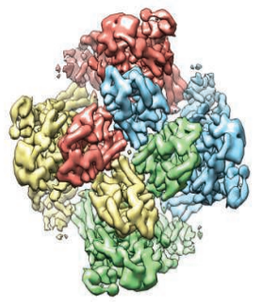
Fitting challenges

David Veesler
University of Washington
Department of Biochemistry
dveesler@uw.edu
https://twitter.com/veeslerlab

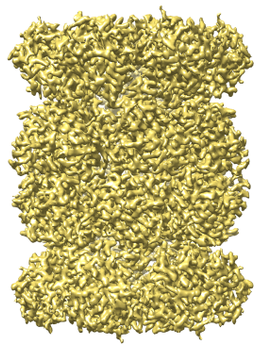
Outline

[Methods for backbone tracing into moderate resolution structures. What are the problems, how are they approached, what are the solutions? How do we validate the methods and the results?]

Single particle cryoEM at near-atomic resolution



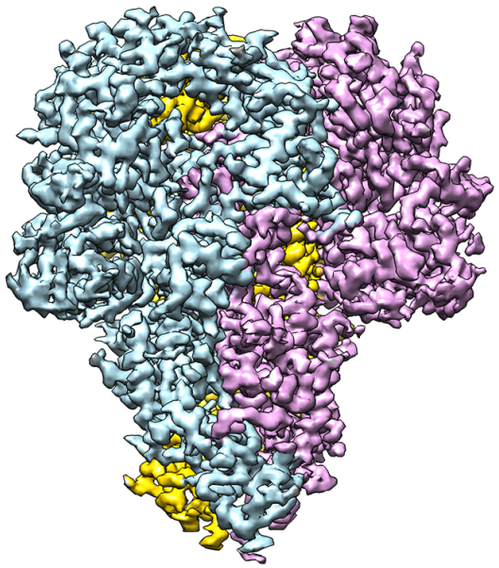
TRPV1
3.4 Å
Liao *et al*, Nature 2013



T20S
2.8 Å
Campbell *et al*, eLIFE 2015

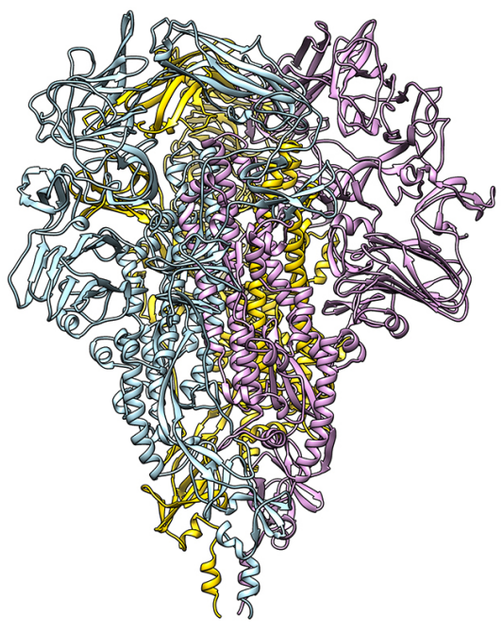
GDH
1.8 Å
Merk *et al*, Cell 2016

Main goal of structure determination



+ MYPRQTEINSEQVENCE

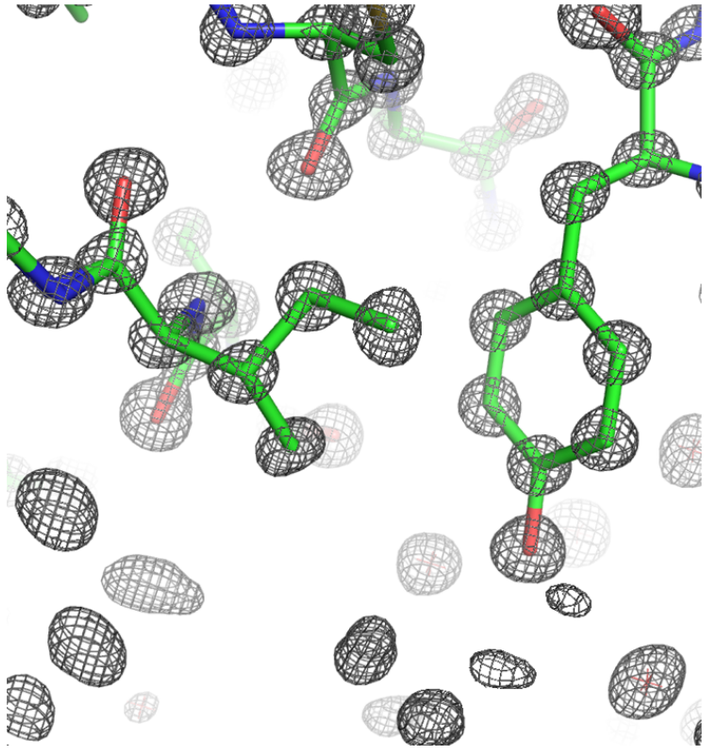
=



Goals of structure determination

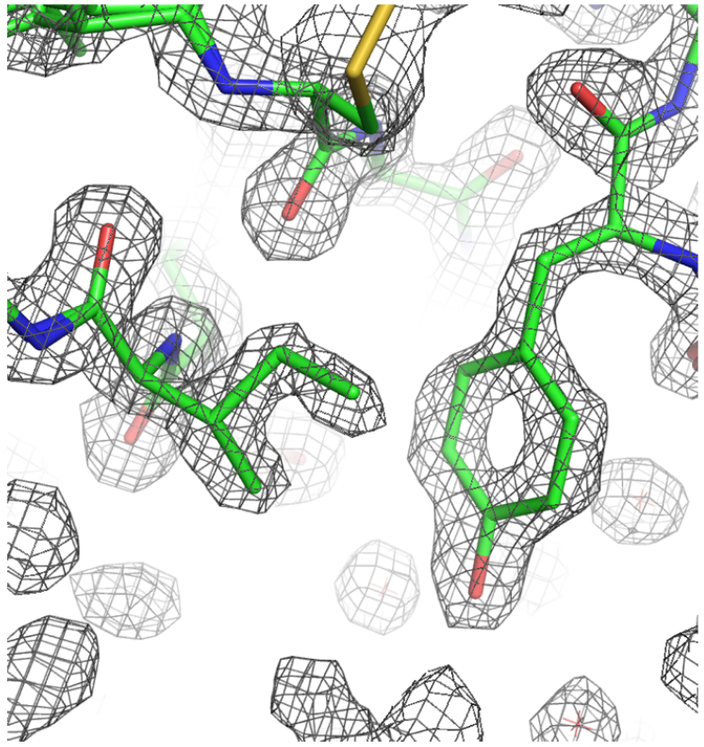
- Determining the position of all the atoms in a structure
- Model refinement aims to:
  - Identify and correct errors
  - Improve fit to data
  - Improve model geometry
- Each atom is described by four parameters:
  - cartesian coordinates (x,y,z) & temperature factor (B)
- Data/parameter ratio >>1 is sought in X-ray crystallography
- Use the structure to infer biological/biochemical properties of the protein studied

Determining structures using “intermediate” resolution data requires more “prior knowledge”



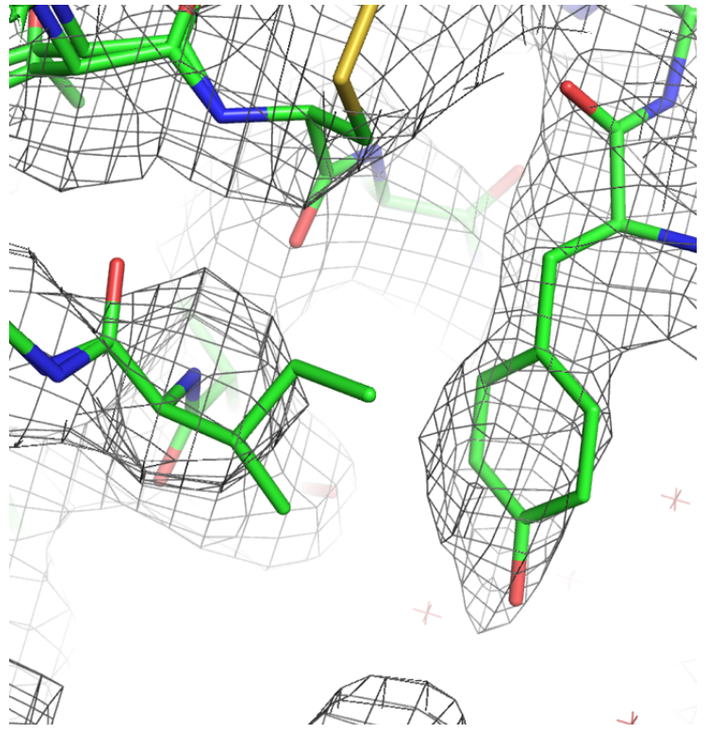
Resolution: 0.5 Å

Atom connectivity



Resolution: 2.0 Å

Bond lengths
Bond angles
Sterics



Resolution: 4.0 Å

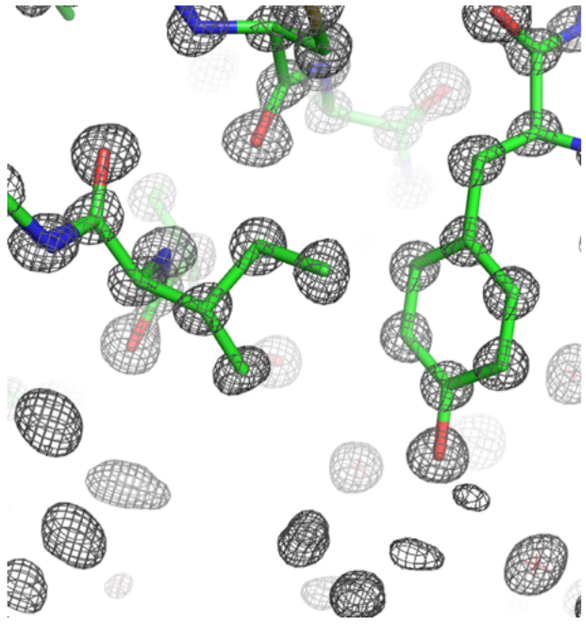
Bond lengths
Bond angles
Sterics
Torsion distributions
Solvation
Hydrogen bonding
Electrostatics

Determining structures using “intermediate” resolution data requires more “prior knowledge”

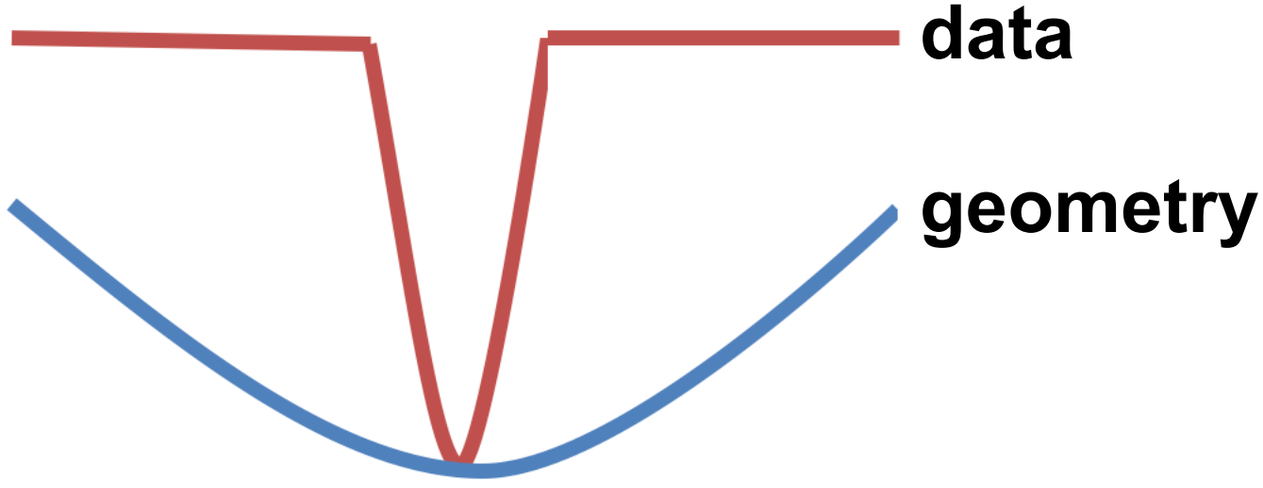
Refinement finds atom position minimizing

$$E(\text{total}) = E(\text{geometry}) + w \cdot E(\text{data})$$

Atom connectivity



Resolution: 0.5 Å

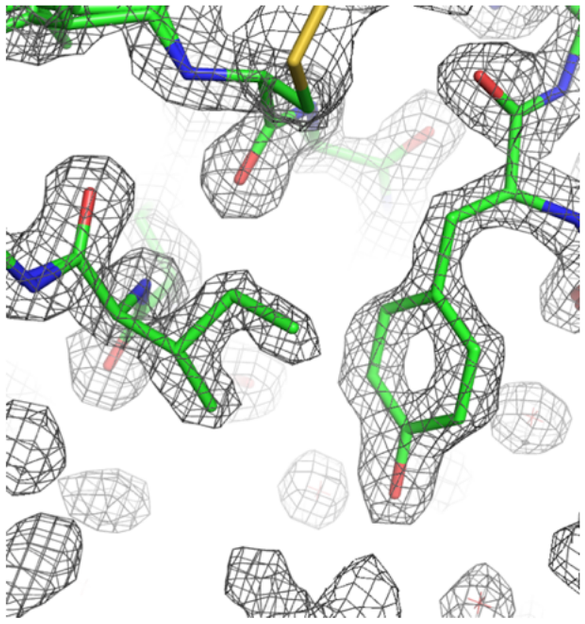


Determining structures using “intermediate” resolution data requires more “prior knowledge”

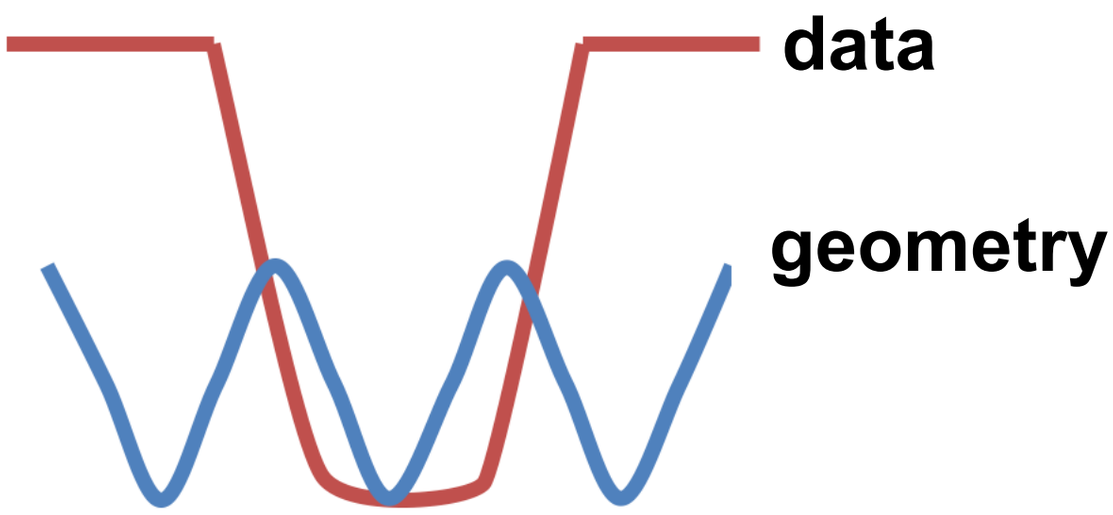
Refinement finds atom position minimizing

$$E(\text{total}) = E(\text{geometry}) + w \cdot E(\text{data})$$

Bond lengths
Bond angles
Sterics



Resolution: 2.0 Å



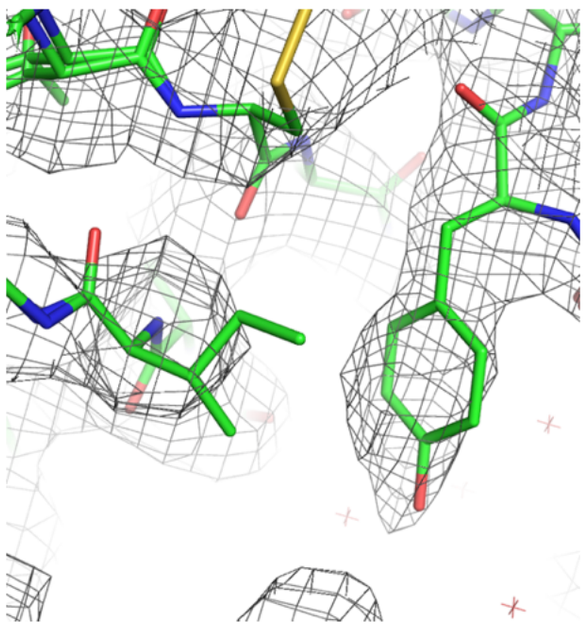
Determining structures using “intermediate” resolution data requires more “prior knowledge”

Refinement finds atom position minimizing

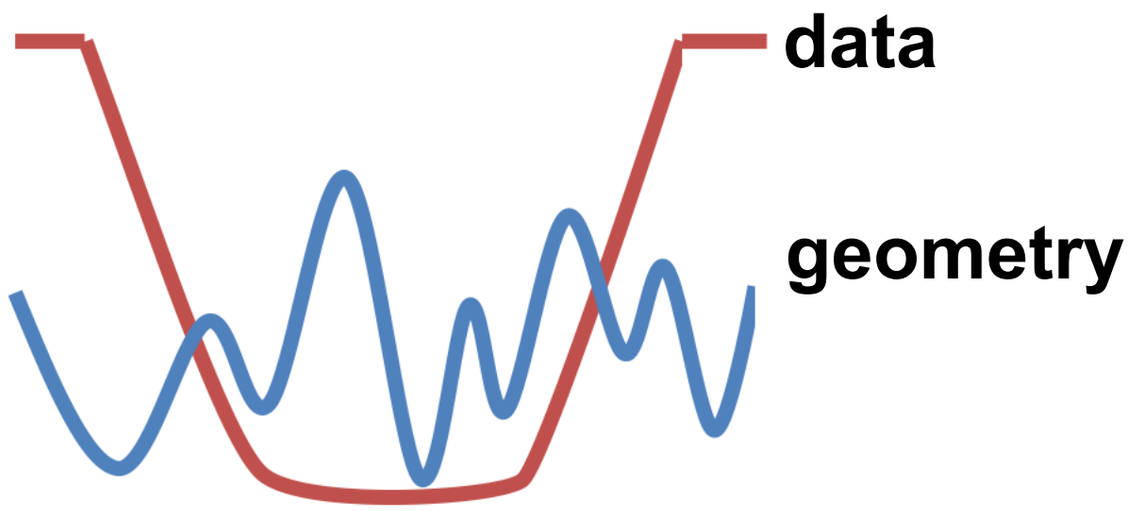
$$E(\text{total}) = E(\text{geometry}) + w \cdot E(\text{data})$$

Bond lengths
Bond angles
Sterics
Torsion distributions
Solvation
Hydrogen bonding
Electrostatics

=Rosetta Energy function



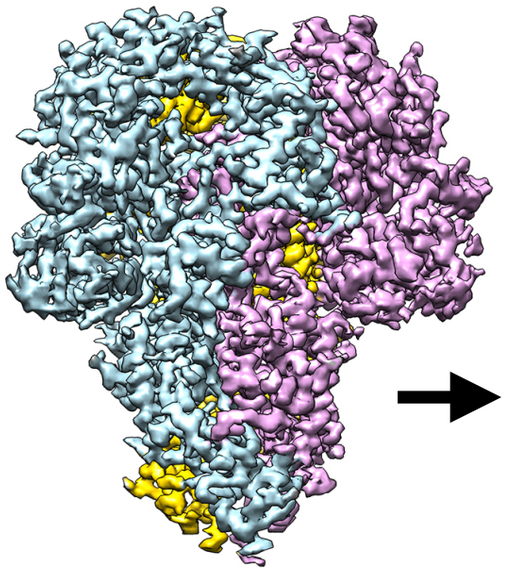
Resolution: 4.0 Å

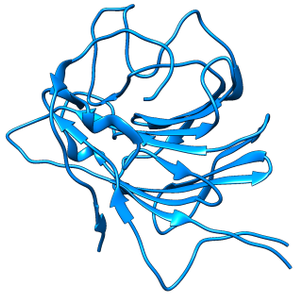


Software for model building

- UCSF Chimera (rigid-body docking of structures into EM maps)
  - Visualizing “overall” cryoEM reconstructions
- Coot (rigid-body docking, manual model building, real-space refinement)
  - Visualizing “local” segment of cryoEM reconstructions
- Refmac (refinement of atomic coordinates)
- Phenix (model building “map_to_model” derived from resolve & refinement of atomic coordinates)
- Buccaneer (model building)
- PathWalker(model building)
- DireX (flexible fitting/refinement of atomic coordinates)
- MDFF (flexible fitting/refinement of atomic coordinates)
- Rosetta (model building & refinement of atomic coordinates)
- …

Rosetta structure determination pipeline



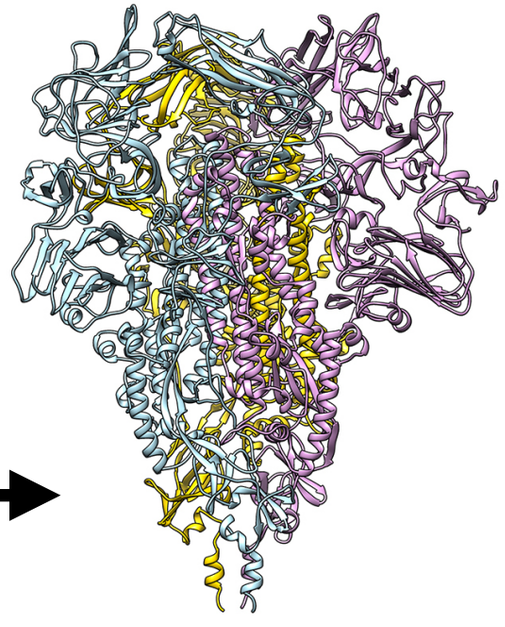


De novo model building

Homology modeling

Model completion

All-atom refinement



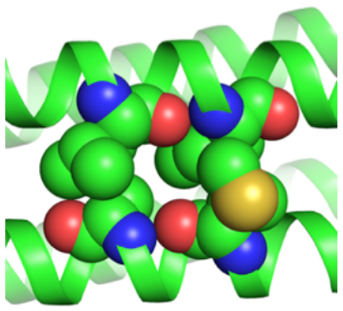
Frank DiMaio

DiMaio F et al (2015) Nature Methods
Wang RYR et al (2015) Nature Methods
Wang RYR et al (2016) eLIFE
Frenz B et al (2017) Nature Methods
Song Y et al (2013) Structure

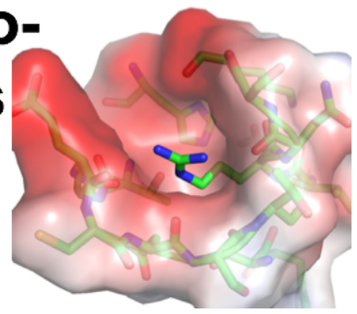
Use resolution of high-resolution structures to extend structure determination at lower resolution

Energy function

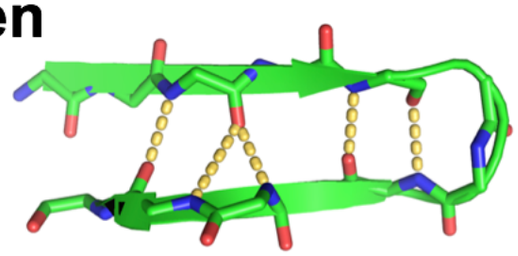
Core packing



Electro-statics

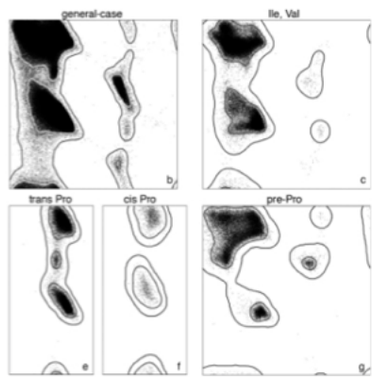


Hydrogen bonding

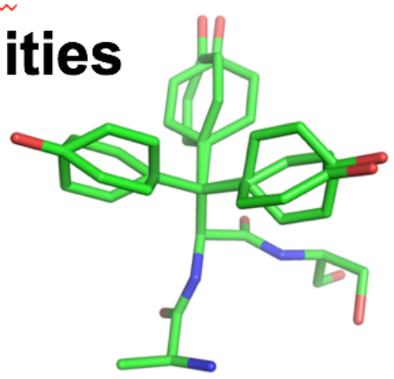


Hydrophobic Effect

Torsional probabilities



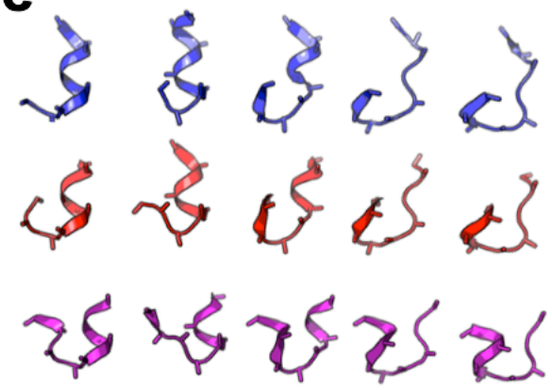
Rotamer probabilities



Sampling

Local sequence restricts local structure

...VTKPLVARAKL...



