Molecular Replacement

Andrey Lebedev
CCP4

CCP4 MR tutorials

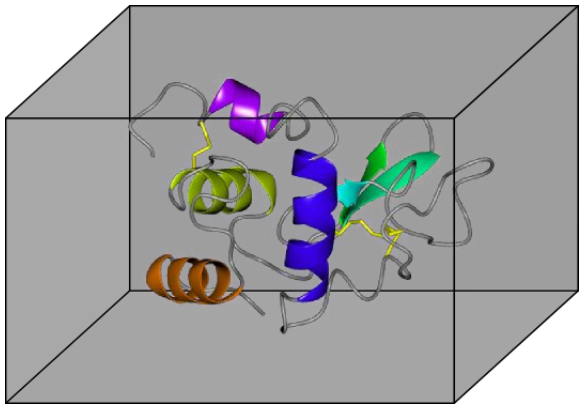
http://www.ccp4.ac.uk/tutorials/

Basic phasing tutorials

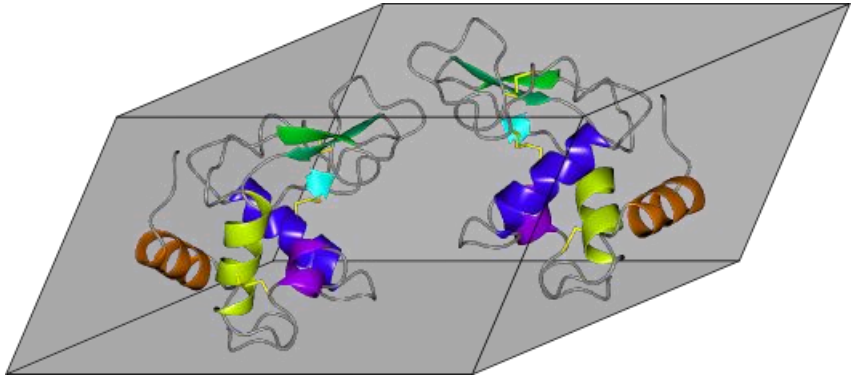
(Includes both MR and Experimental Phasing)

MR Problem

Known crystal structure



New crystal structure

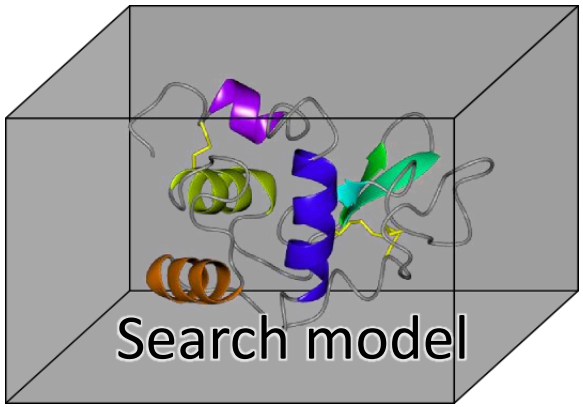


Given:
- Crystal structure of a homologue
- New X-ray data

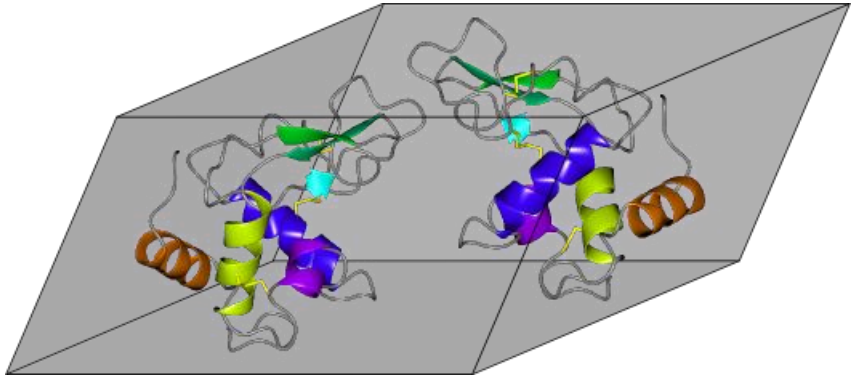
Find:
- The new crystal structure

MR Technique

Known crystal structure



New crystal structure



Method:
- 6-dimensional global optimisation
  - one 6-d search for each molecule in the AU
    >> split further to orientation + translation searches = 3 + 3?

Required:
- Scoring
  - the match between the data and an (incomplete) crystal model
  - ideally: the highest score = the correct model

Real and Reciprocal spaces

Two reasons to discuss the real and reciprocal spaces

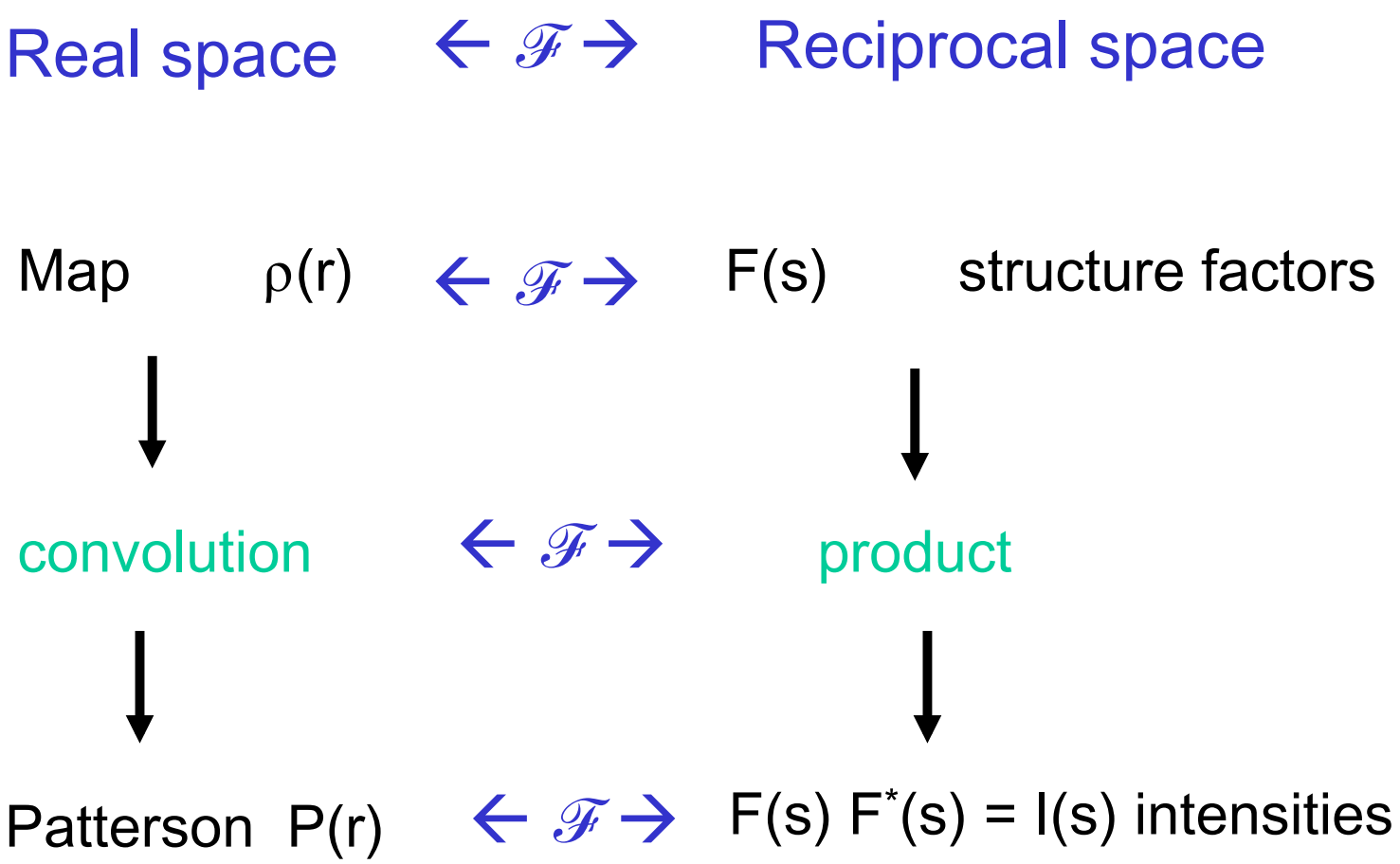
Perception of crystallographic methods

- Many concepts formulated in one space have their counterparts in another space
- The concepts formulated in real space are sometimes more intuitive

Terminology vs. methods

- Sometimes names are taken from real space but assume calculations in the reciprocal space:
  - "Search in the electron density"
  - "Patterson search"

Functions in Real and Reciprocal spaces



Next few slides: *F* and *I* as experimental data

Structure factors and Electron density map

Structure factors $F(h,k,l)$

– A discrete complex function in the reciprocal space

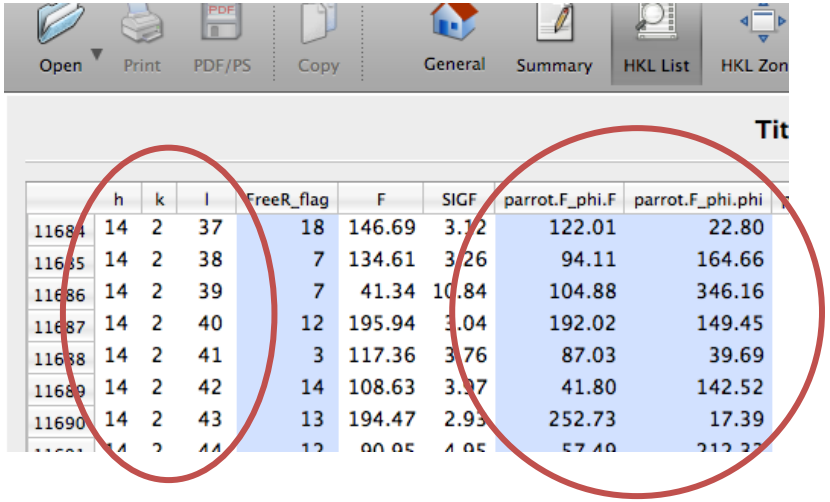
Each *F*:

– Complex number:

$$F = A + iB$$

– Can be expressed via structure amplitude and phase

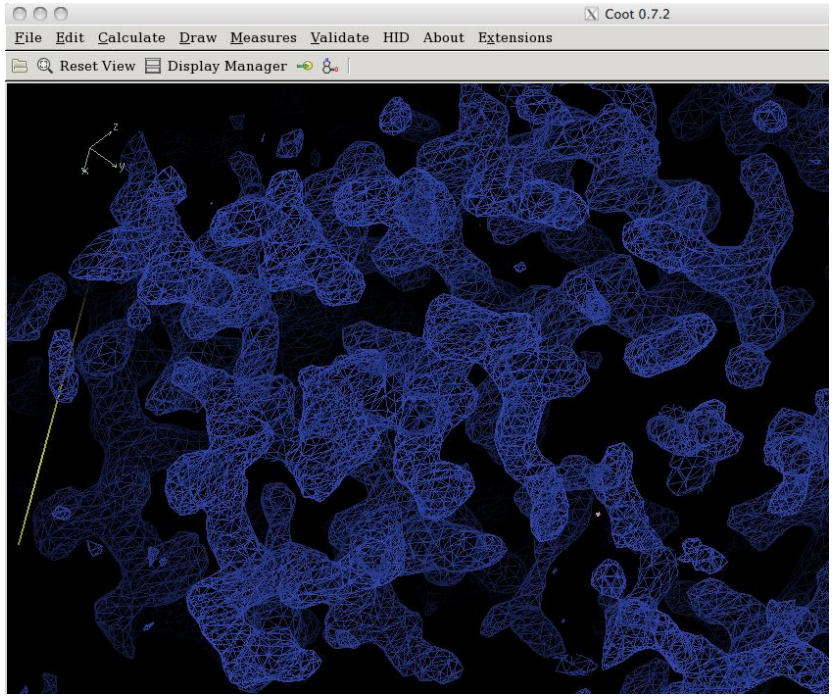
$$F = |F| \exp(if)$$



	h	k	l	FreeR_flag	F	SIGF	parrot.F_phi.F	parrot.F_phi.phi
11684	14	2	37	18	146.69	3.12	122.01	22.80
11685	14	2	38	7	134.61	3.26	94.11	164.66
11686	14	2	39	7	41.34	10.84	104.88	346.16
11687	14	2	40	12	195.94	3.04	192.02	149.45
11688	14	2	41	3	117.36	3.76	87.03	39.69
11689	14	2	42	14	108.63	3.97	41.80	142.52
11690	14	2	43	13	194.47	2.93	252.73	17.39

