Science and Technology Facilities Council

Scientific Computing

Atomic Model Validation

Agnel Praveen Joseph
CCP-EM STFC

08/11/2024
Icknield workshop

CCP-EM

Outline

1 Background

2 Validation: Why is it important?

3 Tools for atomic model validation

4 Some examples

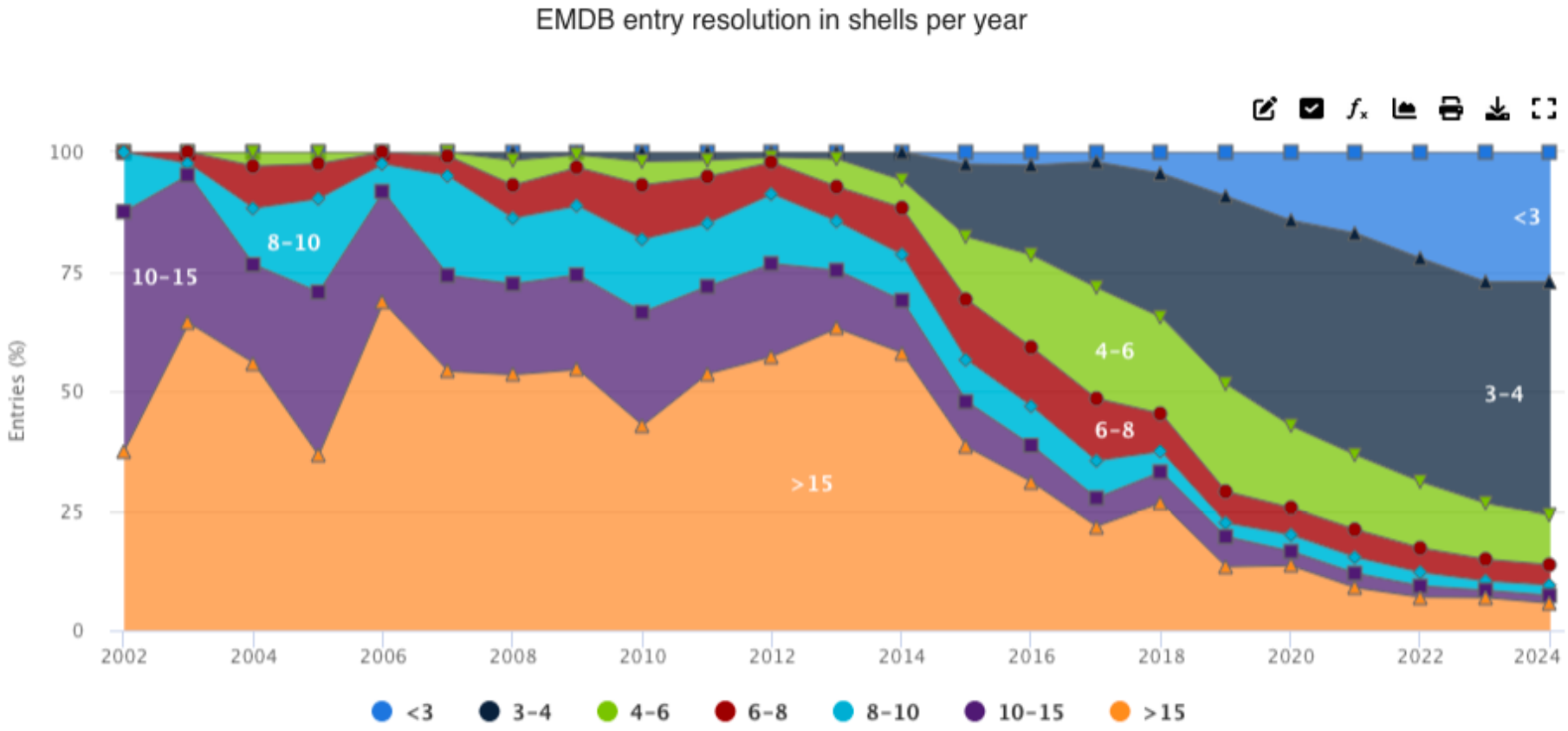
5 Model validation in Doppio

Science and Technology Facilities Council

Scientific Computing

EM DataBank: map resolution

>39k cryo-EM reconstructions in EMDB

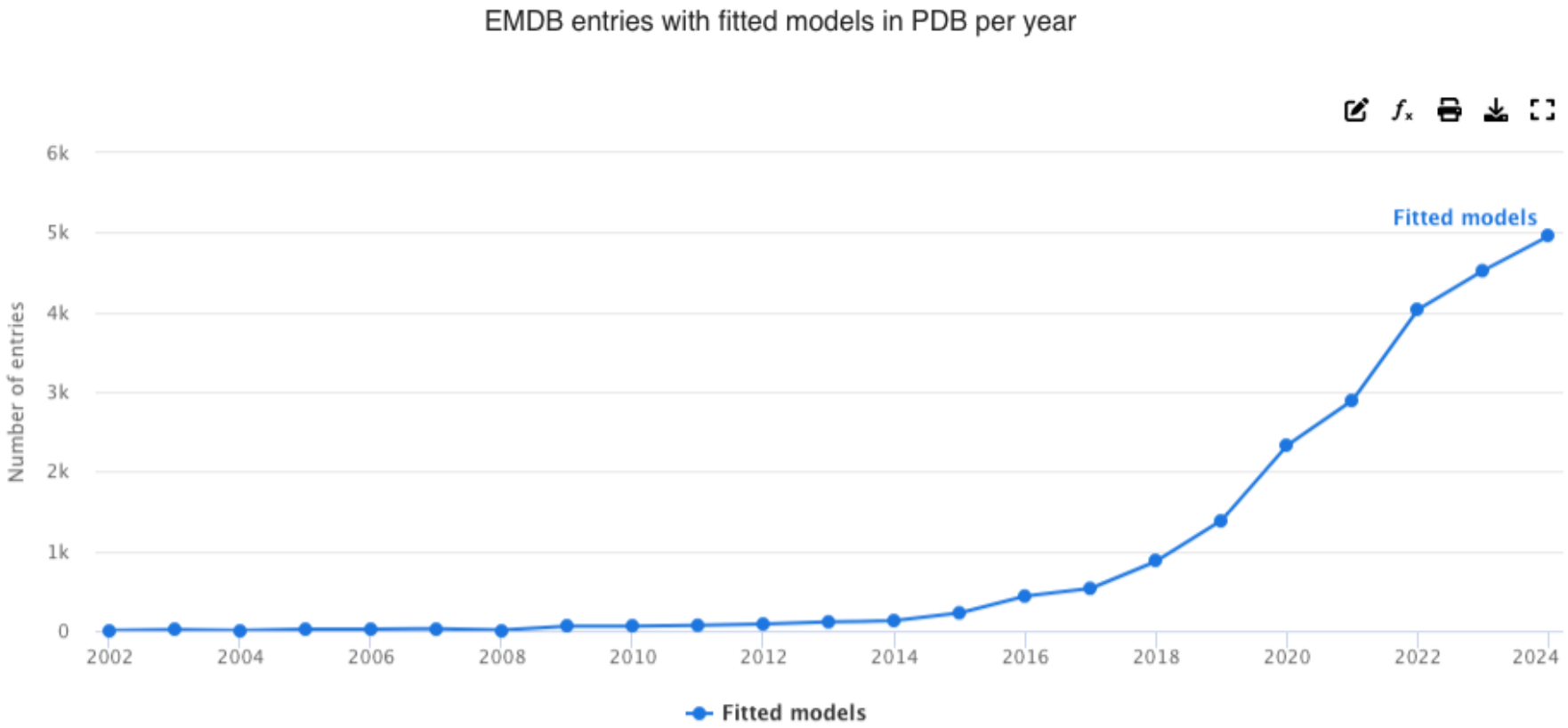


Last 3 years
3-6 Å: **~60%**

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

Science and Technology Facilities Council
Scientific Computing

EMDB: maps with fitted models

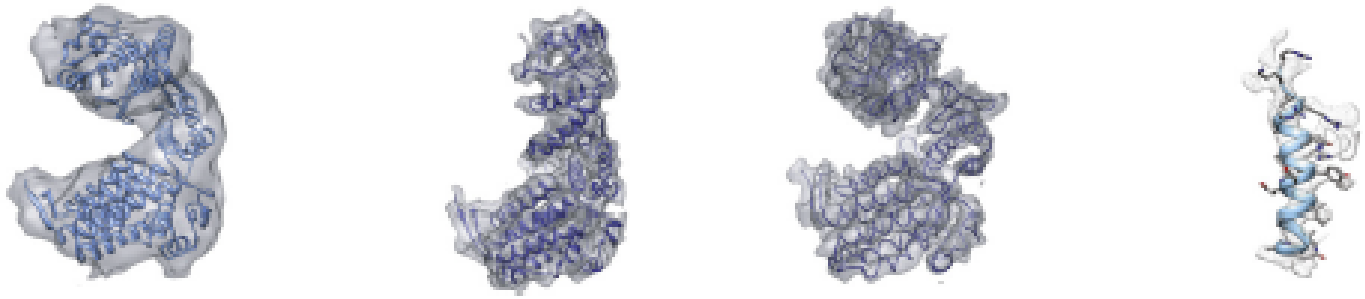


Resolution

- < 2Å [197]
- ≥ 2Å and < 3Å [5360]
- ≥ 3Å and < 4Å [12130]
- ≥ 4Å and < 5Å [2728]
- ≥ 5Å and < 6Å [448]
- ≥ 6Å and < 8Å [865]
- ≥ 8Å and < 12Å [611]
- ≥ 12Å and < 16Å [171]

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

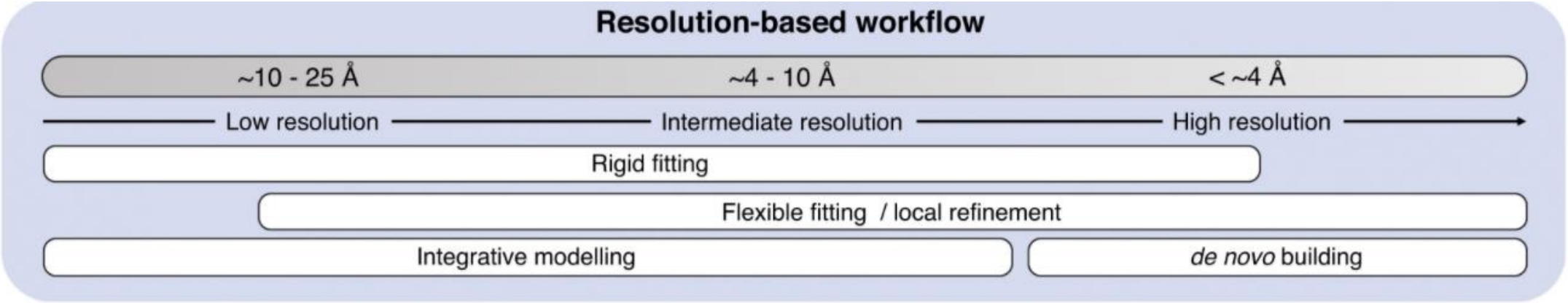
Features interpretable at different resolutions



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
			Individual beta strands
	Alpha helices	Alpha helices	Alpha helices
			Pitch of alpha helices
		Pitch of RNA helices	Pitch of RNA helices
			Phosphate "bumps"
			Side chains

Malhotra S. et al., COSB 2019



Scier
Tech
Facili
Scientific Comp

Science and Technology Facilities Council

Scientific Computing

Model validation

Science and Technology Facilities Council

Scientific Computing

Aspects of validation

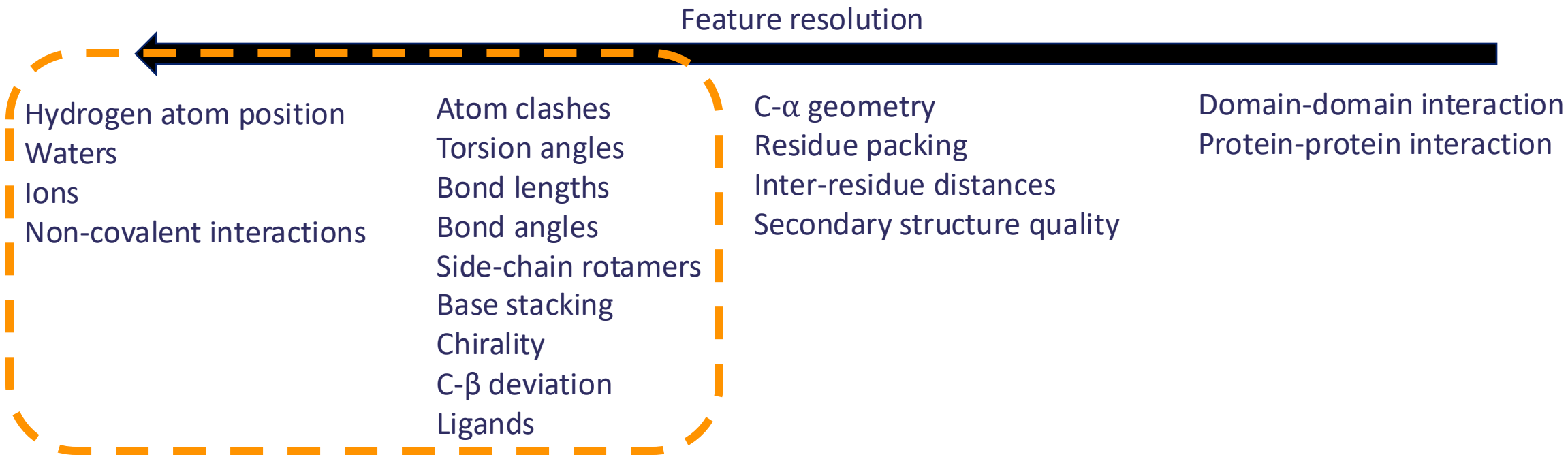
Model quality (geometry)

Fit to data (Global and Local)

Overfitting

Model bias

Model quality (geometry)

