



Science and
Technology
Facilities Council

Scientific Computing

Model Validation

Agnel Praveen Joseph
CCP-EM STFC



Why are structural details important?

Atomic model gives a more interpretable representation of the structure.

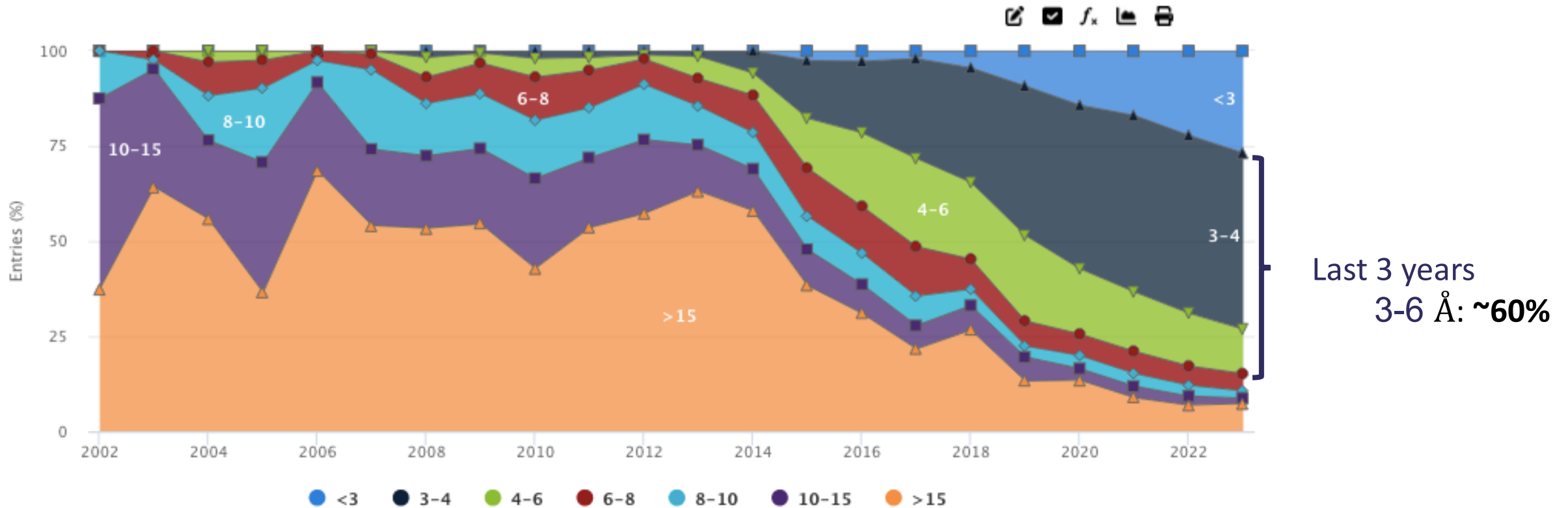
Finer details of function, interaction and dynamics.

Structure based drug-design.

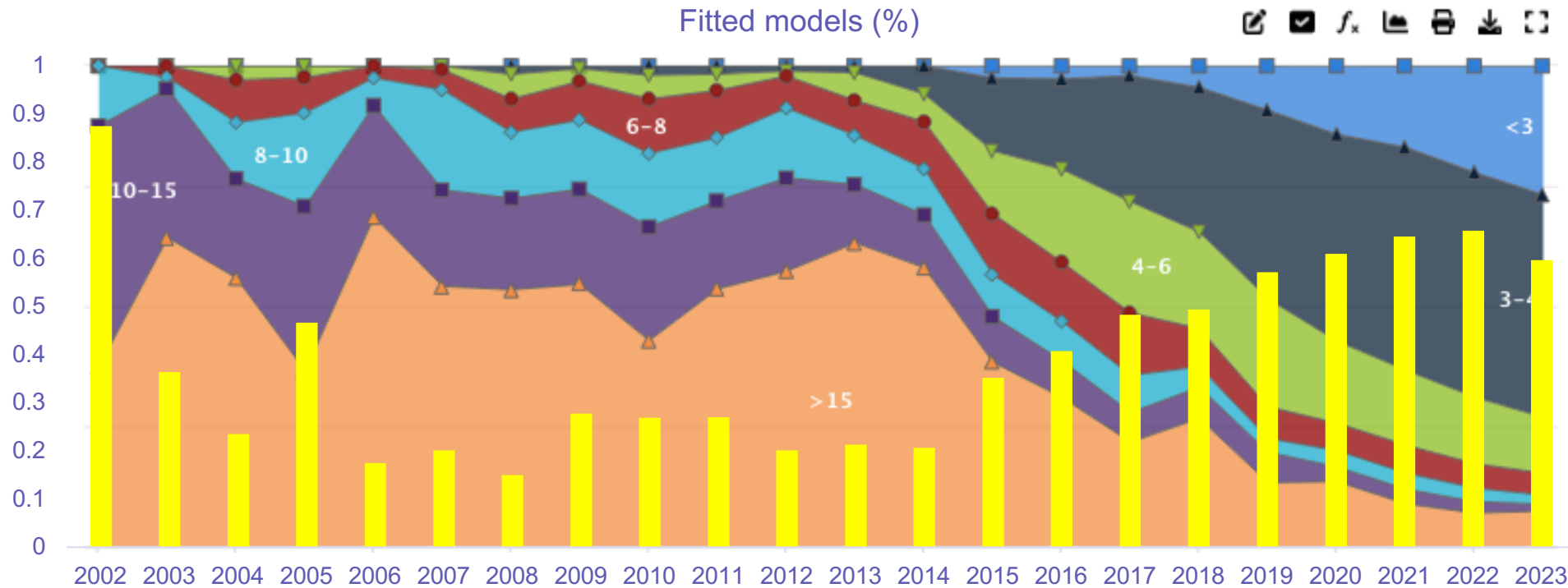
EM DataBank: map resolution

>28k cryo-EM reconstructions in EMDB

EMDB entry resolution in shells per year



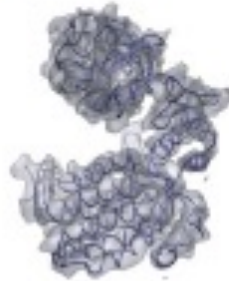
EMDB: maps with fitted models



Resolution

- $< 2\text{\AA}$ [141]
- $\geq 2\text{\AA}$ and $< 3\text{\AA}$ [3170]
- $\geq 3\text{\AA}$ and $< 4\text{\AA}$ [8421]
- $\geq 4\text{\AA}$ and $< 5\text{\AA}$ [2265]
- $\geq 5\text{\AA}$ and $< 6\text{\AA}$ [362]
- $\geq 6\text{\AA}$ and $< 8\text{\AA}$ [732]
- $\geq 8\text{\AA}$ and $< 12\text{\AA}$ [568]
- $\geq 12\text{\AA}$ and $< 16\text{\AA}$ [165]

Features interpretable at different resolutions

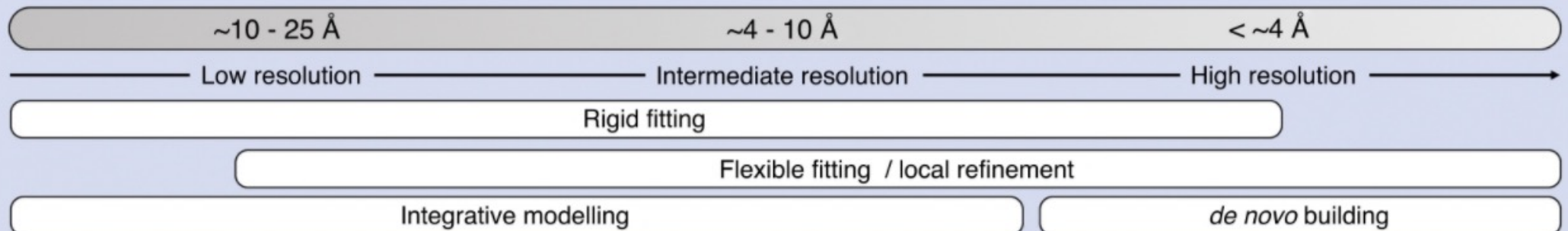


Villa et al. 2014

Up to 20 Å	Up to 9 Å	Up to 6 Å	Up to 4 Å
Conformational changes	Conformational changes	Conformational changes	Conformational changes
Domain boundaries	Domain boundaries	Domain boundaries	Domain boundaries
		Beta sheets	Beta sheets
	Alpha helices	Alpha helices	Individual beta strands
		Pitch of RNA helices	Alpha helices
			Pitch of alpha helices
			Pitch of RNA helices
			Phosphate "bumps"
			Side chains

Malhotra S. et al., COSB 2019

Resolution-based workflow





Science and
Technology
Facilities Council

Scientific Computing

Model validation



Science and
Technology
Facilities Council

Scientific Computing



Aspects of validation

Fit to data (Global and Local)

