TEMPy-ReFF + ChemEM

CCP-EM Icknield Model Building Workshop

05.11.2024

Maya Topf & Aaron Sweeney (CSSB Hamburg)

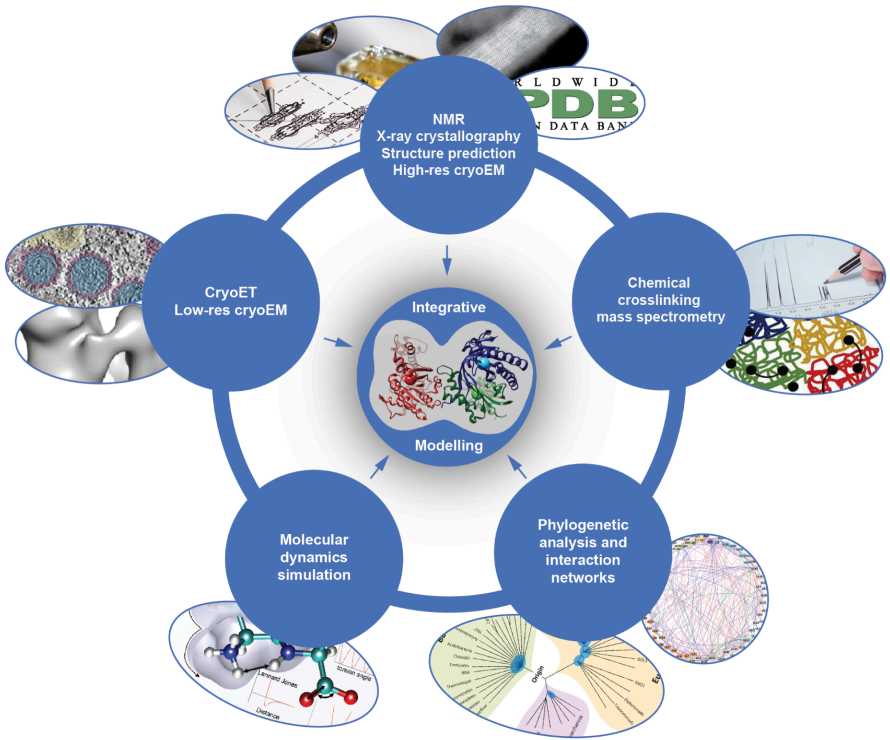
UKE HAMBURG

LIV

CSSB
Centre for Structural
Systems Biology

INTEGRATIVE MODELLING OF MACROMOLECULAR ASSEMBLIES

CSSB
Centre for Structural
Systems Biology

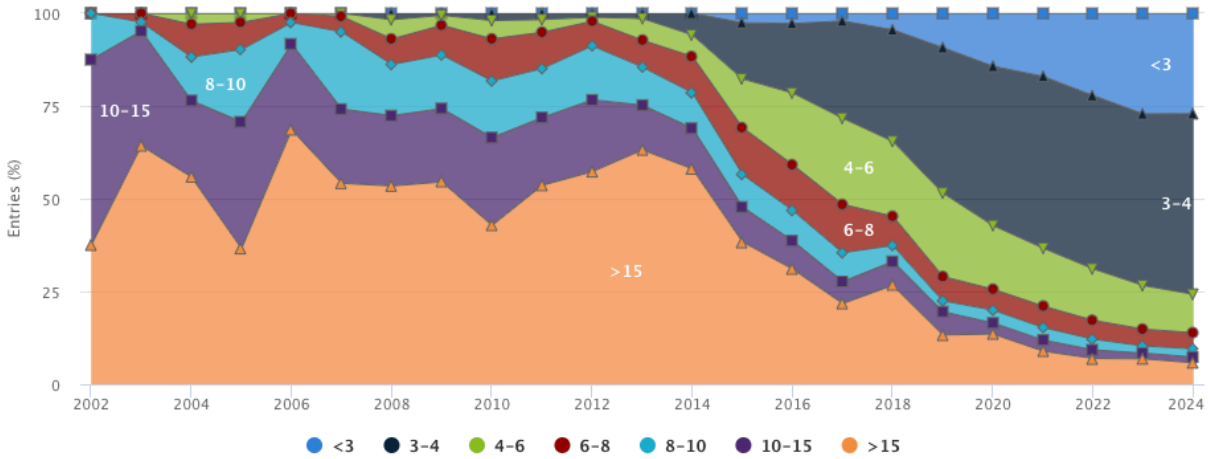


- **Develop** methods for integrative modelling of macromolecular assemblies
- **Apply** methods to model structures of viral assemblies (e.g. herpesviruses, arenaviruses)
- **Analyse** structural models to further understand their function
- **Distribute** and support new software for the structural biology community

UKE HAMBURG
LIV Leibniz Institute of Virology

Joesph et al. *Curr Opin Struct Biol*, 2017

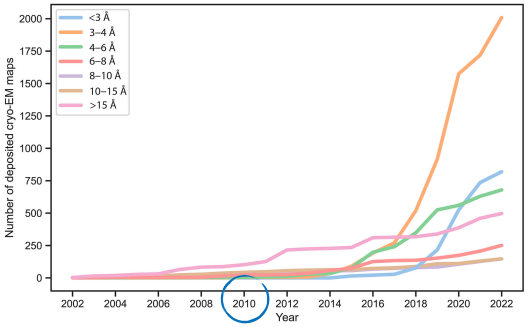
EMDB ENTRY RESOLUTION IN SHELLS PER YEAR

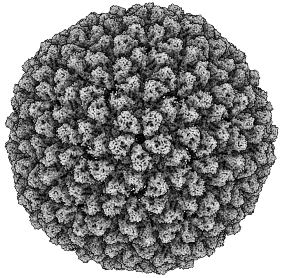


"The Resolution Revolution"
W Kühlbrandt, *Science* 2014

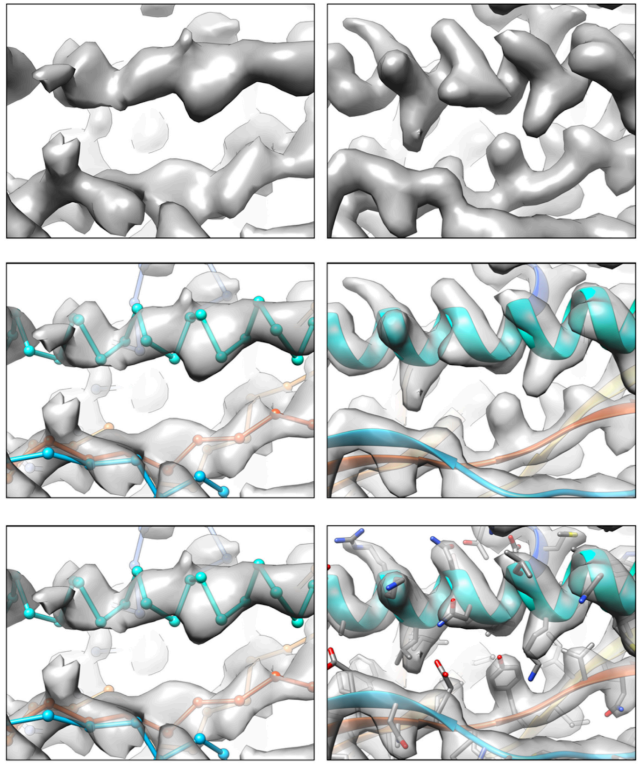
https://www.ebi.ac.uk/emdb/statistics/

THE "RESOLUTION REVOLUTION"





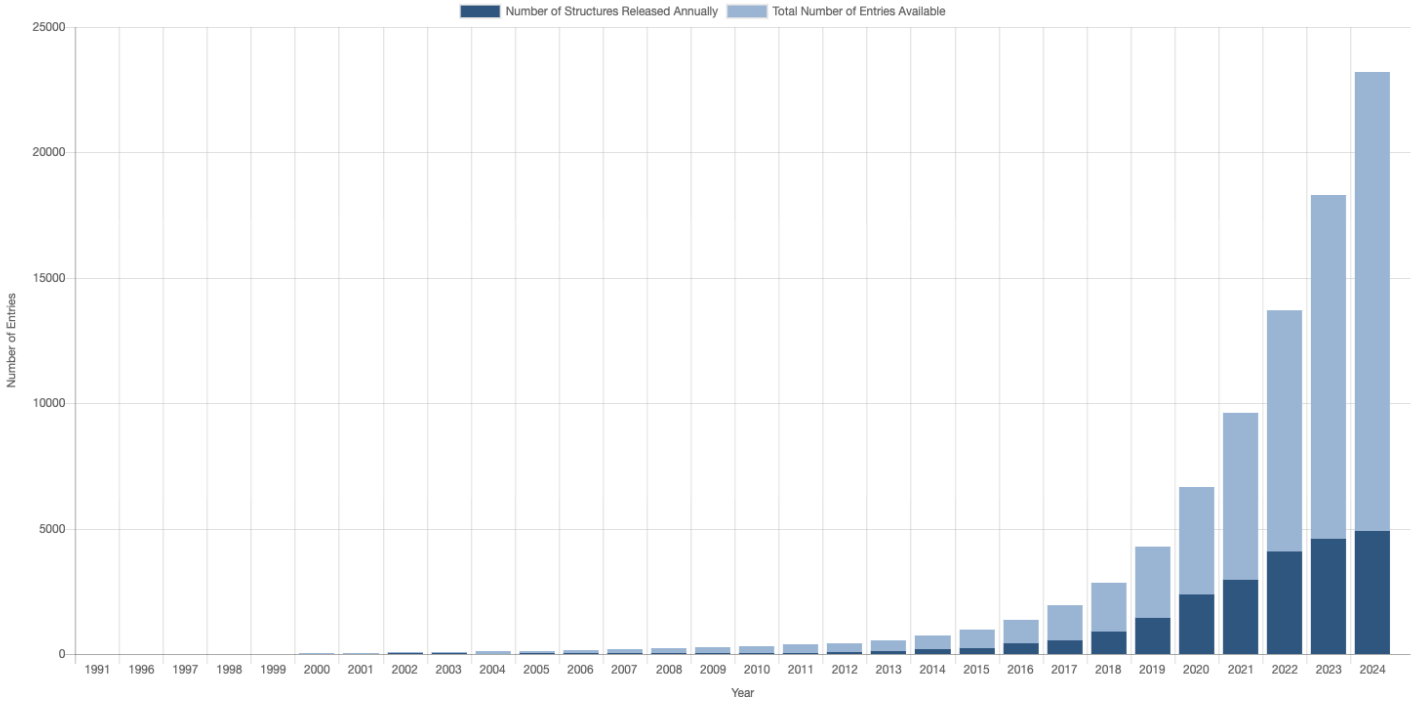
Bacteriophage P22
4Å (actually ~5Å)
EMDB 5137
Chen et al., 2011



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

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

NUMBER OF CRYOEM ENTRIES IN PDB

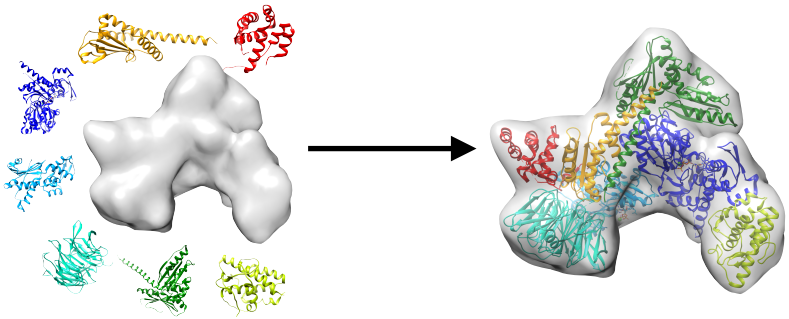


- ~10% of the entries in PDB are associated with maps in EMDB
- These days most entries are associated with either a single model or multiple models

https://www.rcsb.org/stats

What is it?

