Phaser

Molecular replacement

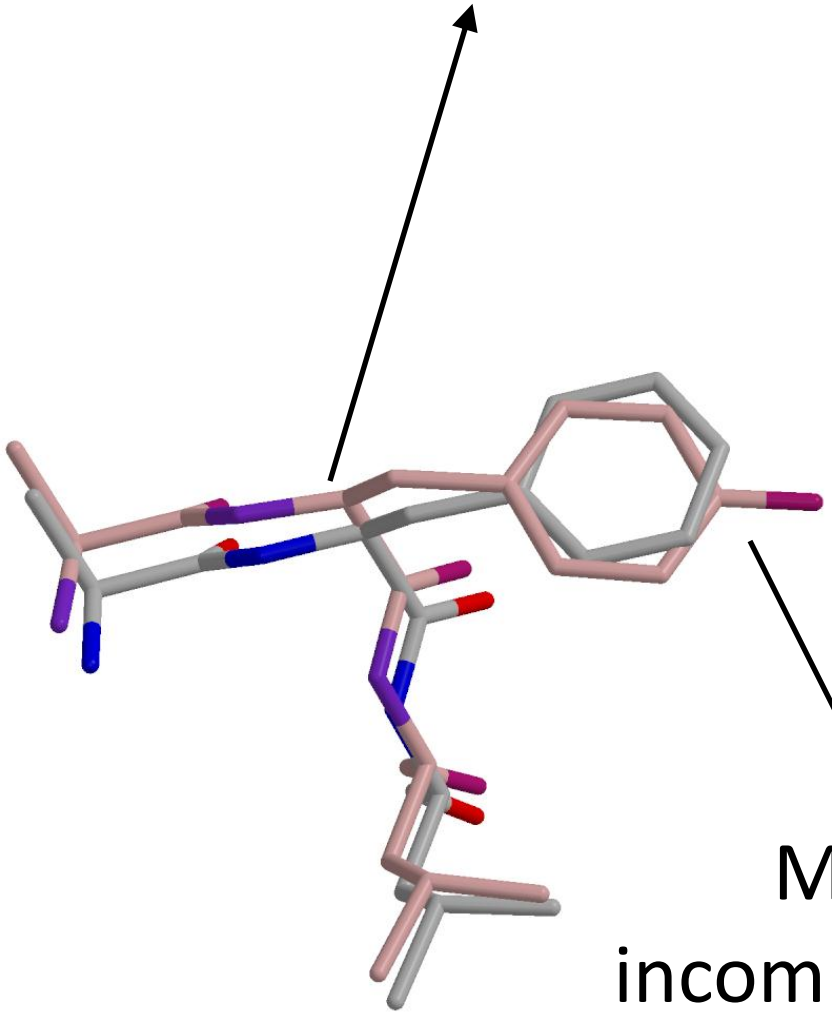
Gábor Bunkóczi

ガボア　ブンコーツイ

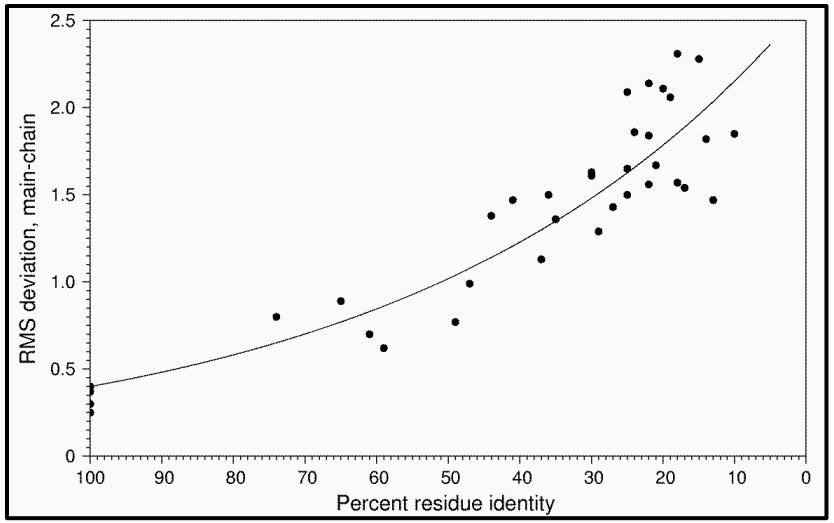
CCP4 Fukuoka School,
30 October 2012

Accounting for model error

Position errors



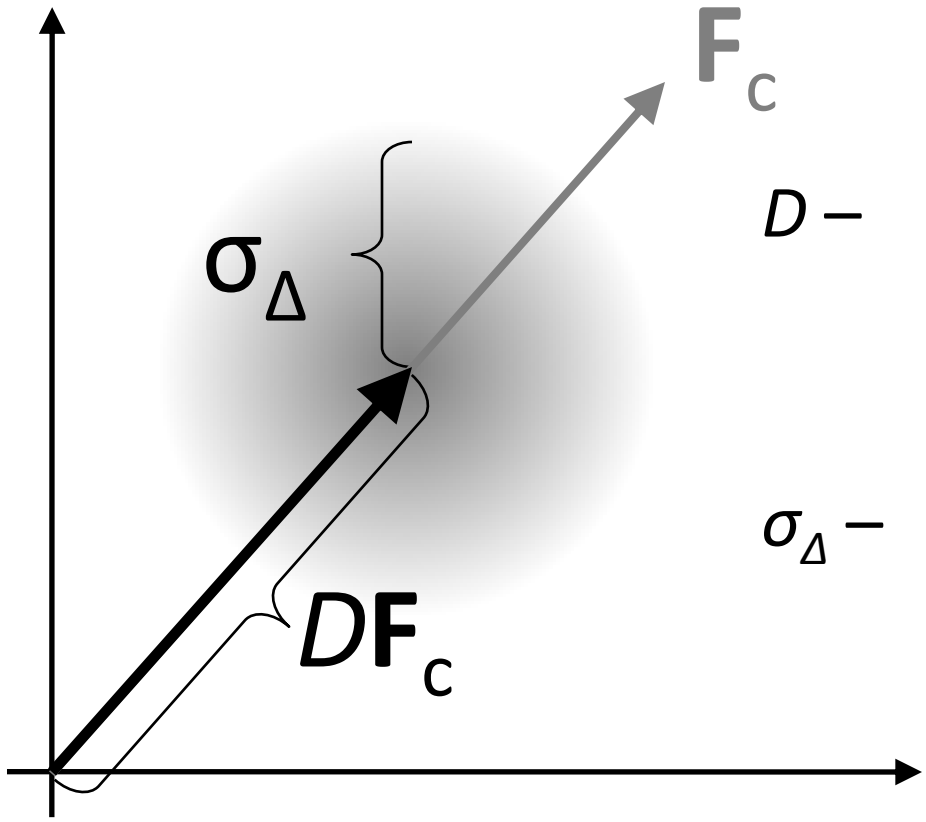
Calculate from sequence identity (Chothia & Lesk)



Model incompleteness

Calculate from unit cell composition

Likelihood function

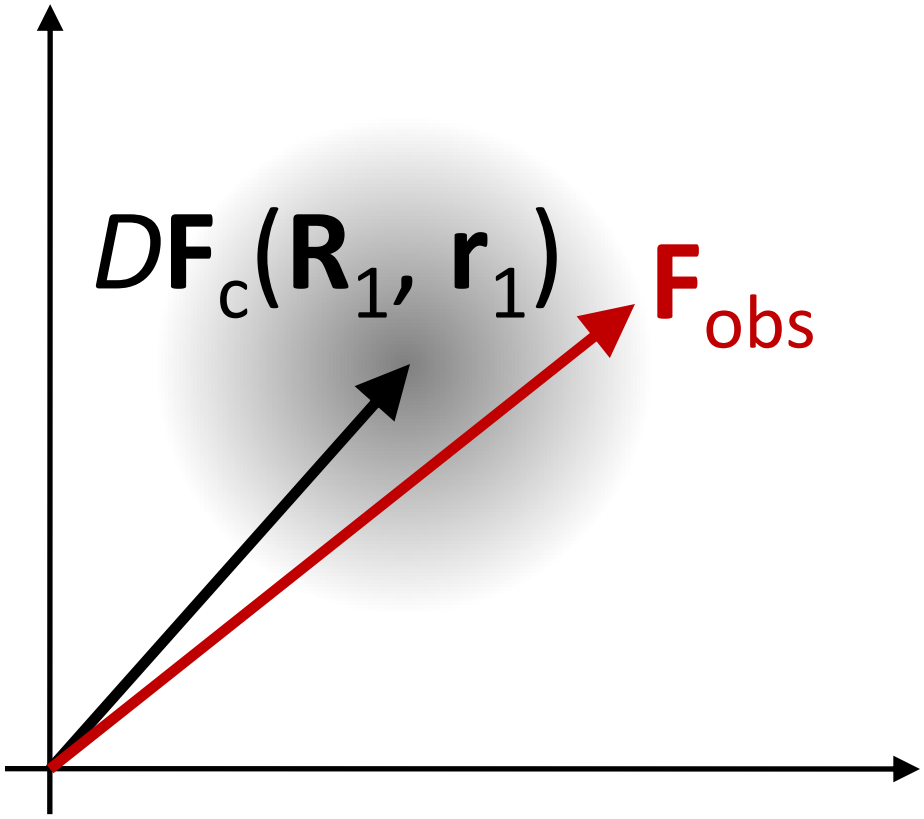


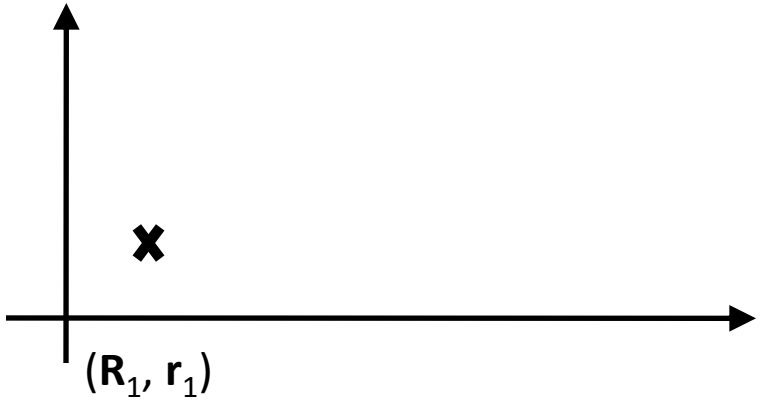
$\mathbf{F}_c$ – model structure factor
Dependent on orientation and position of model

D – Luzzati-factor ($0 \leq D \leq 1$)
Dependent on model errors

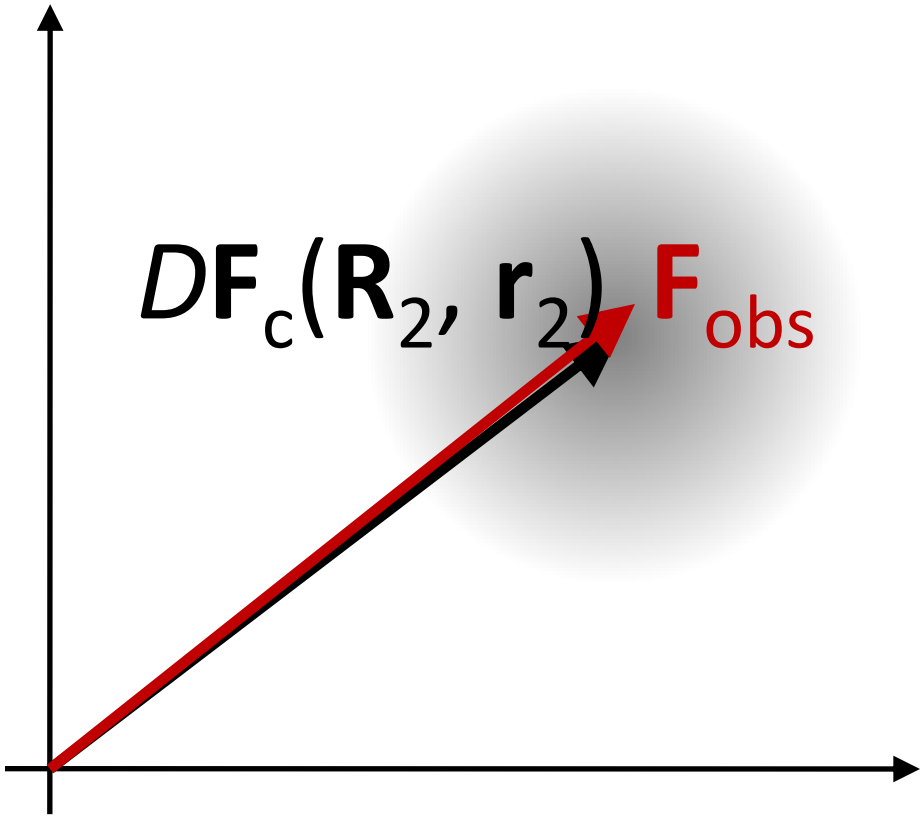
σ_Δ – uncertainty
Dependent on model errors and incompleteness