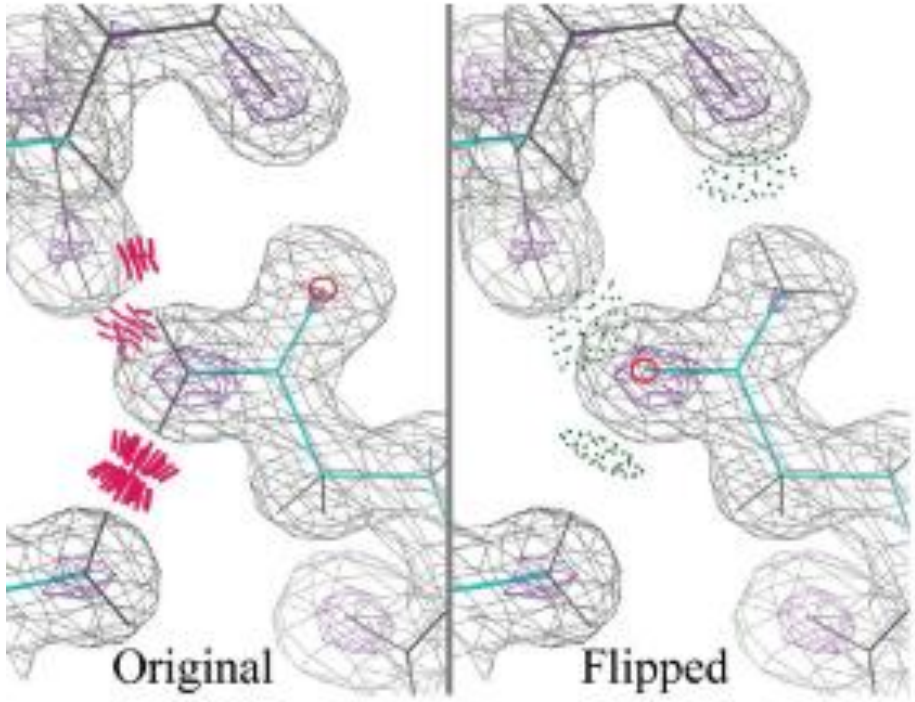


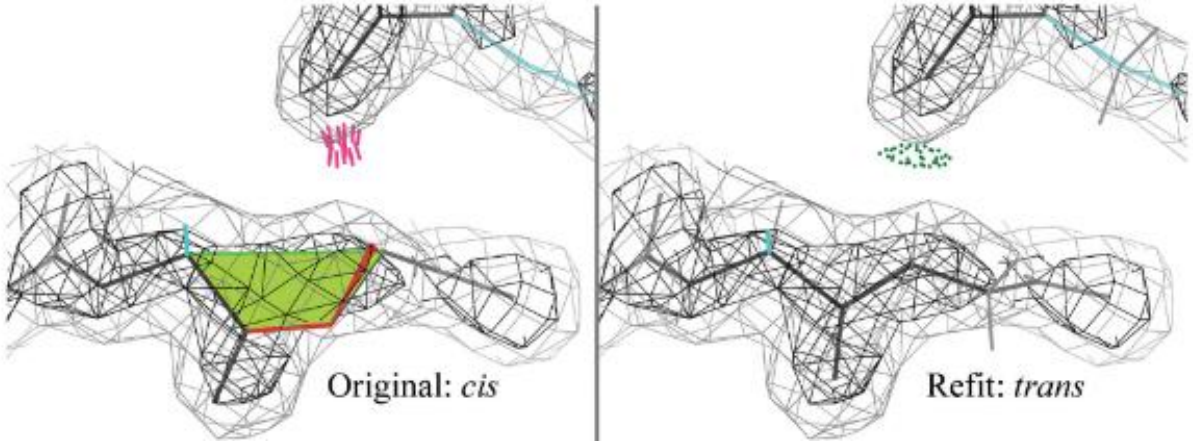
Molprobity, Coot/Moorhen, Isolde, PDB-REDO and others

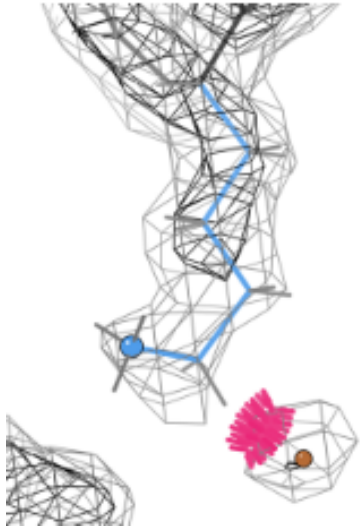
Richardson et al. Acta D, 2018
Prishant et al. Prot Sci., 2020
De Vries et al. IUCrJ, 2021
Casanal et al. Prot Sci. 2020
https://moorhen.org/

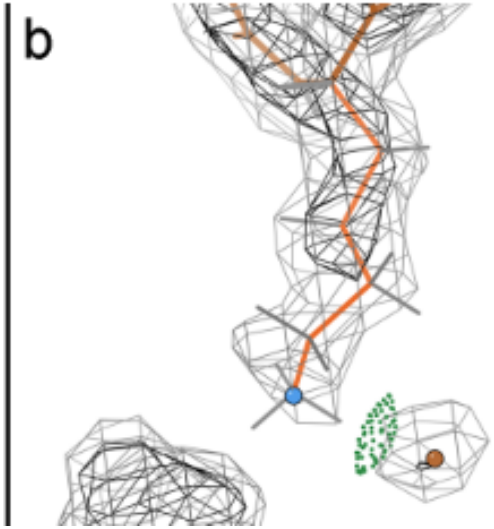
Model quality (geometry)

Clashes









Richardson et al. *Acta D*, 2018
Prishant et al. *Prot Sci*., 2020

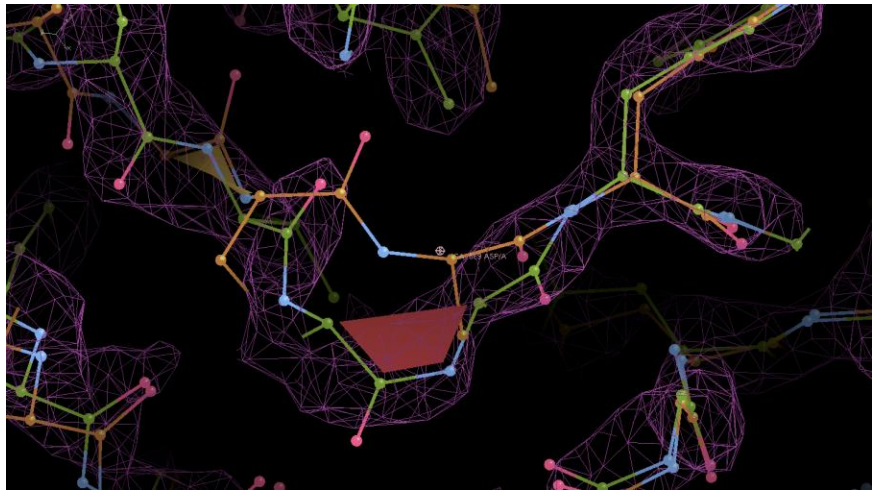
Model quality (geometry)

Cis-peptides

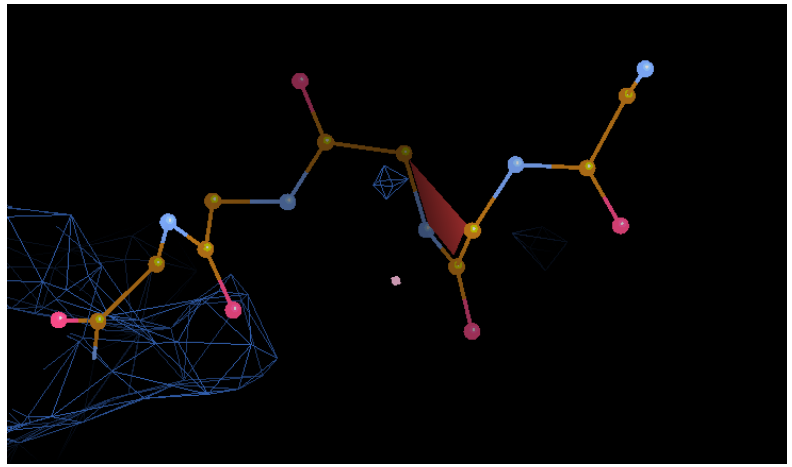
Cis-Pro : 5-6%

Cis-nonPro: ~0.05%

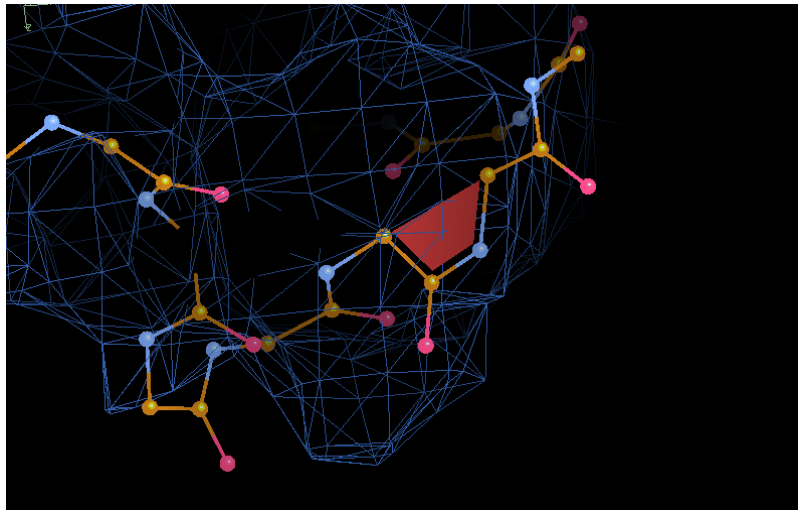
genuine



?



?



Science and Technology Facilities Council

Scientific Computing

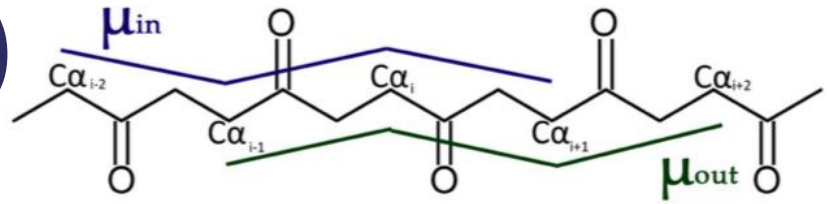
Richardson et al. *Acta D*, 2018

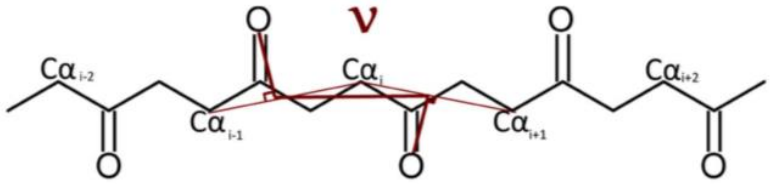
Prishant et al. *Prot Sci.*, 2020

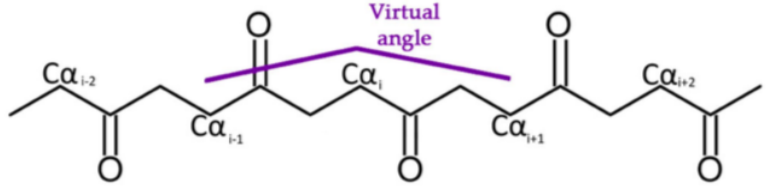
Model quality (geometry)

CaBLAM

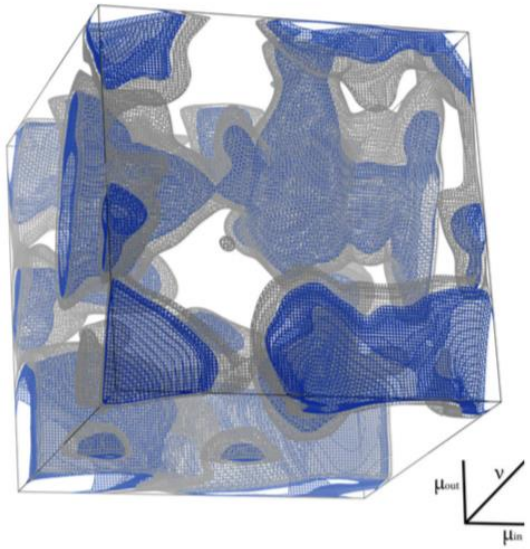
C-Alpha Based Low-resolution Annotation Method

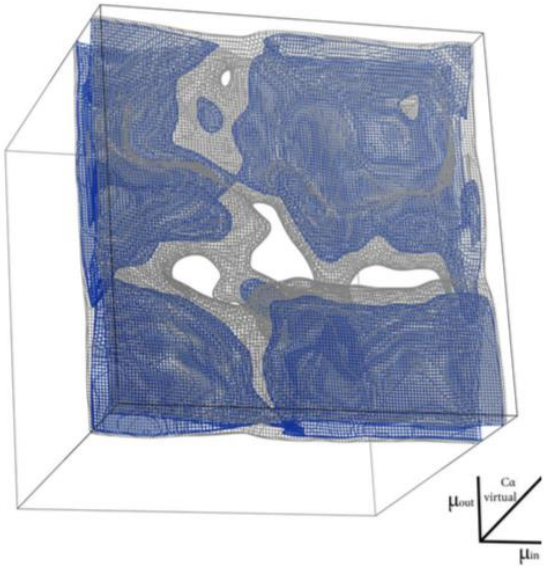


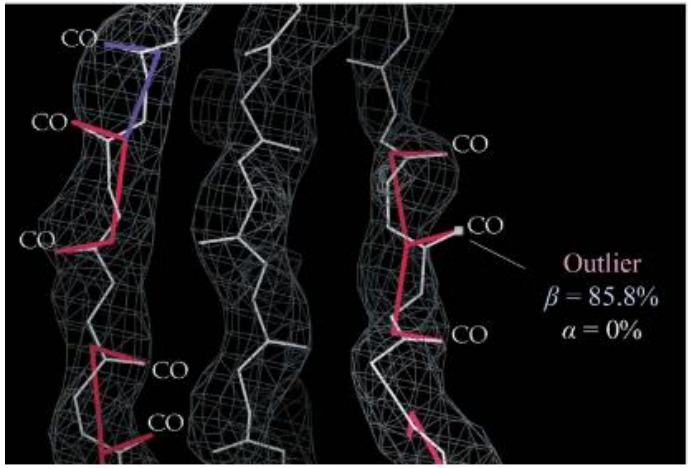




Backbone geometry check







Science and Technology Facilities Council

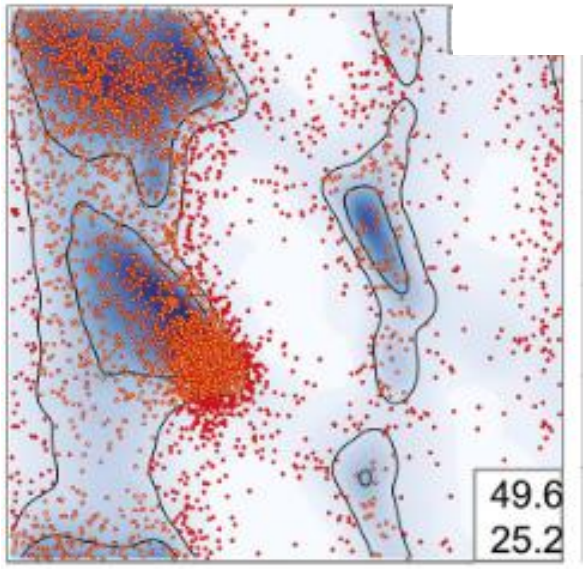
Scientific Computing

Williams et al. *Prot Sci*., 2018

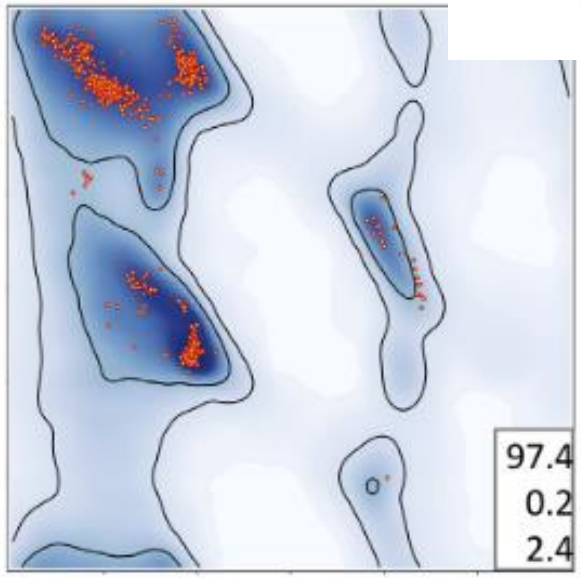
Prishant et al. *Prot Sci*., 2020

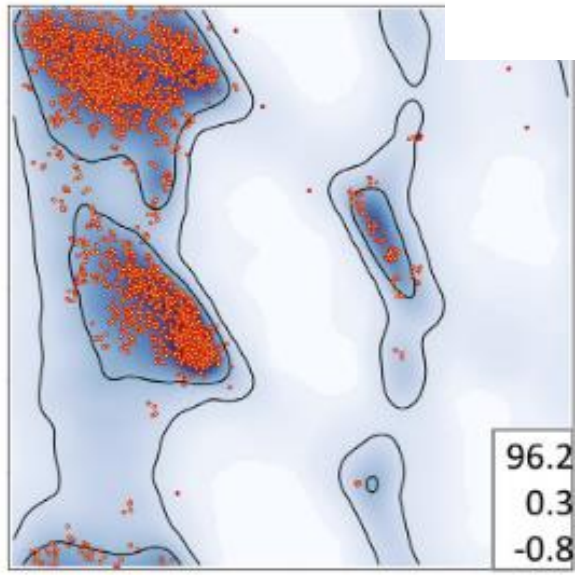
Model quality (geometry)