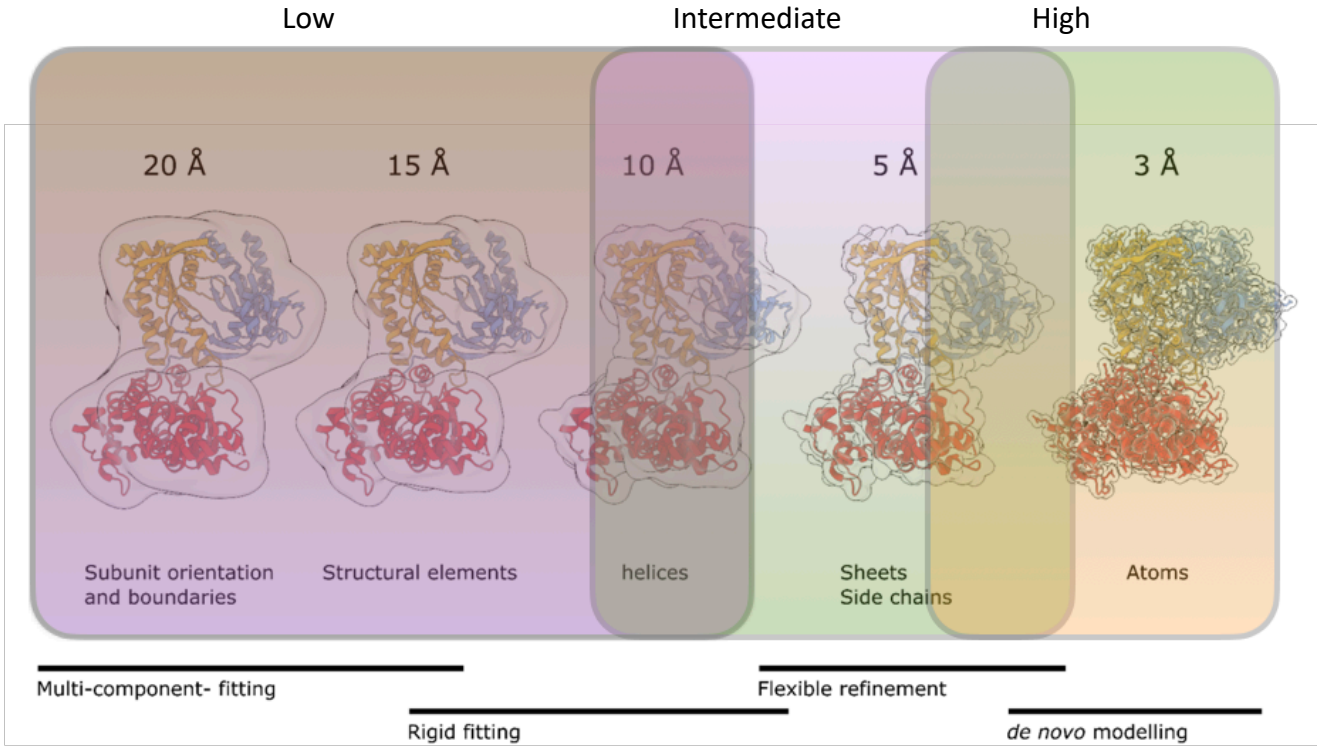
* Generation, representation, “manipulation” of the 3D structure of biological molecules
* A molecular model in cryoEM is a compact interpretation of a density map in light of everything known *a priori* about the structure-composition of the macromolecule of interest



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 small molecules and other macromolecules;
5. To help predict how new related molecules might look.
6. To simulated the dynamics of the model to better understand its function

STRUCTURAL FEATURES AT DIFFERENT RESOLUTION LEVELS



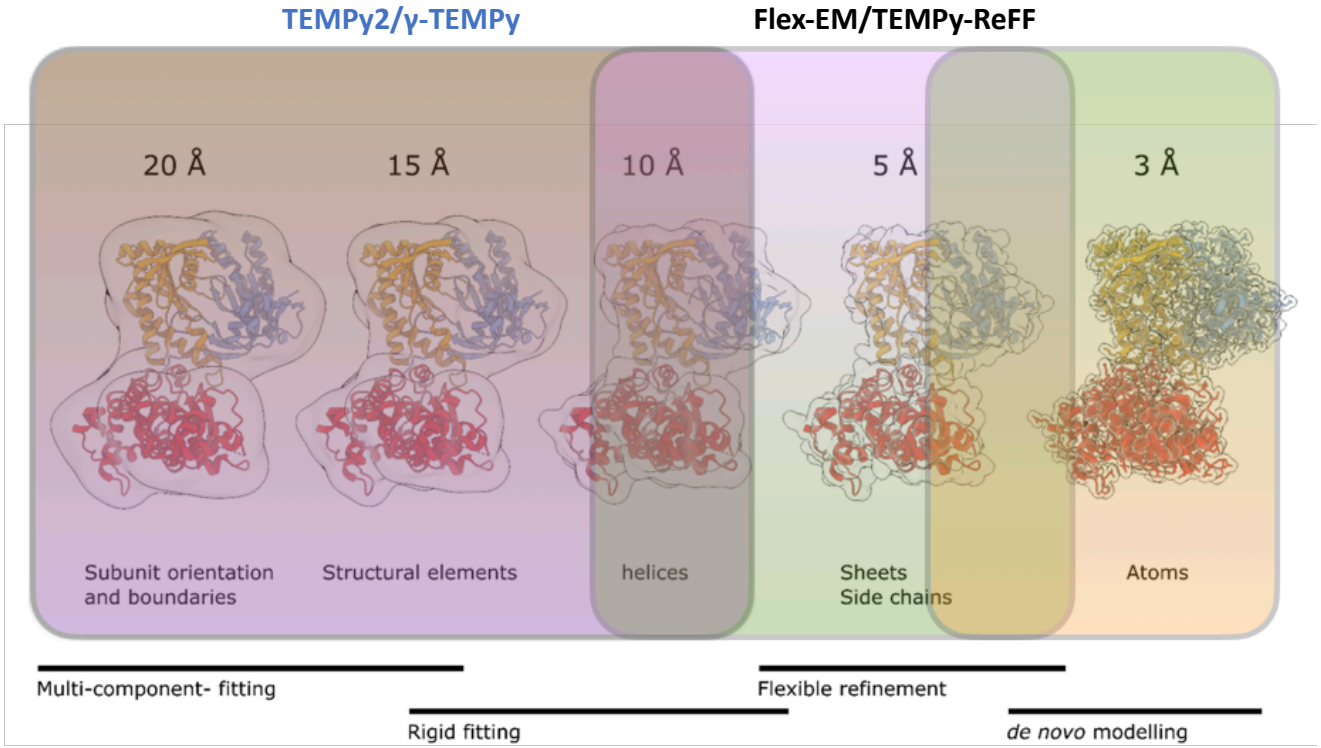
- If a complete experimental structure is available (and/or a reliable predicted model) 10 Å data may be sufficient.

- If no sequence/composition data is available even 3 Å may be challenging

Beton et al. WIREs Comput Mol Sci. 2023

Beton et al. WIREs Comput Mol Sci. 2023

STRUCTURAL FEATURES AT DIFFERENT RESOLUTION LEVELS

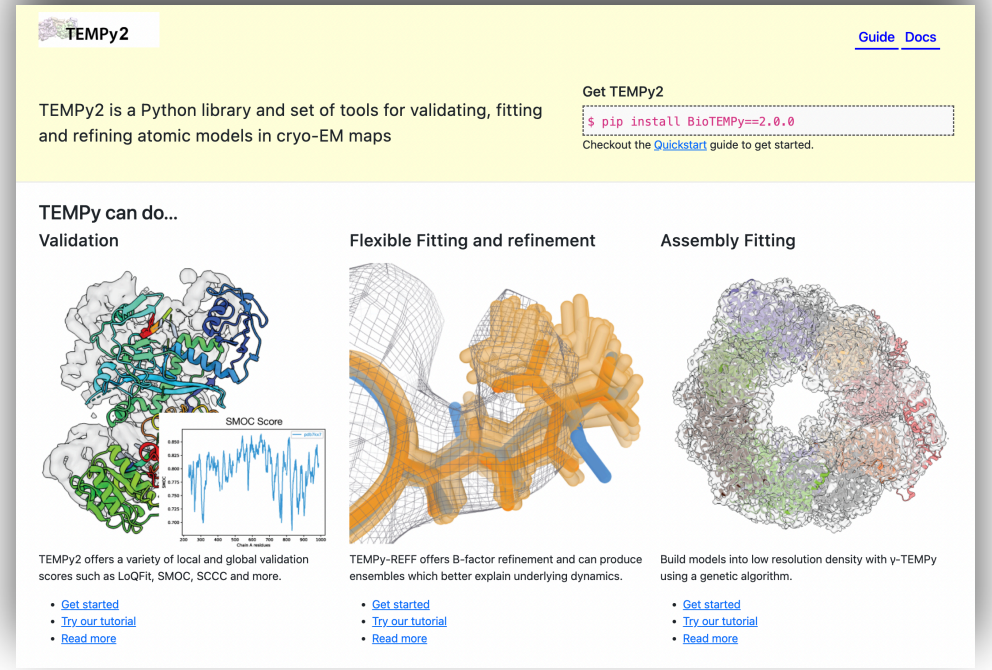


- If a complete experimental structure is available (and/or a reliable predicted model), 10 Å data may be sufficient.

