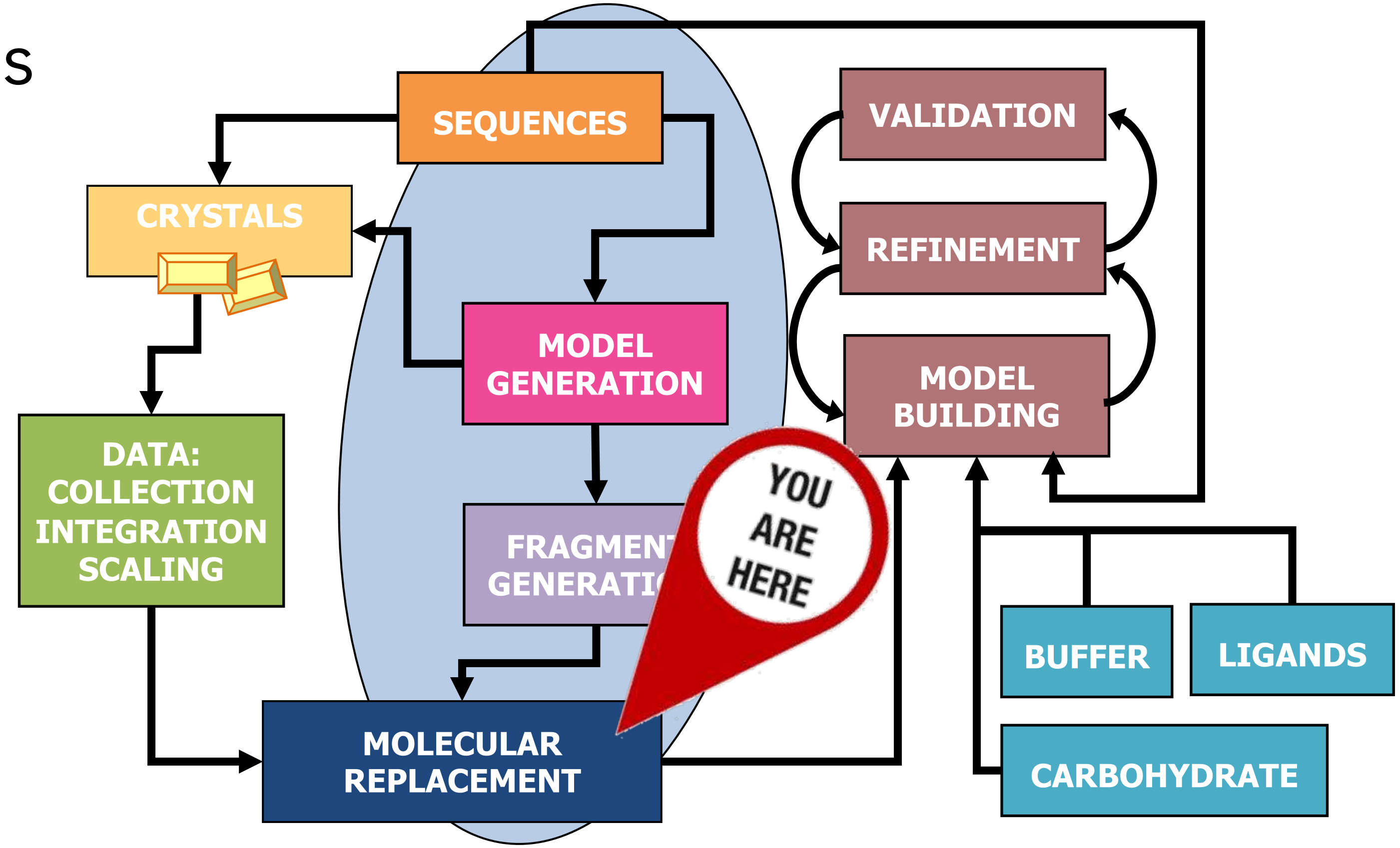
Molecular replacement

Airlie McCoy

UNIVERSITY OF
CAMBRIDGE

Molecular Replacement

1. prepare native crystals
2. collect and prepare data
3. <u>obtain the phases: molecular replacement</u>
4. model building
5. refinement and validation



with Phaser

- some things will be general to molecular replacement and
- some things specific to molecular replacement with Phaser

programs for molecular replacement

Phaser

Maximum Likelihood

MolRep
AMoRe
EPMR
COMO
CNS
XPLOR
SOMoRe

Patterson based

Maximum likelihood has many advantages over Patterson based methods

1. Increased signal by accounting for errors
2. Powerful statistics for predicting and judging results

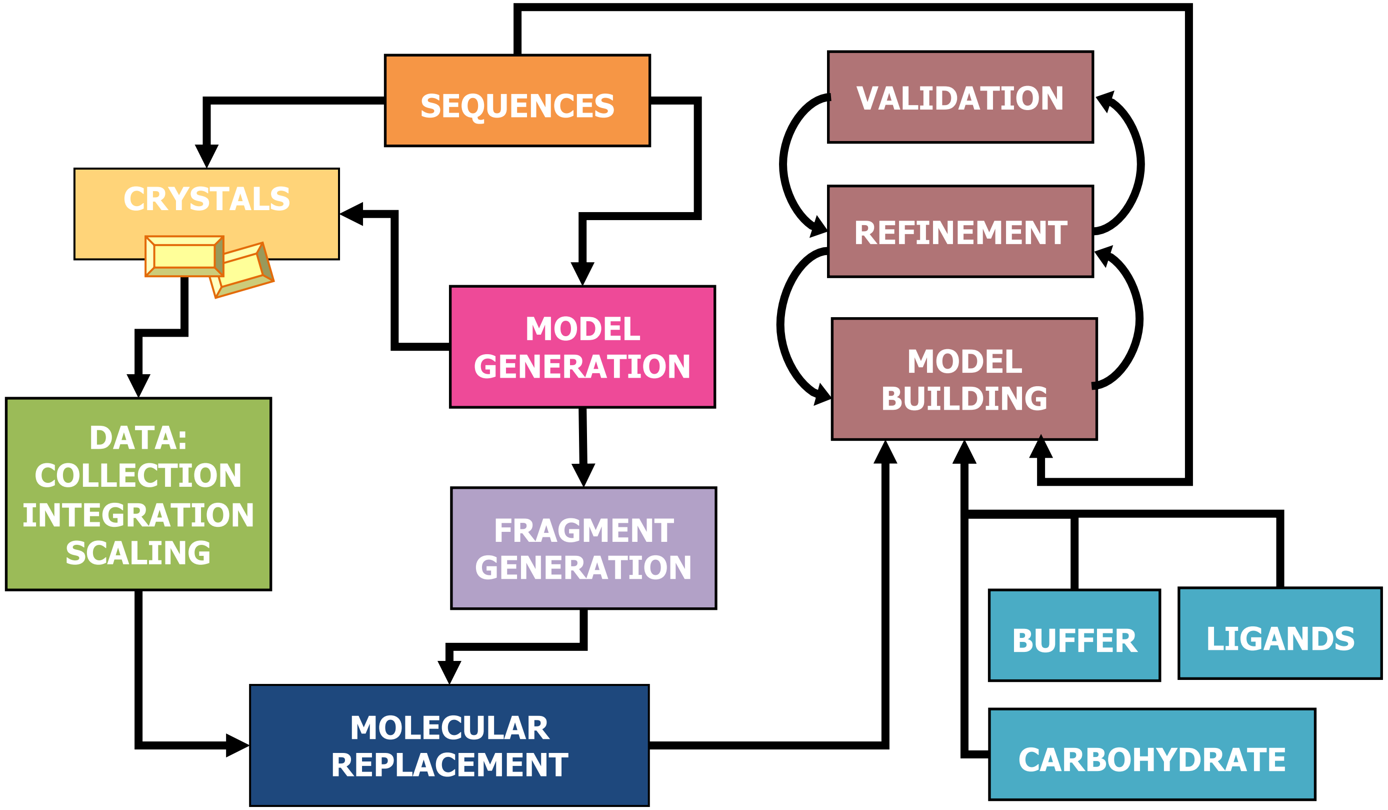
Phaser

- Phaser is software for
  - molecular replacement (MR)
  - single-anomalous dispersion (SAD)
  - molecular replacement with single-anomalous dispersion (MR-SAD)
  - EM docking for phased data ("EMplacement")

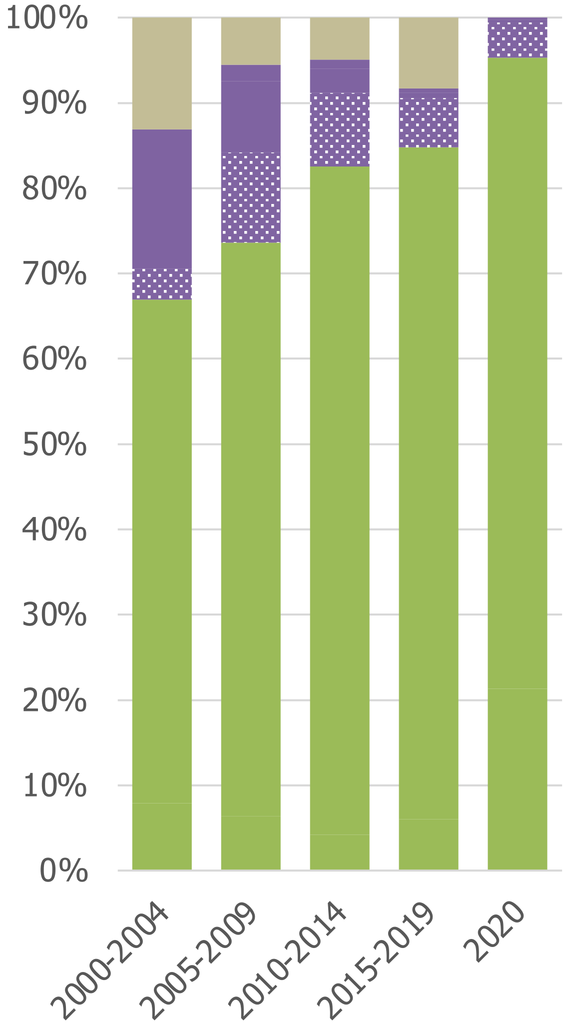
- Distribution
  - CCP4
  - Phenix

Phaser and pipelines

- Phaser performs MR in
    - Arcimboldo
    - MrBUMP
    - Balbes
    - Ample
    - CCP4 cloud
    - phenix.mr_rosetta
    - phenix.automr
    - phaser.MRage
    - MRGrid
- Phaser is also often run independently of pipelines



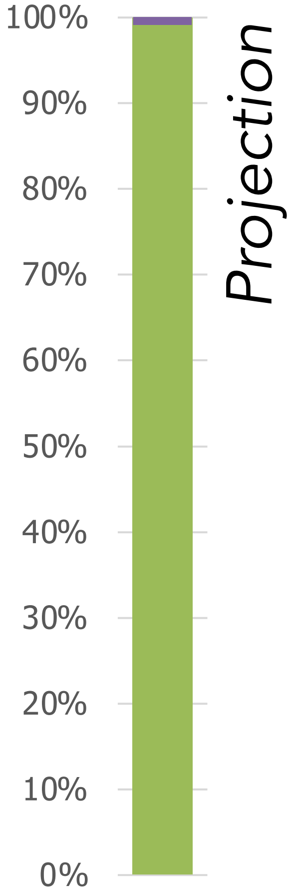
X-ray crystallography phasing methods in the pdb



🟩 Molecular replacement

🟪 Experimental Phasing

▨ SAD



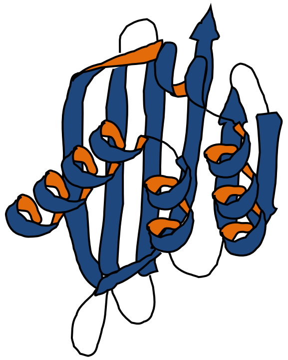
2021

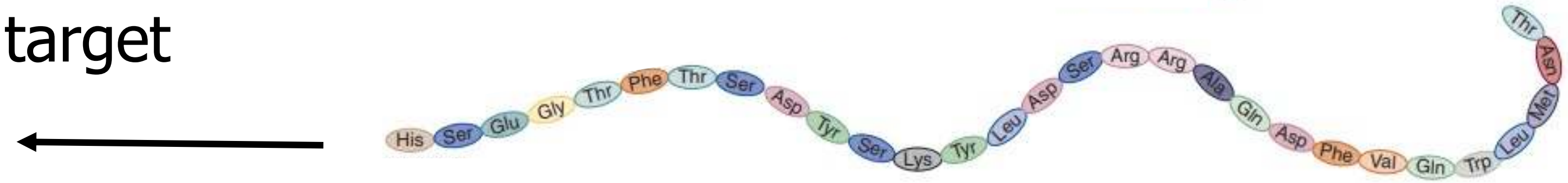
Arrival of Alphafold2 models

phasing by molecular replacement

molecular replacement

requires a model structure
with low rmsd to the target





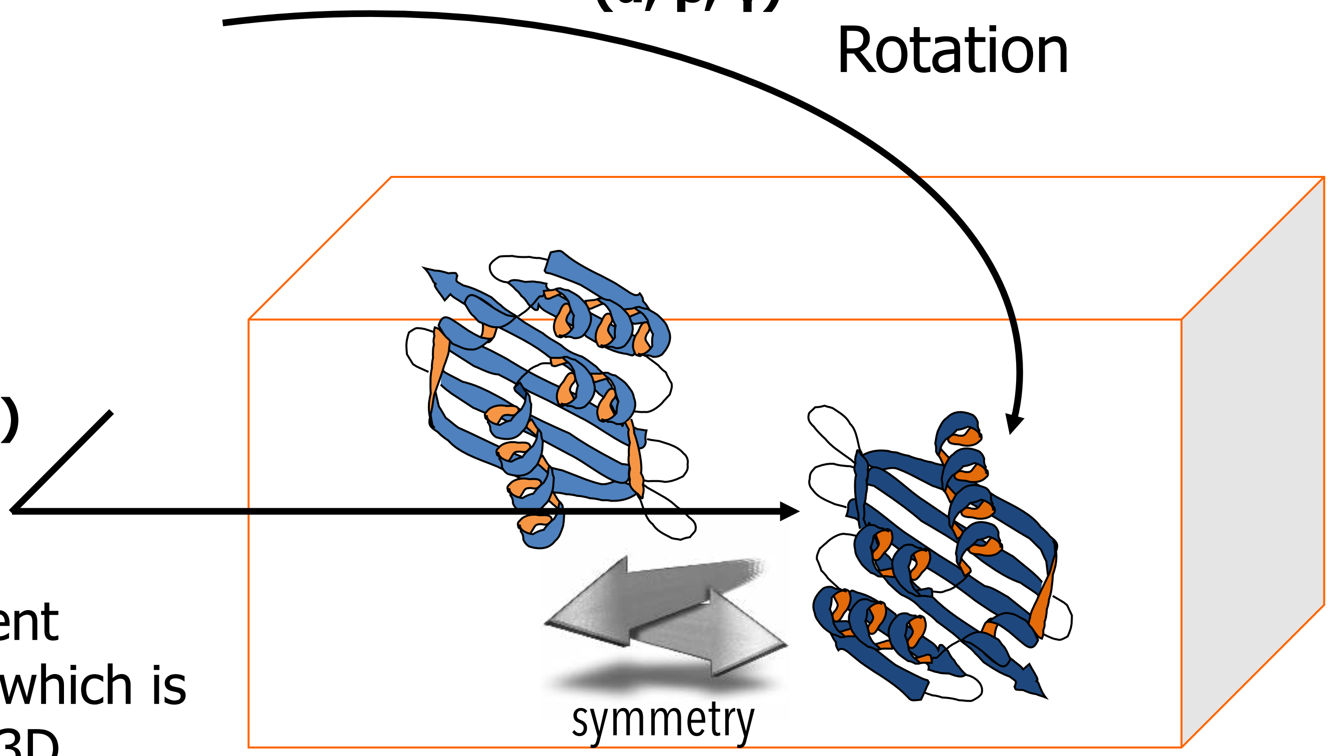
Known
sequence

place the model in the crystal so that it has lowest rmsd to the target structure and use the calculated phases from the model to kickstart refinement

Unknown
crystal
structure

molecular replacement – rotation and translation

requires a model structure
with low rmsd to the target



(α, β, γ)

Rotation

(X,Y,Z)

Translation

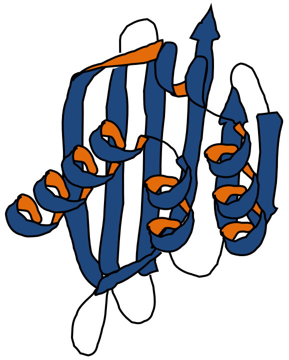
symmetry

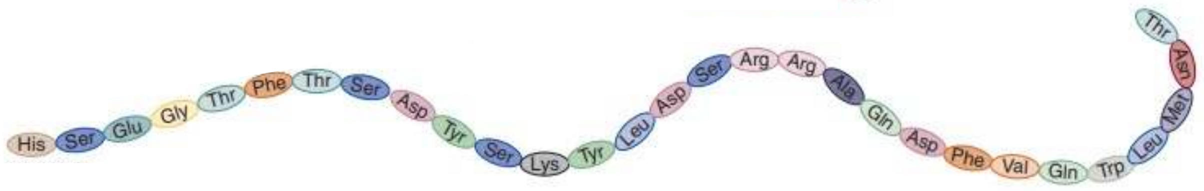
Unknown
crystal
structure

For one component
it is a 6D search which is
divided into two 3D
searches
- rotation
- translation

molecular replacement

requires a model structure
with low rmsd to the target





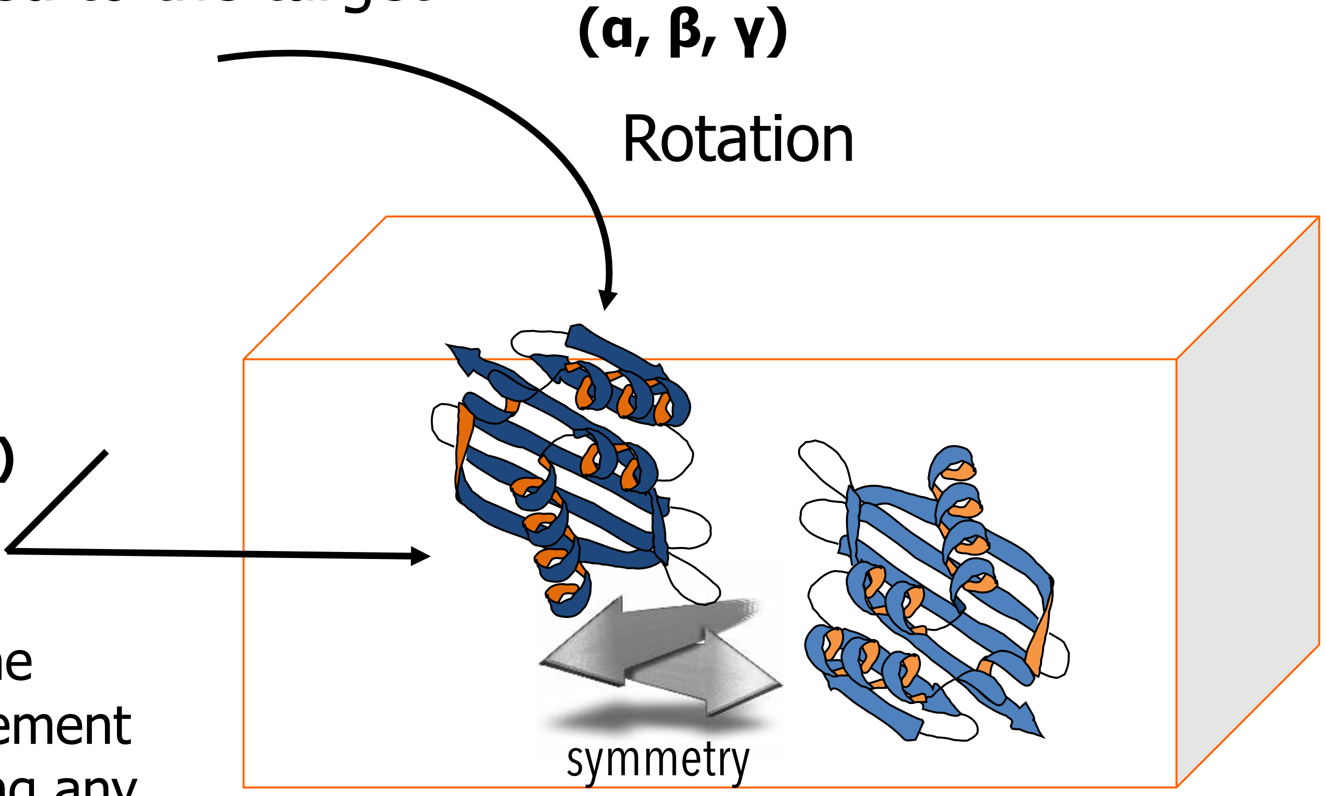
Known
sequence

Unknown
crystal
structure

You can ‘solve’ the molecular replacement problem by finding any one of the (symmetry related) copies in the crystal lattice

molecular replacement – rotation and translation

requires a model structure
with low rmsd to the target



(X,Y,Z)

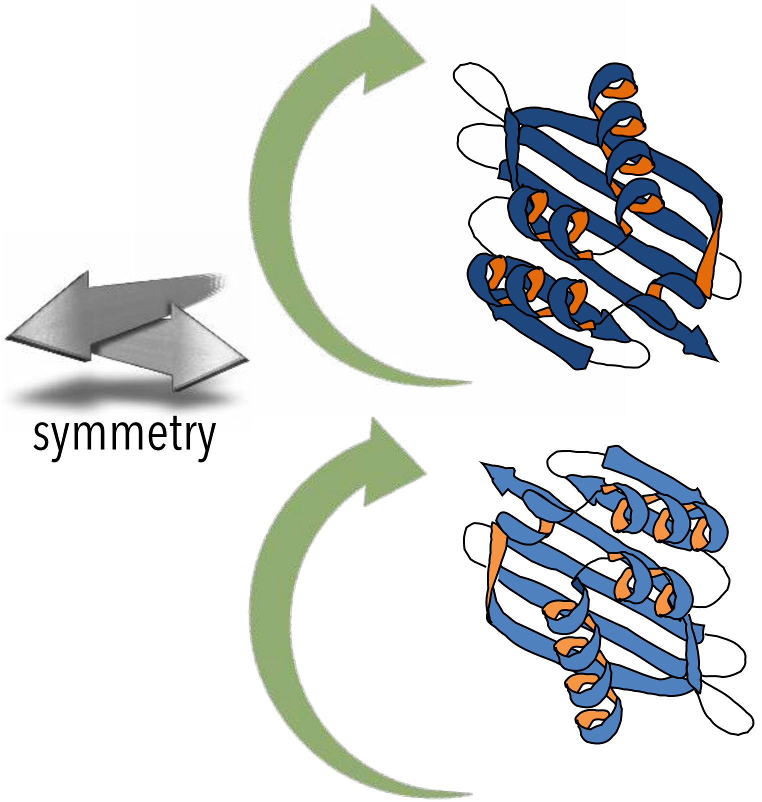
Translation

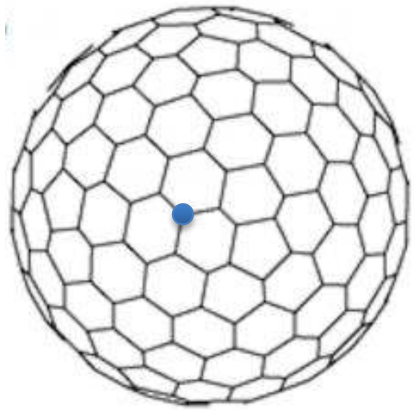
You can ‘solve’ the molecular replacement problem by finding any one of the (symmetry related) copies in the crystal lattice

Unknown
crystal
structure

rotation search (rotation function)

- Conceptually, orient model at in all angles on a grid
- Score each, rank and take the top or best few
- In practise, there are many speed enhancements

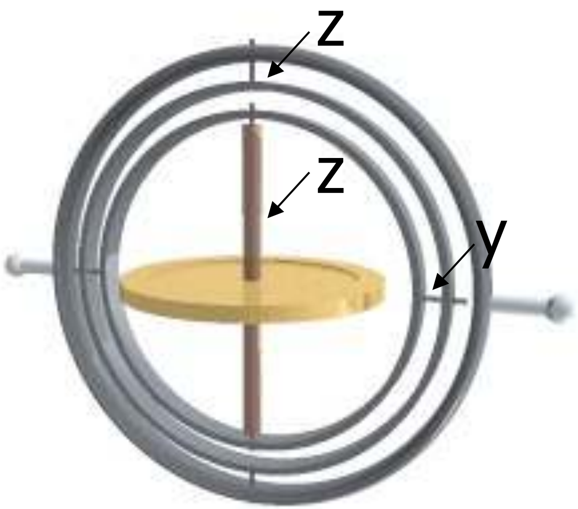




LLG
log-likelihood gain

Euler Angles*

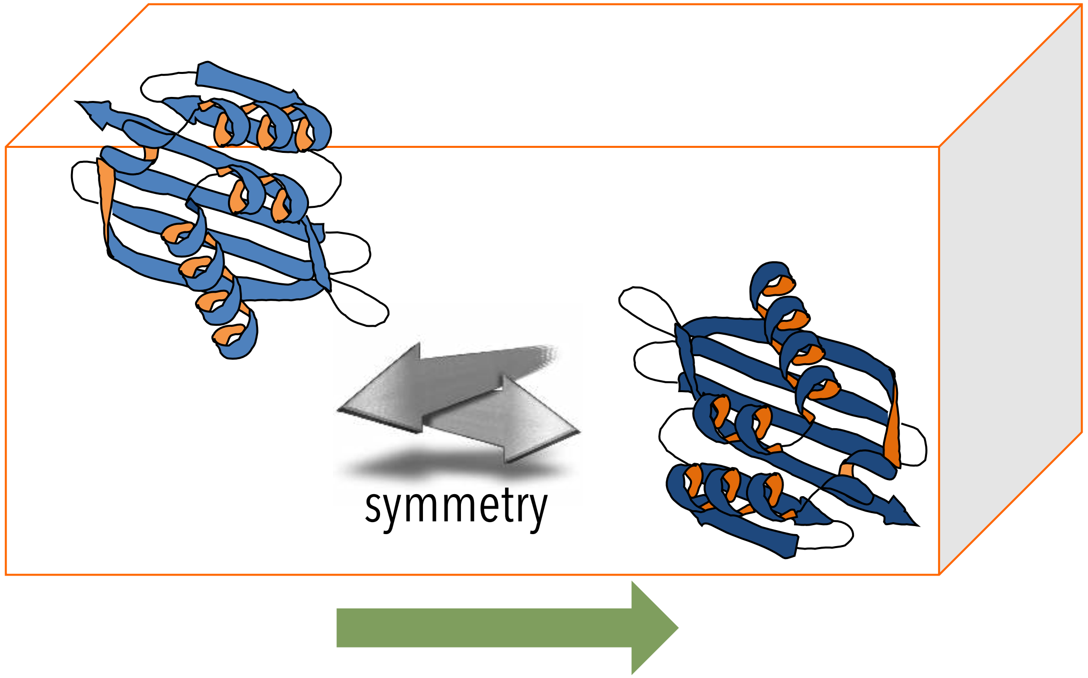
- Phaser rotation angles are reported as Euler angles (**α**, **β**, **γ**)
  - **α** is rotation about the first axis
  - **β** is rotation about the second axis
  - **γ** is rotation about the third axis
- In crystallography, all programs report Euler angles with respect to intrinsic "z,y,z" axes
- You <u>do not need</u> to be able to 'draw' the final rotation axis position by reading Euler angles!
  - Programs will write out the coordinates for you!

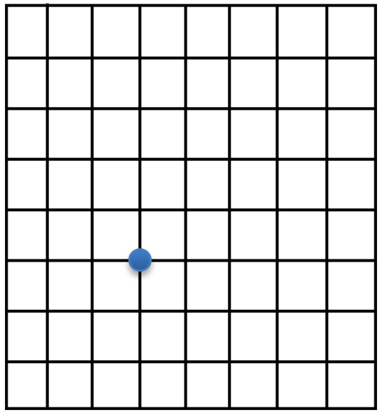


*This slide is for information only

translation search (translation function)

- Conceptually, place model at positions on a grid and score each
- Score each, rank and take the top or best few
- In practise, there are many speed enhancements

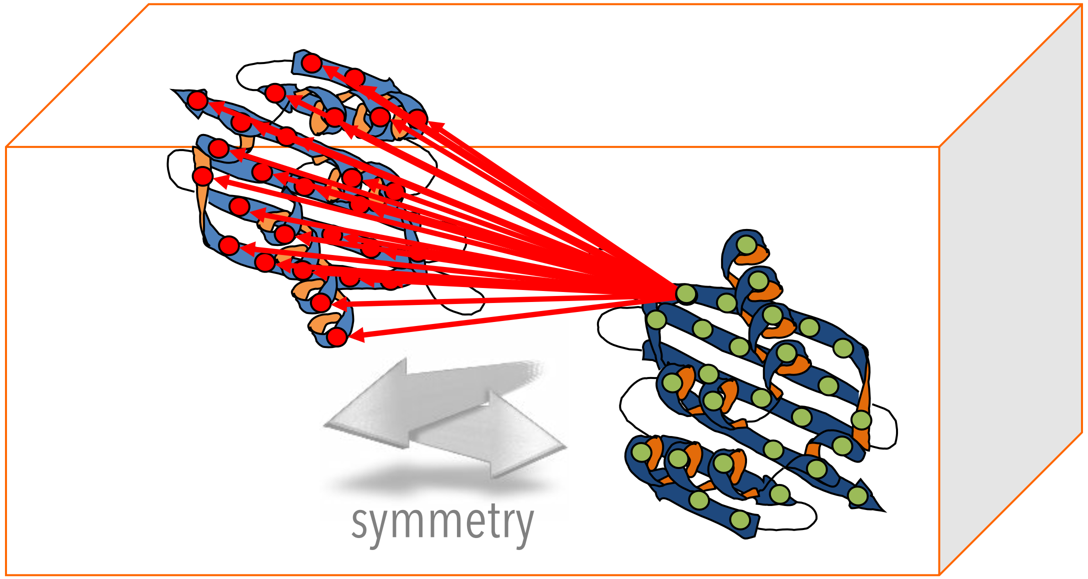




LLG
log-likelihood gain

packing analysis (packing function)

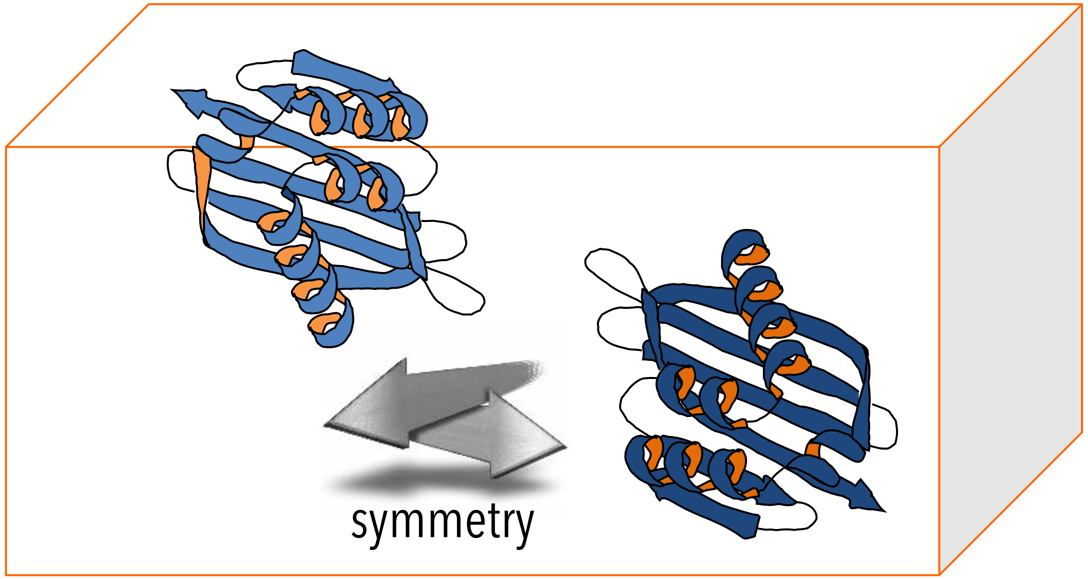
- Cα clash test
- Excludes physically impossible poses, reducing search space

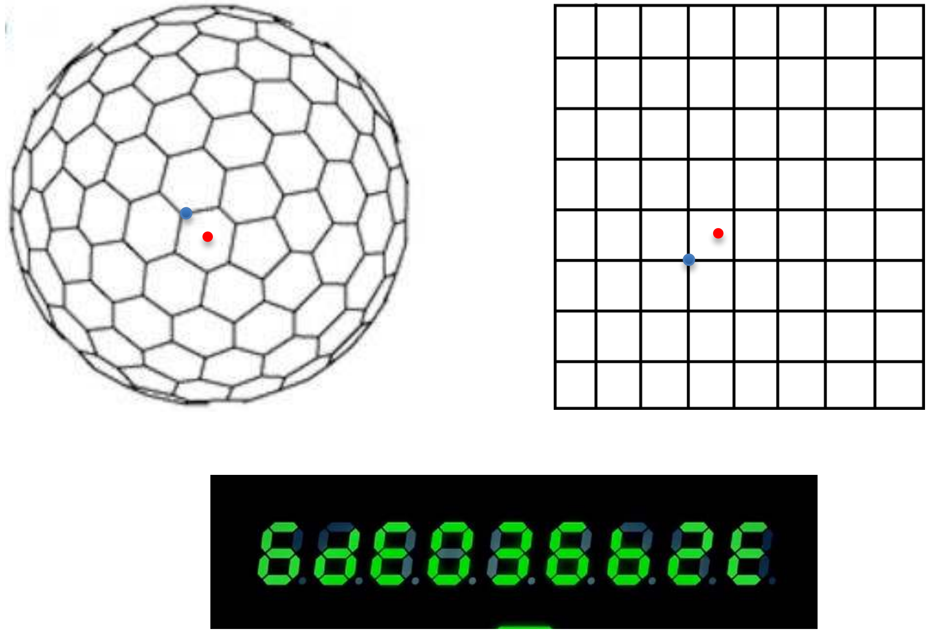


Clash with other contents of unit cell

refinement

- Optimize poses away from grid search locations

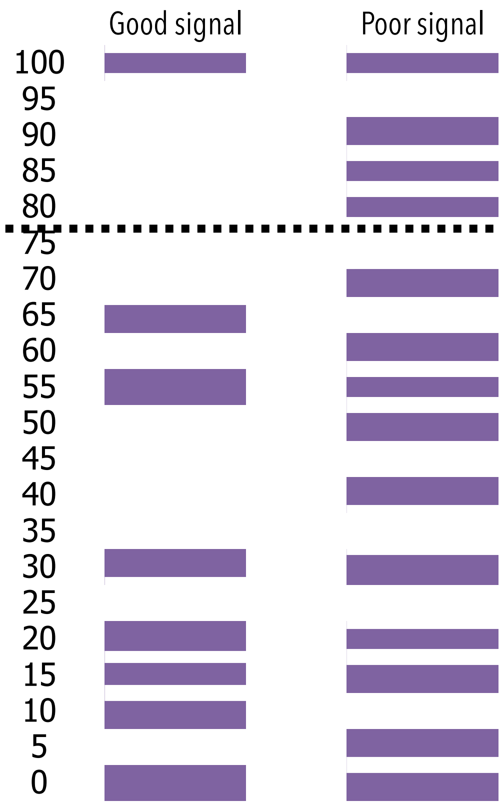




LLG
log-likelihood gain

peak selection in rotation and translation function

- Must chose a selection criteria to carry potential solutions through to the next step
- By default in phaser, solutions over 75% of the difference between the top peak and the mean are selected
  - Good signal, few potential solutions
  - Poor signal, many potential solutions
- The absolute value of the LLG is not used for decision making here*



* spoiler alert: we use percent here but the absolute value is also important!