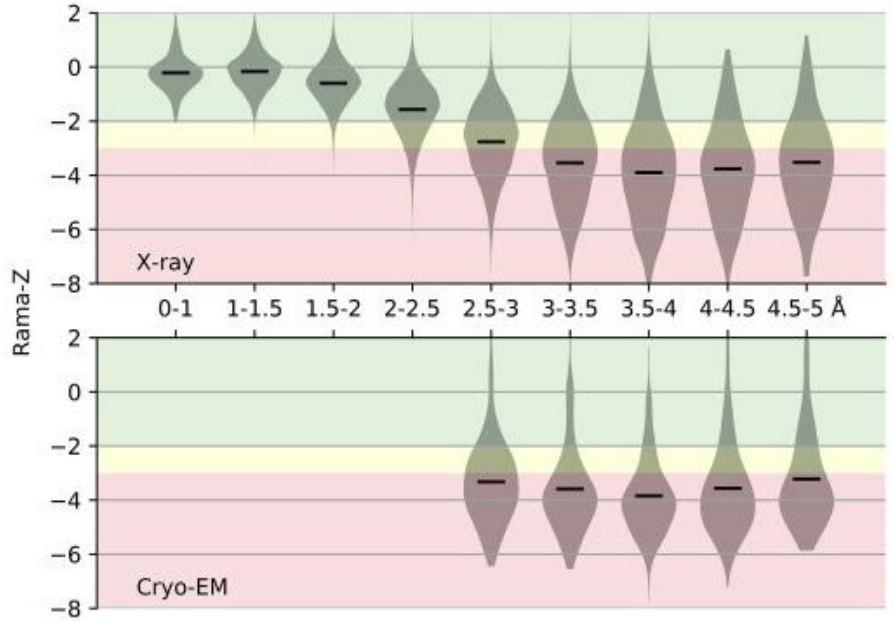
Ramachandran outliers



Ramachandran z-scores







Science and Technology Facilities Council

Scientific Computing

Sobolev et al. *Structure*, 2020

Fit to data

How well the model agrees with data both globally and locally ?

Is the model traced/fitted to the background rather than the molecular volume?

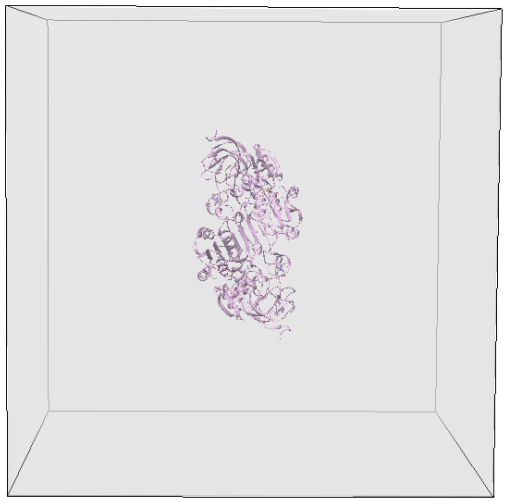
Is the model overfitted (to noise in the map) ?

Global agreement with map

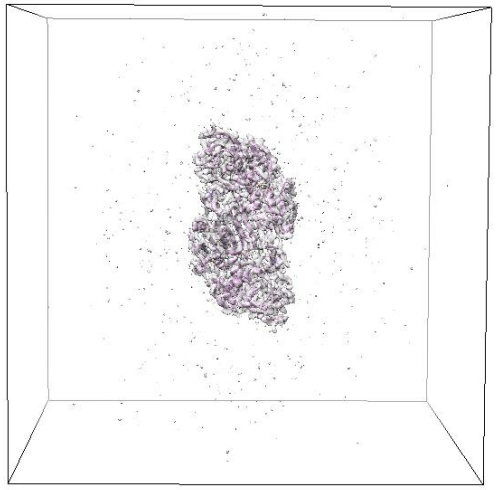
Cross-Correlation Coefficient

Values vary depending on map processing, resolution

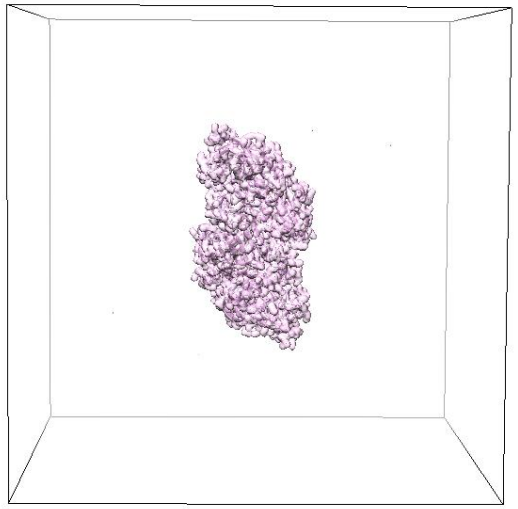
2.9Å full map: 0.32



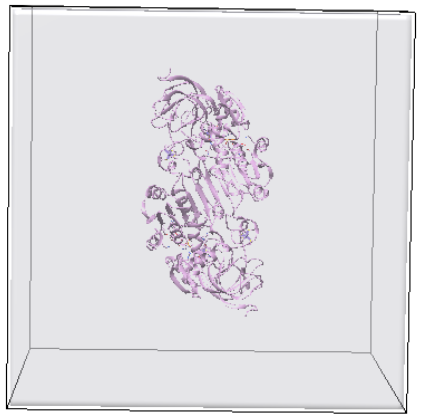
Contoured map: 0.59



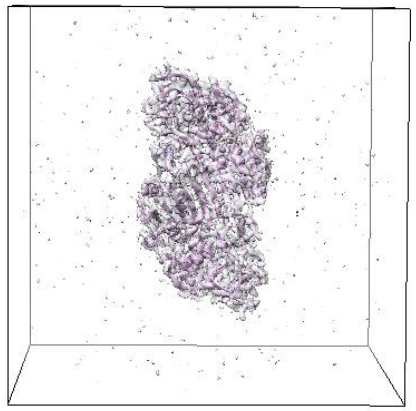
Map-model overlap mask: 0.42



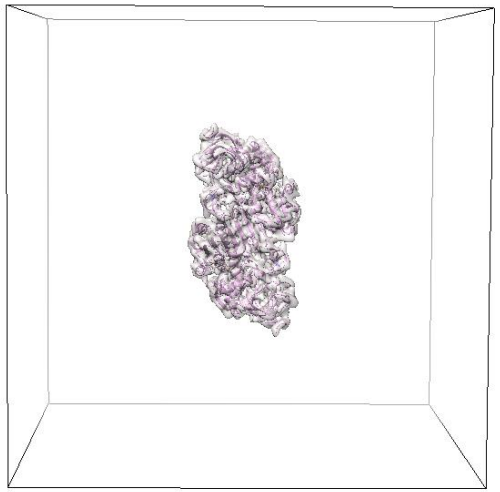
Cropped full map: 0.37



Cropped,contoured map: 0.55



5Å contoured map: 0.76

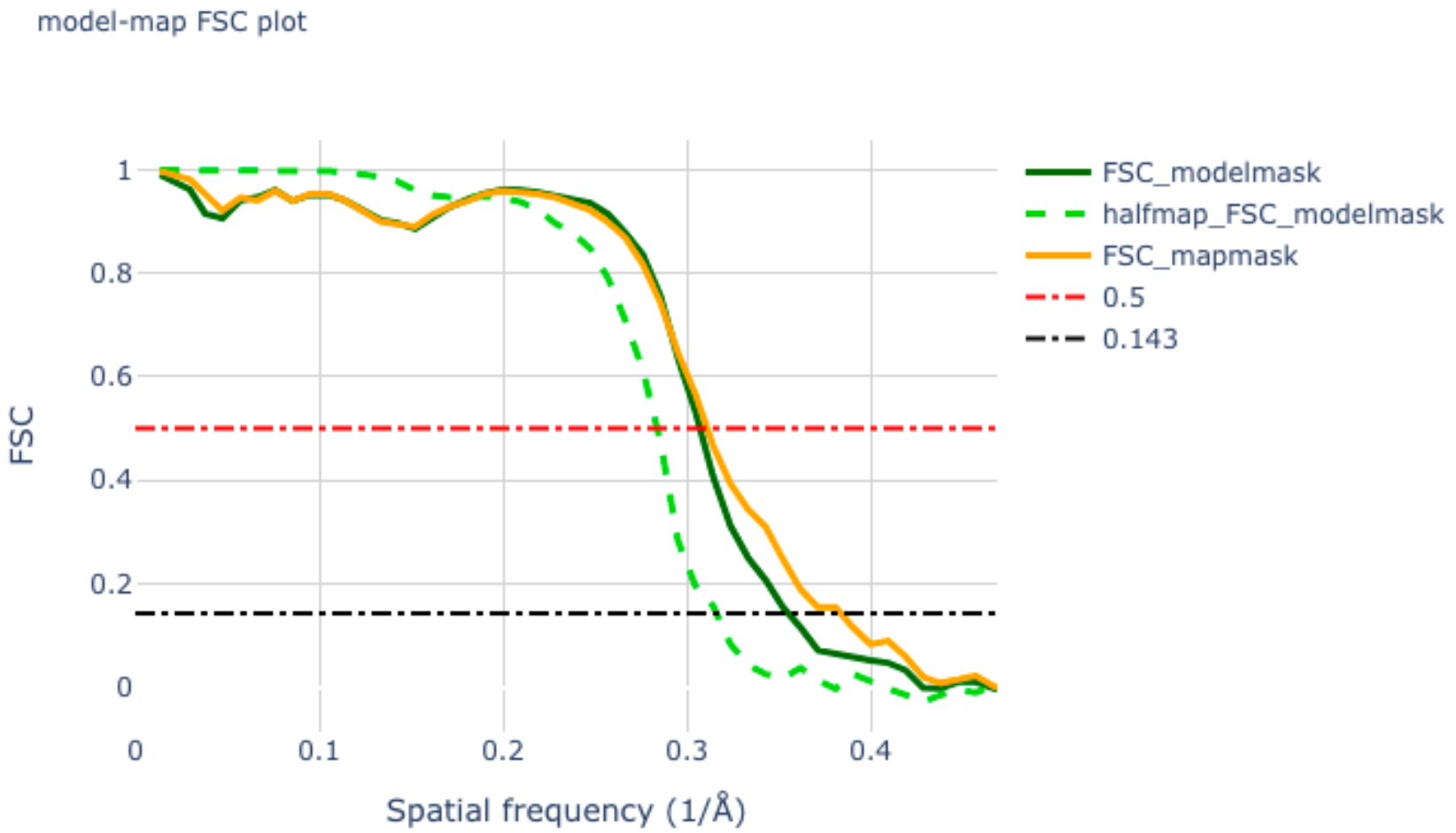


Science and Technology Facilities Council

Scientific Computing

Global agreement with map

Model-map FSC



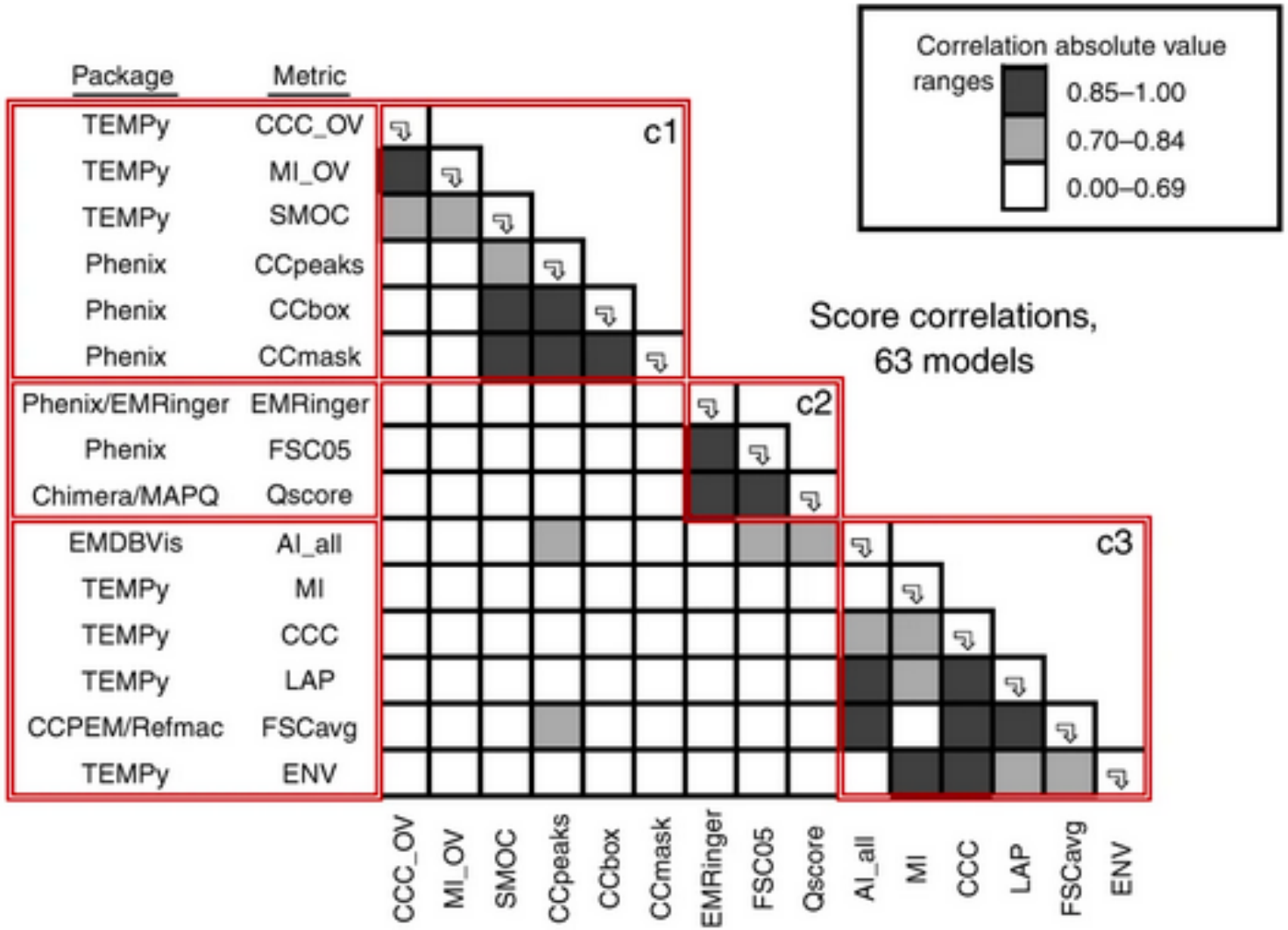
Brown et al. Acta D, 2015
Yamashita et al. Acta D, 2021

