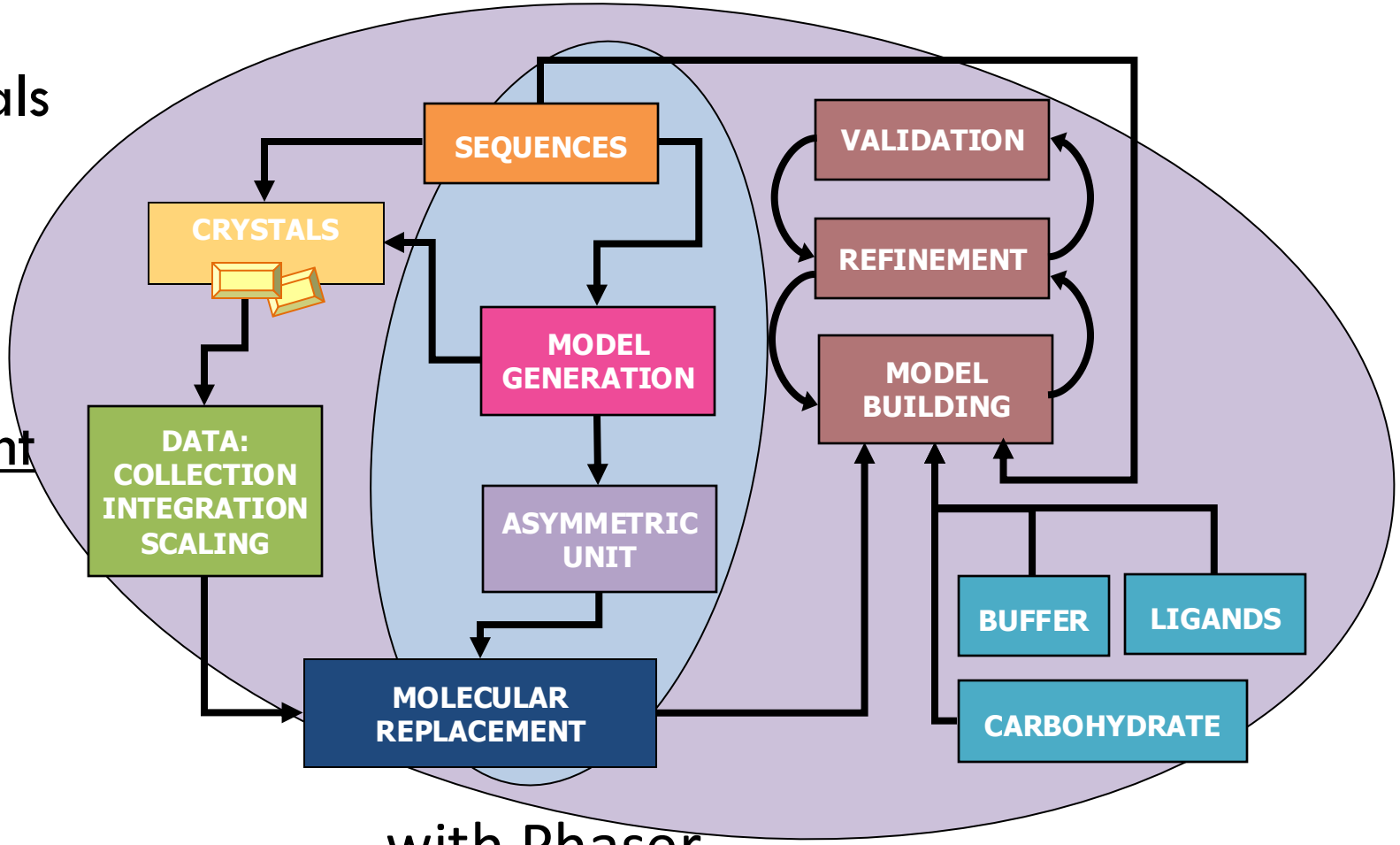
Molecular replacement

Airlie McCoy

Molecular Replacement

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



with Phaser

- some things will be general to molecular replacement

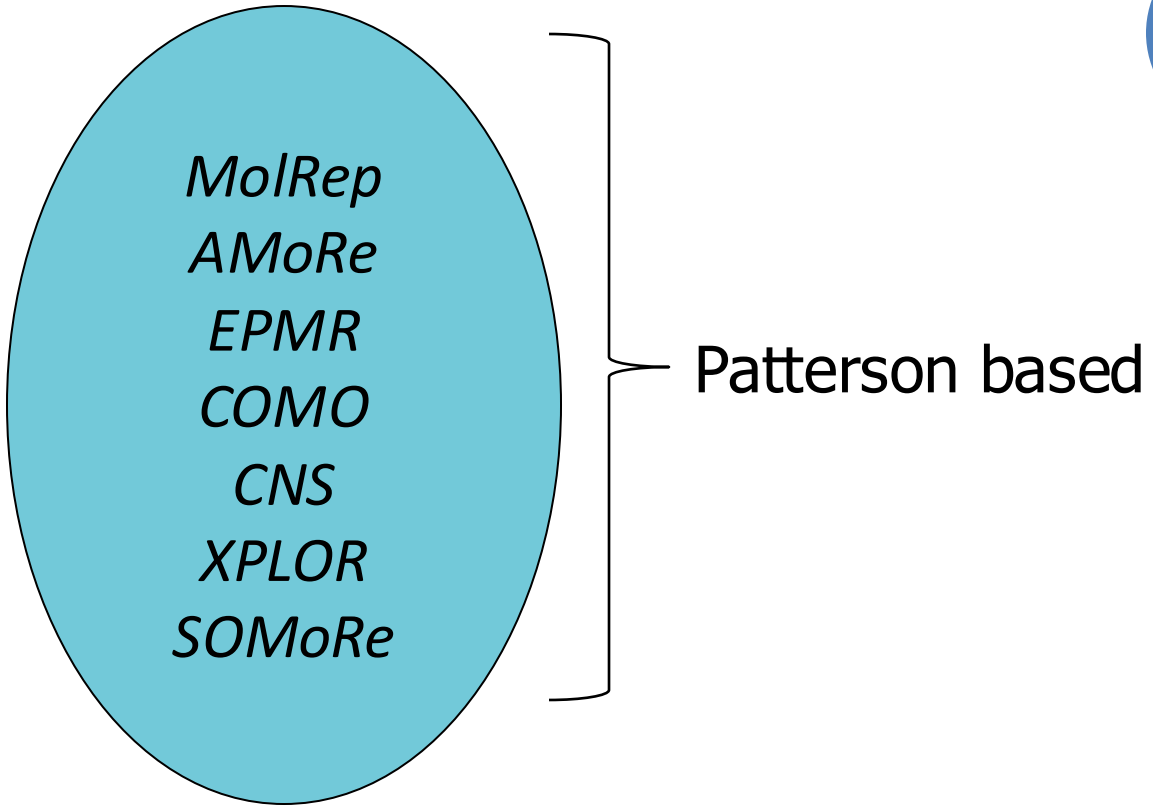
and

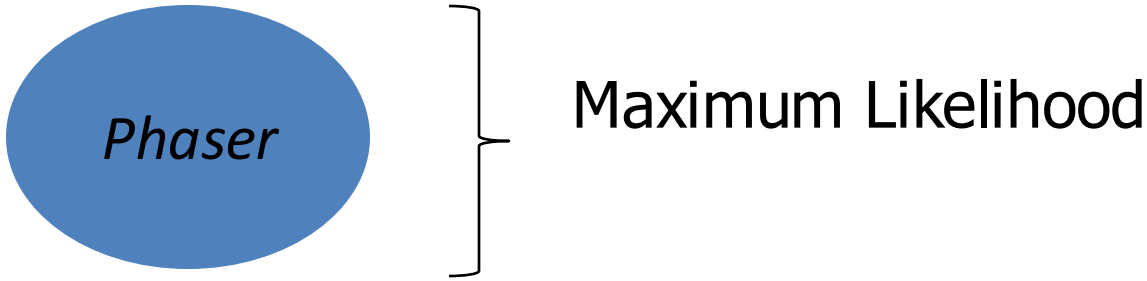
- some things specific to molecular replacement with

molecular replacement

- molecular replacement is a method used to solve the phase problem for a new crystal
- it uses a known, structurally similar molecule (or molecules) called the 'model'
- the initial phases are only approximately correct
- phases are improved by refinement
- the model must be placed within the lattice of the new crystal
  - using what is known about the unit cell, space group and crystallographic intensities (data)

programs for molecular replacement





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")

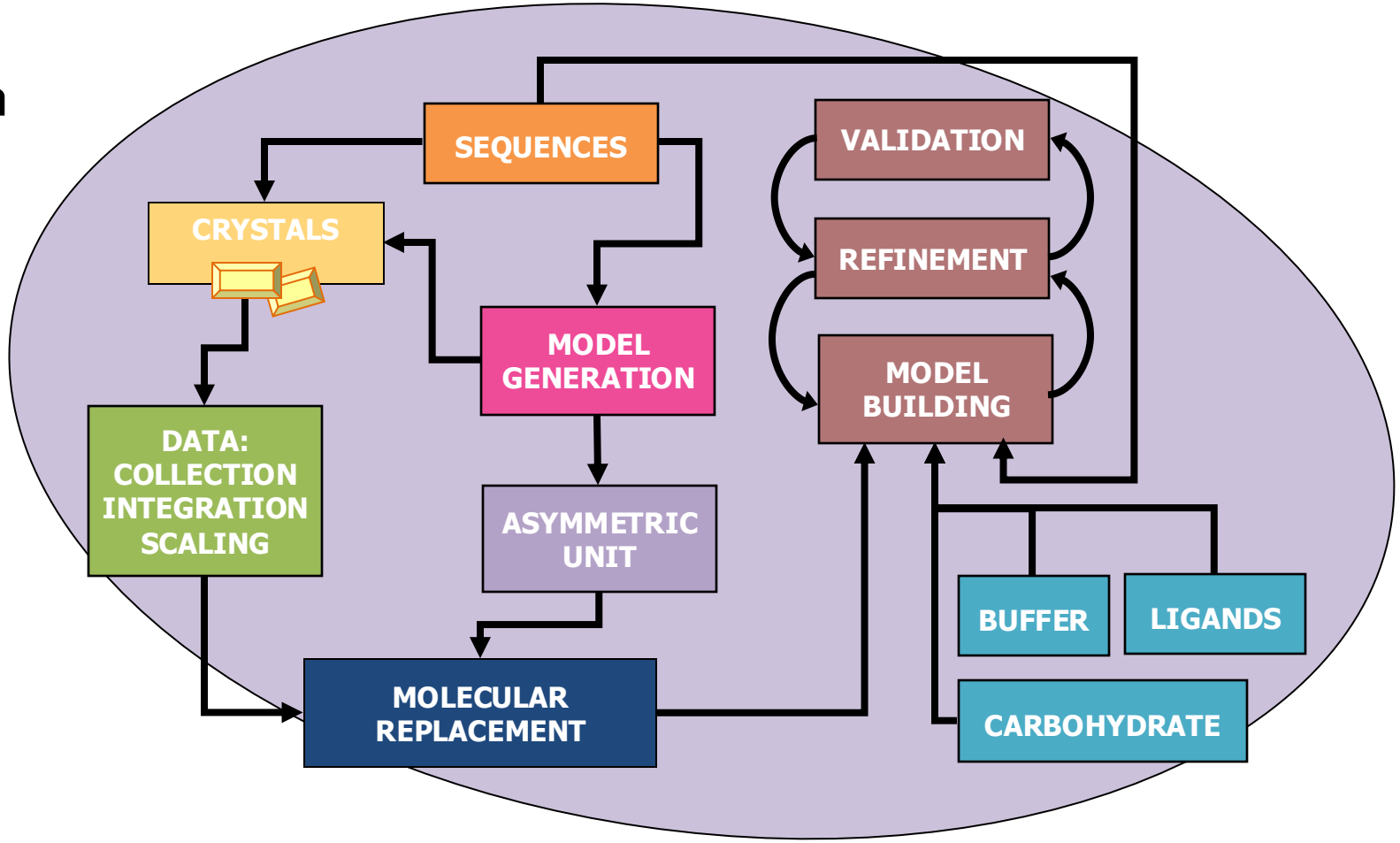
- Distribution
  - CCP4
  - Phenix

UNIVERSITY OF CAMBRIDGE

Cambridge Institute for Medical Research

Phaser and pipelines

- Phaser performs MR in
  - Arcimboldo
  - MrBUMP
  - Balbes
  - Ample
  - CCP4 cloud
  - phenix.mr_rosetta
  - phenix.automr
  - phaser.MRage
  - MRGrid

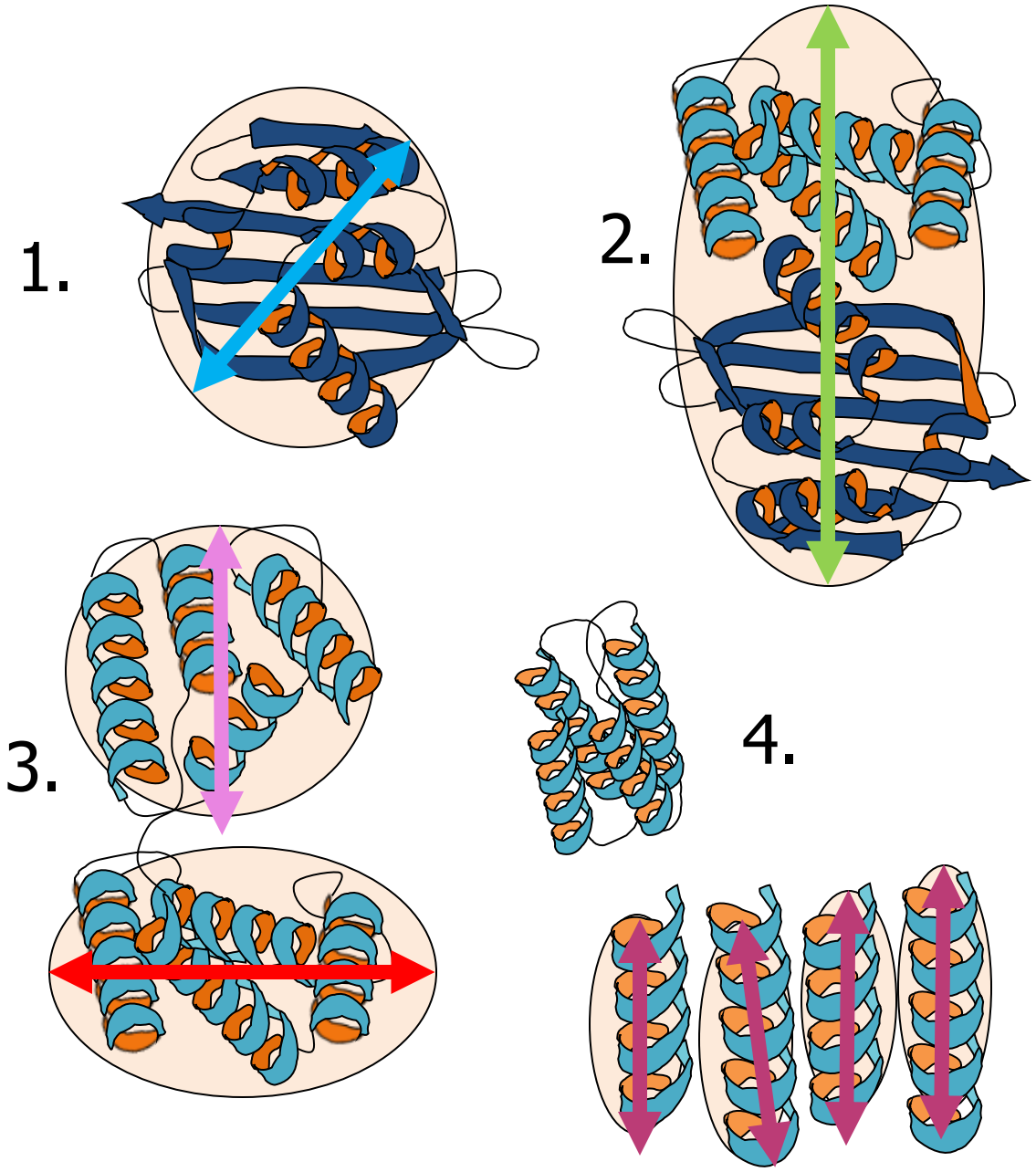


- Phaser is also often run independently of pipelines

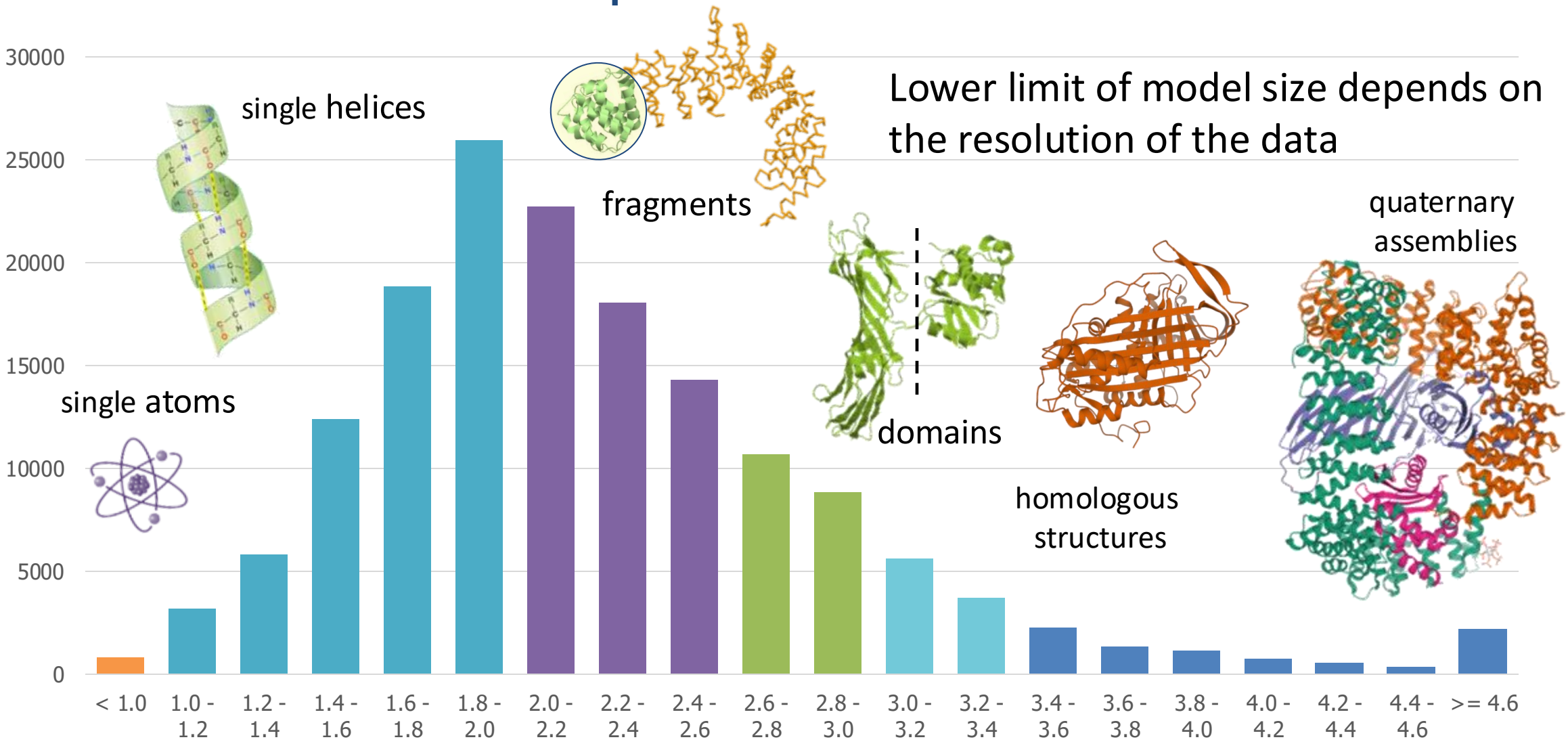
models

”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
- NMR structures and alphafold models tend to have poorer long- range accuracy
- You don’t know (for certain) in advance whether the model meets this criteria… working ‘blind’



models for molecular replacement



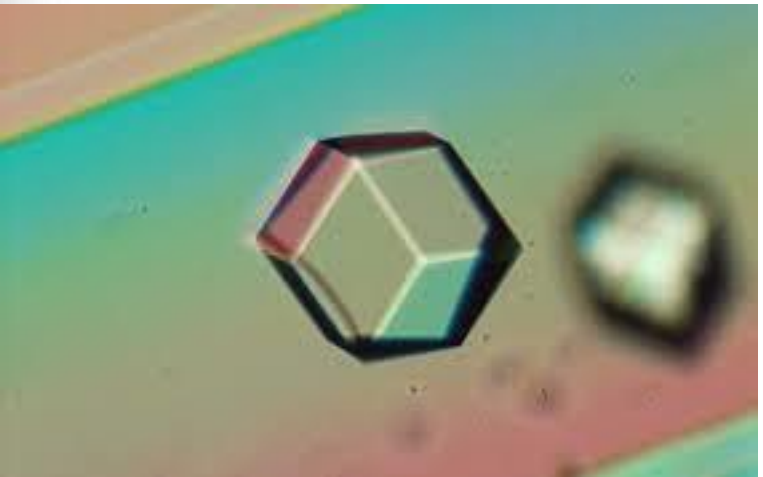
phasing by molecular replacement

File Edit Calculate Draw Measures Validate About Ligand

Reset View | Display Manager | calpha all | atoms all | symm all | symm off | Patterson

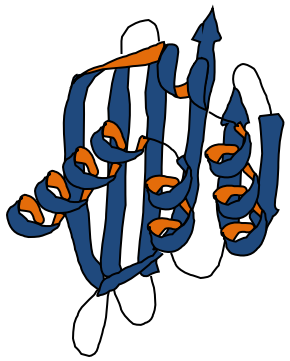
R/RC

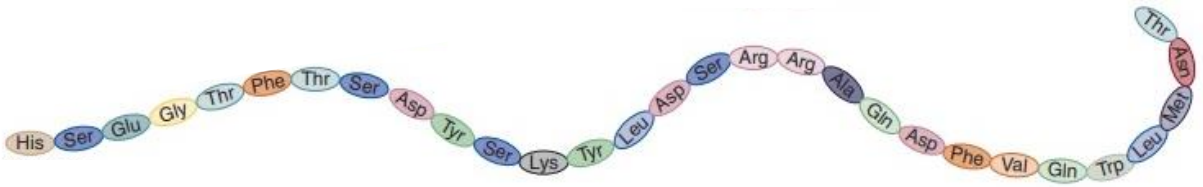
Map



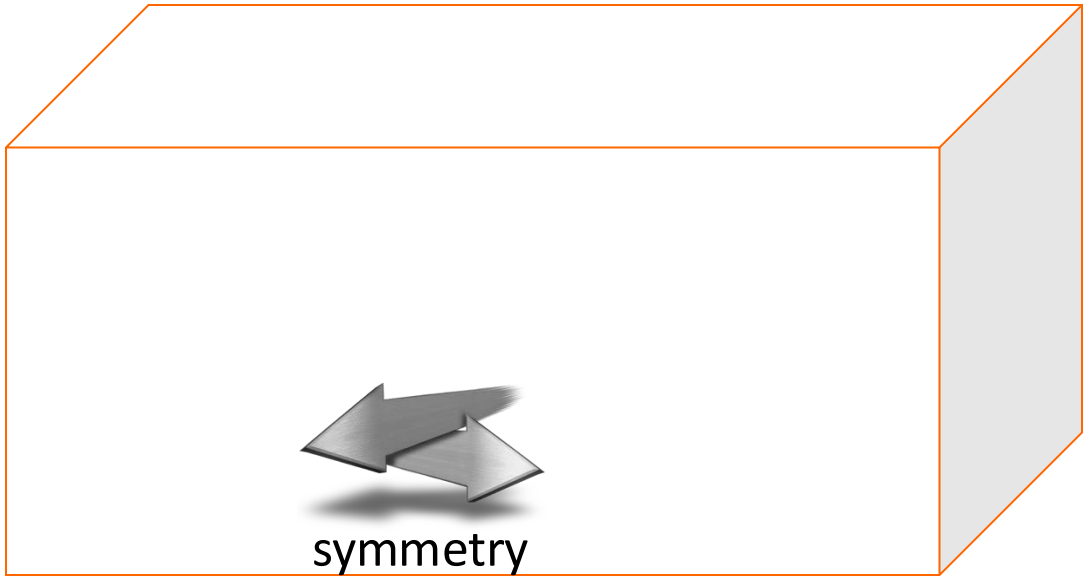
molecular replacement

requires a model structure
with low rmsd to the target





Known
sequence

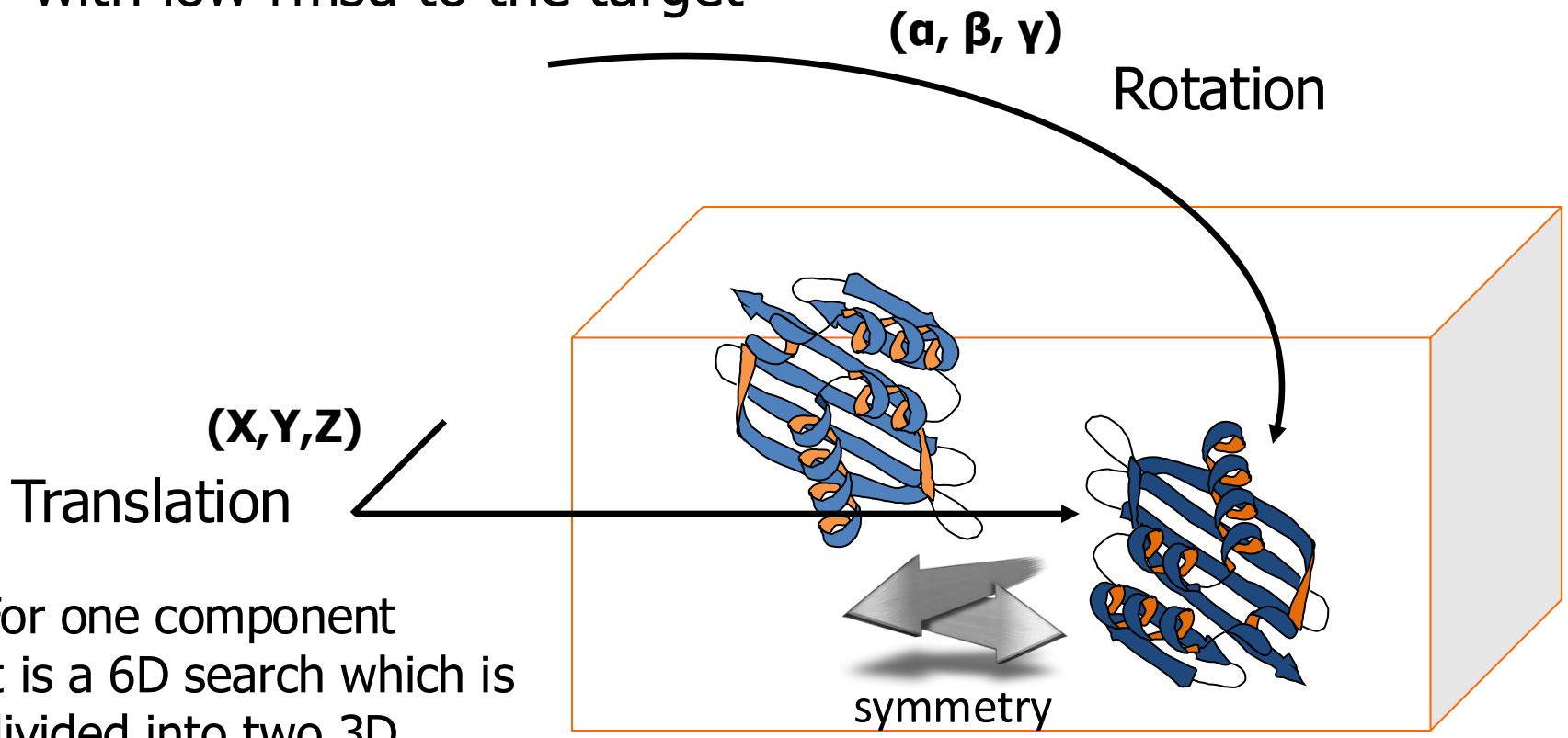


Unknown
crystal
structure

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

molecular replacement – rotation and translation

requires a model structure
with low rmsd to the target

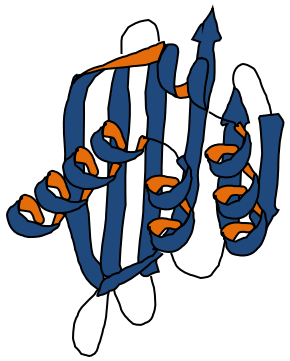


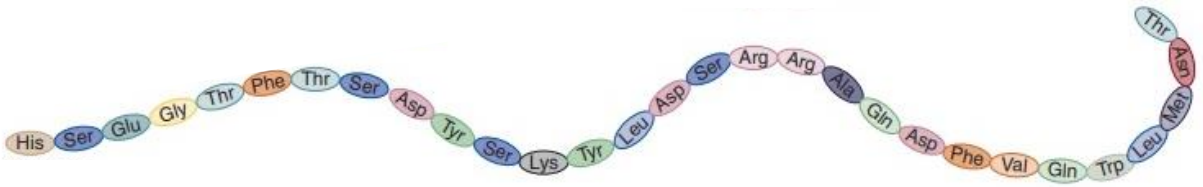
For one component
it is a 6D search which is
divided into two 3D
searches
- rotation
- translation

