Unknown sequences and hidden errors in macromolecular structure models

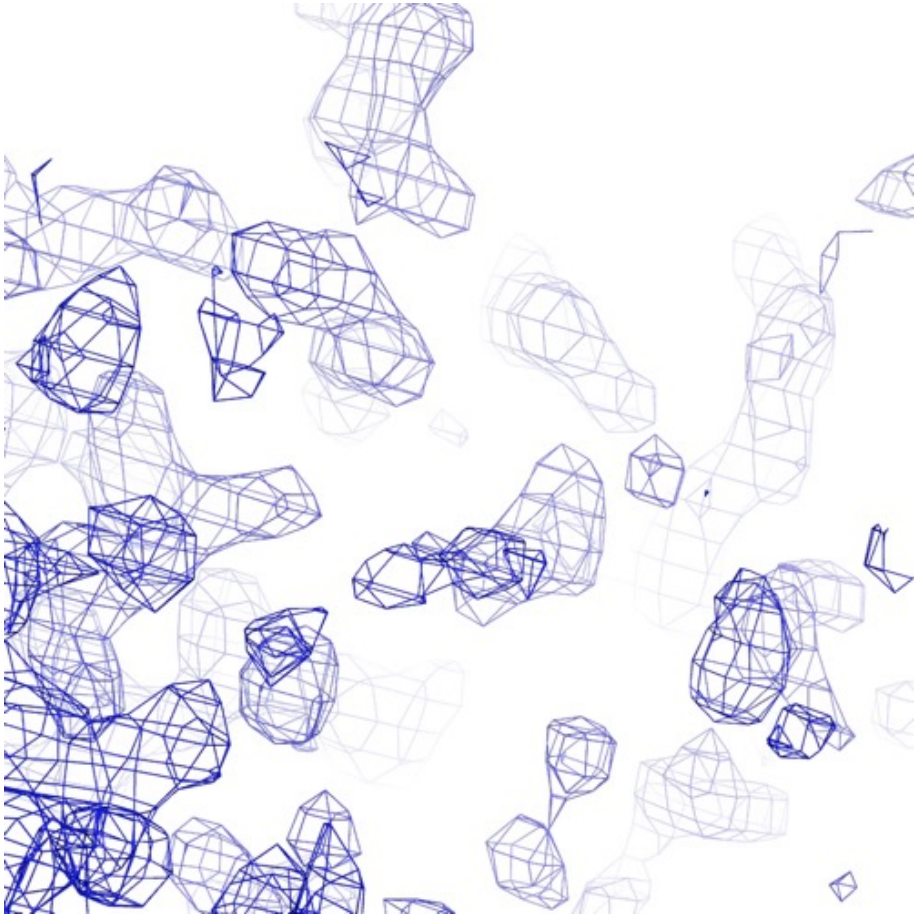
(New developments in model building and validation)