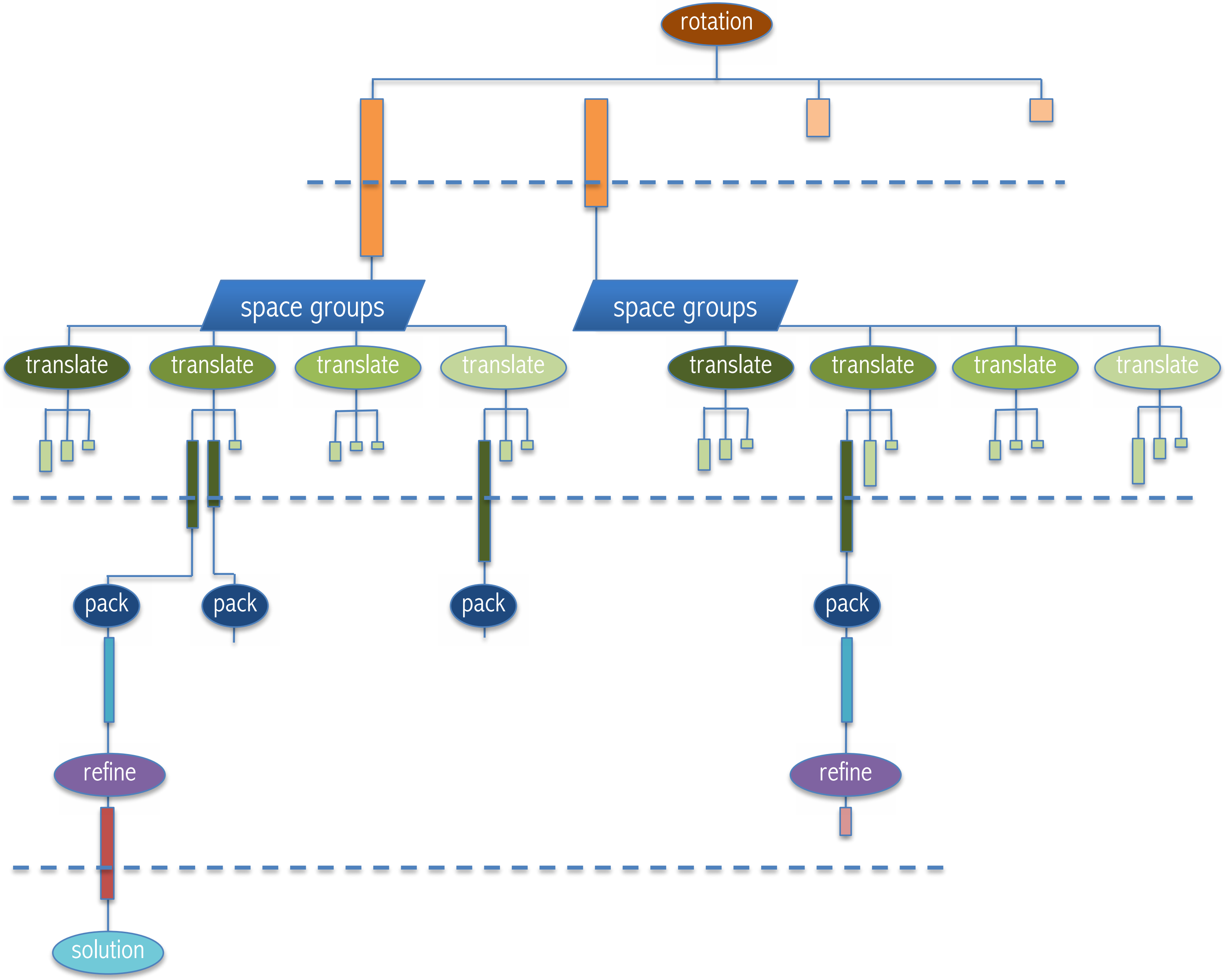
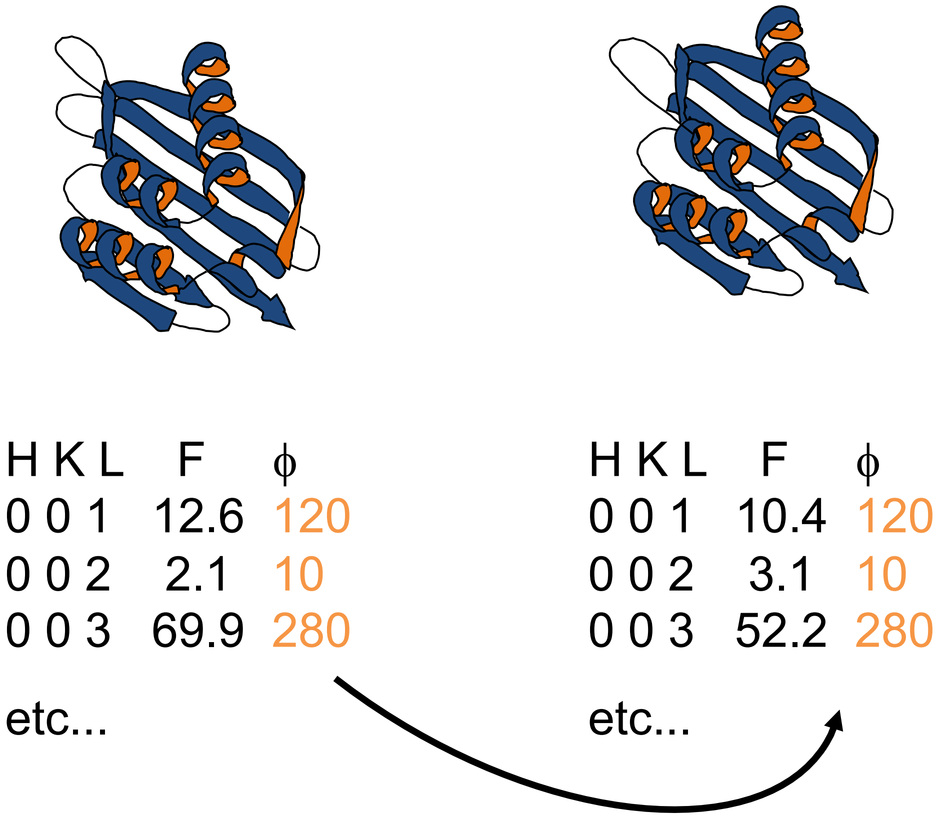


Diagram: Martin Noble

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!



H K L	F	ϕ
0 0 1	12.6	120
0 0 2	2.1	10
0 0 3	69.9	280

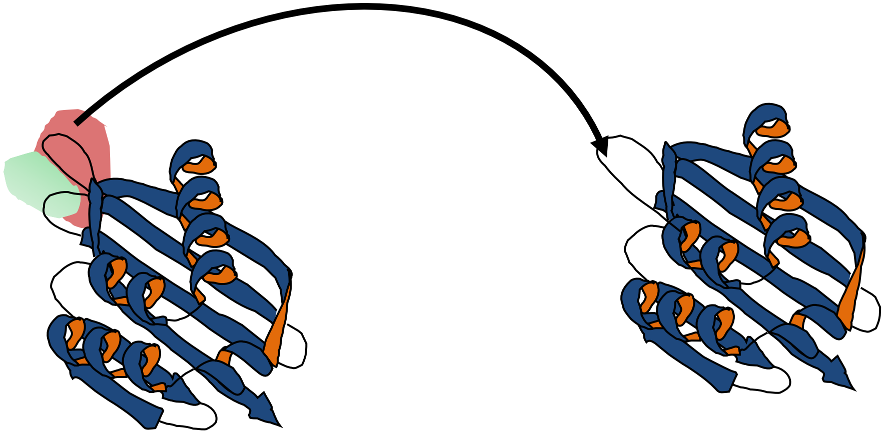
etc...

H K L	F	ϕ
0 0 1	10.4	120
0 0 2	3.1	10
0 0 3	52.2	280

etc...

rebuilding

- After molecular replacement the electron density maps can be inspected to see where the model is wrong
- Rebuilding can be thought of as a type of phase improvement/density modification



H K L	F	ϕ
0 0 1	12.6	120
0 0 2	2.1	10
0 0 3	69.9	280

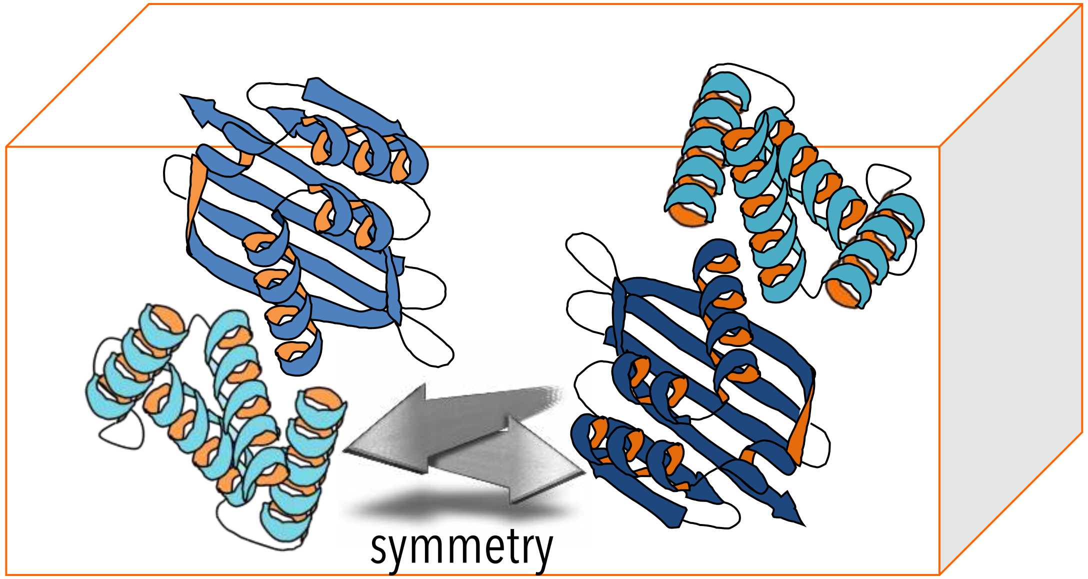
etc...

H K L	F	ϕ
0 0 1	10.4	142
0 0 2	3.1	34
0 0 3	52.2	250

etc...

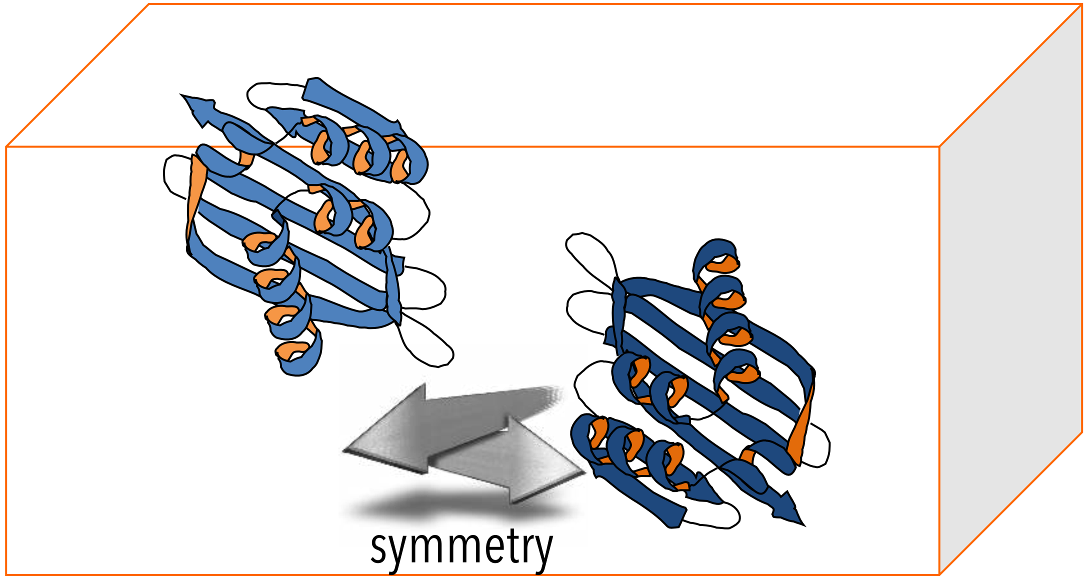
molecular replacement

- Protein complex in the asymmetric unit

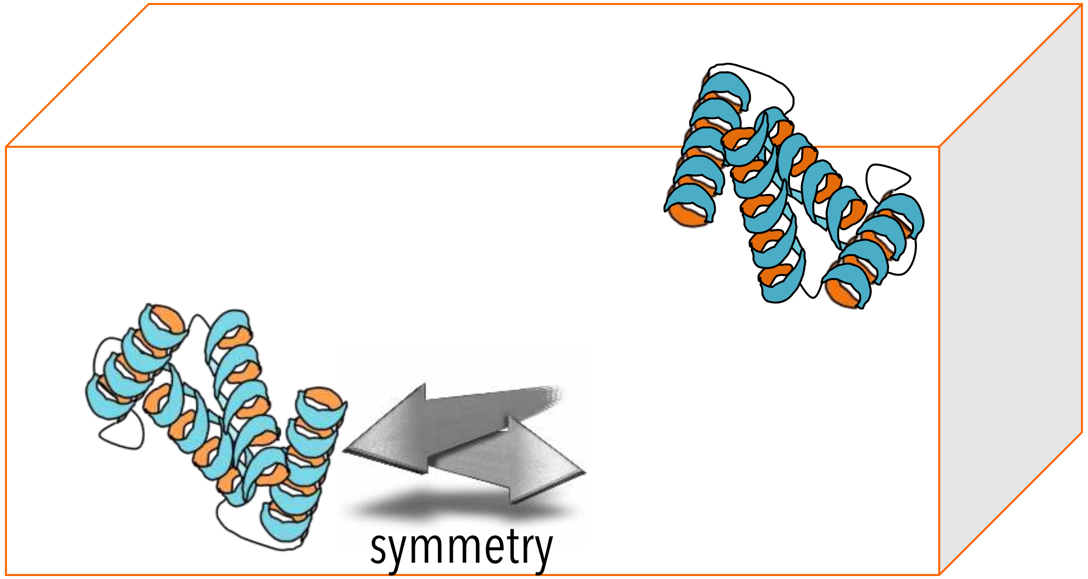


molecular replacement

- Protein complex in the asymmetric unit

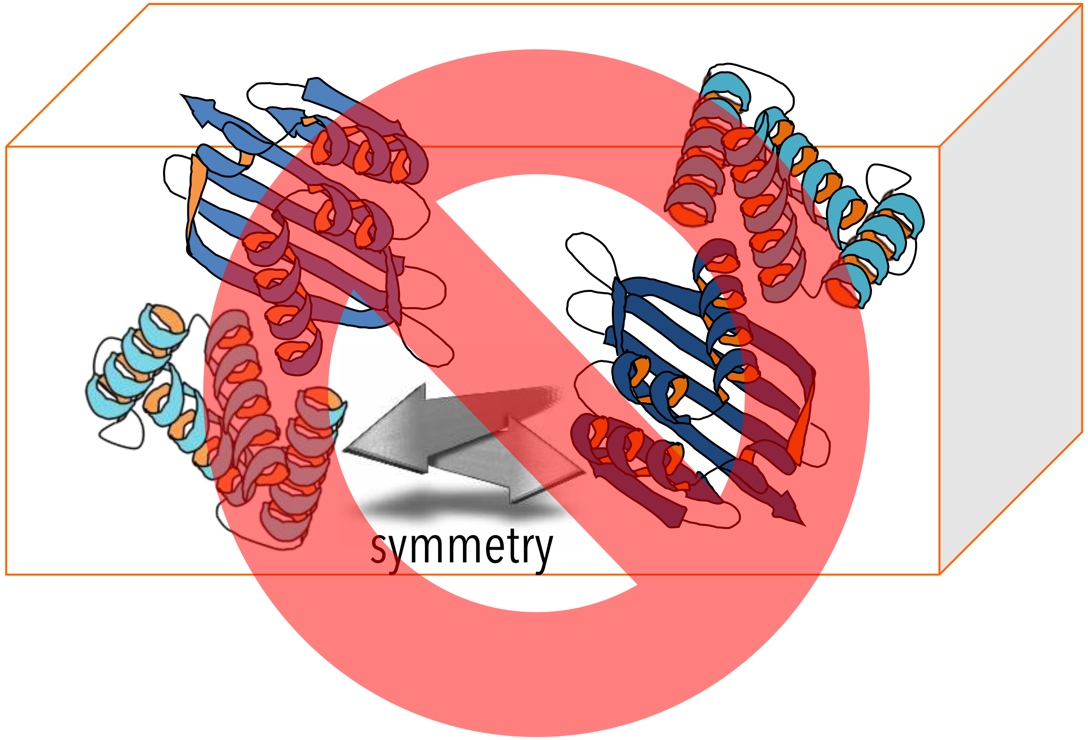


- *Could* search for the two components separately
  - and merge solutions afterwards



molecular replacement

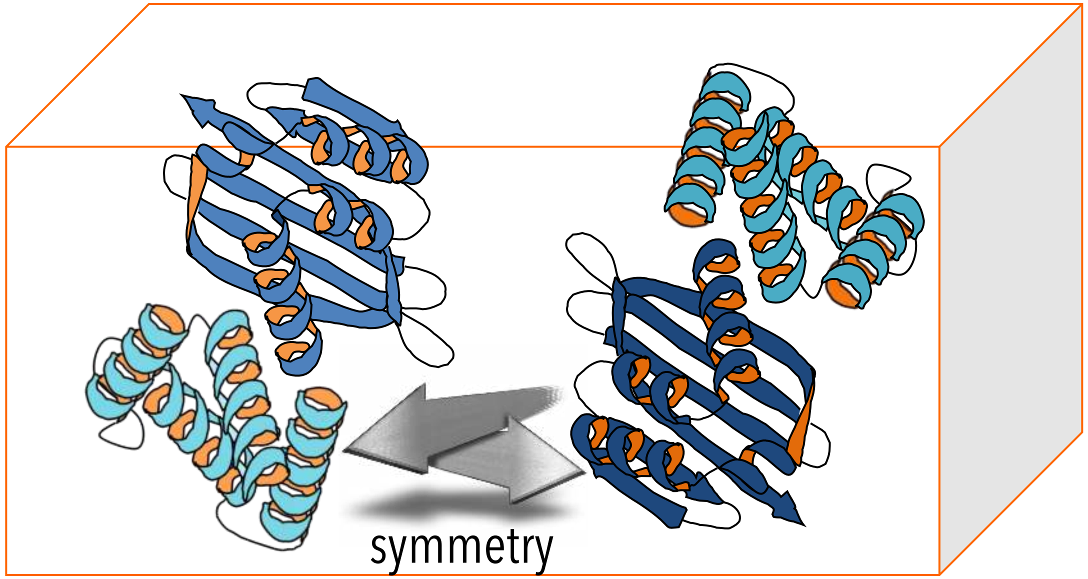
- Protein complex in the asymmetric unit
- Separate solutions could be on different origins*
- Signal may be low for one or more components



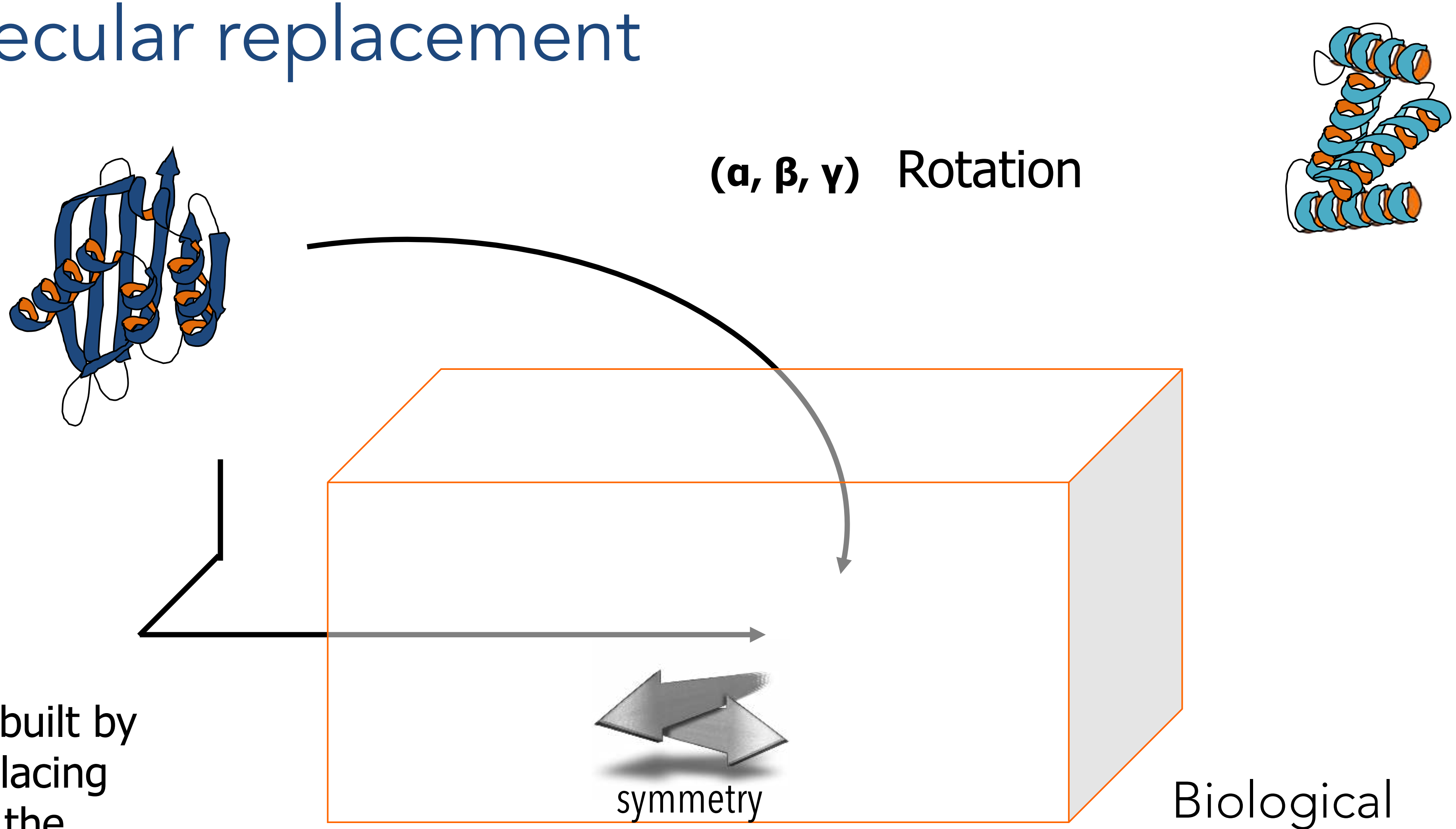
* later in talk!

molecular replacement

- Protein complex in the asymmetric unit
- Complex built by addition



molecular replacement



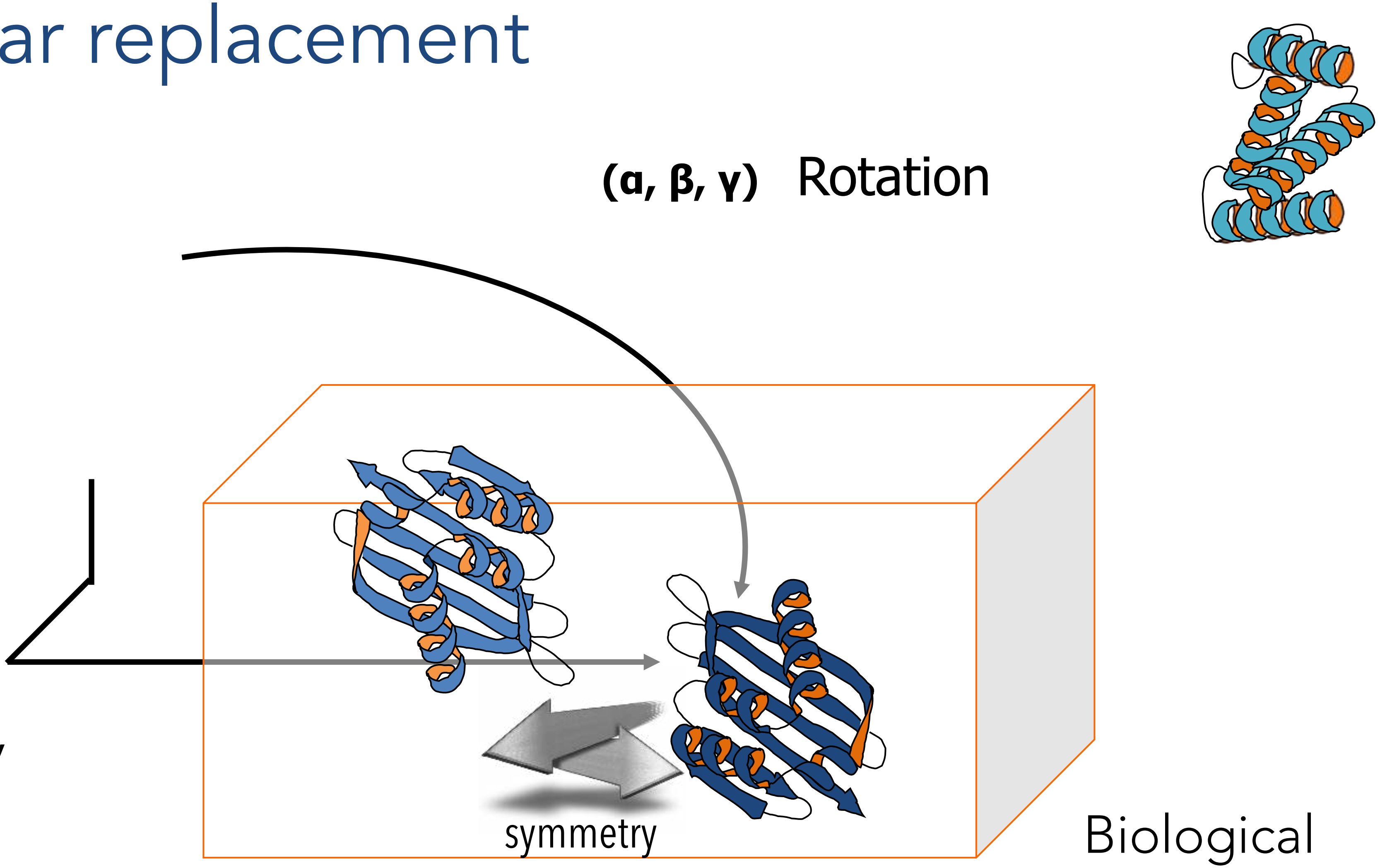
(α, β, γ) Rotation

symmetry

Structure built by addition placing second in the presence of the first

Biological oligomer placement

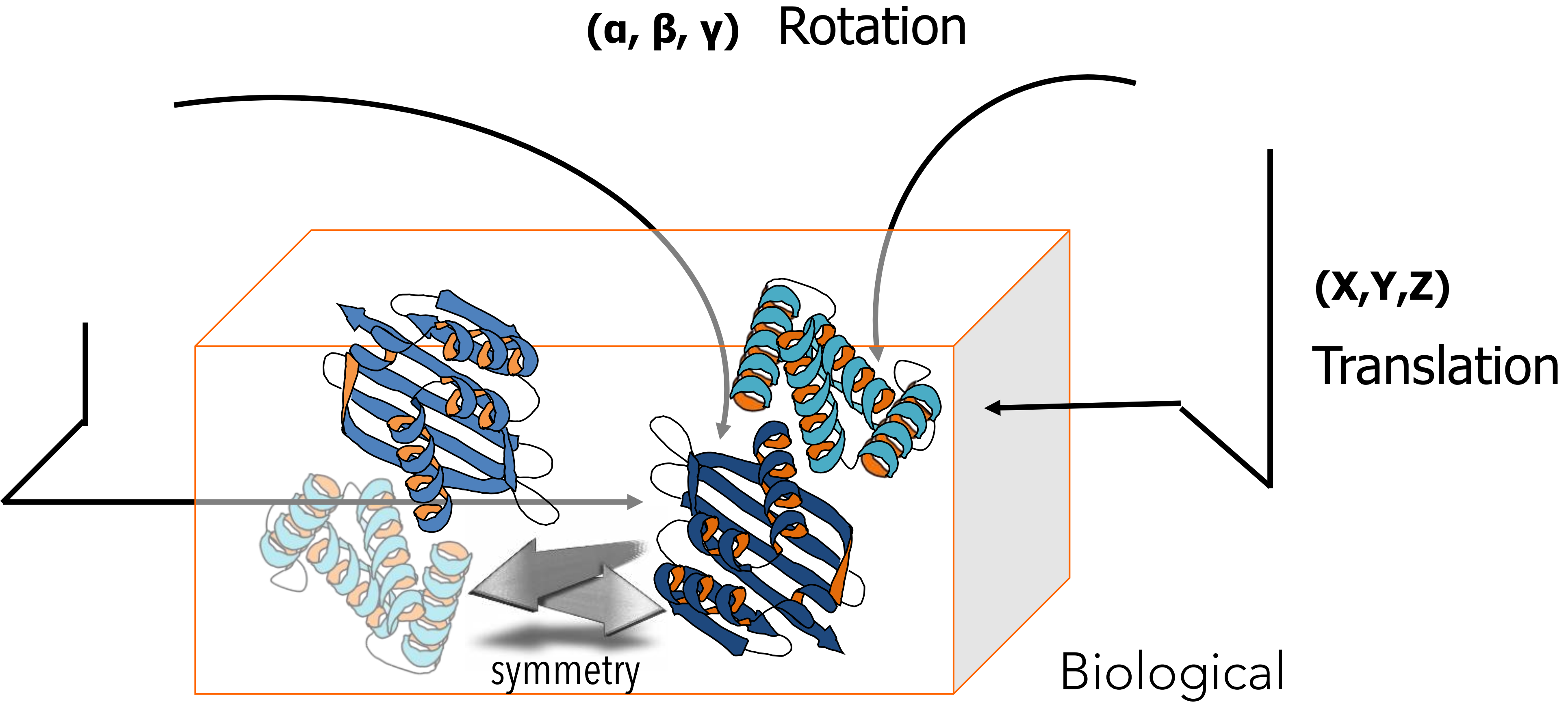
molecular replacement



Structure built by addition placing second in the presence of the first

Biological
oligomer
placement

molecular replacement



(α, β, γ) Rotation

(X,Y,Z)