Unknown
crystal
structure

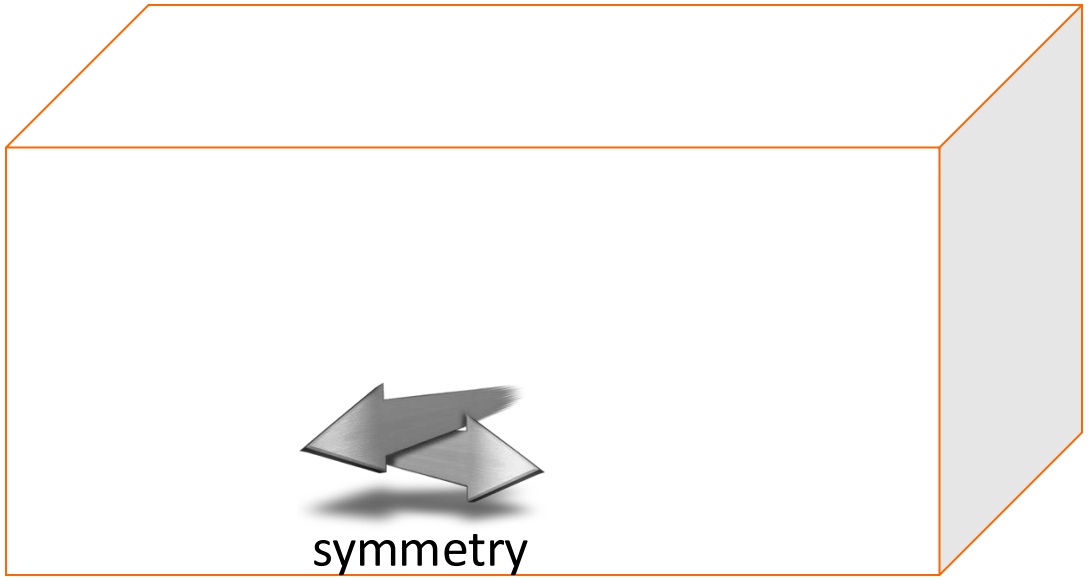
molecular replacement

requires a model structure
with low rmsd to the target





Known
sequence



Unknown
crystal
structure

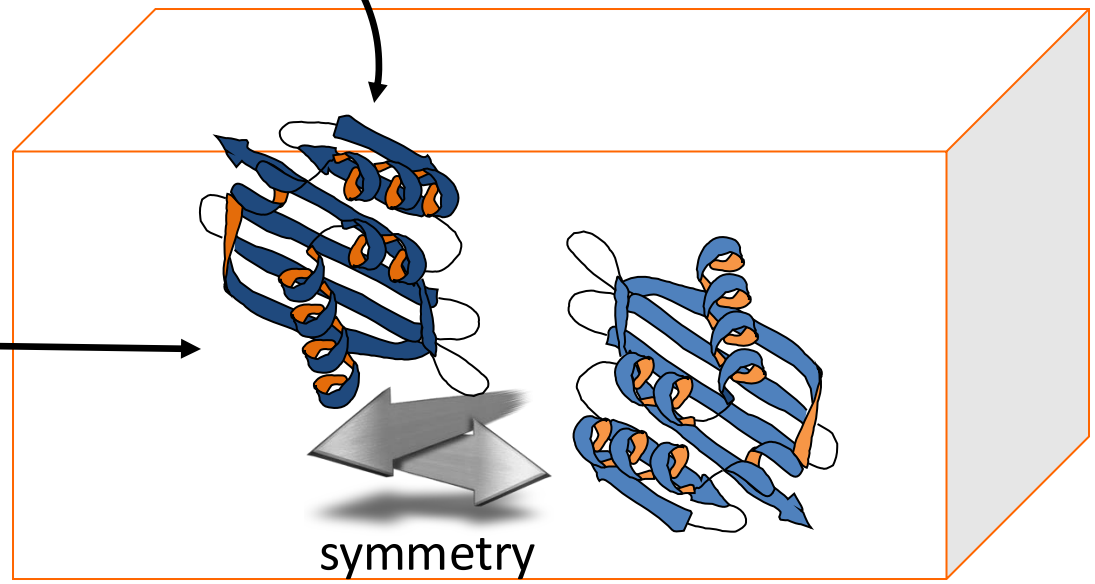
You can 'solve' the molecular replacement problem by finding any one of the (symmetry related) copies in the crystal lattice

molecular replacement – rotation and translation

requires a model structure
with low rmsd to the target

(α, β, γ)

Rotation



(X,Y,Z)

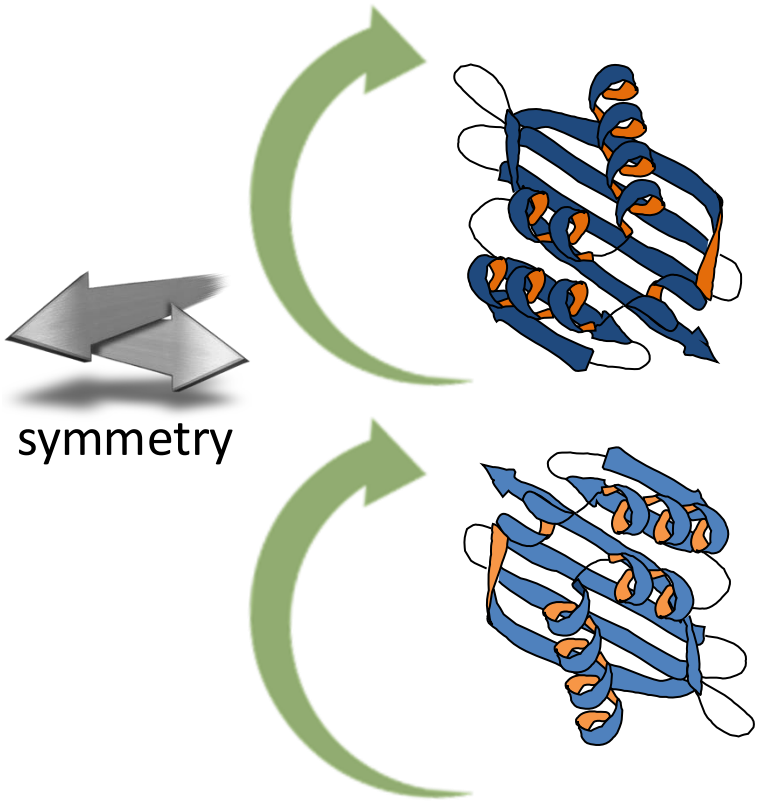
Translation

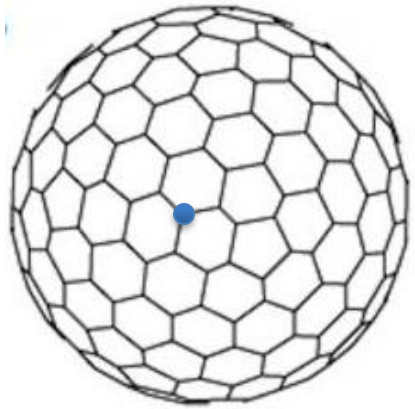
You can 'solve' the molecular replacement problem by finding any one of the (symmetry related) copies in the crystal lattice

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

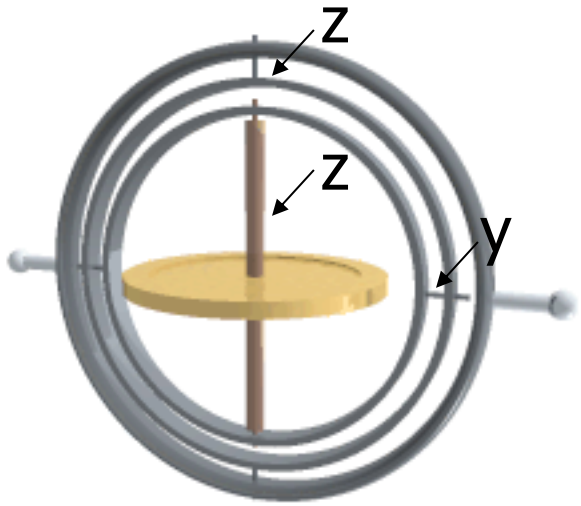




LLG
log-likelihood gain

Euler Angles*

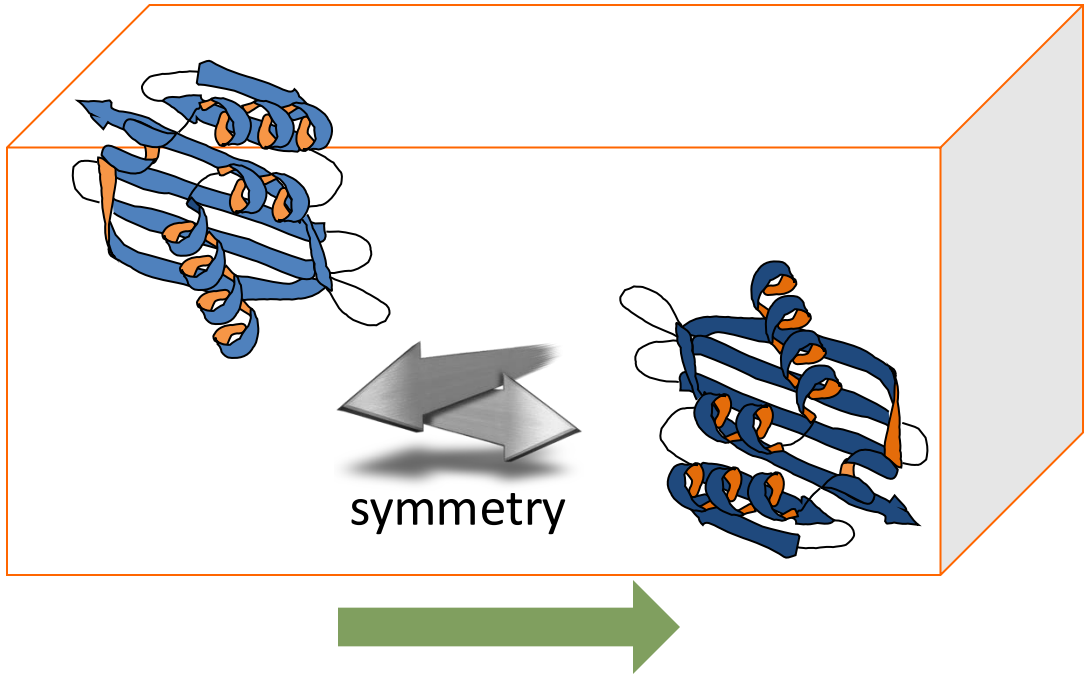
- Phaser rotation angles are reported as Euler angles (α, β, γ)
  - α is rotation about the first axis
  - β is rotation about the second axis
  - γ is rotation about the third axis
- In crystallography, all programs report Euler angles with respect to intrinsic "z,y,z" axes
- You <u>do not need</u>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

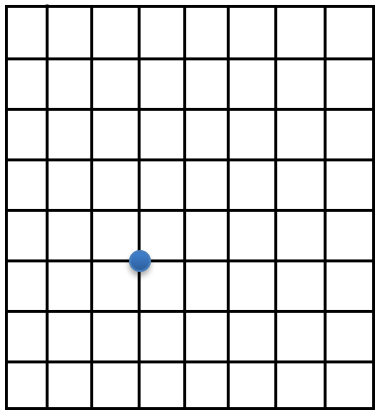


*This slide is for information only

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

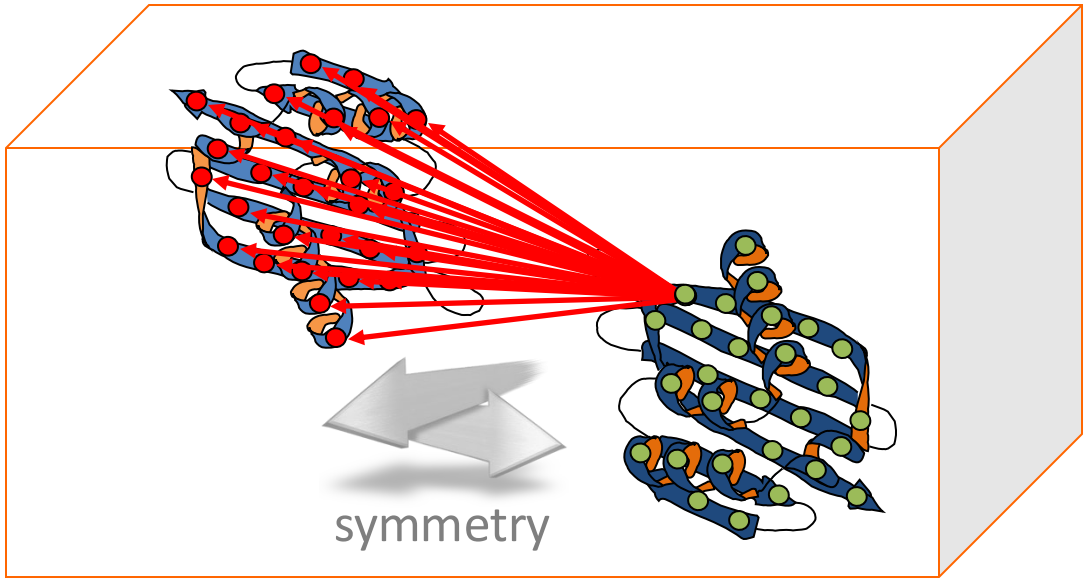




LLG
log-likelihood gain

packing analysis (packing function)

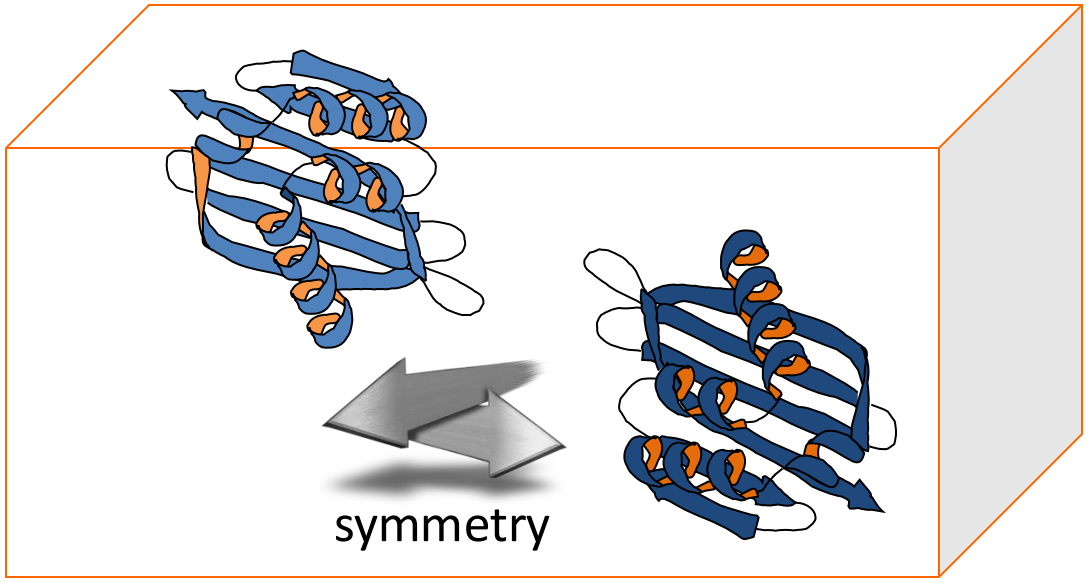
- Cα clash test
- Excludes physically impossible poses, reducing search space

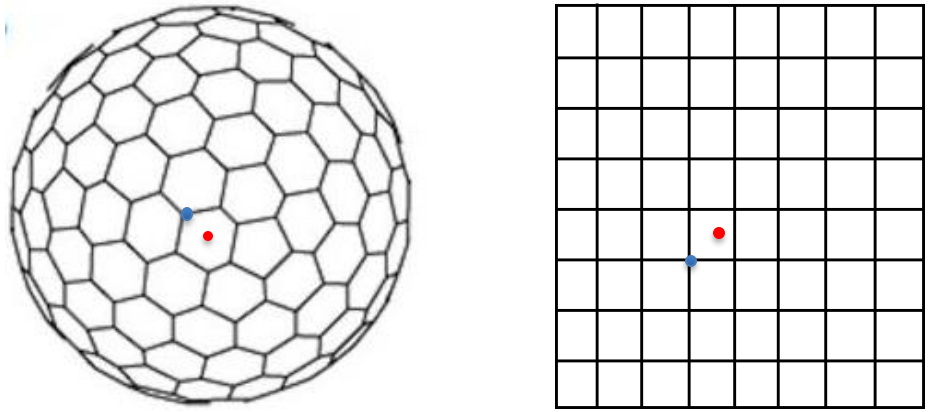


Clash with other contents
of unit cell

refinement

- Optimize poses away from grid search locations





LLG
log-likelihood gain

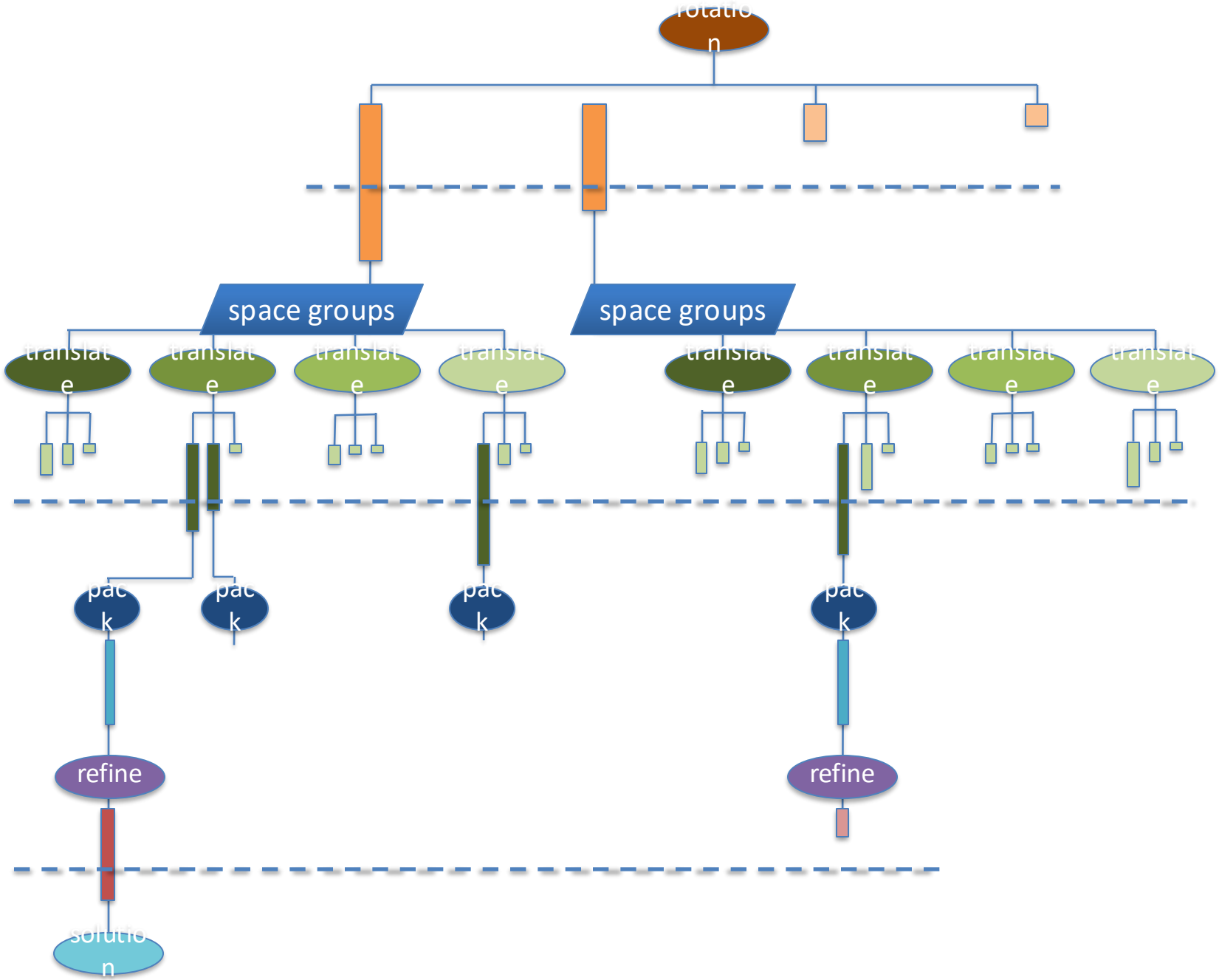
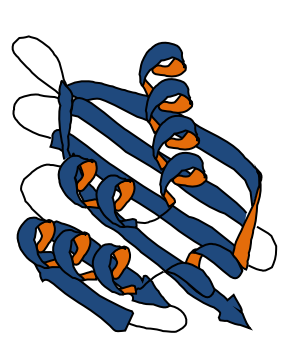
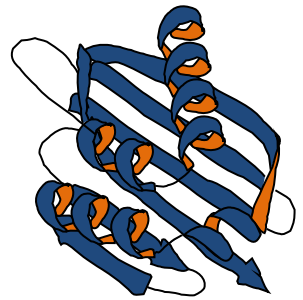


Diagram: Martin Noble

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!



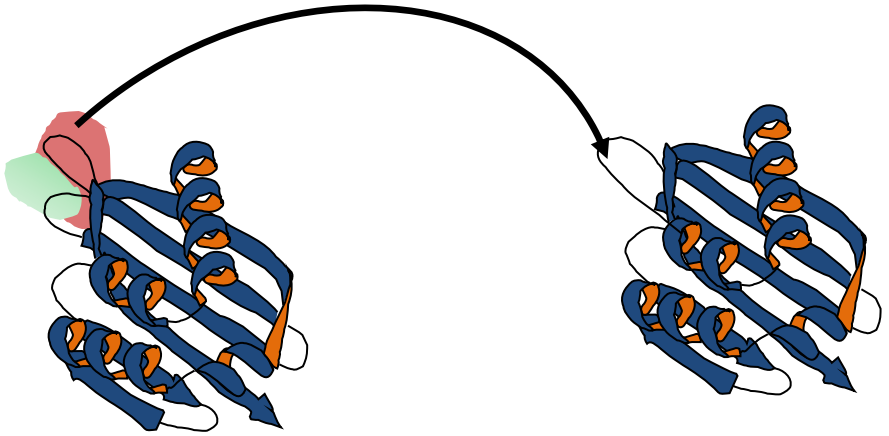


H K L	F	ϕ
0 0 1	12.6	120
0 0 2	2.1	10
0 0 3	69.9	280
etc...		

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



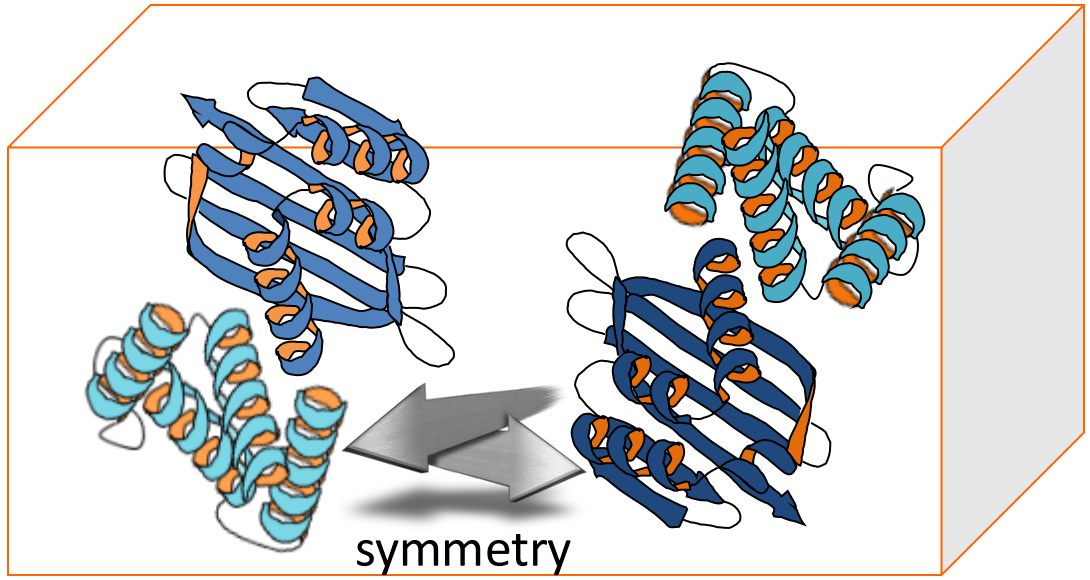
H K L	F	ϕ
0 0 1	12.6	120
0 0 2	2.1	10
0 0 3	69.9	280
etc...		