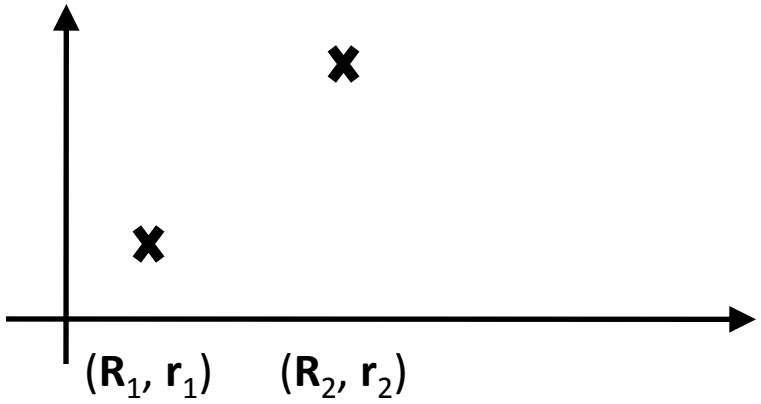
Reflection likelihood



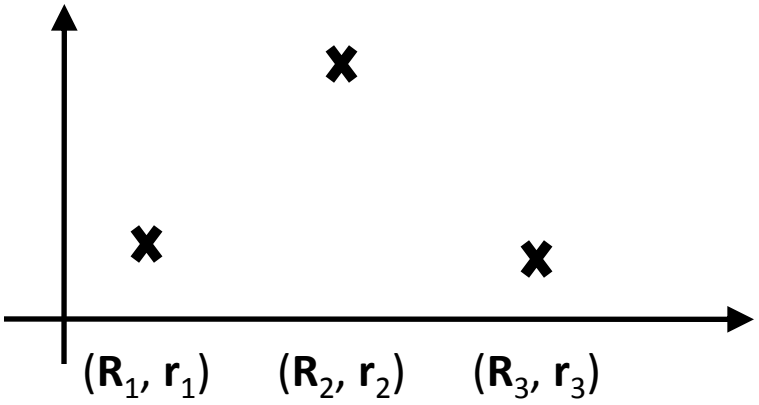


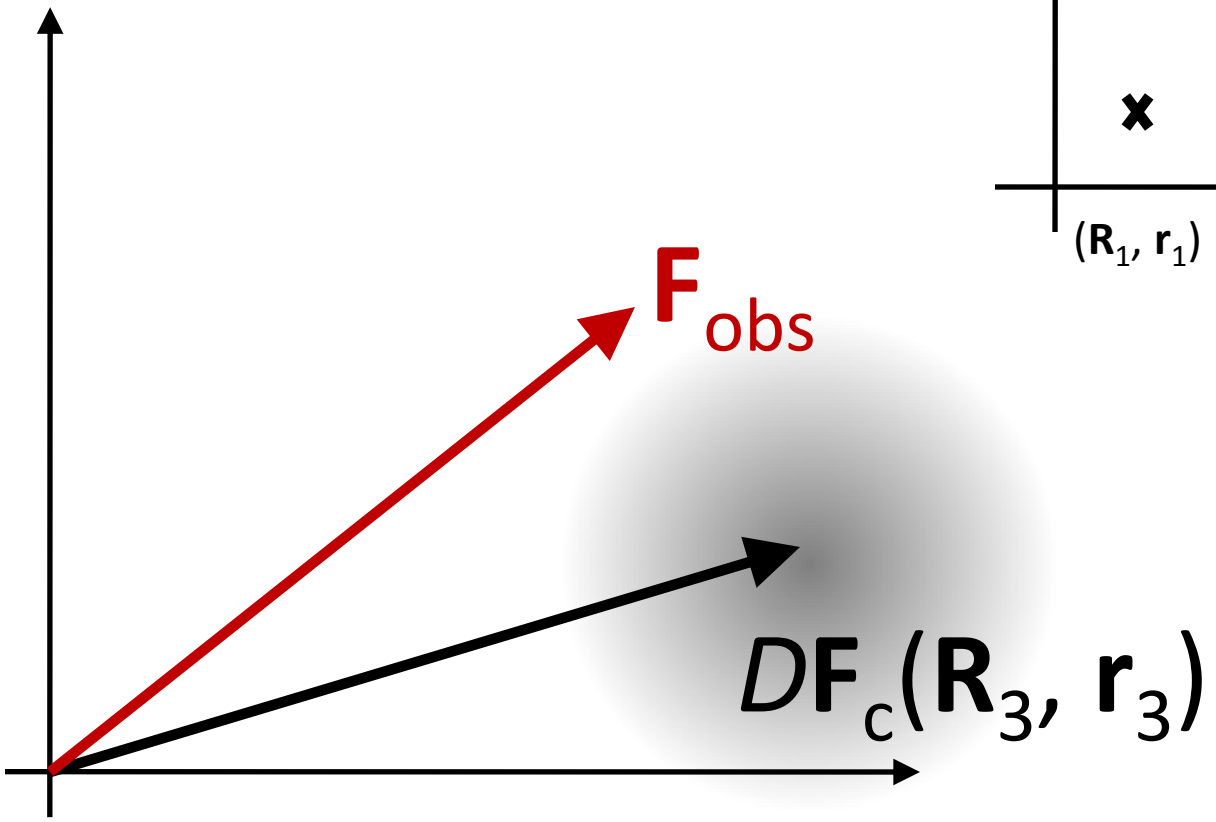
Reflection likelihood



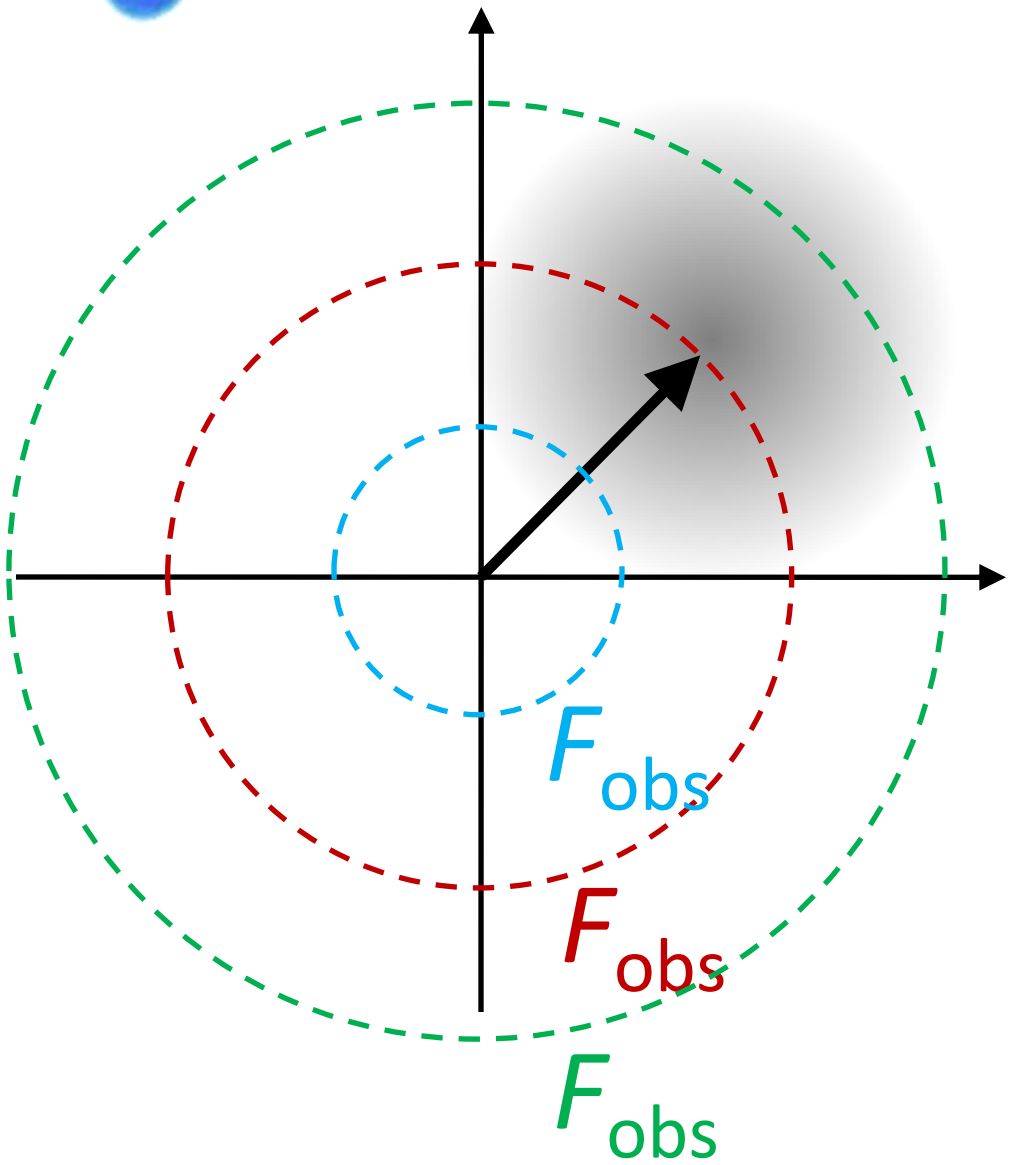


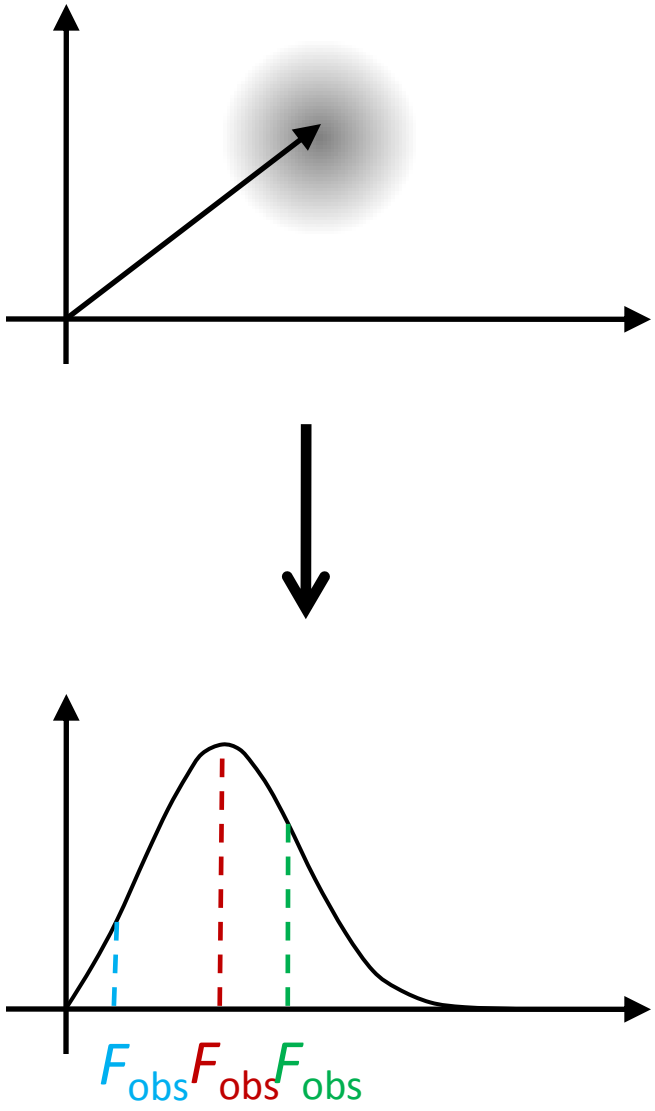
Reflection likelihood



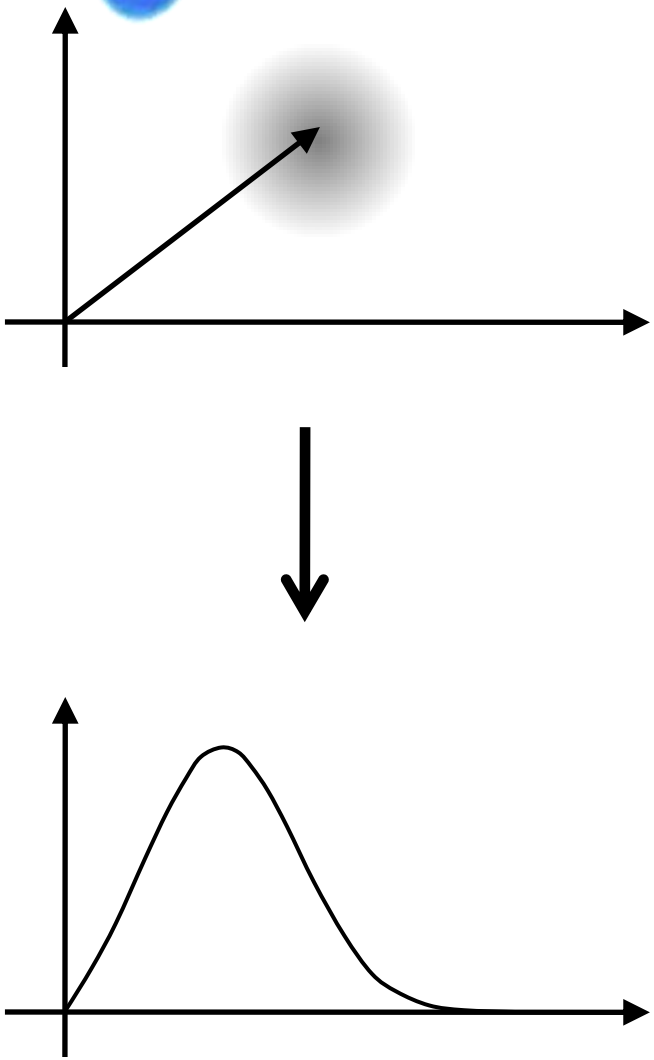


Reflection amplitude likelihood