Grzegorz Chojnowski
EMBL Hamburg

EMBL

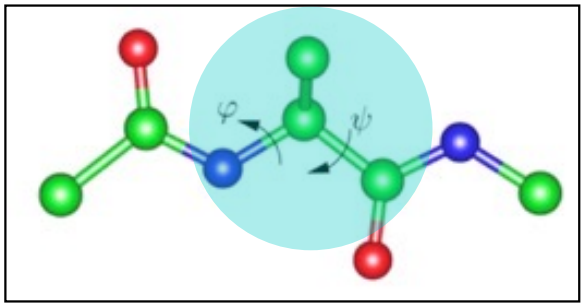
Diamond-CCP4 workshop 03.12.2022

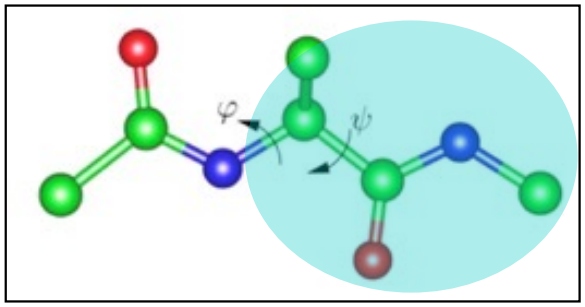
Electron density map interpretation

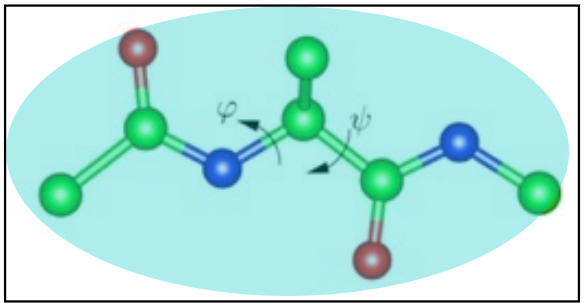


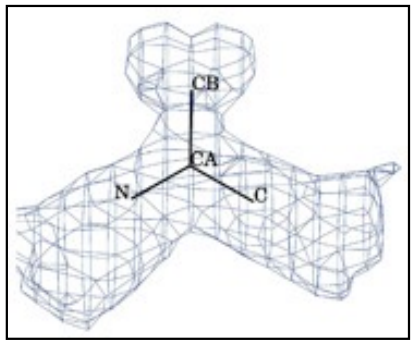
EMBL

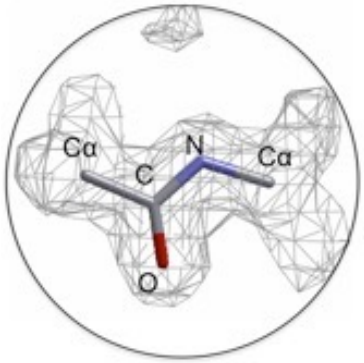
Crystal structure model building

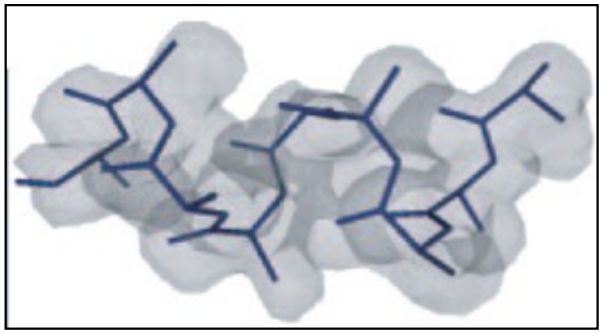












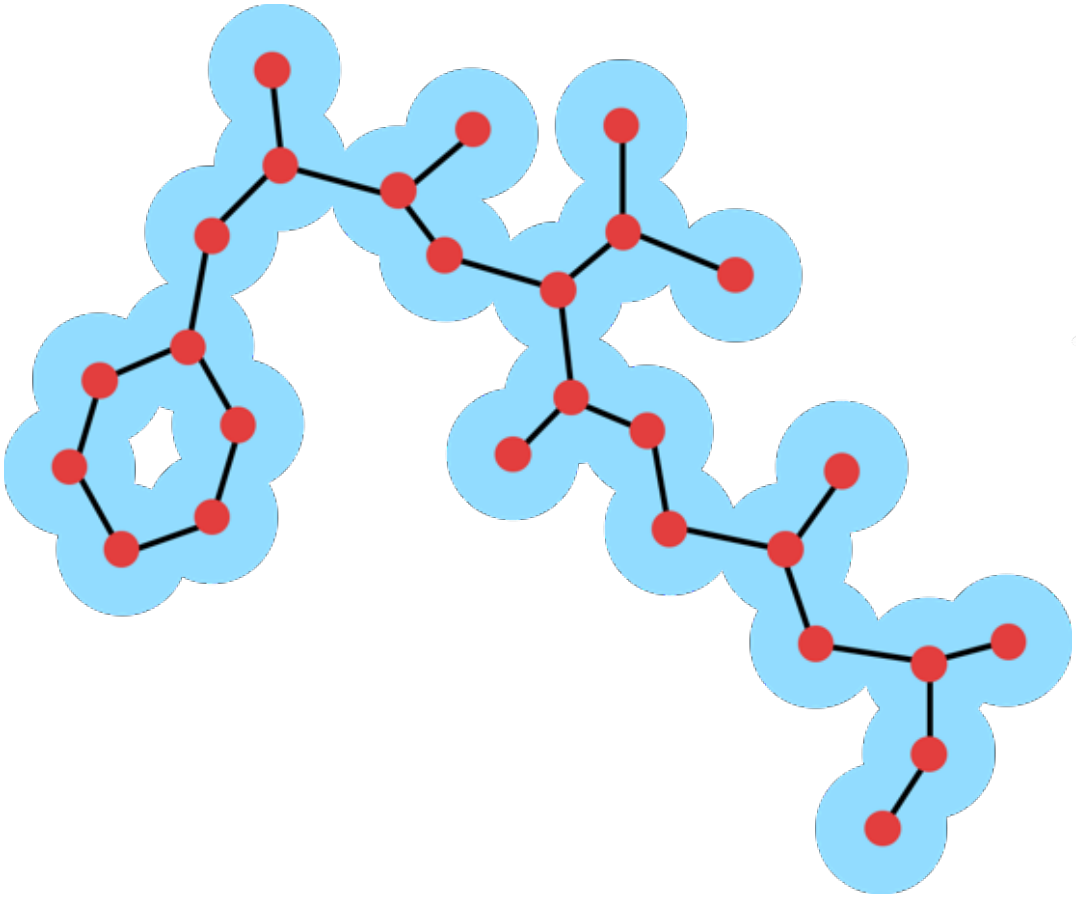
Buccaneer
ModelCraft

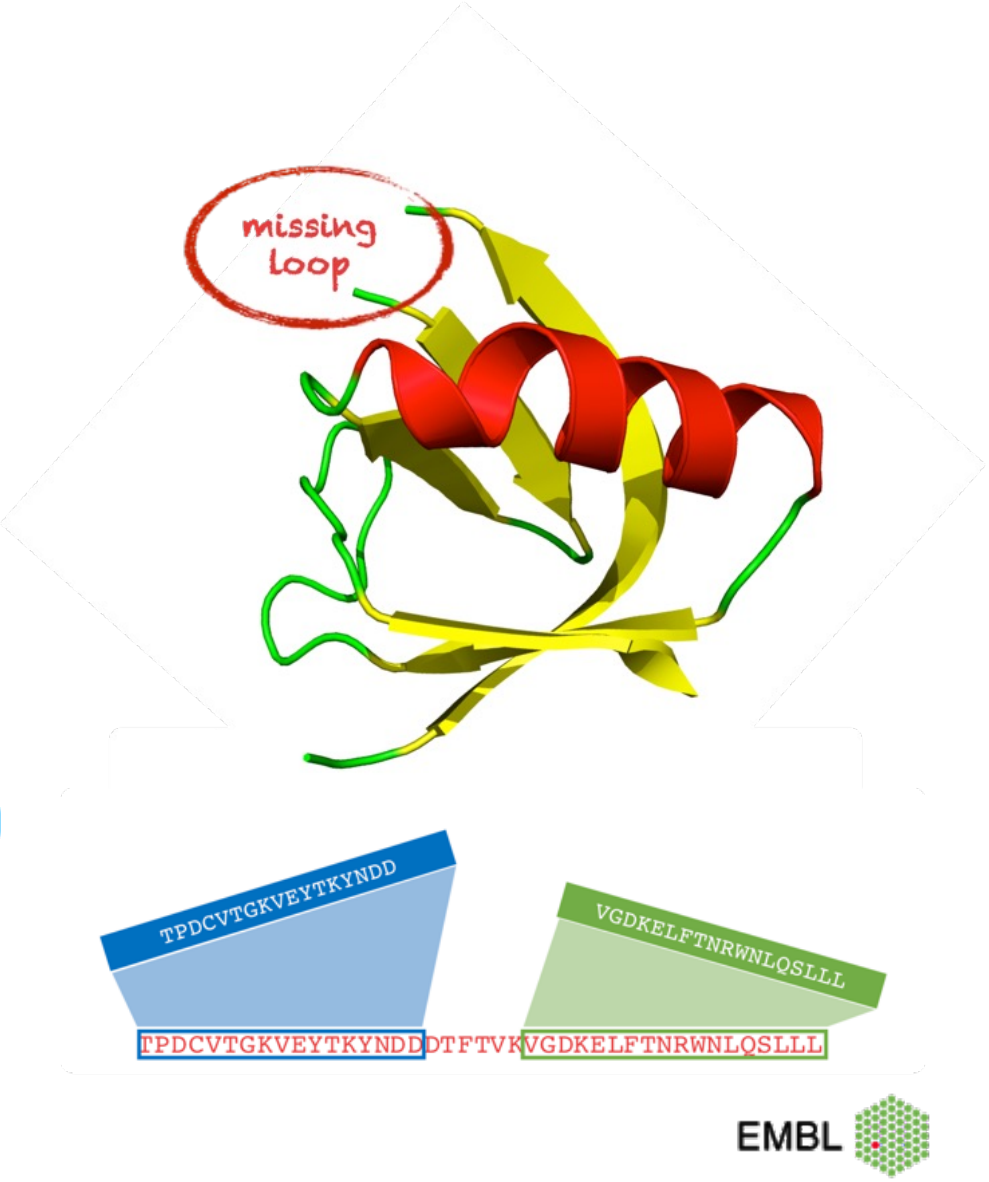
ARP/wARP

Phenix.autobuild

EMBL

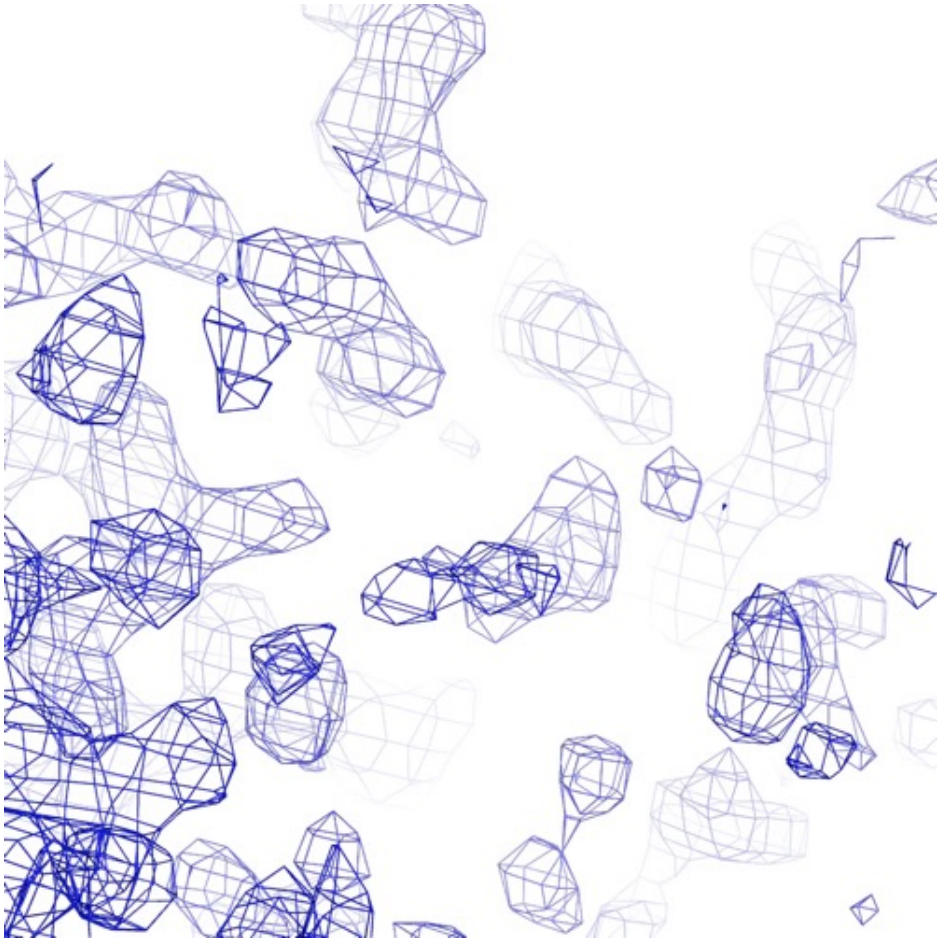
Crystal structure model building

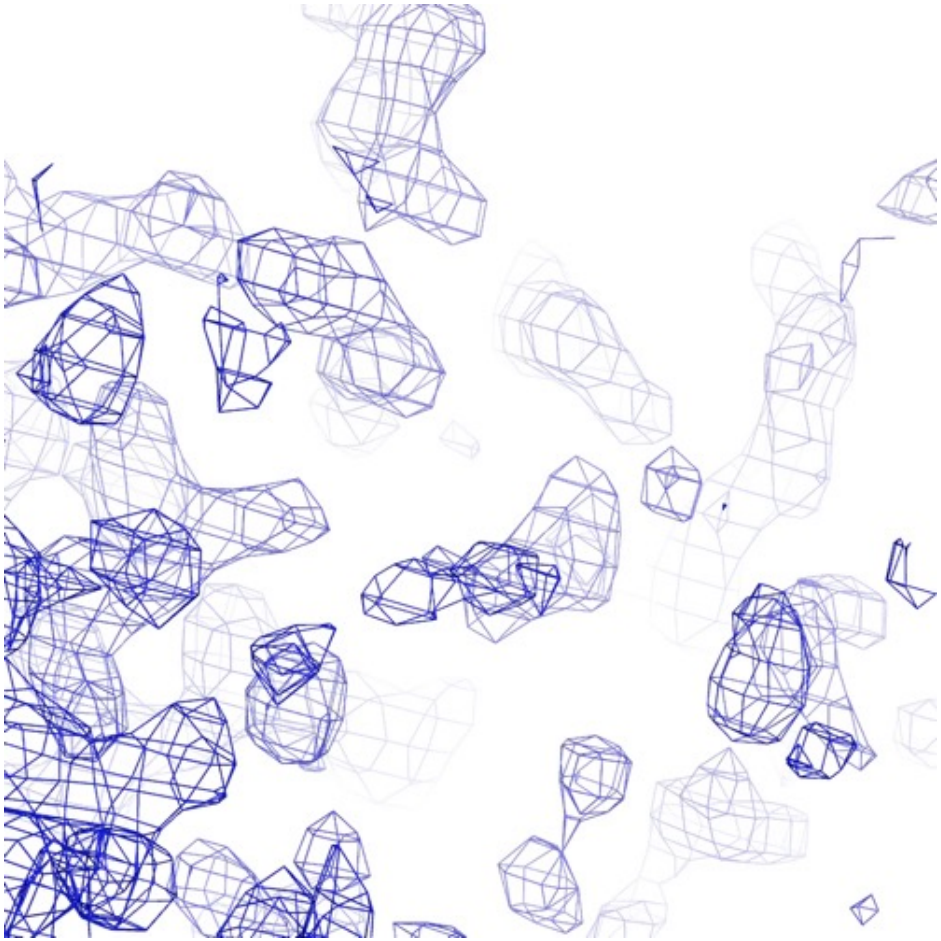




EMBL

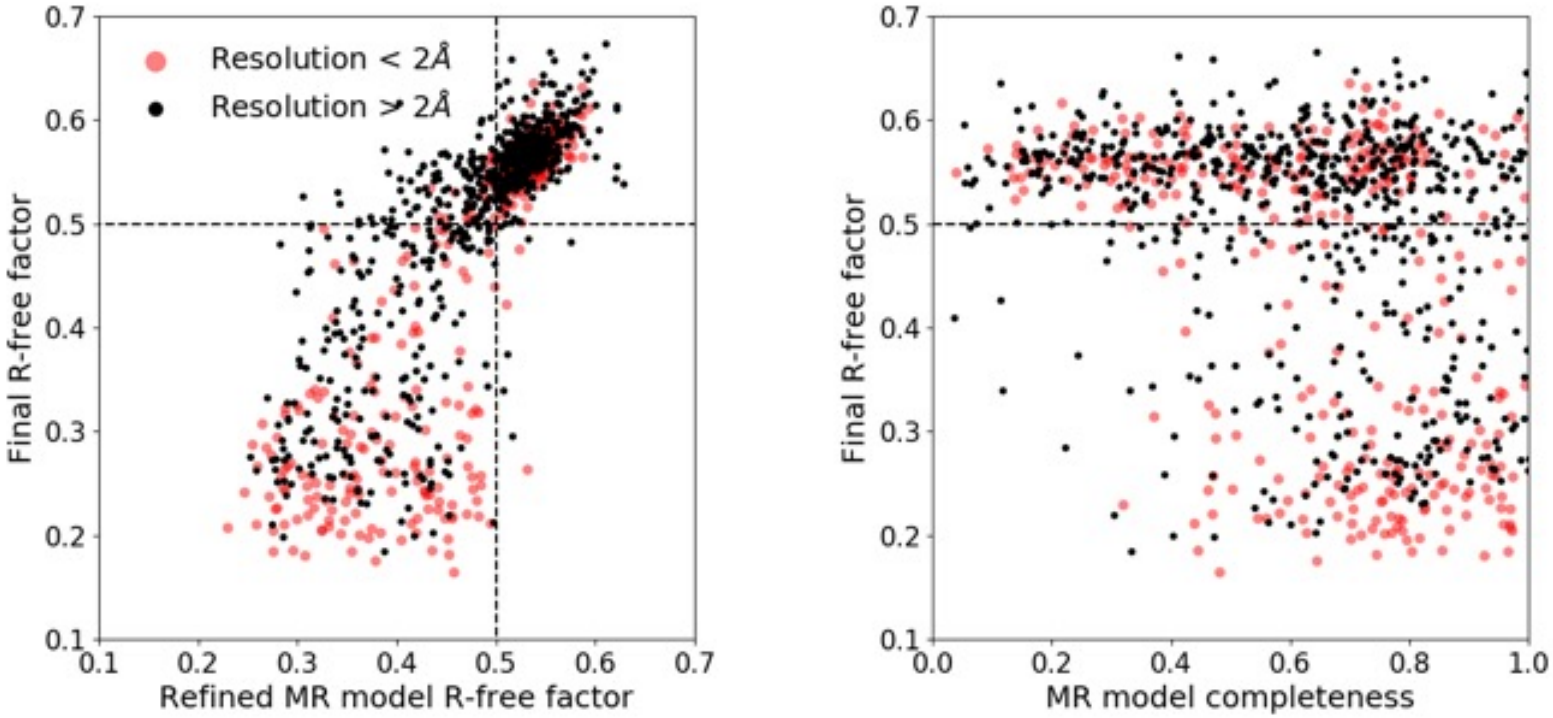
Crystal structure model building





EMBL

Crystal structure model building

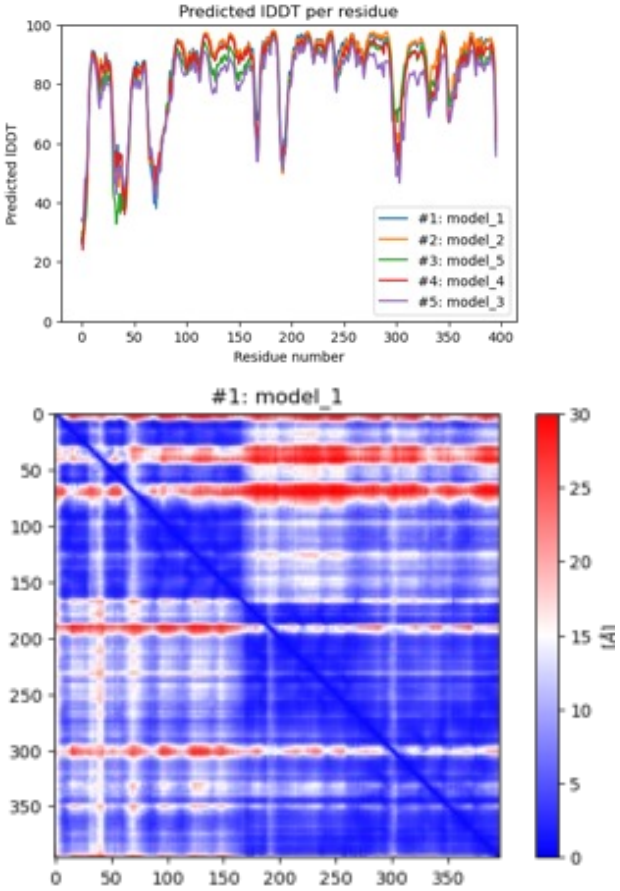


