



phaser and phasing

Airlie McCoy



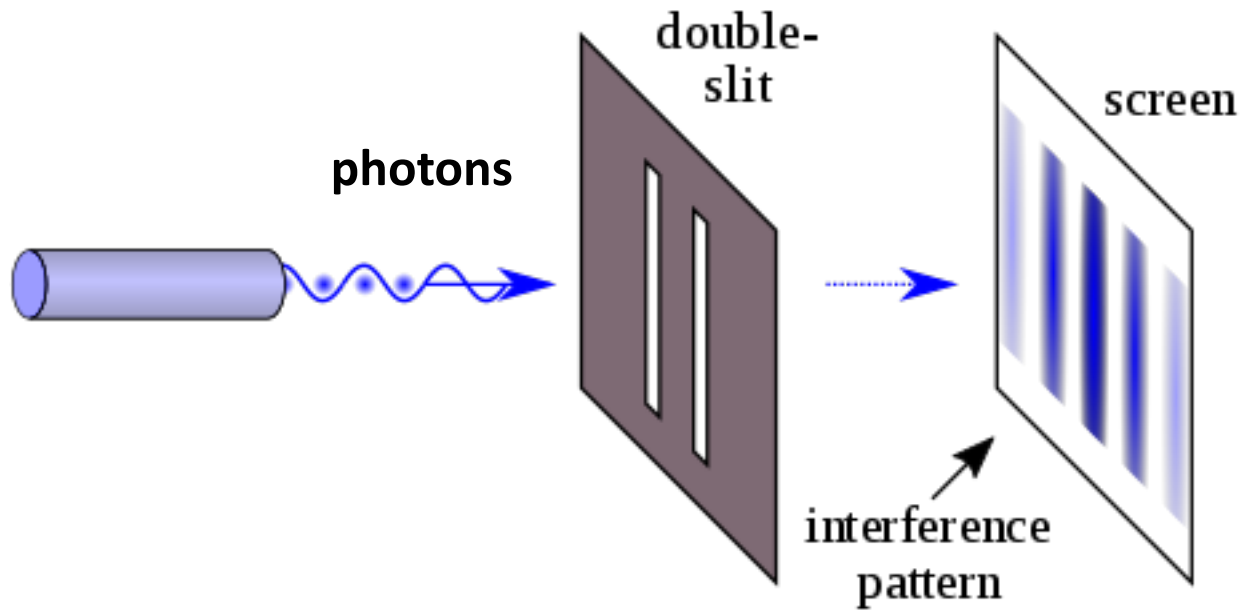
UNIVERSITY OF
CAMBRIDGE

What is phasing?

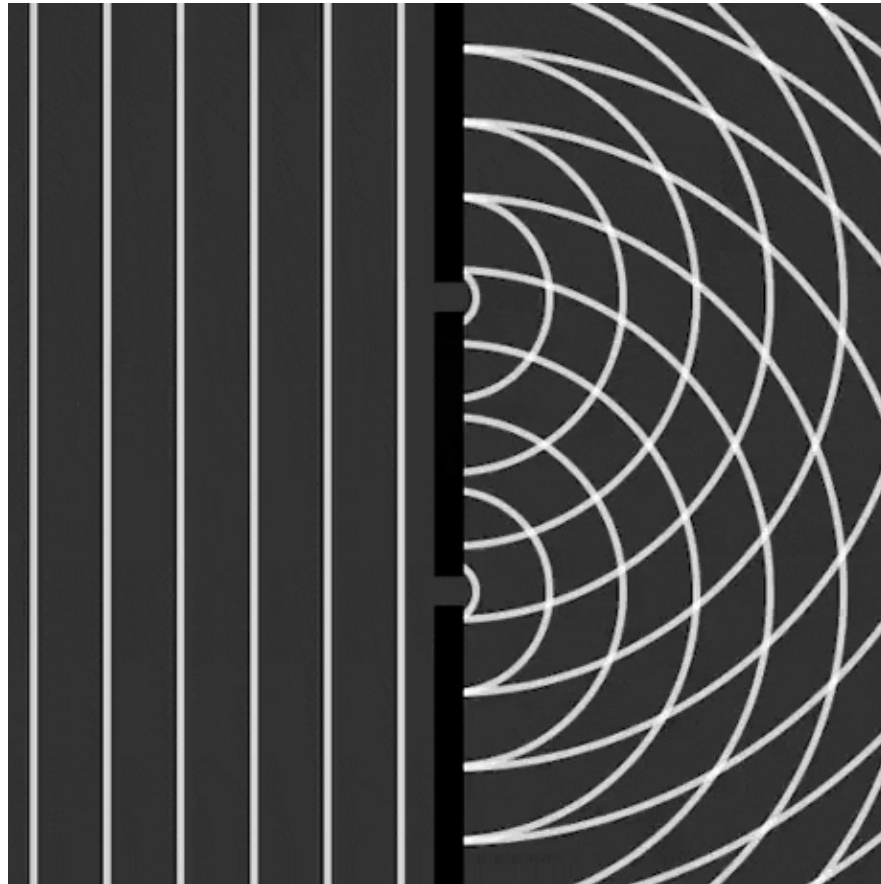
- There isn't a 1-to-1 correspondence between spots and atoms in the protein!
- Each spot contains some information about the position of every atom
- How do we get at this information?



Young's Double Slit Experiment



Young's Double Slit Experiment



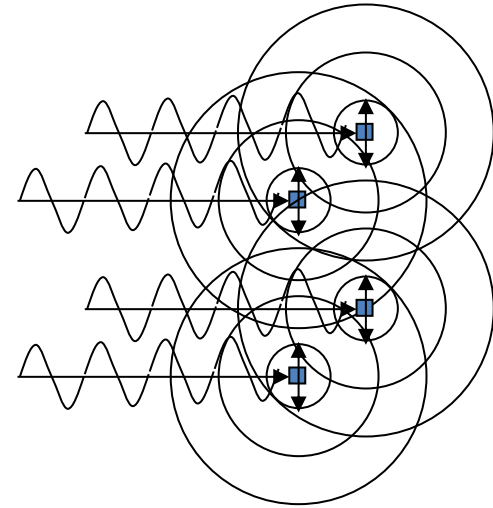
Interference to Diffraction

"No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage... . The best we can do, roughly speaking, is to say that when there are only a few sources, say two, interfering, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used."

Richard Feynman

diffraction

- An electron that interacts with an electromagnetic wave (X-ray) will oscillate
- Oscillating (accelerating) charges act as new sources
- X-rays are scattered (or re-emitted) in all directions from the electrons they encounter

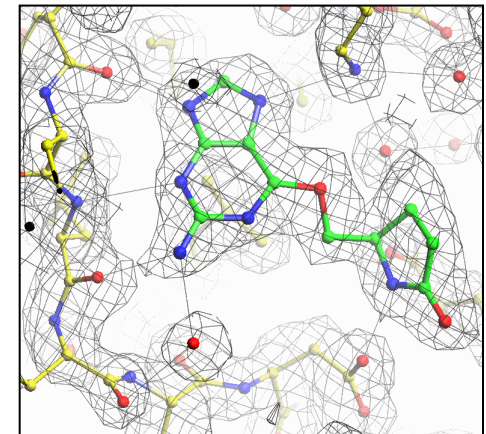
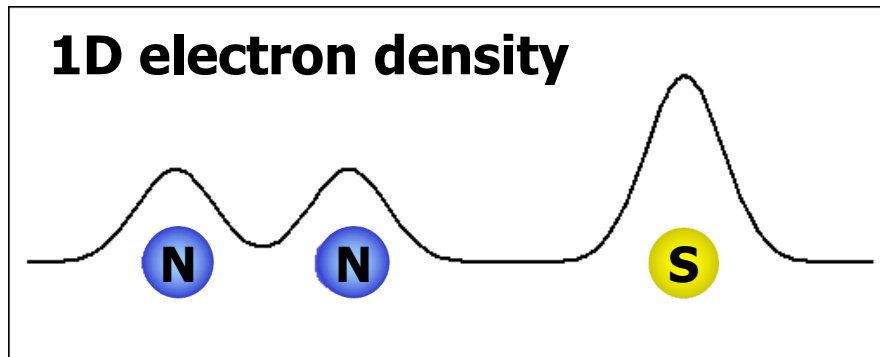


Fourier Transform

It can be shown mathematically that the diffraction from an object is the Fourier Transform on the object

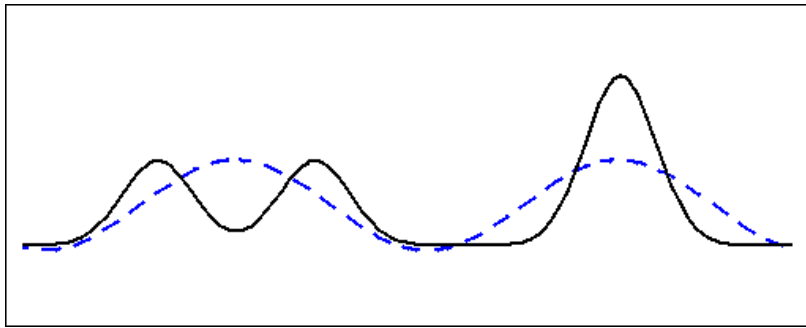
3 atoms in 1 dimension

- The electron density through three atoms is an electron density curve with three peaks

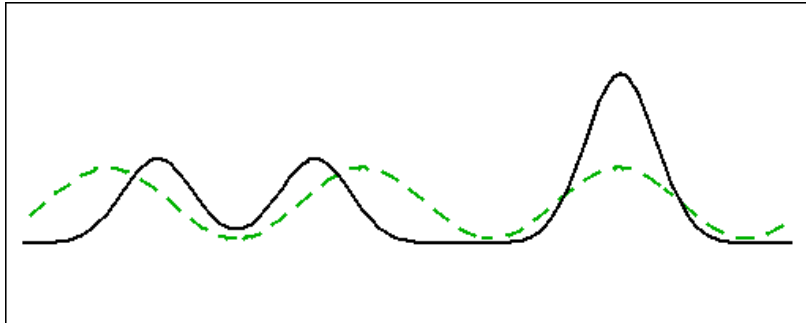


**3D electron
density**

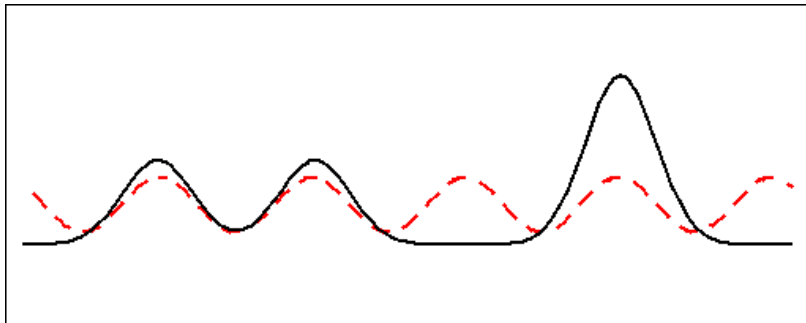
- The electron density curve can be described as the sum of sine and cosine functions



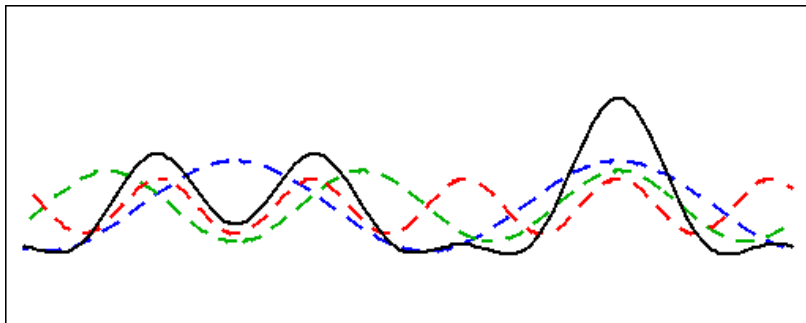
Curve has some
frequency=2 character



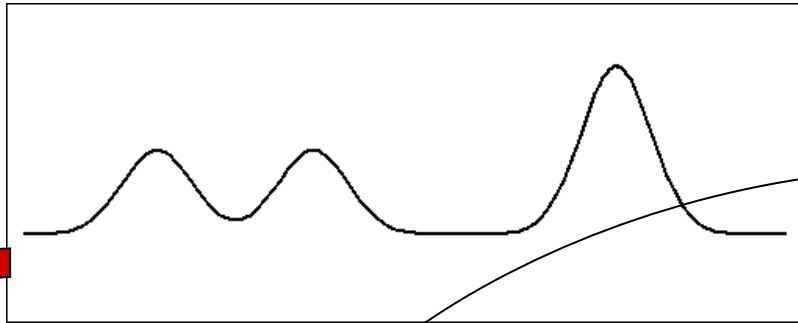
Curve also has some
frequency=3 character



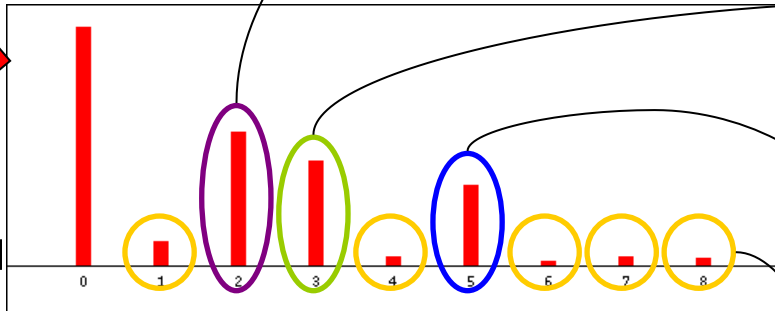
And some
frequency=5 character



Sum of the 3 functions above
gives a curve **approximating**
the 3 atom density

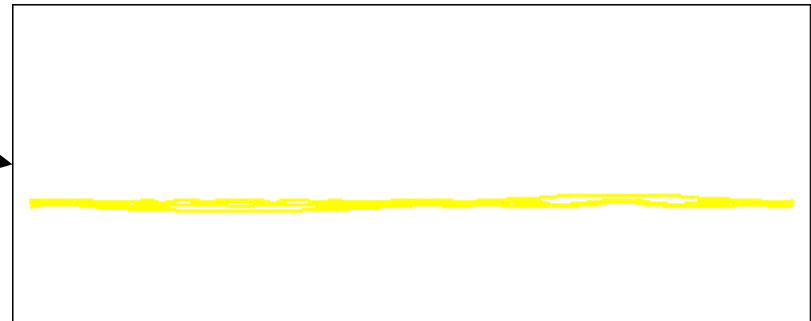
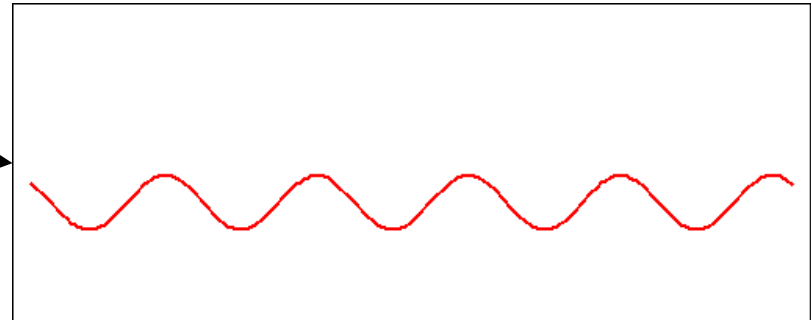
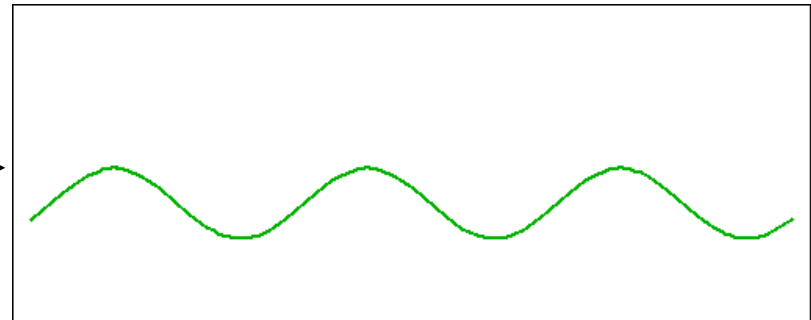
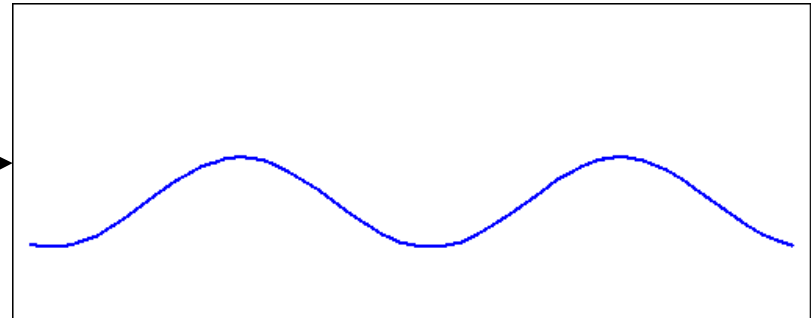
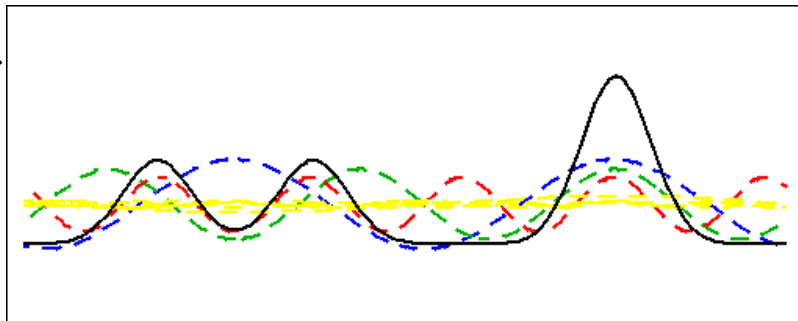


Fourier Transform



The FT actually gives complex numbers and so has phase

Inverse Fourier Transform



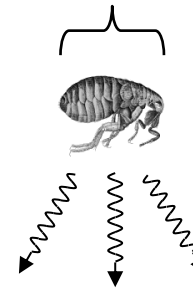
diffraction and imaging



Size 2,000,000nm

Object

*Light Scattered
First Fourier Transform*

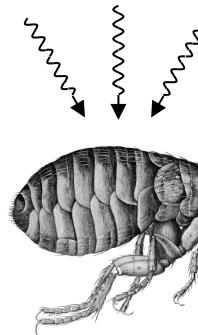


Blue Light
wavelength
500nm is much
shorter than
distances to be
resolved

Lens



*Light Focused
Second Fourier Transform
Image*



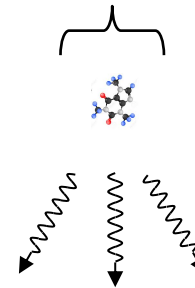
diffraction and imaging

X-ray
wavelength $1\text{\AA}$
is shorter than
(twice) the
distances to be
resolved

Size 1nm

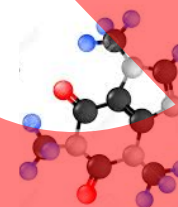
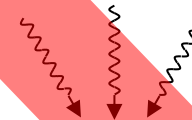
Object