Rotation and translation functions

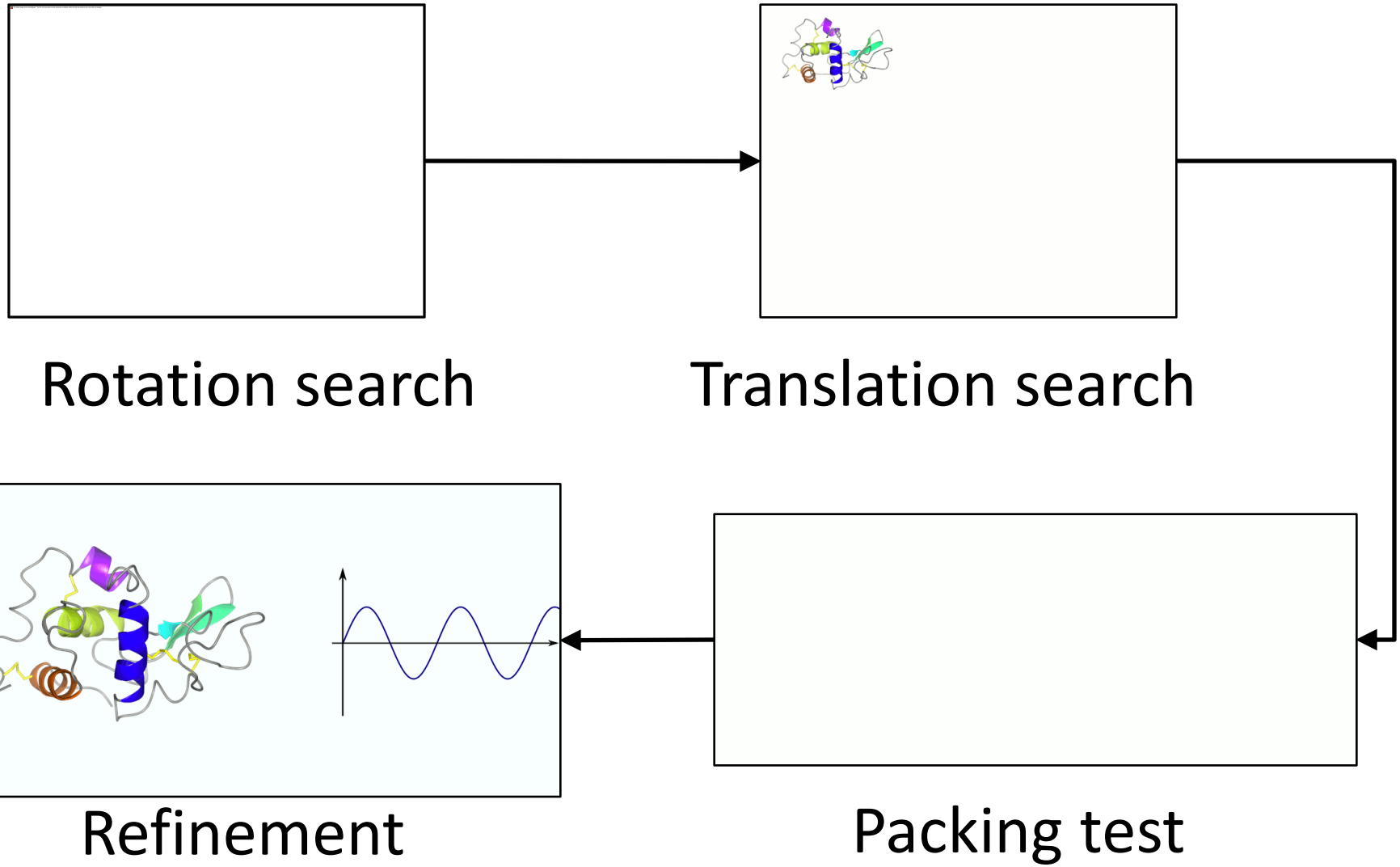


Identical mathematical forms!

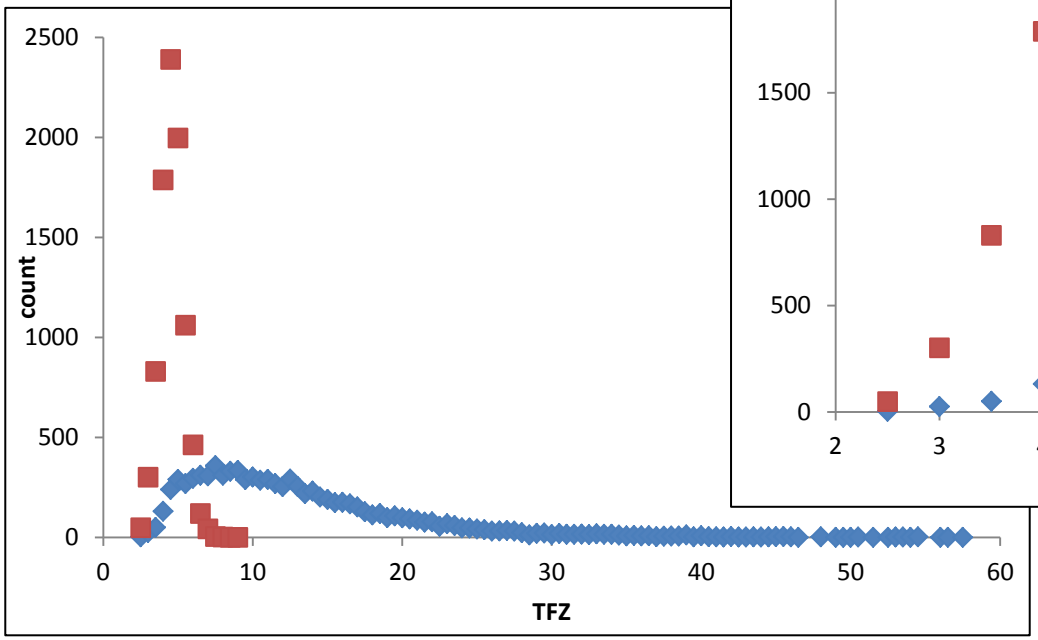
Translation function:
Reflection amplitude likelihood (for all reflections)

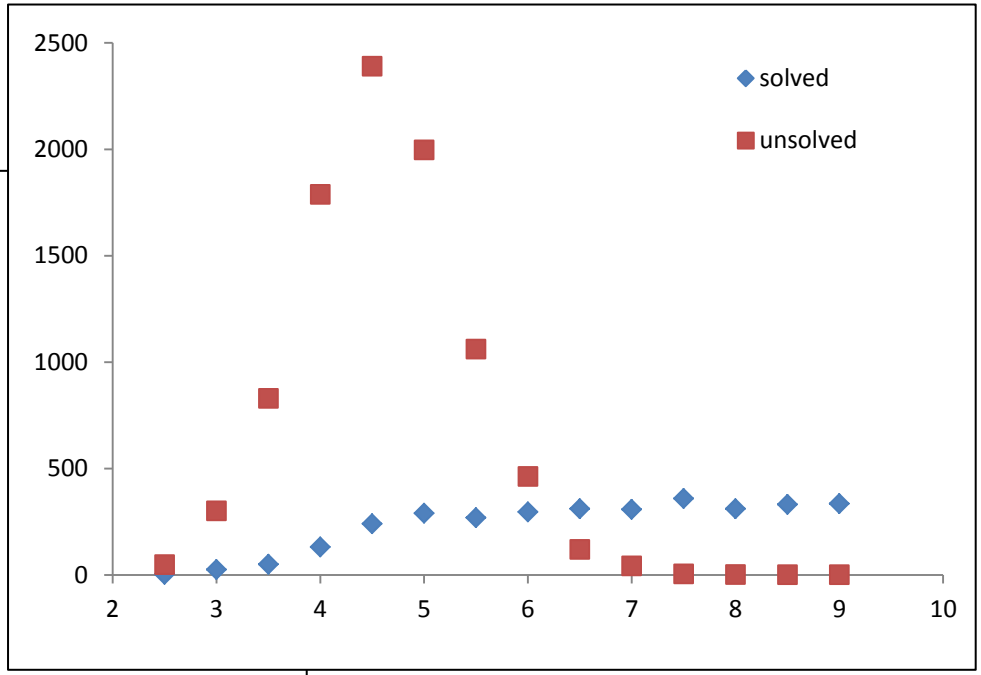
Rotation function:
Reflection amplitude likelihood with σ_Δ is increased to account for unknown translation