- If no sequence/composition data is available even 3 Å may be challenging

Beton et al. WIREs Comput Mol Sci. 2023

TEMPY2 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

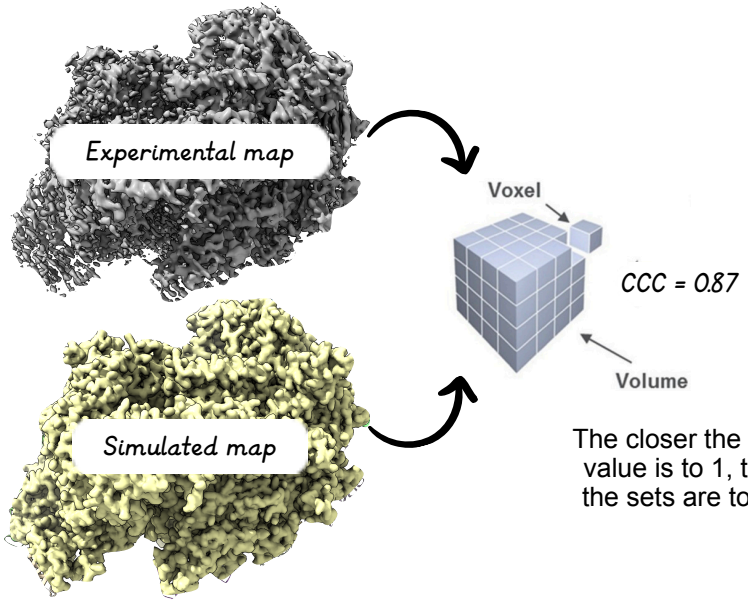
Local scoring
- SCCC
- SMOC
- LoQfit

Map comparison
- Difference maps
- FSC

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

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

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



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

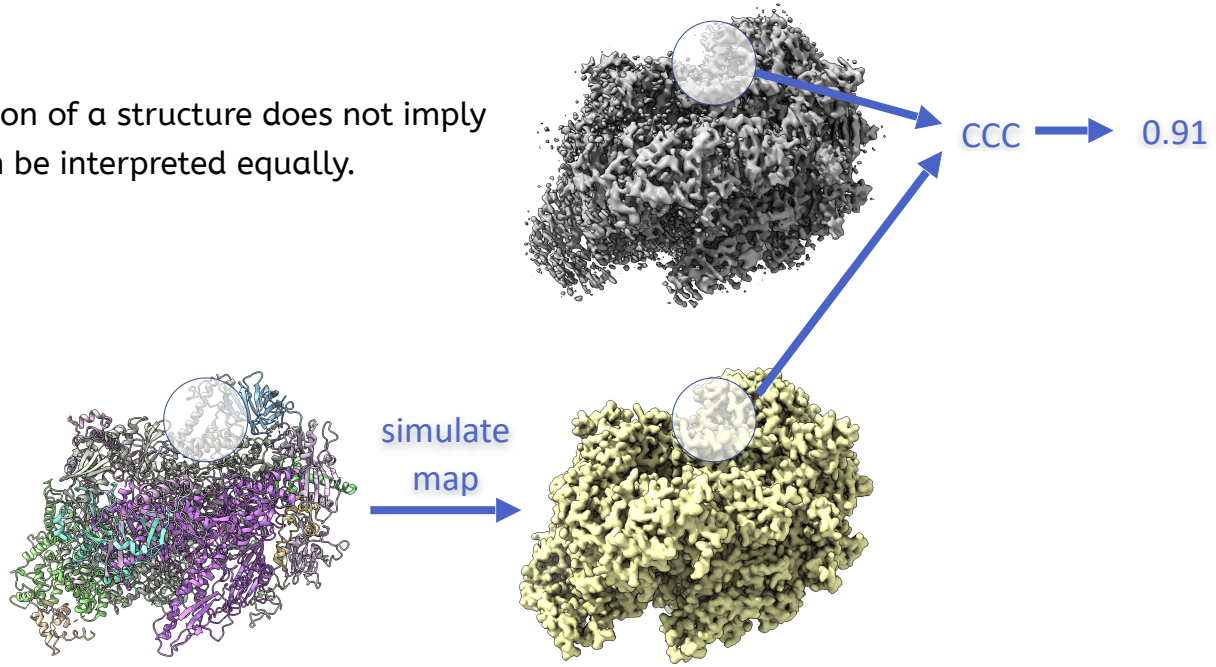
* When possible, a known or pre-calculated model is placed and fitted in the cryo-EM map as a rigid body.
* Often an exhaustive six dimensional (6D) grid search of the 3 translational and 3 rotational degrees of freedom is performed to locate the highest global cross-correlation.

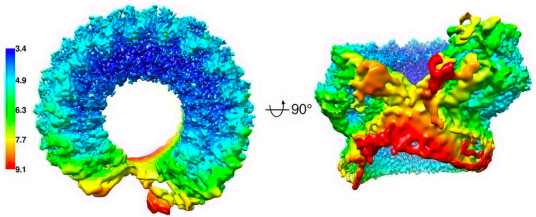
Cross Correlation Coefficient $$\mathbf{CCC}_{\rho,\mathrm{lin}} = \int_{\mathbf{x}} (\rho_{\mathbf{obs}}(\mathbf{x})\rho_{\mathbf{calc}}(\mathbf{x},\mathbf{m}))^2 d^3\mathbf{x}$$

Mutual information score $$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

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

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

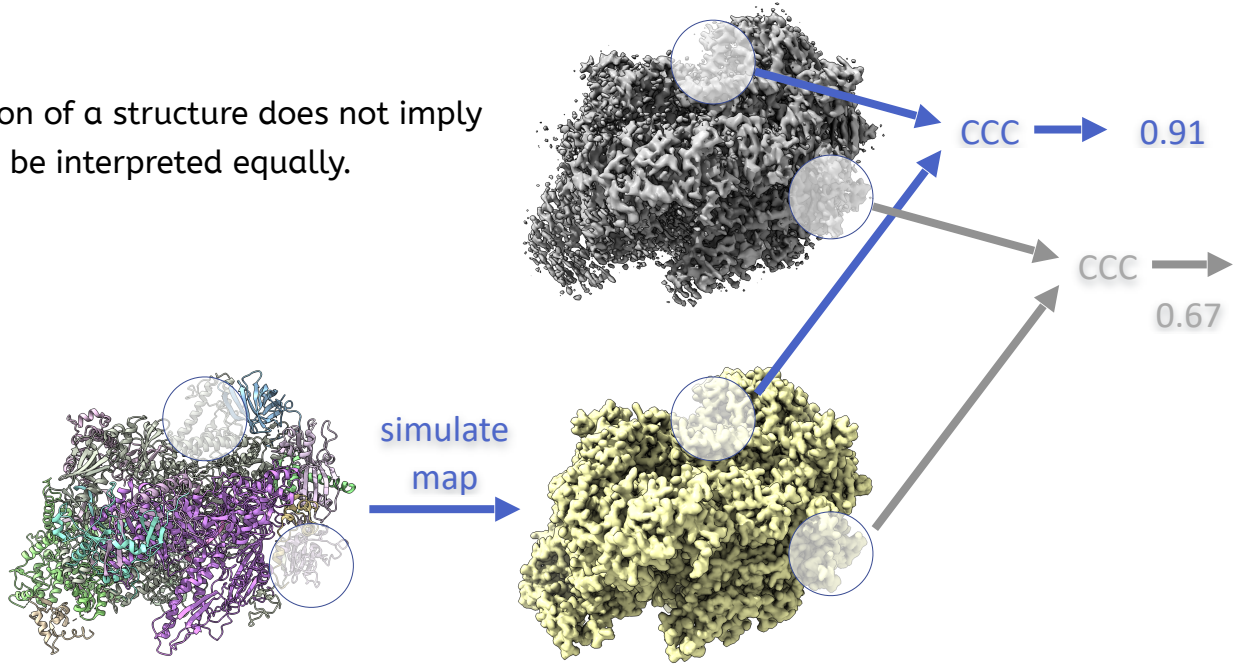


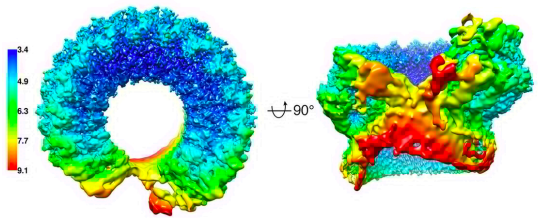


Yu et al., *EMBO J* 2022

* Local resolution maps are useful for estimating resolution variability and map quality in general.

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





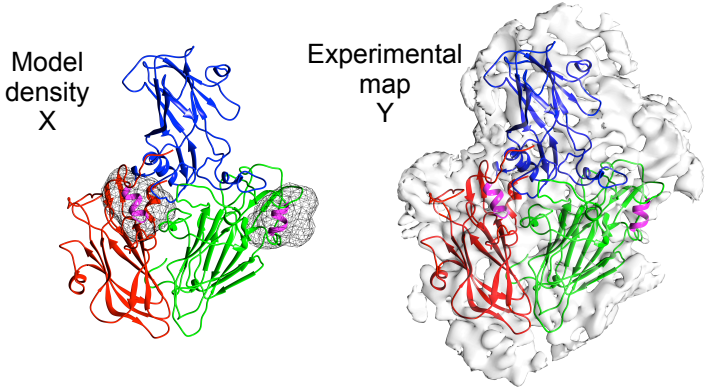
Yu et al., *EMBO J* 2022

* Local resolution maps are useful for estimating resolution variability and map quality in general.

LOCAL SCORING FOR MEDIUM-TO-HIGH RESOLUTION MAPS

SCCC: Segment-based cross-correlation coefficient:

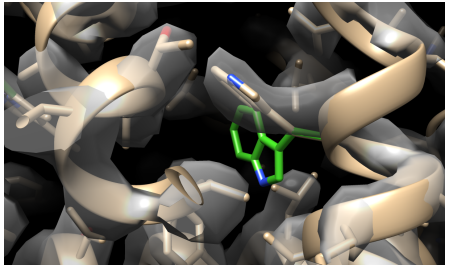
$$\text{SCCC} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$



Agnel Joseph

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

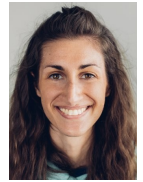
$$SMOC = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^{\,2}}\sqrt{\sum_{i=1}^{n} Y_i^{\,2}}}$$



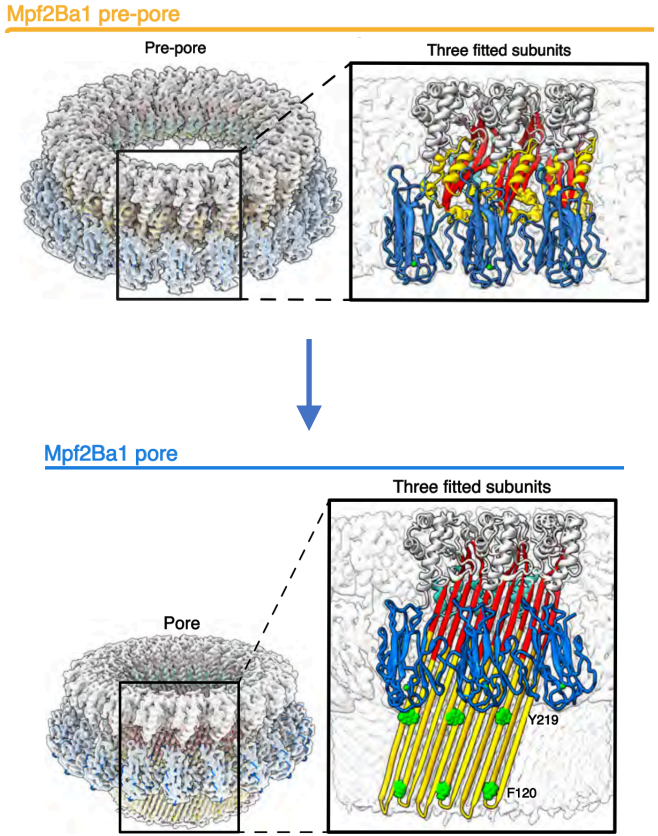
EMD-3488 (3.2Å)
Deposited model PDB: 5NI1

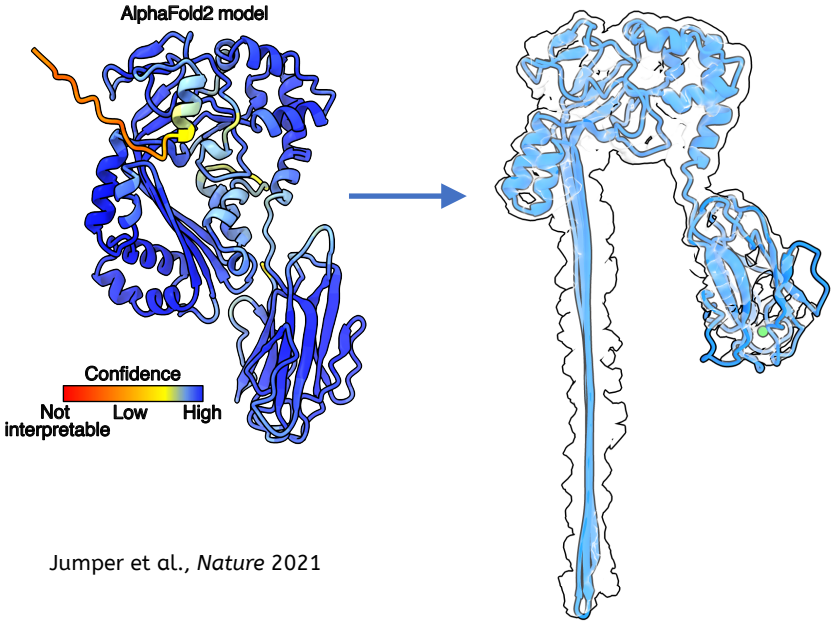
Beton et al. *WIREs Comput Mol Sci.* 2023;
Joseph et al. *Methods* 2016;
Farabella et al. *J Appl. Cryst.* 2015

