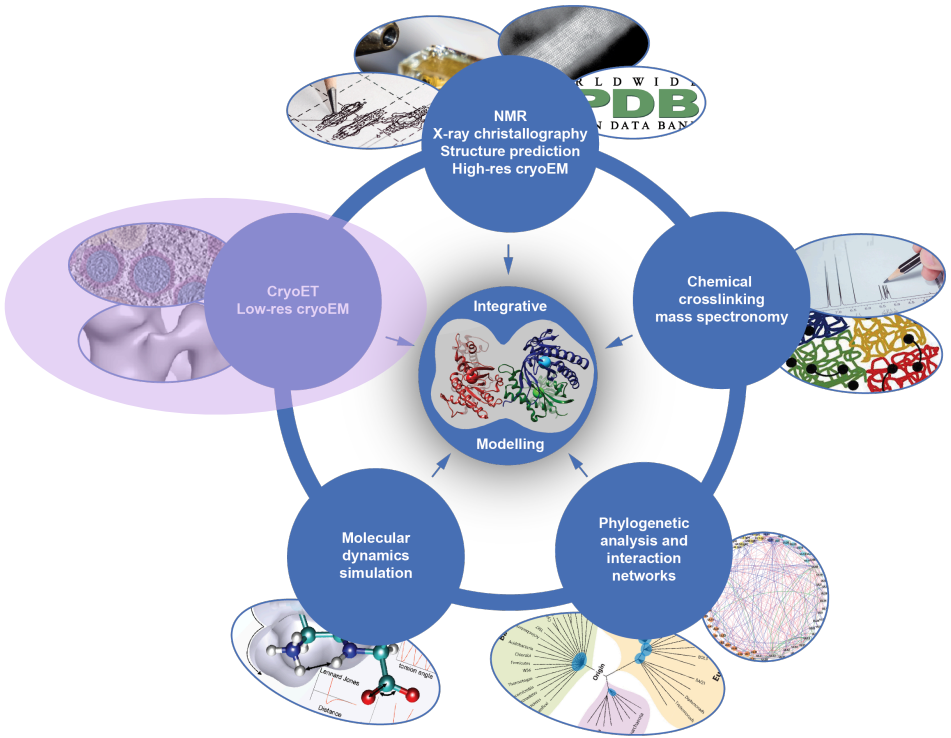
Flex-EM & TEMPy-ReFF

CCP-EM Icknield Model Building Workshop

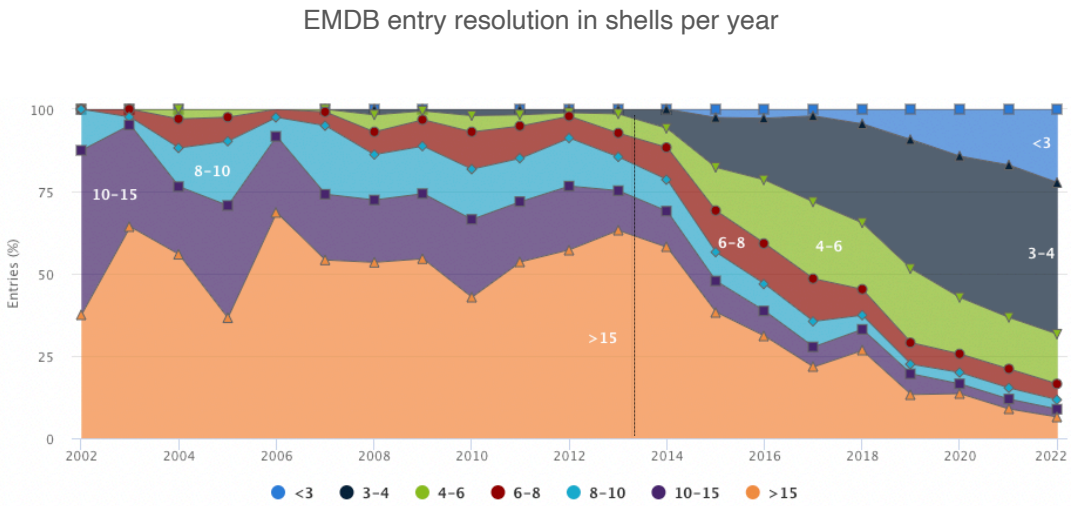
11.10.2023

Maya Topf (CSSB Hamburg)

COMPUTATIONAL METHODS FOR INTEGRATIVE STRUCTURAL BIOLOGY



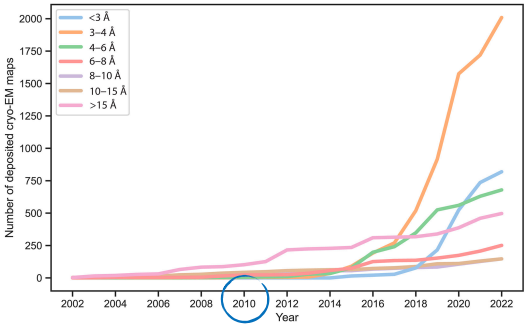
- **Develop** methods for integrative modelling of macromolecular assemblies
- **Apply** the methods to model structures of viral assemblies (e.g. herpesviruses) and other systems (e.g bacterial pore-forming proteins)
- **Analyse** the structures to further understand their function and identify potential drug targets
- **Distribute** and support new software for the structural biology community

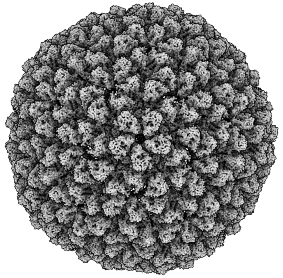


"The Resolution Revolution"
W Kühlbrandt, *Science* 2014

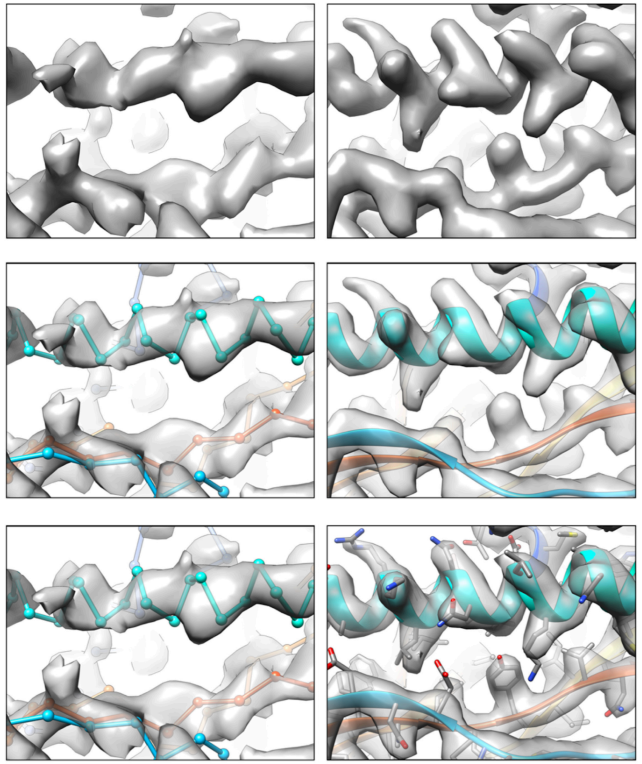
- More than 50% of the maps are now at better than ~5 Å resolution

THE "RESOLUTION REVOLUTION"





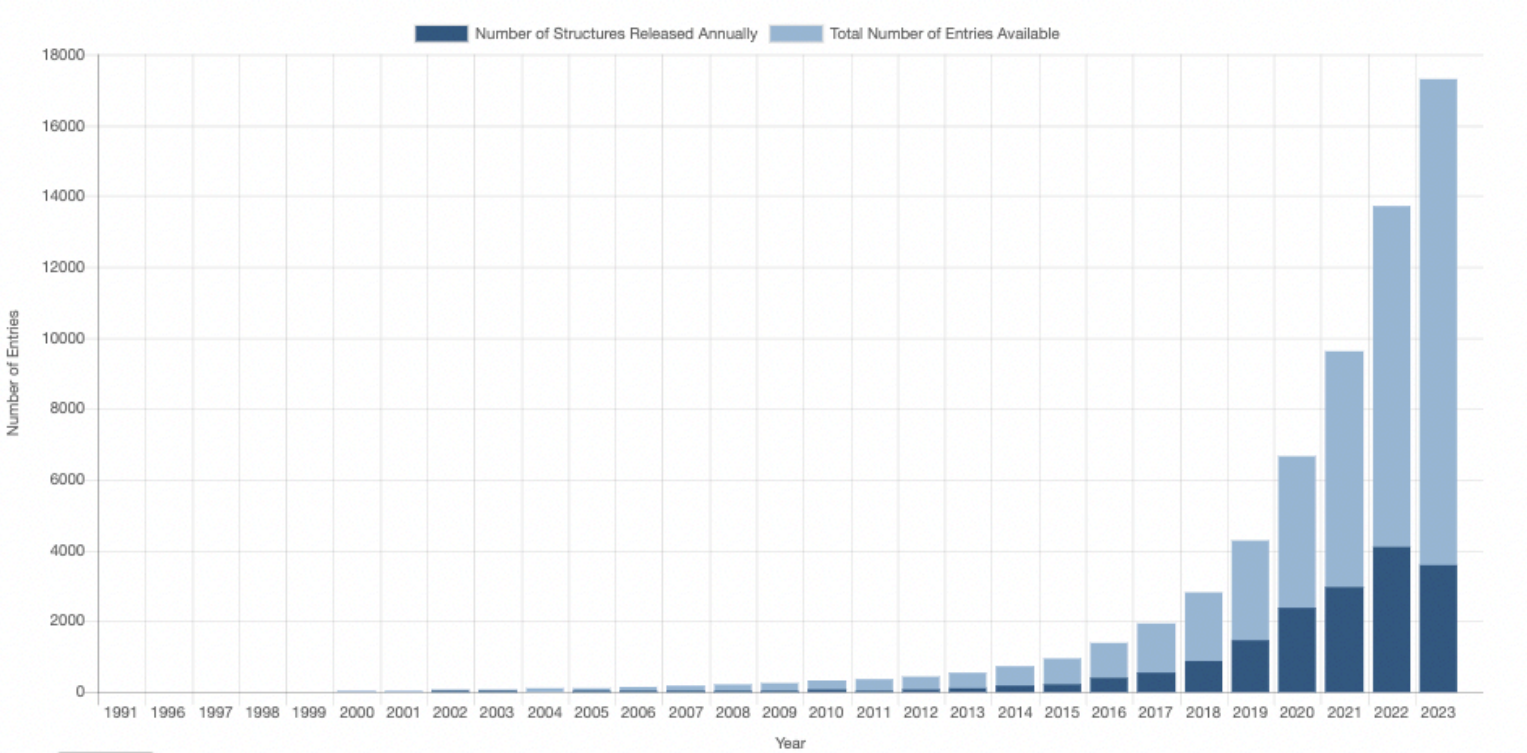
Bacteriophage P22
4Å (actually ~5Å)
EMDB 5137
Chen et al., 2011



Chen, 2011. *PNAS*
Non-validated C-alpha trace

Hryc, 2017. *PNAS*
***de novo* all-atom model**

PROTEIN DATA BANK STATISTICS

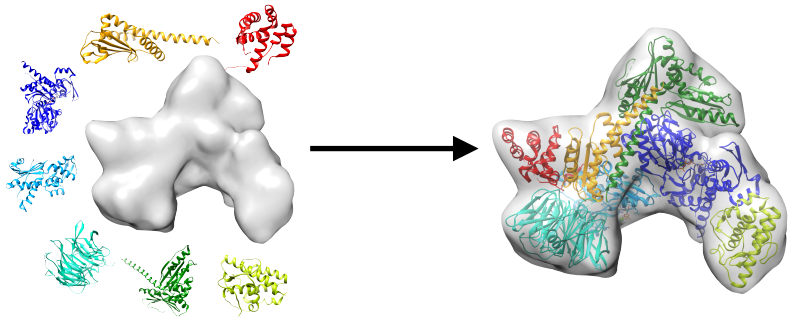


- 7-8% of the structures in PDB are associated with maps in EMDB
- These days most entries are associated with either a single model or multiple models

https://www.rcsb.org/stats/growth/growth-em

What is it?

