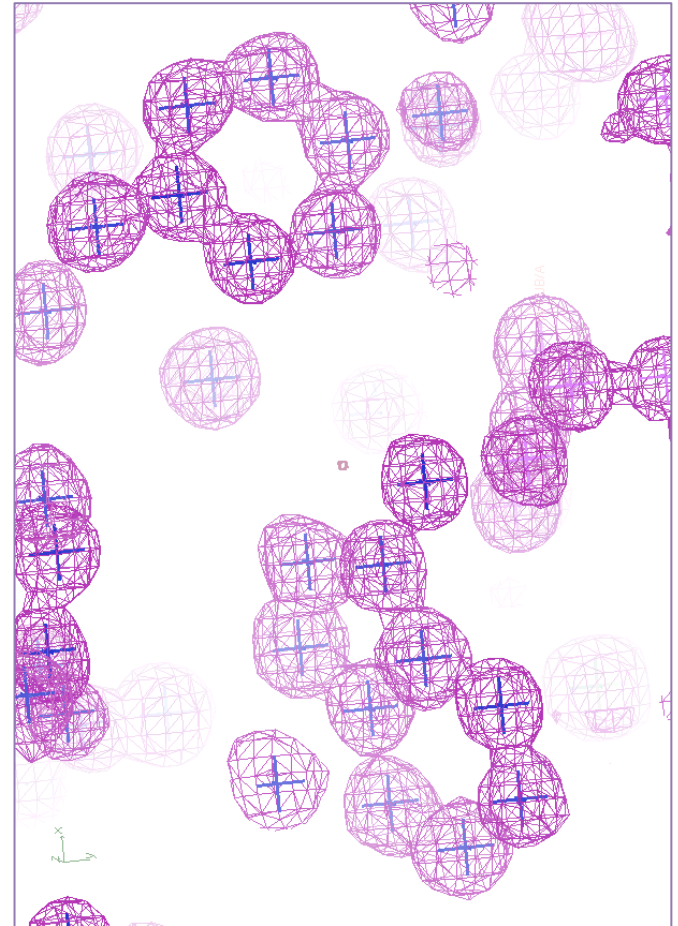
<LLG> and Resolution

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



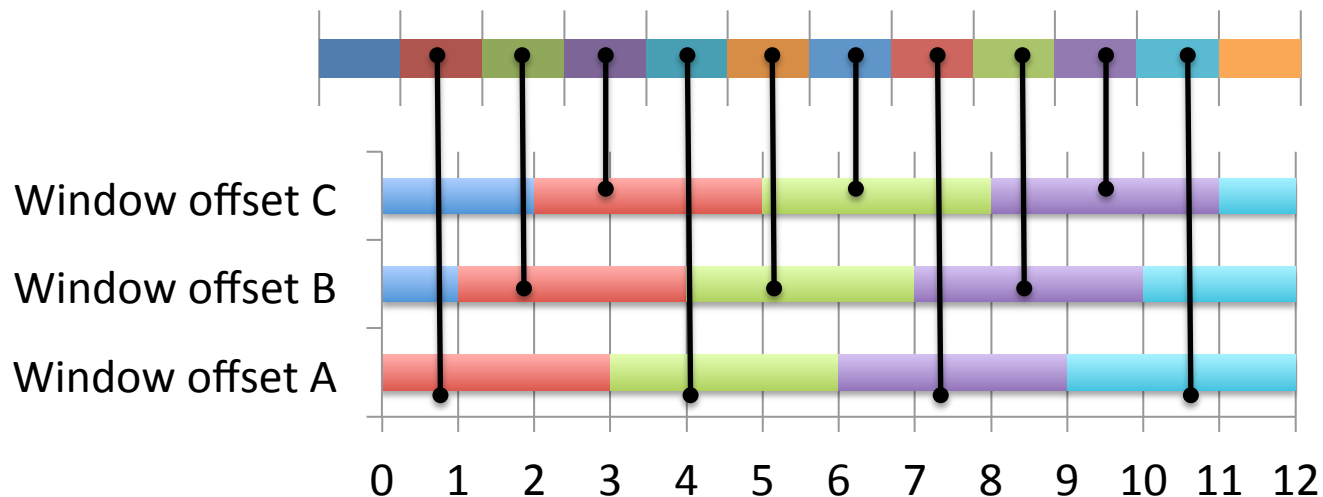
Single Atom MR

- 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 $2 \text{ \AA}^2 < \text{Wilson B}$