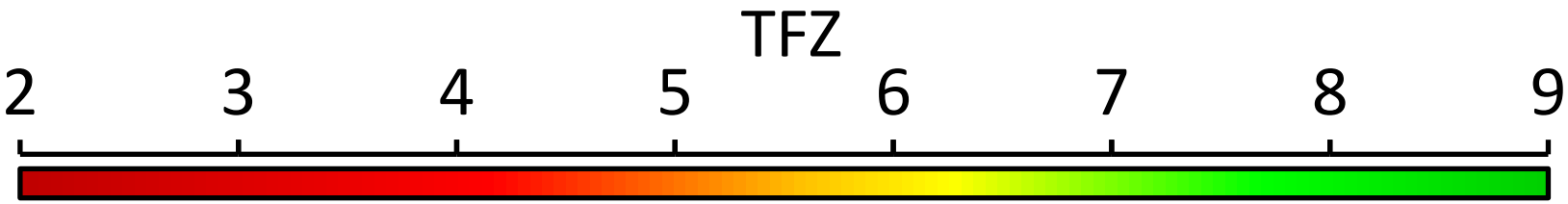
Molecular replacement workflow



Identifying solutions



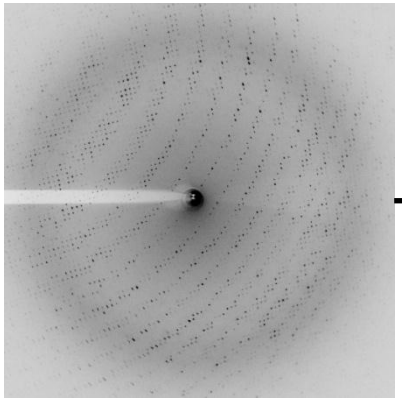




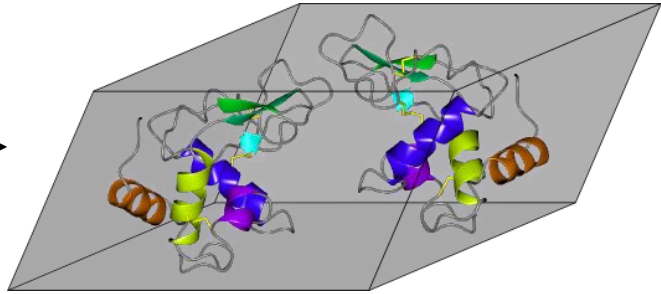
Rob Oeffner

Multiple component searches

Sequential approach



Molecular Replacement

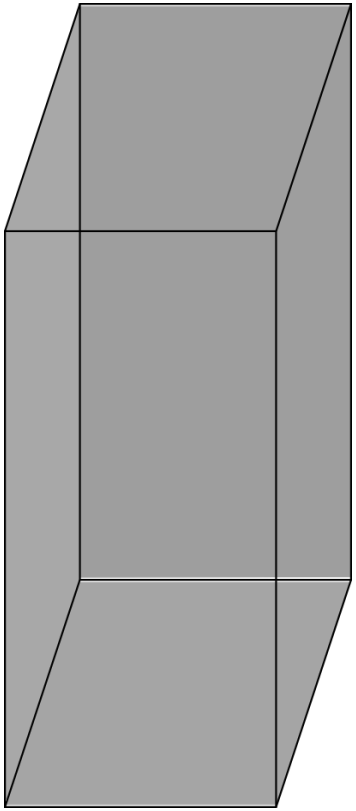


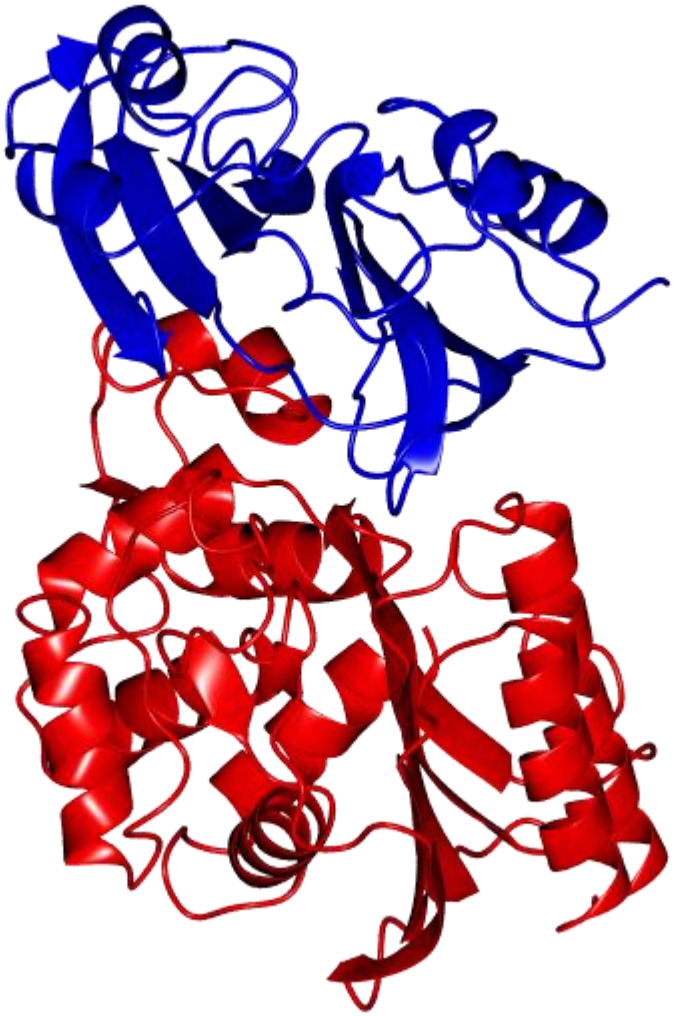
Component1

Component2

Signal increase for subsequent components!

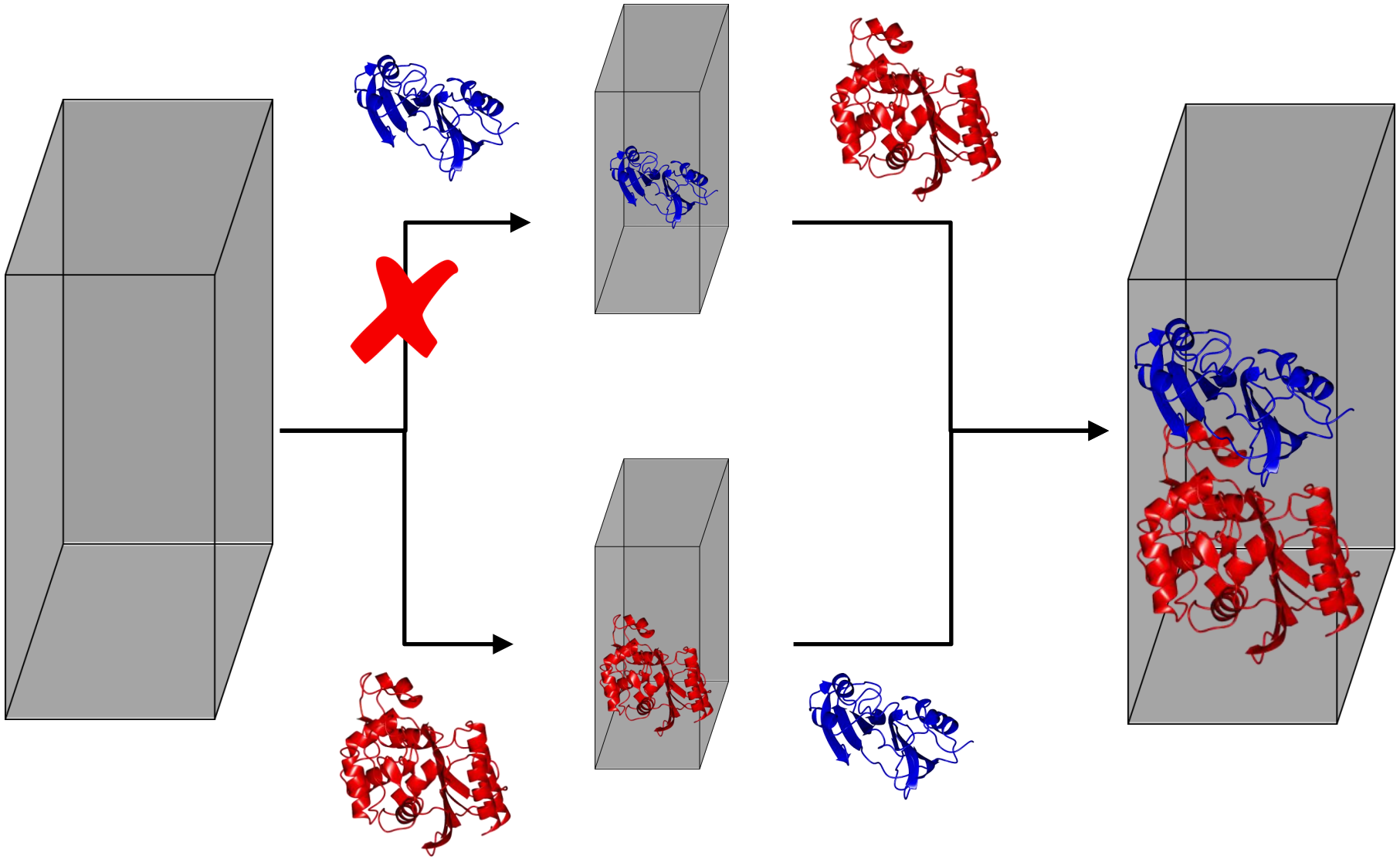
Ordering the search



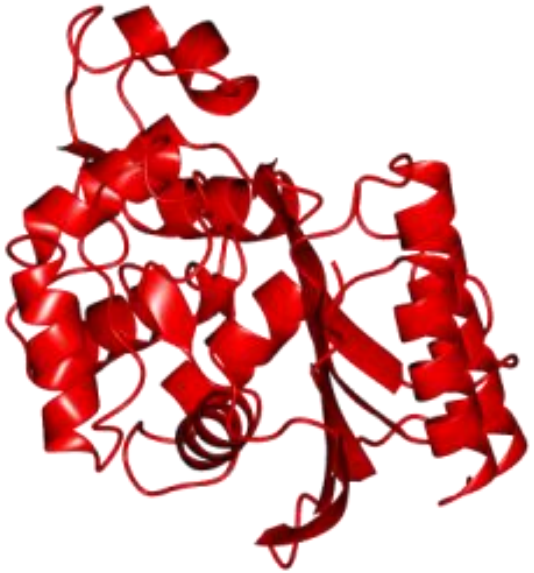


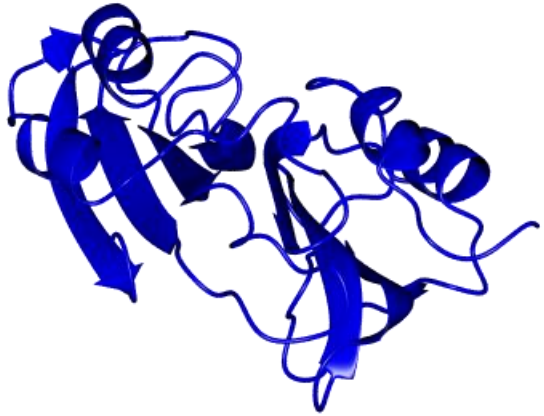
?

Ordering the search



Search order determination





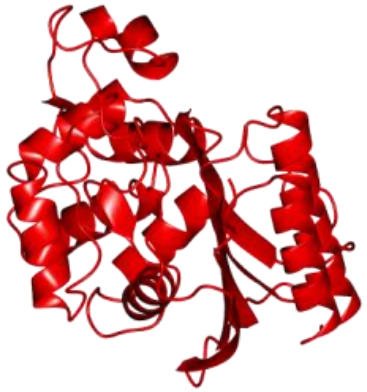
65% of structure
35% identical model

35% of structure
50% identical model

Which model is easier to find?