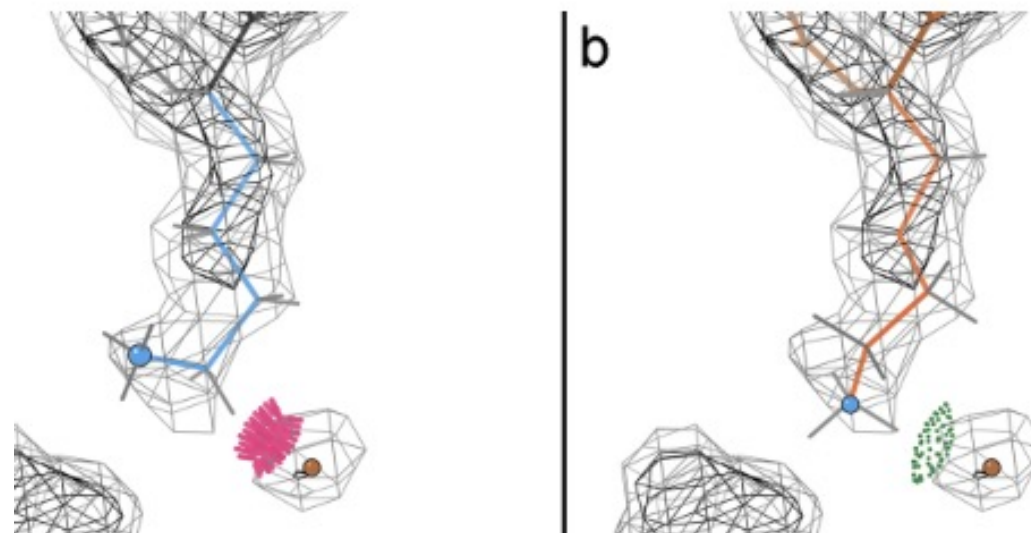
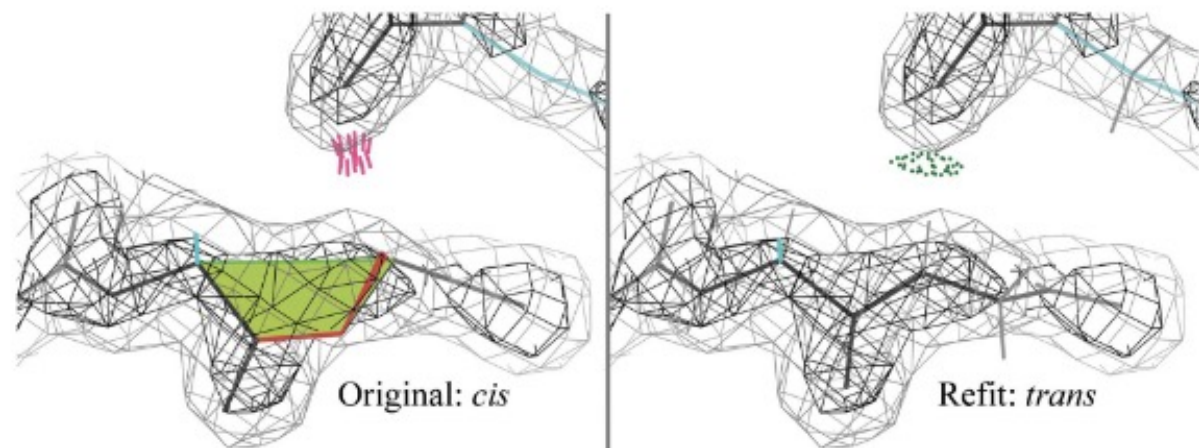
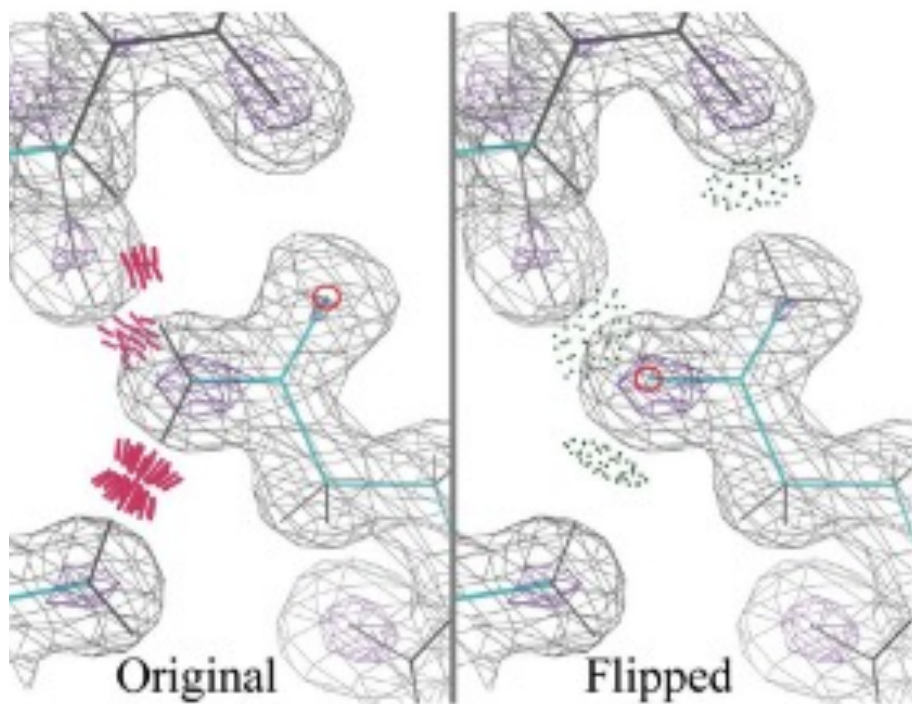
Model quality (geometry)

Overfitting

Model bias

Model geometry

Clashes



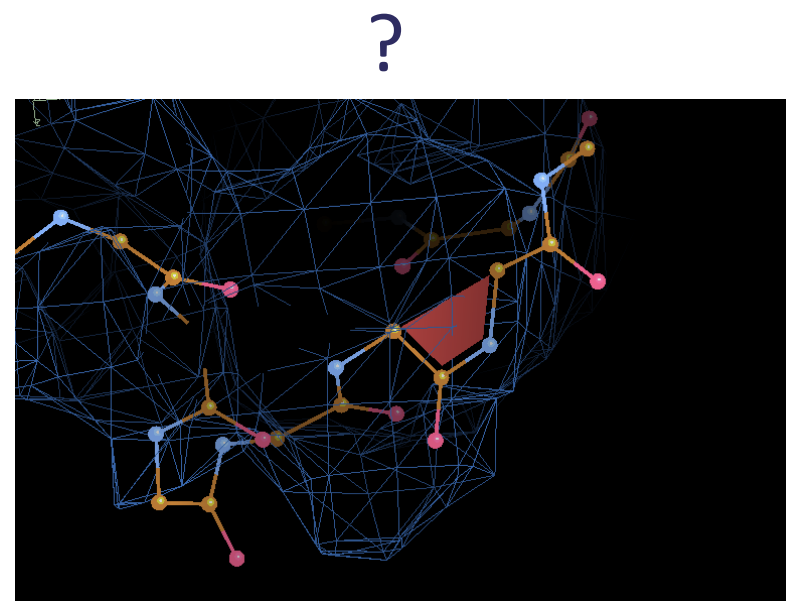
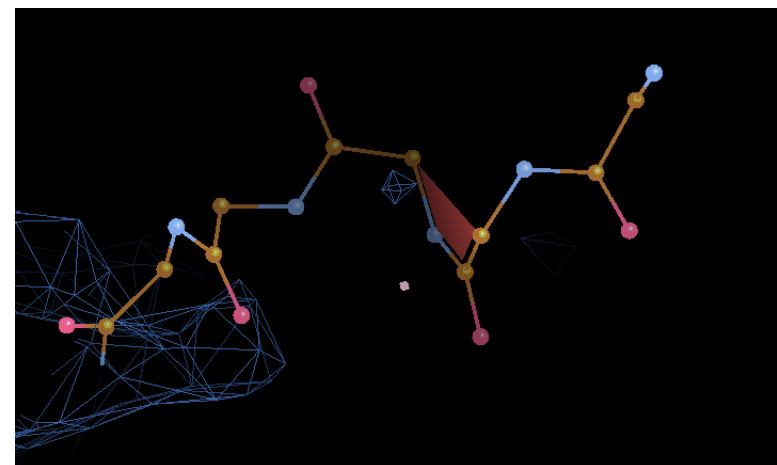
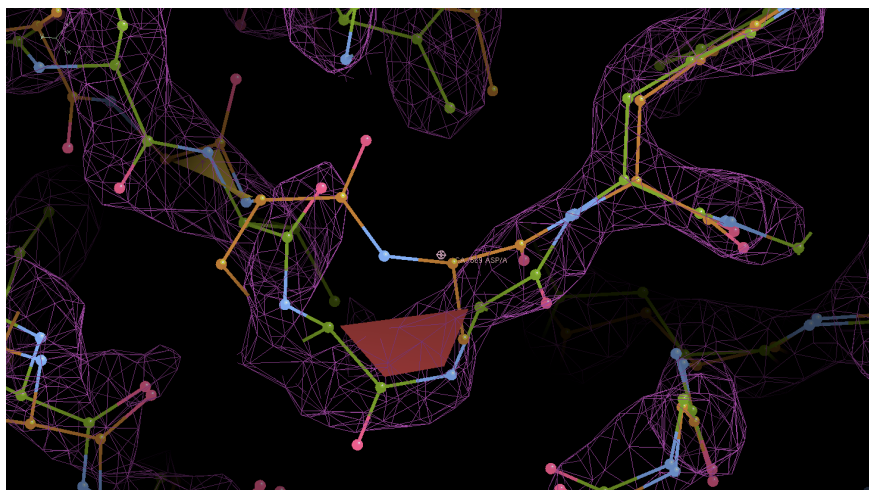
Model geometry

Cis-peptides

Cis-Pro : 5-6%

Cis-nonPro: ~0.05%

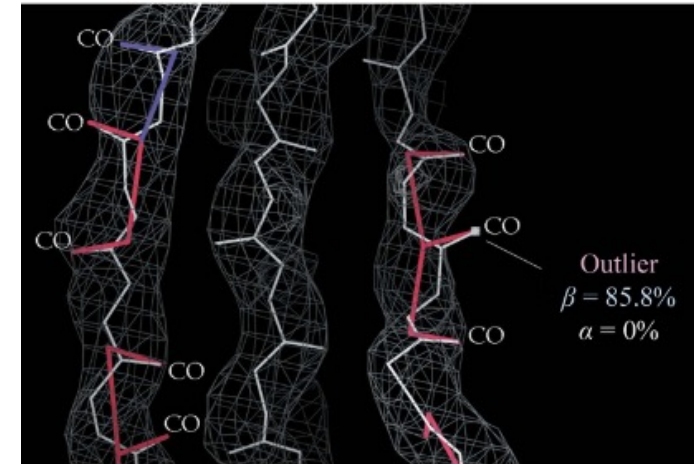
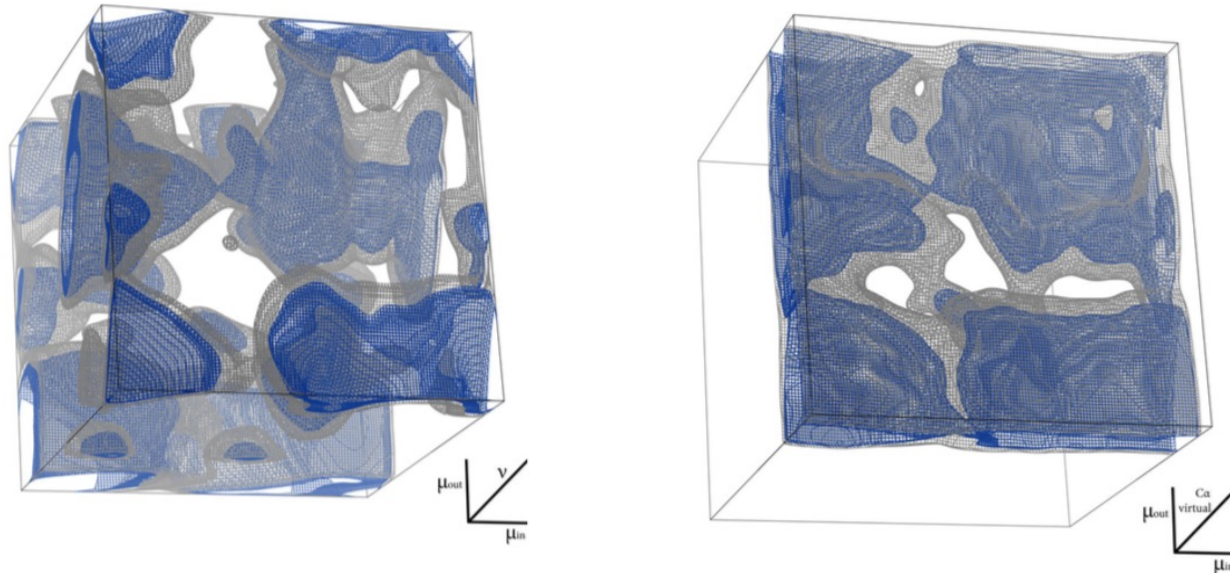
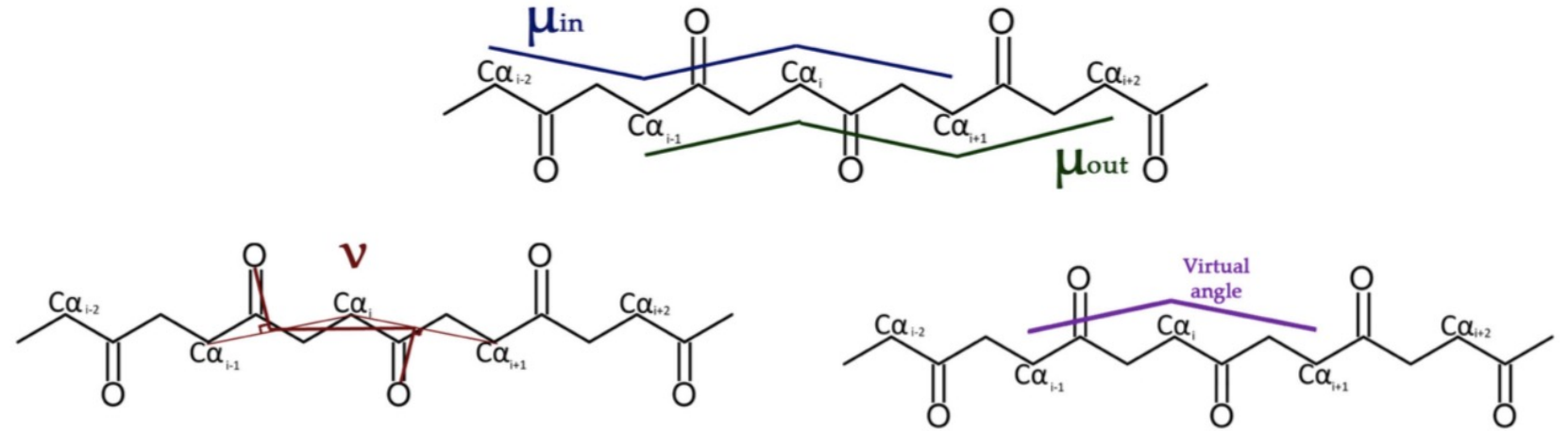
genuine