ENTER THE "AI" ERA: GOOD INITIAL PREDICTIONS BUT…



Gwen Marini



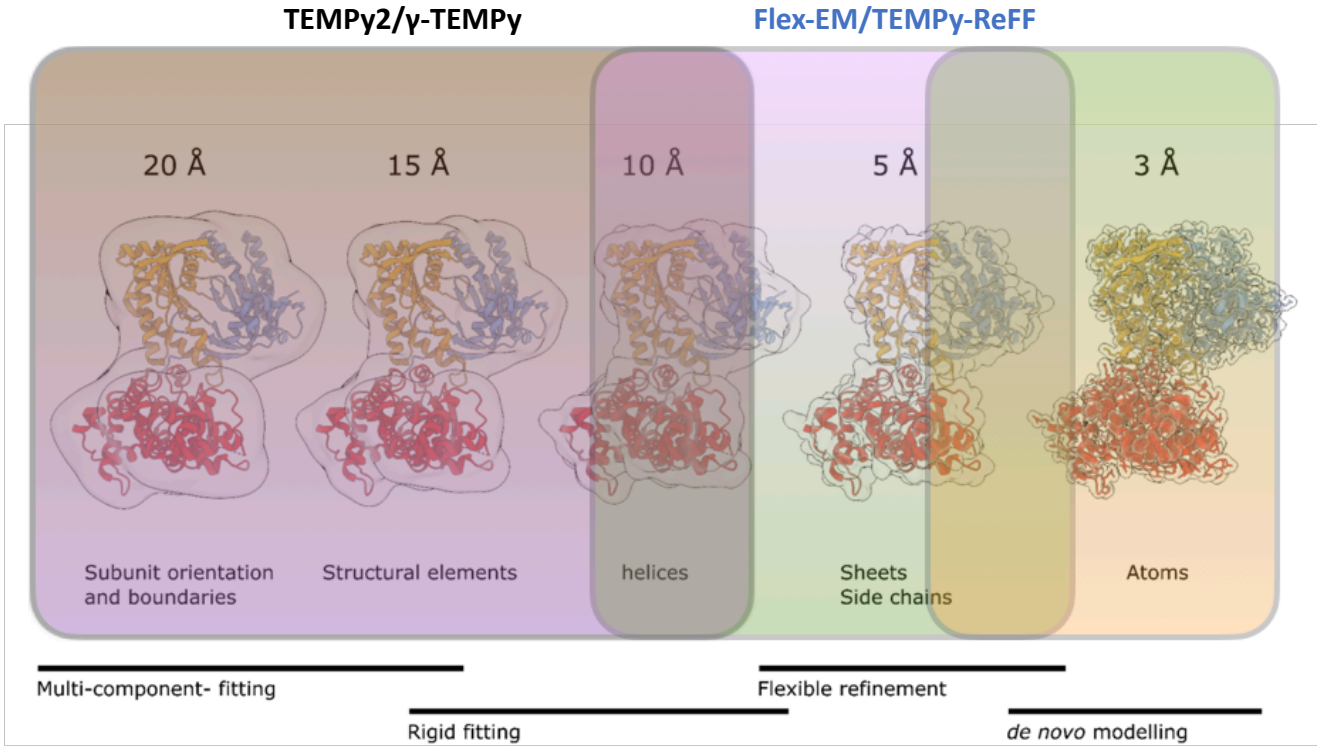


Jumper et al., *Nature* 2021

2-3 Å resolution

Marini et al., *Nat Comms* 2023

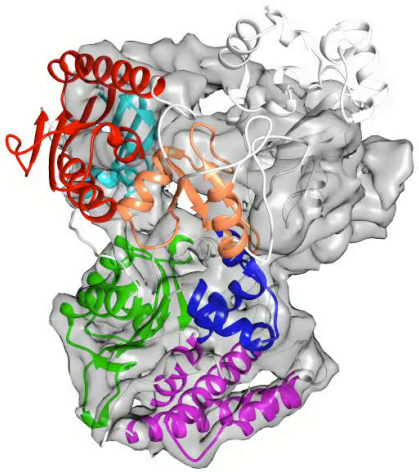
STRUCTURAL FEATURES AT DIFFERENT RESOLUTION LEVELS



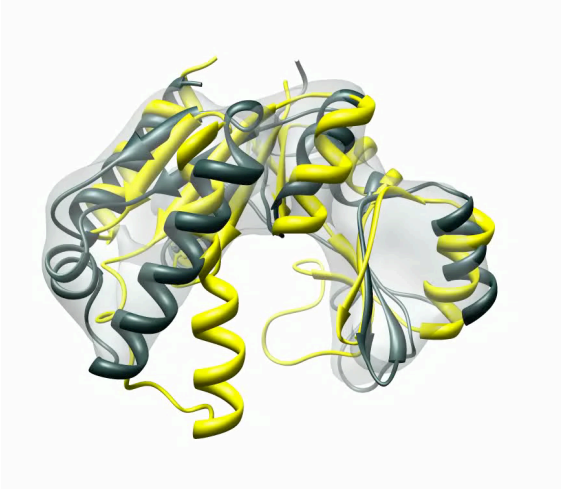
- What about higher resolutions?

Beton et al. WIREs Comput Mol Sci. 2023

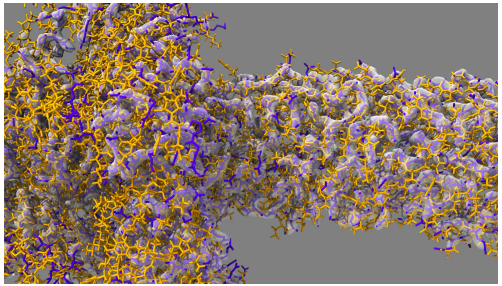
FLEXIBLE FITTING AND REFINEMENT AT LOW TO HIGH RESOLUTIONS



sub-domains

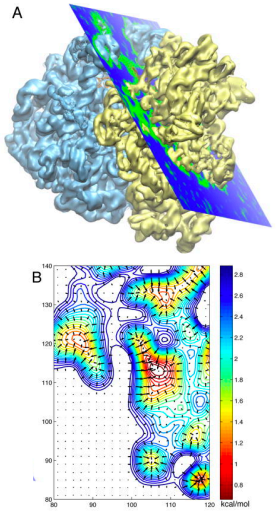


secondary structure elements



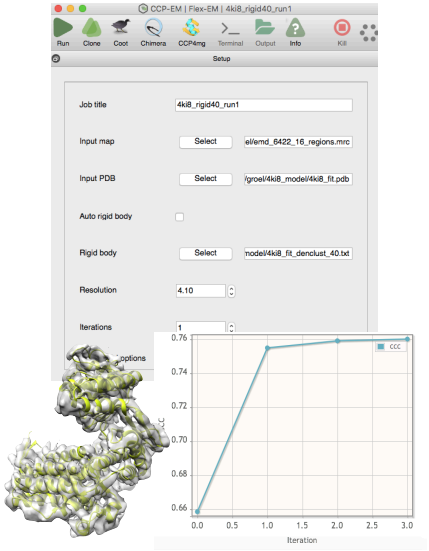
residue / all-atom

FLEXIBLE FITTING AND REFINEMENT



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

Trabuco et al., *Methods* 2008

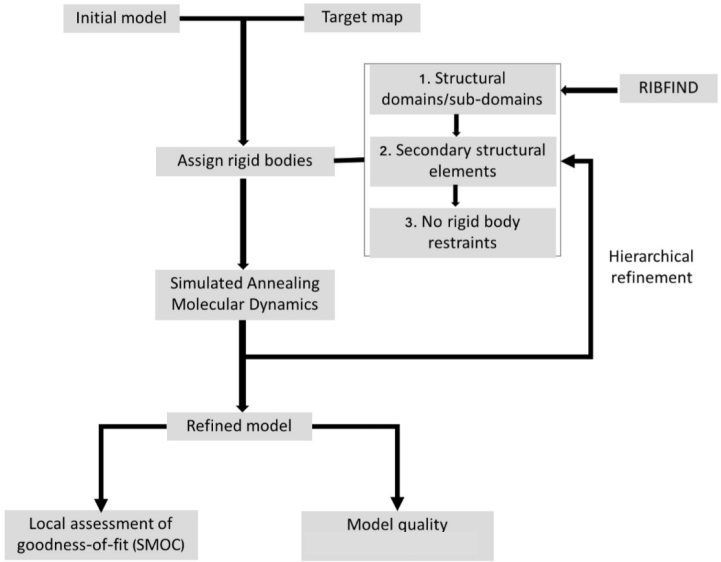


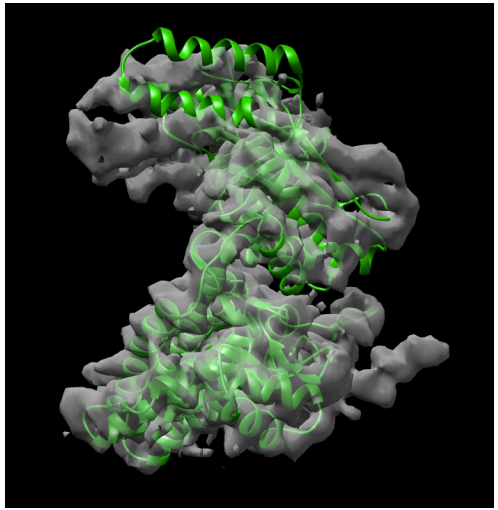
Flex-EM improves the correlation between the model sand the cryo-EM map

Topf et al., *Structure* 2008;
Joseph et al., *Methods* 2016

FLEXIBLE FITTING AT MEDIUM RESOLUTION BY HIERARCHICAL REFINEMENT

Agnel Joseph



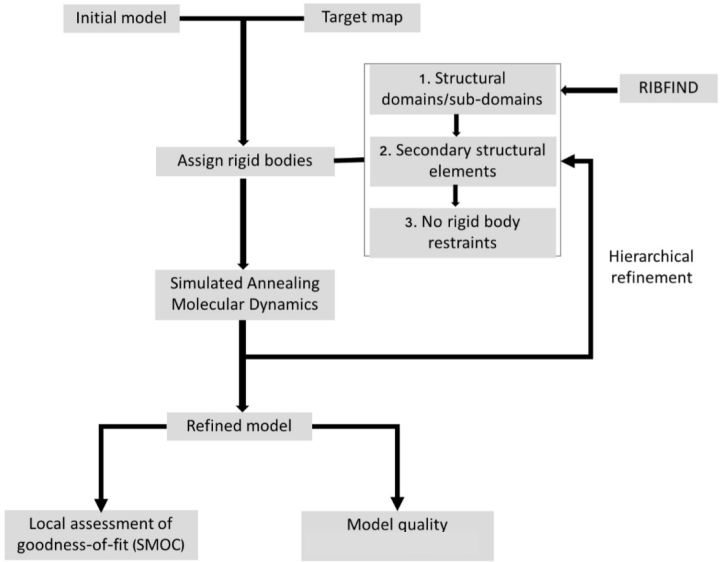


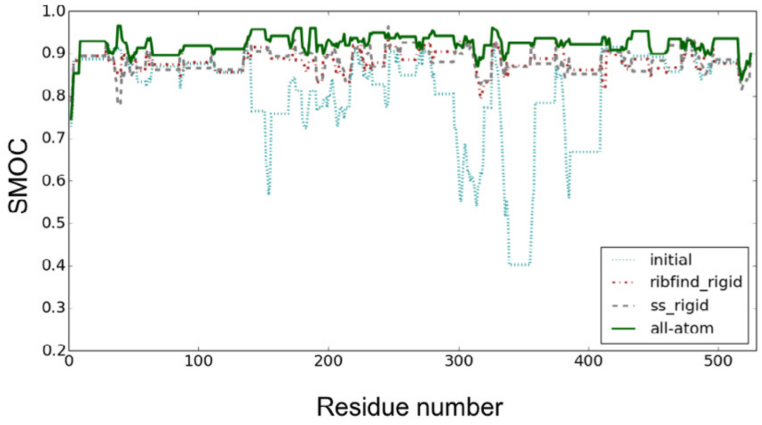
Unliganded GroEL at 4.2 Å resolution (EMD-5001)

