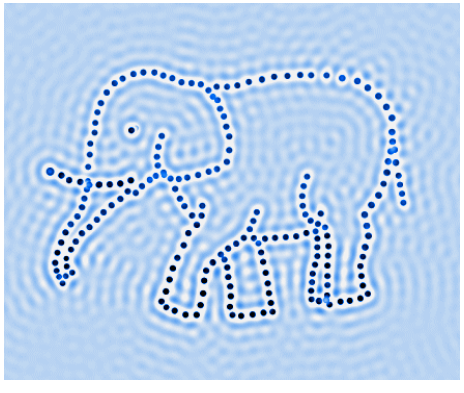
Lecture 15

Maya Topf

Fitting of structures, scoring and assessment



EMBO course on image processing for cryo EM

10 September 2019

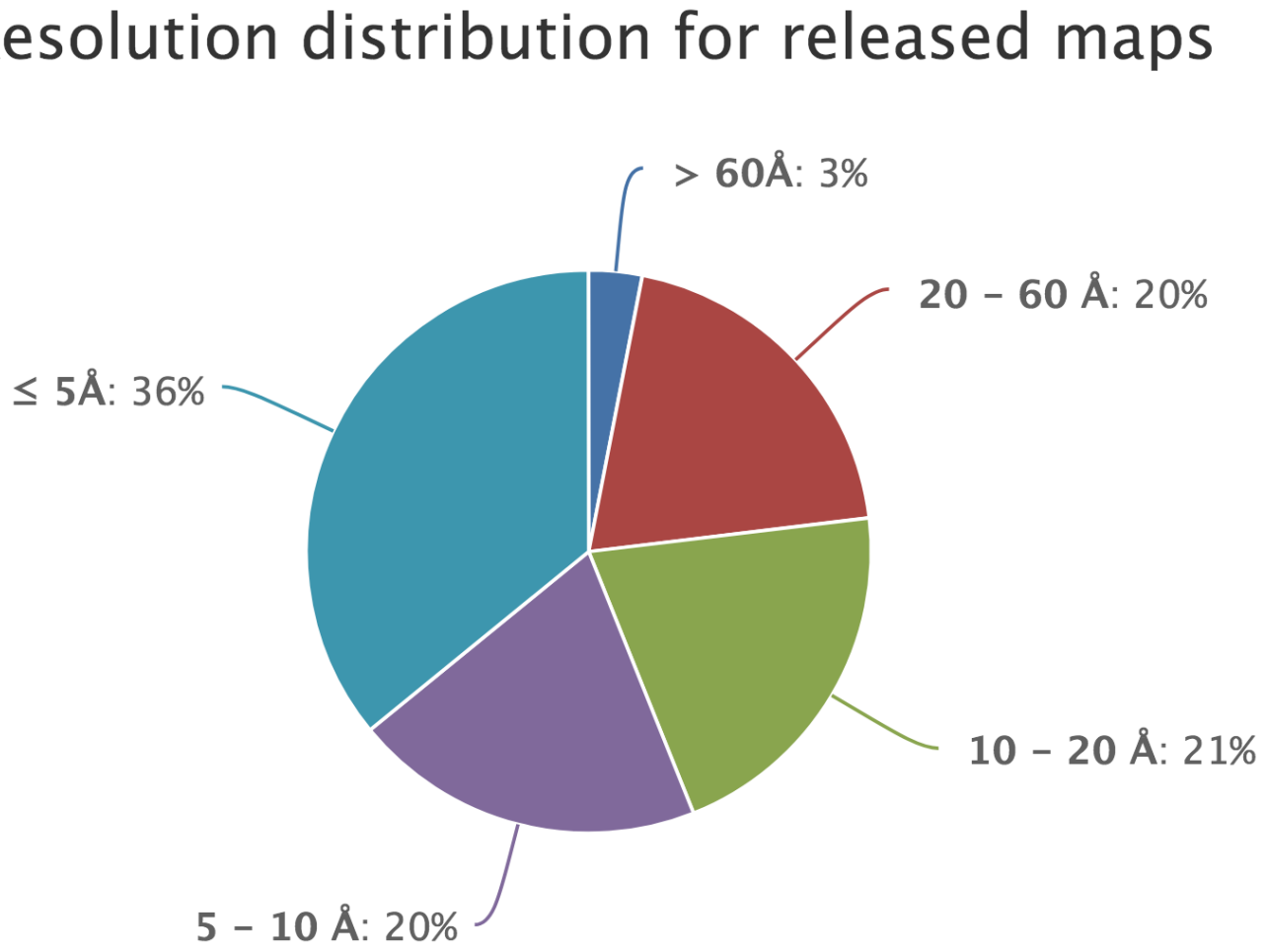
EMBO

ISMB
Institute of Structural and Molecular Biology

Aims of this lecture:

- To understand 3D-EM density fitting and what we can achieve with it.
- To describe the different types of density fitting methods:
  - rigid fitting
  - flexible fitting
  - assembly (multiple) fitting
- To be aware of some software tools used for visualization and density fitting.

EMDB Statistics



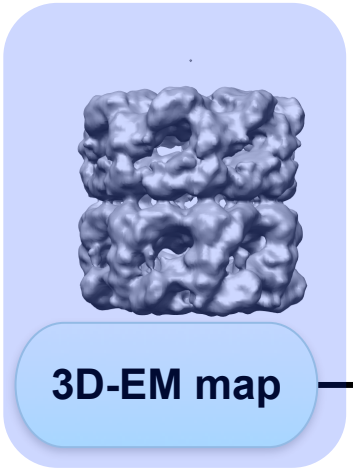
~More than half of the maps are better than ~10 Å resolution!

https://www.ebi.ac.uk/pdbe/emdb/

Q: How can we get more out of these 3D-EM maps?

A: We can use them as a constraint in model building

Model building and refinement is often **interactive** and **iterative**.



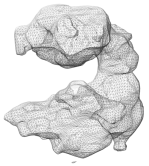
Segmentation

Component structure known?

Yes

No

What resolution?



Component structure

10-20Å

~4.5-10Å

<4.5Å

Component Sequence

Fitting all known folds

SSE assignment

Cα chain tracing, all-atom modelling

Fold assignment from sequence

Template detected?

Yes

No

'template-free' modelling

Homology modelling

Rigid fitting

fit different from map?

Yes

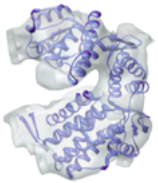
No

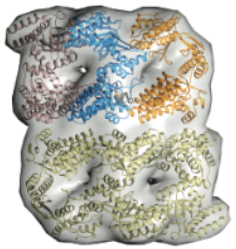
Multiple conformations

ENM / NMA

Real-space refinement

MD-based methods

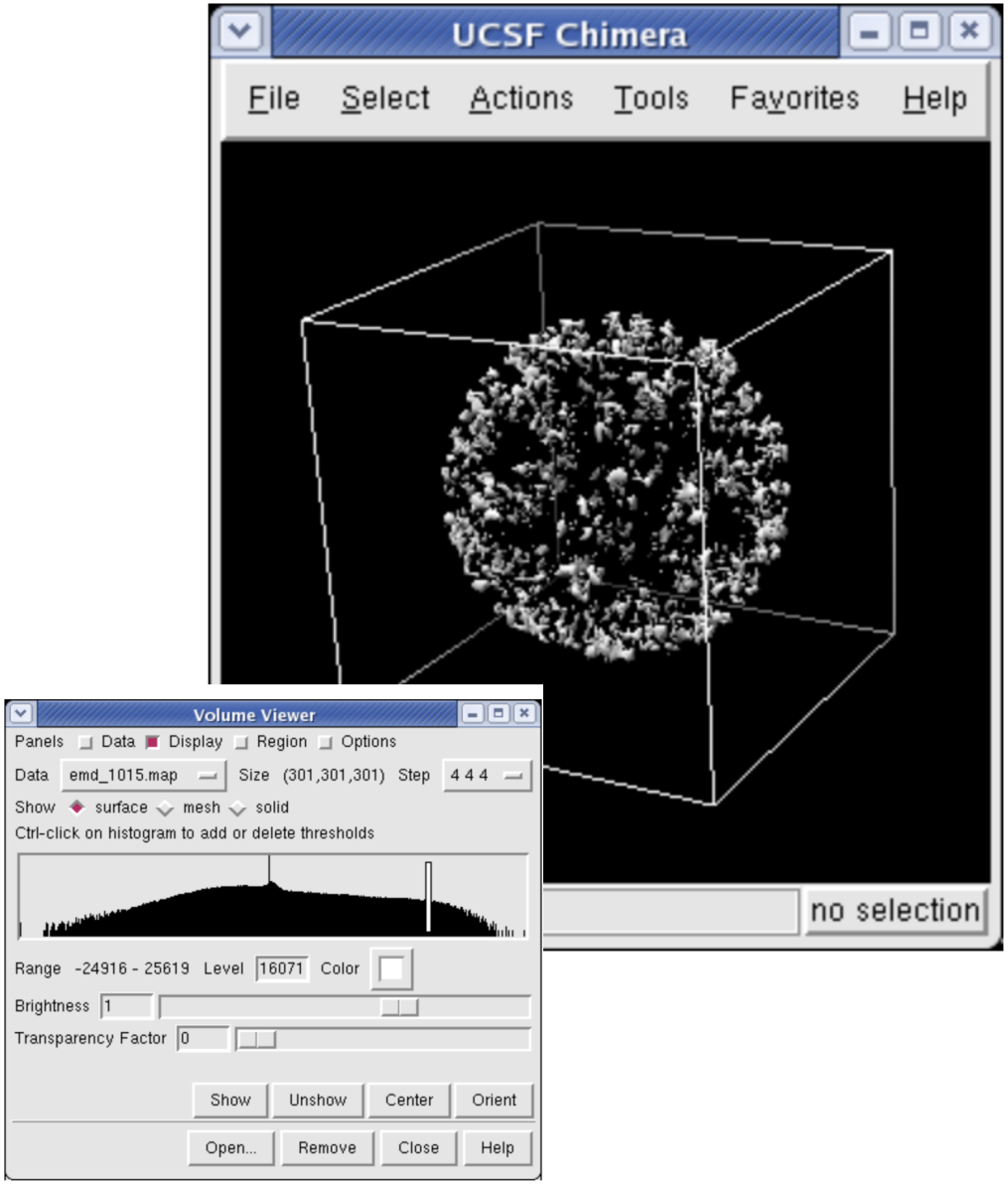


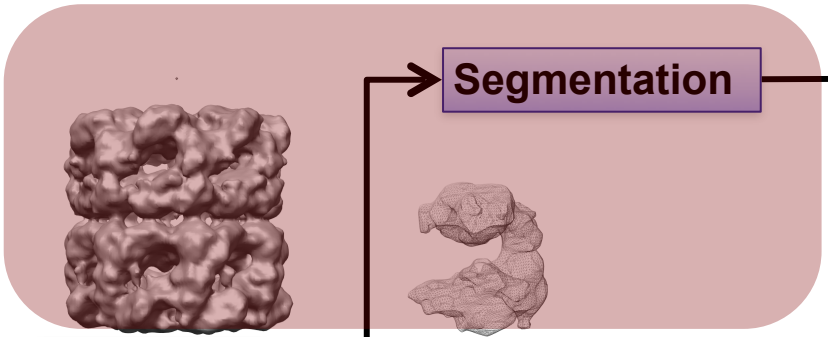


Visualization tools

- **Academic programs:**
- Chimera, ChimeraX
- Coot
- Python Molecular Viewer (PMV)
- VMD
- VolRover
- Gorgon

- **Commercial programs:**
- PyMOL (Schrödinger)
- Amira (Thermo Fisher Scientific)





3D-EM map

Segmentation

Component structure known?

Yes

No

What resolution?

Component structure

Component Sequence

10-20Å

Fitting all known folds

~4.5-10Å

SSE assignment

<4.5Å

Cα chain tracing, all-atom modelling

Fold assignment from sequence

Template detected?

Yes

No

'template-free' modelling

Homology modelling

Rigid fitting

fit different from map?

Yes

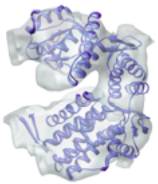
No

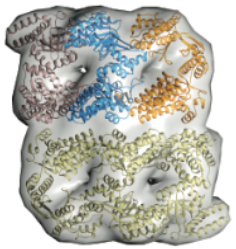
Multiple conformations

ENM / NMA

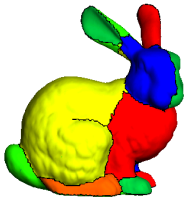
Real-space refinement

MD-based methods





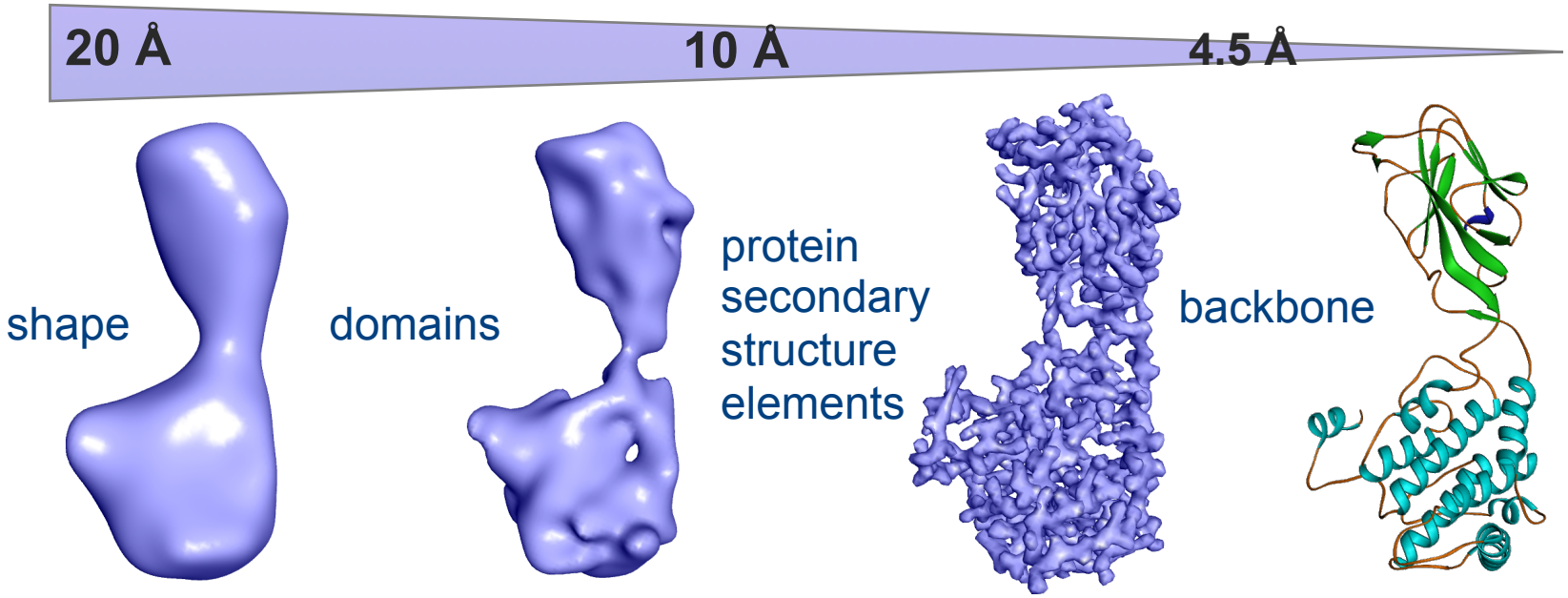
Segmentation



- Identify boundaries between 3D regions that represent structural components in the context of *structural*, *biochemical* and *bioinformatics* knowledge.