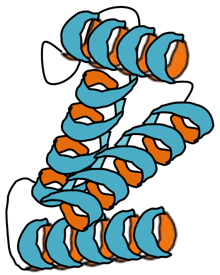
Translation

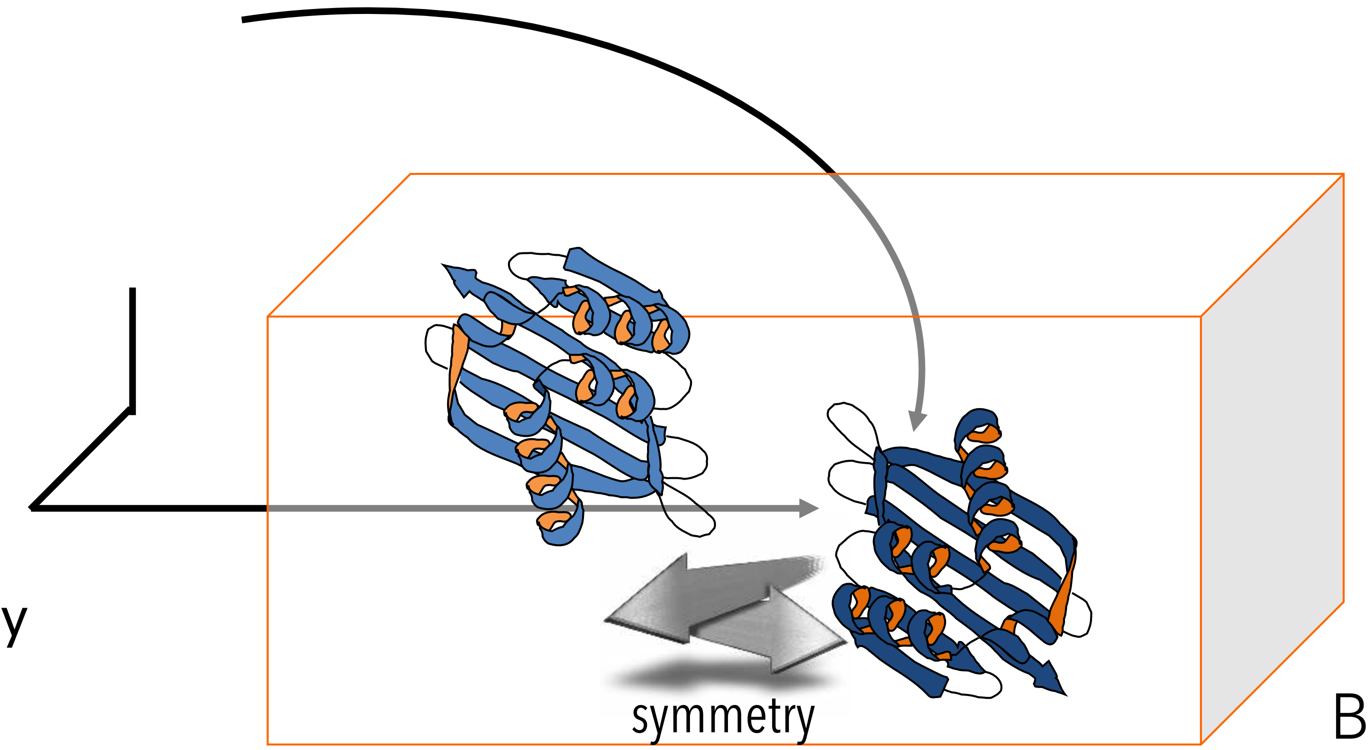
Structure built by addition placing second in the presence of the first

Biological oligomer placement

molecular replacement

(α, β, γ) Rotation

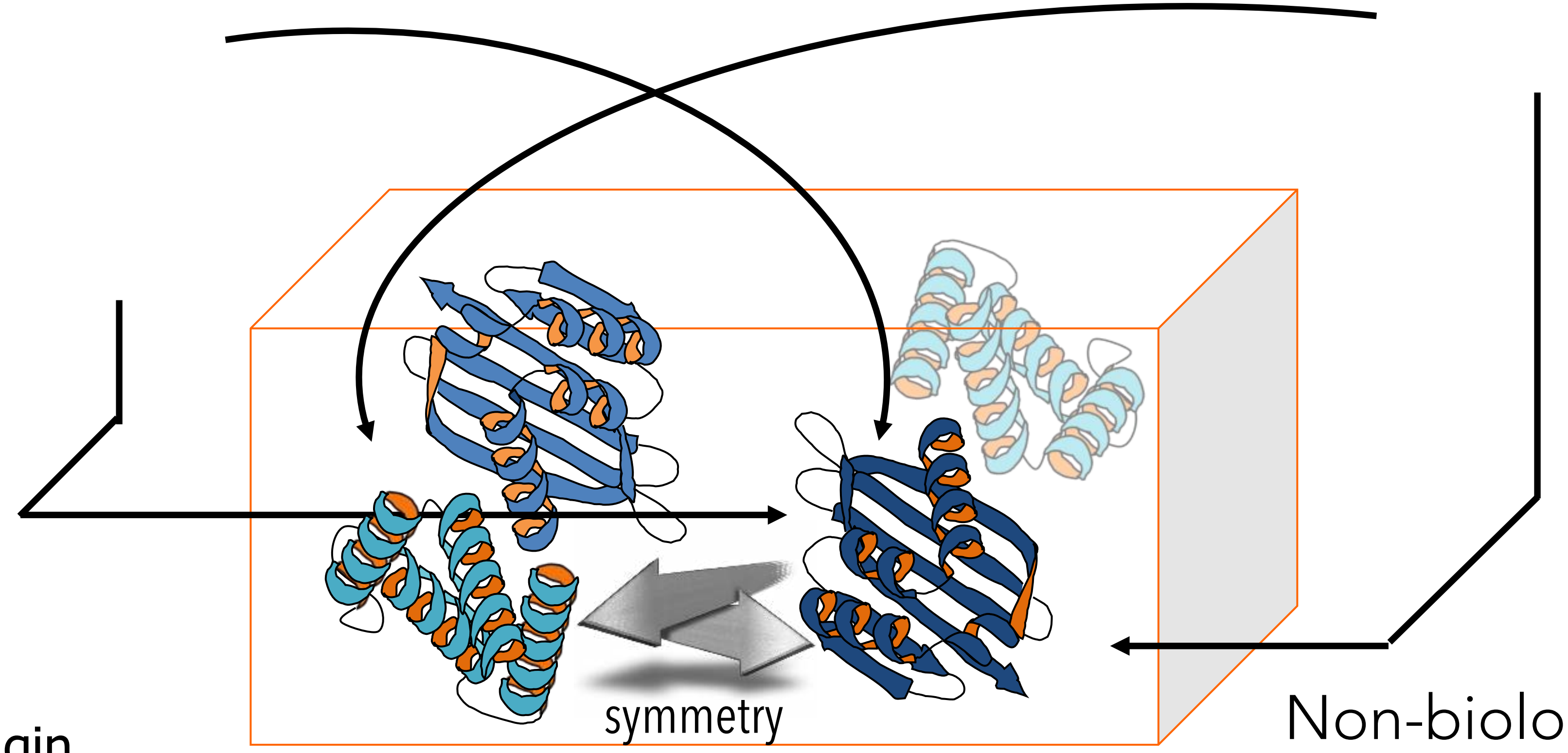




Structure built by addition placing second in the presence of the first

Biological oligomer placement

molecular replacement



(α, β, γ) Rotation

(X,Y,Z)

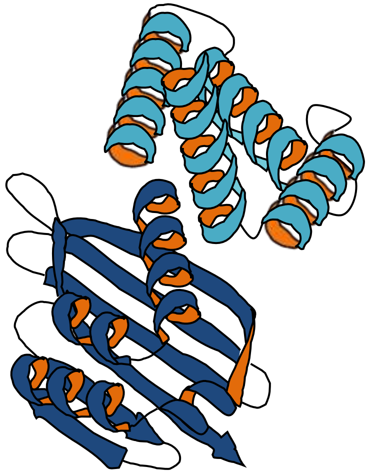
Translation

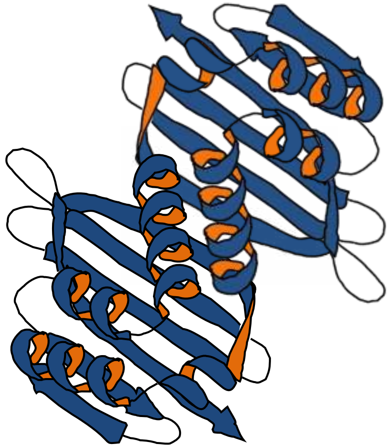
Phaser will place first component closest to the origin and second closest to the centre of mass of the first

Non-biological oligomer placement

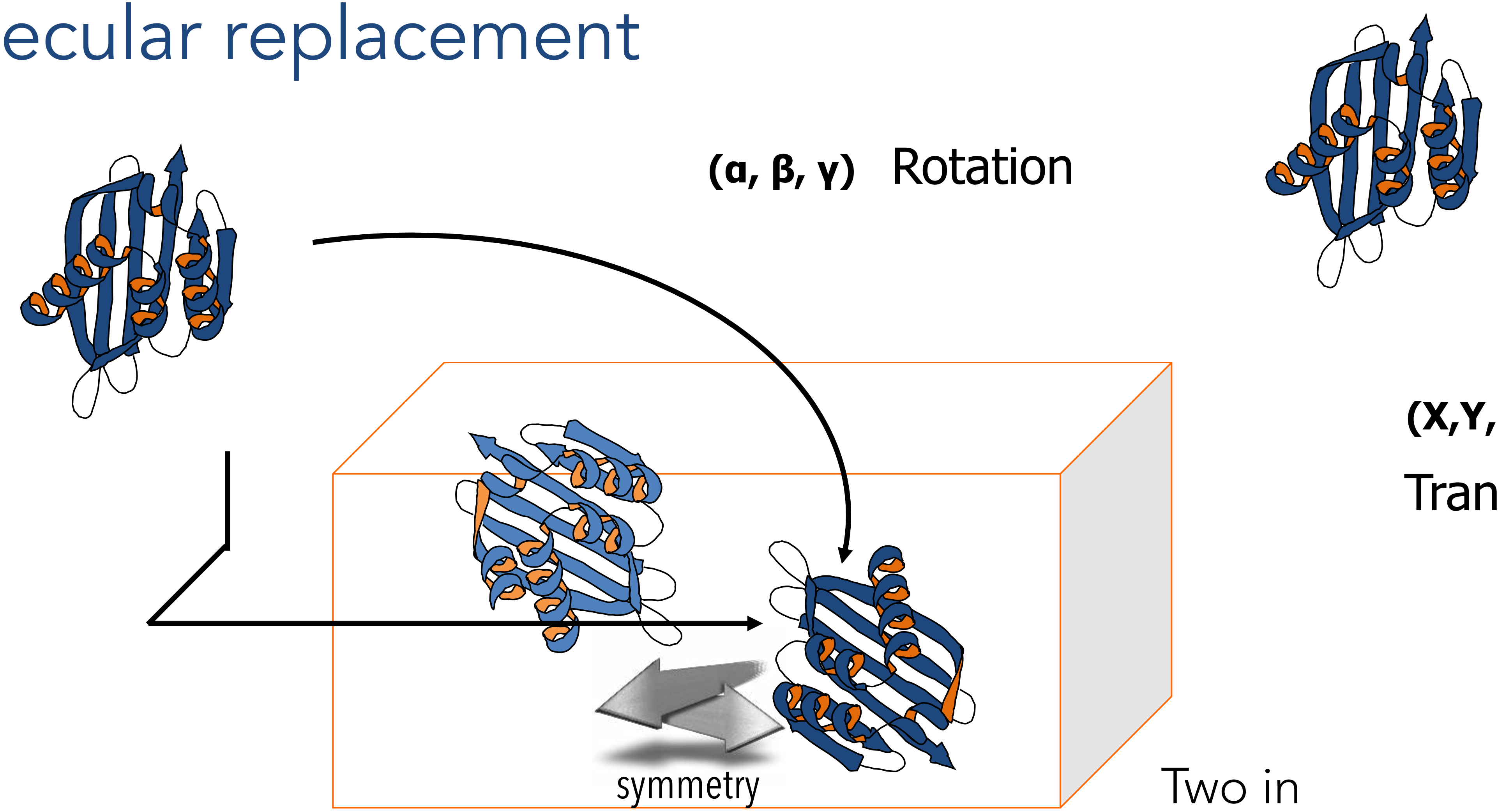
asymmetric unit contents

- non-crystallographic symmetry arises when there is more than one copy of a component in the asymmetric unit (asu)
  - The ‘symmetry’ can be point-group symmetry or equally can be non-point-group symmetry
  - non-crystallographic = does not allow ‘crystal’ to map onto itself
- Multiple copies in the asu and complexes in the asu are equivalent problems in molecular replacement
  - Equivalent problem: Find position of rigid components





molecular replacement



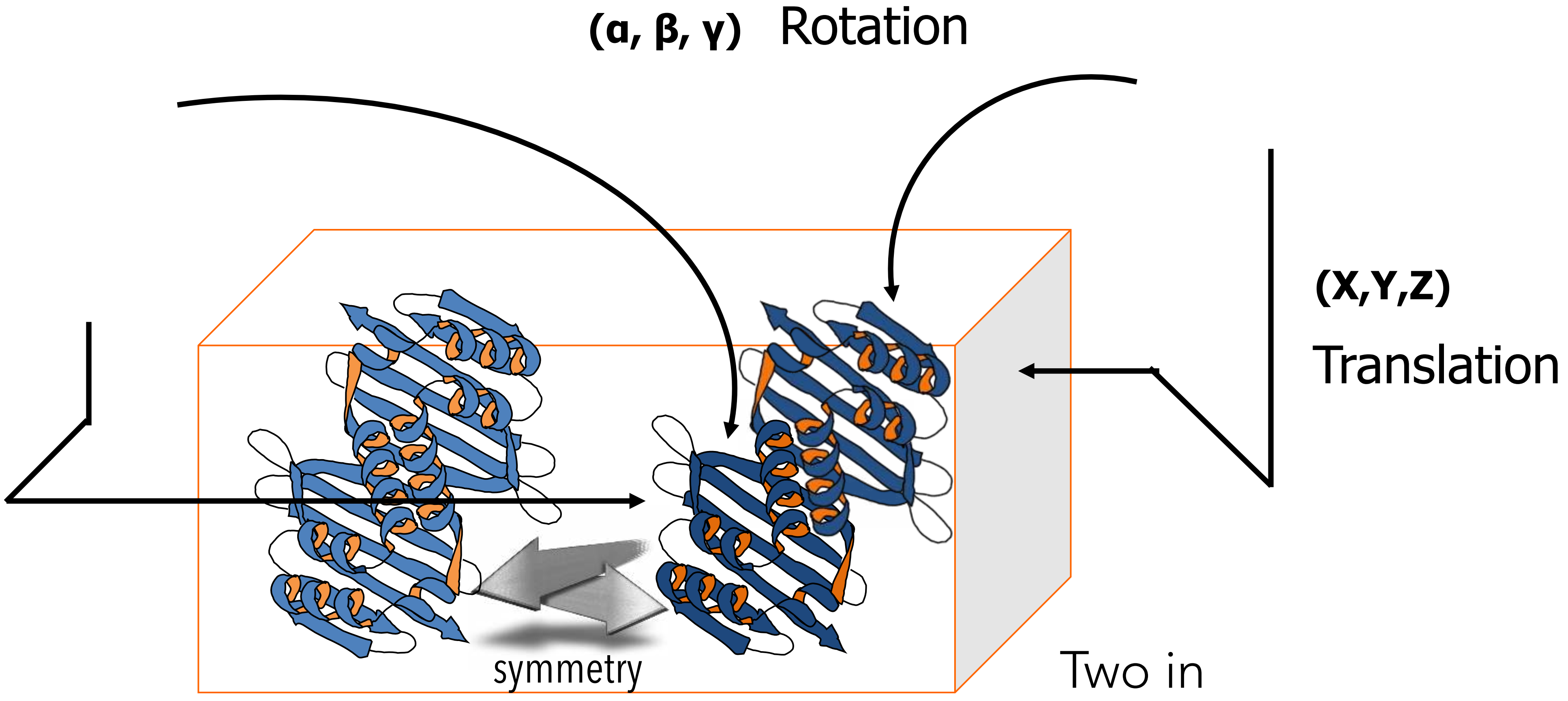
(α, β, γ) Rotation

(X,Y,Z)

Translation

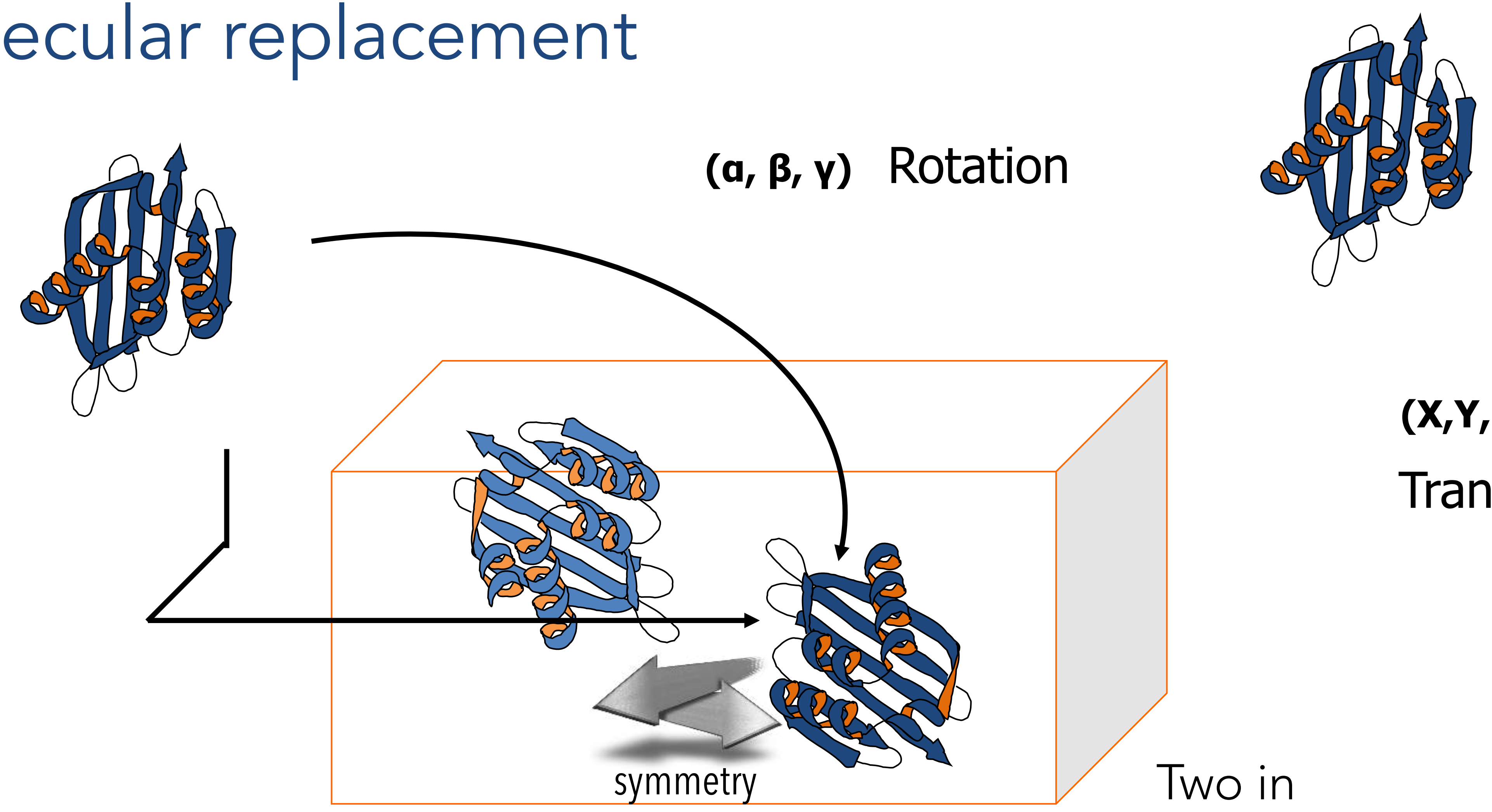
Two in
asymmetric unit

molecular replacement



Two in
asymmetric unit

molecular replacement



(α, β, γ) Rotation

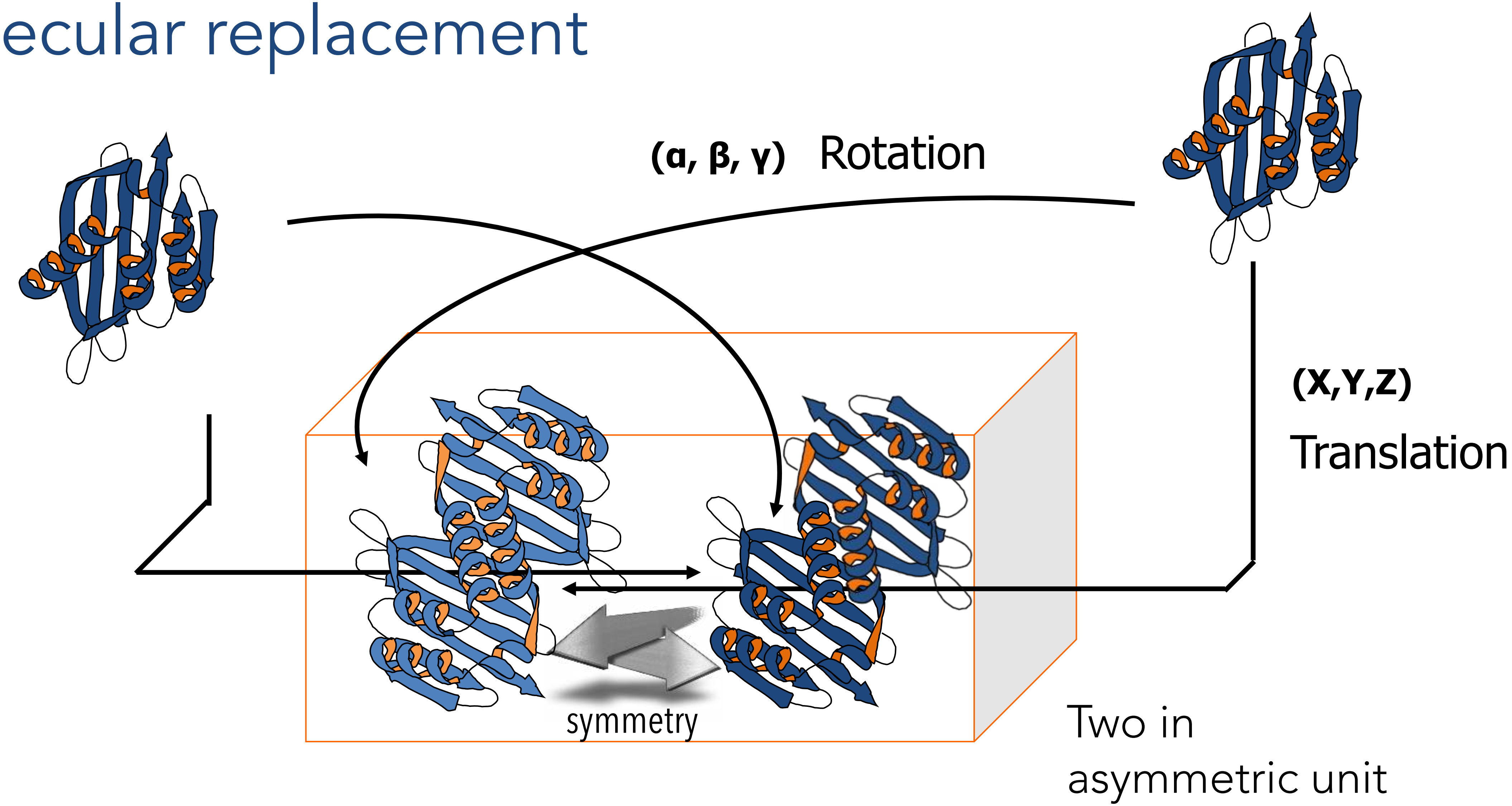
(X,Y,Z)

Translation

symmetry

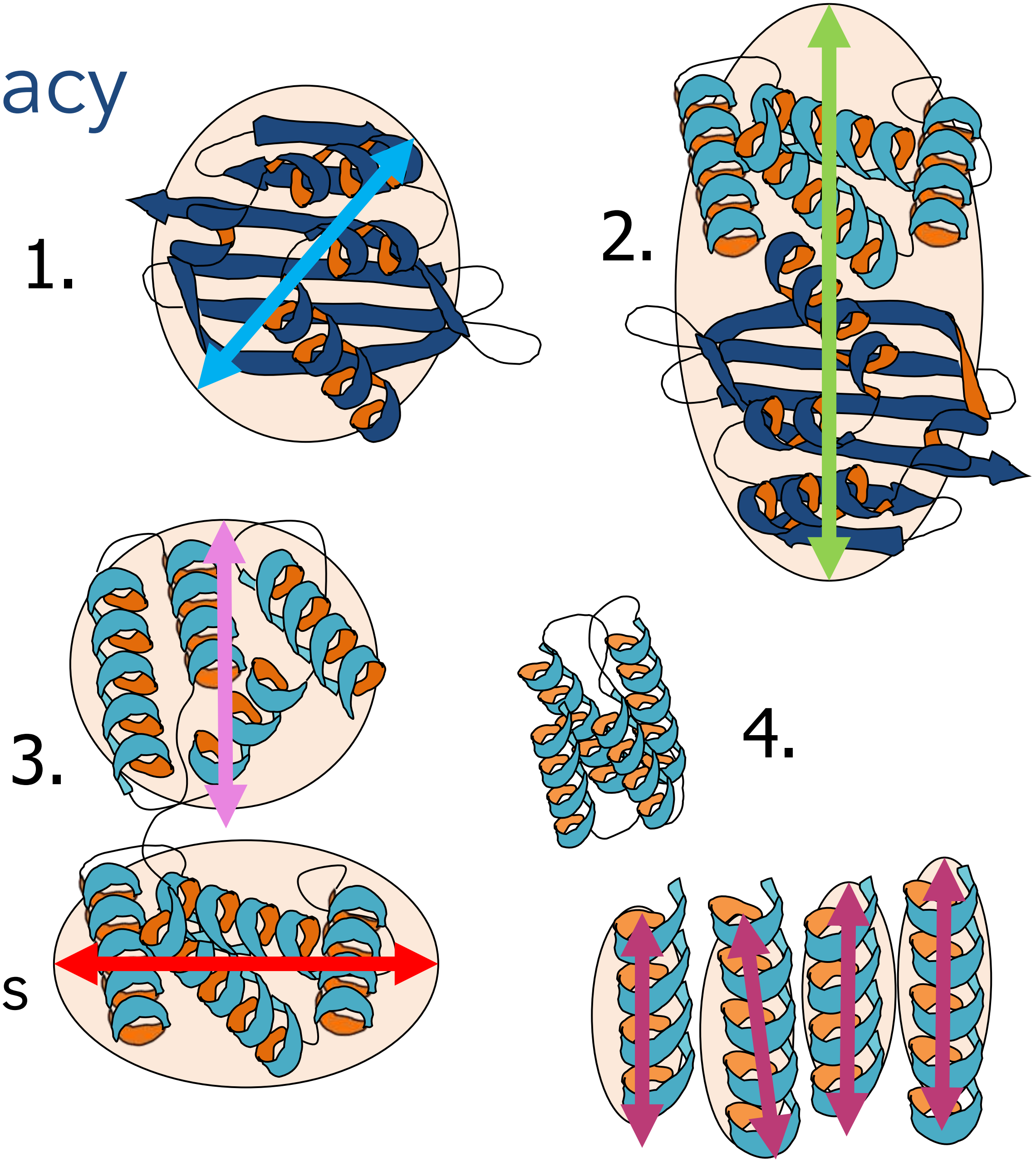
Two in
asymmetric unit

molecular replacement

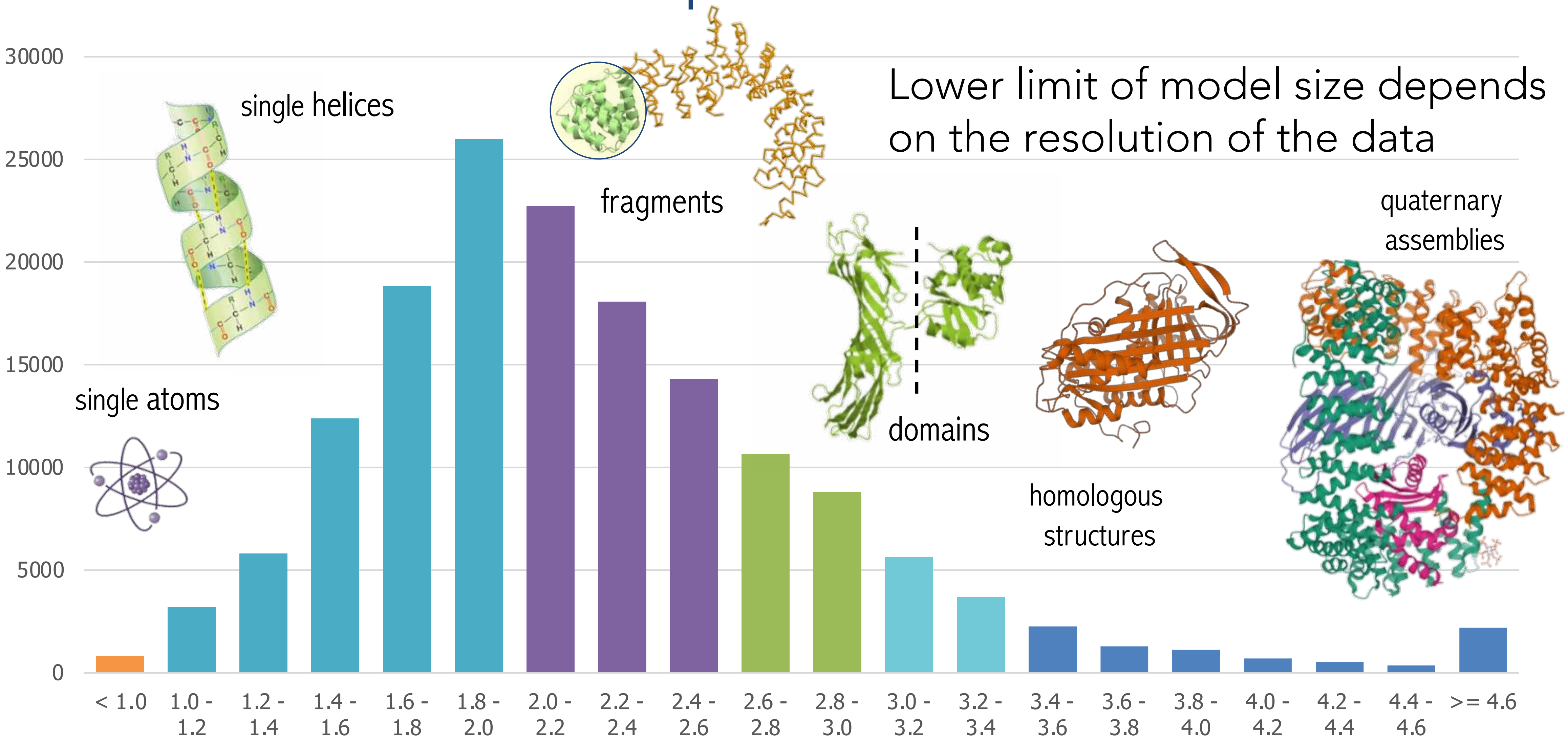


"model" – long range accuracy

- A (good) model is a (large) collection of atoms whose displacement from each other is the same in model and target
    1. Whole protein
    2. Quaternary assembly
    3. Protein domain(s)
    4. Fragments
- You don't know (for certain) in advance whether the model meets this criteria… working 'blind'



models for molecular replacement

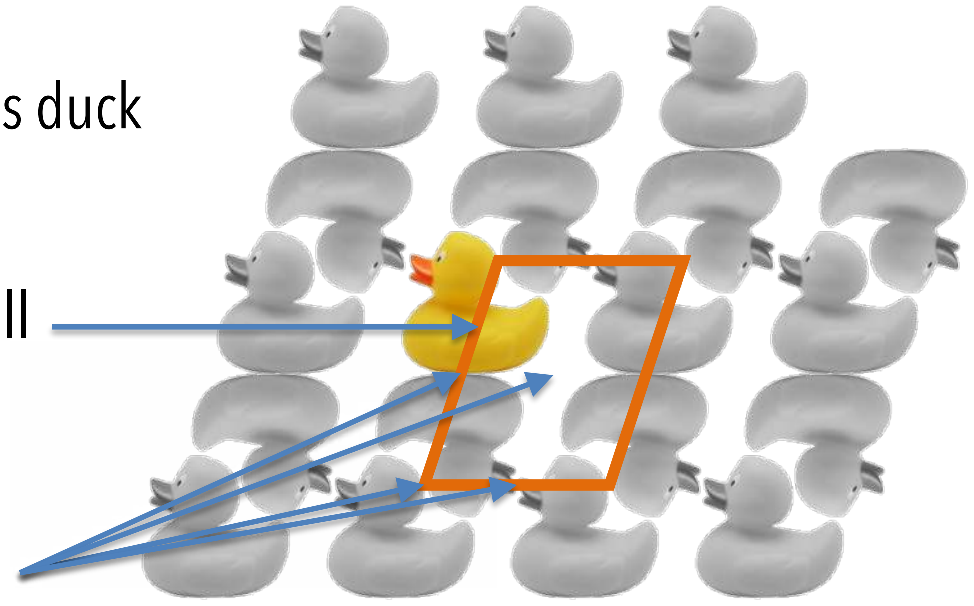


Lower limit of model size depends on the resolution of the data

searching for multiple copies

asymmetric unit

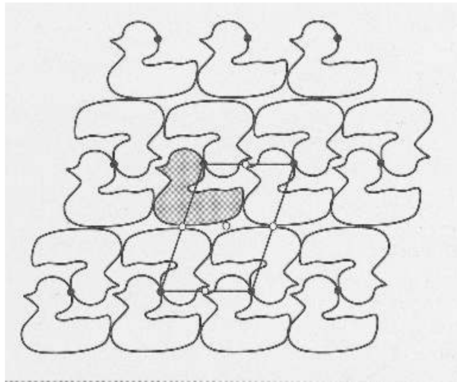
- The asymmetric unit is the smallest unit of structure that can generate the whole crystal after application of the crystal symmetry



Asymmetric unit is duck

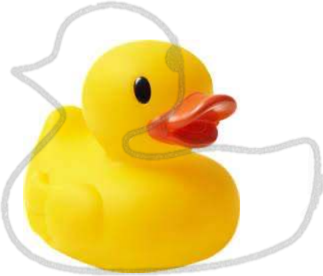
unit cell

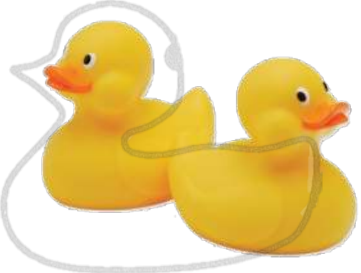
two-fold symmetry axis
(out of the plane)



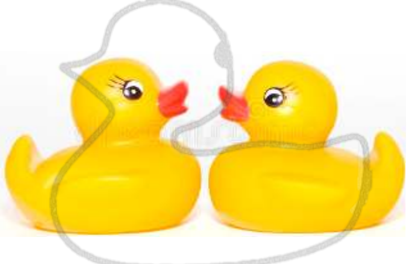
asymmetric unit contents

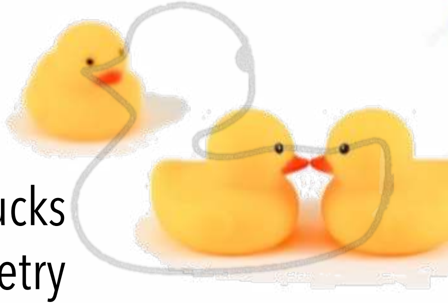
- duplication: non-crystallographic symmetry
- does not propagate through crystal