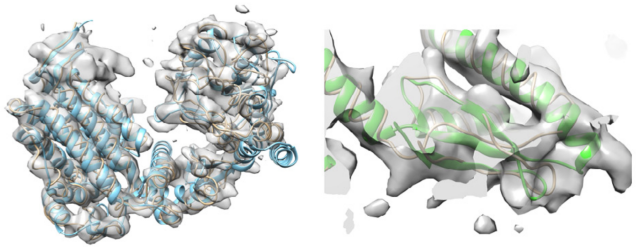
Using validation approaches, we can **iteratively** improve the fit to the density

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)

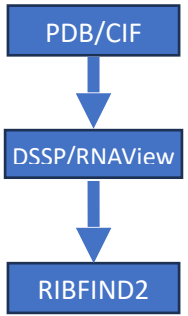
Joseph A.P. et al. *Methods* 2016

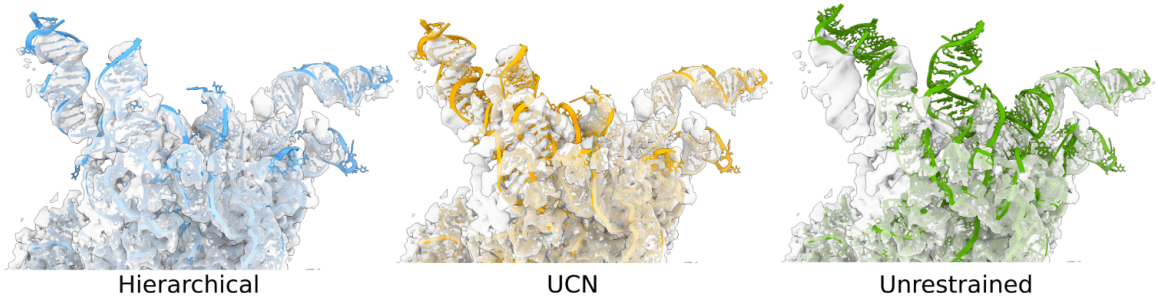
RIBFIND2 - HELPS TO AVOID GETTING "STUCK" IN LOCAL OPTIMA

Thomas Mulvaney

Sony Malhotra

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

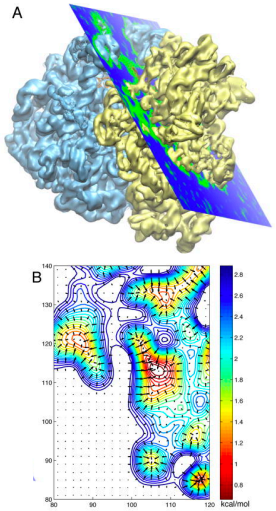




Malhotra, Mulvaney et al. *NAR* 2023;
Pandurangan *et al. J Struct Biol* 2012

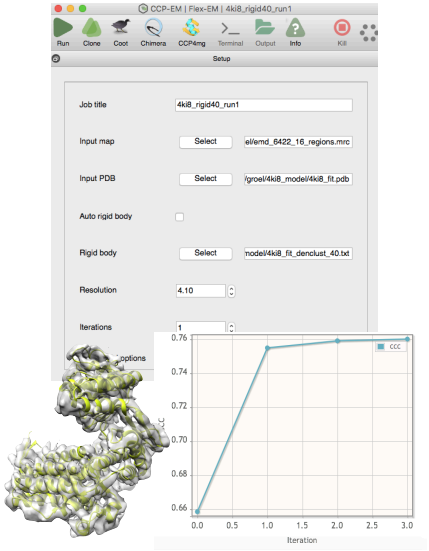
https://ribfind.topf-group.com/

FLEXIBLE FITTING AND REFINEMENT



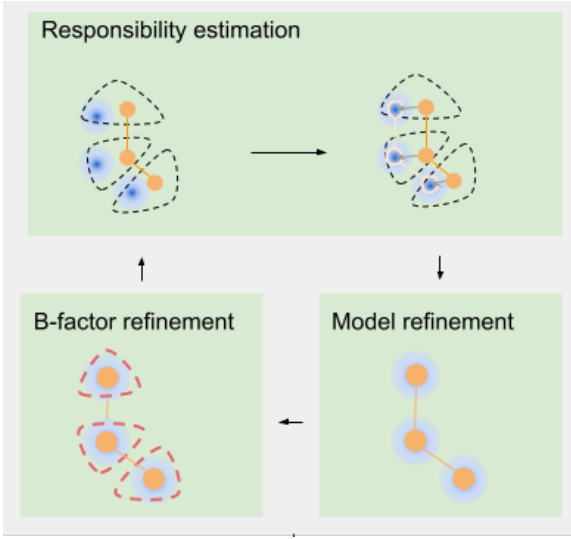
Molecule Dynamics Flexible Fitting (MDFF) uses a standard force field with a potential based on the density

Trabuco et al., *Methods* 2008



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

Topf et al., *Structure* 2008;
Joseph et al., *Methods* 2016



TEMPy-ReFF (Responsibility-based Flexible-Fitting)

Cragnolini et al., *Proteins* 2021;
Beton*, Mulvaney* et al, *Nat Comm* 2024

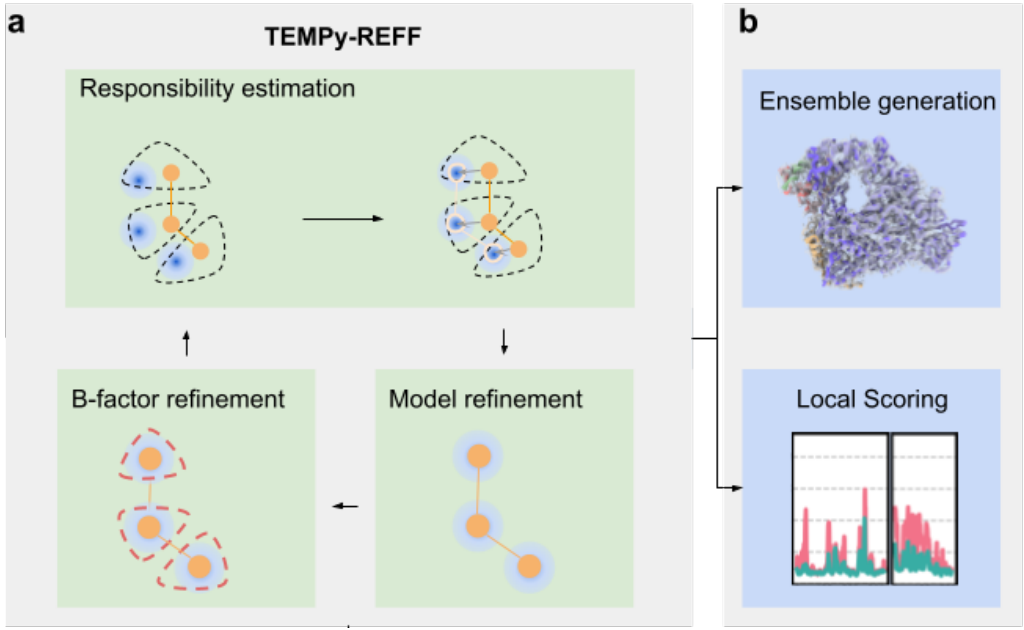
TEMPY-REFF: FLEXIBLE FITTING AND REFINEMENT

Thomas Mulvaney

Joseph Betoni

Tristan Cragnolini

- 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

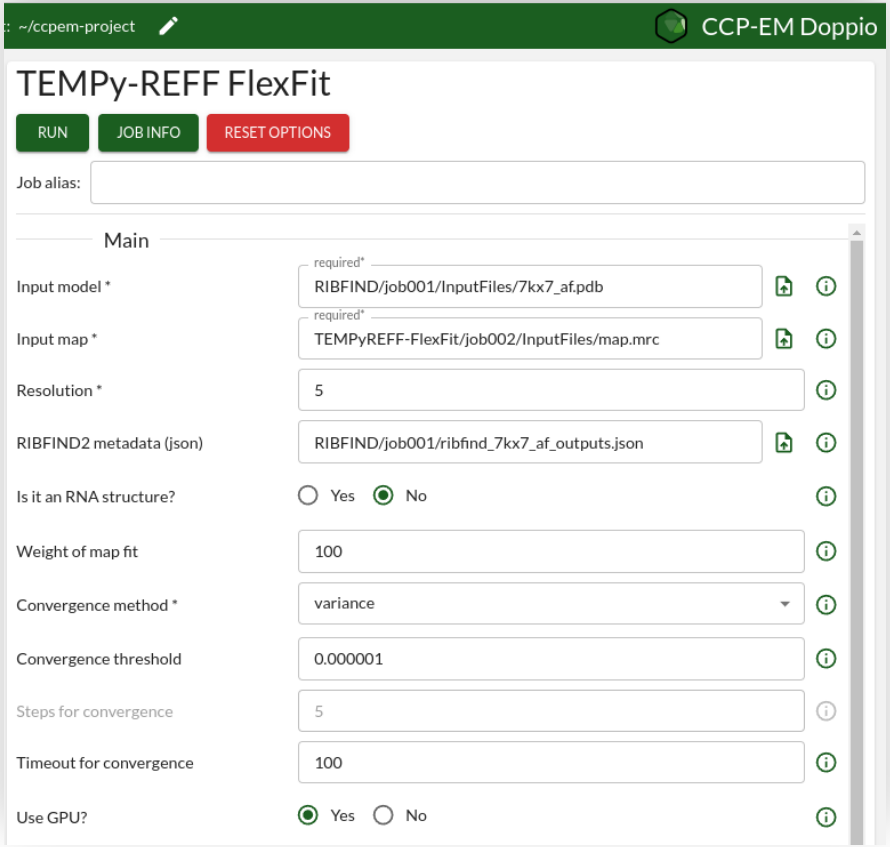
Integration with Langevin dynamics, at 100K

Beton*, Mulvaney* et al, *Nat Comm* 2024

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

TEMPY-REFF: FLEXIBLE FITTING AND REFINEMENT

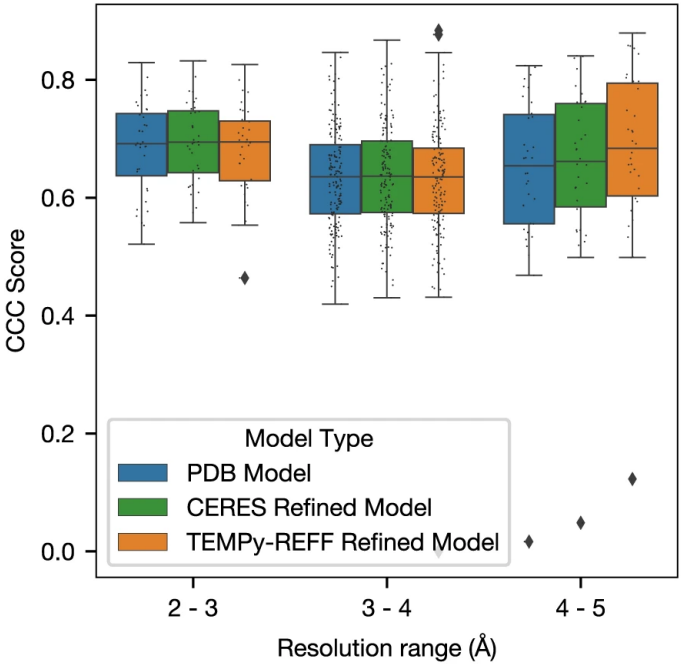
- **Flexible fitting**
  - Flexible fitting of proteins intermediate resolutions using density guided force + AMBER.
  - Faster – built on top of OpenMM, can handle much larger models
  - Hierarchical rigid body protocol using **RIBFIND2**
- **Refinement**
  - Optimises fit of GMM to data using molecular dynamics with OpenMM and AMBER.