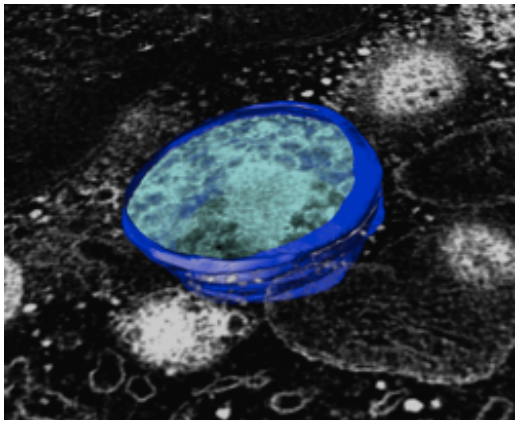
- The identified boundaries can be useful in detecting the positions of known component structures in the map.

- The size of the segmented components is related to the map resolution.

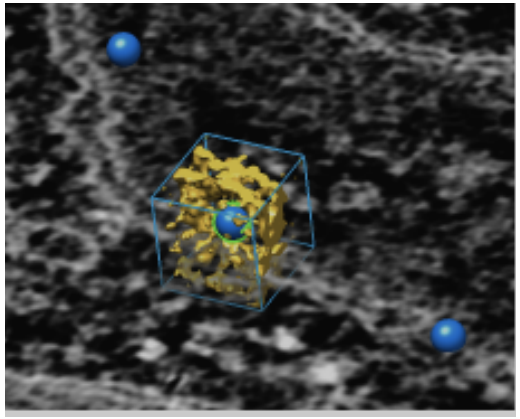


Segmentation

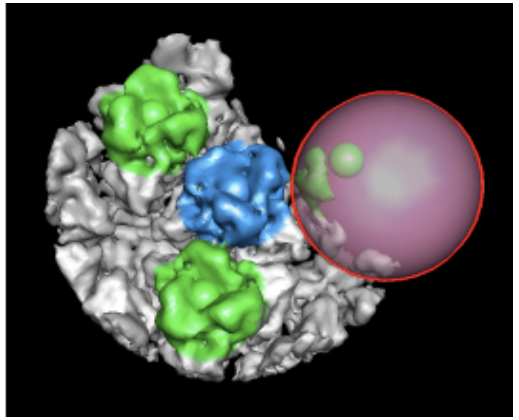
- Manual segmentation



Mask



Box around marker/atoms

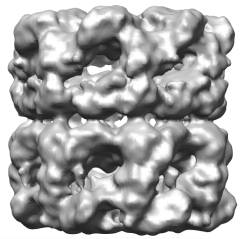


Hand erasing

- Automated segmentation: based on density alone, with *or* without the use of symmetry information. (*e.g.* in VolRover, Segger, Amira, IMOD)

- Knowledge-based segmentation:

- Antibody labelling; gold clusters; subunit/domain deletion (difference mapping).
- Recognition of structural components - **density fitting**.



3D-EM map

Segmentation

Component structure known?

Yes

No

What resolution?

10-20Å

~4.5-10Å

<4.5Å

Fitting all known folds

SSE assignment

Cα chain tracing, all-atom modelling

Component structure

Component Sequence

Fold assignment from sequence

Template detected?

Yes

No

'template-free' modelling

Homology modelling

Rigid fitting

fit different from map?

Yes

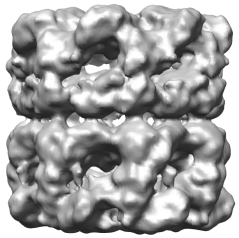
No

Multiple conformations

ENM / NMA

Real-space refinement

MD-based methods



3D-EM map

Segmentation

Component structure known?

Yes

No

What resolution?

10-20Å

~4.5-10Å

<4.5Å

Fitting all known folds

SSE assignment

Cα chain tracing, all-atom modelling

Component structure

Component Sequence

Fold assignment from sequence

Template detected?

Yes

No

'template-free' modelling

Homology modelling

Rigid fitting

fit different from map?

Yes

No

Multiple conformations

ENM / NMA

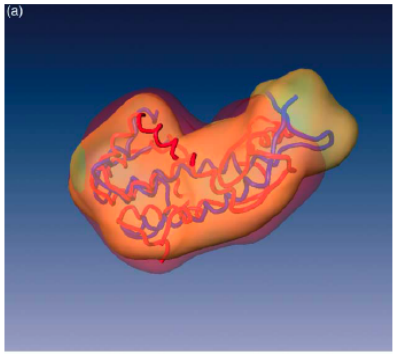
Real-space refinement

MD-based methods

Fold recognition from density

< ~10-20 Å: Fit domains from a non-redundant protein domain database (e.g. CATH);

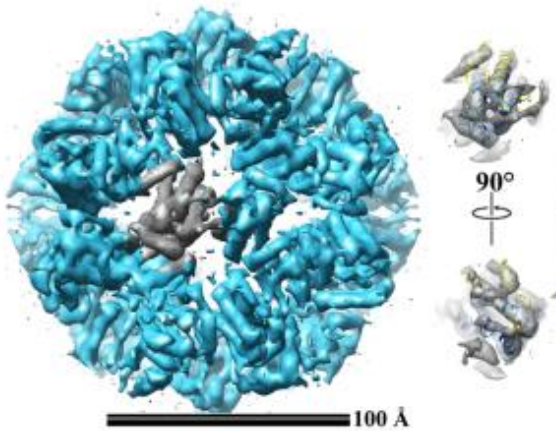
12 Å



Fitting of a domain from1.20.1060.10 (mainly alpha) into 1.10.530.10 (mainly-alpha).

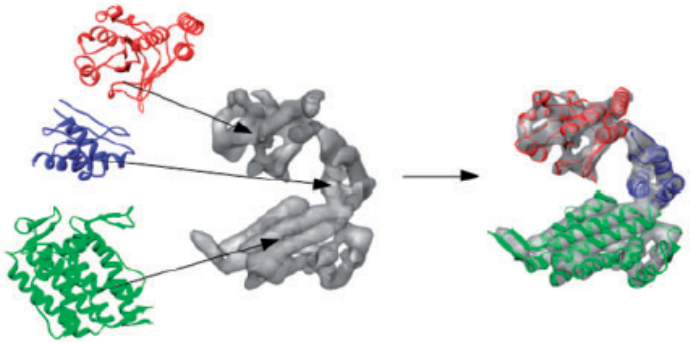
Velazquez-Muriel et al. *JMB 2005*

7 Å



Detection of bacteriophage Lambda

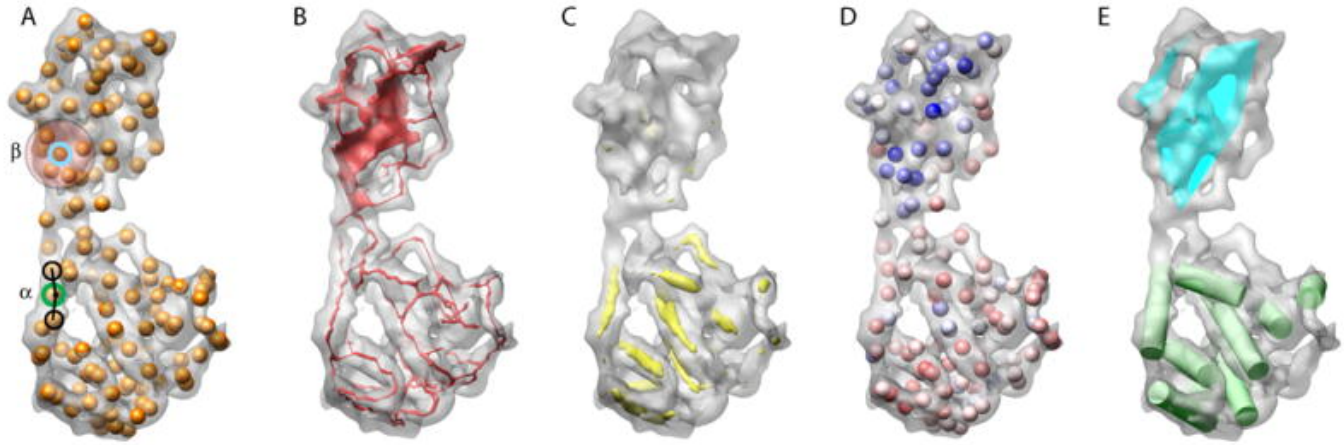
Khayat et al. *JSB 2010*



FOLD-EM: Saha et al. *Bioinformatics 2012*

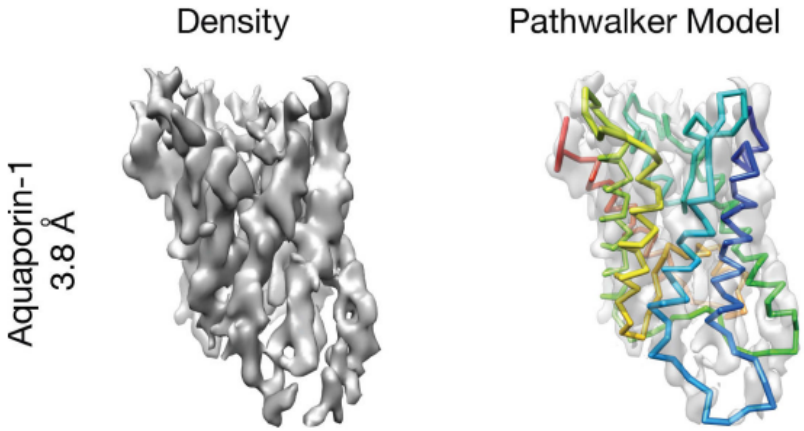
Fold recognition from density

~4.5-10 Å: secondary structure element detection



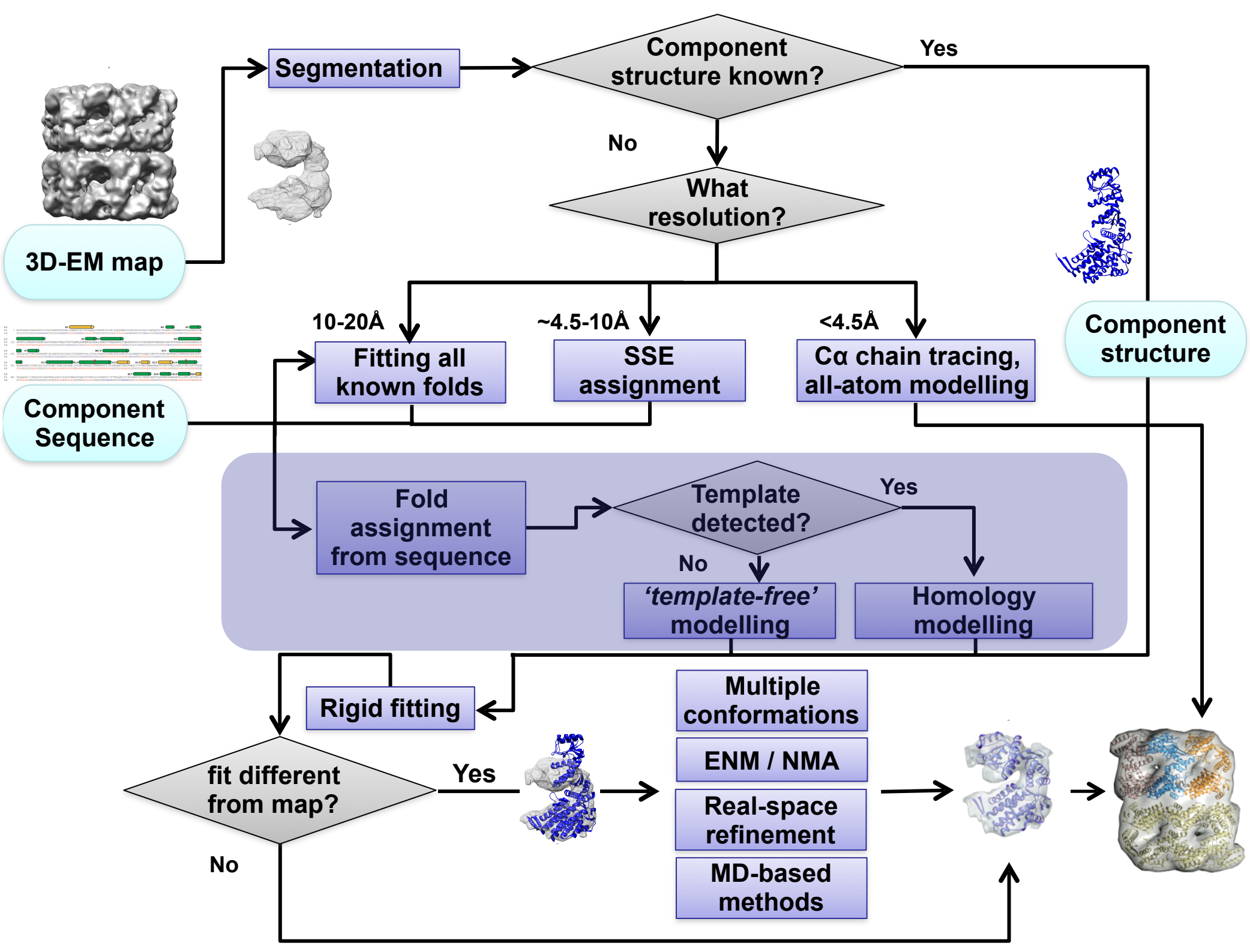
Baker et al. *Structure* 2007

4.5 Å and better: *de novo* Cα tracing and model building



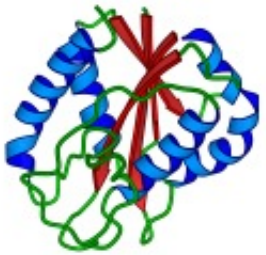
Baker et al. *Structure* 2012

Programs: SSEhunter (Gorgon), SSETracer, Ematch, Pathwalker, **Coot**, **Buccaneer**, EM-fold, Rosetta (sequence information), Phenix autobuild, ARP/wARP, MAINMAST


3D-EM map
Segmentation
Component structure known?
Yes
No
What resolution?
Component structure
Component Sequence
10-20Å
~4.5-10Å
<4.5Å
Fitting all known folds
SSE assignment
Cα chain tracing, all-atom modelling
Fold assignment from sequence
Template detected?
Yes
No
'template-free' modelling
Homology modelling
Rigid fitting
fit different from map?
Yes
No
Multiple conformations
ENM / NMA
Real-space refinement
MD-based methods