Model geometry

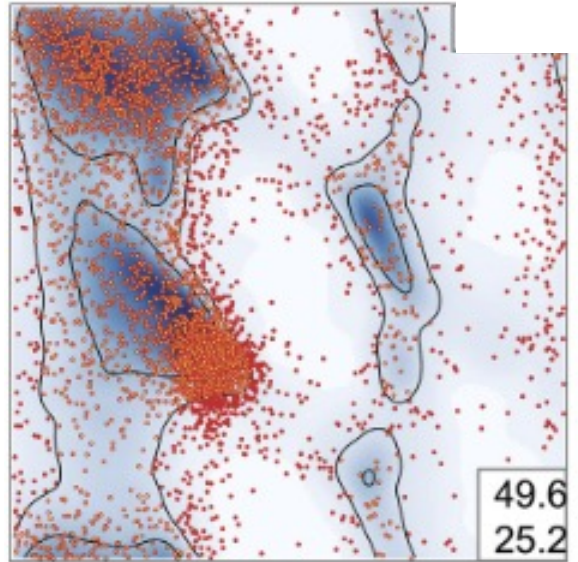
CaBLAM

Backbone geometry check

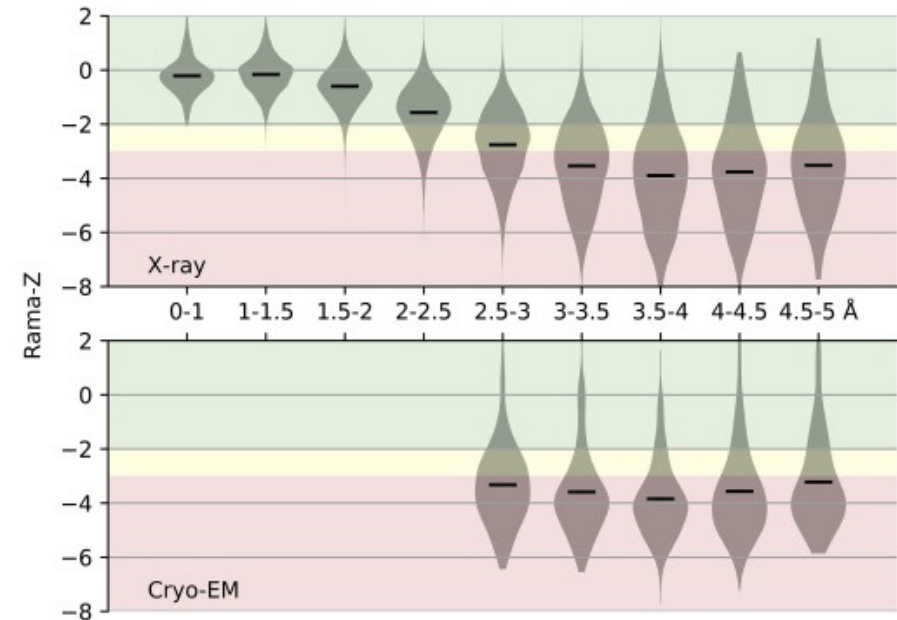
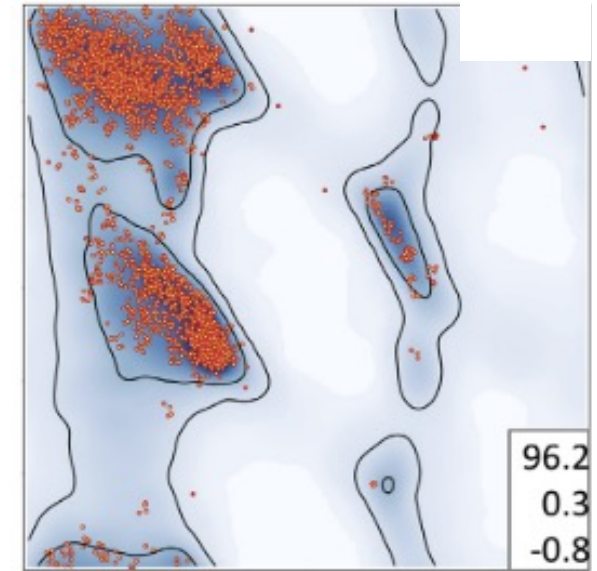
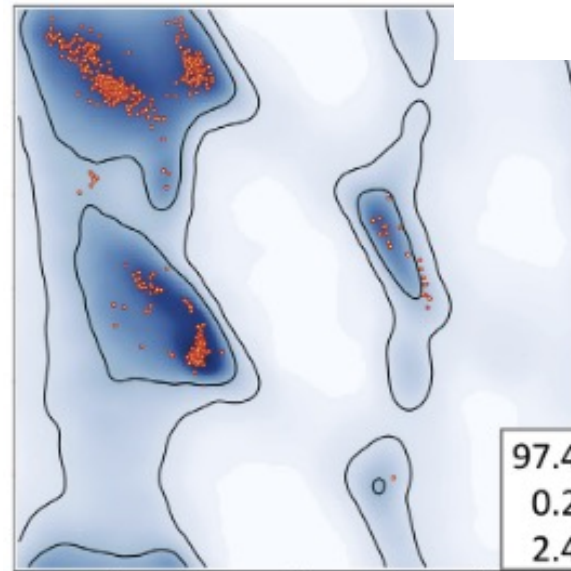