Electron density map

– periodic 3-d function in real space



is directly interpretable

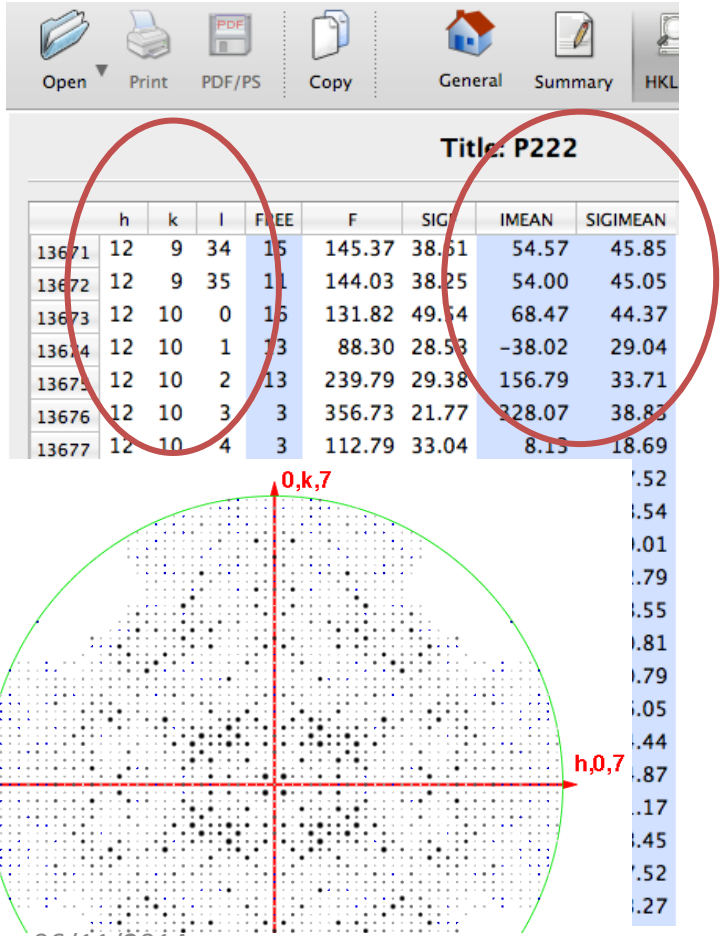
– model building

– real-space fitting of fragments

Intensities and Patterson map

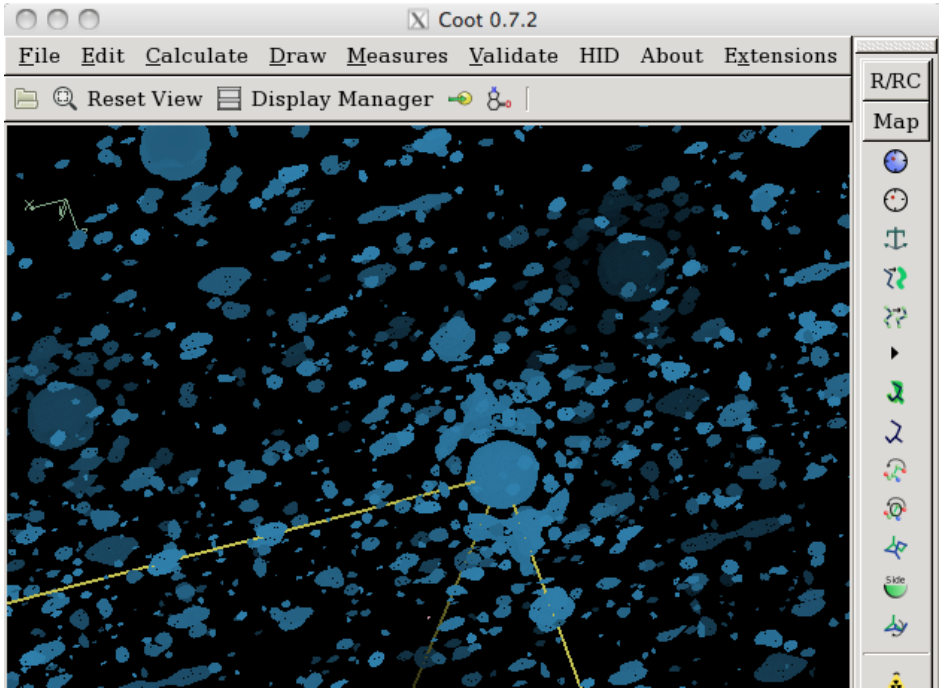
Intensities $I(h,k,l)$

– 3-d discrete real function in the reciprocal space



Patterson map:

– 3-d function in real(*) space



– Nothing reminds a protein molecule

– **Model building**, residue by residue, is **impossible**

Data and MR

MR: Two distinct cases dependent on availability of phases

- Data = structure factors (include phases)
  - "Search in the electron density"
  - Electron density maps are compared: calculated *vs*. observed
  - Can be treated in a more straightforward way (model building)
    - » Useful in special cases

- Data = observed intensities (no phases)
  - "Patterson search"
  - Patterson maps are compared: calculated *vs*. observed
  - Direct model building is impossible in the absence of phases
    - » The most common case of MR

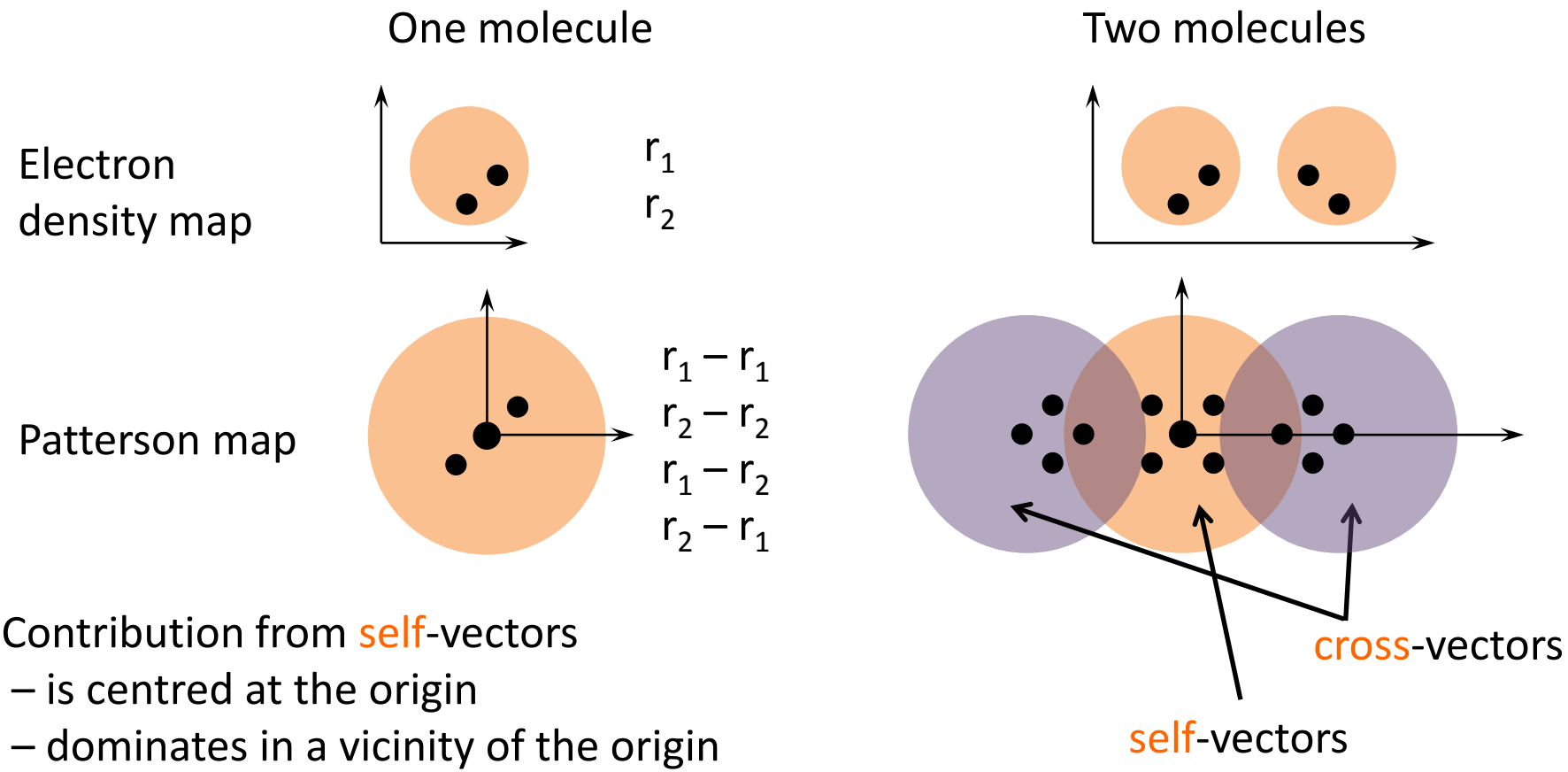
As a rule, all computations are in the reciprocal space

Self and cross vectors

Electron density map = peaks from all atoms

Patterson map = peaks from all interatomic vectors

- self-vectors: vectors between atoms belonging to the same molecule
- cross-vectors: vectors between atoms belonging to different molecules



Contribution from self-vectors
– is centred at the origin
– dominates in a vicinity of the origin

What can be seen to be separated?

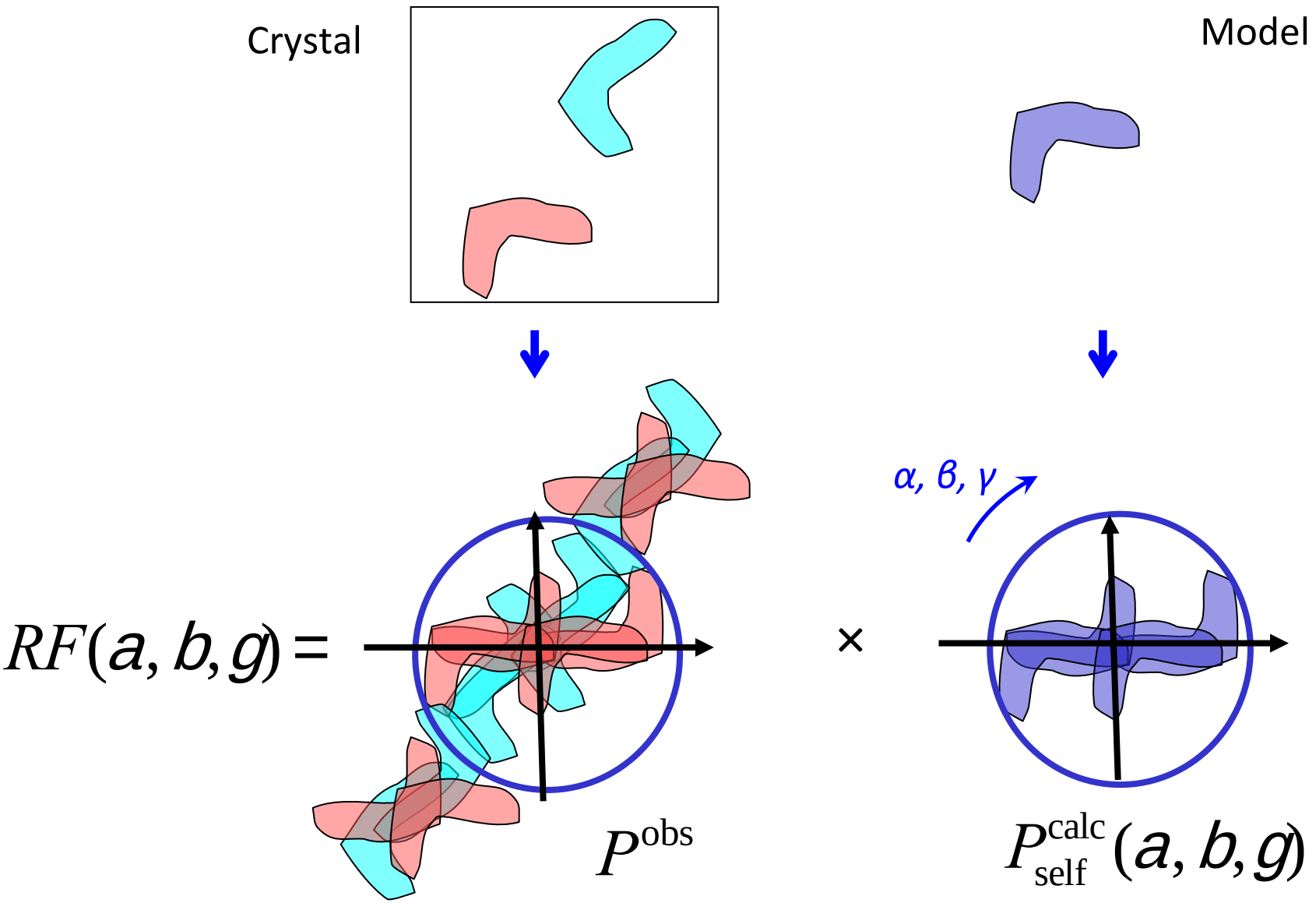
Electron density maps:

- Peaks from atoms or larger fragments are separated in space
  - Model building is possible

Patterson map:

- Contribution from self-vectors is centred at the origin
- Self-vectors are, in average, shorter than cross-vectors
  - Peaks from self-vectors dominates in a vicinity of the origin
  - Peaks from cross-vector dominates away from the origin
- One 6-dimensional search splits into
  - Rotation Function: 3-dimensional search (using self-vectors)
  - Translation Function: 3-dimensional search (using cross-vectors)