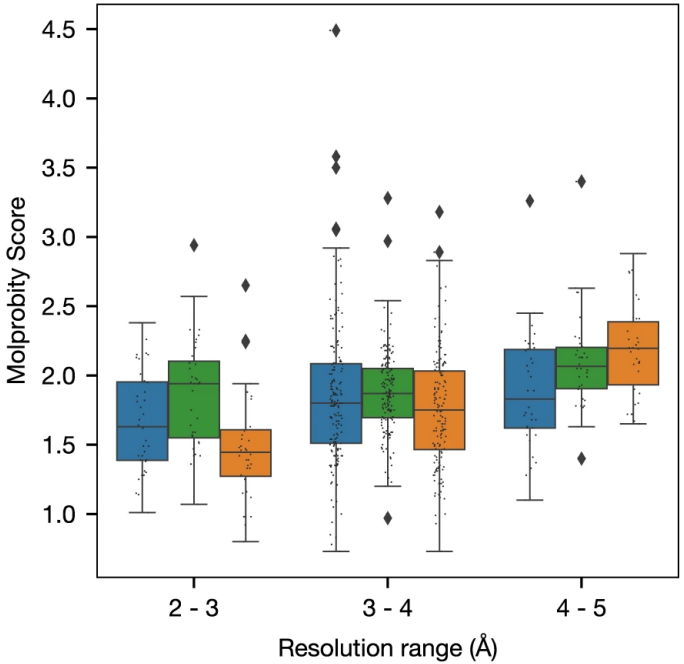


Fancy GUI in CCPEM Doppio

Beton*, Mulvaney* et al, *Nat Comm* 2024

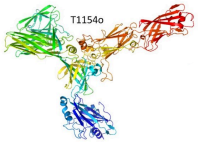
COMPARISON AGAINST 229 CERES MODELS

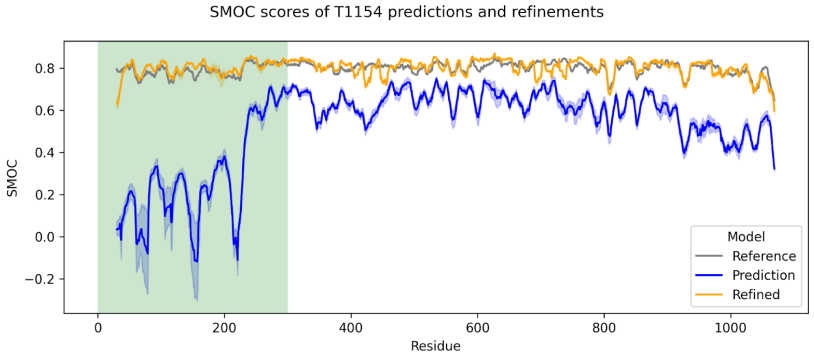


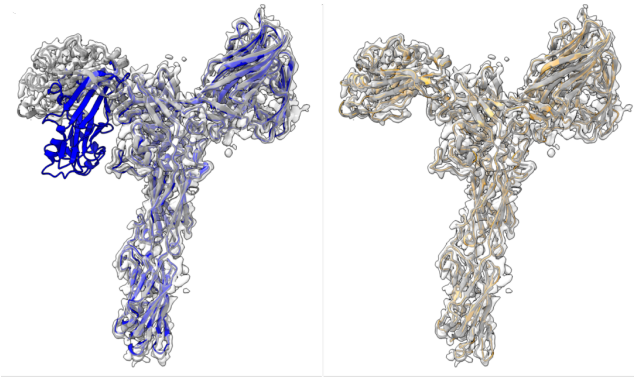


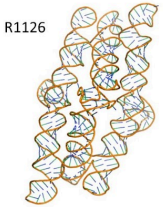
https://cci.lbl.gov/ceres

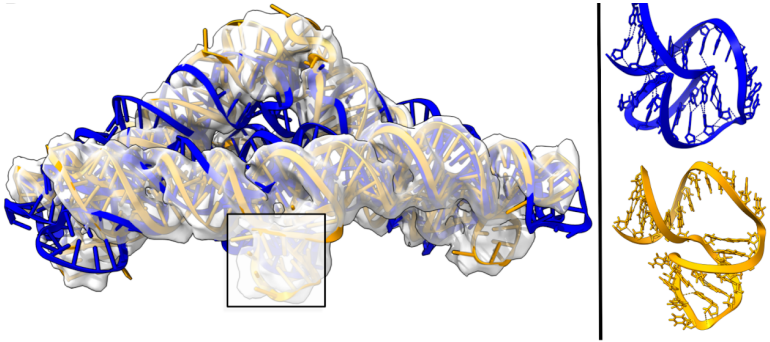
TEMPY-REFF REFINEMENT OF CASP15 MODELS (PROTEIN COMPLEXES AND RNA)







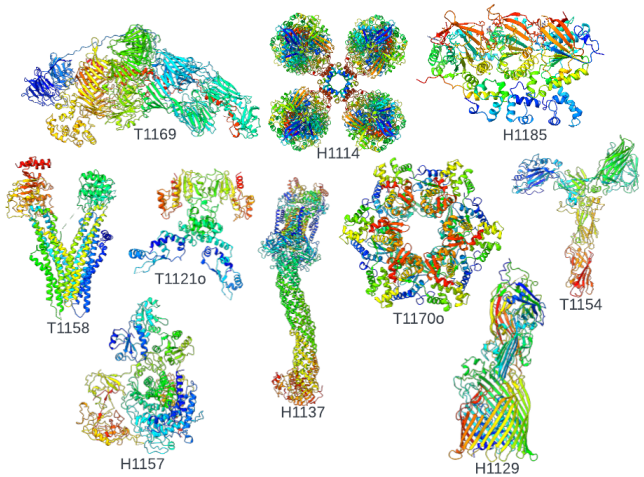


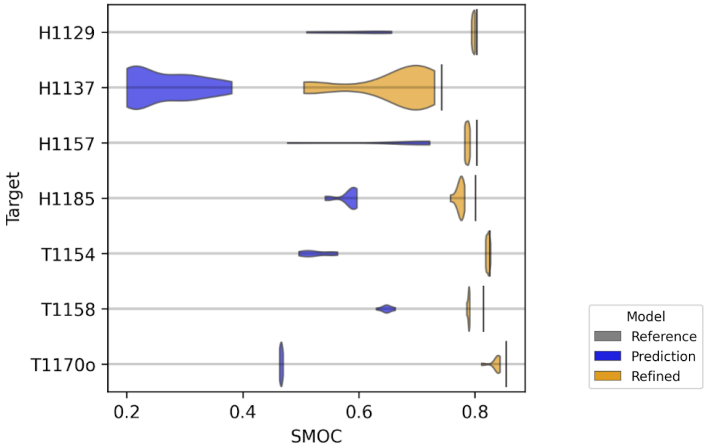


Mulvaney et al, *Proteins* 2023 (CASP15 specital edition)

TEMPy2

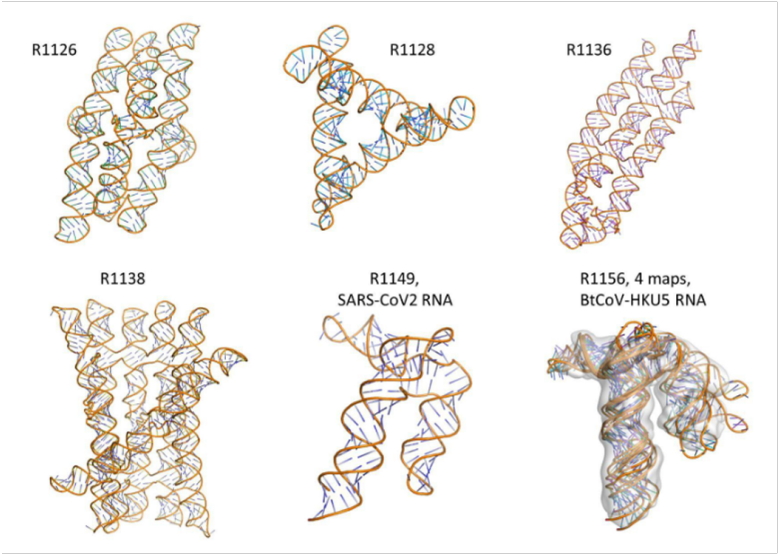
TEMPY-REFF REFINEMENT OF CASP15 MODELS (PROTEIN COMPLEXES AND RNA)

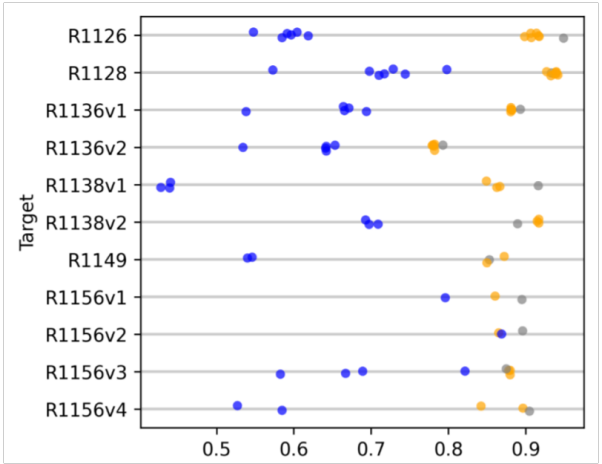




Mulvaney et al, *Proteins* 2023 (CASP15 specital edition)

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

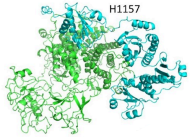


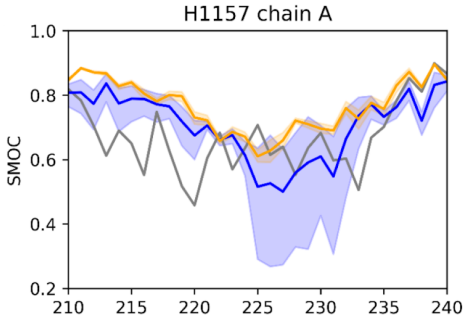


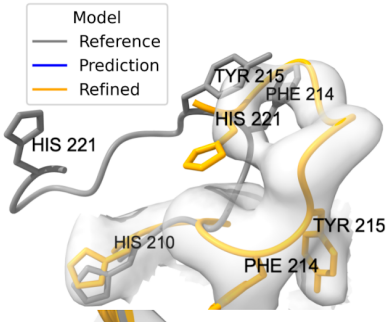
Rachael Kretsch,
Das Lab, Stanford

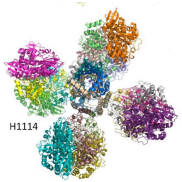
Mulvaney et al, *Proteins* 2023 (CASP15 specital edition)