Model geometry

Ramachandran outliers



Ramachandran z-scores

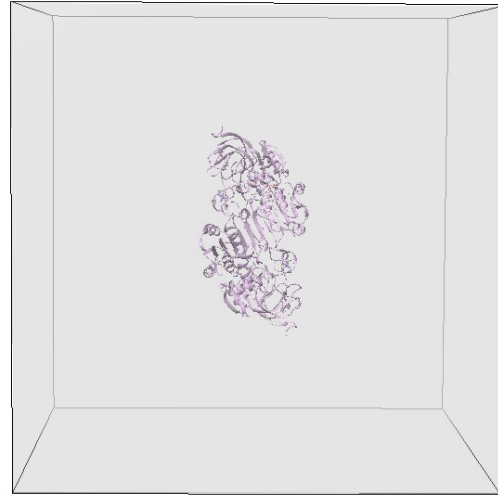


Global agreement with map

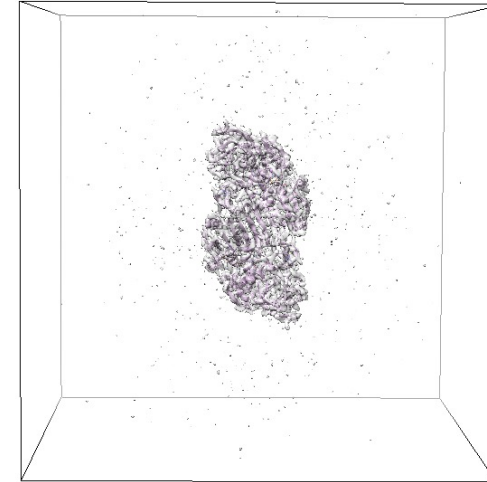
Cross-Correlation Coefficient

Values vary depending on map processing, resolution

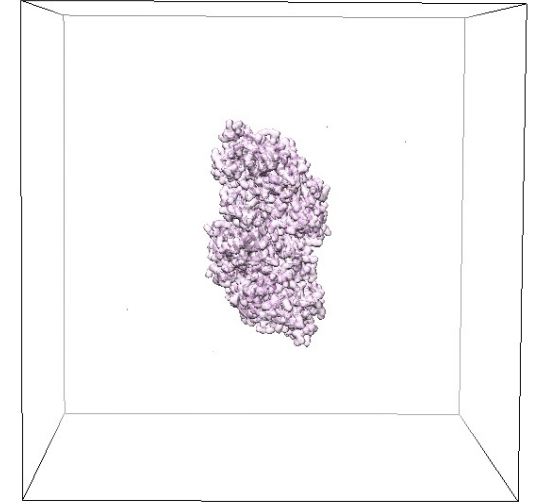
2.9Å full map: 0.32



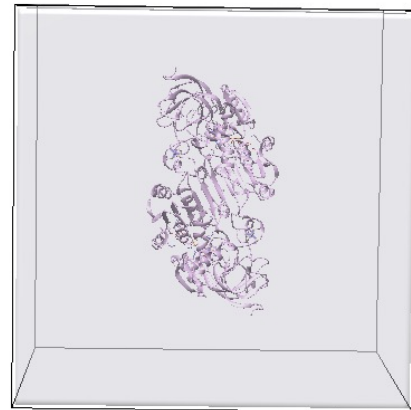
Contoured map: 0.59



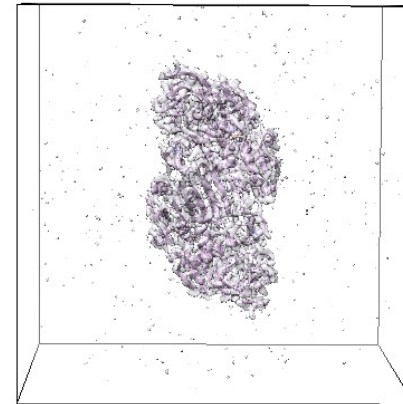
Map-model overlap mask: 0.42



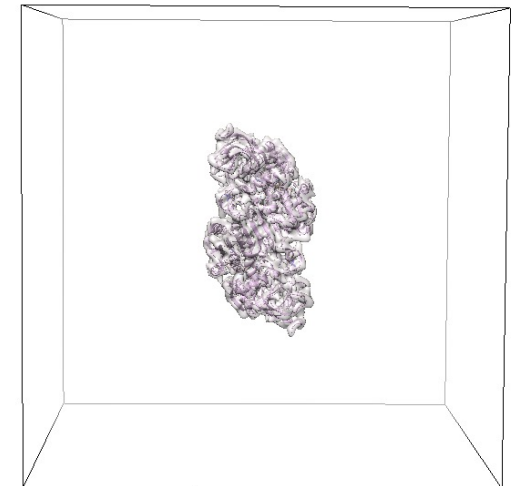
Cropped full map: 0.37



Cropped,contoured map: 0.55



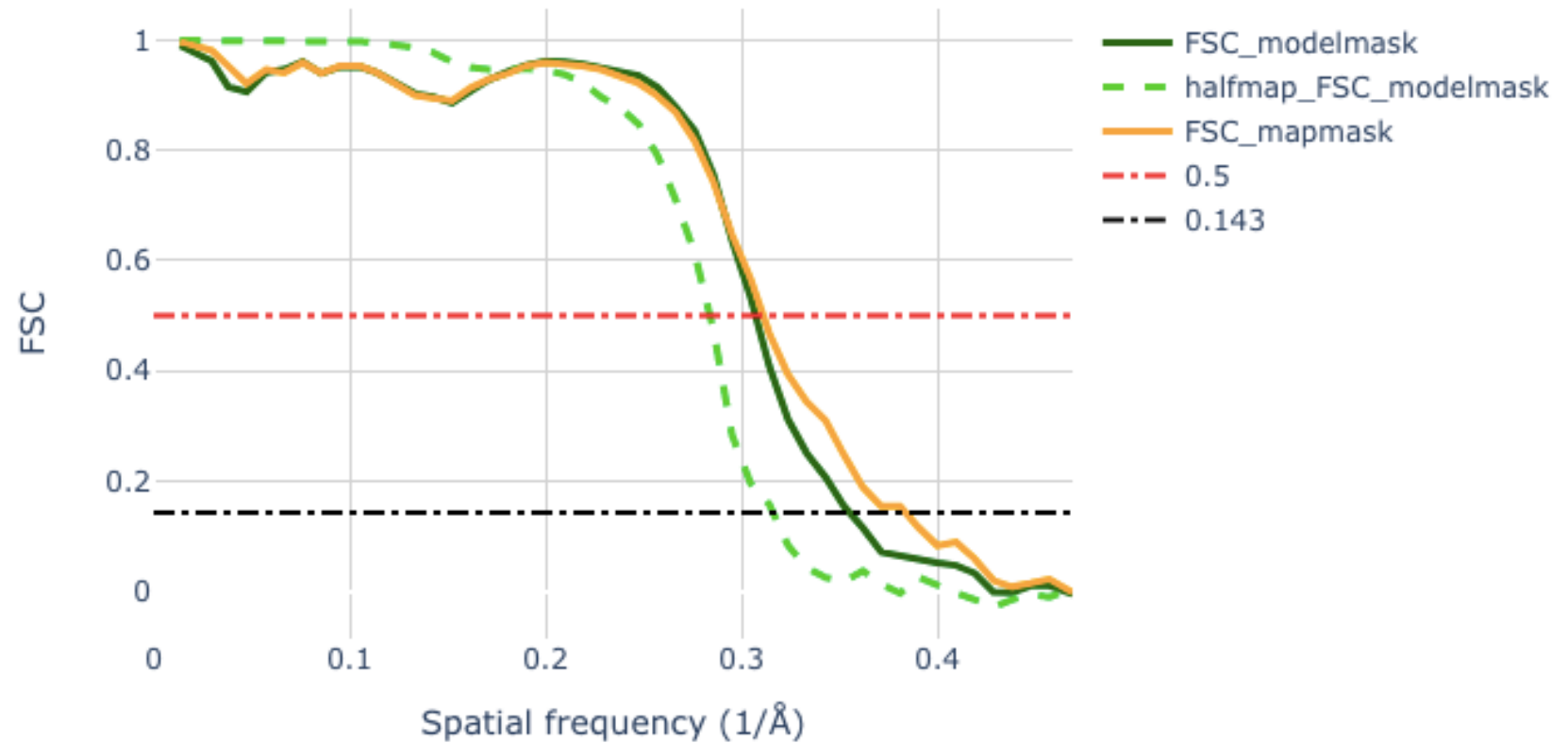
5Å contoured map: 0.76



Global agreement with map

Model-map FSC

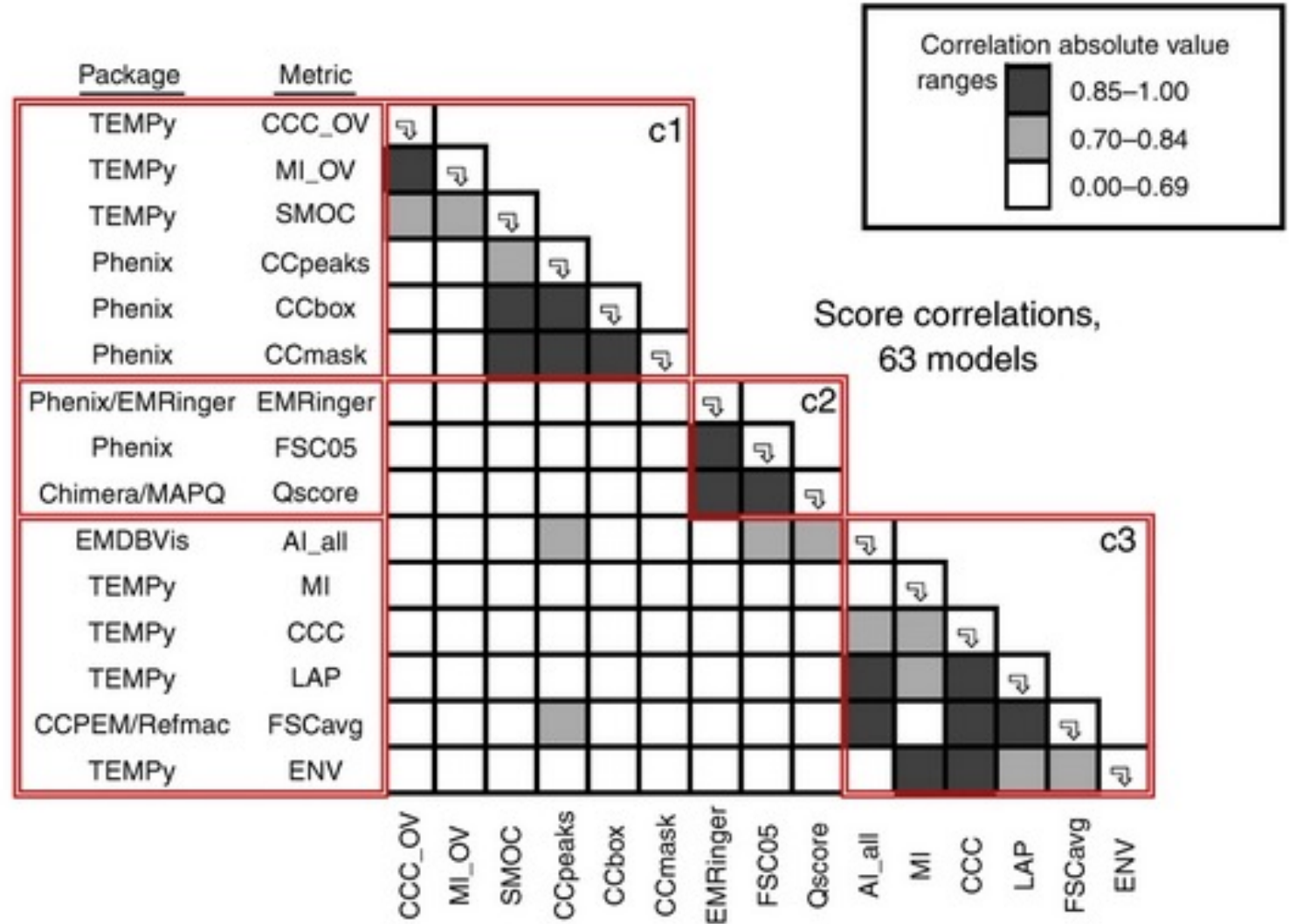
model-map FSC plot



Brown et al. Acta D, 2015

Yamashita et al. Acta D, 2021

Global agreement with map



Lawson et al. *Nature Methods*, 2021

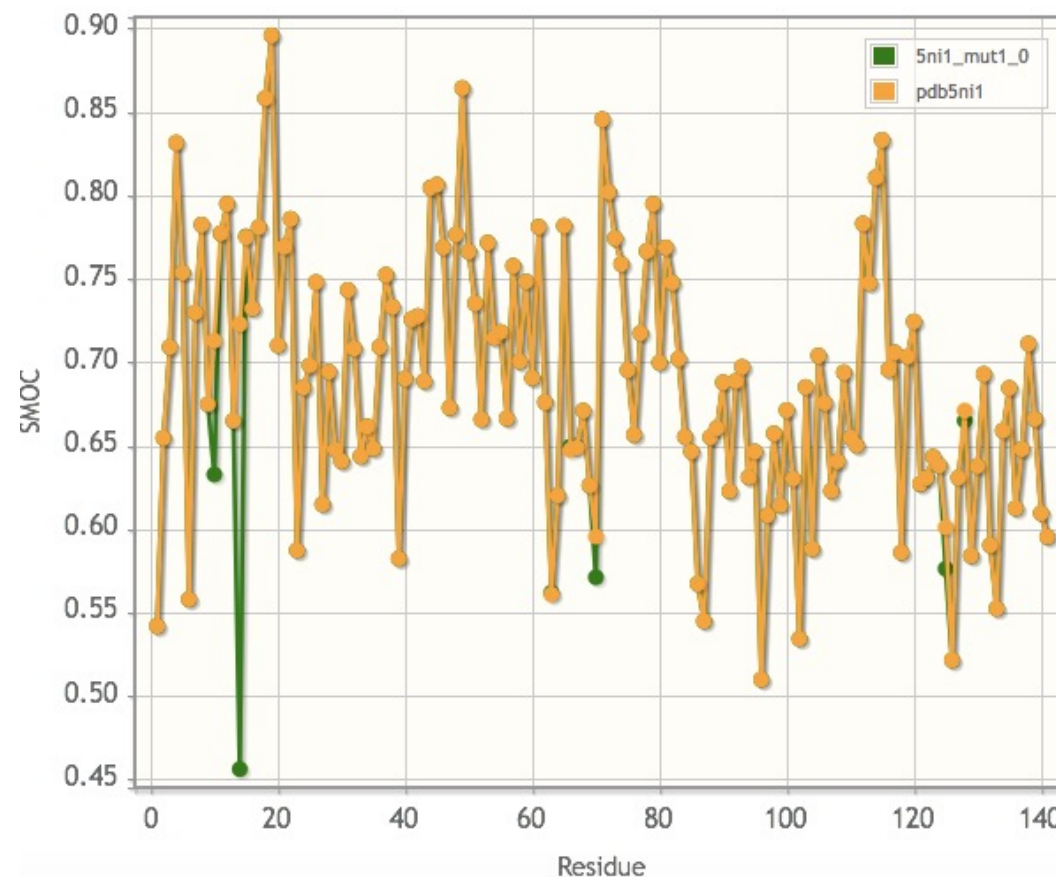
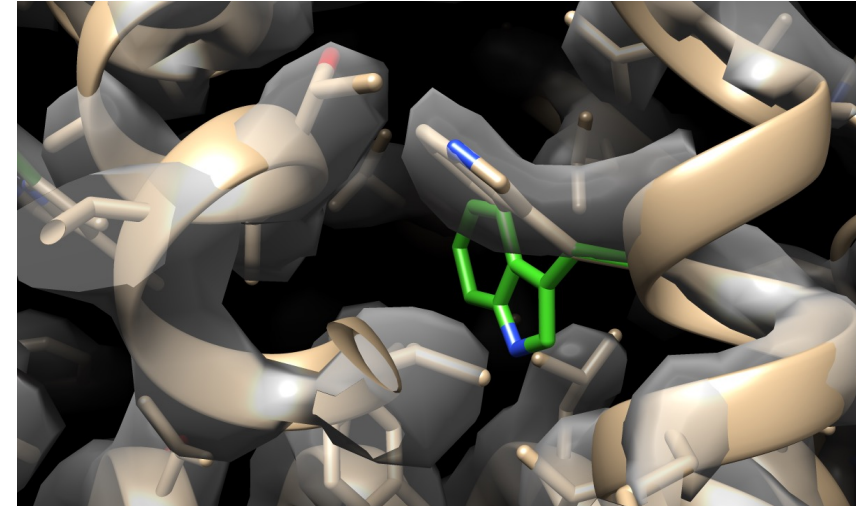
Local agreement with map