Light Scattered
First Fourier Transform

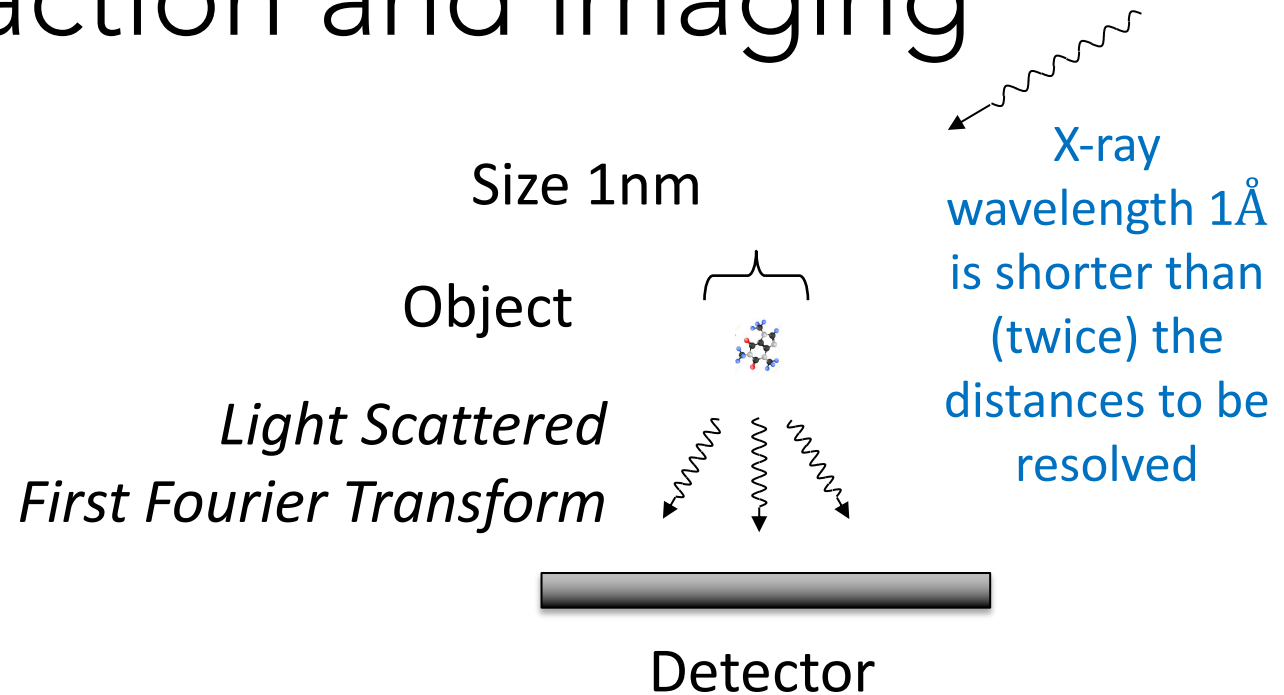


Lens

Light Focused
Second Fourier Transform
Image



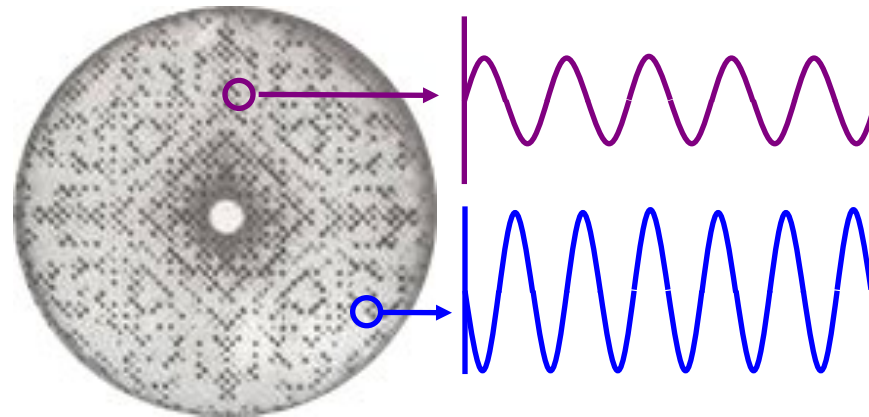
diffraction and imaging



**If you capture the diffraction,
you get the diffraction pattern**

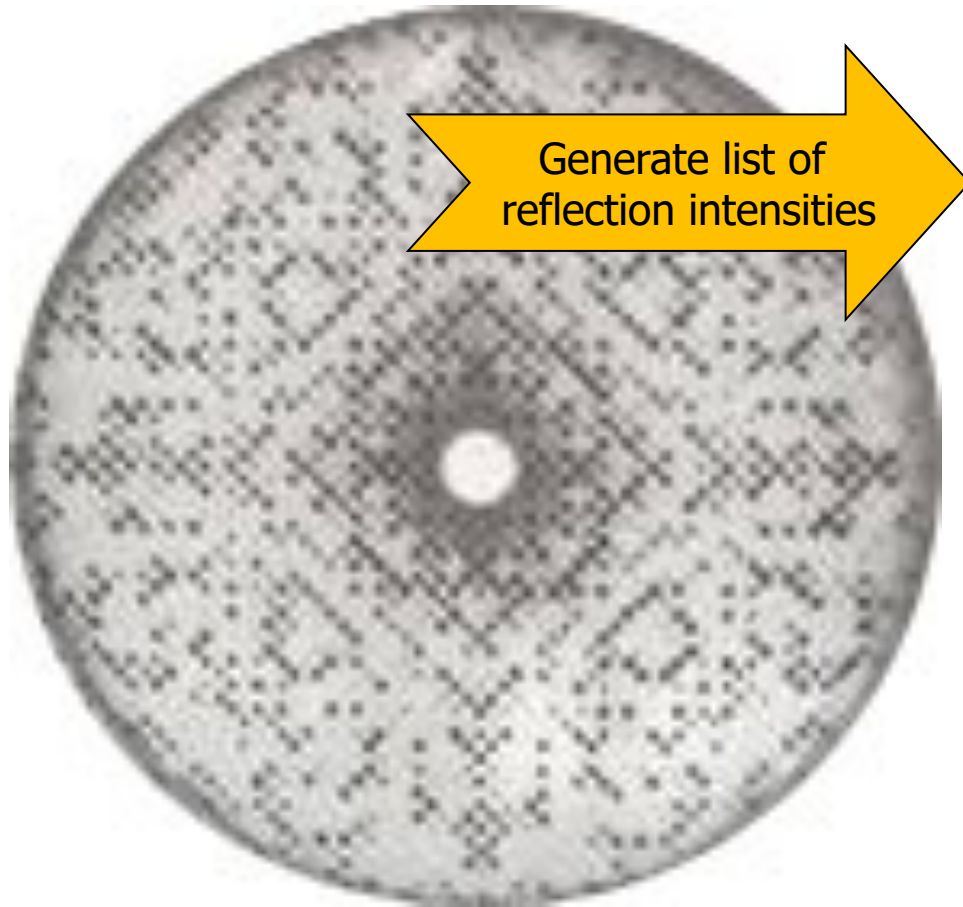
phase problem

- Intensity of spots in the diffraction pattern is nothing to do with the phase



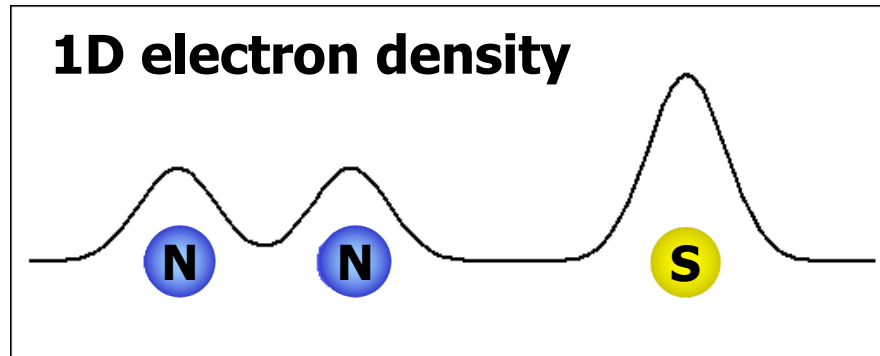
- We do not have all the information required to do the second Fourier Transform

phase problem



H	K	L	F	σF
0	0	1	126	12
0	0	2	21	10
0	0	3	699	280
etc...				
0	1	13	54	3
0	1	14	231	34
etc...				
27	11	14	522	242
27	11	15	447	219
27	11	16	46	35

3 atoms in 1 dimension



phase problem

Summing waves with the amplitudes but different phases gives different curves

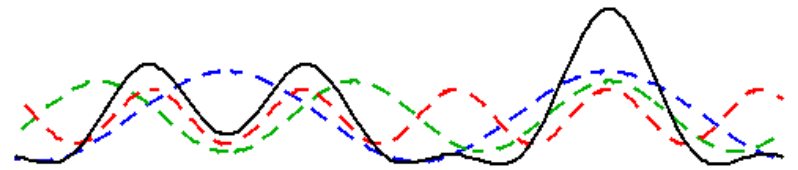
Without the phase the registration of the waves is unknown

Cannot add waves to generate correct electron density

Is this correct?



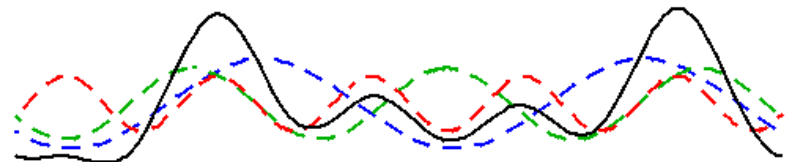
Or is this correct?



What about this one?



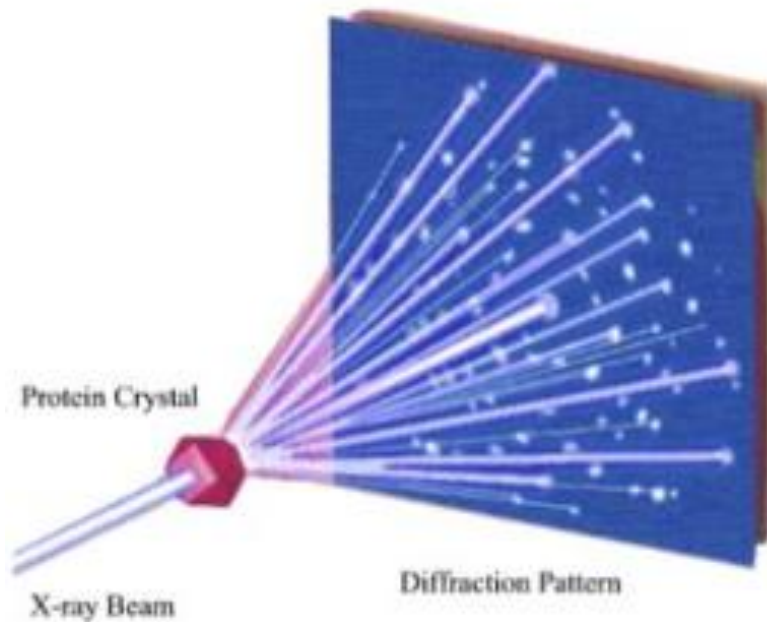
Maybe this one!



diffraction and imaging

First Fourier Transform
Performed by Diffraction

Second Fourier Transform
Performed in Computer



**Solve the
Phase
Problem**



Solutions to the Phase Problem

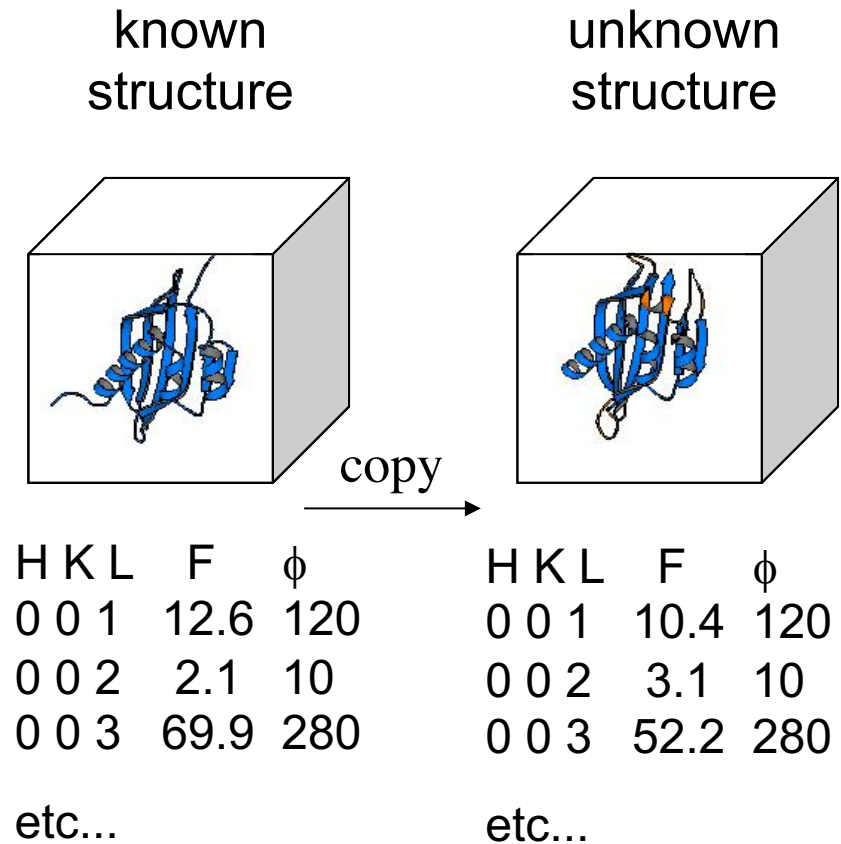
- Experimental Phasing
 - Do an interference experiment to find the phases
 - Incorporate heavy atoms or anomalously scattering atoms into the structure
 - Need to know heavy atom positions
- Molecular Replacement
 - Borrow the phases from known structure(s)
 - Requires a model/models for the target
 - Need to know orientation and position of the model(s)