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

two molecules…

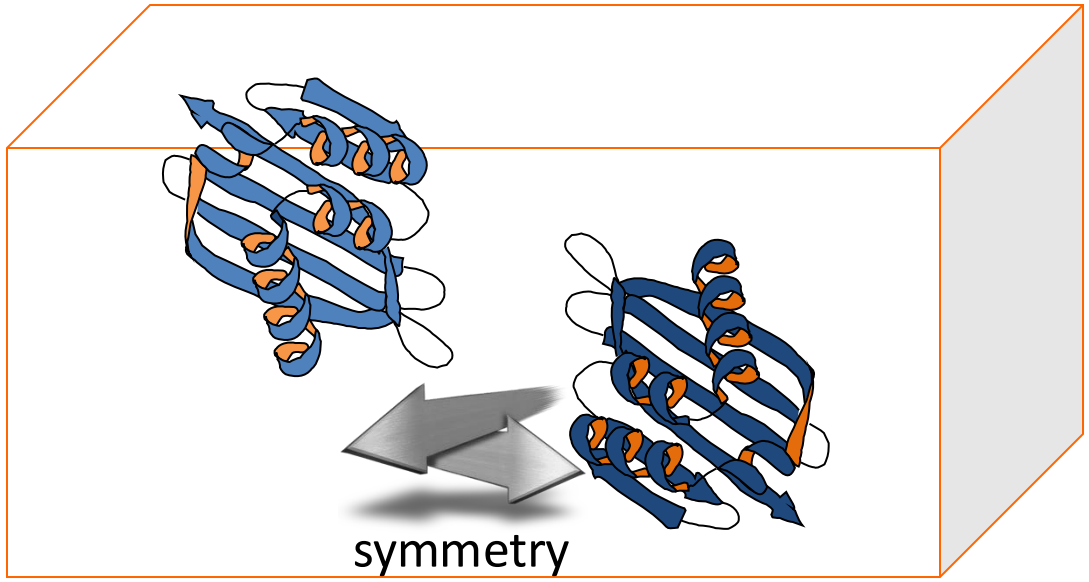
molecular replacement

- Protein complex in the asymmetric unit*

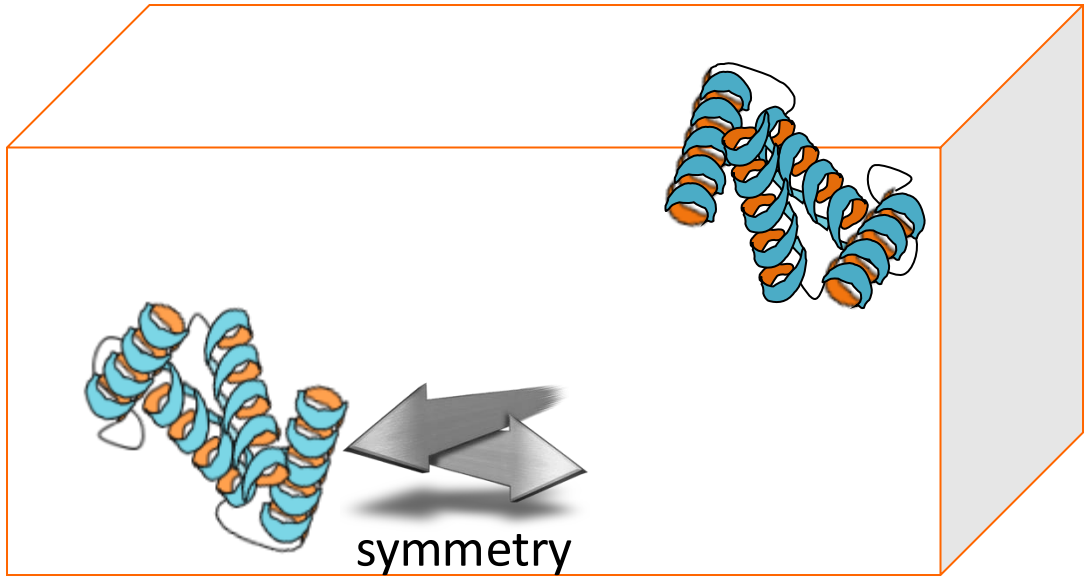


molecular replacement

- Protein complex in the asymmetric unit



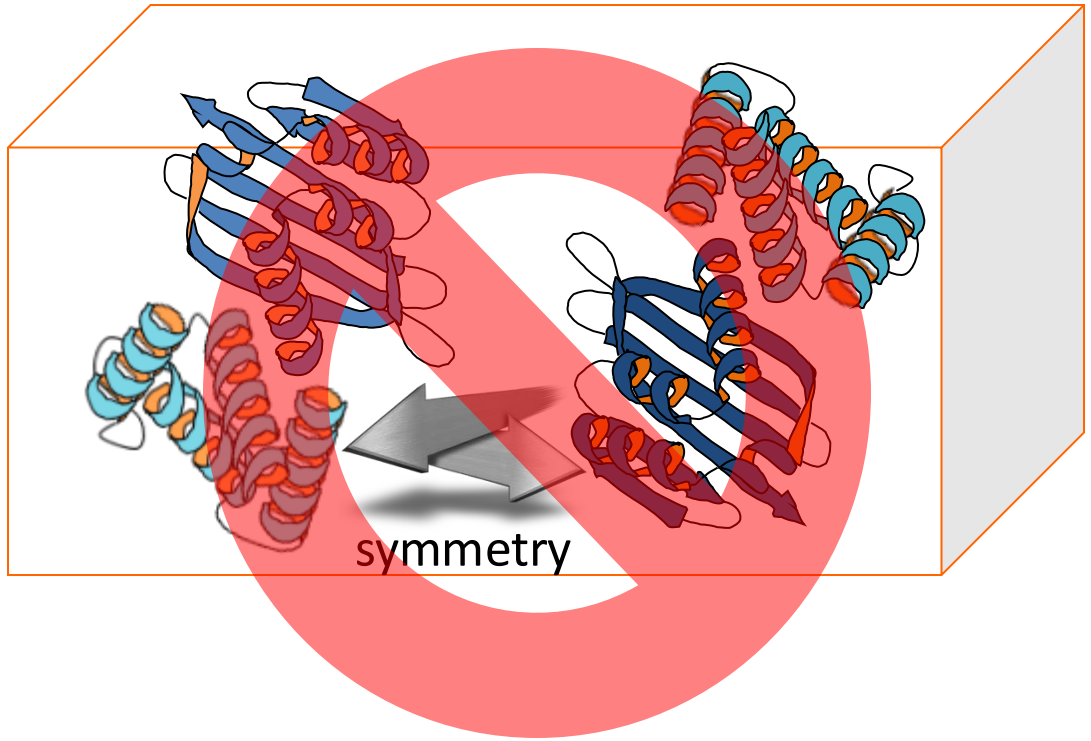
- *Could* search for the two components separately
  - and merge solutions afterwards



molecular replacement

- Protein complex in the asymmetric unit

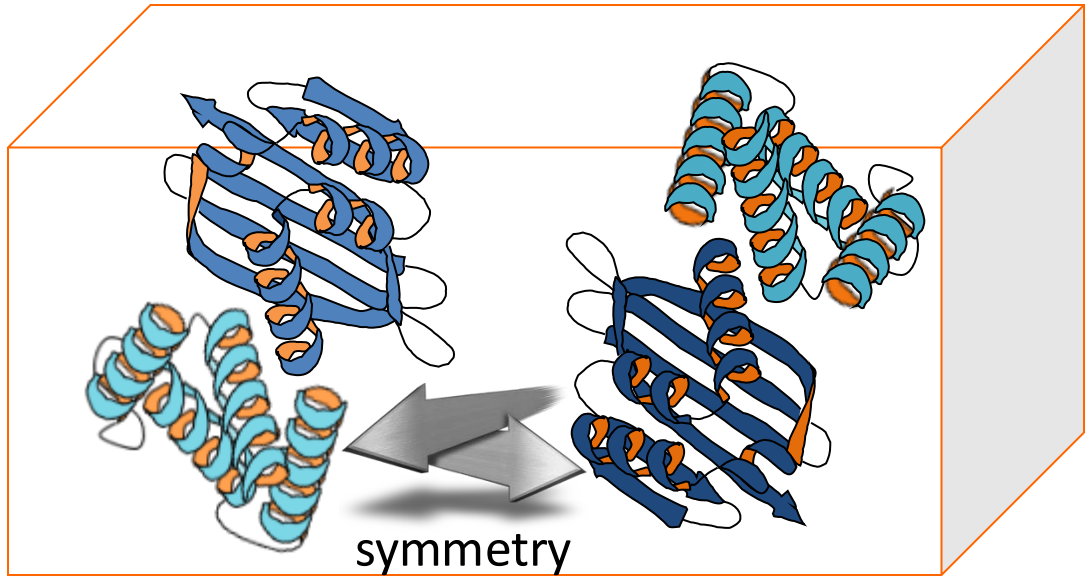
- Separate solutions could be on different origins*
- Signal may be low for one or more components



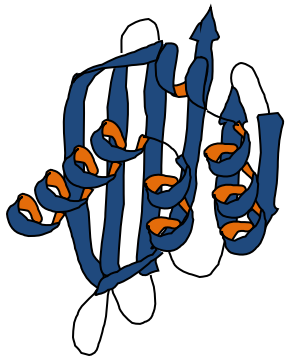
* later in talk!

molecular replacement

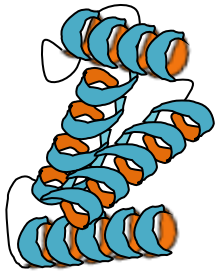
- Protein complex in the asymmetric unit
- Complex built by addition

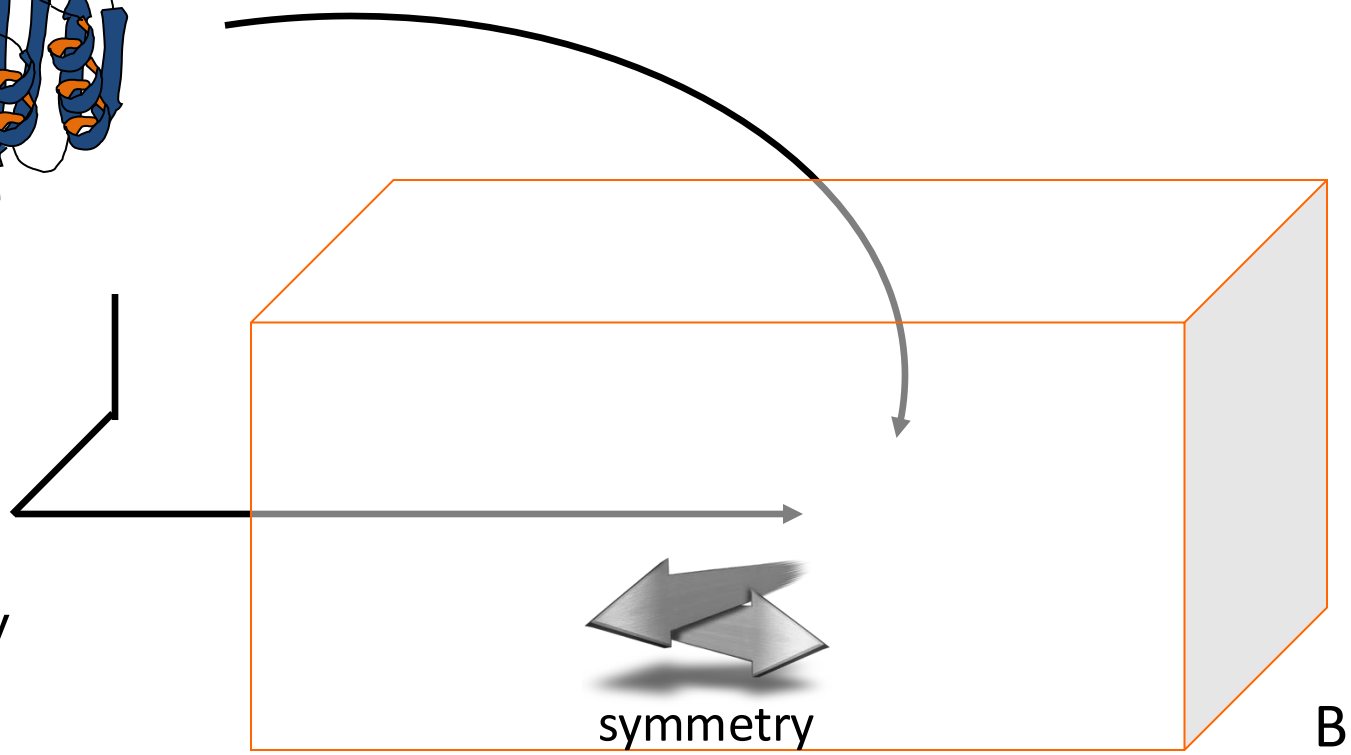


molecular replacement



(α, β, γ) Rotation

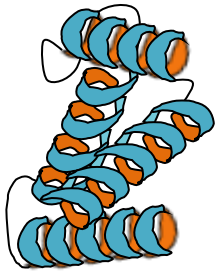




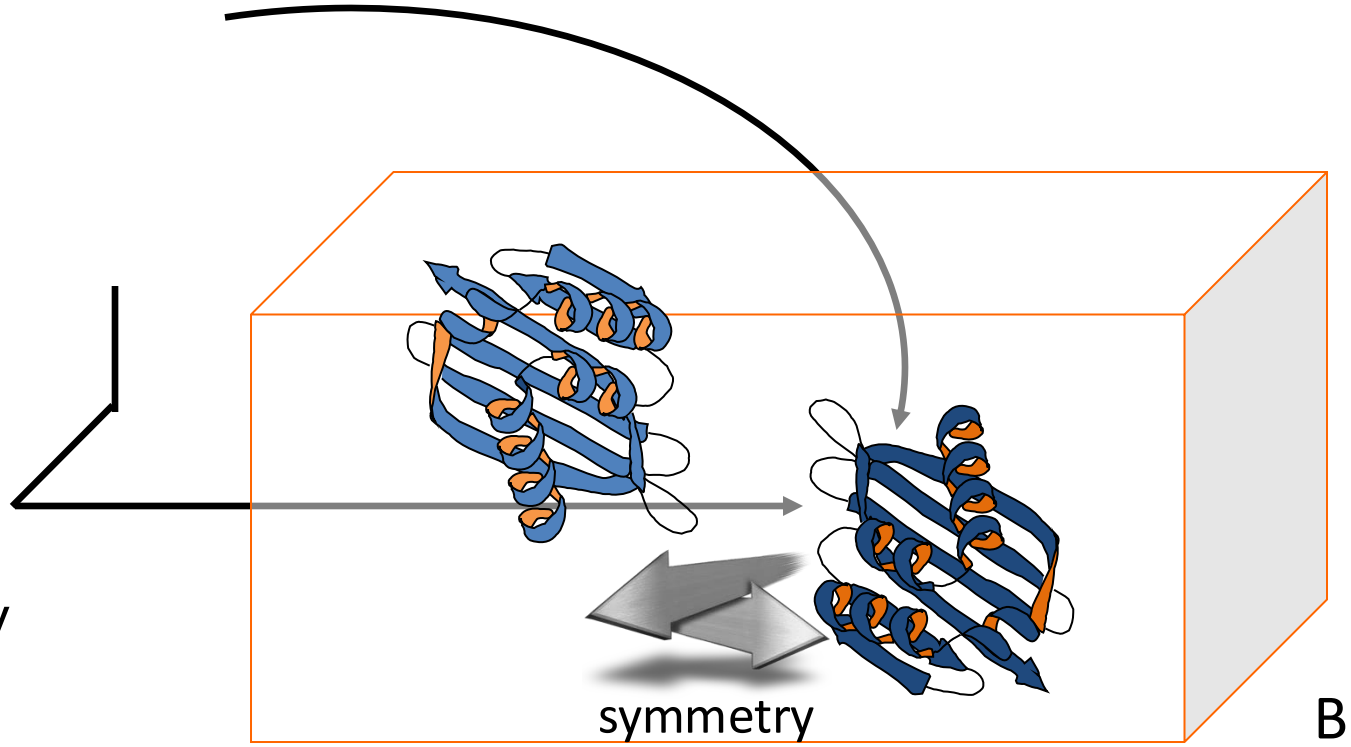
Structure built by addition placing second in the presence of the first

Biological oligomer placement

molecular replacement



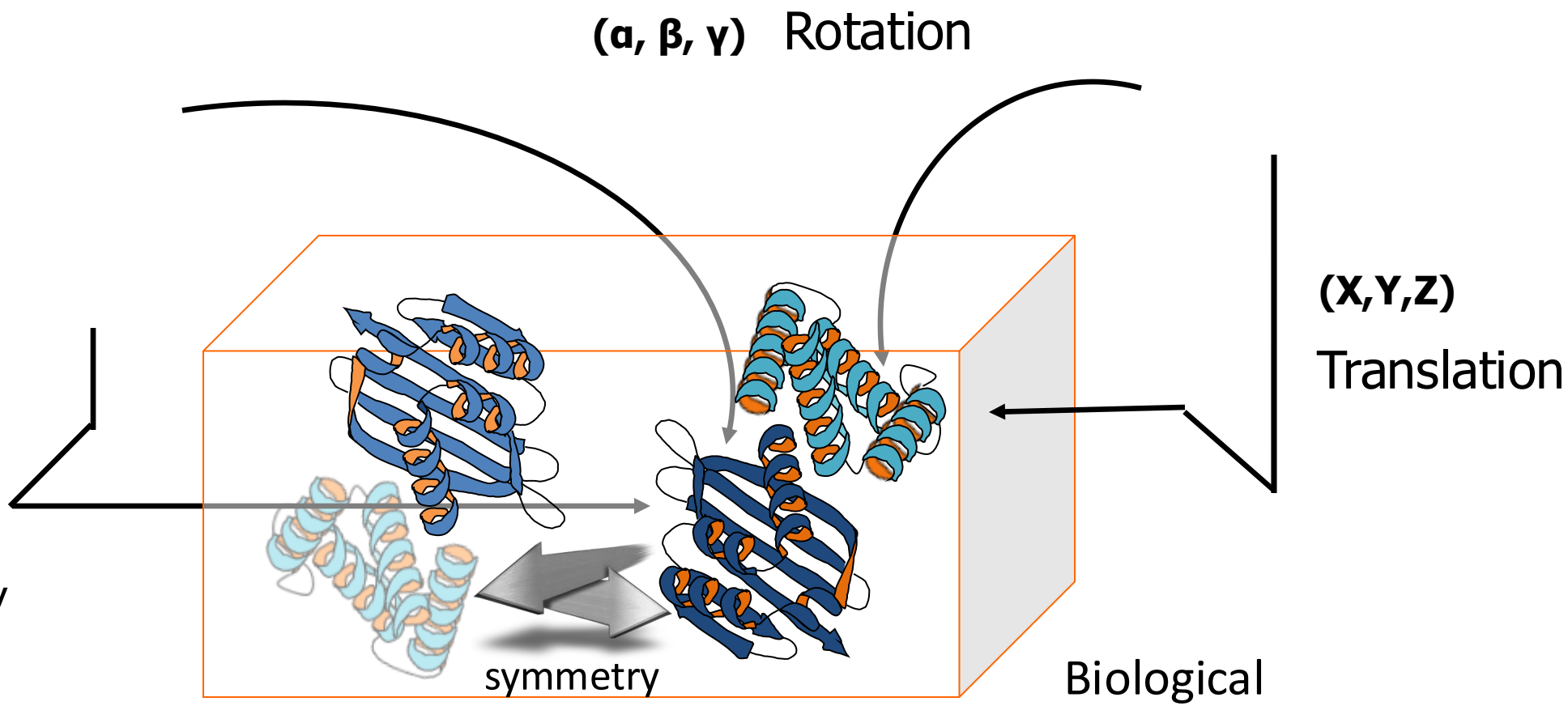
(α, β, γ) Rotation



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

Biological oligomer placement

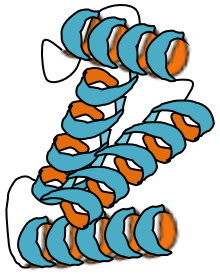
molecular replacement



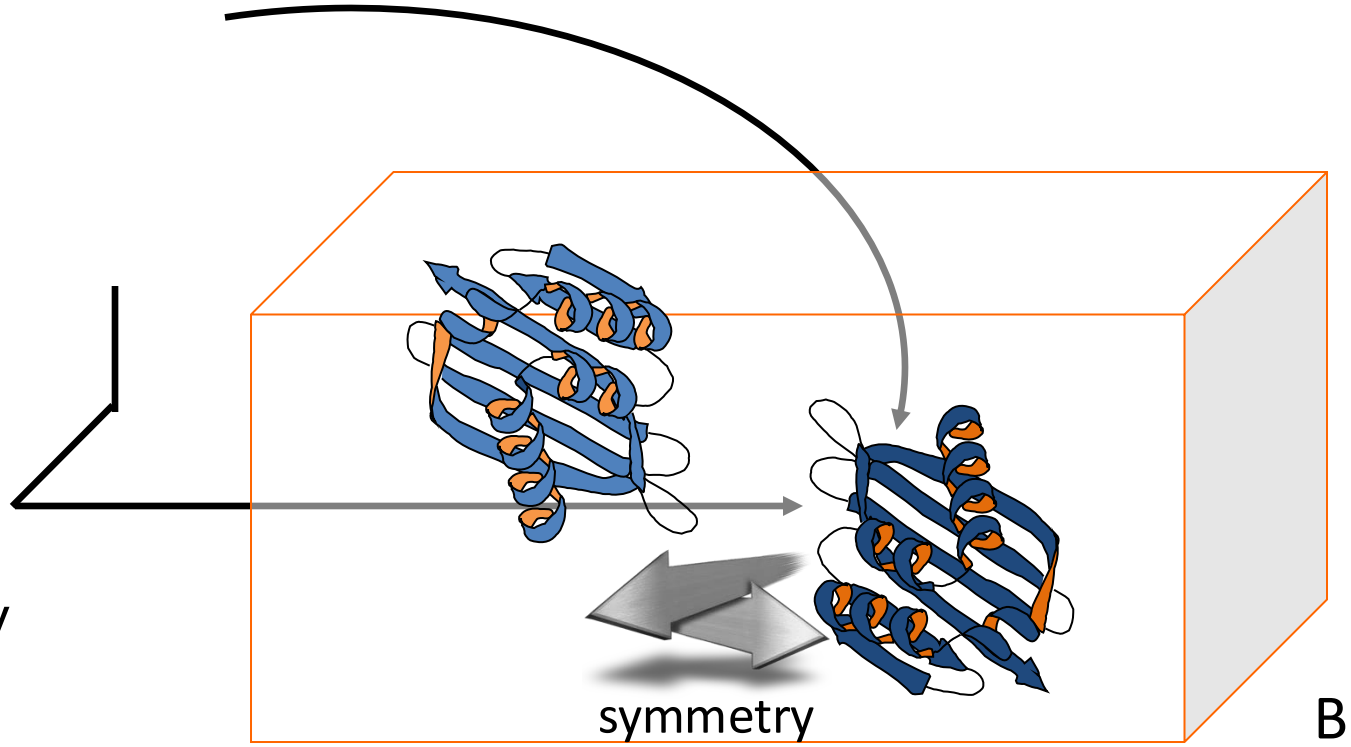
Structure built by addition placing second in the presence of the first

Biological oligomer placement

molecular replacement



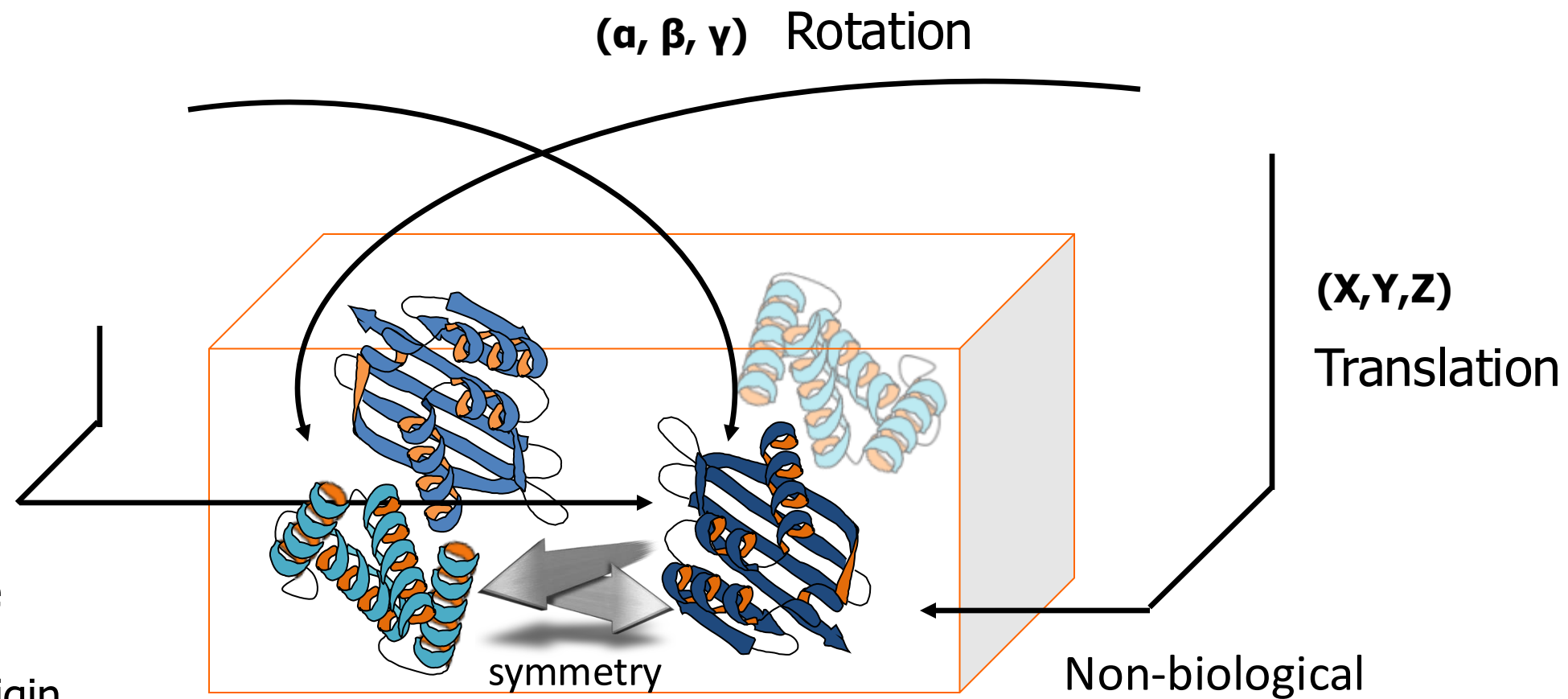
(α, β, γ) Rotation



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

Biological oligomer placement

molecular replacement



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

asymmetric unit contents

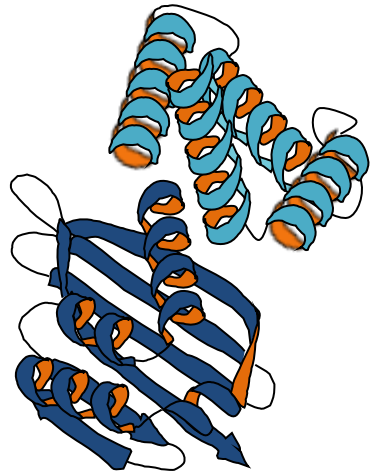
- Even if you never ever ever solve a crystal structure you need to know about asymmetric units
- It is the fundamental concept that distinguishes crystallographic structures from cryo-em, saxs, and the alphafold database
  - Crystallography gives you the structure *in the asymmetric unit*
  - *It is what is deposited!!*
- After you have a structure, you are at liberty to split it up and interpret it ignoring the asymmetric unit/crystal lattice
  - But beware!

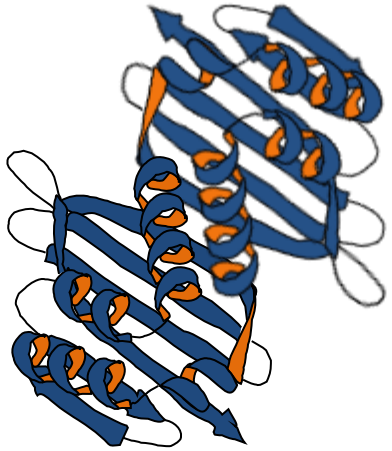
asymmetric unit

- The asymmetric unit is that part of the unit cell which can be used to generate the complete unit cell by the crystal symmetry
- Only the molecules in one asymmetric unit are deposited in crystallographic databases
- The choice of asymmetric unit is not unique
- The choice of an asymmetric unit is normally a black box operation done by a computer program
- The choice of **the** asymmetric unit is not an intrinsic property of the crystal
  - But the presence of **an** asymmetric unit is fundamental!
- We talk about 'the asymmetric unit' but mean 'an asymmetric unit'

asymmetric unit contents

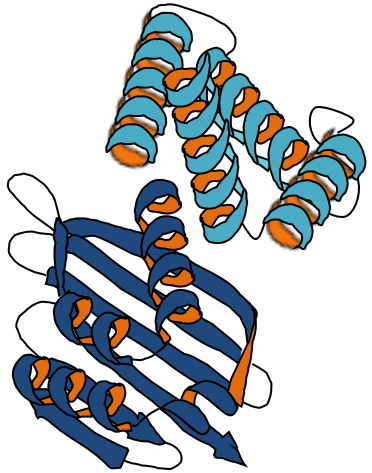
- non-crystallographic symmetry arises when there is more than one copy of a component in the asymmetric unit (asu)
- not all crystals have non-crystallographic symmetry
  - but they all have an asymmetric unit!