Rotation Function



Rotation Function

$$RF(a,b,g) = \iiint P^{obs} \cdot P^{calc}_{self}(a,b,g)\,dr^3$$

$P^{calc}_{self}(a,b,g)$ contains only

- self-vectors

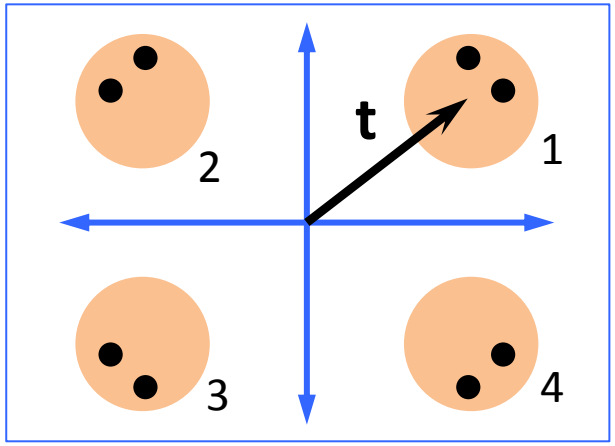
P^{obs} contains

- self-vectors (mostly signal)
- cross-vectors (noise)

Translation Function

Analytical steps

- The centre of molecule 1
  - Parameter t
- Centres of molecules 2, 3 and 4
  - form symmetry operations



- $F_{\mathbf{h}}^{\mathrm{calc}}(\mathbf{t})$, $I_{\mathbf{h}}^{\mathrm{calc}}(\mathbf{t})$
- $TF(\mathbf{t}) = \sum_{\mathbf{h}} I_{\mathbf{h}}^{\mathrm{obs}} \cdot I_{\mathbf{h}}^{\mathrm{calc}}(\mathbf{t})$
- Can be converted to a form:

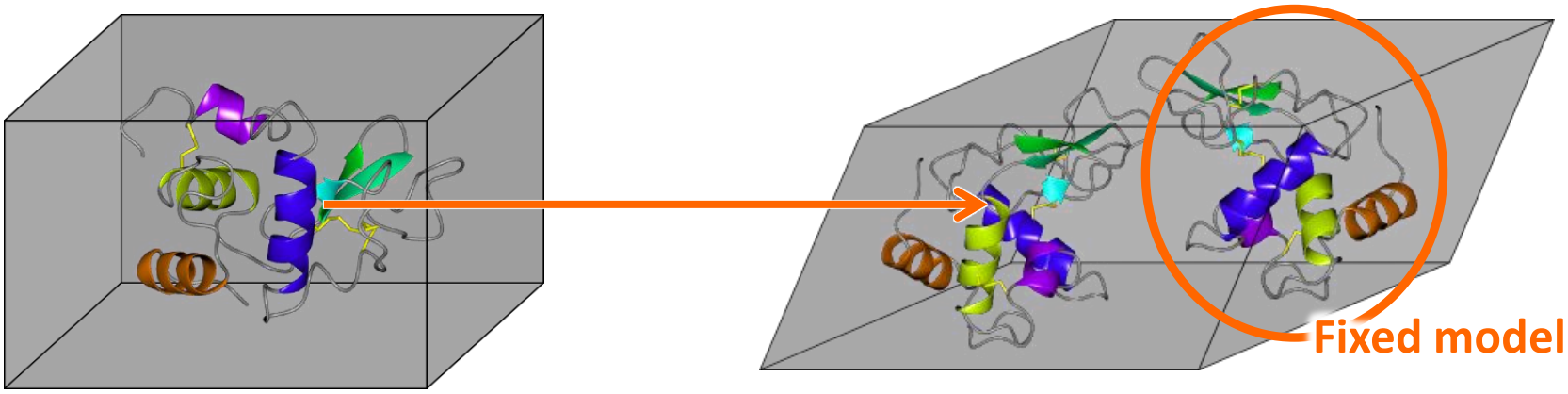
$$TF(\mathbf{t}) = \sum_{\mathbf{h}} G_{\mathbf{h}} \exp(2\pi i\, \mathbf{h} \times \mathbf{t})$$

  - FFT techniques can be applied

Numerical calculations

- $G_{\mathbf{h}}$
- FFT: $G_{\mathbf{h}} \rightarrow TF(\mathbf{t})$
- Peak search in $TF(\mathbf{t})$: best **t**

Fixed partial model



Almost the same equation as for a single molecule search,

$$TF(\mathbf{t}) = \sum_{\mathbf{h}} I_{\mathbf{h}}^{\text{obs}} \cdot \left| F_{\mathbf{h}}^{\text{fixed}} + F_{\mathbf{h}}^{\text{calc}}(\mathbf{t}) \right|^2$$

Again, can be converted to a form:

$$TF(\mathbf{t}) = \sum_{\mathbf{h}} G_{\mathbf{h}} \exp(2\pi i\, \mathbf{h} \times \mathbf{t})$$

and FFT technique can be used

Translation Function

$$TF(\mathbf{t}) = \iiint P^{\text{obs}} \cdot P^{\text{calc}}_{\text{cross}}(\mathbf{t})\,\mathrm{d}r^3$$

$P^{\text{calc}}_{\text{cross}}(\mathbf{t})$ contains only

- cross-vectors

P^{obs} contains

- self-vectors (noise)
- cross-vectors (mostly signal)

Packing considerations

Molecules in the crystal do not overlap

How can we use this information?

» Patterson map does not explicitly reveal molecular packing

Reject MR solutions

- Restrict distance between centres of molecules
- Count close interatomic contacts

Modify TF

- Divide by Overlap Function
- Multiply by Packing function

Fast Packing Function

Estimation of overlap

- Using mask from search model
- FFT

Packing Function

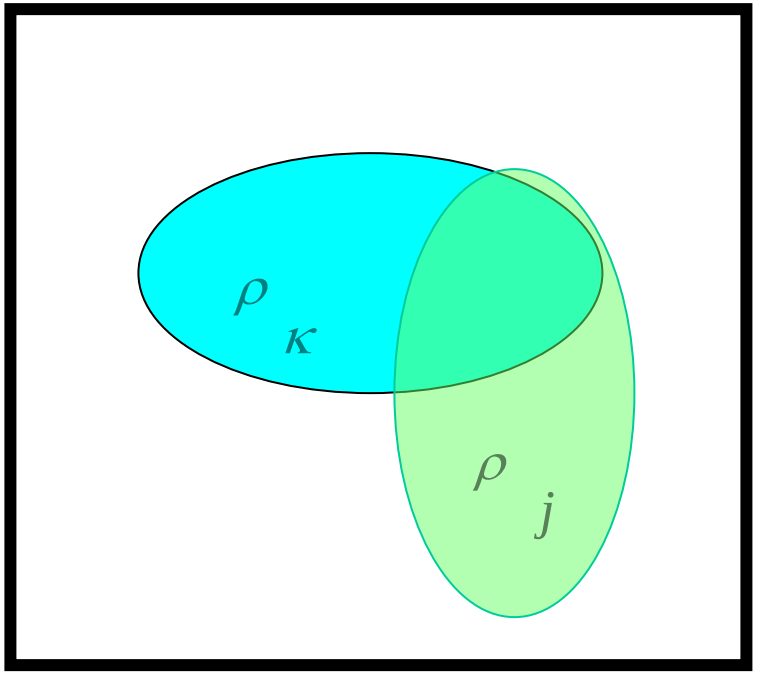
$PF = 1 - \text{overlap}$

Modified Translation Function

$TF' = TF \times PF$

Peak search

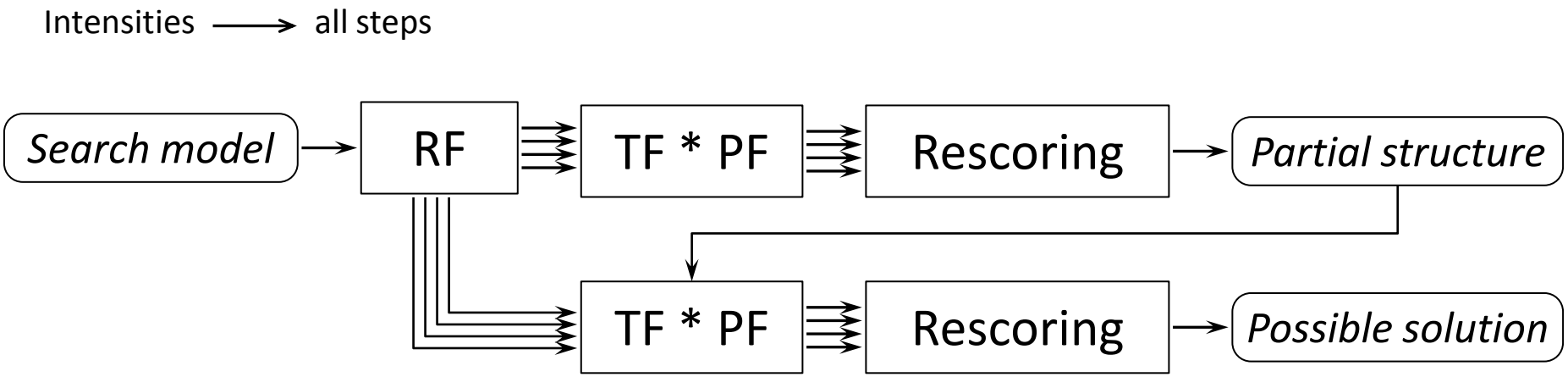
- Using TF'
- No irrelevant peaks are passed to rescoring step