Molecular Replacement

A solution to the phase problem

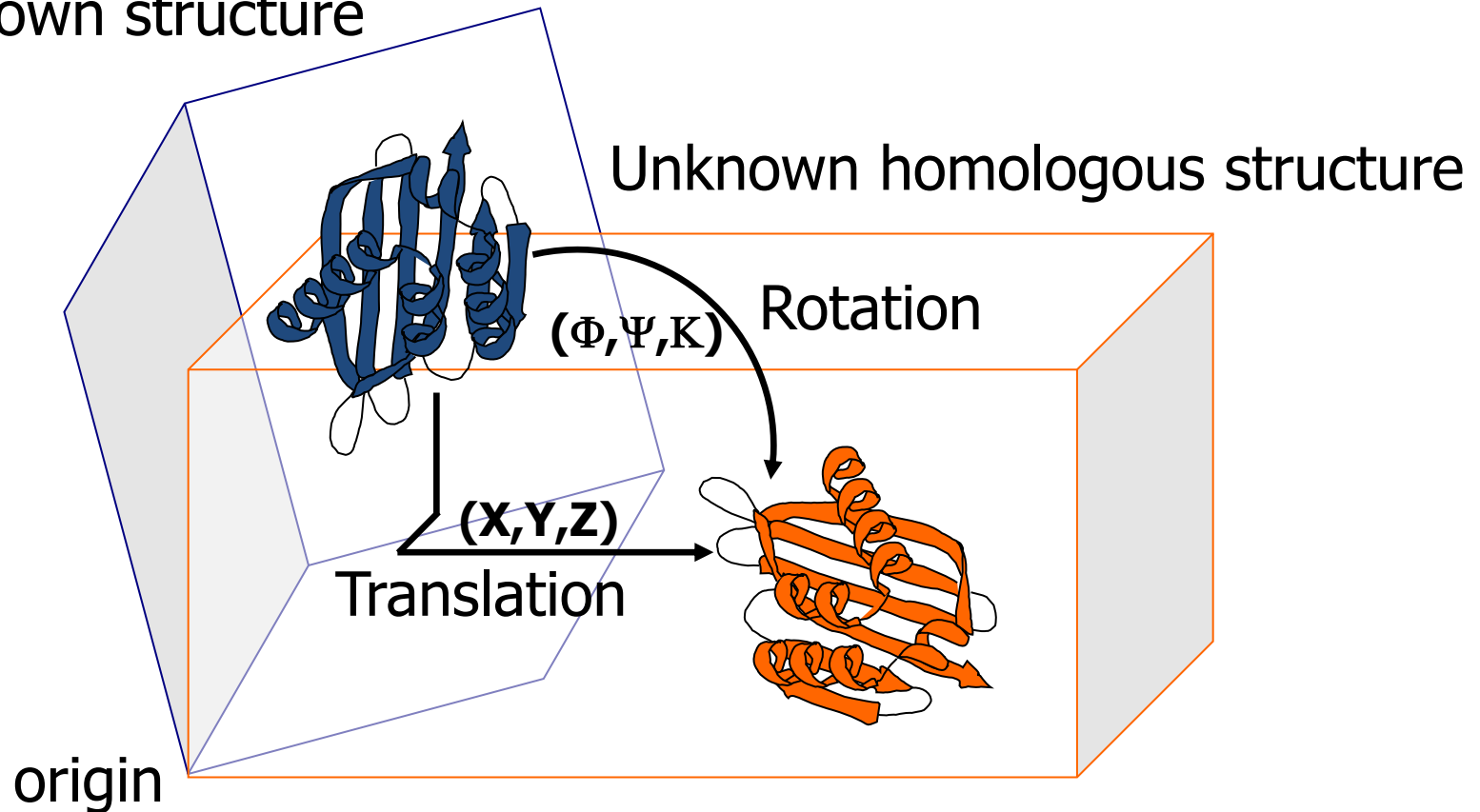
molecular replacement

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



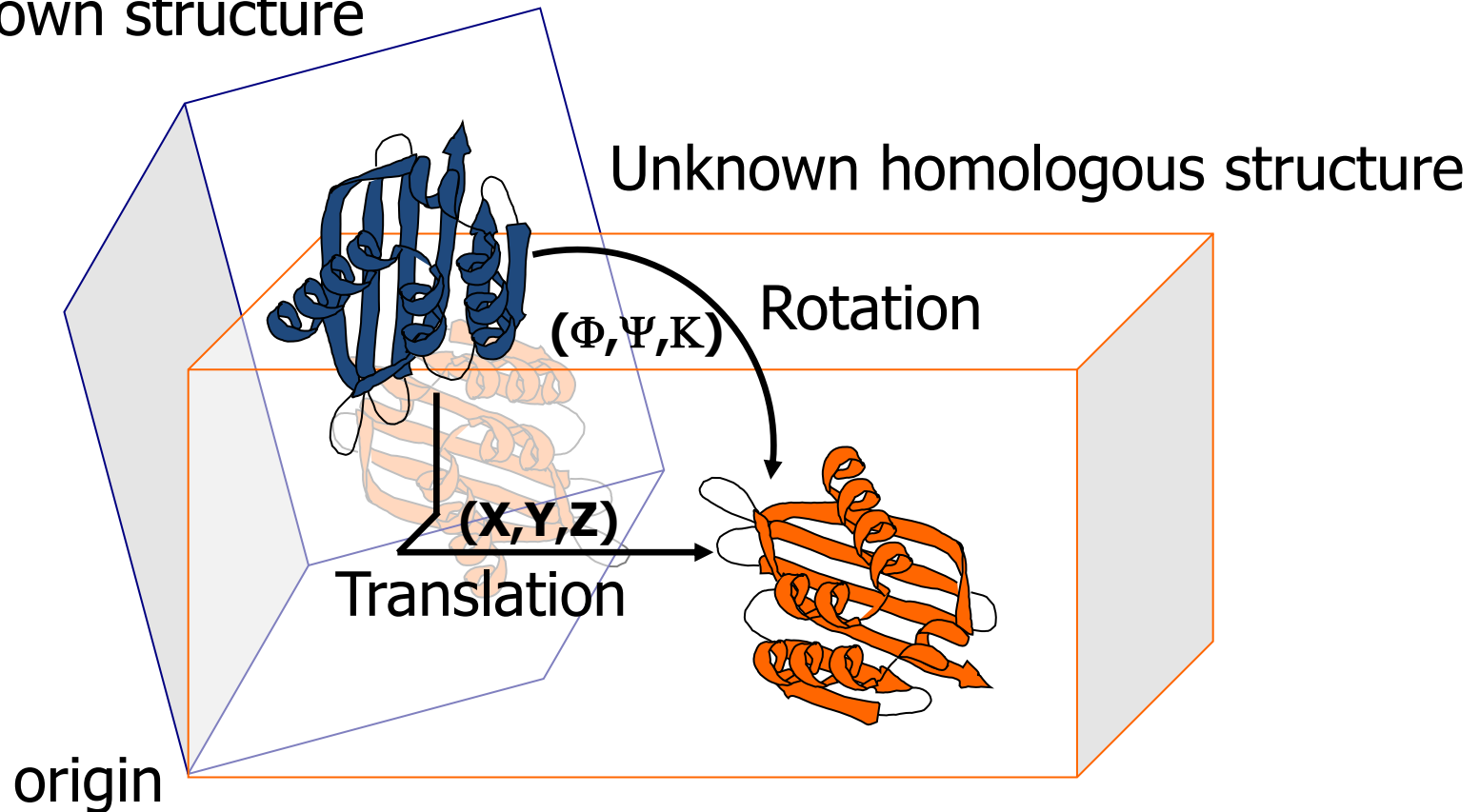
molecular replacement

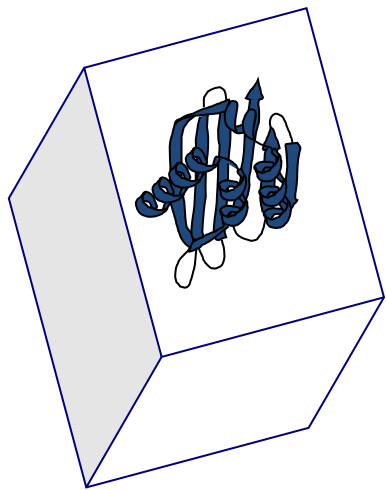
Known structure



molecular replacement

Known structure

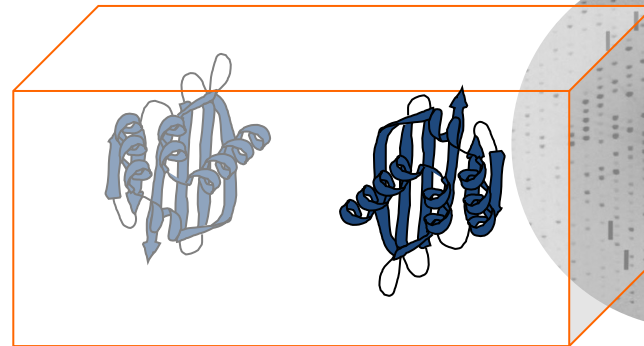




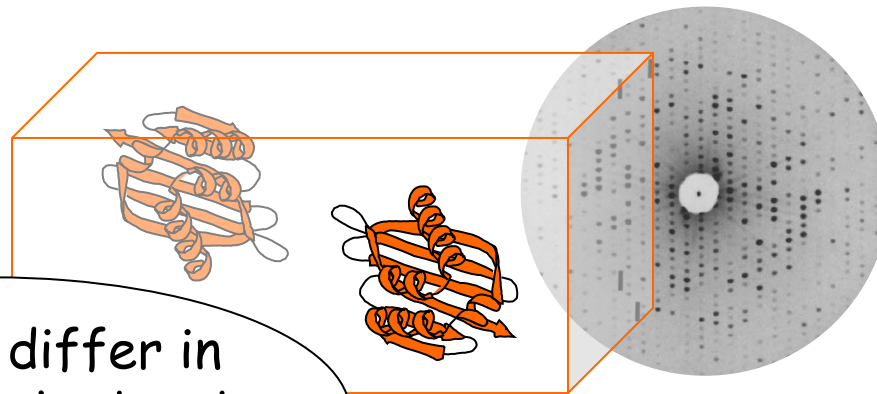
SEARCH

(Φ, Ψ, χ)

(x, y, z)



SCORE



Programs differ in
search method and
scoring function

molecular replacement

pose: oriented and positioned model

- Score each pose to find when the model best fits the target structure
- Search for solutions: strategies for exploring poses

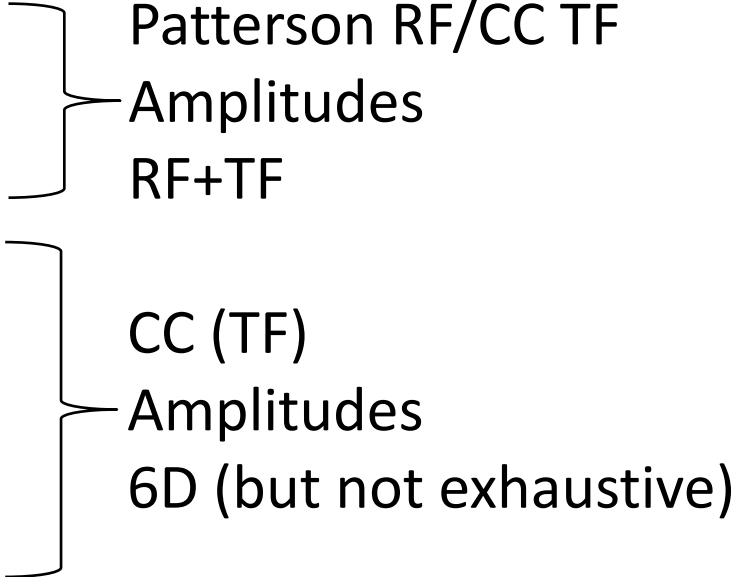
searching

- 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
- 
- Patterson RF/CC TF
Amplitudes
RF+TF
- CC (TF)
Amplitudes
6D (but not exhaustive)
- Phaser Maximum likelihood
 Intensities
 RF+TF+RF+TF

maximum likelihood scoring

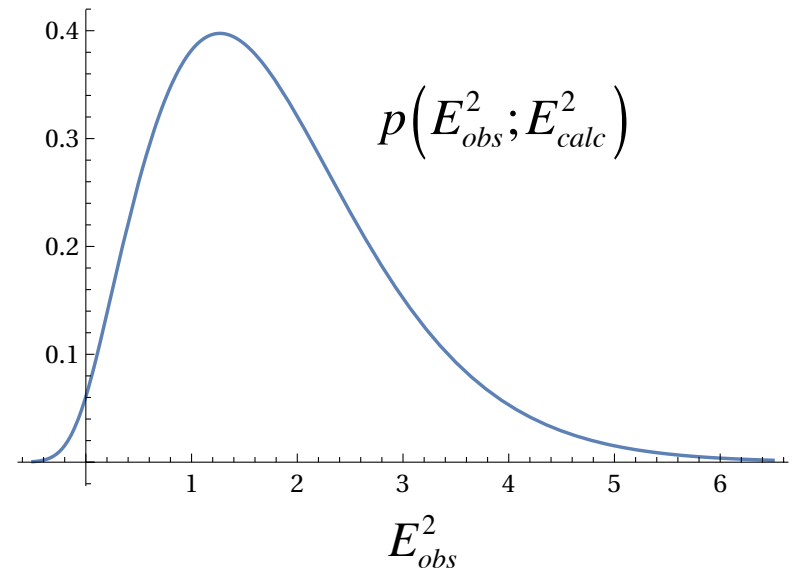
- Use probability
- Probabilities account for errors
 - Patterson methods cannot do this

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

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

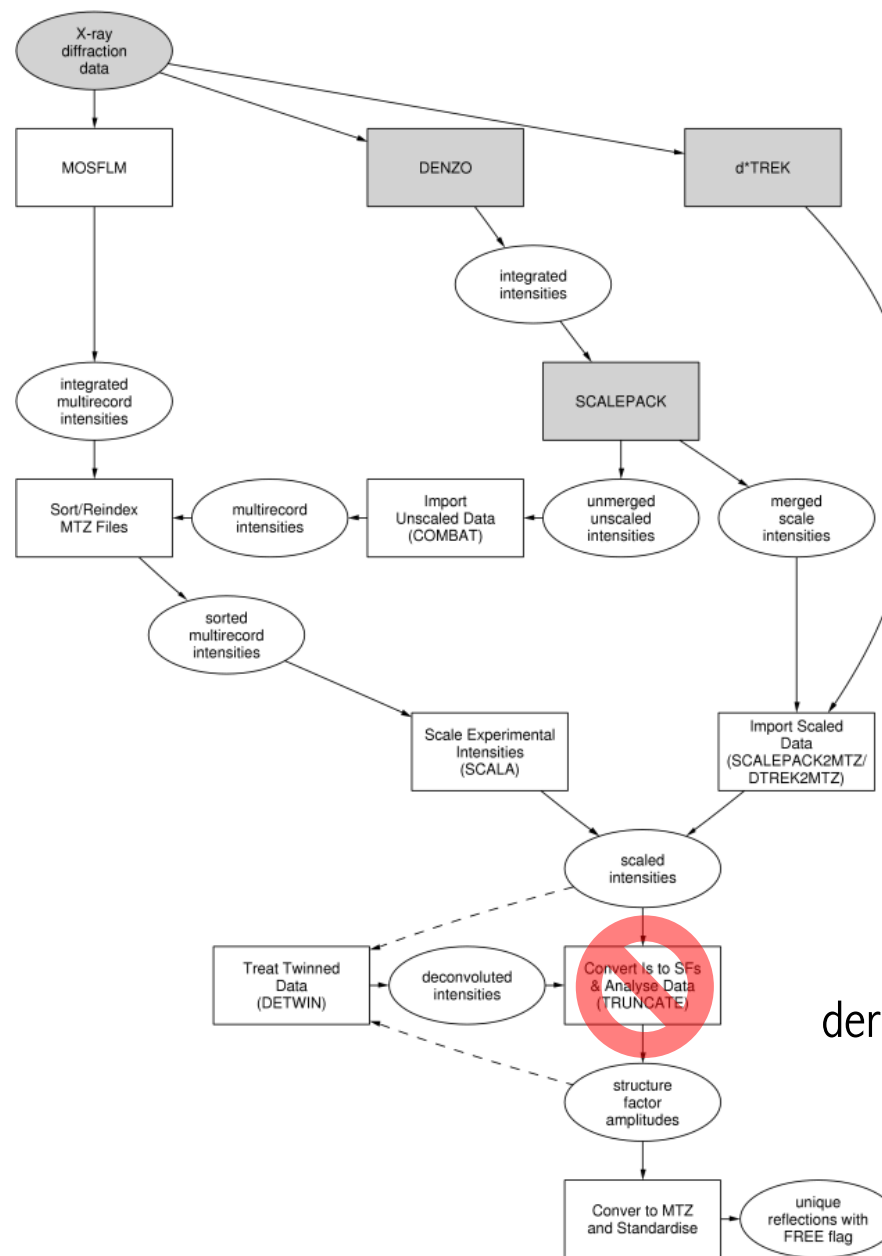
maximum likelihood scoring

- Peak intensity minus background can yield negative net intensity
 - Weak data cannot be excluded with resolution limit for pathologies
- Use intensities in Phaser
 - not French and Wilson truncated structure factors
- *Please*, deposit your intensities at the PDB



Exact intensity likelihood function for
 $E_C = 1.5$, $\sigma_A = 0.8$ and $\sigma_{obs} = 0.2$

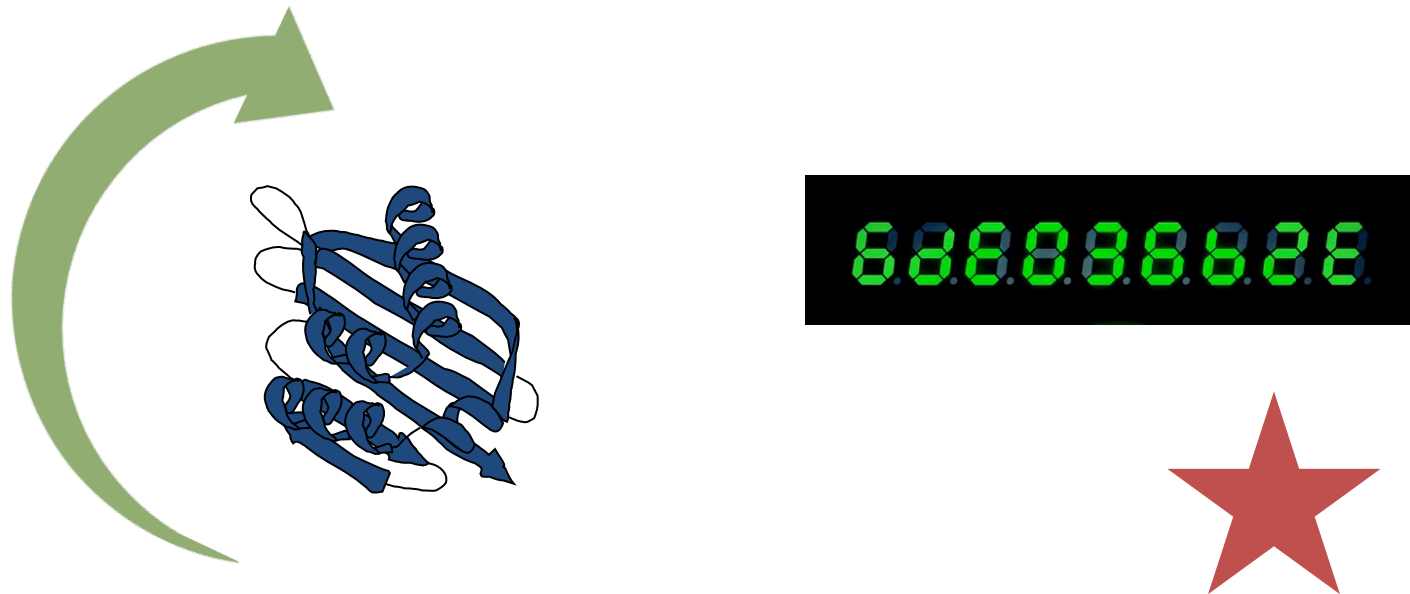
roadmap



Use
 $D_{\text{obs}}/F_{\text{eff}}$
derived from $I/\sigma I$

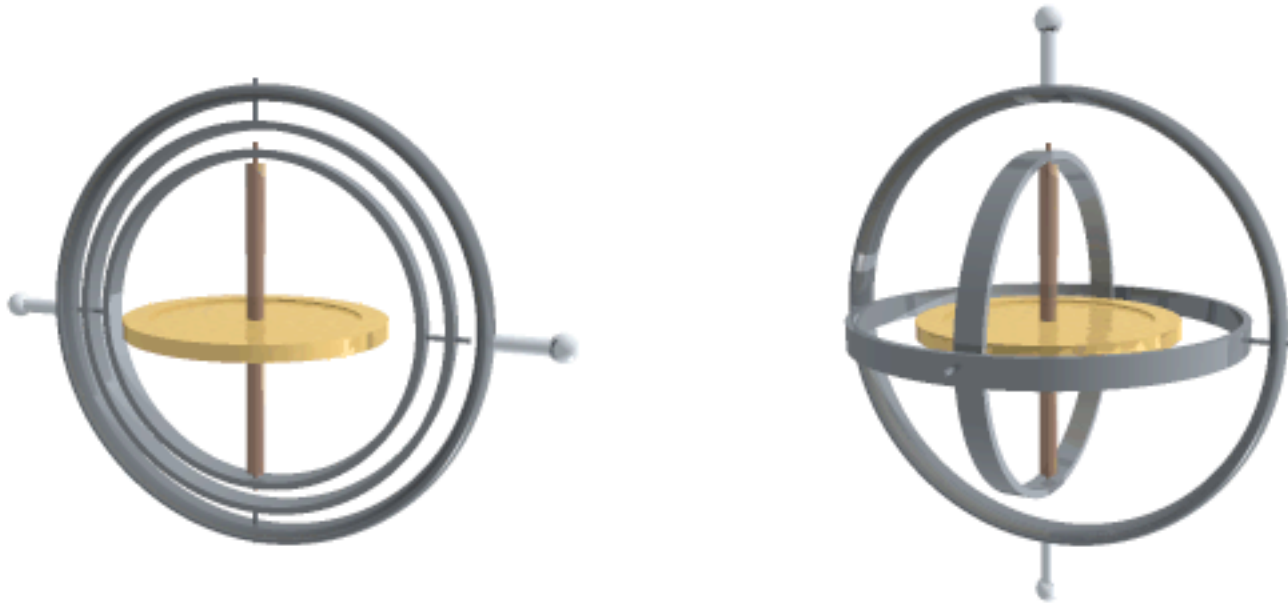
Brute rotation function search

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



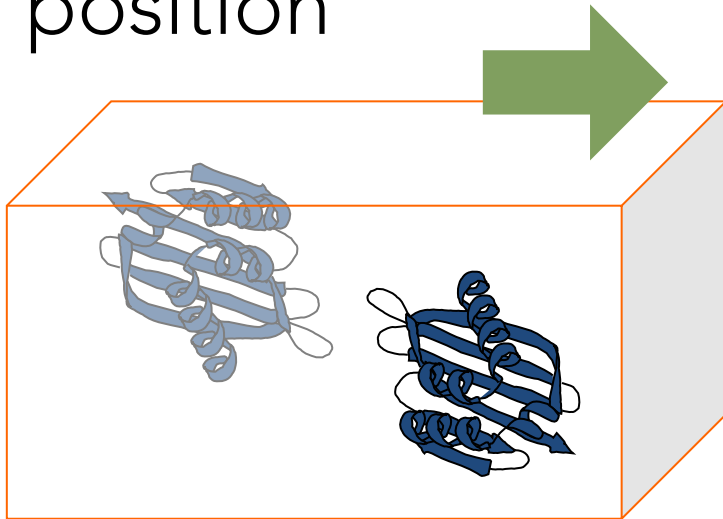
$$LLGI = \sum_{\mathbf{h}} \log \left(\frac{2E_e}{1 - D_{obs}^2 \sigma_A^2} \exp \left(- \frac{E_e^2 + D_{obs}^2 \sigma_A^2 E_C^2}{1 - D_{obs}^2 \sigma_A^2} \right) I_0 \left(\frac{2E_e D_{obs} \sigma_A E_C}{1 - D_{obs}^2 \sigma_A^2} \right) \right)$$