1493 MR solutions submitted to the ARP/wARP web service from automated pipelines (BALBES, MrBUMP, MORDA)

Chojnowski et al. ActaD (2020)

EMBL

Solving crystal structures with AlphaFold2 models



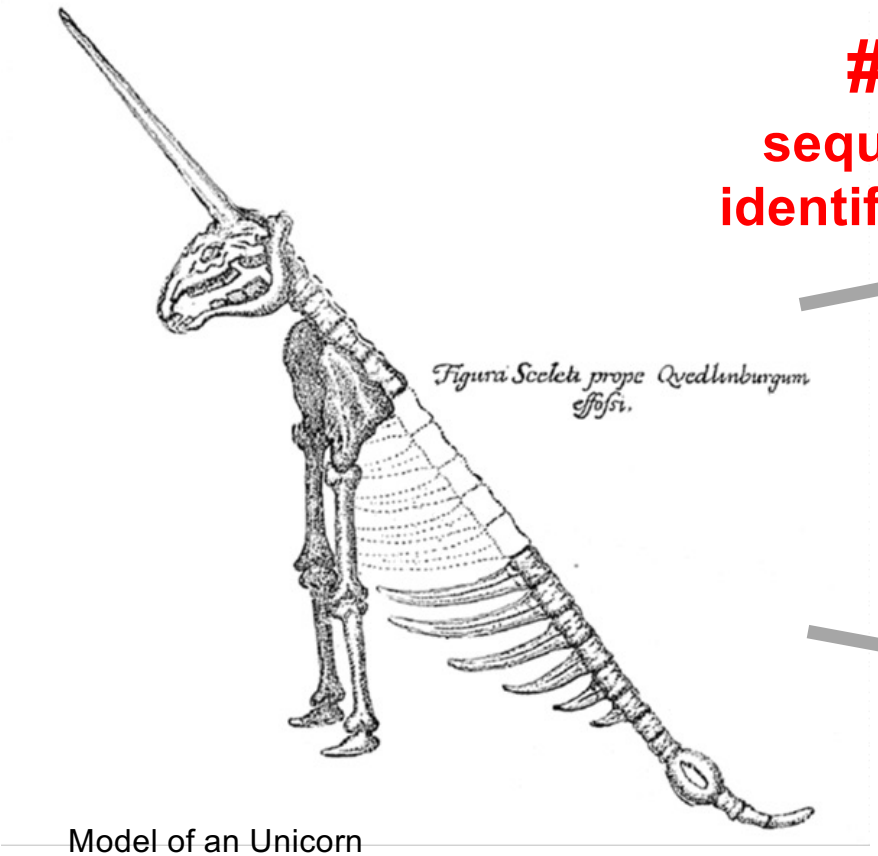
pTM=0.68 <pLDDT>=83

AlpaFold2 prediction

crystal structure model

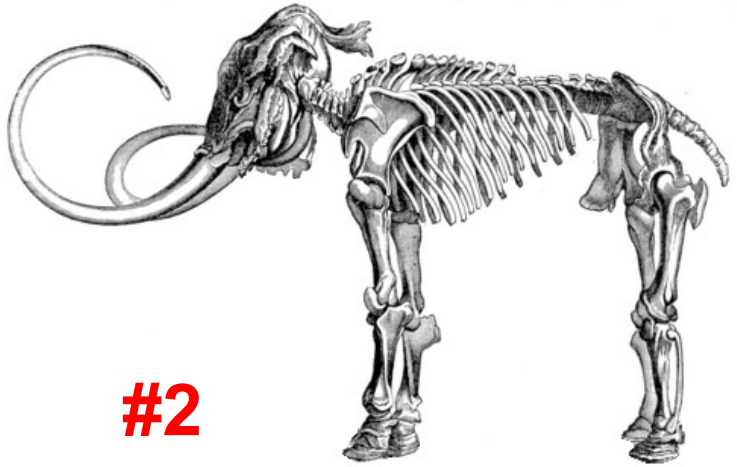
EMBL

Model building traps in MX (and cryo-EM)

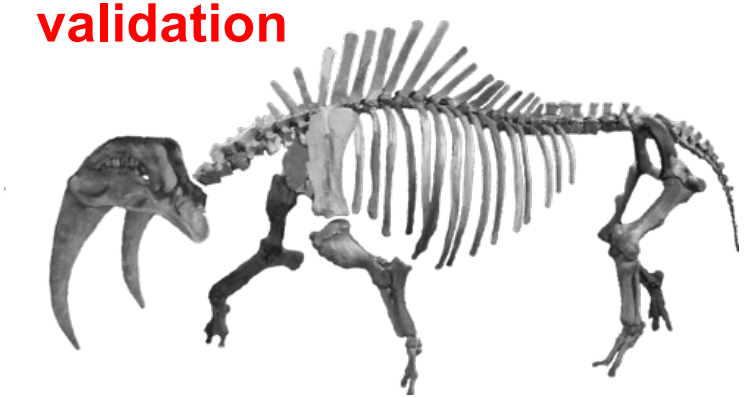


Model of an Unicorn
Gottfried Leibniz after Otto von Guericke, *Protogaea* (1719)

#1
sequence identification

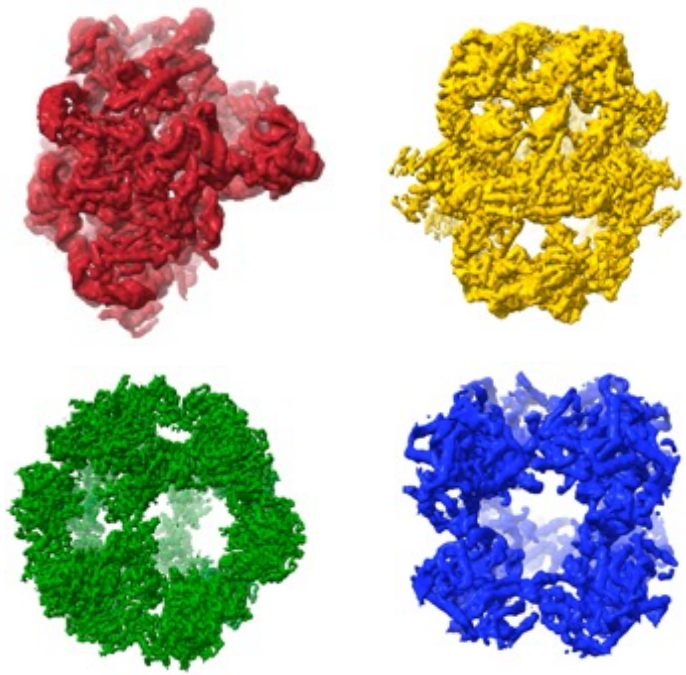


#2
sequence validation

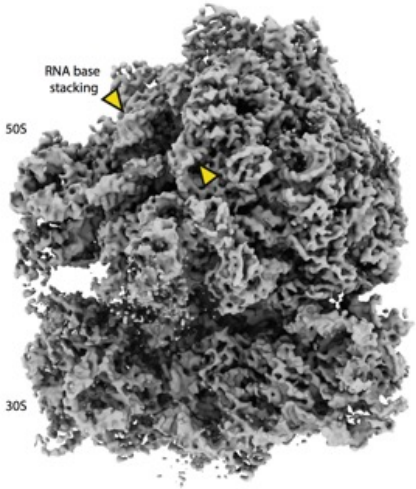


EMBL

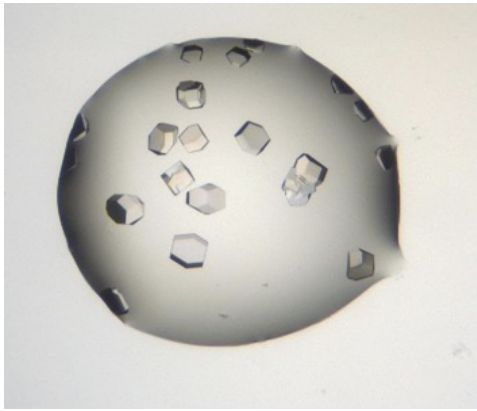
Unknown proteins in EM and MX



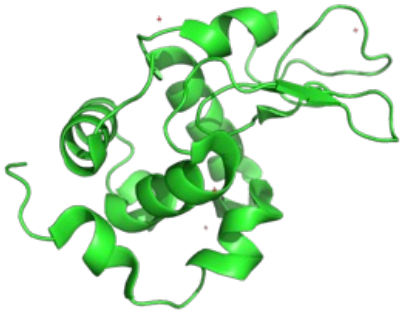
Cryo-EM and artificial intelligence visualize endogenous protein community members
Skalidis et al. Structure 2022