Implemented in *MOLREP*

A conservative MR protocols for two copies of a model

As implemented in *MOLREP*



RF = $\sum_{hkl} w * I_O * I_C(\alpha\beta\gamma)$

TF = $\sum_{hkl} w * I_O * I_C(xyz)$

Rescoring: Correlation Coefficient* PF

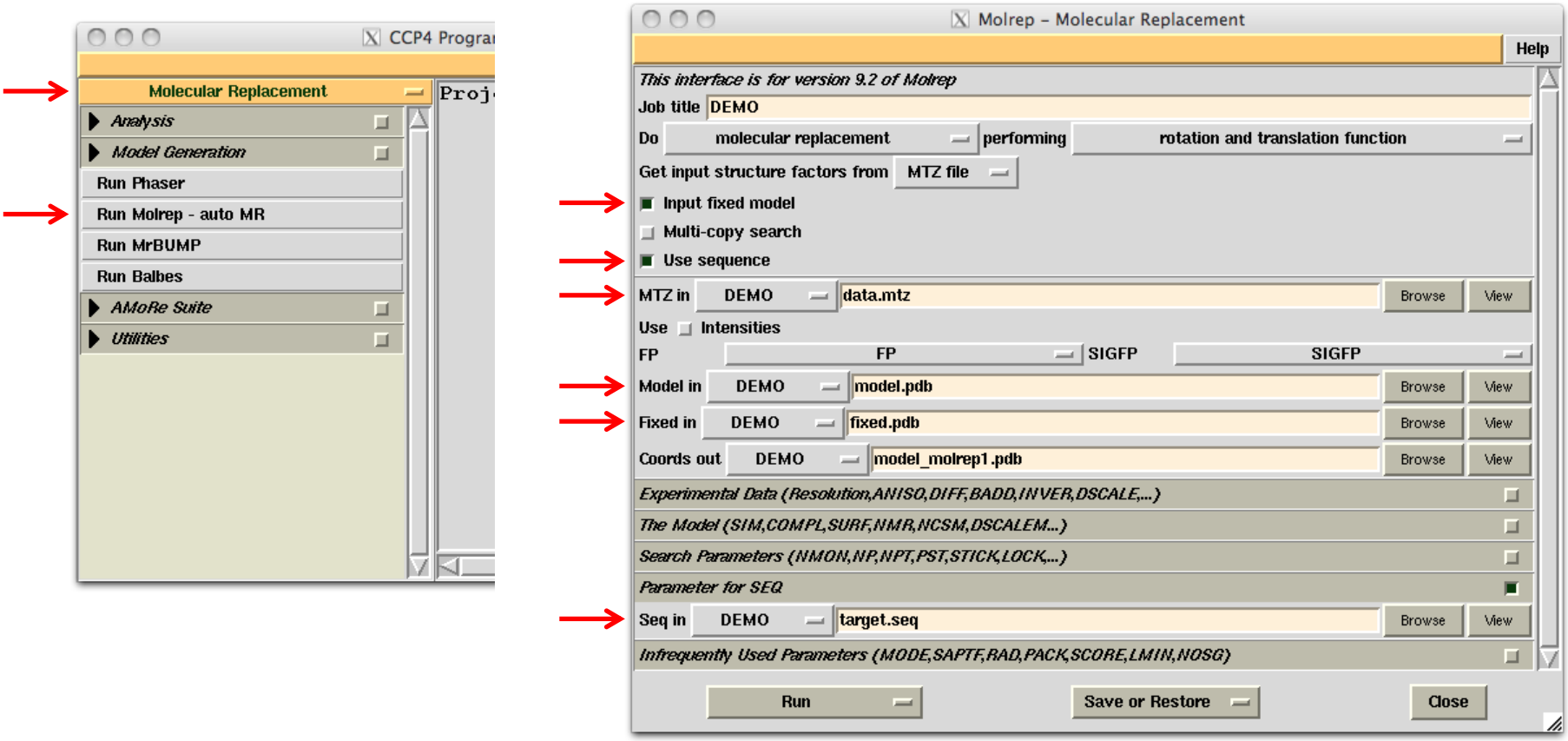
Molrep

Alexey Vagin

YSBL University of York

Molrep

```
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
```



Molrep default protocol

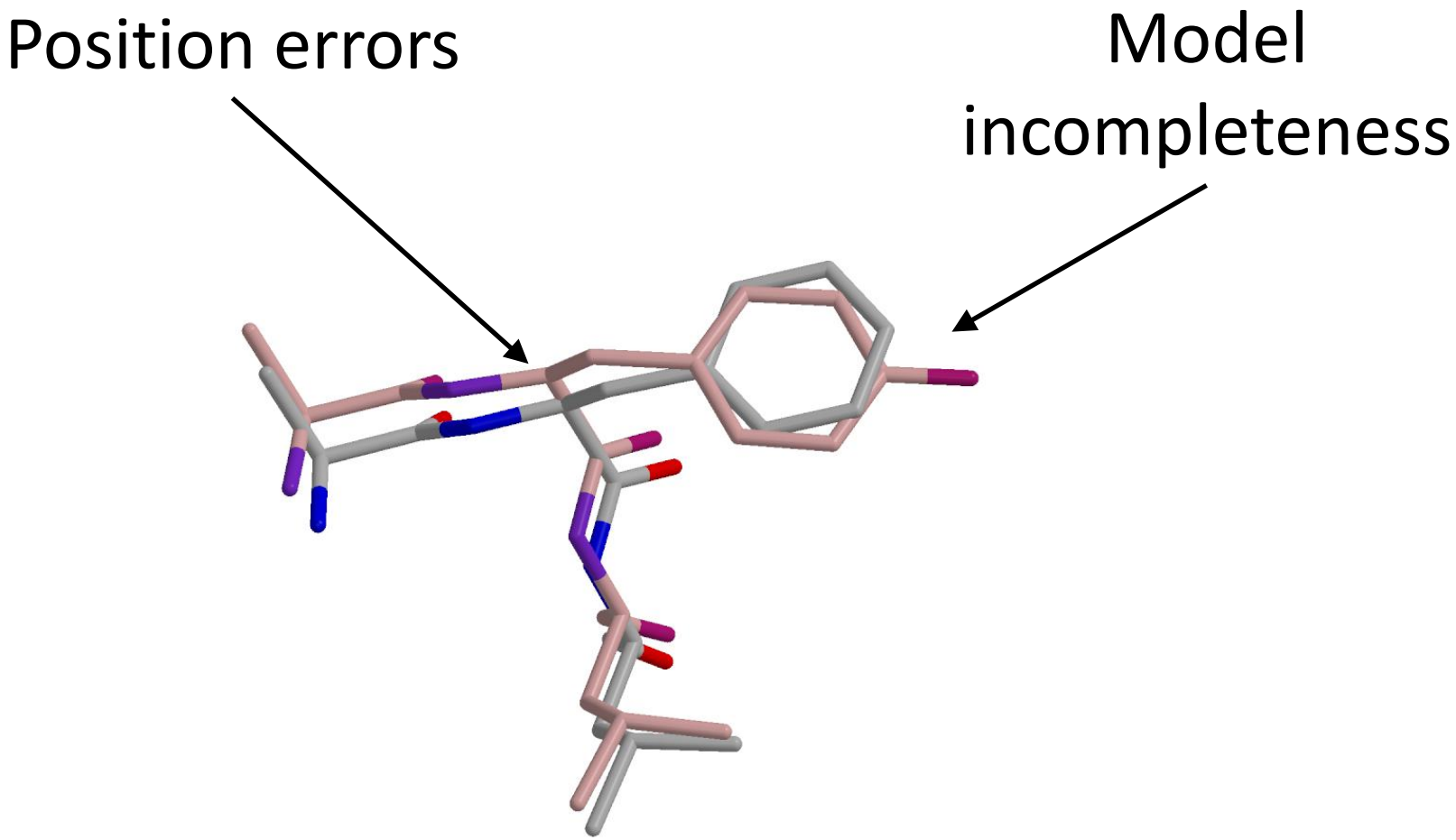
```
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
```

- model correction if sequence provided
- defines the number of molecules per AU
- modification of the model surface
- anisotropic correction of the data
- weighting the data according to model completeness and similarity
- check for pseudotranslation and use it if present
- 30+ peaks in Cross RF for use in TF (accounts for close peaks)
- applied packing function
- make use of partial structure (fixed model)

Molecular Replacement

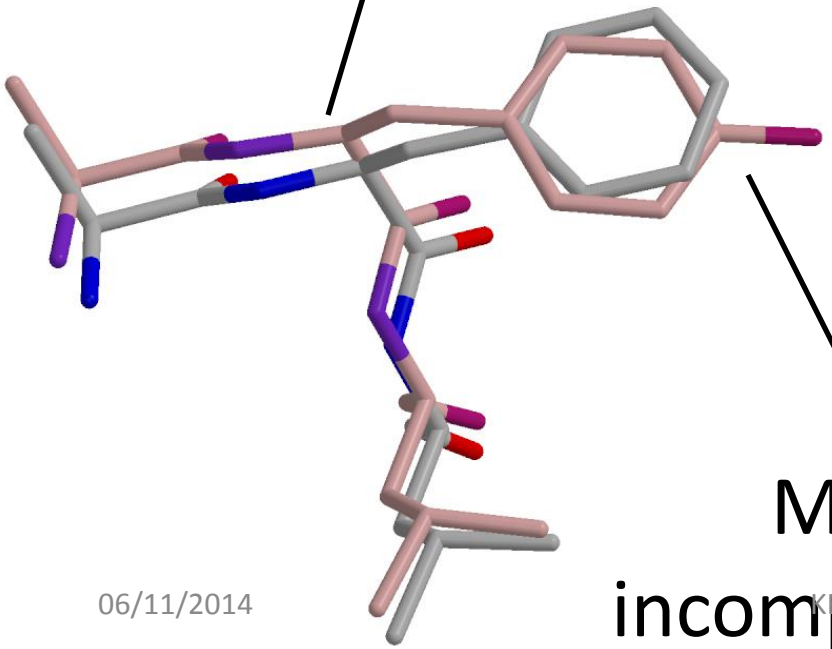
Gábor Bunkóczi

Model error

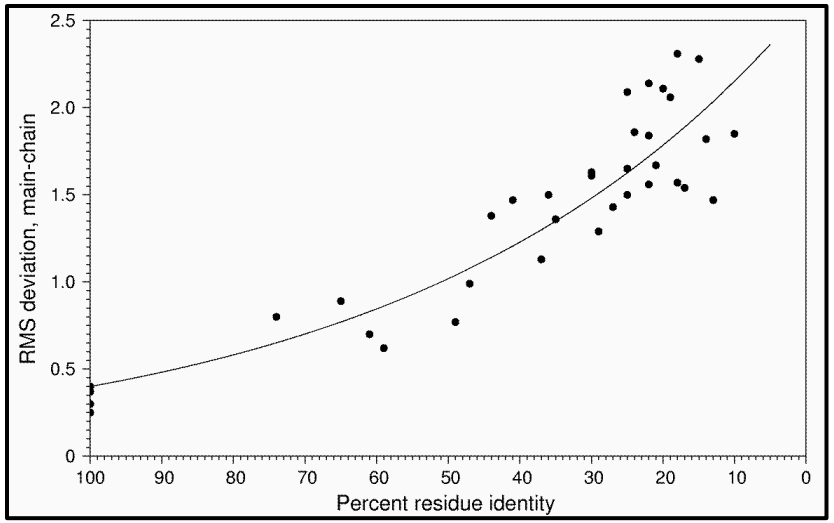


Model error

Position errors



Calculate from sequence identity (Chothia & Lesk)

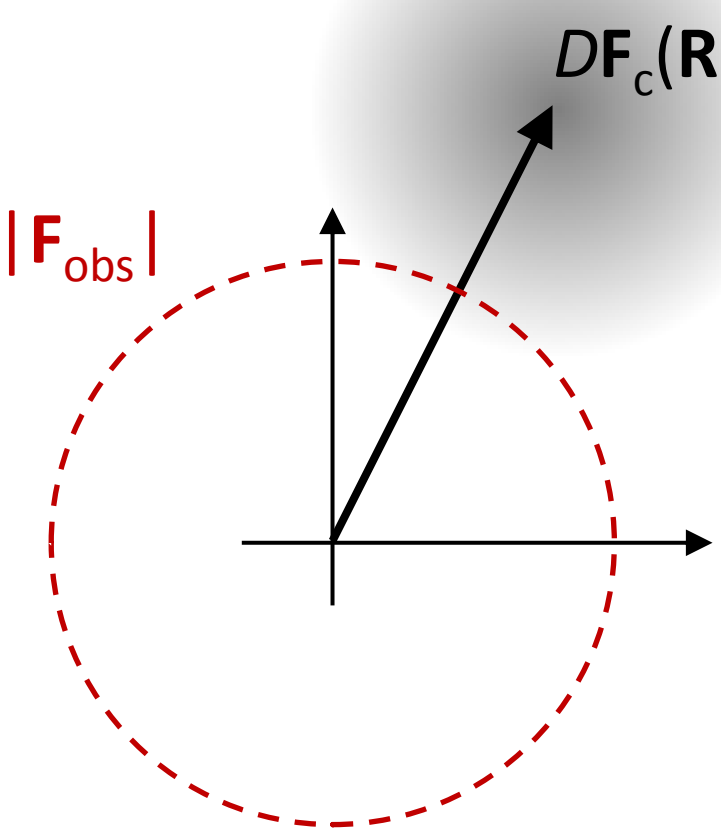


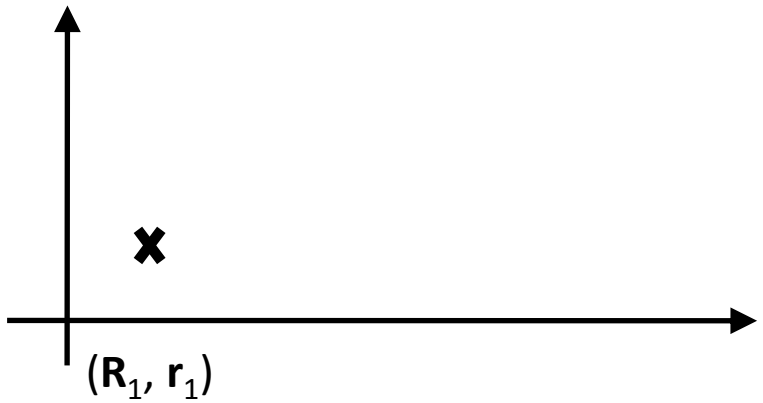
Model incompleteness

Calculate from unit cell composition

Likelihood of $|\mathbf{F}_{obs}|$

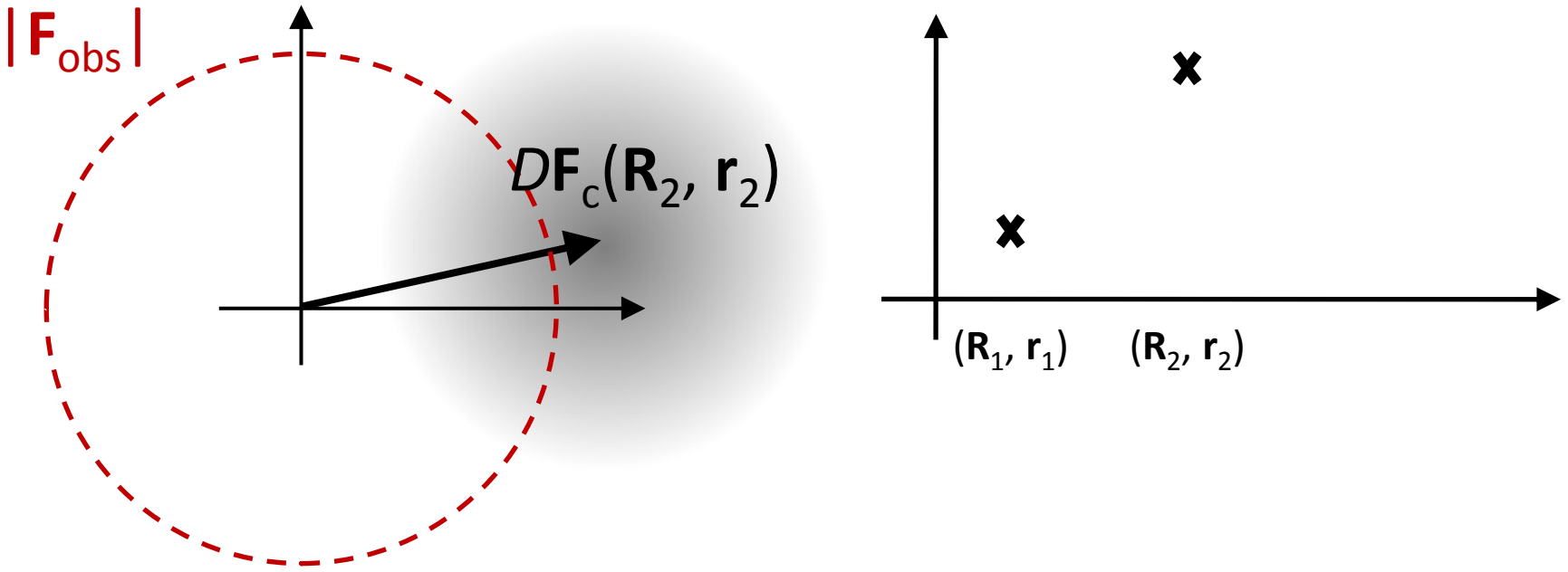
(given $\mathbf{R}_1$, $\mathbf{r}_1$, D, σ_Δ)