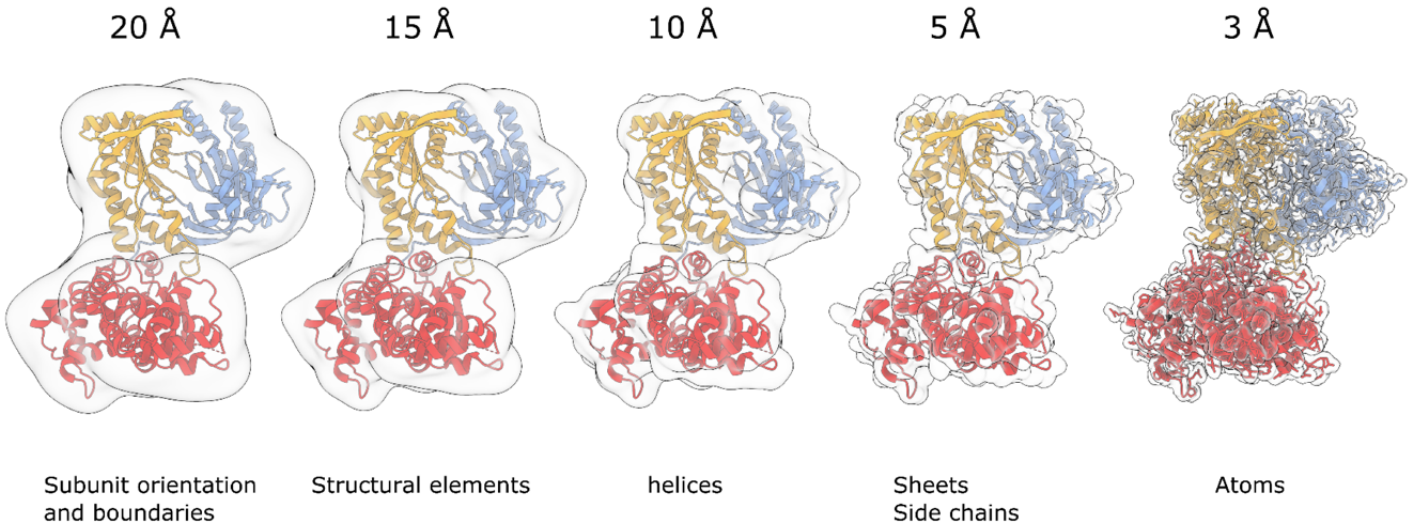
* Generation, representation, “manipulation” of the 3D structure of biological molecules
* A molecular model in cryoEM is a compact interpretation of a density map in light of everything known *a priori* about the structure-composition of the macromolecule of interest



Why we need it?

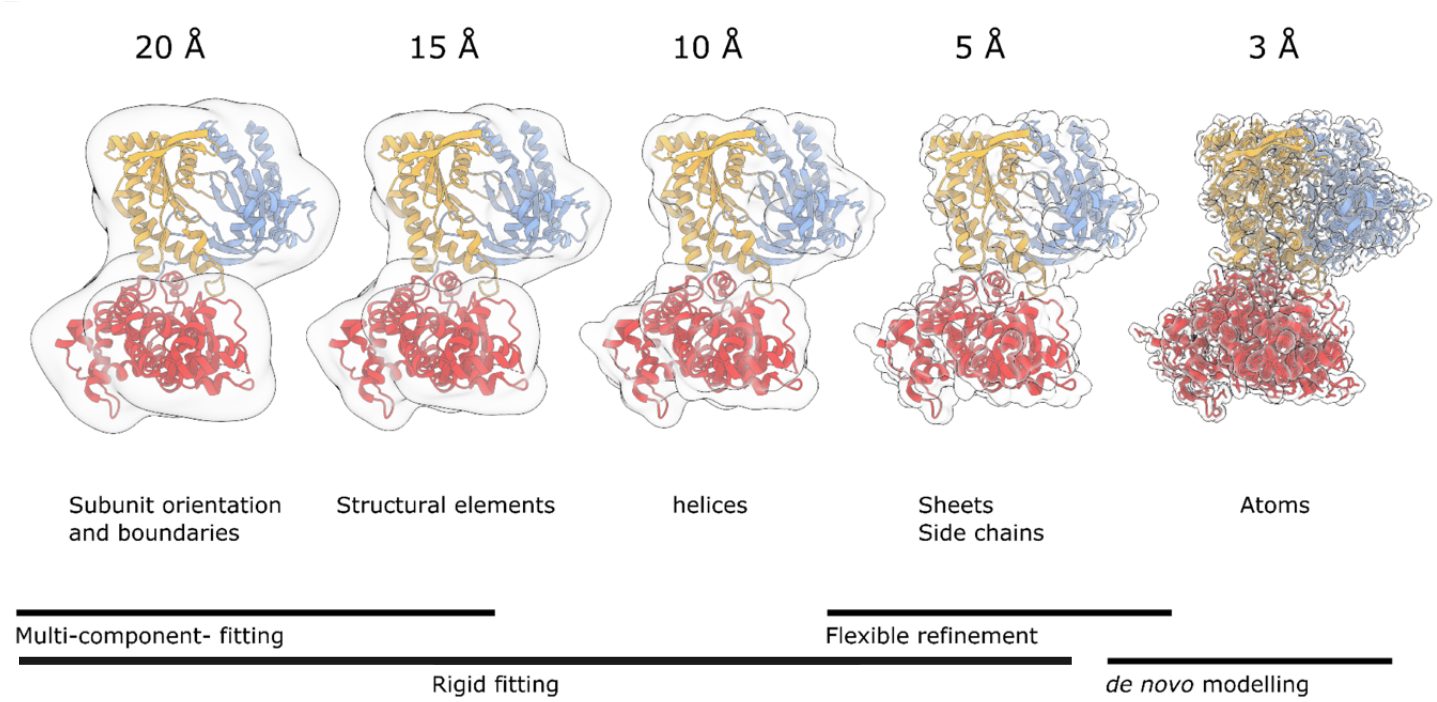
1. To get the atomic 3D structure of the molecules;
2. to know physicochemical characteristics of the molecules;
3. to compare the structure of a molecule with different molecules;
4. to visualize complexes formed between different molecules/macromolecules;
5. and to predict how new related molecules might look.

STRUCTURAL FEATURES AT DIFFERENT RESOLUTION LEVELS

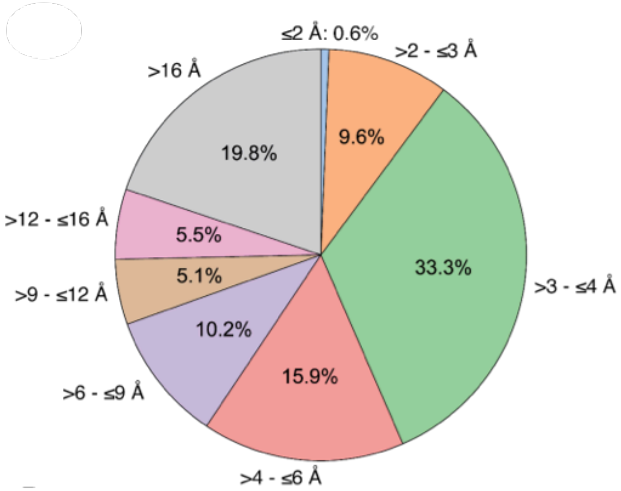


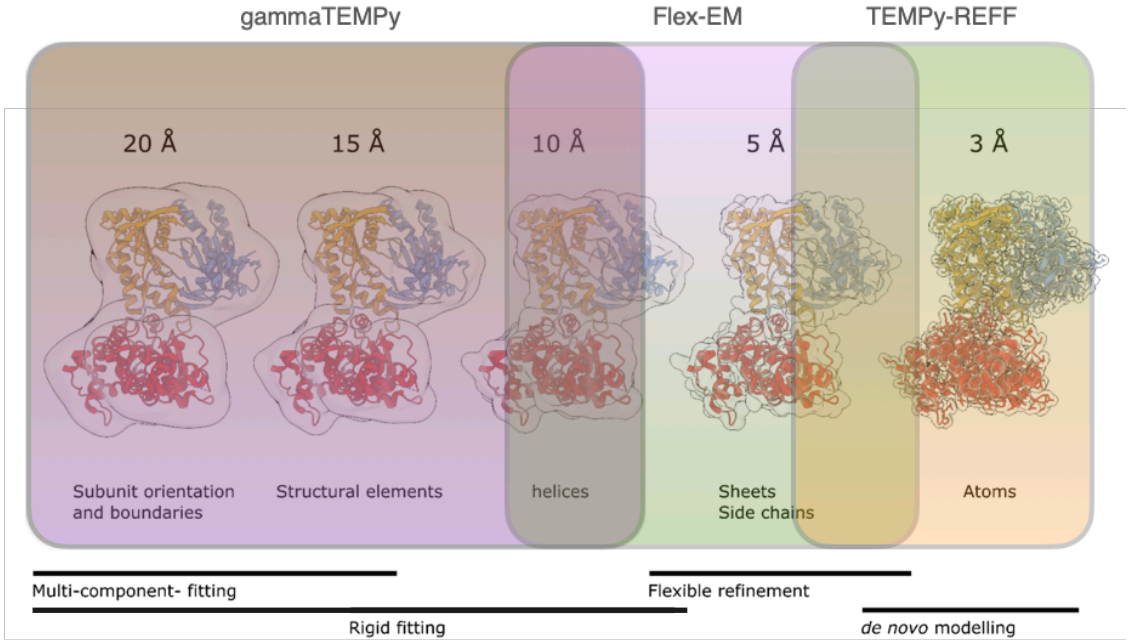
if a complete crystal structure is already available, 10Å data may be sufficient, while if no sequence/composition data is available even 3Å may not suffice

STRUCTURAL FEATURES AT DIFFERENT RESOLUTION LEVELS



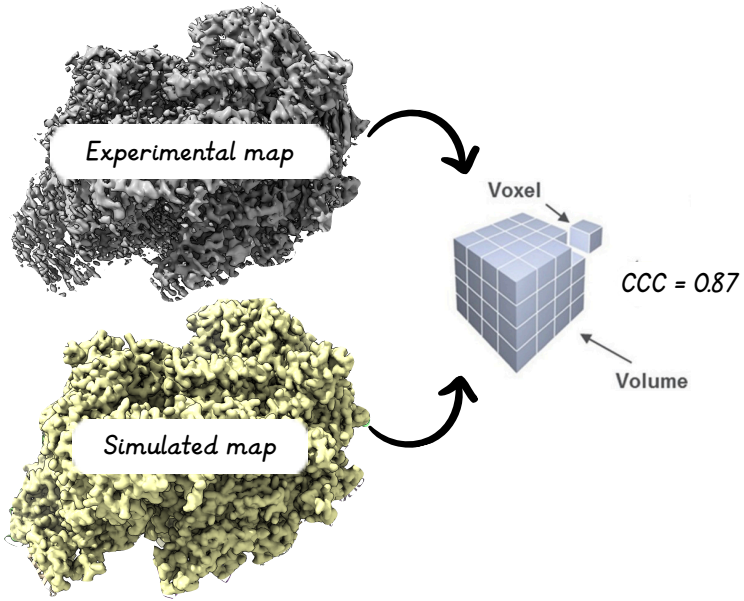
FITTING TOOLS ACROSS THE RESOLUTION SPECTRUM





* Impossible to mention all the modelling softwares developed till now…
  We just want to tell you about some of our tools.

Topf et al. *Structure*, 2008
Pandurangan et al. Structure 2015
Cargnolini et al. *Proteins* 2021
Beton et al. WIREs Comput Mol Sci. 2023

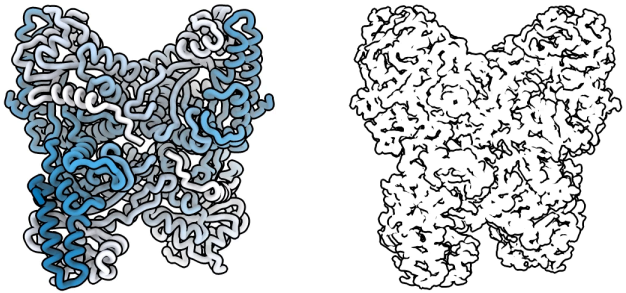


$$\rho(x,y,z) = \sum_N \frac{Z_N}{\left(\sigma\sqrt{2\pi}\right)^3} e^{-\frac{(x-x_n)^2+(y-y_n)^2+(z-z_n)^2}{2\sigma^2}}$$

The closer the cross-correlation value is to 1, the more closely the sets are to being identical.