M. pneumoniae 70S ribosome at 3.5 Å refined from in situ tilt-series data
Tegunov et al. 2021

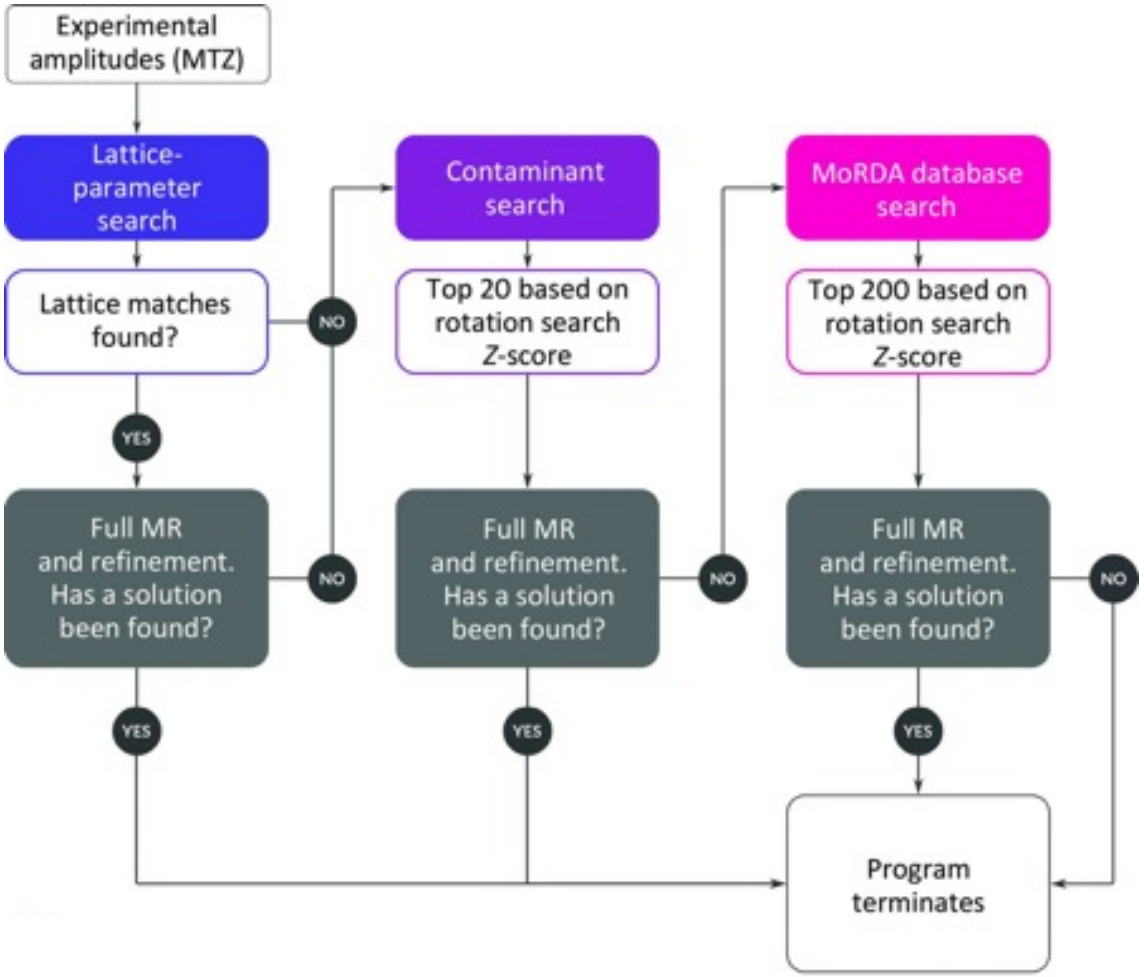


?



EMBL

Sequence-free MR with SIMBAD

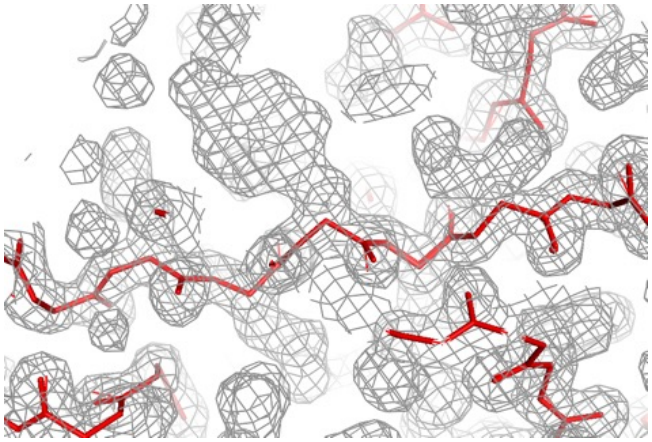


Available in CCP4

Simpkin et al (2018). Acta Cryst. D74, 595-605

EMBL

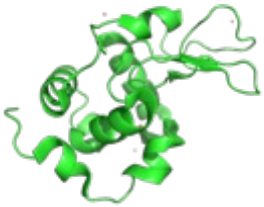
Protein sequence identification from a map



main-chain model and a map

sequence profile

GESAMT, FATCAT, DALI

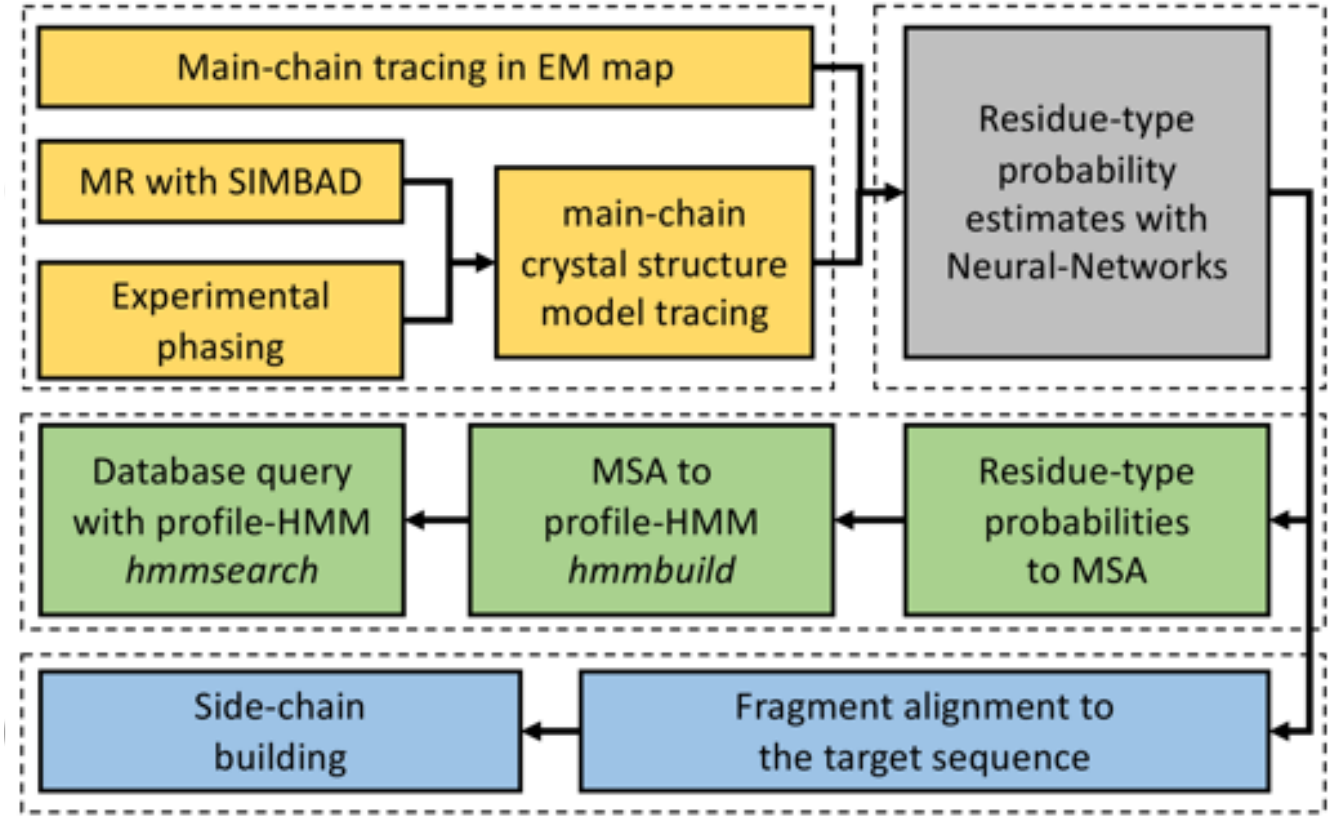


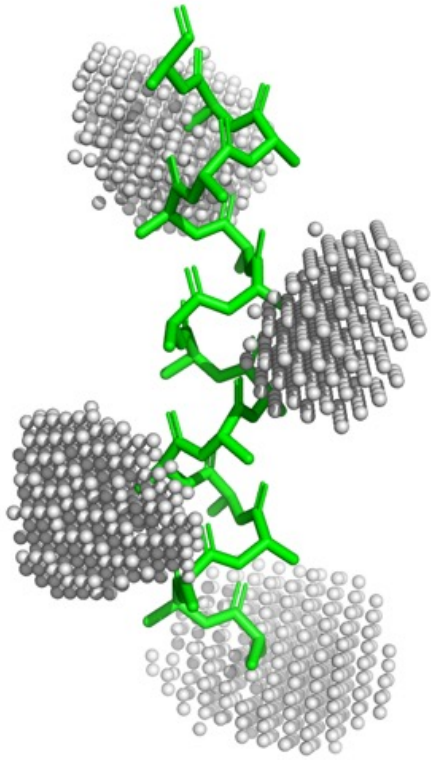
protein structure database

sequence database

EMBL

Protein sequence identification with findMySequence





residue-type NN classifier
328 input features
111,800 parameters

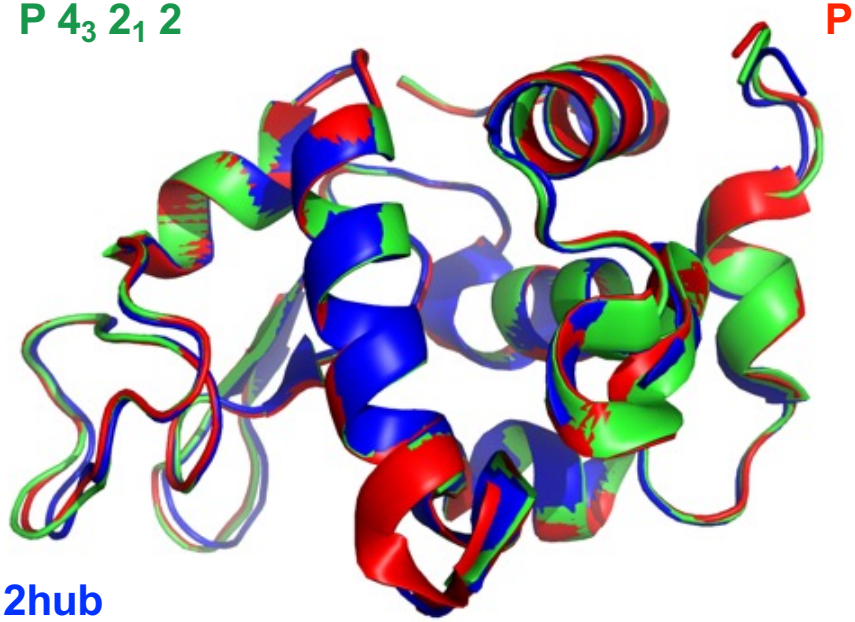
Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX benchmarks: three hen egg-white lysozyme targets