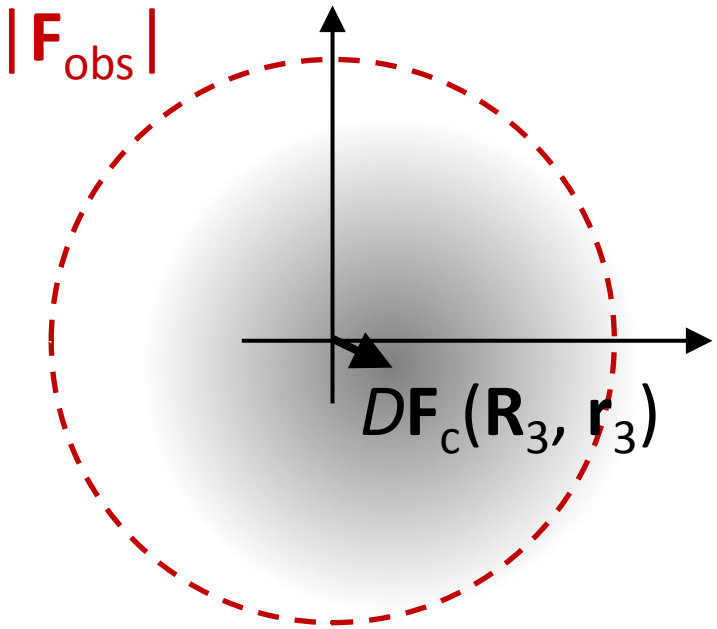
Likelihood of $|\mathbf{F}_{obs}|$

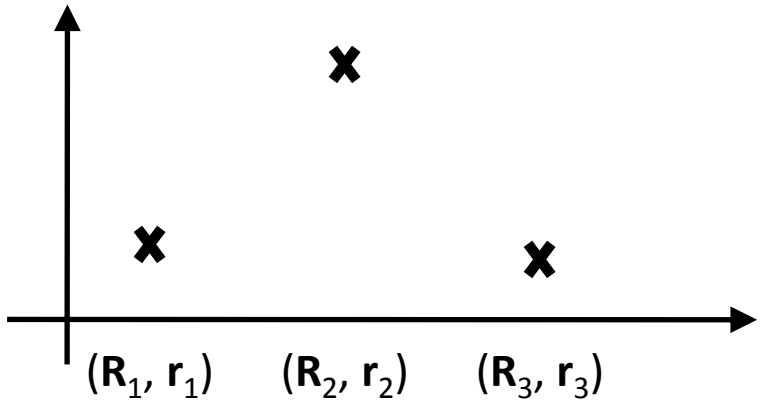
(given $\mathbf{R}_2$, $\mathbf{r}_2$, D, σ_Δ)



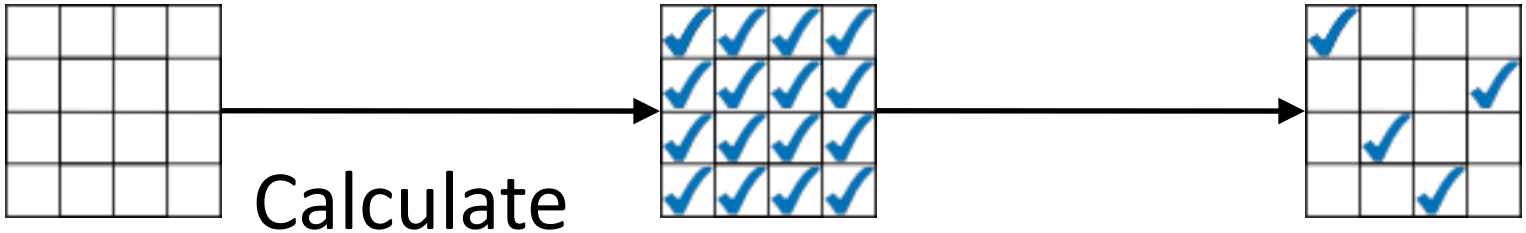
Likelihood of $|\mathbf{F}_{obs}|$

(given $\mathbf{R}_3$, $\mathbf{r}_3$, D, σ_A)



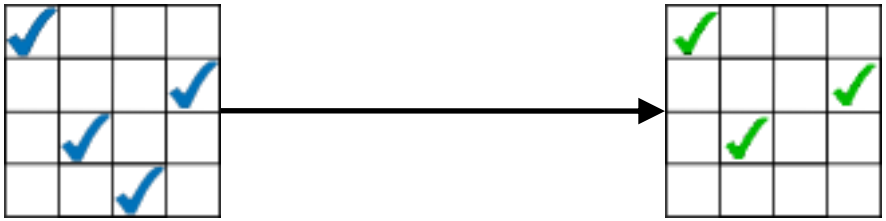


Likelihood usage



Calculate **approximation** by **FFT**

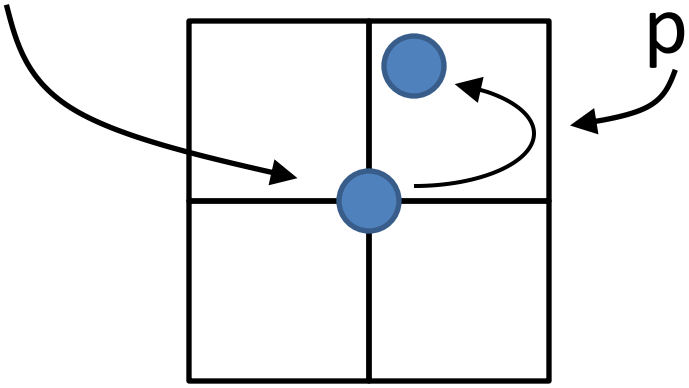
Peak search



Rescore with likelihood

Refinement – Likelihood target

The rotation and translation functions were performed on a (not very fine) grid



The solution can be improved if the grid is taken away and the rotational and translational parameters optimized

Fast search

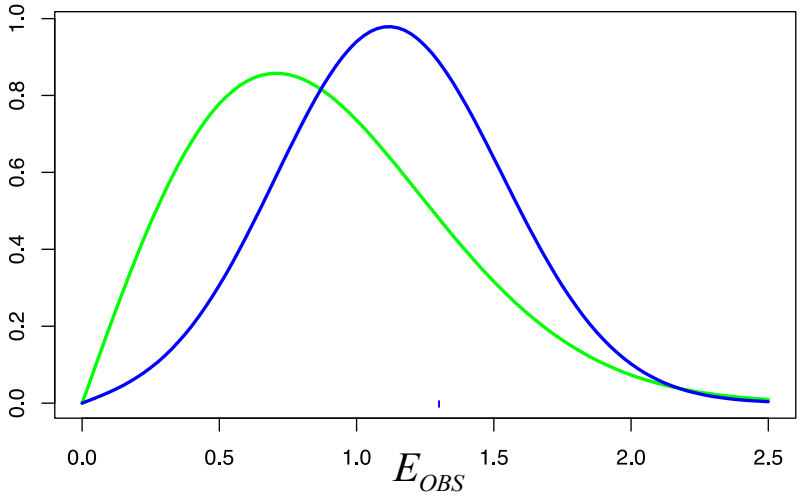
Back to classical MR? Not quite

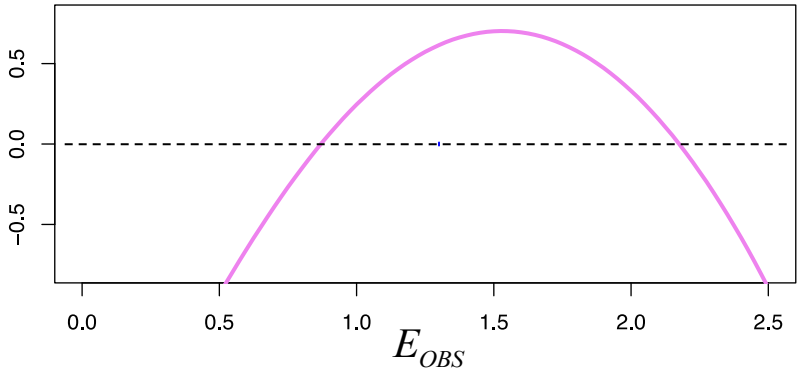
Advantages of using Likelihood approach

- Final scoring is the most robust
- Maximum consistency between preliminary and final scoring

A way to push boundaries of the method

Log Likelihood Gain (LLG)

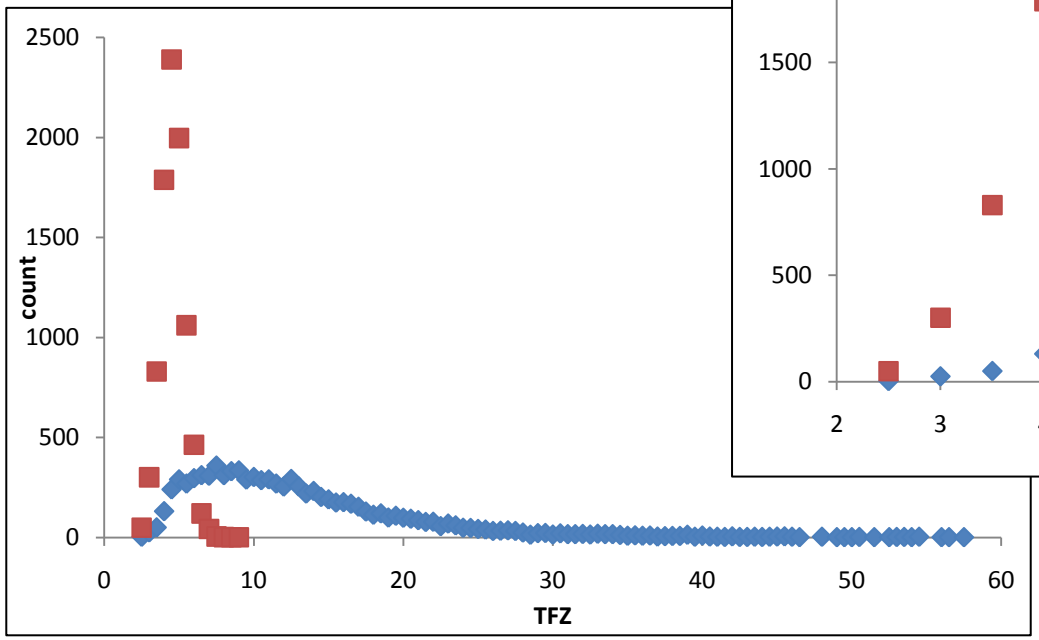


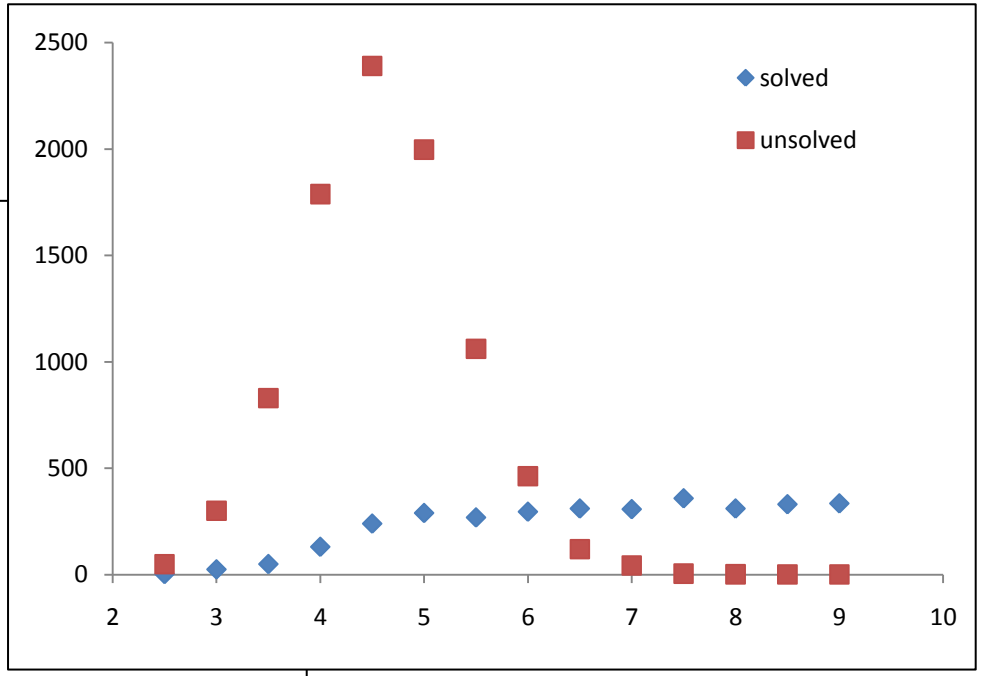


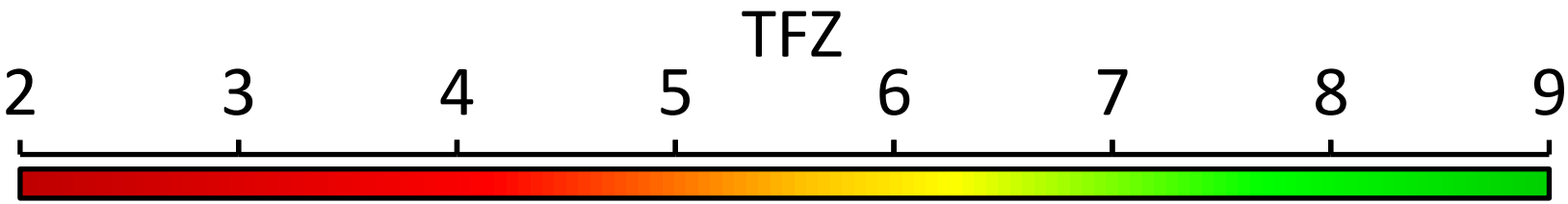
- $p_1 = p(E_{OBS} \mid E_{CALC} = 1.3, s_A = 0.8)$
  The model is assumed to be good
- $p_0 = p(E_{OBS} \mid E_{CALC} = \text{Any}, s_A = 0)$
  The model is assumed to be random
- $\text{LLG}_1 = \log p_1 - \log p_0$
  Log Likelihood Gain (LLG)