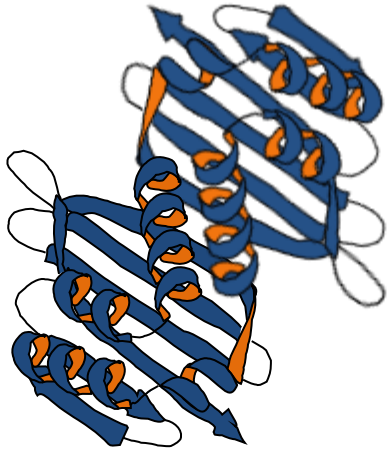




asymmetric unit contents

- The problem in molecular replacement is to place all the models in the asymmetric unit
- Equivalently:
  - Multiple copies of same protein
  - Components of complexes
  - Complexes
  - Fragments of long flexible proteins
  - Fragments of badly modelled proteins
  - Non-protein components
- The more components the harder the problem
  - Some asu symmetries can also make it harder

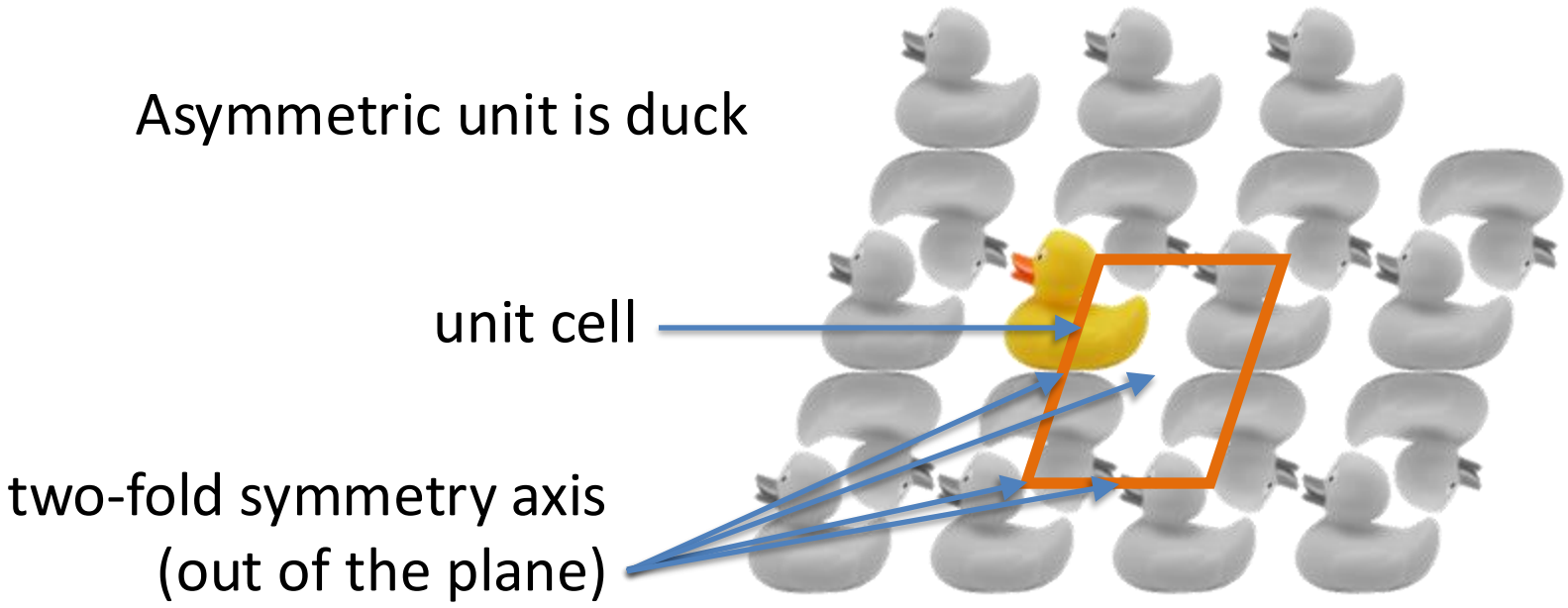


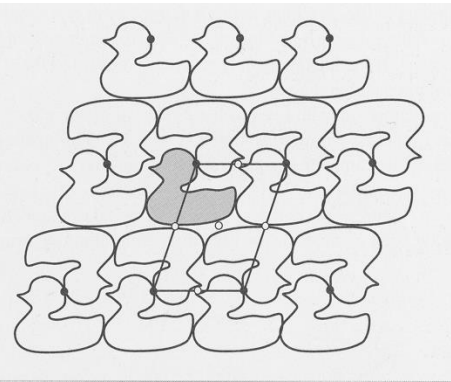


ducks

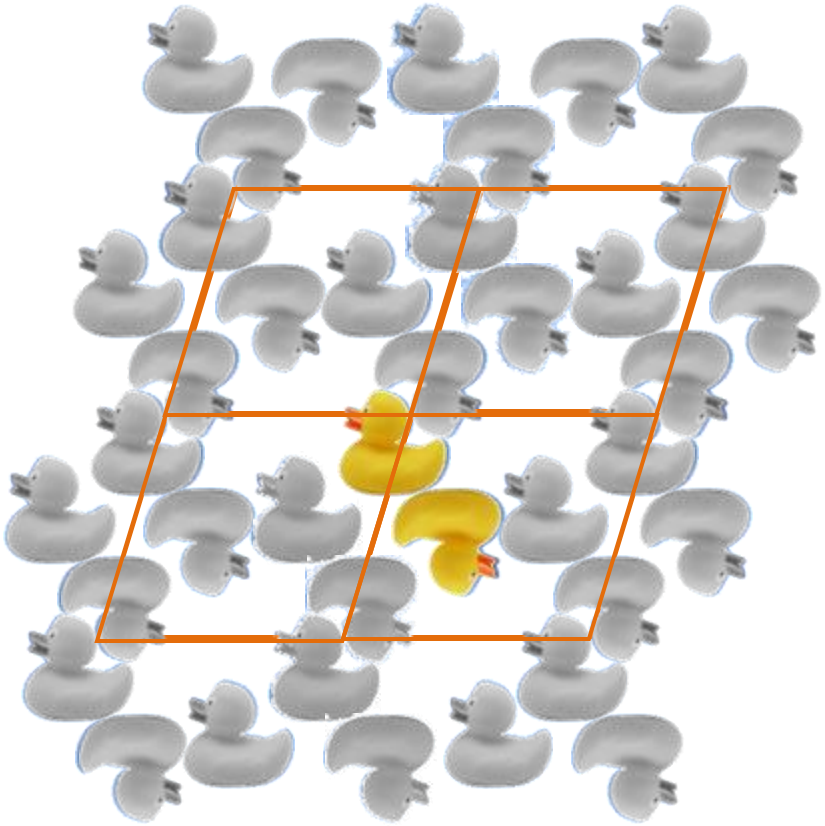
asymmetric unit

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

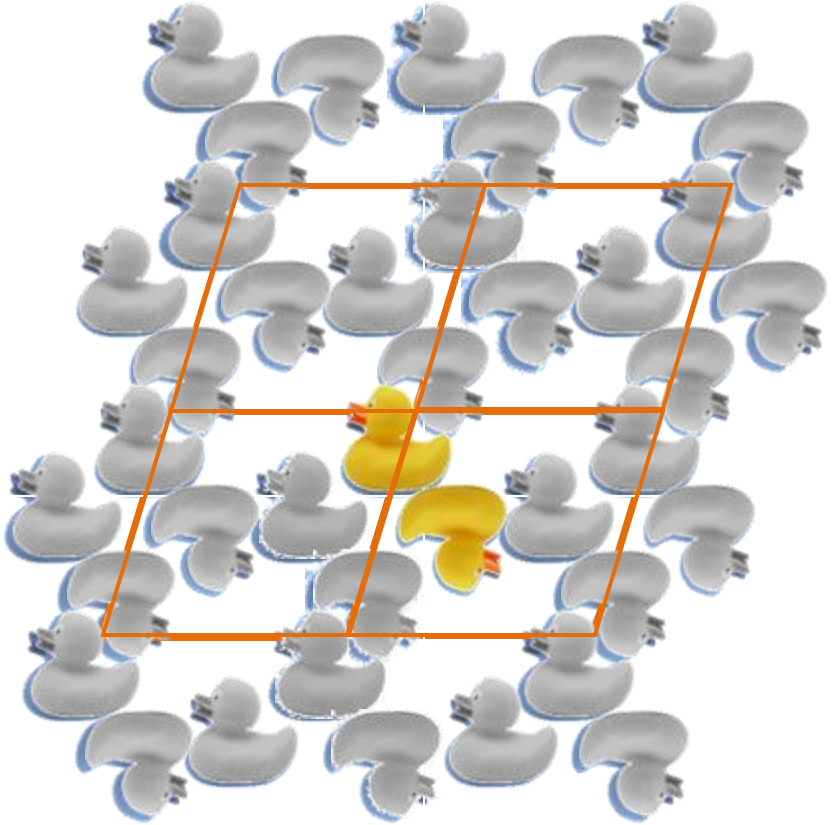




Crystallographic Symmetry

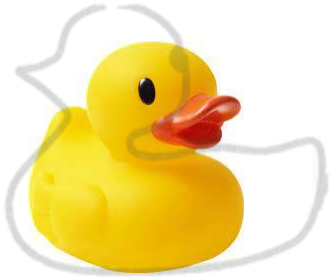


Non-crystallographic Symmetry

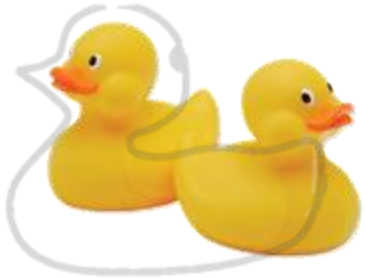


asymmetric unit contents

- duplication: non-crystallographic symmetry



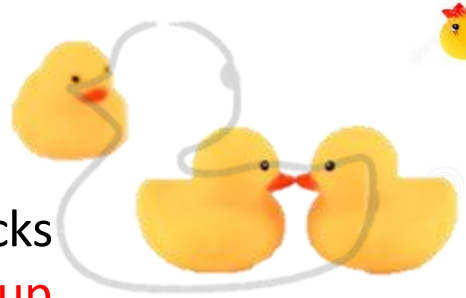
A duck



A brace of ducks
No point group symmetry

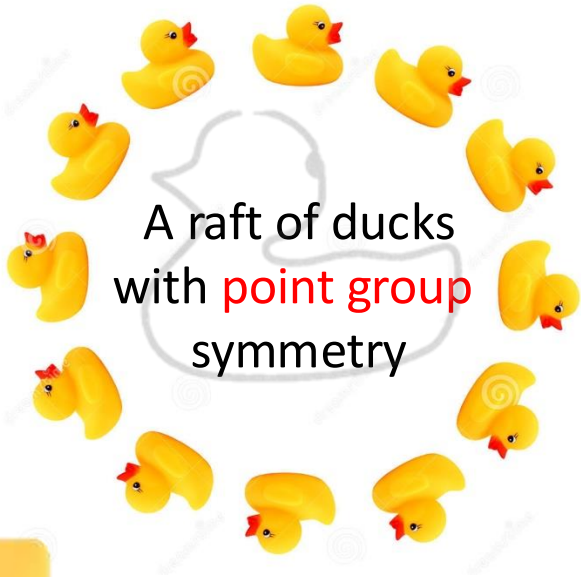
A brace of ducks
with point group
symmetry

A raft of ducks
non-
crystallographic
symmetry



A leash of ducks
with partial point group
symmetry

Four ducks
without translational
symmetry

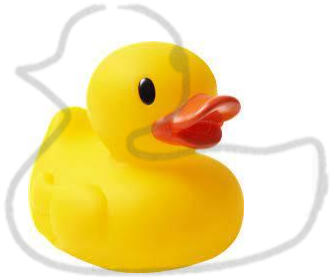


A raft of ducks
with point group
symmetry

Four ducks
with translational
symmetry

asymmetric unit contents

- The duck can also represent a complex



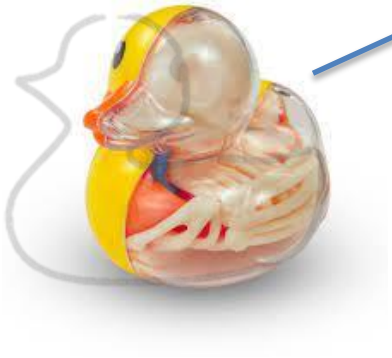
A duck

e.g. protein complex

e.g. RNA/DNA-protein complex

e.g. ribosome

e.g. protein-ligand complex

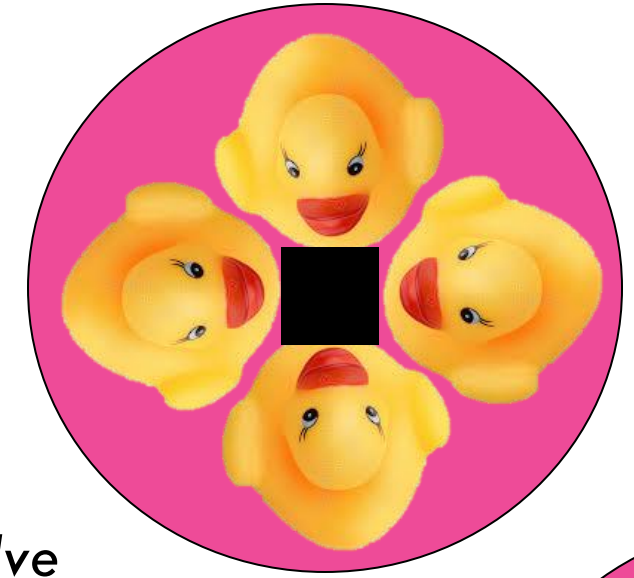


e.g. (secondary) structure elements

asymmetric unit and biological assemblies

Asymmetric unit

Crystallographic 4-fold



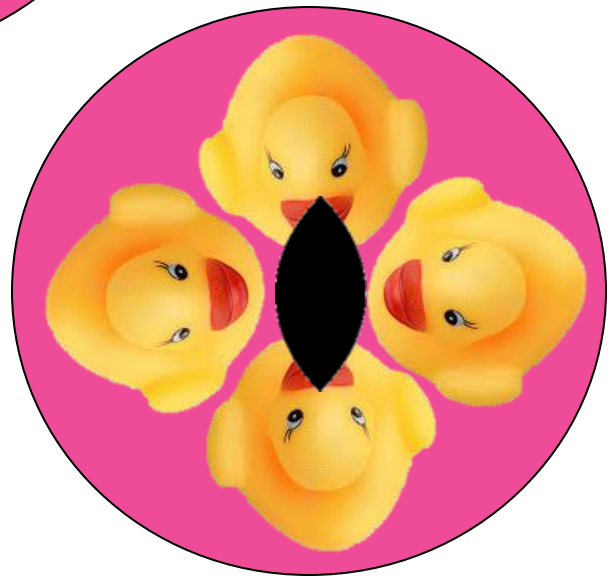
biological assembly (tetramer)

Relationship between asu and biological assembly may involve crystal symmetry

not biological assembly

Asymmetric unit

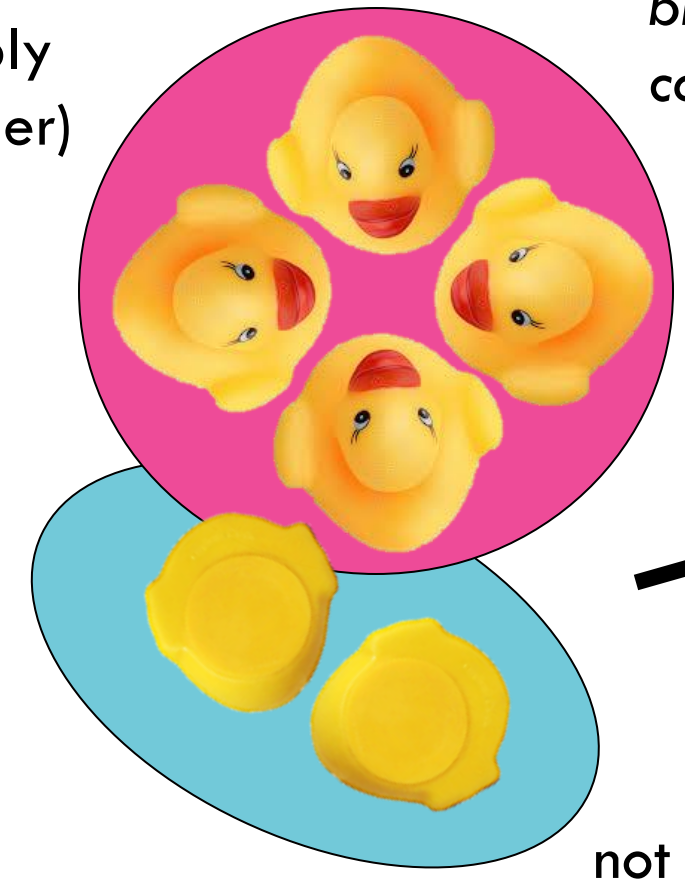
Crystallographic 2-fold



asymmetric unit and biological assemblies

biological assembly (tetramer)

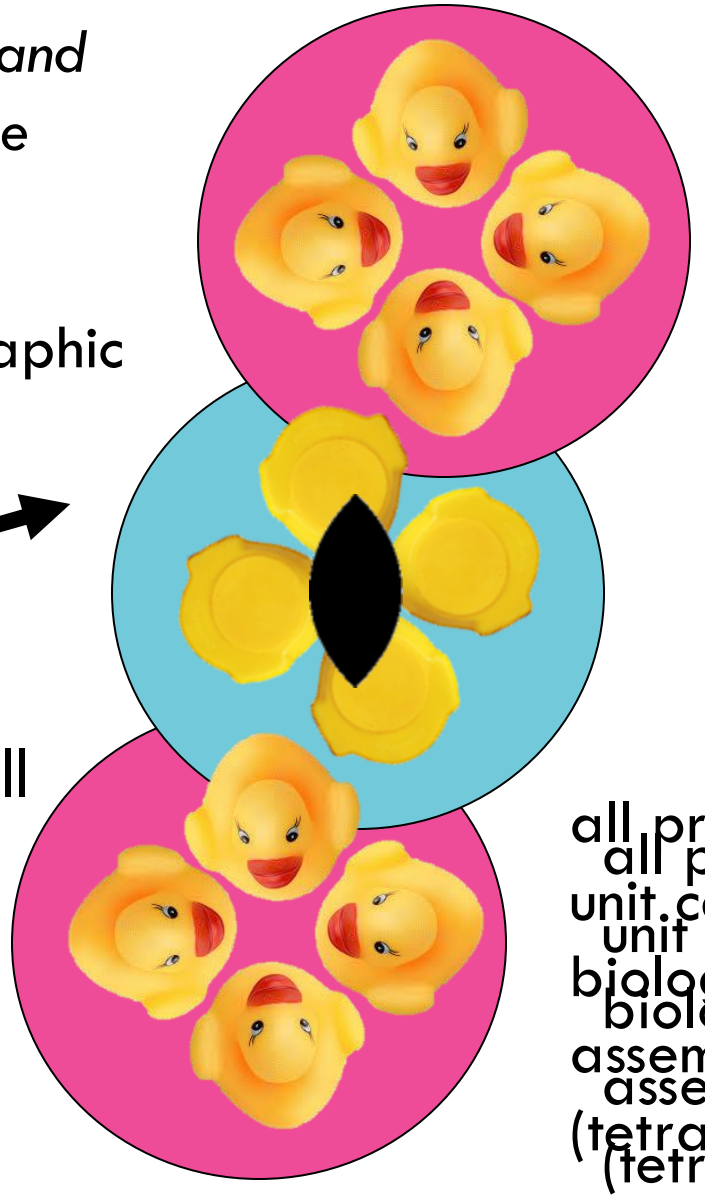
Relationship between asu and biological assembly can be complicated



Asymmetric unit

not biological assembly

Crystallographic 2-fold



Unit cell

all present in unit call as biological assembly (tetramer)

Matthew’s coefficient

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*

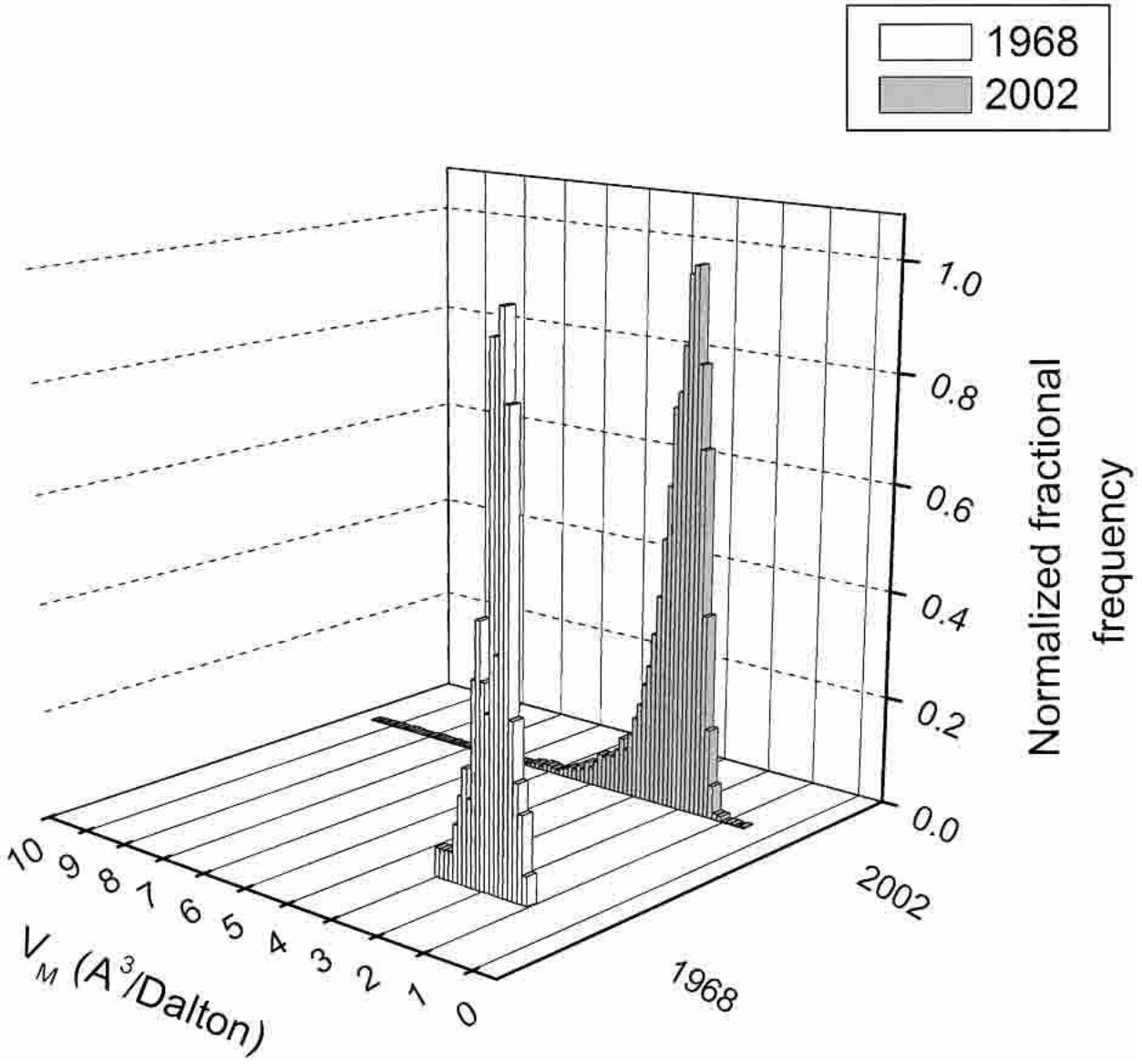
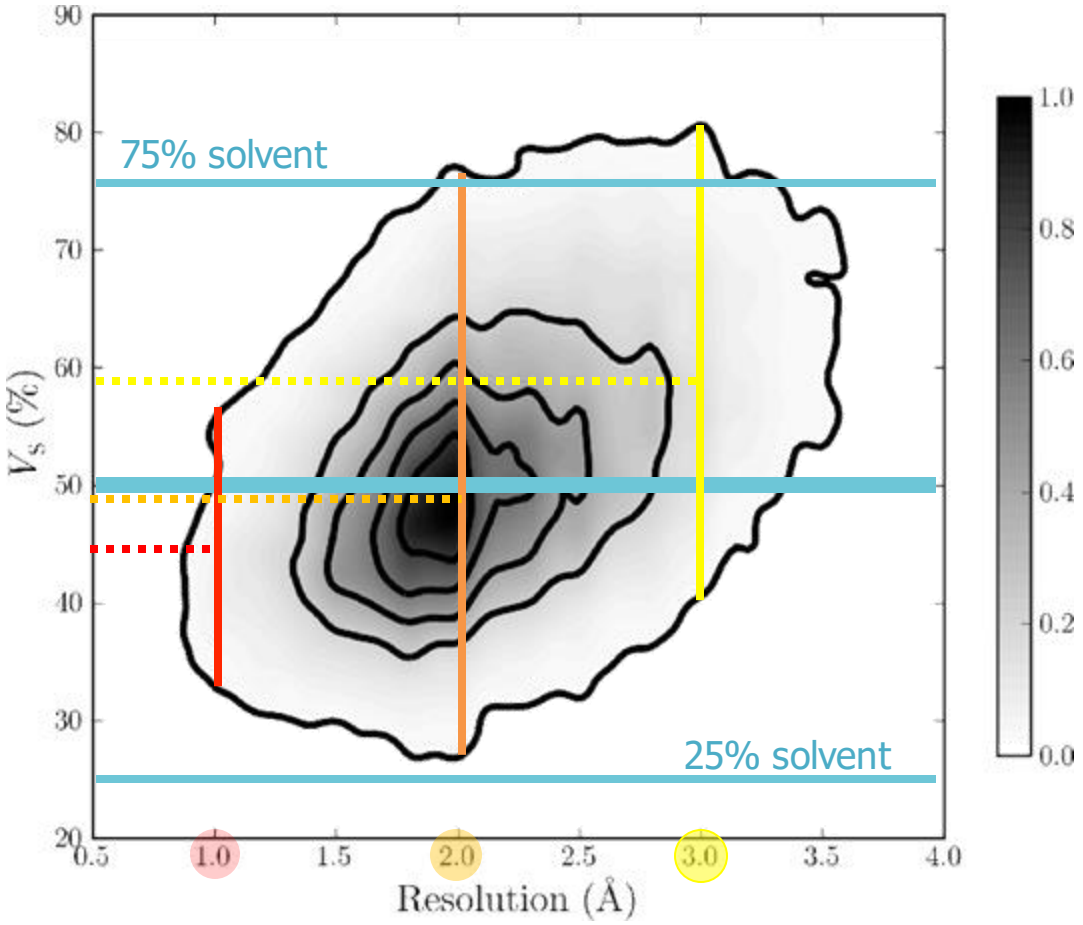


Figure 1: Kantardjieff and Rupp (2003)

Matthew's coefficient

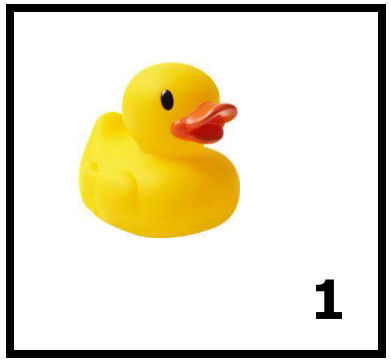
- First calculated by Brian Matthews in 1968
- Most crystals are 50% protein by volume
  - Between 25% and 75% protein
  - Slightly resolution dependent
- Can be used to estimate the contents of the asymmetric unit
  - *c.f. Self Rotation Function*
  - *c.f. TNCS Order*