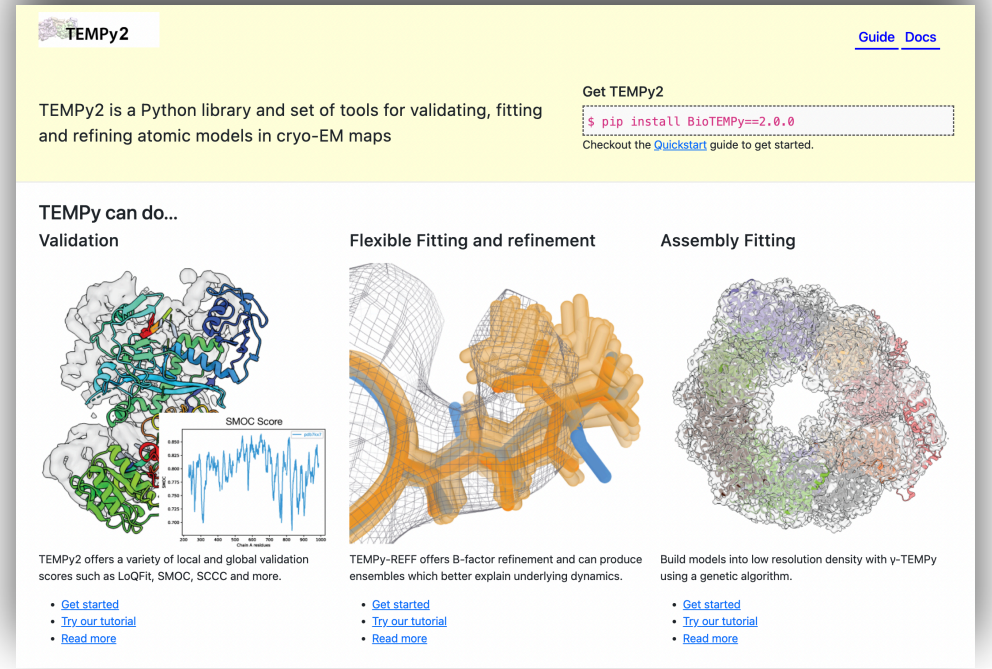
Cross Correlation Coefficient $\mathbf{CCC}_{\rho,\mathrm{lin}} = \int_{\mathbf{x}} (\rho_{\mathrm{obs}}(\mathbf{x})\rho_{\mathrm{calc}}(\mathbf{x},\mathbf{m}))^2 d^3\mathbf{x}$

* When possible, a known or pre-calculated model is placed and fitted in the cryo-EM map as a rigid body.
* Often a full-exhaustive six dimensional (6D) grid search of the three translational and three rotational degrees of freedom is performed.
* Analysis of all possible solutions is performed to locate the global cross-correlation minimum.



Tools > Volume Data > FitInMap
fit #model #map search 10

SOFTWARE



Map processing
- Transformation (rotation/translation)
- Filters

Model processing
- Transformation (rotation/translation)
- Model-to-map
- Ensemble generation
- Clustering

Fitting
- Local random and exhaustive search
- Multicomponent fitting (γ-TEMPy)
- Refinement (TEMPy-REFF)

Global Scoring
- Density based
- Surface based

Local scoring
- SCCC
- SMOC
- LoQfit

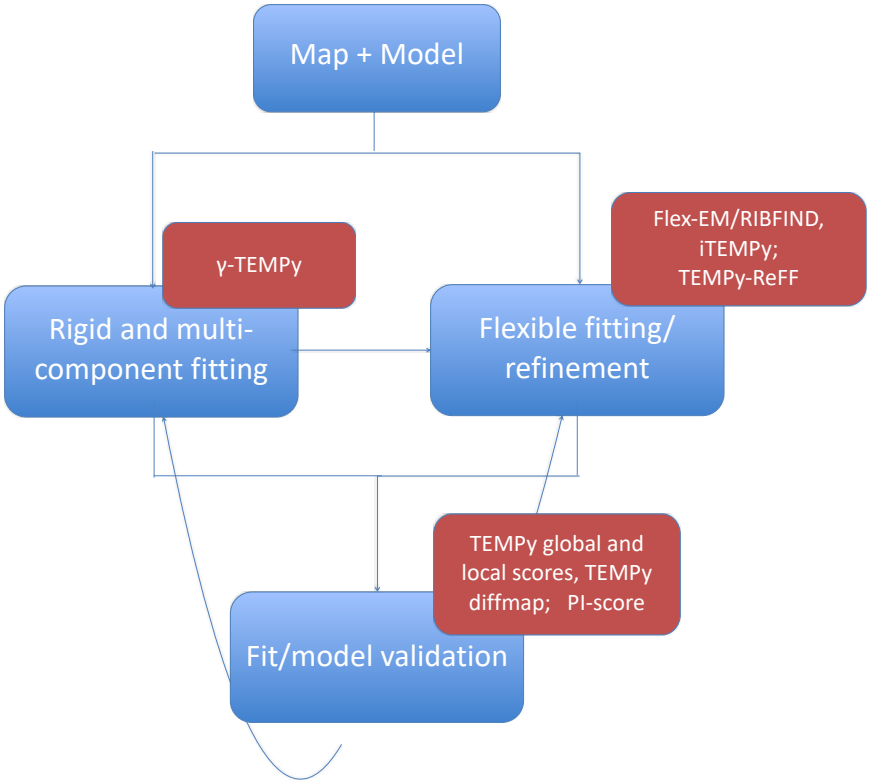
Map comparison
- Difference maps
- FSC

***Chimera* scripts**
Plots
Attribute files
PDB files

Cragnolini et al. *Acta Crystallogr* D (2021)

https://tempy.topf-group.com/

TEMPY2 TOOLS FOR MODELLING STRUCTURES FROM CRYO-EM MAPS

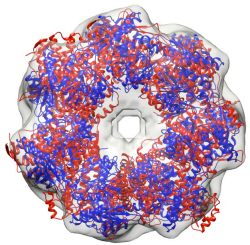


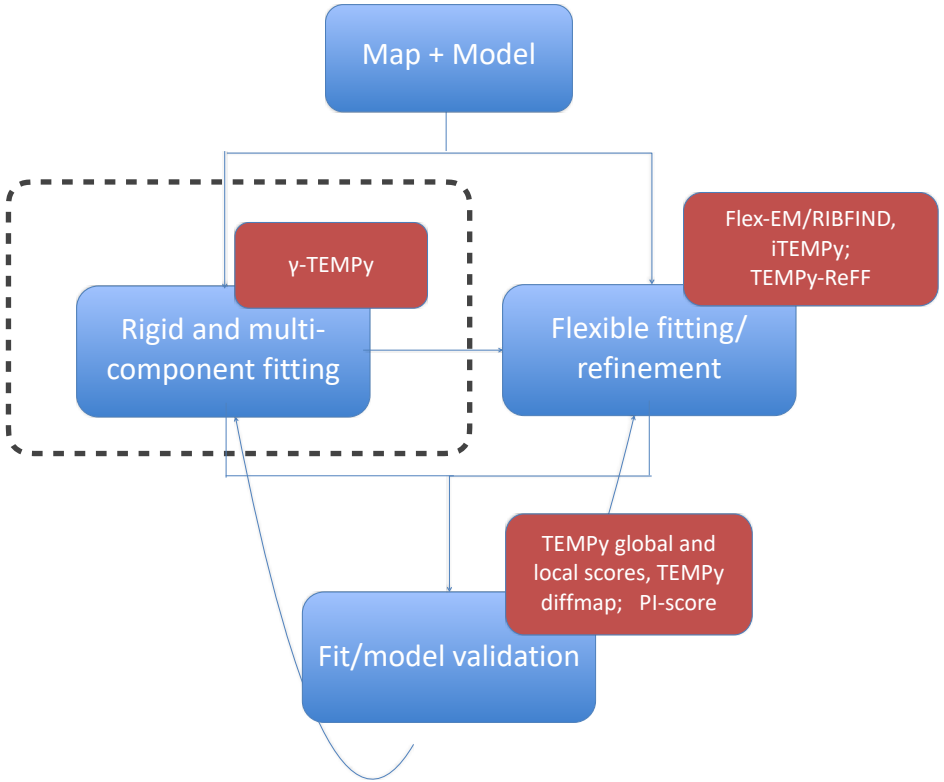
Topf et al. *Structure*, 2008
Vasishtan et al. *J Struct Biol* 2011
Farabella et al. *J App Cryst* 2015
Joseph et al. *Methods* 2016
Cragnolini et al. Acta Cryst D 2021
Malhotra et al. *Nat Comm* 2021
Cargnolini et al. *Proteins* 2021

https://topf-lab.gitlab.io/tempy

TEMPy2: NumPy and SciPy, Gemmi, Matplotlib, OpenMM

TEMPY2 TOOLS FOR MODELLING STRUCTURES FROM CRYO-EM MAPS



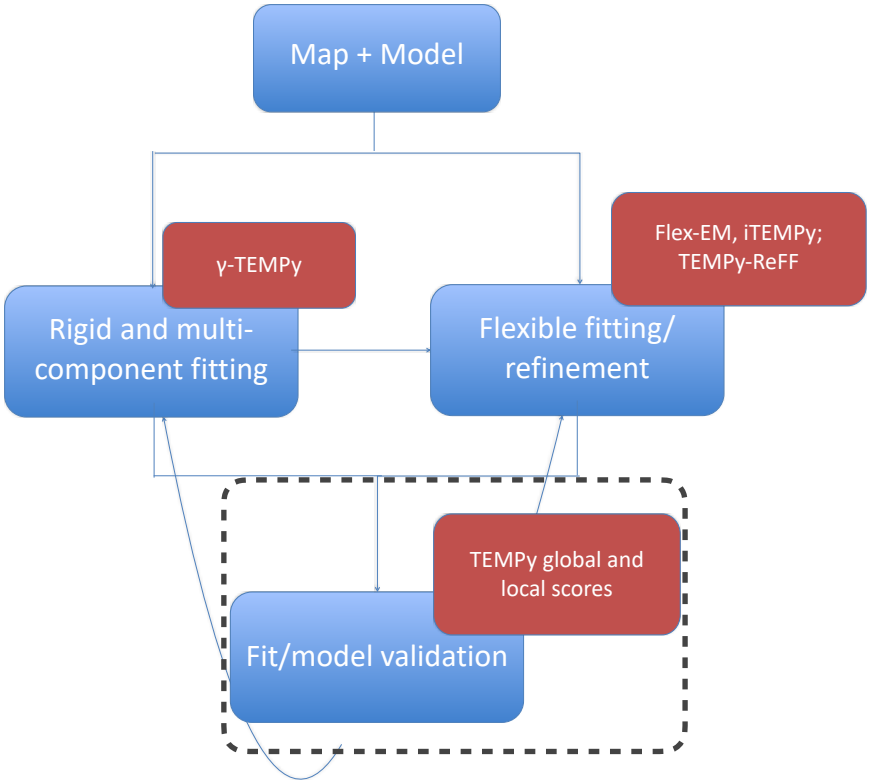


Topf et al. *Structure*, 2008
Vasishtan et al. *J Struct Biol* 2011
Farabella et al. *J App Cryst* 2015
Joseph et al. *Methods* 2016
Cragnolini et al. Acta Cryst D 2021
Malhotra et al. *Nat Comm* 2021
Cargnolini et al. *Proteins* 2021

https://topf-lab.gitlab.io/tempy

TEMPy2: NumPy and SciPy, Gemmi, Matplotlib, OpenMM

TEMPY2 TOOLS FOR MODELLING STRUCTURES FROM CRYO-EM MAPS



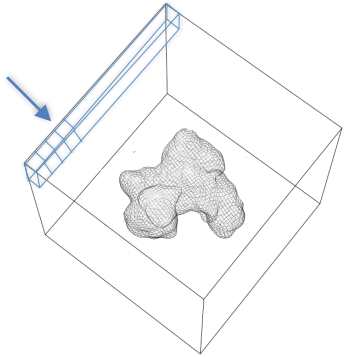
Topf et al. *Structure*, 2008
Vasishtan et al. *J Struct Biol* 2011
Farabella et al. *J App Cryst* 2015
Joseph et al. *Methods* 2016
Cragnolini et al. Acta Cryst D 2021
Malhotra et al. *Nat Comm* 2021
Cargnolini et al. *Proteins* 2021

https://topf-lab.gitlab.io/tempy

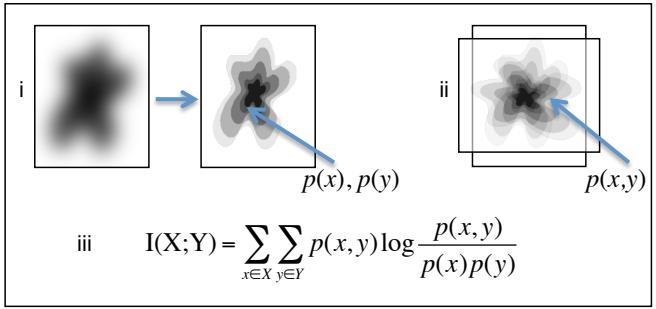
TEMPy2: NumPy and SciPy, Gemmi, Matplotlib, OpenMM

- Cross-correlation coefficient (CCC)