Segment based Manders' Overlap Coefficient (SMOC)

An overlap coefficient is calculated over voxels covered by each residue (and the local neighborhood)

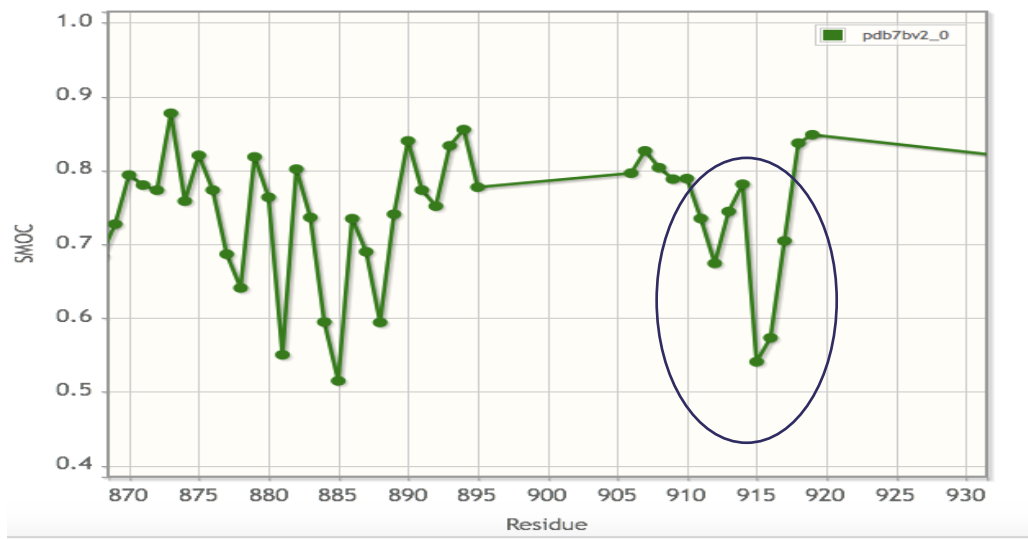
$$SMOC = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \sqrt{\sum_{i=1}^n Y_i^2}}$$

Joseph *et al. Methods* 2016 , Farabella *et al. Acta D* 2015



Local agreement with map

2.5Å SARS-CoV2 RNA pol



Chojnowski G. Acta Cryst 2022

CheckmySequence Chain: A

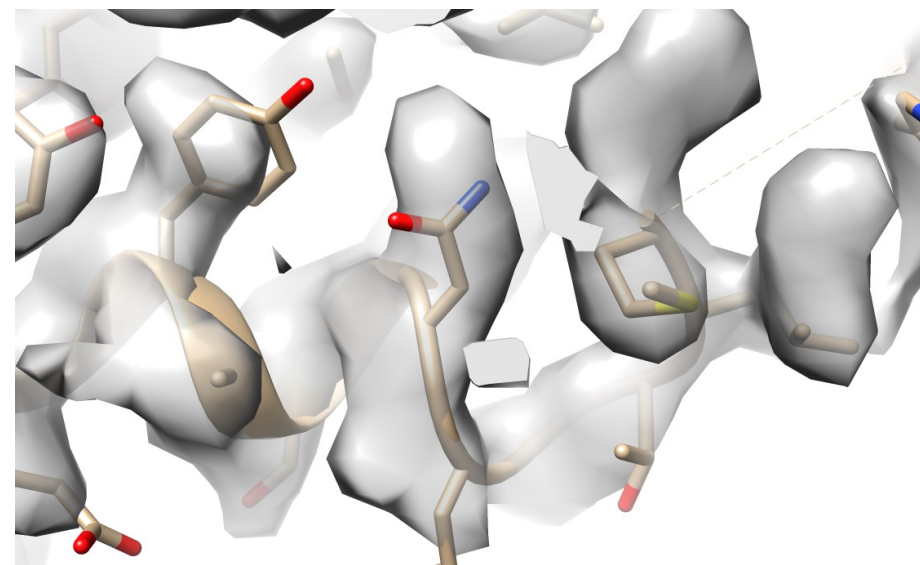
*** Register shift:

protein start: 906 end: 919 start_new: 915 end_new: 928 -log(pvalue): 1.0068789995424352 si: 100.0

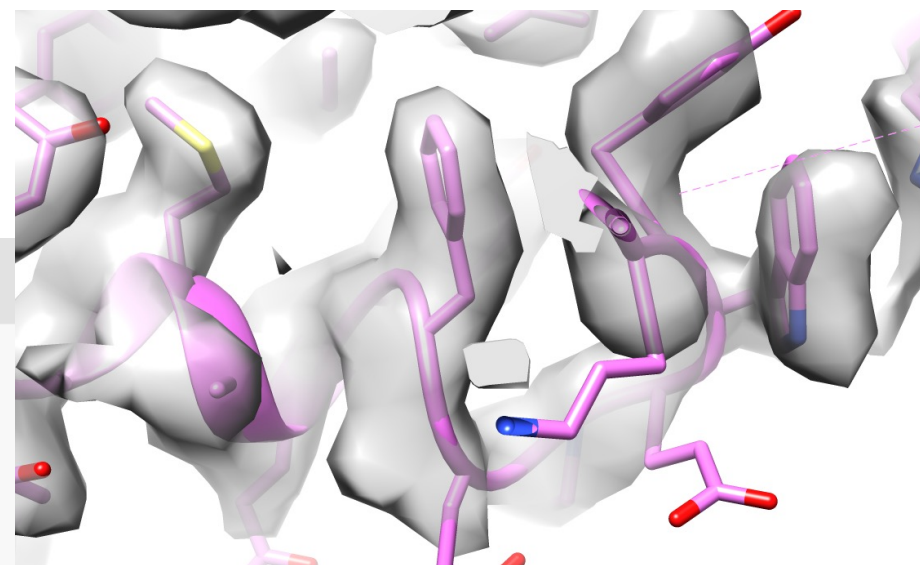
model seq: dmysvMLTNDNTRSRYWEPEfyeamytphtvlqgg

new seq : dmysvmltndntsrYWEPEFYEAMYPHTvlqgg

Deposited

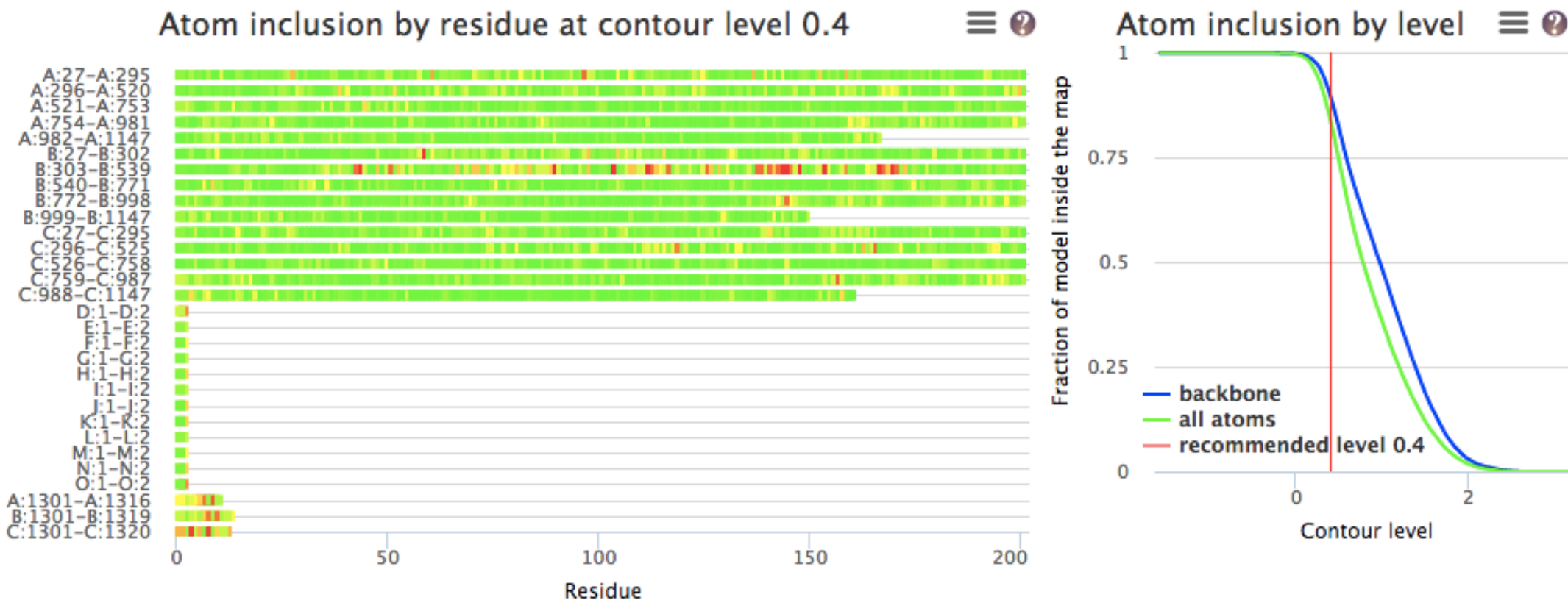


Remodelled



Local agreement with map

Atom inclusion score



Local agreement with map

Lawson et al. *Nature Methods*, 2021

FDR validation score

MapQ	
SMOC	
SCCC	
PHENIX	
FSC_Q	
FDR_score	

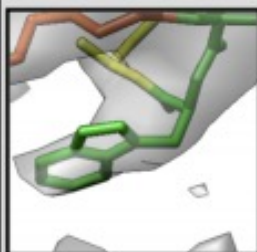
Value	Z-score
0.20	-3.34
0.78	-1.40
0.60	-1.25
0.73	-0.76
1.37	4.89
0.49	-4.85

TRP188

T0007EM192_2



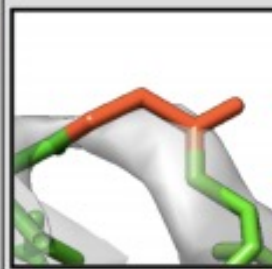
5a63



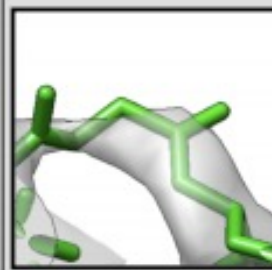
Value	Z-score
0.53	-0.66
0.70	-0.29
0.55	-0.64
0.73	-0.34
0.25	0.11
0.83	-1.23