- An empirical criterion LLG > 70:
  - identification of the MR successful run
  - prediction of sufficient amount of data (resolution cut-off)

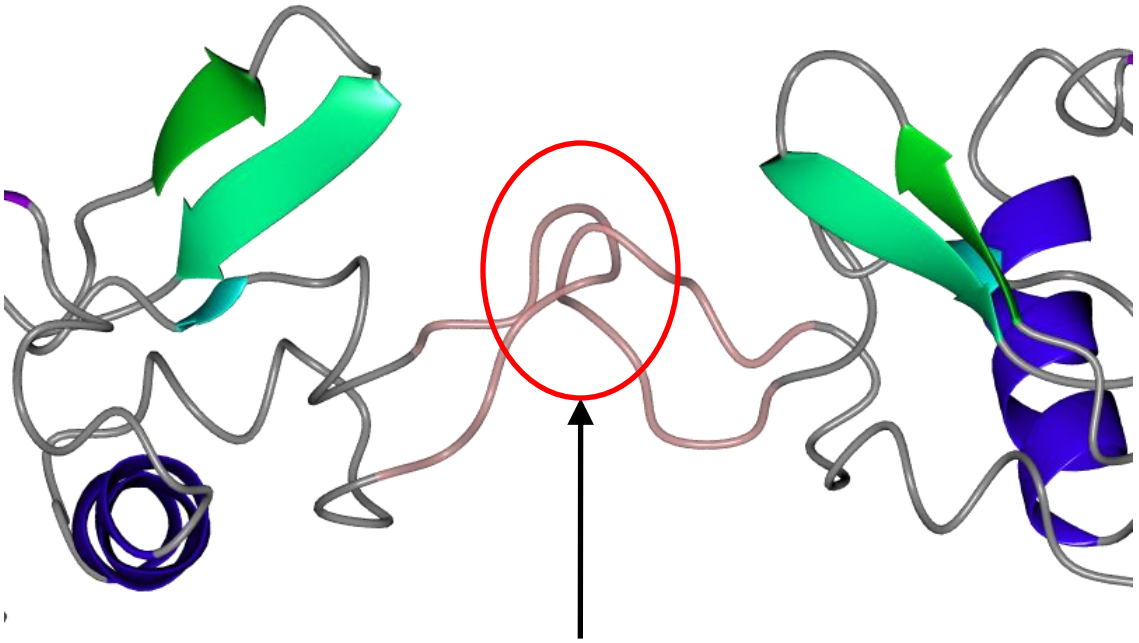
Identifying solutions





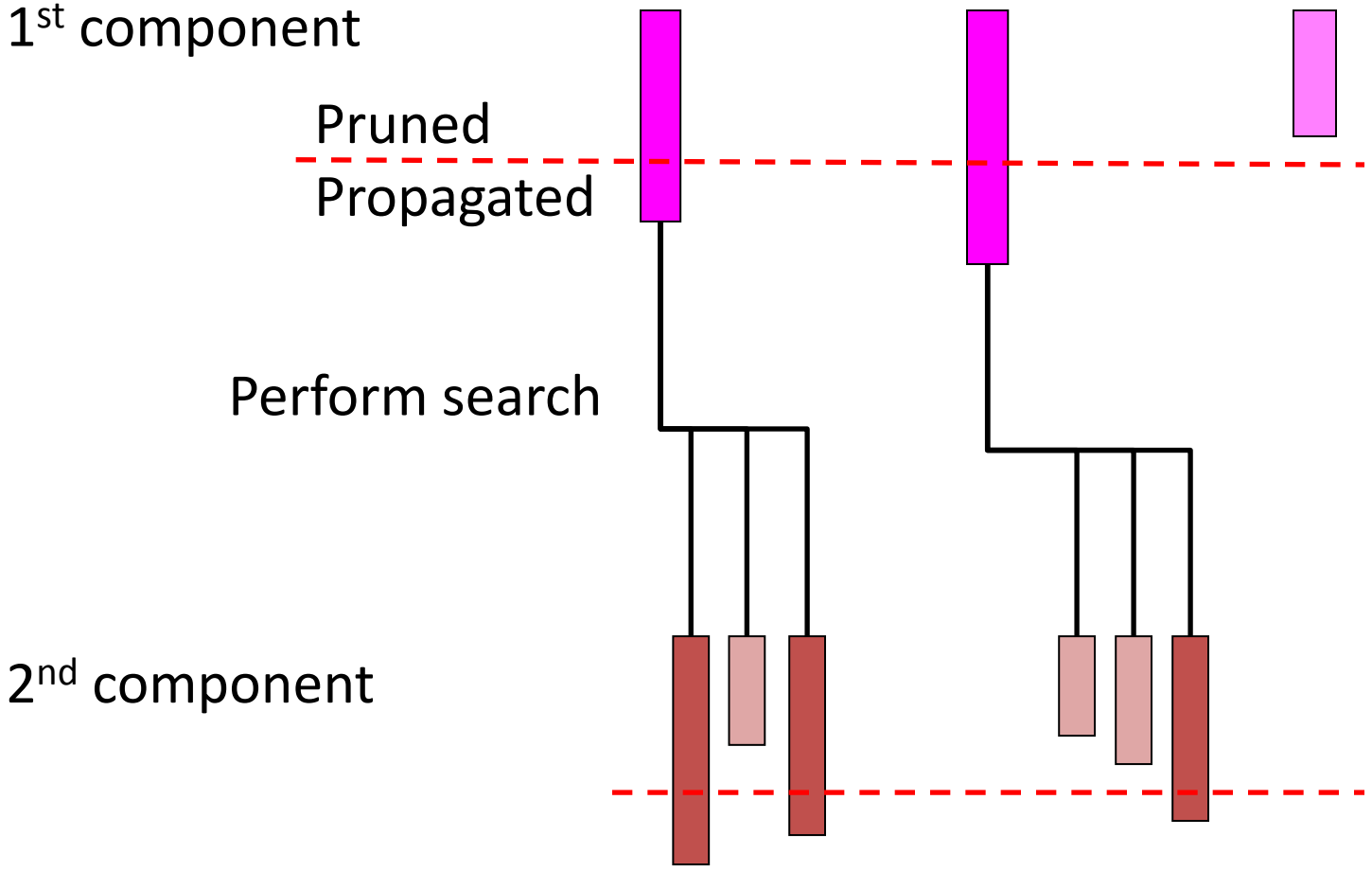


Packing

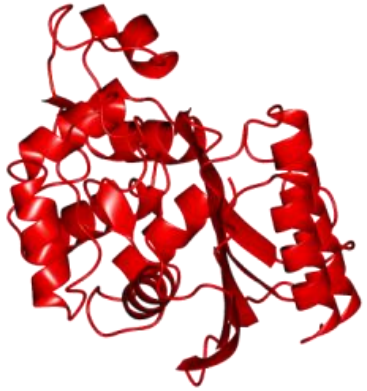


A small number of clashes
is acceptable

Tree search with pruning

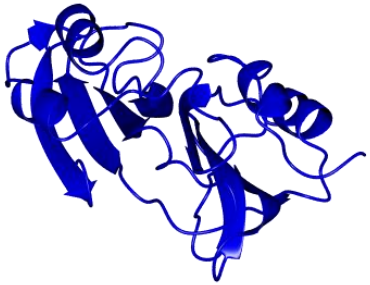


Model search order



65% of structure
35% identical model

Better model if data resolution is low



35% of structure
50% identical model

Better model if data resolution is high

Determined automatically!

If automated MR does not work

MR combination with other techniques

- Automated MR is quite advanced, is convenient, but
  - still may fail to produce complete model
  - may be very slow if there are many components to position
  - limited feedback from user

  In general, automated MR is good for finding partial model, but it is not the most efficient way to generate a nearly complete model.

- Combine MR with
  - refinement of partial model
  - density improvement
  - manual or automated model building

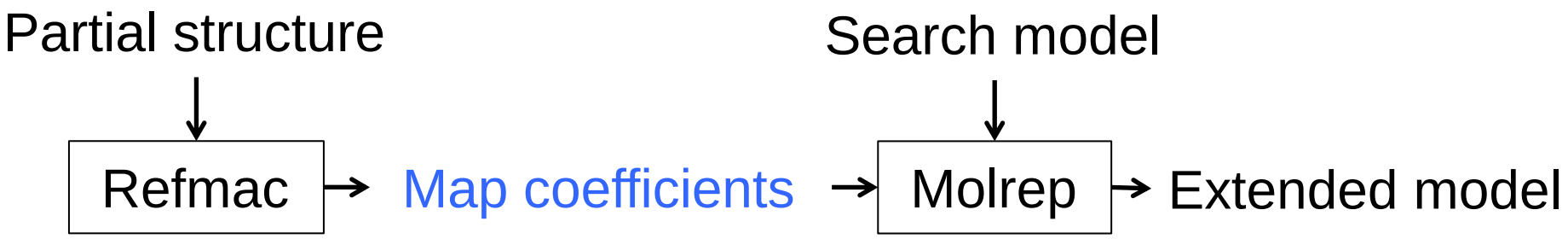
  MR pipelines: combinations with other automated methods

- Specialised MR techniques can be useful

Specialised MR techniques

- Handling Translational Non-Crystallographic symmetry
  - Non-origin peaks in the Patterson map indicate the presence of TNCS
  - Molecules related by TNCS can be found in one go as they have the same orientation
- Locked RF and TF
  - Using point symmetry of oligomers
- Exhaustive and stochastic searches
- Using phase information in MR
- Self Rotation Function

Refinement refinement + Phased MR



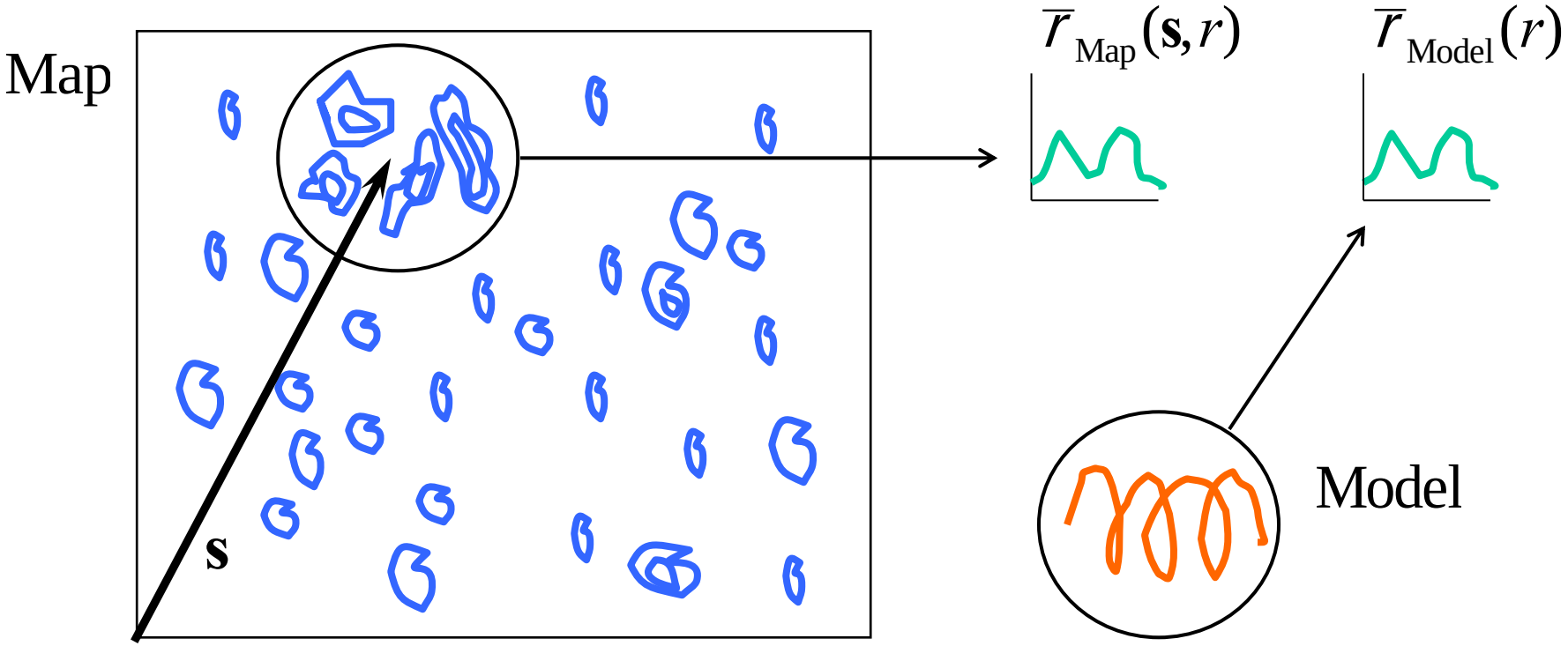
Search in the map

- Calculate 2-1 or 1-1 maps after restrained refinement of partial structure
- Flatten the map corresponding to the known substructure
- Calculate structure amplitudes from this map
- Use these modified amplitudes in Rotation Function
  - Theoretically quite similar to RF with fixed model in *Phaser*
- And finally – Phased TF