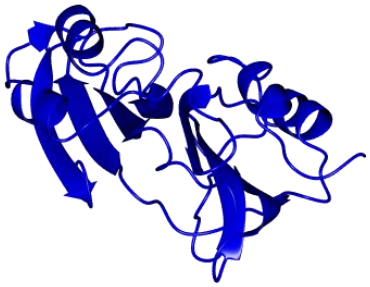
Search order determination

A better model explains more of the data



65% of structure
35% identical model

Better model if data resolution is low



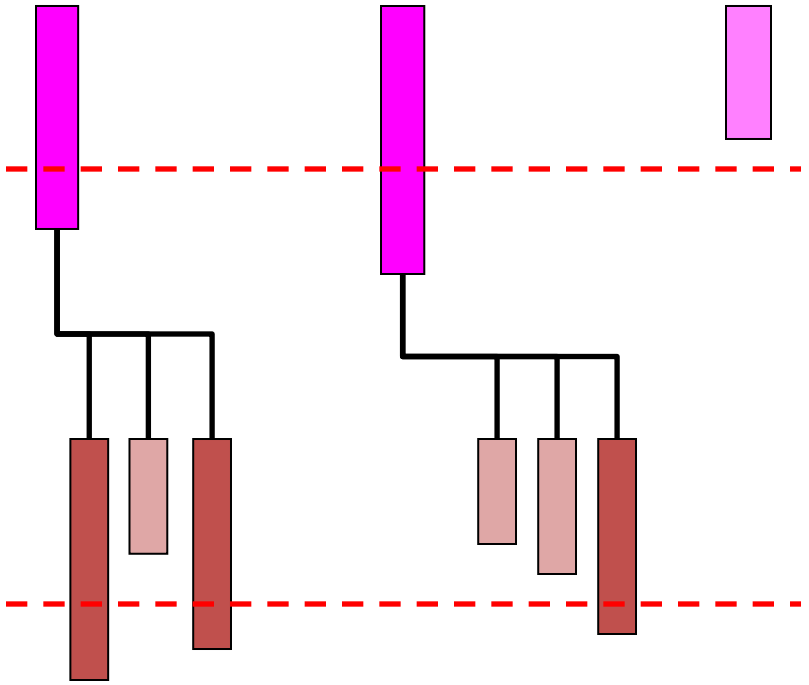
35% of structure
50% identical model

Better model if data resolution is high

Executing sequential searches

1st component

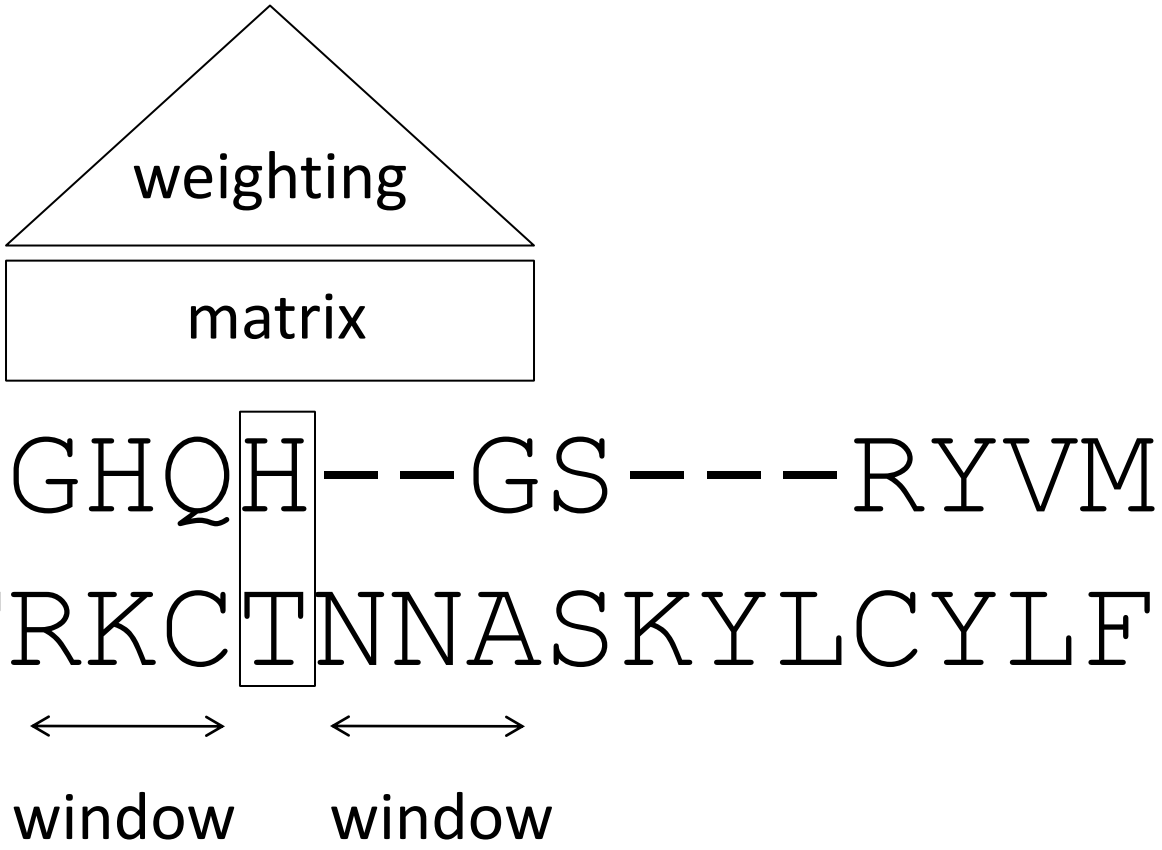
2nd component



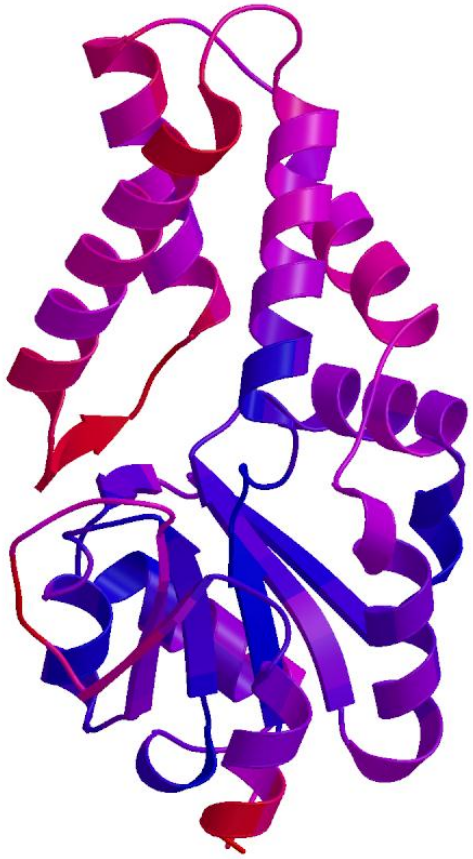
Sculptor: model improvement

- main chain
  Delete residues from the model that are not present in the target (e.g. extra domains)
- side chain
  - Prune side chain back to Cγ if residues are different
  - Add Cβ to GLY that is not GLY in the target
- B-factors
  - weight up segments where sequence is highly conserved (sequence similarity)
  - weight down regions with structural flexibility (solvent accessible surface area)

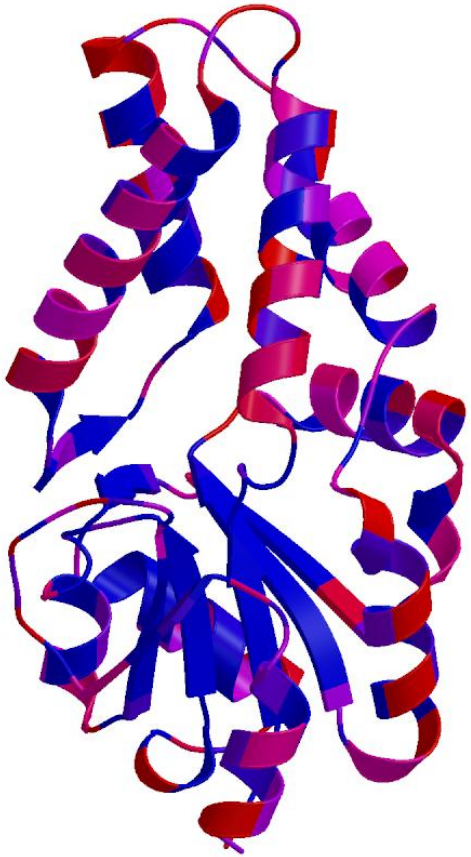
Sequence similarity



B-factor weighting



Sequence similarity
(residue level)



Accessible surface area
(atom level)

B-factor weighting

Solve tetragonal lysozyme structure

Tetragonal lysozyme model

- **Original B-factors:**
  LLG = 806.9
- **Modified B-factors:**
  LLG = 592.02

Monoclinic lysozyme model

- **Original B-factors:**
  LLG = 328.3
- **Modified B-factors:**
  LLG = 330.6

Sculptor: model improvement

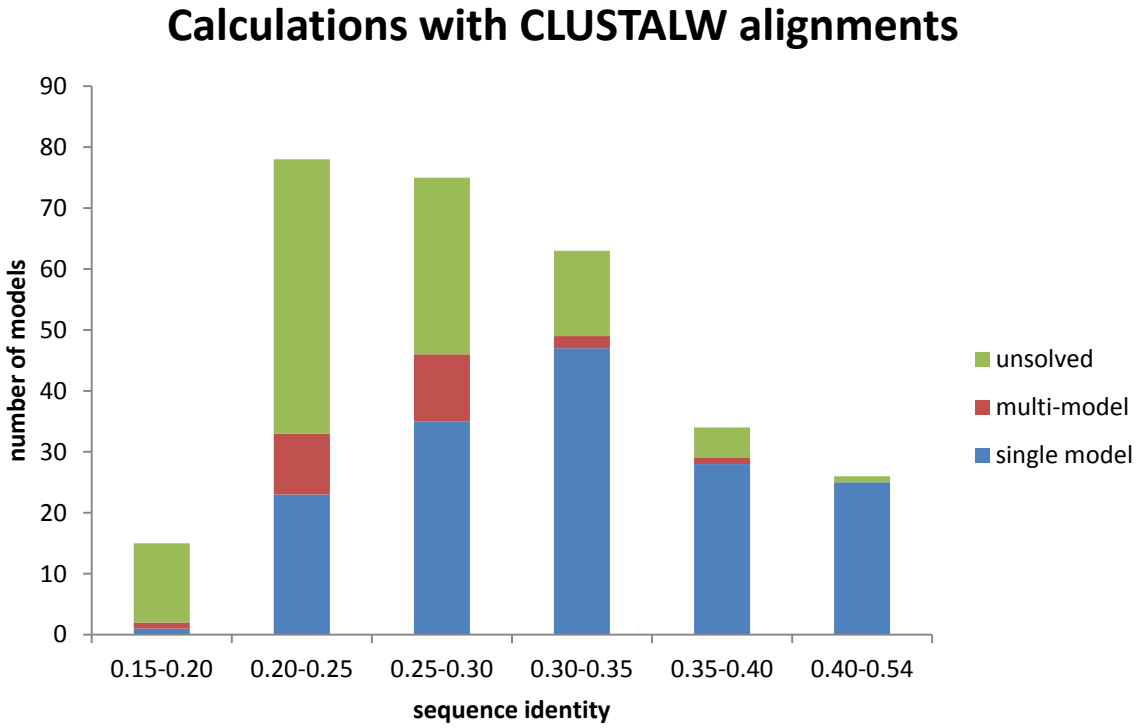
Model
Default

Model
SS mainchain

Model
Default
+ASA B-factors

Model
Default
+sequence similarity B-factors

Model
Default
+sequence similarity B-factors
+ ASA B-factors



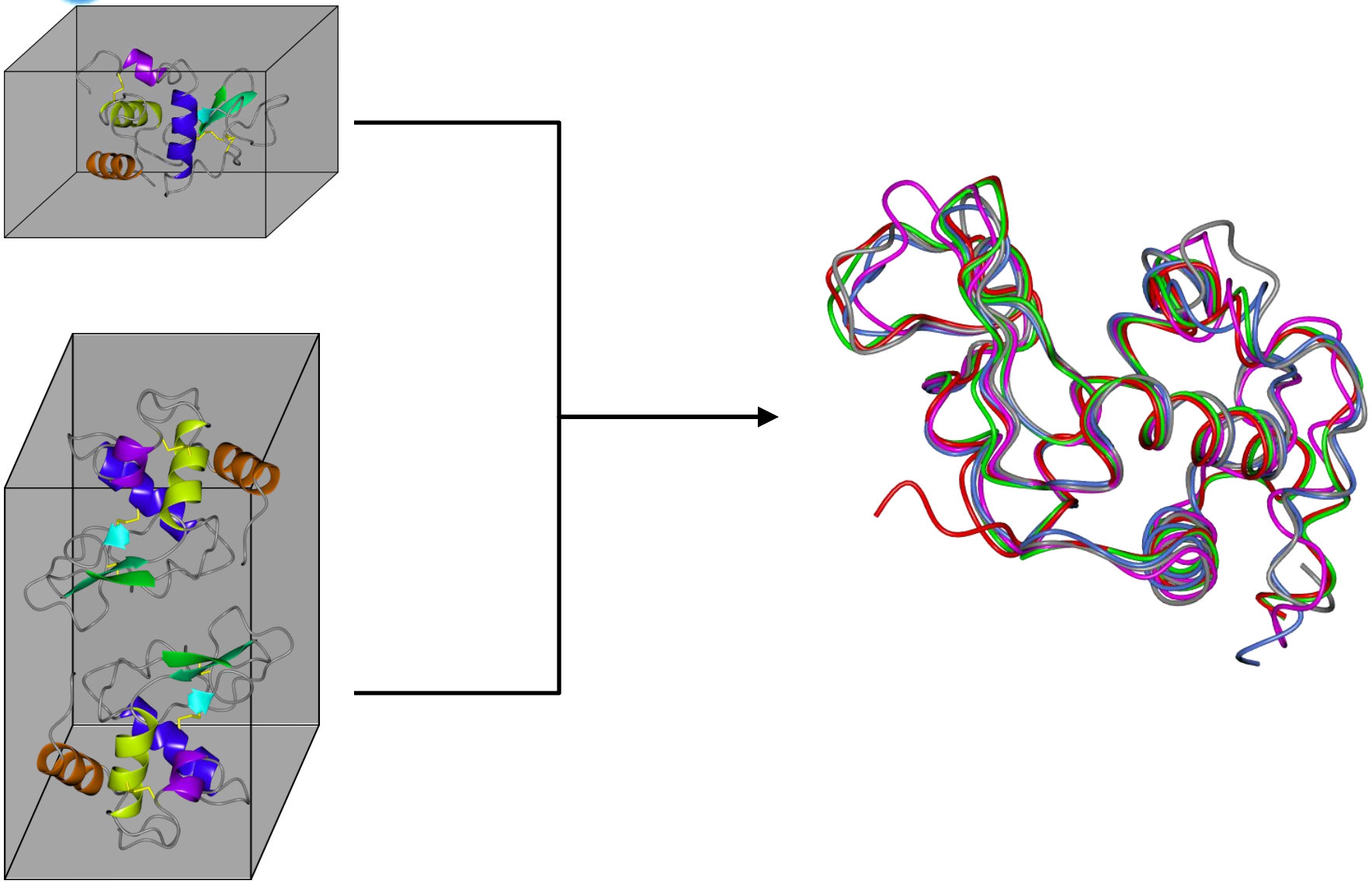
Protocols

Testcase: 1HJ9, 1 molecule/ASU, 32% identical

1.	2B9L_8.pdb	TFZ=7.3	LLG=41.731
2.	2B9L_10.pdb	TFZ=7.1	LLG=40.661
3.	2B9L_12.pdb	TFZ=7.6	LLG=40.644
4.	2B9L_7.pdb	TFZ=6.8	LLG=38.864
5.	2B9L_11.pdb	TFZ=7.3	LLG=38.582
6.	2B9L_9.pdb	TFZ=6.7	LLG=38.419