Rosenthal and Henderson JMB 2003

Science and Technology Facilities Council

Scientific Computing

Global agreement with map



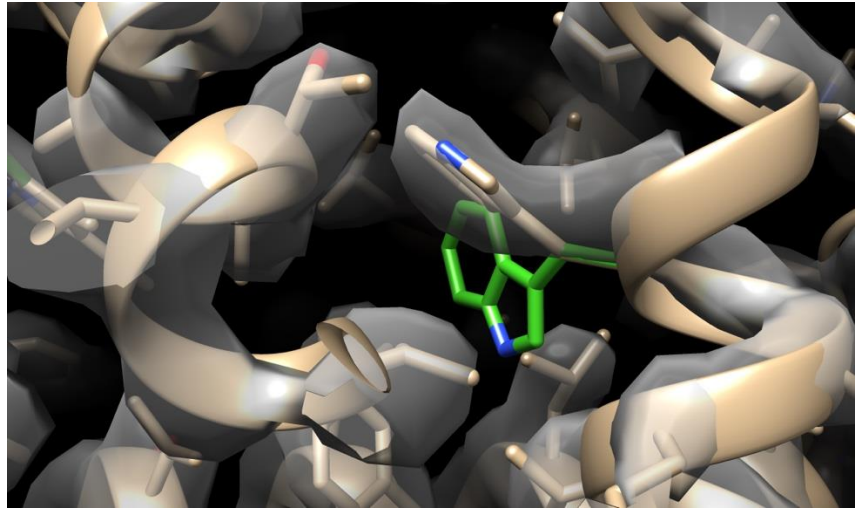
Lawson et al. *Nature Methods*, 2021

Science and Technology Facilities Council

Scientific Computing

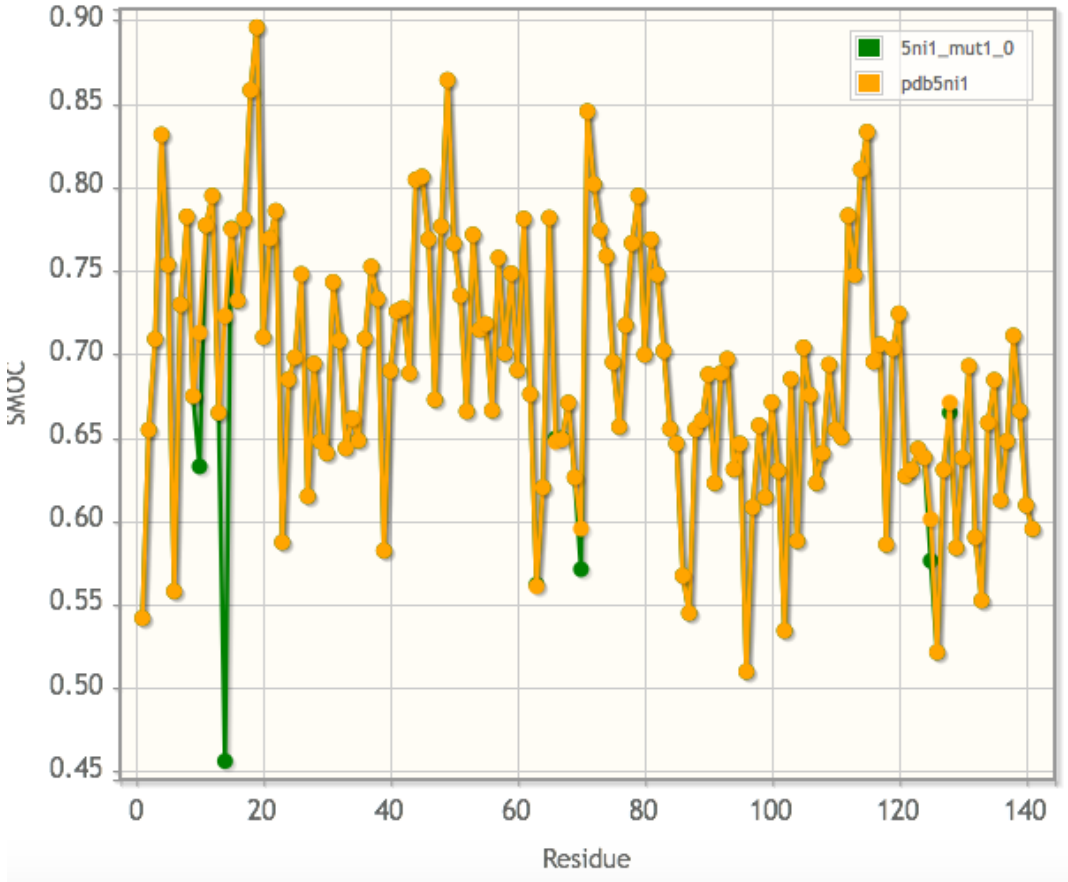
Local agreement with map

Segment based Manders' Overlap Coefficient (SMOC)



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

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



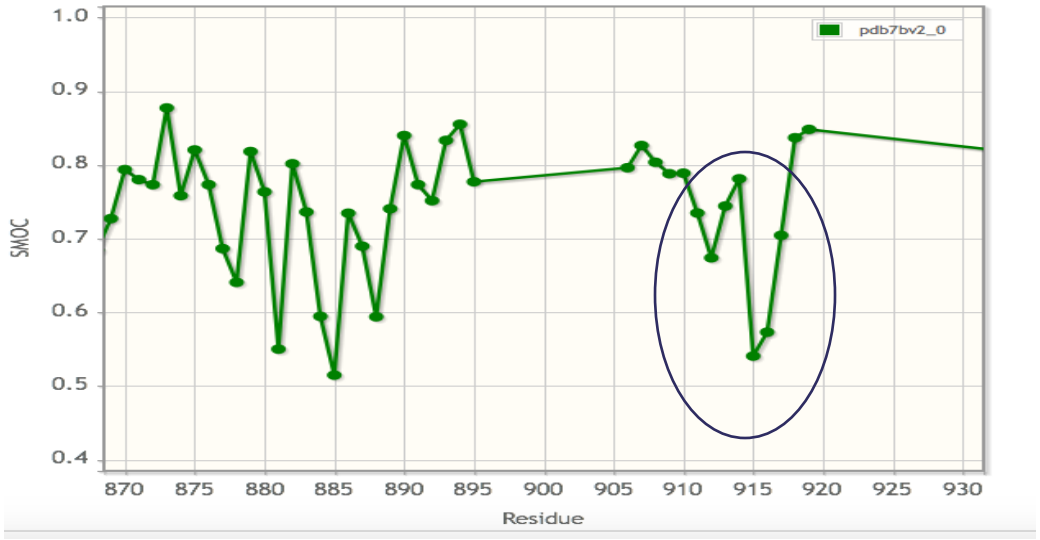
Joseph *et al. Methods* 2016 , Farabella *et al. Acta D* 2015

Science and Technology Facilities Council

Scientific Computing

Local agreement with map

2.5Å SARS-CoV2 RNA pol

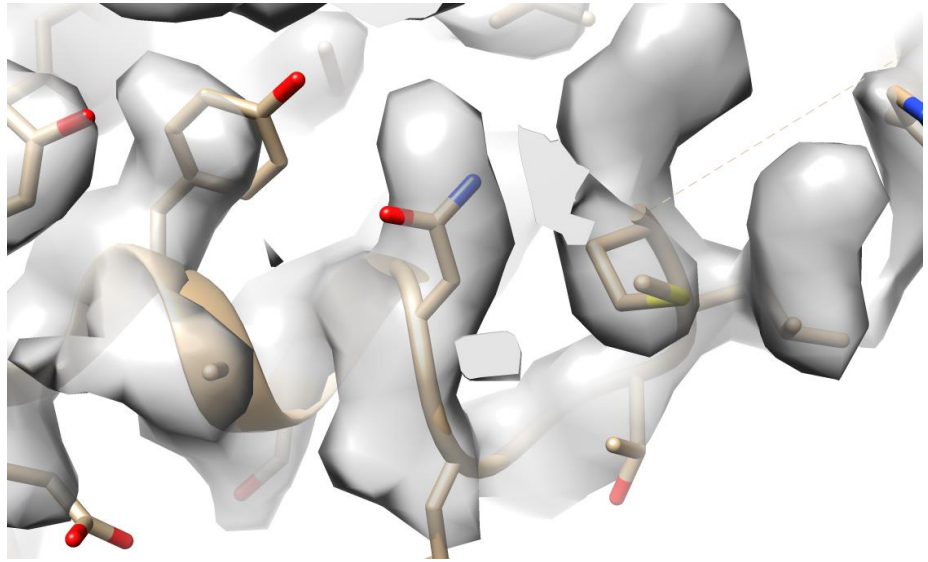


Chojnowski G. Acta Cryst 2022

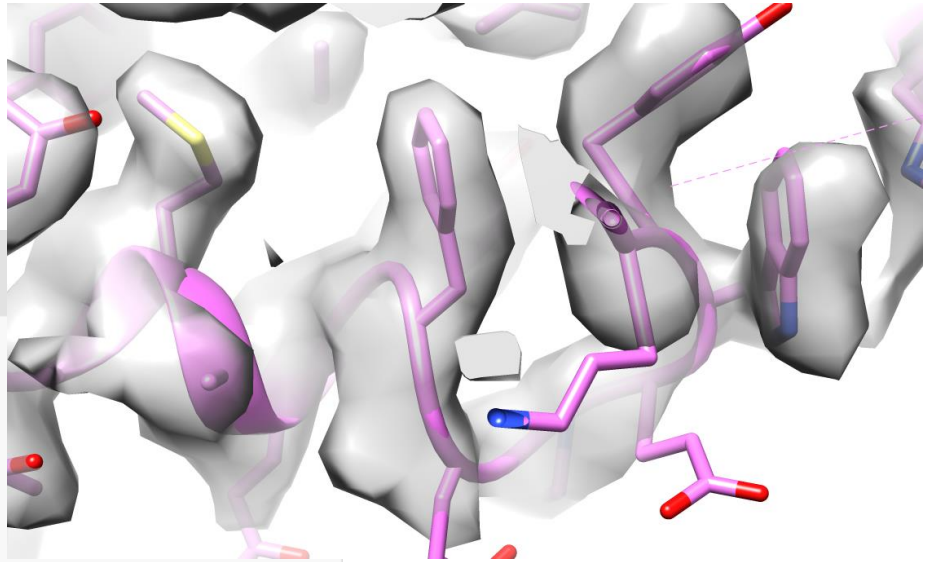
CheckmySequence Chain: A

```
*** Register shift:

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

model seq: dmysvMLTNDNTSRYWEPEfyeamytphtvlqgg
new seq  : dmysvmltndntsrYWEPEFYEAMYTPHtvlqgg
```