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

components of asymmetric unit

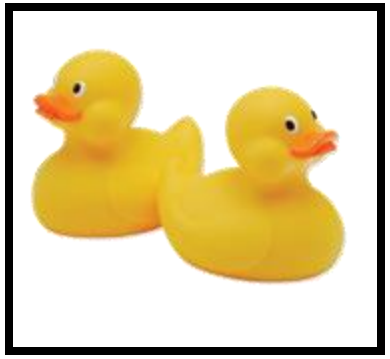
- About 50% solvent

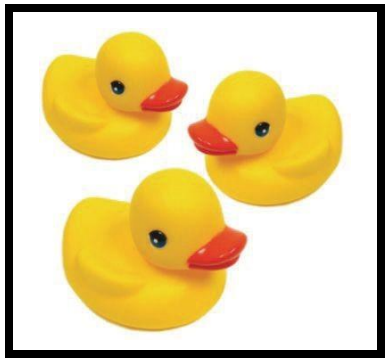


With low numbers of possible copies, options are low

1 can only be 1

For 1, 2 or 3,
1 is unlikely

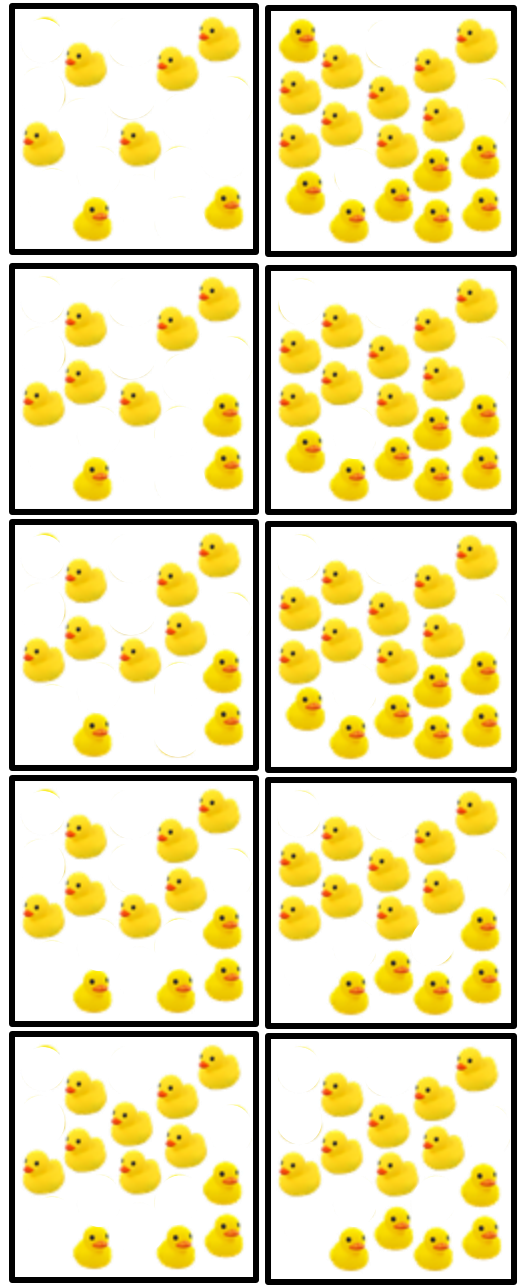




1
2
3

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

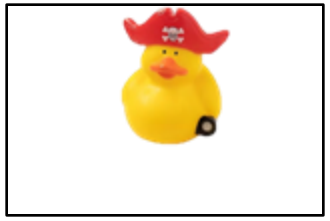


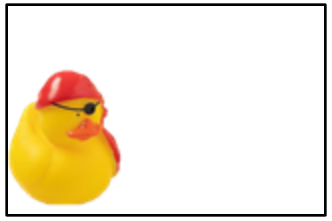
search order

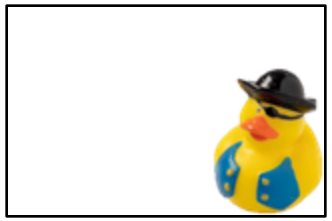
search order permutations

- 3 components
- 6 search orders

Protein present with high B-factor and can only be placed when other components are placed, which increases the signal



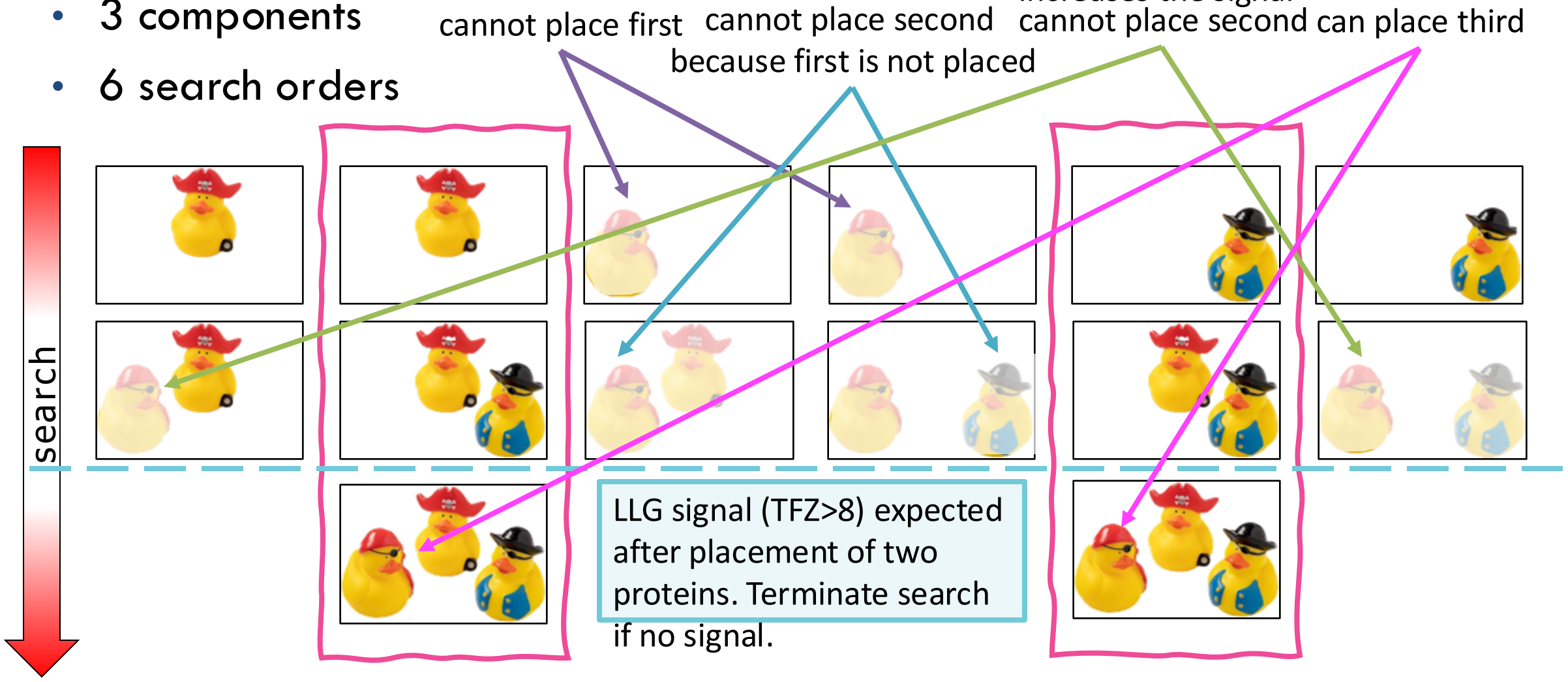




search order permutations

- 3 components
- 6 search orders

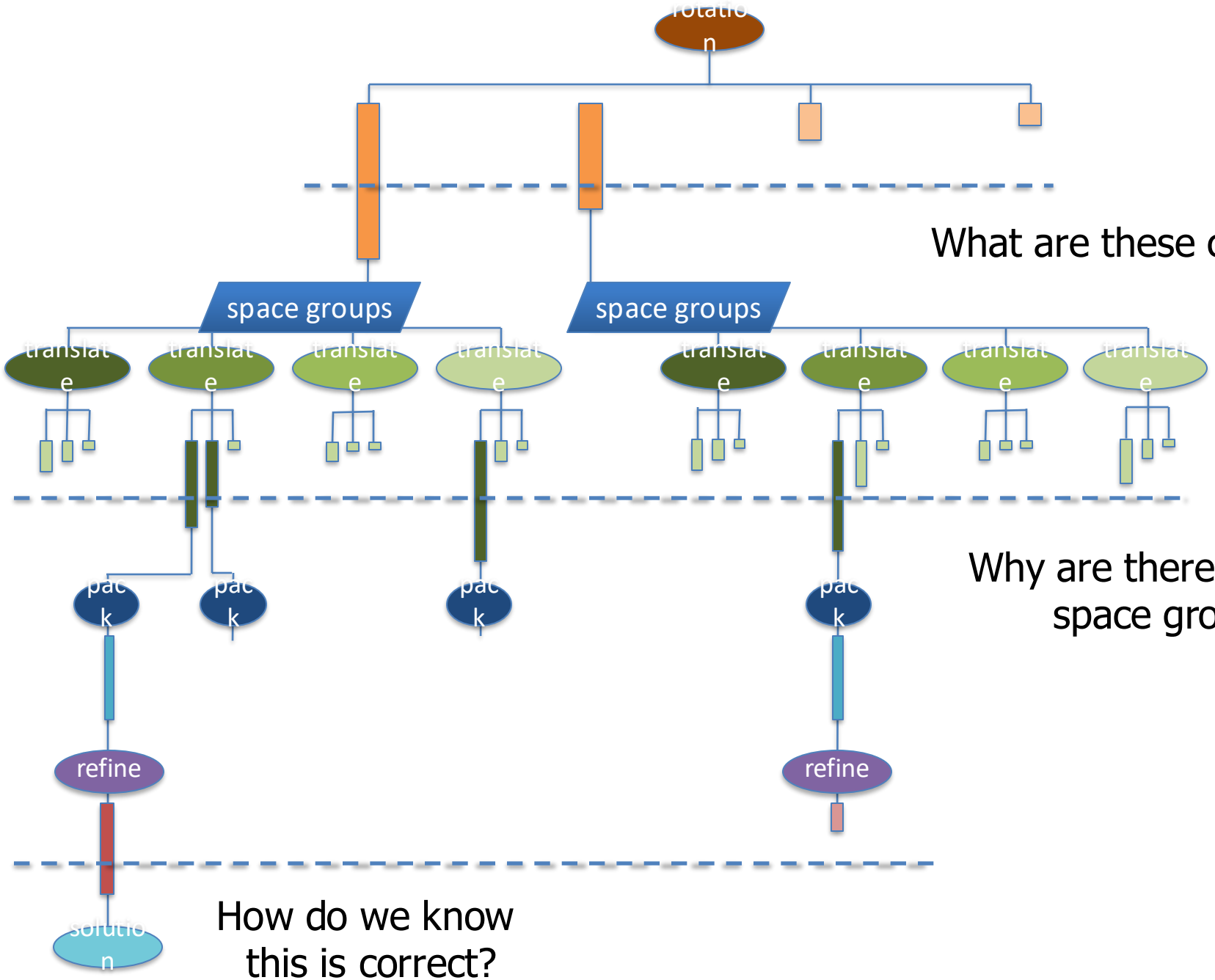
Protein present with high B-factor and can only be placed when other components are placed, which increases the signal



search order permutations

- Search permutations, combinations and search order may be handled by molecular replacement pipelines
- If you don’t get a ’full’ solution you will need to understand how to add extra components to complete the solution

Phaser



What are these cutoffs?

Why are there different space groups?

How do we know this is correct?

Diagram: Martin Noble

maximum likelihood scoring in phaser

- Use probability
- Probabilities account for errors
  - Errors in the data, using I and σI (intensities
  - Errors in the model (rms error)

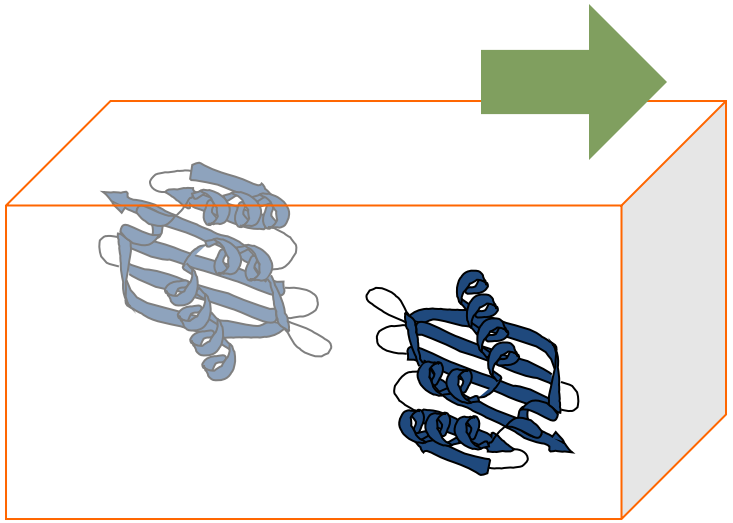
$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1 - D_{obs}^2 \sigma_A^2} \exp\left(- \frac{E_e^2 + D_{obs}^2 \sigma_A^2 E_C^2}{1 - D_{obs}^2 \sigma_A^2} \right) I_0\left(\frac{2E_e D_{obs} \sigma_A E_C}{1 - D_{obs}^2 \sigma_A^2} \right) \right)$$

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

example: translation function search

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

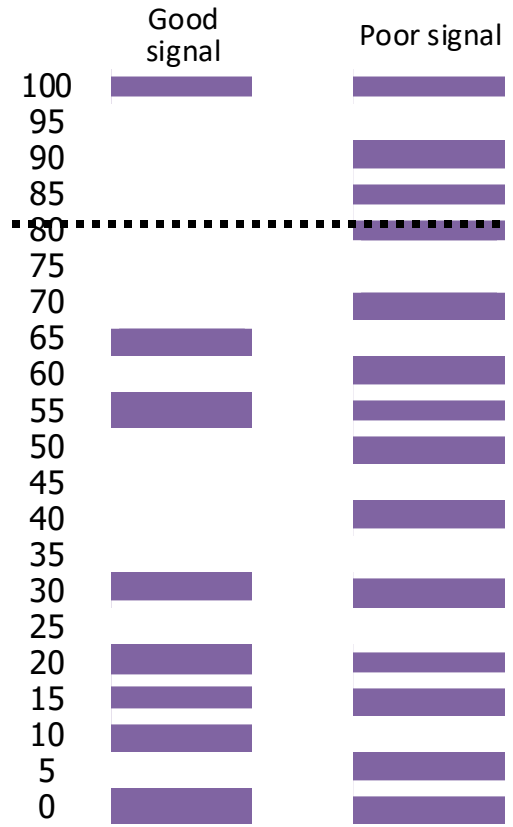
P2



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

peak selection in rotation and translation function

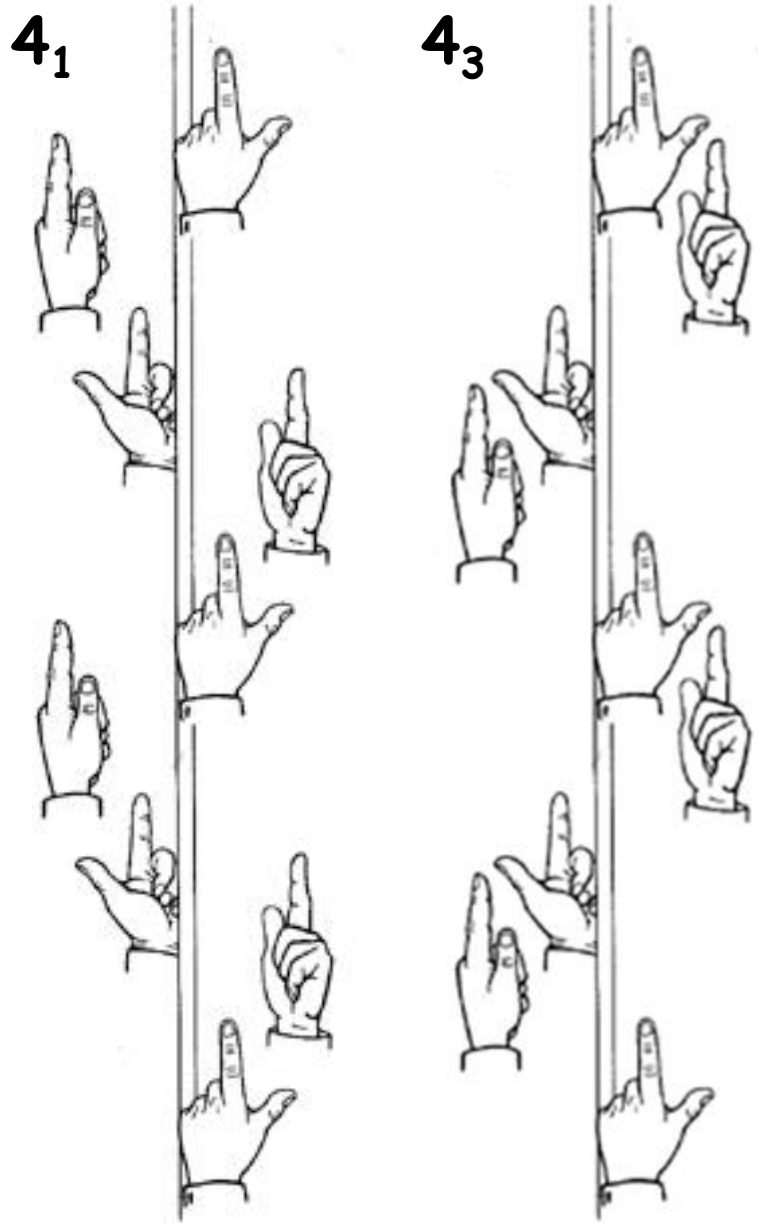
- Must chose a selection criteria to carry potential solutions through to the next step
- By default in phaser, solutions over 75% of the difference between the top peak and the mean are selected
  - Good signal, few potential solutions
  - Poor signal, many potential solutions
- The absolute value of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



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

space group determination

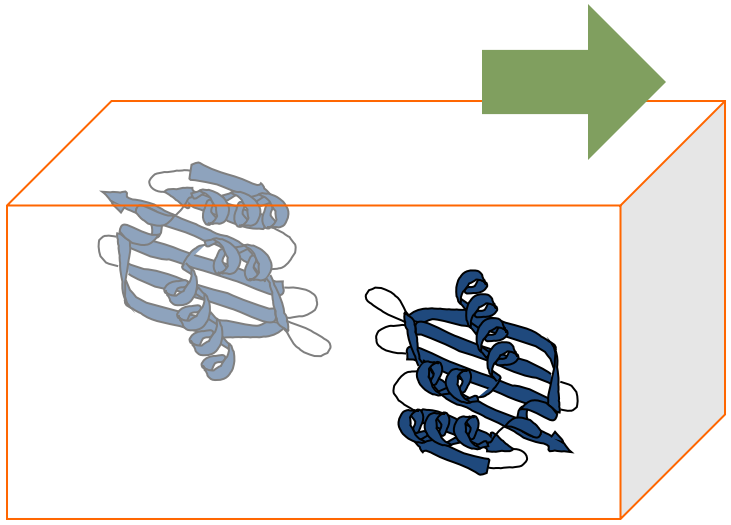
- Space groups that come in enantiomorphic pairs (e.g. $P4_1$, $P4_3$) cannot be distinguished at the data processing stage
- The space group is only confirmed when the structure is solved
- Phaser will run enantiomorphic pairs in the point group in the translation function
  - Or all space groups in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



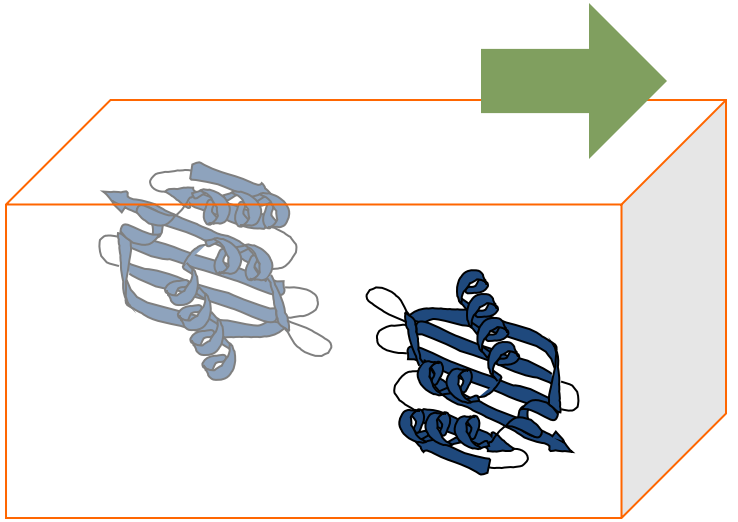
example: translation function search

- Different space groups

$P2_1$



P2

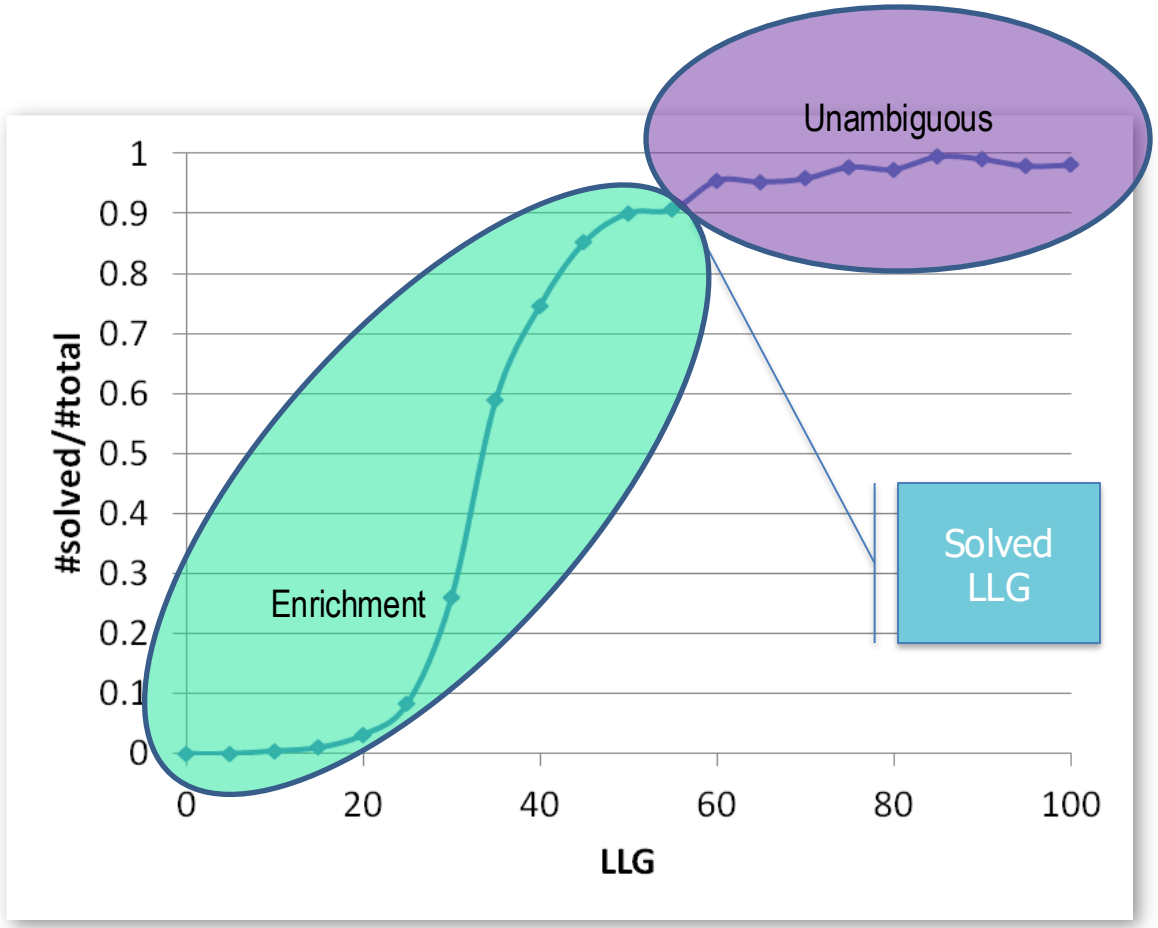


$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1 - D_{obs}^2 s_A^2} \exp\left(-\frac{E_e^2 + D_{obs}^2 s_A^2 E_C^2}{1 - D_{obs}^2 s_A^2}\right) I_0\left(\frac{2E_e D_{obs} s_A E_C}{1 - D_{obs}^2 s_A^2}\right)\right)$$

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

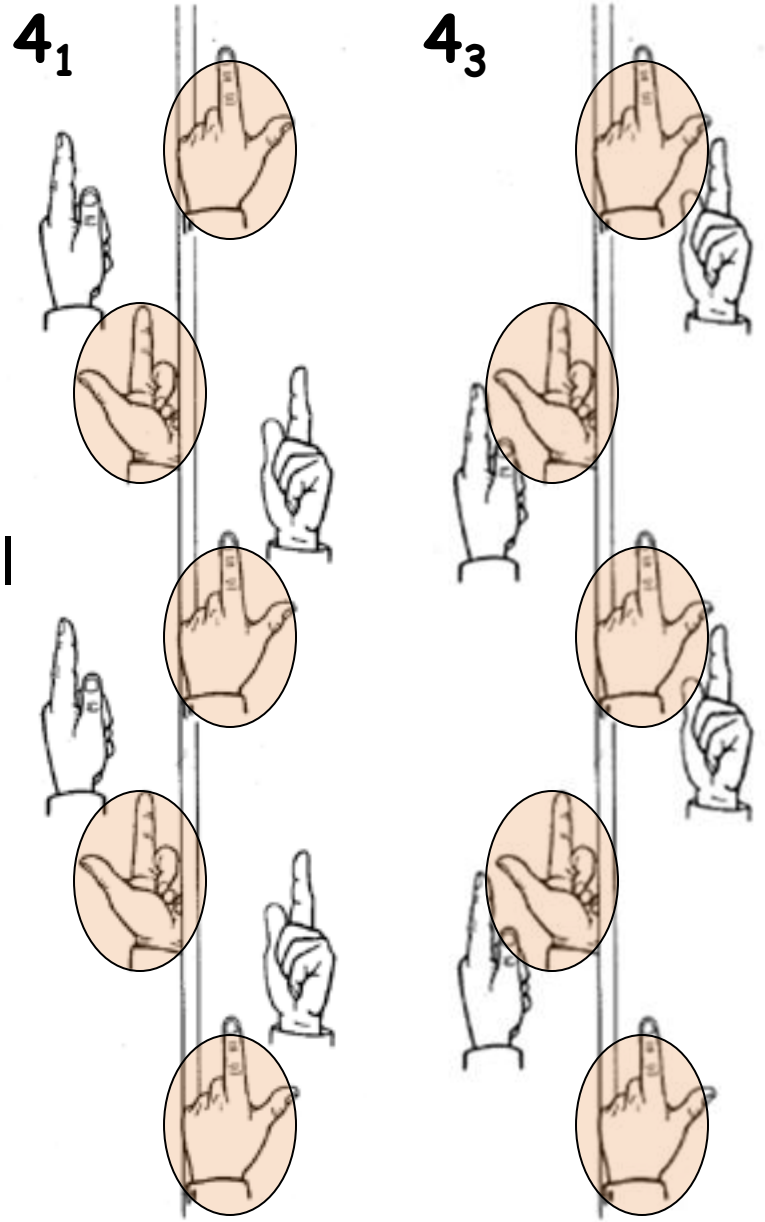
When is a model correctly placed?

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

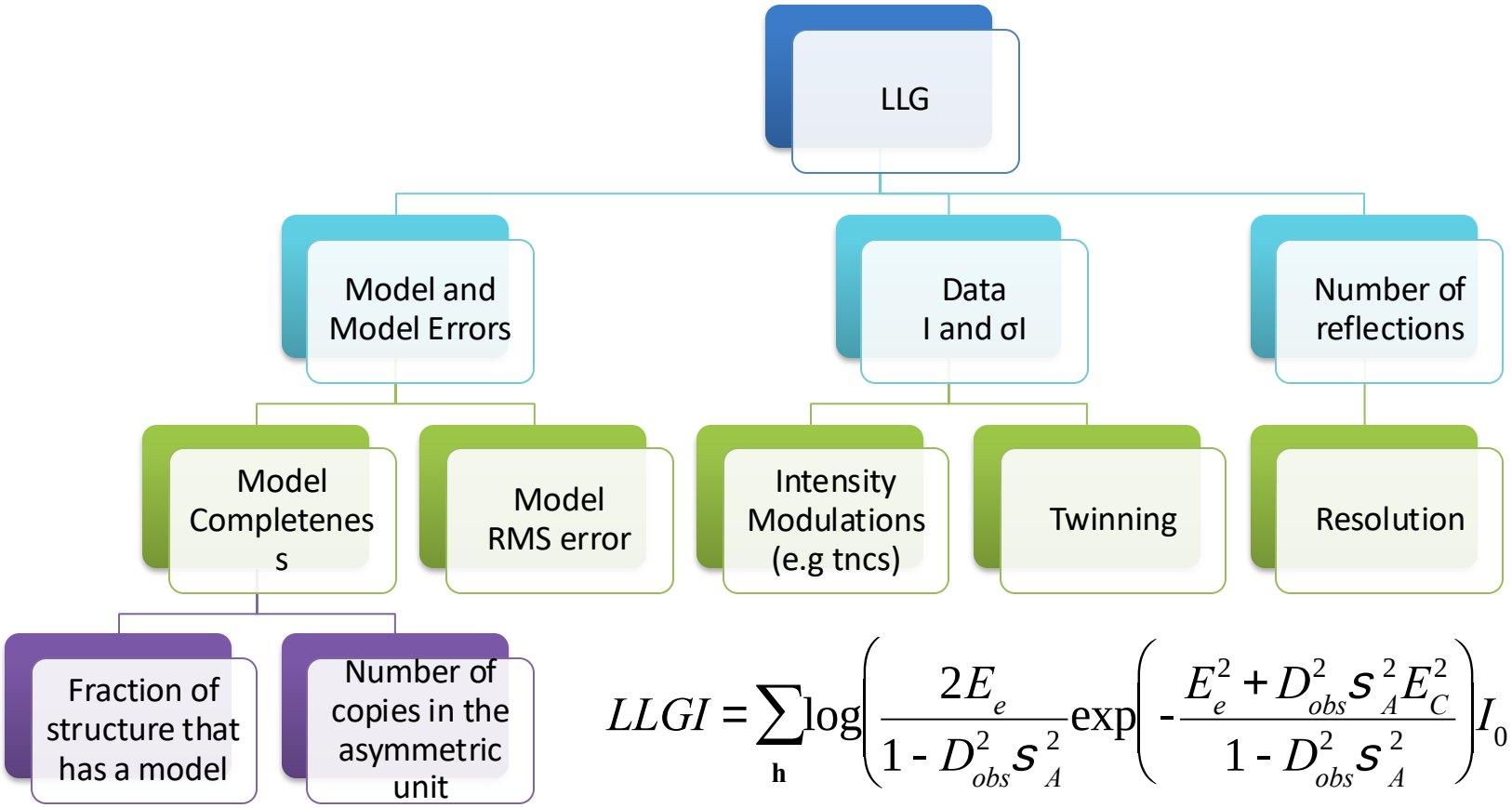
how not to judge a solution

- 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



What degrades the signal?