SAPTF

Spherically Averaged Phased Translation Function
(FFT based algorithm)

$$\mathrm{SAPTF}(\mathbf{s}) = \int \bar{r}_{\mathrm{Map}}(\mathbf{s}, r)\, \bar{r}_{\mathrm{Model}}(r)\, r^2\, \mathrm{d}r$$



Search in the map with SAPTF

1. Find approximate position:
   Spherically Averaged Phased Translation Function

2. Find orientation:
   Local Phased Rotation Function
   – Local search of the orientation in the density

3. Verify and adjust position:
   Phased Translation Function

Search in the map with SAPTF

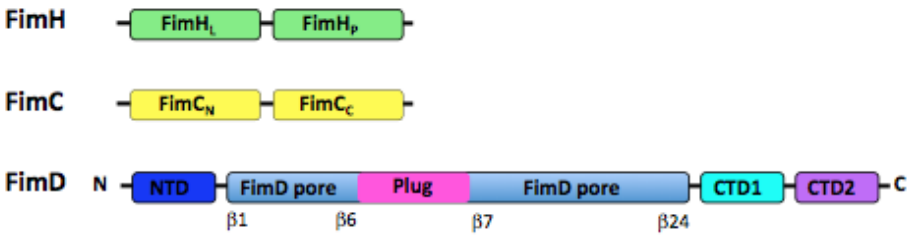
1. Find approximate position:
   Spherically Averaged Phased Translation Function

2. Find orientation:
   Local Rotation Function
   – Structure amplitudes from the density within the SAPTF sphere

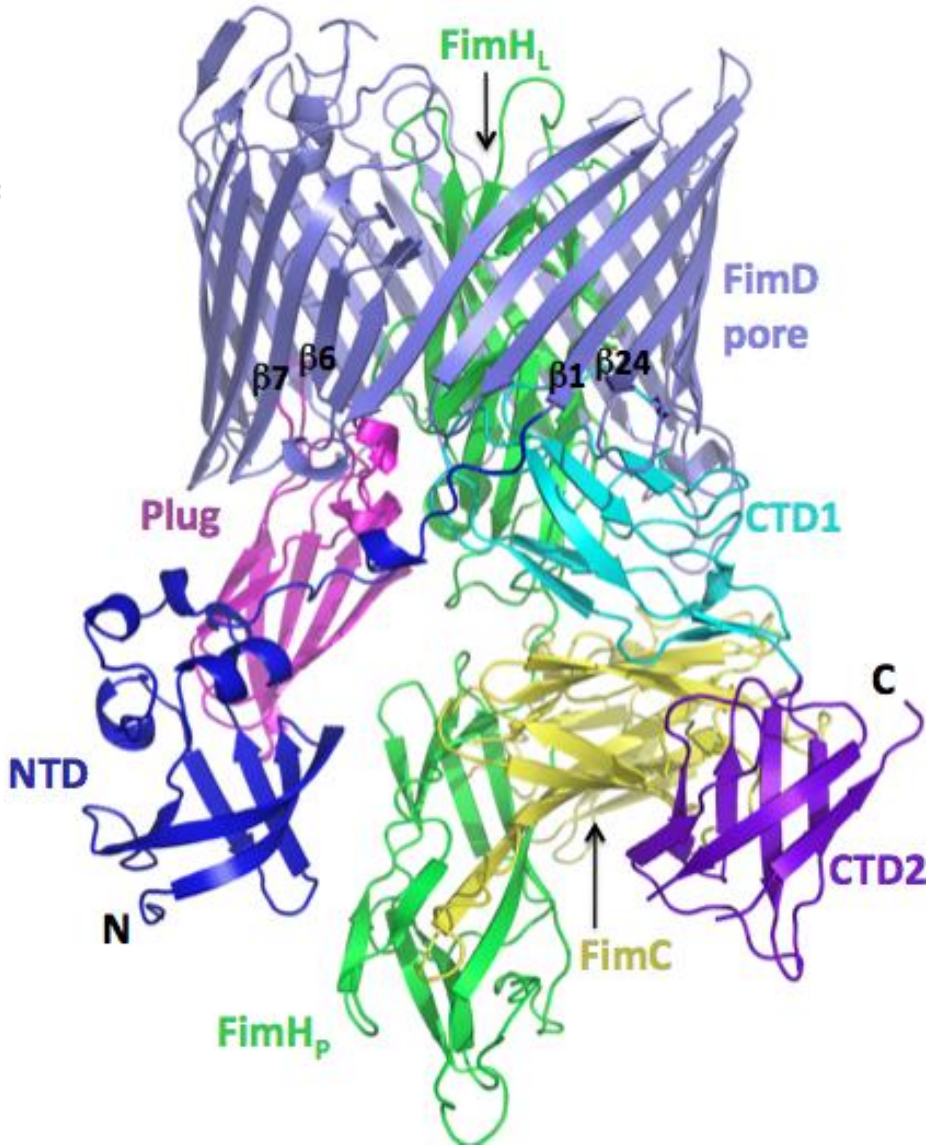
3. Verify and adjust position:
   Phased Translation Function

- Local RF is less sensitive than Phased RF to inaccuracy of the model position

Example



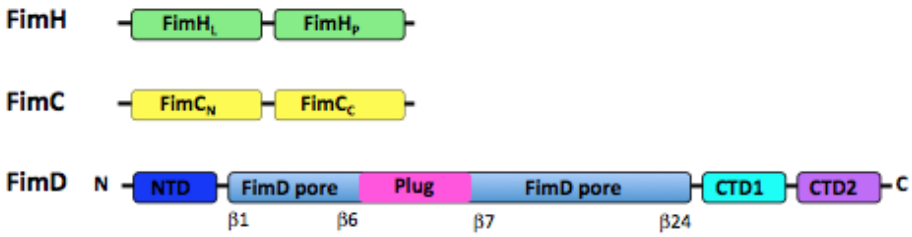
- Asymmetric unit two copies
- Resolution 2.8 Å



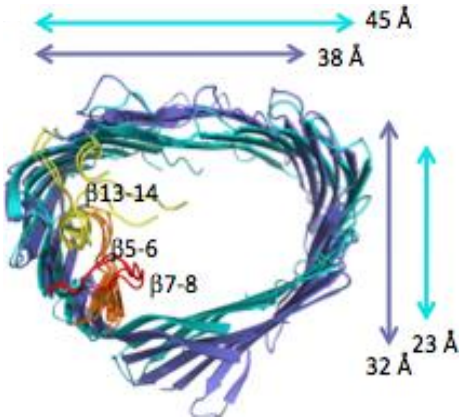
Phane et. al (2011) Nature, 474, 50-53

Usher complex structure solution

1. Conventional MR
   - FimC-N + FimC-C
   - FimH-L + FimH-P
   - FimD-Pore



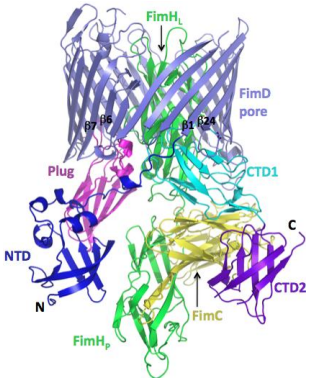
2. Jelly body refinement (Refmac)
   - FimD-Pore

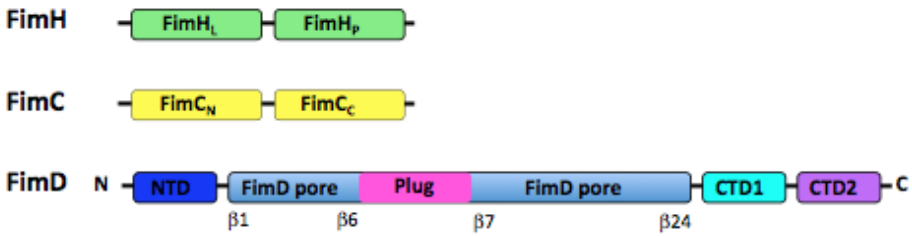


3. <u>Fitting into the electron density</u>
   - FimD-Plug
   - FimD-NTD
   - FimD-CTD-2

4. Manual building
   - FimD-CTD-1

Performance of fitting methods





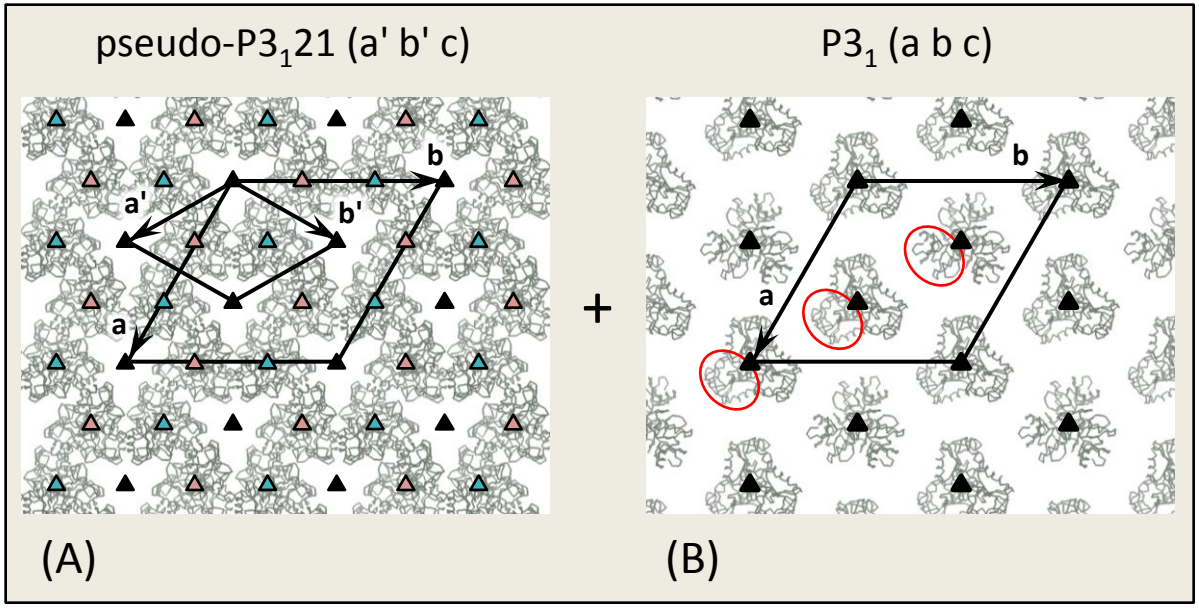
	search model	sequence identity	"Masked" RF PTF *prf n*	SAPTF PRF PTF *prf y*	SAPTF Local RF PTF *prf s*
FimD-Plug	3fip_A	38.5%	2 (2)	– (–)	1 (2)
FimD-NTD	1ze3_D	100%	2 (2)	1 (2)	2 (2)
FimD-CTD-2	3l48_A	33.3%	– (–)	2 (2)	– (–)

Trying several methods is a good practice (also because of cross-validation)

Twinned crystal with pseudo-symmetric substructure

Human macrophage receptor CLEC5A for dengue virus

Watson, A. A. et al. (2011). J Biol Chem 286, 24208-18.