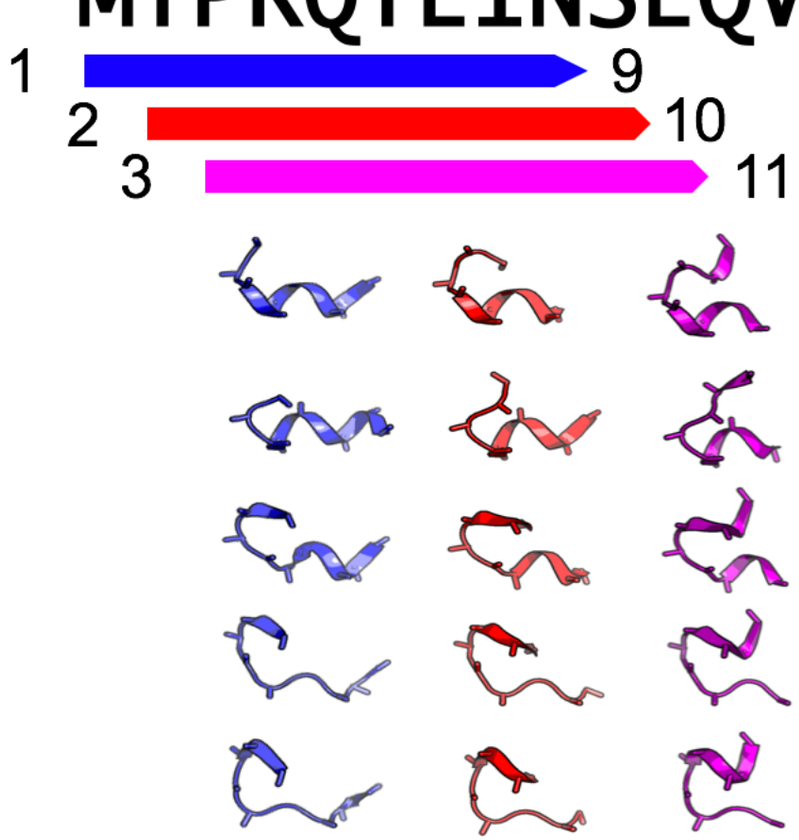
De novo structure determination

- Obtaining a structure with no other information than the sequence of the protein(s) and a cryoEM reconstruction
- Challenging at 3-5 Å resolution
- Hand tracing is time consuming
- Hand tracing is sometimes not possible
- Why do crystallographic model building softwares fail?
  - Usually work well at resolution ≤3 Å
  - Assign sequence primarily using side chain density

The Rosetta *de novo* building strategy

Take advantage of knowledge-based sequence/structure information:

Local sequence confers conformational preferences

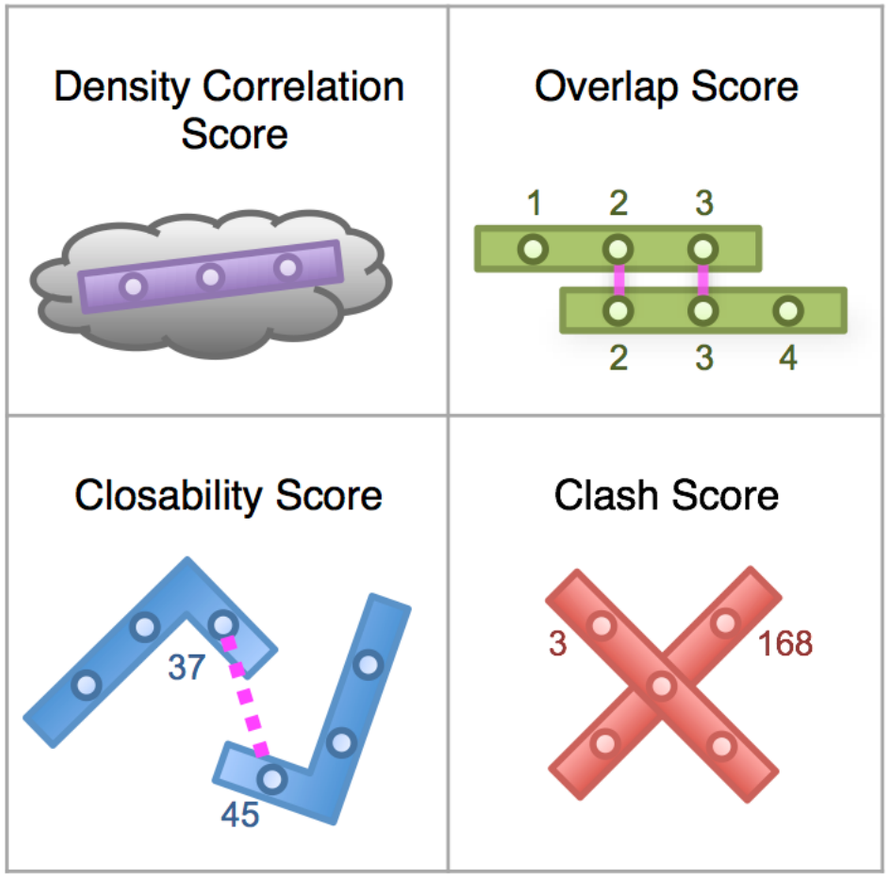


For <u>9-residue windows</u> centered at each residue position, representative backbone conformations (<u>fragments</u>) are predicted using local sequence.

Wang RYR et al (2015) Nature Methods

The Rosetta *de novo* building strategy

Use fragment consistency to select correct fragments



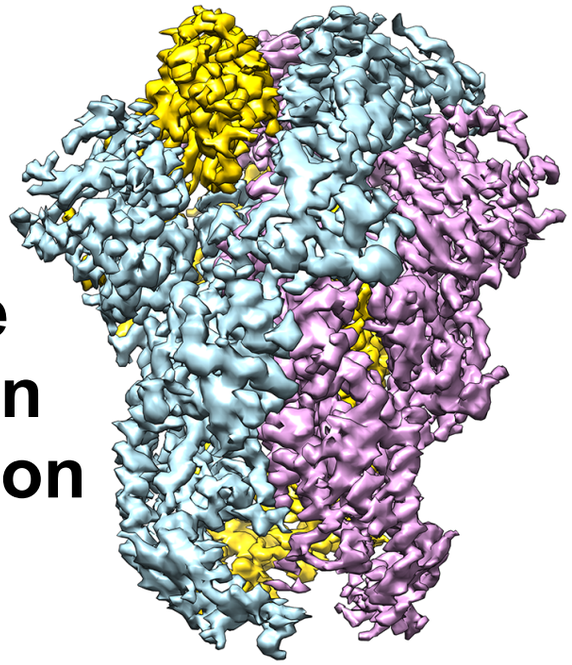
$$\text{score}_{\text{total}}(\boldsymbol{F}) = w_{\text{dens}} \sum_{f_i \in \boldsymbol{F}} \text{score}_{\text{dens}}(f_i) + w_{\text{overlap}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{overlap}}(f_i, f_j) + w_{\text{close}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{close}}(f_i, f_j) + w_{\text{clash}} \sum_{f_i, f_j \in \boldsymbol{F}} \text{score}_{\text{clash}}(f_i, f_j)$$

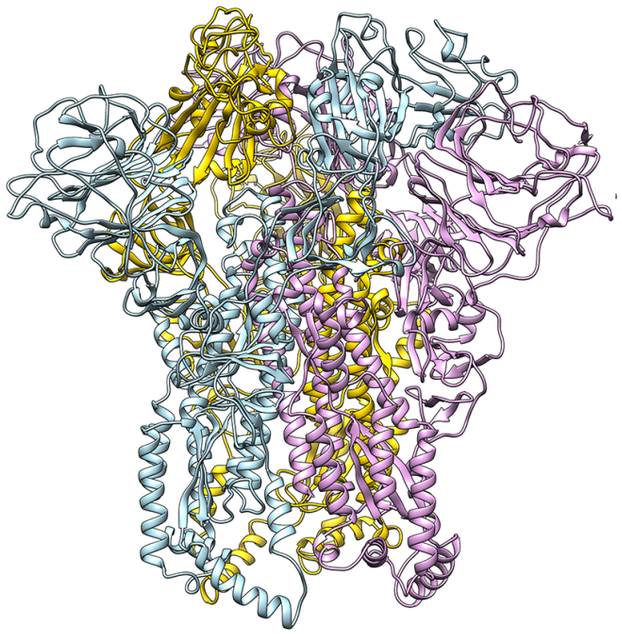
- **Near-native fragments match the density well and are consistent with one another:**
  - **Non-clashing**
  - **Assign the same residue to the same position**
  - **Nearby residue in sequence are nearby in cartesian space**
- **Score function evaluates the consistency of a set of fragments**
- **Monte-Carlo sampling finds fragment set optimizing the score function.**
- **Works up to 4.8Å resolution**

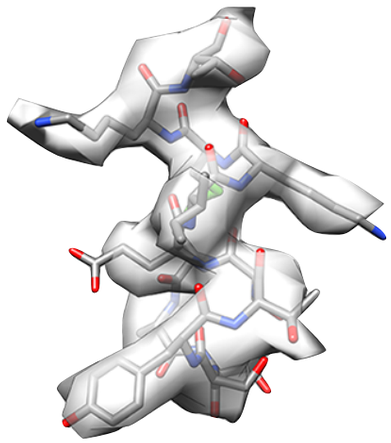
Wang RYR et al (2015) Nature Methods

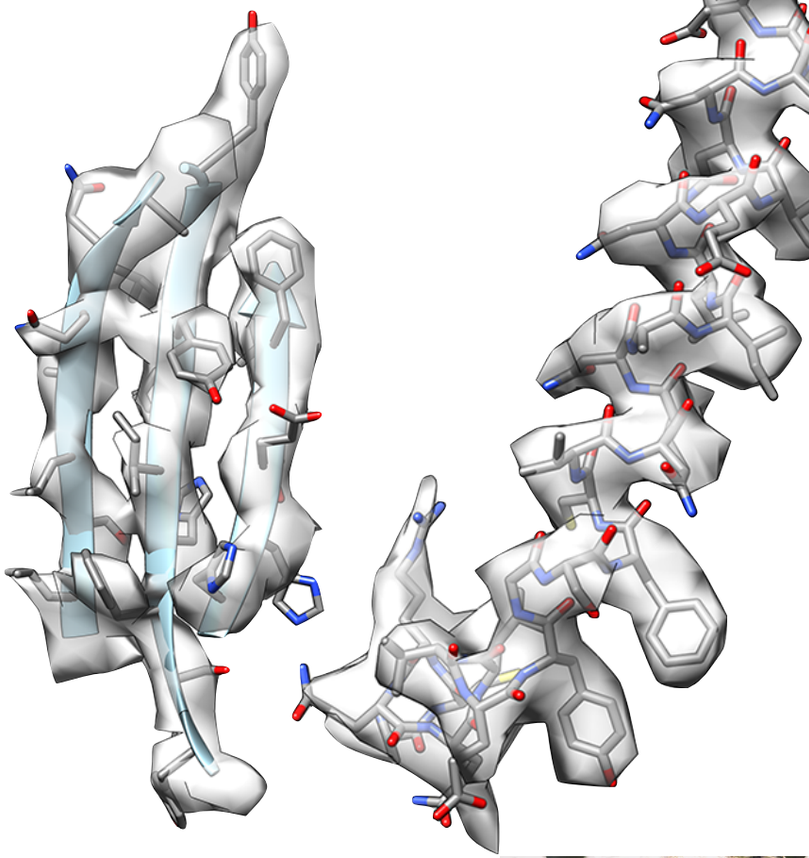
The Rosetta *de novo* building strategy

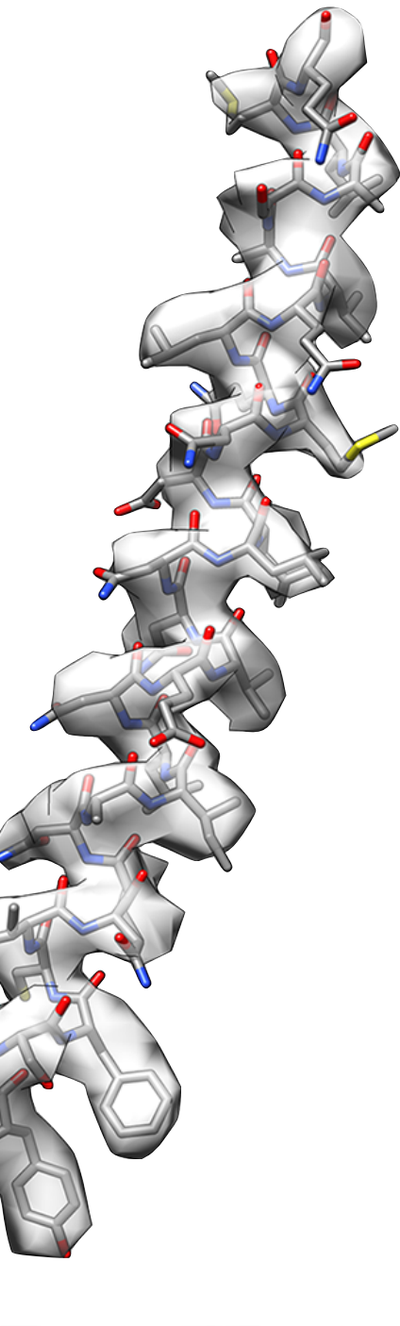
Prefusion MHV spike glycoprotein 4.0Å resolution



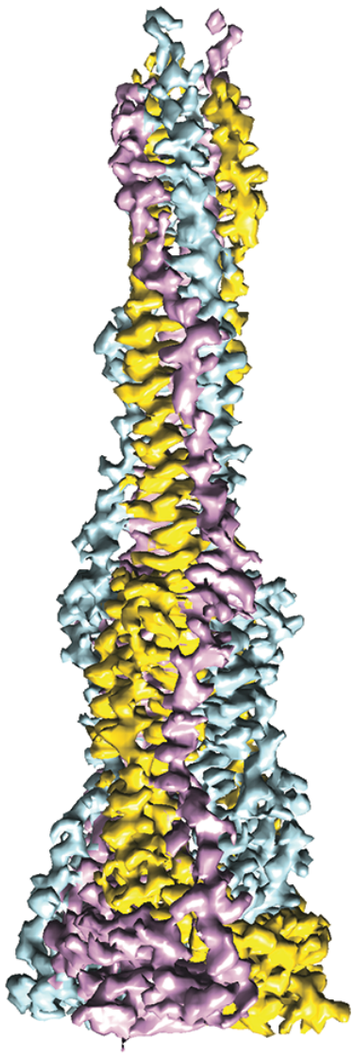


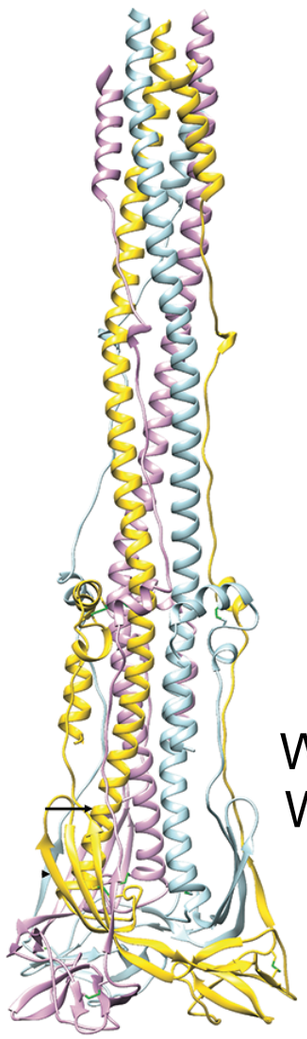






Postfusion MHV spike glycoprotein 4.1Å resolution





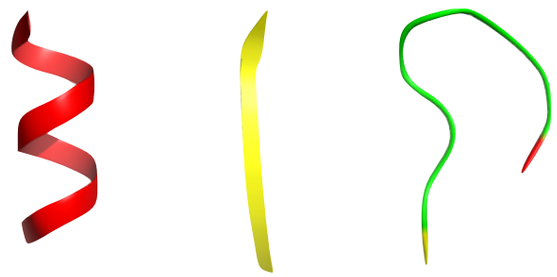
Walls AC et al (2016) Nature
Walls AC et al (2017) PNAS

Lexi Walls

The Rosetta *de novo* building strategy

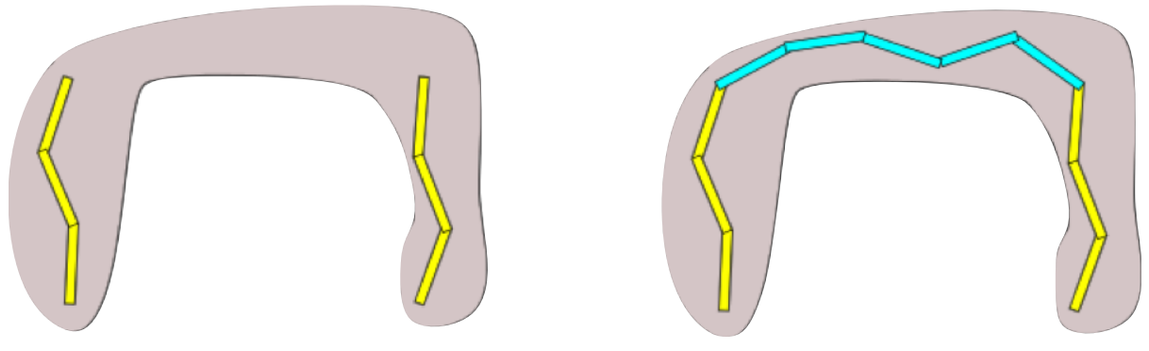
- **In placement: long fragments do not necessarily capture structure accurately in regions of high local variation**

low variability



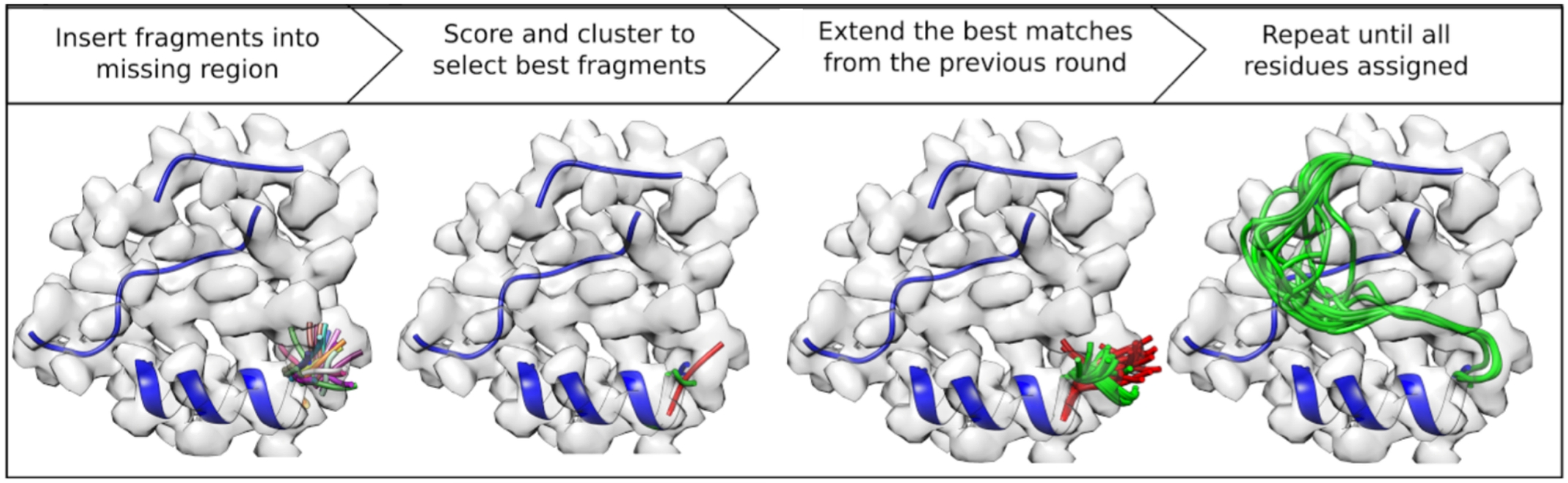
high variability

- **In rebuilding: sampling long segments requires all the residues to be correct in order to see an energy signal**