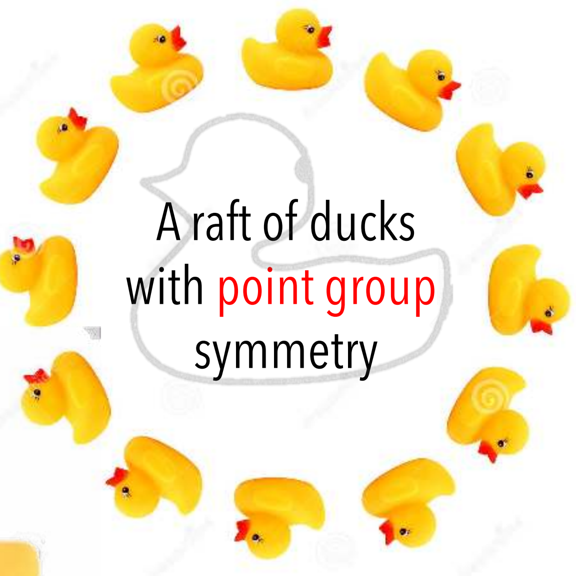

A duck

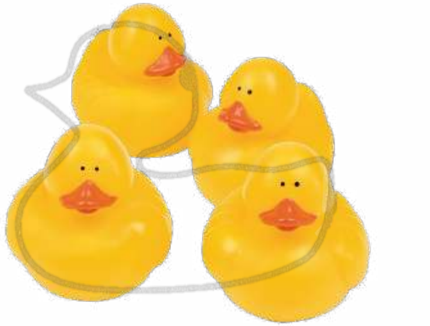

A brace of ducks
No point group symmetry

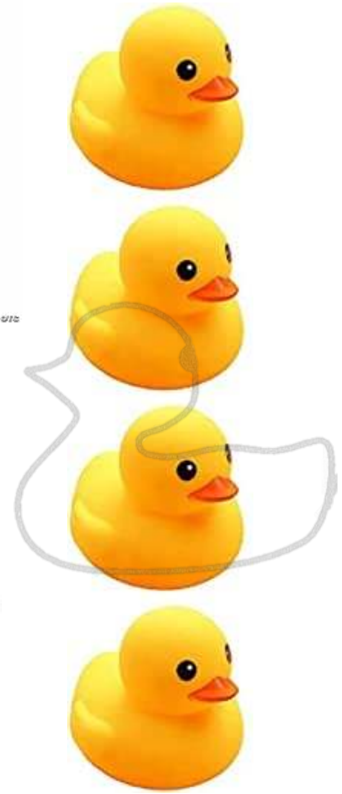

A raft of ducks
non-crystallographic
symmetry


A brace of ducks
with point group
symmetry


A leash of ducks
with partial point group symmetry

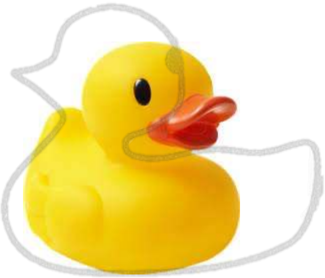

A raft of ducks
with point group
symmetry


Four ducks
without translational
symmetry


Four ducks
with translational
symmetry

asymmetric unit contents

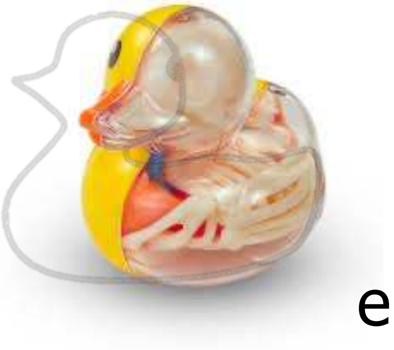
- no duplication: no non-crystallographic symmetry



A duck

e.g. complex

e.g. ribosome



e.g. (secondary) structure elements
pseudo- non-crystallographic symmetry?

asymmetric unit and biological relevance

- It can be difficult to distinguish between biologically relevant and artifactual associations between molecules
- Non-biological ("crystallographic") associations should not be called 'dimers', 'trimers' etc.

biologically relevant

crystal artefact

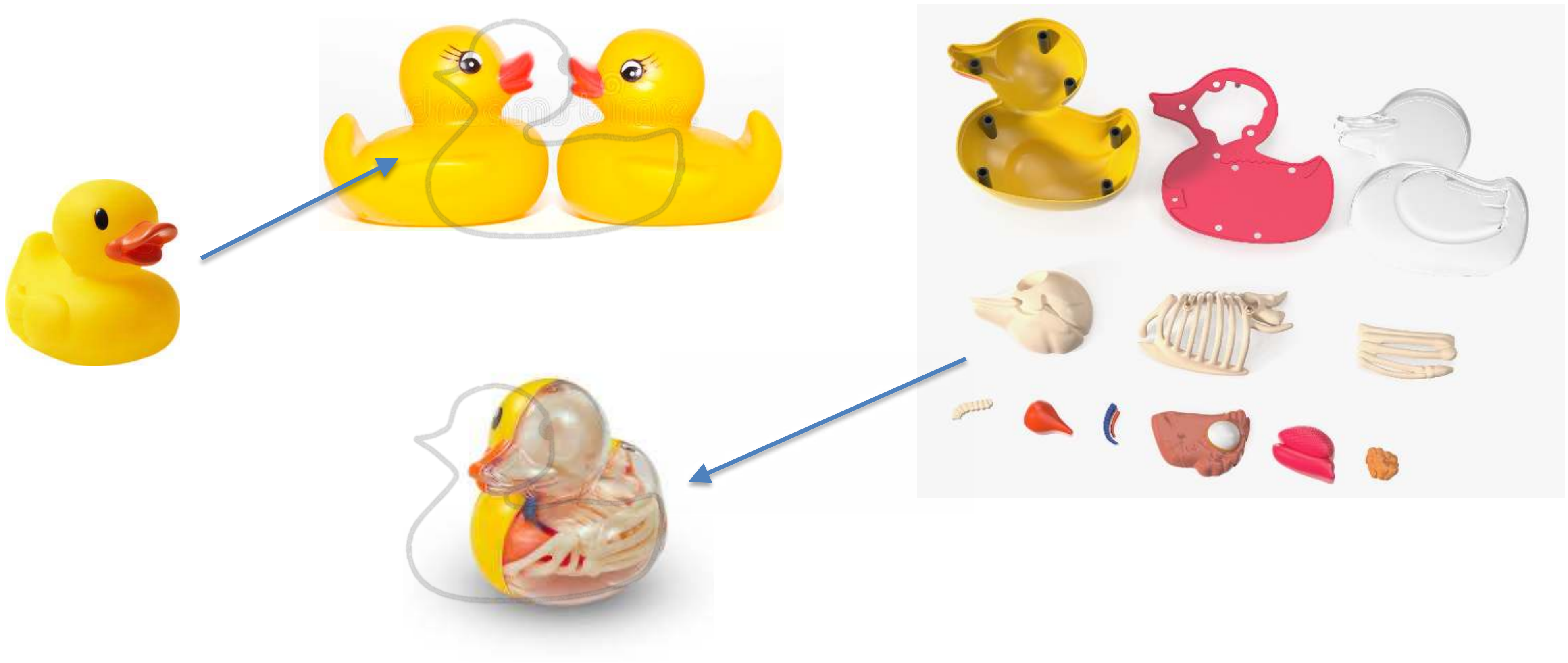
homo-dimer

hetero-dimer

Buried surface area can indicate biologically relevant associations

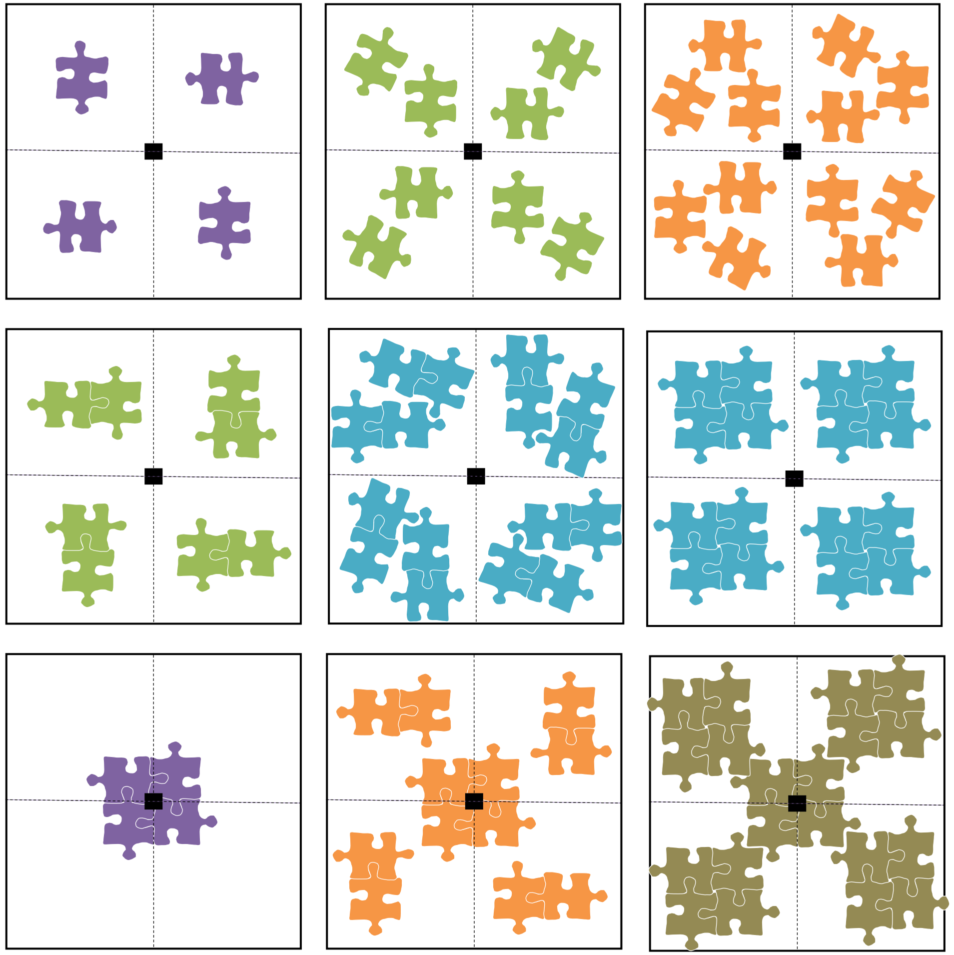
asymmetric unit contents

- in molecular replacement, we build up the asymmetric unit from the smallest units of rigid components

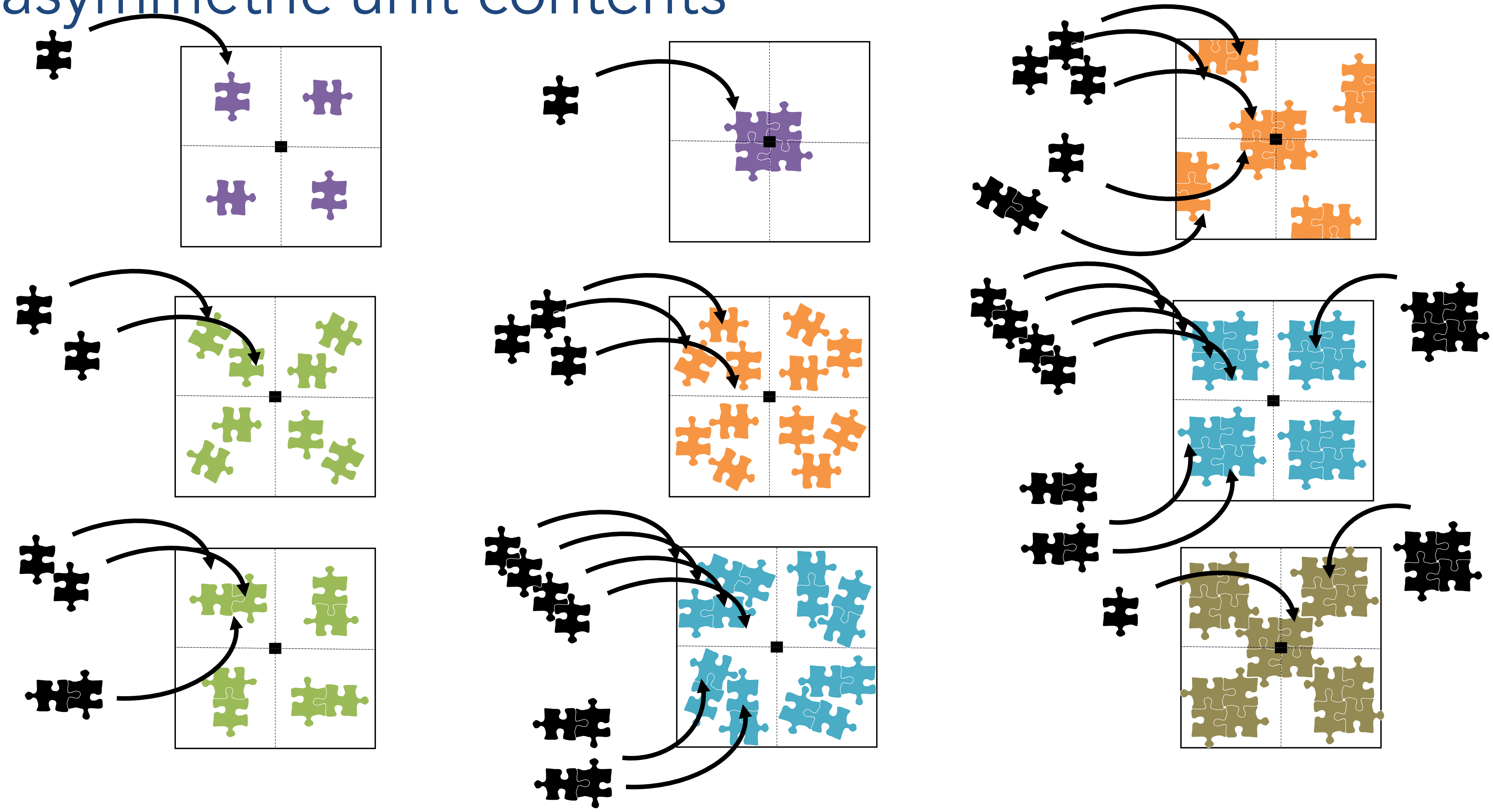


asymmetric unit contents

- Many crystals have more than one copy of your biological unit in the asymmetric unit
- Other crystals have only part of the biological unit in the asymmetric unit, with the full structure generated by the crystal symmetry

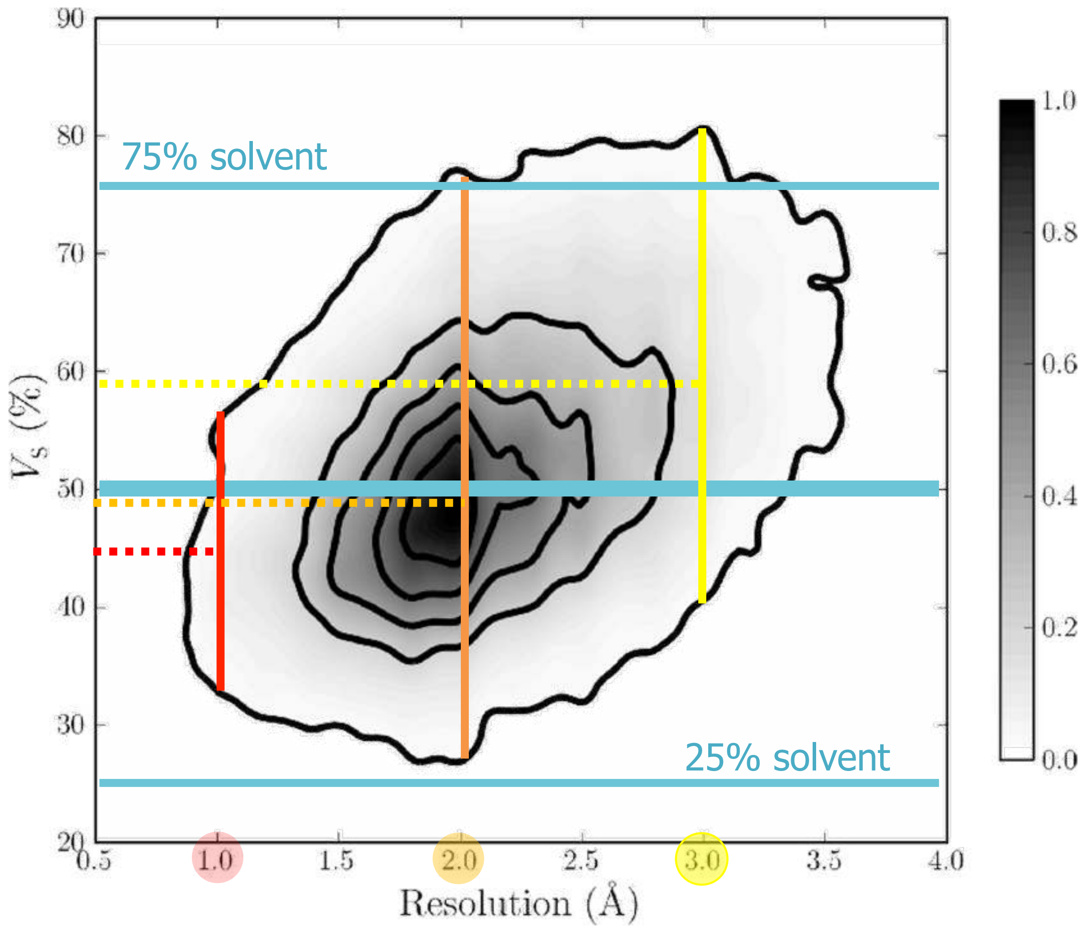


asymmetric unit contents



Matthew’s coefficient

- First calculated by Brian Matthews in 1968
- Most crystals are 50% protein by volume
  - Between 25% and 75% protein
  - Slightly resolution dependent
- Can be used to estimate the contents of the asymmetric unit
  - *c.f. Self Rotation Function*
  - *c.f. TNCS Order*

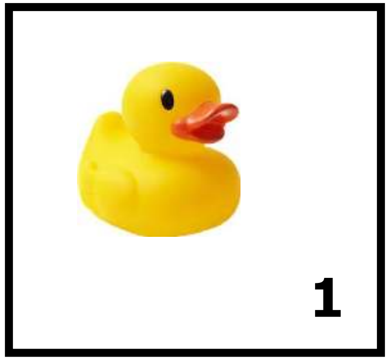


Acta Cryst. (2015). D71, 1023-1038

components of asymmetric unit

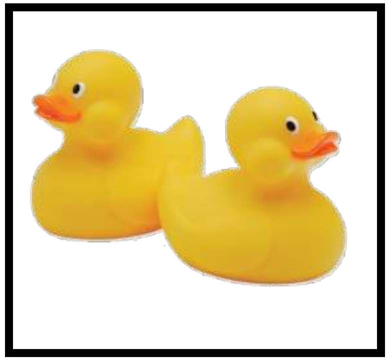
- About 50% solvent

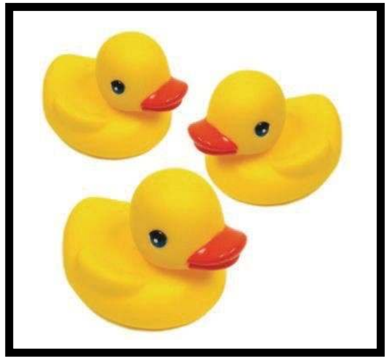
With low numbers of possible copies, options are low



1 can only be 1

For 1, 2 or 3,
1 is unlikely

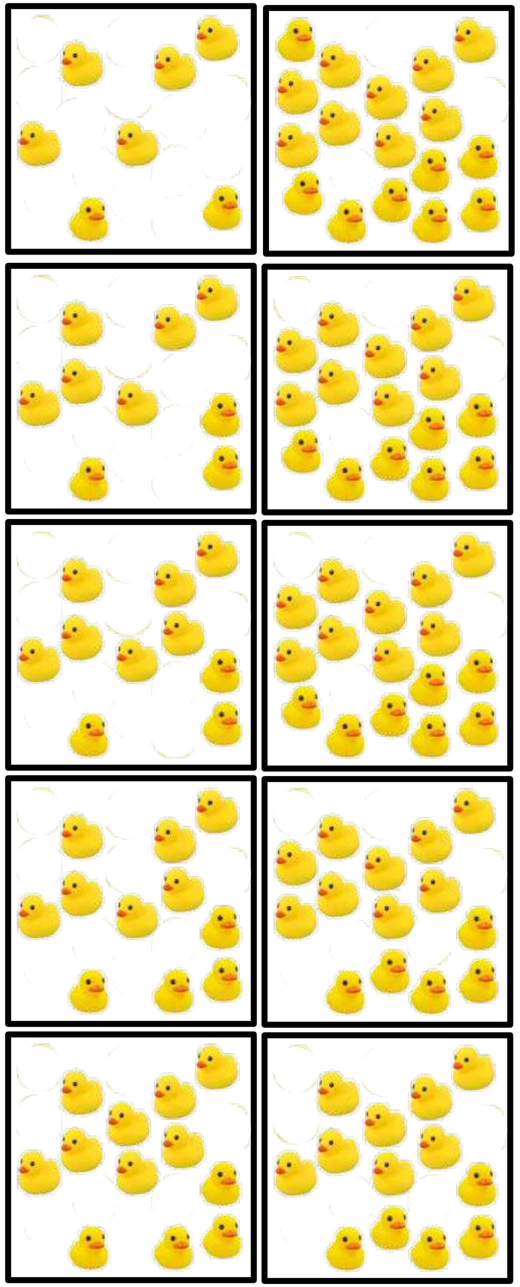




1
2
3

1	7	13	19
2	8	14	20
3	9	15	21
4	10	16	22
5	11	17	23
6	12	18	24

With high numbers of possible copies, options are much greater



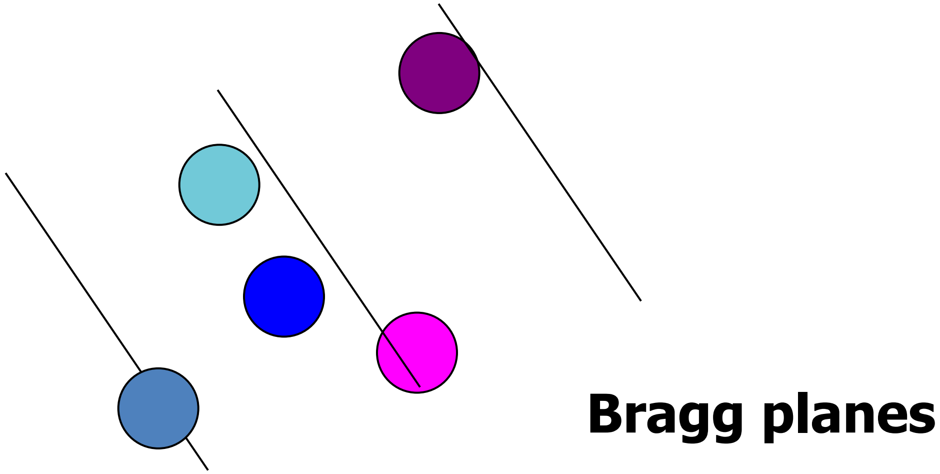
Liking Likelihood

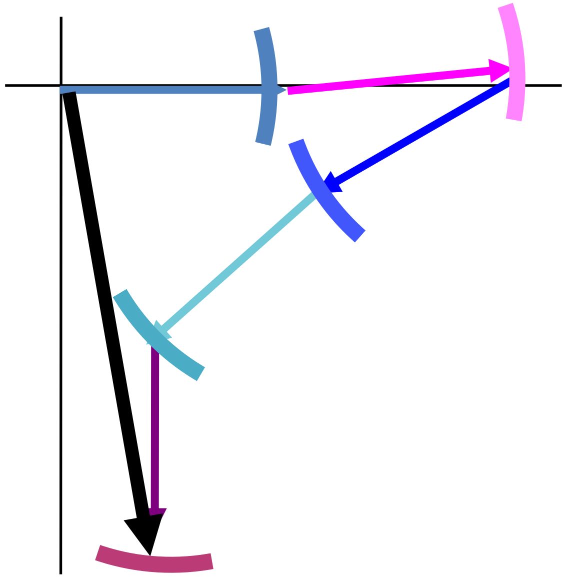
Phaser – likelihood theory

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

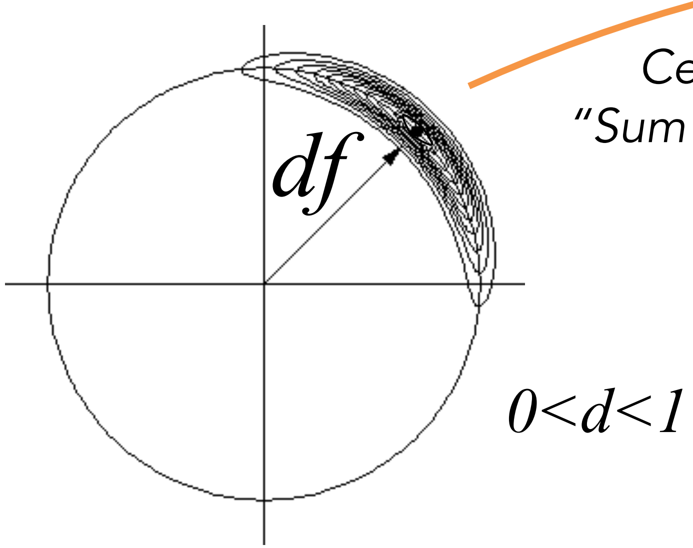
Summing structure factors with errors

- The model has errors
- The error in the atomic position manifests as a phase error in the structure factors

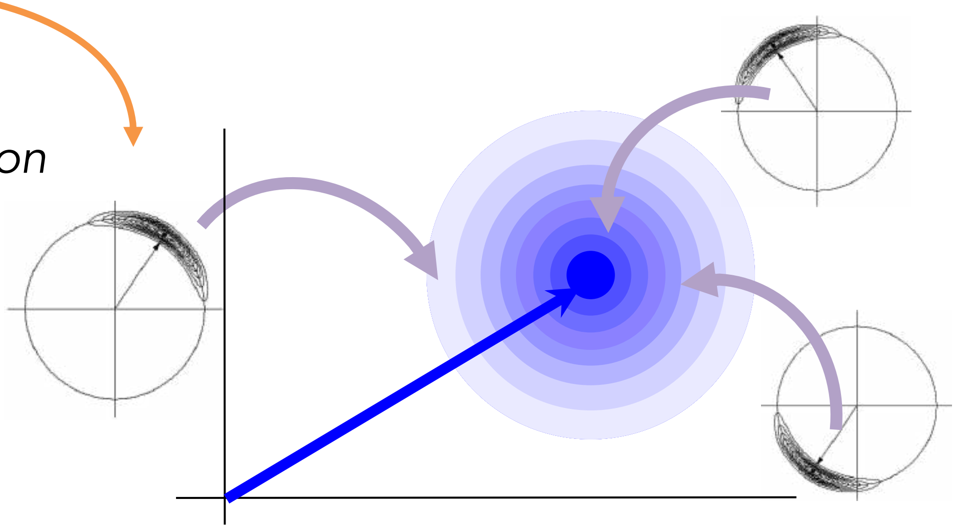


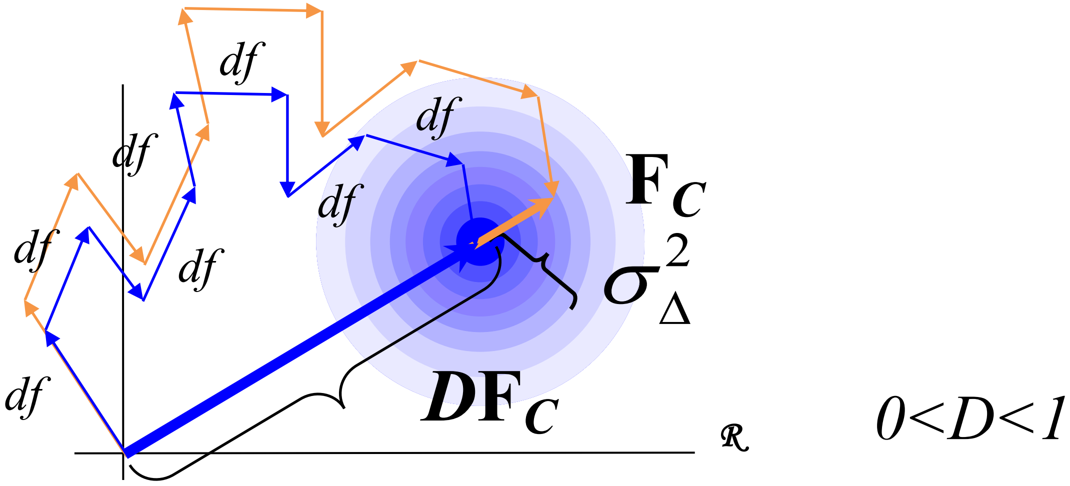


What is the probability distribution for the total structure factor with errors?