7.	2B9L_4.pdb	TFZ=6.2	LLG=26.755
11.	2B9L_1.pdb	TFZ=5.2	LLG=23.545

Ensembler: multiple superposition

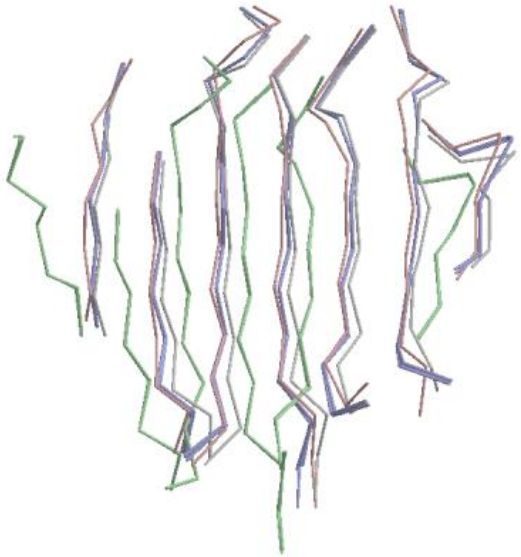


Ensembler: workflow

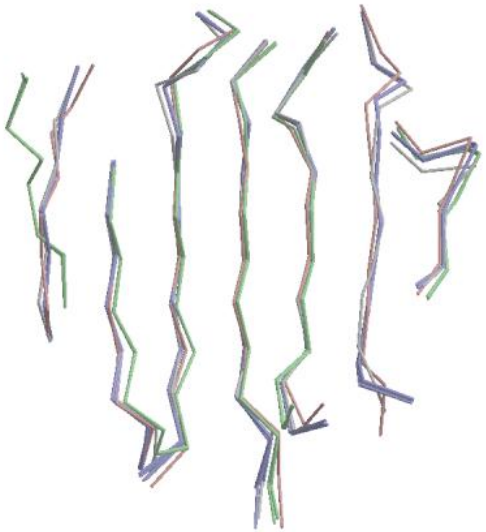
- Find protein chains (discard solvent)
- Align residues
  - SSM
  - MUSCLE
  - Alignments
- Match atoms (by name)
- Superpose
  - gapless: use atoms present in all structures
  - gapped: use sites present in at least two structures

Ensembler: weighting

- Robust-resistant weighting to superpose structurally similar regions
- Less sensitive to alignment errors

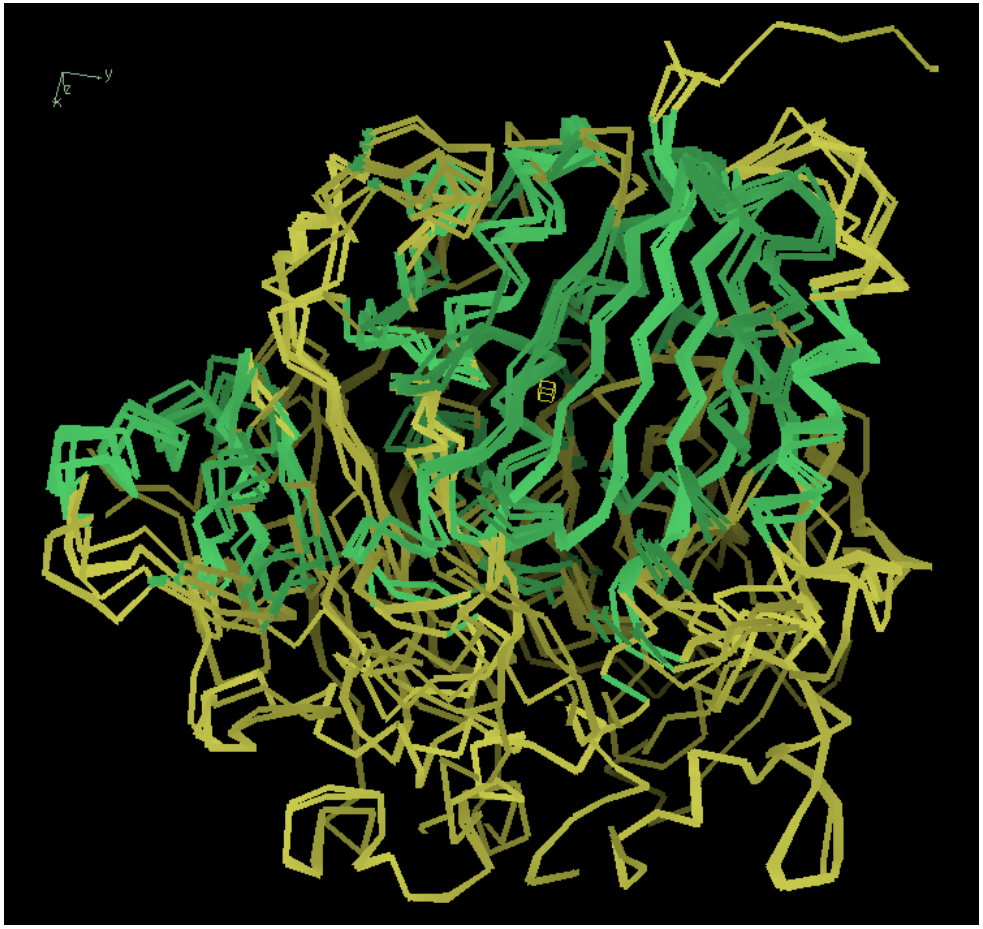


unweighted



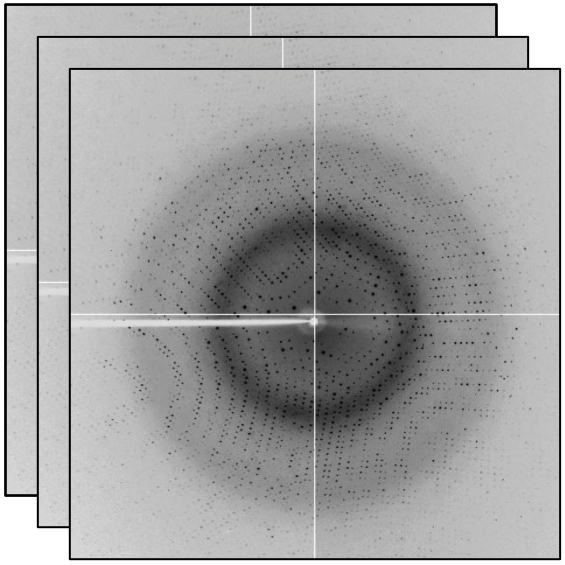
weighted

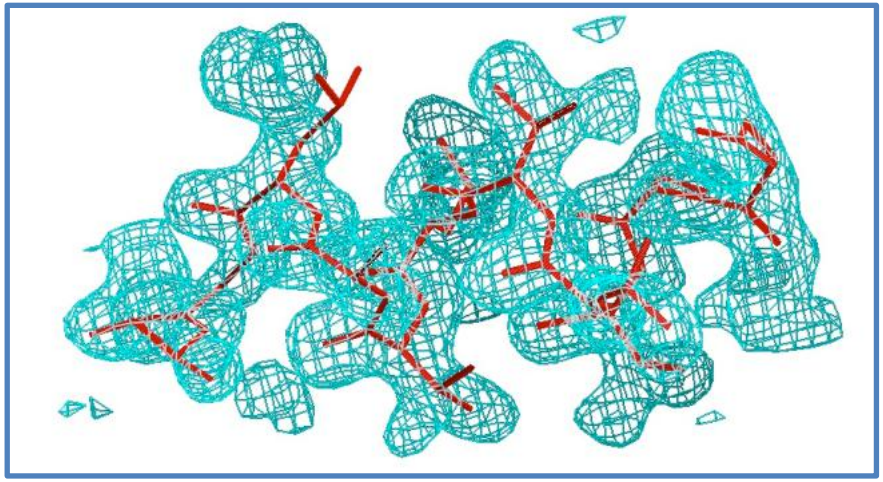
Ensembler: trimming



- based on unweighted positional rmsd
- superposed positions only

Molecular replacement pipeline





MRage

Goals

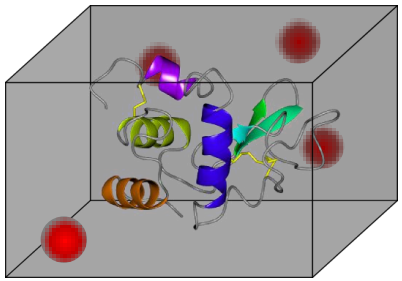
- **Solve easy cases quickly**
- **Solve difficult cases**
- **Simple user interface**
  Minimal input: sequence, data
- **Sufficient customisation**
- **Efficient use of computational resources**
  - Workstations and clusters
  - Scale with search complexity
- **Optimise for large number of models**

Search models

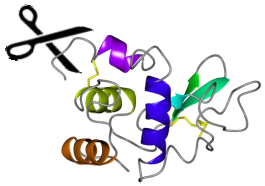
Homology search

HCYKS------GIQVR
HCVNSYQSNLDAIKIR

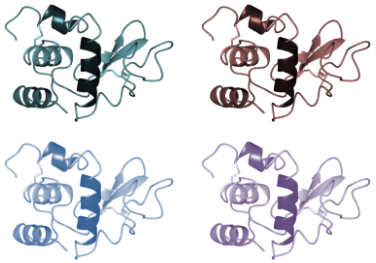
Alignment



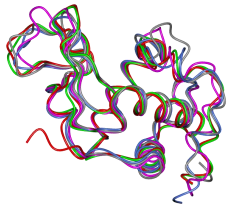
Template

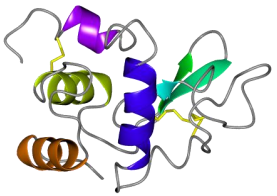


sculptor

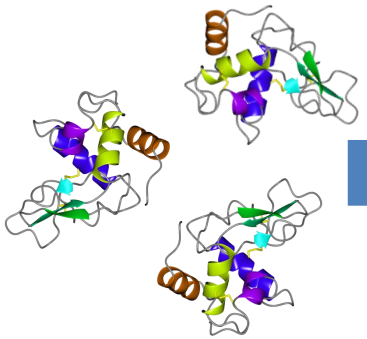


Models

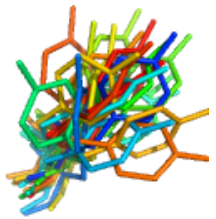




Ensemble Model



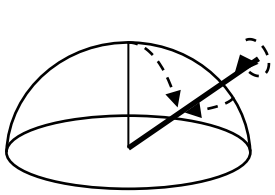
Models



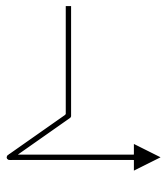
ensembler

BLAST

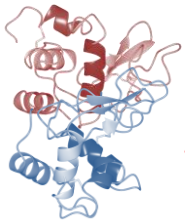
Molecular replacement

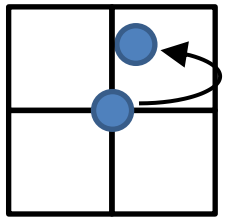


(φ, ψ, κ)

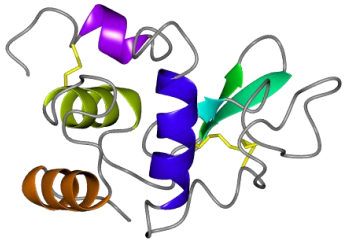


(x, y, z)





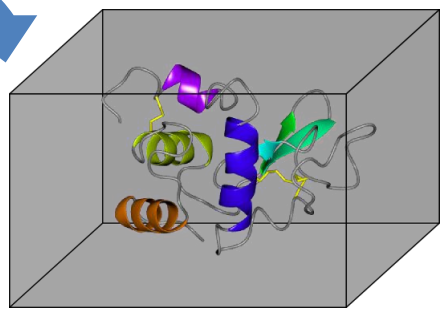
Default route



Model

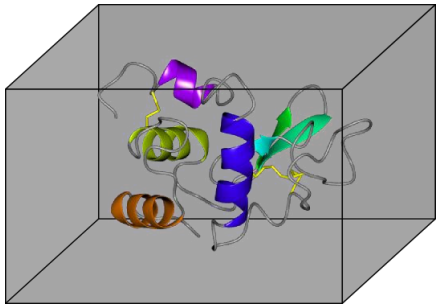
“Shortcut” routes

After a correct solution with a high Z-score has been found

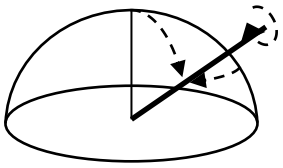


Solution

Superposition

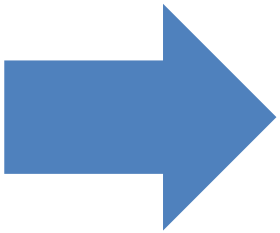


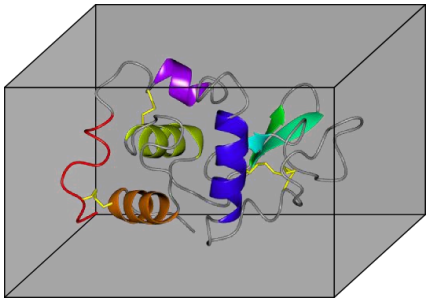
solution



(φ, ψ, κ) + (x, y, z)

superposition with alternative model (SSM)