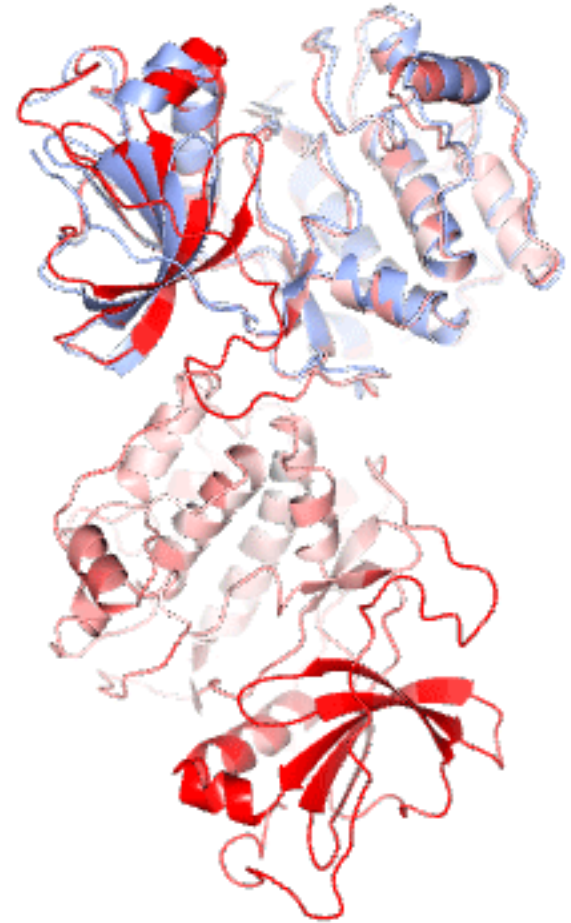
<LLG> Pruning

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



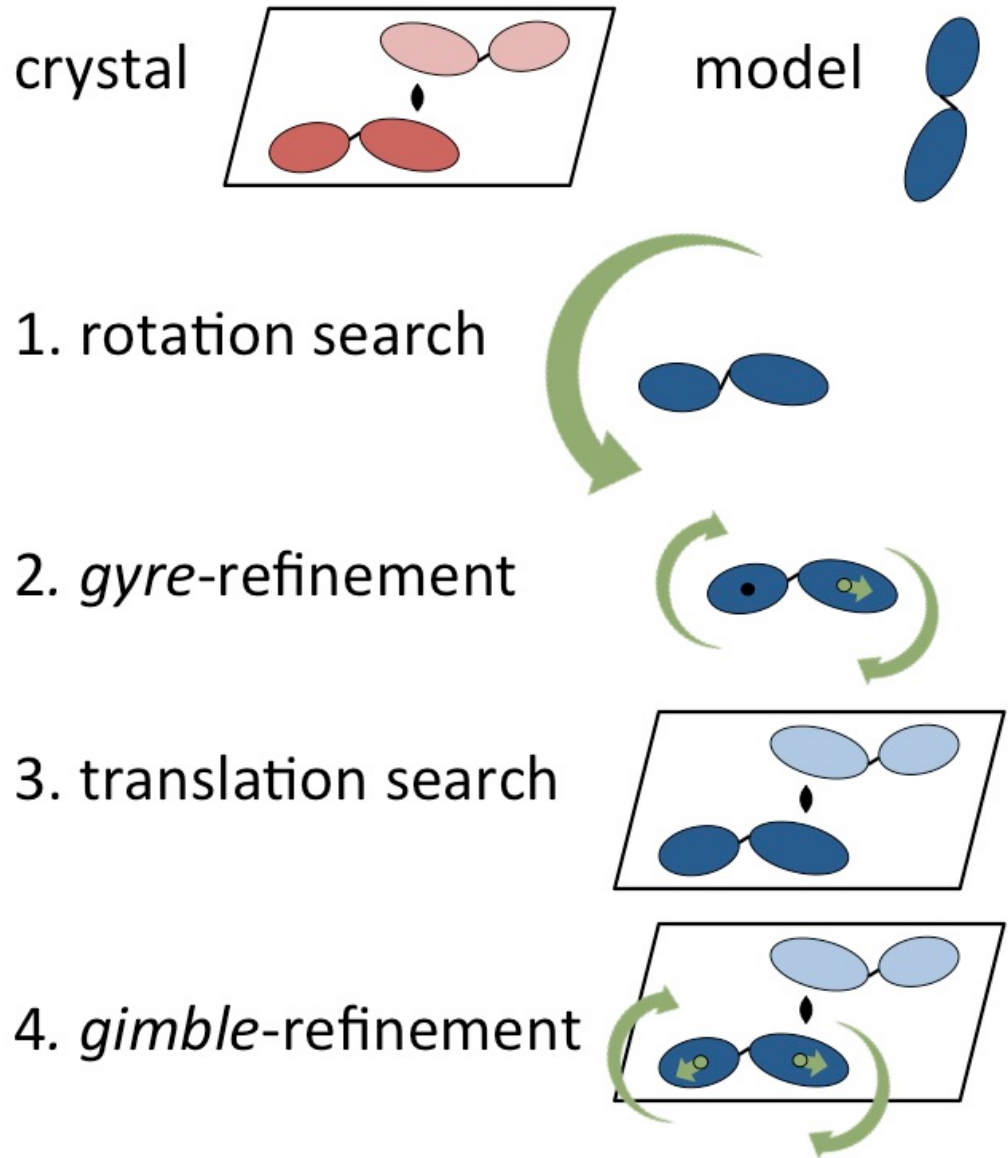
<LLG> Pruning

- 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



Gyre and Gimble

ML replacement for PC- refinement



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)