GLY86

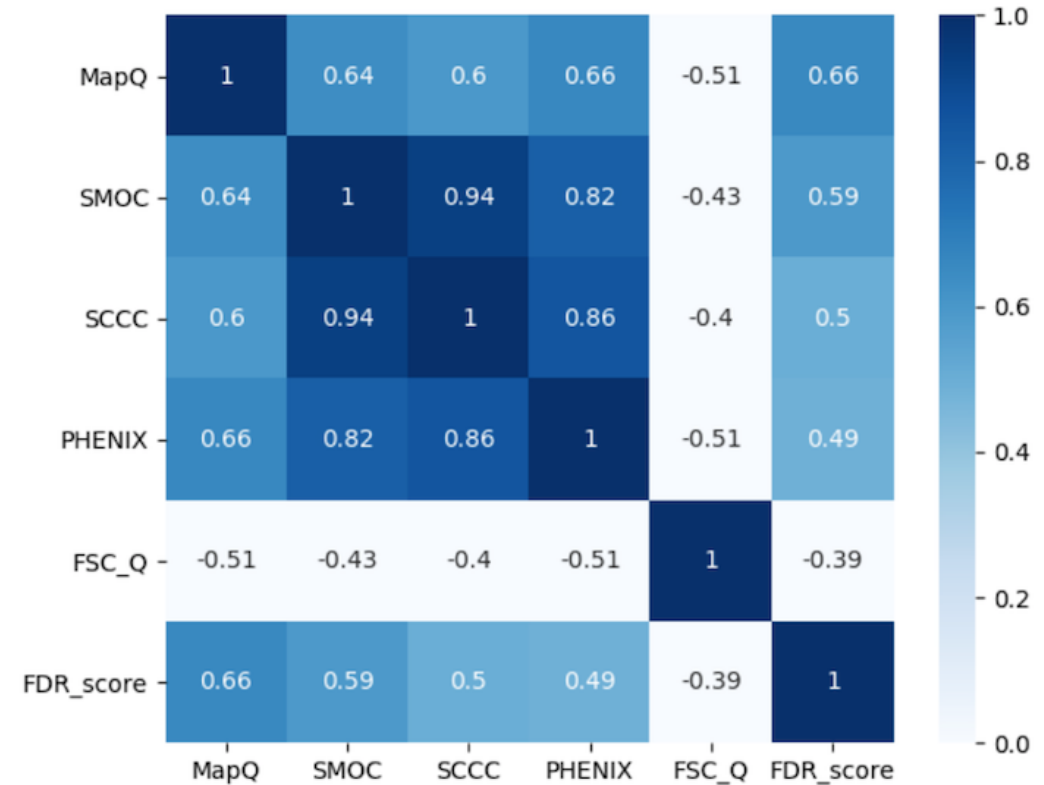
T0104EM060_2



6nbb.2

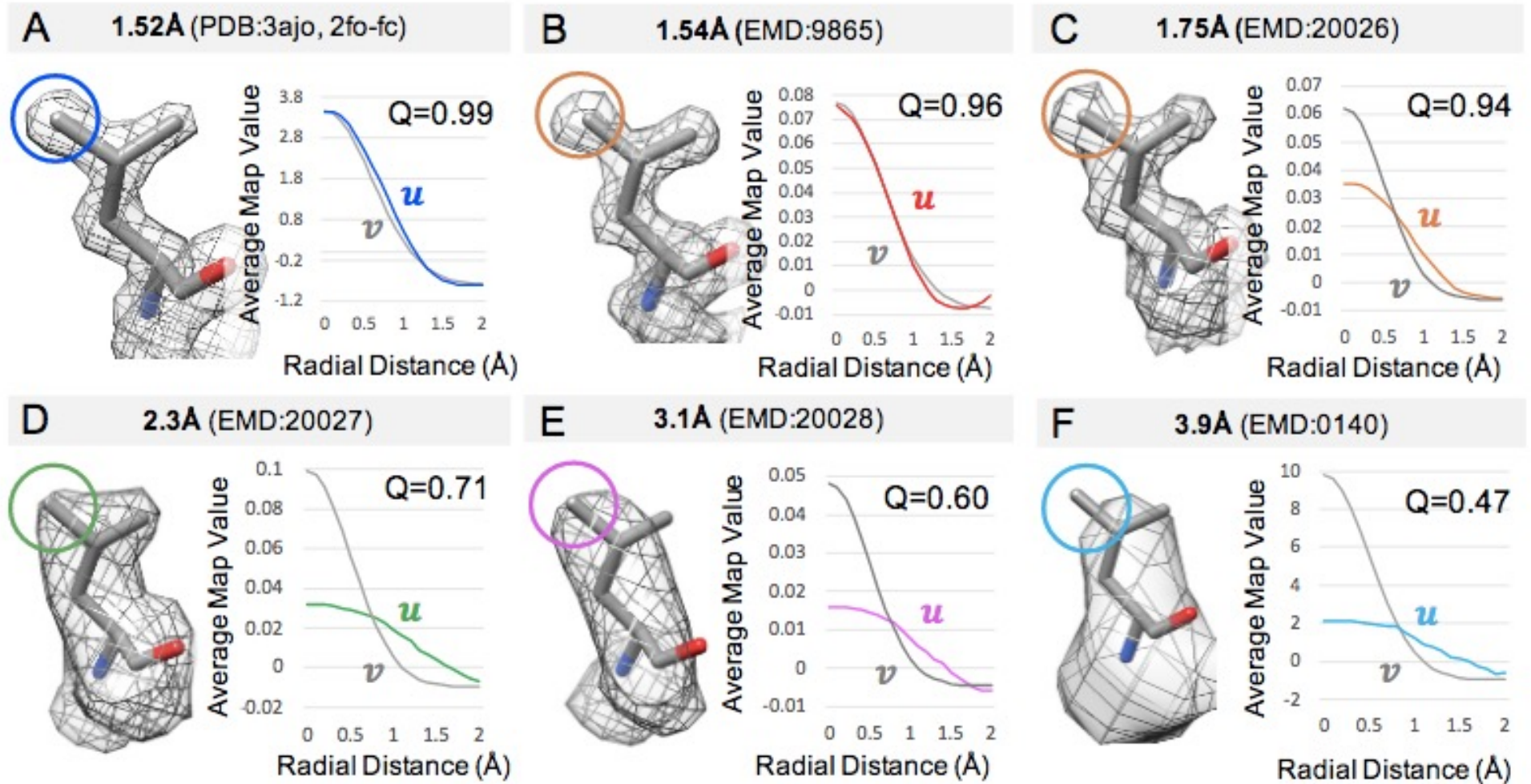


T0104EM060_2



Local agreement with map: resolvability

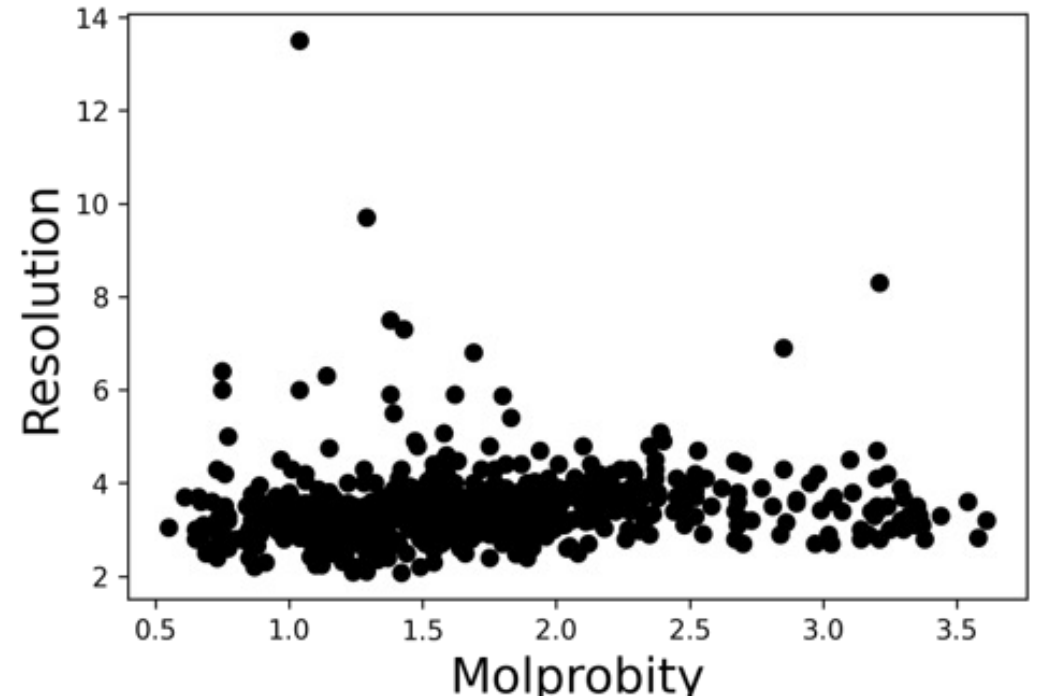
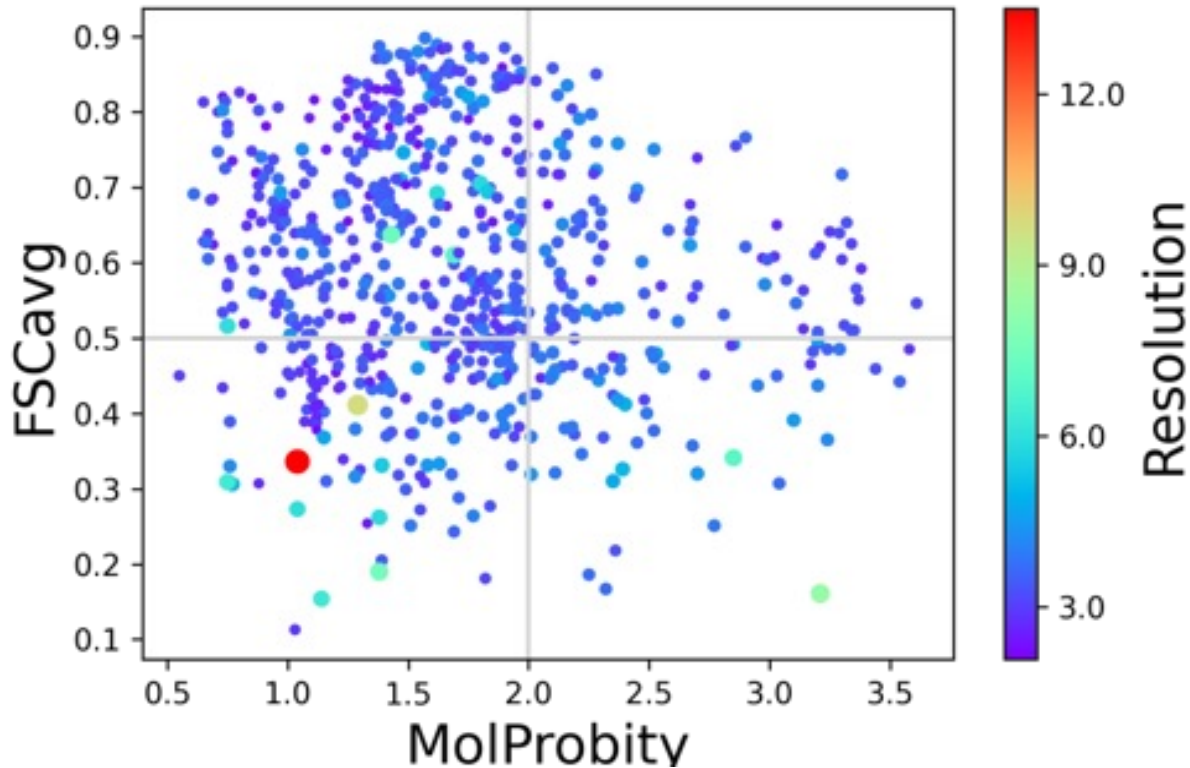
Q score



Model geometry vs fit-to-data

Do the models represent the map data well?

Do we need more validation metrics for publication?



Mean score of structures worse than 3.5Å resolution is 1.8 (< 3.5Å is 1.6).

31.2% of the structures had FSCavg scores worse than 0.5.