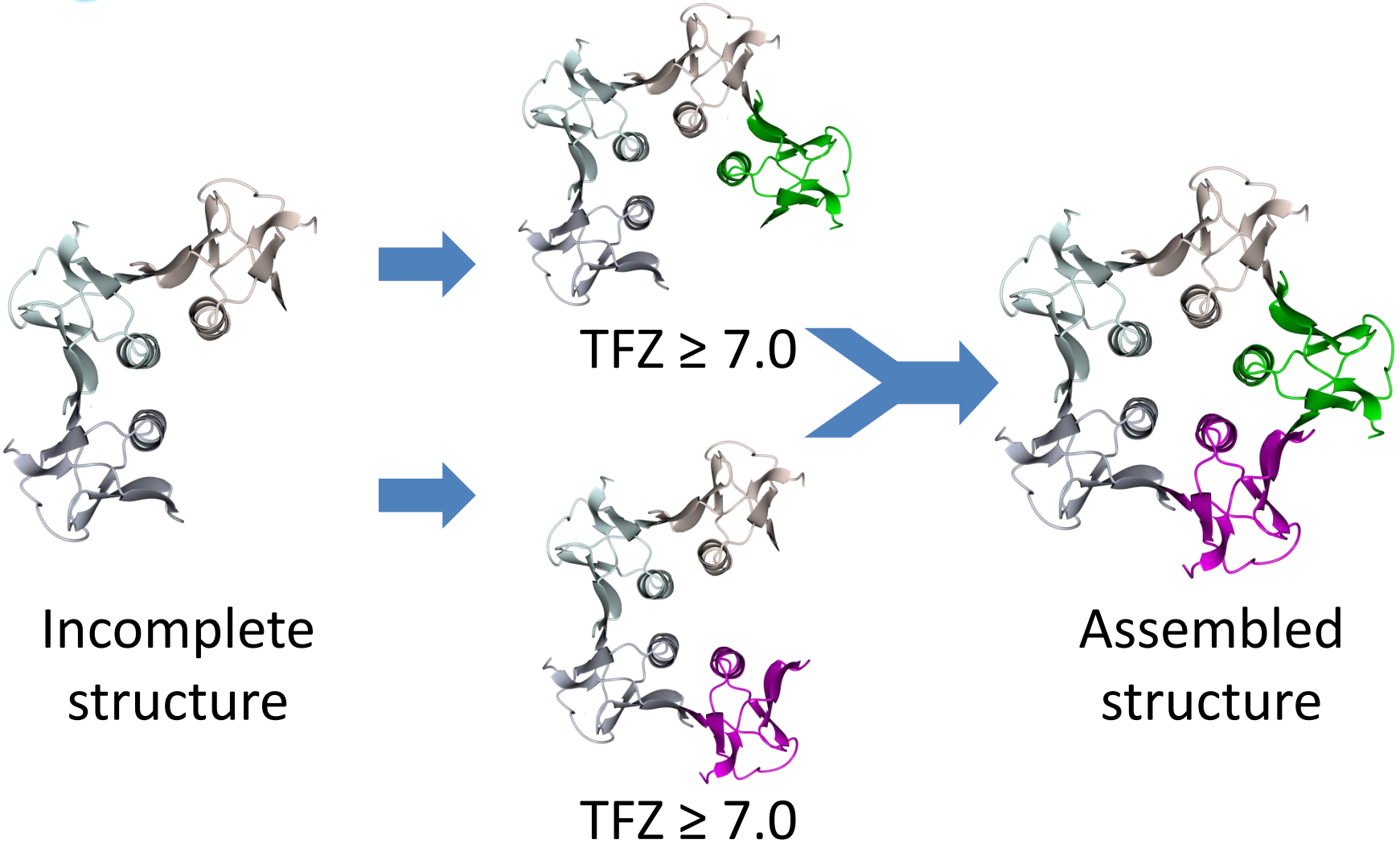




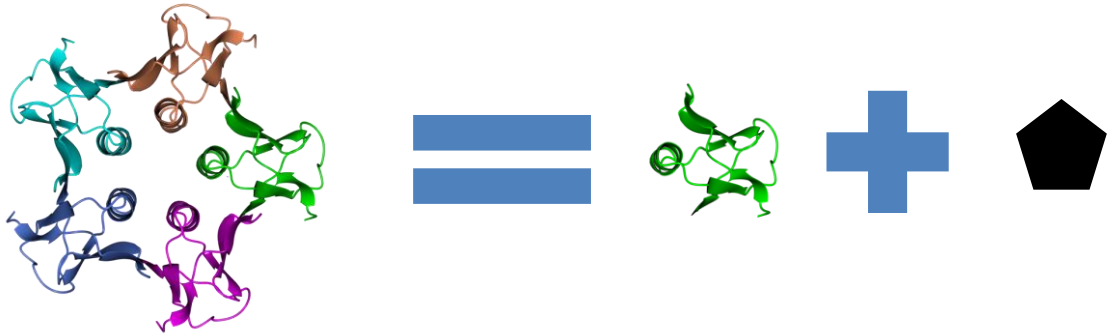
solution with alternative model

Find best

Amalgamation



Assembly recognition



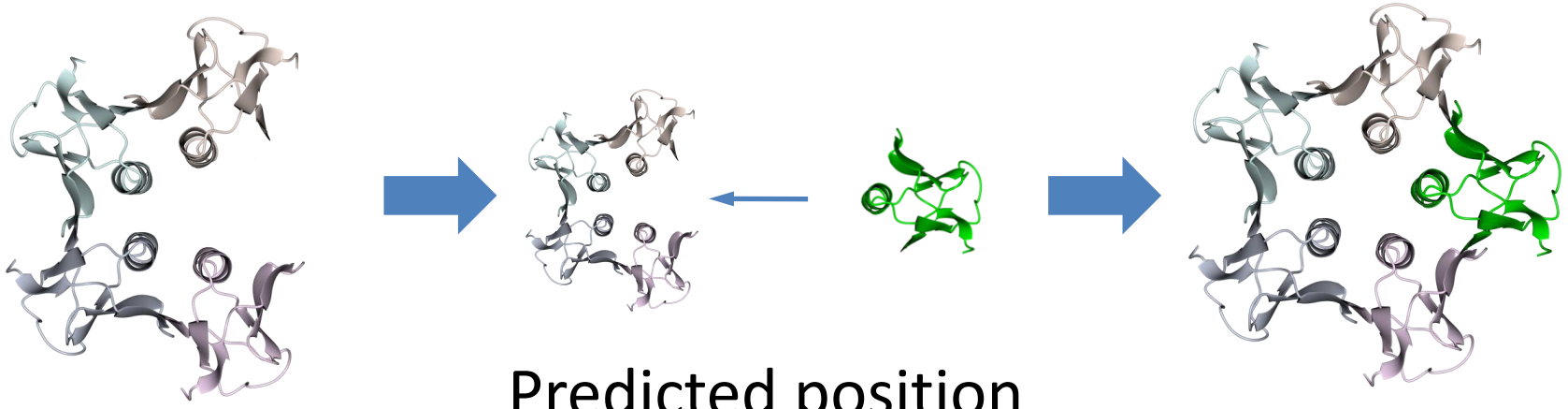
- identify (local) point group symmetry
- use full assembly in subsequent search
  better signal
- refine as set of individual molecules
  correct for local differences

Assembly recognition

Testcase: Shiga-like toxin, 1 model, 4 pentamers/ASU

Index	Model	TFZ	LLG	ΔLLG
1	monomer	5.3	43.1	43.1
2	monomer	11.9	154.7	111.6
	…			
10	monomer	22.3	2005.9	293.3
11	**5*monomer**	**42.6**	**3889.3**	**1883.4**
16	monomer	33.5	4545.4	656.1
	…			
19	monomer	38.1	6557.1	673.9
20	monomer	9.2	7322.9	765.8

Assembly completion

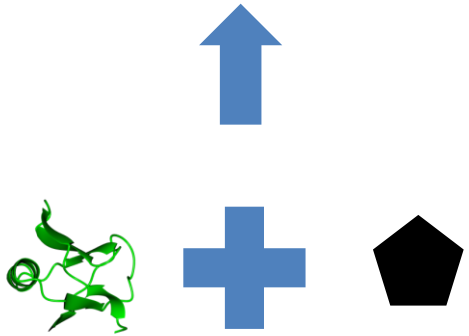


Incomplete assembly

Predicted position

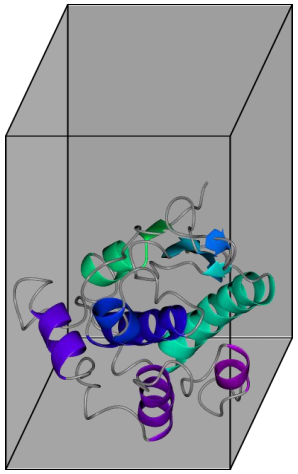
Completed assembly

calculate TFZ and LLG

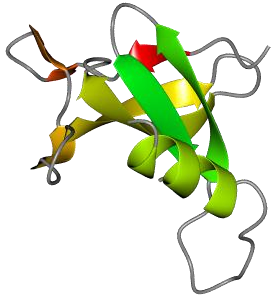


Known assembly

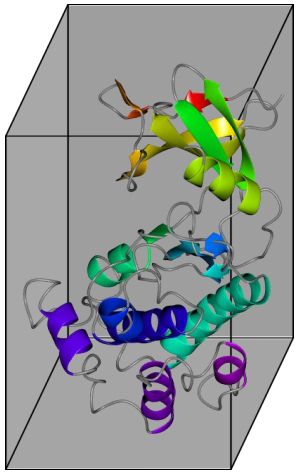
Assembly completion



Incomplete assembly

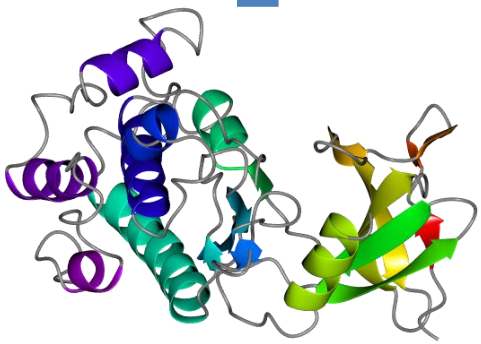


Predicted position (approximately correct)