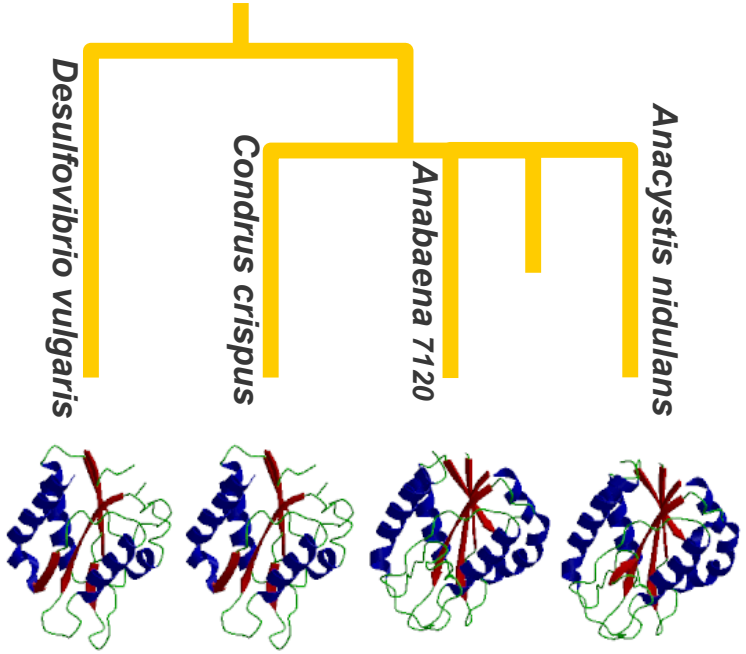
Fold recognition from sequence

GFCHIKAYTRLIMVG…



Template-free

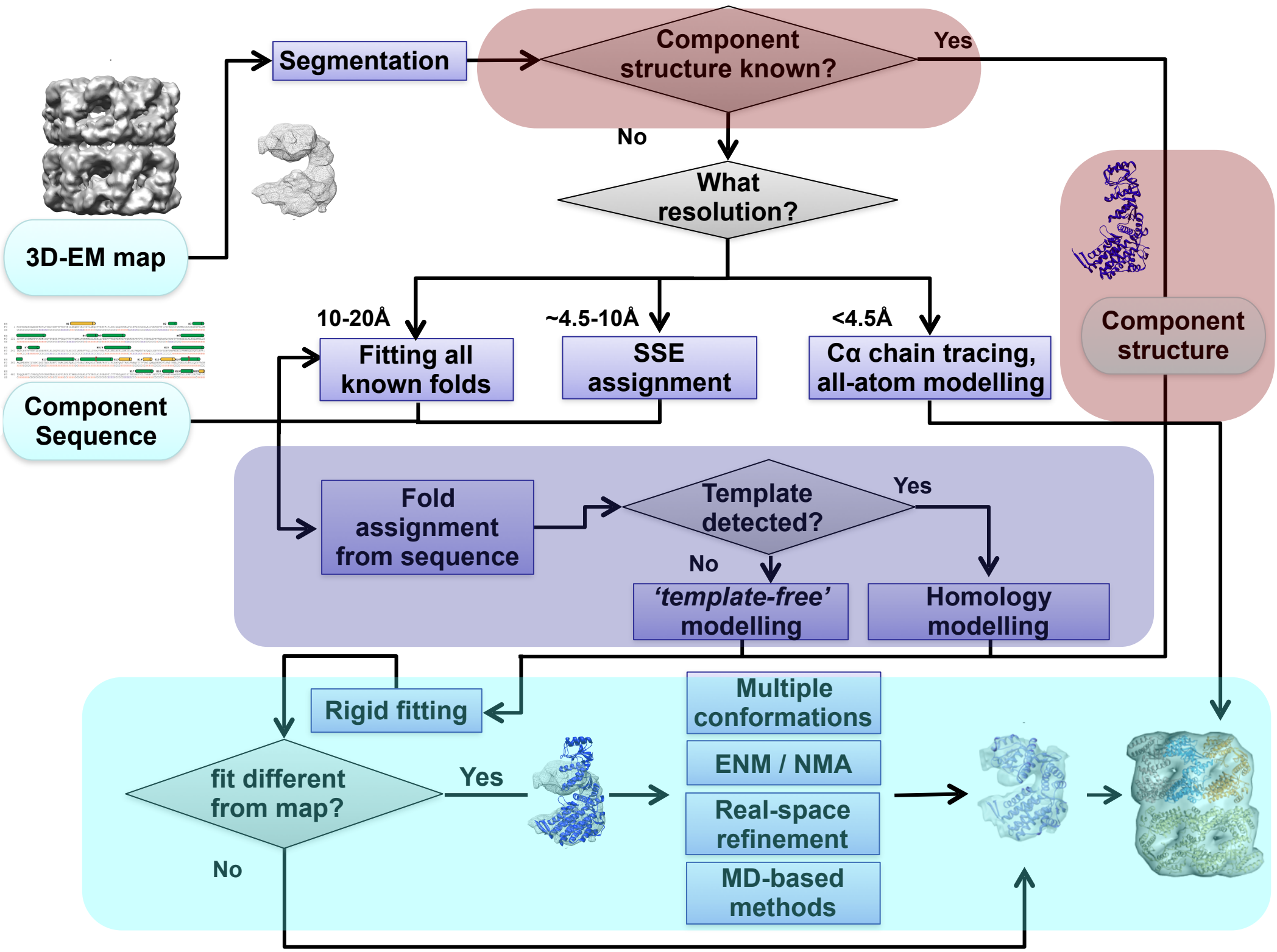
Ab initio (*de novo*) prediction
Fragment Assembly
Evolutionary Couplings



Template-based

Threading
Comparative (Homology) Modelling

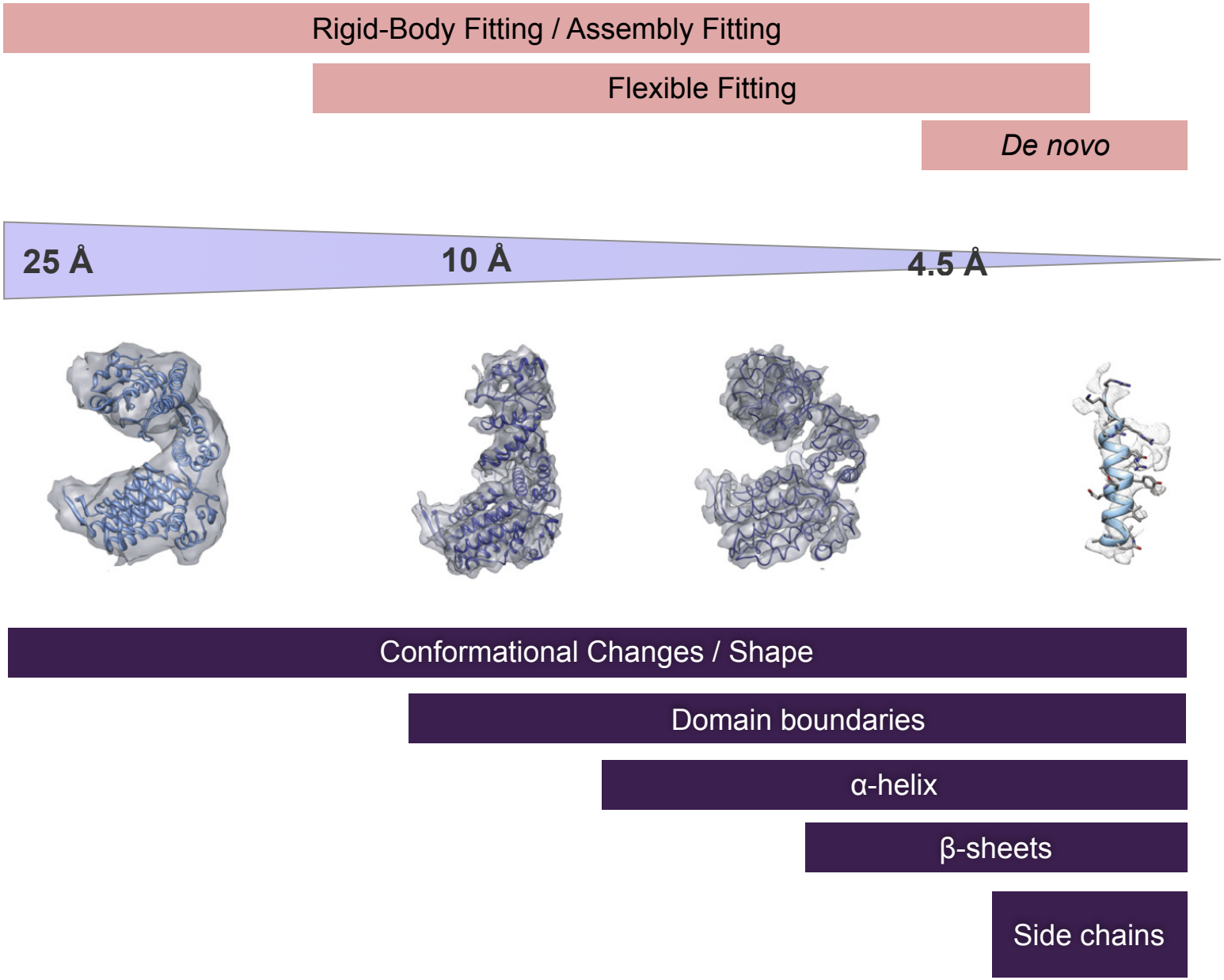
Programs: MODELLER, SWISS-MODEL, Phyre2, RaptorX, I-TASSER, Rosetta, EVfold…


3D-EM map
Segmentation
Component structure known?
Yes
No
What resolution?
Component Sequence
10-20Å
~4.5-10Å
<4.5Å
Fitting all known folds
SSE assignment
Cα chain tracing, all-atom modelling
Component structure
Fold assignment from sequence
Template detected?
Yes
No
'template-free' modelling
Homology modelling
Rigid fitting
fit different from map?
Yes
No
Multiple conformations
ENM / NMA
Real-space refinement
MD-based methods

Density fitting

Find the best “match” between the atomic model and the 3D-EM density map

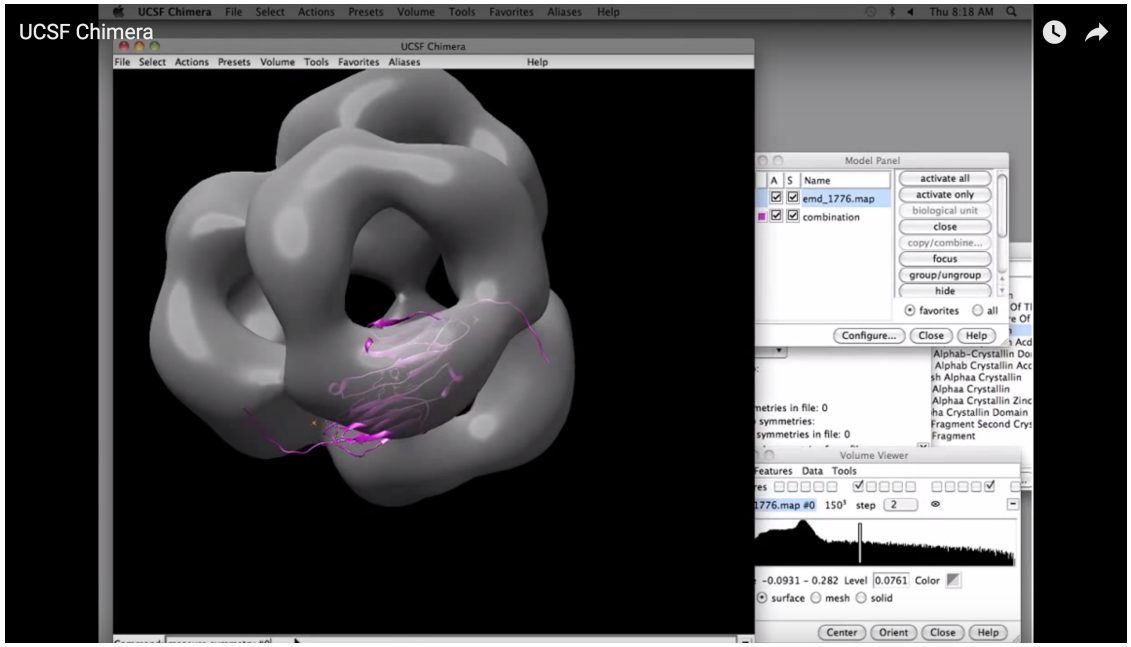
Density fitting



Villa & Lasker, *Curr Opin Struct Biol,* 2014,
Cassidy et al, *Curr Opin Microbiology* 2018

Manual fitting

Fitting an atomic structure within the envelope (an isocontour) of the density using visualisation programs.



Pros:

-Human brain in efficient in certain pattern recognition tasks.

-Immediate feedback and intelligent choices by the user.

-Often good for the initial placement of the component in the map.

Cons:

-High level of subjectivity may lead to error, especially if the map does not have sufficient distinctive features for an unambiguous placement of the component.

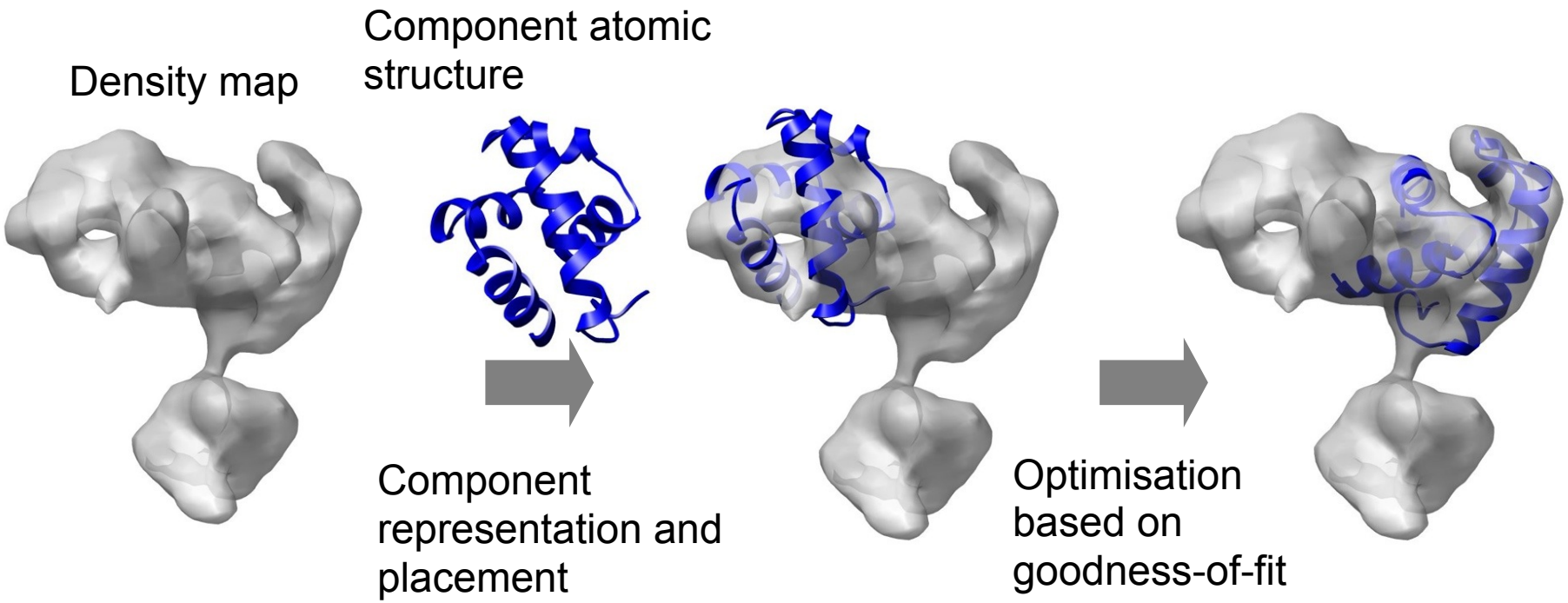
-Depends on contour level.

-Conformational rearrangements cannot be modelled (misfits and steric clashes).