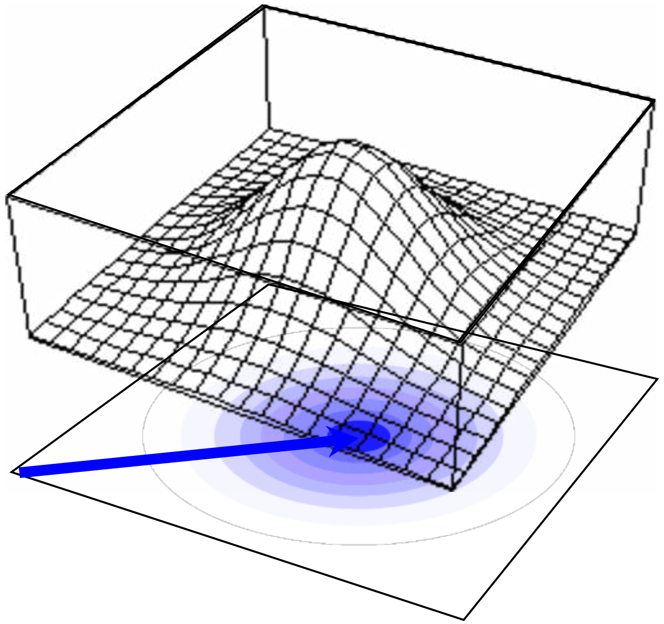


Central Limit Theorem:
"Sum of any error distribution
is a Gaussian"





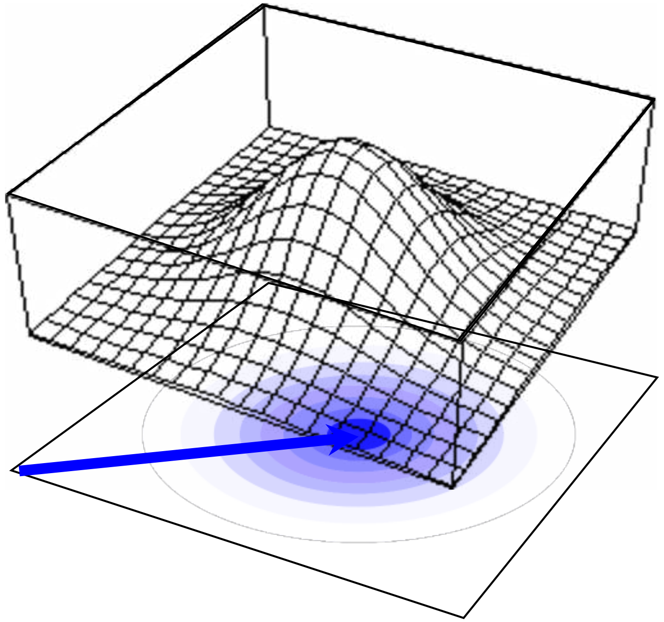
Structure Factor (F,D,σ_Δ)

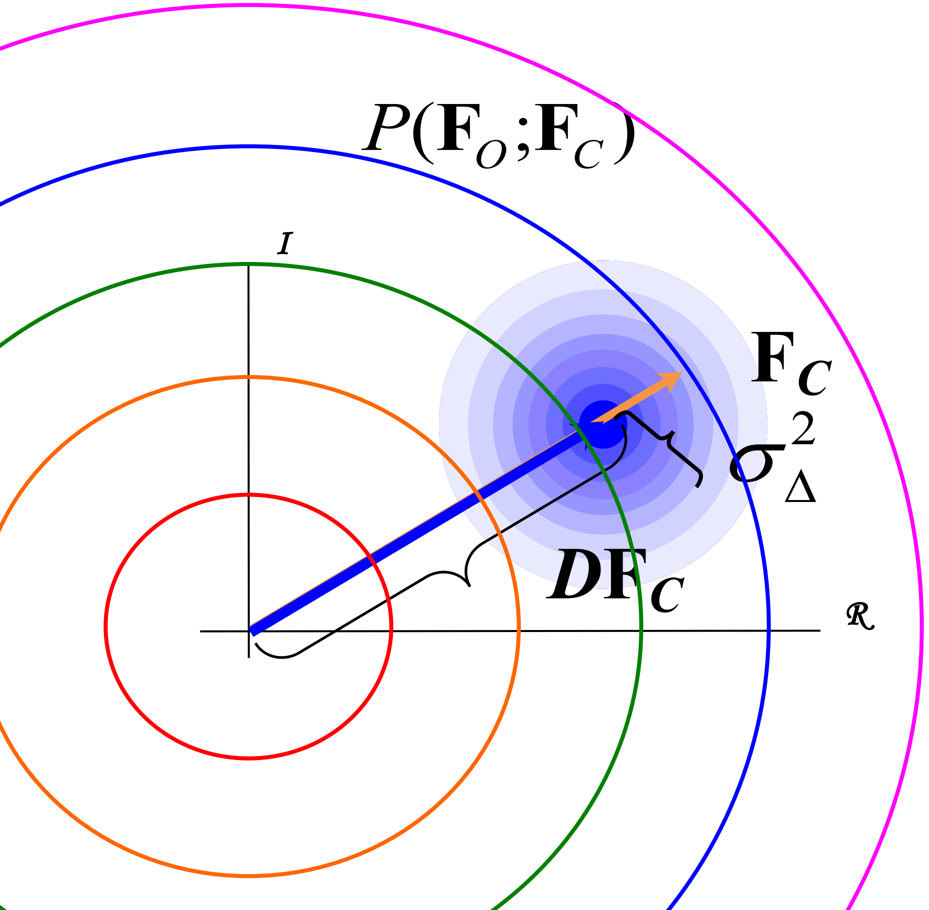


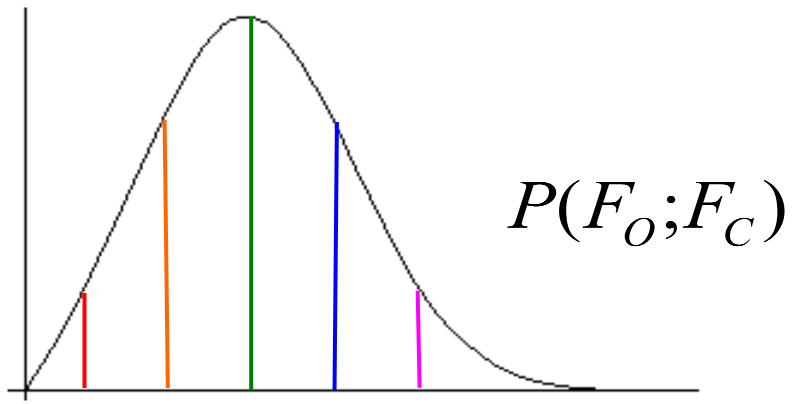
$$a\,exp\left(-\frac{|x-b|^2}{c^2}\right)$$

functional form for 2D Gaussian

$$P(\mathbf{F}_o;\mathbf{F}_C) = \frac{1}{\pi\sigma_\Delta^2}\,exp\left(-\frac{|\mathbf{F}_O - D\mathbf{F}_C|^2}{\sigma_\Delta^2}\right)$$

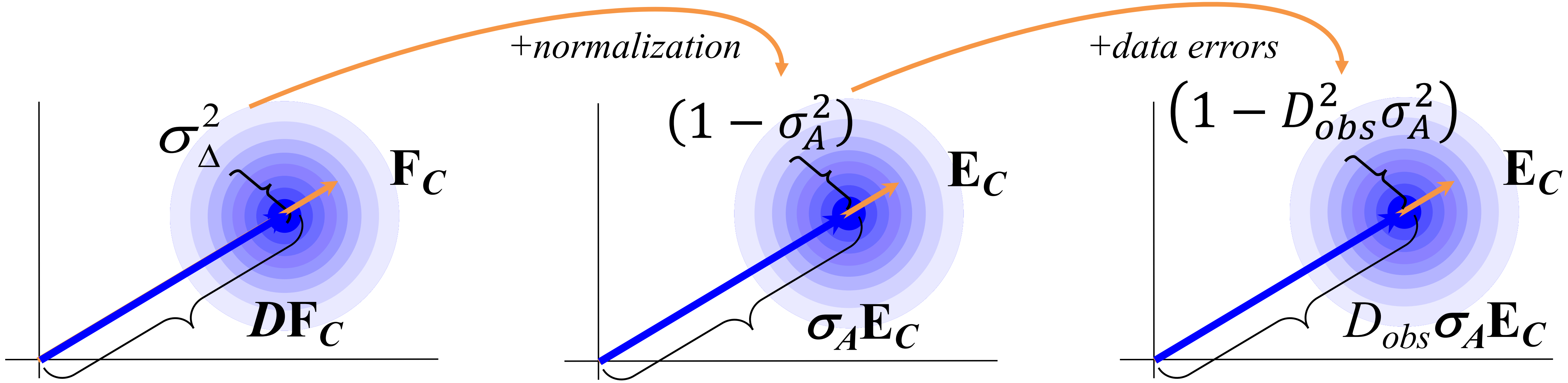






$$P(F_o;F_C) = \frac{2F_O}{\sigma_\Delta^2}\,exp\left(-\frac{F_O^2 + D^2F_C^2}{\sigma_\Delta^2}\right) I_0\left(\frac{2F_oDF_C}{\sigma_\Delta^2}\right)$$

2D Gaussians



Data: $|F_{obs}|$ Data: $|E_{obs}|$ Data: $|E_e|$

Advanced information: D_{obs} *and* E_e *come from an approximation to a log-likelihood gain based on intensities (which does not have an analytical solution) cast in terms of a function that implies complex errors. The values are determined once, before phasing.*

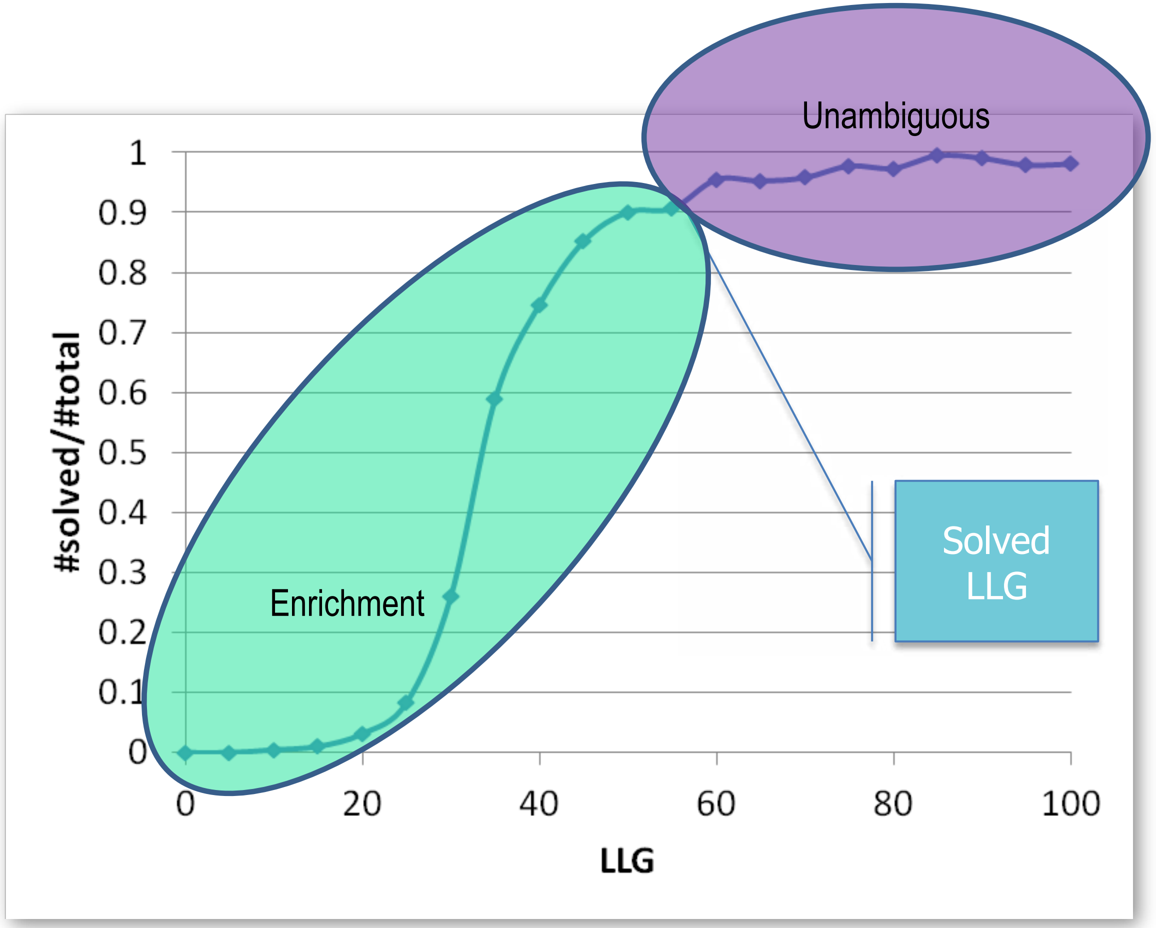
Read RJ, McCoy AJ. 2016 Acta Cryst D72:375

Phaser – likelihood theory

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

Do I have a solution?
Will I get a solution?

log-likelihood gain for solutions



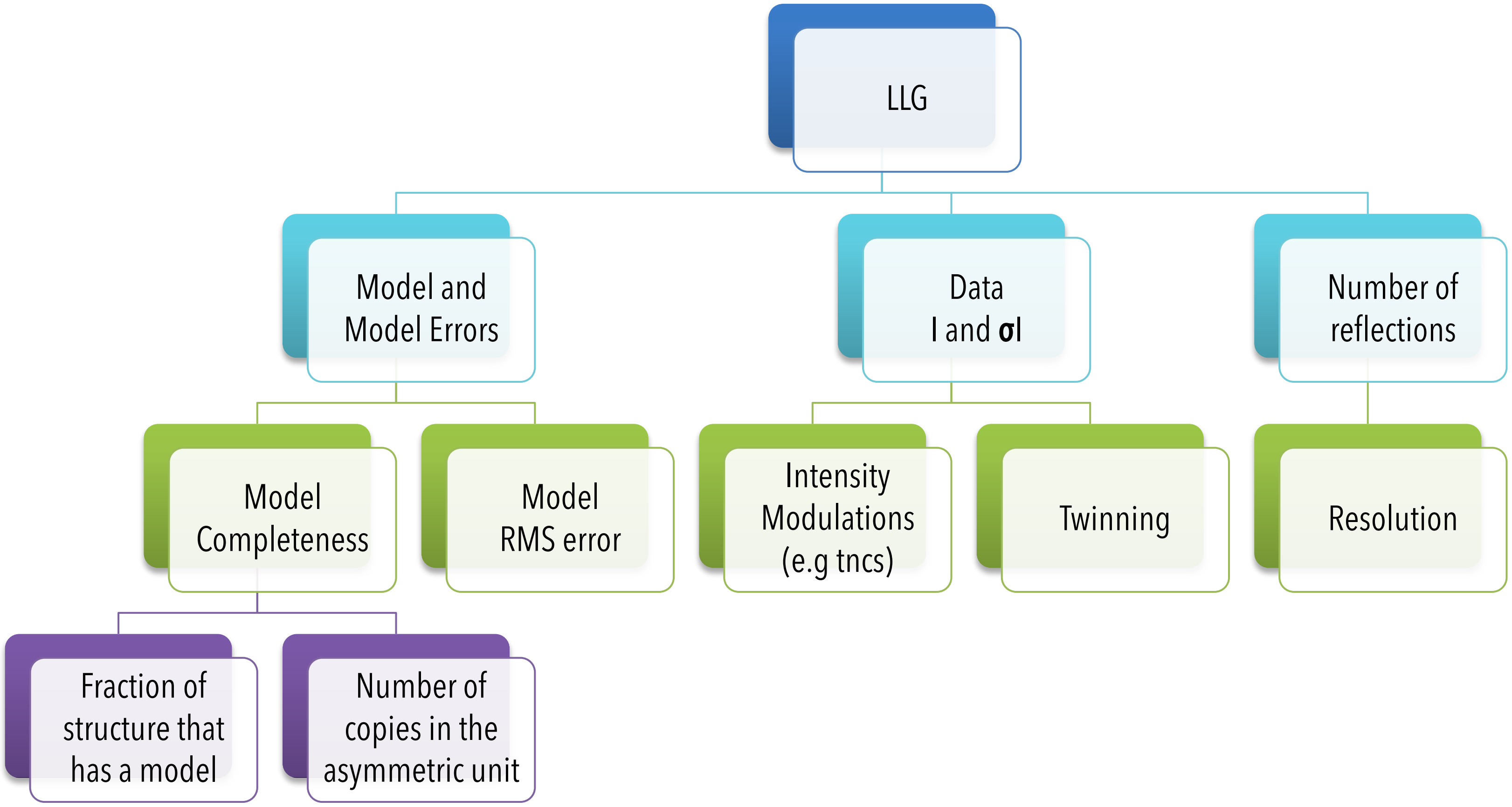
Plot of LLG versus success in structure solution

Database of
over 23000
MR problems

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


LLG

Model and Model Errors

Data I and σI

Number of reflections

Model Completeness

Model RMS error

Intensity Modulations (e.g tncs)

Twinning

Resolution

Fraction of structure that has a model

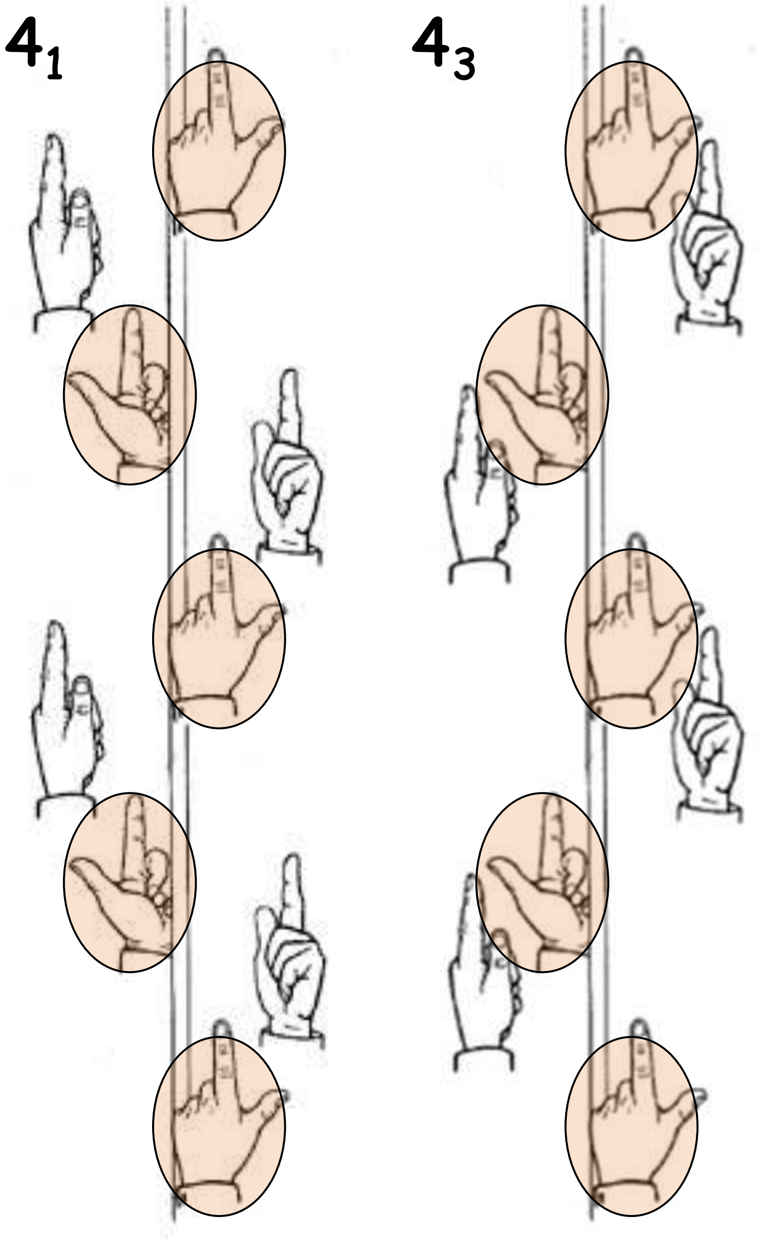
Number of copies in the asymmetric unit

- Phaser’s target function allows us to account for many of the problems that occur in molecular replacement

how not to judge a solution

- Don’t use “I can see features in the difference map”
- Most incorrect solutions with phaser are partially correct
  - the phases are partially correct
- Therefore there will be things to see in the difference maps – just not as clearly as with a fully correct solution
- Correct solutions can only be judged against other possibilities

An incorrect solution in space group P4$_1$ for space group P4$_3$ will have half the molecules in the correct place



expected log-likelihood gain for solutions

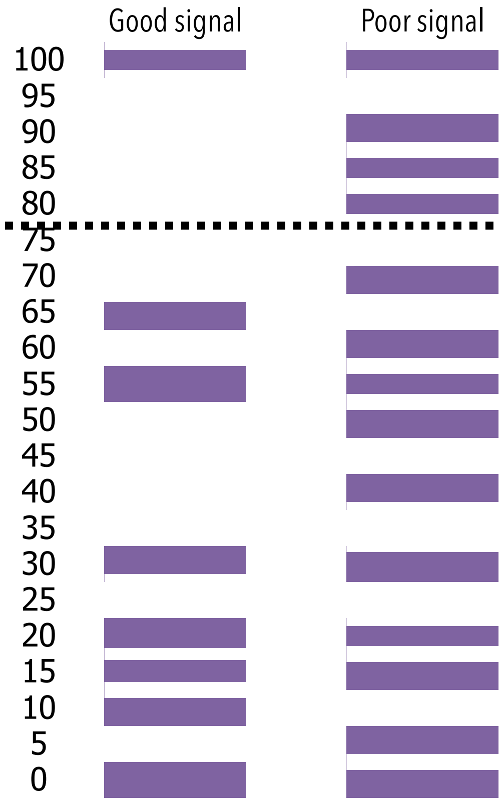
- So if you can predict the LLG…
  - You know how easy/difficult will be MR
  - You can prioritize structure solution strategies
- The expected LLG of a model predicts the LLG for a placed model
- Removes uncertainty in MR
  - Knowing when to start/stop has always been the problem with MR
  - Prioritize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

When will a model be placed?

eLLG score	Will it place?
< 25	no
25 - 36	unlikely
36 - 49	possibly
49 - 64	probably
> 64	definitely

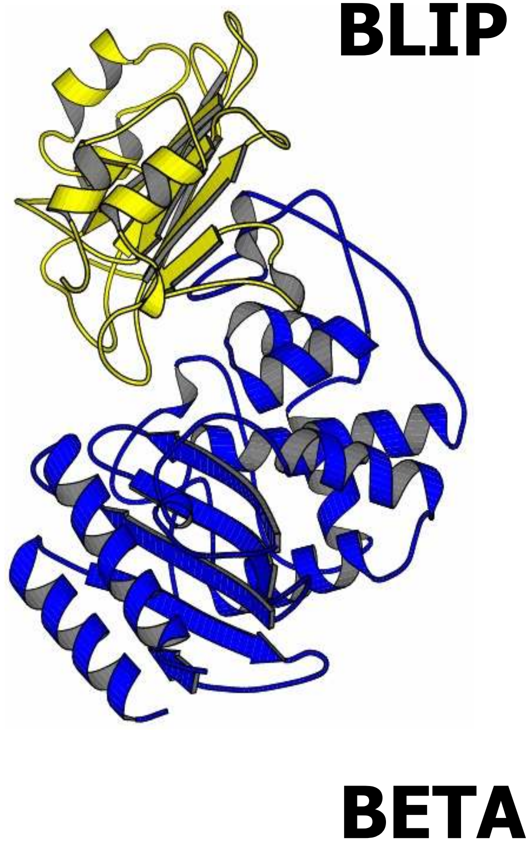
expected log-likelihood gain for solutions

- If there are multiple copies in the asymmetric unit then the placement of the first molecule may not give a signal in MR
- As long as the correct pose for the first component is in the list then second and subsequent components become easier to place
  - and selects the correct first placement from the list
- There is also an eLLG calculation for how many components need to be placed to find a clear solution



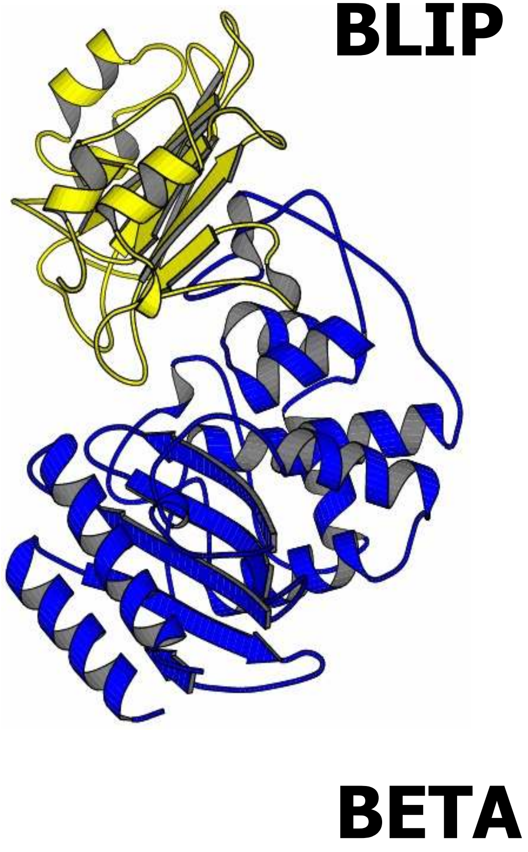
β-lactamase/β-lactamase inhibitor

- BETA and BLIP previously solved separately
- One complex in asymmetric unit
    - BETA 62% residues
    - BLIP 38% residues
- Data are very anisotropic

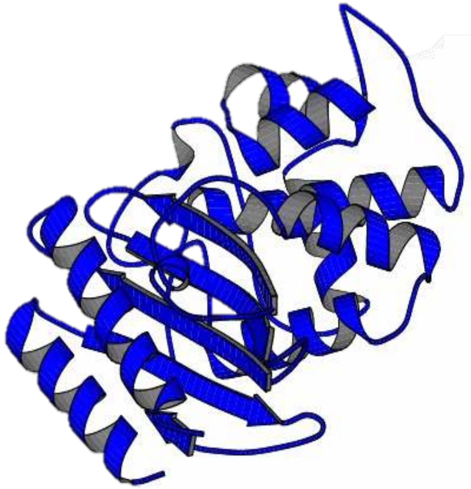


Search Order

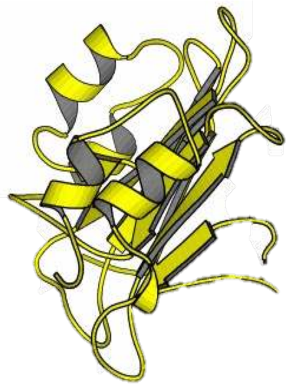
- Finding a MR solution is dependent on search order when complexes are built by addition
- Search for the model that explains the most scattering first
- This is a function of
  - (known) model completeness
  - (estimated) RMSD of coordinates
  - (unknown) relative target B-factors
  - Number of reflections (resolution)



Which model is easiest to find?



Completeness:
62% of
structure
RMSD: 1.2Å



Completeness:
38% of
structure
RMSD: 0.8 Å

Better model if data resolution is low

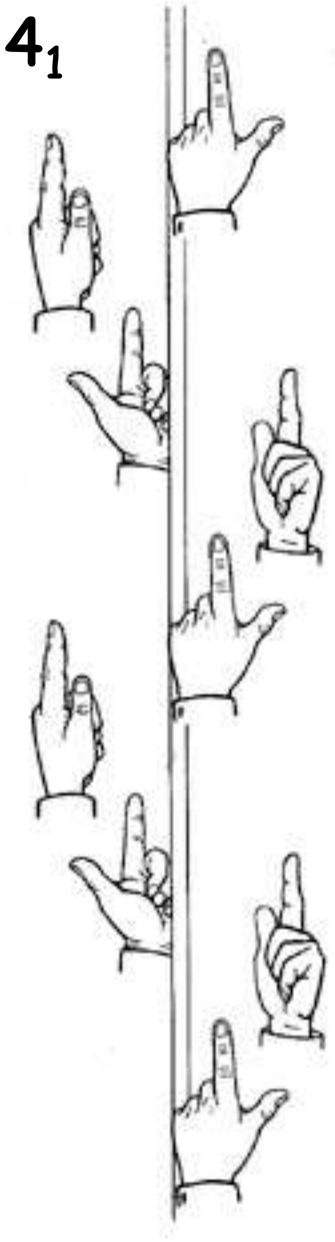
Better model if data resolution is high

space group alternatives

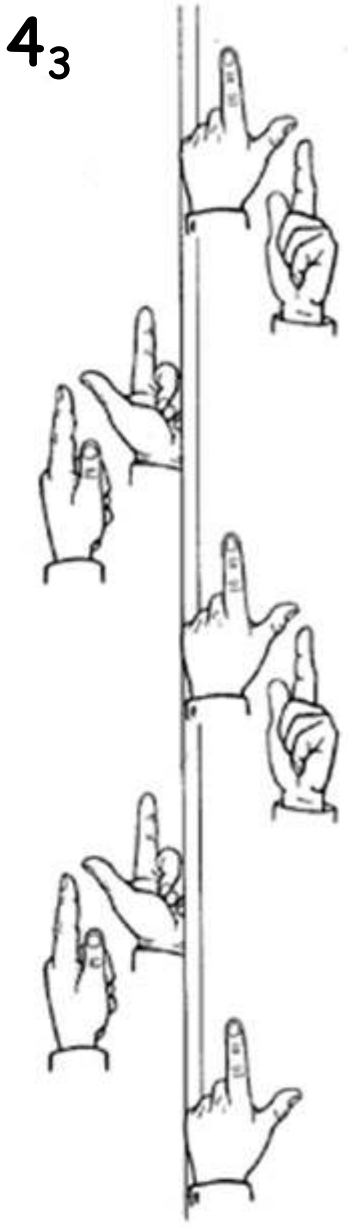
space group determination

- Space groups that come in enantiomorphic pairs (e.g. $P4_1$, $P4_3$) cannot be distinguished at the data processing stage
- The space group is only confirmed when the structure is solved
- Phaser will run enantiomorphic pairs in the point group in the translation function
  - Or all space groups in the point group

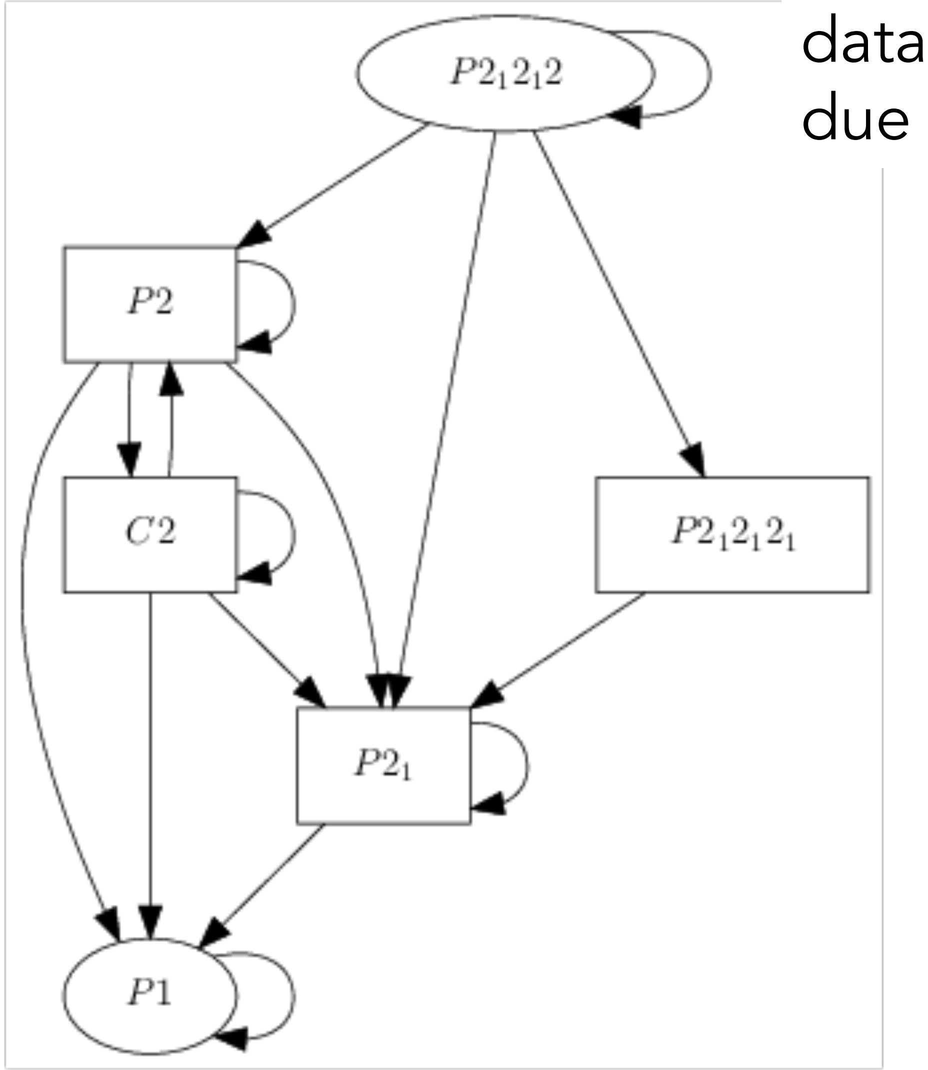
4_1



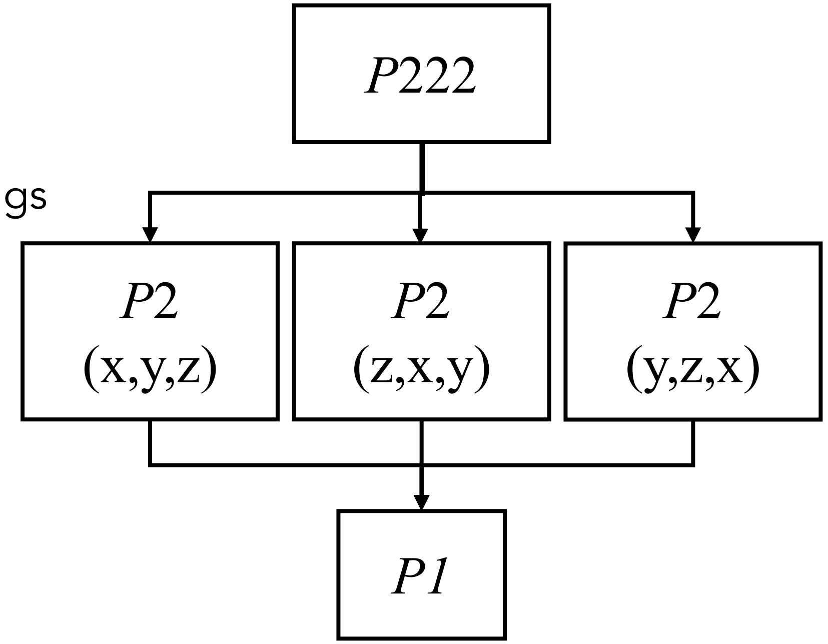
4_3



subgroups



data over-merged
due to twinning



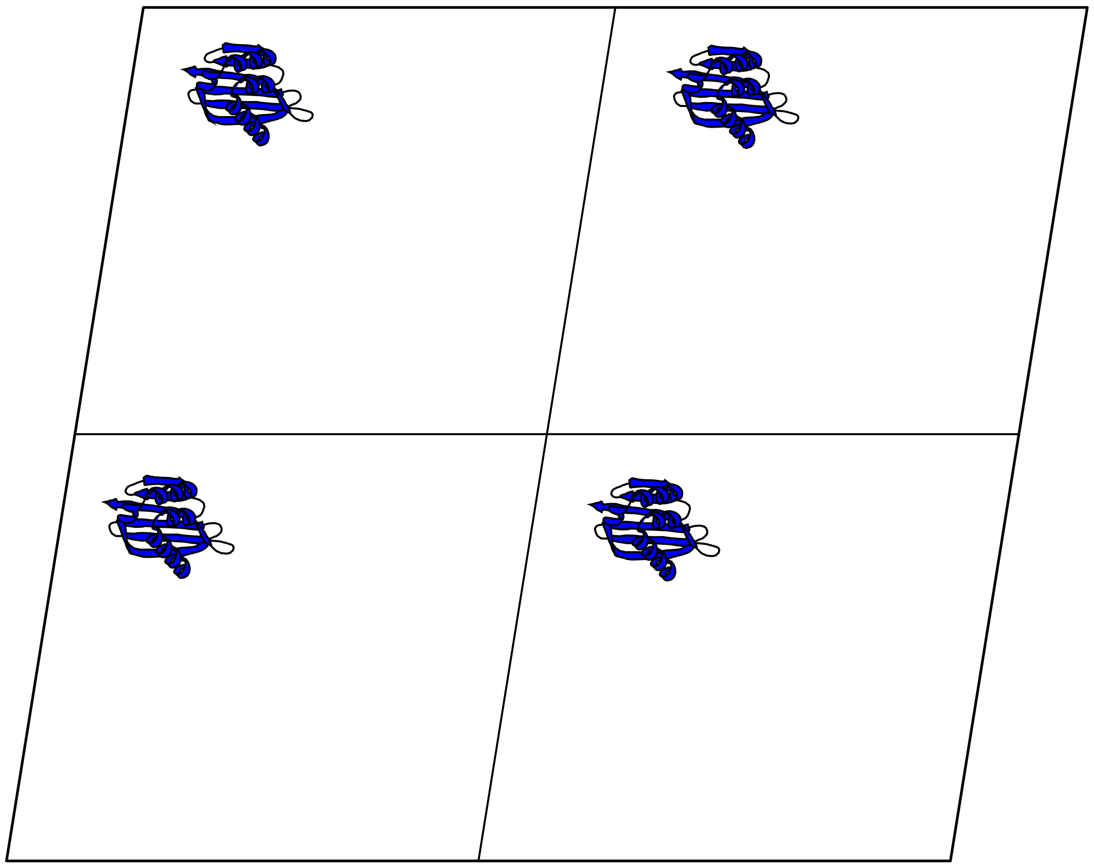
P2 standard settings
axis permutations

subgroups in Phaser
for molecular replacement

alternative origins

P1

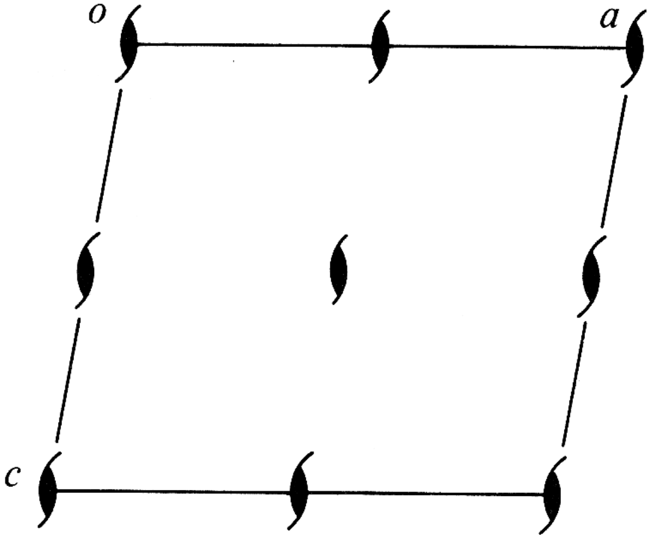
- Origin arbitrary

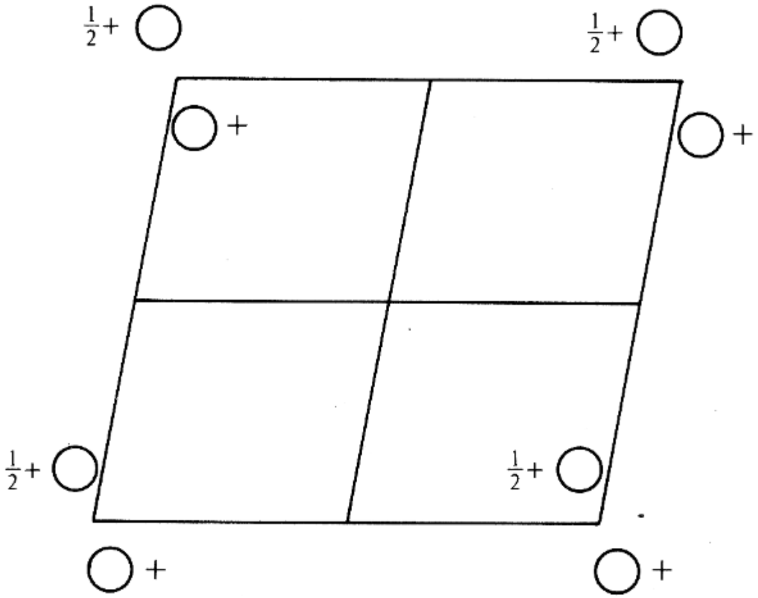


$P\,2_1$

No. 4

UNIQUE AXIS b





Origin on 2_1

Asymmetric unit $0 \leq x \leq 1;\quad 0 \leq y \leq 1;\quad 0 \leq z \leq \frac{1}{2}$