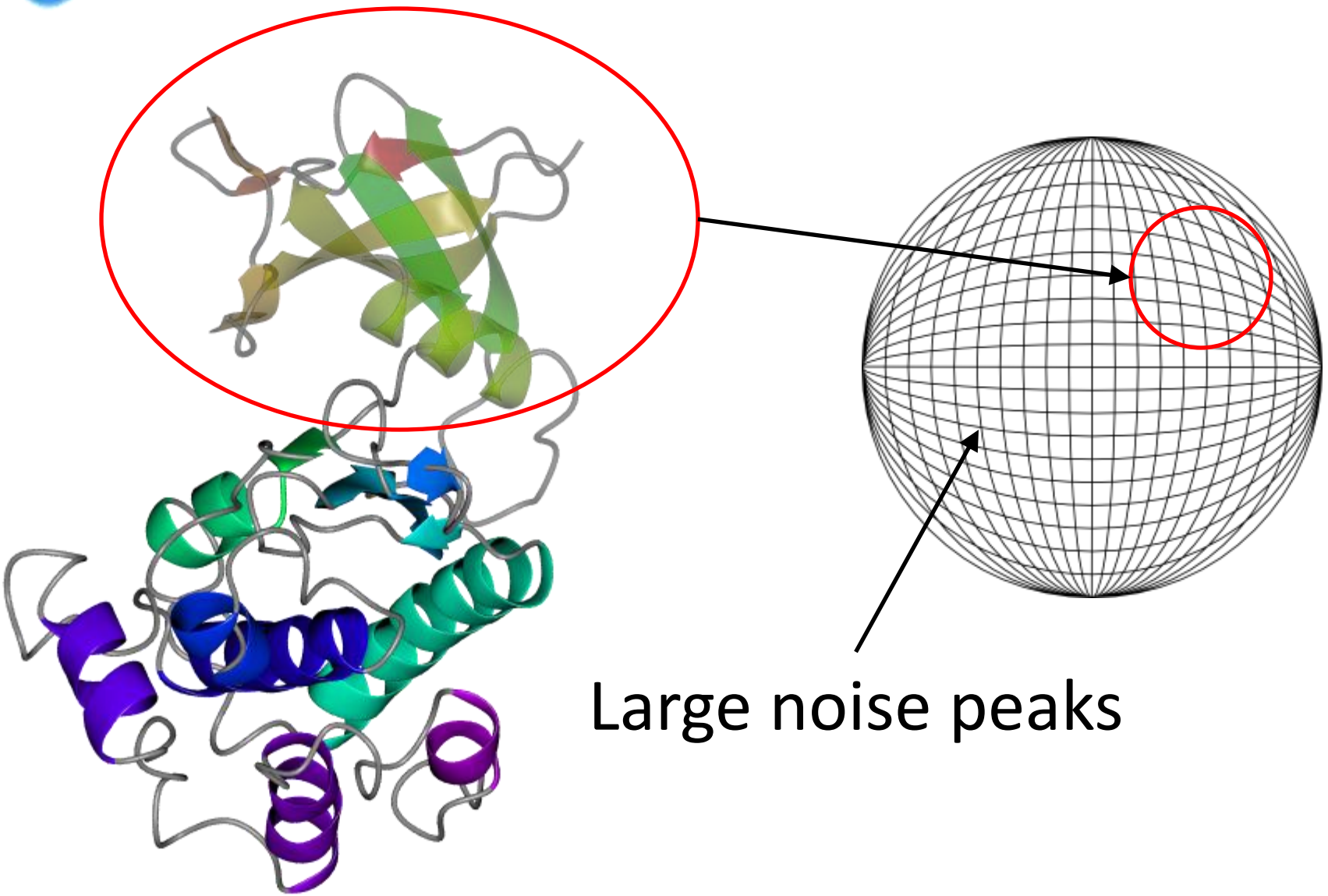
Completed assembly

calculate TFZ and LLG

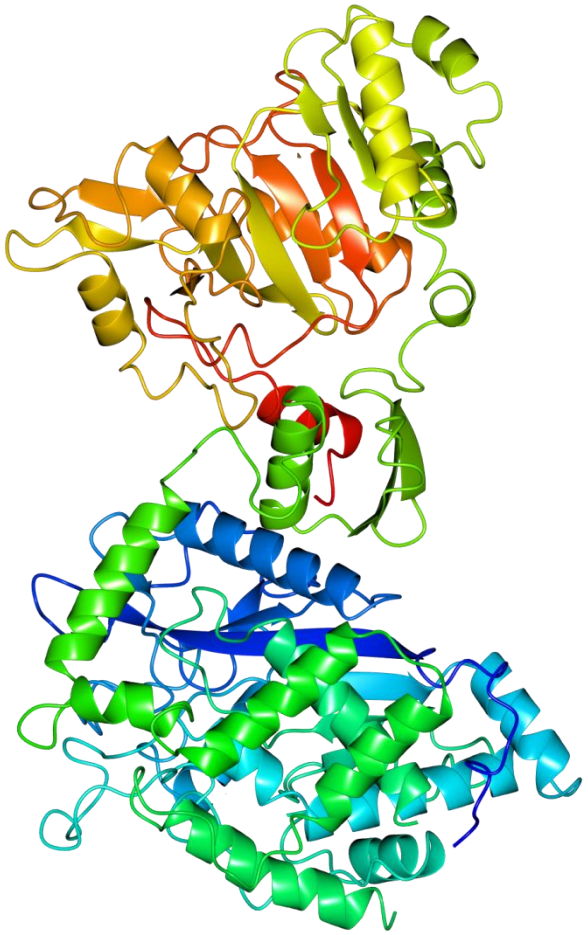


Known assembly

Local search



Fully automated solution



3LN6

- 750 residues, 1 chain/asu
- 2.95 Å, I 4_1 2 2
- Solved by SIRAS

- 82 HHPRED hits

Fully automated solution

Input

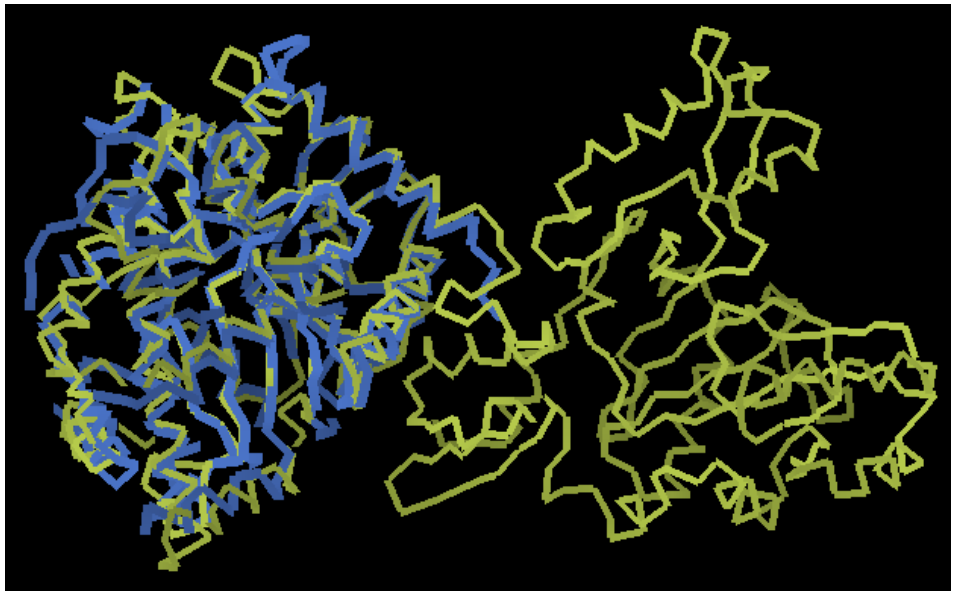
```
component
{
   sequence=3ln6.seq

   homology
   {
      file_name=3ln6.hhr
   }
}
```

After 5 minutes (50 CPUs)

Significant hit:

3NZT_9, TFZ=10.0, LLG=80



Model is only partial!

Sequence-based assembly

MIIDRLLQRSHSHLPILQATFGLERESLRIHQPTQRVAQTPHPKTLGSRNYHPYIQTDYSEPQLE

1. hit:

2. hit:

negligible overlap

N. hit:

negligible model

Fully automated solution

After one day (50 CPUs)

Solution:

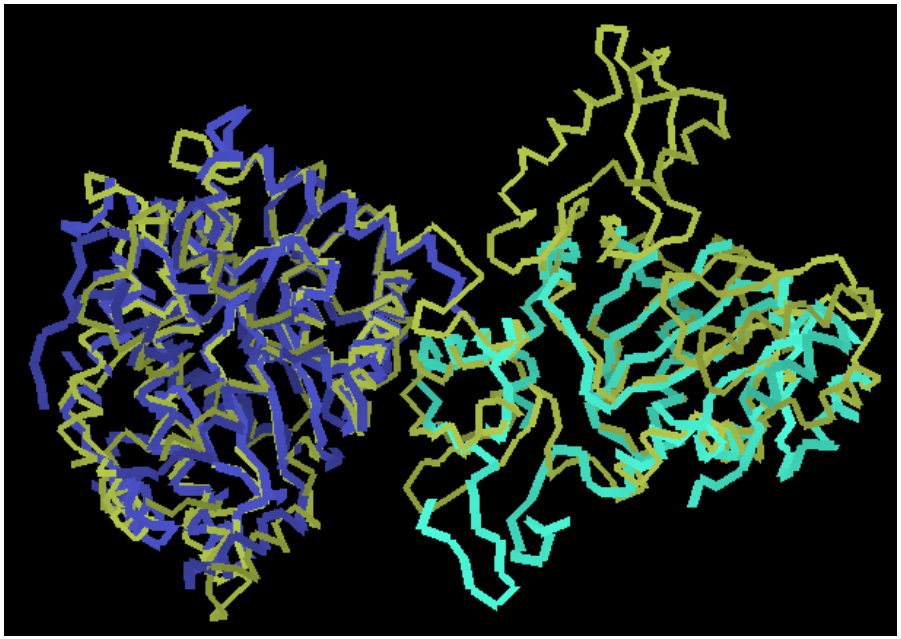
3NZT_9, LLG=80

1UC8_11, TFZ=6.4, LLG=143

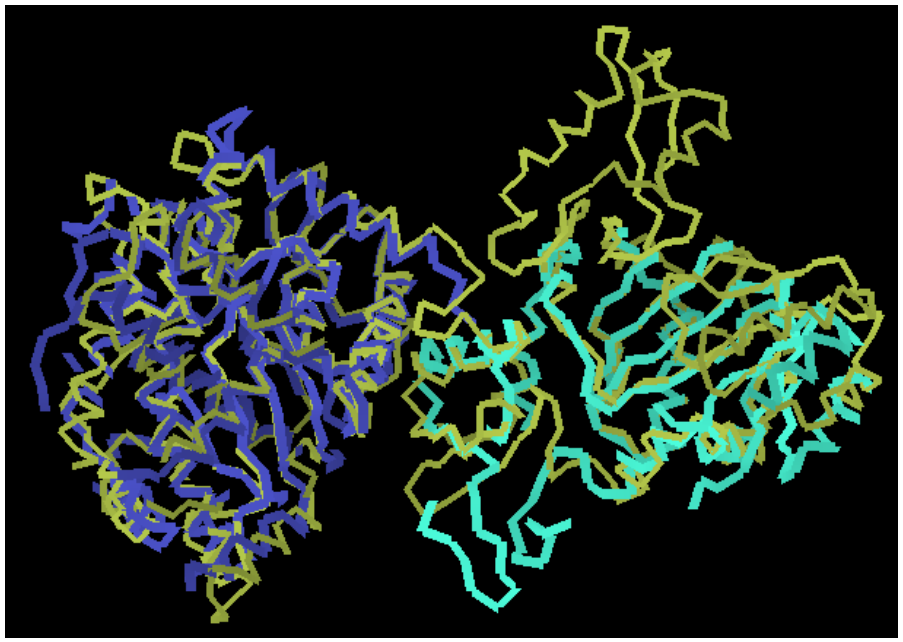
Processed:

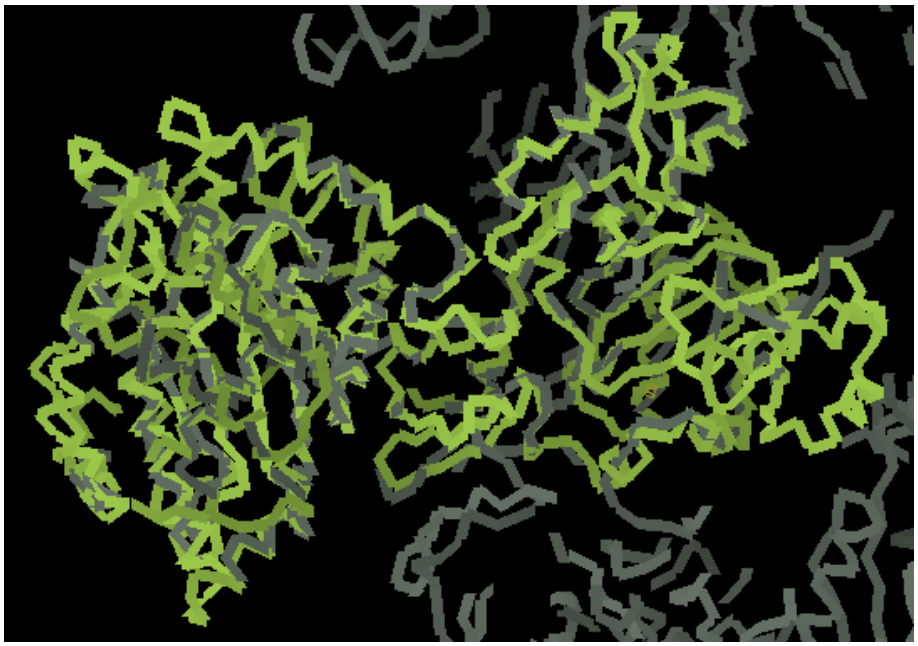
82 templates
845 models
…
1 000 000+ translation peaks

Solution



Fully automated solution





Solution

After AutoBuild

431 built, 92 placed

R=0.4058/0.4740

Acknowledgements

PHENIX:

- Paul Adams
- Nat Echols

CIMR:

- Stephen Graham
- Nikol Simecek

Industrial Consortium:

- Joel Bard
- Herb Klei
- Bob Nolte
- Steven Sheriff

CCP4 team

Acknowledgements

Python-based Hierarchical Environment for Integrated Xtallography

Computational Crystallography Initiative / LBNL
Paul Adams, Ralf Grosse-Kunstleve, Pavel Afonine, Nathaniel Echols, Jeffrey Headd, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Los Alamos National Lab
Tom Terwilliger, Li-Wei Hung

University of Cambridge
Randy Read, Airlie McCoy, Robert Oeffner

Duke University
Jane and Dave Richardson, Vincent Chen