Euler Angles



Brute translation function search

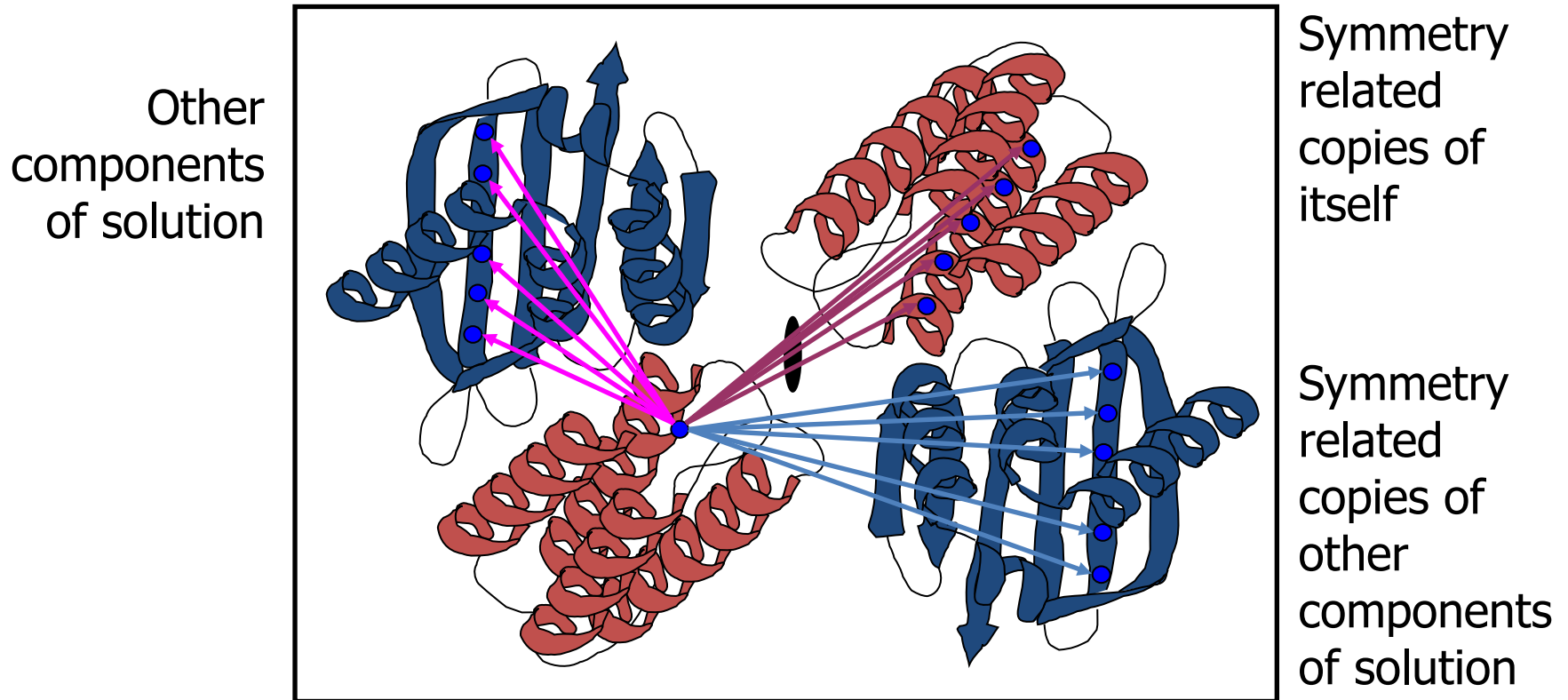
- Place model at points in unit cell and calculate probability that it is in each position



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

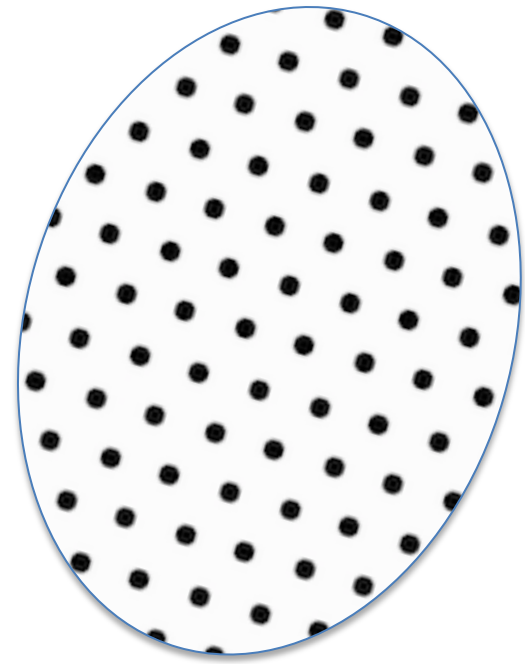
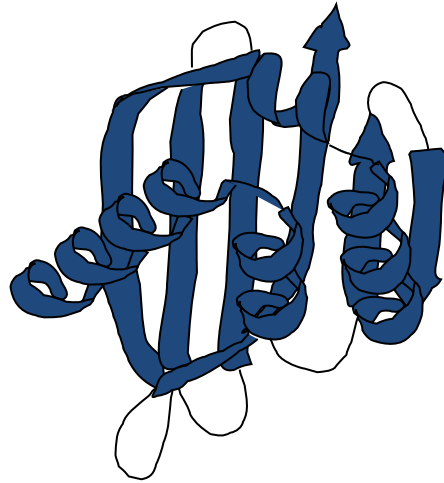
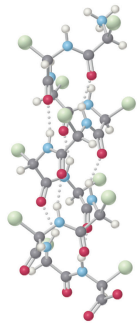
Packing Function

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



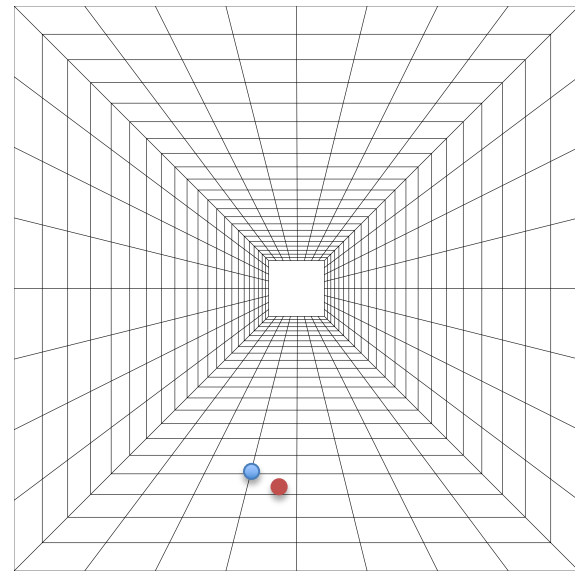
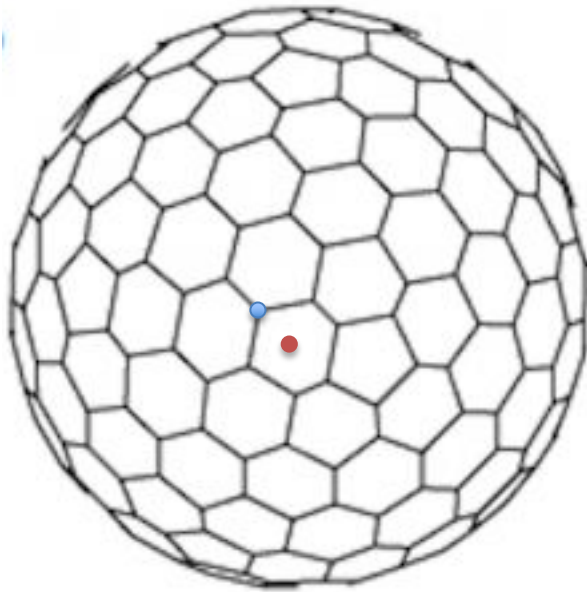
Packing Function

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

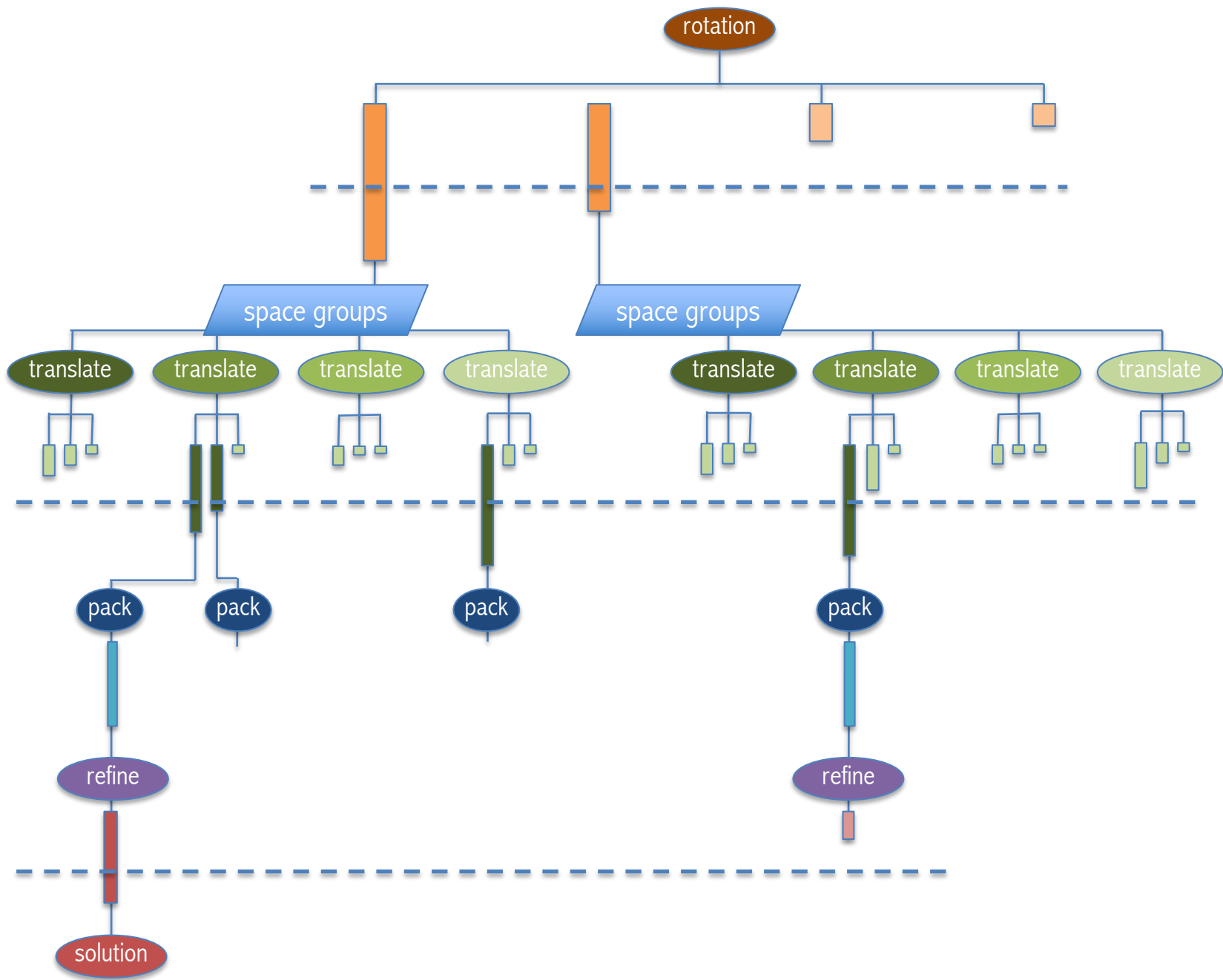


Refinement

- Rotation and Translation searches are scored on a grid

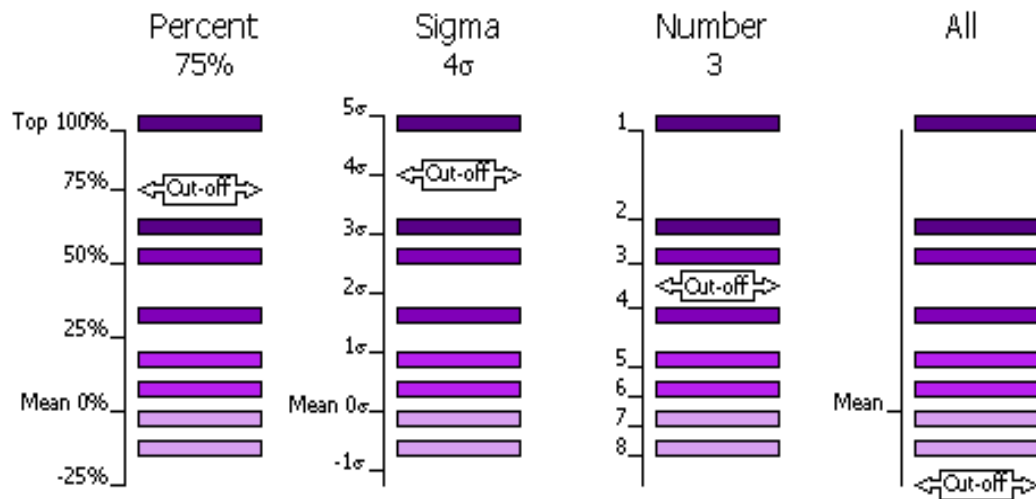


$$LLGI = \sum_{\mathbf{h}} \log \left(\frac{2E_e}{1 - D_{obs}^2 \sigma_A^2} \exp \left(- \frac{E_e^2 + D_{obs}^2 \sigma_A^2 E_C^2}{1 - D_{obs}^2 \sigma_A^2} \right) I_0 \left(\frac{2E_e D_{obs} \sigma_A E_C}{1 - D_{obs}^2 \sigma_A^2} \right) \right)$$



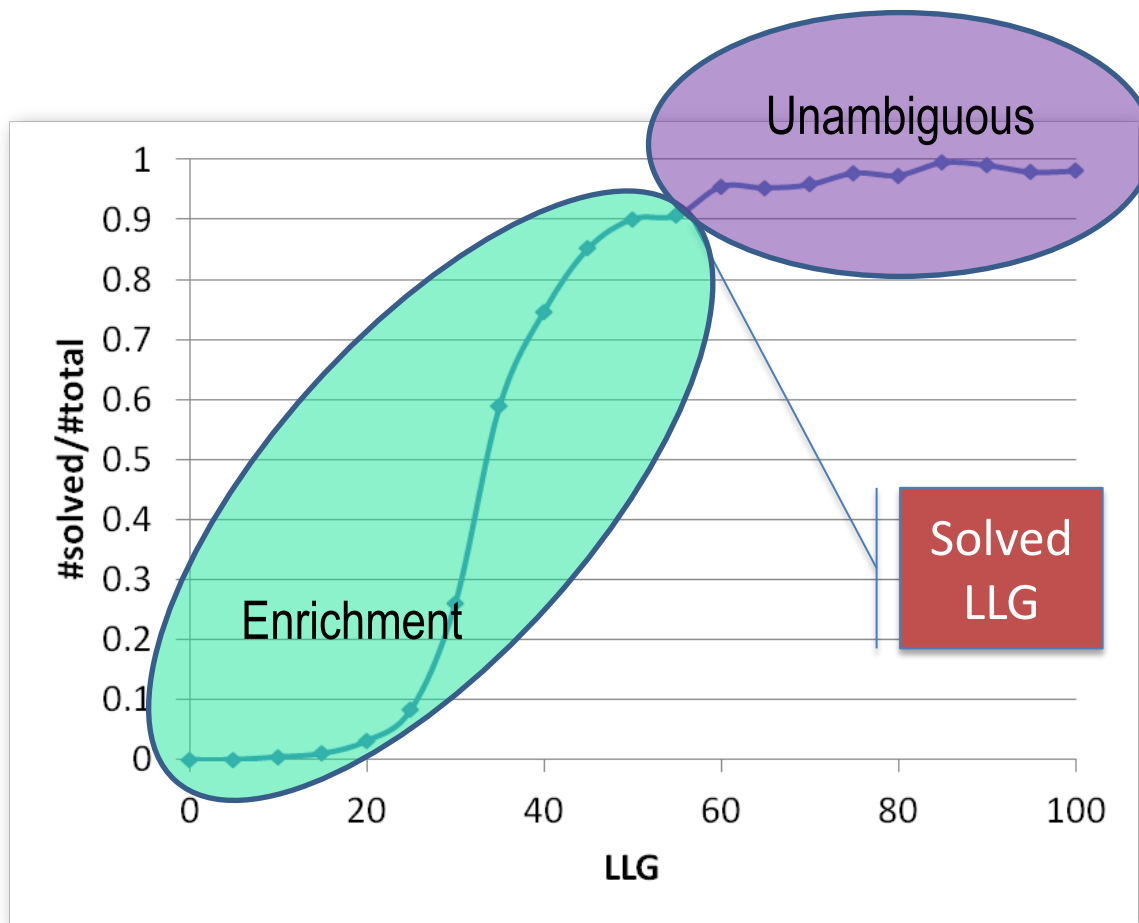
Peak selection

- 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

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

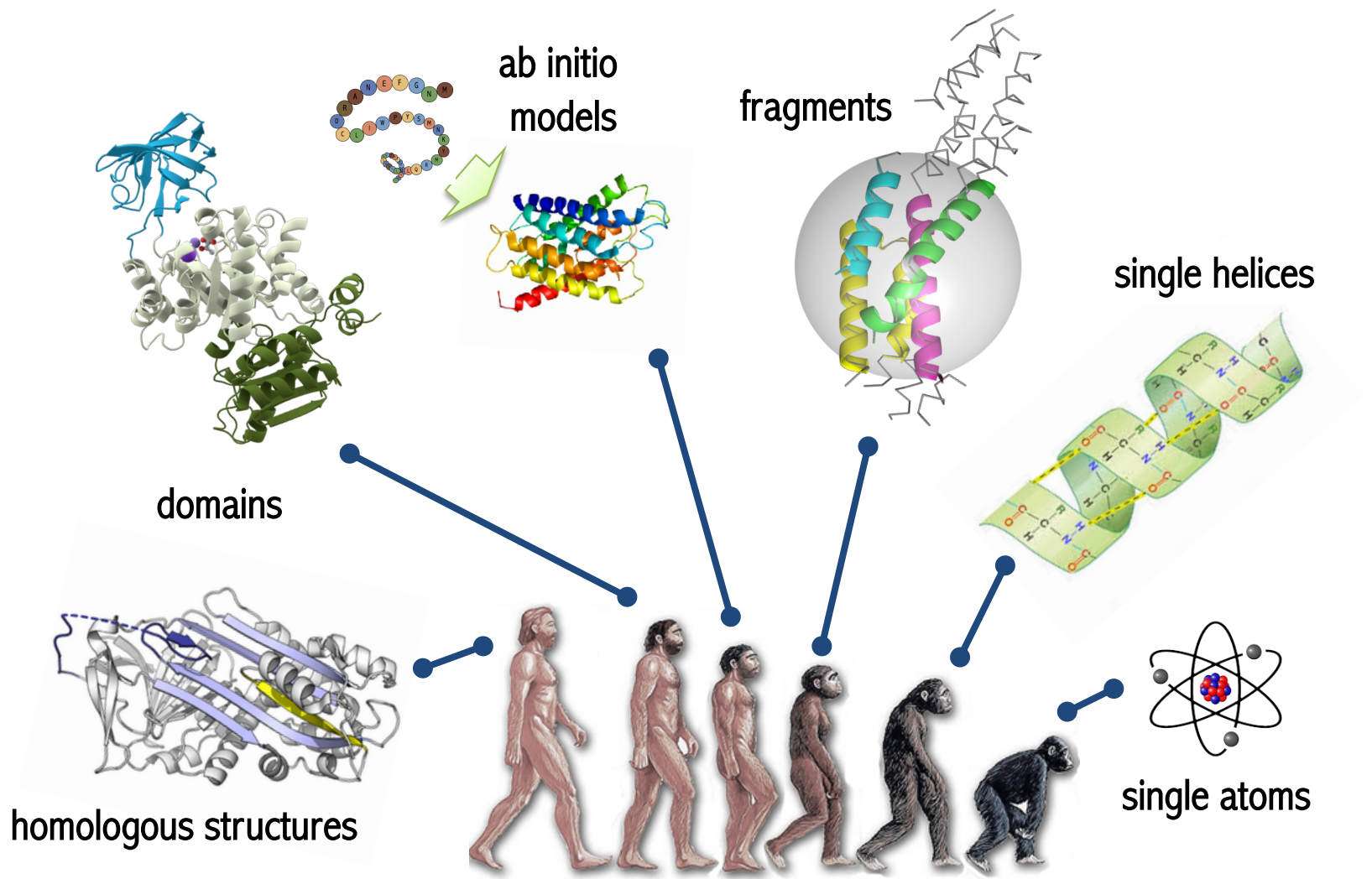
Models

All shapes and sizes

What chance of success?