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



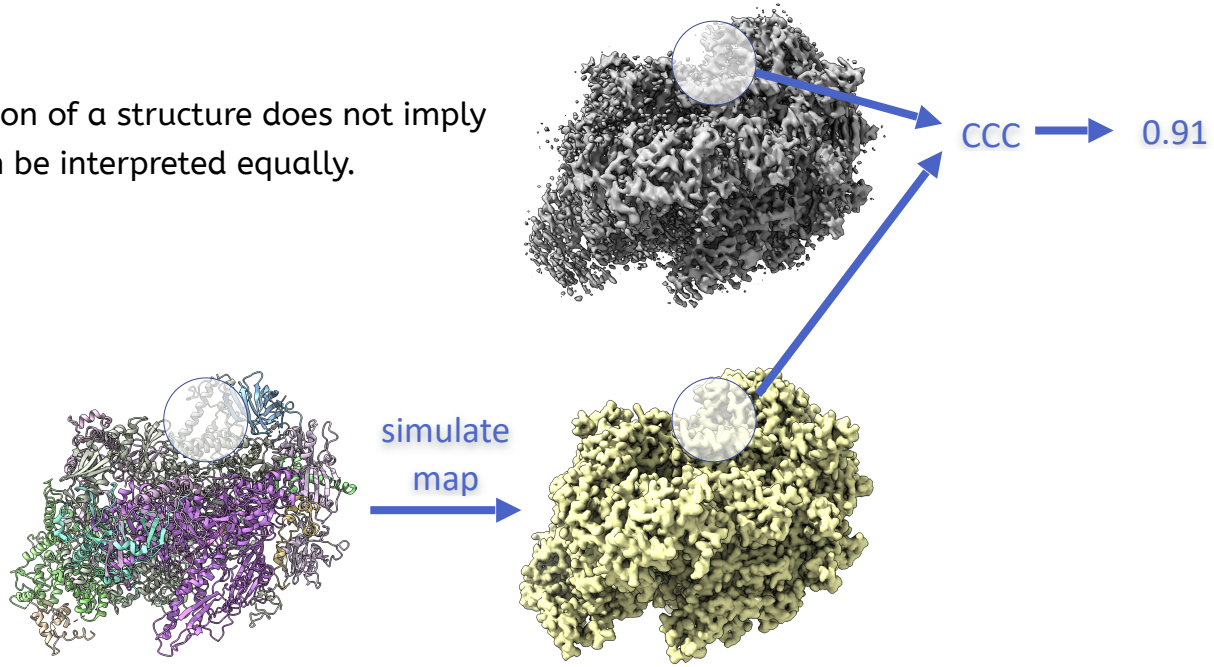
- Mutual information-based score (MI)



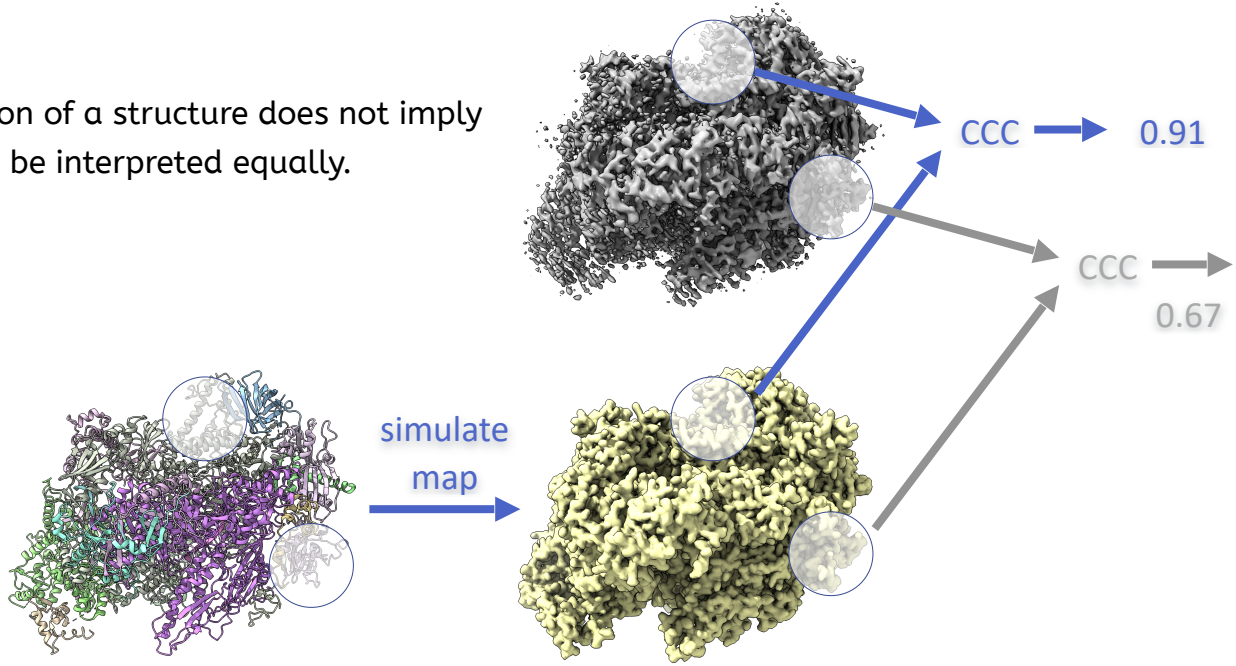
Useful at intermediate resolutions; noisy maps; less sensitive to relative intensity levels

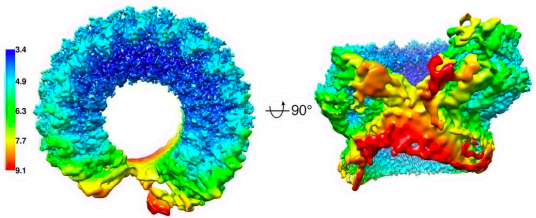
Vasishtan & Topf, *J Struct Biol* 2011, Joseph et al *J Struct Biol* 2017

* The overall resolution of a structure does not imply that all regions can be interpreted equally.



* The overall resolution of a structure does not imply that all regions can be interpreted equally.



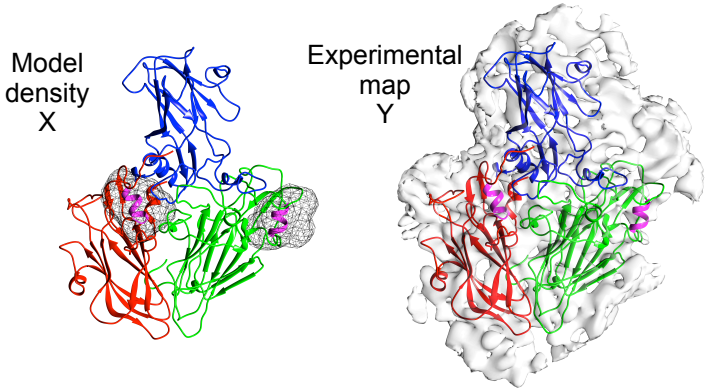


Yu et al., *EMBO J* 2022

* Local resolution maps are useful for estimating resolution variability, but visual inspection is always essential to assess map quality.

SCCC: Segment-based cross-correlation coefficient:

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

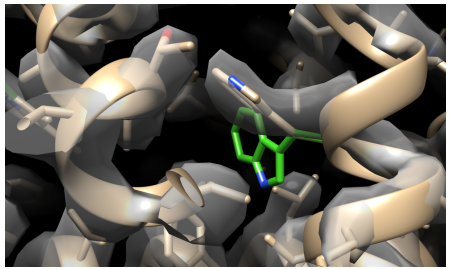


Farabella et al. *J Appl. Cryst.* 2015

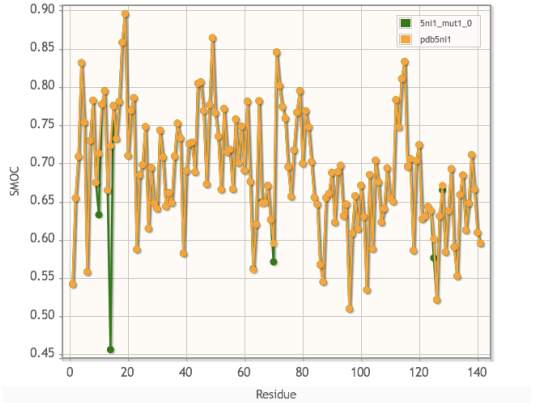
SMOC: An overlap coefficient is calculated over voxels covered by each **Residue** (and the local neighbourhood):

Joseph et al. *Methods* 2016

$$SMOC = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} {X_i}^2}\sqrt{\sum_{i=1}^{n} {Y_i}^2}}$$

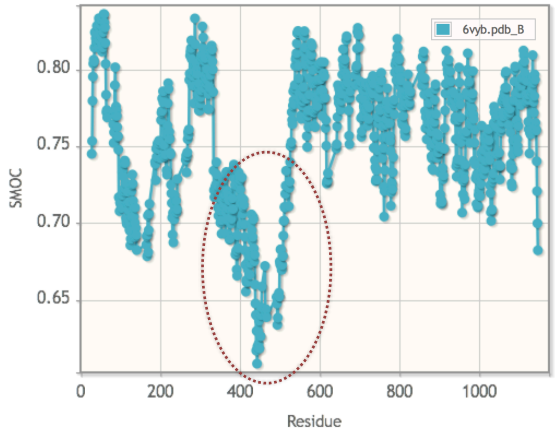


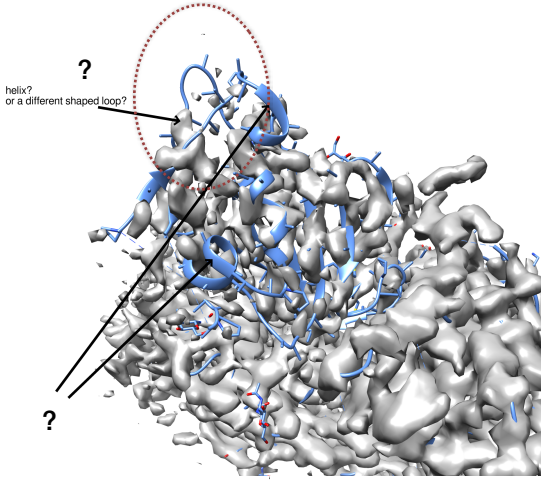
EMD-3488 (3.2Å)
Deposited model PDB: 5NI1



EXAMPLE

$$SMOC = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}} \sqrt{\sum_{i=1}^{n} Y_i^{\,2}}}$$

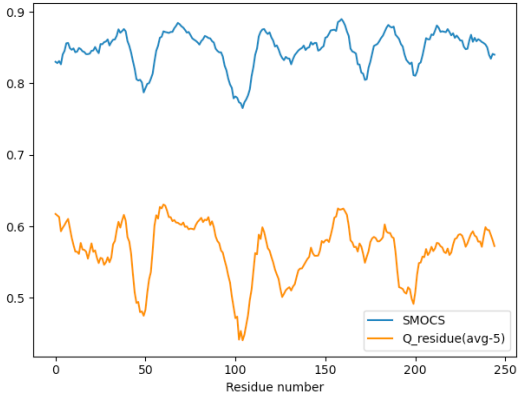


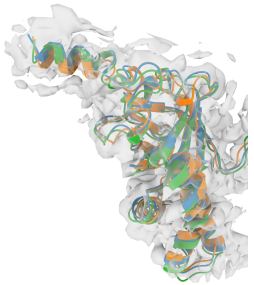


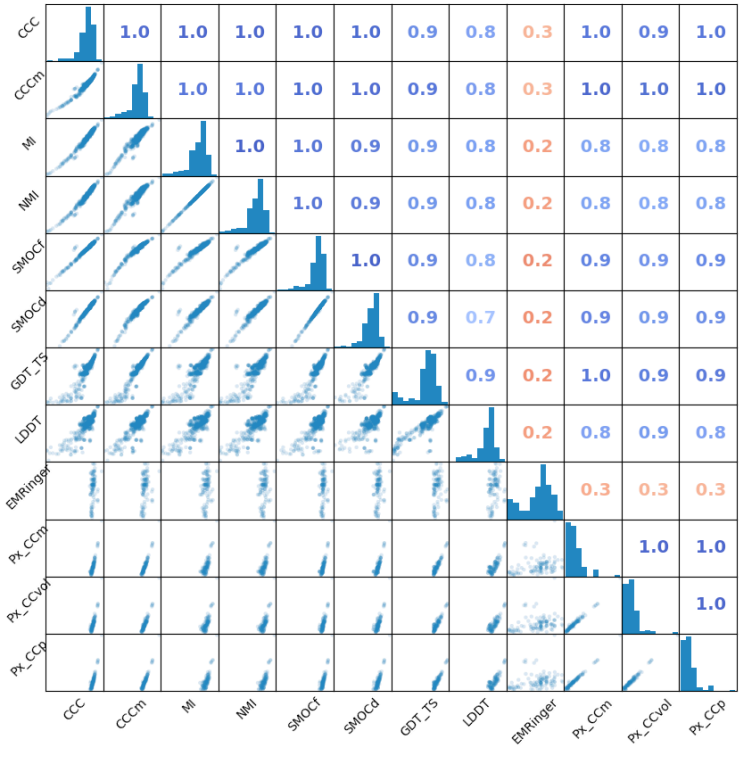
SARS-CoV-2 spike ectodomain (3.2 Å resolution)

Nearly all scores exhibit a relatively high degree of correlation and this is true across all targets