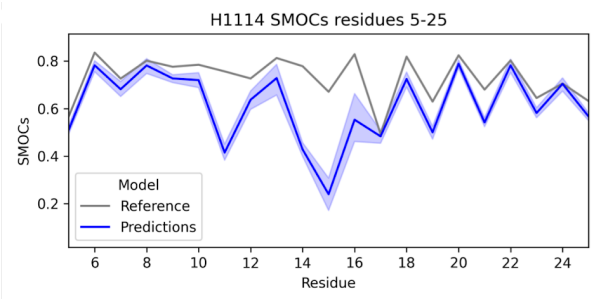
EXPERIMENTAL VALIDATION AND REFINEMENT BY CRYO-EM (CASP15)

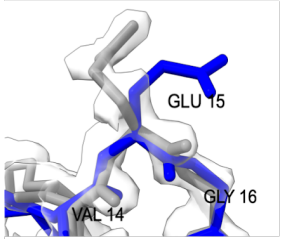


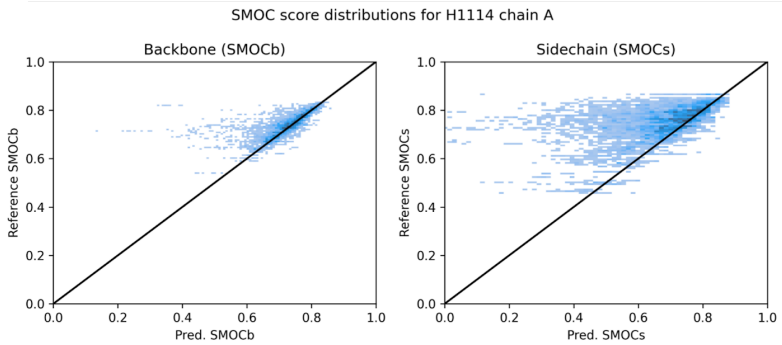




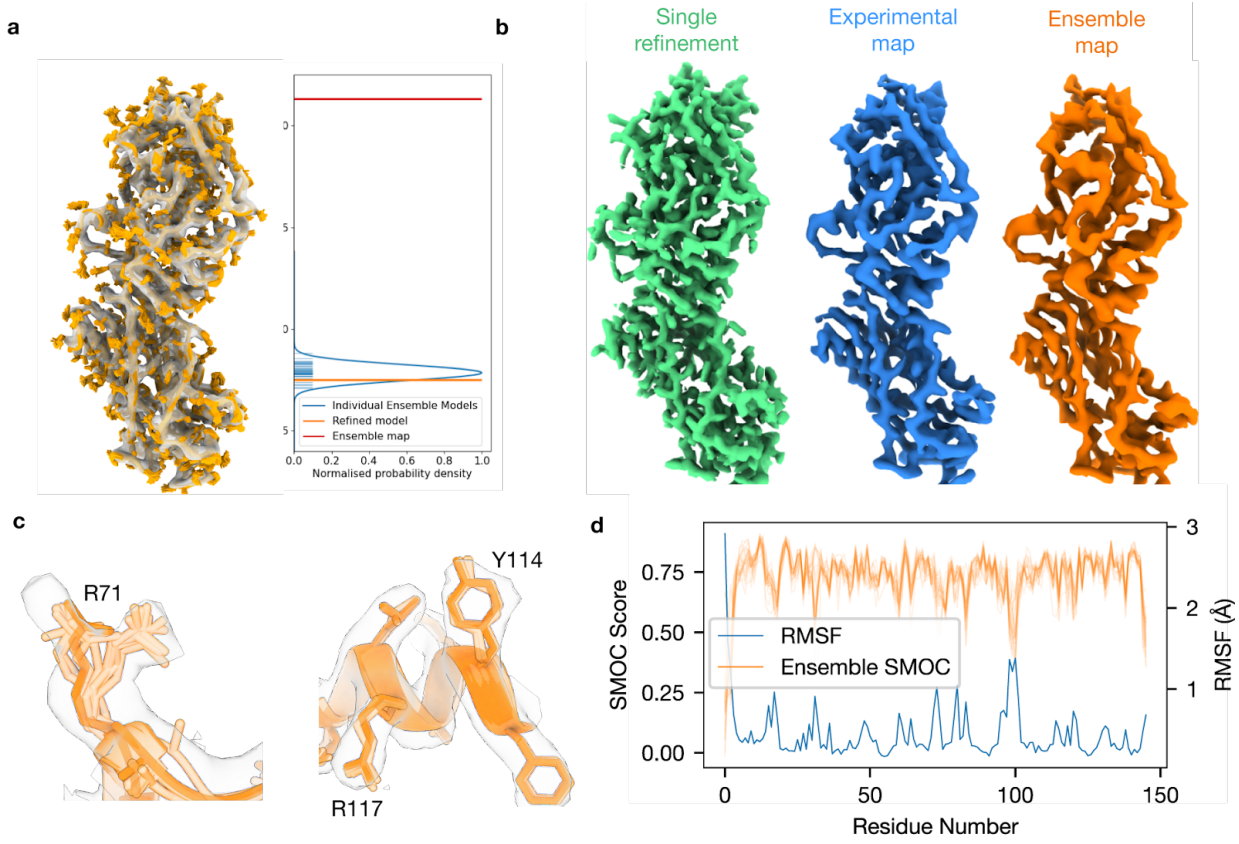




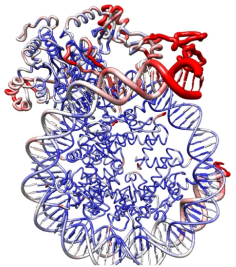




TEMPY-REFF CAN GENERATE ENSEMBLES

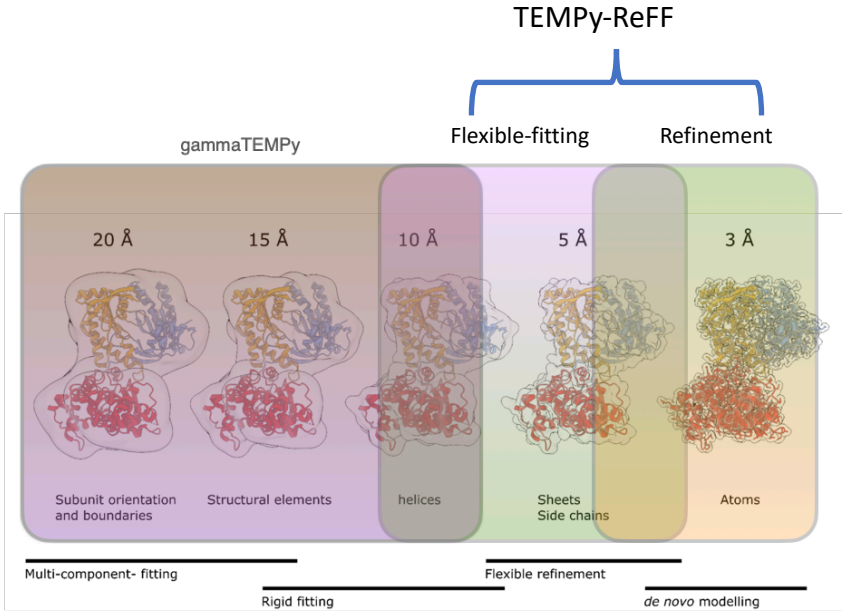


Color: TEMPy-ReFF Bfactor
Width: ResMap



Beton, Mulvaney et al. *Nat. Comm* 2024

- 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
- **TEMPy-ReFF** is useful across a range of resolutions.
- **RIBFIND2** can generate rigid bodies for Protein and RNA.
- **TEMPy-ReFF** ensembles can model uncertainty/heterogeneity.
- **CASP15** predictions were good starting points for models for flexible fitting and refinement.



- **More attention is needed for small molecules…**

ChemEM: flexible docking of small molecules in Cryo-EM structures

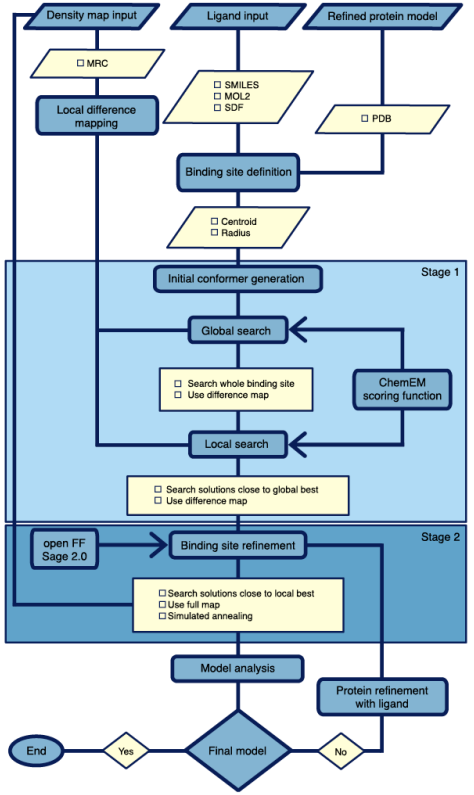
ChemEM

Uses Molecular docking, Molecular Dynamics and Density Difference mapping.

Fits ligands and refines binding site residues.

Acurate ligand fitting at 2.2-5.6 Å resolution.

Can fit multiple ligands in parralel to a single site.



Using ChemEM : Stage 1

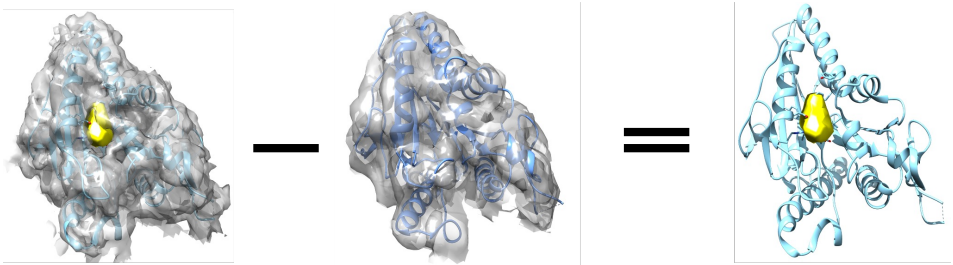
ChemEM: Ligand Placement and Refinement Workflow

- **Stage 1: Initial Ligand Docking**
  - Segment ligand density from the electron density map.
  - Perform molecular docking using ligand density as a ‘restraint’.
  - Score solutions by combining:
    - **Docking Score** for physicochemical reasonableness.
    - **Mutual Information Score** for alignment with density.