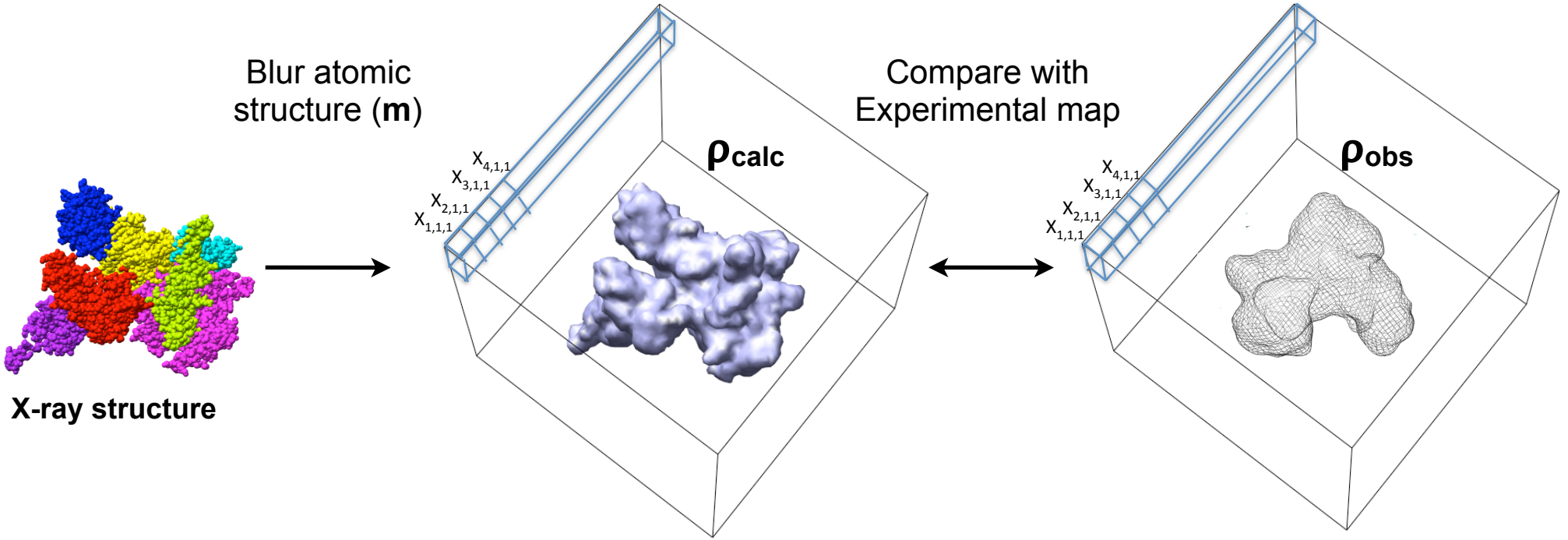
Automated fitting

All automated fitting methods require:

1. a way of representing both the structure and the density map (**representation**).
2. a way of measuring the goodness-of-fit (**scoring**).
3. a method of finding the best fit (**optimisation**).



Representation and scoring



$$R_{\rho,lsq} = \int_x (\rho_{obs}(\mathbf{x}) - \lambda_\rho \rho_{calc}(\mathbf{x},\mathbf{m}))^2 d^3\mathbf{x}$$

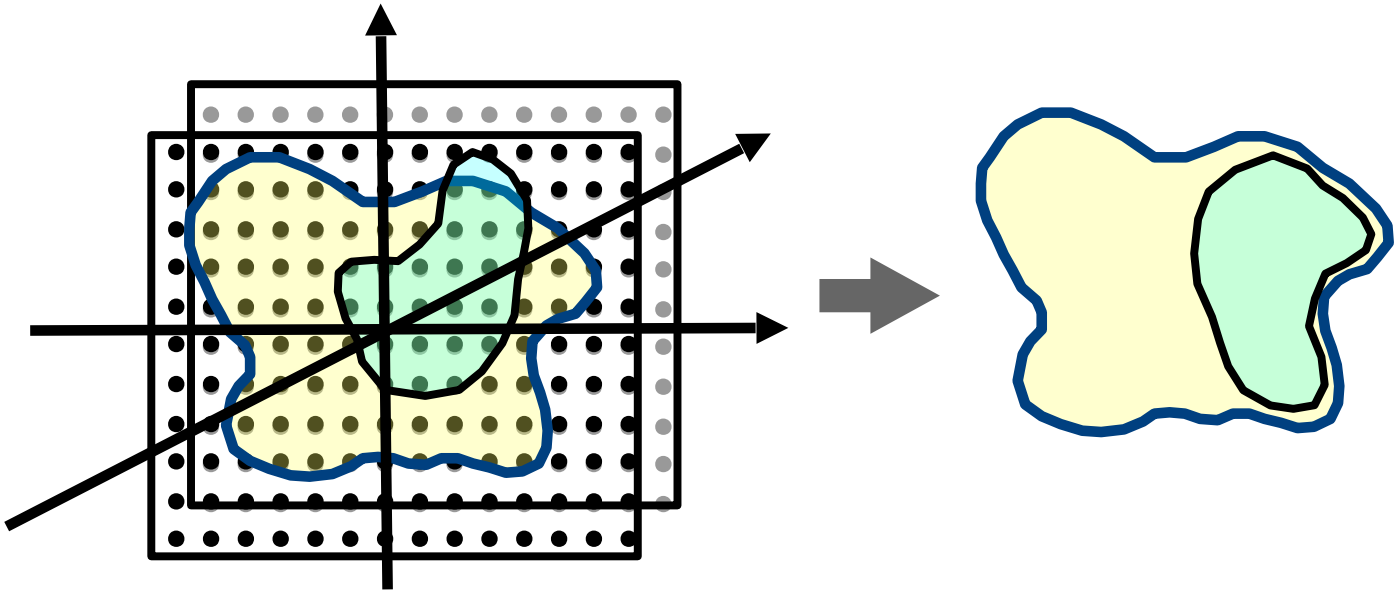
Cross Correlation Coefficient

$$CCC_{\rho,lin} = \int_x (\rho_{obs}(\mathbf{x}) \rho_{calc}(\mathbf{x},\mathbf{m}))^2 d^3\mathbf{x}$$

Exhaustive search

Pros: Get the global solution in respect to a given scoring function.

Cons: The search in real space is too large for most scores (very expensive).



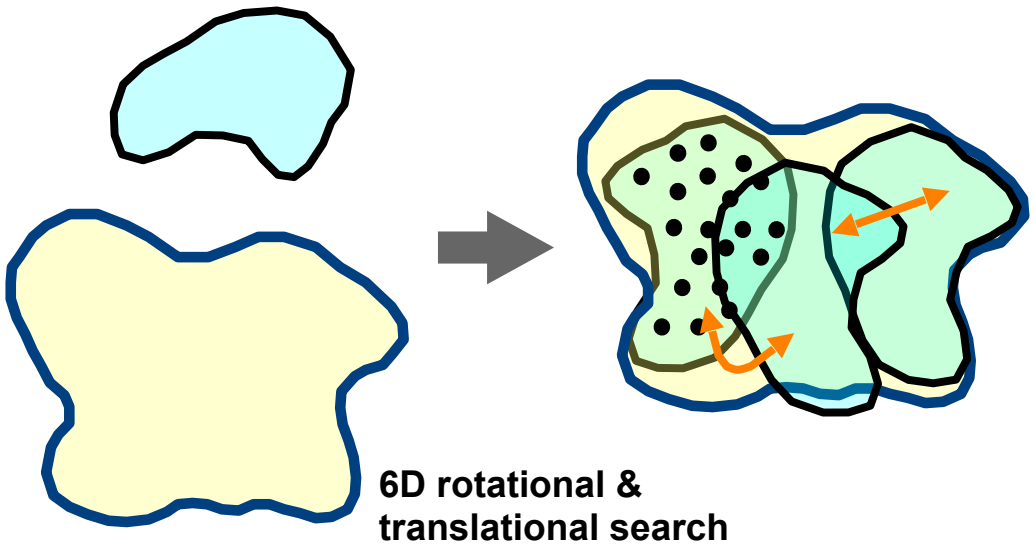
- Acceleration: FFT (translational moves); Spherical harmonics (rotational moves)
COLORES, DOCKEM, ADP-EM, PowerFit, gEMfitter (GPU acceleration)…

- Local fitting - Search exhaustively a given sub-region in the map (Mod-EM, Chimera)

Stochastic/random and gradient methods

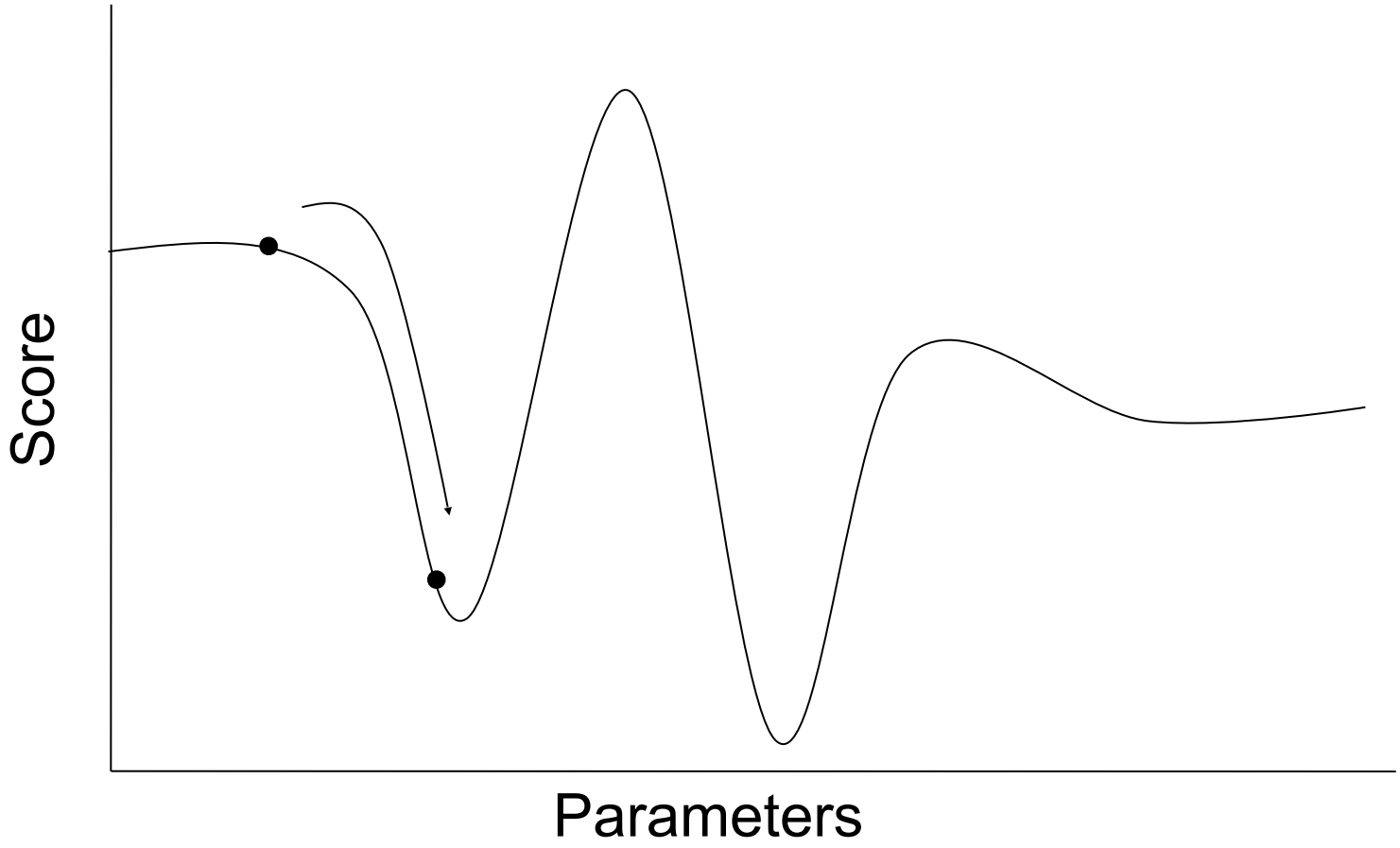
Pros: Fast; easy to implement different scoring functions.

Cons: The model can be “trapped” in a local minimum



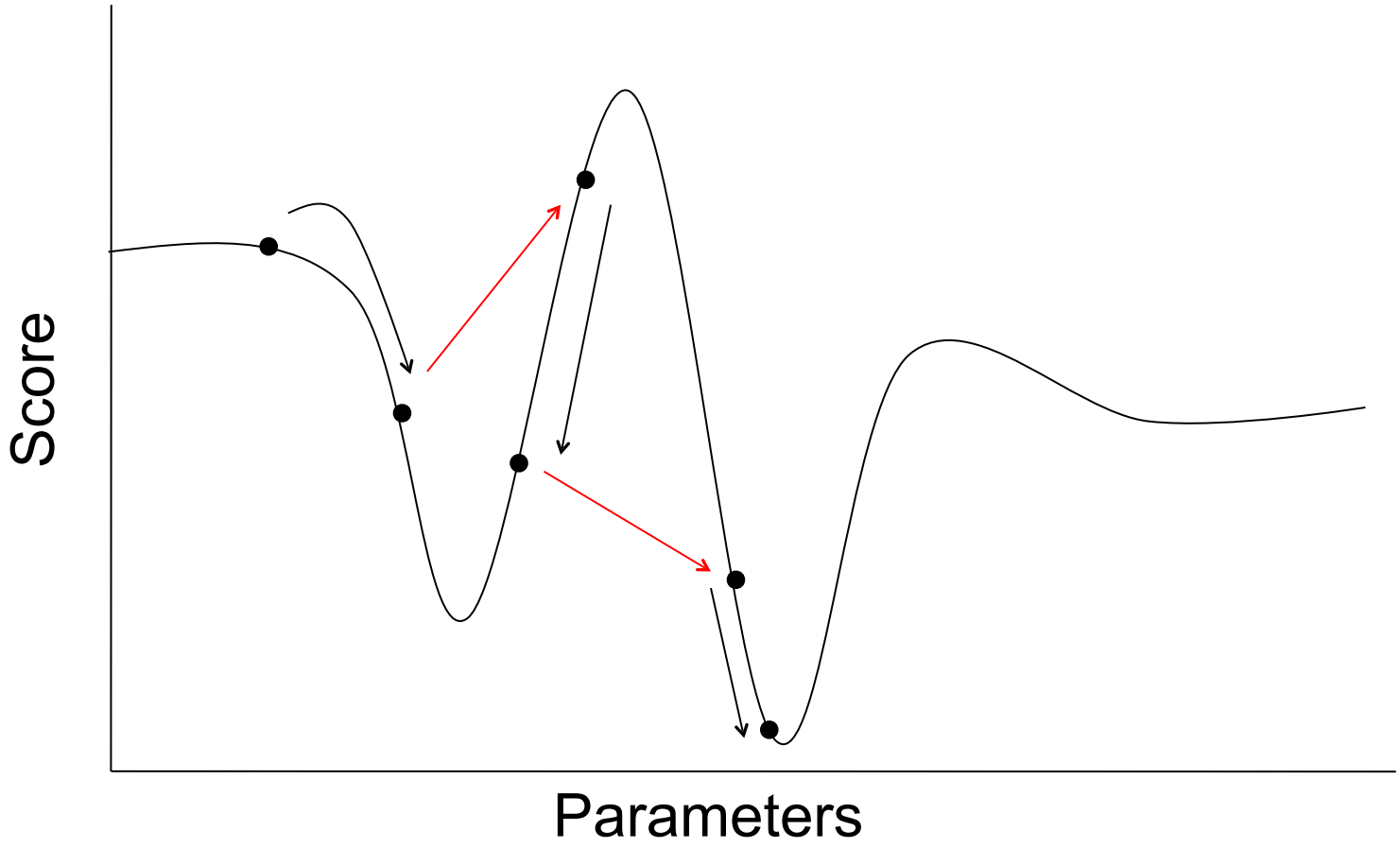
Programs: Mod-EM, Chimera, Rosetta, IMP, HADDOCK, GMfit (gaussian approximation)

Gradient-based methods



Optimisation follows steepest gradient

Stochastic methods



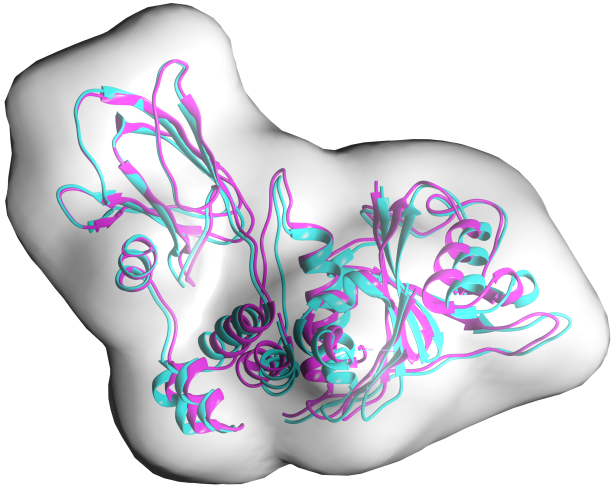
Example: simulated annealing the optimisation follows a gradient method, with ‘jumps’ to avoid local minima