Symmetry operations

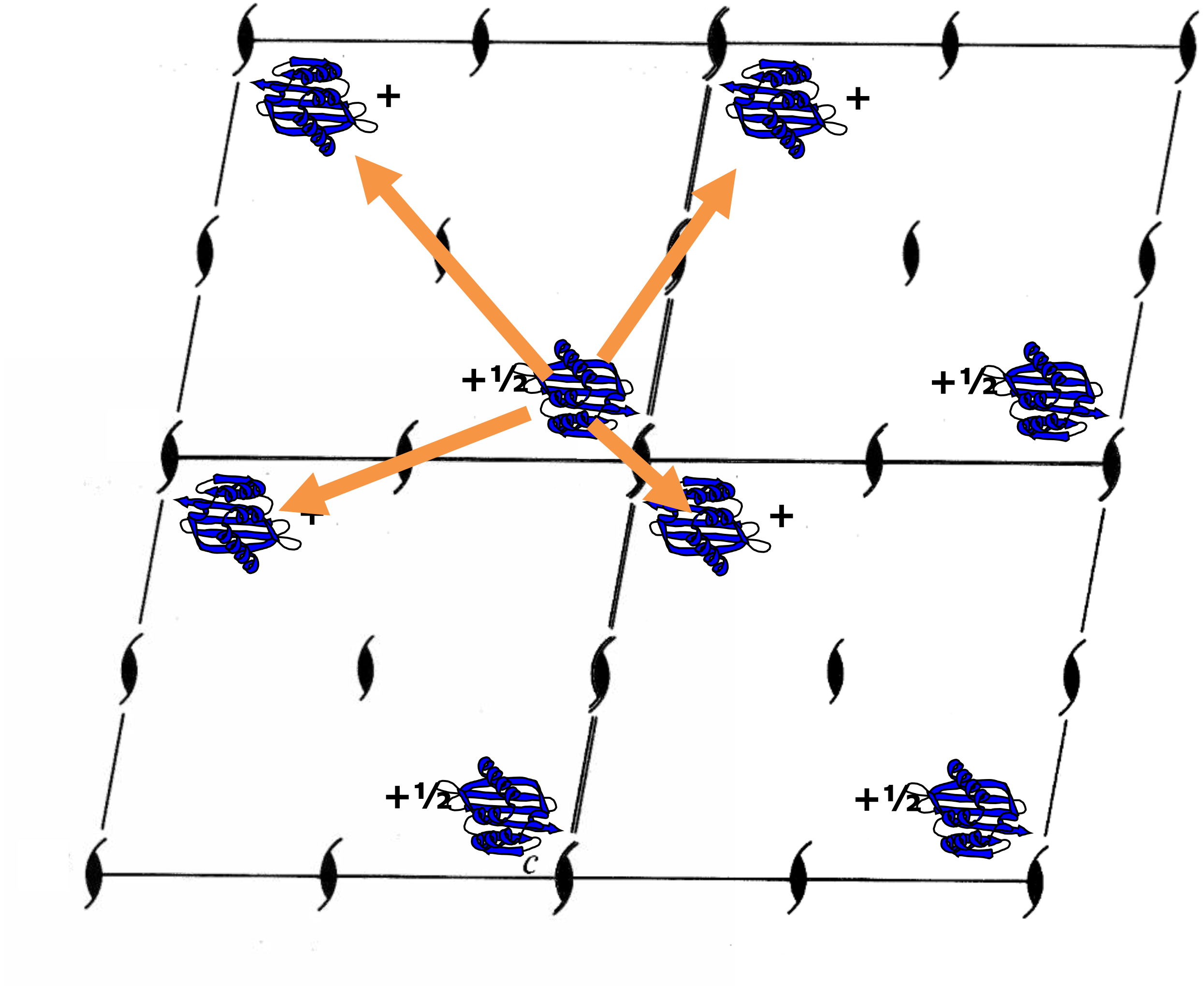
(1) 1 (2) $2(0,\frac{1}{2},0)\quad 0,y,0$

Positions

Multiplicity, Wyckoff letter, Site symmetry			Coordinates	
2	a	1	(1) x,y,z	(2) $\bar{x},y+\frac{1}{2},\bar{z}$

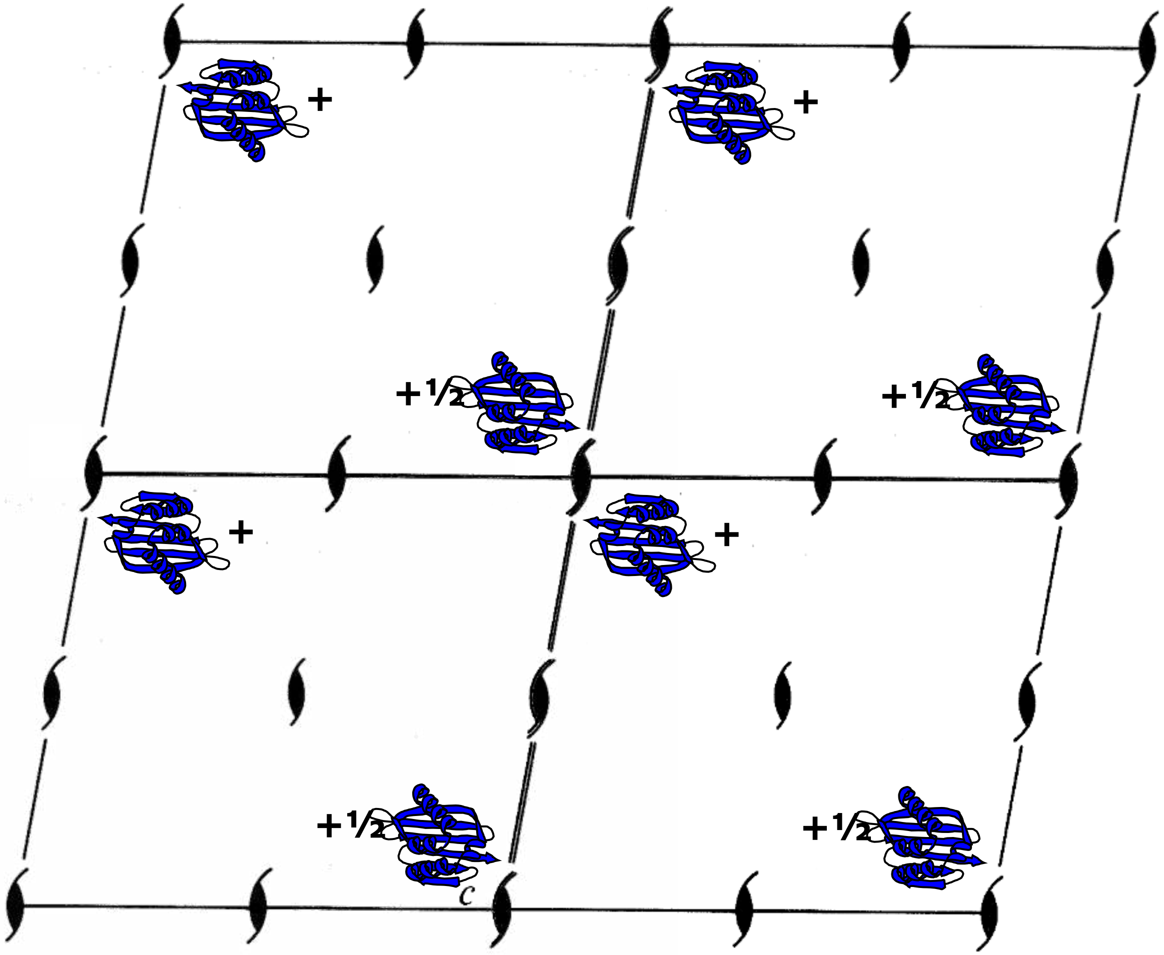
$P2_1$

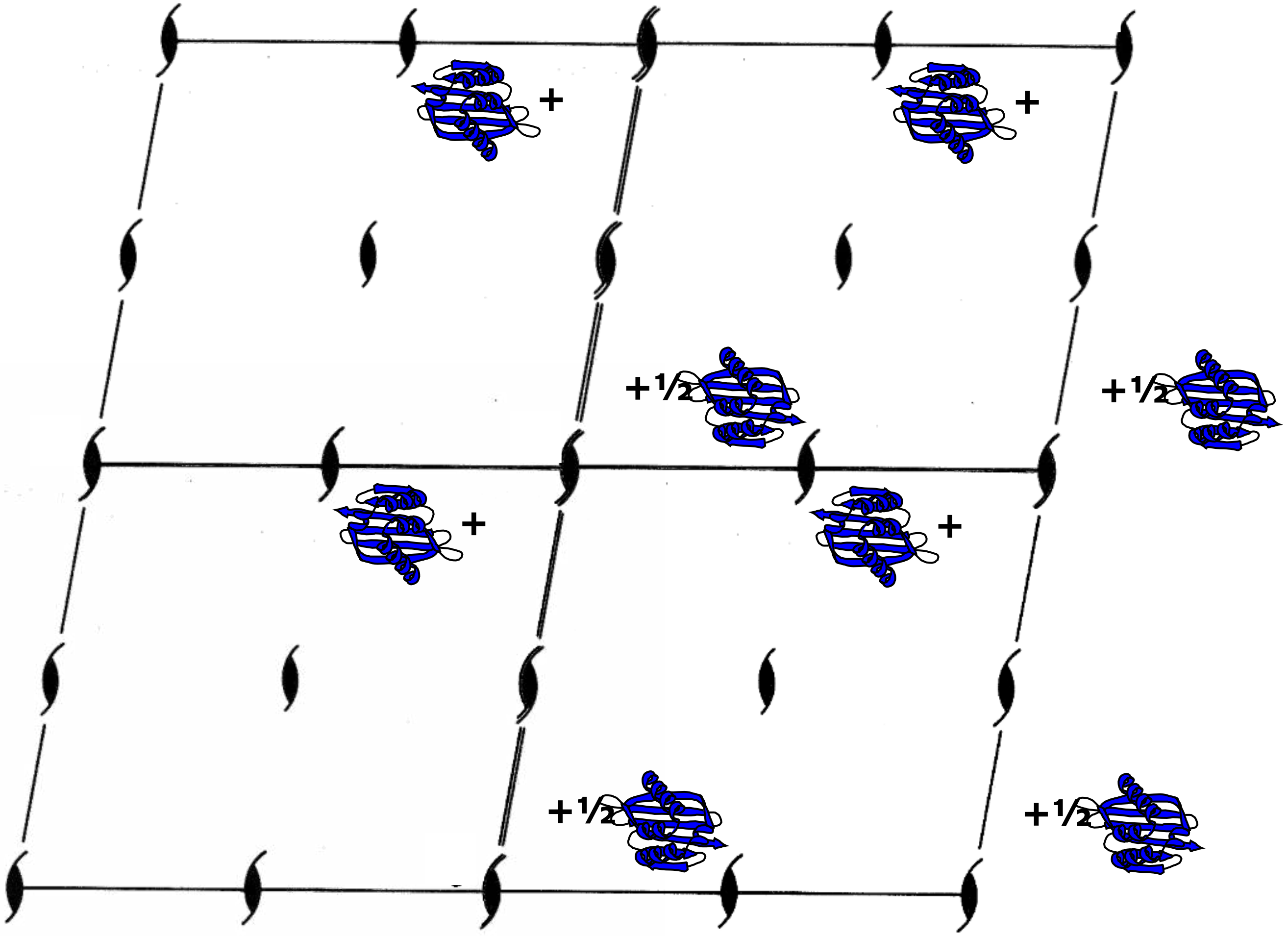
- Origin anchored at symmetry operations
- Symmetry operations (x,y,z), (-x,y+½,-z)

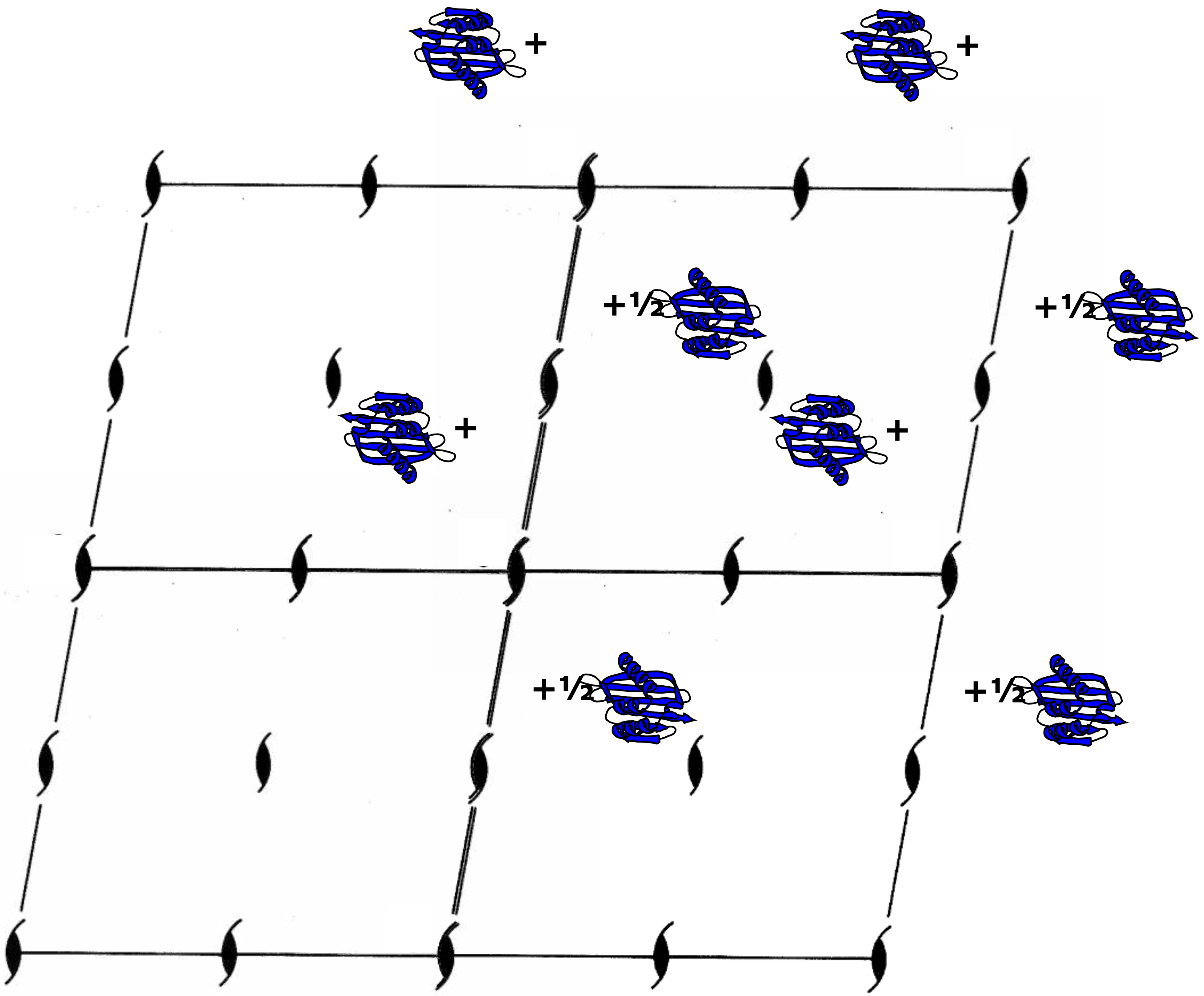


origins

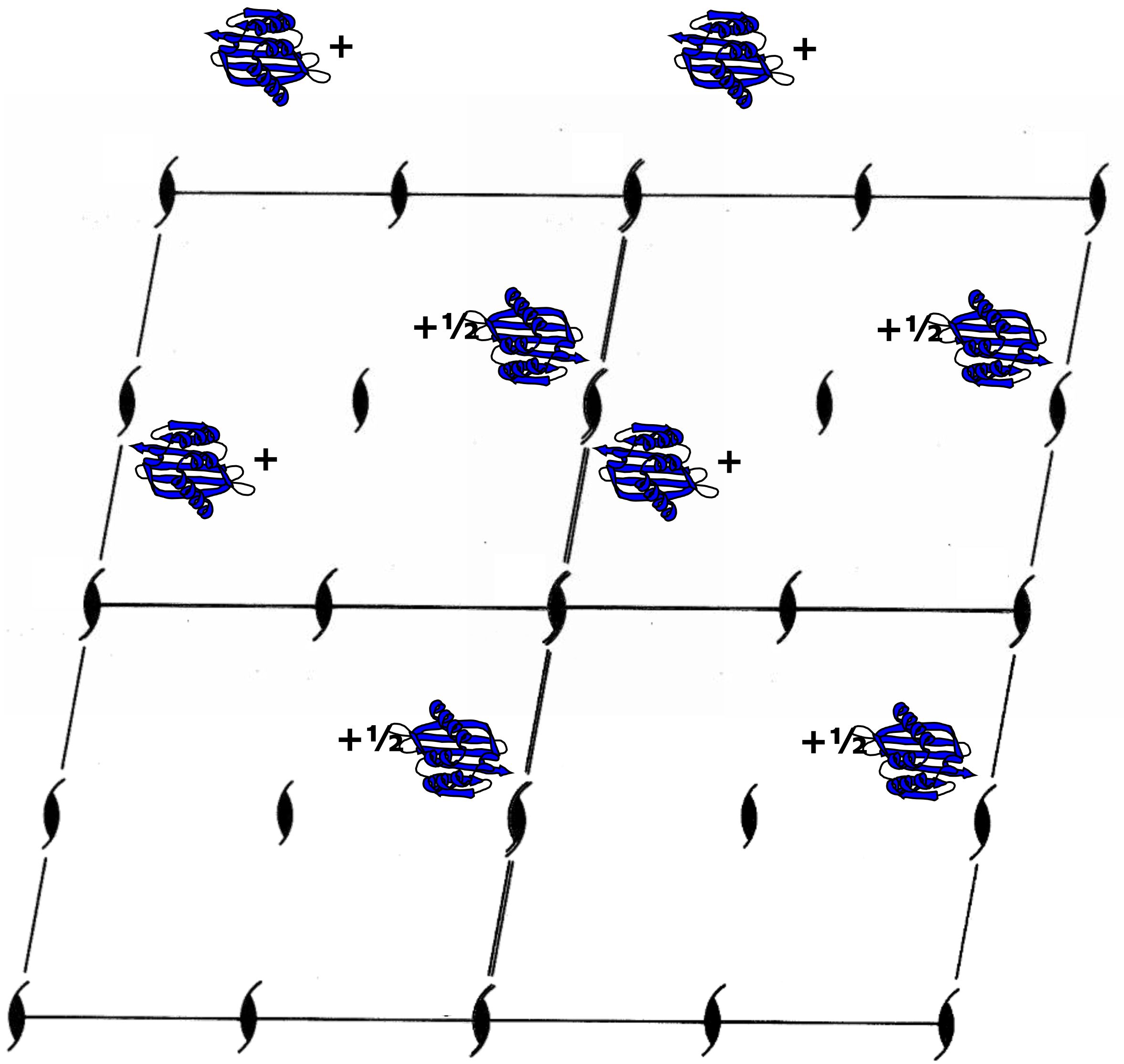
- Can only find the translation perpendicular to a rotation axis
  - no rotation symmetry, no translation to find!
- If there are multiple symmetry axes of the same order of rotation in a plane then the translation can be defined with respect to any one of these
  - These are equivalent to different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


+
+
+½
+½
+
+
+½
+½
c


+
+
+½
+½
+
+
+½
+½



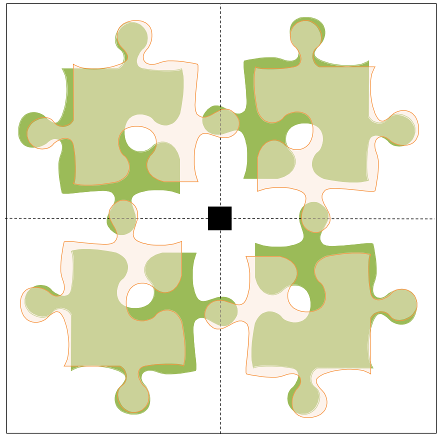
+
+
+½
+½
+
+
+½
+½


+
+
+½
+½
+
+
+½
+½

pathologies

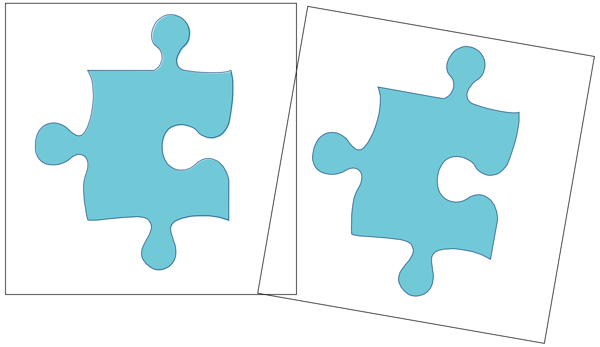
Pathologies

1.Twinning

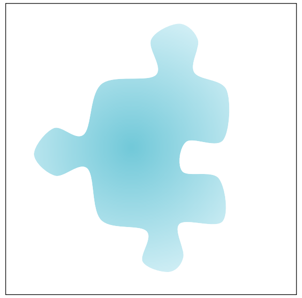


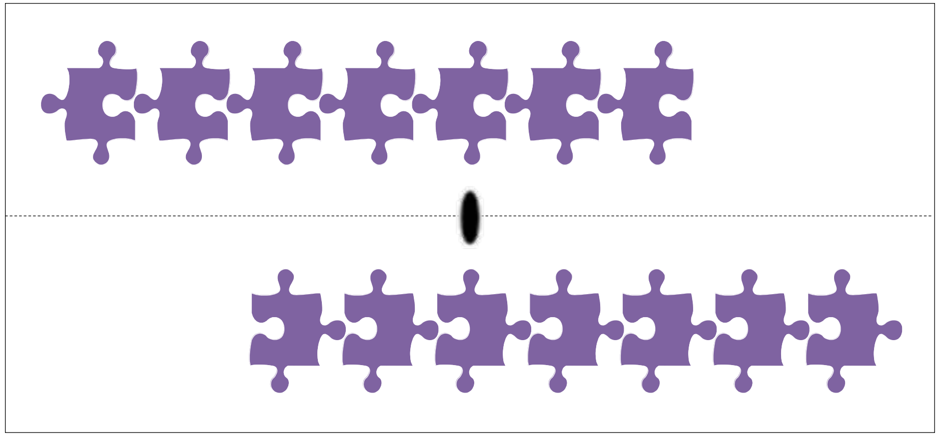
3. Anisotropy

2. High Mosaicity



4. Low Resolution





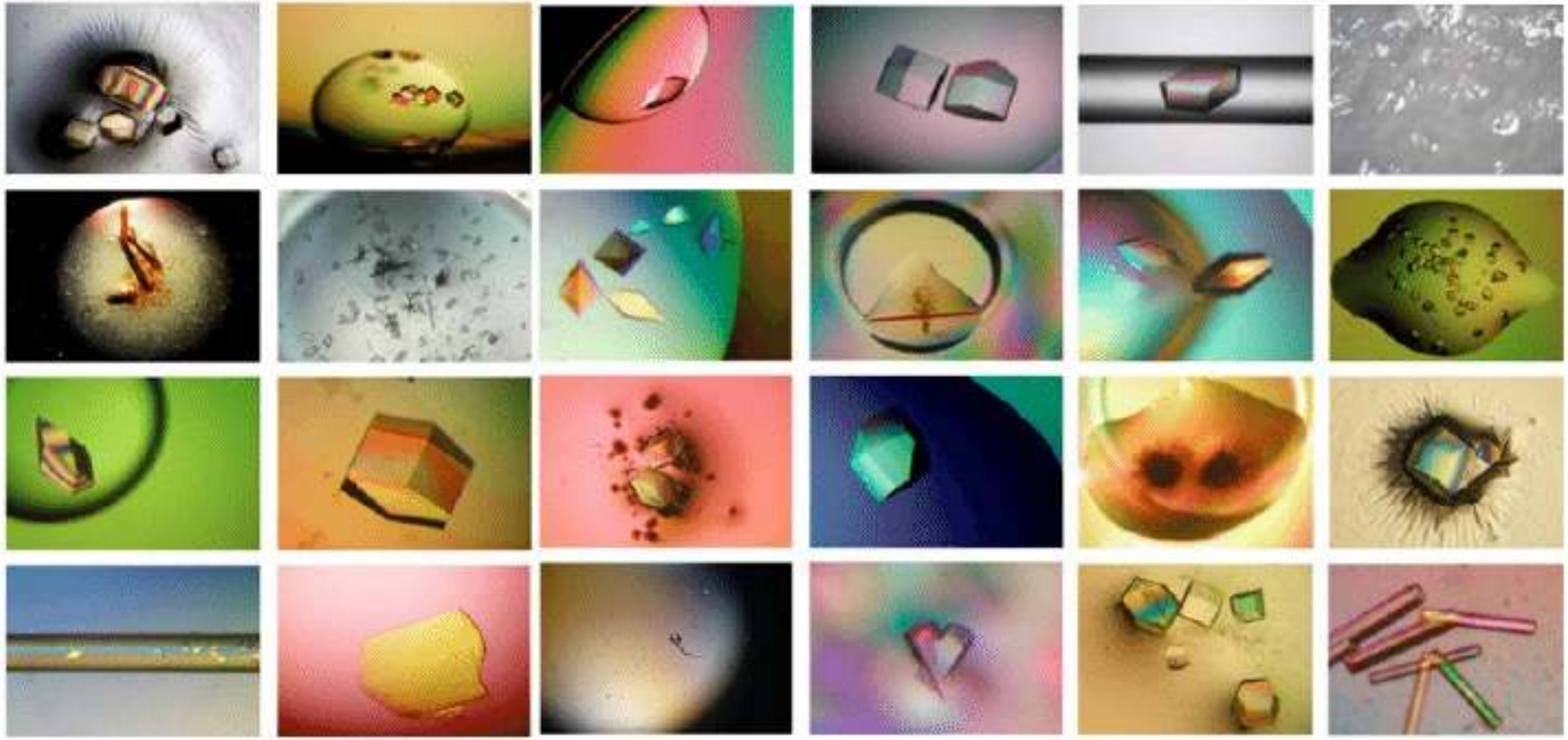
5. Translational non-crystallographic symmetry

Practical Pathologies

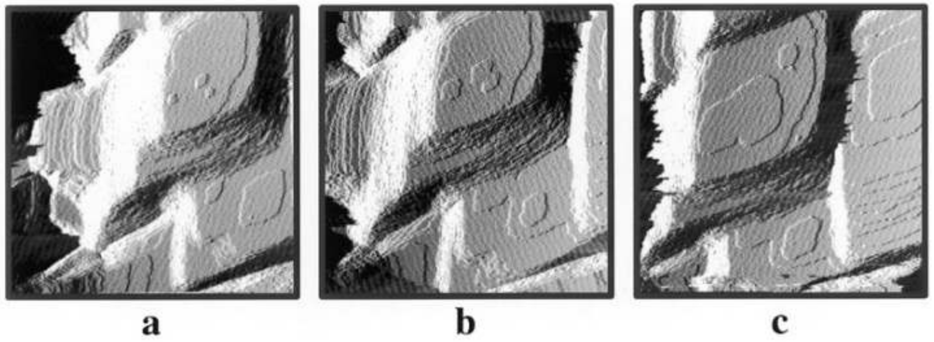
- Some pathologies can only be cured with better crystals or better crystal handling or better data collection strategies
  - High mosaicity
  - Low resolution

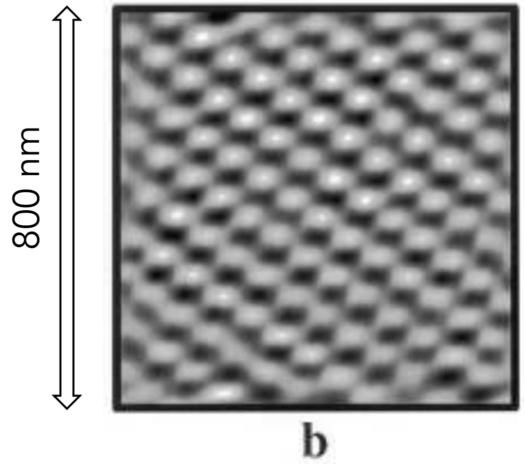
- Other pathologies can be cured using computational methods after data collection
  - Anisotropy
  - Twinning
  - Translational non-crystallographic symmetry