Model geometry vs fit-to-data

Geometry ✓

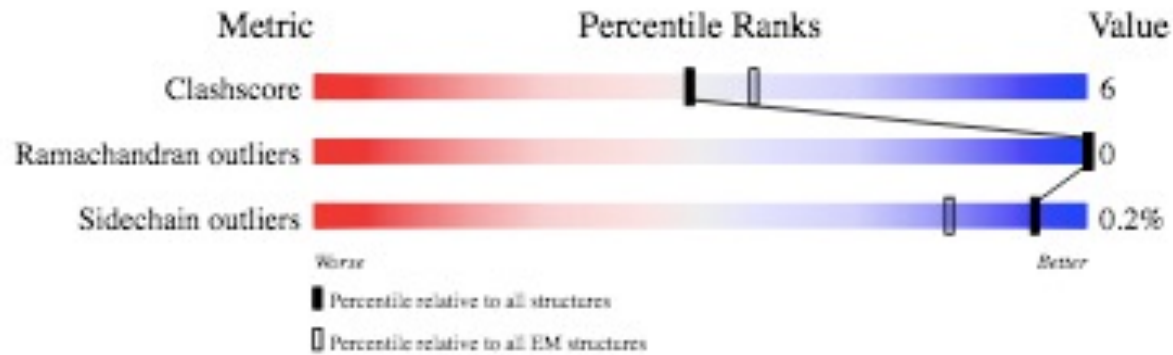
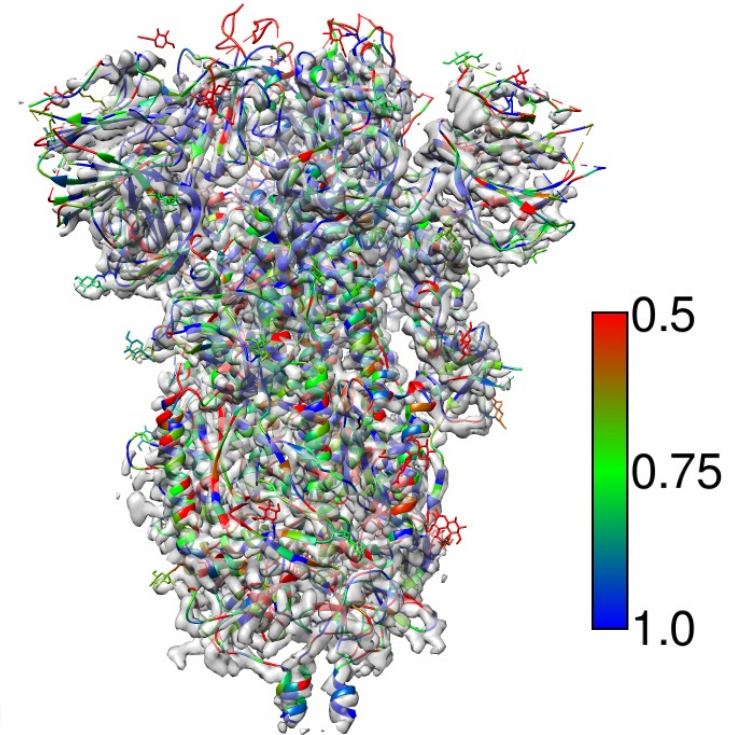


Image:
PDBE/wwPDB

Map agreement ?



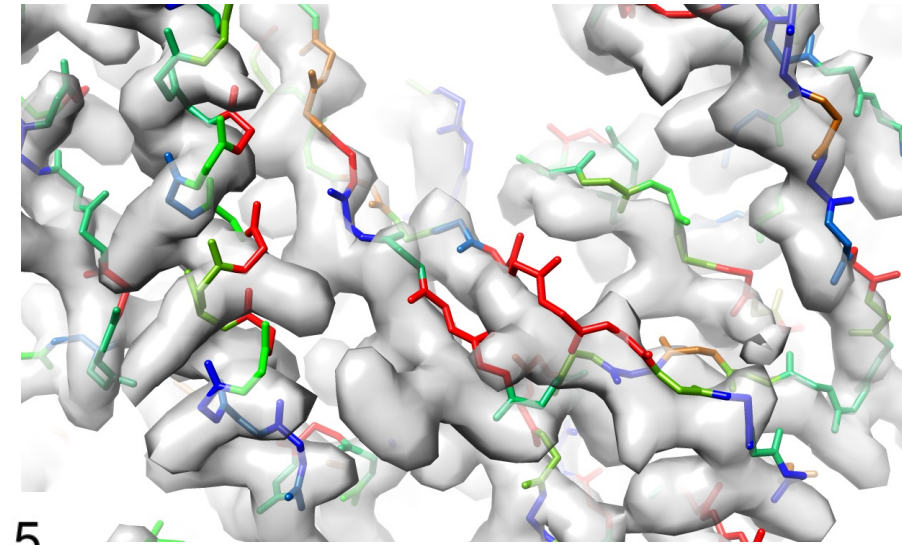
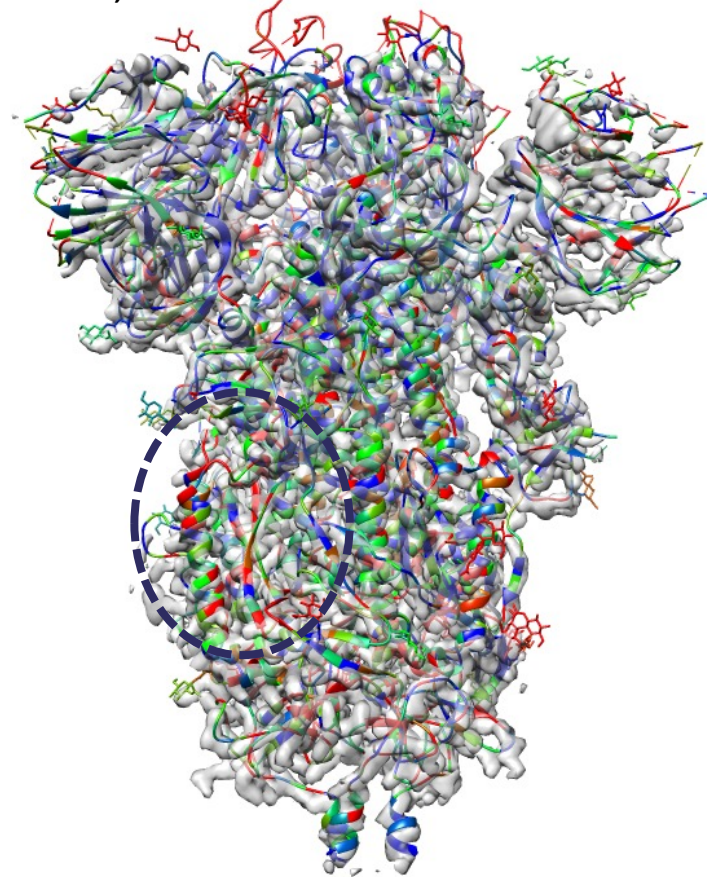
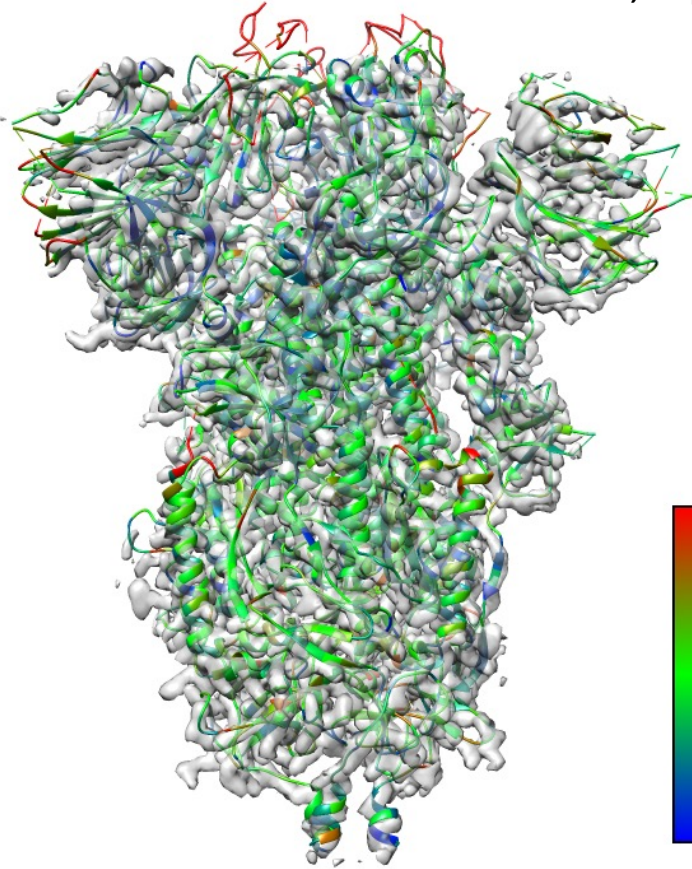
Overfitting ?

Poor fit

~28% of structures have potential issues with backbone trace (FDR-backbone score)