Refinement at intermediate resolution

Before refinement

Cα RMSD from native: 7.5 Å

After refinement



Cα RMSD from native: 2.1 Å

1VCB, 10 Å resolution
native
best predicted fit

http://topf-group.ismb.lon.ac.uk/flex-em/

Problems of density fitting

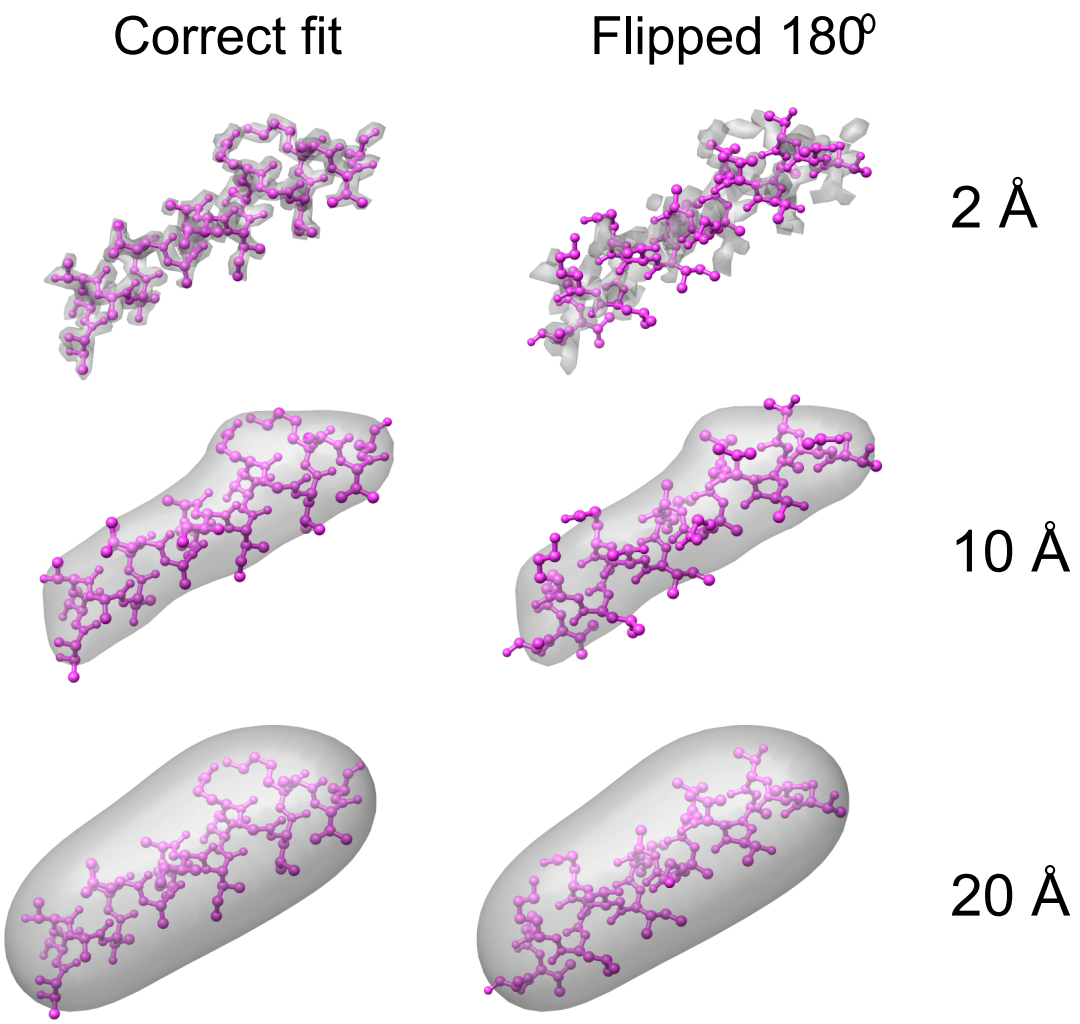
i. Limitations of resolution

Problems:

- At low resolution: many local optima with similar numerical values.
- Local resolution, noise, scaling, filtering, masking.
- Blurring of the atomic structure.

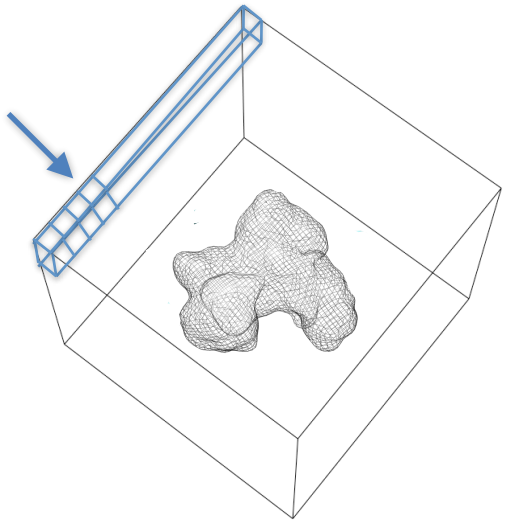
Solutions:

- (Improve your resolution!)
- Improve scoring for goodness-of-fit.
- Coarse-graining (change representation)
- Fit/model validation



Density-based scoring functions

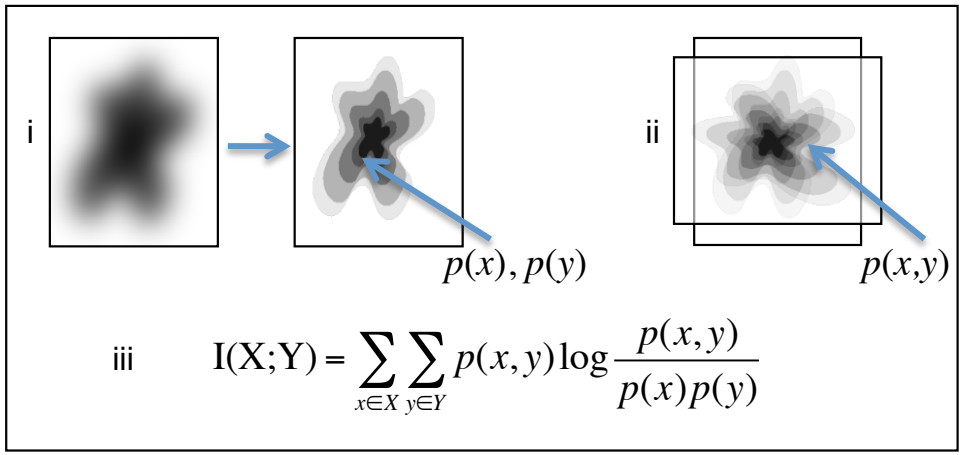
TEMPy: http://tempy.ismb.lon.ac.uk/



- Cross-correlation coefficient (CCC)

$$CCC = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Mutual information-based score (MI)



Useful at intermediate resolutions; noisy maps;
less sensitive to relative intensity levels

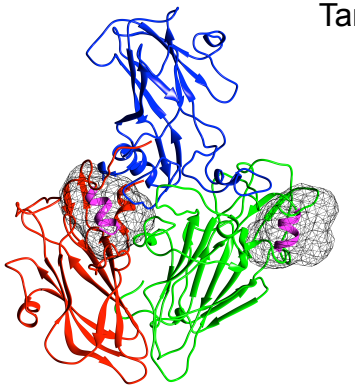
Wriggers & Chacon, *Structure* 2001; Shatsky et al, 2009, Vasishtan & Topf, *J Struct Biol* 2011

Local scoring

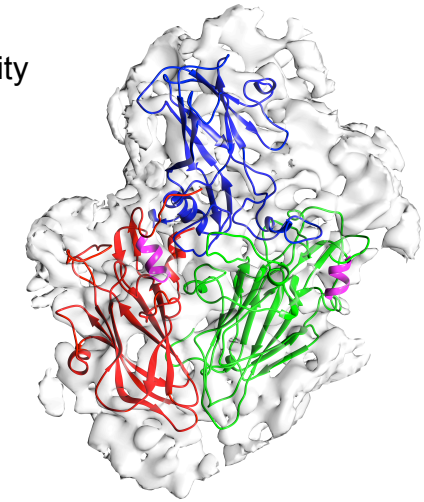
$$SCCC = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

scores local segments in structure

Probe density
X



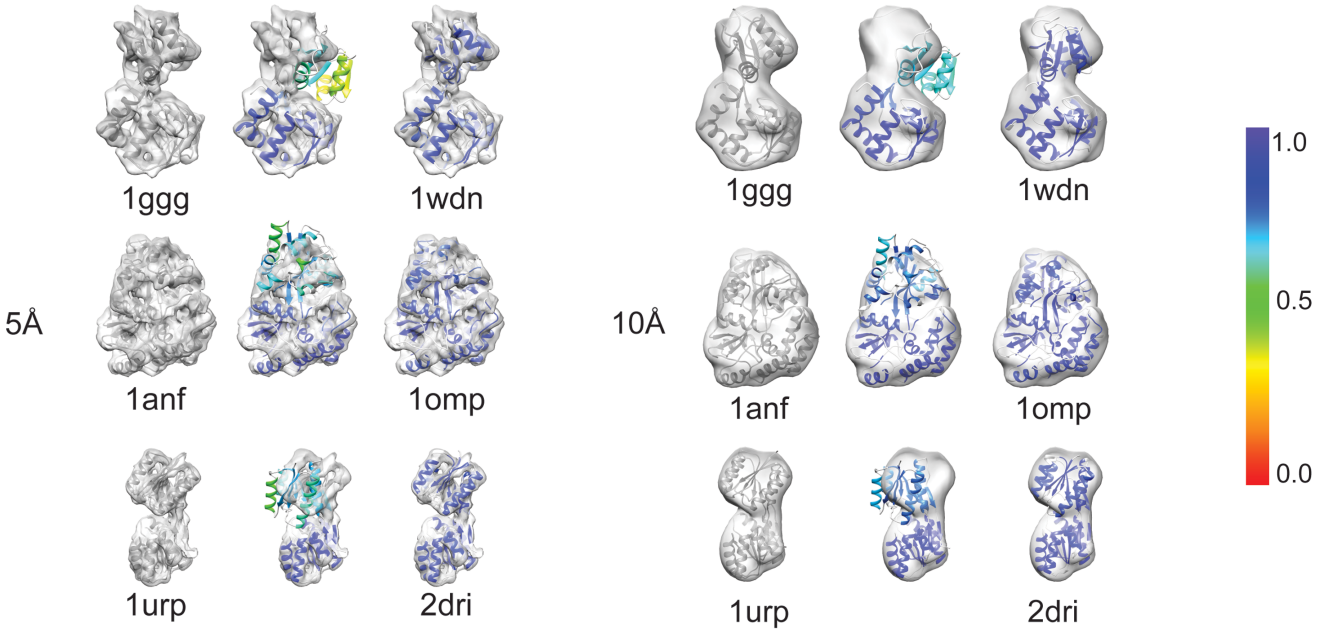
Target density
Y



Roseman, *Acta Crystallogr* D 2000; Pandurangan *et al., J Struct Biol* 2014

TEMPy + Chimera attribute files

Useful for calculating CCC on any defined local segment

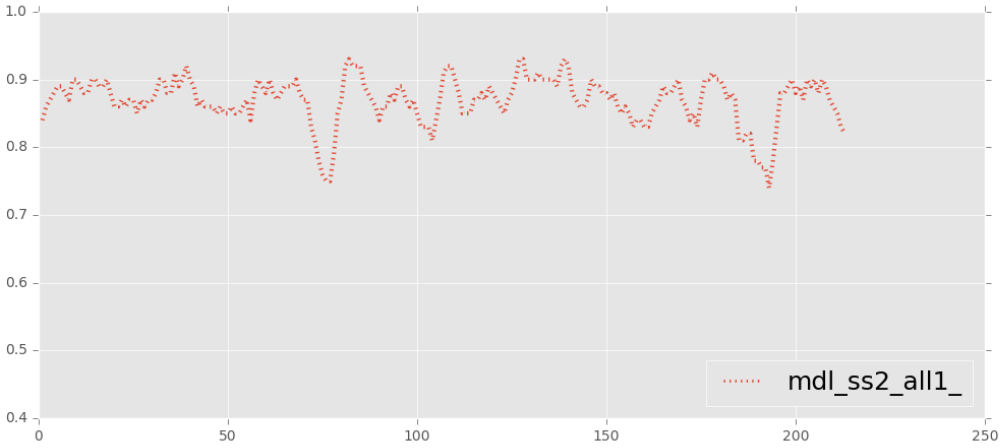


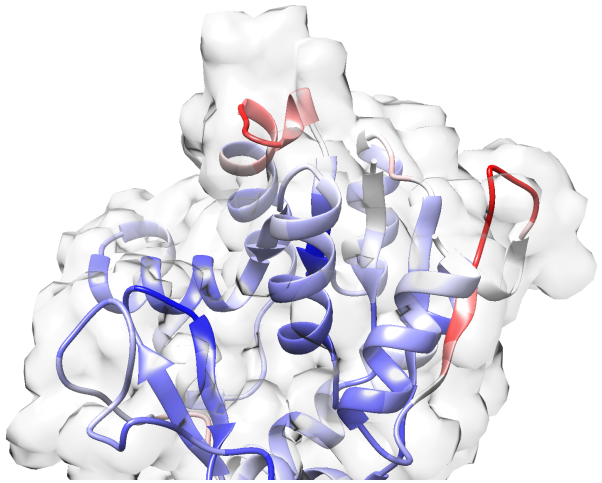
Farabella et al. *J App Cryst* 2015

Local scoring

- Segment-based manders' overlap coefficient (SMOC):

Calculated on overlapping segments along the sequence and assigned to central residue so that each residue has a score.





$$SMOC = \frac{\sum_{i \in vox\_sr} \rho_i^{EM} \times \rho_i^{P}}{\sqrt{\sum_{i \in vox\_sr} (\rho_i^{EM})^2 \times \sum_{i \in vox\_sr} (\rho_i^{P})^2}}$$

Useful to calculate local fit per residue (segment)

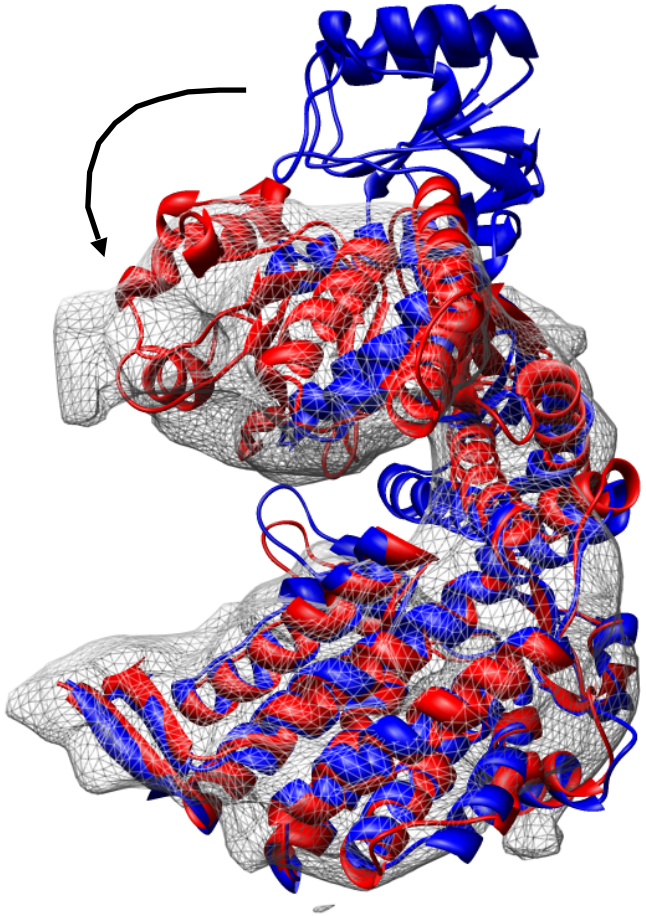
Joseph A.P. et al. *Methods* 2016

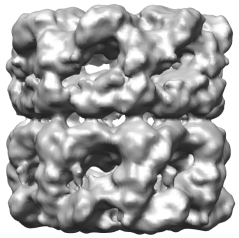
ii. Conformational variability

Problem: Conformations observed by 3D EM often deviate from the conformations of the atomic models we fit.