- Better than 40% identity: usually easy, unless large conformational changes are involved.
- 30-40%: MR possible, but sometimes more difficult.
- 20-30%: MR difficult, careful model preparation required.
- Less than 20%: MR unlikely to work in most cases.
- In terms of RMSD, above 2.5Å is very unlikely to work, while 1.5Å or less is preferable.

attribution withheld



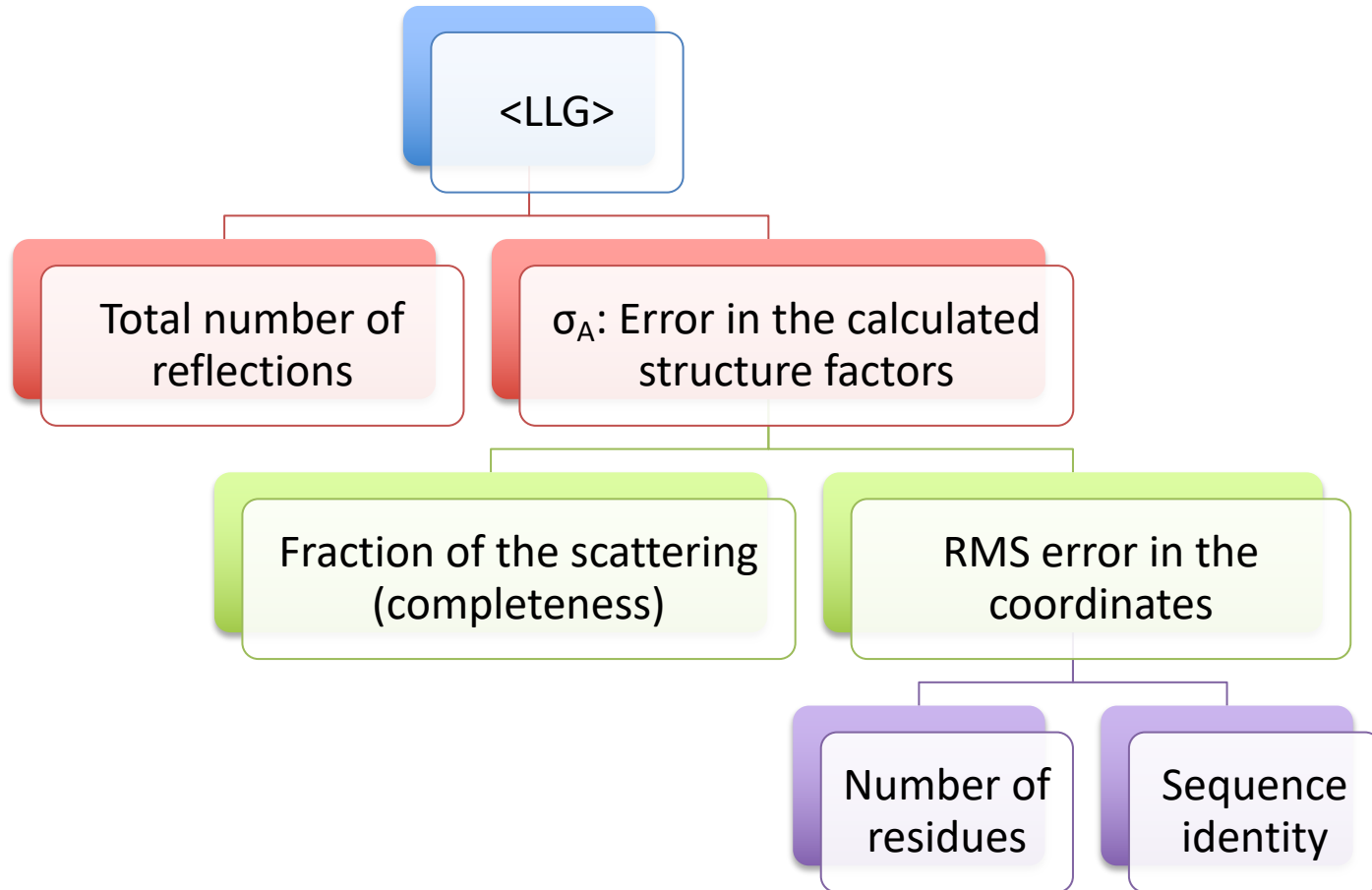
When is a model correctly placed?

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

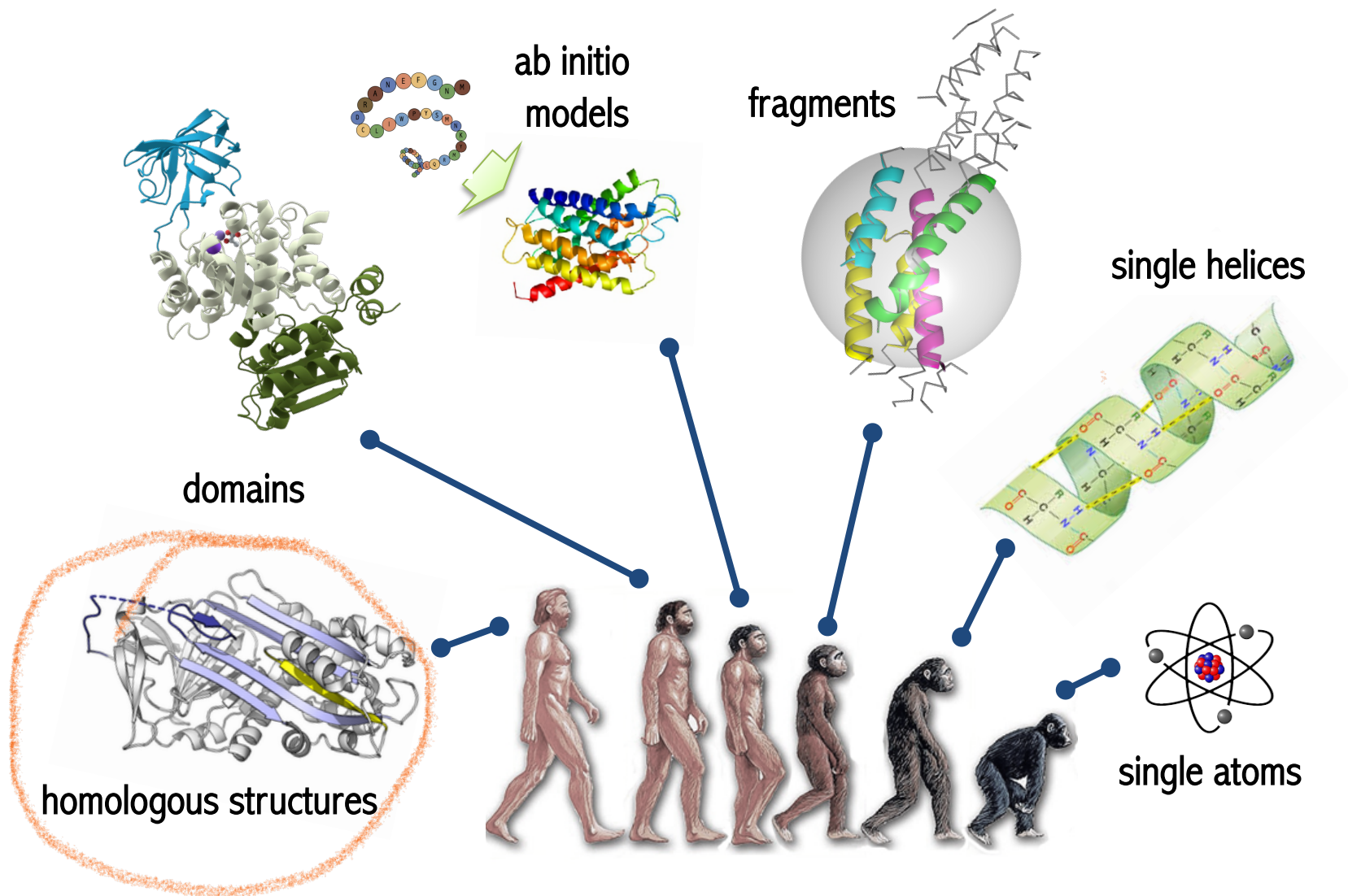
- So if you can predict the LLG...
 - You know how easy/difficult will be MR
 - You can prioritize structure solution strategies
- The expected LLG of a model predicts the LLG for a placed model
- Removes uncertainty in MR
 - Knowing when to start/stop has always been the problem with MR
 - Prioritize resources

Expected LLG

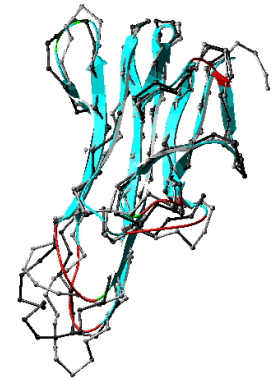


Expected LLG

- $\text{ellg} > 120$: you are likely to solve your structure easily, with a clear single peak
- $\text{ellg} > 60$: you are likely to solve your structure with the solution being the top in the list
- $\text{ellg} > 30$: you are likely to have the correct solution in the list of solutions
- Smaller ellg values mean that your solution list will be less enriched

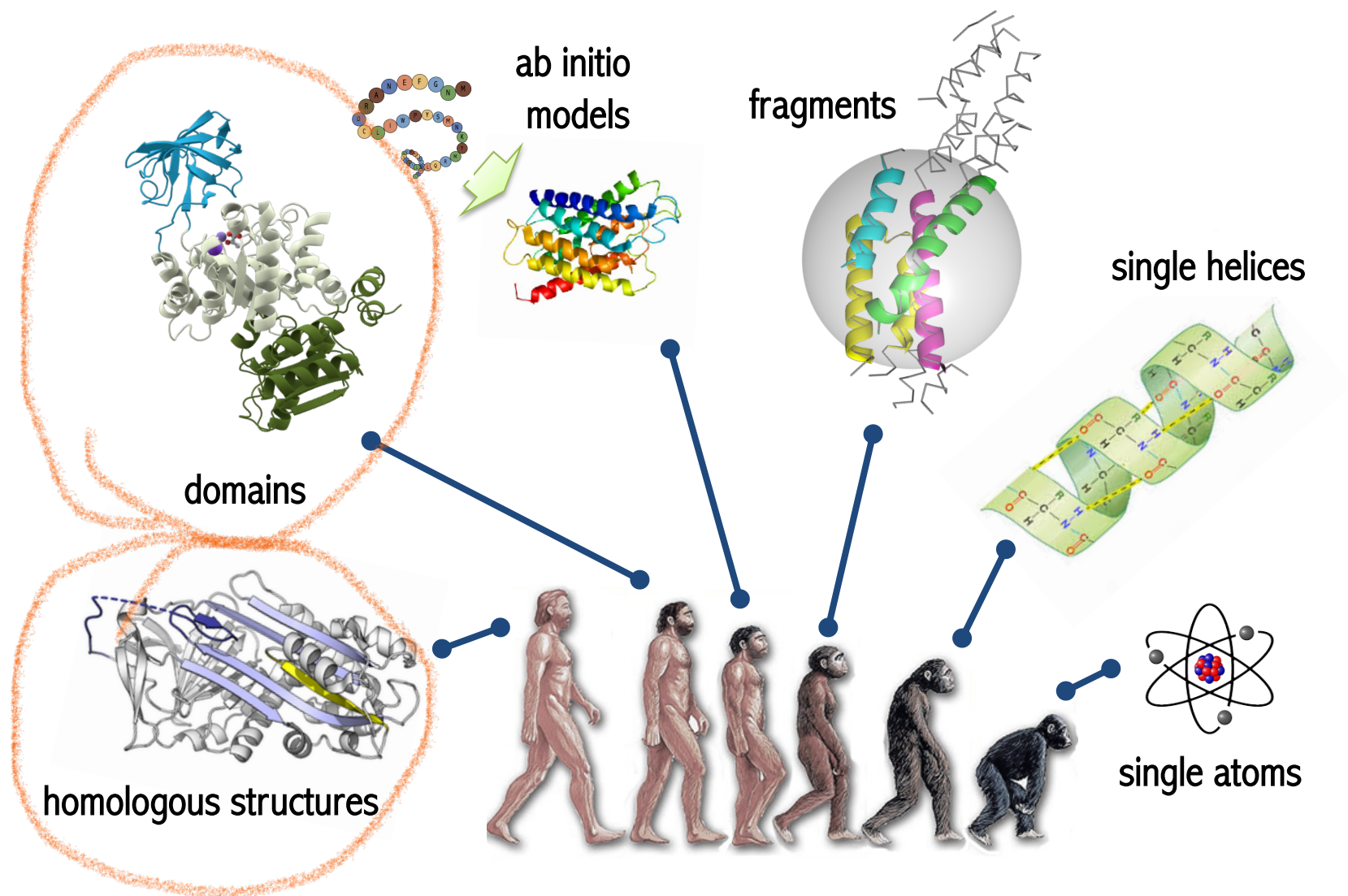


Ensembles

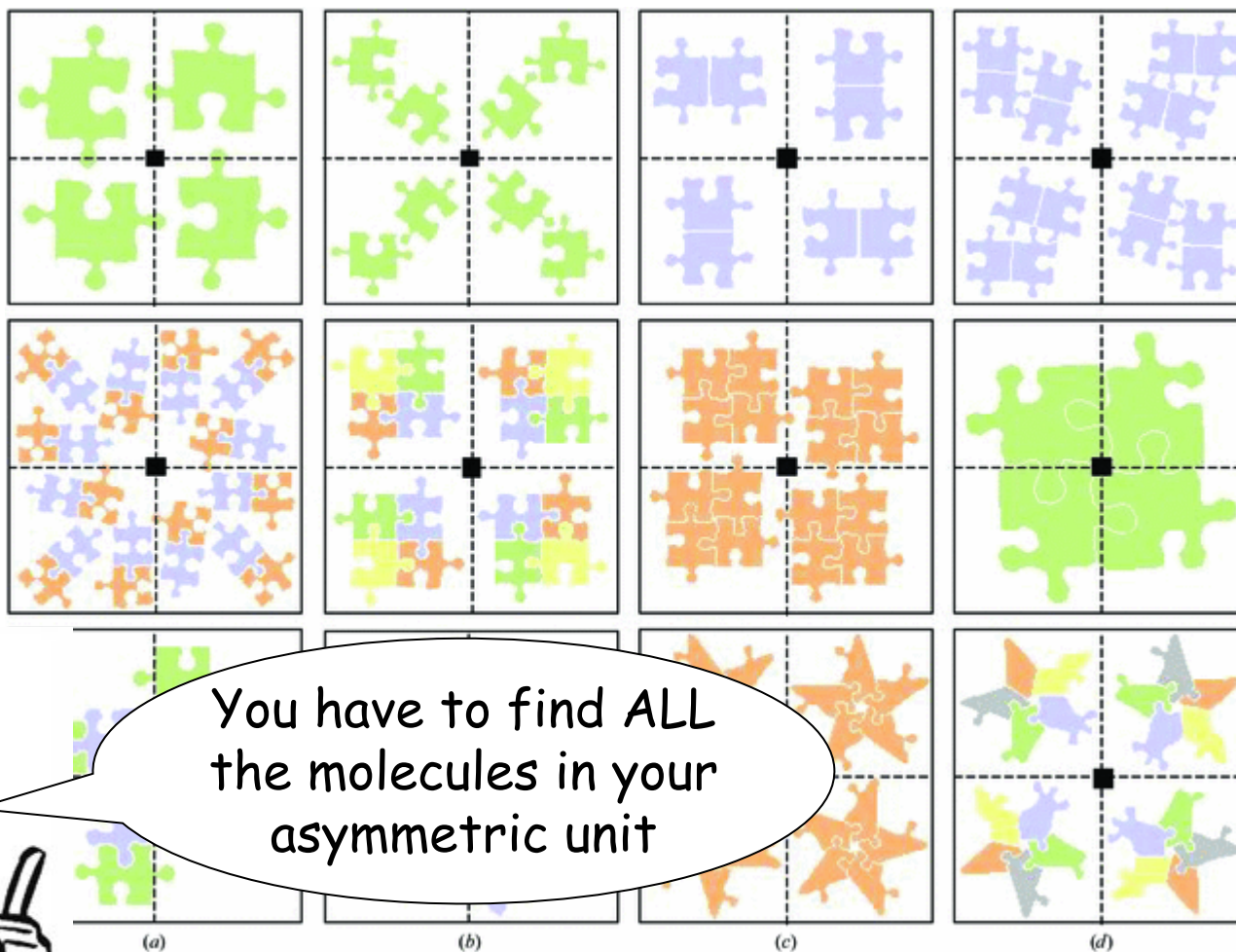


- Align homologs
- Places where the atoms overlap are reinforced
- Structures weighted by expected rmsd to target