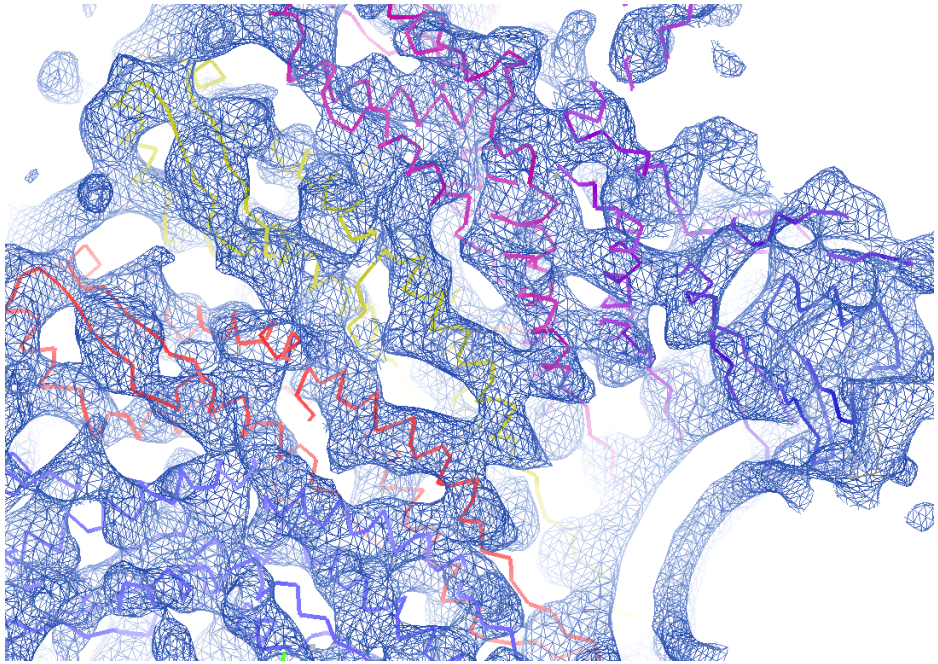


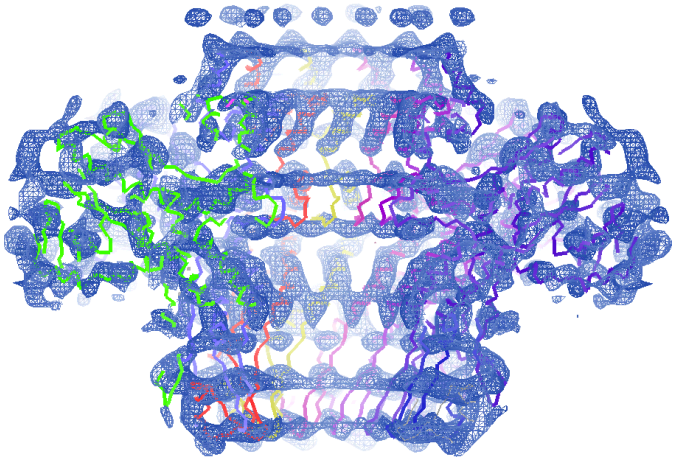
3-fold axes with respect to the true structure:

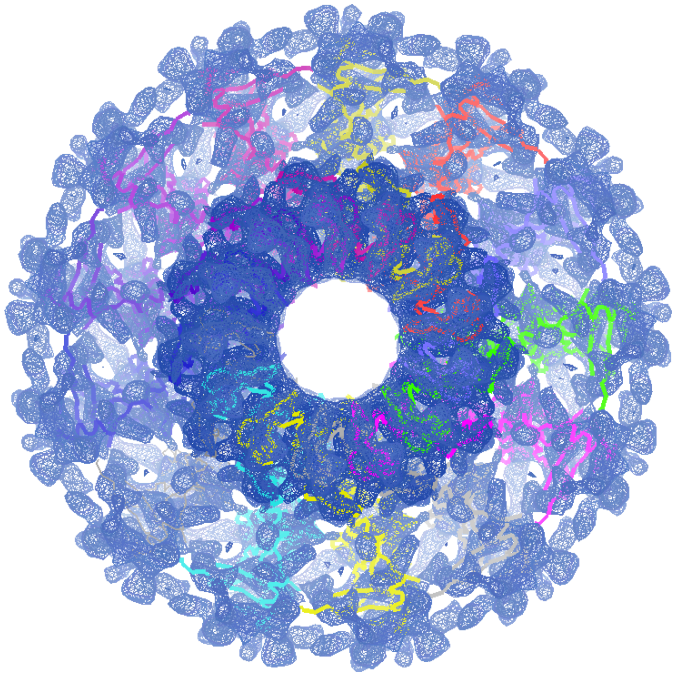
- ▲ crystallographic
- ▲▲ pseudosymmetry for (A)

- Substructure (A): Patterson search (4 copies), search in density (2 copies)
- Substructure (B): Search in the density (3 copies)

Fitting into EM maps

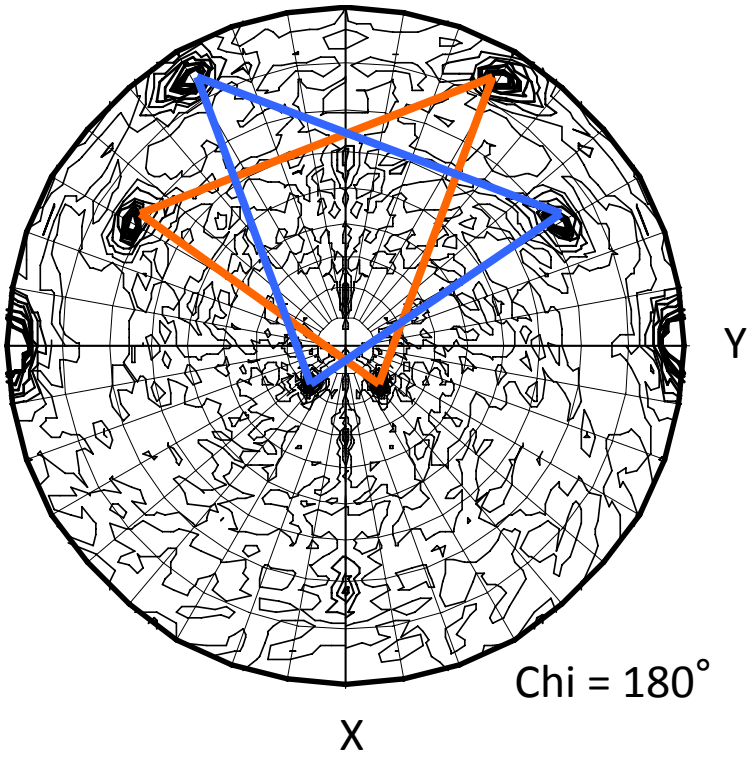






SPP1 portal protein

Self Rotation Function (SRF)



Example of SRF

- Space group P21
- One 222-tetramer in the AU

Preliminary analysis of X-ray data

- Oligomeric state of the protein in crystal
- Selection of oligomeric search model

Limited use

- No clear interpretation or even artifact peaks in high symmetry point groups (e.g. 622)
- different oligomers with the same symmetry

Locked Rotation Function

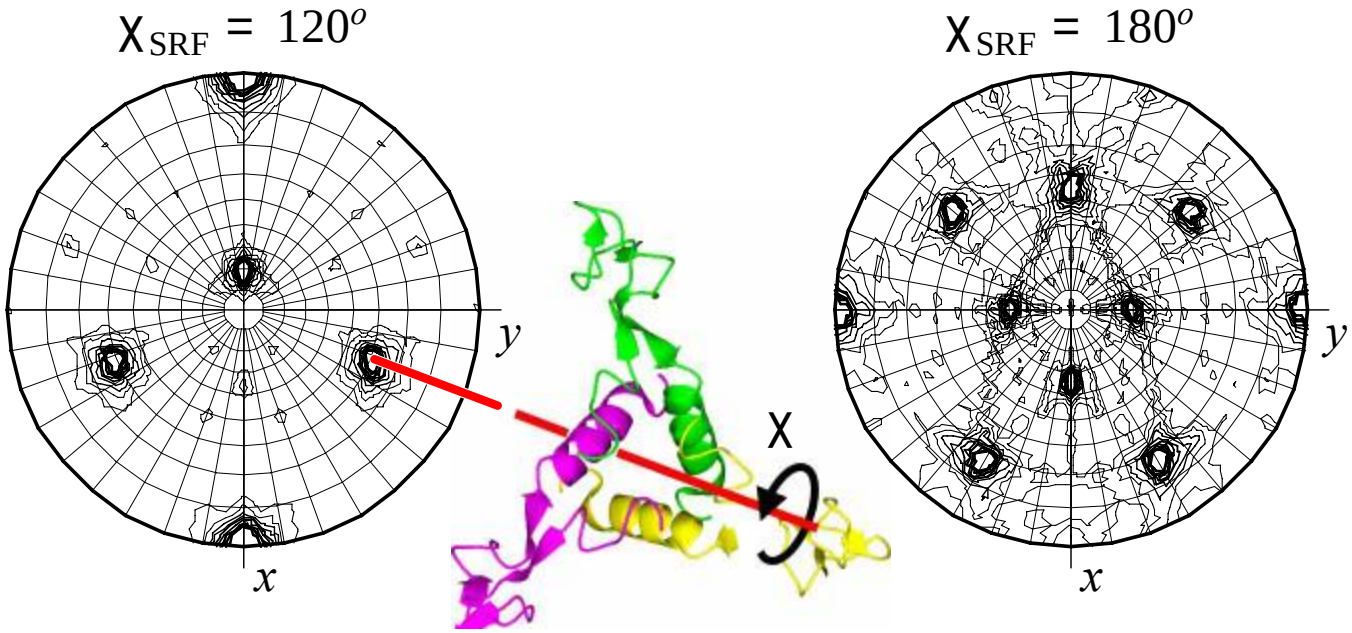
- Uses SRF to derive NCS operations
- Averages RF over NCS operations
- In favorable cases Improves signal to noise ratio in RF

Automatic mode:

```
molrep -f s100.mtz -m monomer.pdb -s s100.seq –i <<+
lock y
+
```

There is an option of selecting specific SRF peaks

One-dimensional exhaustive search (exotic case)

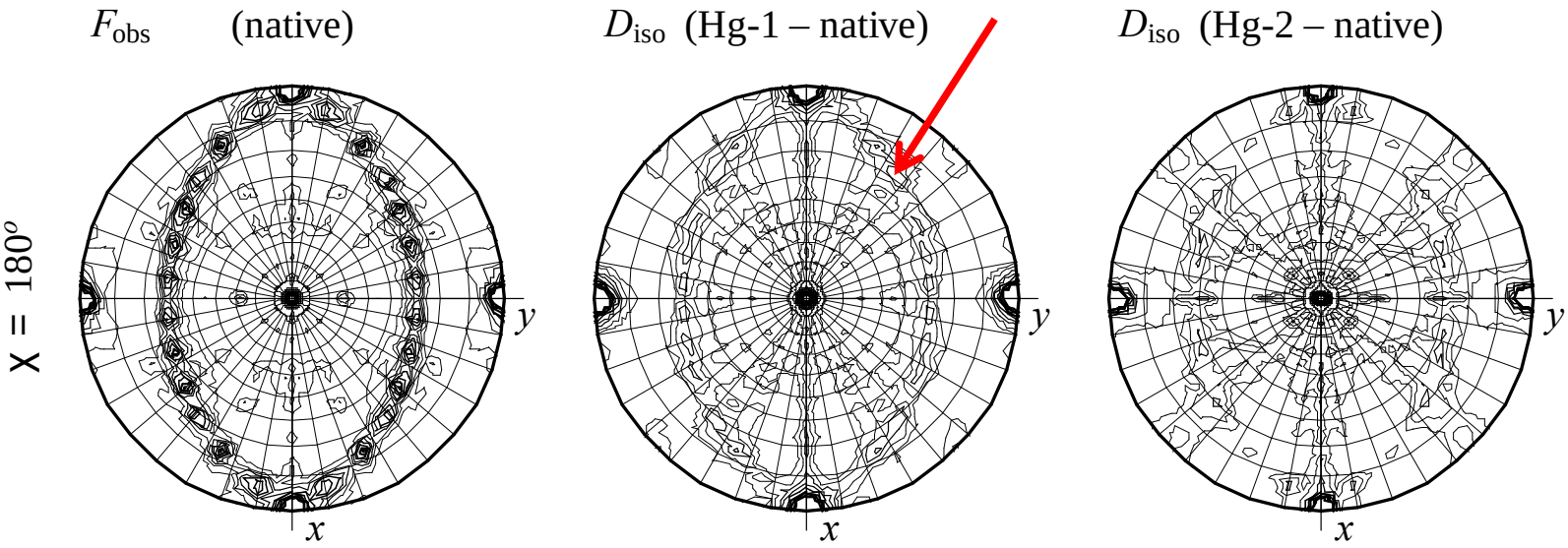


SRF helps restrict dimensionality in an exhaustive search

- Orientation of the ring is known from the analysis of SRF
- Unknown parameter: rotation about 3-fold axis
- One-parametric exhaustive search using TF as score function

MR substructure solution (exotic case)

- Select isomorphous derivative
  - by comparing native SRF and SRF from D-iso



- Hg-substructure is a 13-atom ring (from native SRF analysis)
  - Orientation of the ring is known from the analysis of SRF
  - Unknown parameters: radius of the ring, rotation about 13-fold axis
- Two-parametric exhaustive search

Which direction does MR go?

Automation. Therefore:

✖ Collection of tricks

✔ Improvement of "standard" methods
✔ Better scoring system

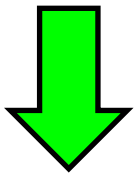
✔✔ Models

Model improvement

✓ Structural information in MR models

✓ Extra information in sequence alignment

✓ Many homologous structures



- Remove residues that do not align
- Remove "excessive" atoms from aligned residues
- Ensemble models (several superposed models)

CCP4 programs for model preparation

Single model correction:

- Chainsaw
- Molrep
- Sculptor

Preparation of ensemble models – fitting models:

- Lsqkab
- ProSMART
- SSM (also in Coot)
- Gesamt

Automated preparation of ensemble models:

- Ensembler

Model modification in MOLREP

```
molrep -f data.mtz -m model.pdb -s target.seq
```

- Performs model correction:
  - Identifies secondary structure in the model
  - Aligns target and model sequences
    - » no deletions or insertions in α-helixes or β-strands
  - Retains aligned residues
  - Retains "aligned" atoms in aligned residues
- Adds B-factor to residues exposed to solvent
- Uses sequence identity to down-weight high resolution data

Molecular Replacement in CCP4

MR Programs

- AMoRe
- Molrep
- Phaser

MR Pipelines

- MrBUMP
- Balbes
- AMPLE

Acknowledgements

Molrep

Alexey Vagin	University of York, UK
Garib Murshudov	MRC LMB, UK

Phaser

Randy Read	University of Cambridge
Airlie McCoy	University of Cambridge
Gabor Bunkozci	University of Cambridge

FimD-FimC-FimH

Gilles Phan	ISMB, London, UK
Gabriel Waksman	ISMB, London, UK