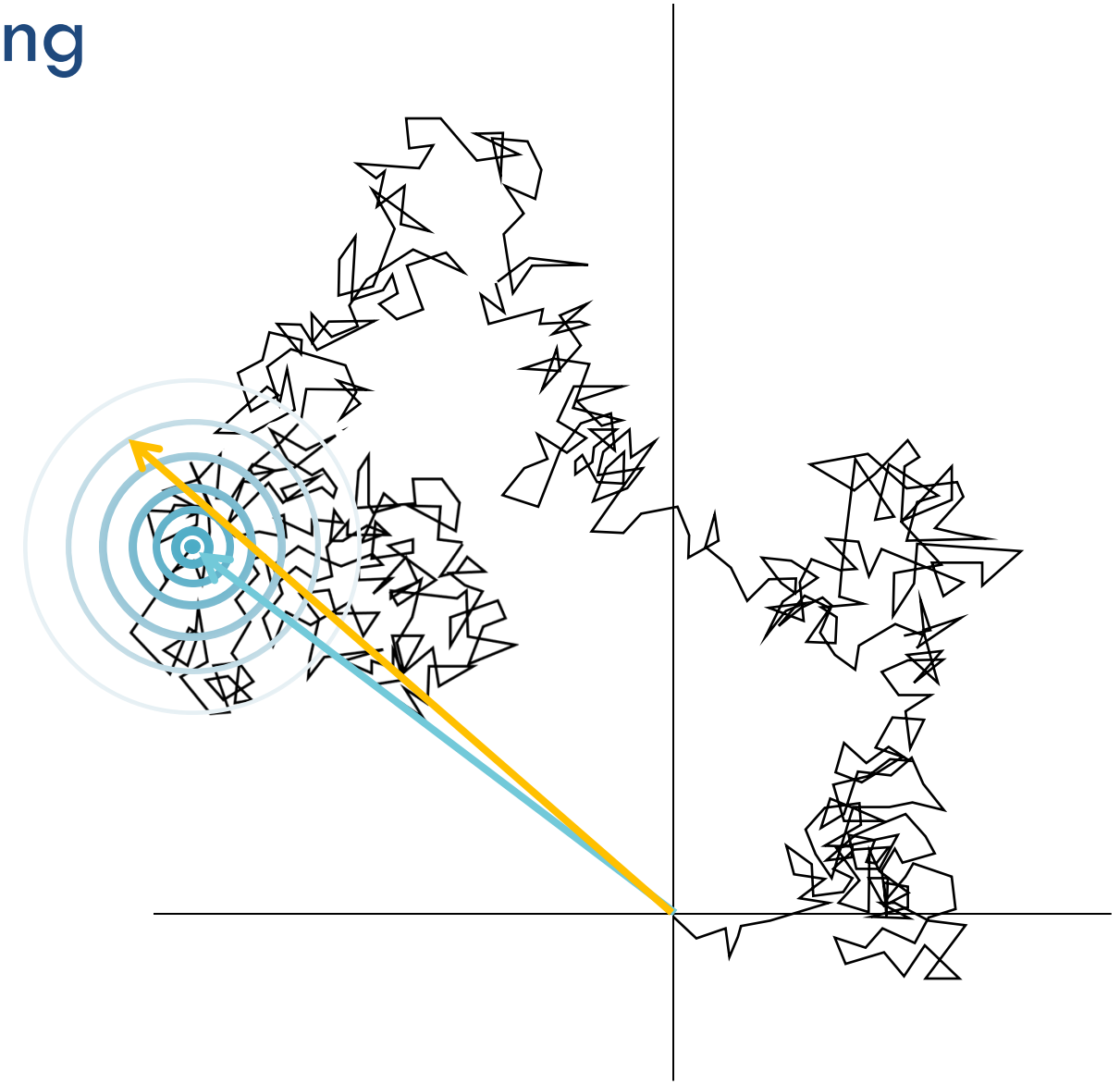


$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1 - D_{obs}^2 s_A^2} \exp\left(-\frac{E_e^2 + D_{obs}^2 s_A^2 E_C^2}{1 - D_{obs}^2 s_A^2}\right) I_0\left(\frac{2E_e D_{obs} s_A E_C}{1 - D_{obs}^2 s_A^2}\right)\right)$$

- Phaser's target function allows us to account for many of the problems that occur in molecular replacement
  - But not explicitly twinning

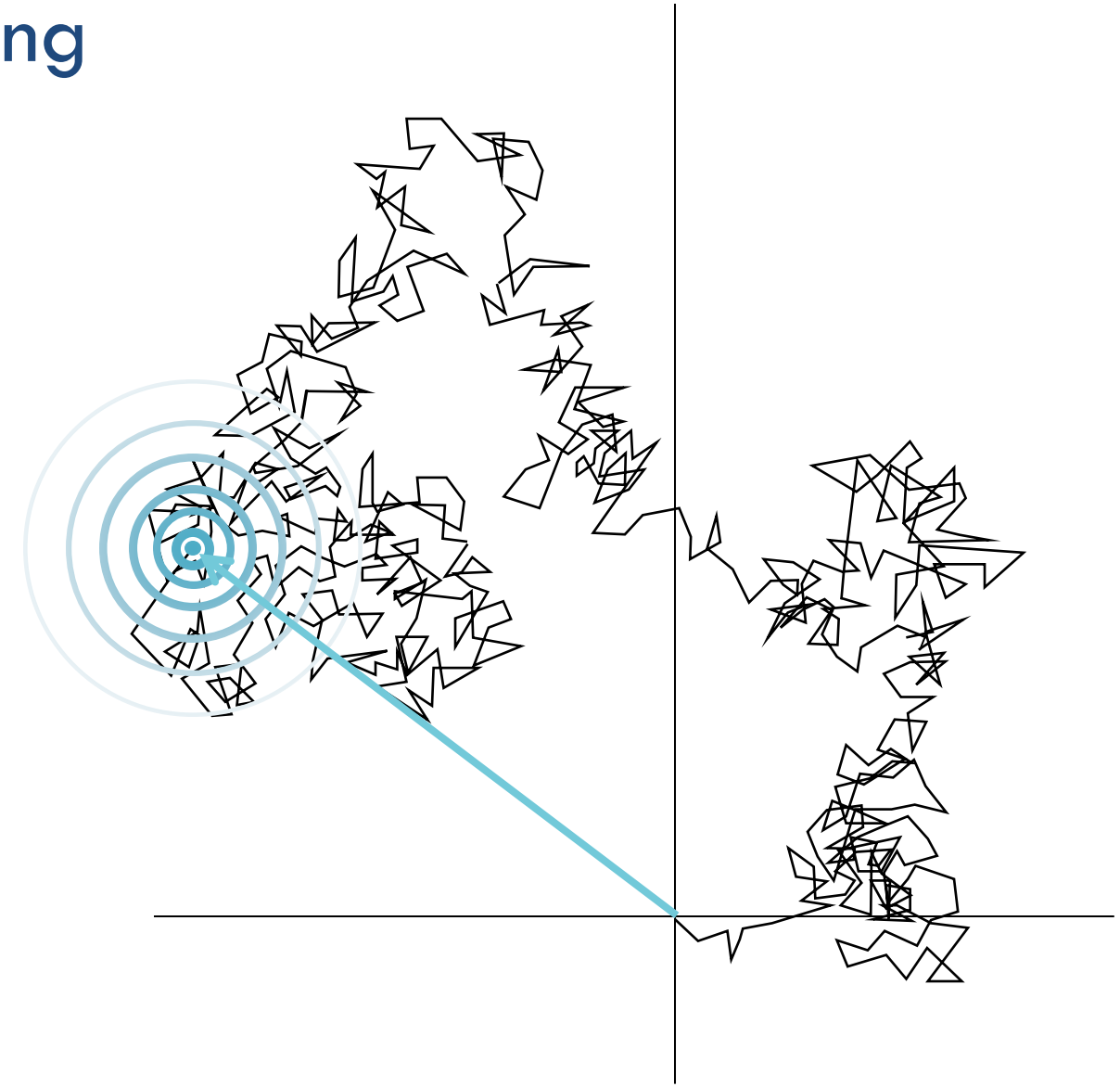
bad models – poor modelling

- Atoms in the wrong place will degrade the calculated structure factor
- Domains should be split



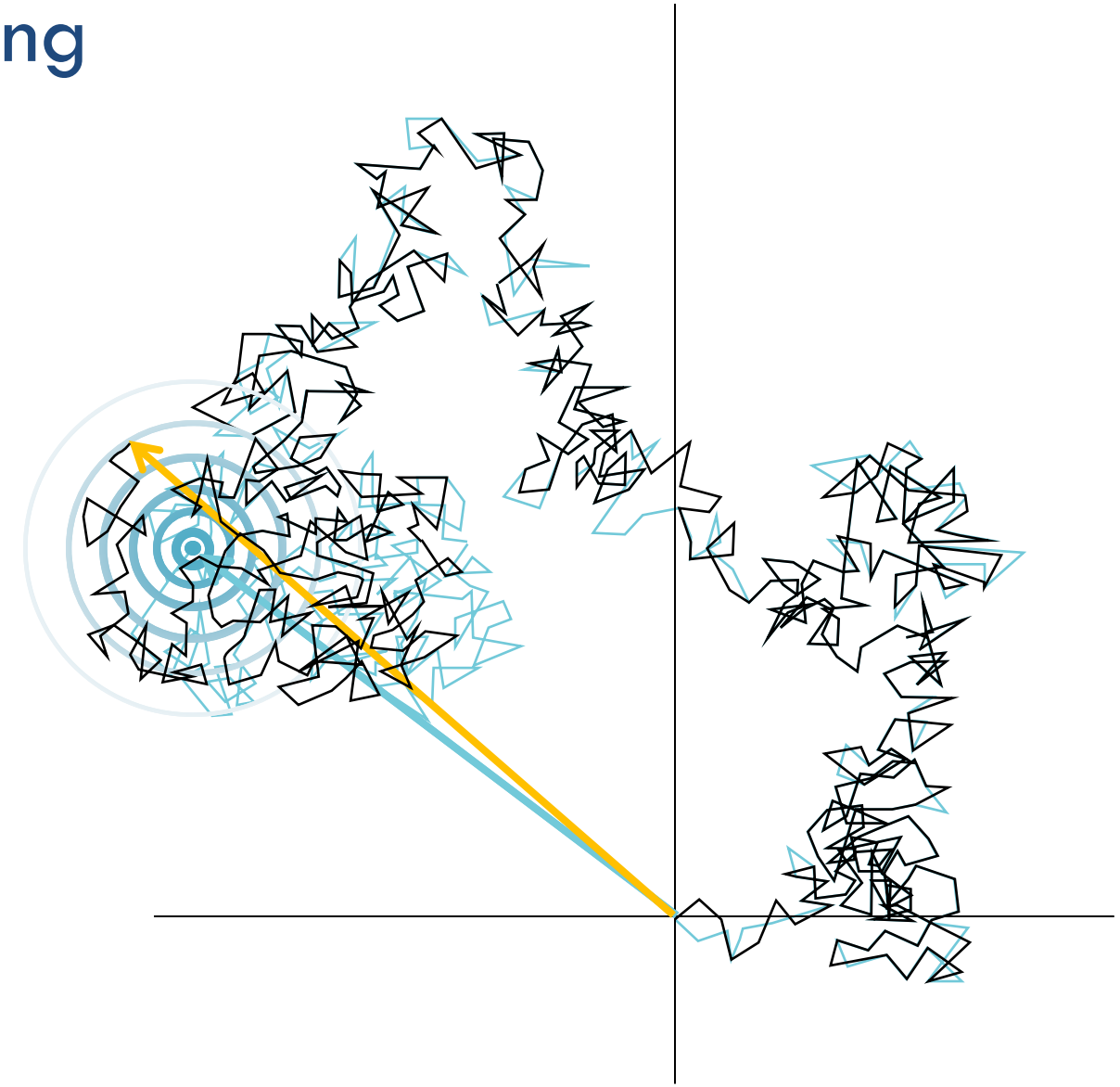
bad models – poor modelling

- Atoms in the wrong place will degrade the calculated structure factor
- Random error in each coordinate



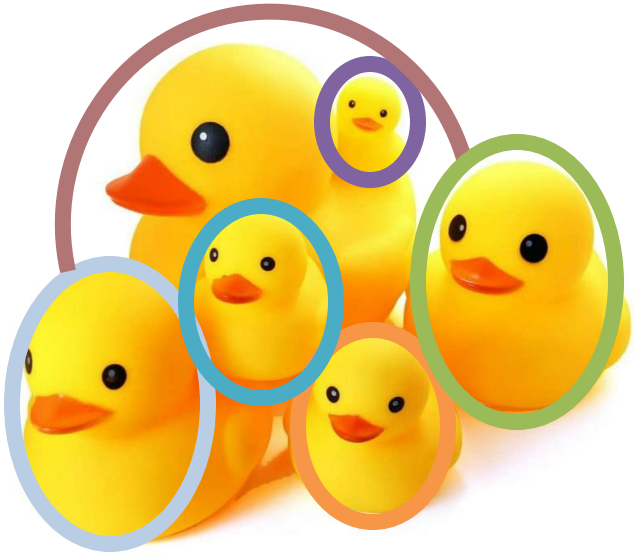
bad models – poor modelling

- Atoms in the wrong place will degrade the calculated structure factor
- Random error in each coordinate

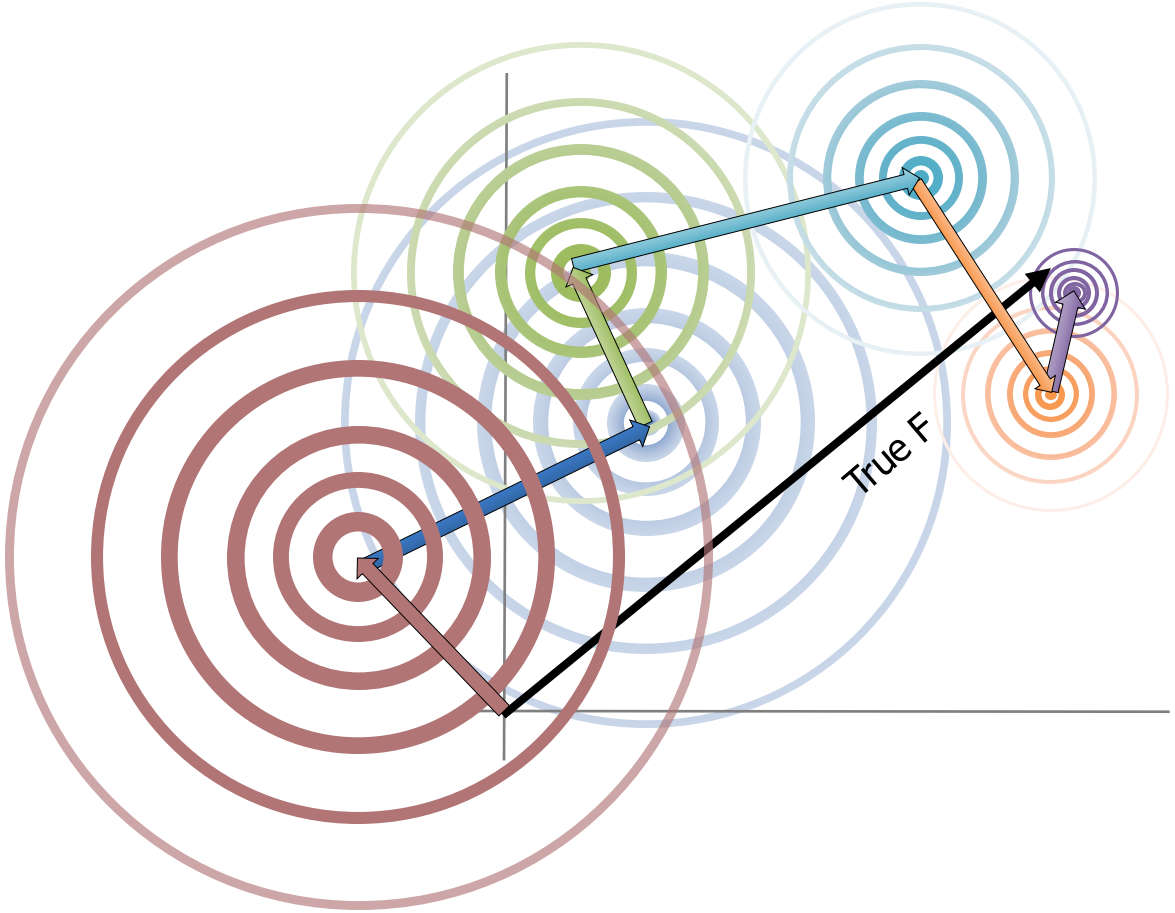


bad models – fraction scattering

- Missing atoms will also degrade the calculated structure factors
- This is the biggest source of error **during** molecular replacement

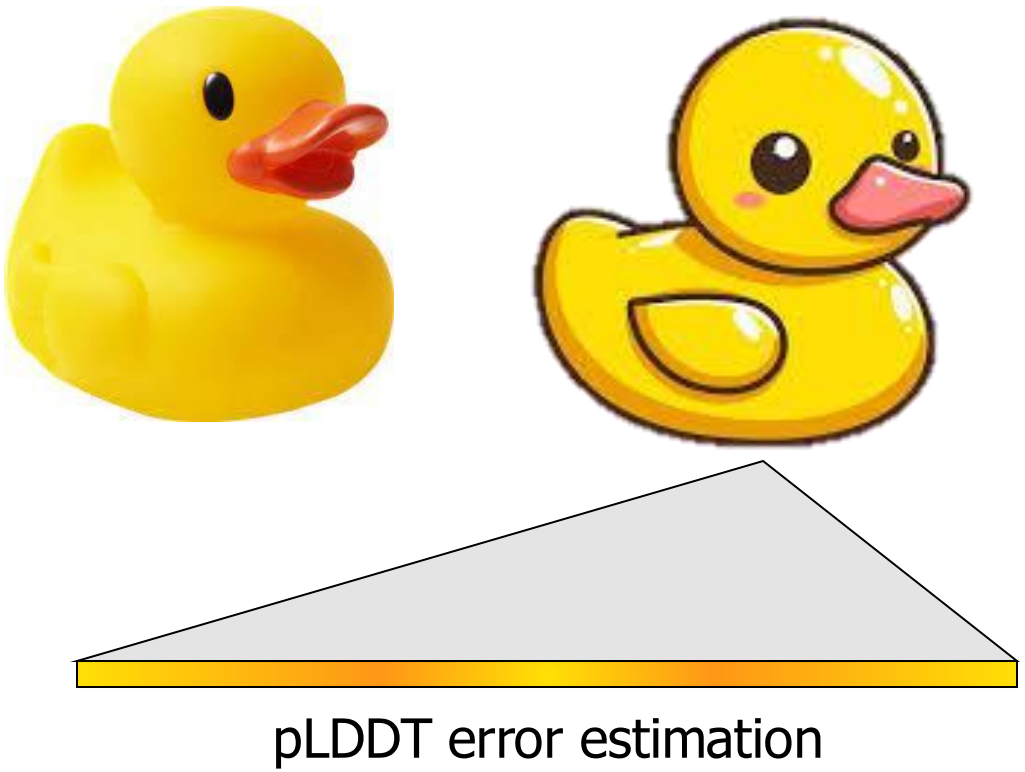


Whenever a component is placed, the variances are reduced, thus increasing the signal-to-noise of the search for the next component



accuracy of the error estimation

- The estimate of the rms between model and target can be critical in structure solution
- Error estimation via pLDDT and incorporated into B-factors of the model
- There is also an error in the error estimation, which requires different estimates of σ_A



pLDDT error estimation

low resolution or incomplete data

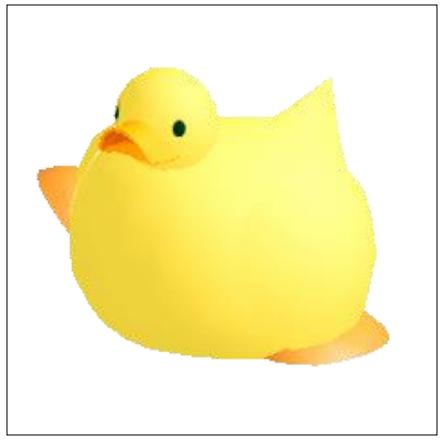
- The likelihood is summed over all reflections
- The fewer the number of reflections the lower the signal

pathologies

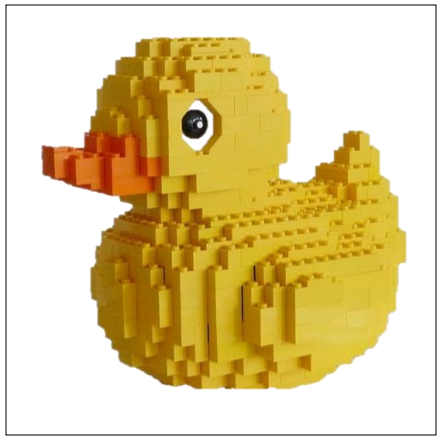
Pathologies

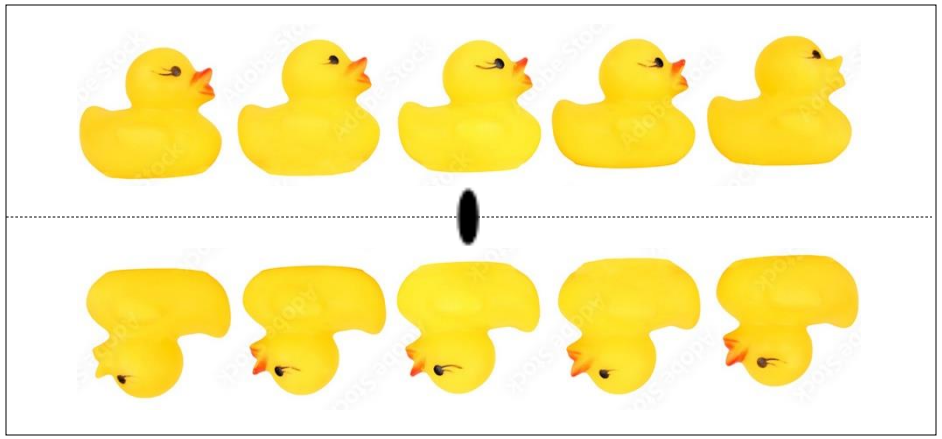
1.Twinning

3. Anisotropy



2. High Mosaicity





4. Low Resolution

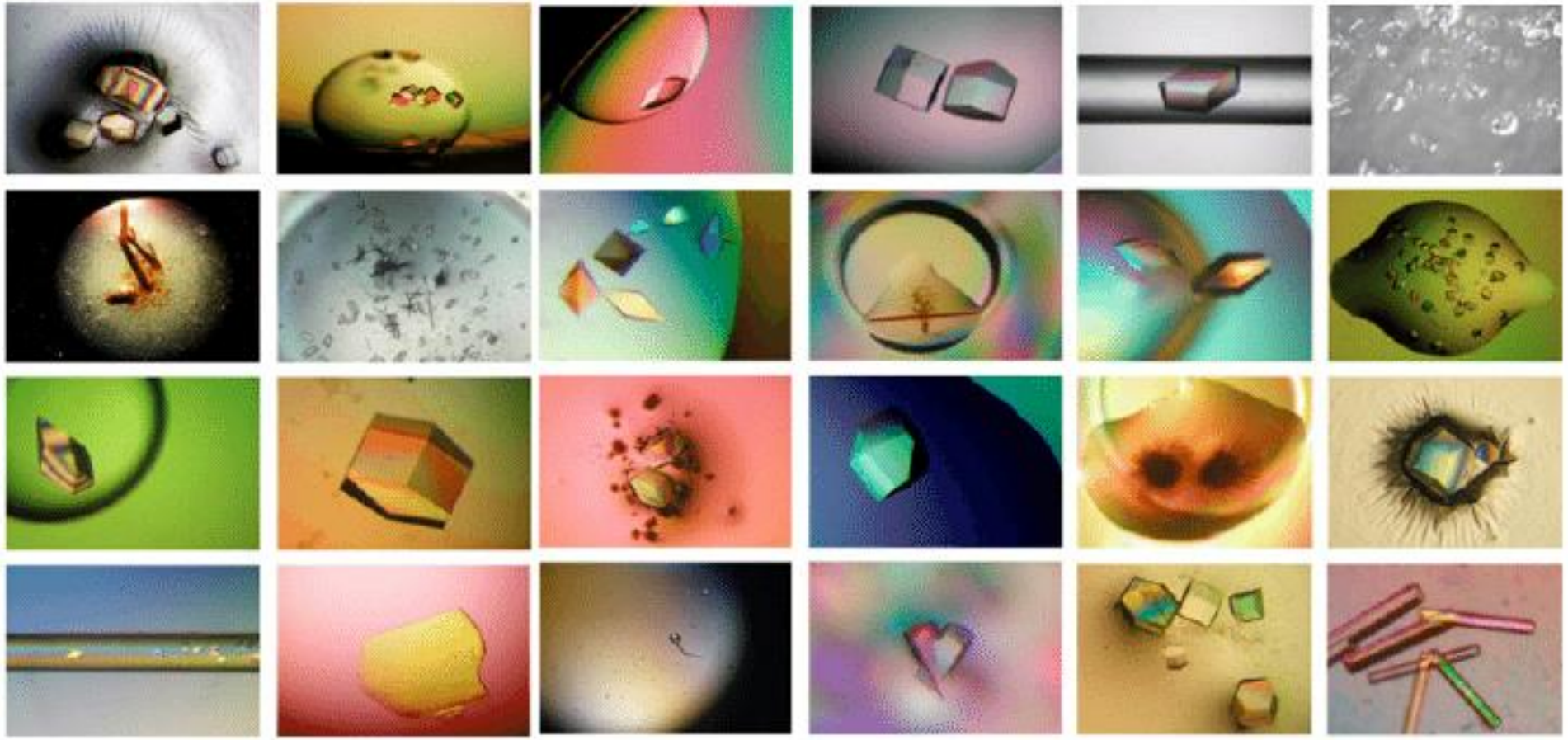
5. Translational non-crystallographic symmetry

Practical Pathologies

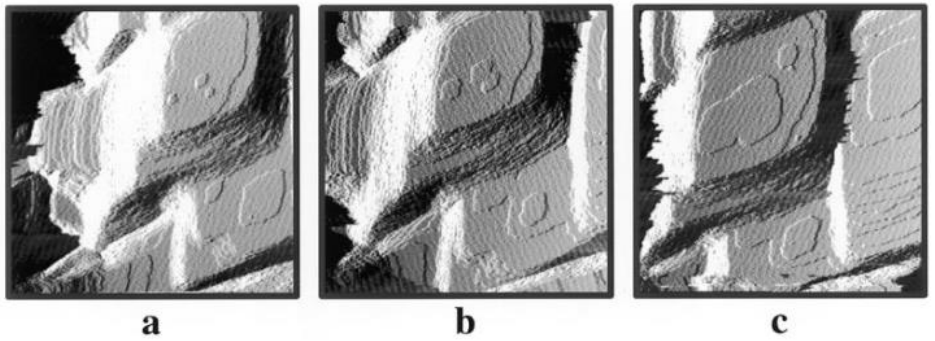
- Some pathologies can only be cured with better crystals or better crystal handling or better data collection strategies
  - High mosaicity
  - Low resolution