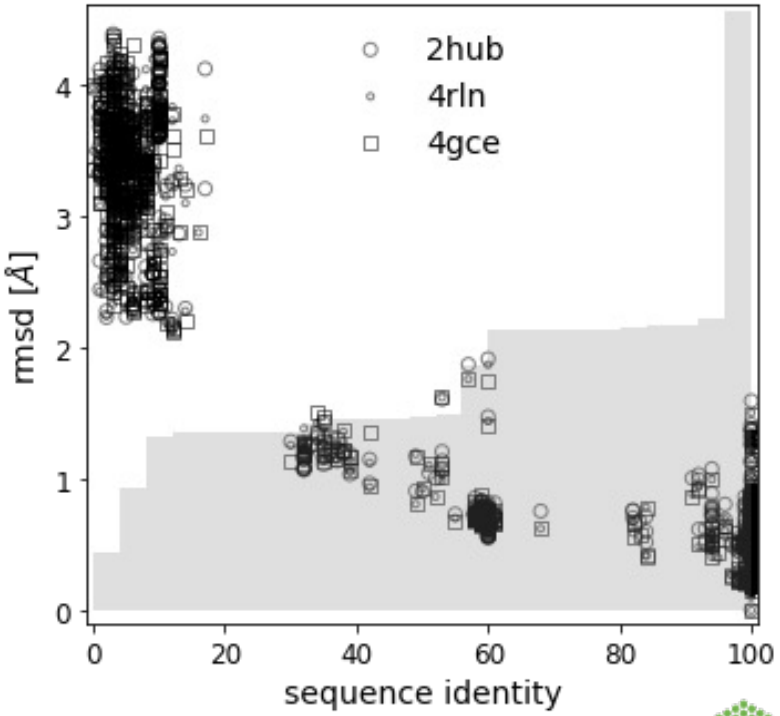
4gce
2.9Å
P 4_3 2_1 2

4rln
2.2Å
P 4_3 2_1 2



2hub
1.2Å
P 4_3 2_1 2

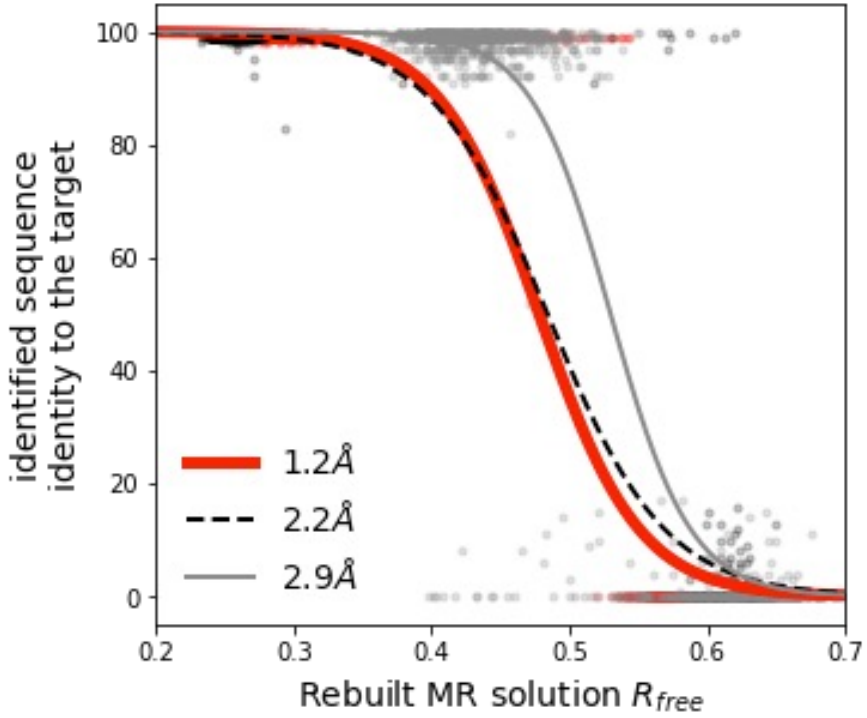
1,500 artificial
MR search models

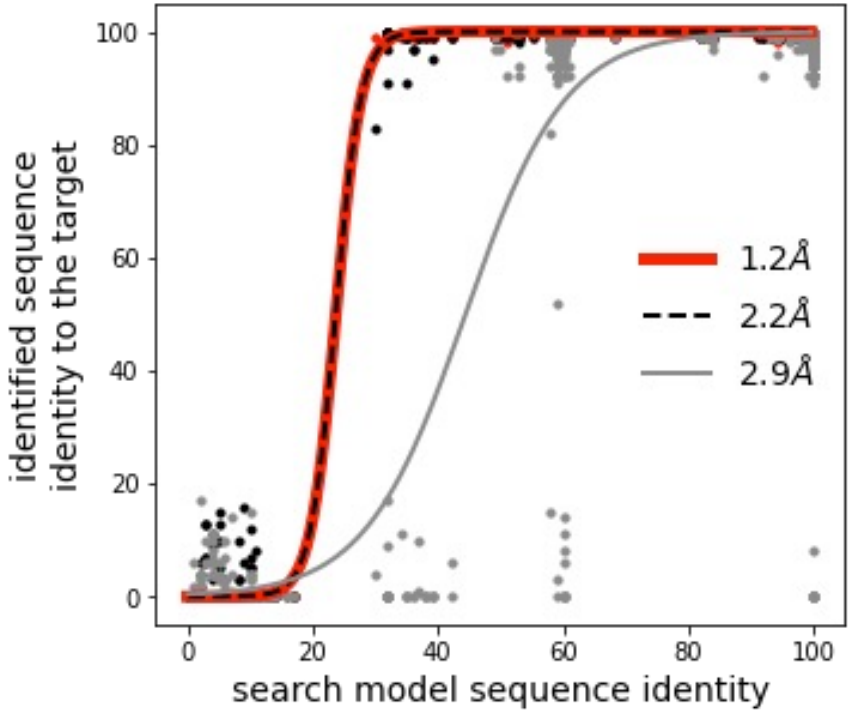


Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX benchmarks: three hen egg-white lysozyme targets

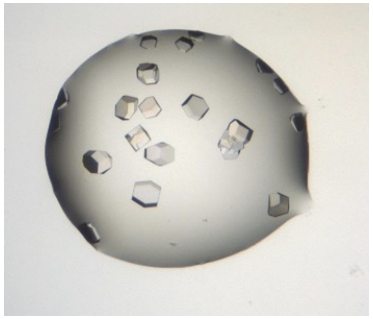


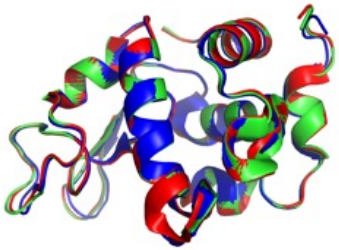


Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX use-case: *Bothrops atrox* venom proteins



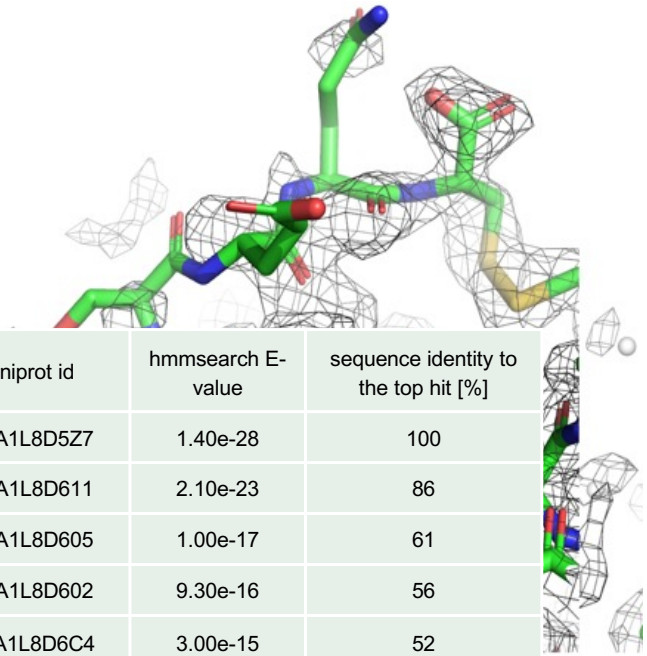


Diego A. Leonardo
Wolfram Seifert-Davila
Dan E. Vivas-Ruiz

EMBL

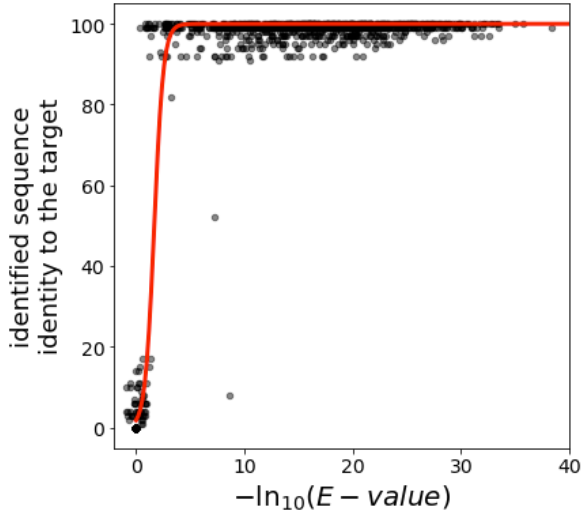
MX use-case: *Bothrops atrox* venom proteins

initial MR solution

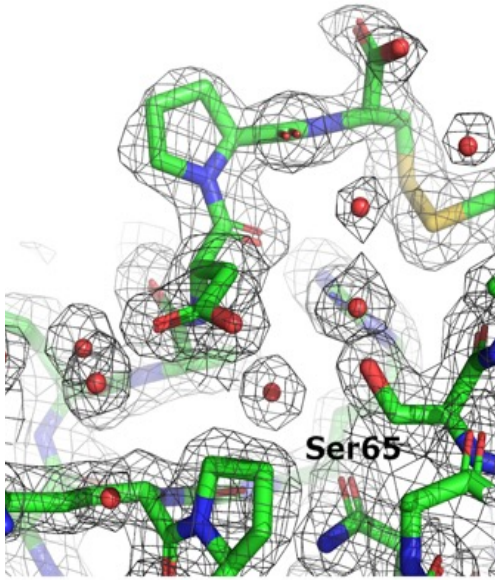


uniprot id	hmmsearch E-value	sequence identity to the top hit [%]
A0A1L8D5Z7	1.40e-28	100
A0A1L8D611	2.10e-23	86
A0A1L8D605	1.00e-17	61
A0A1L8D602	9.30e-16	56
A0A1L8D6C4	3.00e-15	52
A0A1L8D5Z4	4.10e-12	50

model rebuilt
wout sequence



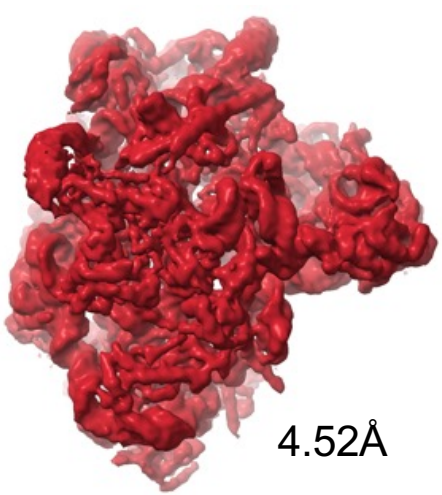
final model



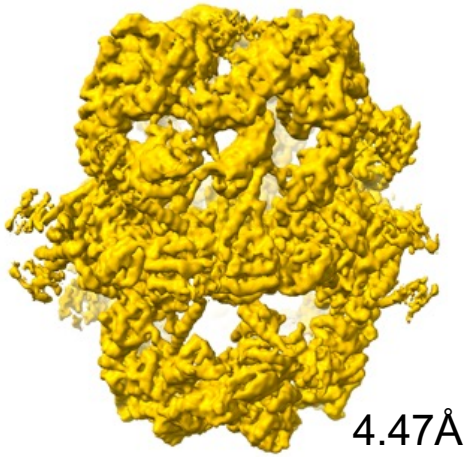
Chojnowski et al IUCrJ 9.1 (2022)