Find best
model

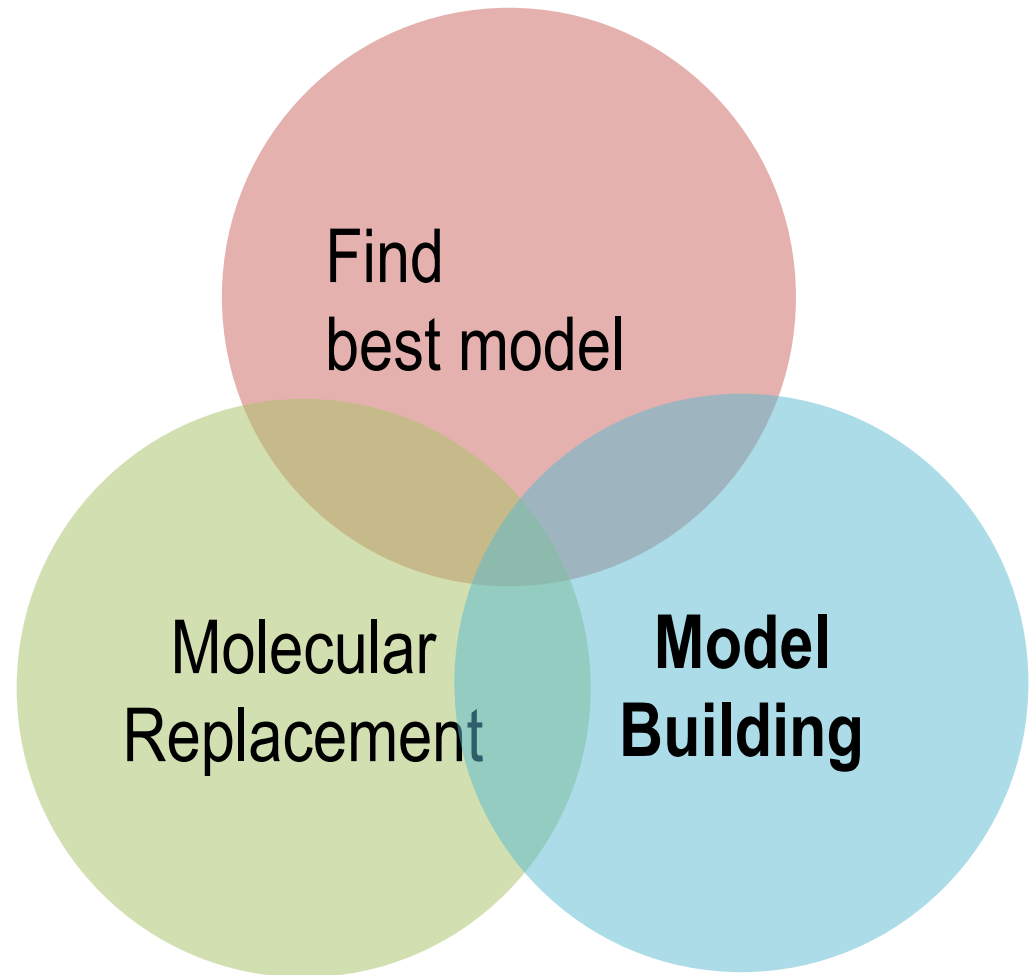
```
graph TD; A[Find best model] --> B[Molecular replacement]; B --> C[Model building];
```

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



phaser.mrage

- Fetches models and processes using sculptor
- For each partial structure model MR is farmed out to a cluster in a highly parallel manner
 - Calculations are performed in the order of sequence identity or LLG score at each stage
- Exploration continues until a solution is found
- All alternative models are superposed onto the solution and refined. This allows the quick evaluation of model quality for a potentially large number of alternative models.

Phaser.MRage: automated molecular replacement
Bunkoczi G, Echols N, McCoy AJ, Oeffner RD, Adams PD, Read RJ
Acta Cryst. (2013). D69, 2276-2286

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



**UNIVERSITY OF
CAMBRIDGE**

**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***