- Dynamics.
- Crystal packing effects.
- Errors in structure prediction.

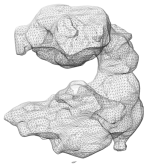
Solution: change the conformation of the atomic model during the fitting process — **flexible fitting.**





3D-EM map

Segmentation



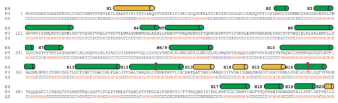
Component structure known?

Yes

No

What resolution?

Component structure



Component Sequence

10-20Å

Fitting all known folds

~4.5-10Å

SSE assignment

<4.5Å

Cα chain tracing, all-atom modelling

Fold assignment from sequence

Template detected?

Yes

No

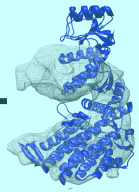
'template-free' modelling

Homology modelling

Rigid fitting

fit different from map?

Yes



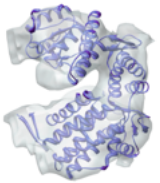
No

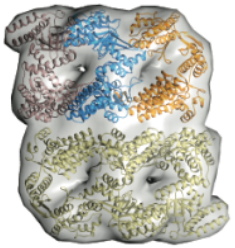
Multiple conformations

ENM / NMA

Real-space refinement

MD-based methods





Model refinement

Without any restraints a model may fit well with a high score in near-atomic-to-low resolution density: “perfectly overfitted model” (e.g. Faulkner et al. 2013)

The resulting model however will not have standard protein geometry:

backbone torsions: phi/psi (Ramachandran space), peptide planarity, chirality (trans/cis), bond lengths and angles, side chain torsions / rotamers

Refinement methods try to maintain standard geometry while fitting the model in density. These geometry restraints reduce the degrees of freedom (sampling space).

Approaches to refinement

- Elastic Network Model (ENM)

- Normal Mode Analysis (NMA): A collection of harmonic oscillators; those with low frequency and large amplitude motions often correlate with experimentally observed conformational changes.

- Geometry-based conformational sampling using harmonic restraints

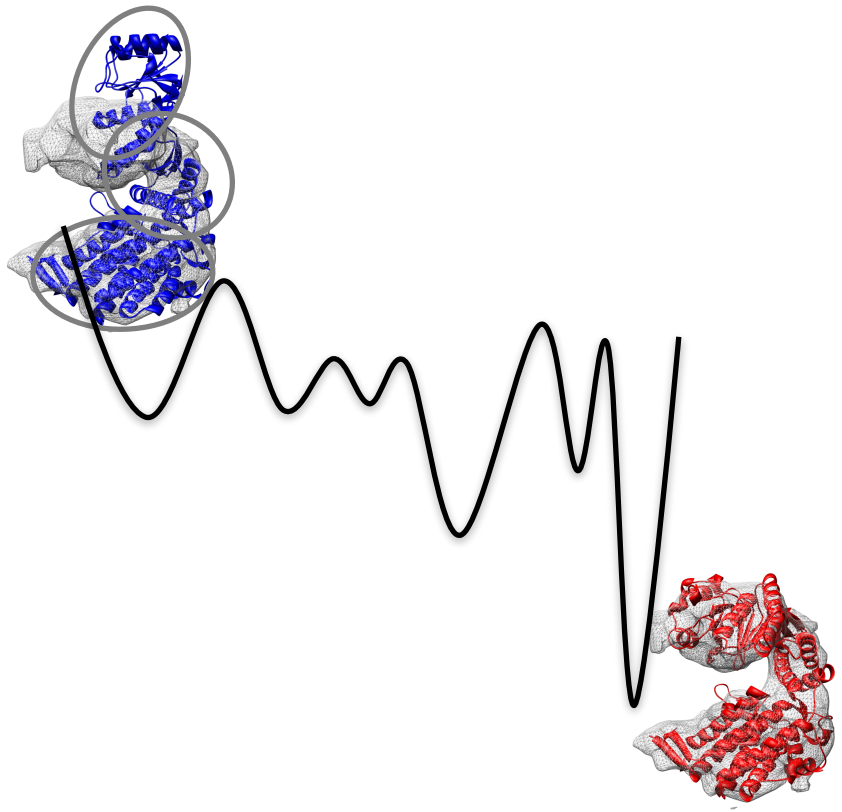
Real-space refinement - Flex-EM

- The fit of the probe structure is optimised simultaneously with the stereo-chemical properties by the minimisation of a scoring function, such as:

$$E = w_1 * E^{CC}(P) + w_2 * E^{SC}(P) + w_3 * E^{NB}(P)$$

- Optimisation is performed on **rigid bodies (b)** by energy minimisation and molecular dynamics.

$$\vec{F}(b_I) = -\sum_{j \in Atom(b_I)} \frac{\partial E(b_I)}{\partial \vec{r}_j}$$



Chen & Chapman, *JSB 2003;* Topf et al., *Structure, 2008;* Joseph *et al., Methods 2016 ;*Trabuco et al. *Structure 2008;*

Rigid-body restraints

A cluster of atoms that form a compact structural segment through a network of contacts can be restrained :

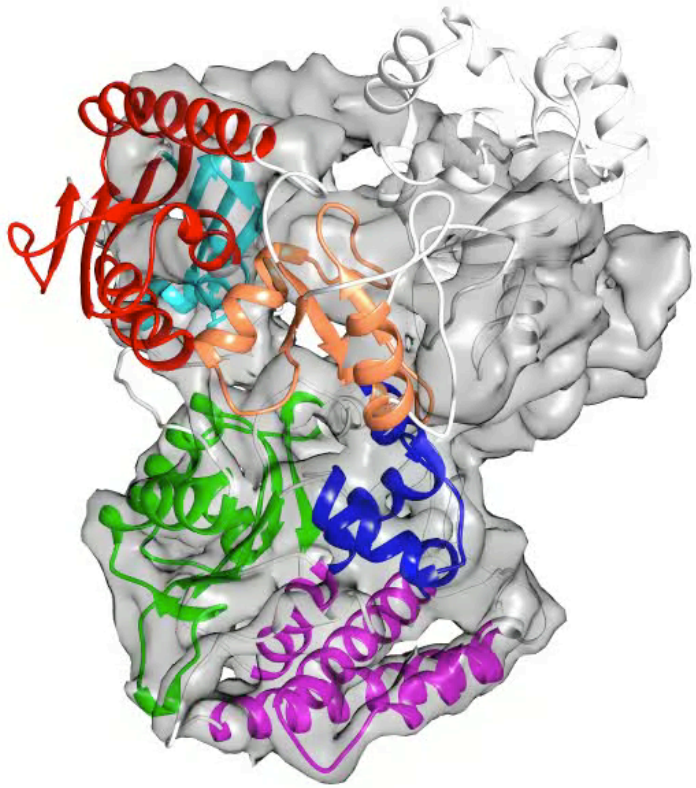
- when the resolution of density map is insufficient to fit smaller entities like individual residues or atoms.

- to allow faster large body movements in the initial stages or refinement

Flex-EM can use **RIBFIND** cluster segments based on secondary structure contacts. Long range distance restraints can be also added using MODELLER

Pandurangan & Topf, *J Struct Biol* 2012;

Refinement at intermediate resolutions



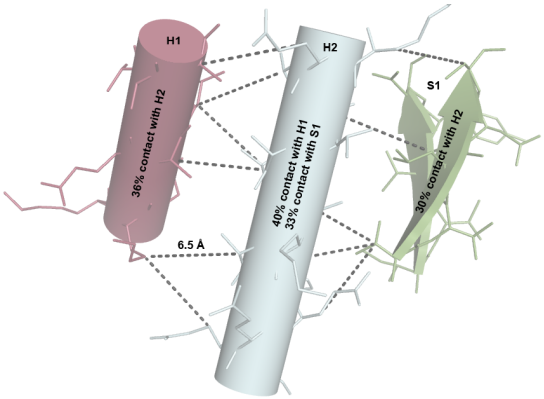
Rigid bodies: sub-domains

5 Å resolution

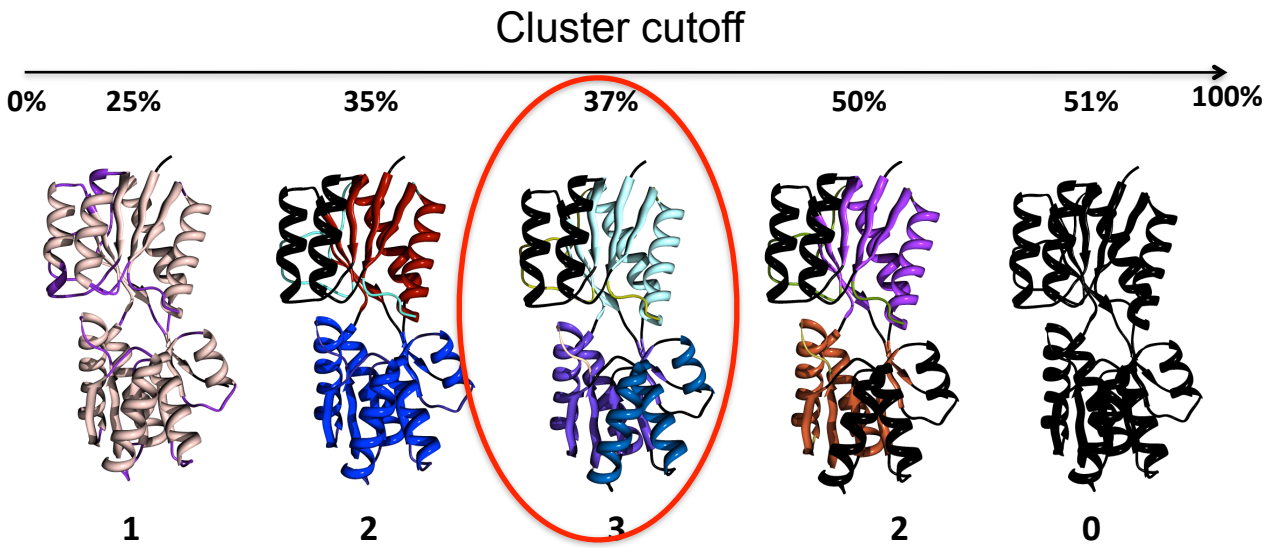
secondary structure elements

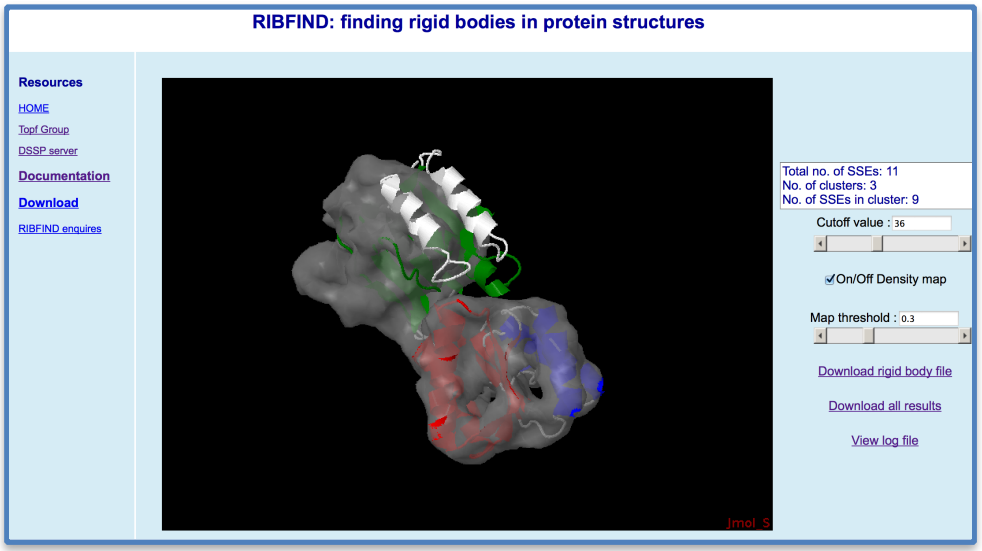
10 Å resolution

RIBFIND: identify sets of rigid bodies



SSE-based clustering

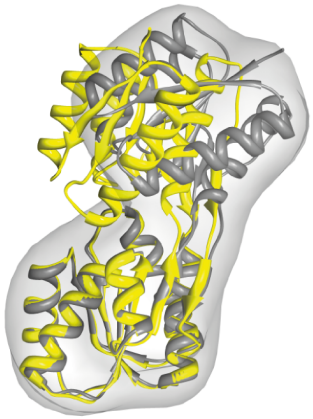
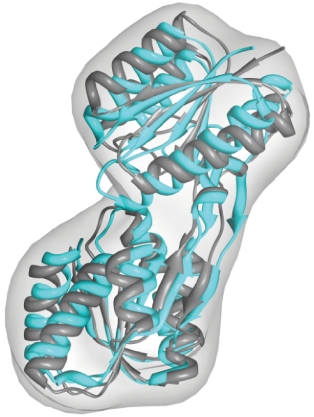
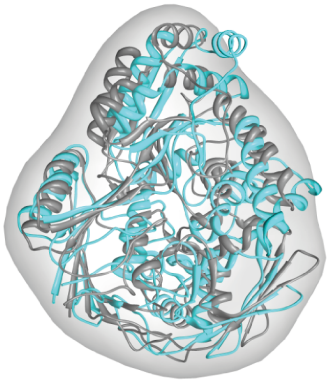
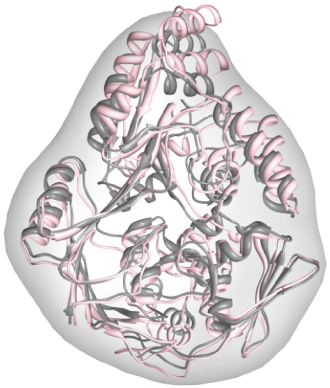




http://ribfind.ismb.lon.ac.uk/

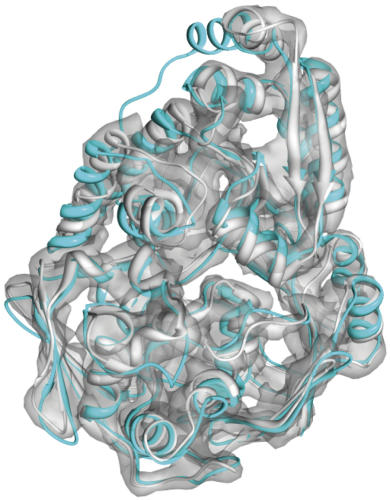
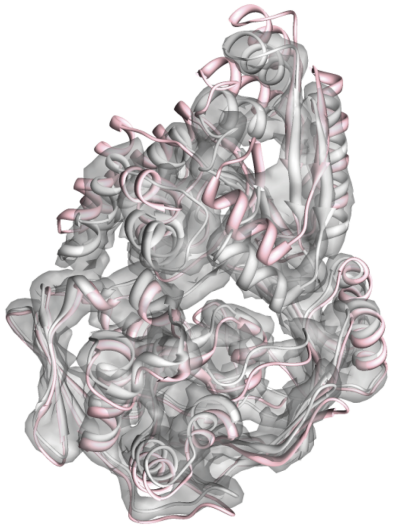
Pandurangan & Topf, *J Struct Biol* 2012, Bioinformatics 2012