Deposited

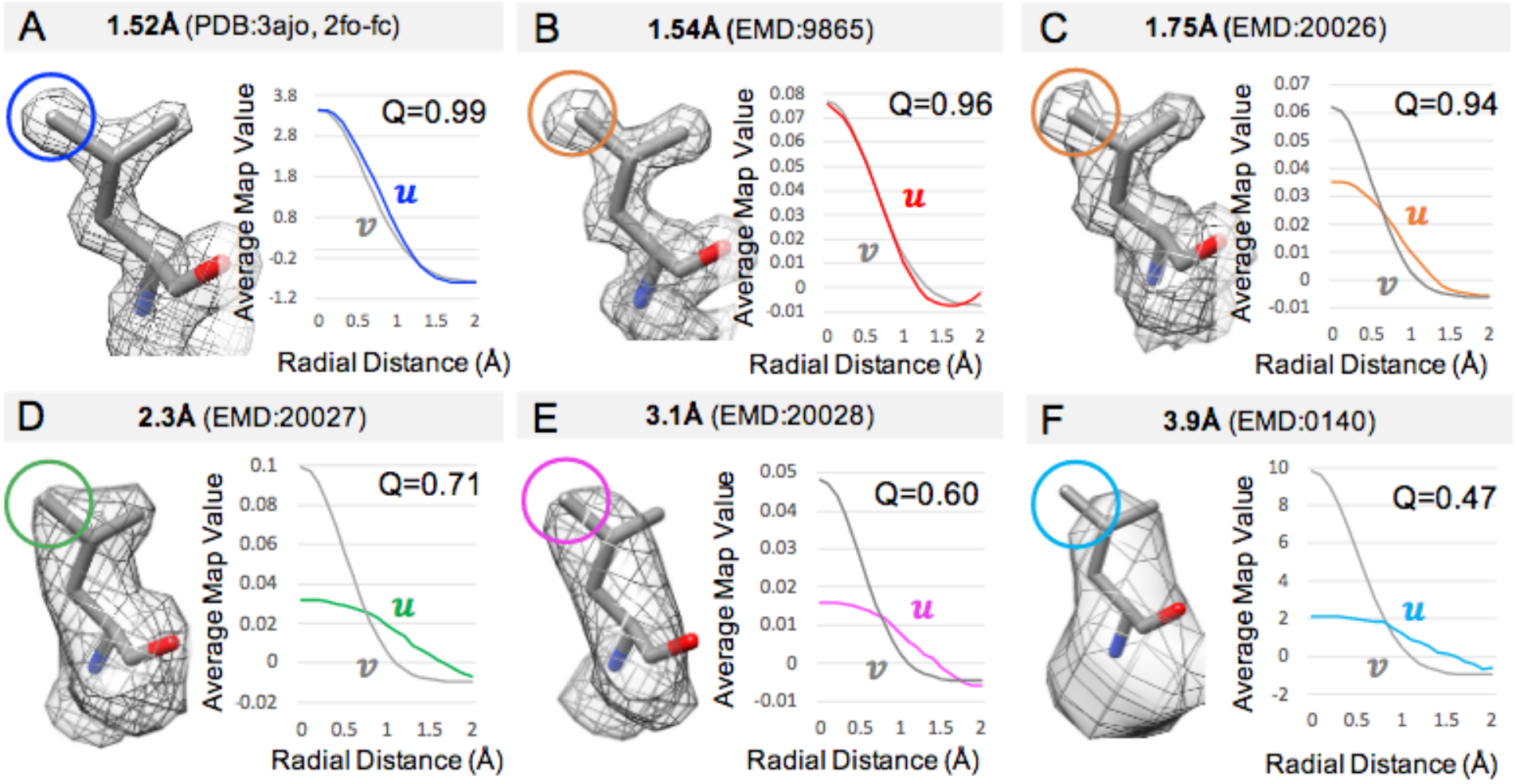


Remodelled



Local agreement with map: resolvability

Q score



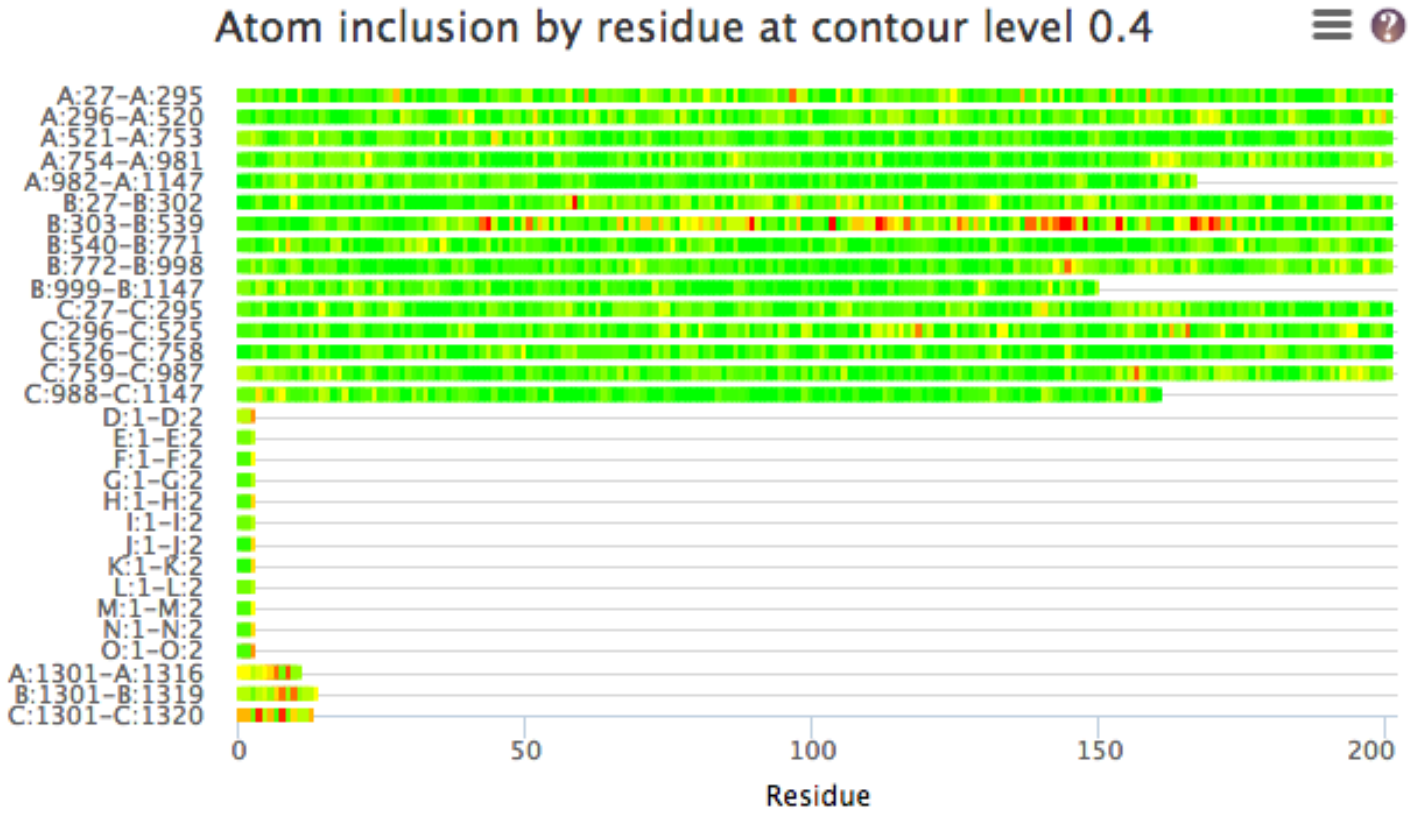
Science and Technology Facilities Council

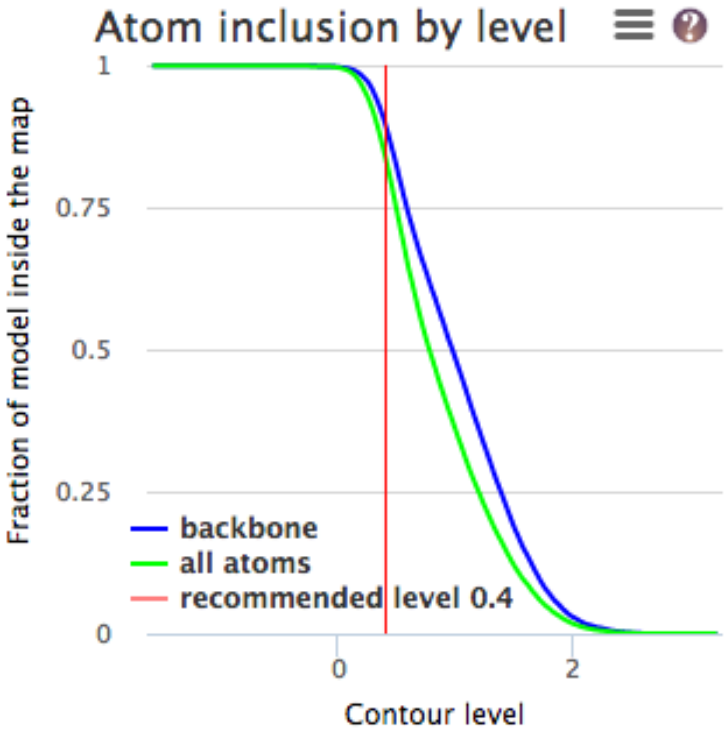
Scientific Computing

Pintile et al. 2020

Residue position/trace

Atom inclusion score





Science and Technology Facilities Council

Scientific Computing

https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-21457/analysis

Backbone trace?

Lawson et al. *Nature Methods*, 2021

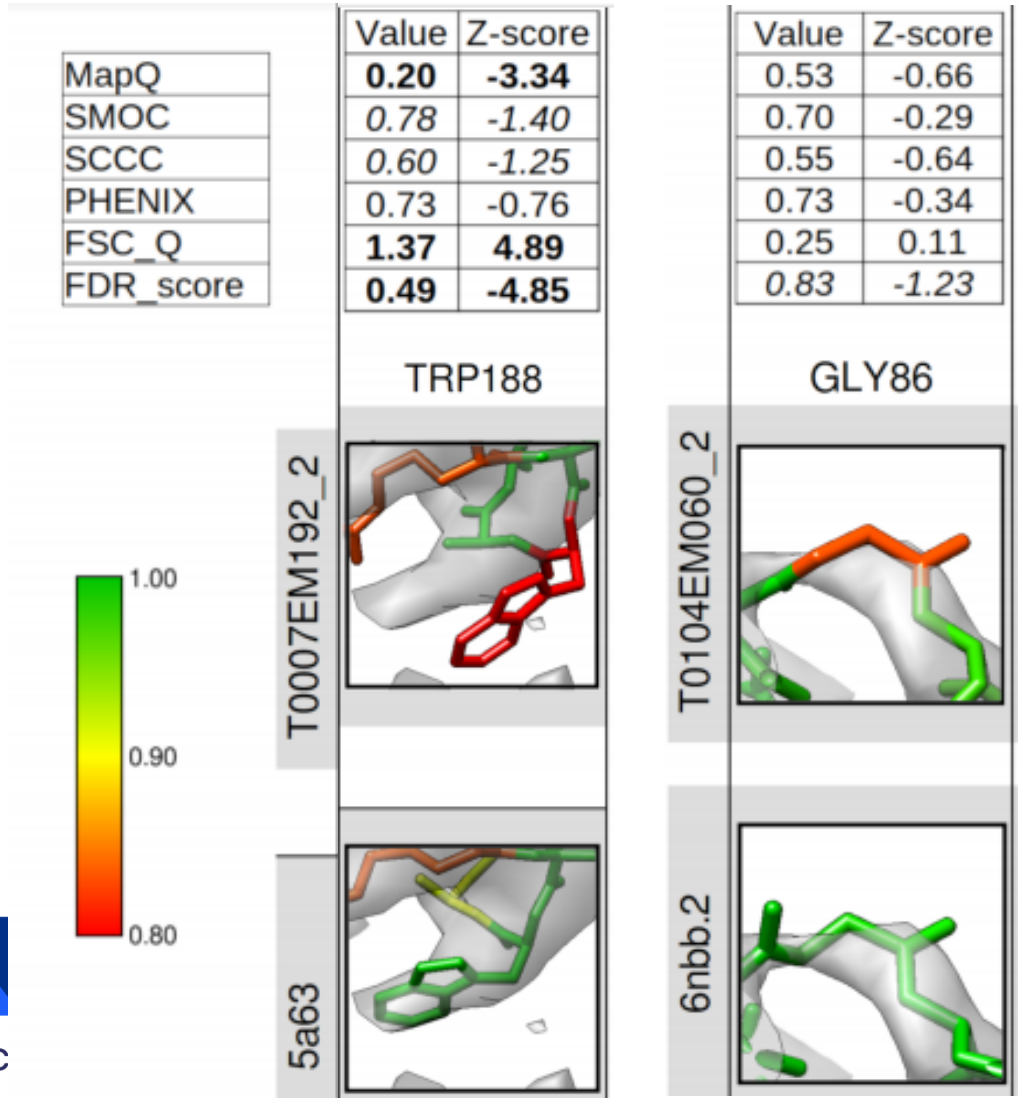
FDR validation score

	Value	Z-score
MapQ	**0.20**	**-3.34**
SMOC	*0.78*	*-1.40*
SCCC	*0.60*	*-1.25*
PHENIX	0.73	-0.76
FSC_Q	**1.37**	**4.89**
FDR_score	**0.49**	**-4.85**

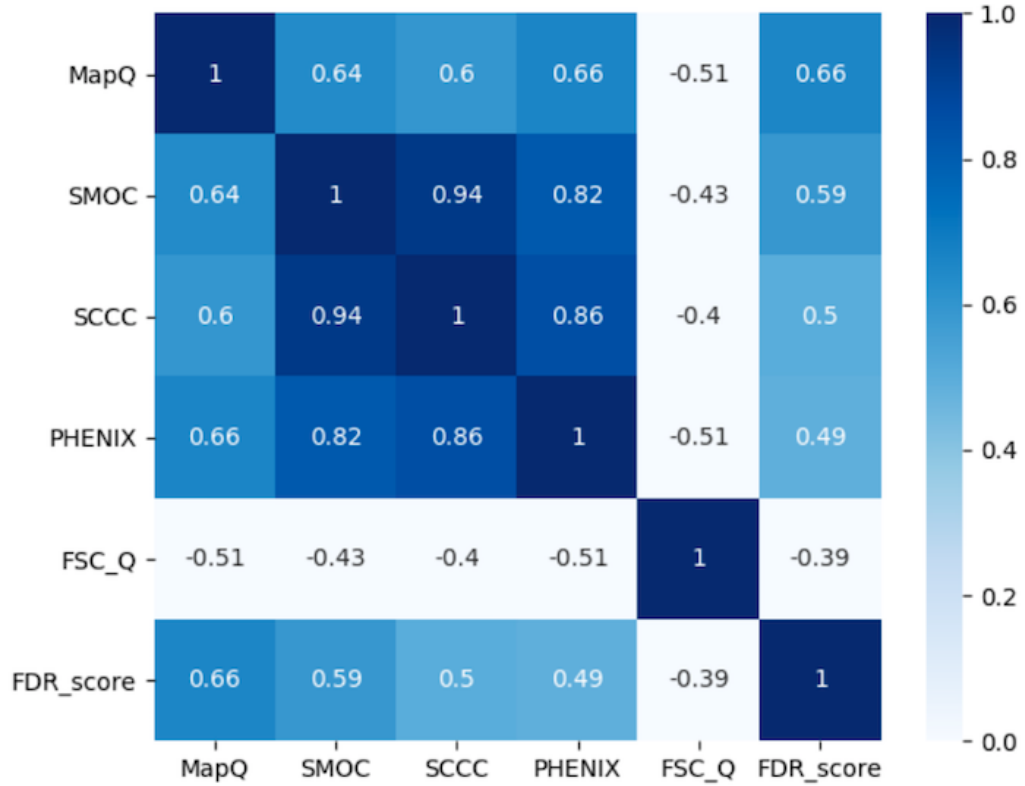
TRP188

	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



Maxmillian Beckers, Arjen Jakobi and Carsten Sachse. 2016

Olek Mateusz
Kathryn Cowtan
Peijun Zhang

Model geometry vs fit-to-data

Geometry

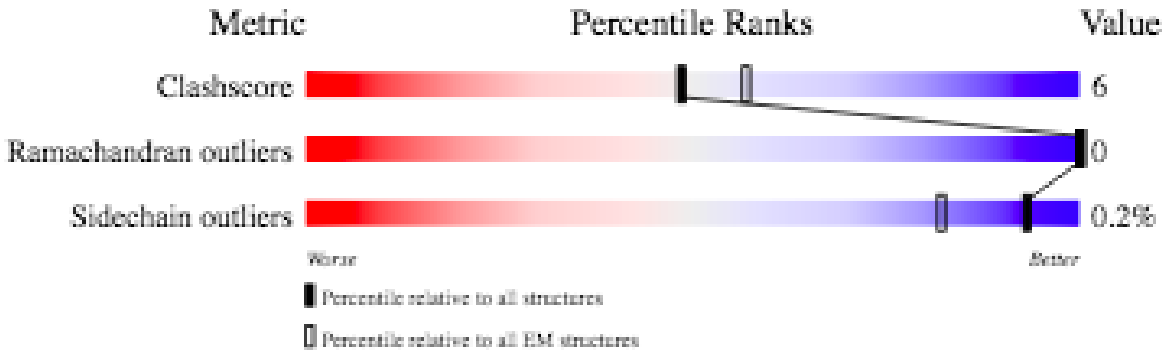
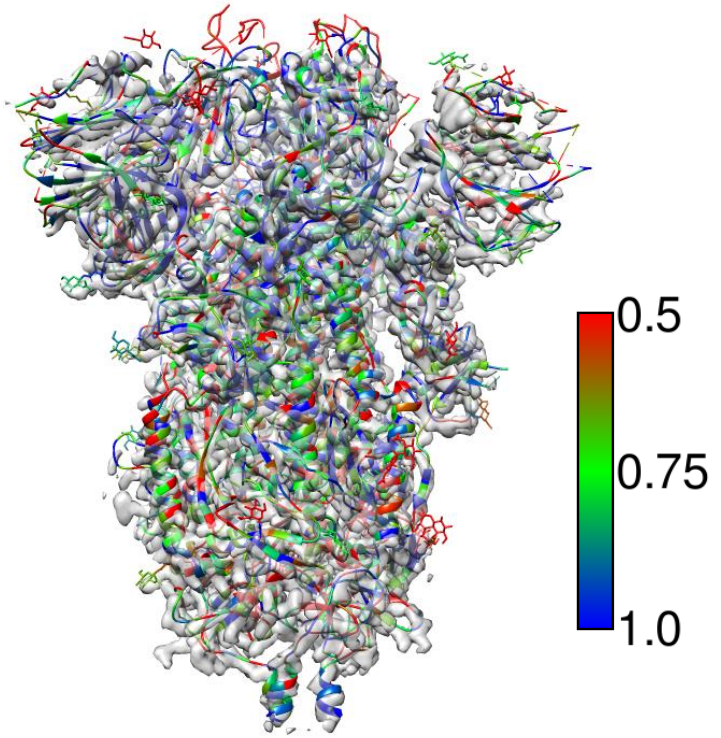


Image: PDBe/wwPDB

Map agreement ?