EMBL

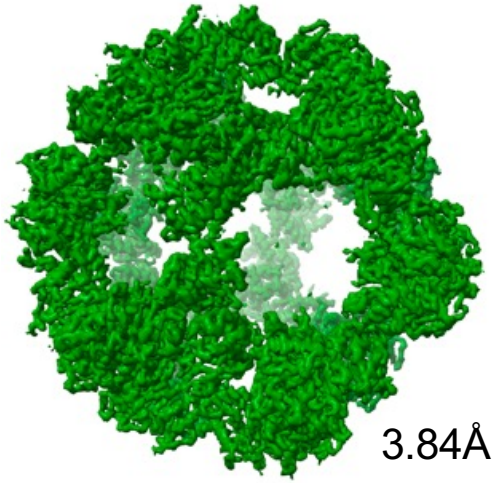
cryo-EM of *C. thermophilum* native cell extracts



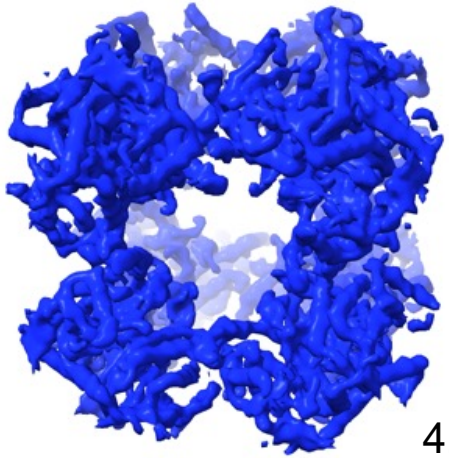
4.52Å



4.47Å



3.84Å



4.38Å

pre 60S

omokage search
AF2
findMySequence

Fatty Acid Synthase

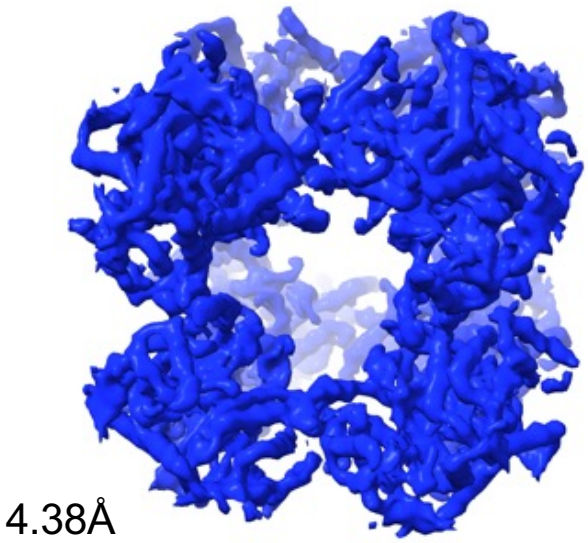
de novo
findMySequence

pyruvate dehydrogenase

with Panos Kastritis and Ioannis Skalidis
Skalidis et al Structure 2022

EMBL

two dehydrogenases in *C. thermophilum* native cell extracts



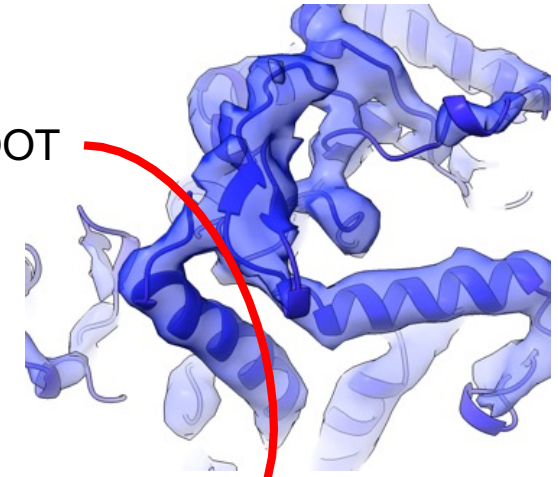
4.38Å

omokage search

>E2b ketoacid
MRSWFF
SRVLRQ
QWRVVR

(~20% SI)

AF2+COOT



>E2o oxo-acid
MLYRGL
RMAARA
APKFAV

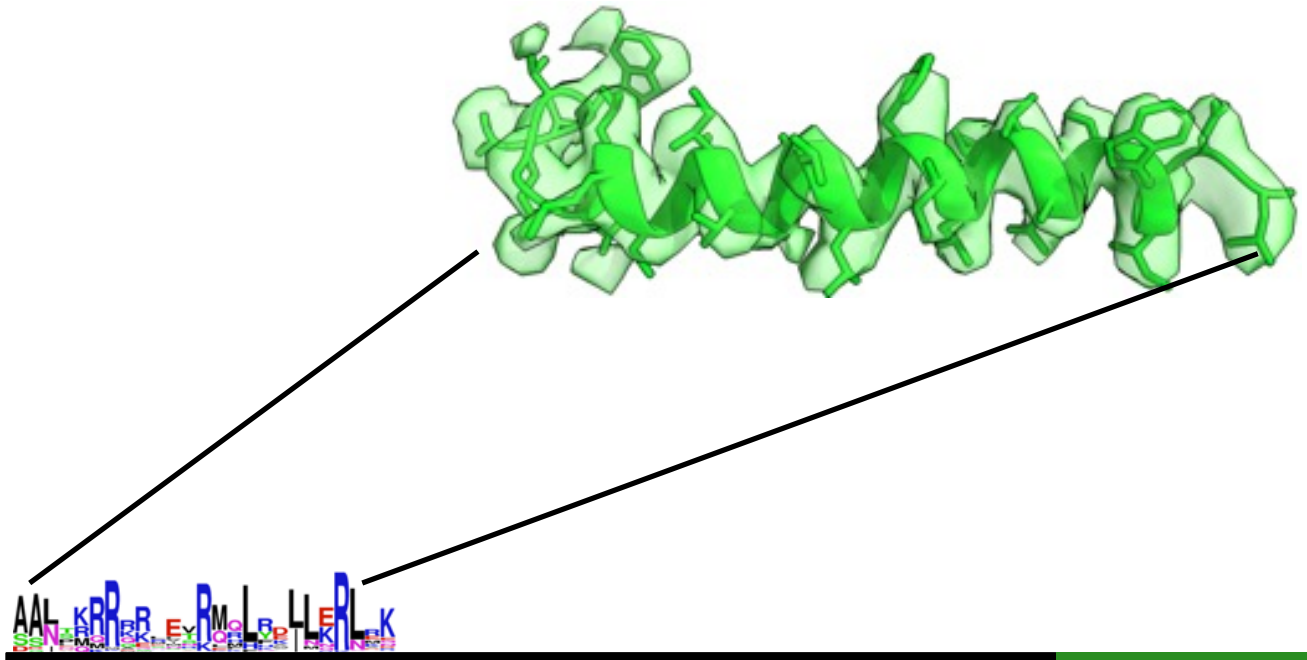
AF2+COOT

findMySequence

Chojnowski et al IUCrJ 2022
Skalidis et al Structure 2022

EMBL

Sequence re-assignment as a model validation



target sequence

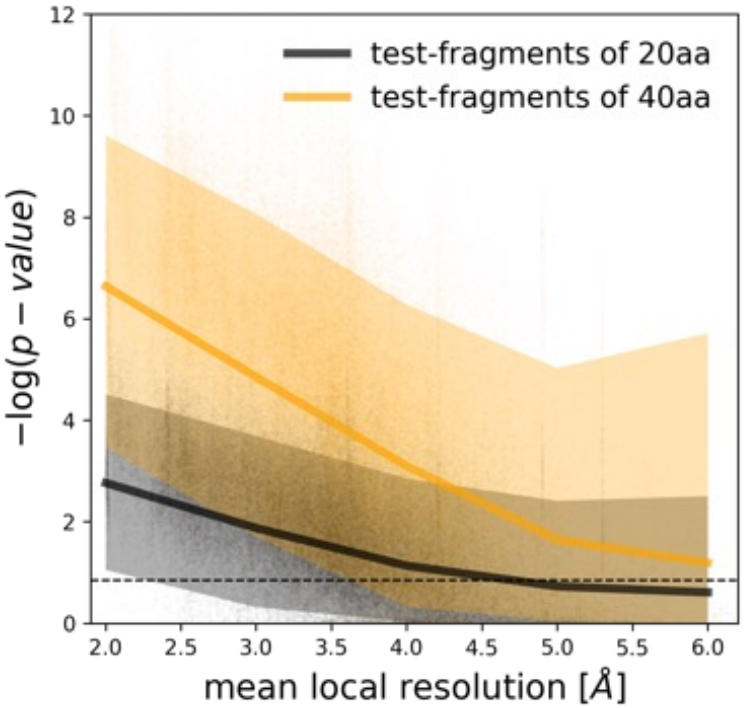
alignment p-value

	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

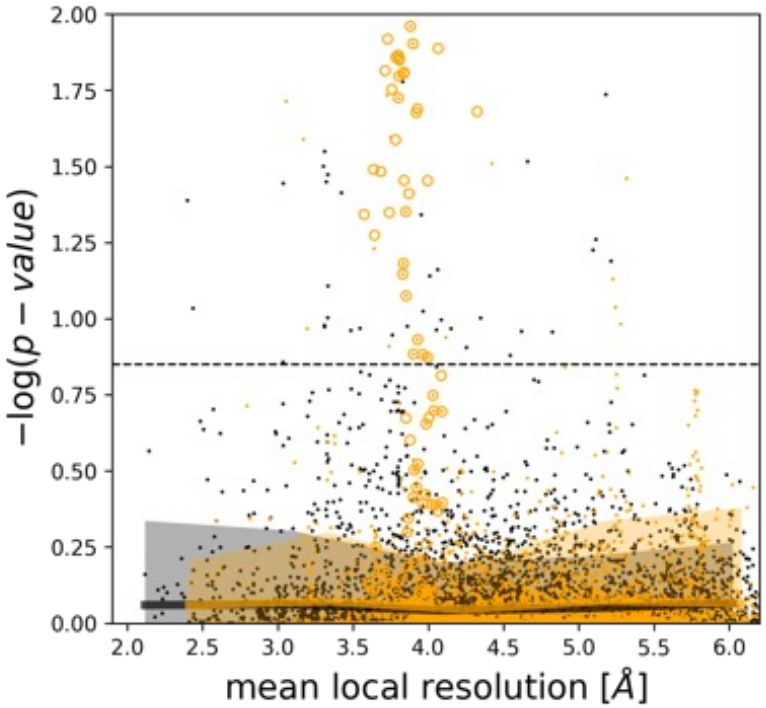
EMBL

Is the sequence assignment p-value a reliable score?