Refinement with Flex-EM/RIBFIND

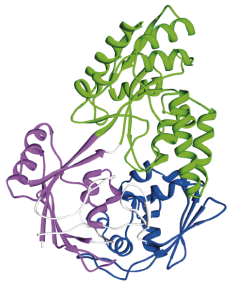
RIBFIND rigid bodies	Initial	Final non-clustered	Final clustered
2driA, 10 Å 			
RMSD from target:	7.55 Å	3.04 Å	1.71 Å
1dpeA, 15 Å 			
RMSD from target:	12.28 Å	7.00 Å	4.17 Å

Hierarchical refinement

1dpe, 5 Å

	Initial	Final un-clustered	Final clustered
			
RMSD from native:	**12.28 Å**	**3.69 Å**	**3.92 Å**
CCC:	**0.814**	**0.891**	**0.867**

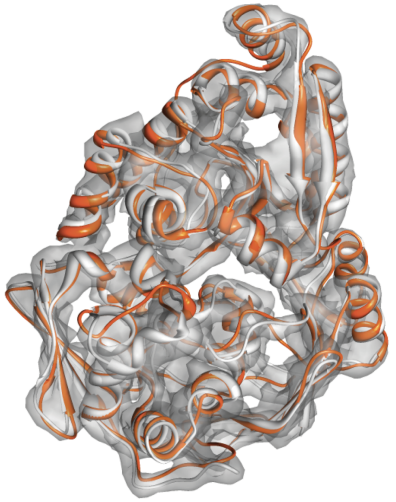
un-cluster



Final two-stage refinement

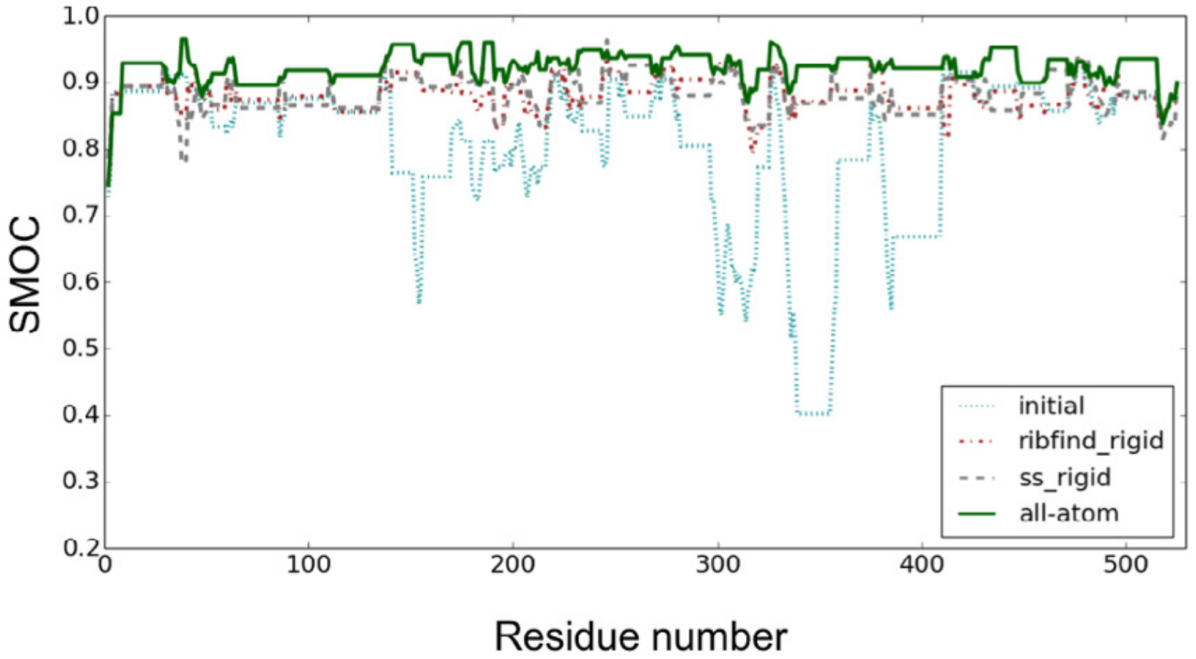
RMSD: 2.12 Å

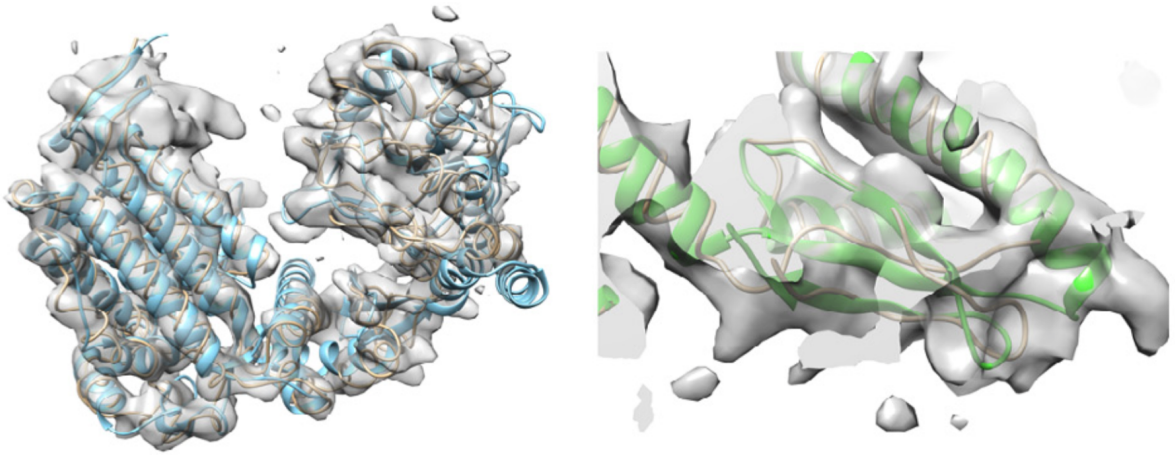
CCC: 0.92



SMOC plots for subnanometer resolution

ADP-bound GroEL (PDB: 4KI8) refined in the density of the unliganded form of GroEL solved at 4.2 Å resolution (EMD-5001).





Joseph A.P. et al. *Methods* 2016

Refinement methods

MDFF: Molecular Dynamics (*Trabuco et al. 2008; Singharoy et al. 2016*)

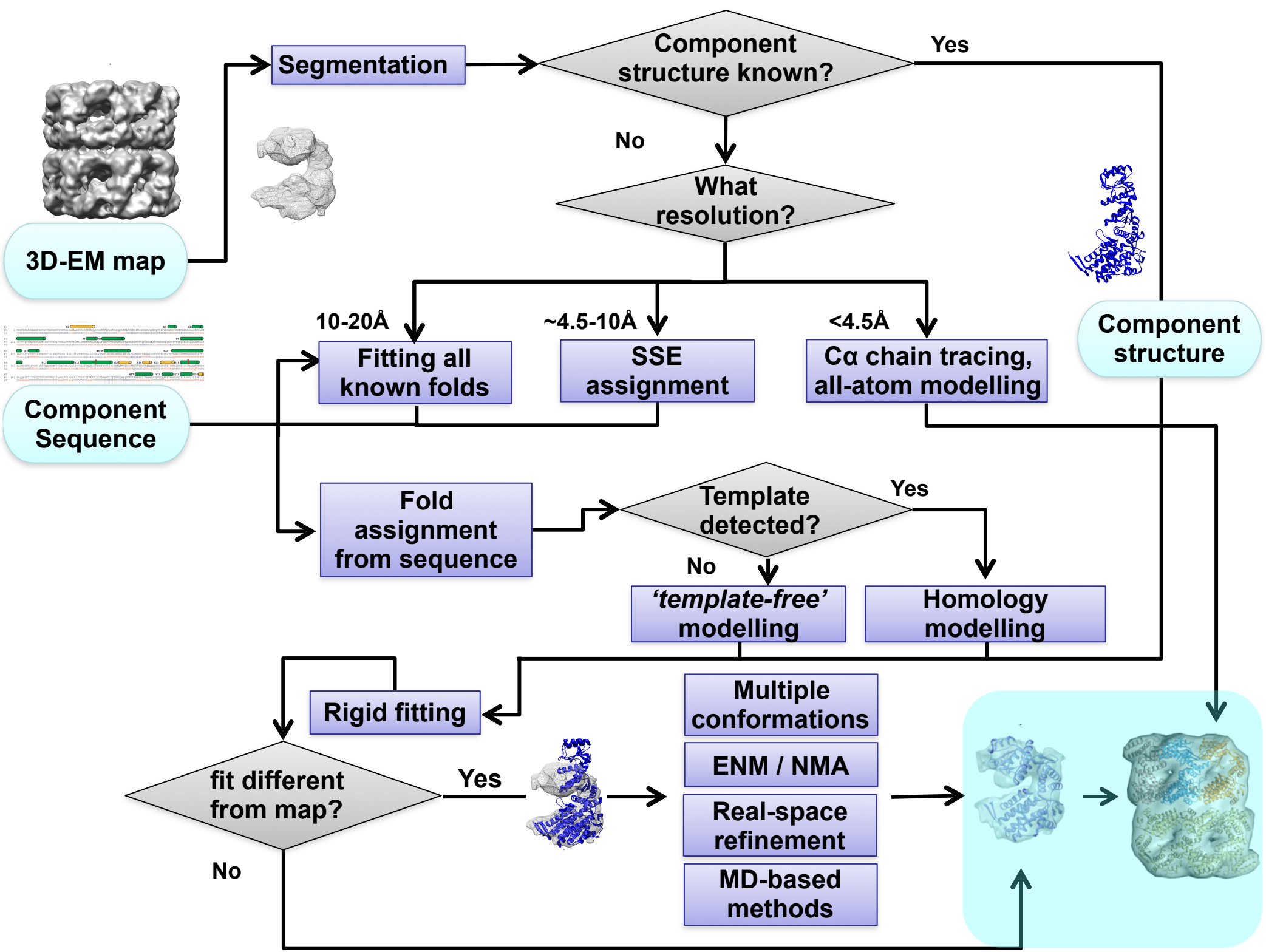
Direx, NMFF, iMODFIT: Normal modes and geometric constraints (*Wang and Schroder 2012; Tama et al. 2004; Blanco and Chacon 2013*)

Rosetta: Monte-Carlo/stochastic (*Wang et al 2016; DiMaio et al. 2015*)

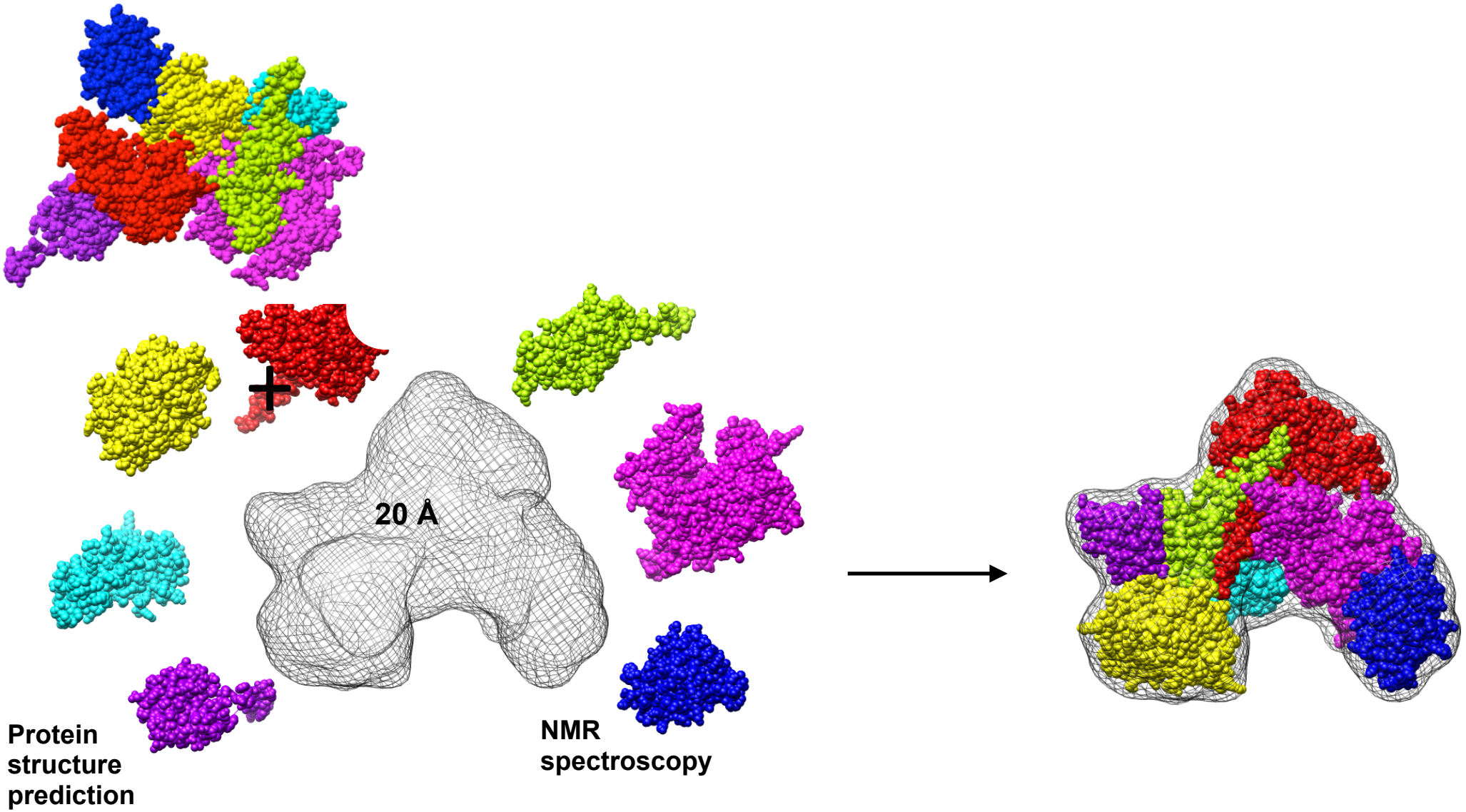
Refmac: Maximum likelihood (*Murshudov 2011; Brown et al. 2015, Nicholls et al. 2018*)

Coot: Interactive/stochastic/exhaustive/gradients (*Emsley et al. 2010; Brown et al. 2015*)

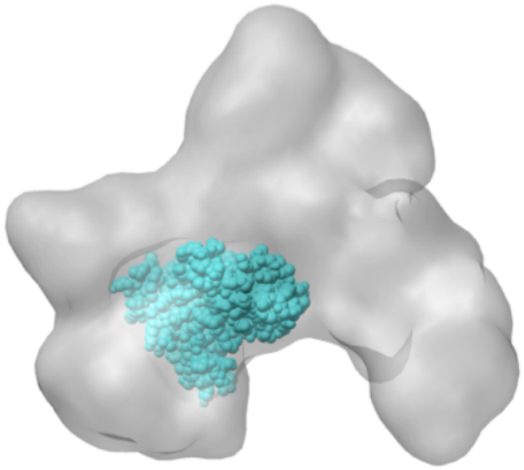
Phenix: Gradient/Simulated annealing MD/exhaustive (*Afonine et al. 2012, 2018*)


3D-EM map
Segmentation
Component structure known?
Yes
No
What resolution?
Component structure
10-20Å
~4.5-10Å
<4.5Å
Component Sequence
Fitting all known folds
SSE assignment
Cα chain tracing, all-atom modelling
Fold assignment from sequence
Template detected?
Yes
No
'template-free' modelling
Homology modelling
Rigid fitting
fit different from map?
Yes
No
Multiple conformations
ENM / NMA
Real-space refinement
MD-based methods