Strong correlations with GDT_TS





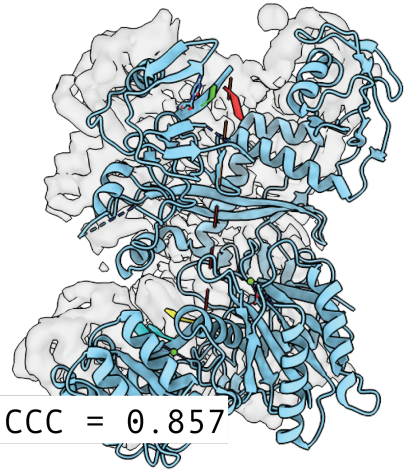


T1092-D2

AR9 RNA polymerase

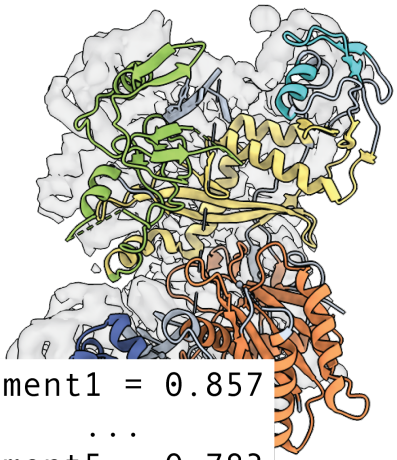
- There are many different scoring functions in TEMPy, which can be grouped into:

Global Scores
(one number per model)



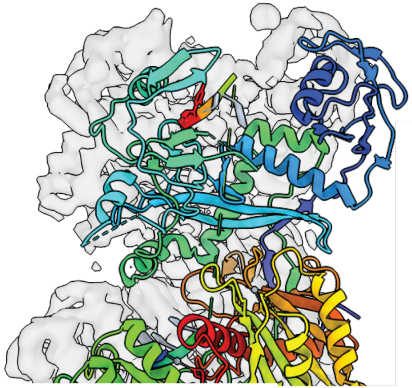
```
CCC = 0.857
```

Segmented Scores
(one number per segment)

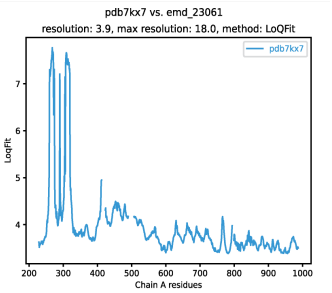


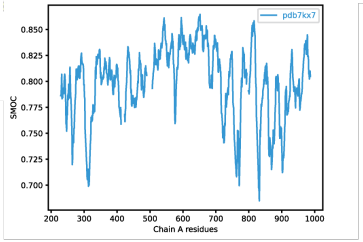
```
Segment1 = 0.857
      ...
Segment5 = 0.783
```

Residue Scores
(one number per residue)



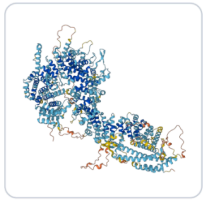
```
Residue1 = 0.857
      ...
ResidueN = 0.783
```





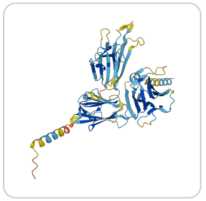
https://topf-lab.gitlab.io/tempy

Enter the "AI" era



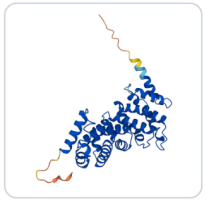
Nuclear pore complex protein Nup205

Part of a large complex that acts as a gateway in and out of the cell nucleus



Gametocyte surface protein P45/48

From the malaria parasite; a candidate protein for including in vaccines

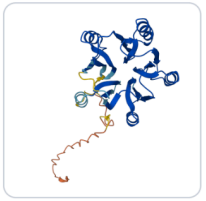


CCR4-NOT transcription complex subunit 9

Regulates an important cellular process (the rate of mRNA degradation)

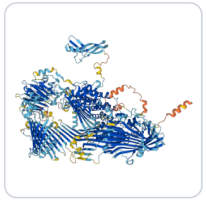
Ice nucleation protein

Bacterial protein that can trigger ice formation at relatively high temperatures, causing frost damage to plants



F2OH23.2 protein

Plant protein; represents a potential new structural superfamily unlike anything seen before

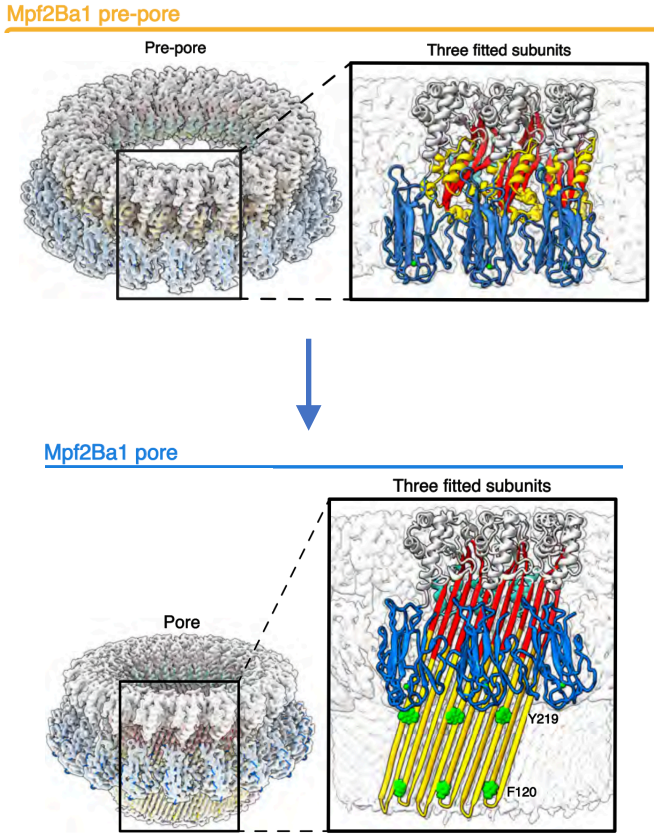


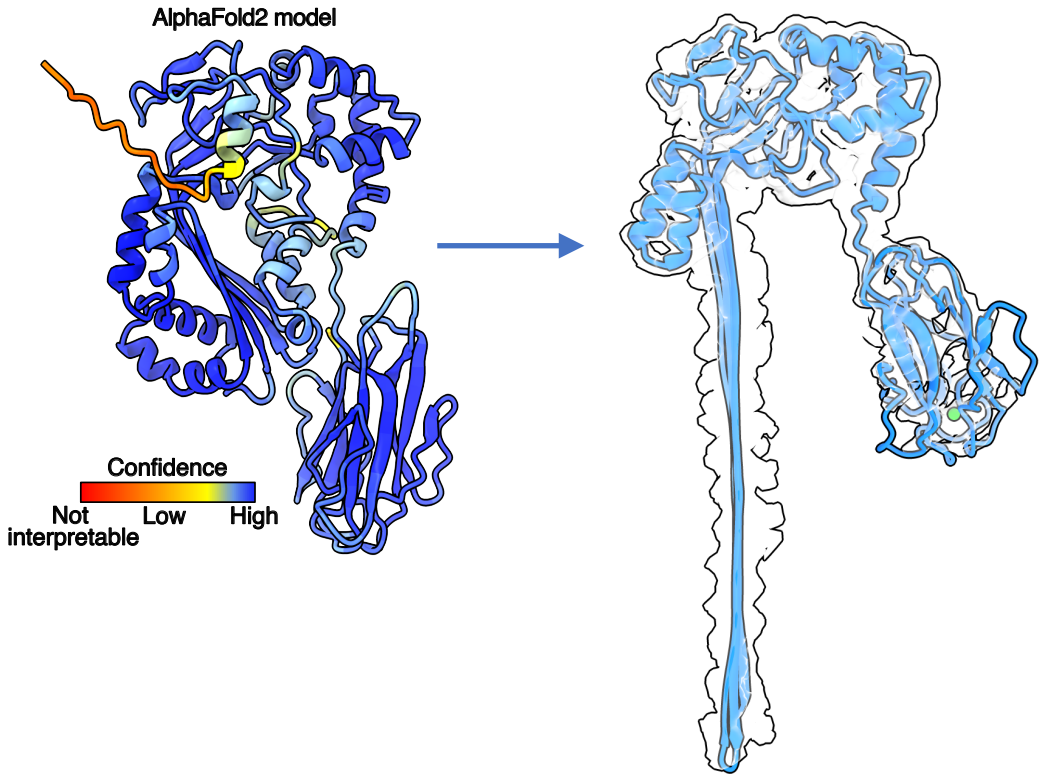
Vitellogenin

Involved in the immune system of egg-laying animals including honeybees

Jumper et al., *Nature* 2021 Baek et al., *Science* 2021 Lin et al., *BioRxiv* 2022 Zhou et al., *Nat protoc* 2022

GOOD INITIAL PREDICTIONS FOR PROTEIN ASSEMBLIES BUT…

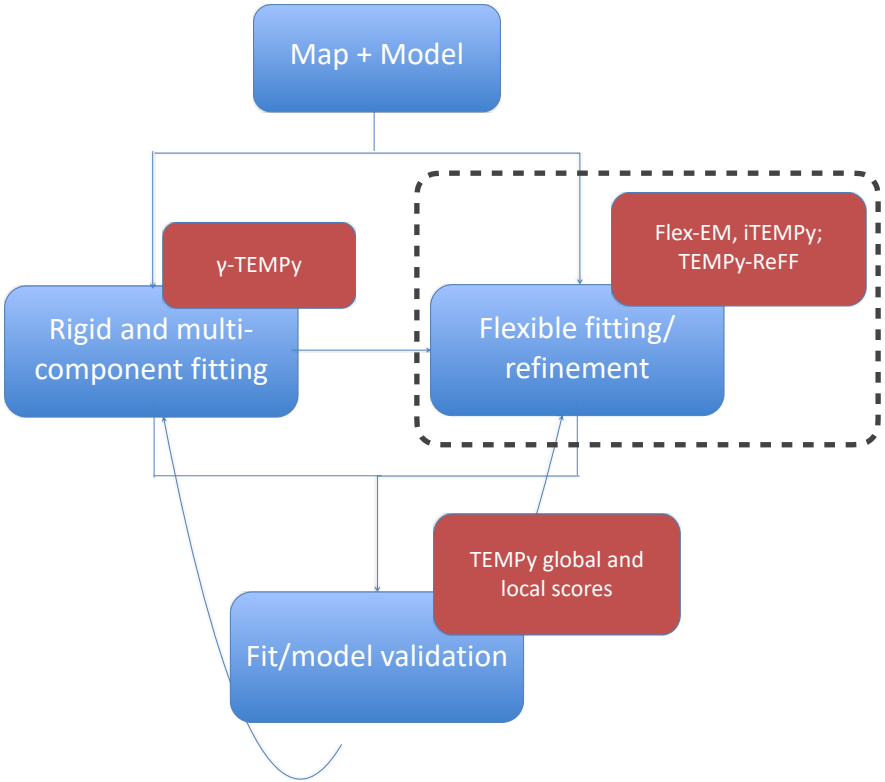




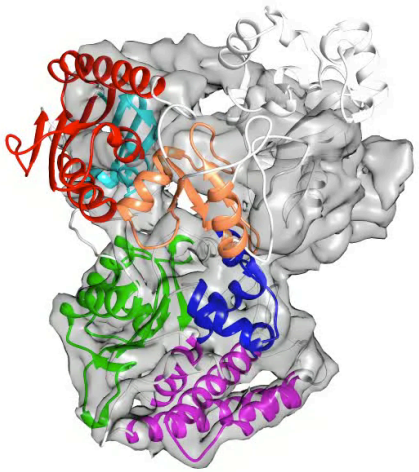
Marini et al., *Nat Comms* 2023

Using validation approaches, we can *iteratively* improve the fit to the density

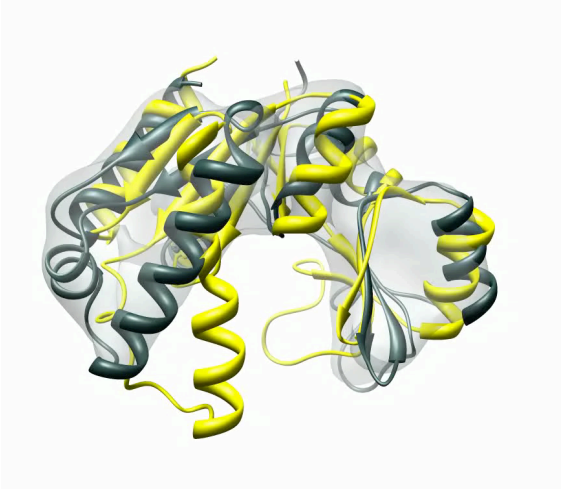
TEMPY2 TOOLS FOR MODELLING STRUCTURES FROM CRYO-EM MAPS



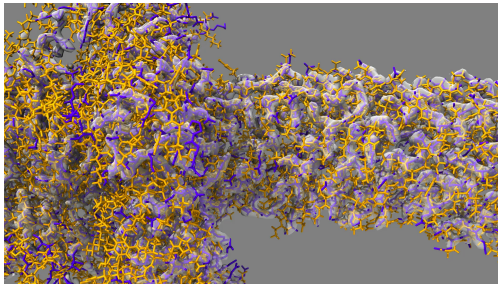
FLEX-EM REFINEMENT AT MEDIUM TO LOW RESOLUTION



sub-domains

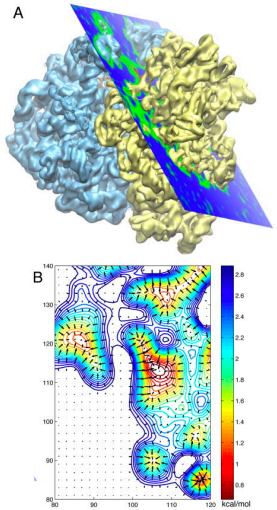


secondary structure elements



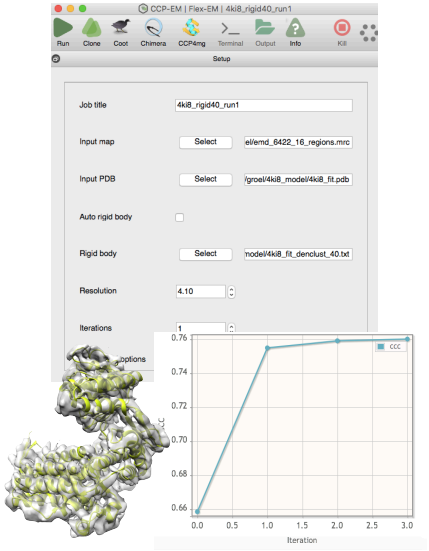
residue / all-atom

Different approaches for flexible fitting have been developed over the years.



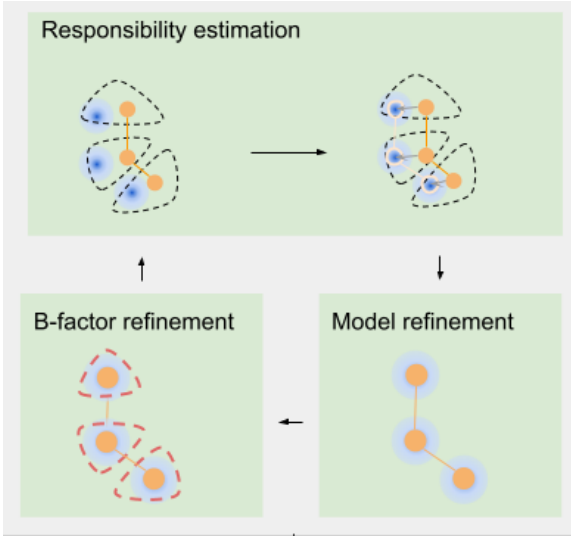
Molecule Dynamics Flexible Fitting (MDFF) uses a standard force field with a potential based on the density

Trabuco et al., *Methods* 2008



Flex-EM improves the correlation between the model sand the cryo-EM map

Topf et al., 2008
Joseph et al., 2016



TEMPy-ReFF (Responsibility-based Flexible-Fitting) utilizes a gaussian mixture model with expectation-maximisation to improve the model in the map

Cragnolini et al., 2021, BioRxiv 2022
Malhotra et al., 2023

REAL-SPACE REFINEMENT WITH FLEX-EM/MODELLER

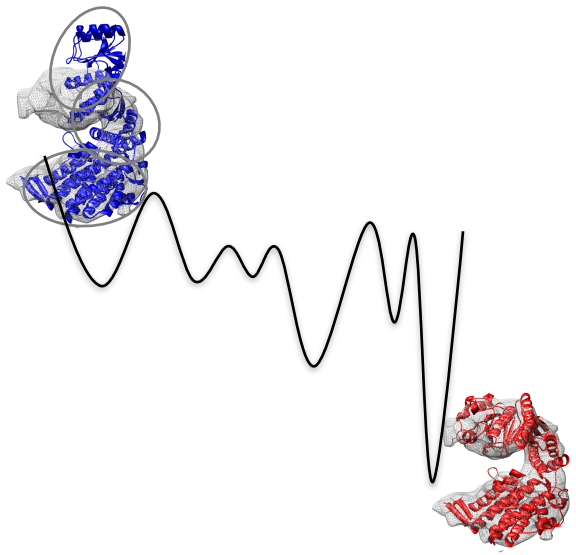
- During the refinement, the atoms are displaced in the direction that maximizes their cross-correlation with the cryoEM density map (E^{CC}) and minimizes the violations of the stereochemical (E^{SC}) and non-bonded contacts (E^{NB}):

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

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

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

- Atoms in a rigid body move together during the course of refinement.



Topf et al. *Structure* 2008

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

TEMPY-REFF: REFINEMENT USING MIXTURE MODELS

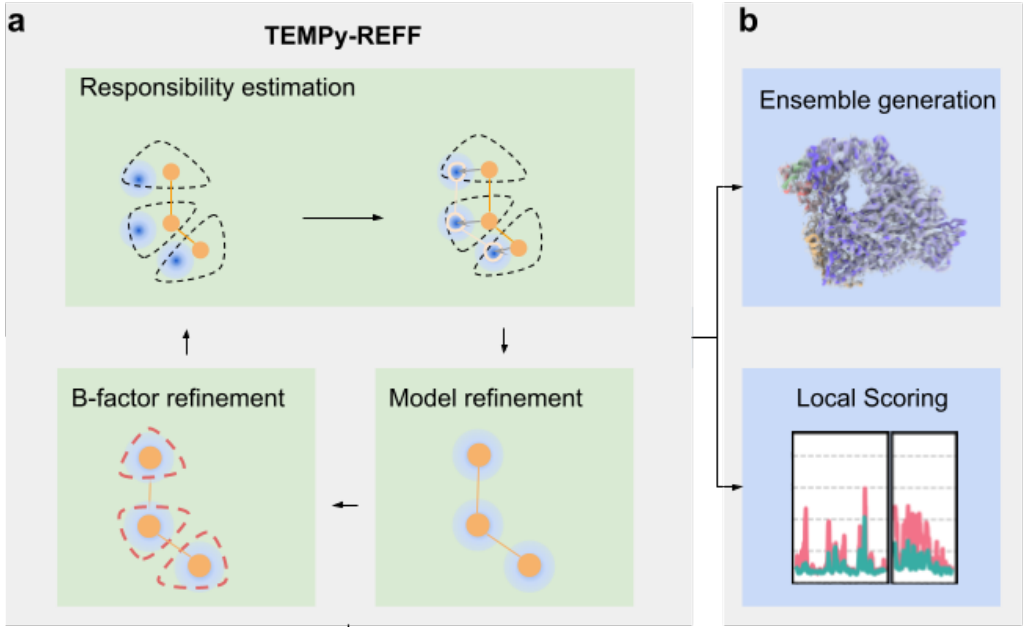
- Use an expectation-maximisation approach
- Use one gaussian per atom, and a background noise term
- Improves atomic position and variance (B-factor)

TEMPy and OpenMM

Amber14 forcefield

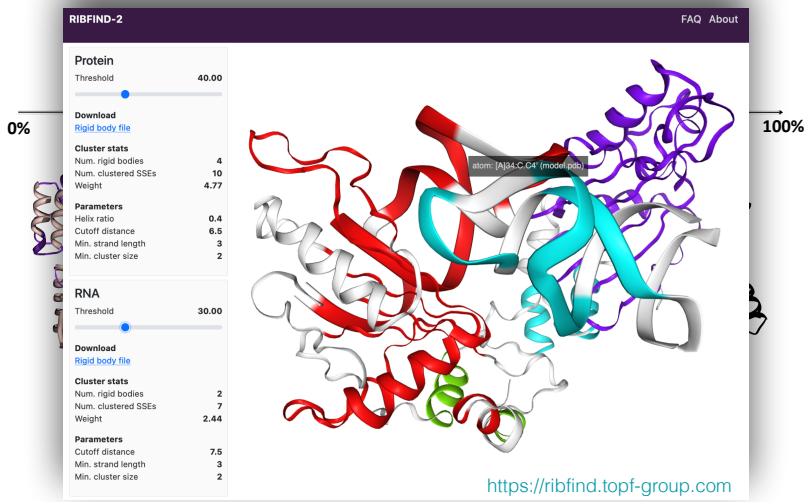
Generalised Born implicit solvent model

Integration with Langevin dynamics, at 100K



RIBFIND2 - IDENTIFY RIGID BODIES IN PROTEIN AND RNA STRUCTURES

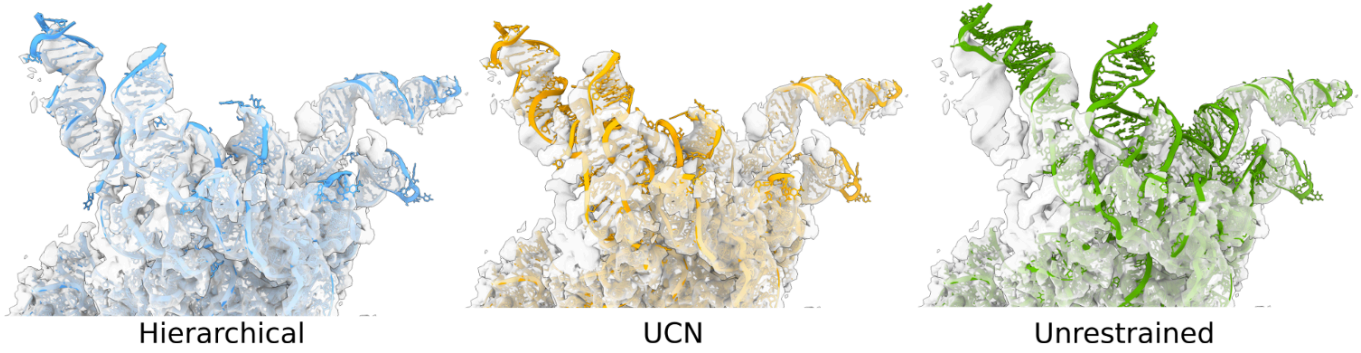
- When the resolution of density map is insufficient to fit smaller entities like individual residues or atoms, we use RIBFIND-defined rigid bodies.
- Clustering of secondary structures together (SSEs computed using DSSP or RNAView).
- Allows faster large body movements in the initial stages or refinement.



Pandurangan *et al. J Struct Biol and Bioinformatics* 2012
Malhotra, Mulvaney et al *NAR* 2023

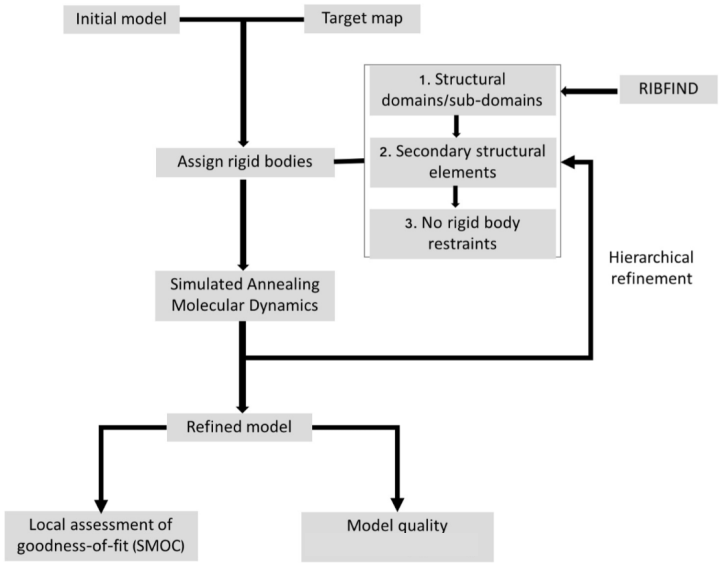
RIBFIND2 - HELPS TO AVOID GETTING "STUCK" IN LOCAL OPTIMA

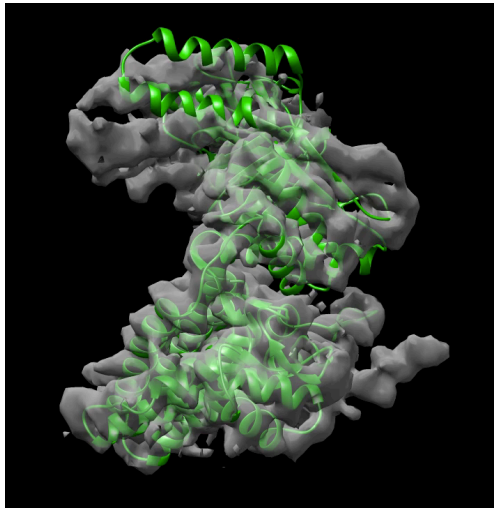
- **human SSU processome**
- **pre-A1 into state post-A1**
- **EMD-23938, 2.7 Å resolution**