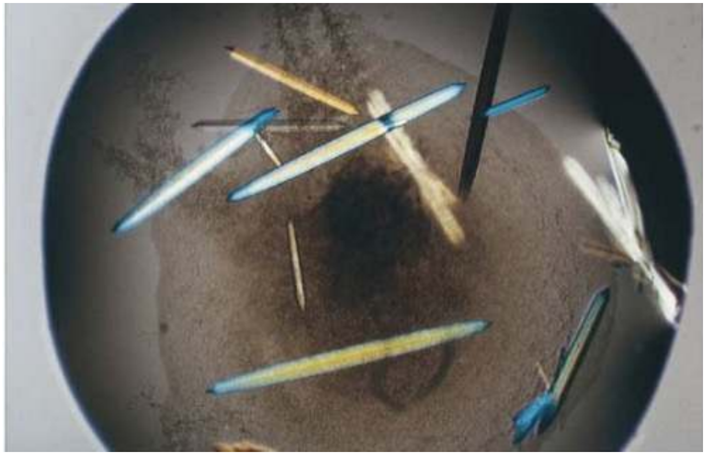
Rupp’s crystal gallery

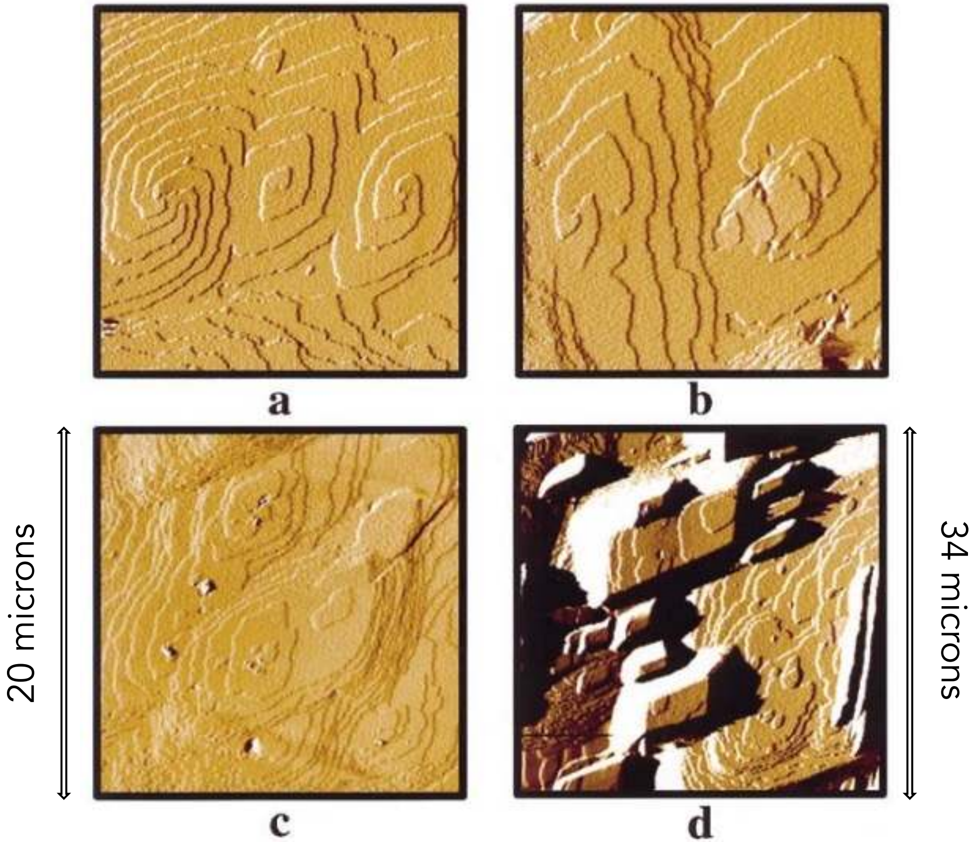


Crystals





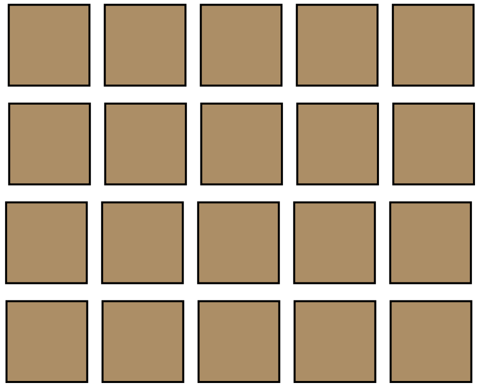




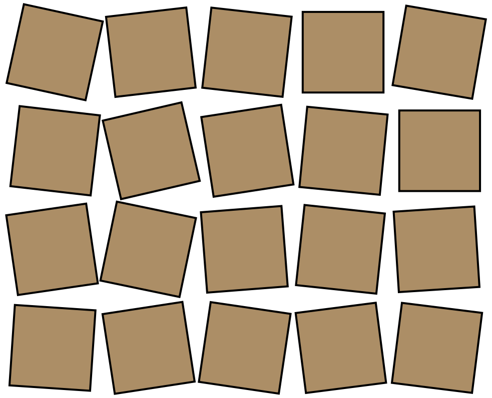
AFM images of yeast tRNA^{Phe}

FEBS Journal 280 (2013) 6456–6497

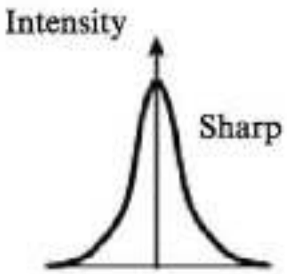
mosaicity and low resolution

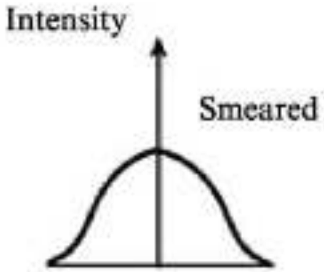


(Almost) Perfect crystals



Real crystals
Mosaic blocks
Highly exaggerated

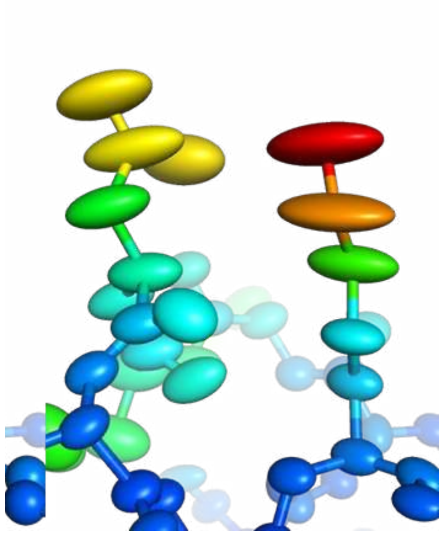




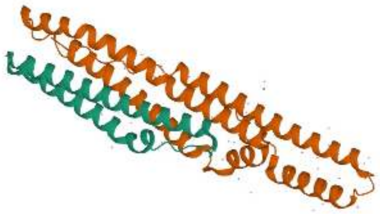
pathologies:
anisotropy

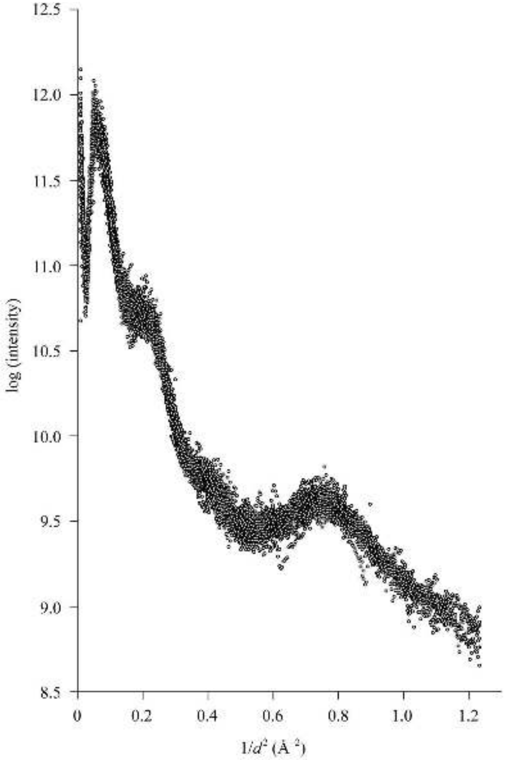
anisotropy correction

- make intensity distribution the same in all directions by refining parameters of the anisotropy tensor to give best fit to ‘BEST’ curve



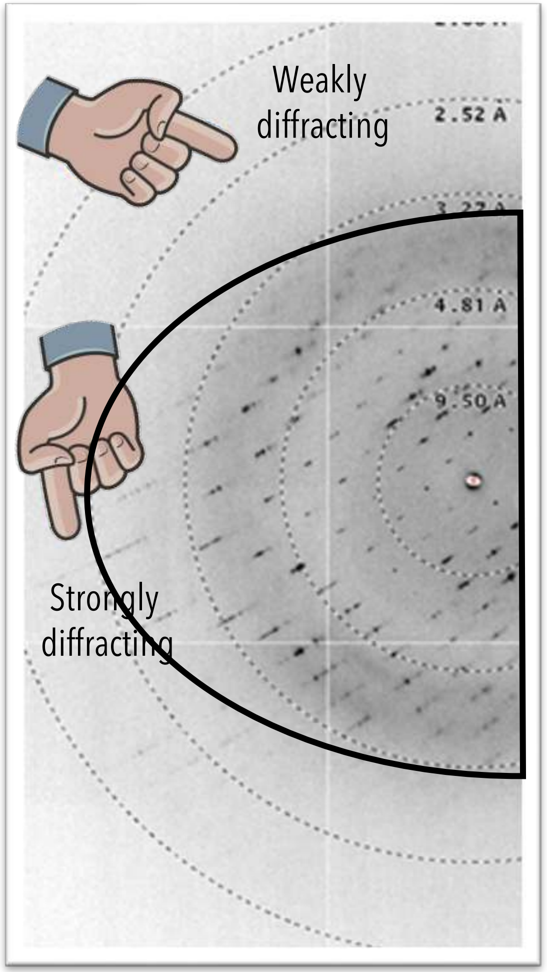
2G38



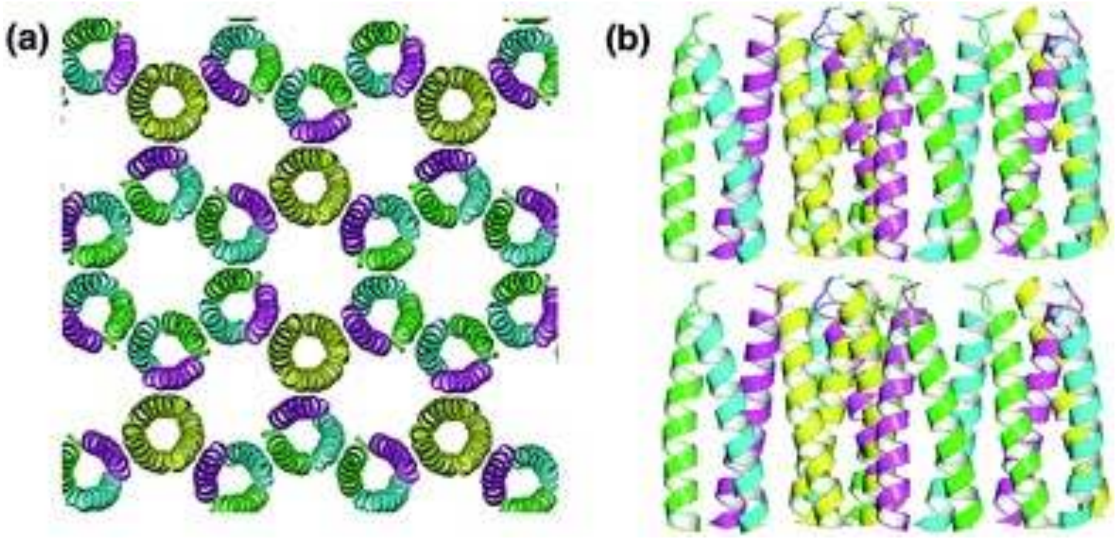


BEST curve
The average intensity versus resolution (from 12 to 0.9 A) for 72 protein crystals.

Popov & Bourenkov (2003)
Acta Cryst. D59, 1145–1153



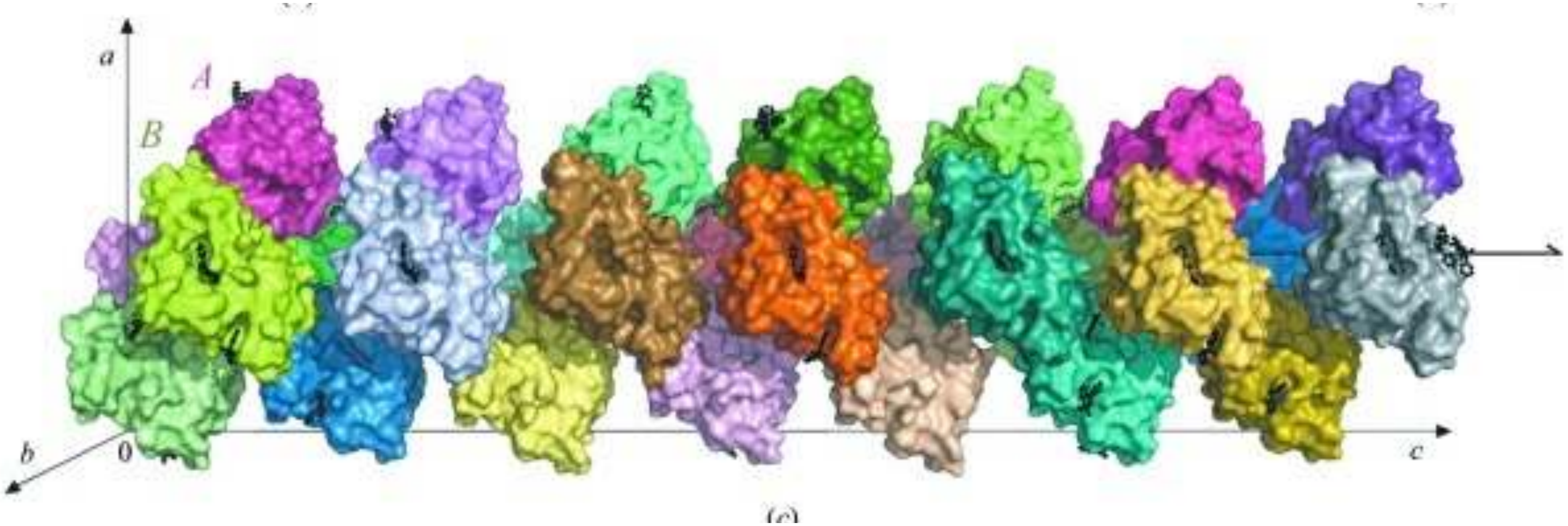
Anisotropy in coiled coils

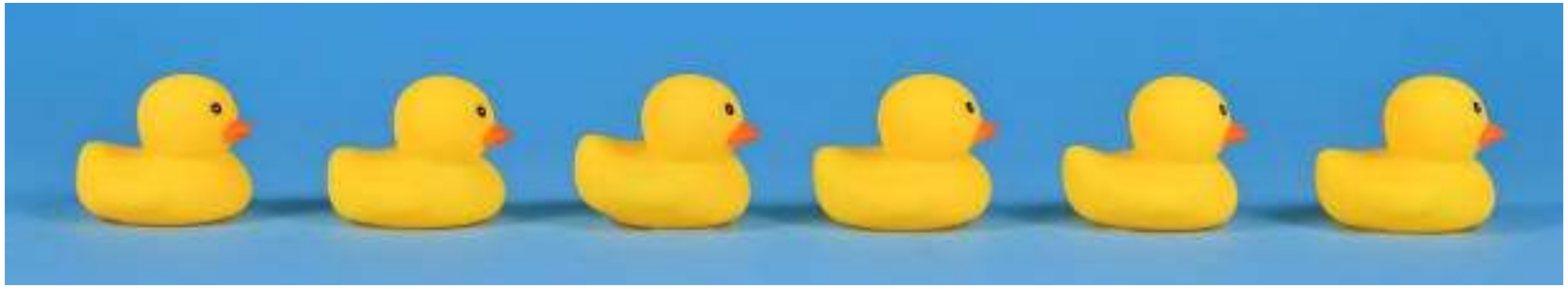


(a) Honeycomb-like P6 crystal lattice of a computationally-designed coiled-coil protein building block. (b) The layer structures were stacked by hydrogen bond interactions between layers. 4DAC

pathologies:
translational non-crystallographic symmetry

translational non-crystallographic symmetry



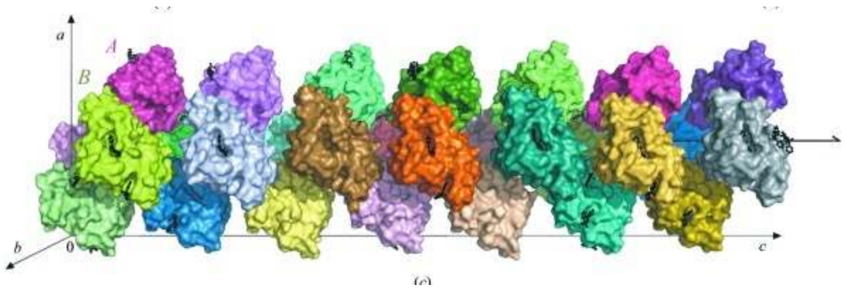


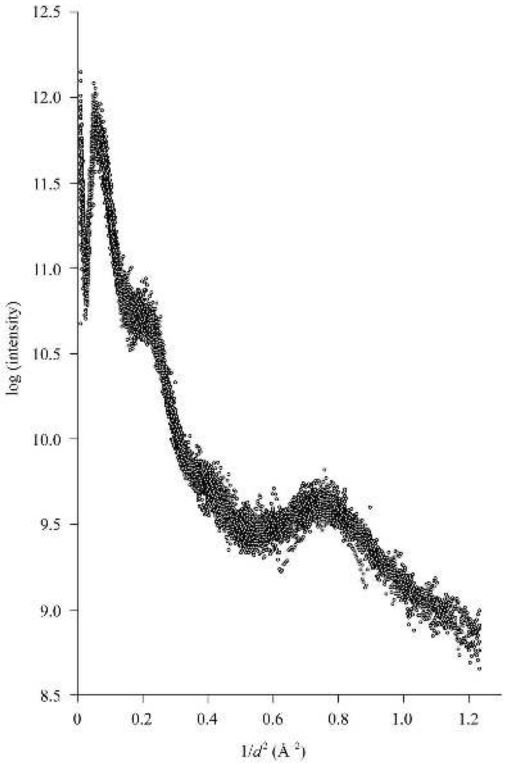
translational non-crystallographic symmetry

- If TNCS is not accounted for then TFZ > 8 does not indicate a correct placement
  - TFZ always higher
  - TFZ > 12 can be wrong
- Assumptions of maximum likelihood, for the LLG calculation, are violated
- When TNCS is ‘accounted for’ then the TFZ values are those expected of data without TNCS
  - ‘accounted for’ means that we have expected intensity factors for each reflection that characterize and correct for the TNCS modulations

translational non crystallographic symmetry

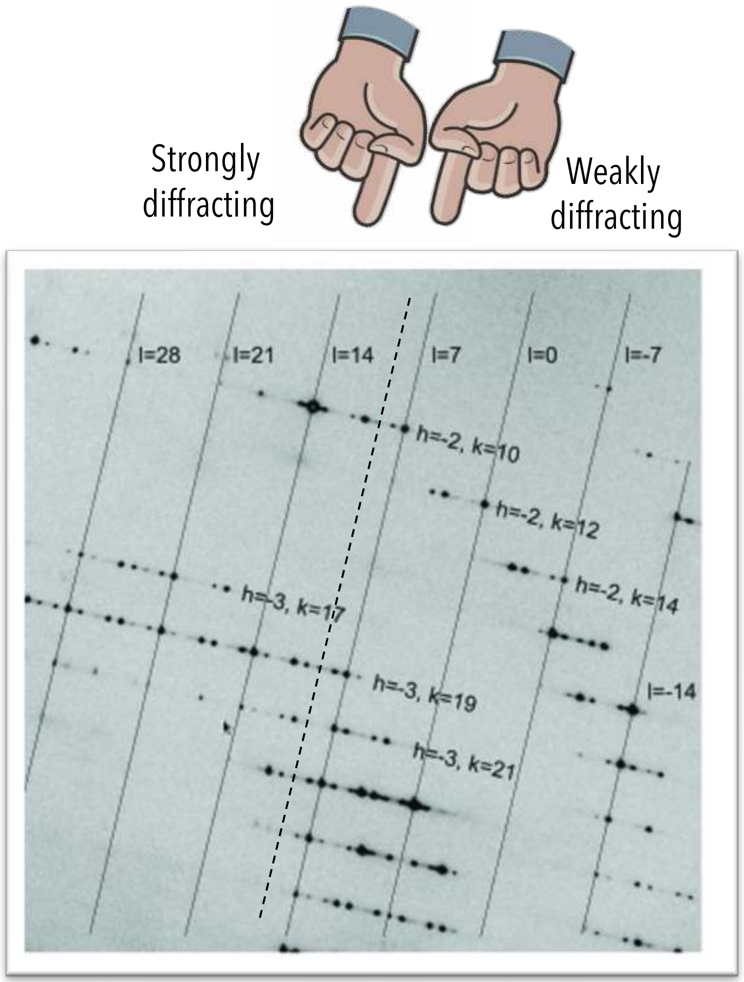
- Make intensity distribution the same in all layers by refining parameters for the expected intensity factors





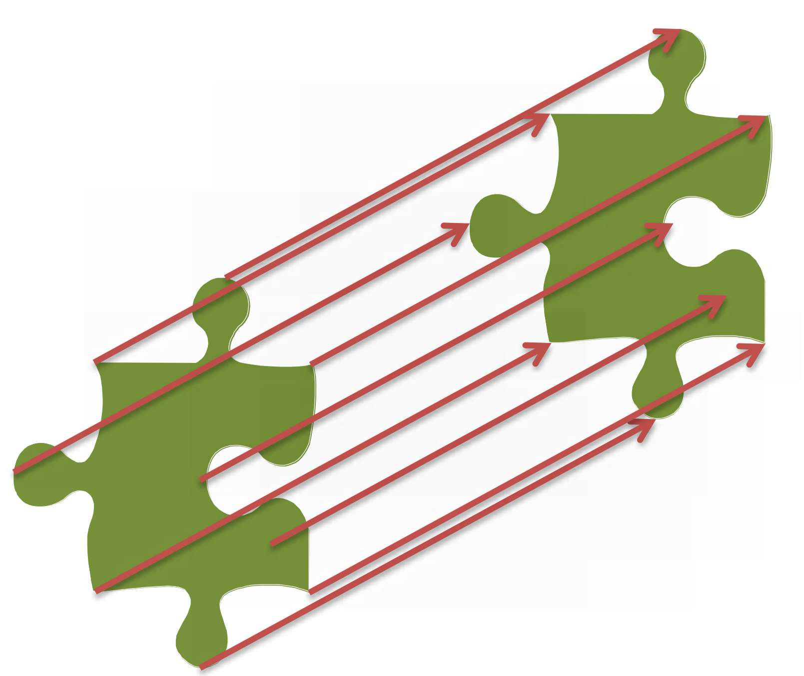
BEST curve
The average intensity versus resolution (from 12 to 0.9 A) for 72 protein crystals.

Popov & Bourenkov (2003)
Acta Cryst. D59, 1145–1153

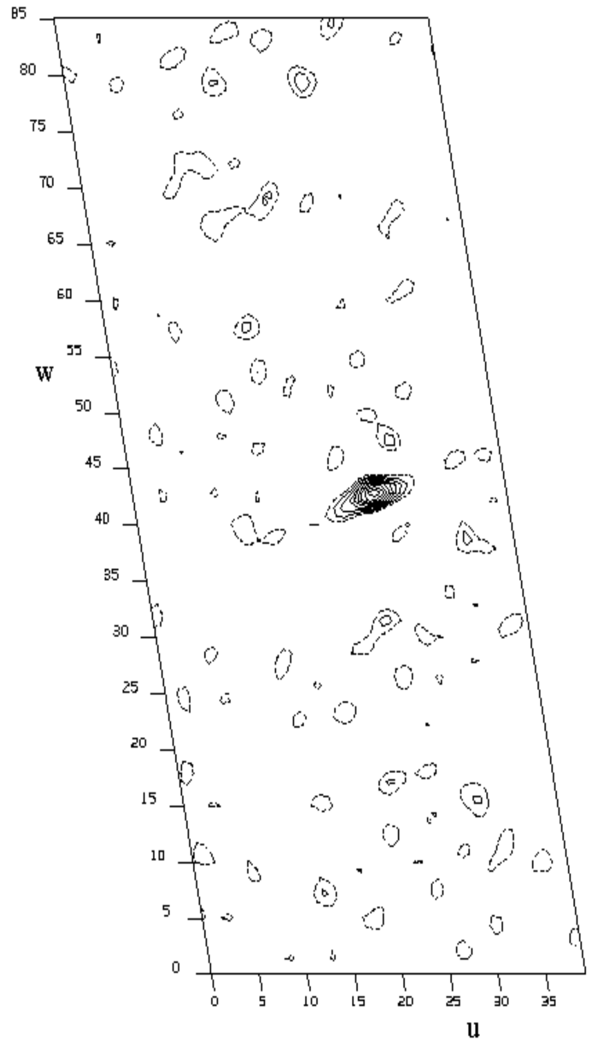


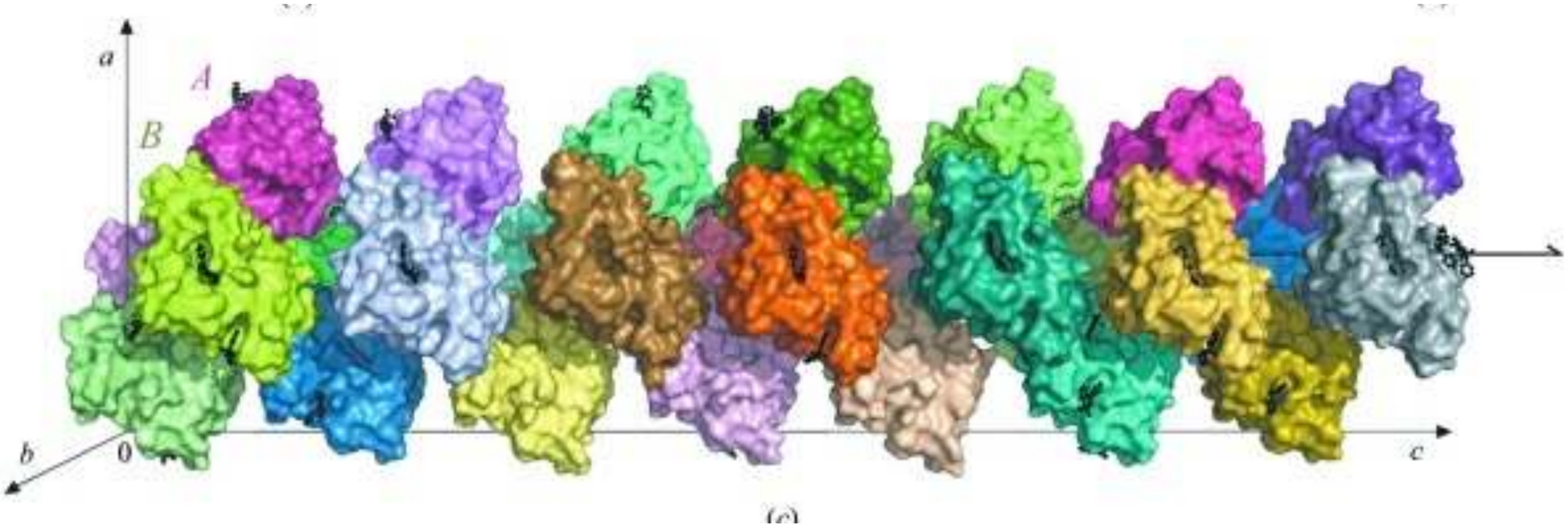
translational non-crystallographic symmetry

- TNCS indicated by Patterson Peak
- 16% origin at 5Å

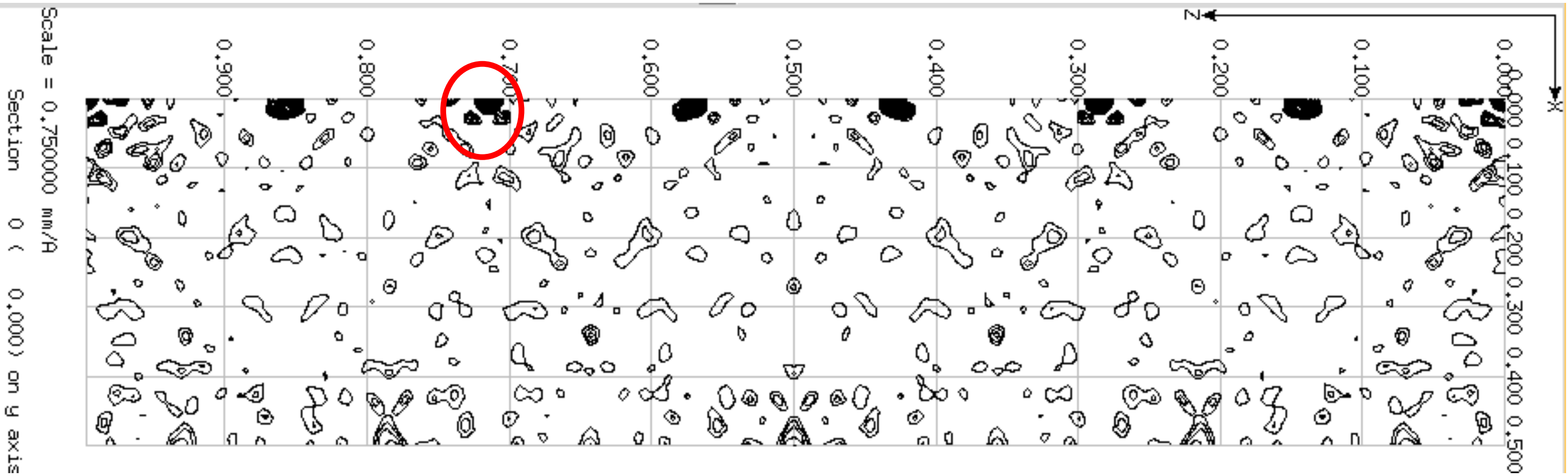


Native Patterson of mouse renin.



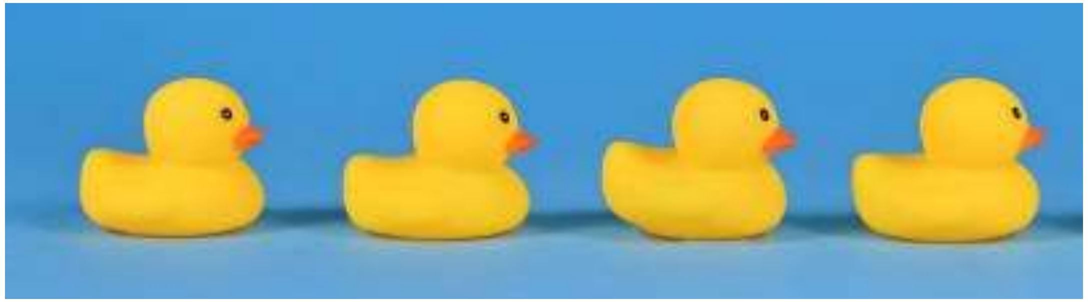


fine structure?



translational non-crystallographic symmetry

- Molecules related by a vector translation
- But the translational symmetry is never exact
  - Differences from mean vector
  - Angular perturbations
- The differences from perfect translation and usually significant
- Can lead to ambiguous tncs
  - when is tncs not tncs?

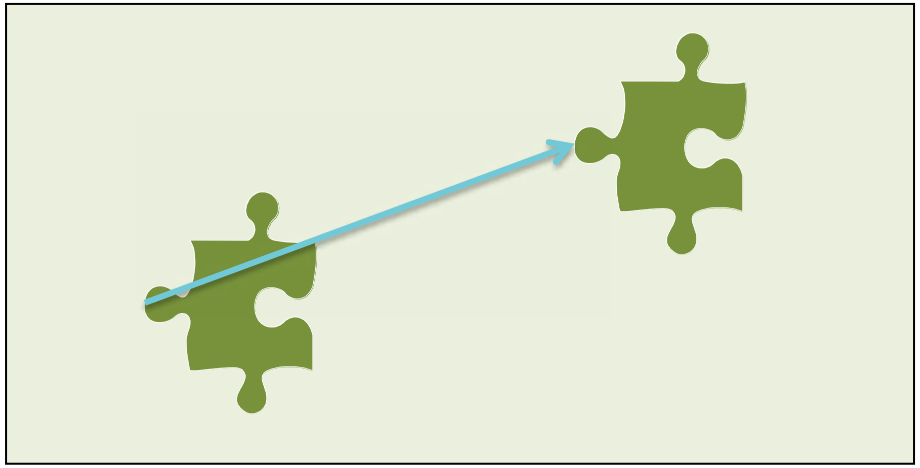


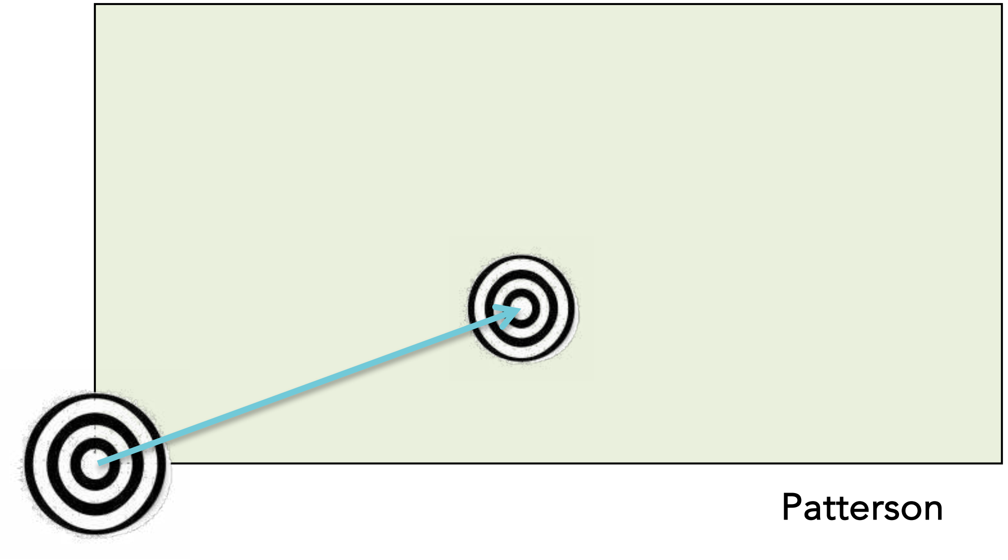
perfect translation

imperfect translation

translational non-crystallographic symmetry

- Molecules related by a vector translation
- Patterson shows a single vector
  - symmetry copies
- Orientation difference and translation vector refined to generate expected intensity factors

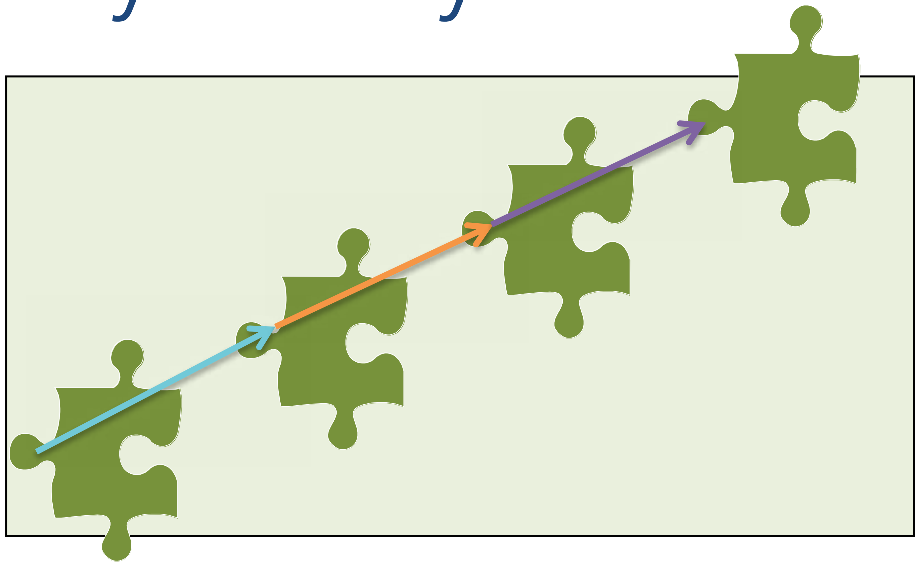


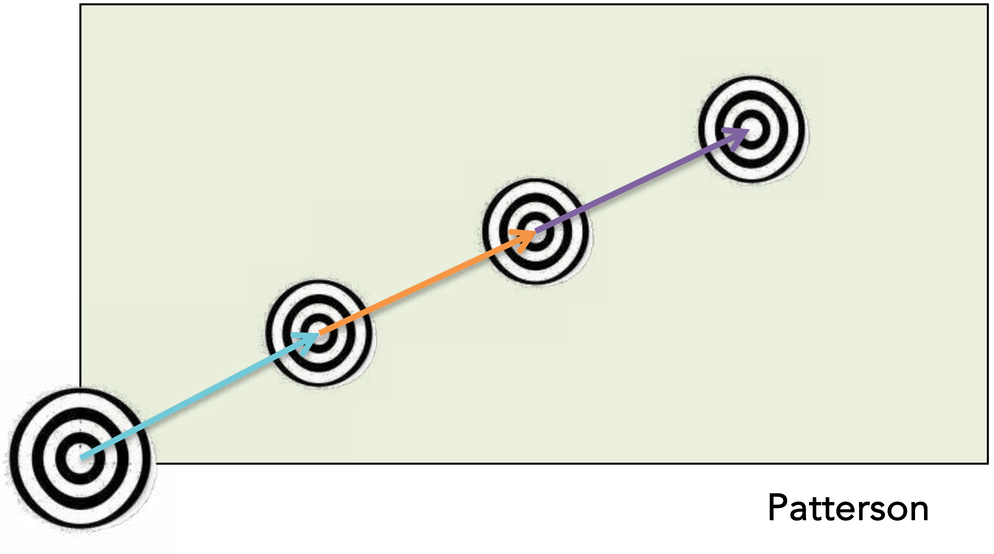


Patterson

translational non-crystallographic symmetry

- Molecules may be related by multiples of same vector
  - this gives a higher 'order' of tncs
- Patterson shows a family of vectors
- This TNCS is often 'commensurate'
  - Vectors are often simple fractions of the cell
  - 'failed' crystal formation with lower cell
    - e.g. (¼,¼,0)
- Only translation vector refined