sequence OK



sequences differ



	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

p-values for 30k protein chain fragments re-assigned to target sequence

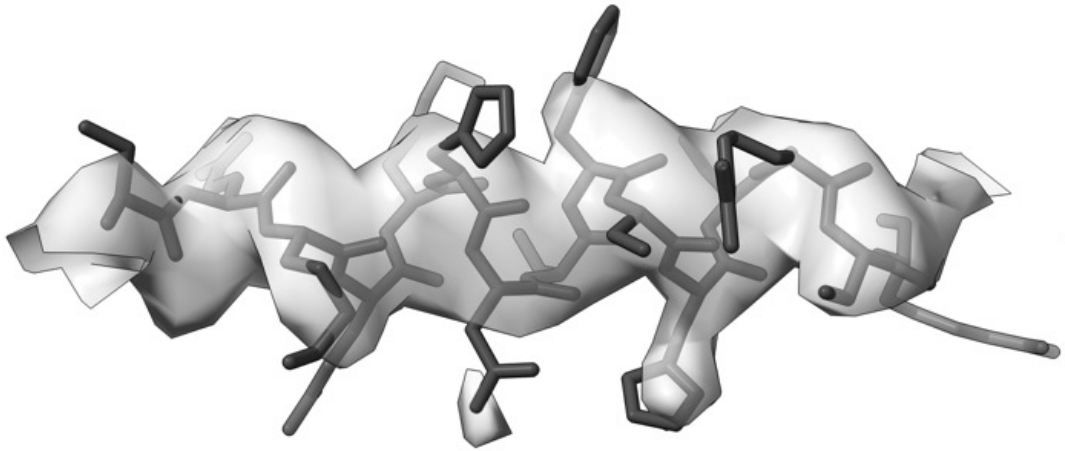
Chojnowski ActaD 2022

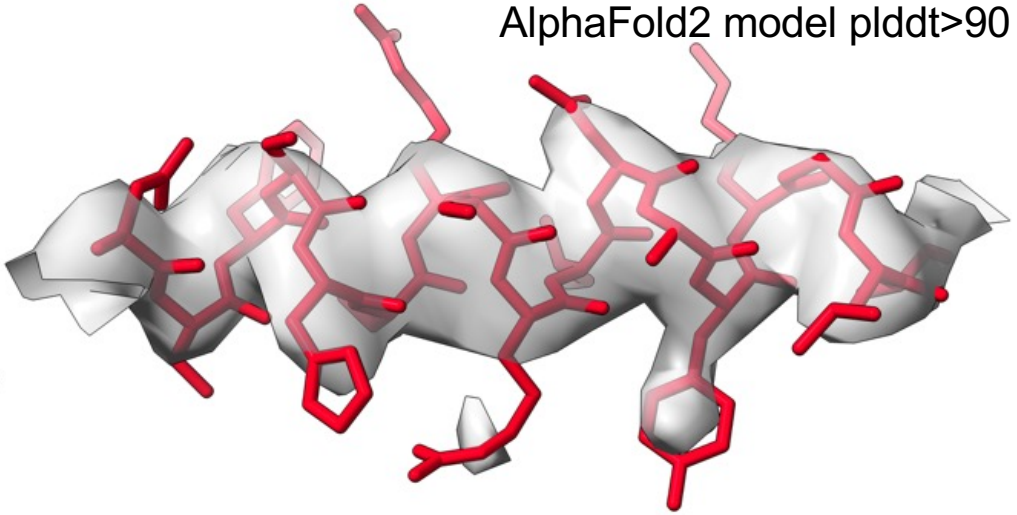
EMBL

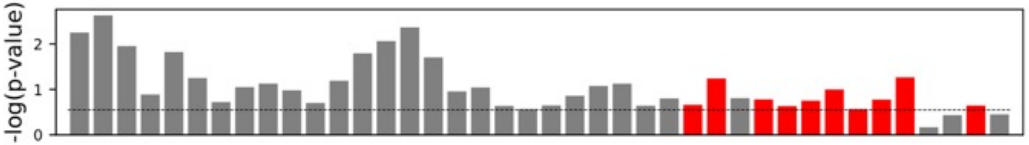
checkMySequence: finding a better hypothesis

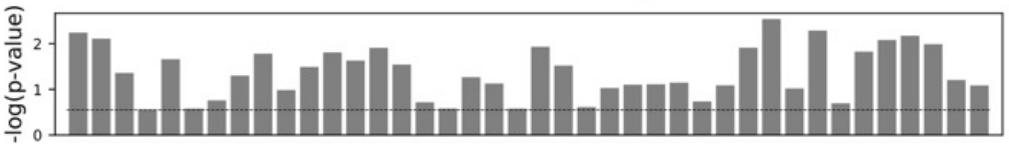
cytoplasmic domain of a cation channel at 3.8Å resolution

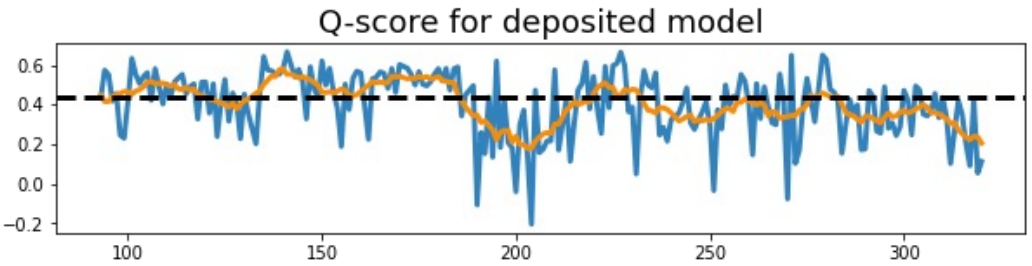
AlphaFold2 model plddt>90

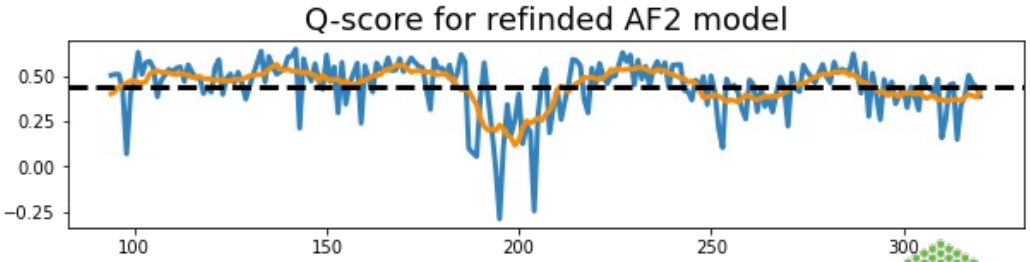












Chojnowski ActaD 2022

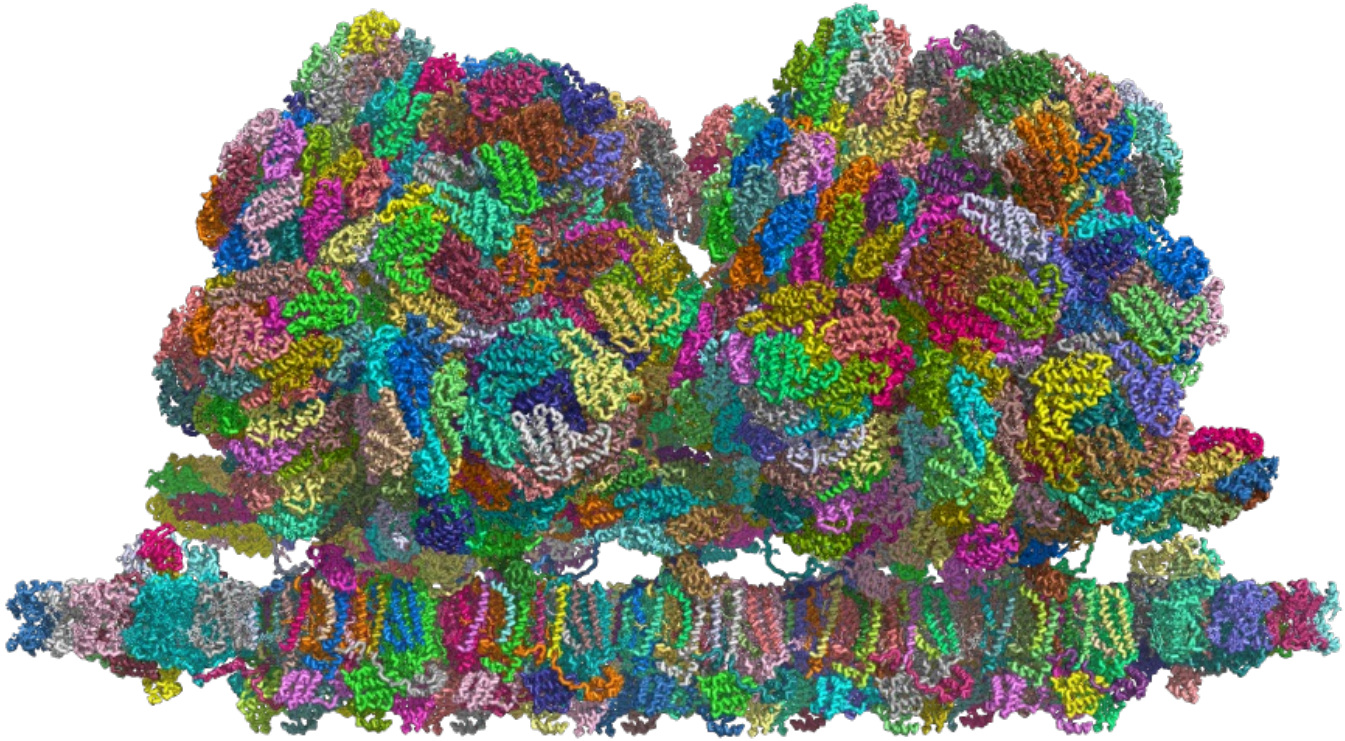
EMBL

checkMySequence: complete sequence assignment validation

```
*********************************************************************************
*********************************   SUMMARY   *********************************
*********************************************************************************

 ==> Unidentified chains; check input sequences and model-to-map fit

             e/2:51
             g/3:39

 ==> Chains with sequence mismatches; you will have to fix them first!

  model       KDNVVQMMNEKKSFDVSDFPKVYLTTAVEEDLDT--
              |||||||||||||||||||||||||*|||||||
  refseq      KDNVVQMMNEKKSFDVSDFPKVYLTTTVEEDLDTRG

 ==> Possible sequence assignment issues

  - Fragment N/5-24 has a chain break at N/11 and no residue indexing gap, check!

  model       ---------VELTEEE-LYISKKNLLFKRF----------
                       ||||||| |||||||||||||
  refseq      AAAAAAMPRVELTEEEKLYISKKNLLFKRFVEPGRLCLIE


- Fragment F/356-395 is shifted by -4 residues [-log(p-value)=1.99]
  model seq 356-395
      sknkkEKRVQKQIQKKELQKINHDYYKGVAKAVKKKKKREEKKAKskktanqavi
    new seq 360-399
      sknkkekrvQKQIQKKELQKINHDYYKGVAKAVKKKKKREEKKAKSKKTanqavi
```