Malhotra, Mulvaney et al. *NAR* 2023

FLEXIBLE FITTING AT MEDIUM RESOLUTION BY HIERARCHICAL REFINEMENT

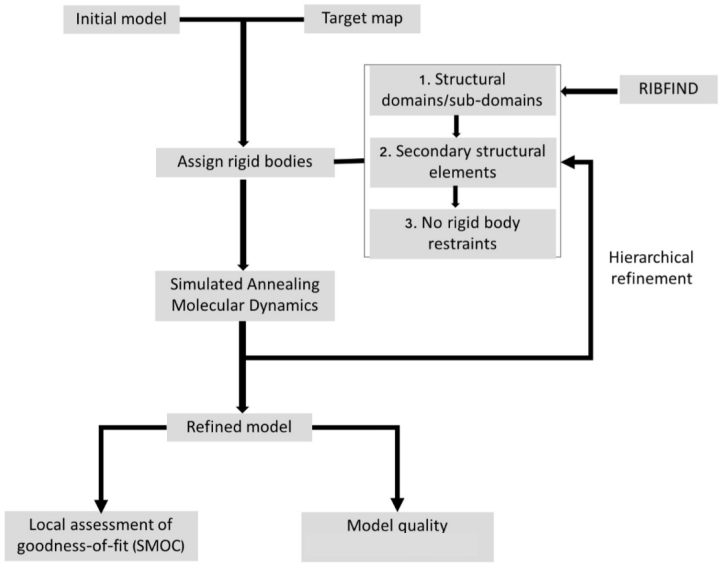


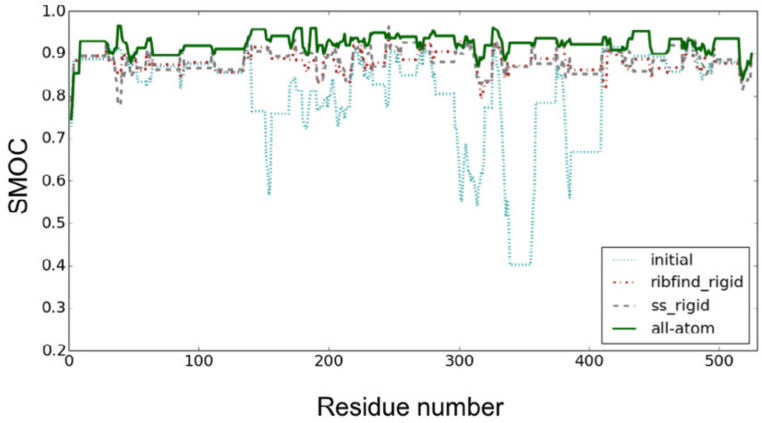


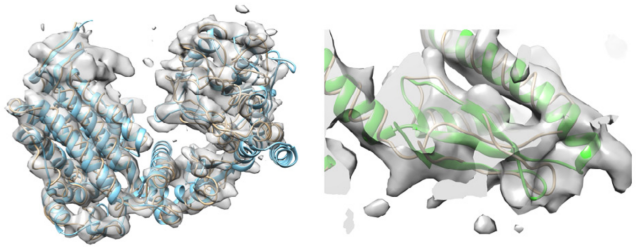
Unliganded GroEL at 4.2 Å resolution (EMD-5001)

Joseph et al. *Methods* 2016

FLEXIBLE FITTING AT MEDIUM RESOLUTION BY HIERARCHICAL REFINEMENT



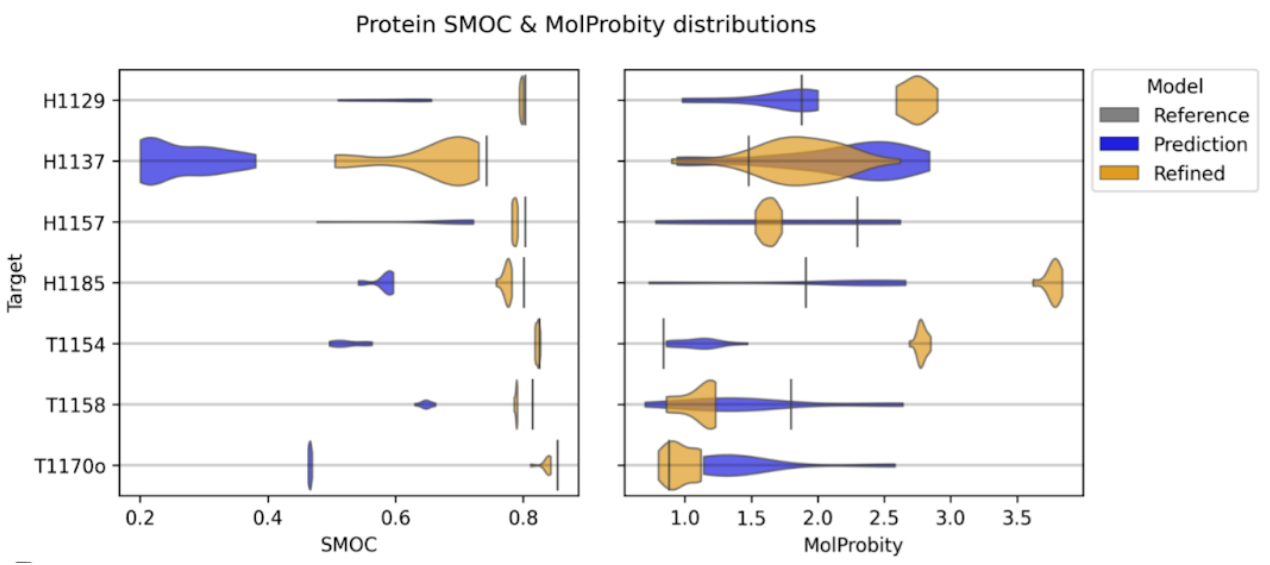




Unliganded GroEL at 4.2 Å resolution (EMD-5001)
Initial model: ADP-bound GroEL (PDB: 4KI8)
Refined model
Deposited model (PDB: 3CAU)

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

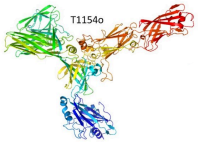
TEMPY-REFF REFINEMENT OF CASP15 MODELS IN CRYO-EM MAPS

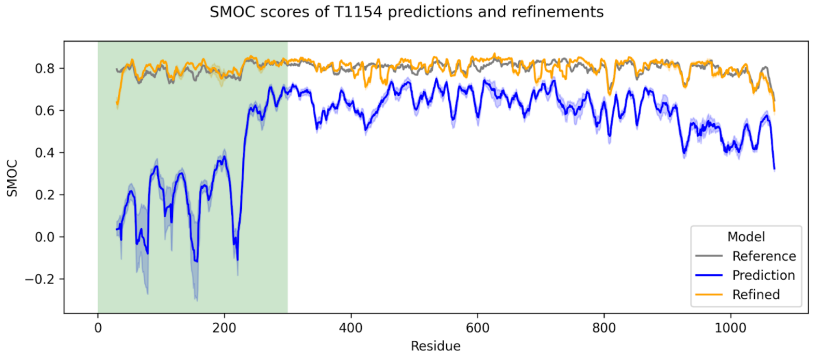


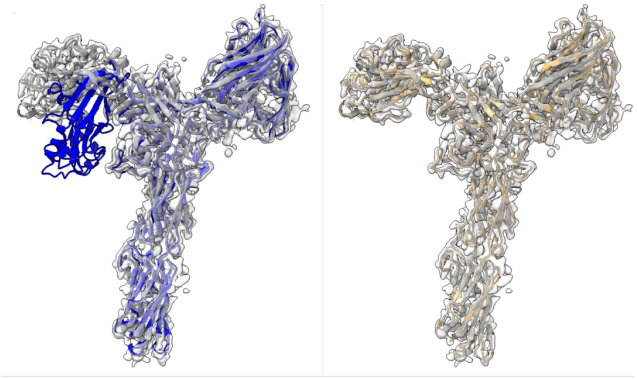
Mulvaney et al. *BioRxiv* 2023

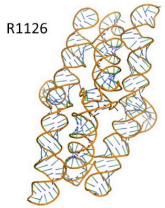
TEMPy2

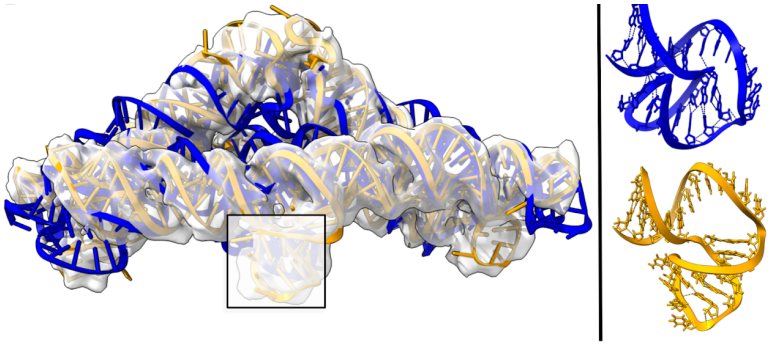
TEMPY-REFF REFINEMENT OF CASP15 MODELS (PROTEIN COMPLEXES AND RNA)







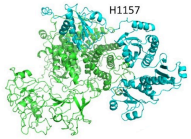


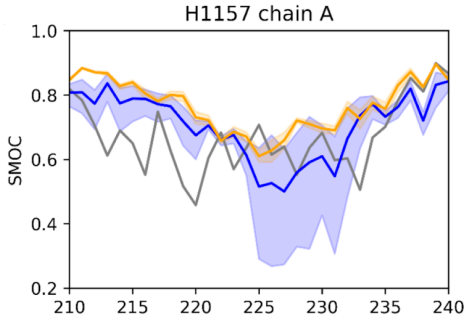


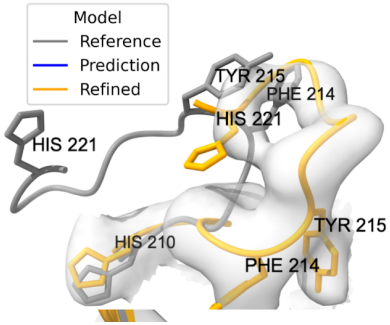
Mulvaney et al, BioRxiv 2023

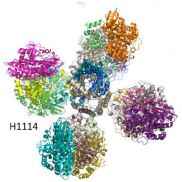
TEMPy2

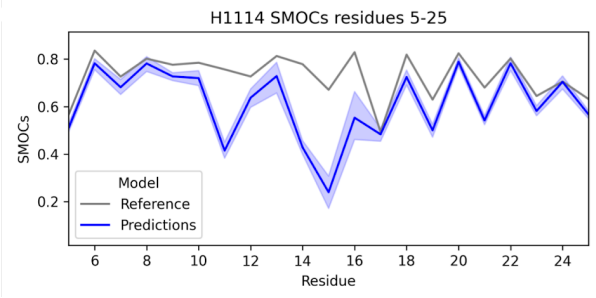
EXPERIMENTAL VALIDATION AND REFINEMENT BY CRYO-EM (CASP15)

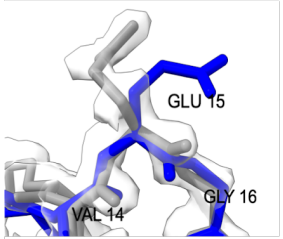


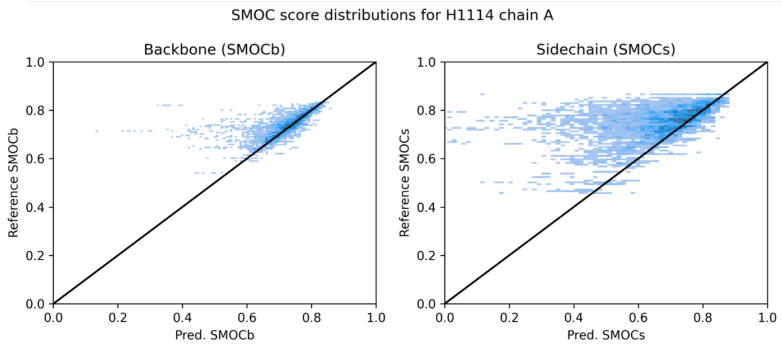








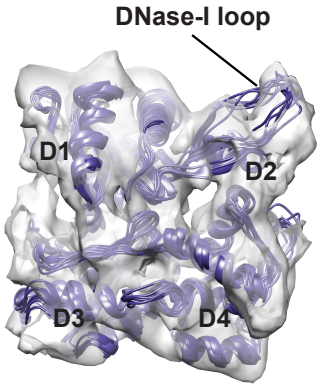


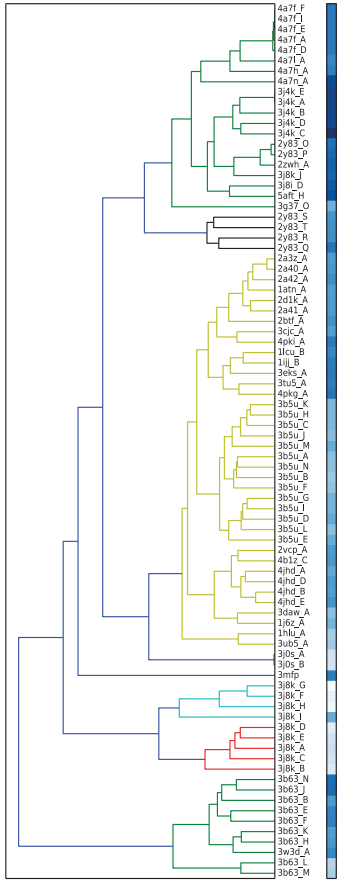


Mulvaney et al, BioRxiv 2023

- The quality of the fit to the density often varies in different parts of the cryoEM map
- Proteins are dynamic
- Some of these parts may be better represented by multiple models