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*

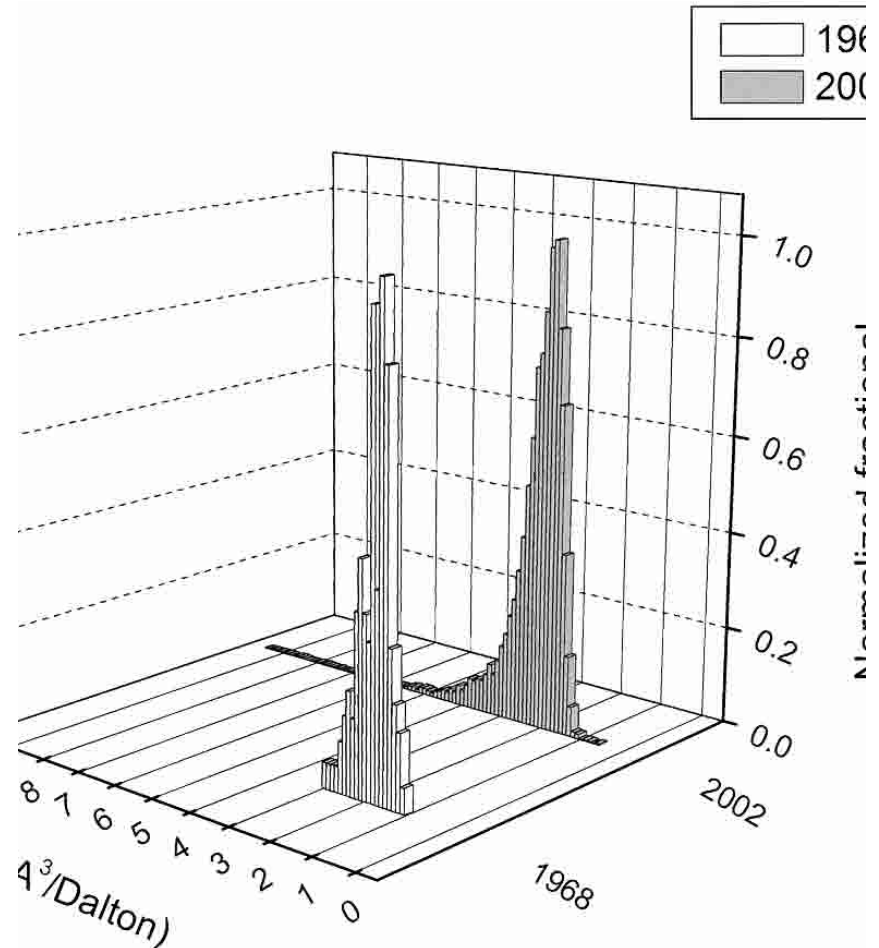


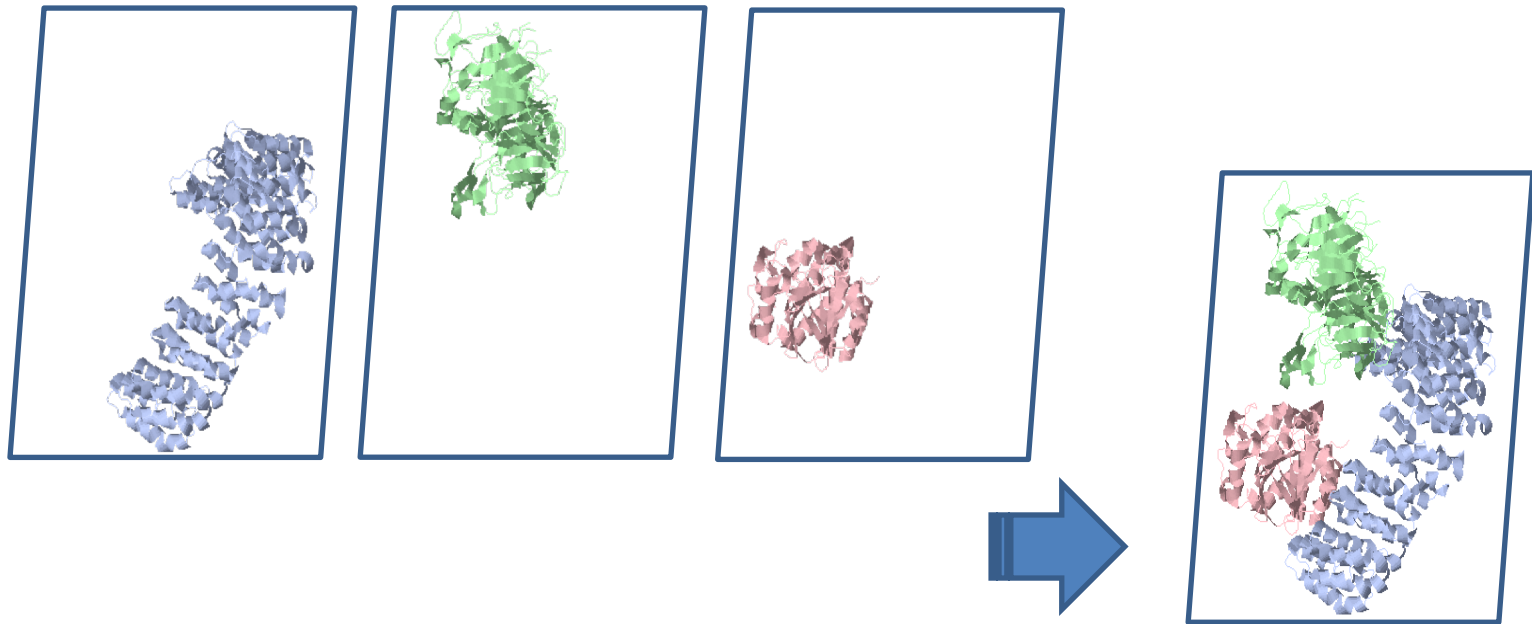
Figure 1: Kantardjieff and Rupp (2003)

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

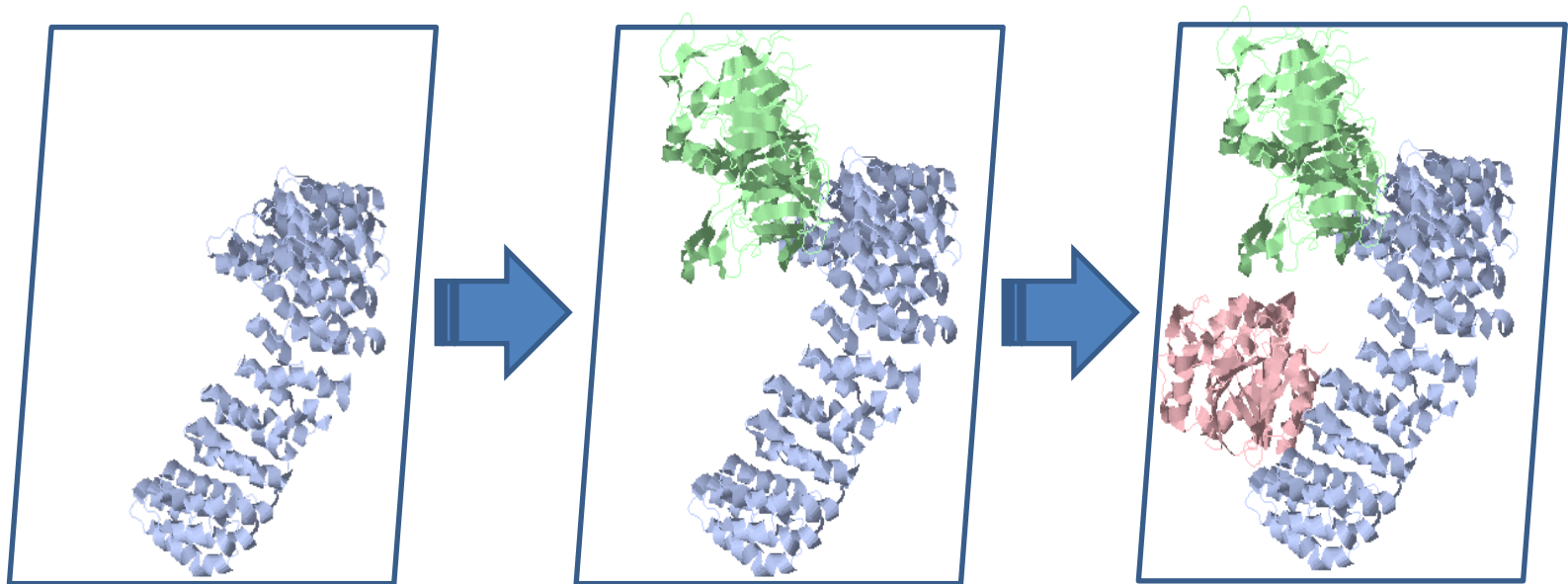
Searches for multiple components

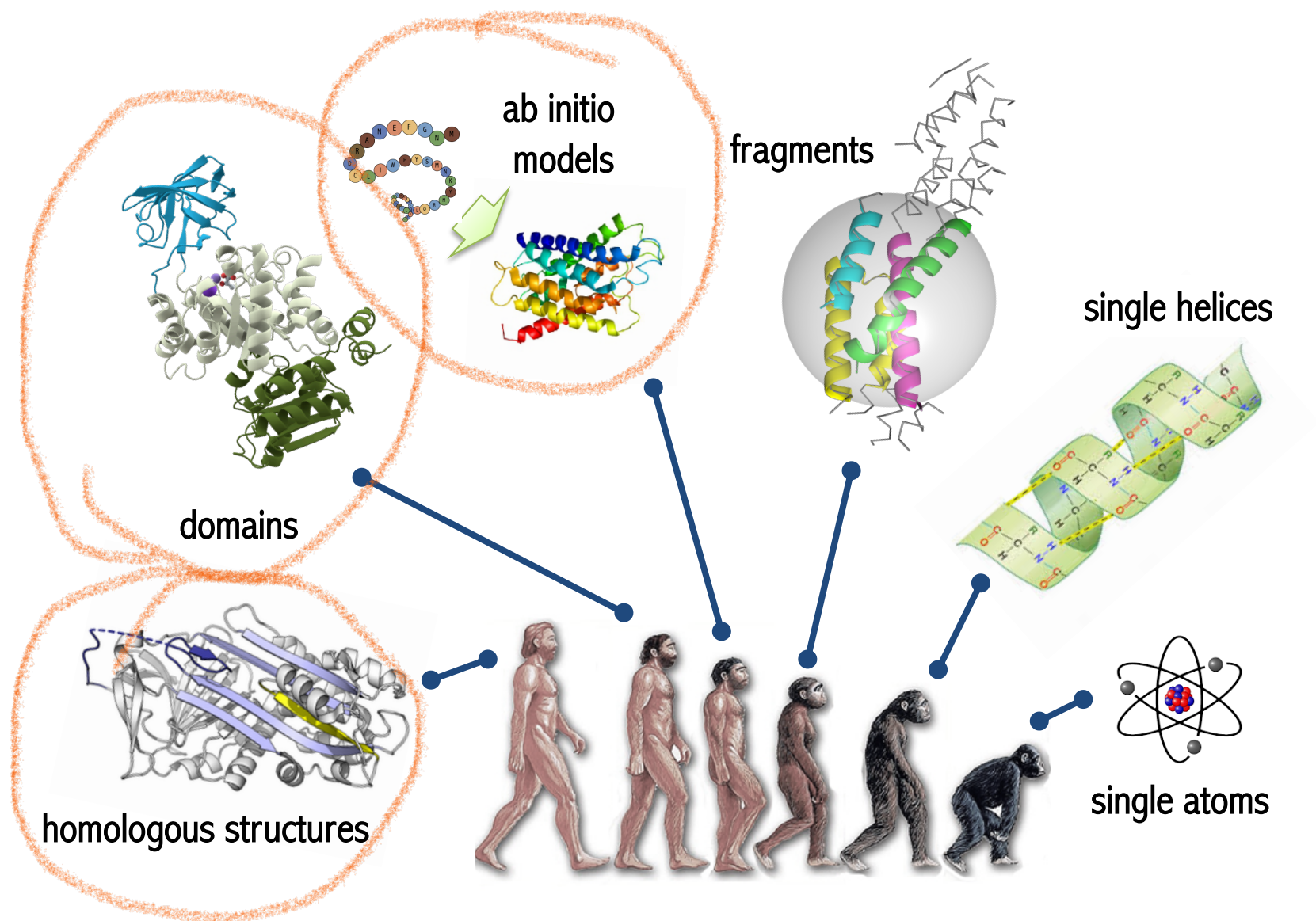
- It may not be possible to find each component in a separate search



Searches for multiple components

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





Accounting for errors

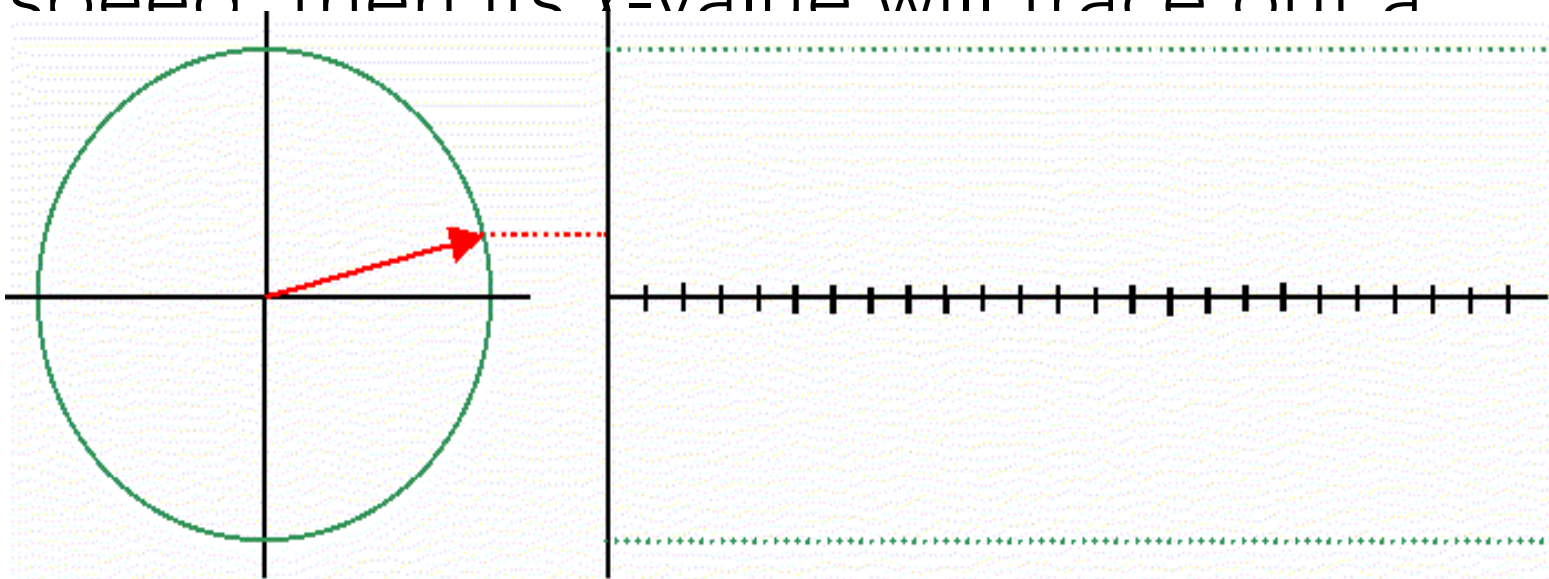
Basis for understanding

Molecular replacement

Experimental phasing

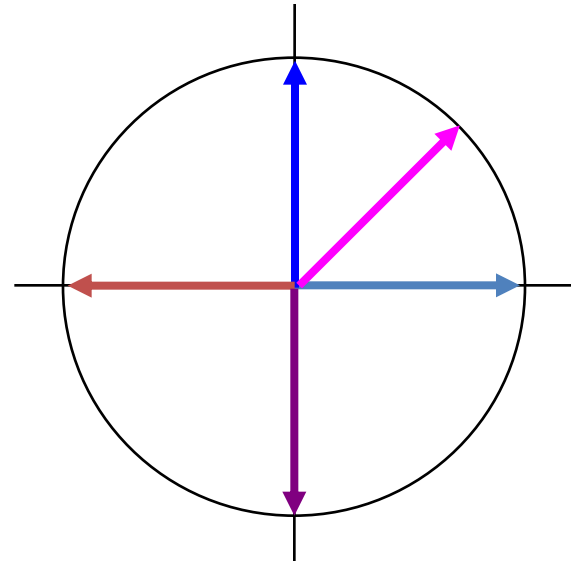
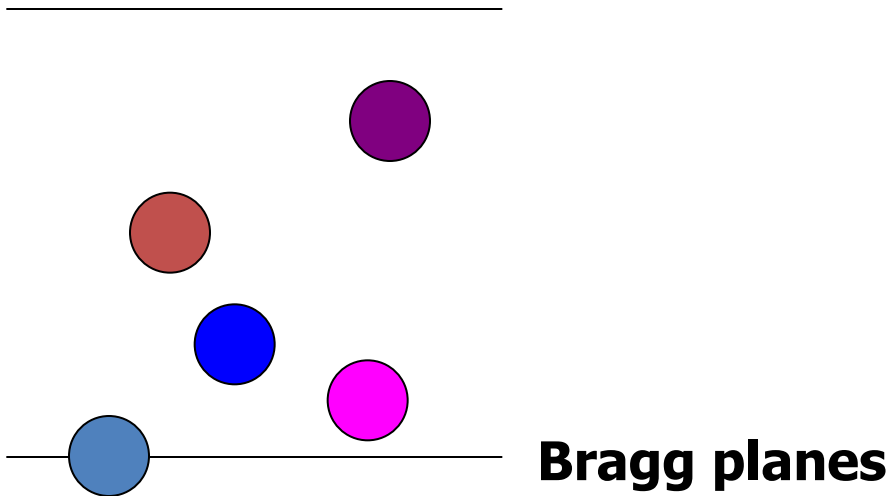
Waves as Vectors

- The cosine of an angle is the x component of a unit vector after it is rotated by that angle
- If the vector is rotated at some constant speed, then its x -value will trace out a



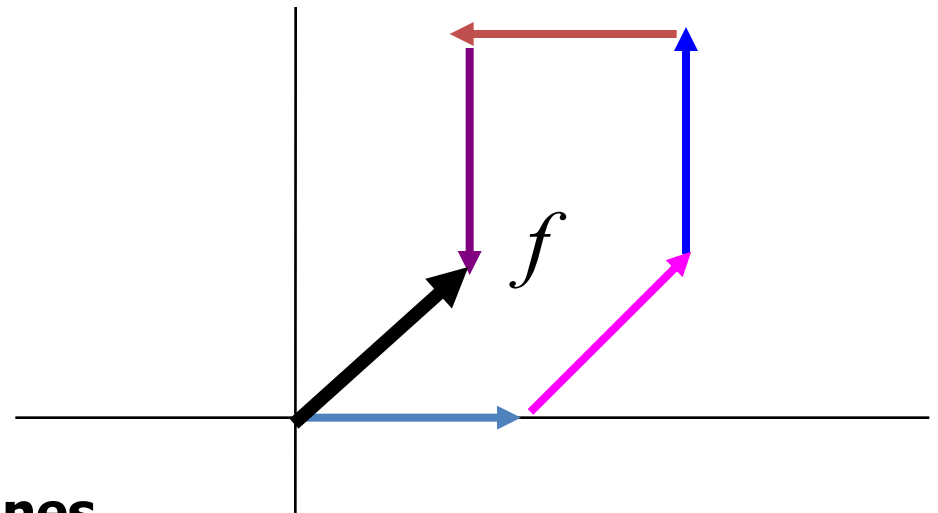
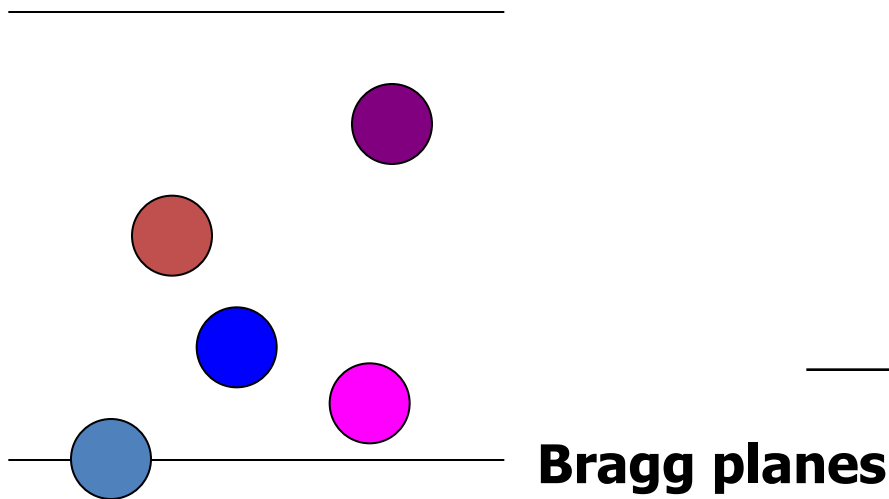
Atom distribution

- The diffraction wave from an atom can be represented by a vector (structure factor)
- Amplitude depends on the atom-type
- Phase depends on fractional distance between Bragg planes