Overfitting ?

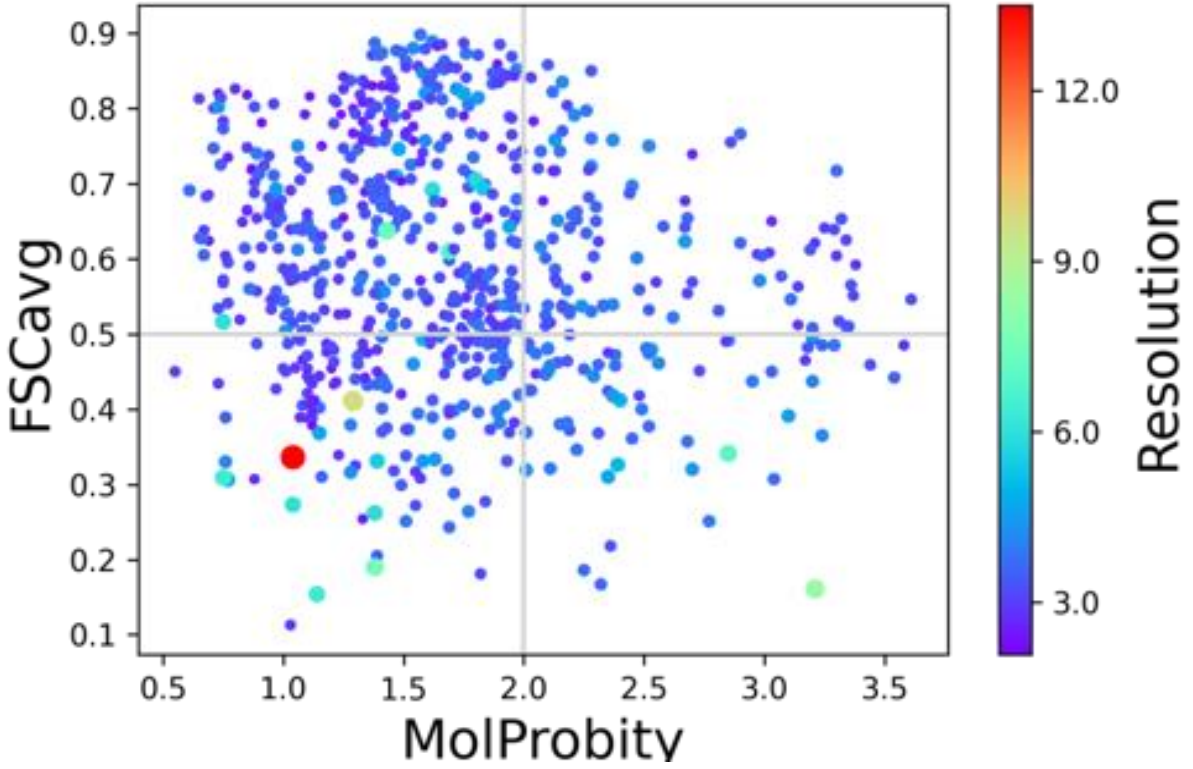
Science and Technology Facilities Council

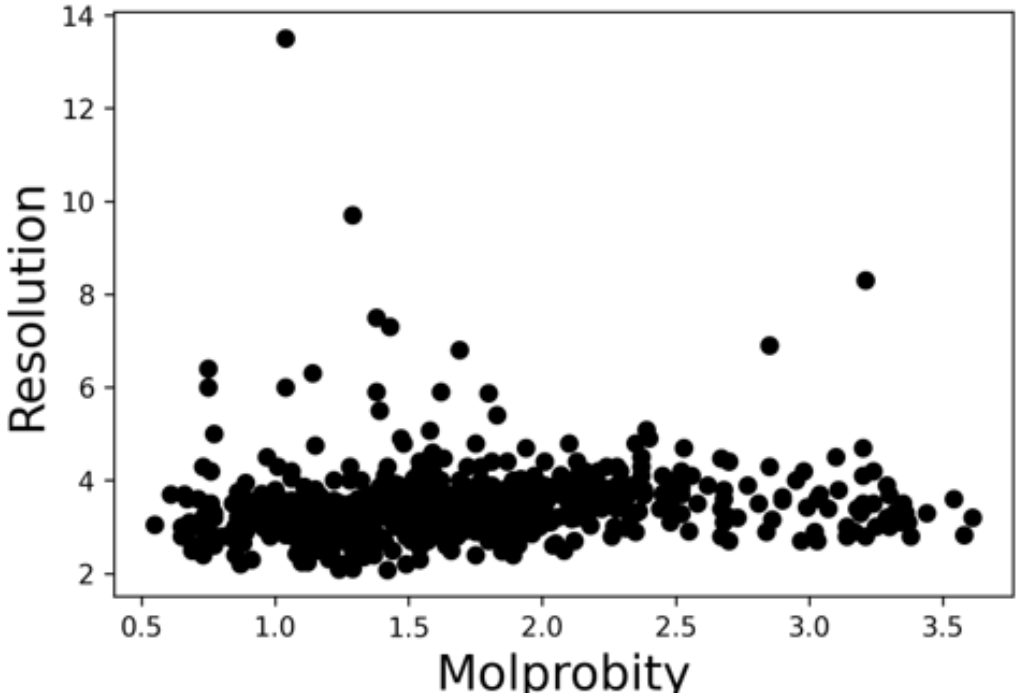
Scientific Computing

Model geometry vs fit-to-data

Do the models represent the map data well?

Do we need more validation metrics for publication?





Science and Technology Facilities Council

Scientific Computing

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

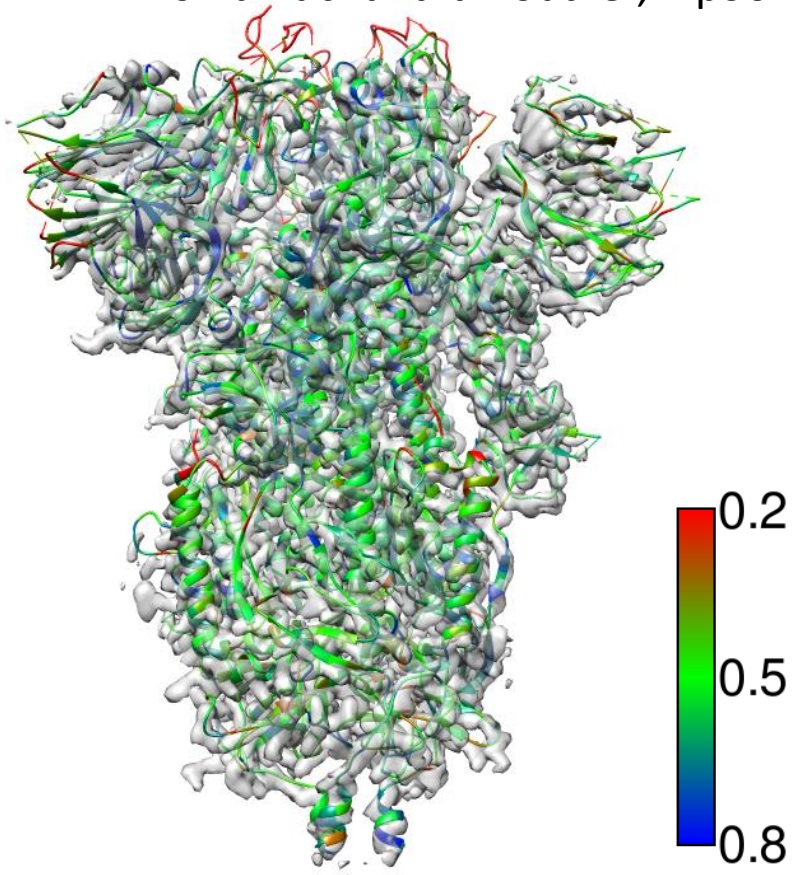
31.2% of the structures had FSCavg scores worse than 0.5.

Poor fit

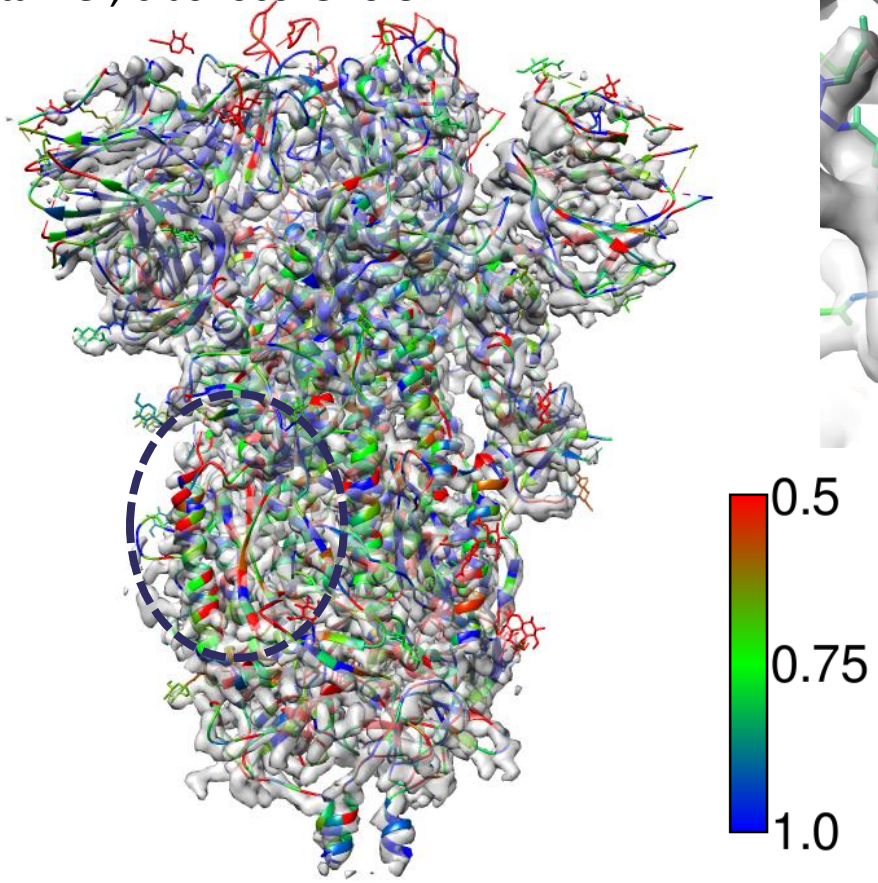
~28% of structures have potential issues with backbone trace (FDR-backbone score)

Molprobity score : 1.39

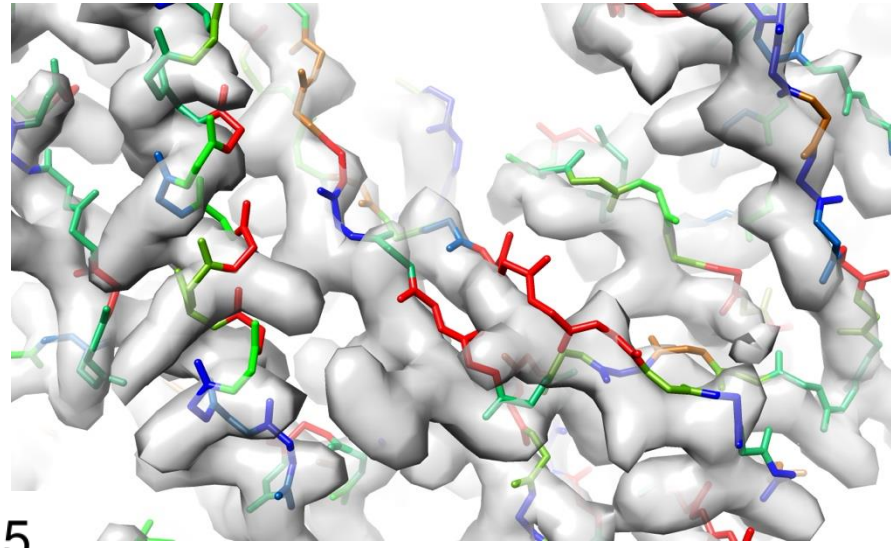
0 Ramachandran outlier, 1 poor rotamer, clashscore: 6.3



SMOC score
(Joseph et al. 2016)



FDR-backbone score
(Olek et al. 2021)