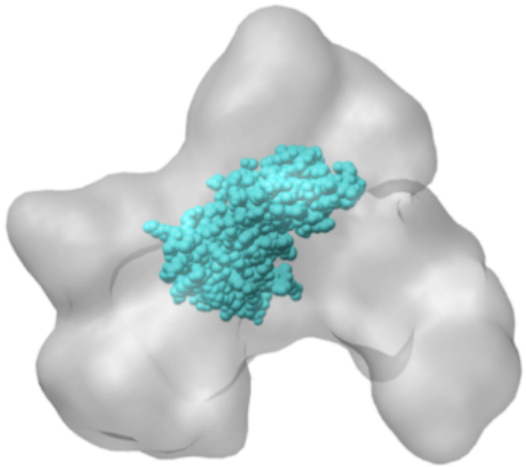
iii. Assembly (multi-component) fitting



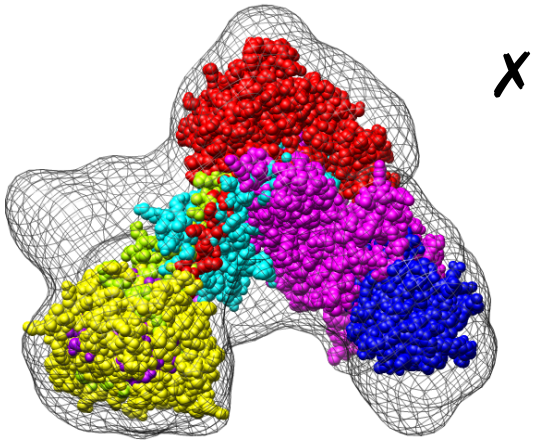
Sequential fitting



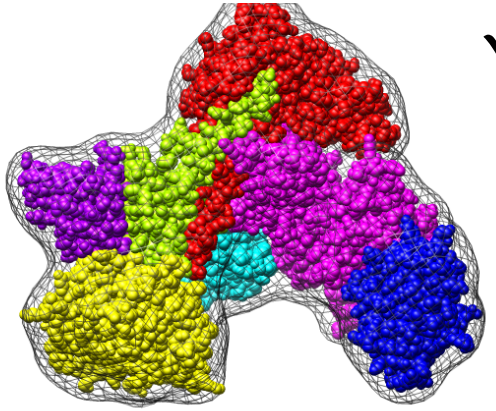
20 Å correct position



20 Å best scoring position



✗



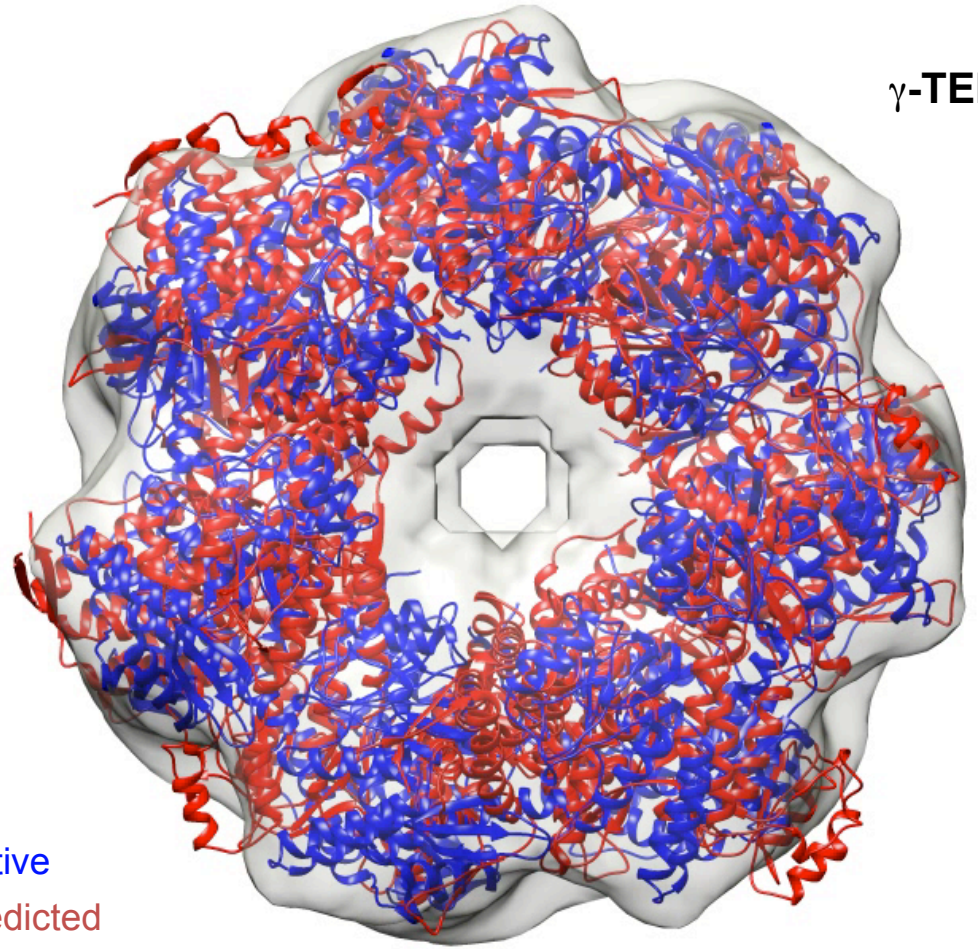
✓

Problem: Components may migrate toward the centre of the map or to a different local maxima.

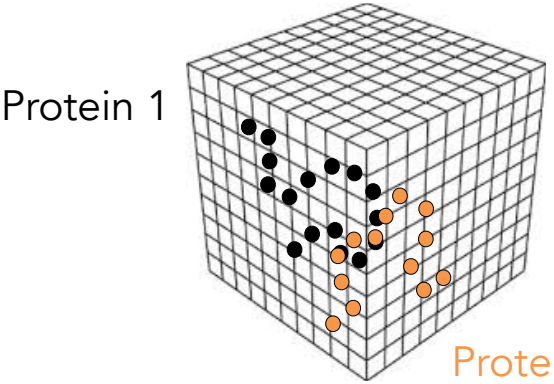
Solution:

- Core-weighted cross correlation (Wu et al, Zundert & Bonvin 2015)
- Simultaneous (assembly) fitting
- Multiple scores (additional constraints - integrative modelling).

Multi-component (assembly) fitting



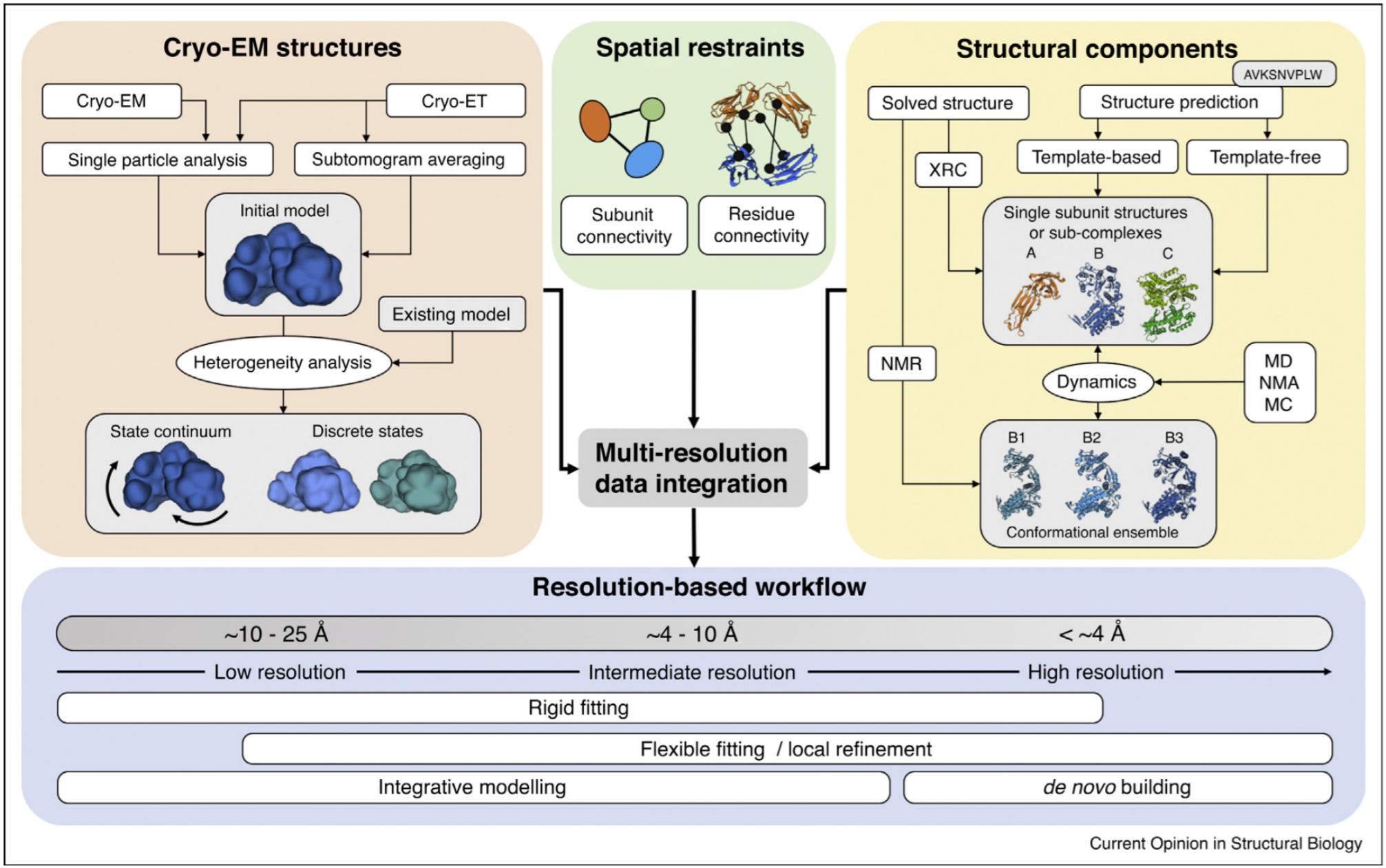
1GRU, EMD-1046 (23.5Å)



$$PS = \sum_{i=1}^{M} (Vol_{Overlap} / (Vol_{Component1} + Vol_{Component2}))$$

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Determining structures of macromolecular complexes using cryo-EM

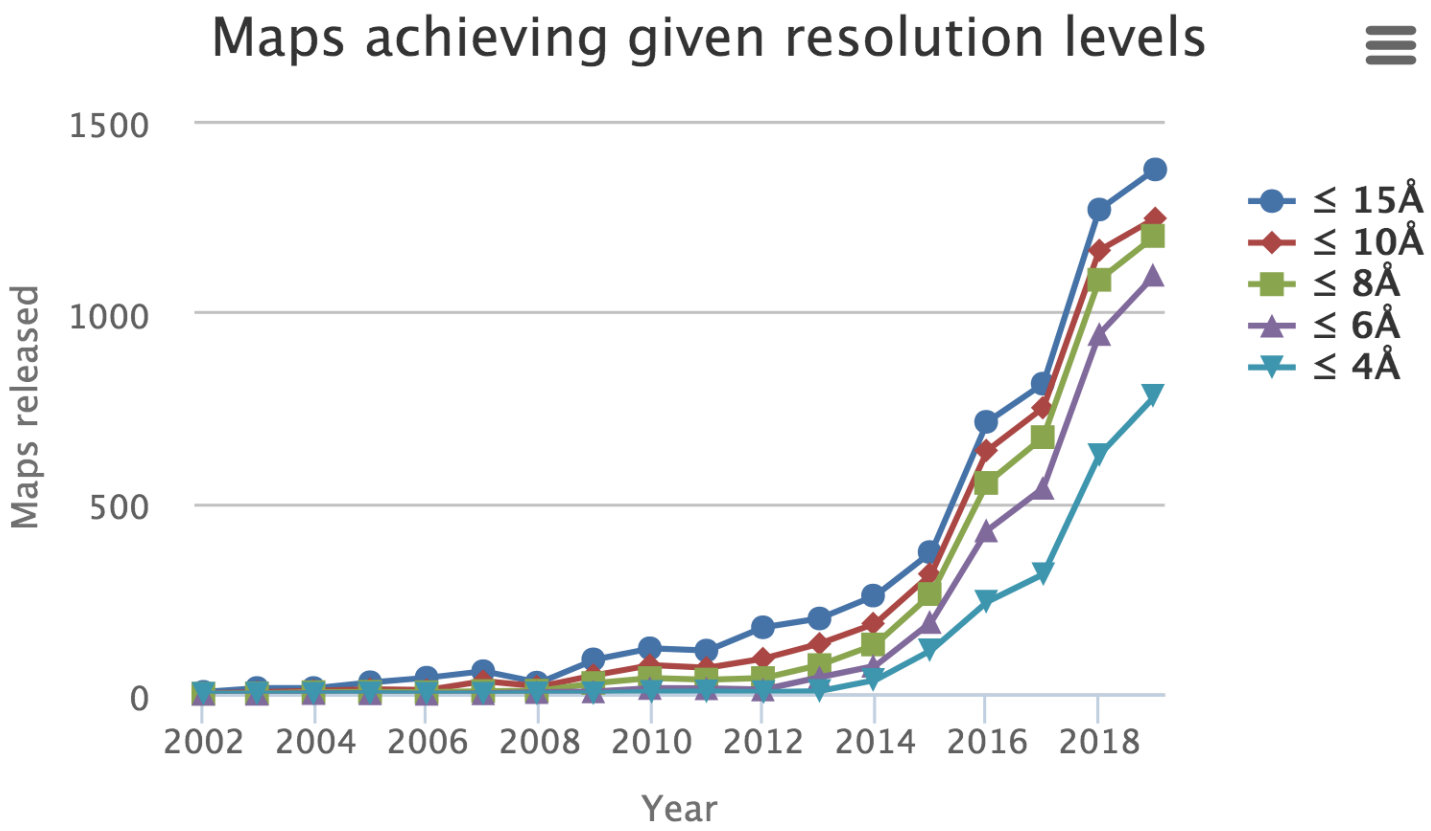


Malhotra *et al.*, 2019

Fit / Model assessment and Validation

Model Validation

~9000 maps in EMDB.
~3683 fits in PDB.



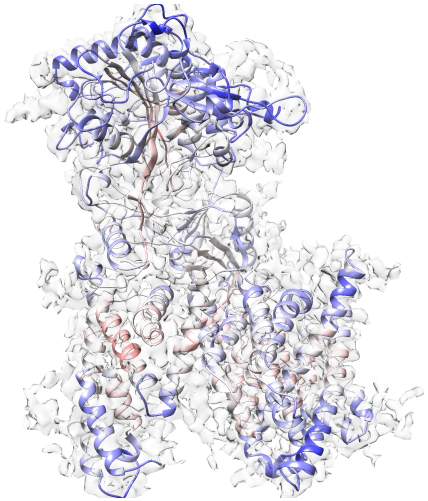
Models are fitted across resolutions but so far there are no systematic assessment and validation pipeline across resolutions. As the structural and biological interpretation is based on the atomic models, the importance of validation is increasingly being realised.

Henderson et al. *Structure* 2012.

Model fit

Model geometry

peptide planarity
backbone torsions (Ramachandran)
bond lengths
bond angles
side chain rotamers



Molprobity: http://molprobity.biochem.duke.edu/
What check: http://swift.cmbi.ru.nl/gv/whatcheck/
PROCHECK: http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

Model Validation

Many tools available for assessment and validation, but there is no systematic pipeline

Goodness of fit

TEMPy
Coot/Refmac
Phenix
EMringer
…

Model geometry

Molprobity
Coot
What-check
…

Secondary structure

MolProbity, Coot, Qmean…
Psipred, …

Tertiary structure

Verify-3D, ProQ3, Prosa,
DOPE (MODELLER), ModFold, ..

Validation

Cross-validation:
Half map (Refmac, Rosetta)
Ensemble assessment with multiple scores (TEMPy)
Resolution shells (Direx)

Experimental validation

mutations, cross-links, …