ChemDock Score: Includes terms for Bonding, Pi-Pi stacking and hydrophobic interactions

$$\begin{aligned} Docking\ Score = & Bias + w1 * \mathrm{HB}_{dist} + w2 * \mathrm{HB}_{angle} + w3 * \mathrm{Vina}_{g1} + w4 * \mathrm{Vina}_{g2} + w5 * \mathrm{Vina}_{steric} + \\ & w6 * \mathrm{LogP}_{HPI} + w7 * \mathrm{Aromatic}_{dist} + w8 * \mathrm{Aromatic}_{angle} + w9 * \mathrm{Intra}_{vg1} + w10 * \mathrm{Intra}_{vg2} + w11 * \mathrm{Intra}_{v\,steric} \end{aligned}$$

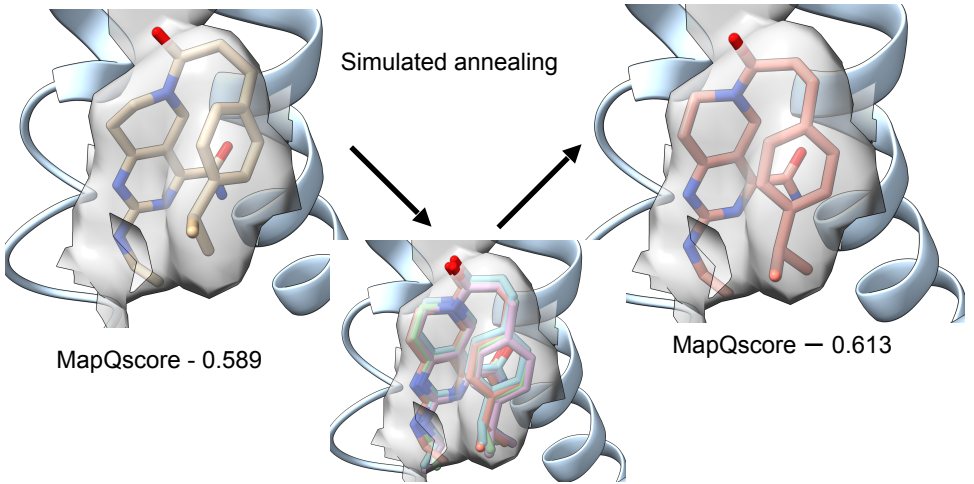
Segmenting Ligand density: Currently using a density difference map method.



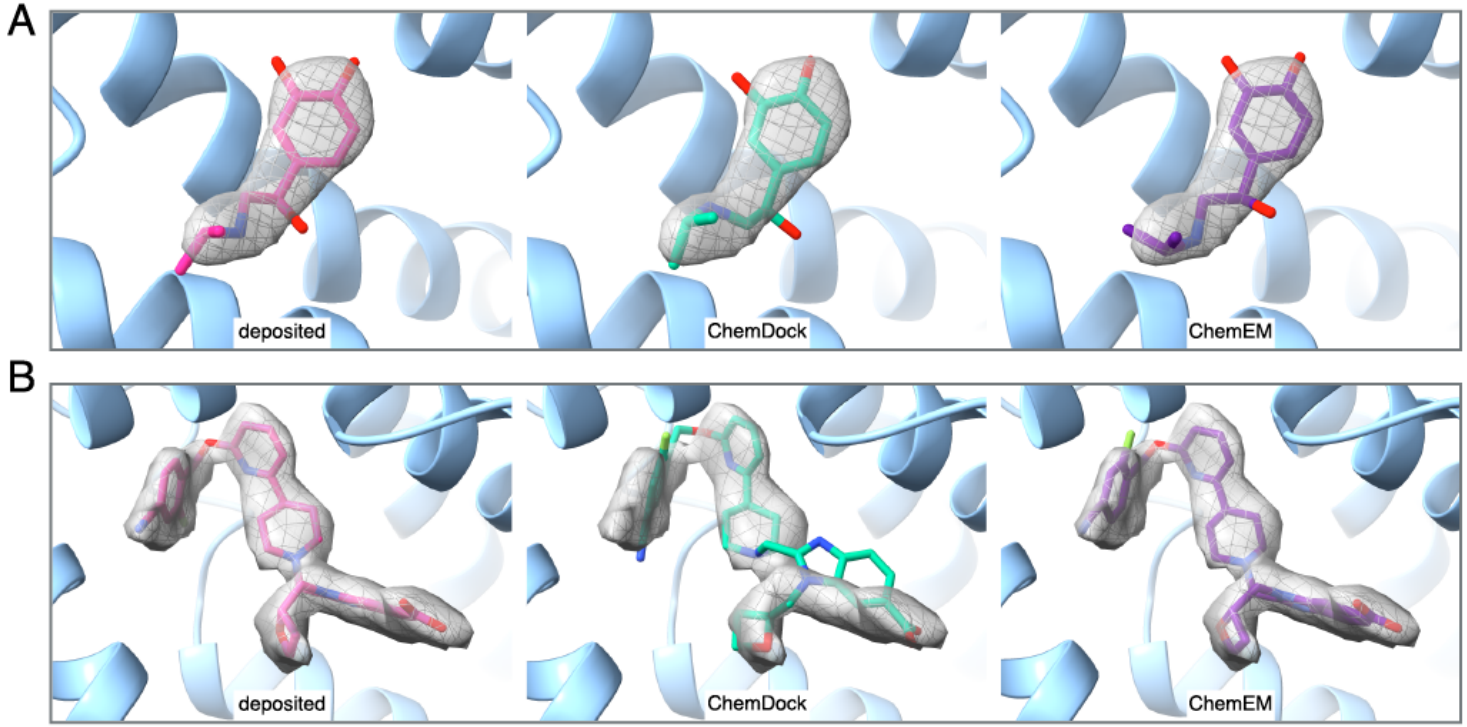
Using ChemEM : Method Overview

Stage 2: Molecular Dynamics Refinement

- Select top solutions from Stage 1.
- Refine ligand placement using molecular dynamics:
  - **Amber Parameters** for protein atoms
  - **OpenFF Parameters** for ligand molecules.



Using ChemEM : Examples



Using ChemEM : Examples

CSSB Centre for Structural Systems Biology

- ChemEM can be downloaded from
- It runs in terminal and all data is passed in from a configuration file.
- The enable stages block defines what protocols to run
- The rest defines the data and output locations
- ChemEM is also capable of fitting multiple ligand to the same site.

```
#ChemEM config file

#enable stages
pre_process = 1
pre_process_split_density = 1
auto_split_point = 1
auto_split_zone = 0
fitting = 1
dock_only = 0
post_process = 0
rescore = 0

#data and working directory
protein = ~/test_data/6tti_protein.pdb
ligand = CC(=O)Nc1cc(no1)C(F)(F)F
ligand = CC(=O)Nc1cc(no1)C(F)(F)F
ligand = CS(=O)C
ligand = CS(=O)C
densmap = ~/maps/6tti.mrc
resolution = 2.5
map_contour = 0.07
centroid = ( 43.908,  54.415,  49.755)
output = ~/multi_fit
```

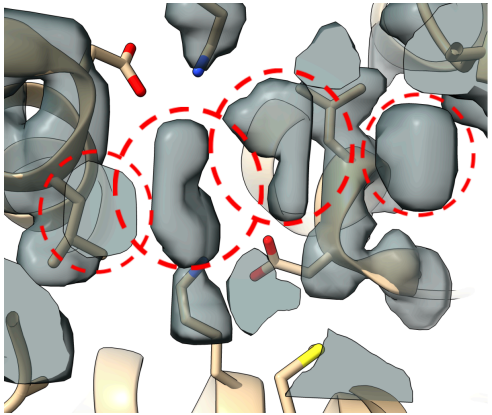
HLIV Leibniz Institute of Virology

UKE HAMBURG

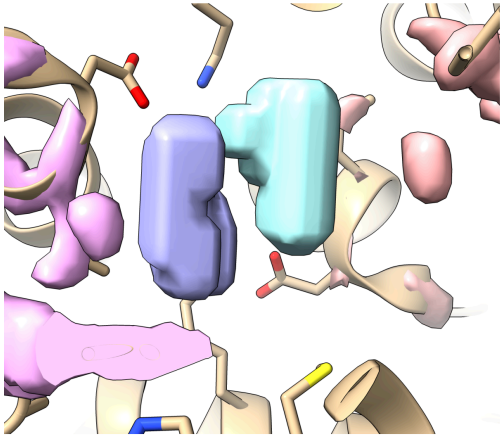
Using ChemEM : Examples

- The density is segmented by the difference mapping technique. This yields a single density map (That has a lot of extra density)

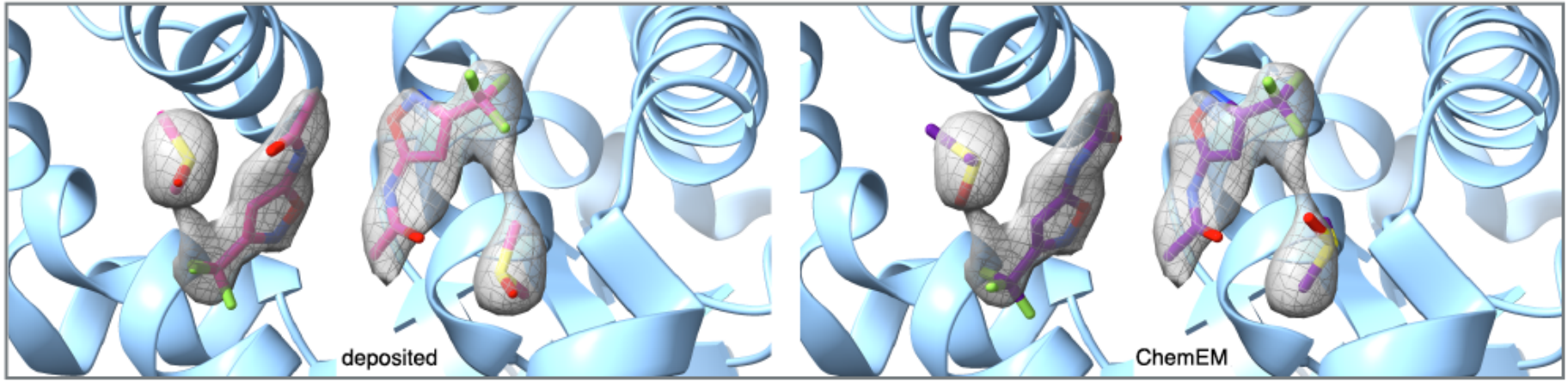
- The results can be improved and calculations sped up by specifying multiple centroid using this initial map. And assigning a ligand to each ‘density’.



```
...
centroid = ( 43.908,   54.415,   49.755)
centroid = ( 39.049,   58.586,   49.760)
centroid = ( 40.625,   62.434,   52.641)
centroid = ( 42.391,   50.558,   52.642)
...
```



Using ChemEM : Multi-ligand fitting



ChemEM:

- **Recent improvement to ChemEM include:**
  - Ions can be included to the docking calculation
  - Previously known ligand positions can be included
  - Protonation state assignment
  - Improved post-processing algorithms enable selective refinement of specific ligand atoms.
- **Coming soon:**
  - Better handling of water molecules and binding site solvation.
  - Lead optimization tools to predict the impact of R-group modifications on ligand affinity.

THANKS…

Topf group (current)

Sanjana Nair
Thomas Mulvaney
Karen Manalastas
Luca Genz
Aaron Sweeney
Laetitia Adeler-Ohde
Matthias Pfeifer
Birgit Märtens
Mauro Maiorca
Guendalina Marini
Annika Rammelt
Pasquale Lamagna
Natan Nagar
Fabian Hausmann
Gabriele Diana

Previous members:

Sony Malhotra (STFC)
Joseph Beton (Biontech)
Agnel Joseph (STFC)
Tristan Cragnolini (Birkbeck)

CCP-EM (STFC, UK)

Tom Burnley
Agnel Joseph
Sony Malhotra
Martyn Winn

DAS group (Stanford,CA)

Rhiju Das
Rachael Kretsch

Birkbeck (UK)

Helen Saibil

CASP (UC Davis, CA)

Andriy Kryshtafovych

UKE HAMBURG
LIV
CSSB Centre for Structural Systems Biology
UCL
Birkbeck UNIVERSITY OF LONDON
wellcome trust
Institute of Structural and Molecular Biology ISMB
InterACt Leibniz ScienceCampus
CCP-EM
Leibniz Association
Bundesministerium für Bildung und Forschung
CRC 1648
DFG
Hamburg-X