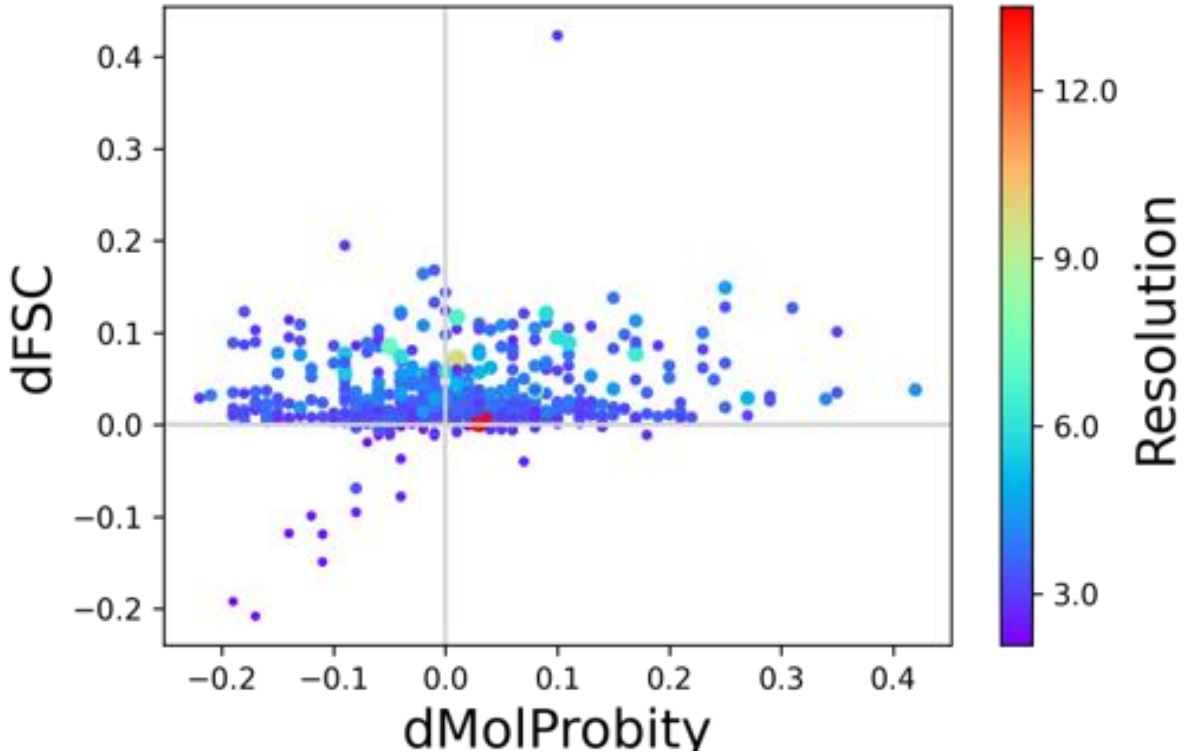


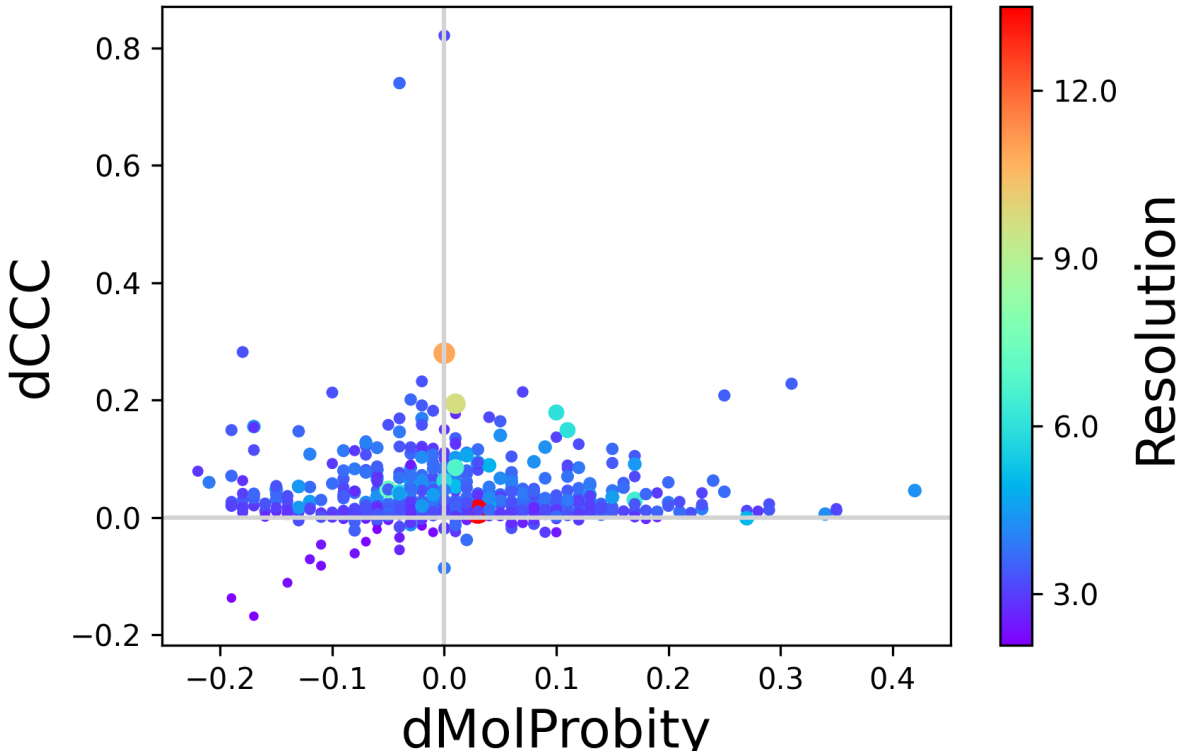
3.4Å

Science and Technology Facilities Council
Scientific Computing

Model geometry vs fit-to-data

Can the map fit / representation be improved ?





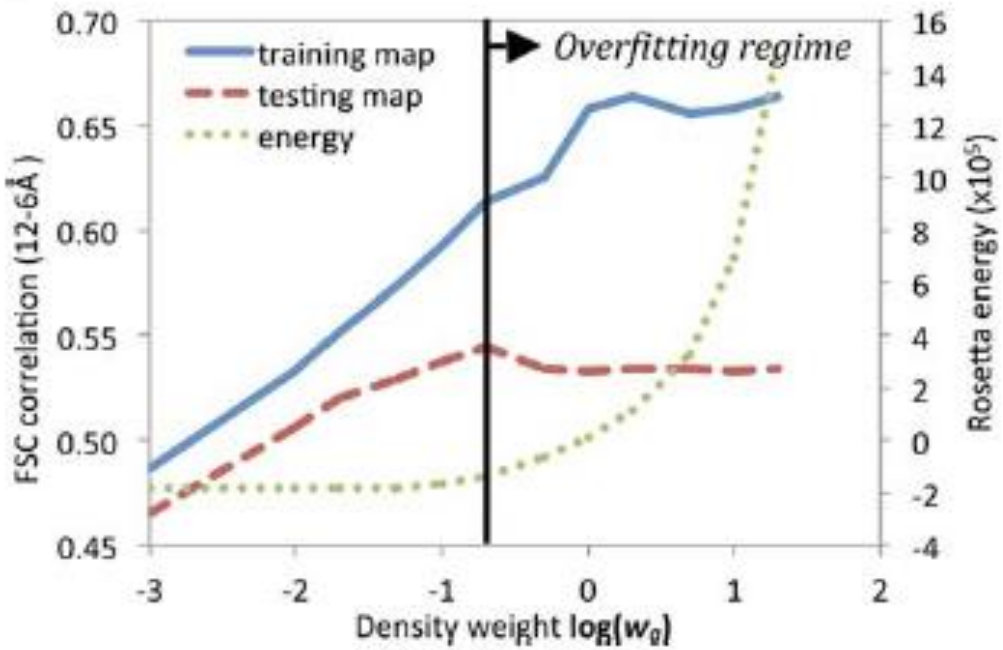
Science and Technology Facilities Council

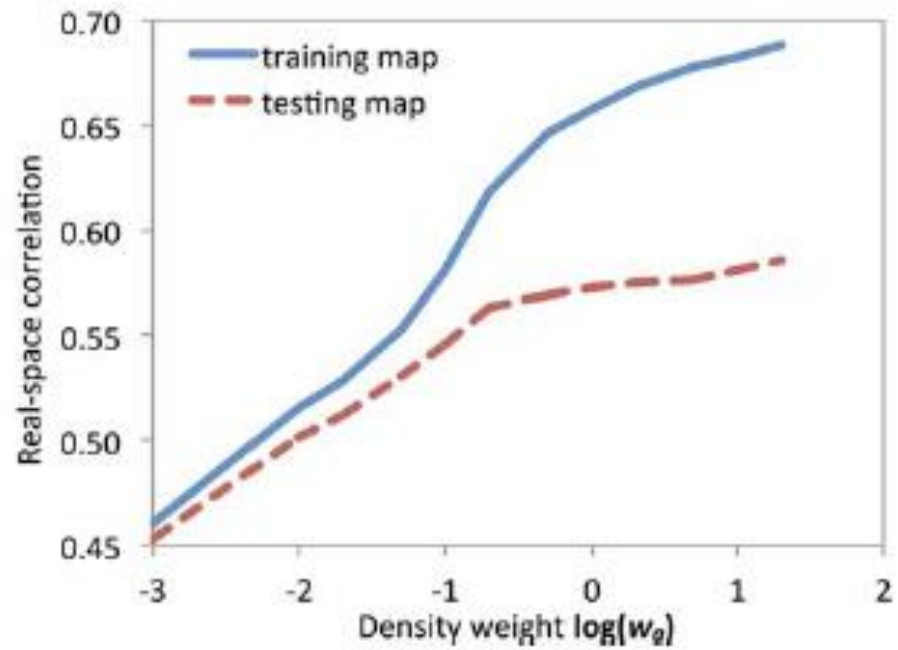
Scientific Computing

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

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

Test against equivalent but independent data

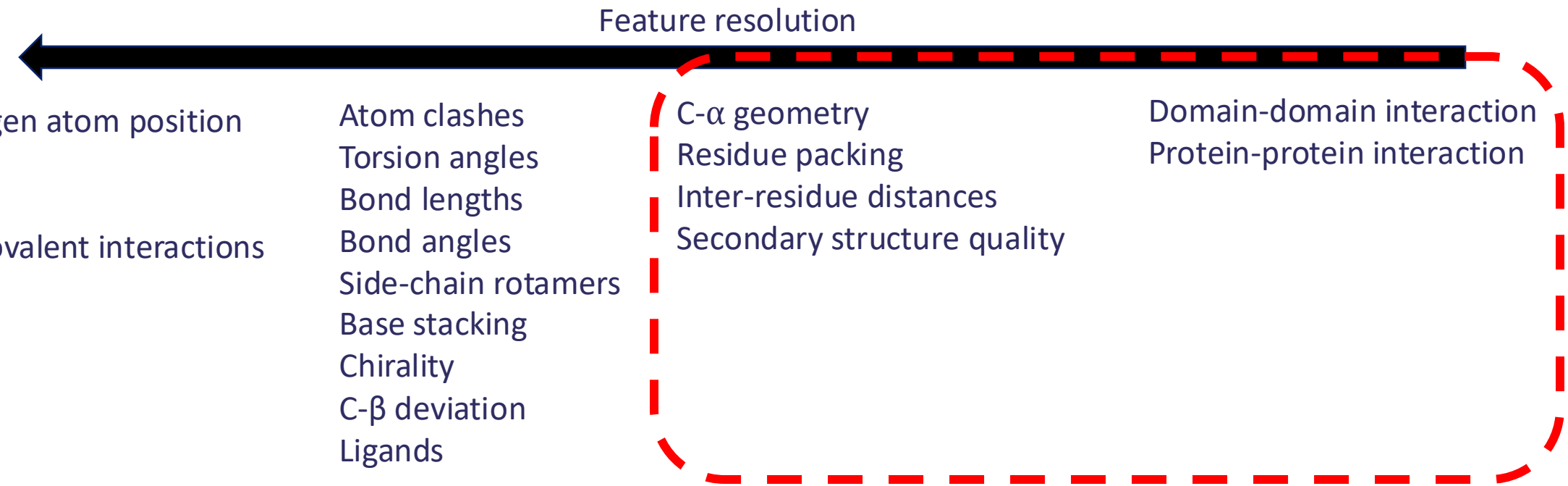




Mm-cpn in the ATP/AlFx induced closed state
4.3 A resolution

DiMaio et al 2013; DiMaio and Chiu 2016

Model quality



Many of the deposited models are optimized more towards model geometry than fit to data.

Evaluation of low-resolution features is not common!

Fit to data

We need scores normalised for resolutions or reference distributions for different resolution bins

Better validation approaches especially to assess overfitting

EMDB statistics (geometry)

Molprobity score

Resolution	Q1	Q2	Q3
1.0-2.5	1.2	1.39	1.67
2.5-3.5	1.27	1.53	1.8025
3.5-4.5	1.4025	1.695	1.96
4.5-6.5	1.32	1.77	2.1175
> 6.5	1.08	1.68	2.05

Clashscore

Resolution	Q1	Q2	Q3
1.0-2.5	3.1575	4.485	6.79
2.5-3.5	3.6175	5.765	8.17
3.5-4.5	4.23	7.165	11.075
4.5-6.5	3.33	7.82	13.803
> 6.5	2.26	6.57	12.65

Cis-proline (%)

Resolution	Q1	Q2	Q3
1.0-2.5	0	1.02	5.41
2.5-3.5	0	0	2.8
3.5-4.5	0	0	2.4675
4.5-6.5	0	0	2.94
> 6.5	0	1.22	2.46

Ramachandran Z-score

Resolution	Q1	Q2	Q3
1.0-2.5	-0.955	0.35	1.44
2.5-3.5	-1.32	-0.12	0.91
3.5-4.5	-1.705	-0.445	0.71
4.5-6.5	-2.12	-0.615	0.11
> 6.5	-1.73	-0.68	0.31

Ramachandran outlier (%)

Resolution	Q1	Q2	Q3
1.0-2.5	0	0	0.04
2.5-3.5	0	0	0
3.5-4.5	0	0	0.1
4.5-6.5	0	0.045	0.405
> 6.5	0	0.09	0.86

EMDB statistics (fit to map)

FSCavg (upto FSC0.5)

Resolution	Q1	Q2	Q3
1.0-2.5	0.746	0.773	0.8
2.5-3.5	0.738	0.7725	0.797
3.5-4.5	0.7	0.747	0.778
4.5-6.5	0.706	0.738	0.767
> 6.5	0.7055	0.754	0.8015

MI (mask)

Resolution	Q1	Q2	Q3
1.0-2.5	0.004	0.012	0.029
2.5-3.5	0.015	0.025	0.041
3.5-4.5	0.019	0.032	0.055
4.5-6.5	0.0245	0.05	0.07525
> 6.5	0.02675	0.0635	0.10825

CCC (mask)

Resolution	Q1	Q2	Q3
1.0-2.5	0.57225	0.615	0.6585
2.5-3.5	0.565	0.616	0.658
3.5-4.5	0.55925	0.615	0.657
4.5-6.5	0.55	0.626	0.693
> 6.5	0.5365	0.635	0.7105

SMOC Z-score < -2 (%)

Resolution	Q1	Q2	Q3
1.0-2.5	2.49675	2.845	3.27925
2.5-3.5	2.353	2.8	3.216
3.5-4.5	2.64725	3.1345	3.55925
4.5-6.5	3.116	3.552	3.995
> 6.5	2.6775	3.2875	3.60225

SMOC average

Resolution	Q1	Q2	Q3
1.0-2.5	0.6	0.6555	0.69325
2.5-3.5	0.6855	0.737	0.779
3.5-4.5	0.775	0.817	0.847
4.5-6.5	0.85975	0.8905	0.906
> 6.5	0.89	0.914	0.935