Rosetta enumerative sampling (RosettaES)

- **Use shorter fragments for more accuracy in variable regions while still maintaining information on sequence preference**
- **Employ a greedy conformational strategy to handle the large conformational space**
- **Model completion**
- **Works with unsegmented density & symmetry**



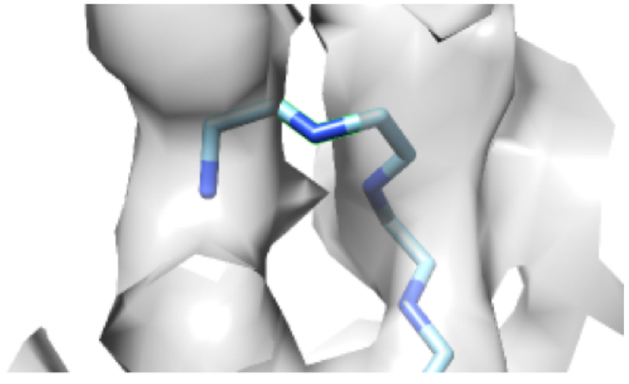
Frank DiMaio

Brandon Frenz

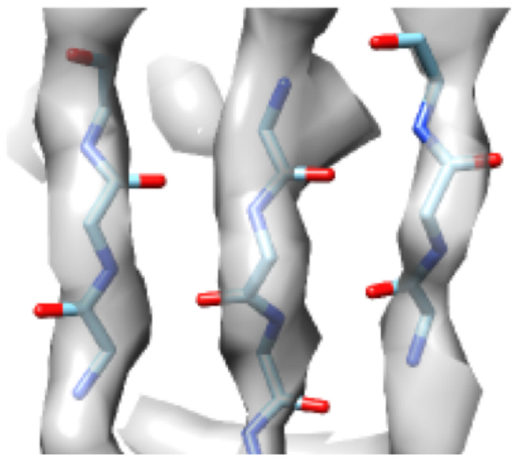
Frenz B et al (2017) Nature Methods

Rosetta enumerative sampling (RosettaES)

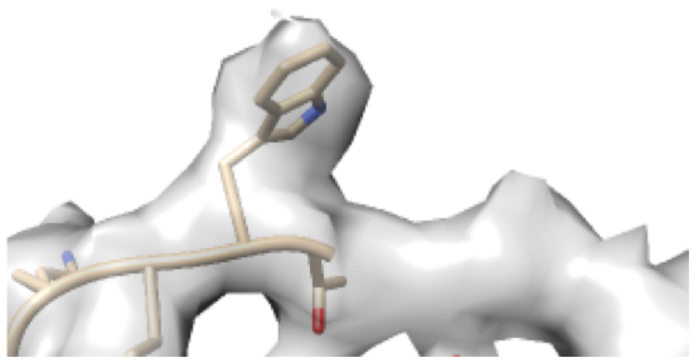
- **Reduction of search space leads to improvements**
  - **Penalizing density discontinuities**



  - **Explicit modeling of β-sheets**

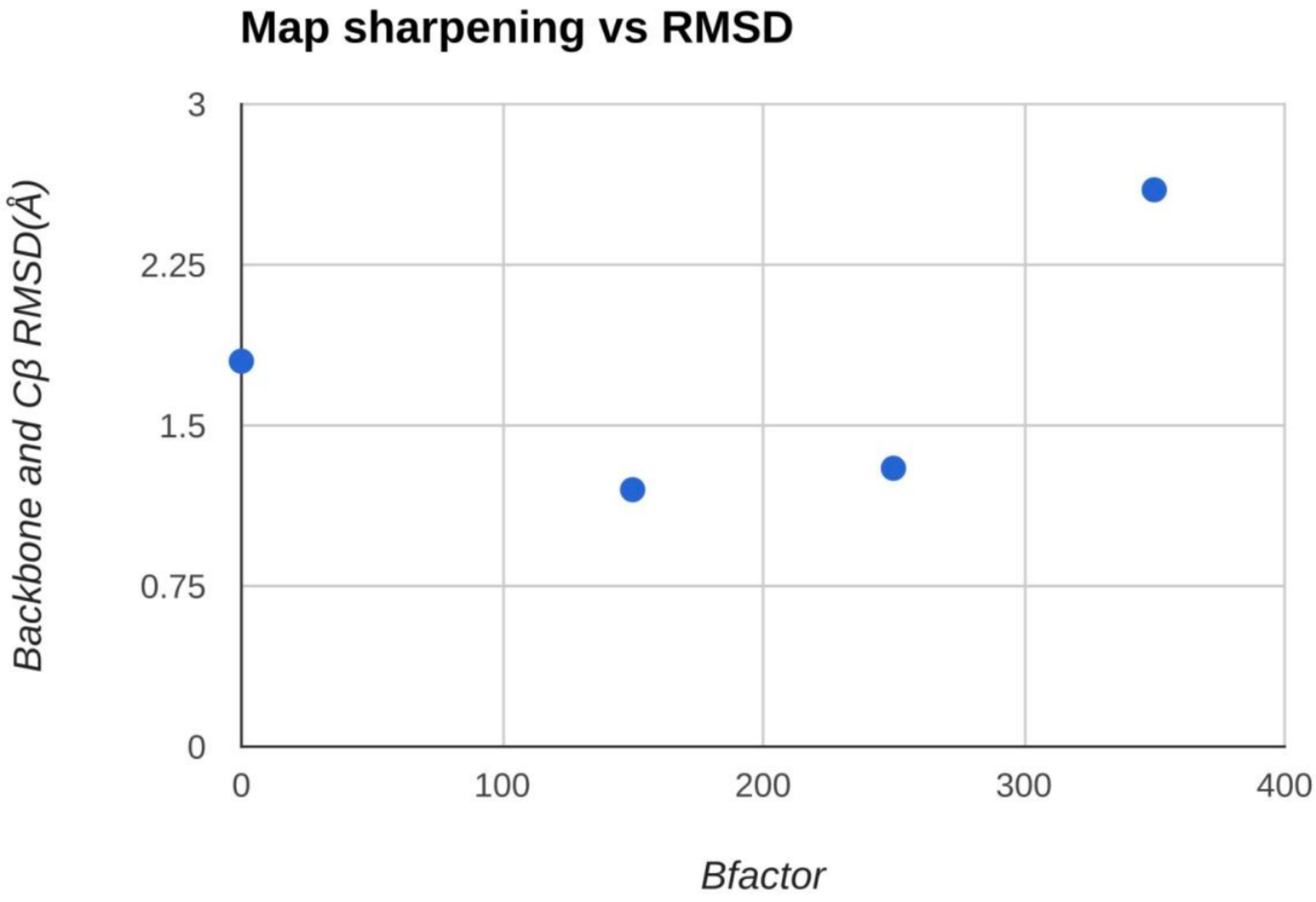


- **Rotamer-like density matching**



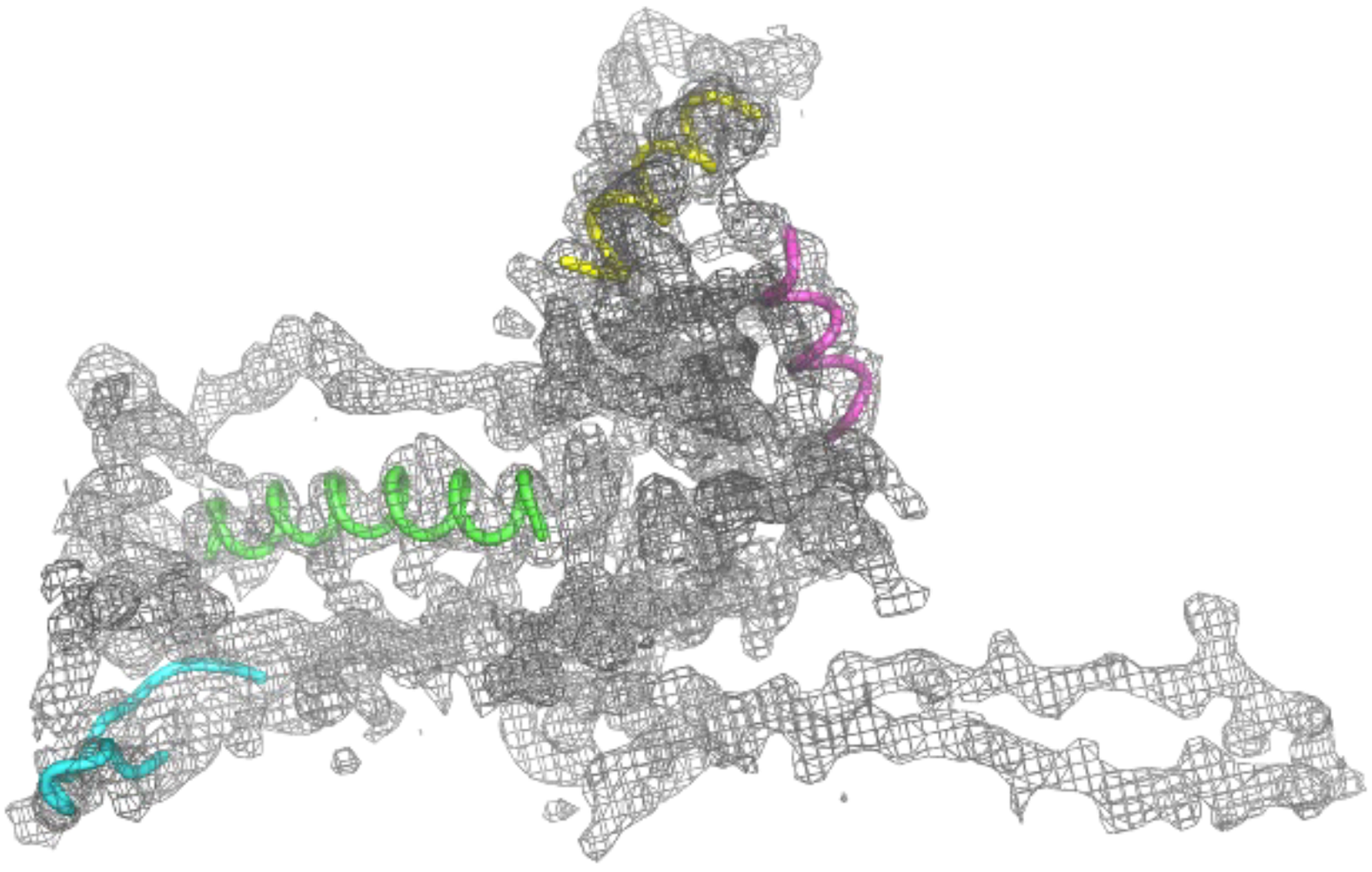
Frenz B et al (2017) Nature Methods

Rosetta enumerative sampling (RosettaES)



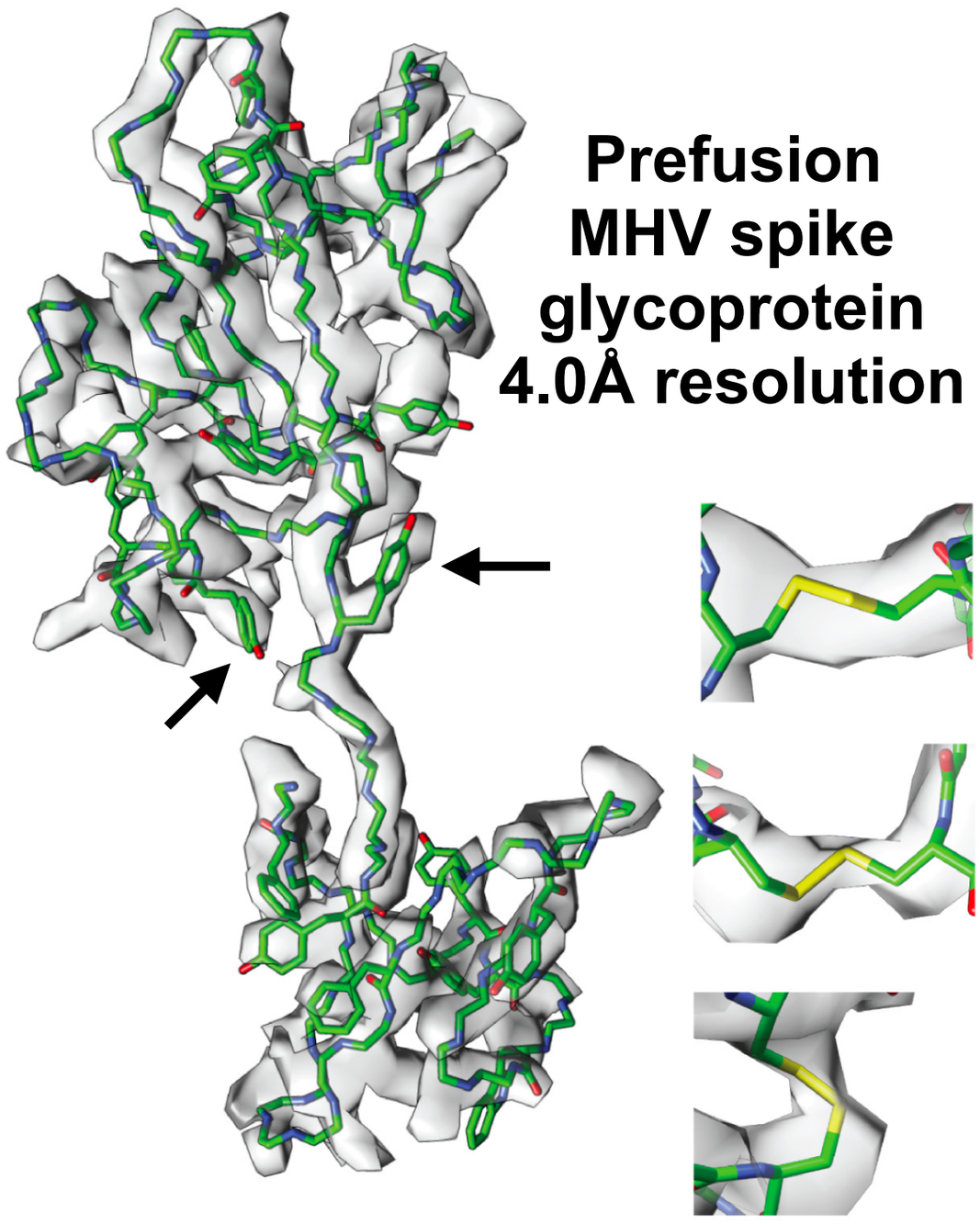
Frenz B et al (2017) Nature Methods

Rosetta enumerative sampling (RosettaES)

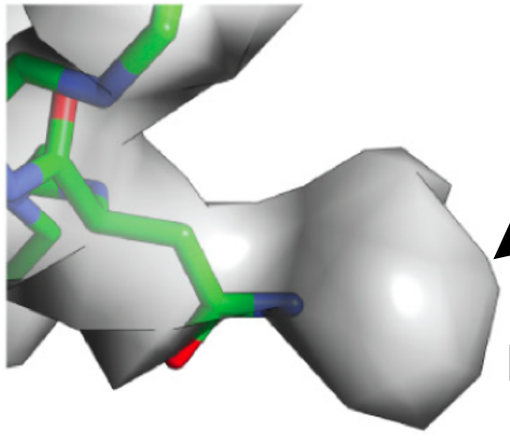


Frenz B et al (2017) Nature Methods
Zhang X et al (eLIFE2013)

Rosetta enumerative sampling (RosettaES)

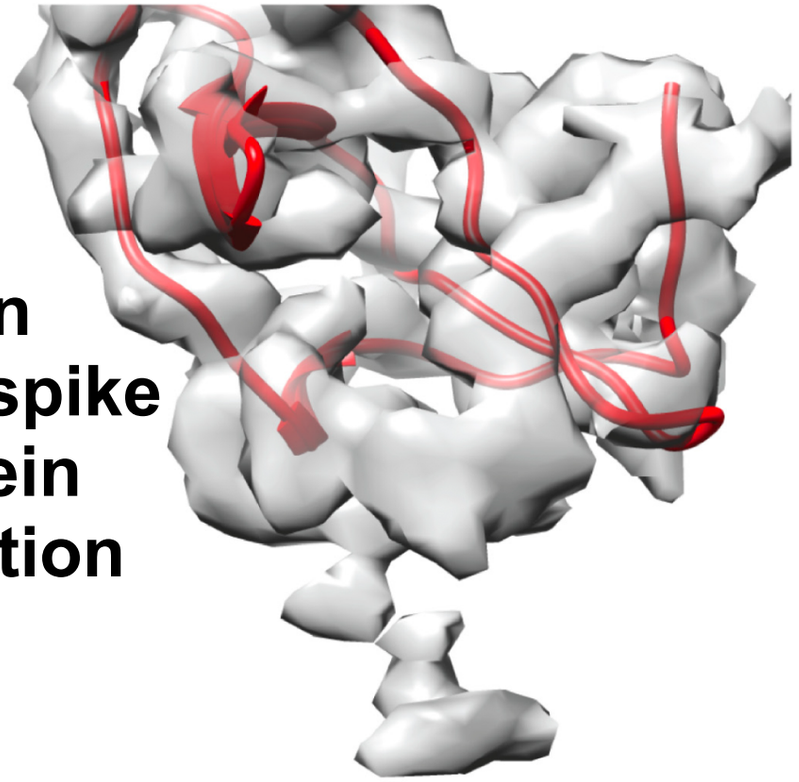


Prefusion MHV spike glycoprotein 4.0Å resolution

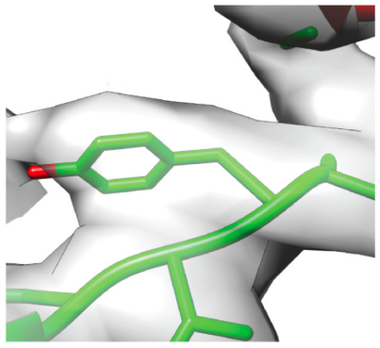


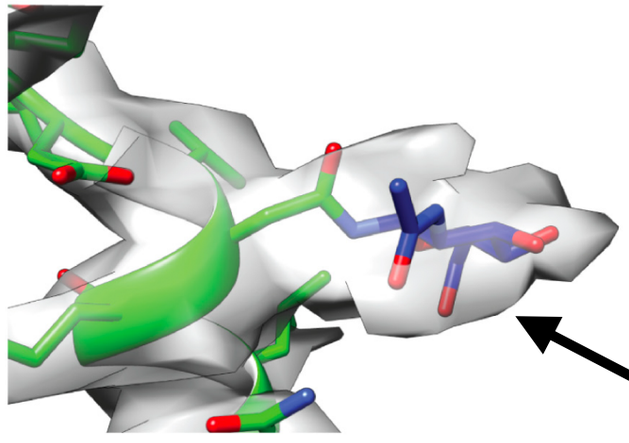
Glycosylation site (confirmed using MS)

Frenz B et al (2017) Nature Methods
Walls AC et al (2016) Nature

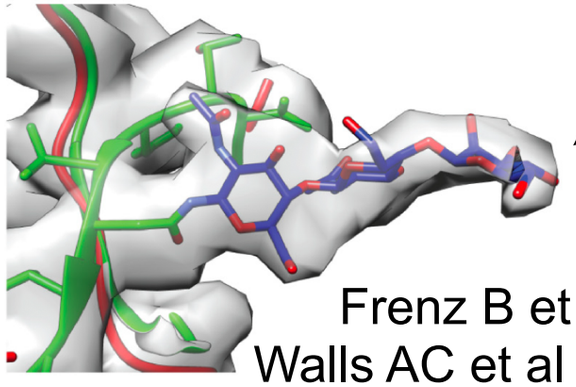


Prefusion HCoV-NL63 spike glycoprotein 3.4Å resolution



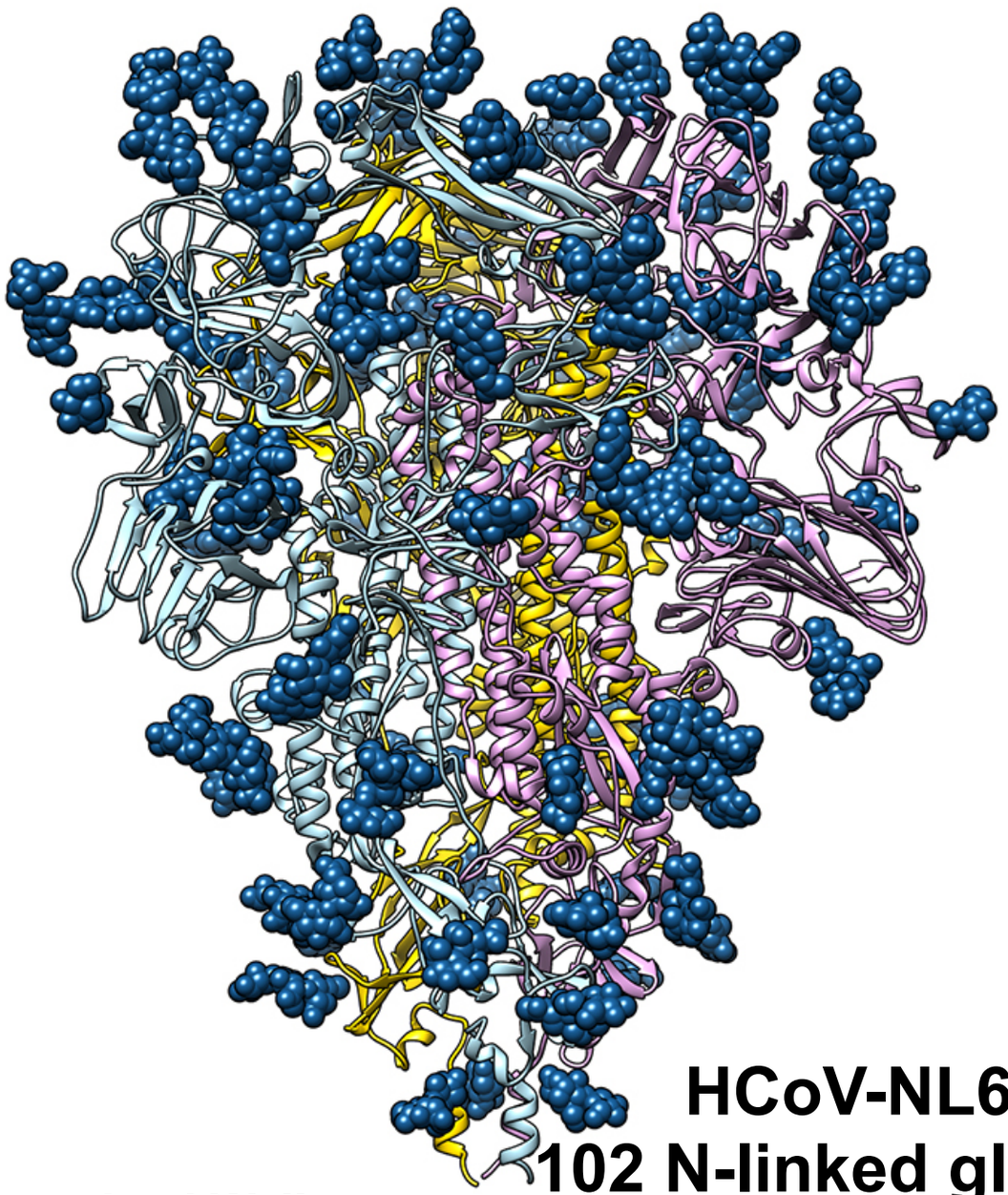


Glycosylation sites (confirmed using MS)



Frenz B et al (2017) Nature Methods
Walls AC et al (2016) Nature Struct Mol Biol

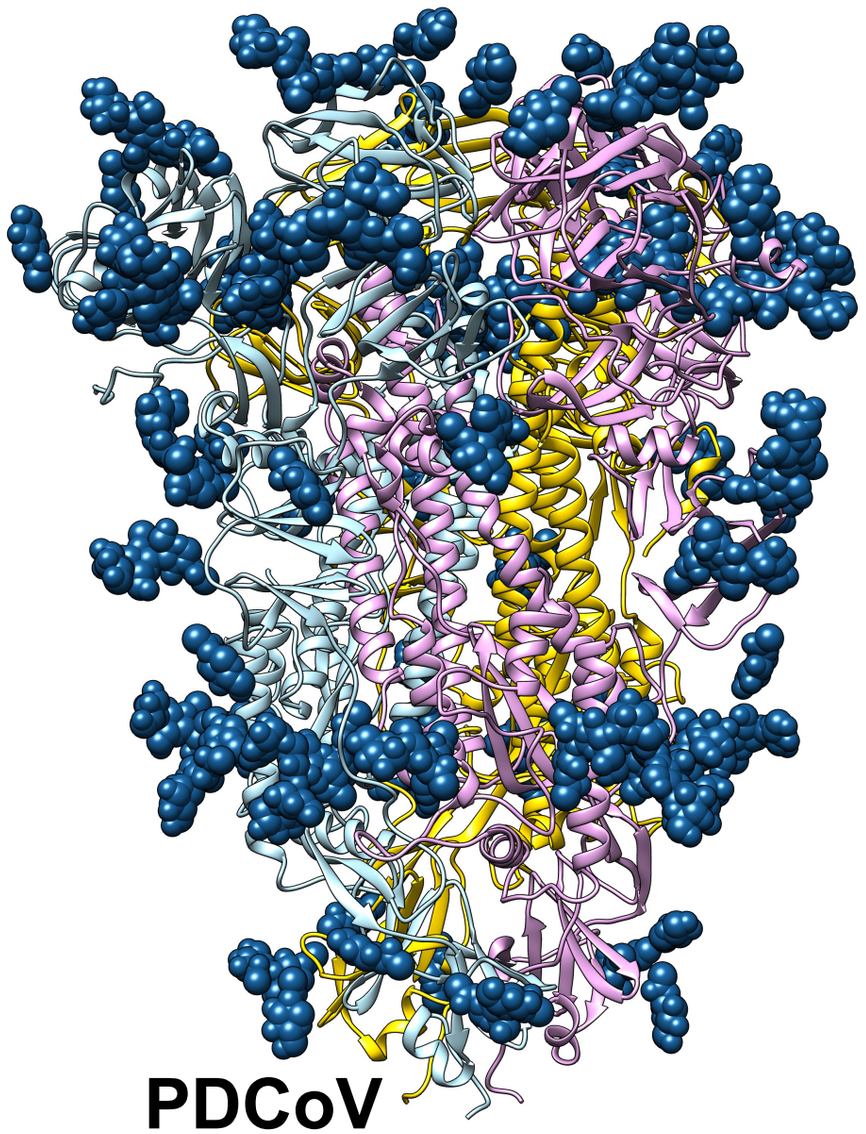
Rosetta glycan refinement



HCoV-NL63
102 N-linked glycans

Frank DiMaio

Brandon Frenz



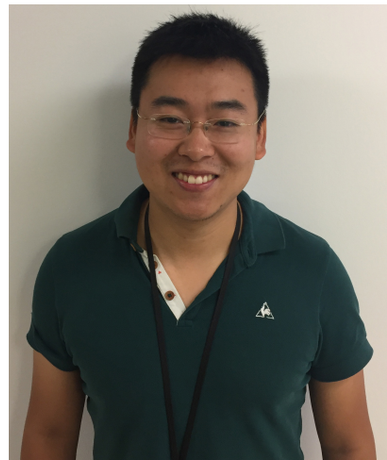
PDCoV
78 N-linked glycans

Lexi Walls

Automated glycan detection
Refinement using stereochemical restraints
(chair conformation)

Walls AC et al (2016) Nature Struct Mol Biol
Xiong X et al (2017) J Virol
Frenz B et al, unpublished

Alex Xiong



Carbohydrate anomalies in the PDB

To the Editor: The importance of carbohydrates both to fundamental cellular biology and as integral parts of therapeutics (including antibodies) continues to grow. The presence of the correct glycans is important for the beneficial effects of therapeutic glycoproteins and is likely to be increasingly required by regulatory agencies. However, carbohydrates (and other small molecules) are handled poorly in macromolecular structural biology. When such small molecules are present in macromolecule structures, they are often reported with stereo- and regiochemical errors and in unlikely conformations. Stereo- and regiochemistry should always be correct, and although conformational distortions may reflect interactions taking place in a complex[1], most are also likely to be erroneous—resulting from poor chemical understanding and lack of appropriate stereochemical restraints in refinement, often against low-resolution data[2].

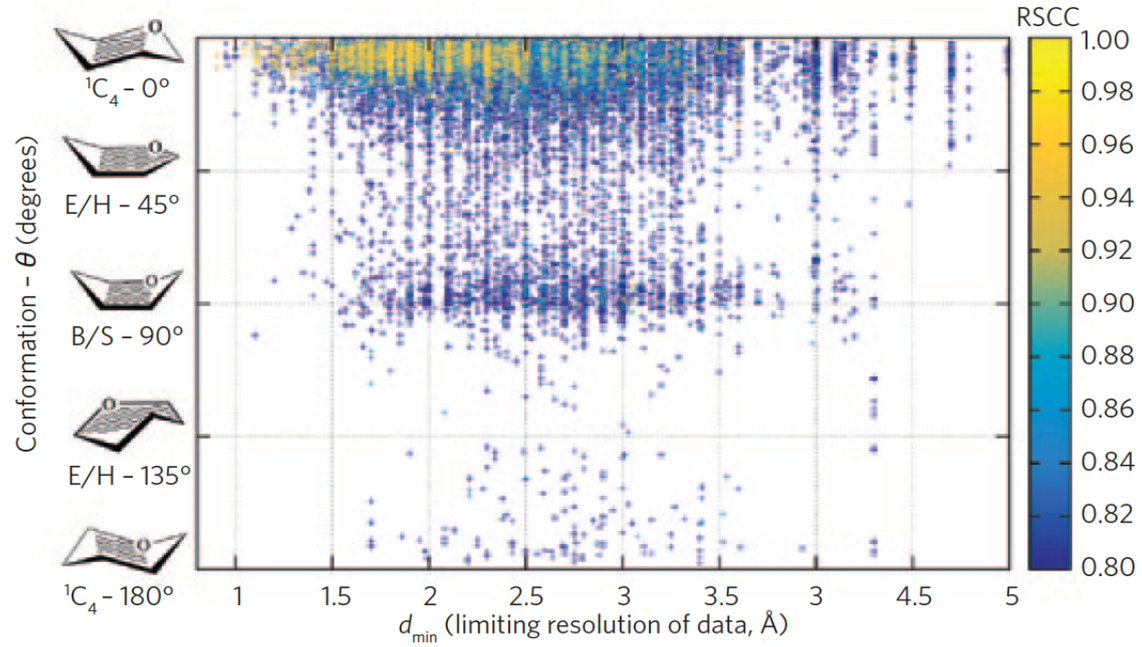
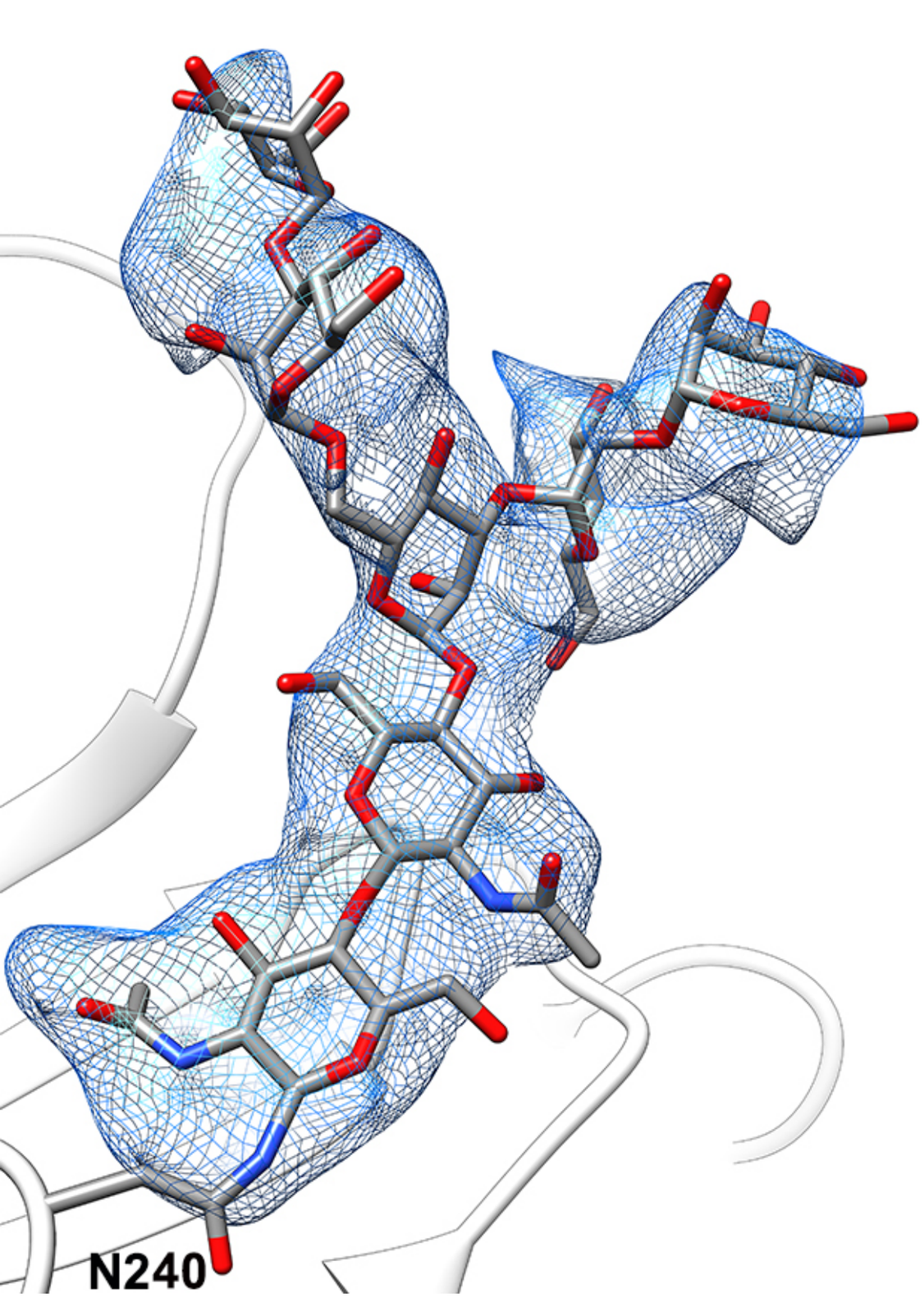
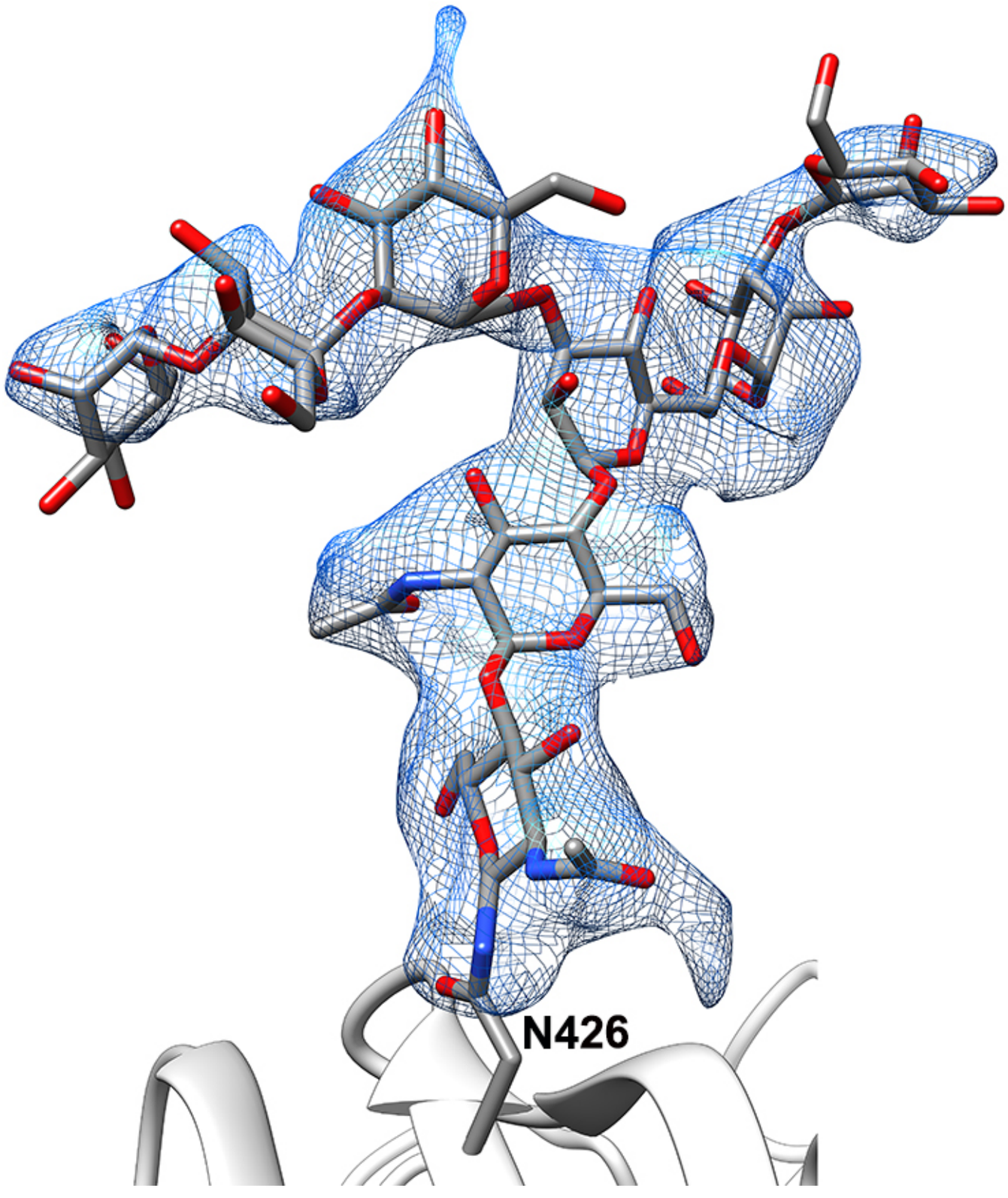


Figure 1 | Distribution of D-pyranoside ring conformations as a function of resolution for all N-linked sugars (at distance <2.0 Å) in the PDB as of January 2015, identified by their Chemical Component Dictionary IDs: NAG, NDG, MAN, BMA, BGC, GLC, GAL and GLA. E/H, envelopes and half-chairs; B/S, boats and skew-boats; wavy lines denote the main ring plane. For clarity, an envelope is depicted at $\theta = 45°$ and a half-chair at $\theta = 135°$, and skew-boat is omitted from the equator.

Agirre J et al (2015) Nature Chem Biol

Rosetta glycan refinement



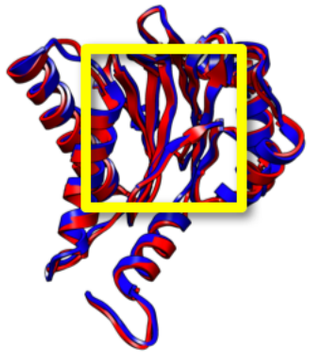


Walls AC et al (2016) Nature Struct Mol Biol

Rosetta comparative modeling (RosettaCM)

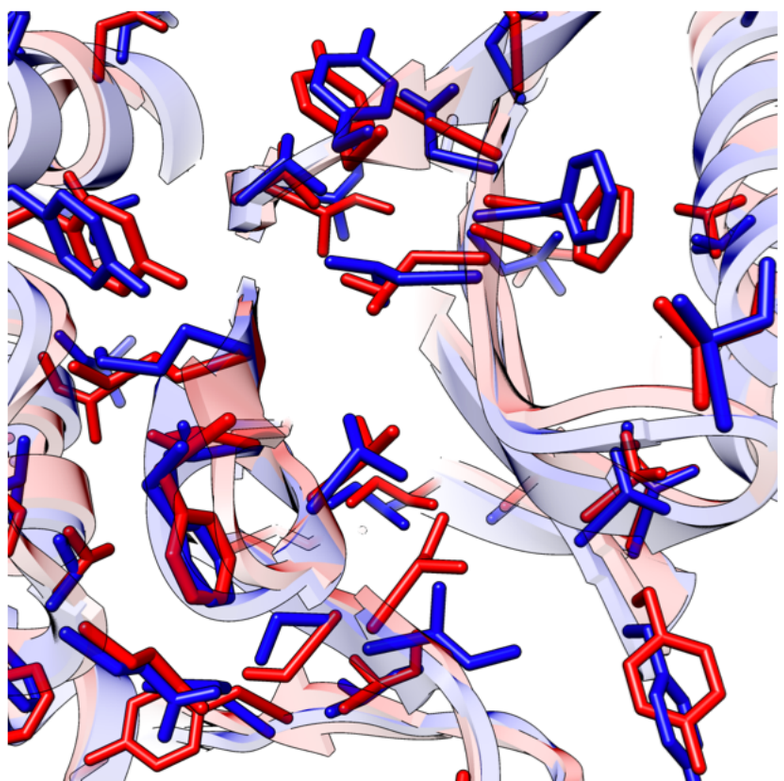
- **Structure(s) sharing sequence similarity to the target structure**
  - **Partial structures**
  - **Multiple structures**
- **Fragments predicted using local sequence**
- **Tuning of the frequency with which the two sources of information are used**
- **Also allows model completion when 70% built model available (although RosettaES outperforms RosettaCM in most cases)**
- **Recovers core packing from crystal structures**

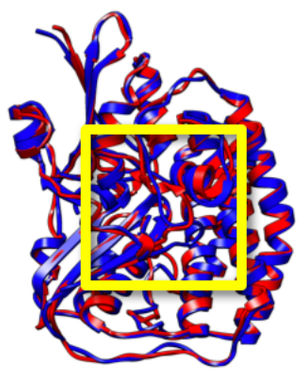
Song Y et al (2013) Structure



T20S α-subunit 4.8 Å map

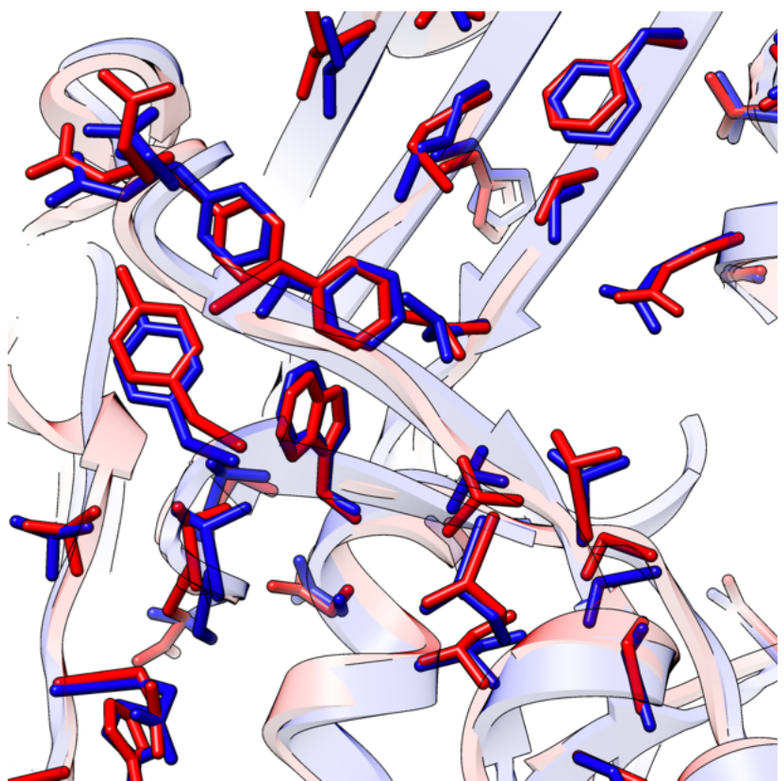
2.0 Å all-atom r.m.s.d.





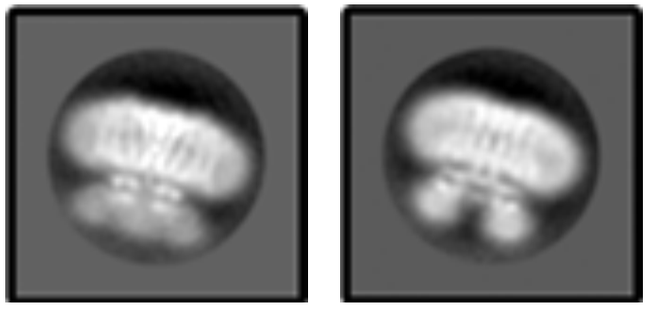
FrhA 3.4 Å map

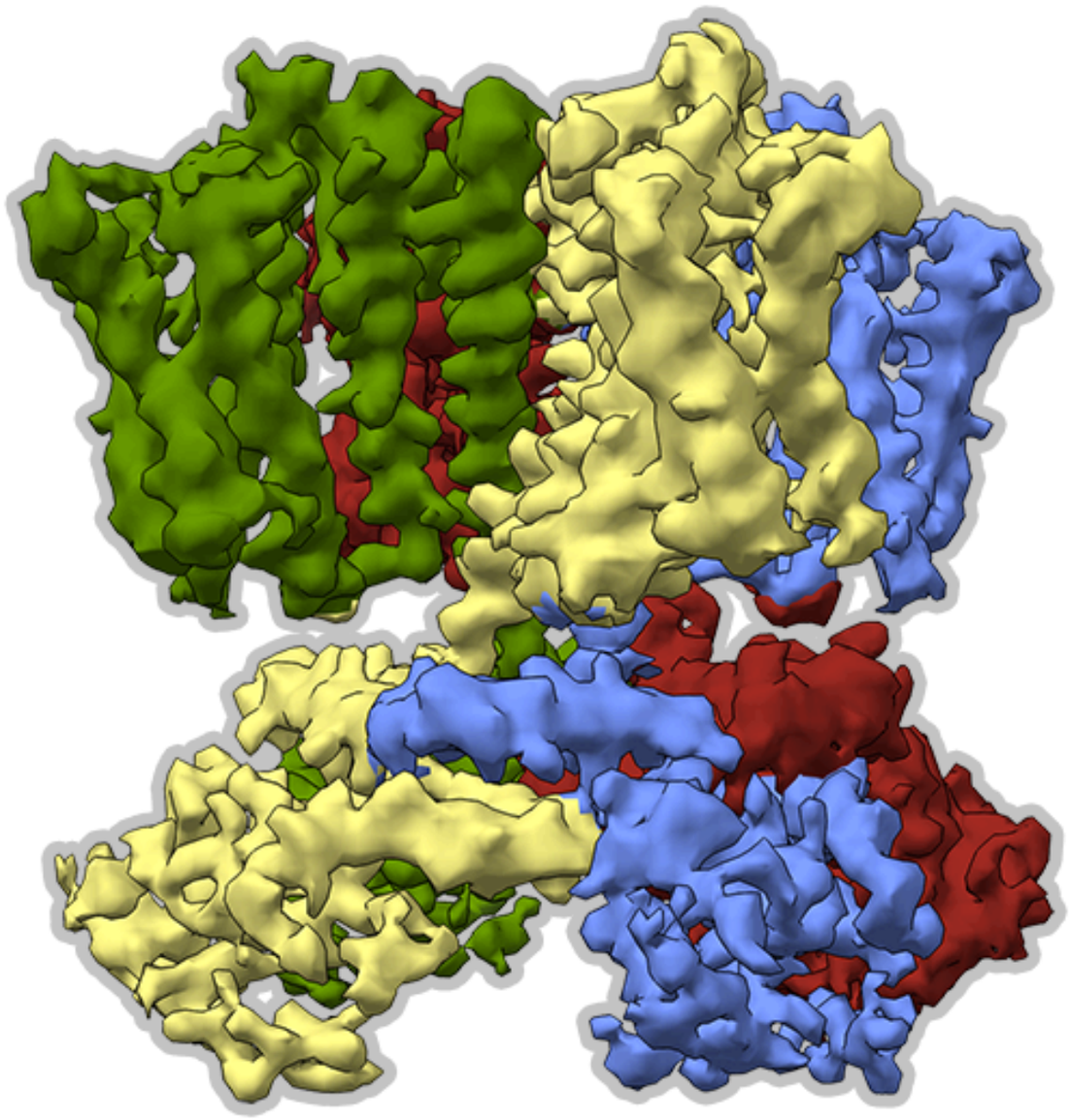
2.1 Å all-atom r.m.s.d.



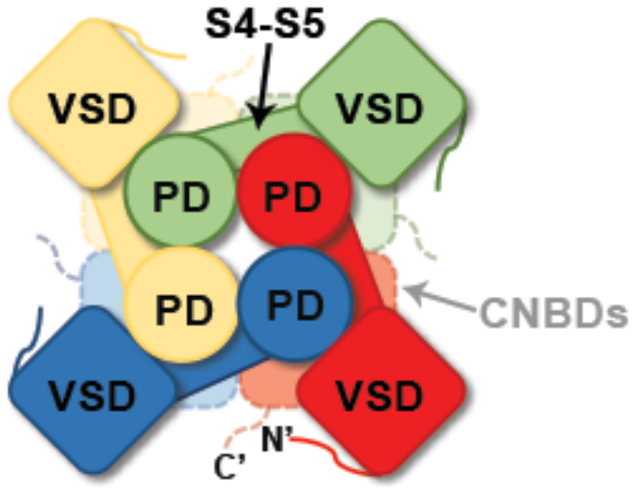
Rosetta comparative modeling (RosettaCM)

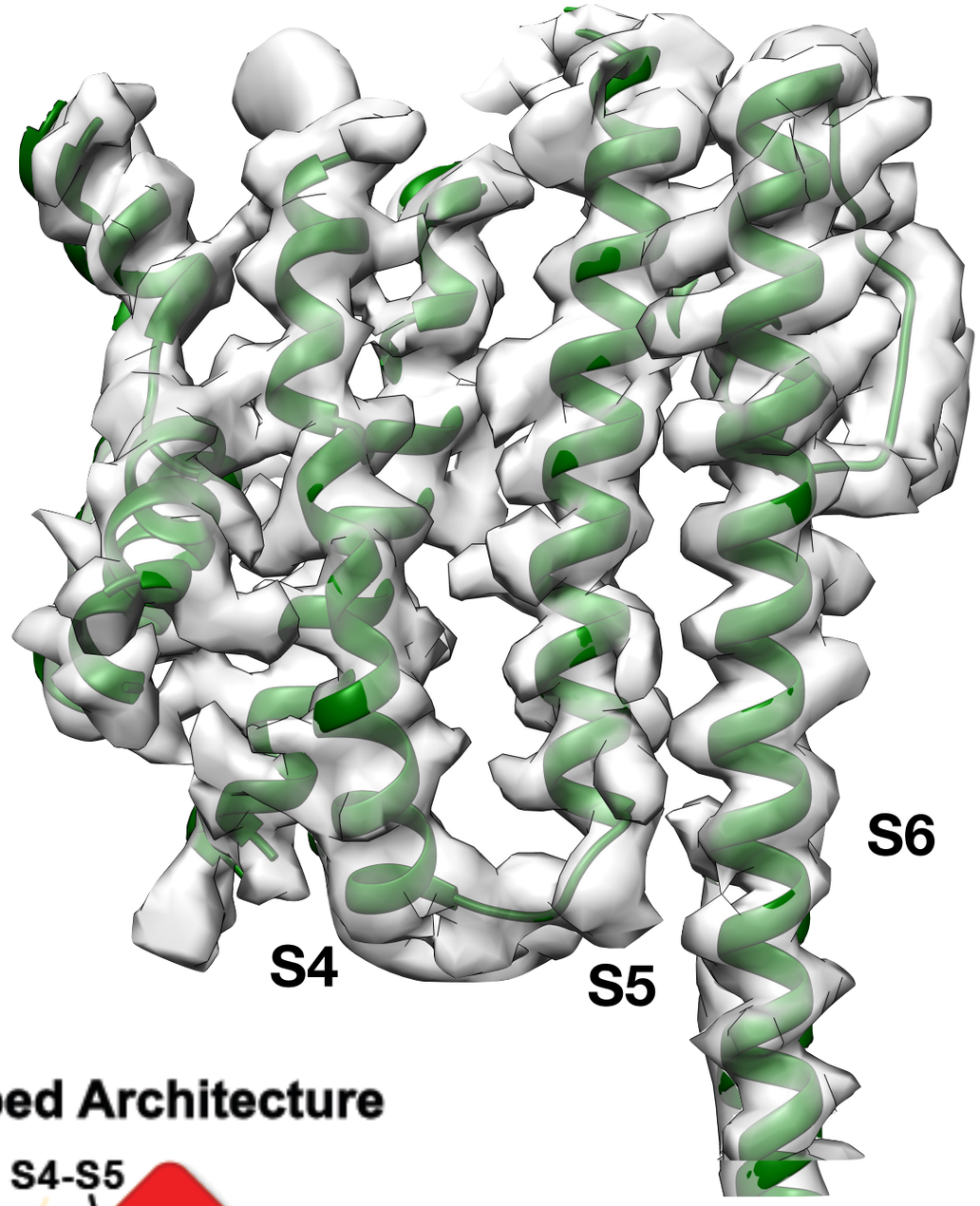
CNG channel
4.2Å resolution



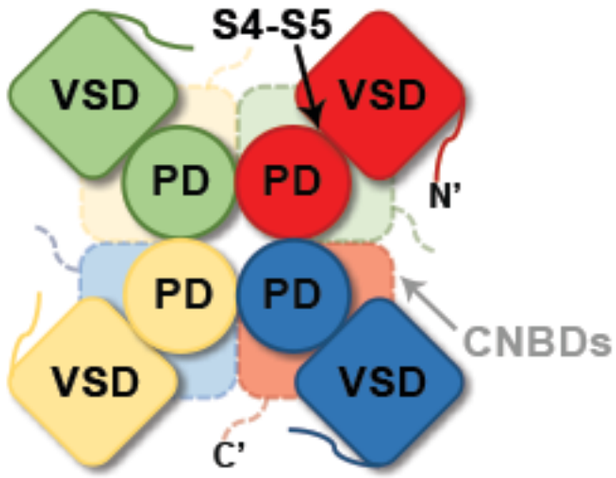


Swapped Architecture





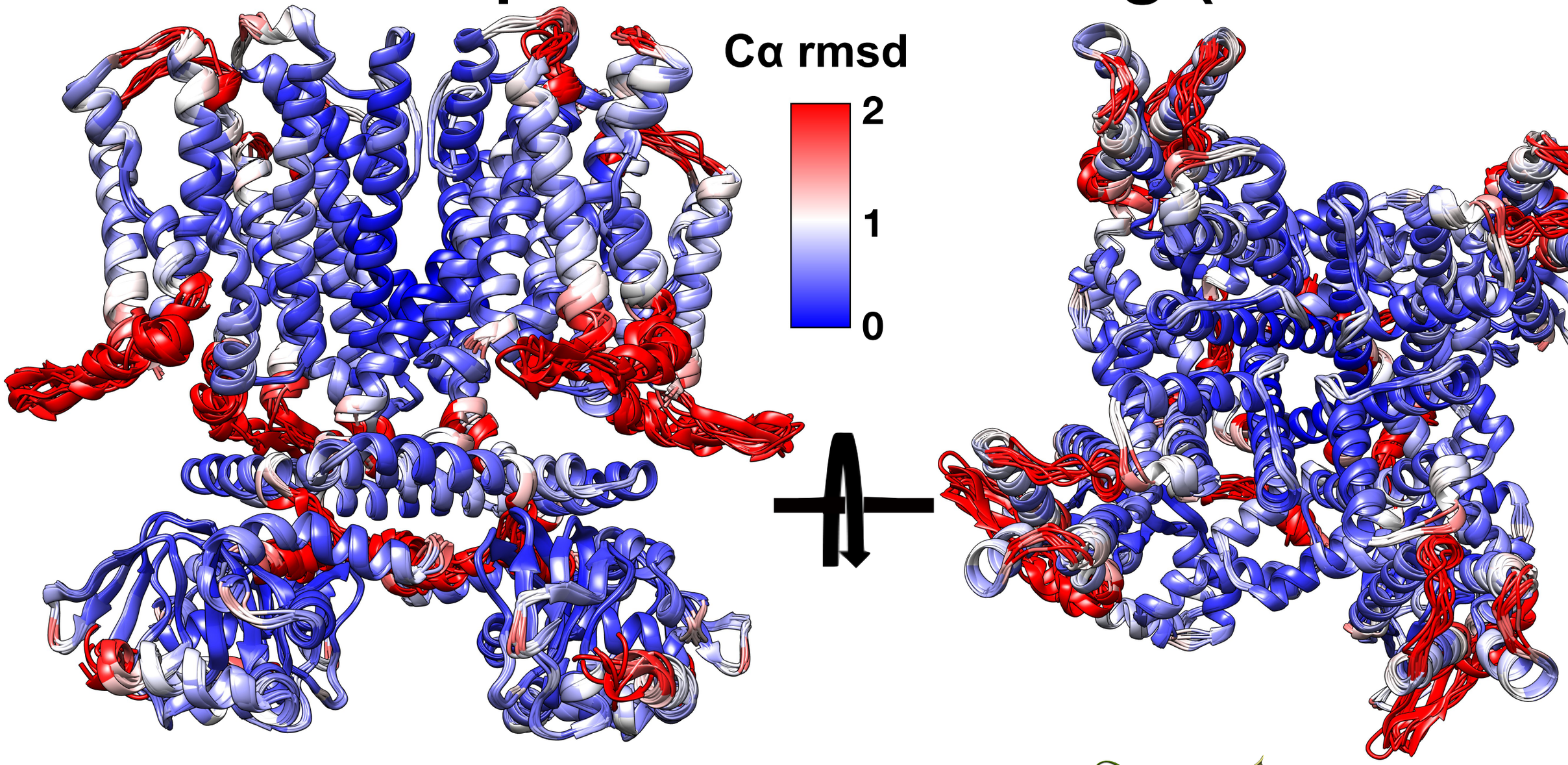
Non-Swapped Architecture



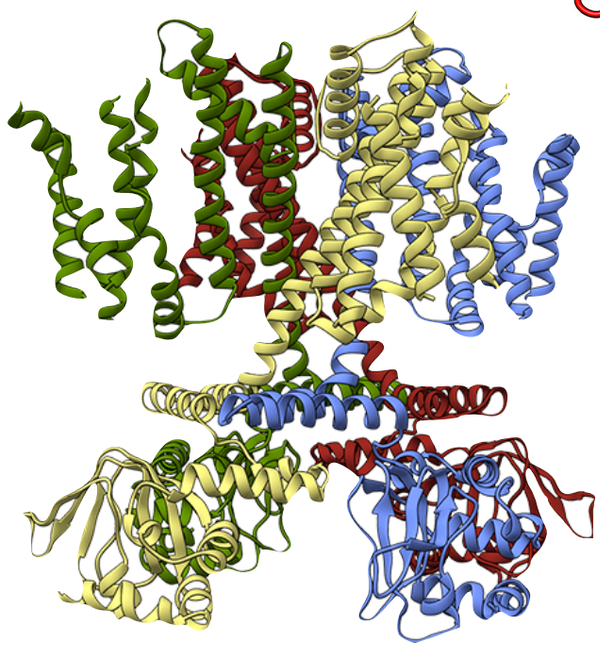
James ZM, Borst AJ et al (2017) PNAS

Andrew Borst

Rosetta comparative modeling (RosettaCM)



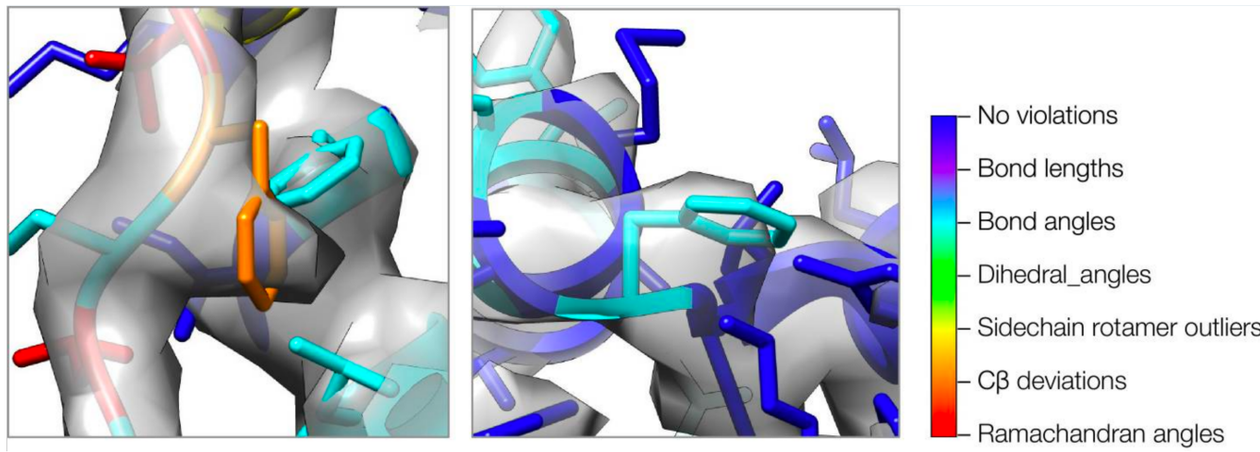
Deleted disordered loops between transmembrane helices
&
Truncated residue to Cβ in the C-terminal CNBD due to dampened resolution



James ZM, Borst AJ et al (2017) PNAS

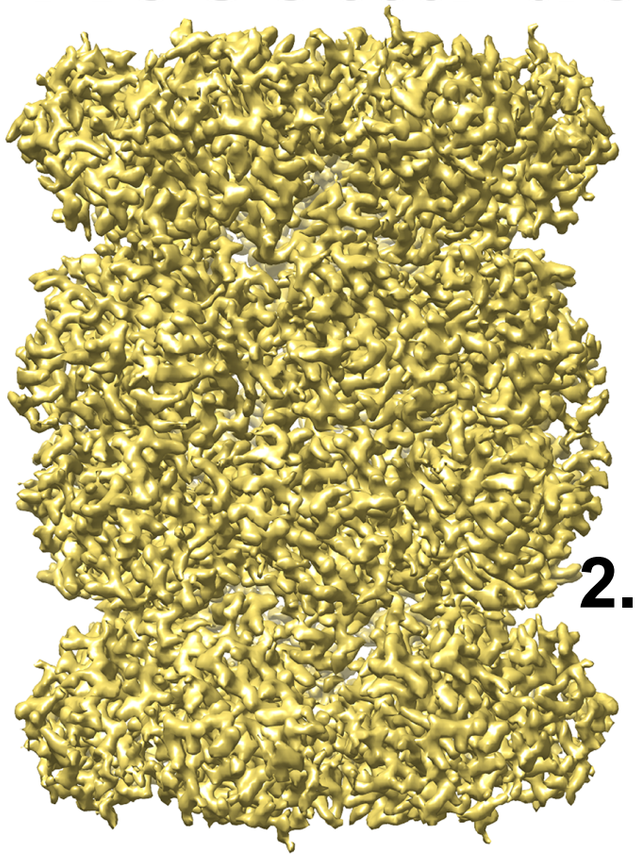
Rosetta density-guided iterative refinement

- Fragments predicted using local sequence
  - Size can be tuned: the longer the fragment, the more divergence is allowed
  - Targets:
    - Regions with conformational strains and poor fit to density
    - Random segments of the chain(s)
      - User-defined regions
- 100-1000 models
- Produces atomic-level accuracy models using maps determined at 4.5 Å resolution or better
- Large radius of convergence
- Voxel size refinement
- B-factor refinement after model completion



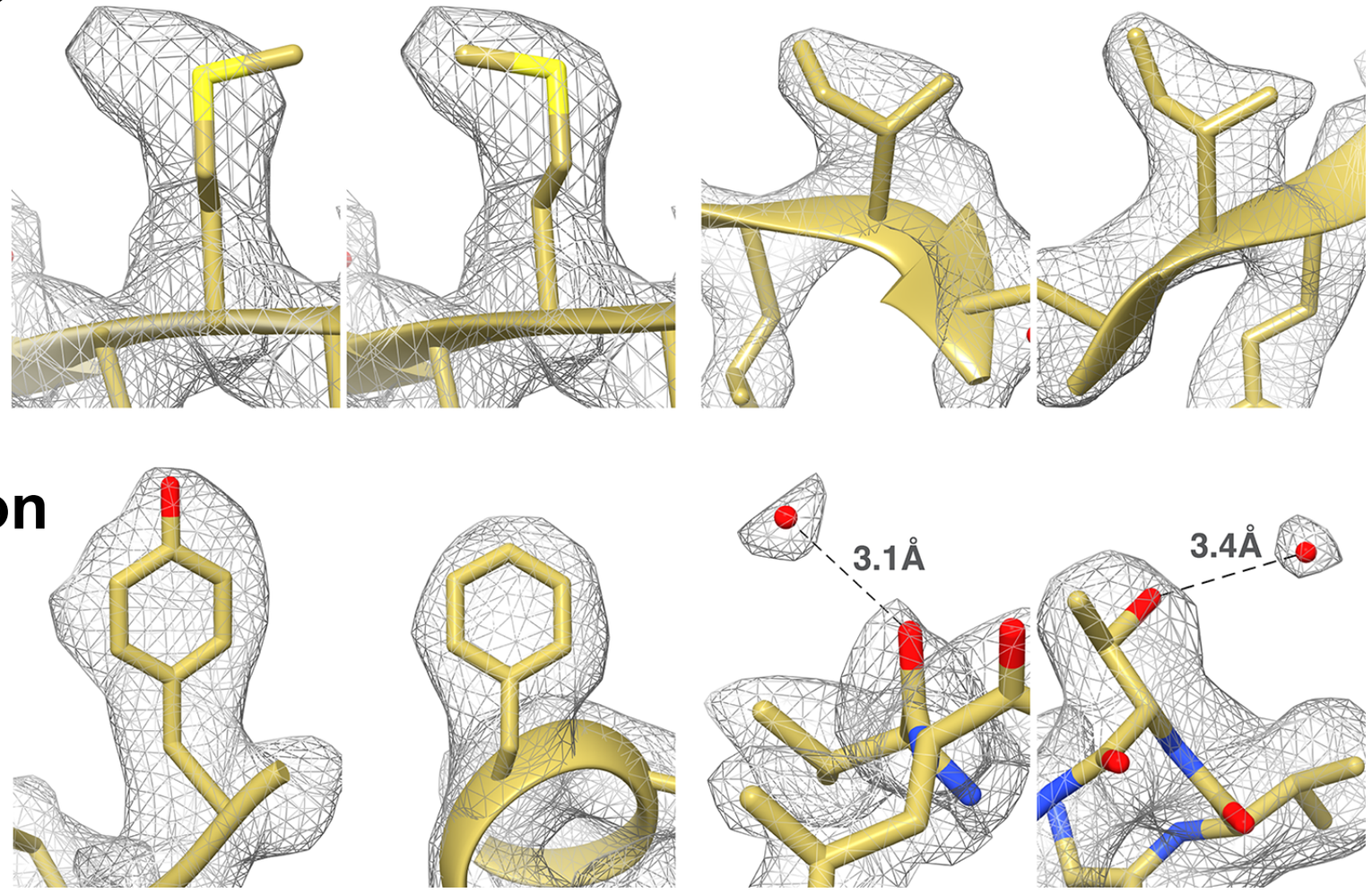
DiMaio F et al (2015) Nature Methods
Wang RYR et al (2015) Nature Methods

Rosetta density-guided iterative refinement

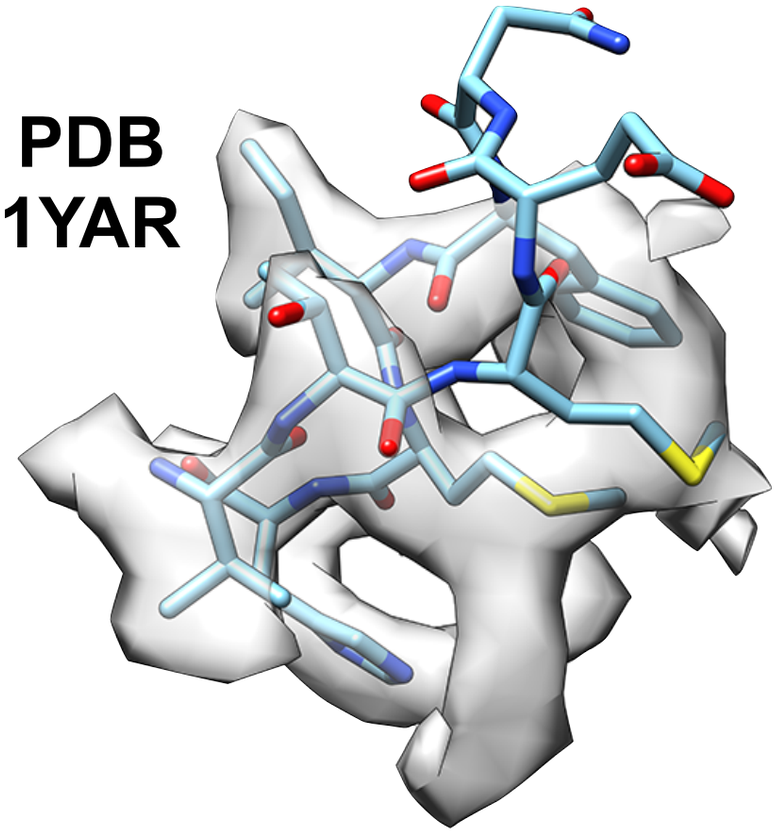


Campbell MG et al (2015) eLIFE

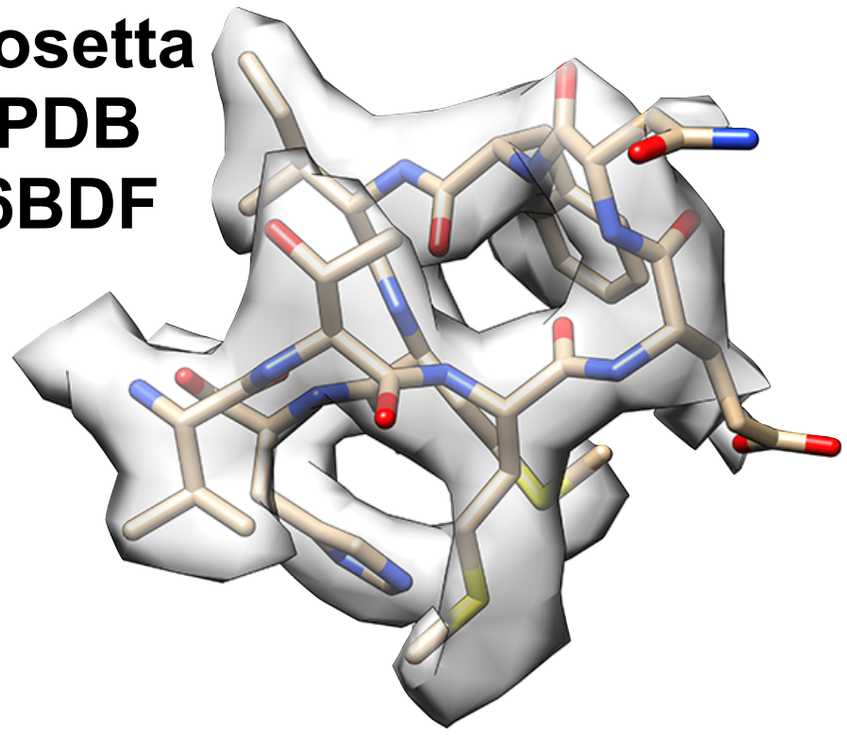
T20S
2.8Å resolution

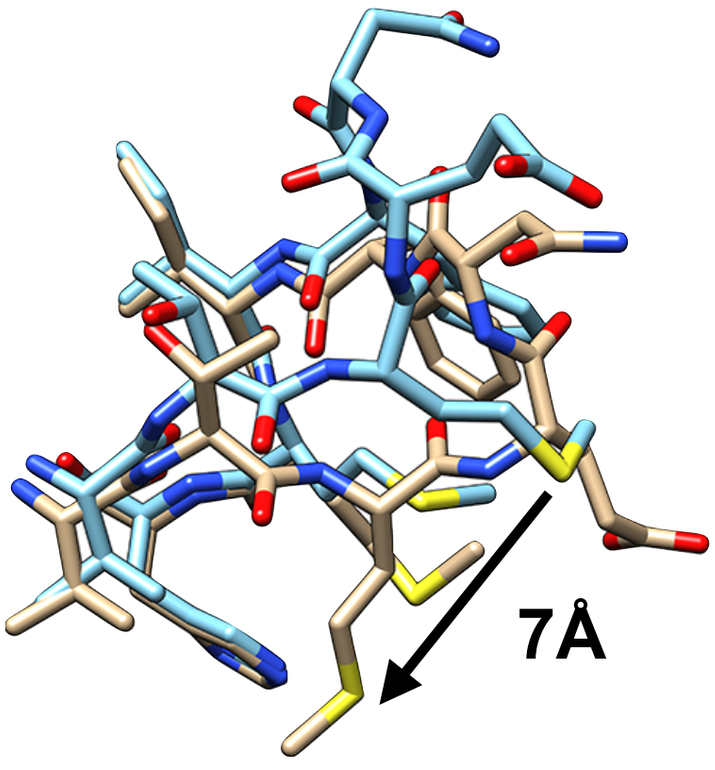


PDB
1YAR



Rosetta
PDB
6BDF





Computational cost of running Rosetta

- Several hundreds to thousands of trajectories
  - One cpu/trajectory
  - Selection based on Rosetta energy function and/or map/model FSC
- Run time depends of size of model to be built
  - 1.5h for a 30-residue fragment for RosettaES using 16 cores
  - 2-4h/model for RosettaCM, Rosetta density-guided iterative refinement (75-120 kDa/protomer applying symmetry)
  - 15 min/model for Rosetta relax
- Backfill queue
  - Makes use of free cpu cycles
  - Typical of super-computer centers

Rosetta structure determination pipeline

- Rosetta energy function and conformational sampling are valuable at bridging the resolution gap between for near-atomic resolution reconstructions

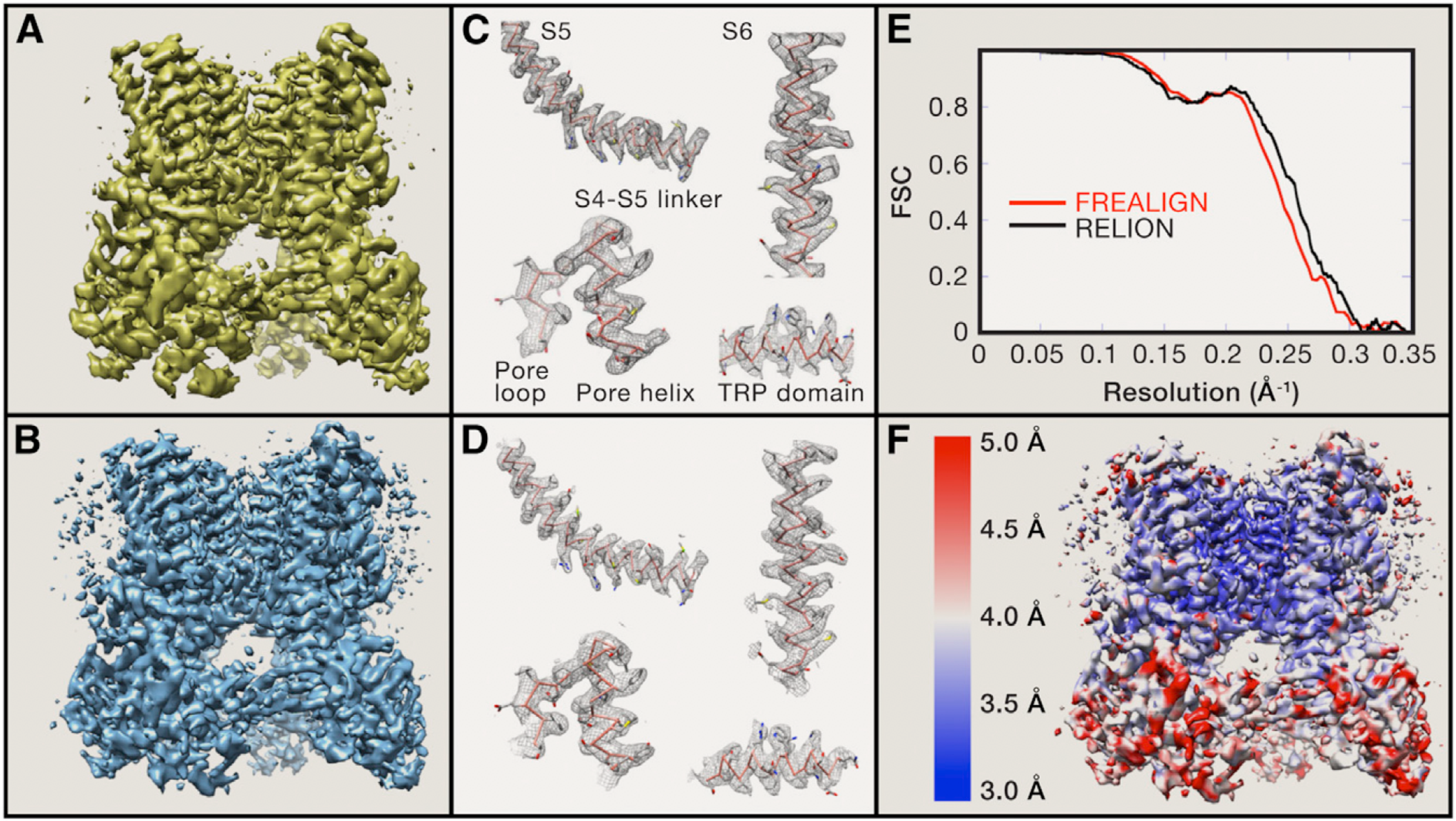
- Automatic de novo structure determination at 3.5-4.5Å is possible (up to 4.8A!)

- Versatile pipeline

Model building using several maps

- Different subsets of particles (classification)

- Different softwares

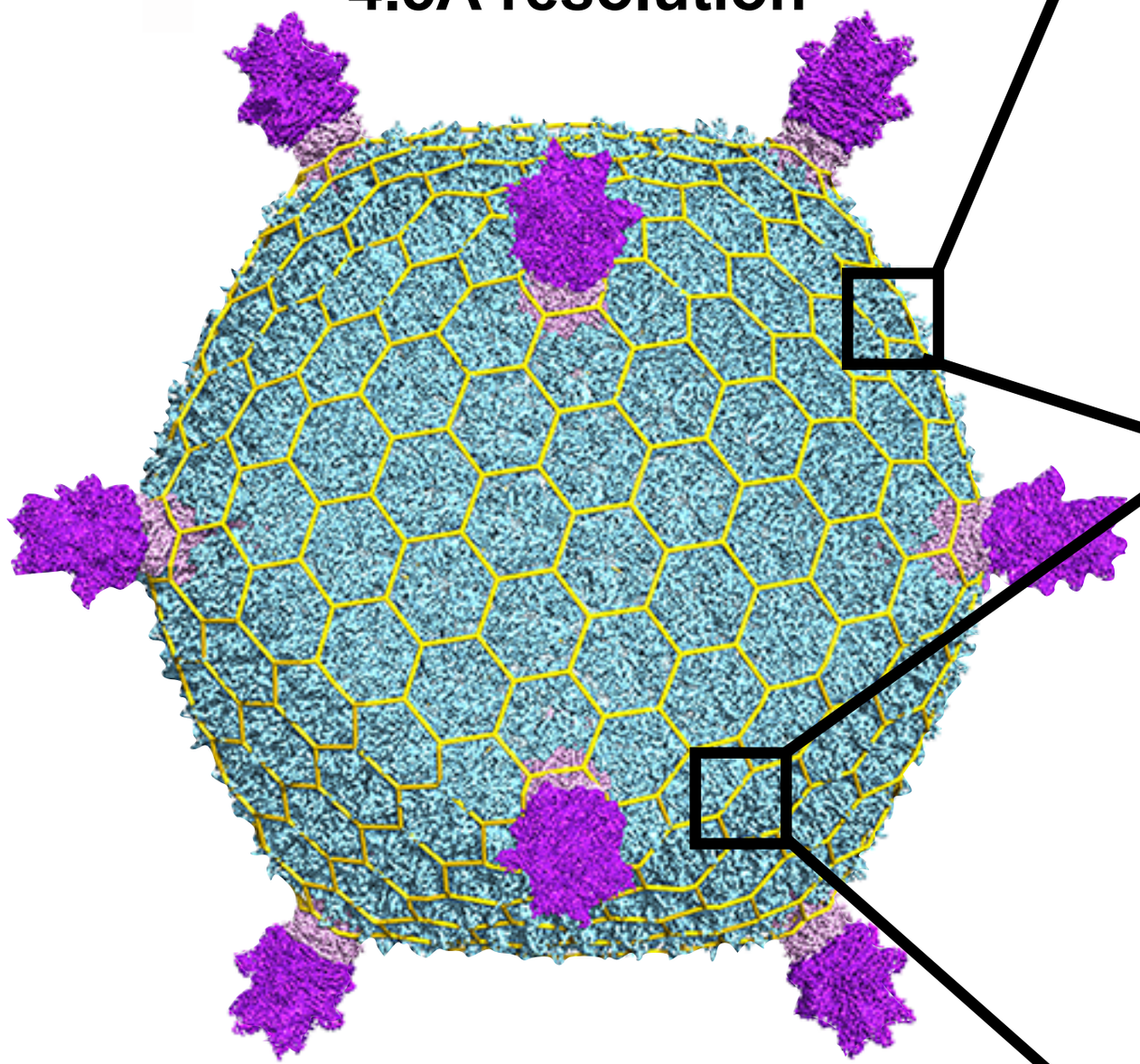


Cheng Y et al (2015) Nature Methods

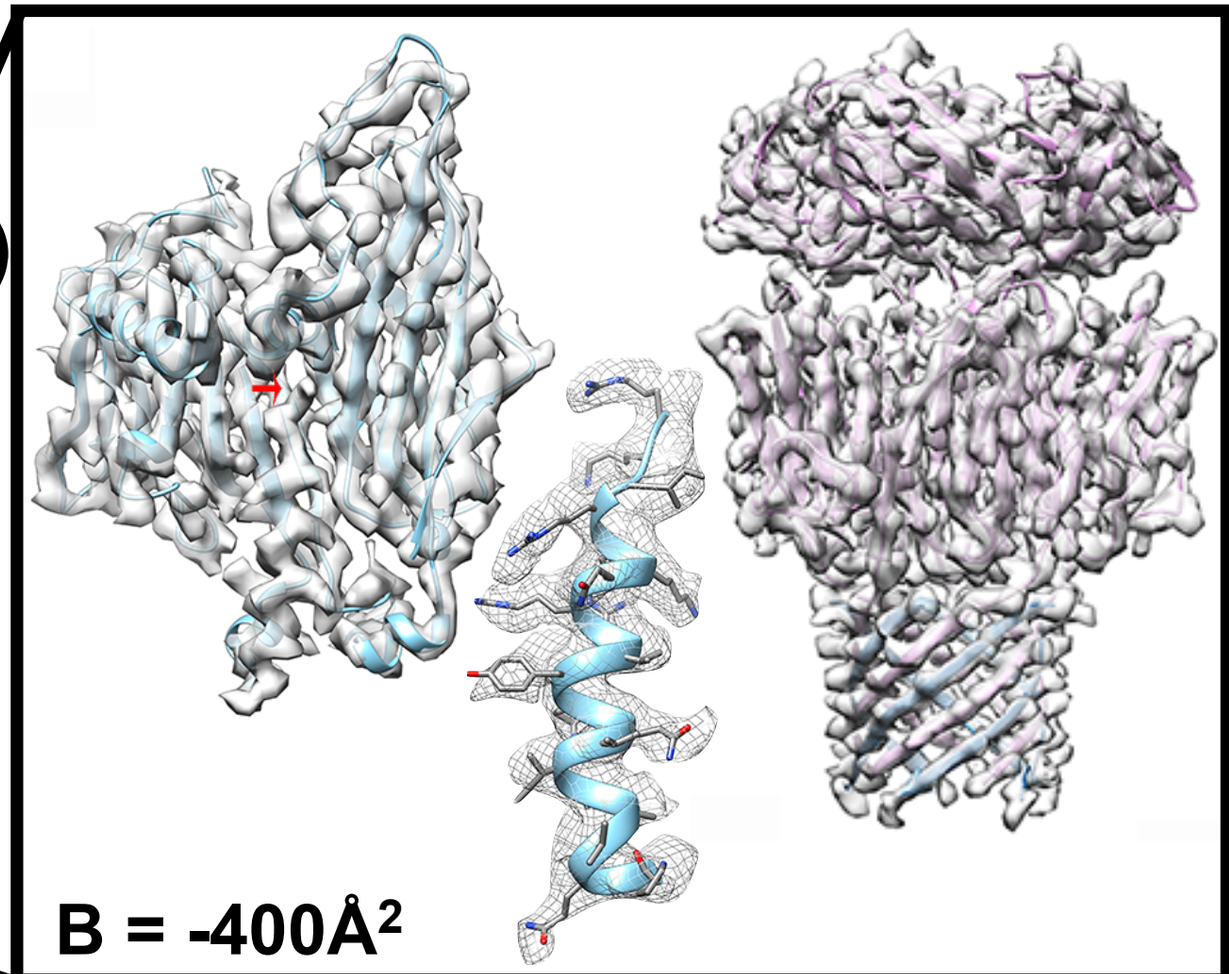
Model building using several maps

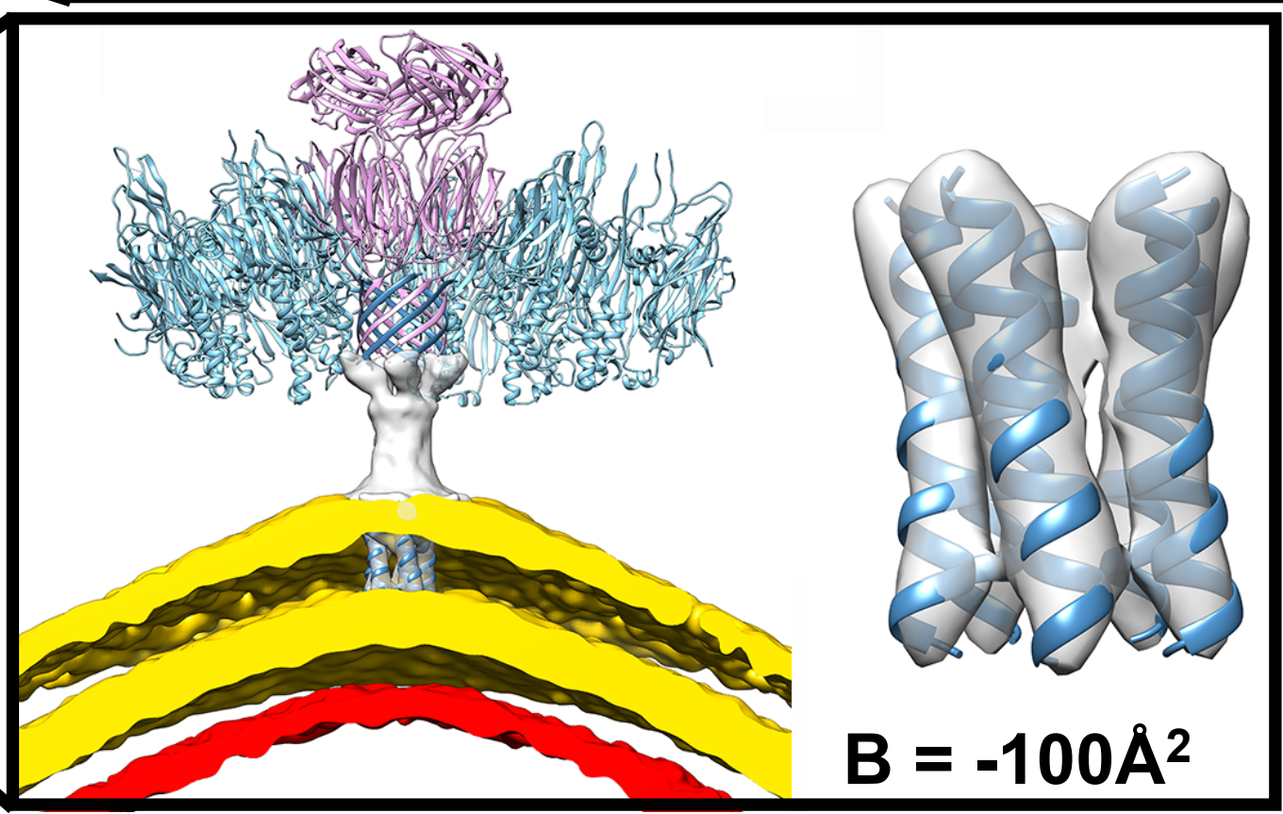
- Different B-factor sharpening
- Or local filtering (sxfilterlocal, …)

STIV
4.5Å resolution



Veesler D et al (2013) PNAS





Which maps should we deposit?

- Recent 3dem/ccpem discussion
- Full map sharpened (current standard)
- Unsharpened map (TRPV1, LliK…)
  - Caveat: end-user needs got know how to sharpen a map
- Half maps
- Mask used for resolution estimation

Model validation

- The model should be stereochemically sound (Molprobity, Privateer)
  - http://molprobity.biochem.duke.edu/
    - Ramachandran statistics
    - Clashes (steric overlaps)
    - Distribution of favored rotamers

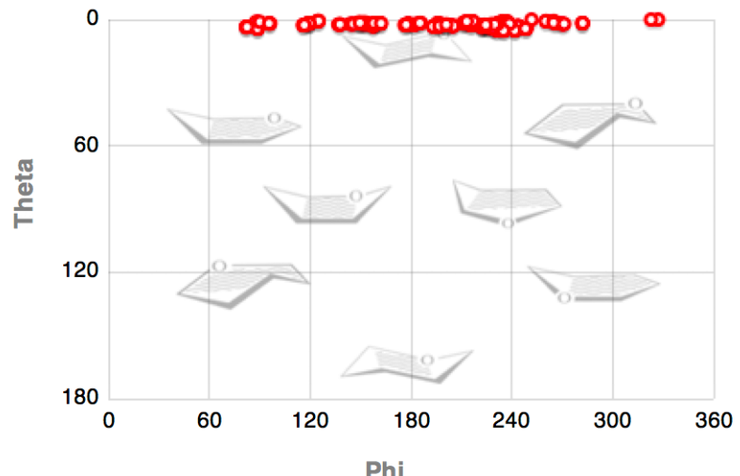
All-Atom Contacts	Clashscore, all atoms:	1.2		100th percentile* (N=744, 1.94Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	0	0.00%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	151	97.42%	Goal: >98%
	MolProbity score^	0.95		100th percentile* (N=11856, 1.94Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad backbone bonds:	0 / 1288	0.00%	Goal: 0%
	Bad backbone angles:	0 / 1734	0.00%	Goal: <0.1%

All-Atom Contacts	Clashscore, all atoms:	135.29		2nd percentile* (N=37, 3Å - 9999Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	4681	46.69%	Goal: <0.3%
	Favored rotamers	3431	34.22%	Goal: >98%
	Ramachandran outliers	3282	28.73%	Goal: <0.05%
	Ramachandran favored	4982	43.62%	Goal: >98%
	MolProbity score^	4.86		1st percentile* (N=342, 3.50Å ± 0.25Å)
	Cβ deviations >0.25Å	9	0.08%	Goal: 0
	Bad bonds:	7 / 94333	0.01%	Goal: 0%
	Bad angles:	490 / 128344	0.38%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	50 / 707	7.07%	Expected: ≤1 per chain, or ≤5%

Chen VB et al (2010) Acta Cryst D

Barad BA et al (2015) Nature Methods

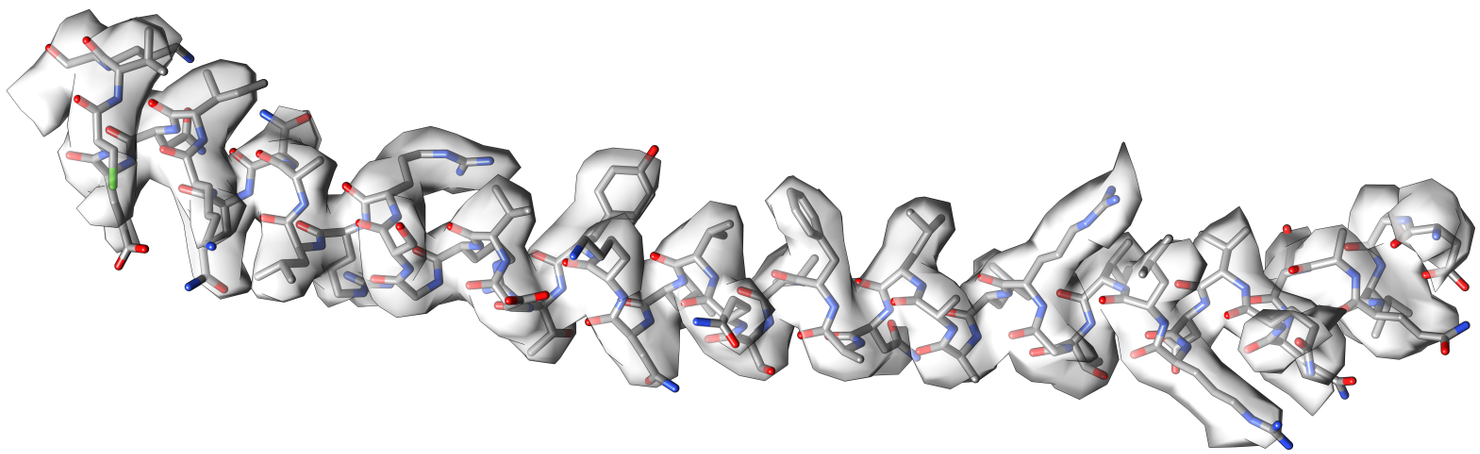
- EMRinger (χ1)
- http://www.ccp4.ac.uk/html/privateer.html



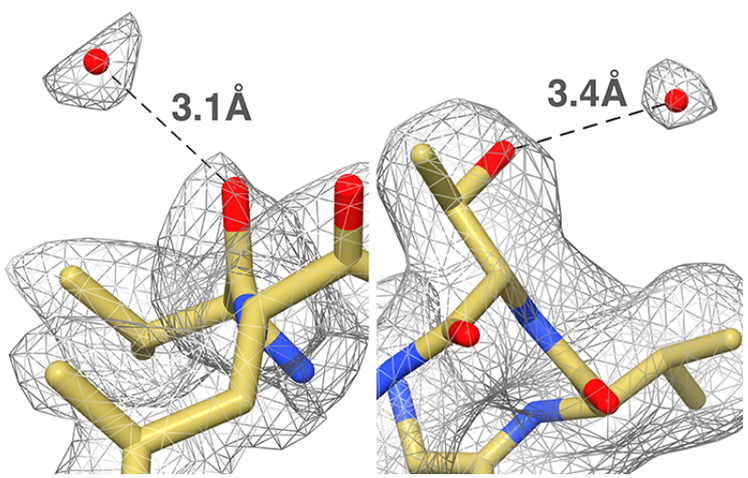
Agirre J et al (2015) Nature Struct Mol Biol

Model validation

- The model should agree with the density



- Half maps are useful



Campbell MG et al (2015) eLIFE

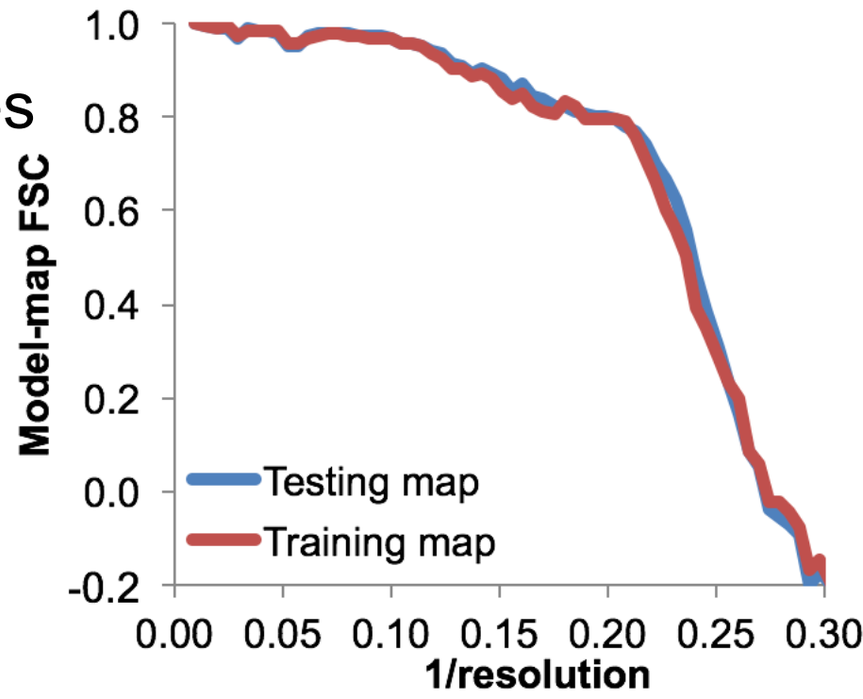
Model validation

- The model should be **<u>cross-validated</u>**
  - Using information that was not used during refinement!
- X-ray crystallography: reciprocal space data
  - Set aside a few percent of reflections (test set: Rfree)
  - Use the rest for model refinement (Rwork)
  - Measures agreement between Rwork and Rfree

Brunger AT (1992) Nature

- CryoEM data: real space data
  - Gold-standard refinement divides dataset in two halves
  - One half used as training (work) map
  - Other half used as testing map (free)
- Last iteration uses all the data (minimization only!)

DiMaio F et al (2013) Protein Science

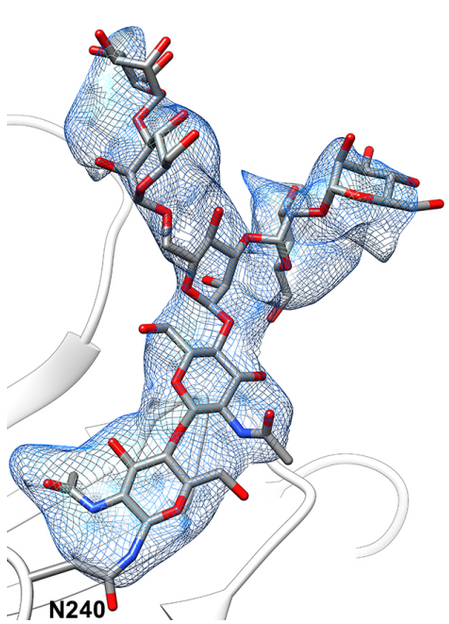


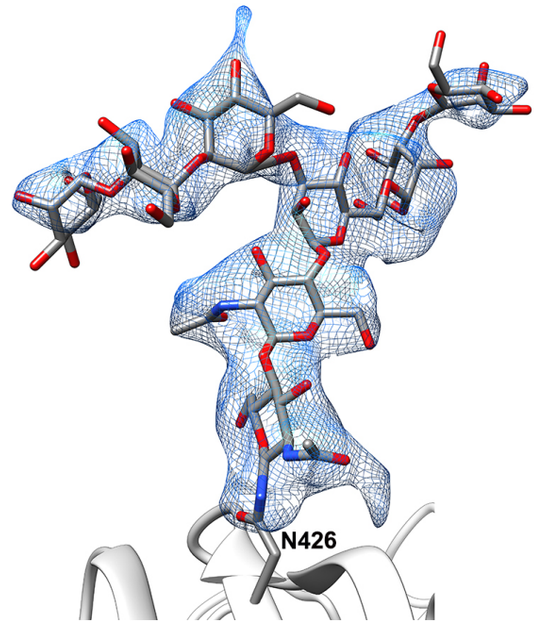
Model validation

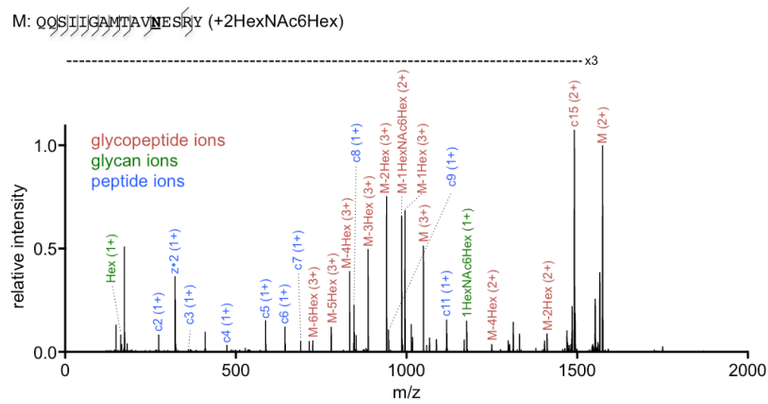
- At low resolution (3-5Å), any additional source of information is most welcome!

- Glycosylation sites

Walls AC et al (2016)
Nature Struct Mol Biol

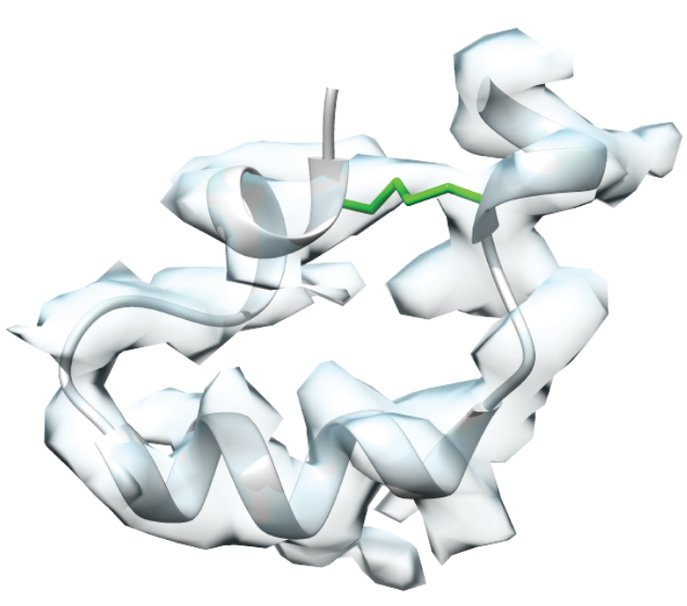


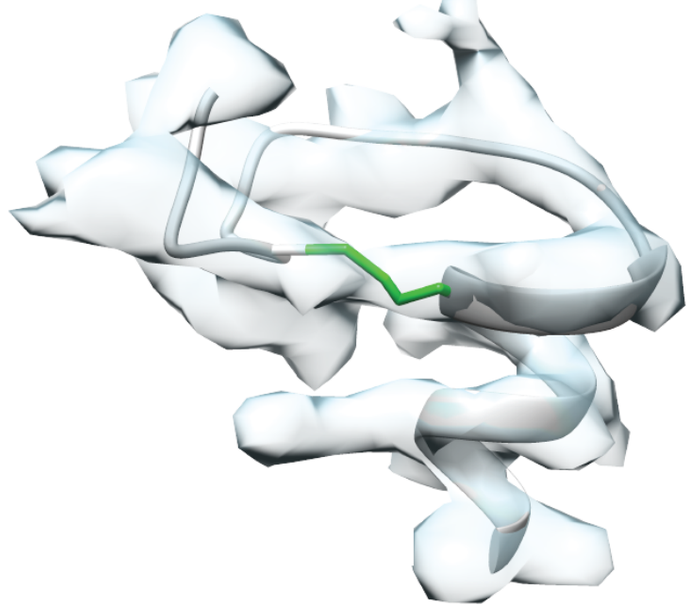




- Disulfide bonds

Walls AC et al (2017) PNAS



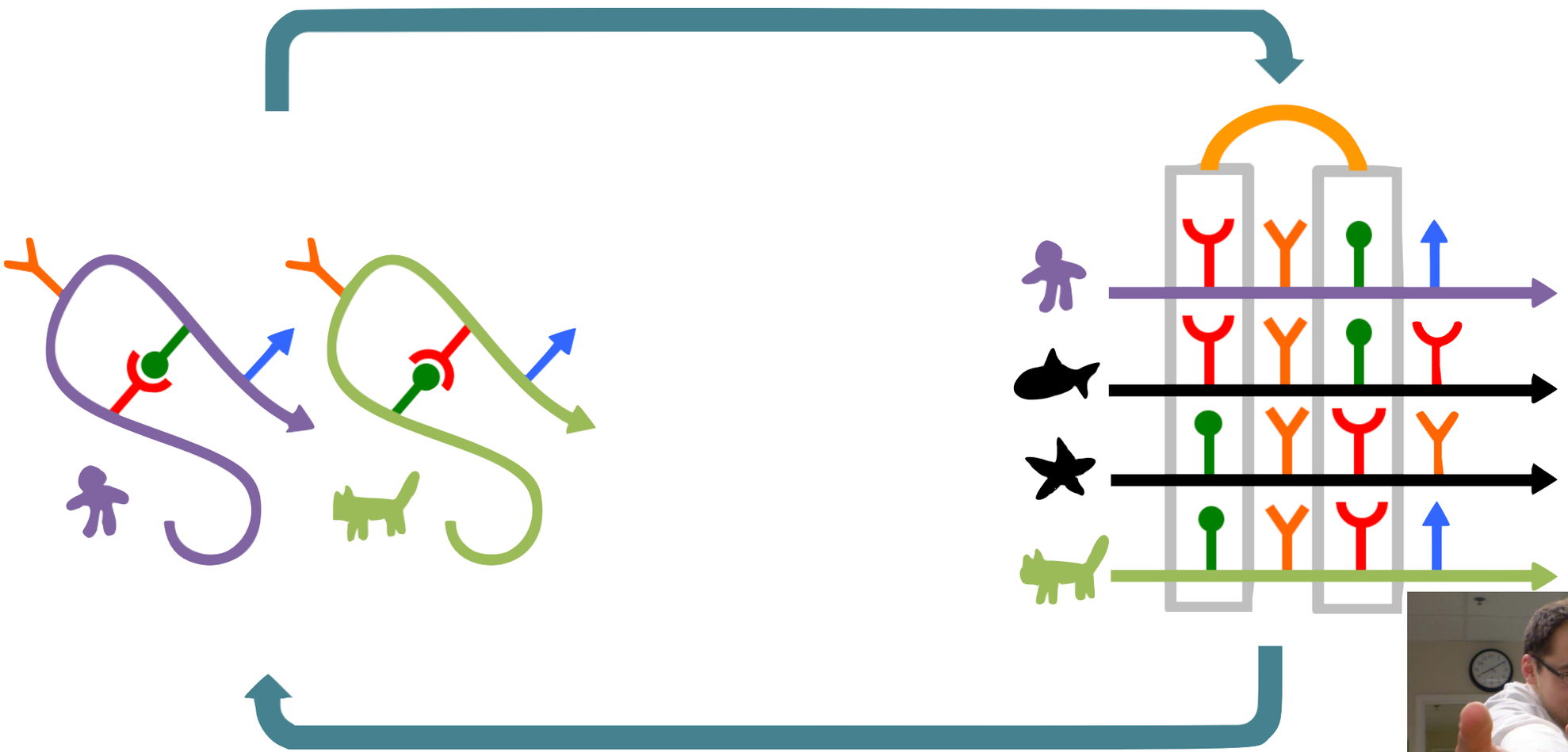


- Cross-linking/MS (can be incorporated in RosettaCM)

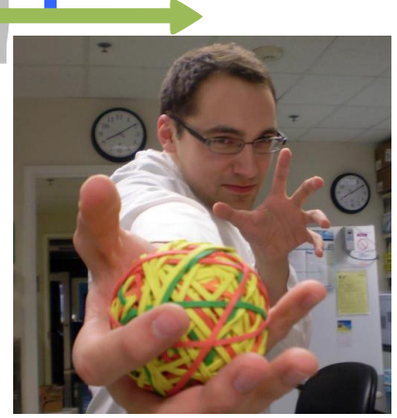
- …

Model validation

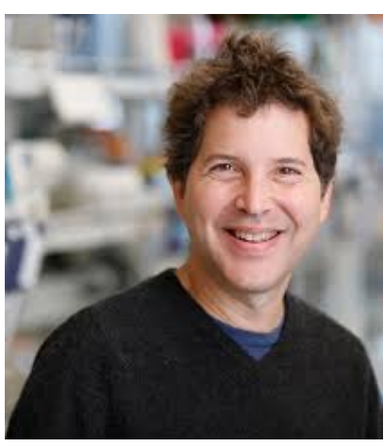
- Coevolution encodes structural information
- Contacts in proteins are evolutionary conserved and encoded in a multiple sequence alignment due to coevolution
- By measuring coevolution, one can infer contacts in proteins



Ovchinnikov S et al (2014) eLIFE
Ovchinnikov S et al (2017) Science

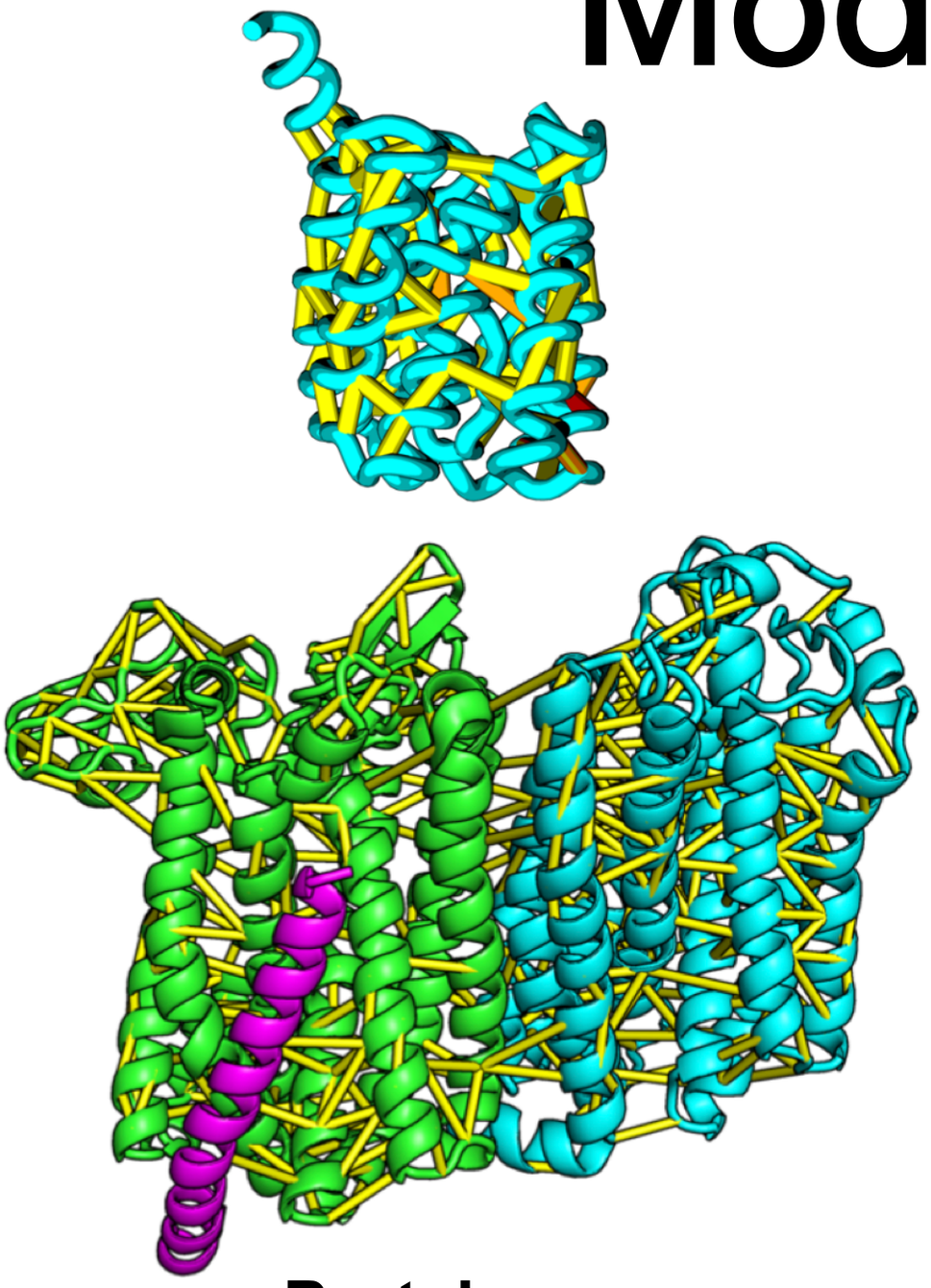


Sergey Ovchinnikov

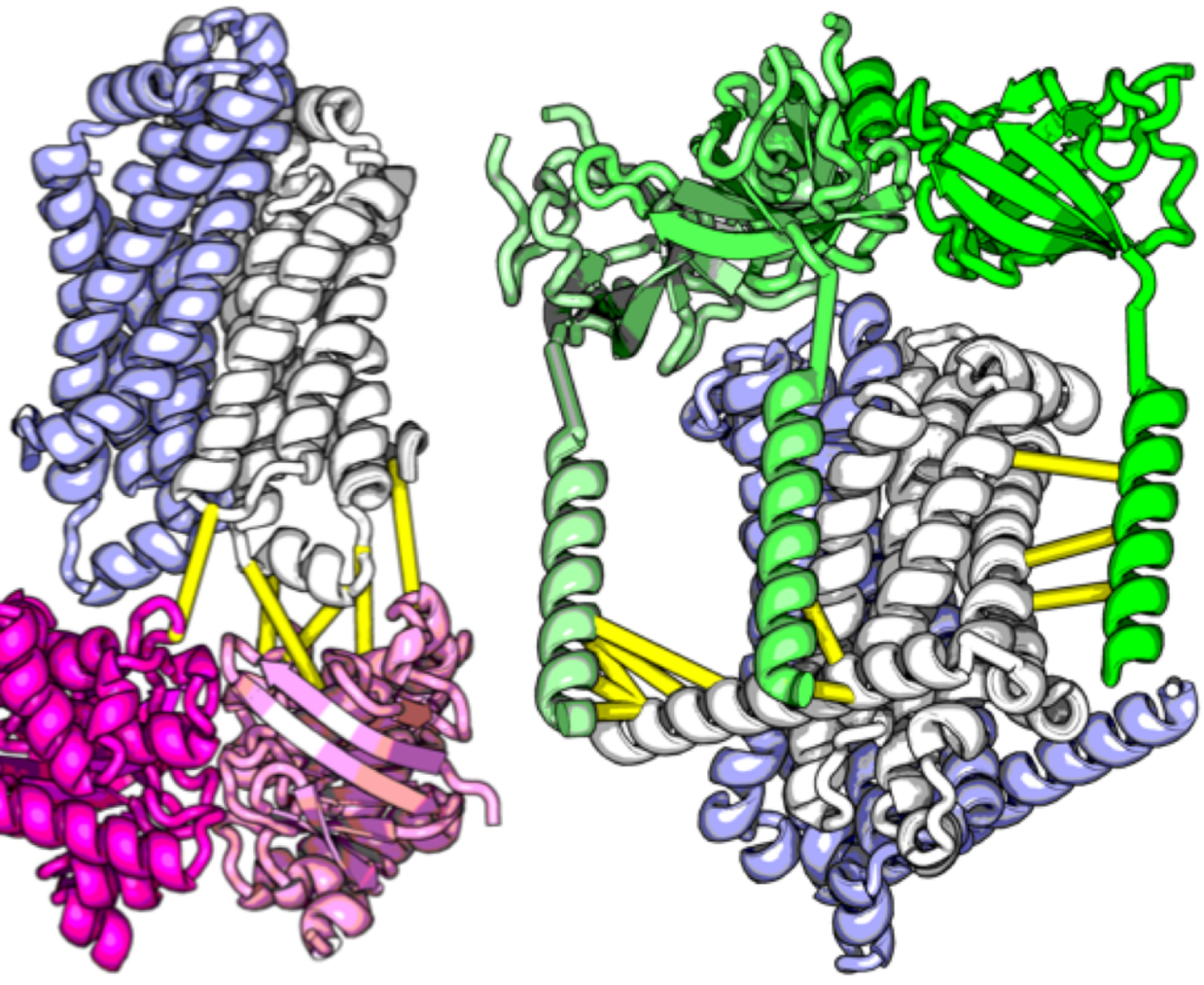


David Baker

Model validation



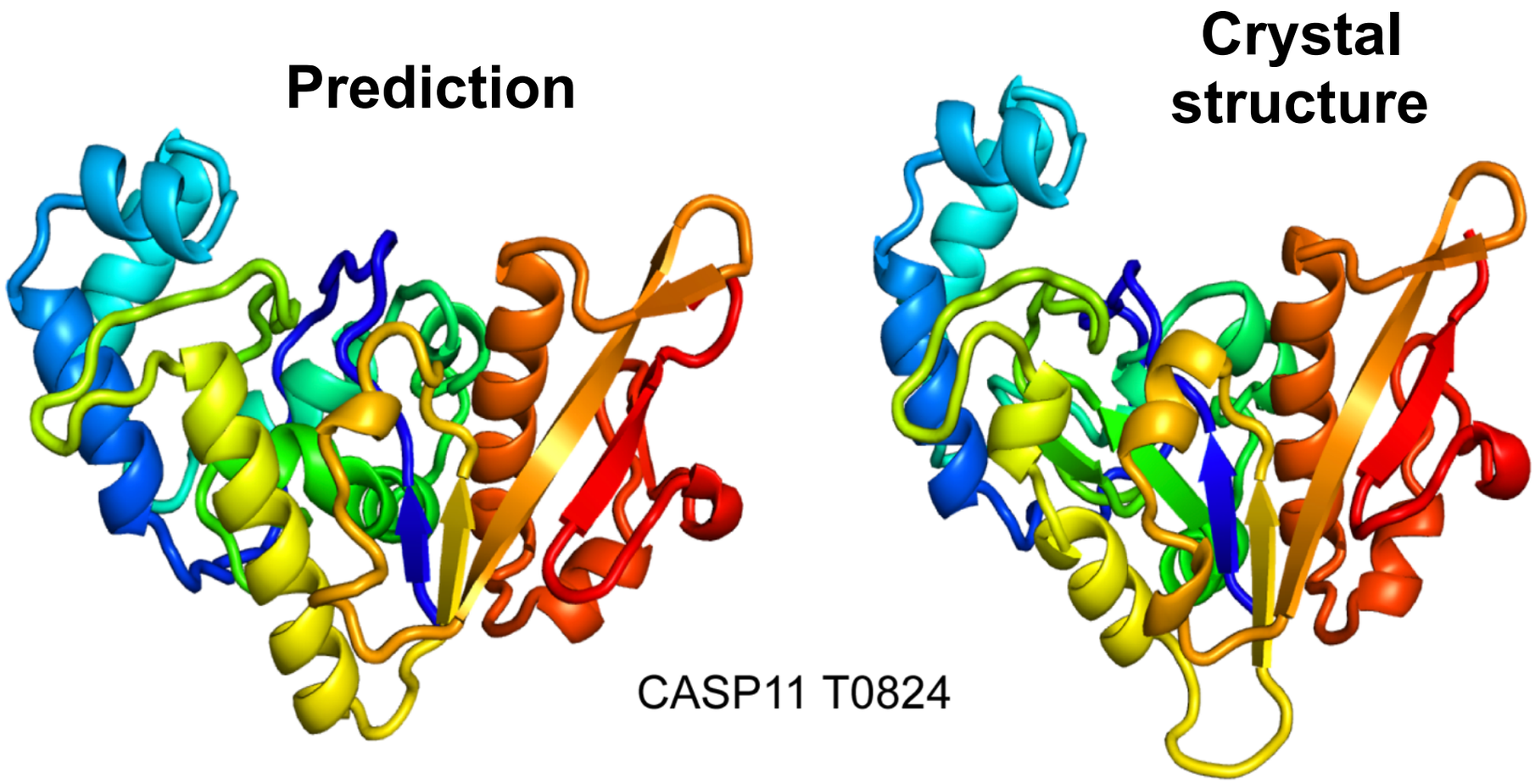
Protein structure prediction



Protein/protein interaction prediction

Ovchinnikov S et al (2014) eLIFE
Ovchinnikov S et al (2017) Science

Model validation



Ovchinnikov S et al (2014) eLIFE
Ovchinnikov S et al (2017) Science

- Could provide an additional metric for cross-validation
  - Currently restricted to prokaryotic sequences

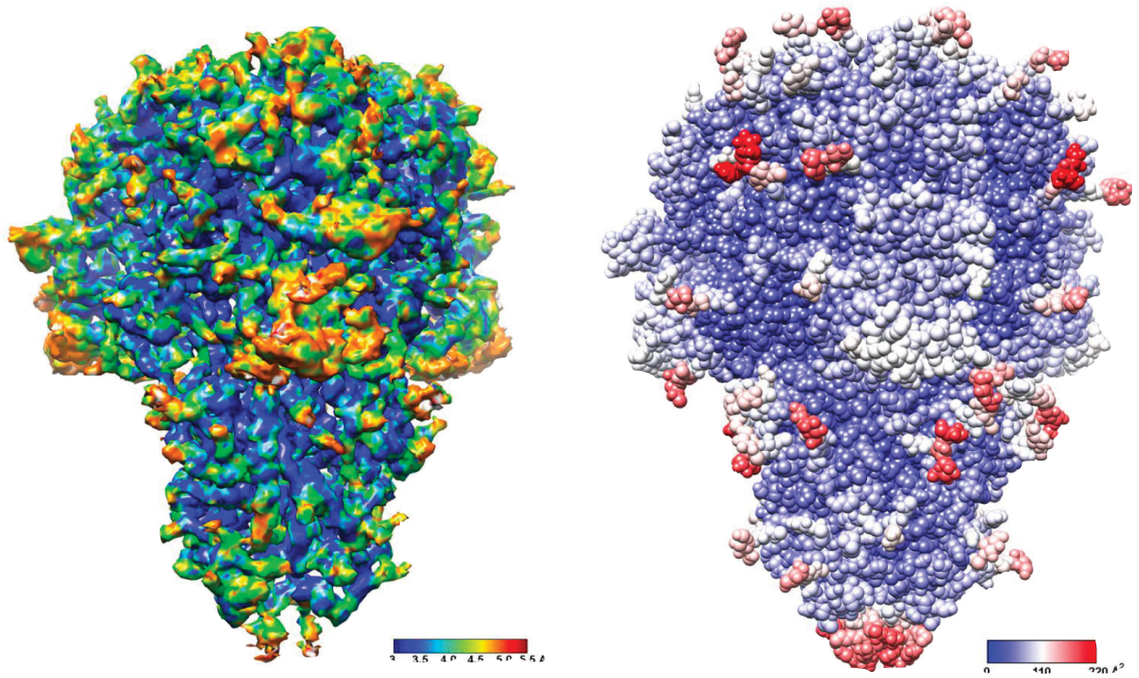
How to deal with uncertainty in model building

- Poorly ordered regions of a map can be:
  - Left unmodeled
  - Modeled and let B-factor account for it
  - Modeled but truncated at Cα or Cβ
  - Modeled with an occupancy of 0
  - A combination of the above
- Best strategy depends of the situation (*de novo* vs rebuilding)
- In any case, write in the manuscript what you have done

Model interpretation

- Model convergence
  - Rosetta ES/CM/density-guided iterative refinement
  - Mark Herzik & Gabe Lander's convergence server
    - https://doi.org/10.1101/128561
    - http://www.lander-lab.com/convergence/

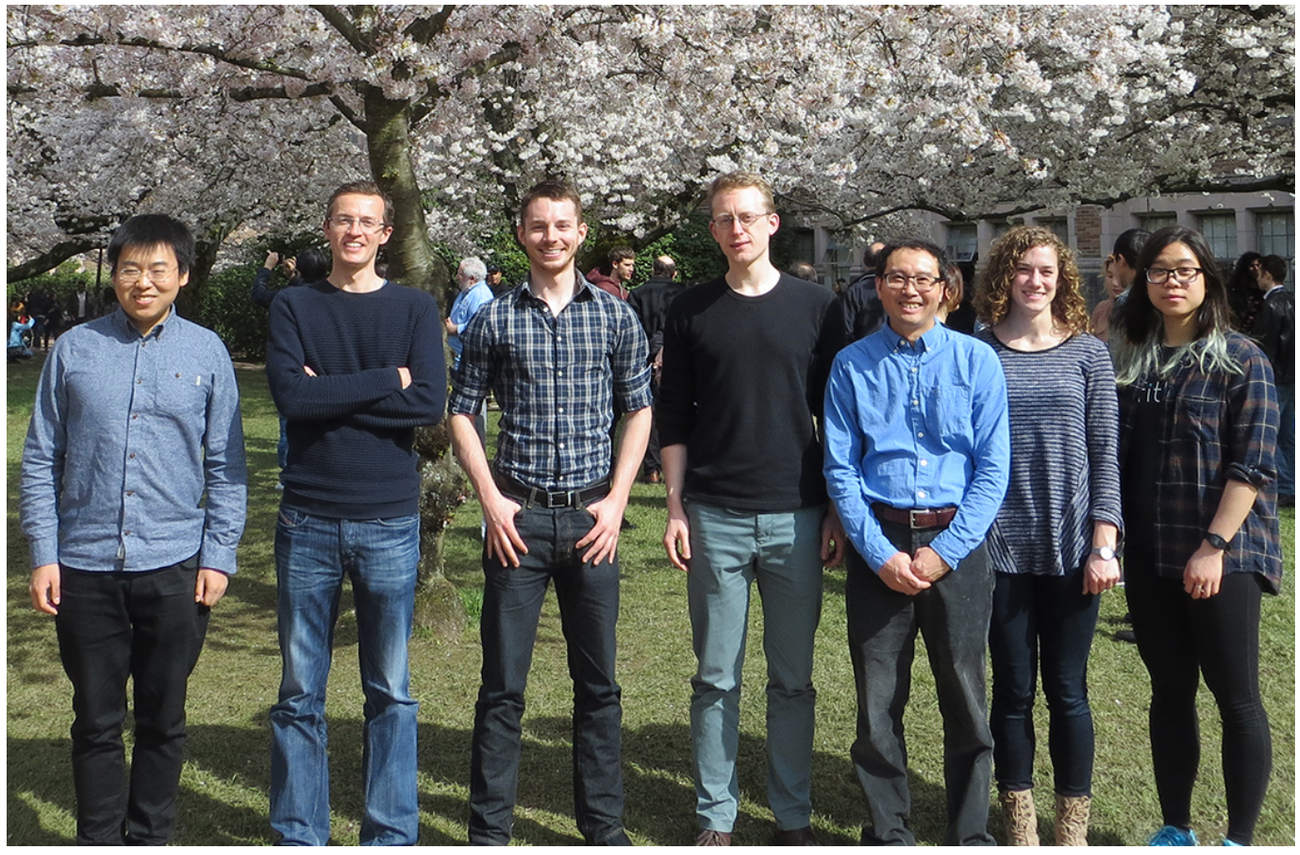
- Agreement between local resolution and B-factor analysis



Walls AC et al (2016) Nature Struct Mol Biol

Acknowledgements

Veesler lab



Collaborators

Frank DiMaio
Brandon Frenz
Felix Rey
M. Alejandra Tortorici
Berend-Jan Bosch
Bill Zagotta
Zachary James
David Baker
Sergey Ovchinnikov

NRAMM
National Resource for Automated Molecular Microscopy

THE PEW CHARITABLE TRUSTS

NIH National Institute of General Medical Sciences

NIH National Institute of Allergy and Infectious Diseases

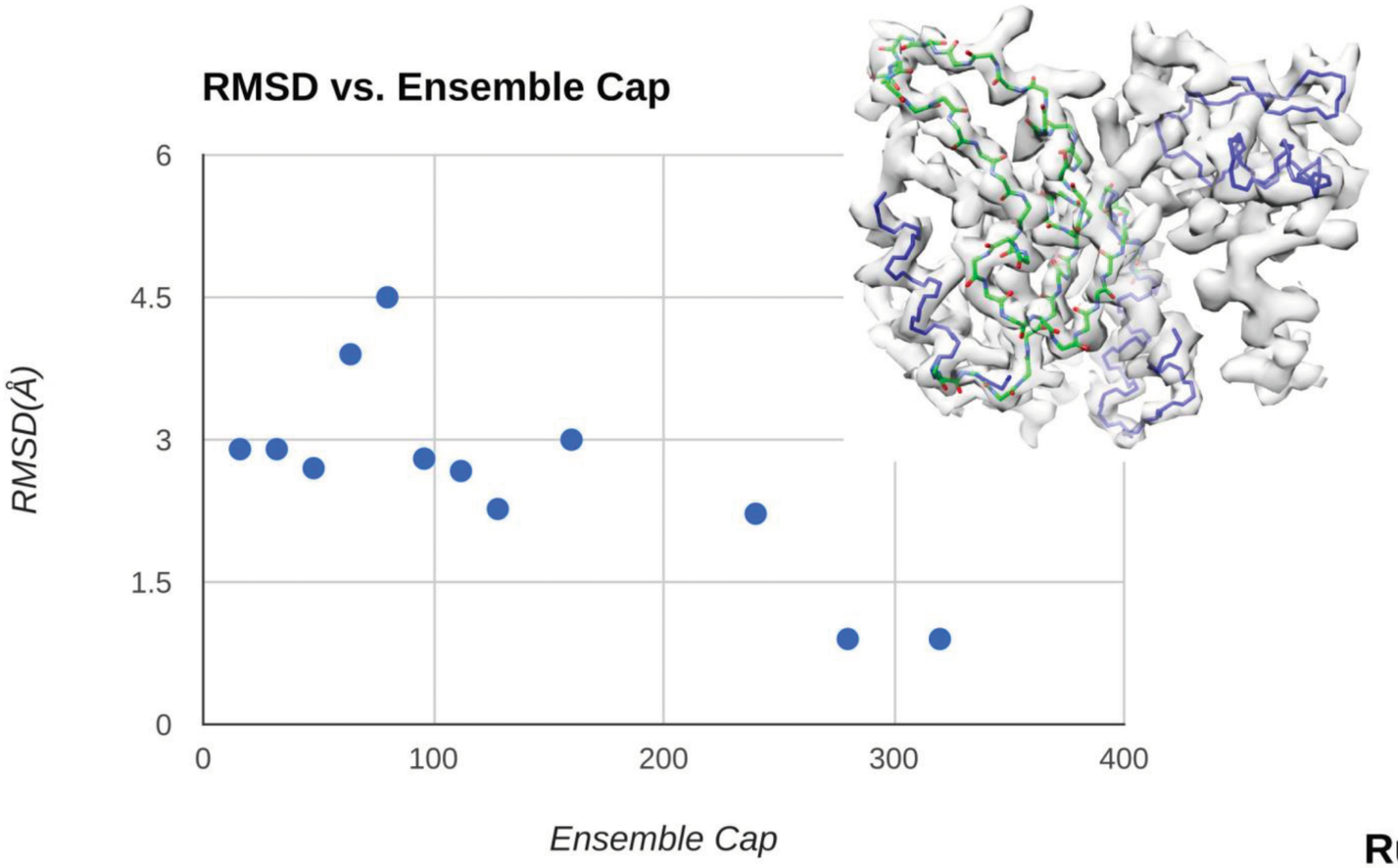
W

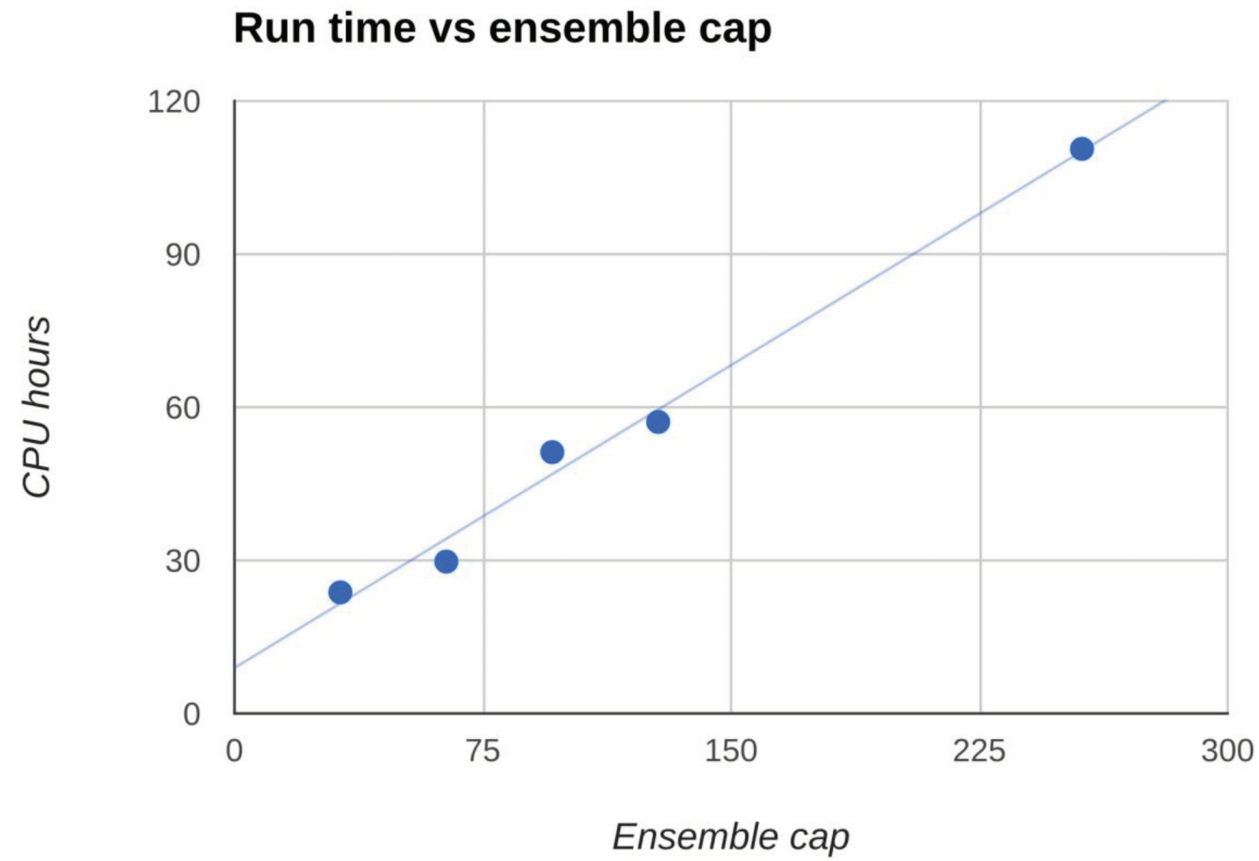
NWO Rubicon

UWPR Advancing Proteomics

eicn Electron Imaging Center for Nanomachines

Rosetta enumerative sampling (RosettaES)





Real-space *B*-factor refinement. To better model the density maps and generate more accurate models, we refined atomic *B* factors against the maps optimizing the real-space correlation between model and map. Given that atom i has a B factor B_i, we calculate the density of the model as

$$\rho_c = \sum_{\text{atoms } i} \left(\frac{\pi}{f_i + B_i/4} \right)^{\frac{3}{2}} \exp\left(-\frac{\pi^2}{f_i + B_i/4} \|x - x_i\|^2 \right)$$

Here, f is a scattering factor fit to each element. Our implementation makes use of a single-Gaussian scattering for each atom type, but it is straightforward to extend this to a standard five-Gaussian scattering model[27].

B-factor refinement is carried out using quasi-Newton optimization, with the gradient of the B factor of atom i (located at coordinates x_i) given in real space by

$$\frac{\partial \text{RSCC}}{\partial B_i} = \frac{1}{\sigma_c^2} \left(\sigma_c \frac{\partial \sum \rho_c \rho_o}{\partial B_i} - \sum \rho_c \rho_o \frac{\partial \sum \rho_c^2}{\partial B_i} \right)$$