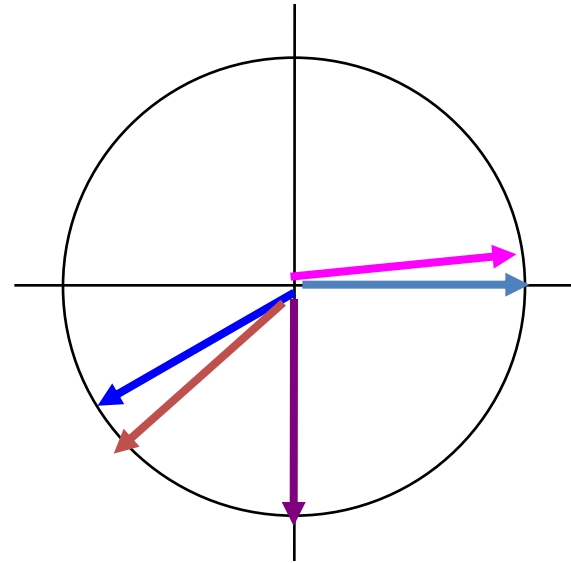
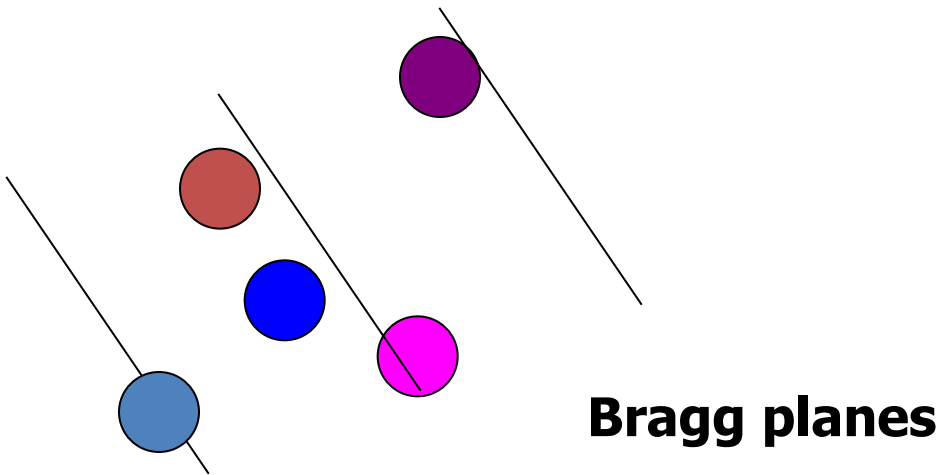
Summing structure factors

- To get the phase and amplitude of a wave diffracting from a Bragg plane, sum the structure factor vectors for the atoms
- Intensity of diffraction spot is square of the amplitude



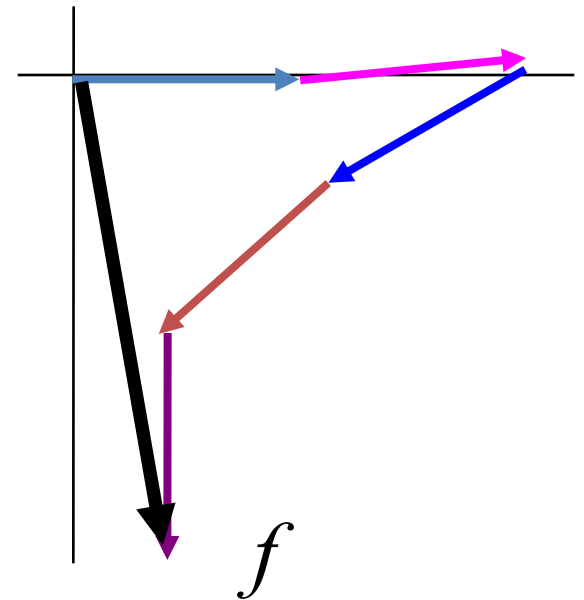
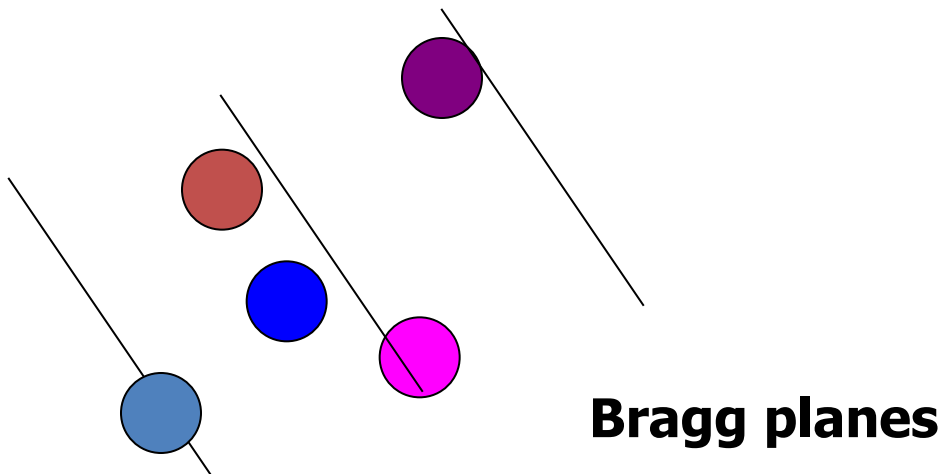
Atom distribution

- A different Bragg plane will give different phases for the atoms
- Amplitudes of atomic structure factors will be the same (same atomtype)

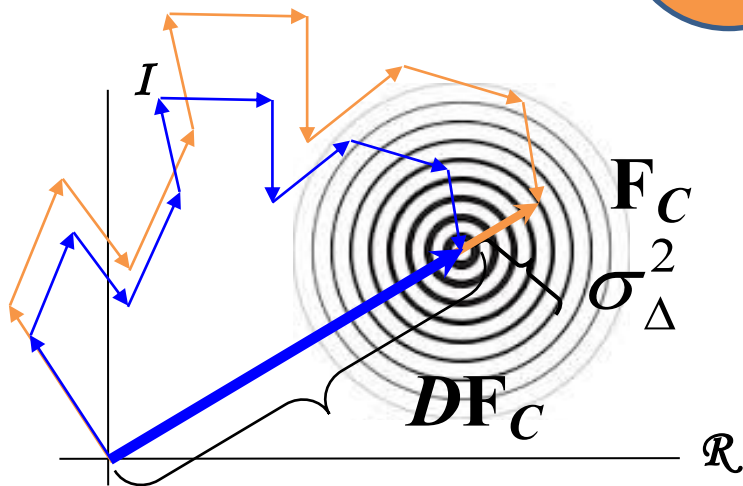
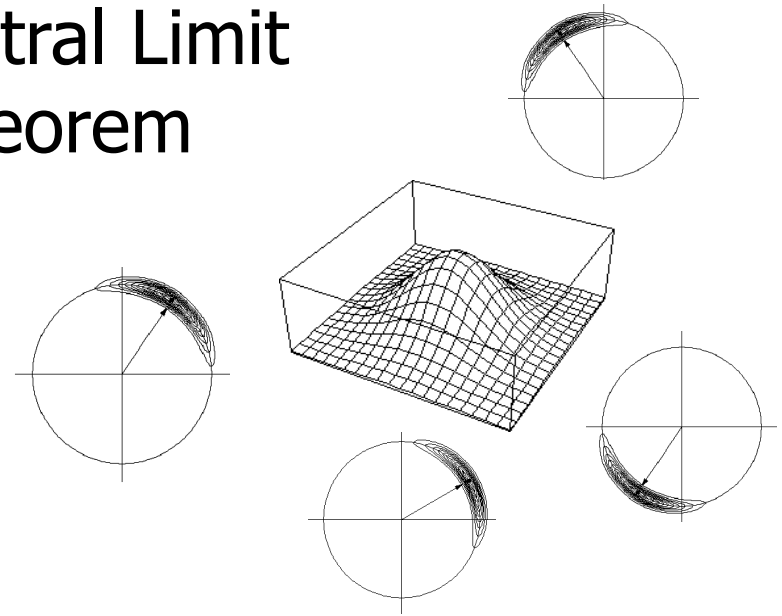
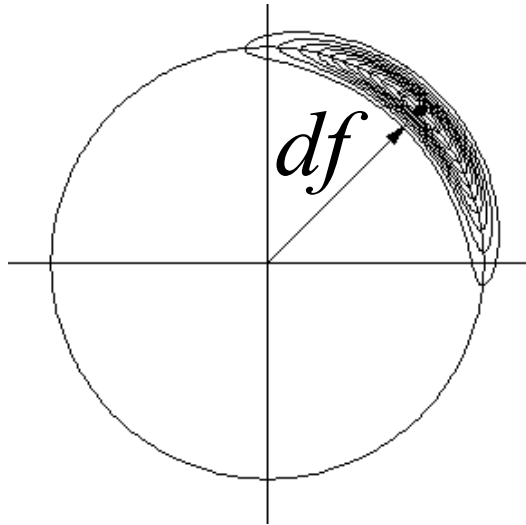


Summing structure factors

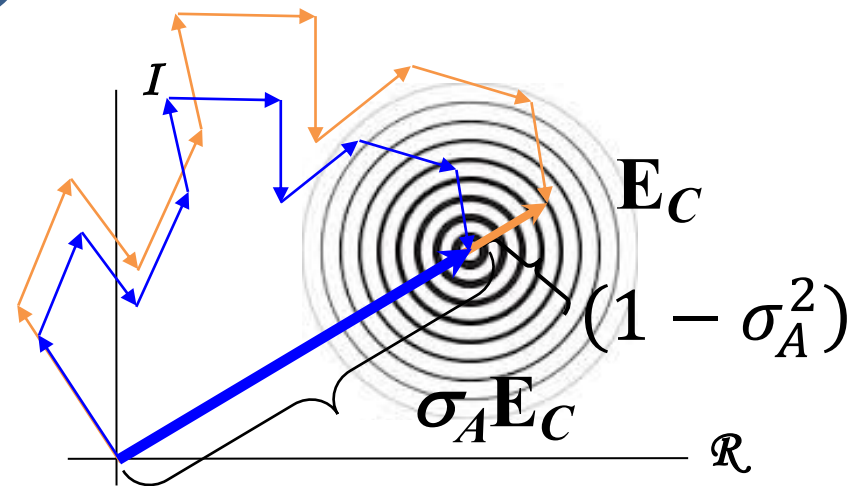
- A different Bragg plane will give different phases for the atoms and hence phase and amplitude for the resulting diffracted wave
- Different intensity



Central Limit Theorem



Structure Factor (F, D, σ_Δ)



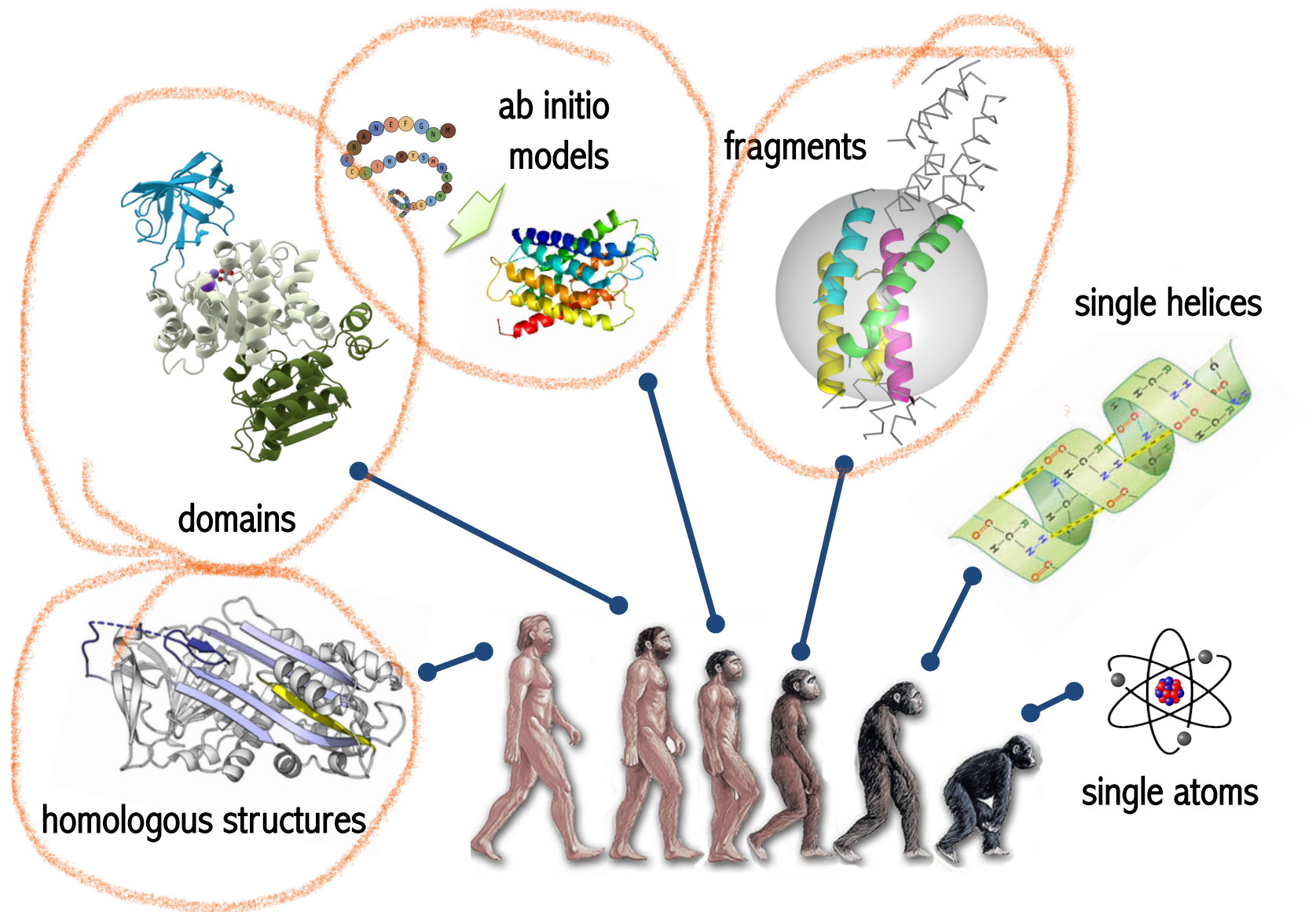
Normalized Structure Factor (E, σ_A)

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



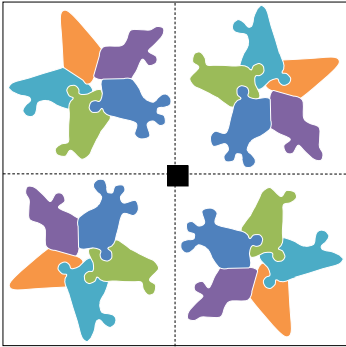
Fragments

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

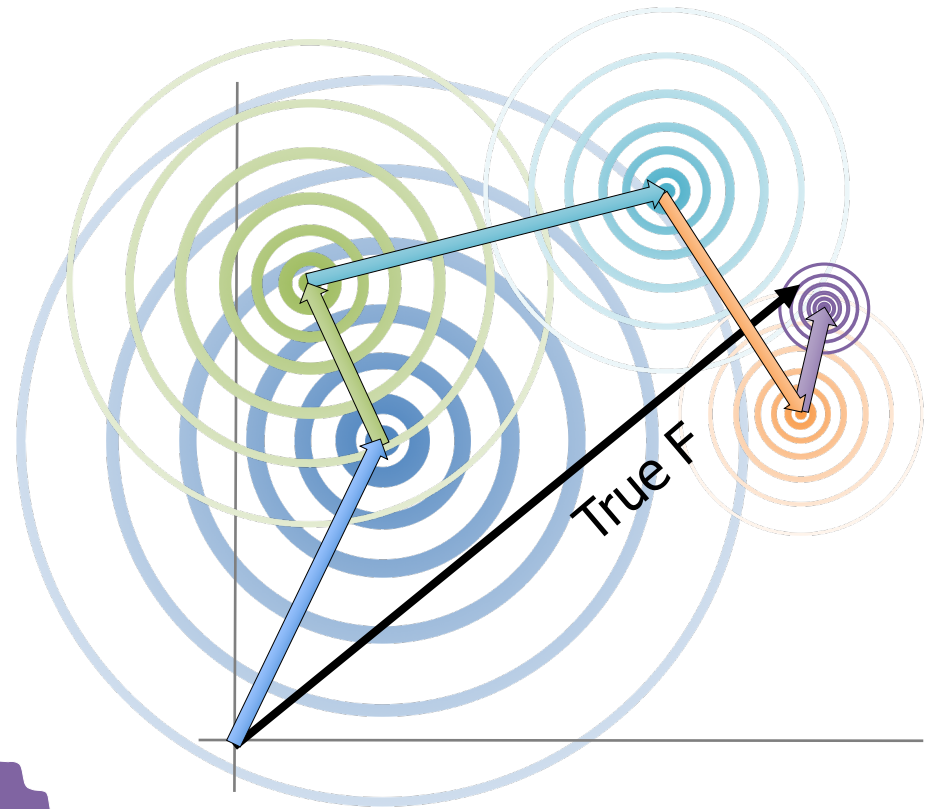
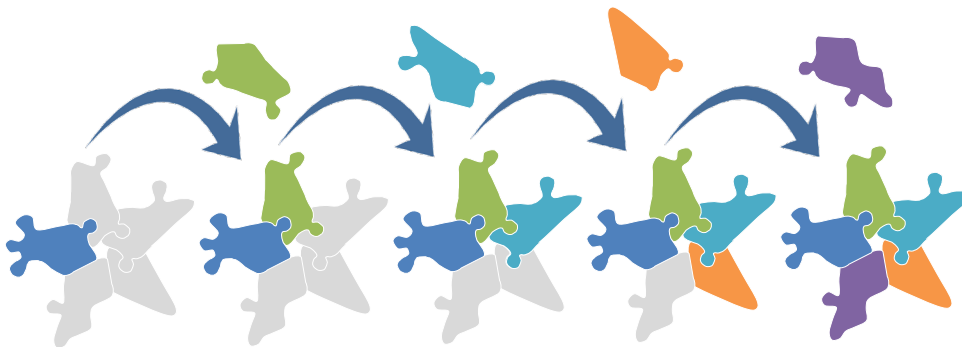
Success does NOT depend on

- The resolution (strictly)
- The type of fragment

σ_A and Fragments



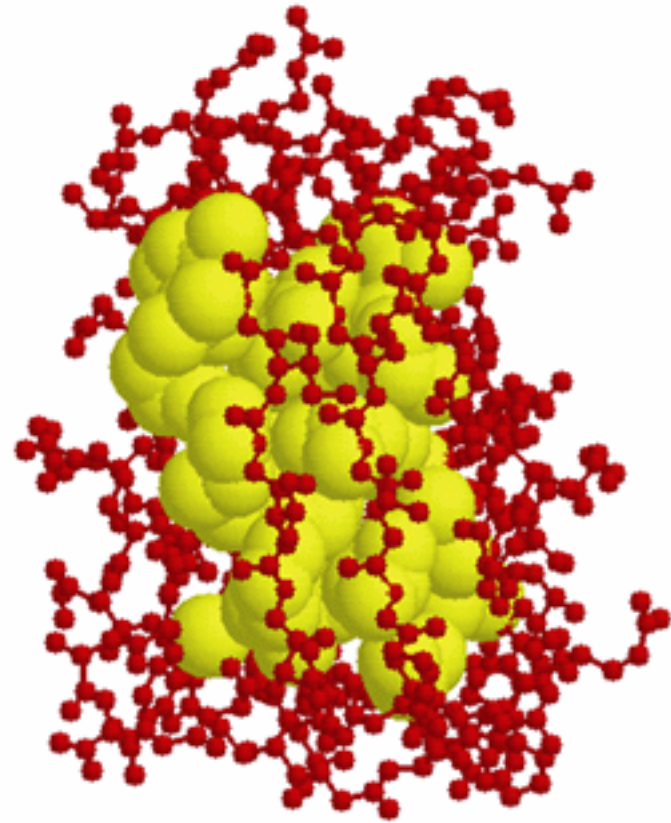
Structure builds
by addition



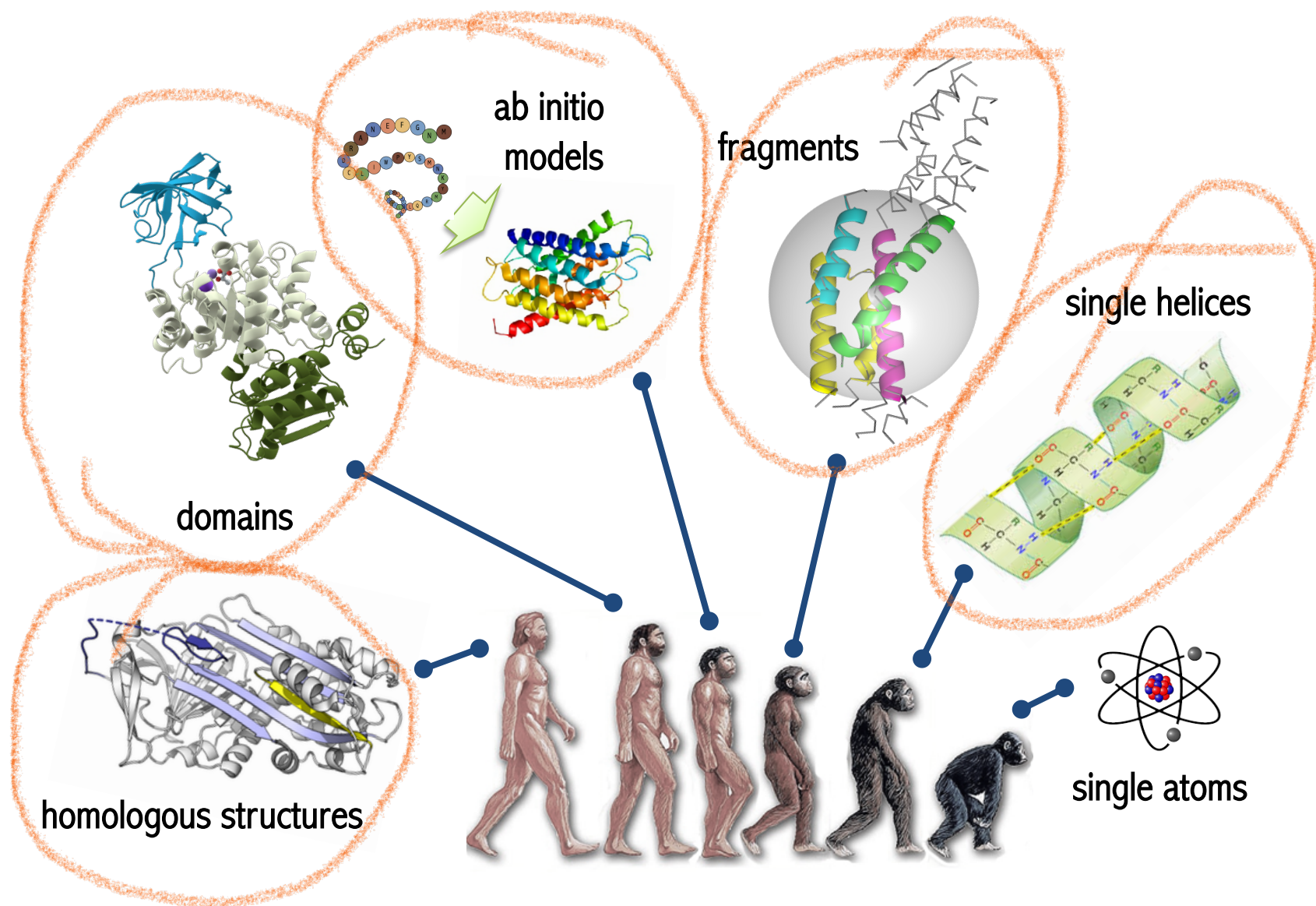
Whenever a component is placed,
the variances are reduced, thus
increasing the signal-to-noise of
the search for the next component

fragments

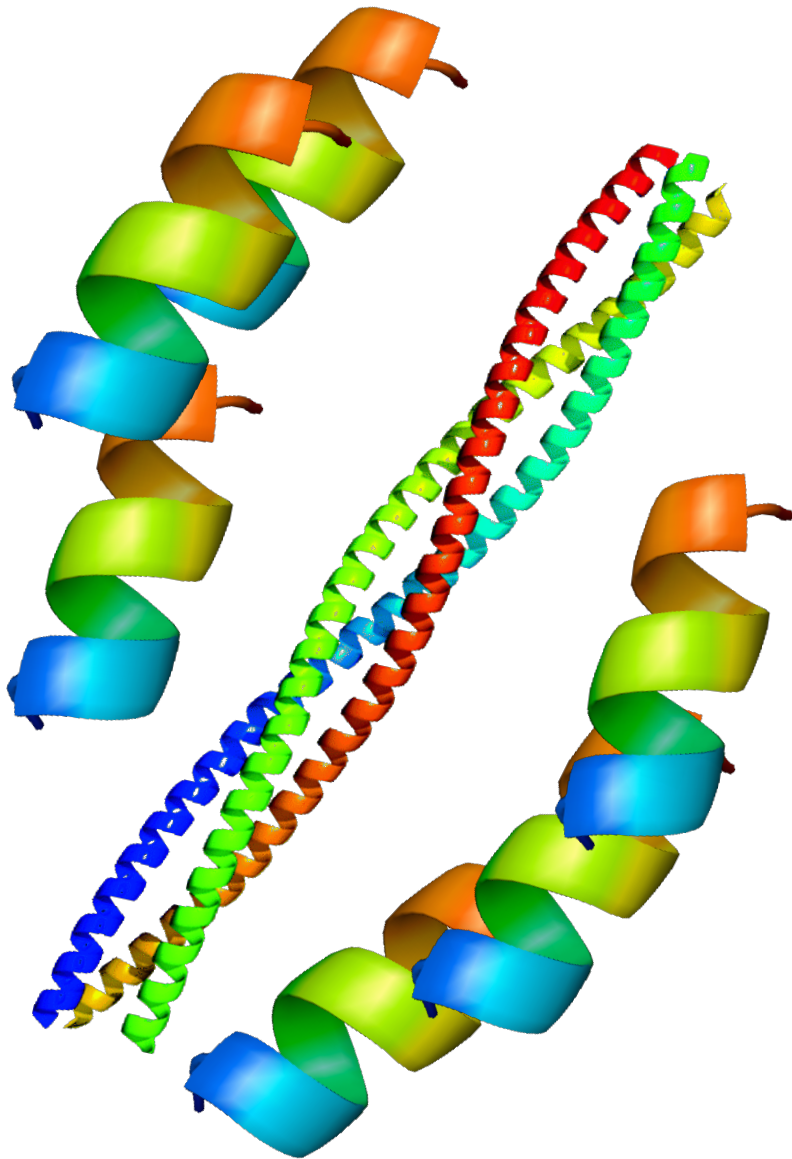
- If there is enough data and the model has low sequence identity to the target it can be advantageous to cut the model down to a small, accurate core
- ARCIMBOLDO-SHREDDER correlates solutions from overlapping fragments to extract signal



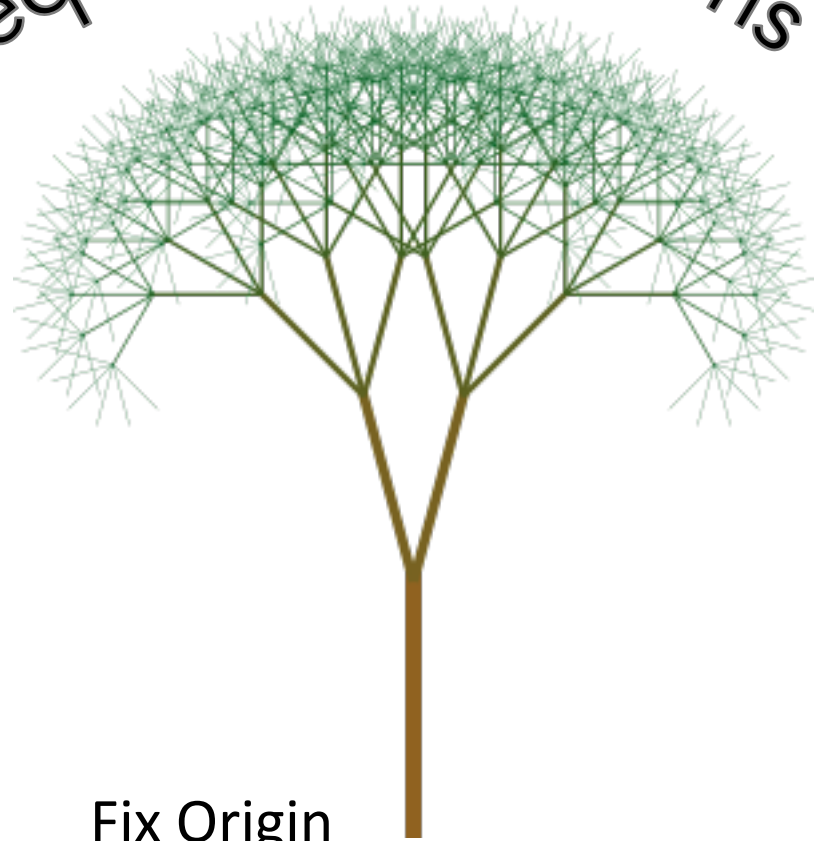
barstar
hydrophobic core in yellow
(Robert W. Hartley)



Helices

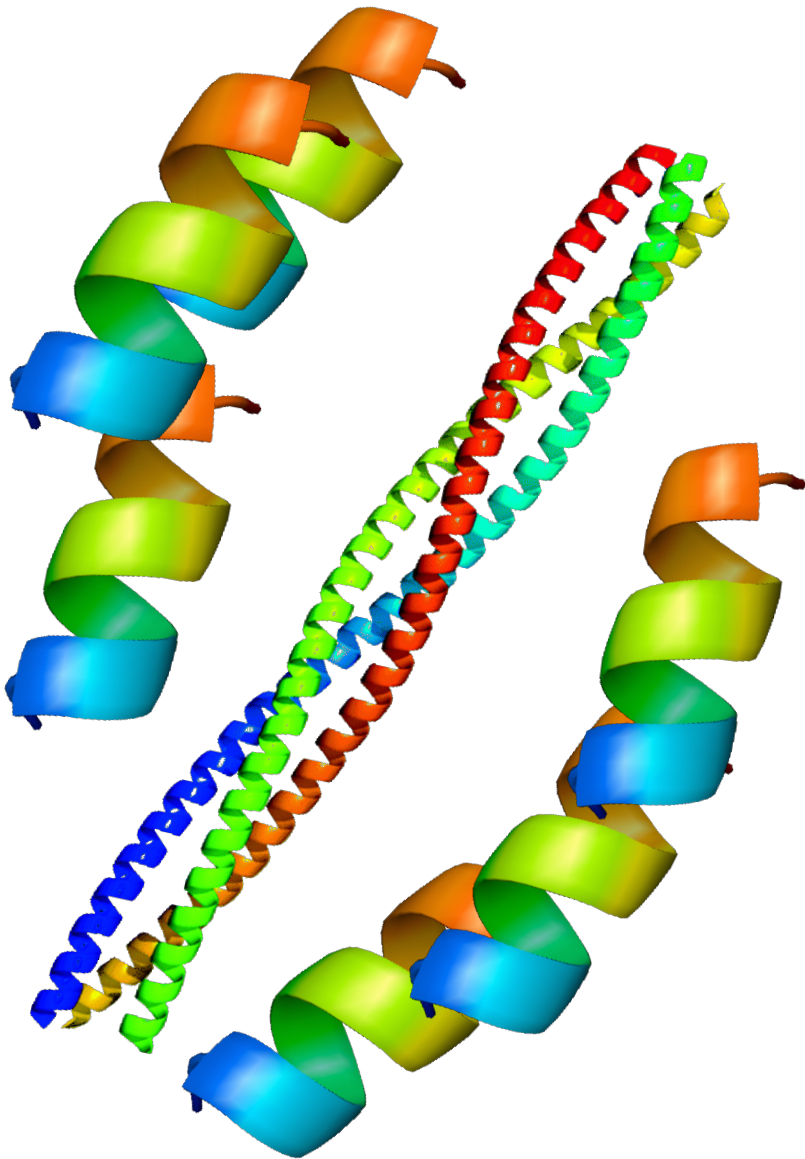


equivalent solutions



Fix Origin

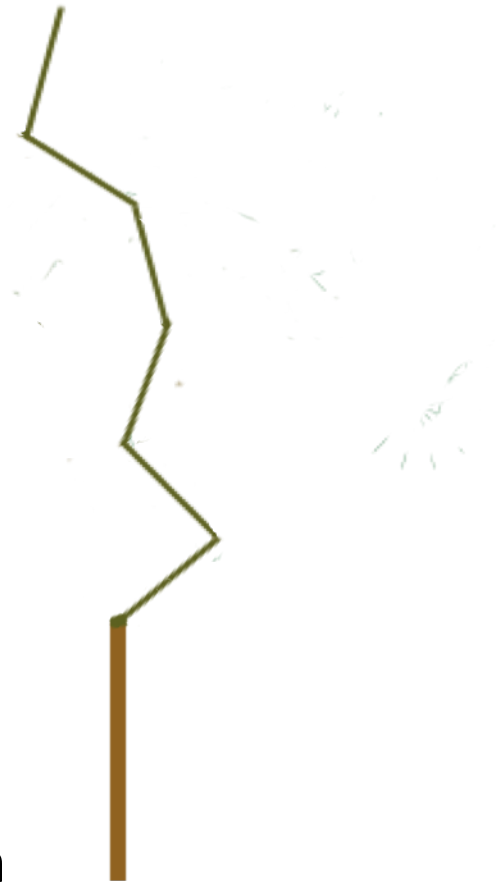
Helices

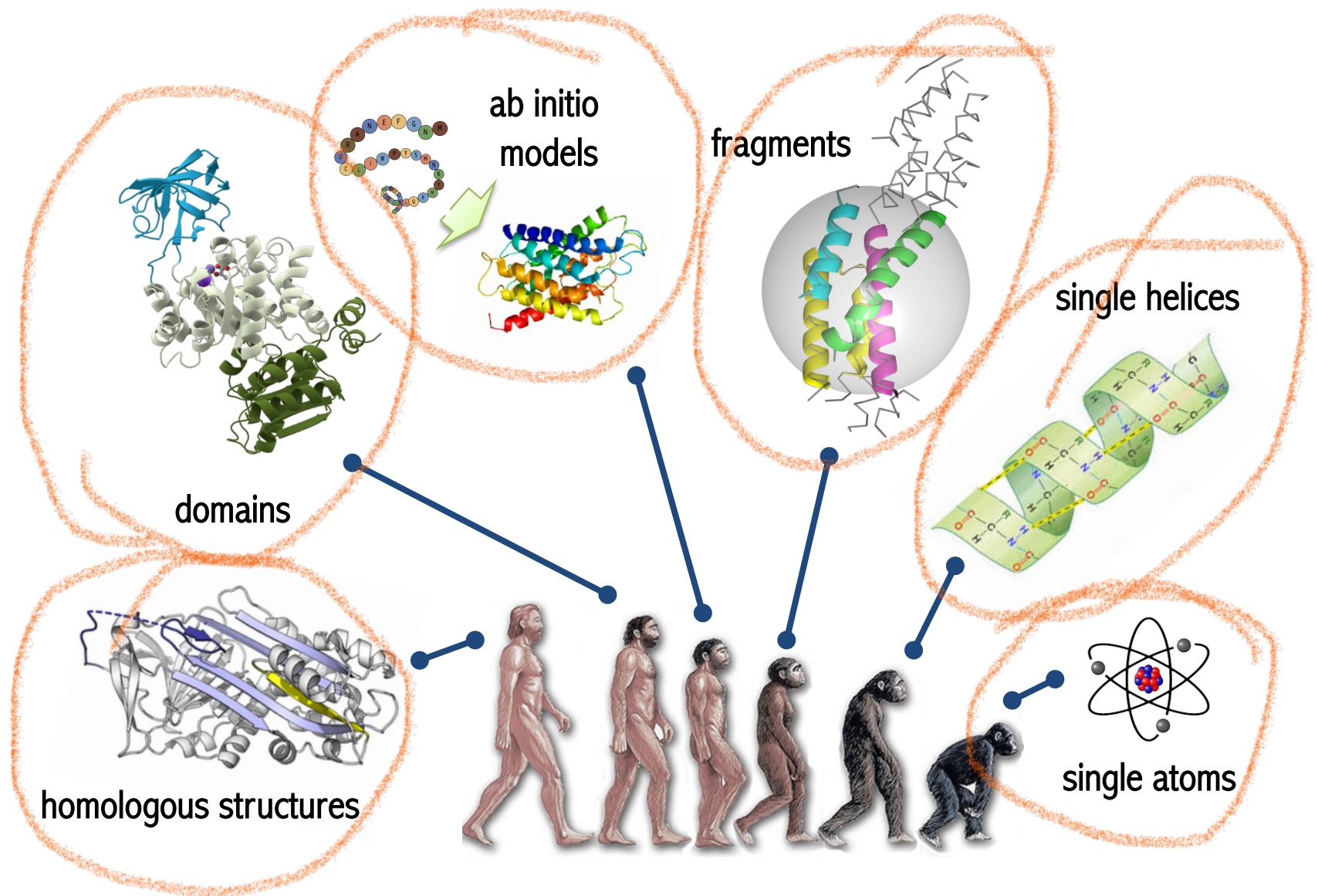


one solution

Amalgamation is possible because LLGI shows which intermediate structures are correct

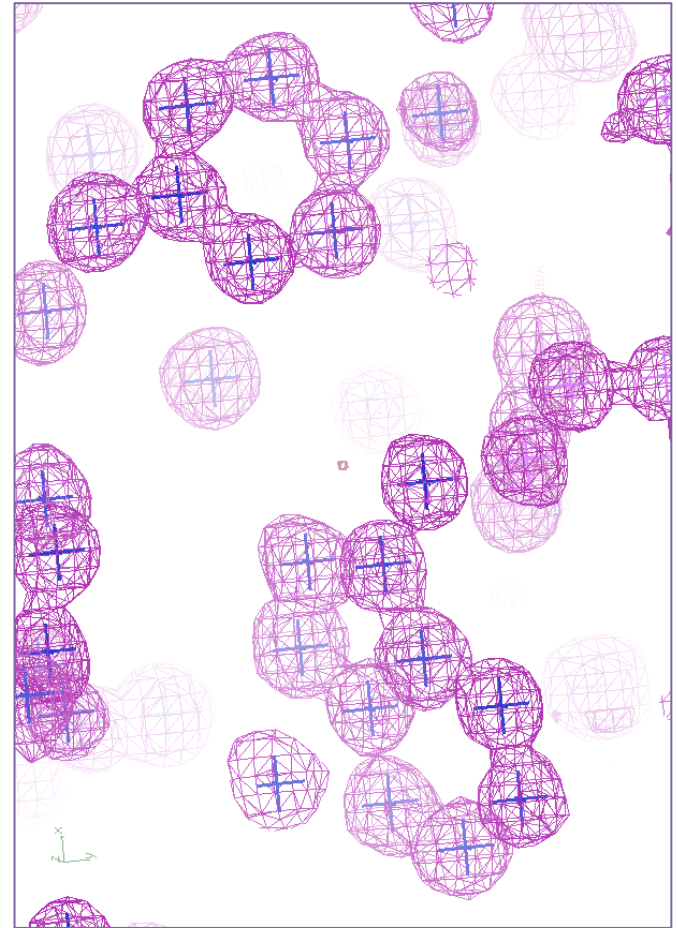
Fix Origin





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



Summary

- Fourier Transforms
 - and the phase problem
- SigmaA
 - rmsd
 - partial structure
- phaser
 - LLG and solution probability
 - expected LLG
 - Model selection criteria
 - Use Intensities
- Origins
- Translational non-crystallographic symmetry
- SAD phasing
 - Harker diagrams
 - Best phase
 - HL coefficients