ENSEMBLE OF FITS

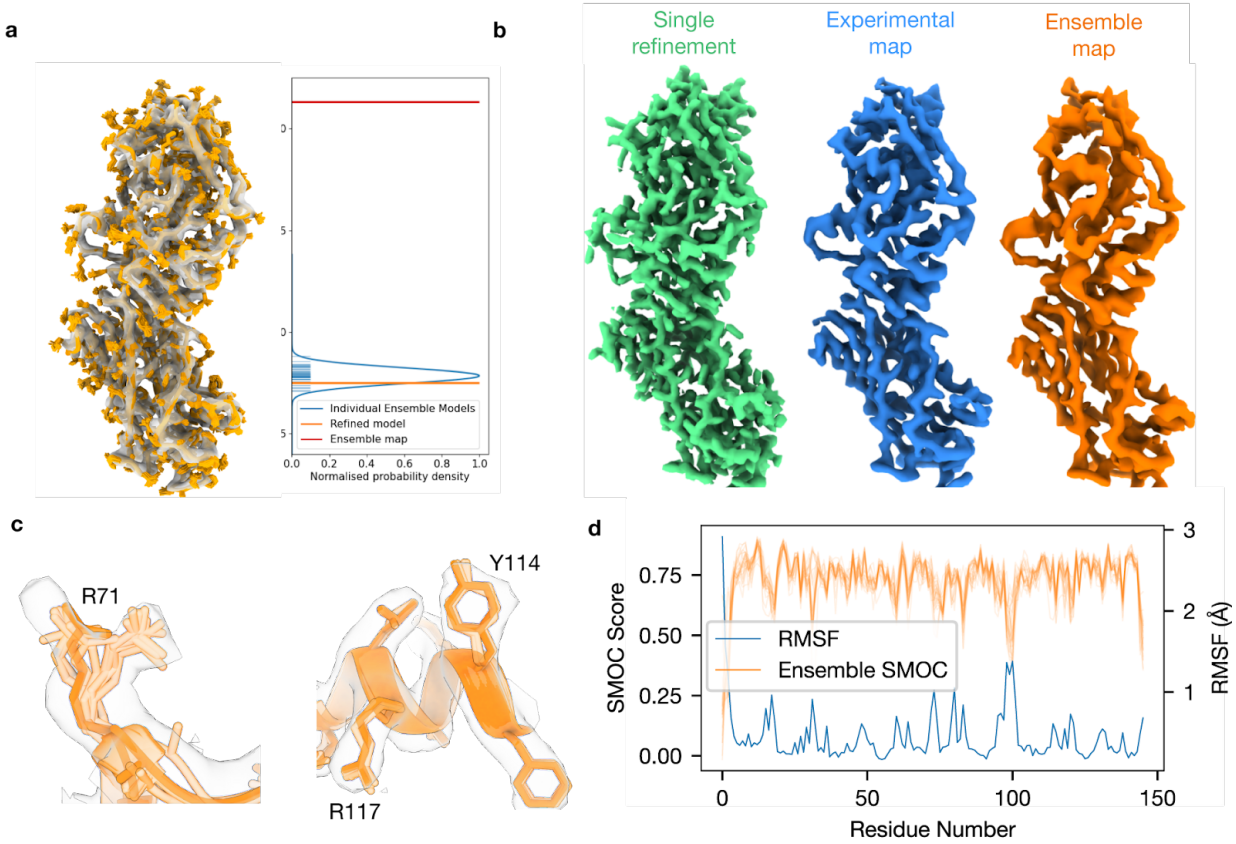




EMD: 1990 (8.9 Å resolution)
84 actin subunits from PDB

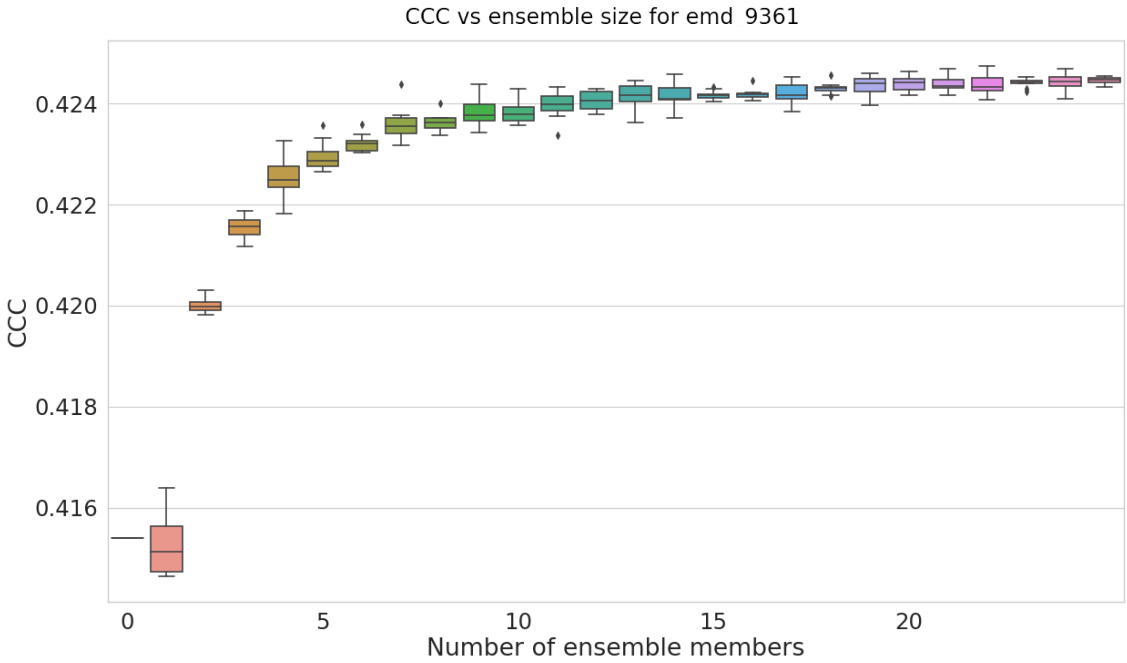
Farabella et al. *J Appl. Cryst.* 2015

TEMPY-REFF CAN GENERATE ENSEMBLES

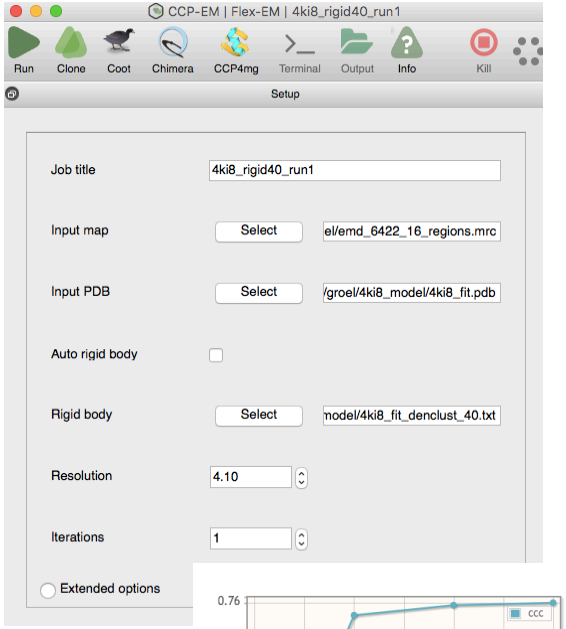


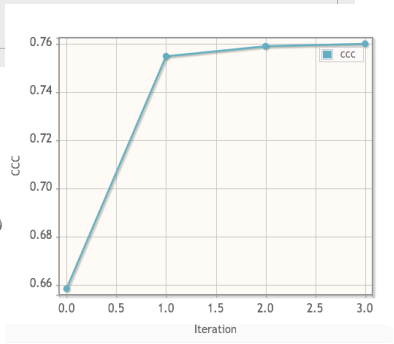
http://cci.lbl.gov/ceres

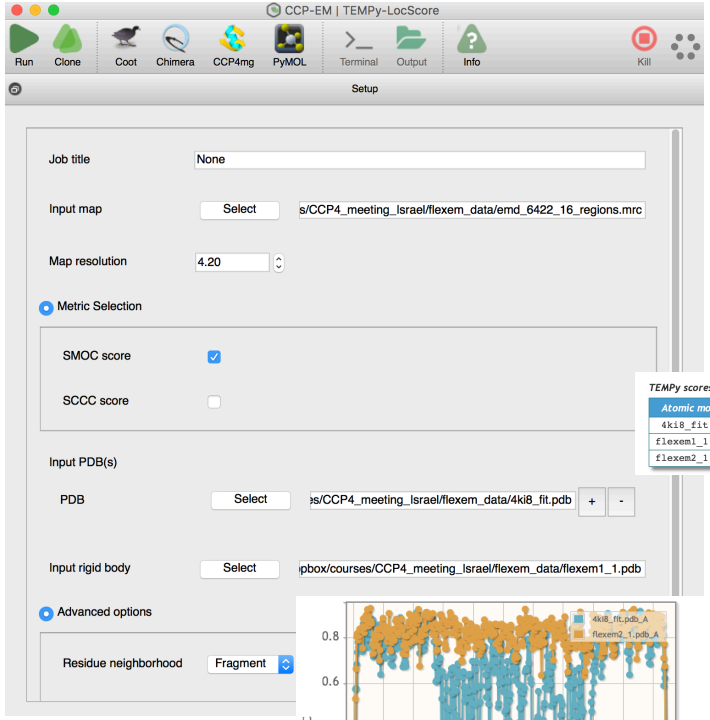
ENSEMBLE SIZE

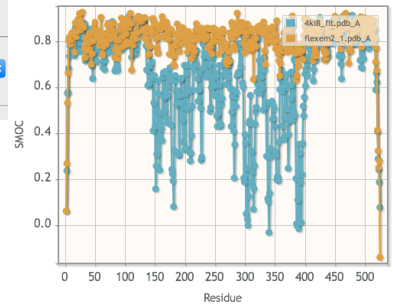


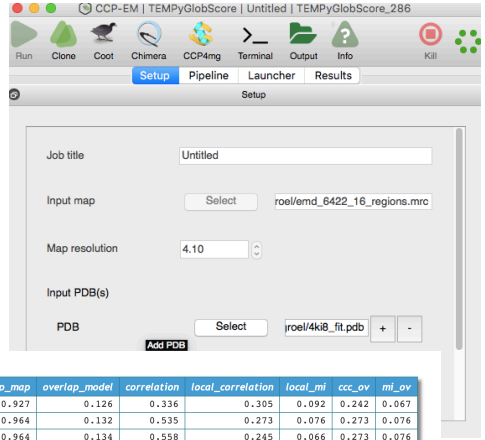
SOFTWARE: CCP-EM IMPLEMENTATION OF FLEX-EM AND TEMPY











TEMPy scores

Atomic model	overlap_map	overlap_model	correlation	local_correlation	local_mi	ccc_ov	mi_ov
4ki8_fit.pdb	0.927	0.126	0.336	0.305	0.092	0.242	0.067
flexem1_1.pdb	0.964	0.132	0.535	0.273	0.076	0.273	0.076
flexem2_1.pdb	0.964	0.134	0.558	0.245	0.066	0.273	0.076

Global scores (CCC, MI…)

https://www.ccpem.ac.uk/

THANKS…

Topf group (current)

Sanjana Nair
Rebecca Brooker
Luca Genz
Thomas Mulvaney
Joseph Beton
Karen Manalastas
Aaron Sweeney
Laetitia Adeler-Ohde
Matthias Pfeifer
Birgit Märtens
Mauro Maiorca
Guendalina Marini
Manaz Kaleel
Aylin del Moral
Maryam Nikooei

previous members:

Sony Malhotra (STFC - CCP-EM)
Agnel Joseph (STFC - CCP-EM)
Arun Prasad Pandurangan (Cambridge)

CASP

Andriy Kryshtafovych

Birkbeck

Helen Saibil

STFC

Martyn Winn
Tom Burnley
Colin Palmer