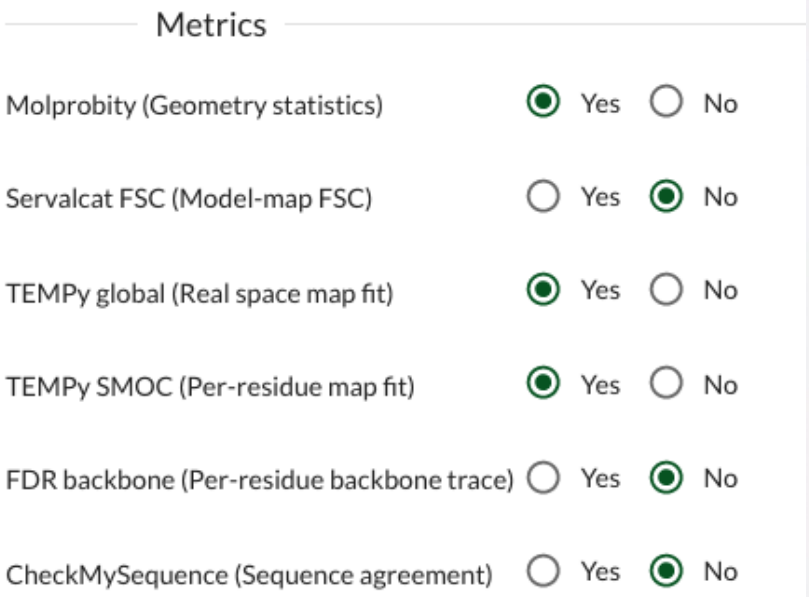
Model validation

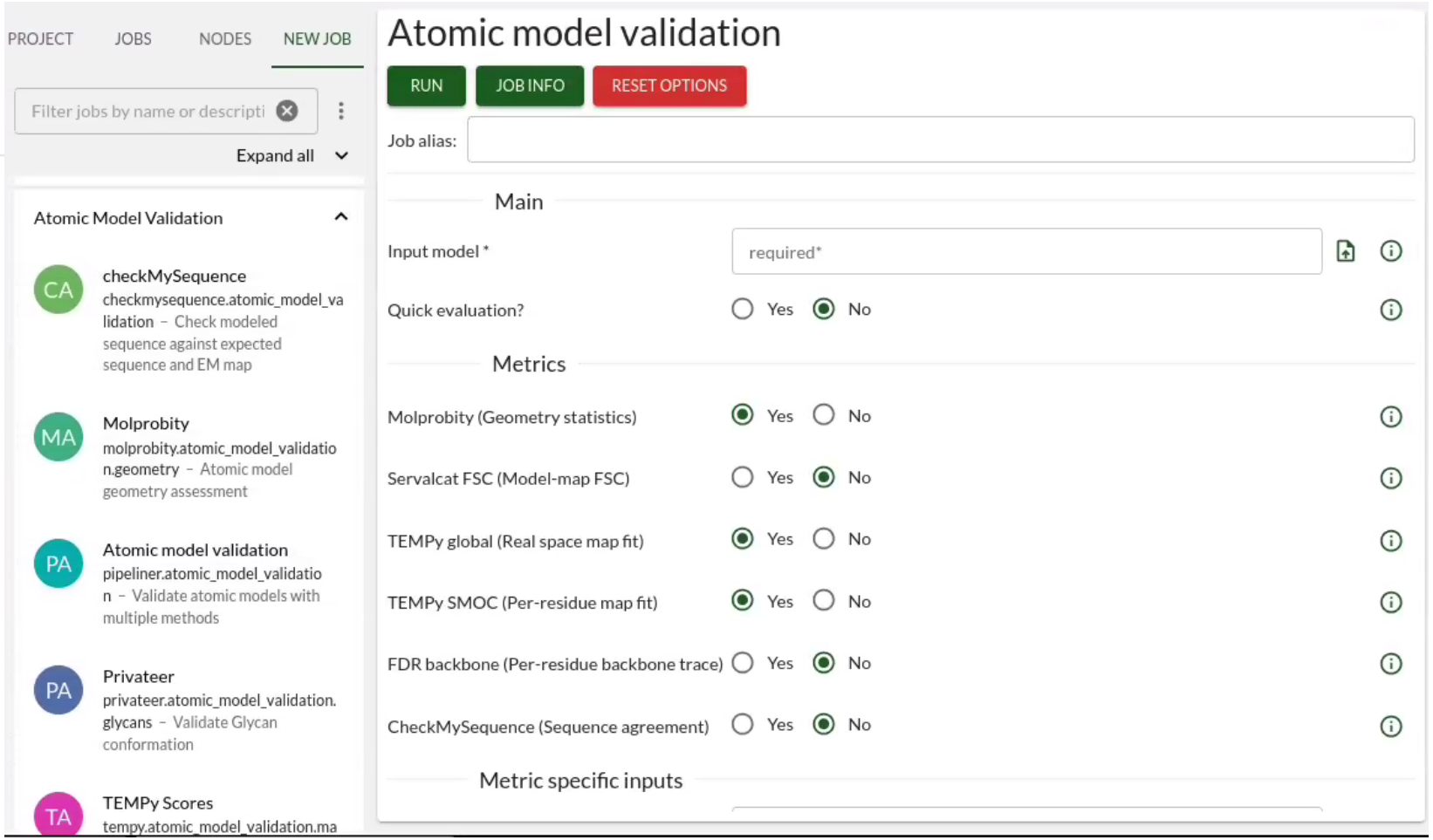
Multiple complementary validation metrics are often useful than a single score

Faster validation scores

Better link between model building, refinement and validation

Doppio model validation





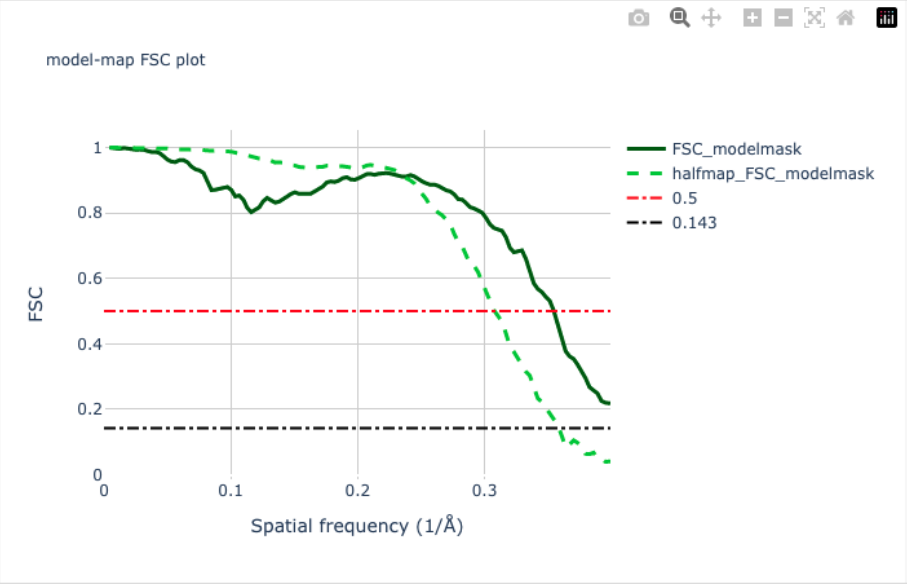
Willams et a. 2018, Chen et al. 2015, Yamashita et al. 2021, Joseph et al. 2016; Grzegorz C Acta D 2022

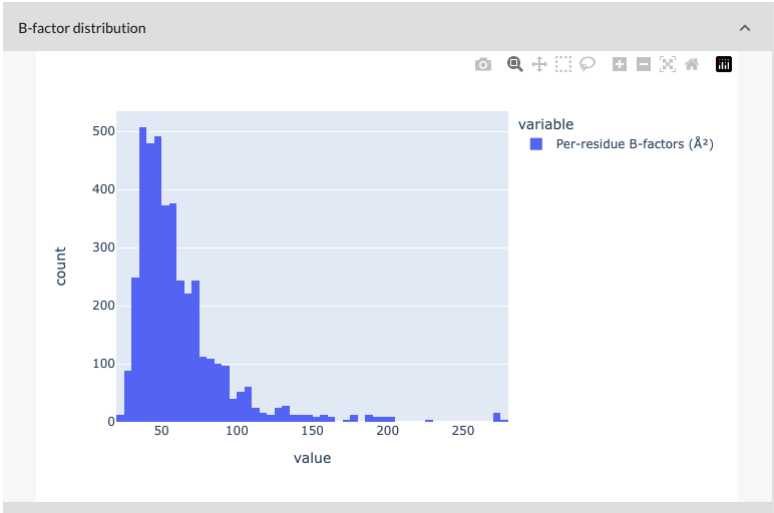
Doppio model validation

Global map-fit scores

Metric	Score
FSCavg_modelmask_FSC0.5	0.784
FSCavg_modelmask	0.639
FSCavg_h1_modelmask	0.568
FSCavg_h2_modelmask	0.568
CCC overlap mask	0.657
Model overlap fraction	0.733
CCC contoured map	0.573
MI overlap mask	0.049

Model-map FSC





Global geometry scores

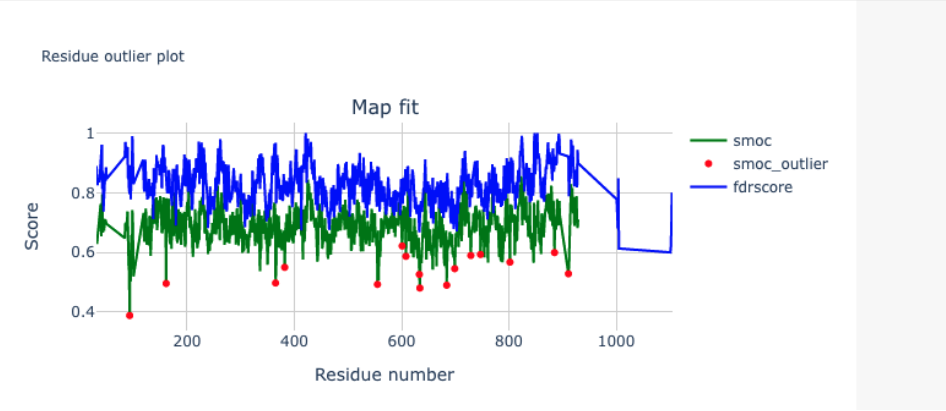
Metric	Score	Expected/Percentile
Ramachandran_outliers	0.10%	< 0.05%
Ramachandran_favored	97.26%	> 98%
RamachandranZ_whole	-0.72	bad |Rama-Z| > 3, suspicious 3 > |Rama-Z|...
RamachandranZ_helix	-1.49	bad |Rama-Z| > 3, suspicious 3 > |Rama-Z|...
RamachandranZ_sheet	0.12	bad |Rama-Z| > 3, suspicious 3 > |Rama-Z|...
RamachandranZ_loop	-0.48	bad |Rama-Z| > 3, suspicious 3 > |Rama-Z|...
Rotamer_outliers	0.34%	< 0.3%
CBeta_deviations	0	0
Clashscore	3.07	76.7 (percentile)
Molprobity_score	1.24	83.0 (percentile)
Cis_proline	8.06	0%
Cis_general	0.31	0%
Rms_bonds	0.0085	< 0.02
Rms_angles	1.42	< 2.0

Doppio model validation

Outlier clusters

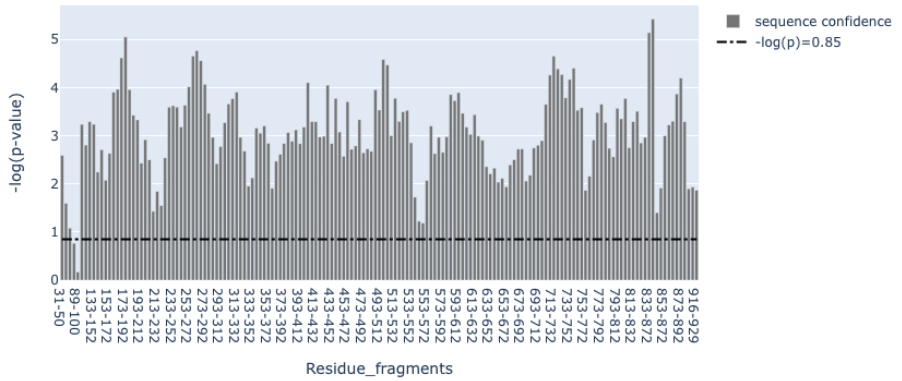
Cluster	Chain_Residue	Outlier Type(s)
1	A_462	ramachandran_outlier
1	A_423	omega_outlier | bad_clash (atom:CG - D_280_H...
1	A_424	bad_clash (atom:N - D_281_OD2) | bad_clash (a...
1	A_463	bad_clash (atom:HB3 - A_428_HG21)
1	D_280	bad_clash (atom:HD11 - A_423_CG) | bad_clas...
1	D_281	bad_clash (atom:OD2 - A_424_N) | bad_clash (a...
1	D_283	bad_clash (atom:NH1 - A_104_HG11) | bad_cla...
2	A_280	bad_clash (atom:HD11 - D_423_CG) | bad_clas...
2	A_281	bad_clash (atom:OD2 - D_424_N) | bad_clash (a...
2	A_283	bad_clash (atom:NH1 - D_104_HG11)
2	D_462	ramachandran_outlier
2	D_423	omega_outlier | bad_clash (atom:CG - A_280_H...
2	D_424	bad_clash (atom:N - A_281_OD2) | bad_clash (a...
2	D_463	bad_clash (atom:HB3 - D_428_HG21)
3	B_462	ramachandran_outlier
3	B_423	omega_outlier | bad_clash (atom:CG - C_280_H...
3	B_424	bad_clash (atom:N - C_281_OD2) | bad_clash (a...
3	B_463	bad_clash (atom:HB3 - B_428_HG21)
3	C_280	bad_clash (atom:HD11 - B_423_CG) | bad_clas...
3	C_281	bad_clash (atom:OD2 - B_424_N) | bad_clash (a...
3	C_283	bad_clash (atom:NH1 - B_104_HG11)

Outliers: chain A



CheckMySequence sequence p-values, protein chain A

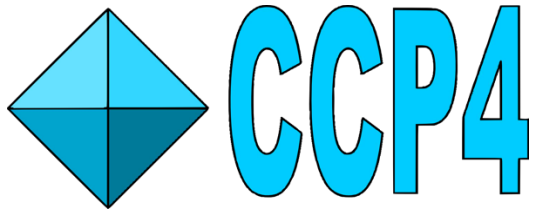
The bars are proportional to the -log(p-value) and the higher they are, the more reliable the sequence of a fragment. Grey bars: Expected fragment sequence and model sequence match.



Science and Technology Facilities Council

Scientific Computing

Acknowledgements



Scientific Computing

Scientific Computing

Questions?

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC