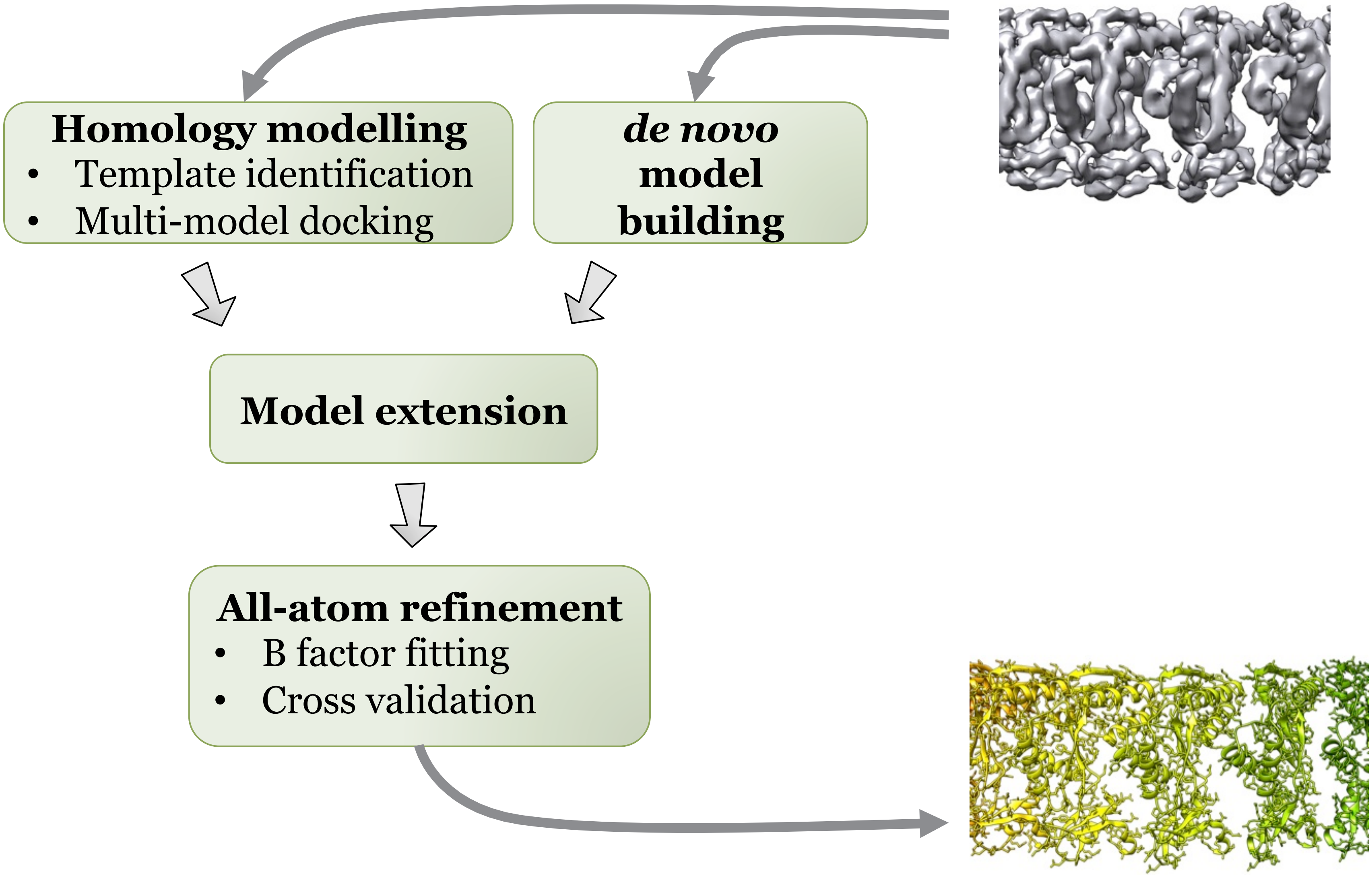
Toward automated structure determination from near-atomic resolution data

Frank DiMaio
University of Washington
Institute for Protein Design

November 2014

Accurate structure determination with RosettaEM



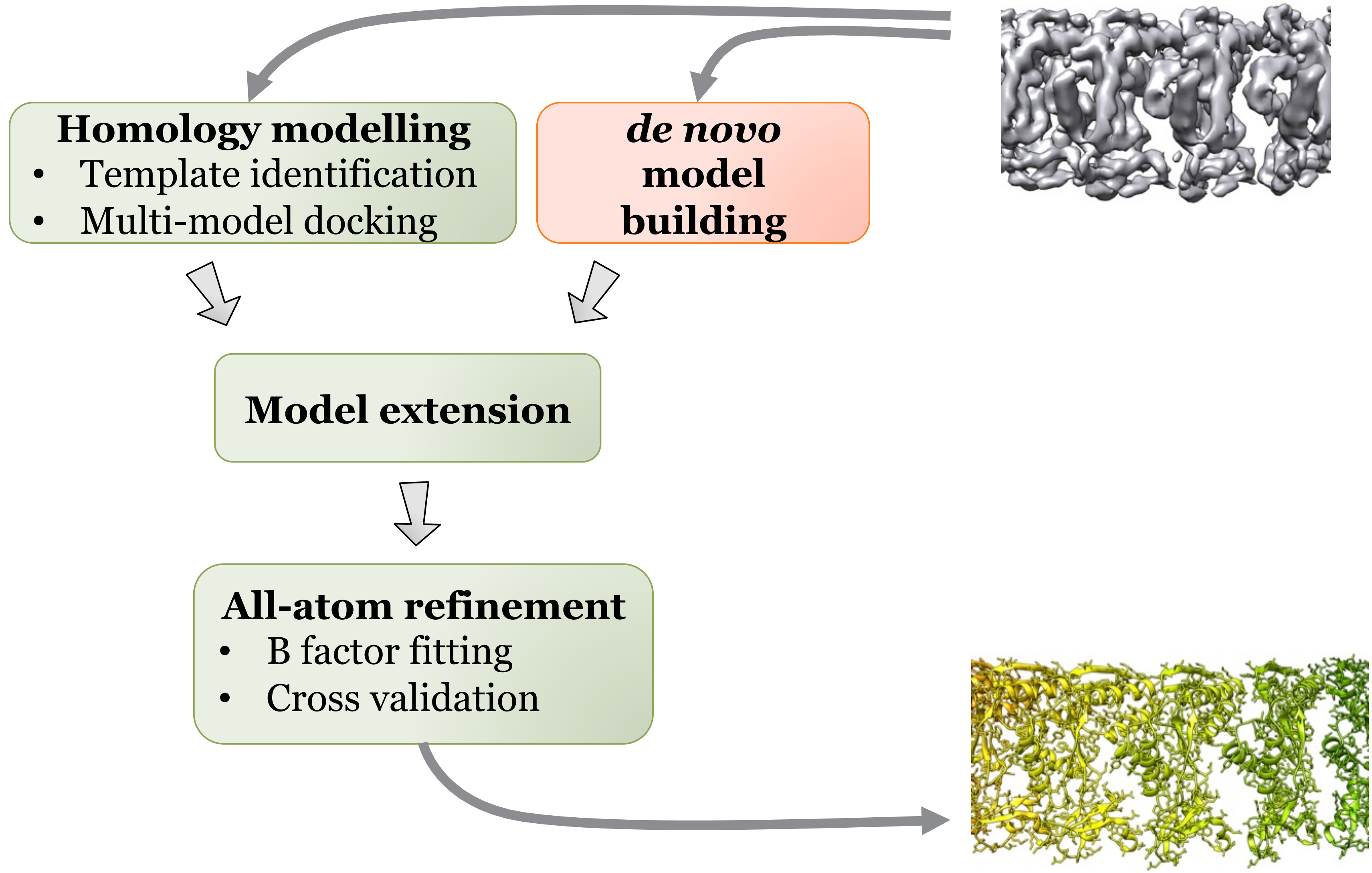
Homology modelling
- Template identification
- Multi-model docking

***de novo* model building**

Model extension

All-atom refinement
- B factor fitting
- Cross validation

Accurate structure determination with RosettaEM



Homology modelling
- Template identification
- Multi-model docking

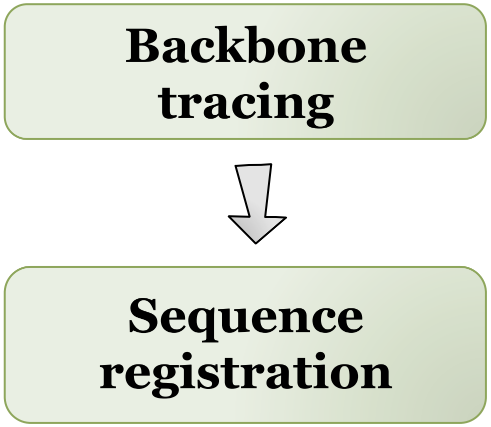
***de novo* model building**

Model extension

All-atom refinement
- B factor fitting
- Cross validation

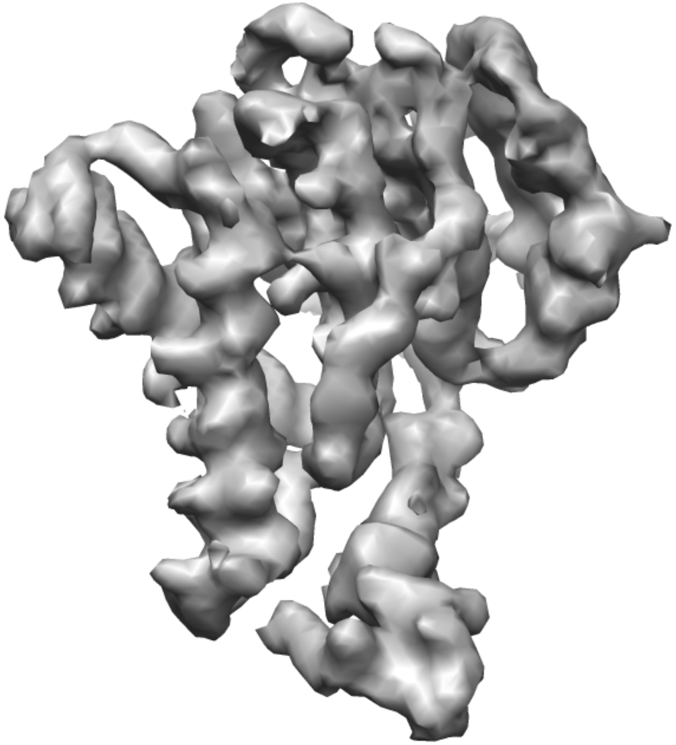
Lack of sidechain detail makes identifying sequence difficult

Crystallographic “autotracing”:



Backbone tracing

Sequence registration

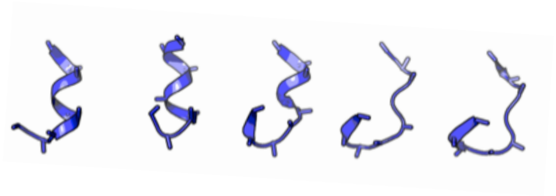


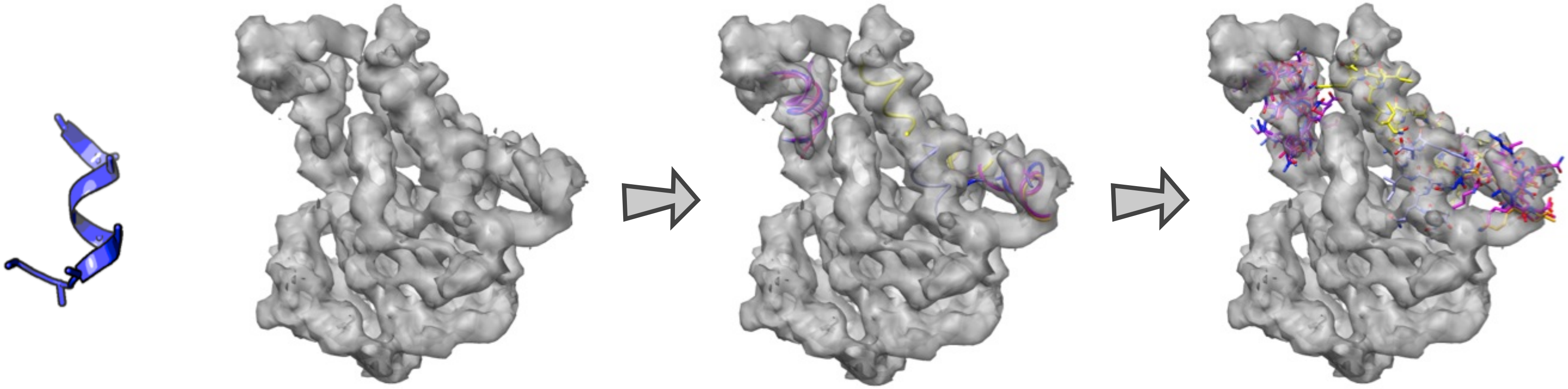
4.8Å reconstruction
20S proteasome
(courtesy Yifan Cheng & Xueming Li)

Searching density for local backbone conformations

Local sequence restricts local structure

…CVKVTKPLVARAKL…





6-dimensional search

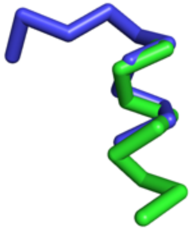
sidechain building & refinement

Ray Wang (in review)

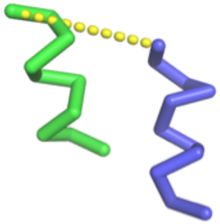
Selecting a maximally consistent set of fragments

Idea: The correct placements must all be consistent

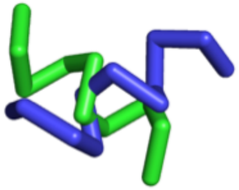
- adjacent fragments must assign the same residue to the same location



- residues close in sequence must be close in space



- no two residues can occupy the same space

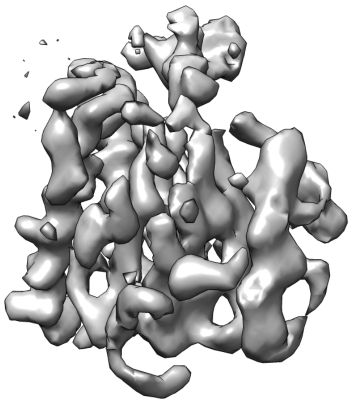


$$score(\boldsymbol{F}) =$$

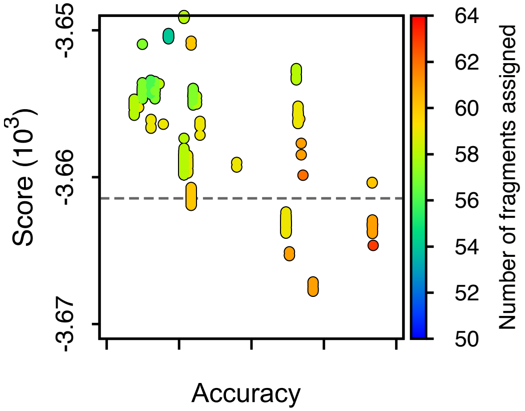
$$\sum_{f_i \in \boldsymbol{F}} sc_{dens}(f_i) + \sum_{f_i, f_j \in \boldsymbol{F}} sc_{overlap}(f_i, f_j) + \sum_{f_i, f_j \in \boldsymbol{F}} sc_{close}(f_i, f_j) + \sum_{f_i, f_j \in \boldsymbol{F}} sc_{clash}(f_i, f_j)$$

Monte Carlo sampling correctly identifies sequence

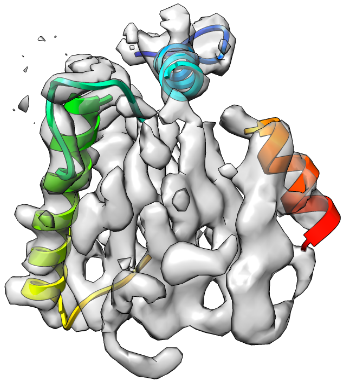
Density Map



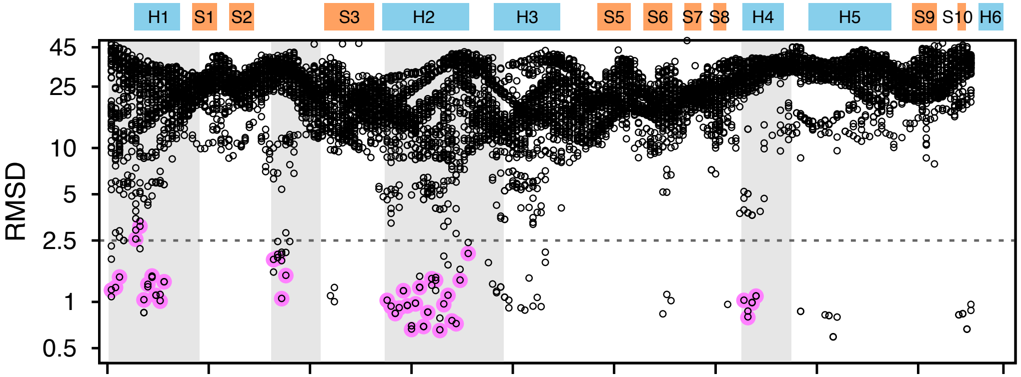
Monte Carlo Sampling



Partial Model

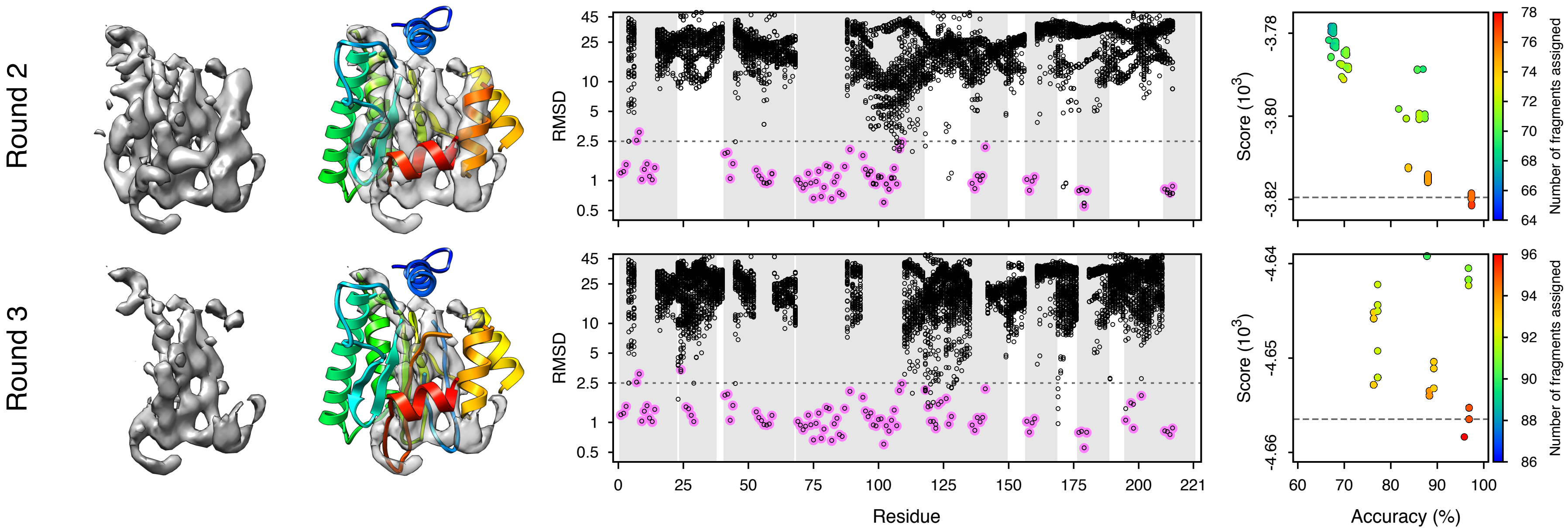


Fragment Placement



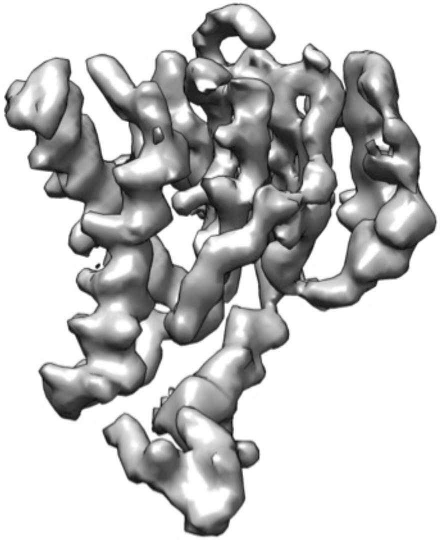
Multiple rounds of sampling completes model

Density Map Partial Model Fragment Placement Monte Carlo



20S proteasome α-subunit at 4.8 Å

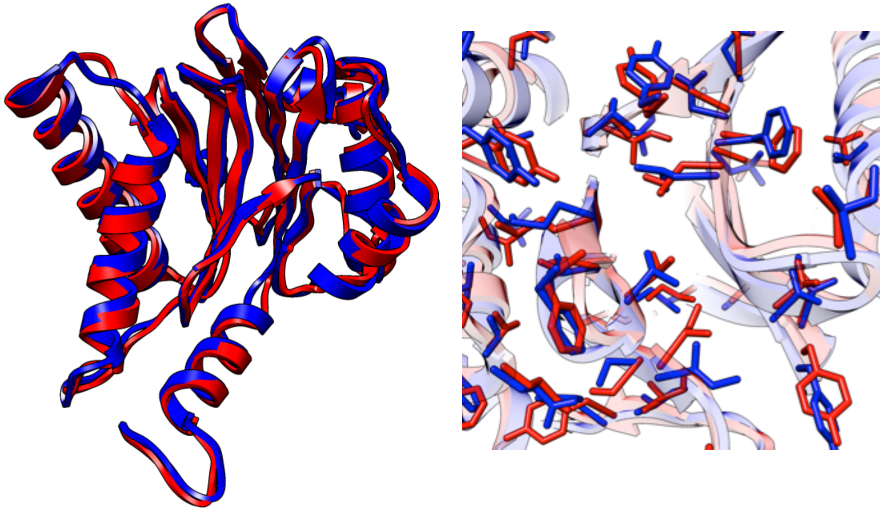
Density



Final Partial Model

1.28 A
196/213 rsds

Overlay of the fulllength model (red) to the native (blue)



1.19 A

Automatic structure determination is accurate in 6 of 9 cases

Target	PDB ID (chain)	EMDB ID	Reported resolution (Å)	Length (aa)	Partial model Cα RMSd [Å] (%)	Cα RMSd [Å]
TMV	3j06 (A)	5185	3.3	155	1.3 (**81**)	**1.7**
TRPV1	3j5q (A)	5778	3.4	310	1.1 (**76**)	**1.4**
FrhA	4ci0 (A)	2513	3.4	385	2.3 (**91**)	**1.3**
FrhB	4ci0 (C)	2513	3.4	280	1.4 (**85**)	**1.7**
FrhG	4ci0 (B)	2513	3.4	228	1.6 (**73**)	**2.2**
BPP1	3j4u (A)	5764	3.5	327	17.2 (42)	-
VP6	1qhd (A)	1461	3.8	397	1.6 (52)	-
20S-α	1pma (A)	TBD	4.8	221	1.3 (**88**)	**1.2**
STIV	3j31 (A)	5584	3.9	344	21.9 (26)	-

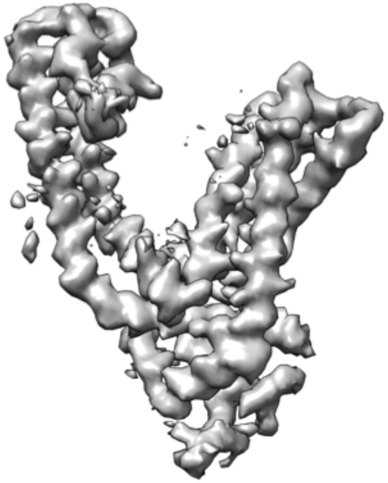
Automatic structure determination is accurate in 6 of 9 cases

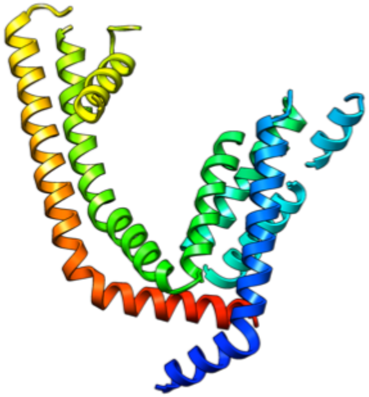
Density

Final Partial Model

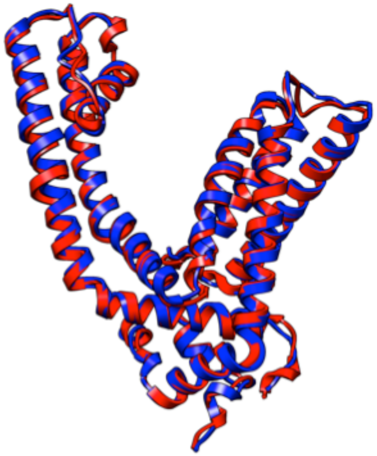
Overlay of the fulllength model (red) to the native (blue)

TRPV1
3.4 Å

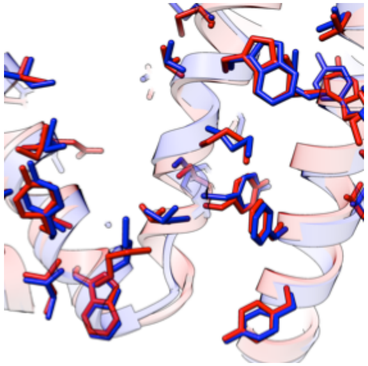




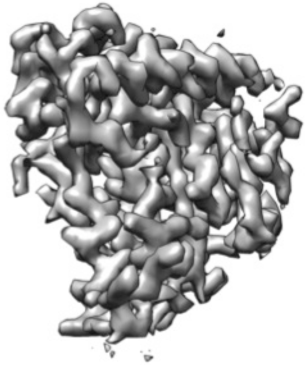
1.26 A
74.9% (236/315 rsds)

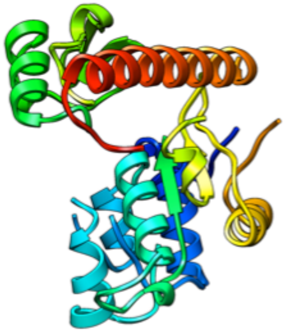


1.43 A

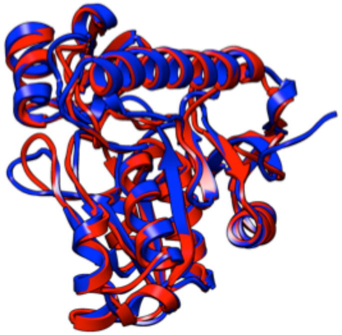


FrhB
3.4 Å

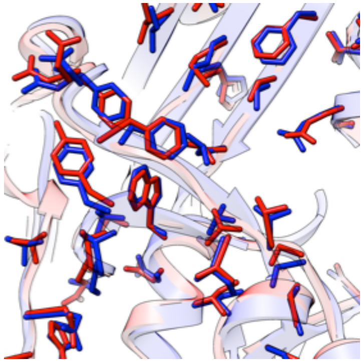




1.40 A
85.1% (239/281 rsds)



1.62 A



Crystallographic chain tracing is generally unable to register sequence

Using Buccaneer:

Target	PDB ID (chain)	Length (aa)	Cα atom placed	Sequence registered	Correctly registered
TMV	3j06 (A)	155	145	56	0
TRPV1	3j5q (A)	315	257	190	0
FrhA	4ci0 (A)	386	382	367	185 (**48%**)
FrhB	4ci0 (C)	281	192	186	126 (**45%**)
FrhG	4ci0 (B)	228	242	190	63 (**27%**)
BPP1	3j4u (A)	327	339	162	0
VP6	1qhd (A)	397	405	155	0
20S-α	1pma (A)	221	224	135	7 (**3%**)
STIV	3j31 (A)	345	553	259	0

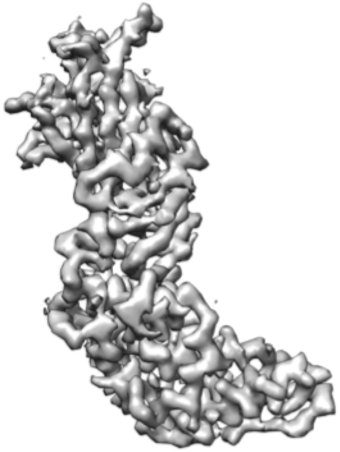
Failures are primarily in sheets

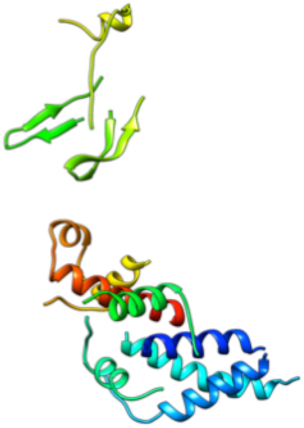
Density

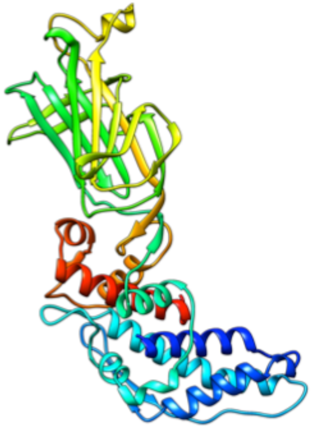
Partial Model

Native

Rotavirus-vp6
3.8 Å

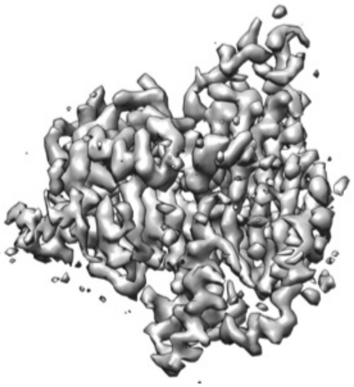


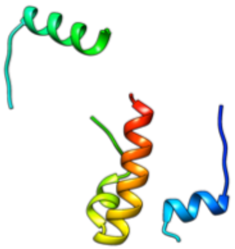


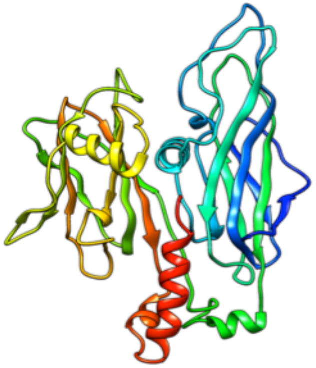


1.62 A
52.1% (207/397 rsds)

STIV
3.9 Å



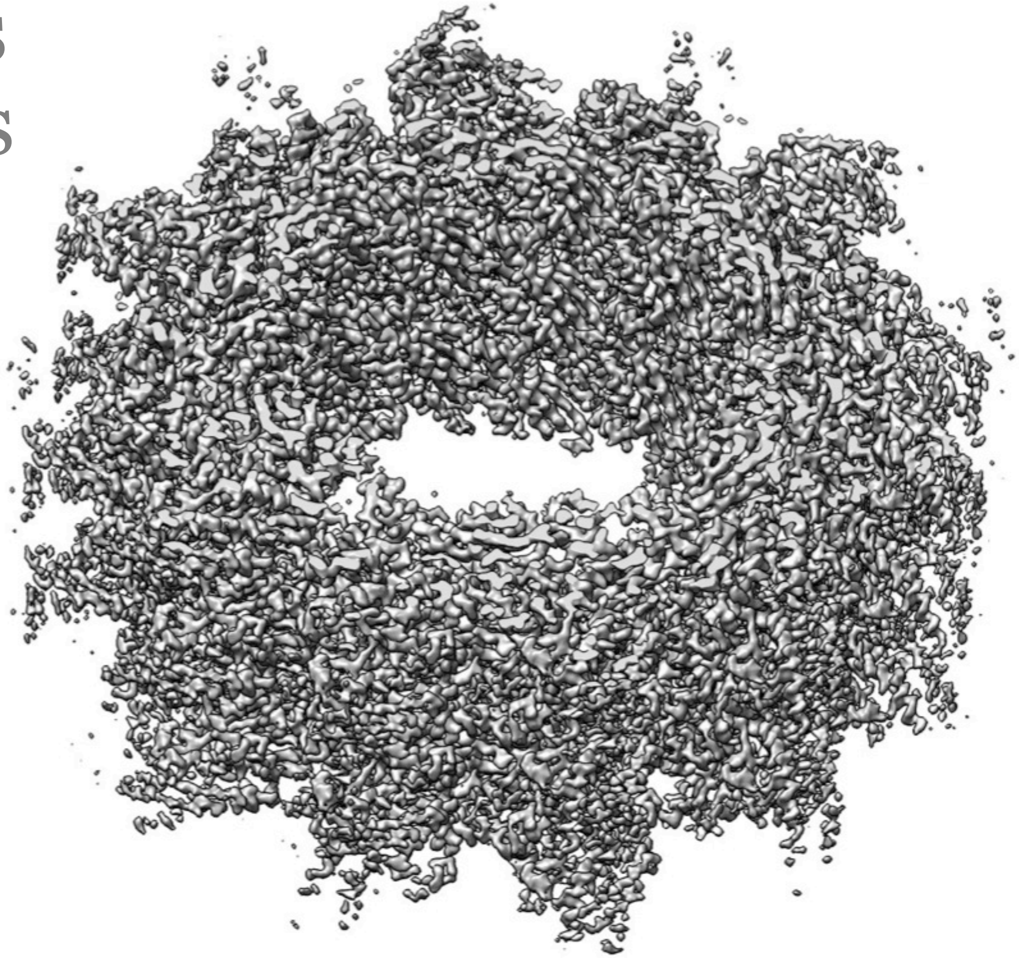




2.46 A
20.0% (69/345 rsds)

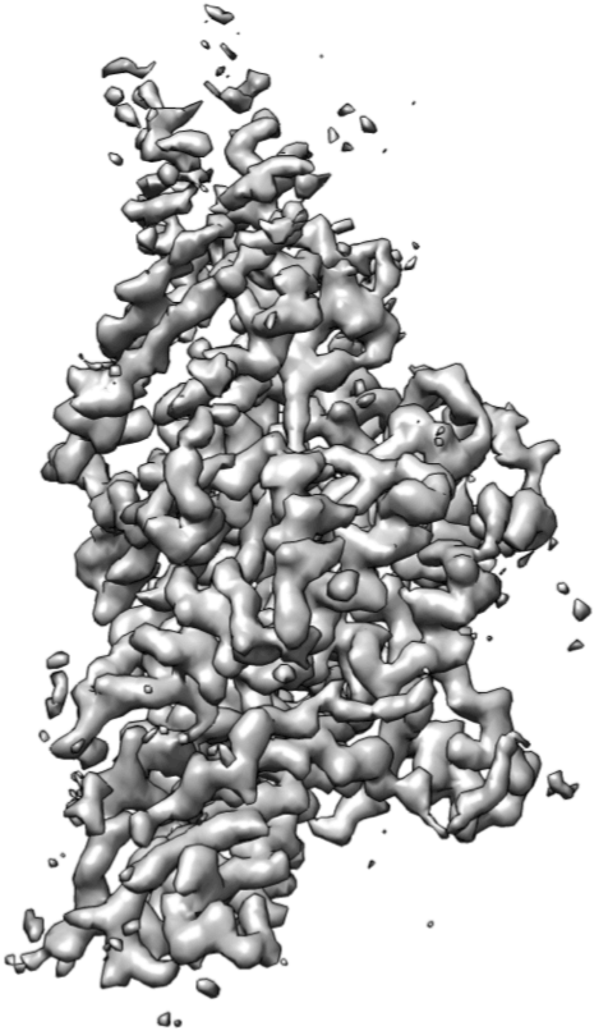
VipAB structure determination

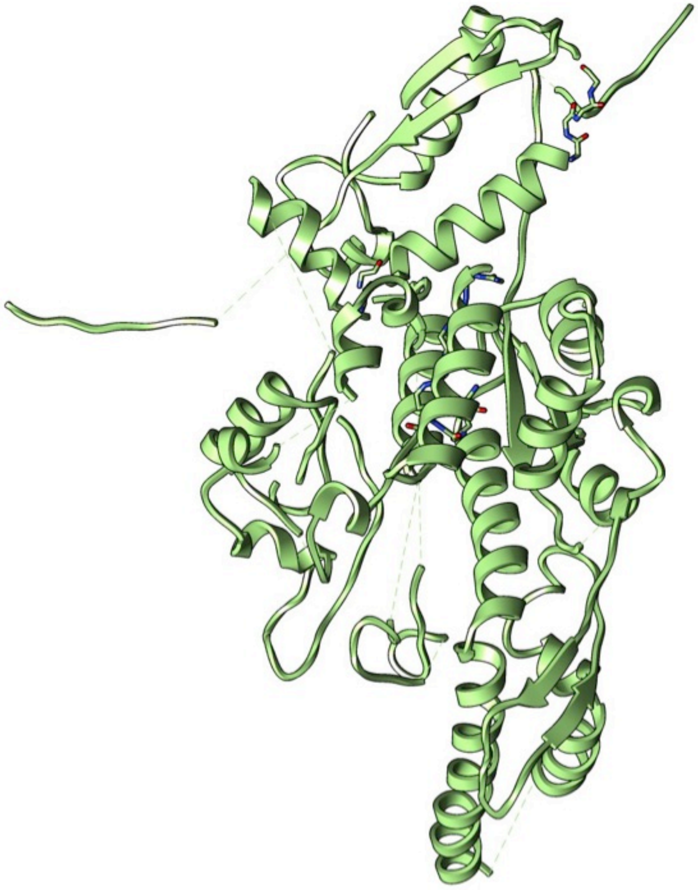
VipA: 168 residues
VipB: 492 residues



with Misha Kudryashev, Marek Basler, Ed Egelman (*in review*)

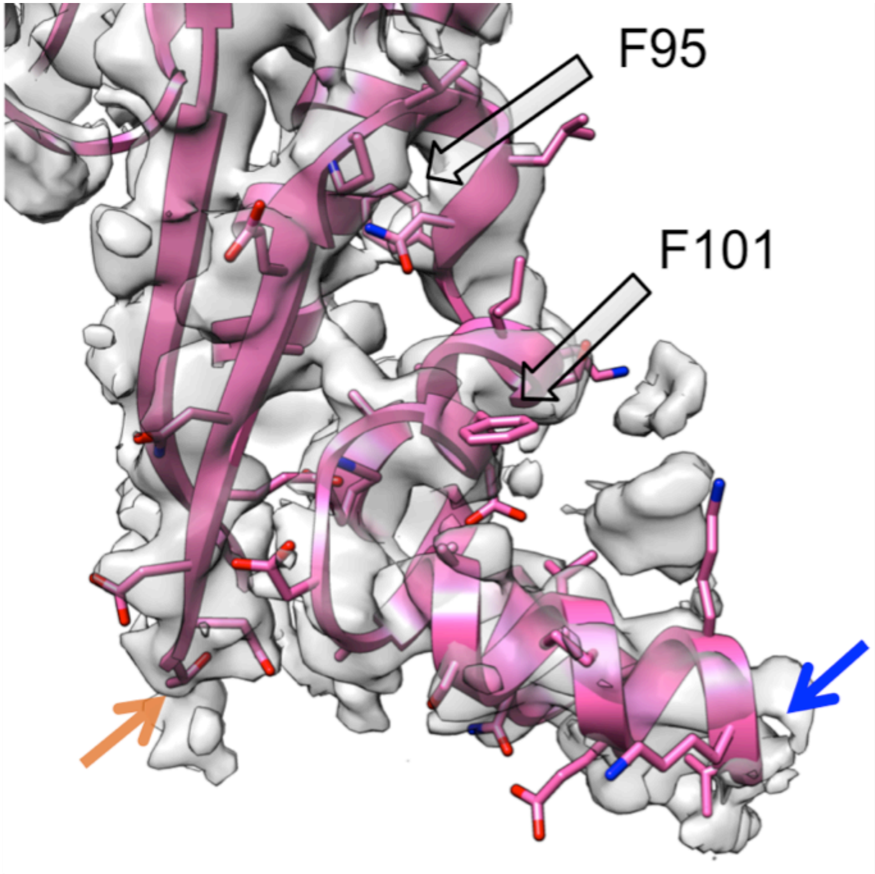
VipAB structure determination



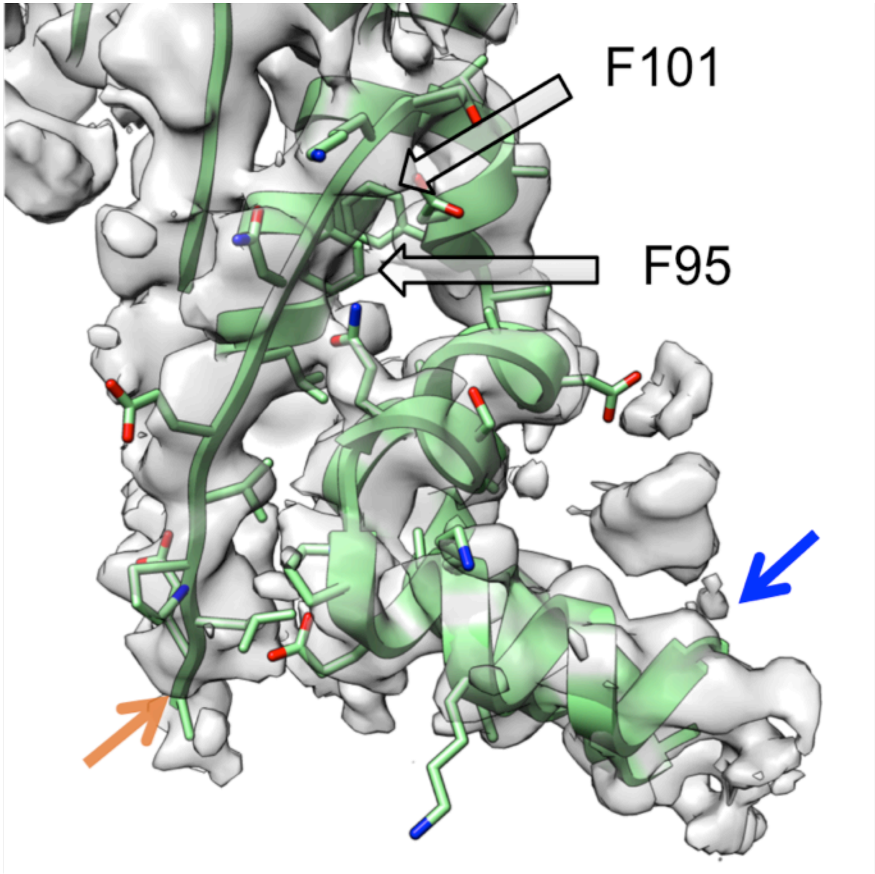


446/660 residues

Our method corrects errors from the manually traced model

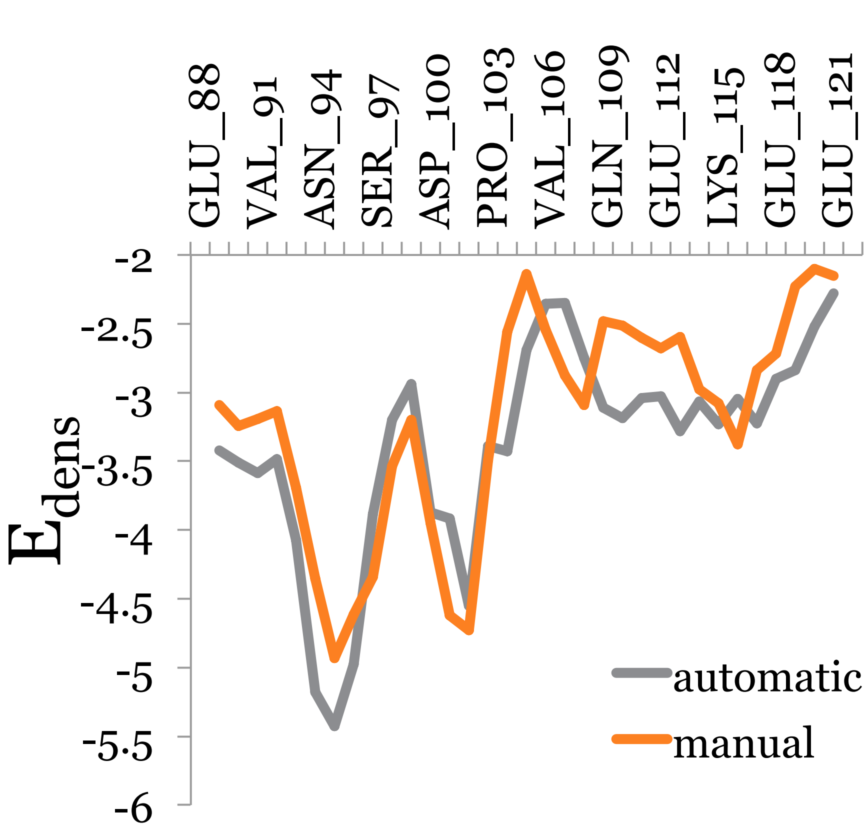


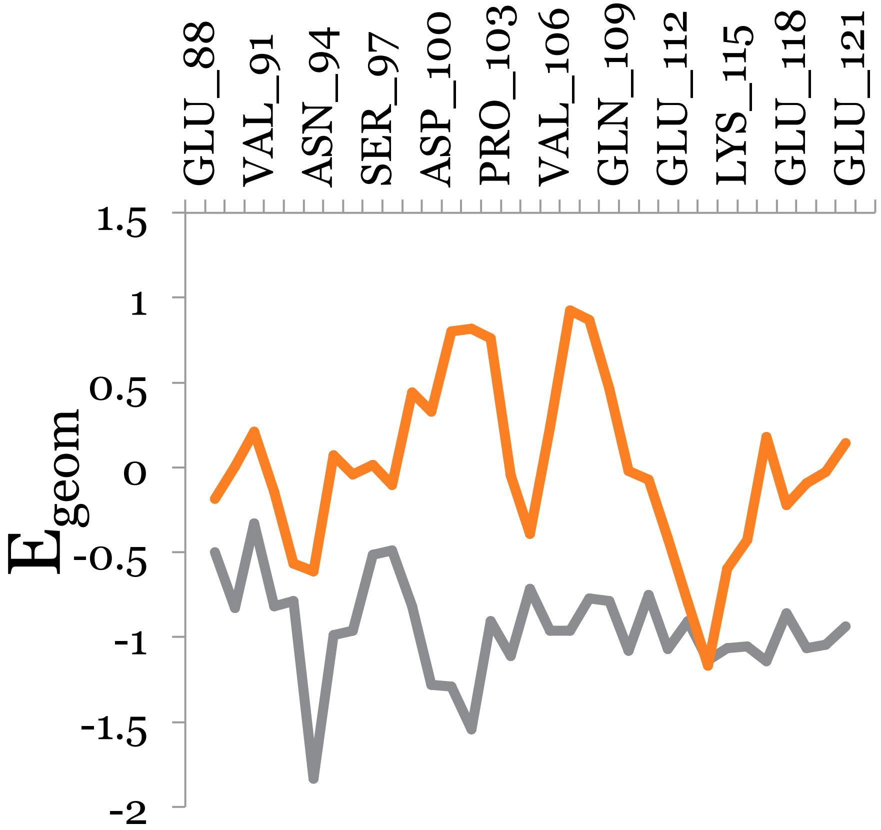
manual model



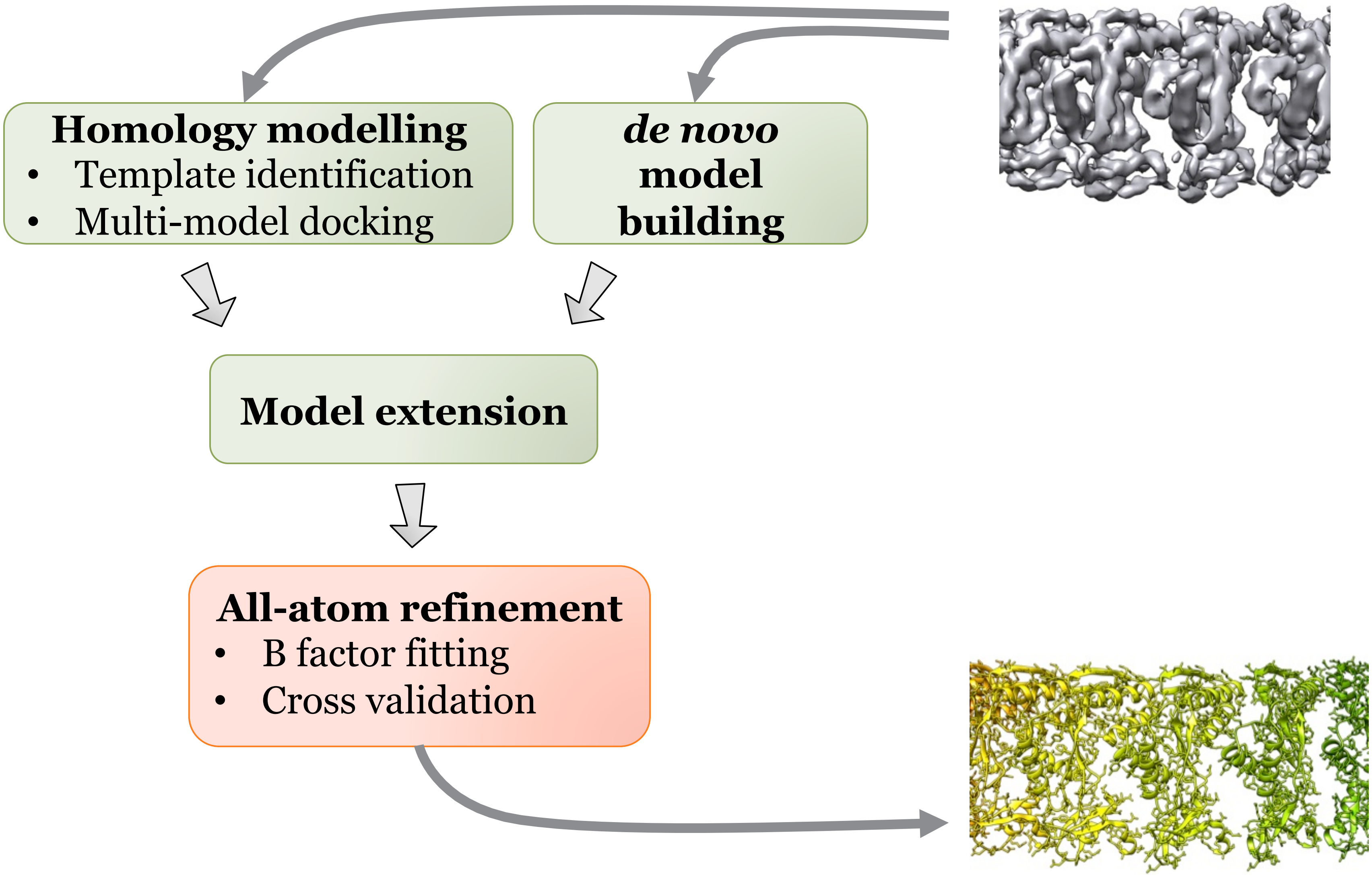
Automated model

Our method corrects errors from the manually traced model





Accurate structure determination with RosettaEM



Homology modelling

- Template identification
- Multi-model docking

***de novo* model building**

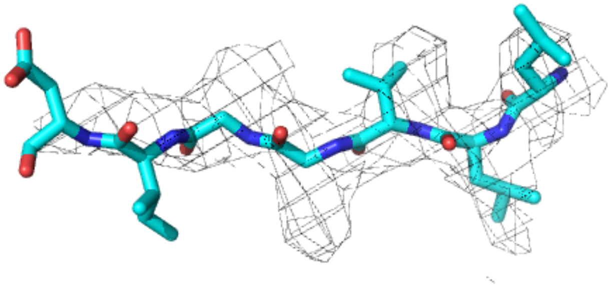
Model extension

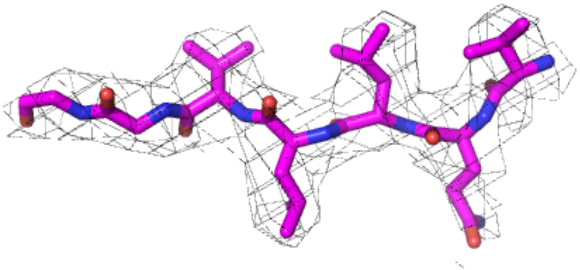
All-atom refinement

- B factor fitting
- Cross validation

Refinement against EM density

- Refinement
  - identify (and correct) errors in the initial model
  - improve fit to data
  - improve model geometry

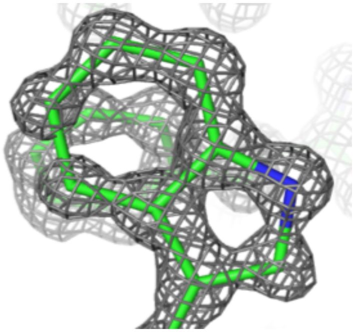




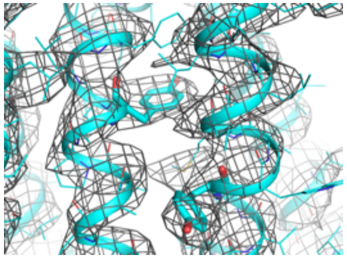
Refinement at low resolution requires a better geometry potential

Refinement: find atom positions optimizing:

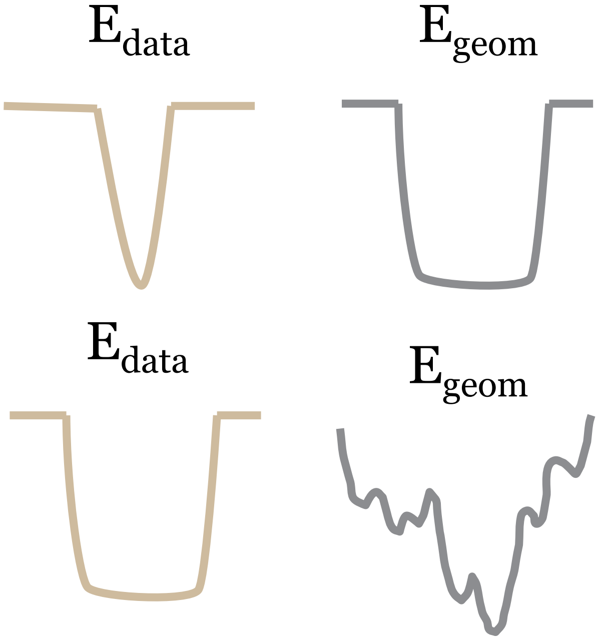
$$E = E_{geom} + w \cdot E_{data}$$



High-resolution



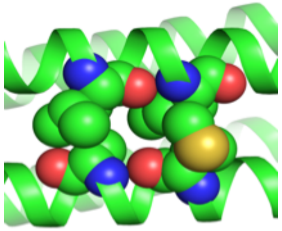
Low-resolution



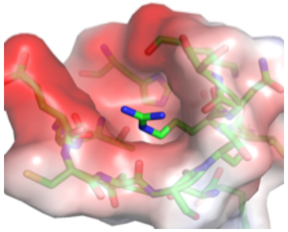
Rosetta forcefield disambiguates low-resolution solutions

Information from known structures reduces conformational space

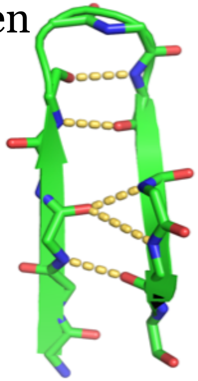
Core packing



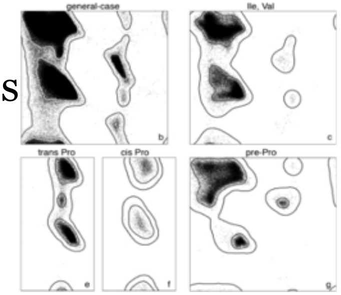
Electro-statics



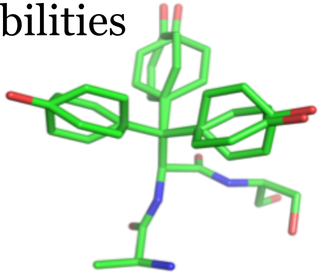
Hydrogen bonding



Torsional probabilities

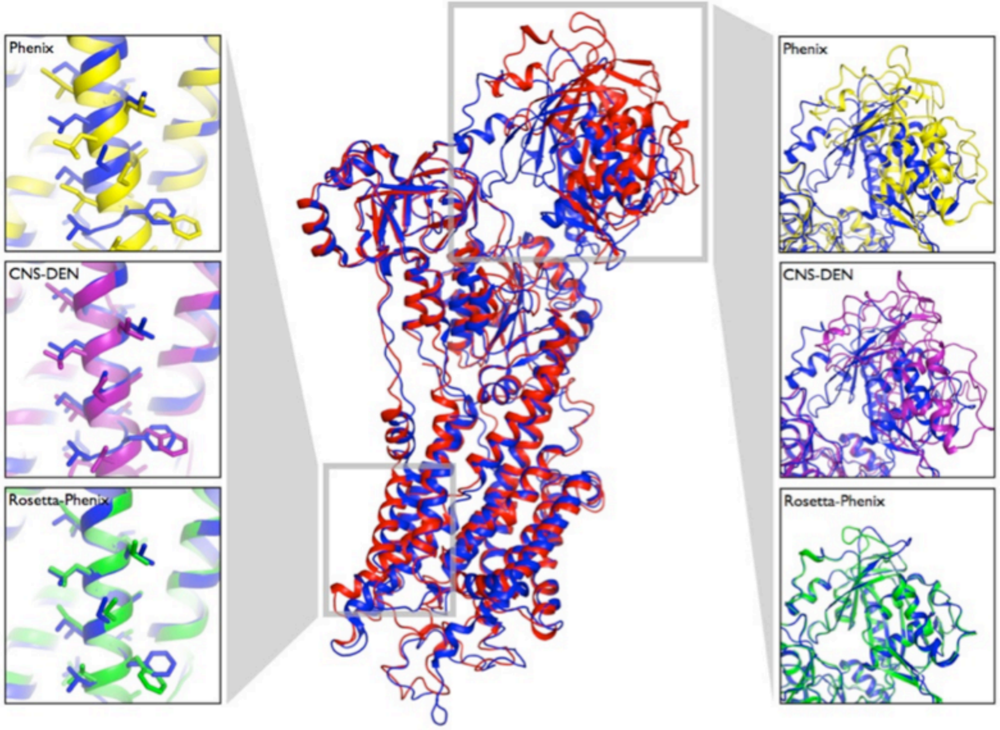


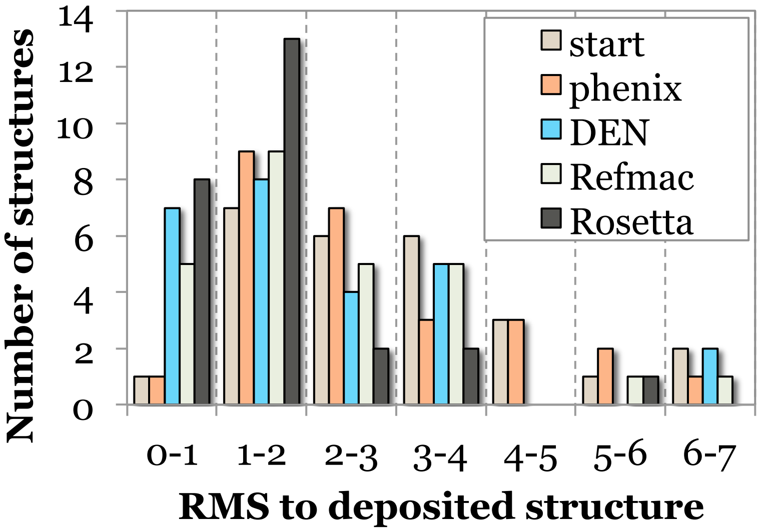
Rotamer probabilities



+ tools for improved optimization
(discrete sidechain optimization,
torsion and Cartesian space minimization, dynamics**)**

Our approach improves refinement against low-resolution crystallographic data

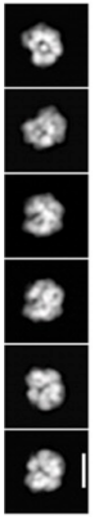


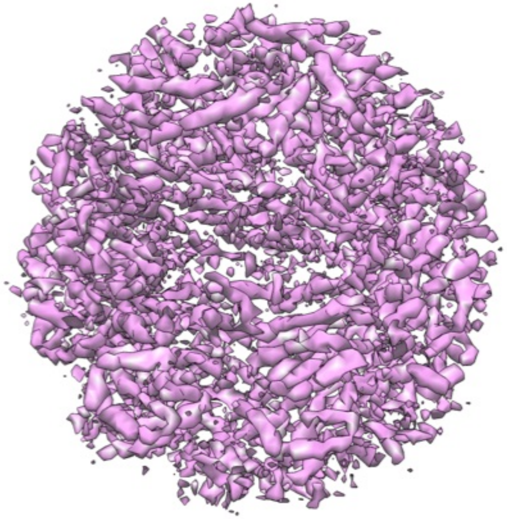


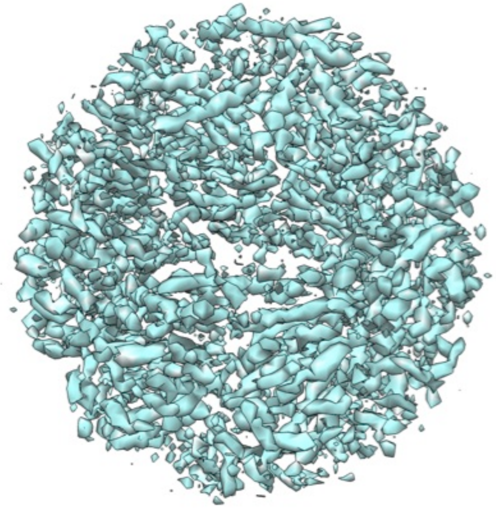
Key components for refinement against cryoEM

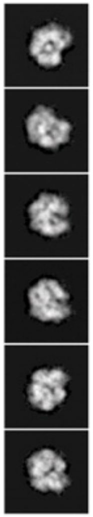
- Model validation
  - Independent map agreement over high-resolution shells
- Variations in local resolution
  - Atomic B factors describing how spread the density is around each atom
- Small radius of convergence
  - Discrete backbone optimization in refinement

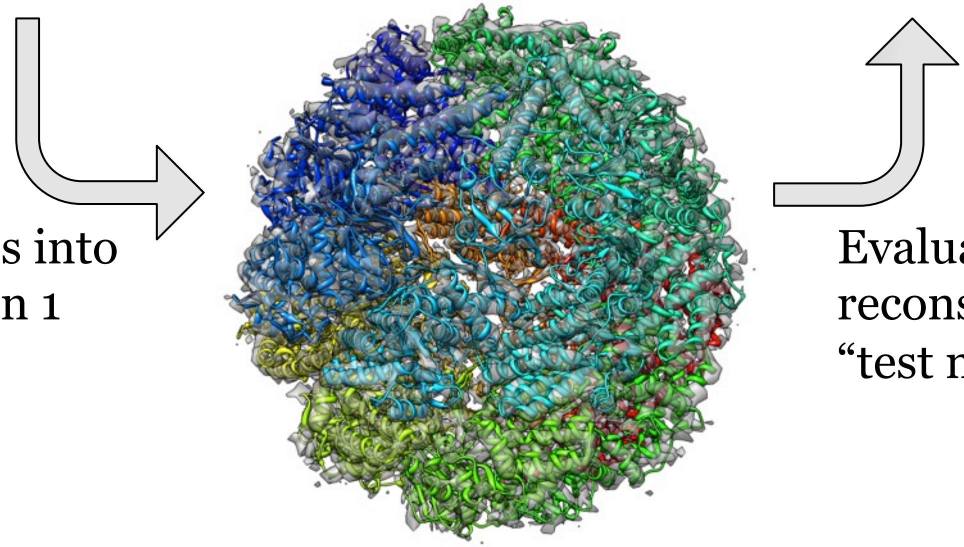
Independent validation







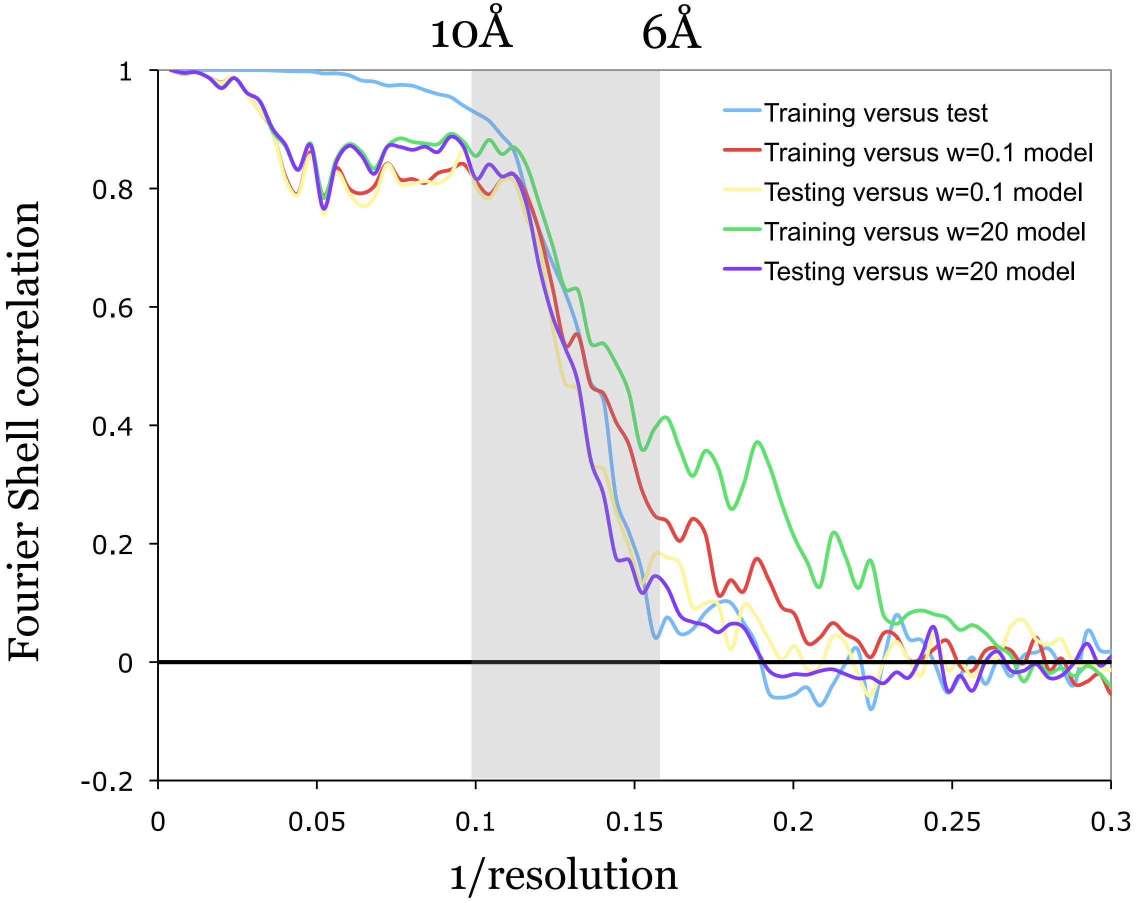




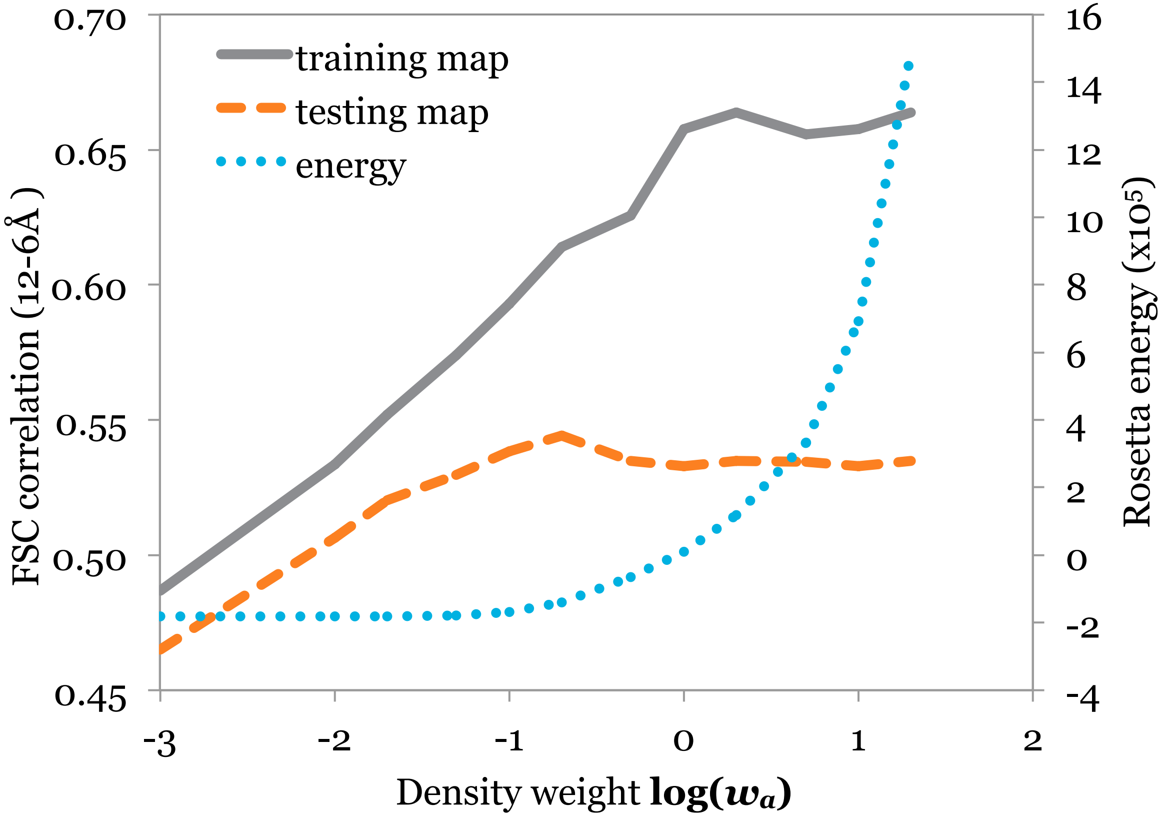
Refine models into
reconstruction 1
“train map”

Evaluate models against
reconstruction 2
“test map”

Independent validation

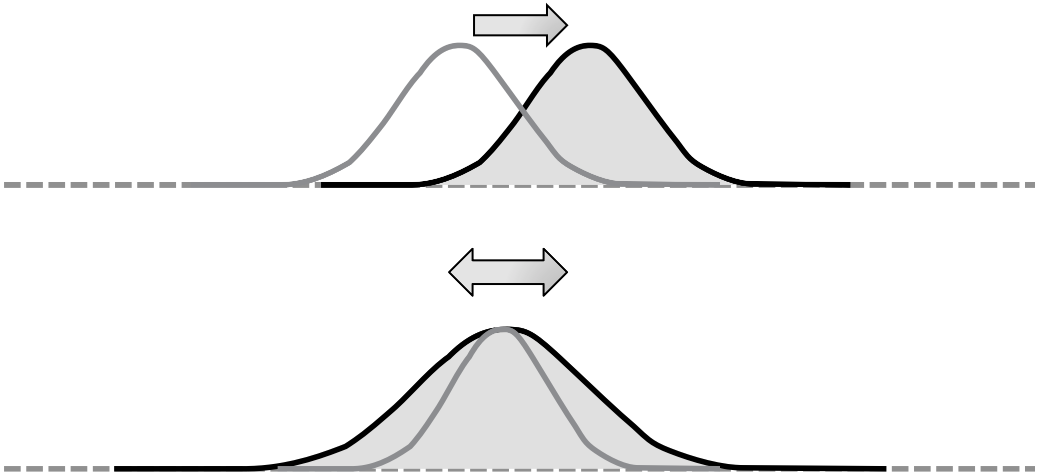


Independent validation



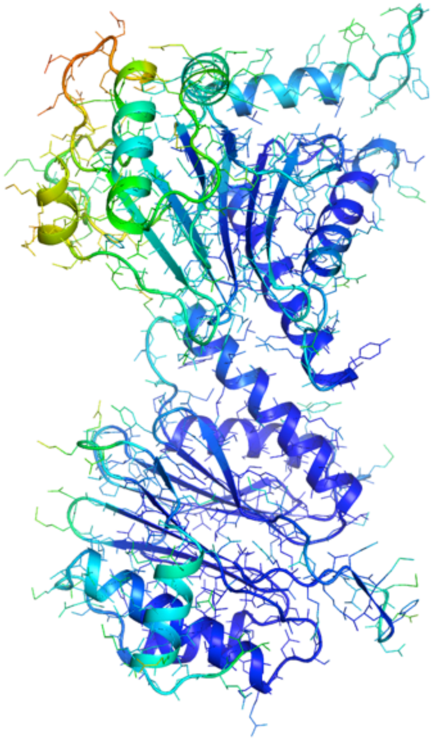
Fitting atomic B factors

- In addition to refining atomic coords, refine per-atom B factors (in real space)

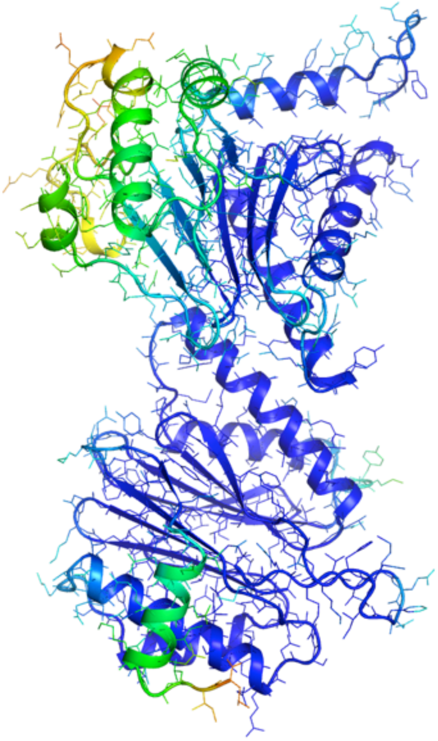


  - Alternate coordinate refinement and B factor refinement
  - Constraint function keeps B factors of nearby atoms close

Model B's have good agreement with crystallographic Bs

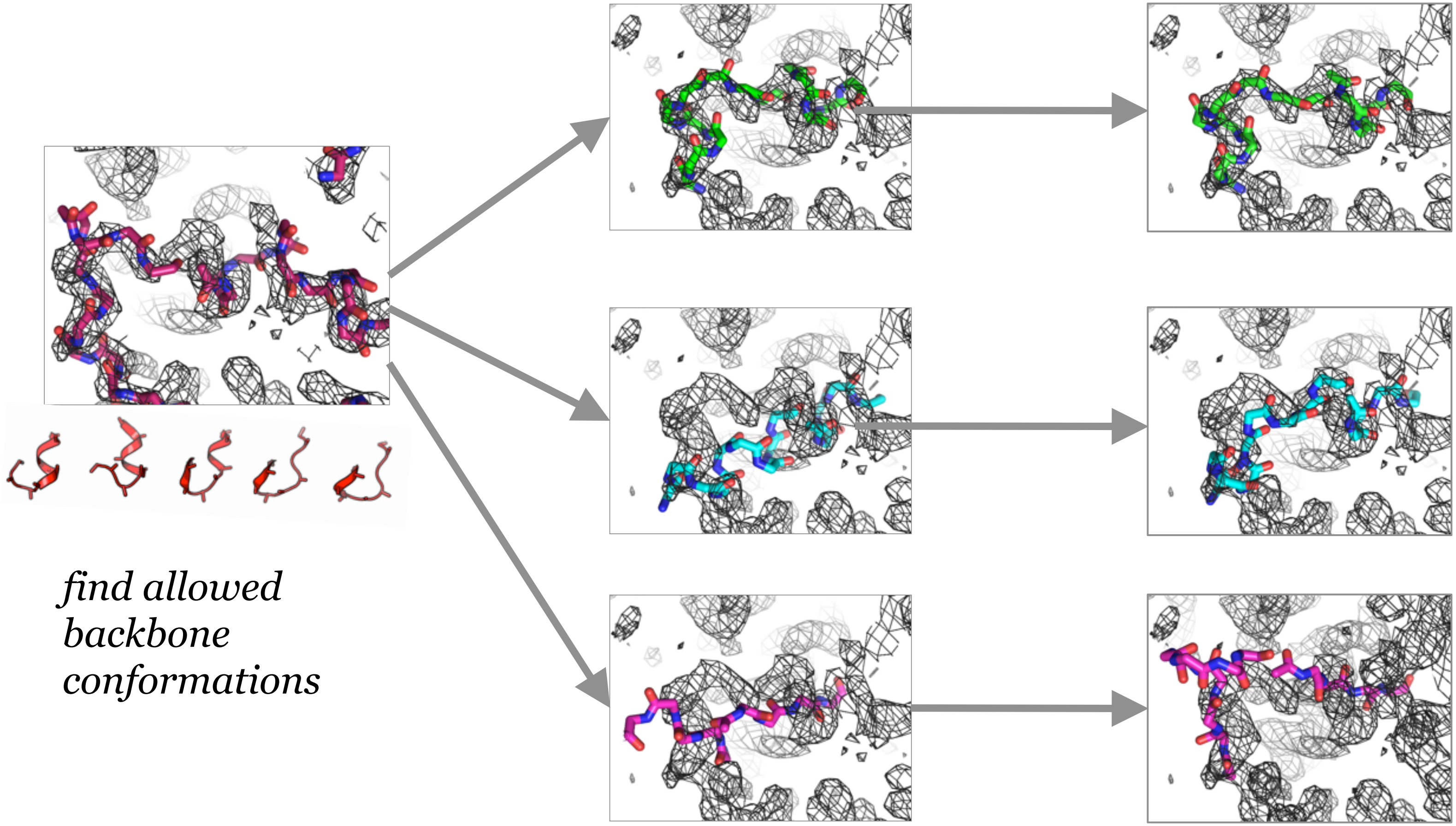


Deposited crystal structure (1pma)



CryoEM map, real-space B factors

Iterative density-guided conformational sampling



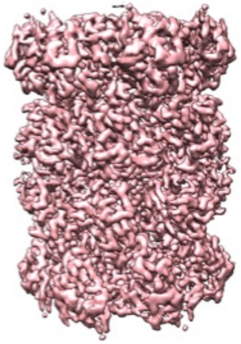
find allowed backbone conformations

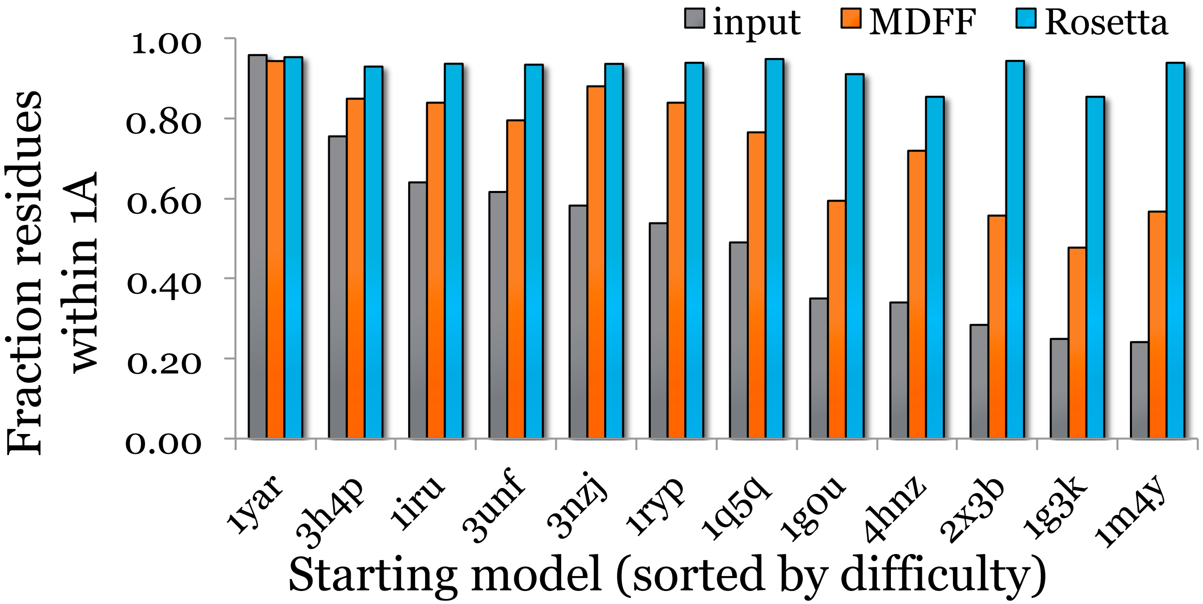
optimize into density with minimal forcefield

Assessing the role of starting-model quality on structure determination

Template	Sequence ID
1yar	100%
3h4p	50%
3nzj	32%
1iru	30%
1ryp	30%
1q5q	26%
3unf	25%
1m4y	20%
2x3b/2z3b	19%
4hnz	17%
1g3k	17%
1g0u	17%

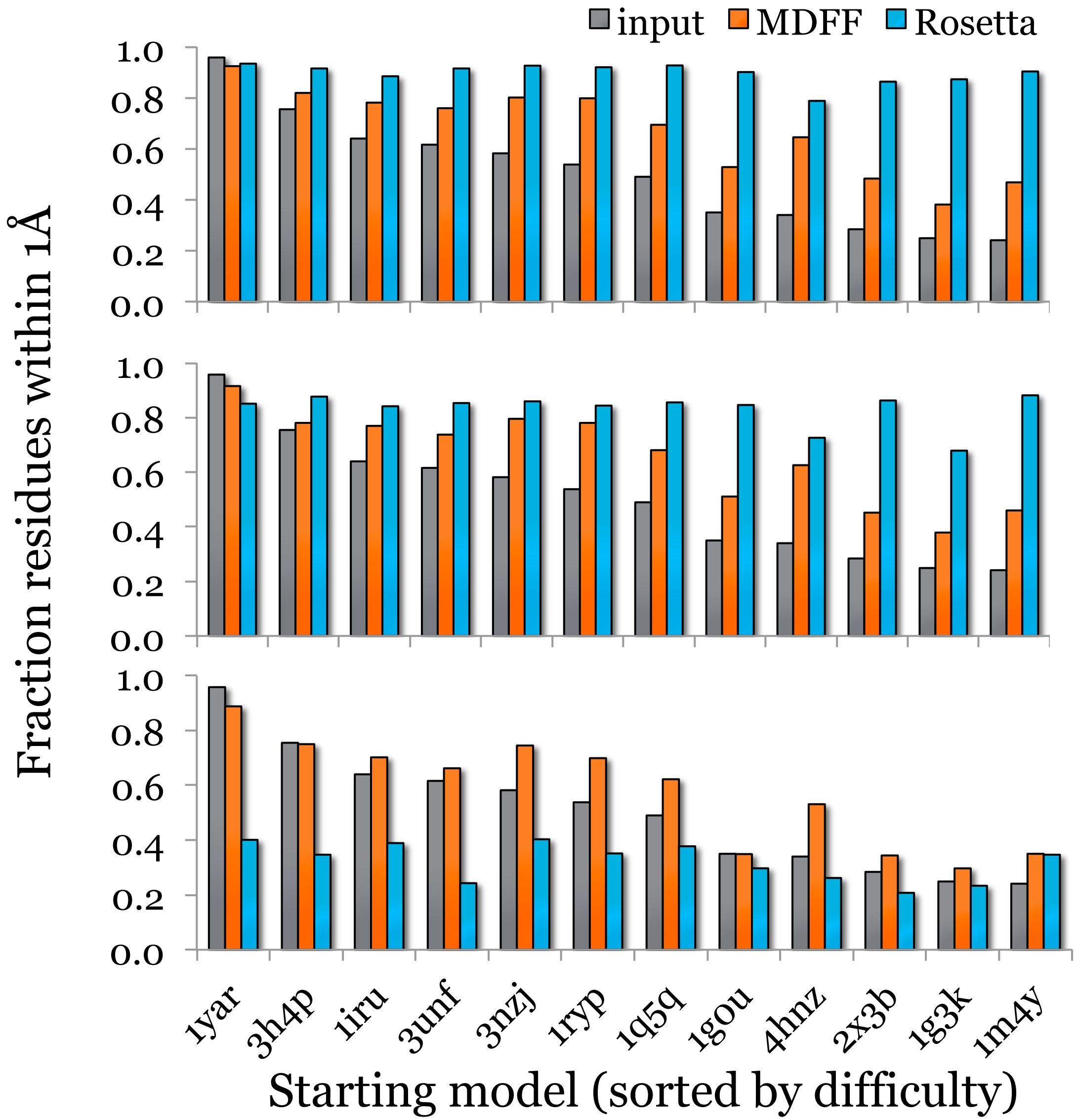
20S proteasome at 3.3Å resolution



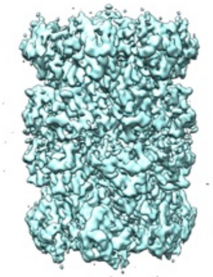


with Yifan Cheng, Xueming Li

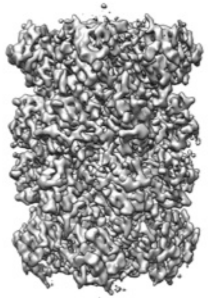
We can accurately determine structures to atomic resolution at 4.4Å or better



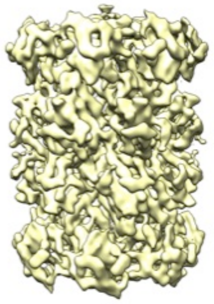
4.1Å
5k particles



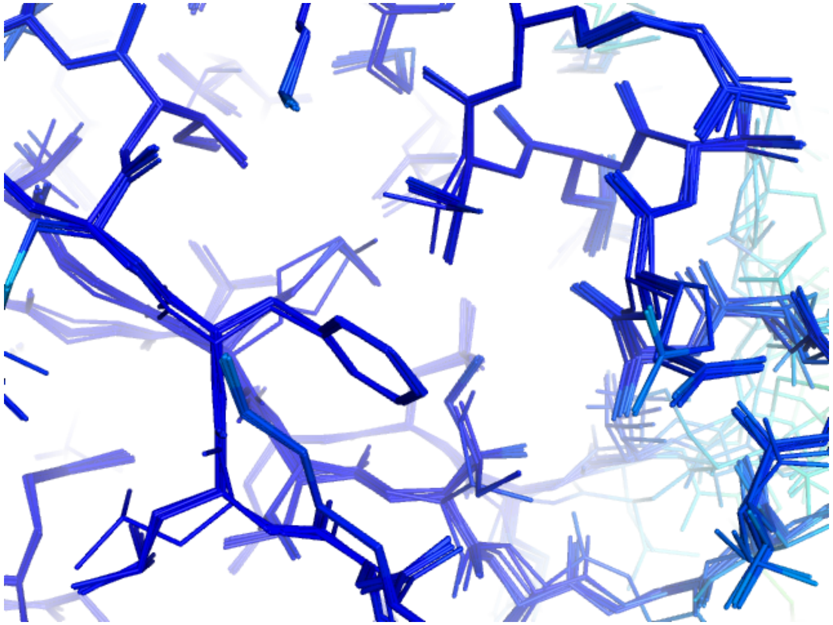
4.4Å
3k particles



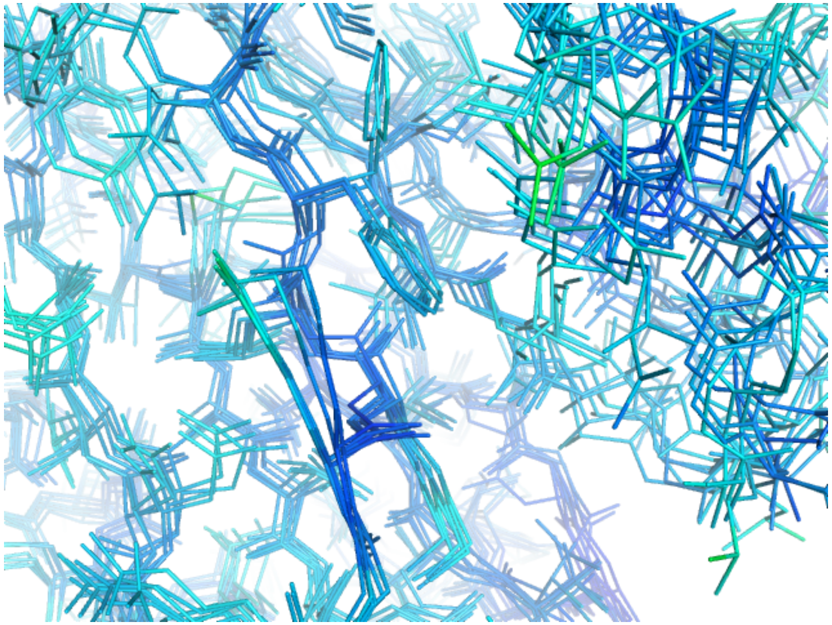
6.0Å
1k particles



Model convergence is an indicator of accuracy

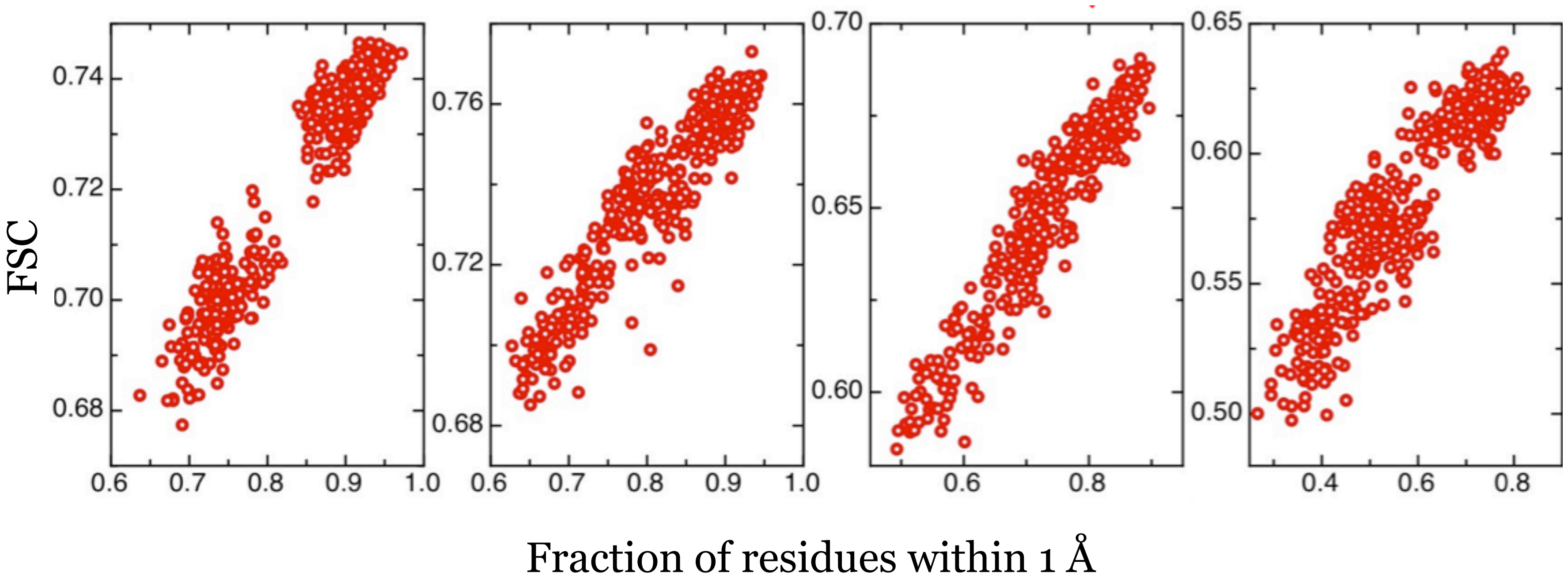


1gou (3.3Å)

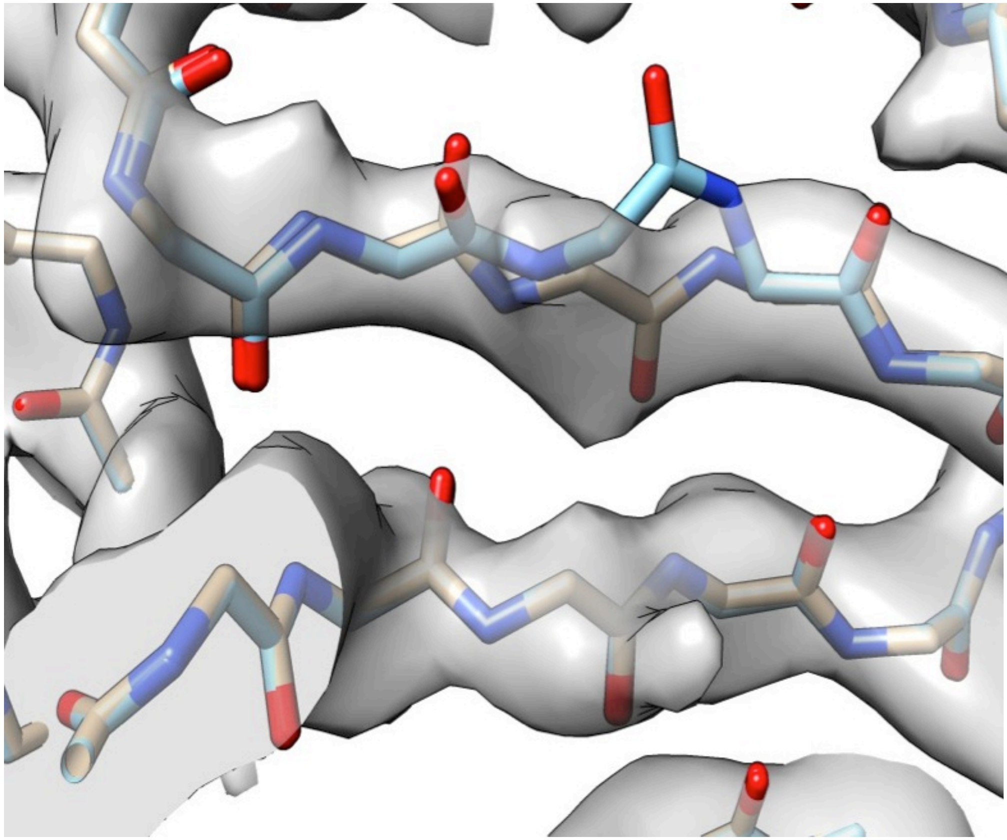


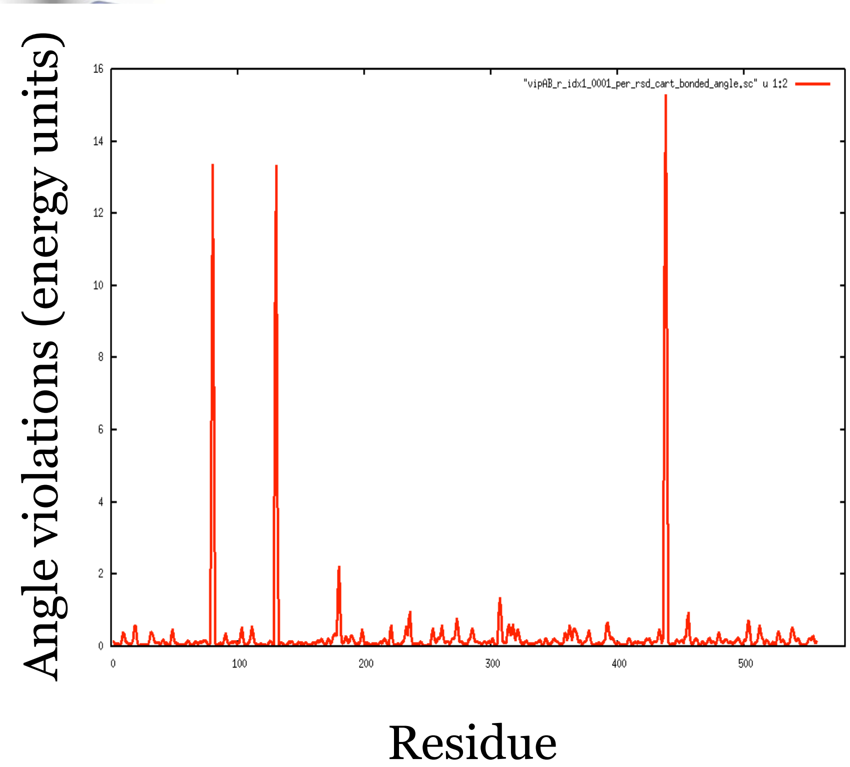
1gou (6.0Å)

Independent FSC is an indicator of accuracy (though not absolute)

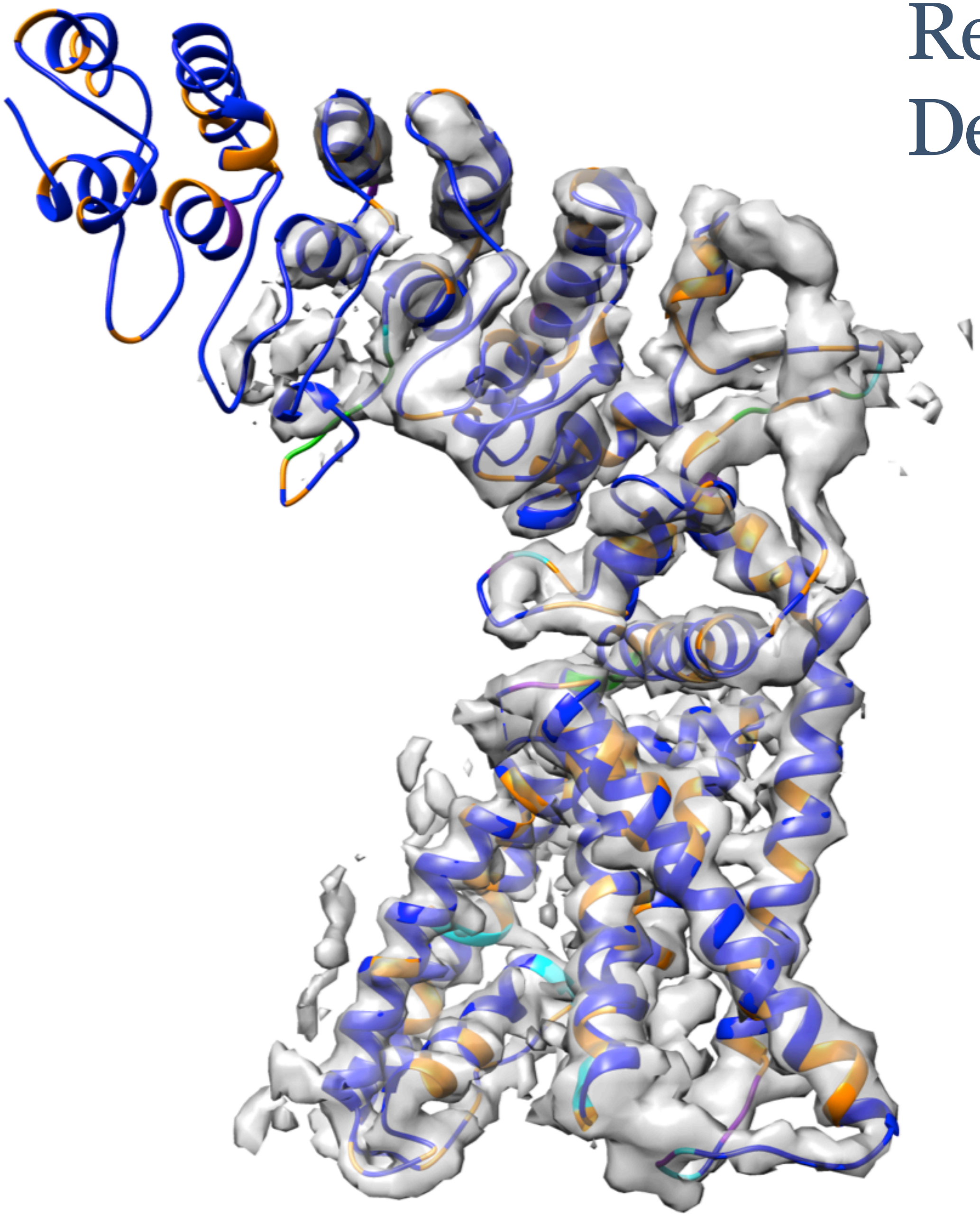


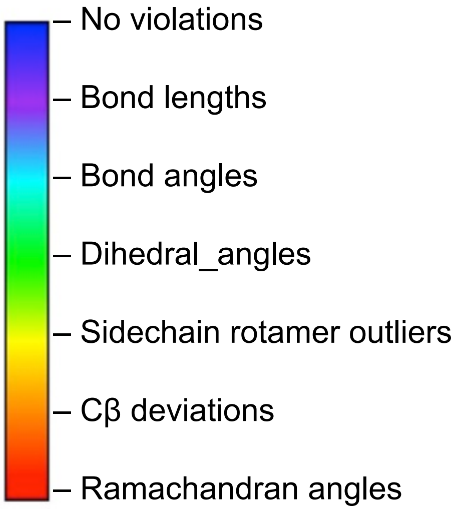
Model strain also can indicate errors



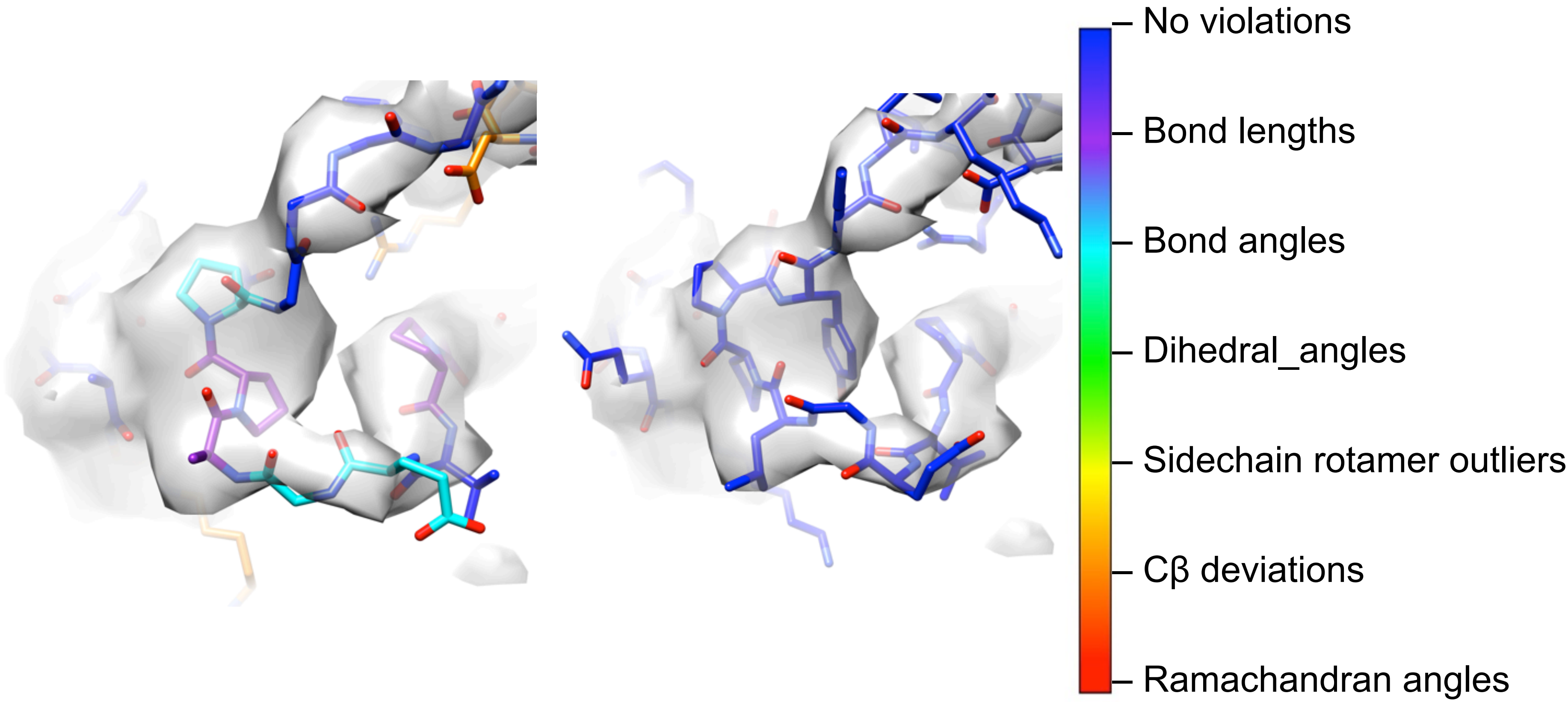

"vipAB_r_idx1_0001_per_rsd_cart_bonded_angle.sc" u 1:2
Angle violations (energy units)
16
14
12
10
8
6
4
2
0
0
100
200
300
400
500
Residue

Refinement of TRPV1: Deposited structure

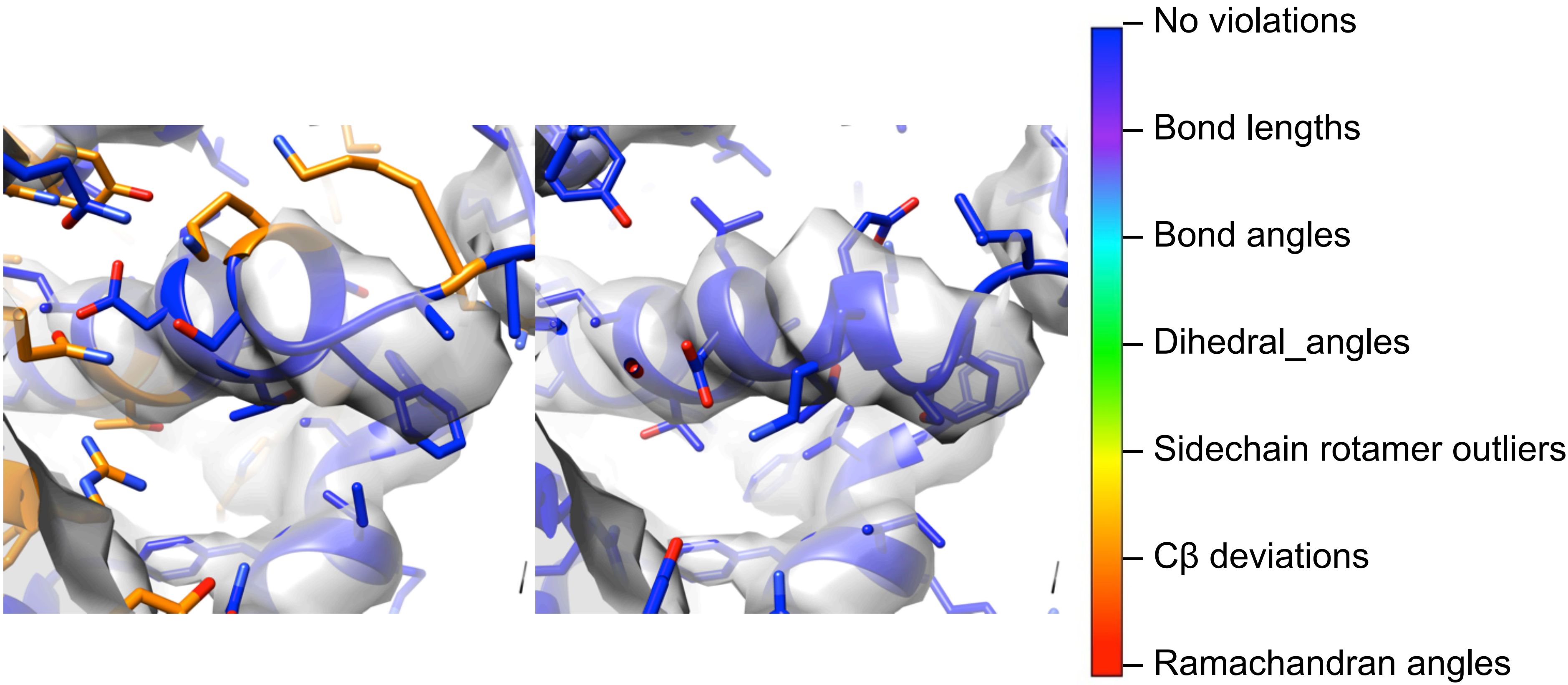




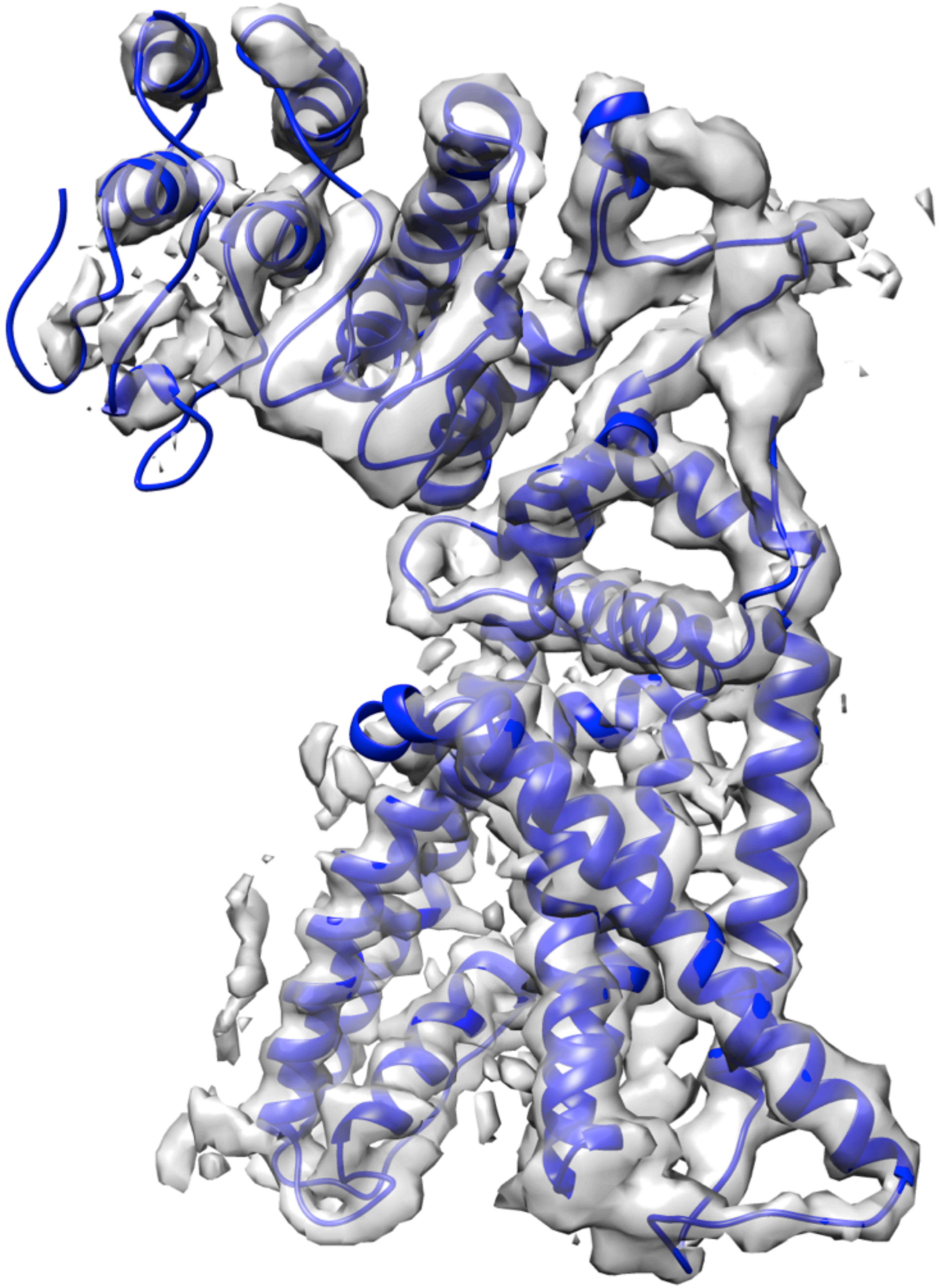
Local strain reveals errors

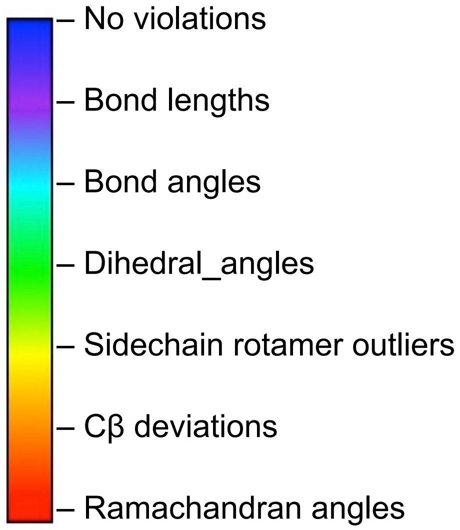


Local strain reveals errors



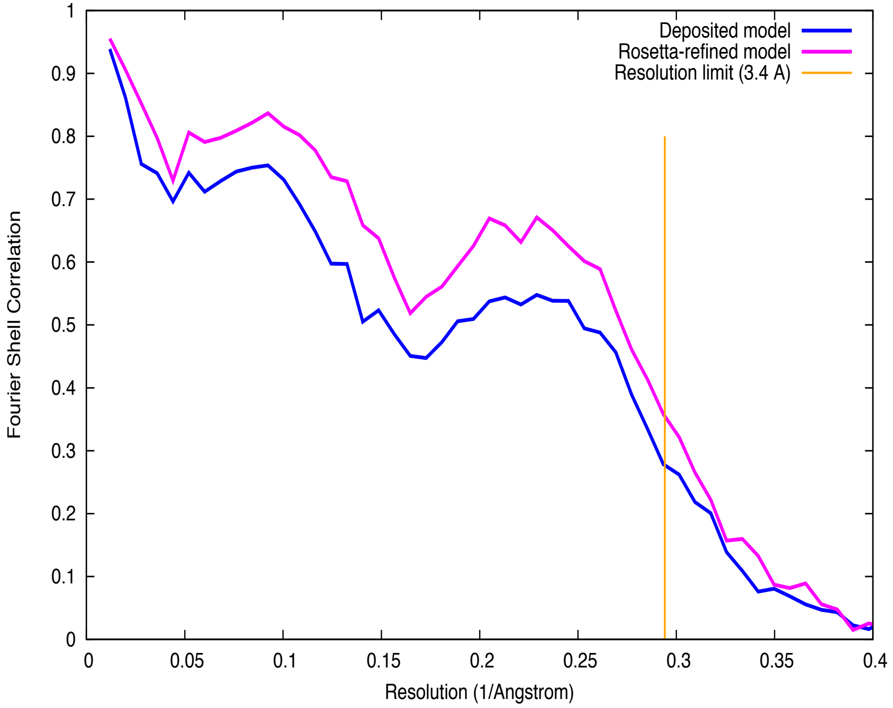
Final refined model



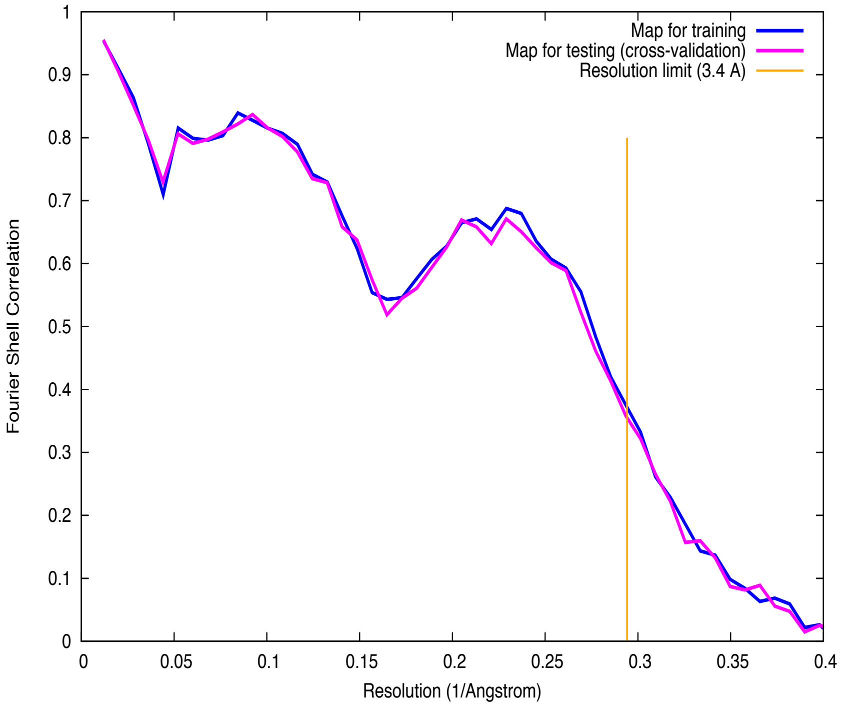


Cross-validation – low/no overfitting

model-map FSC
deposited versus refined



model-map FSC
train versus test

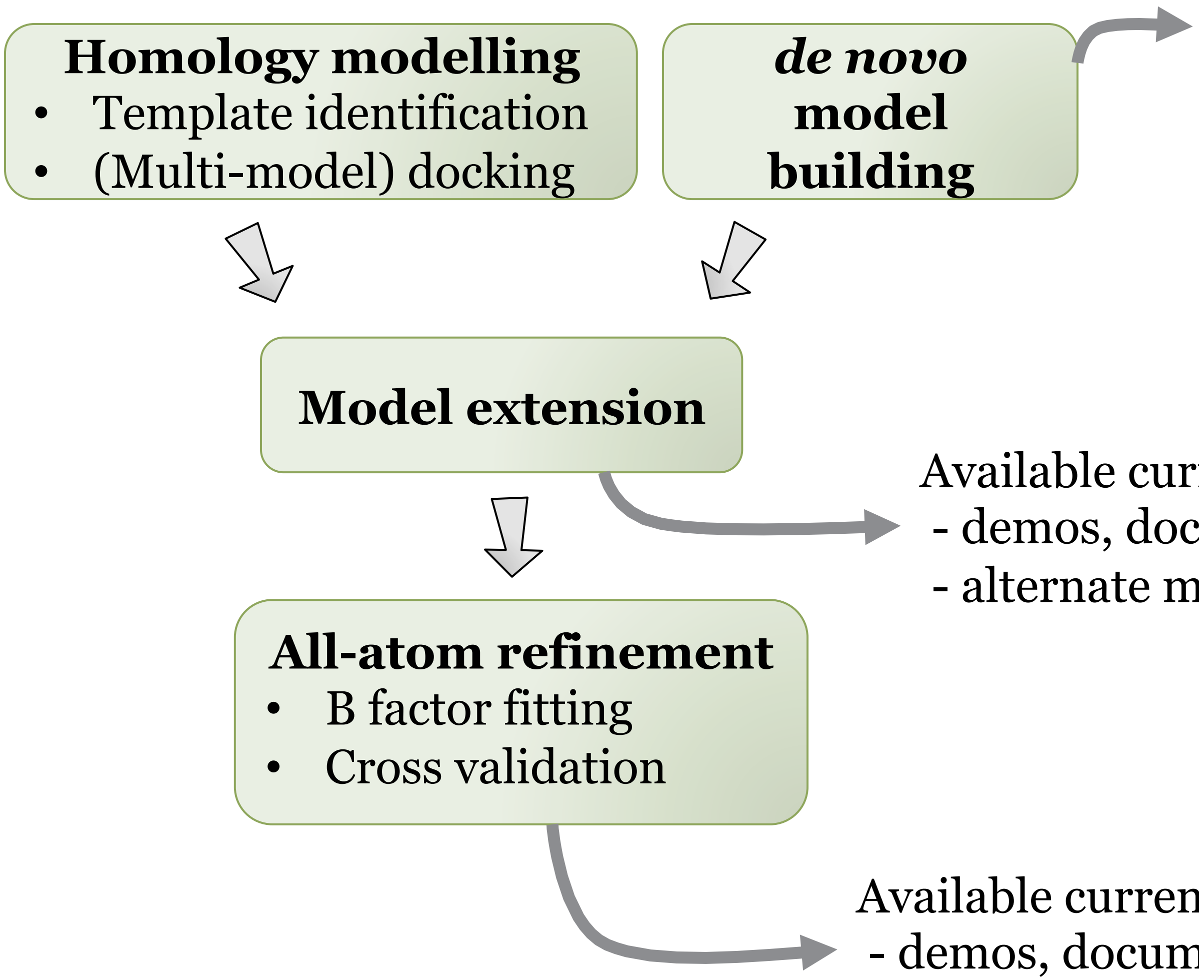


Conclusions

- Atomic accuracy is possible from near-atomic resolution (up to 4.5Å) data

- Have we solved it? Do we have…
  - Good fit to independent data (locally and globally)?
  - No model strain / molprobity outliers?
  - Well converged ensemble of solutions satisfying the above two?

Method availability



Homology modelling

- Template identification
- (Multi-model) docking

***de novo* model building**

Available ~Dec 2014

Model extension

Available currently (as RosettaCM)
- demos, documentation
- alternate methods under development

All-atom refinement

- B factor fitting
- Cross validation

Available currently
- demos, documentation

Acknowledgements

- Collaborators
  - Wah Chiu (Baylor), Junjie Zhang (Texas A&M)
  - Tom Marlovits (IMBA, Austria)
  - Ed Egelman (U. Virginia)
  - Misha Kudryashev, Marek Basler (U. Basel)
  - Xueming Li, Yifan Cheng (UCSF)
- Students & Postdoc
  - **Ray Wang**
  - Patrick Conway
  - Brandon Frenz
  - Zibo Chen
  - Ryan Pavlovicz