Molprobity score : 1.39

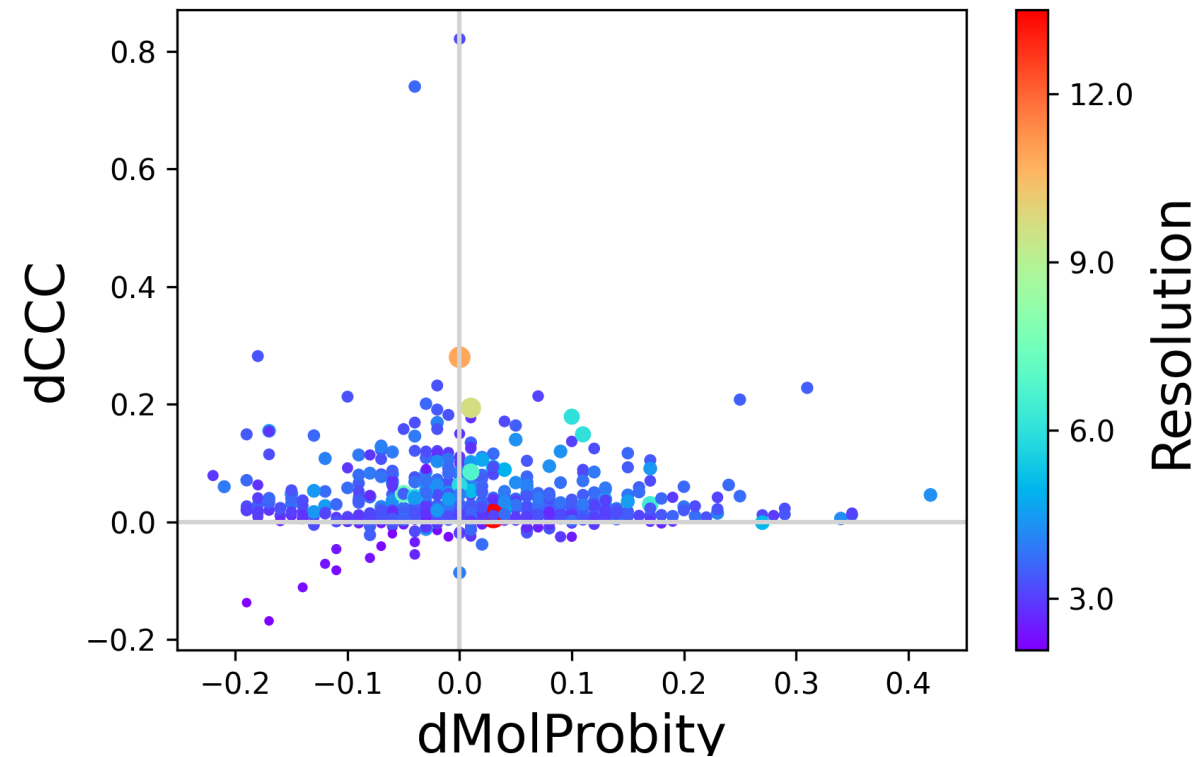
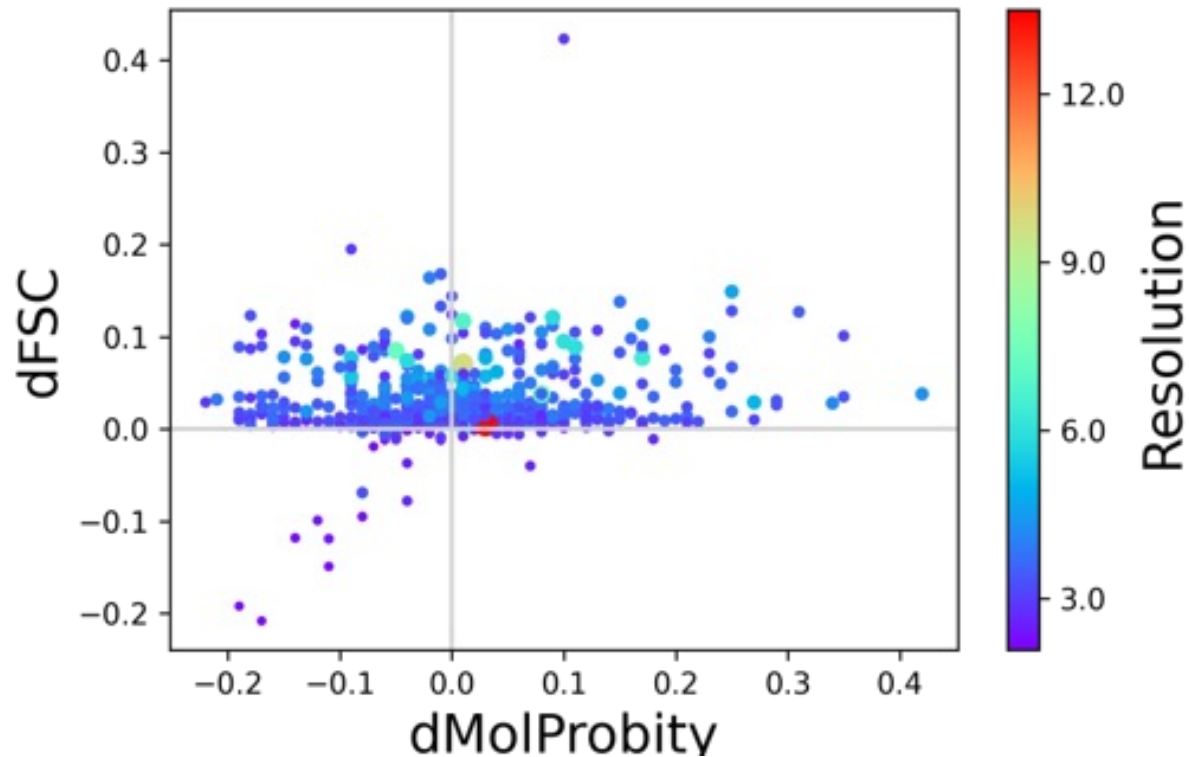
0 Ramachandran outlier, 1 poor rotamer, clashscore: 6.3



3.4Å

Model geometry vs fit-to-data

Can the map fit / representation be improved ?



FSCavg of 94% of structures in the dataset improved with further refinement

~45% of the dataset had improved MolProbity scores as well

Goodness of fit

TEMPy, Coot, Refmac, Phenix,
Emringer, SMOC,
Q-score, FSC-Q

Secondary structure

Molprobity, Coot, CaBLAM, Qmean,
Jpred, ...

Quarternary structure

PI-score, PISA,
Docking scores

Experimental validation

Model geometry

Molprobity, Coot, CaBLAM, Tortoise
What-check
...

Tertiary structure

Verify-3D, ProQ2 (Rosetta),
Prosa, DOPE (Modeller), ModFold, ..

Cross-Validation

Half map (Refmac, Rosetta)
Resolution shells (Direx)

Ensemble assessment (TEMPy)

CCP-EM model validation

Filter jobs by name

Expand all

Reproject

Select

Subtract

Validation

PV

Atomic model sequence check

pipeliner.validation.check_model_sequence.evaluate - Check modeled sequence against expected sequence and EM map

PV

EMDB Validation Analysis

pipeliner.validation.emdb_validation.evaluate - EMDB map and model validation

PV

Atomic model validation

pipeliner.validation.model_validation.evaluate - Validate atomic models with multiple methods

Atomic model validation

RUN

JOB INFO

RESET PARAMETERS

Job alias:

Main

Input model *

required*
pdb5ni1.ent

Input map

emd_3488.map

Resolution *

3.2

Input halfmap 1

run_half1_class001_unfil_5me2.mrc

Input halfmap 2

run_half2_class001_unfil_5me2.mrc

Input map mask

Contour Level

0.08

Input FDR confidence map

Input sample sequence *

required*
5ni1_entry.fasta

Molprobability

Yes No

CCP-EM model validation

Molprobity Yes No

Servalcat FSC Yes No

TEMPy global scores Yes No

TEMPy SMOC Yes No

FDR backbone score Yes No

Confidence map not provided - will be calculated.Requires unmasked input map, preferably raw-map

TORTOIZE Yes No

'tortoise' from CCP4 is currently broken on Mac

CheckMySequence Yes No

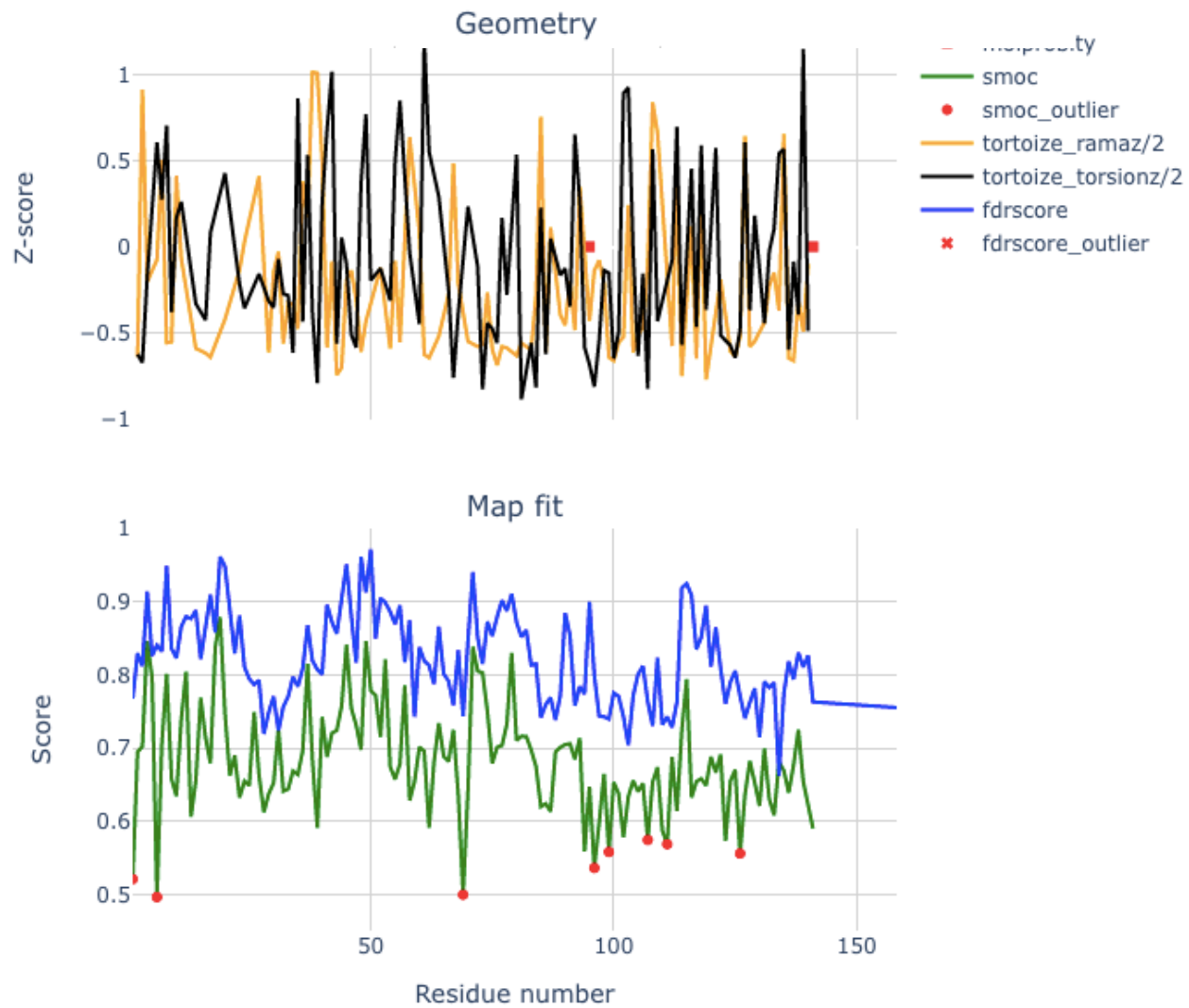
CCP-EM model validation

Outlier clusters



Cluster	Chain_Residue	Outlier Type(s)
6	B34	smoc (raw:0.514;Z:-1.88)
6	B301	fdrscore (raw:0.558;Z:-2.2)
6	B305	fdrscore (raw:0.525;Z:-2.826)
7	B104	smoc (raw:0.559;Z:-1.981)
7	B303	fdrscore (raw:0.6;Z:-2.221)
7	B304	fdrscore (raw:0.548;Z:-2.6)
8	C95	dihedral_angles (1_C_94_CA-1_C_94_C-1_C_95_N-1_C_95_...
8	C96	smoc (raw:0.536;Z:-1.854)
8	C99	smoc (raw:0.558;Z:-1.66)
0	A95	dihedral_angles (1_A_94_CA-1_A_94_C-1_A_95_N-1_A_95_...
0	A96	smoc (raw:0.511;Z:-2.218)
1	A63	smoc (raw:0.562;Z:-1.596)
1	A61	torsion_z (2.438)
2	A86	smoc (raw:0.569;Z:-1.772)
2	A87	smoc (raw:0.542;Z:-2.165)
3	A102	smoc (raw:0.535;Z:-1.909)
3	A302	fdrscore (raw:0.584;Z:-2.035)
4	A38	rama_z (2.007)
4	A20	rama_z (2.014)

CCP-EM model validation



CheckmySequence Chain: C

*** Sequence mismatch:

protein start: 1 end: 141 evalue: 2.1e-63 seq2ref_si: 98.54014598540147

1234567890123456789012345678901234567890123456789012345678901234567890

model -----VLSPADKTNVKAAWGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV

refseq AAAAAAAAAAVLSPADK-NVKAAWGKVGASAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV

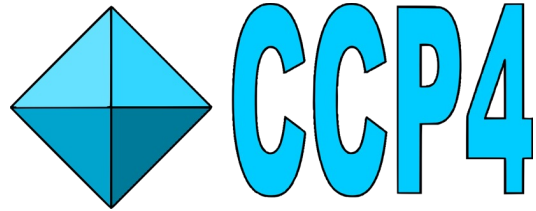
model AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLSDKFLASVSTVLTSKYR-----

refseq AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLSDKFLASVSTVLTSKYRAAAAAAAAAA

model -

refseq A

Acknowledgements





Science and
Technology
Facilities Council

Scientific Computing

Questions?





Science and
Technology
Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

 [@SciComp_STFC](https://twitter.com/SciComp_STFC)