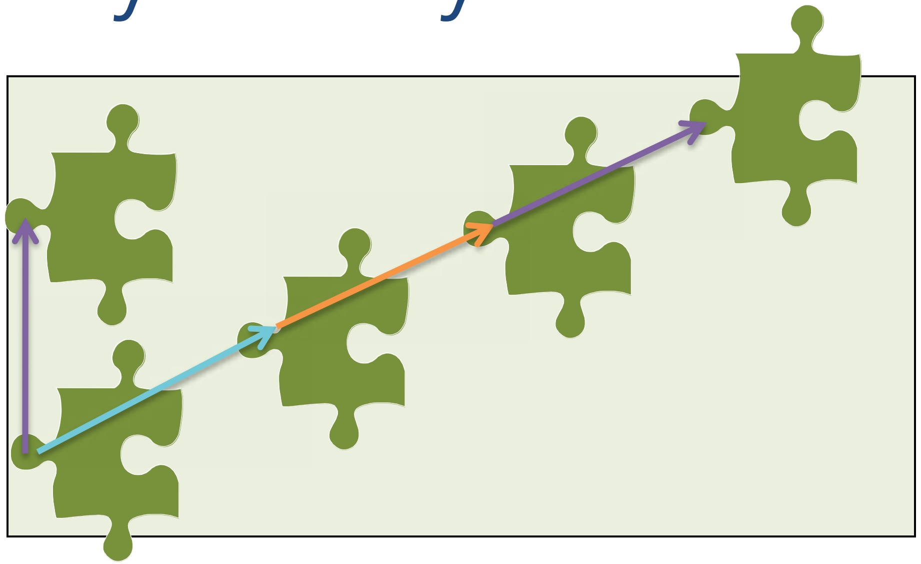


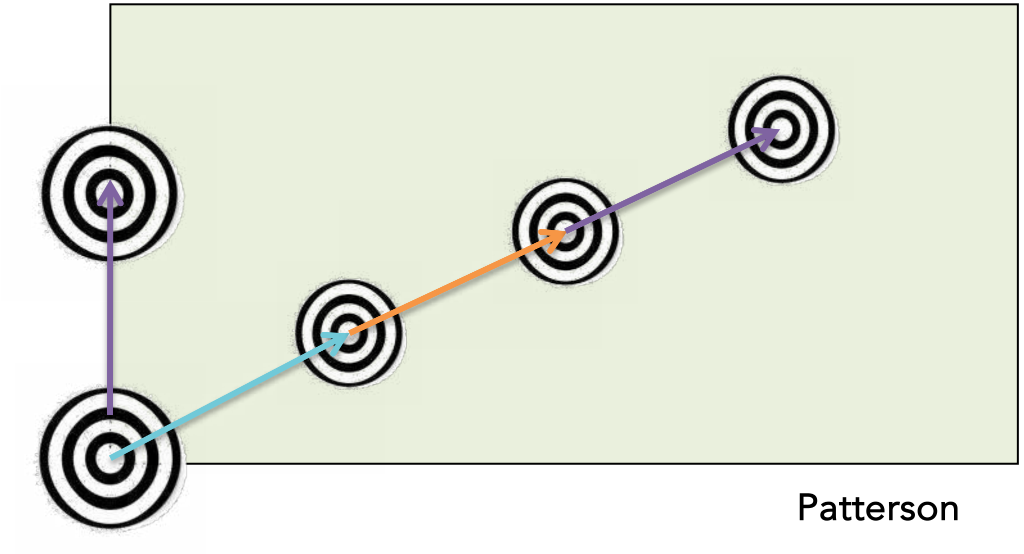


Patterson

translational non-crystallographic symmetry

- Can be more complex than multiples of the same vector
- Halving in different cell directions
  - e.g. (½,0,0), (0,½,0), (½,½,0)
- Break TNCS down into components that can be dealt with separately



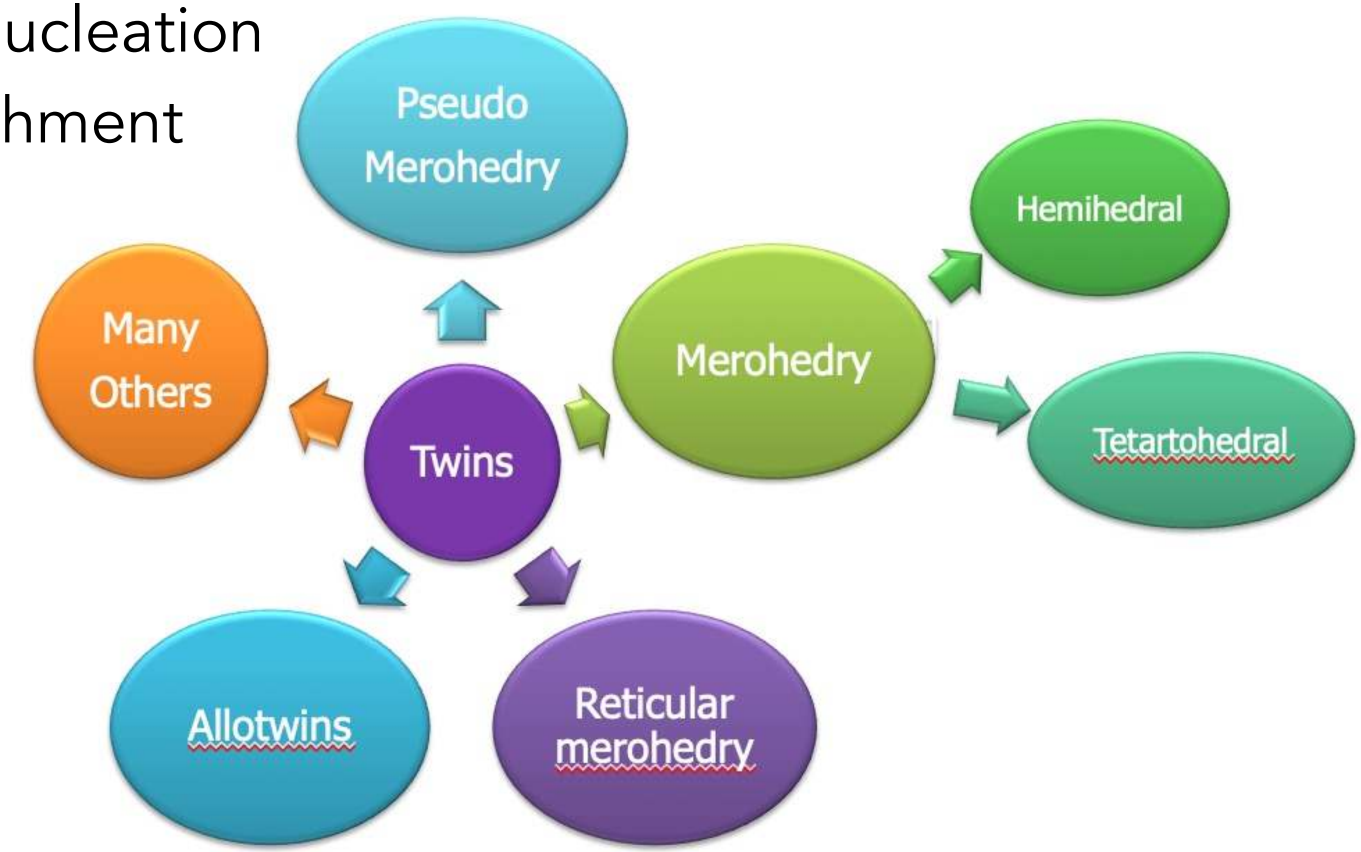


Patterson

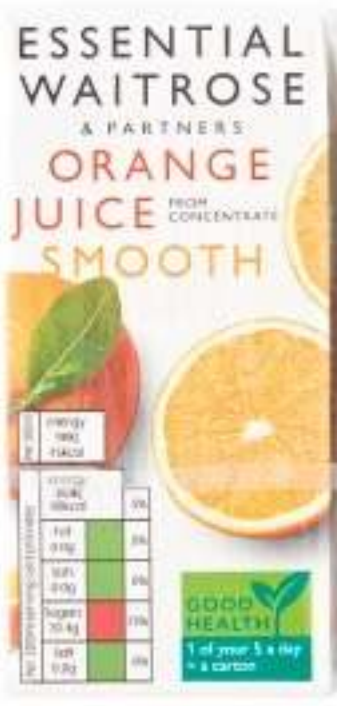
pathologies:
twinning

What is Twinning?

- Twinning is the association of two or more individuals of the same crystalline phase
  - At the early stages of crystal nucleation
  - As post-growth oriented attachment
  - Following a phase transition
  - A mechanical action
- There are many
  - many
    - many
      - many
        - many different types



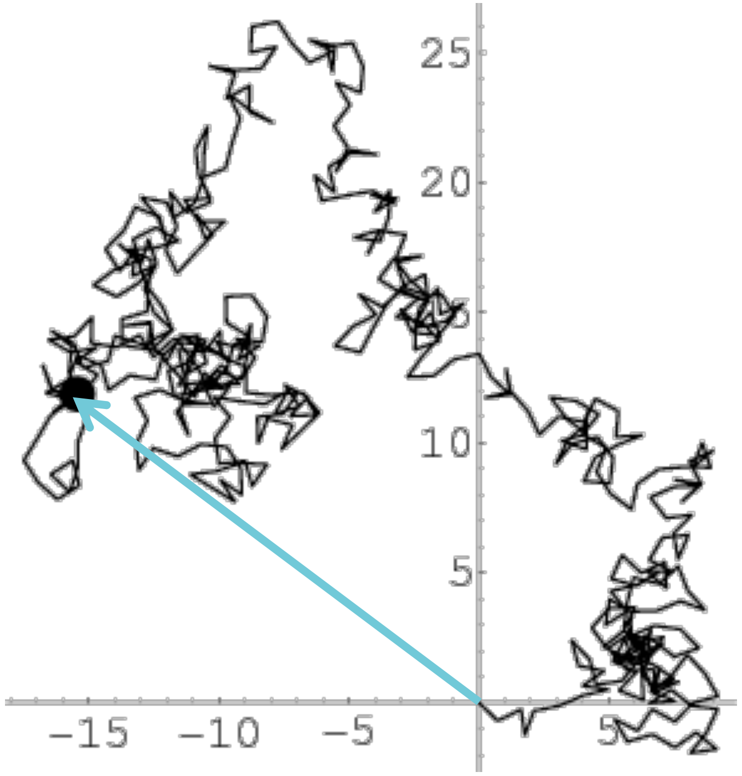
orange juice carton



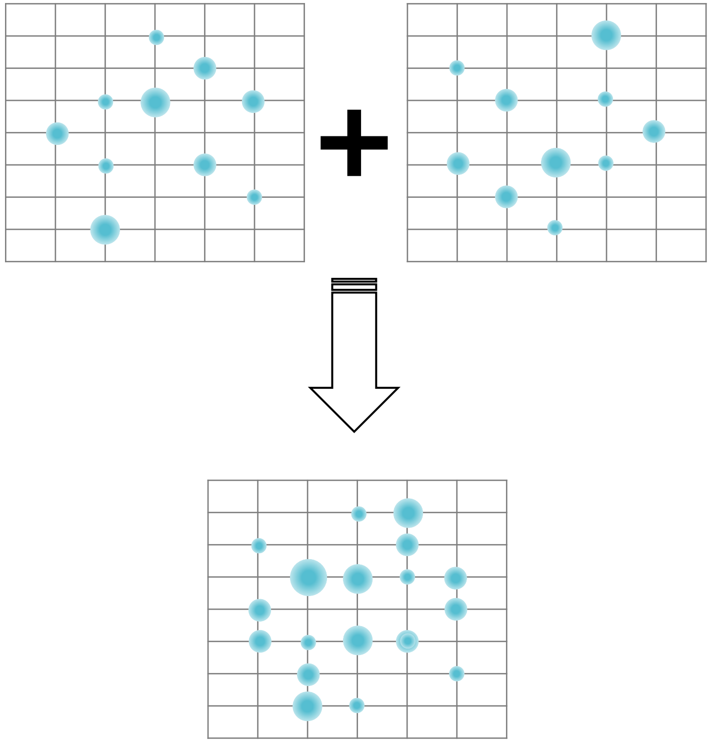
https://elements.envato.com/juice-box-0-5l-slim-KBY64FC

Twinning versus Disorder

Sum of F's



Sum of I's

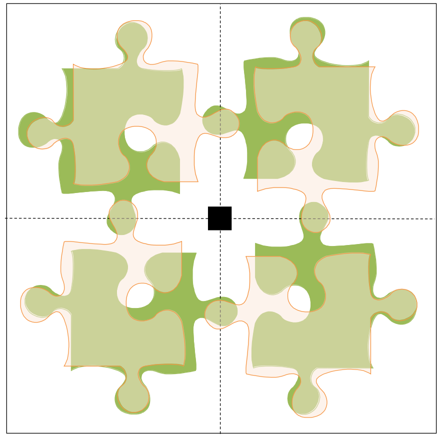


Twinning versus Disorder

- For twinning, the size of twin domains is large compared with the cell dimensions
    - Diffracted X-rays do not interfere
    - Sum of I's not F's
- This is in contrast to disorder in the crystal where the differences are between neighbouring cells
    - Diffracted X-rays represent the spatially and/or temporally averaged content of the unit cells
    - Sum of F's not I's
- There are some nasty intermediate cases

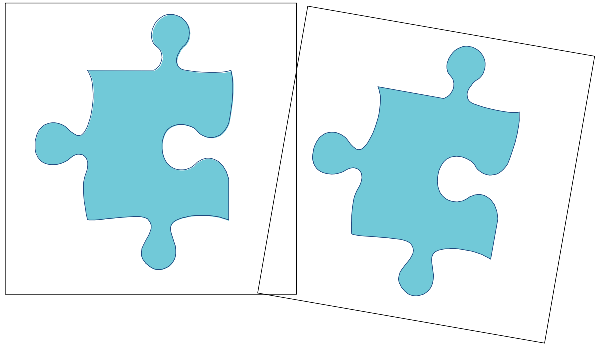
Pathologies

1. Twinning

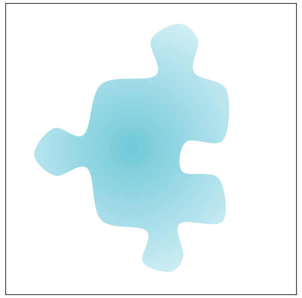


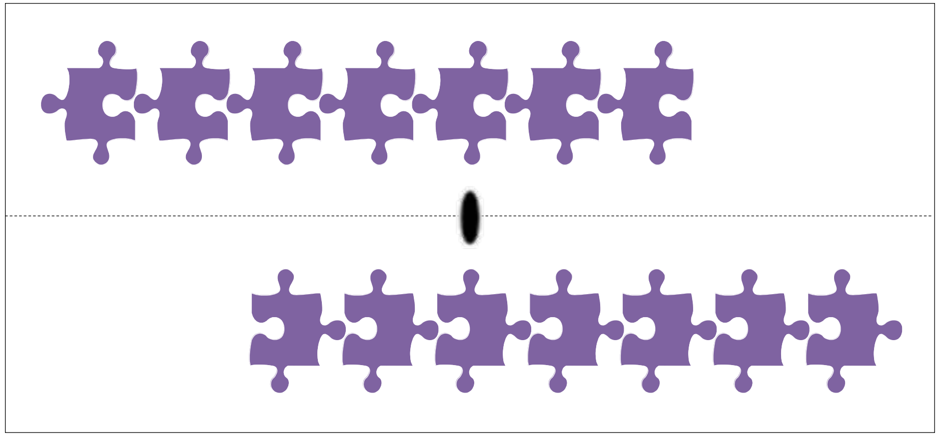
3. Anisotropy

2. High Mosaicity



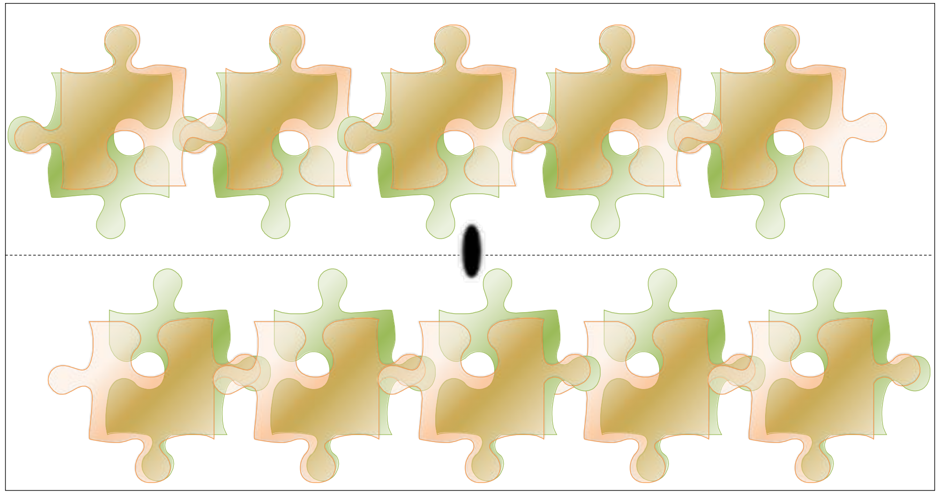
4. Low Resolution





5. Translational non-crystallographic symmetry

Pathologies

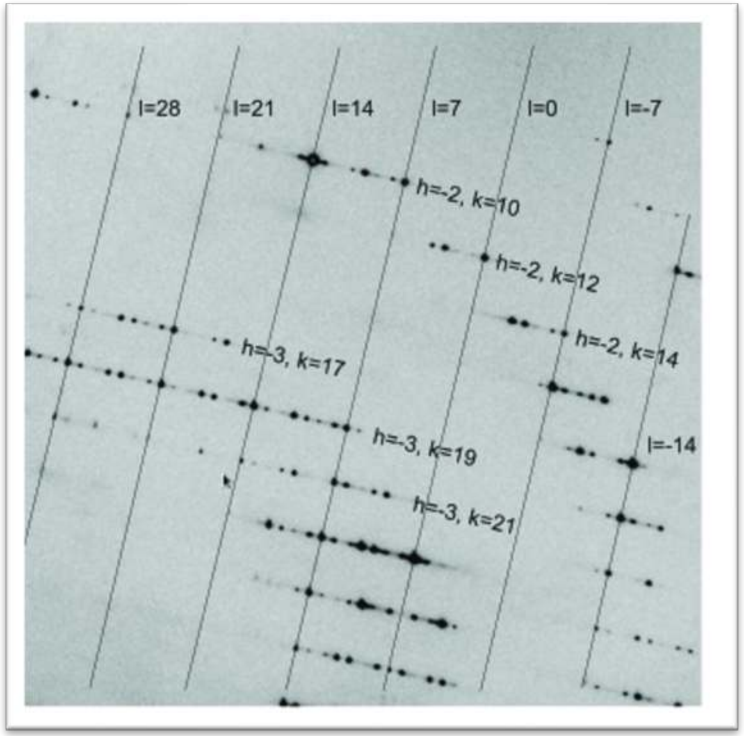


Anisotropy +
Translational NCS +
Twinning

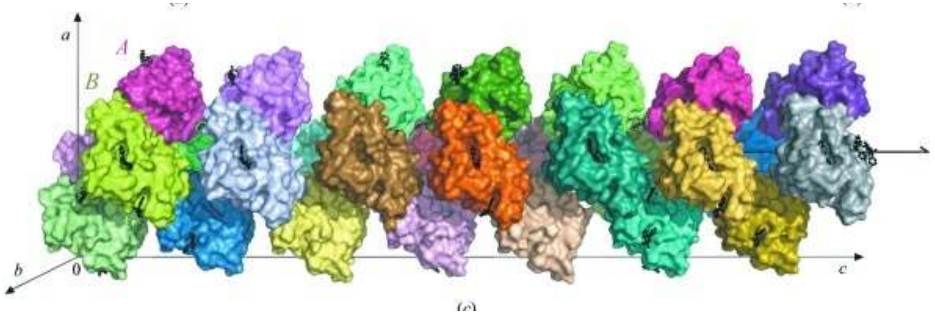
- Translational NCS masks twinning
  - Correcting for translational NCS unmasks twinning
- Phaser gives a P-value for there being twinning in the presence of translational NCS

unmasking twinning

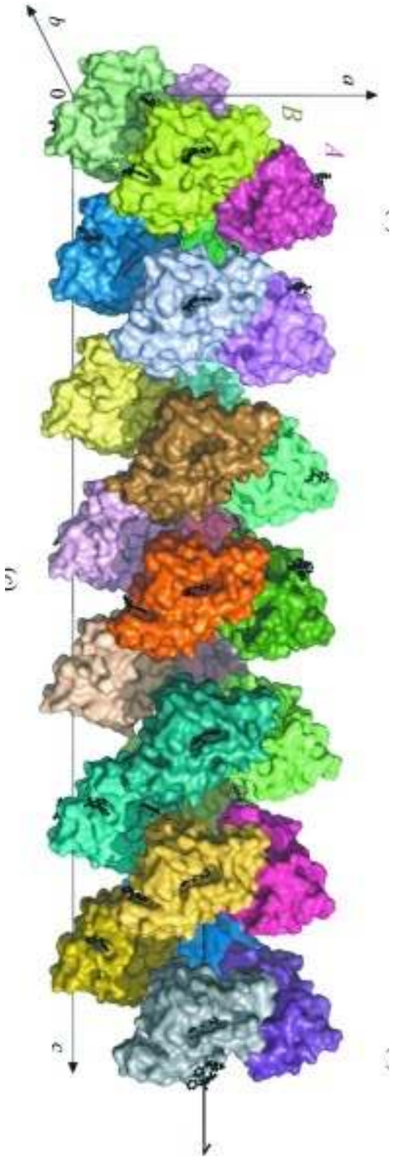
- translational non-crystallographic symmetry masks twinning
- correcting the data for translational non-crystallographic symmetry unmasks twinning
- Twin analysis gives P-value for twinning



Merohedral Tetartohedral Twinning
7-fold TNCS



ANS complex of St John's wort PR-10 protein



- Known monomer structure; good model
- Data merged in P422 with strong $00l \neq 4n$ reflections
  - i.e. $4_1/4_3$ screw axis 'excluded'
- MR in all space groups consistent with point group 422 yielded solutions in space group $P4_122$
  - i.e. with 4_1 screw axis included
- Crystal was tetartohedrally twinned with 7-fold TNCS
- MR was performed successfully in P1 searching for 56 copies
- The correct space group C2 was found by analysing the symmetry of the calculated structure factors
- The MR solution in C2 was obtained by searching for four copies of the first set of seven molecules from the P1 solution

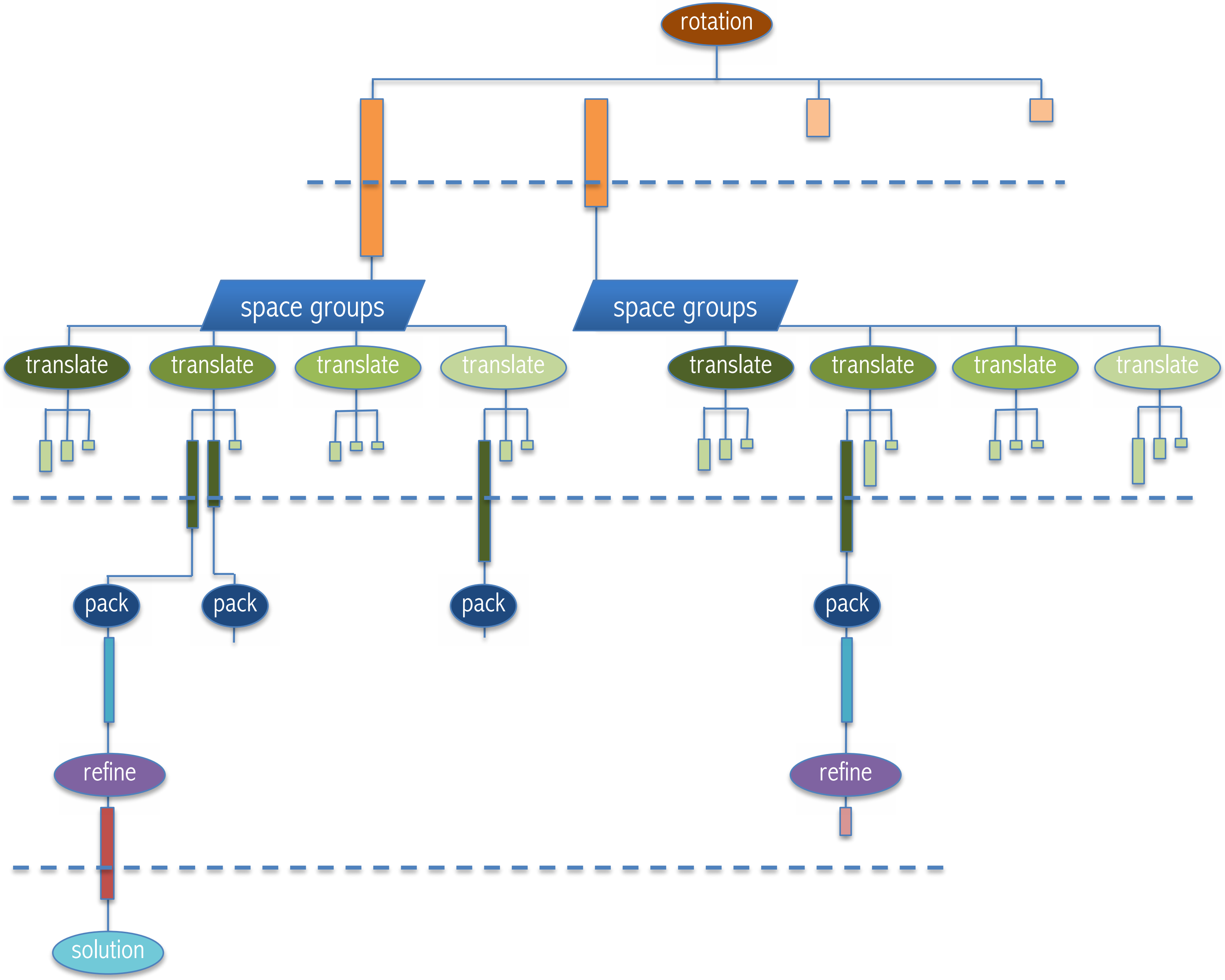


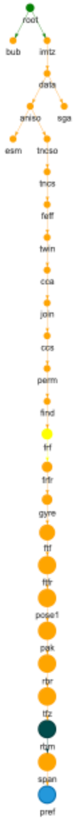
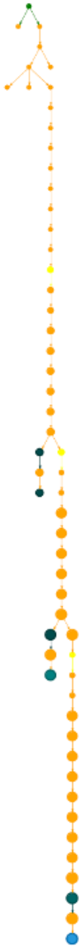
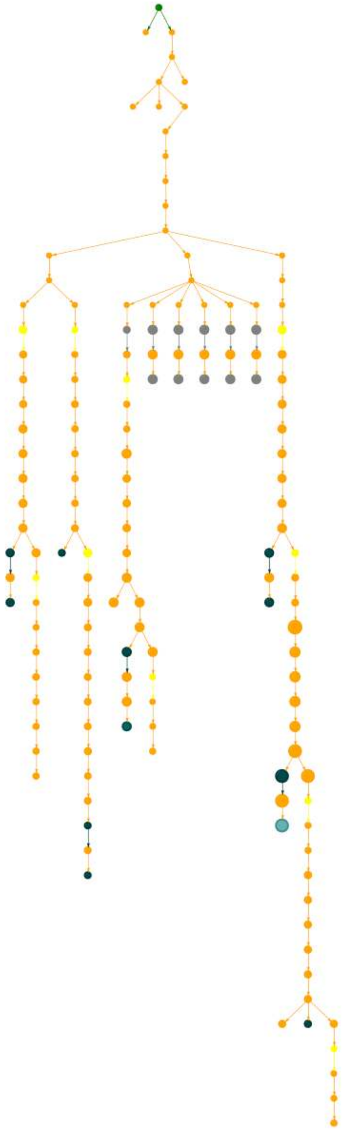
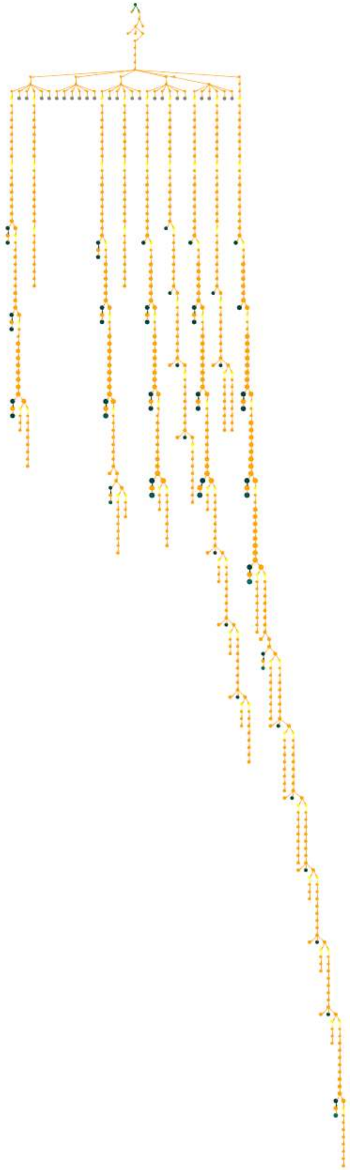
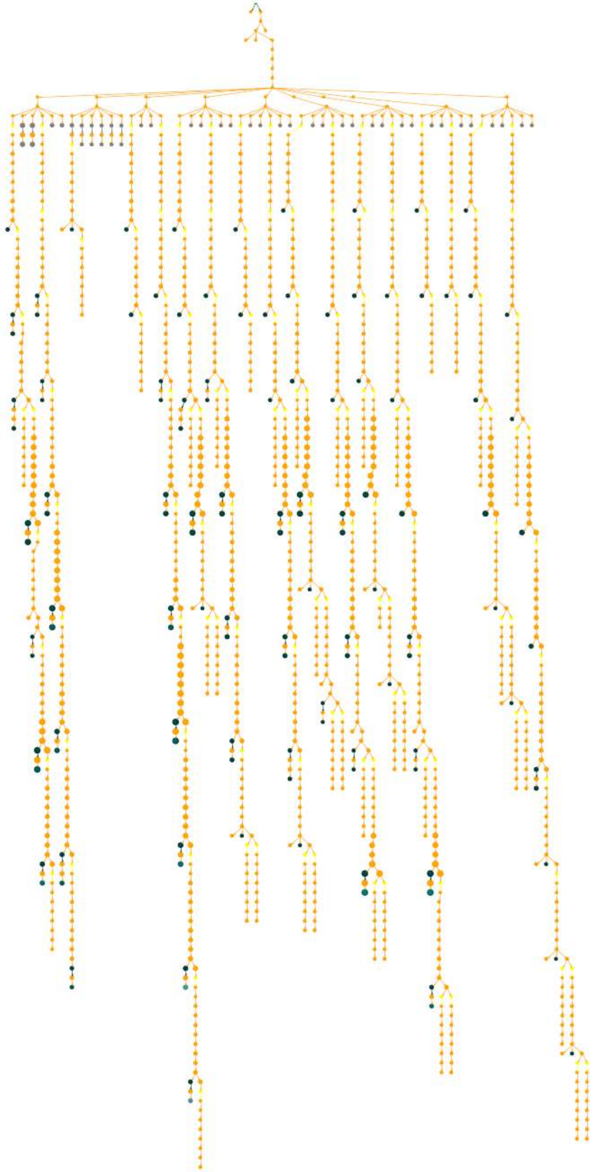
Diagram: Martin Noble

summary –
molecular replacement search dependency

- Biological unit (complex) components
- Number of copies in the asymmetric unit
- How the components are broken into ‘rigid’ fragments
    - conformational change
    - resolution
- Search order
- Translational non-crystallographic symmetry corrections
- Space group translational elements (=systematic absences)
- Space group subgroups in the presence of twinning
- Judge ‘correctness’ by comparing with other options=BEST

phaser-voyager

automates
increasingly
complicated
search strategies

7EJG	7E8J	7ETY	7EWJ	7V9G
				

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Dorothee Liebschner, Nigel Moriarty, Billy Poon, Christopher Schlicksup, Oleg Sobolev

University of Cambridge

Randy Read, Airlie McCoy, Alisia Falini

An NIH/NIGMS funded Program Project

New Mexico Consortium
Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Baylor College of Medicine

Matt Baker

Duke University

Jane & Dave Richardson, Vincent Chen, Michael Prisant, Christopher Williams

Liebschner et al., Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst.* 2019 D75:861-877