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

Rupp's crystal gallery

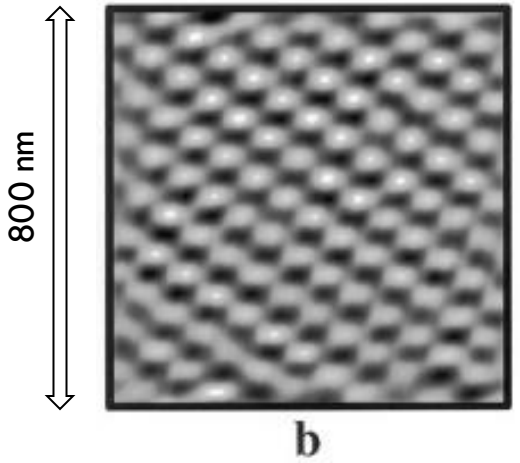


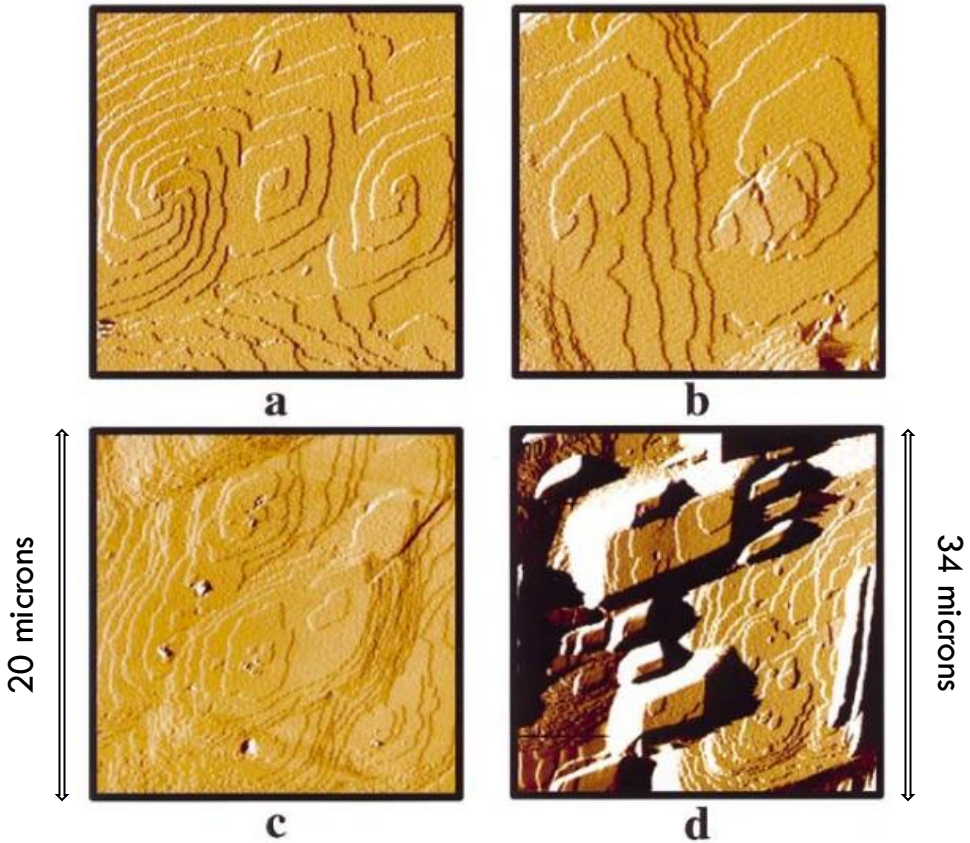
Crystals



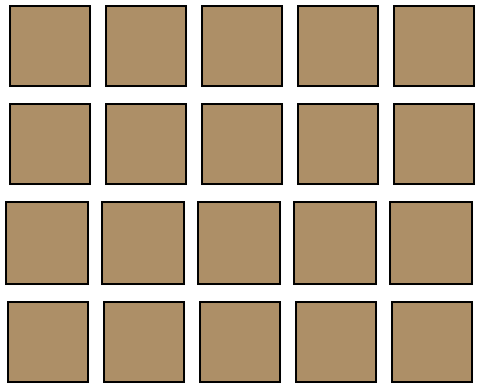
AFM images of yeast $tRNA^{Phe}$

FEBS Journal 280 (2013) 6456–6497

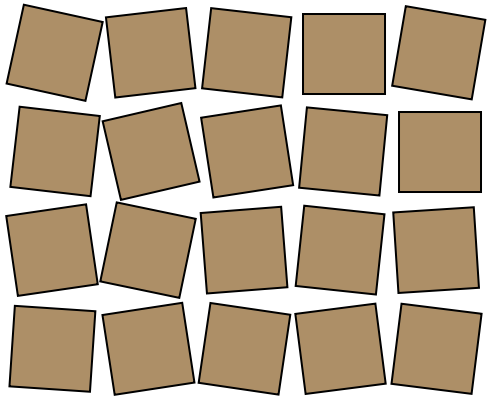




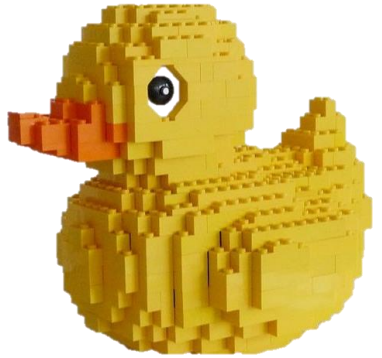
mosaicity and low resolution

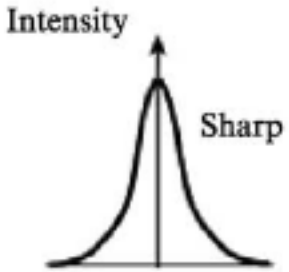


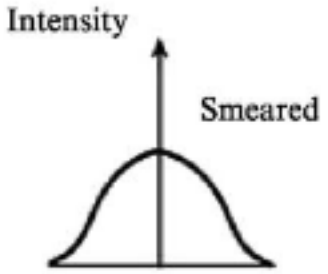
(Almost) Perfect crystals



Real crystals
Mosaic blocks
Highly exaggerated





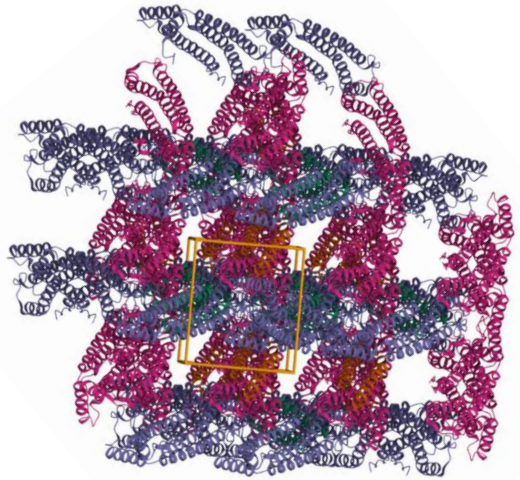


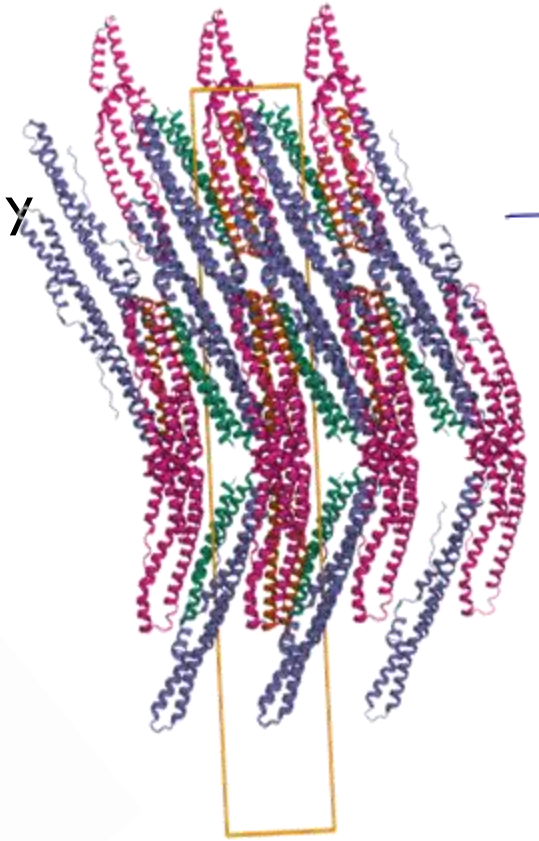
pathologies:
anisotropy

anisotropy correction

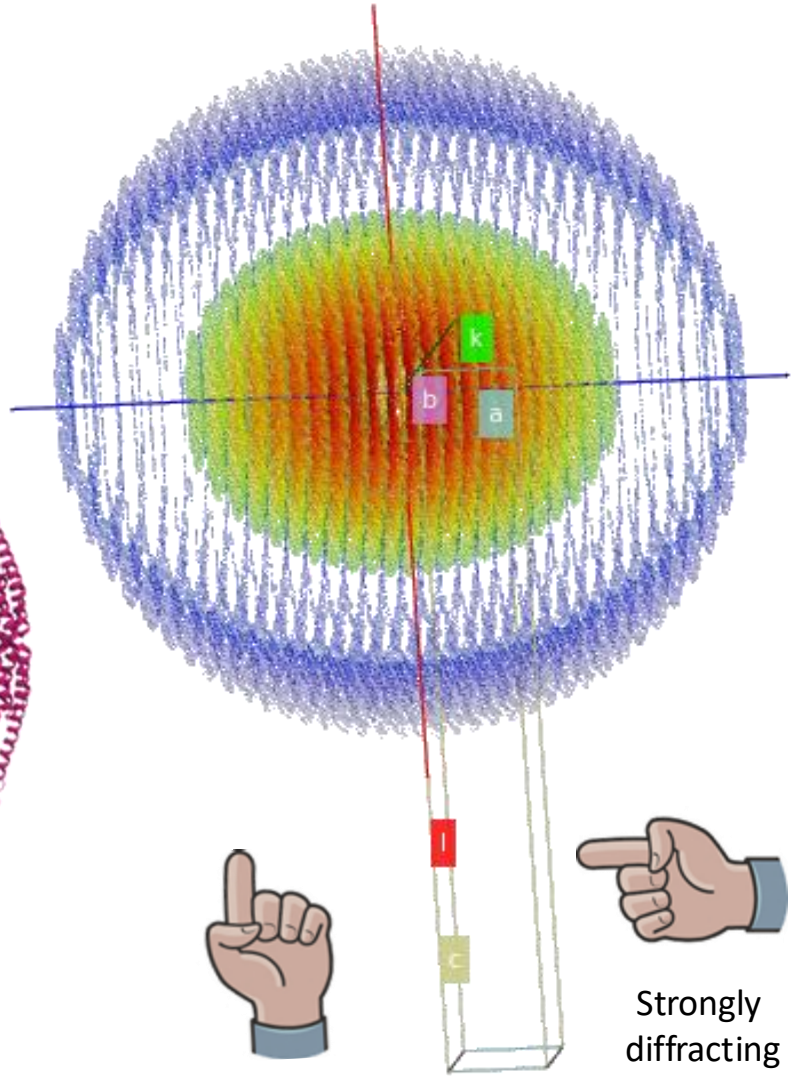
- Many crystals diffract to different resolutions in different directions
- Correct this by making intensity distribution the same in all directions by refining parameters of the anisotropy tensor

2G38





More lateral movement in crystal

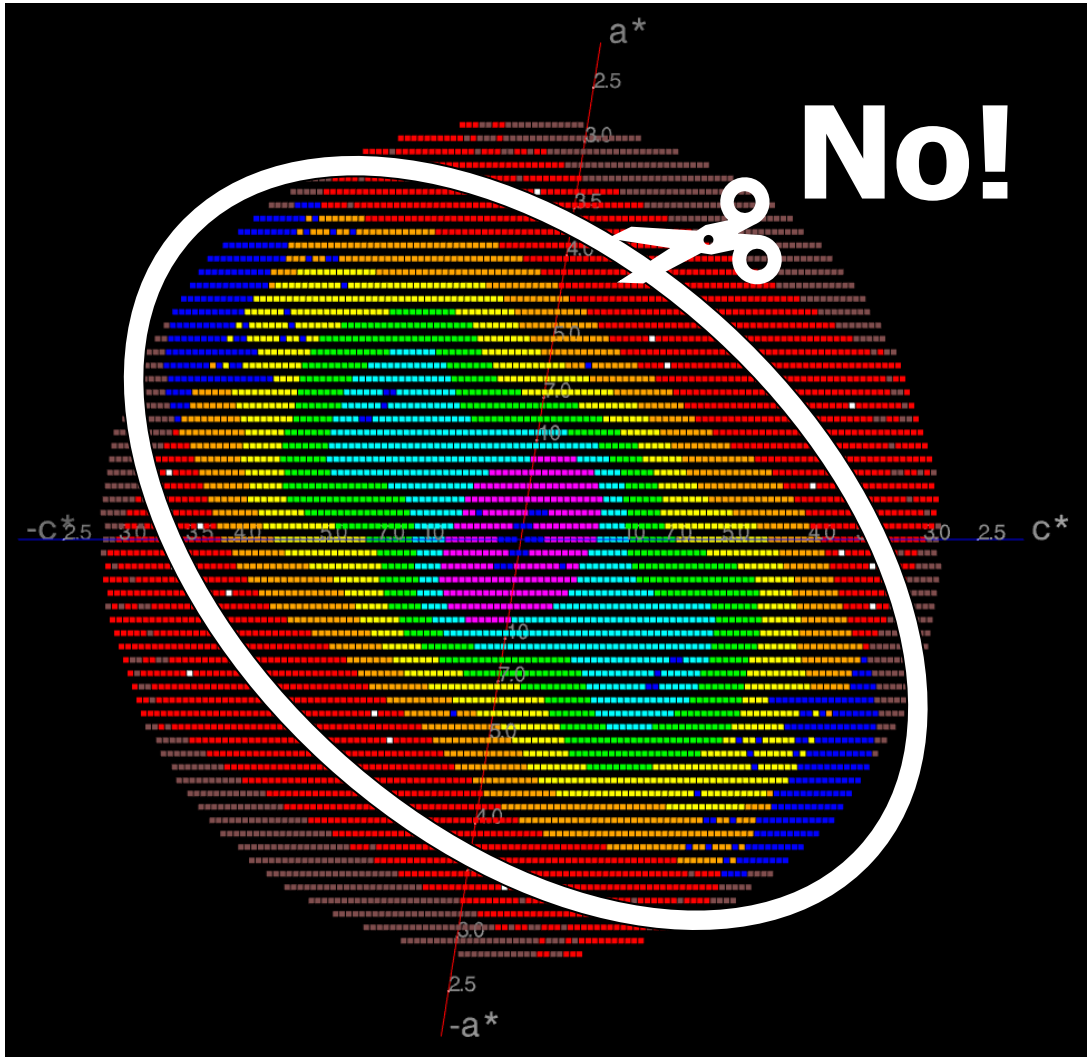


Strongly diffracting

Weakly diffracting

anisotropic data in phaser

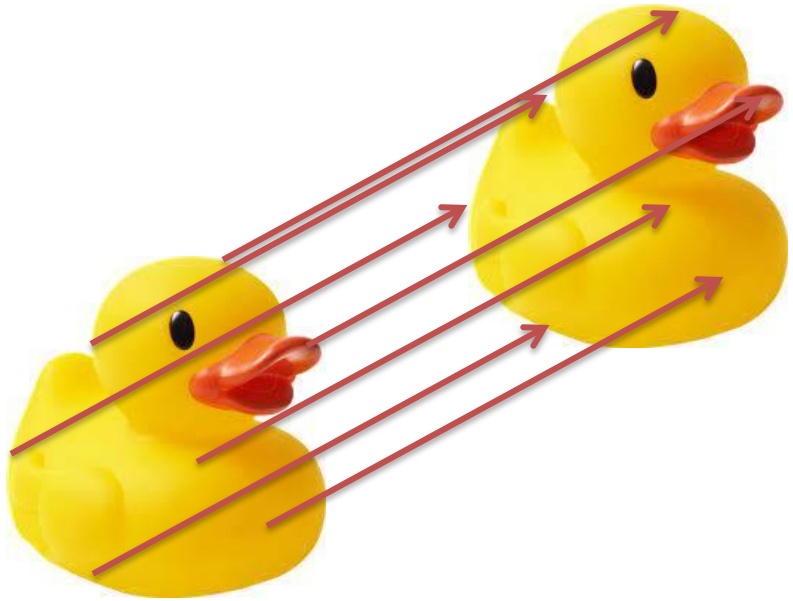
- Do not use anisotropically truncated (staraniso) data in phaser for molecular replacement

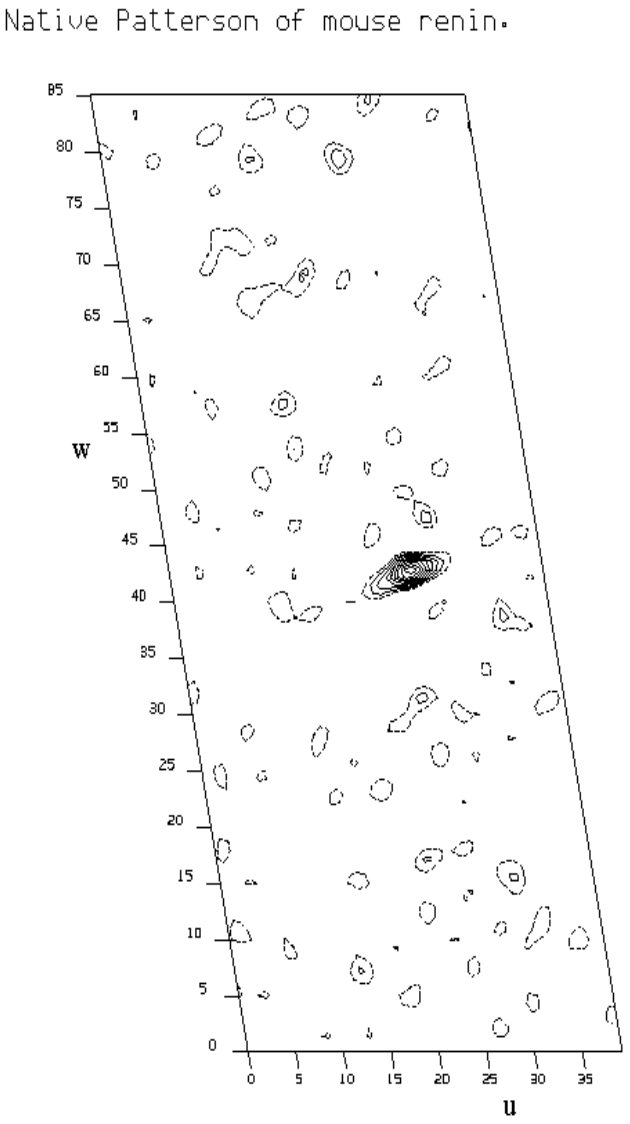


pathologies:
translational non-crystallographic symmetry

translational non-crystallographic symmetry

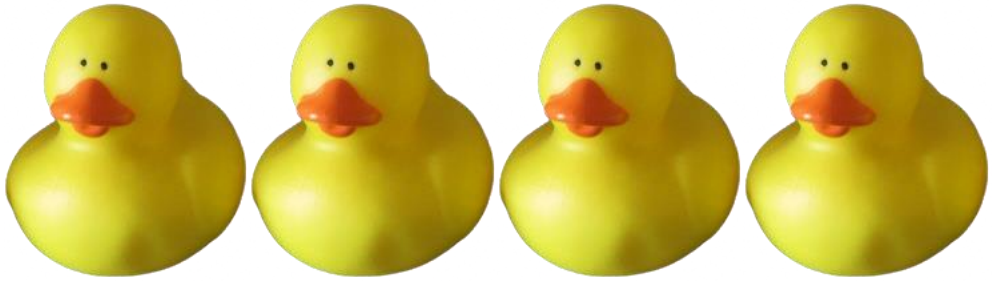
- Patterson is a vector map of the crystal
- Calculated as FT of unphased intensities
- Large origin peak
- TNCS indicated by Patterson Peak
- 16% origin at 5Å



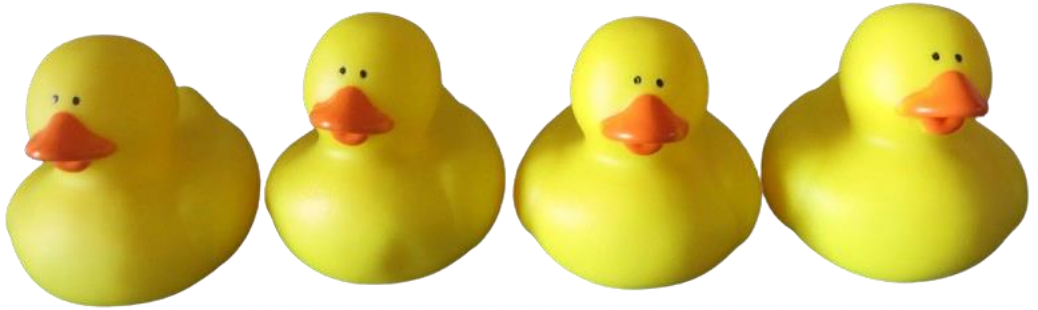


translational non-crystallographic symmetry

- 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?



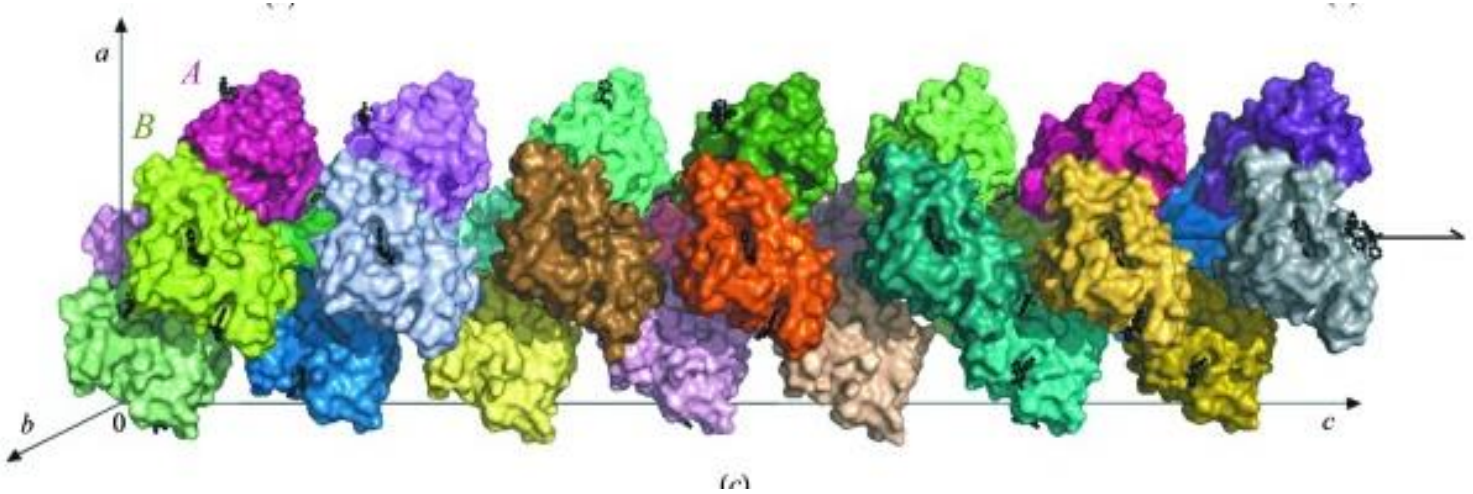
perfect translation

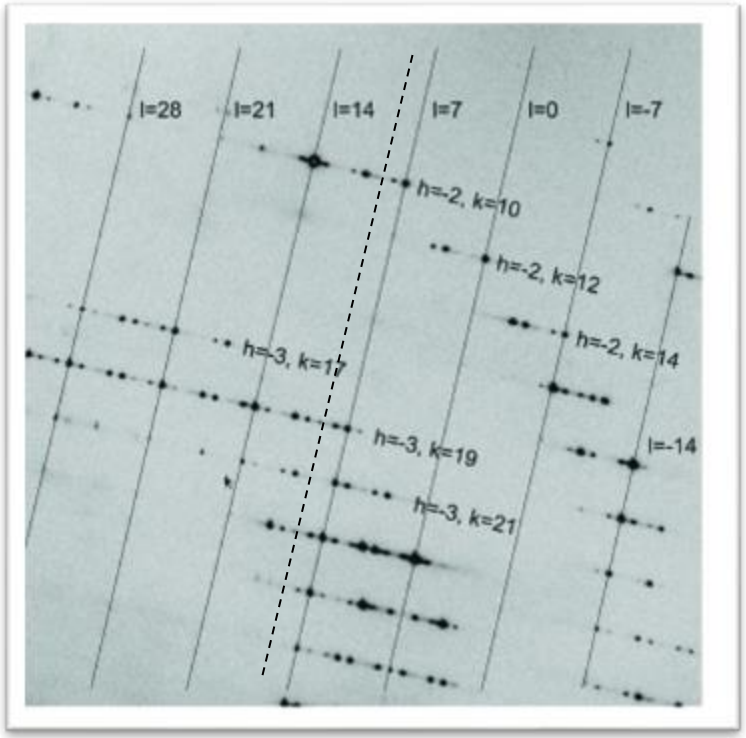


imperfect translation

translational non crystallographic symmetry

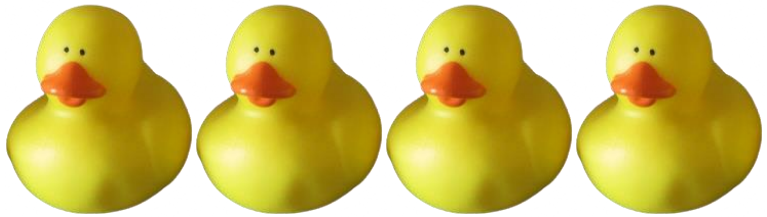
- Make intensity distribution the same in all layers by refining parameters for the expected intensity factors





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

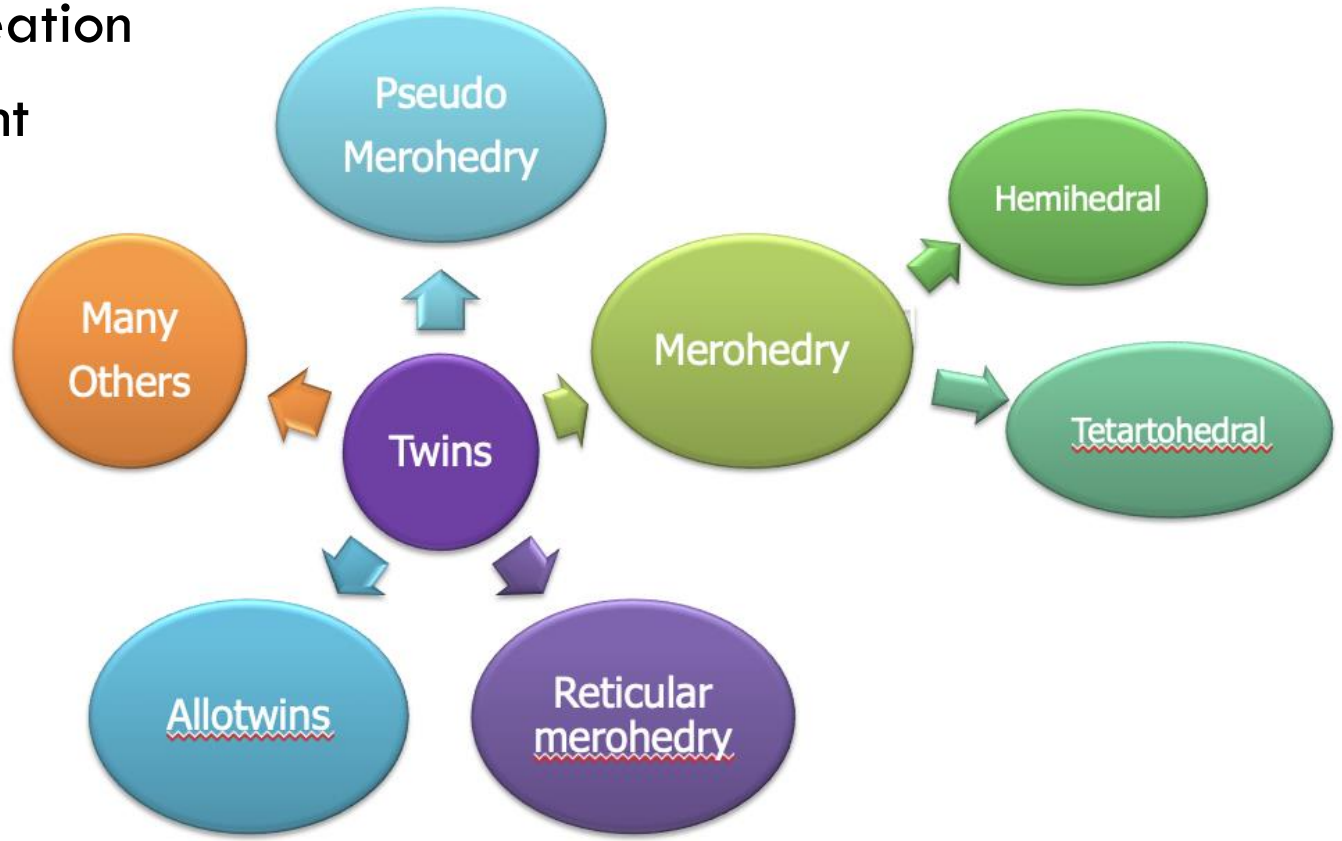
translational non-crystallographic symmetry

- Large Patterson peaks can be caused by order-disorder pathologies so no tncs is always a possibility when there is no peak
- The guess of the tncs order (2,3,4 etc) can be wrong
  - Try other possibilities
- Often associated with twinning
- Often associated with space group ambiguities
  - Space group expansion may be required

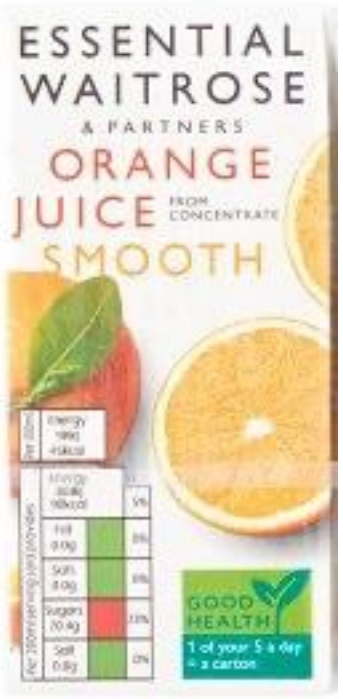
pathologies:
twinning

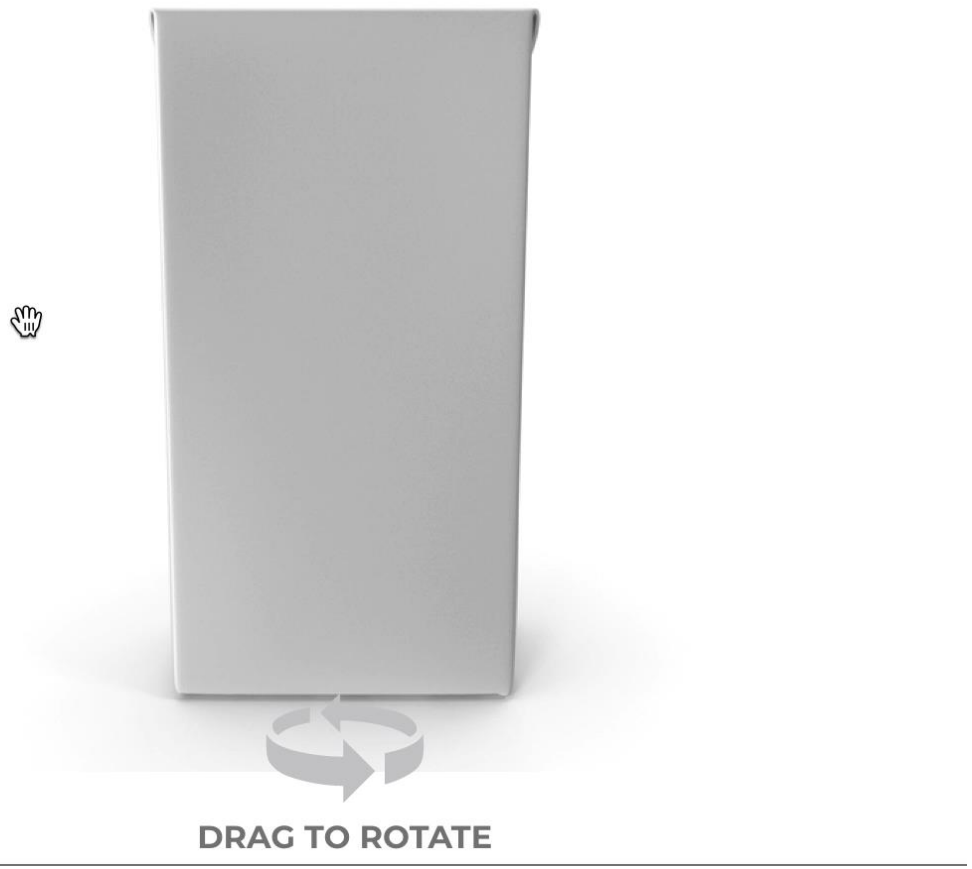
What is Twinning?

- 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

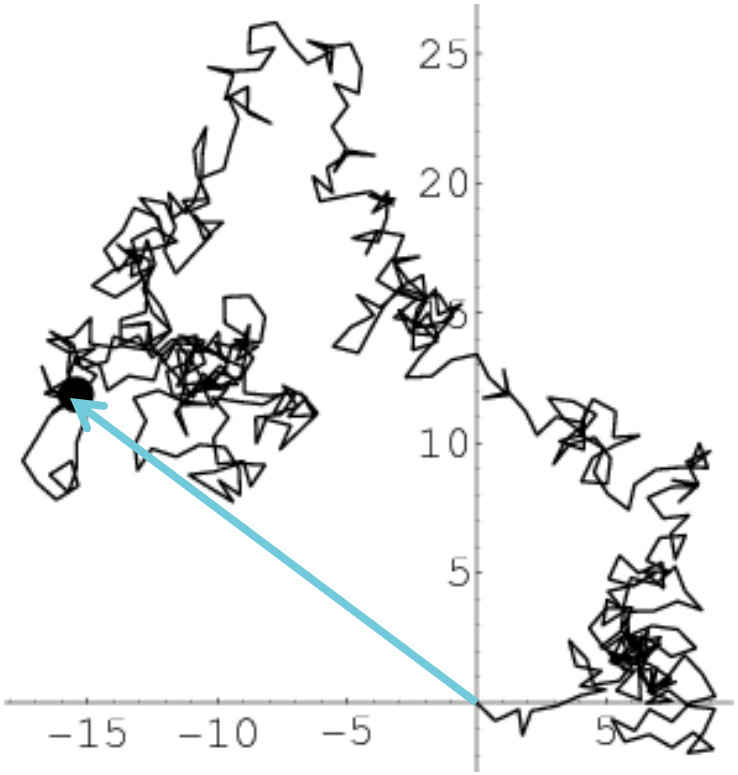




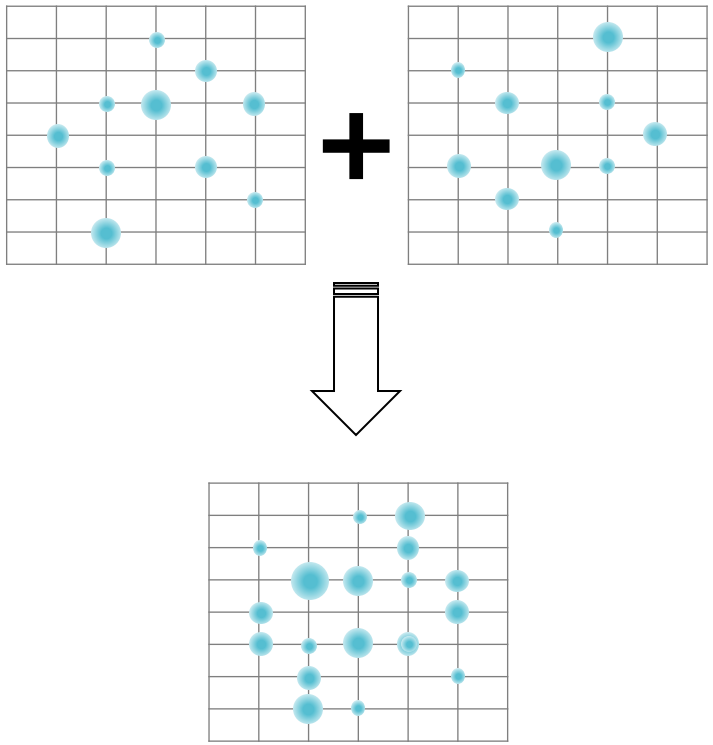
https://elements.envato.com/juice-box-0-5l-slim-KBY64FC

Twinning versus Disorder

Sum of F's



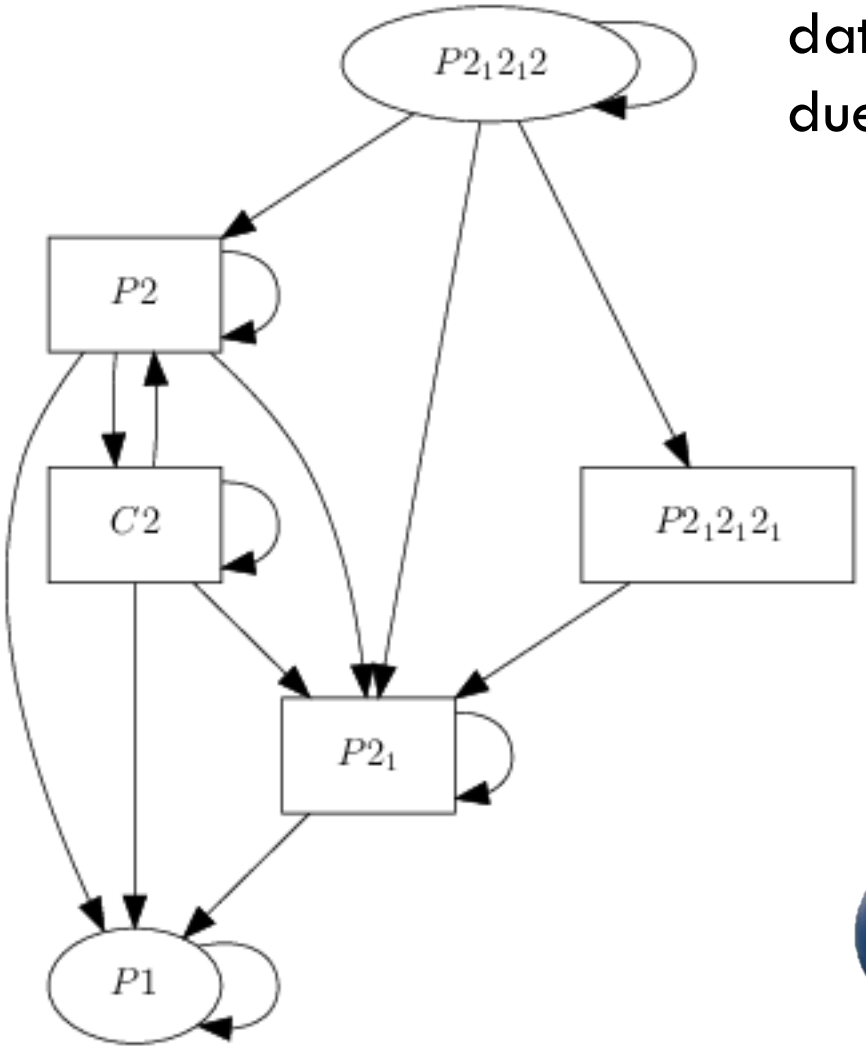
Sum of I's



Twinning versus Disorder

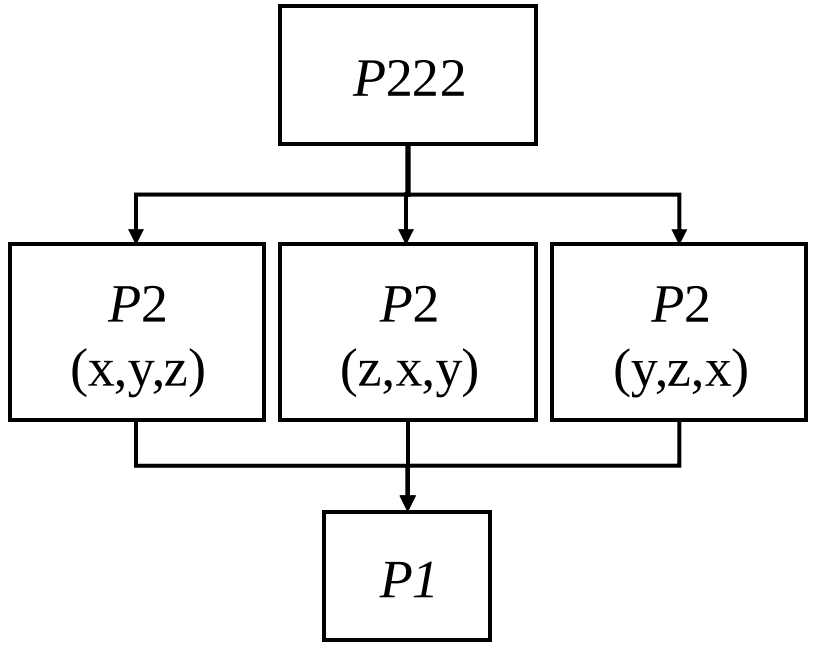
- For twinning, the size of twin domains is large compared with the cell dimensions
  - Diffracted X-rays do not interfere
  - Sum of I's not F's
- This is in contrast to disorder in the crystal where the differences are between neighbouring cells
  - Diffracted X-rays represent the spatially and/or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

subgroups



data over-merged
due to twinning

P2 standard settings
axis permutations

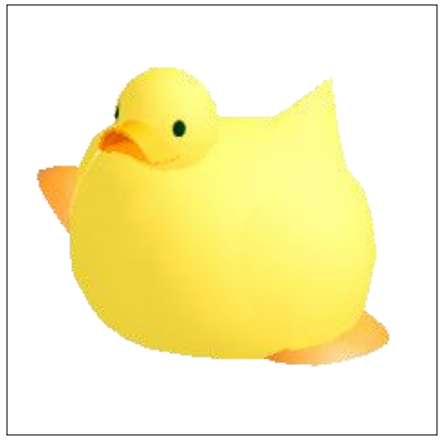


subgroups in Phaser
for molecular replacement

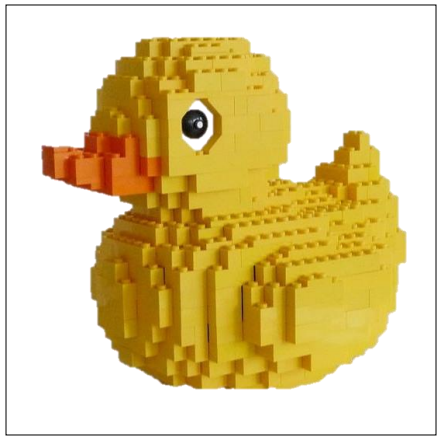
Pathologies

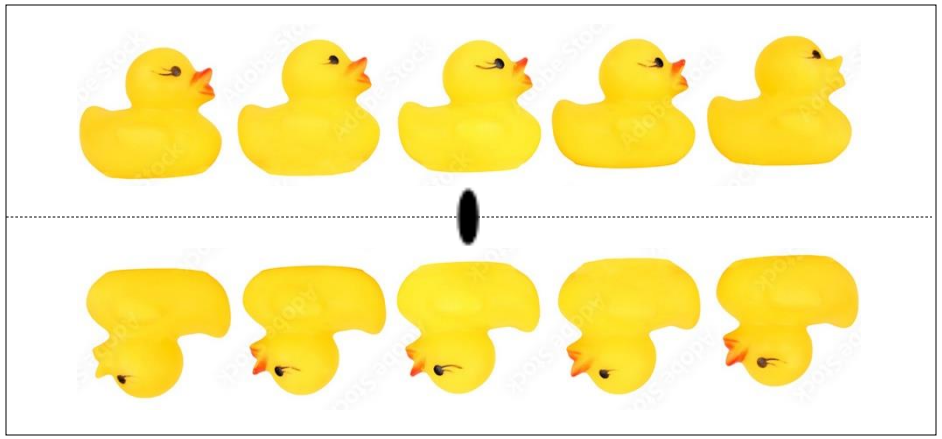
1.Twinning

3. Anisotropy

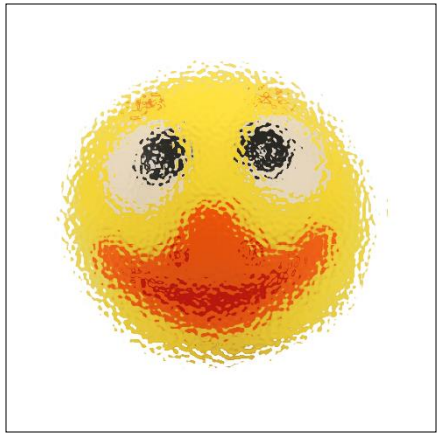


2. High Mosaicity



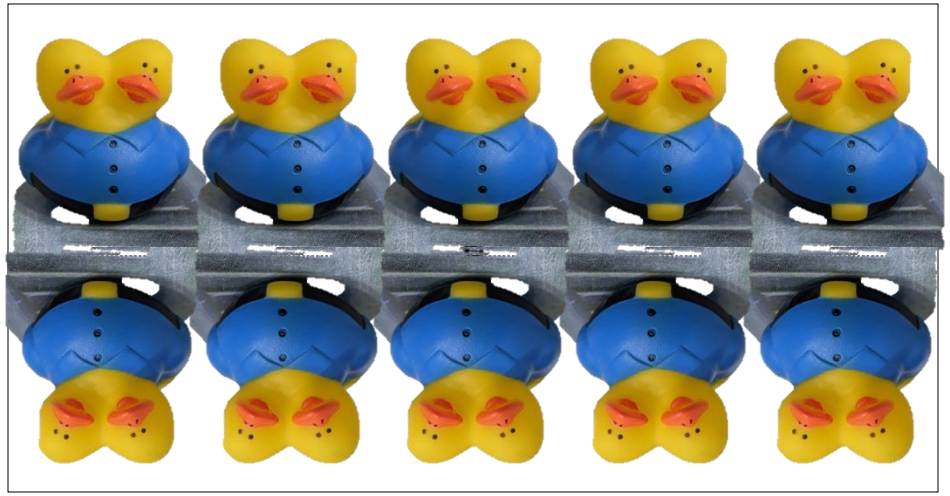


4. Low Resolution



5. Translational non-crystallographic symmetry

Pathologies



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

Phaser Voyager

- Coming soon…

Voyager

Voyager Components

- phasertng
  - “modes”/nodes
  - in c++11 or python
- the dag
  - (directed acyclic graph) schema
- automation scripts (python)

Voyager Scripts

- Scotty
- Changeling
- Xtricorder
- Picard
- Nomad
- Bones
- Riker
- Nanite
- Spock

Picard

Picard is the master molecular replacement protocol. The protocol explores space groups within the same point group, space group subgroups, different translational non-crystallographic symmetry, different cell contents and different estimates of model error. The search space can be configured.

Jean-Luc Picard is commanding officer on Enterprise-D,
the starship of Star Trek the Next Generation

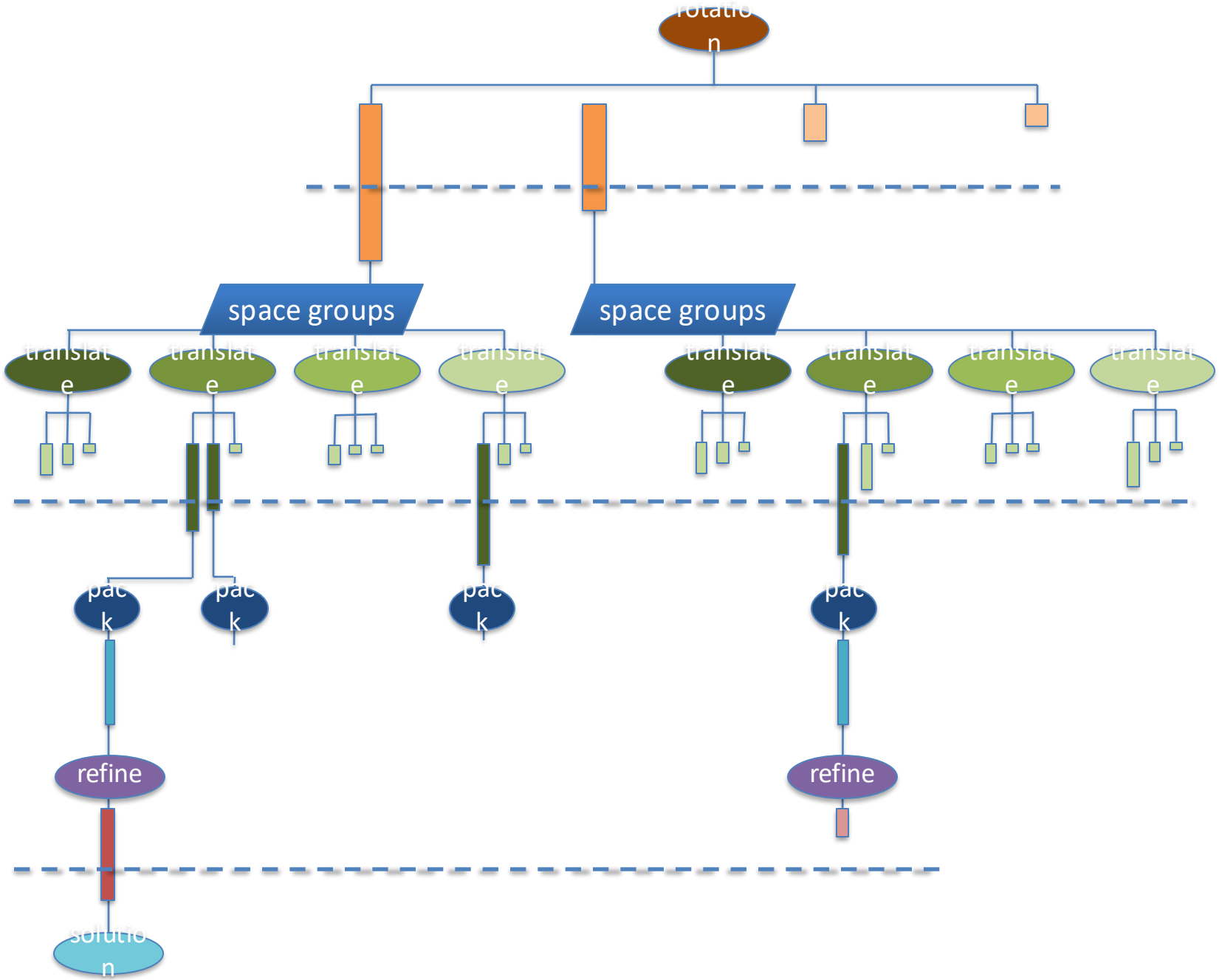


Diagram: Martin Noble

Coming soon…
Phasertng

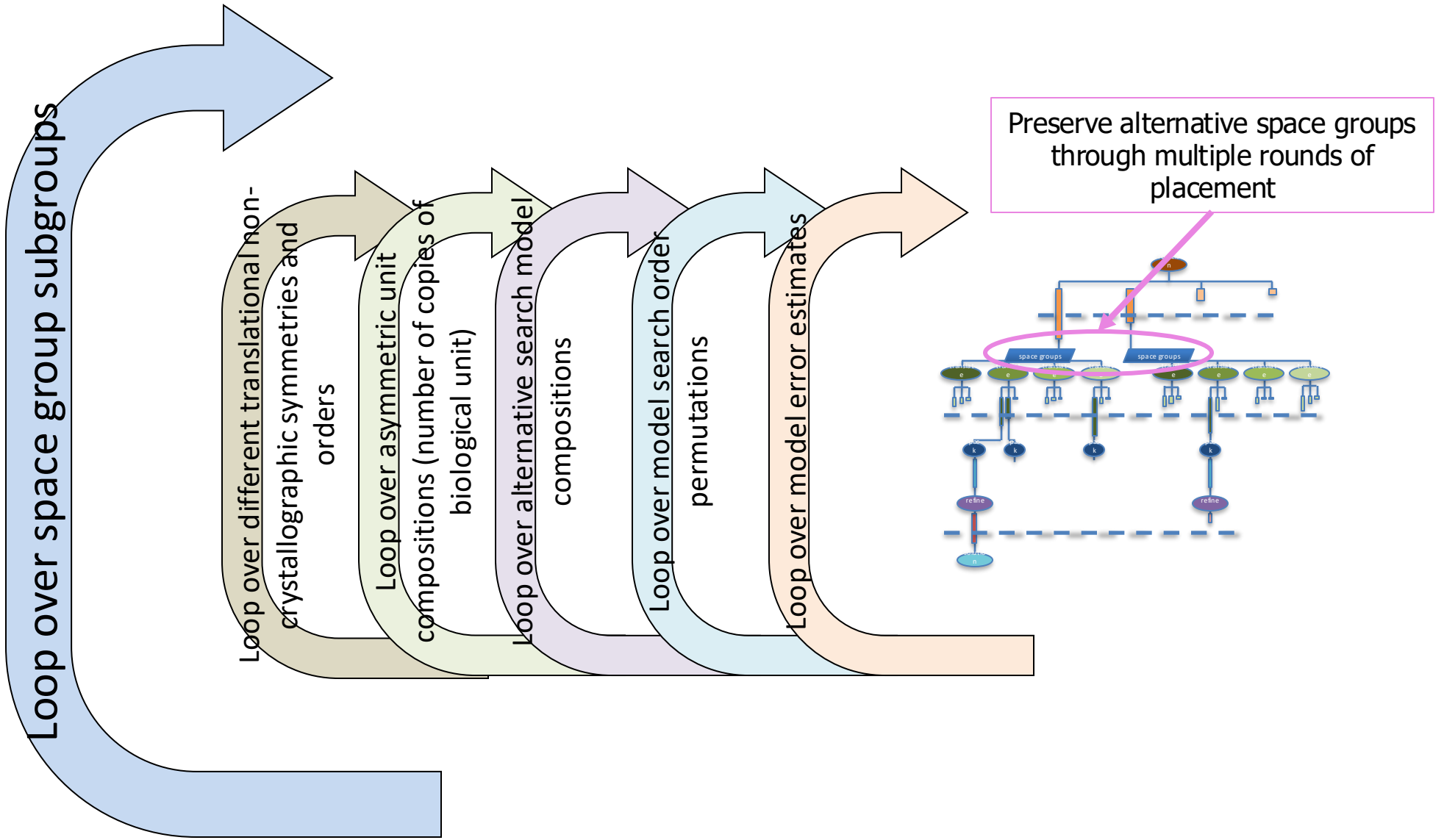
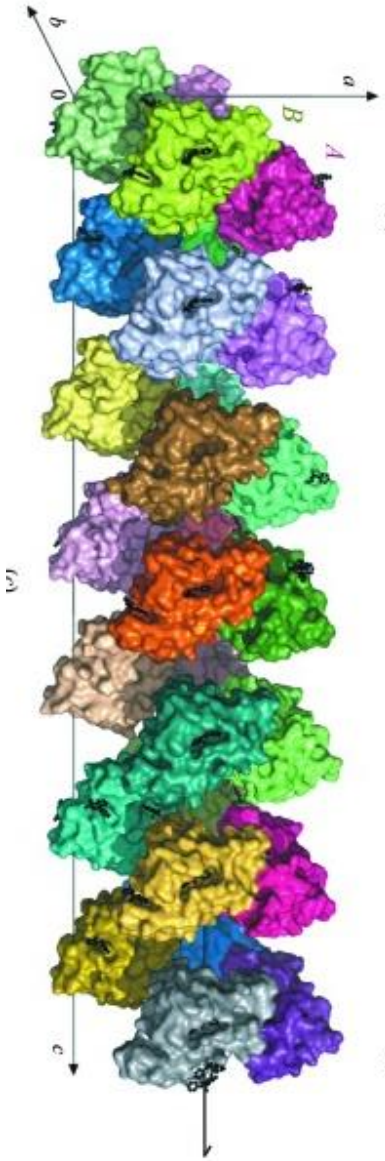


Diagram: Martin Noble

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

- Known monomer structure; good model
- Data merged in P422 with strong 00l ≠ 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



questions