When you think you've seen it all…

In situ **double**-PBS-PSII-PSI-LHCs megacomplex at 4.3Å



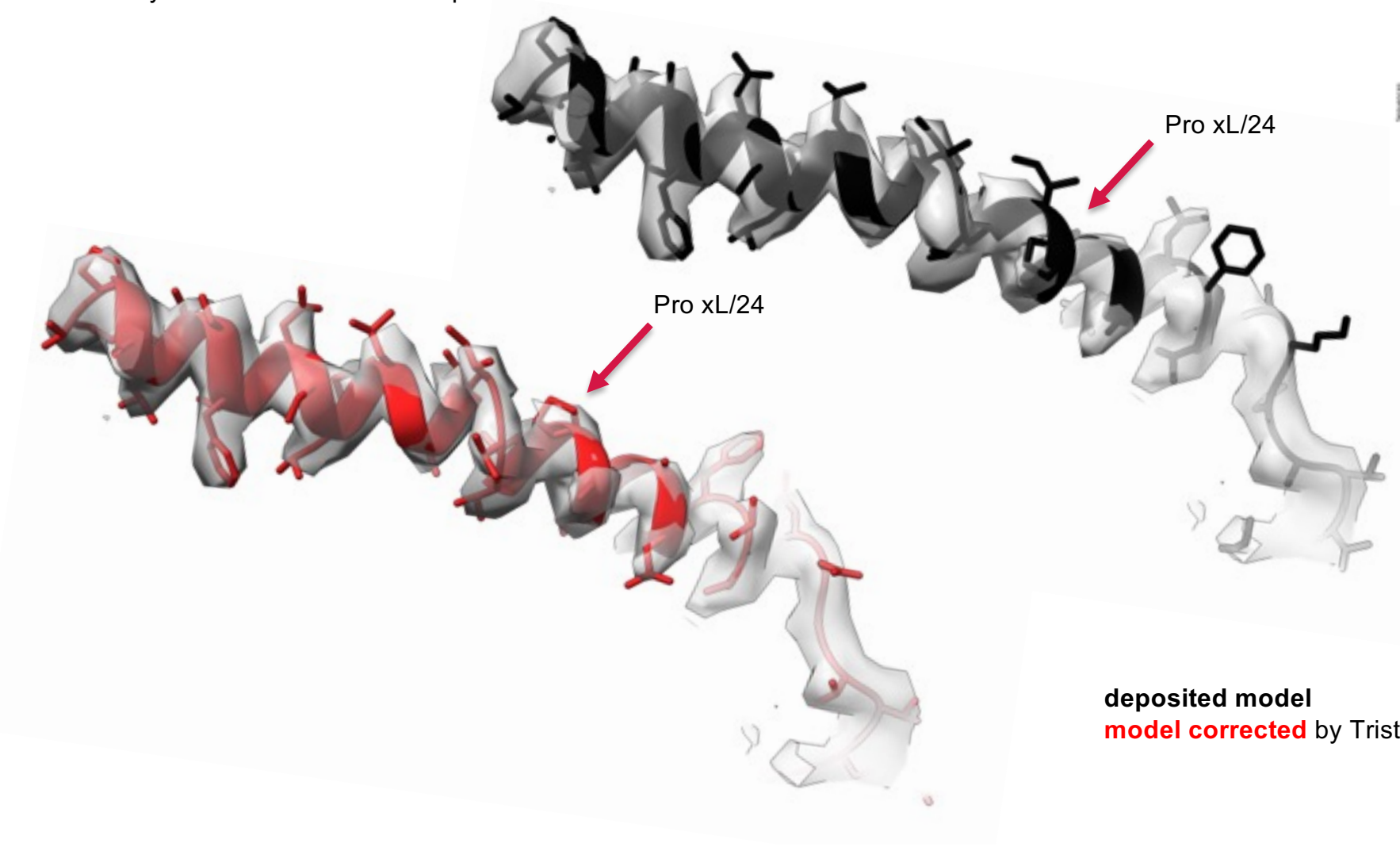
37MDa/305,000 residues/1,792 aa chains

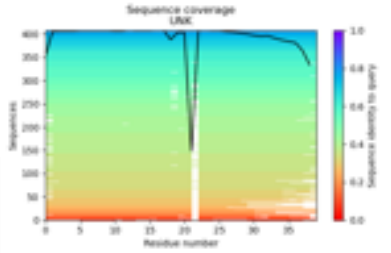
7y5e/EMD-33618
You, X., Zhang, X., Cheng, J., Xiao, Y.N., Sui, S.F. *Nature 2023*

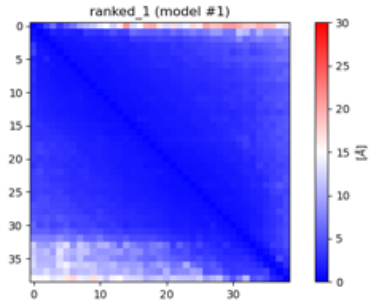
EMBL

In situ photosystem megacomplex

Photosystem II reaction center X protein







pTM=0.57
<pLDDT>=96

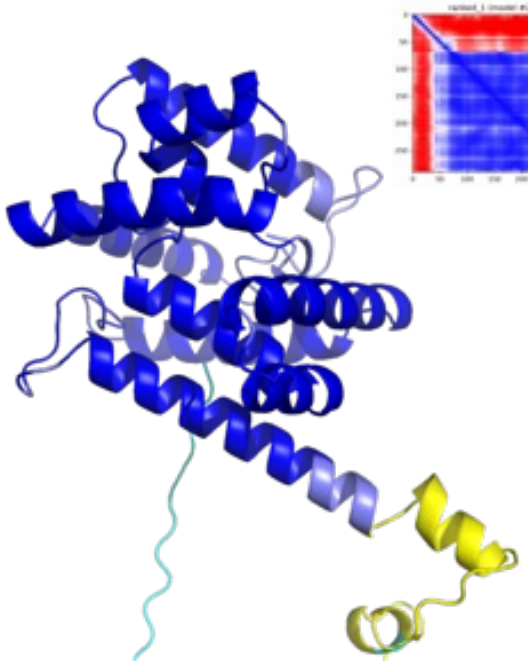
deposited model
model corrected by Tristan Croll

EMBL

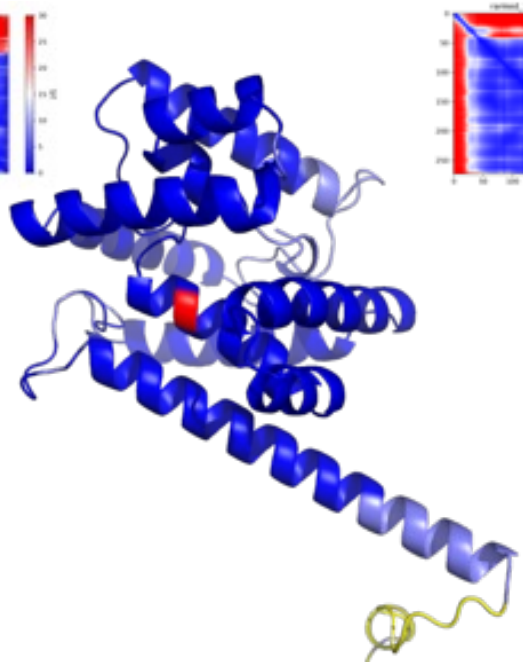
In situ photosystem megacomplex

two R-phycoerythrin gamma chain sequences

deposited model
(7y7a/Y5)



AlphaFold2 prediction for
deposited model sequence
(A0A5J4YX19)



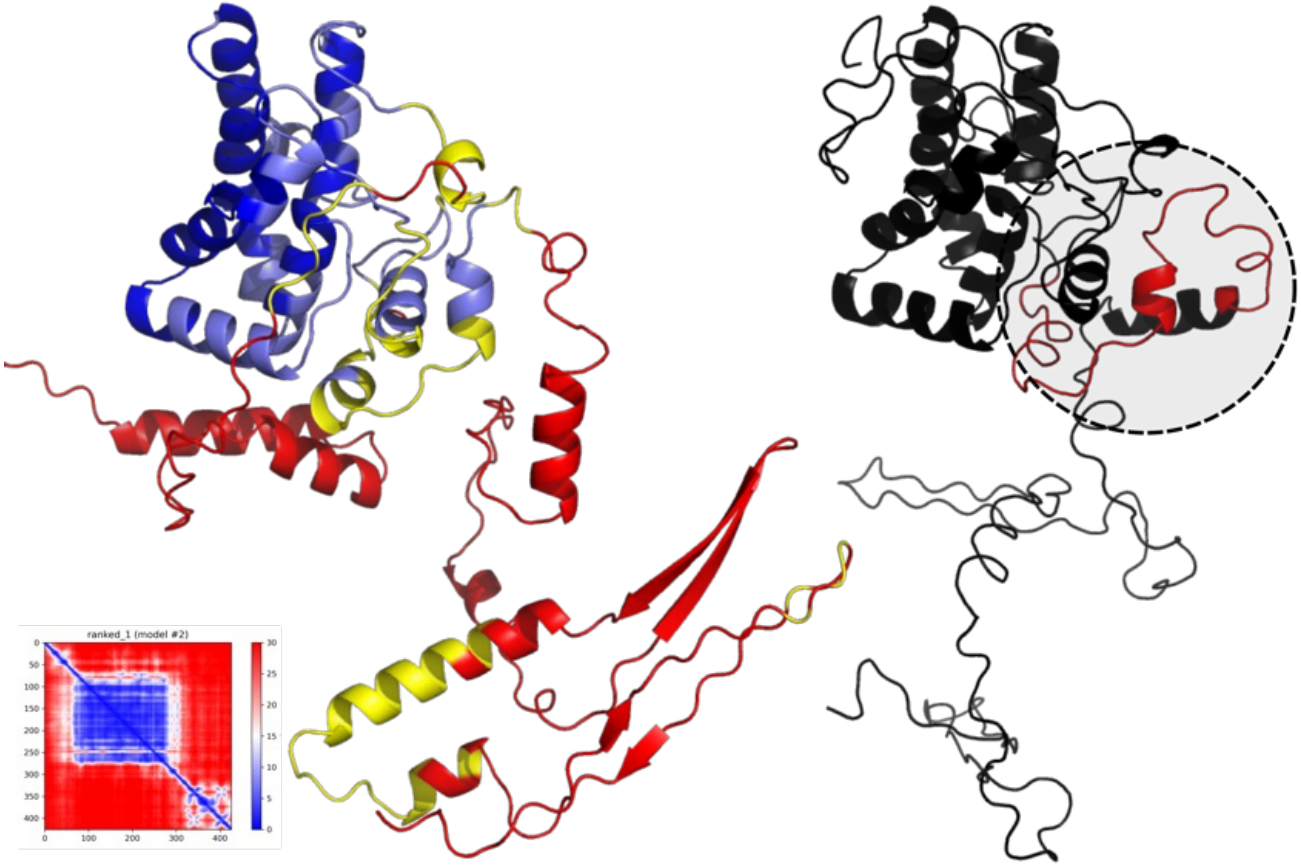
AlphaFold2 model for
alternative sequence
(A0A5J4YZM7)

65% sequence identity
0.5Å CA RMSD

EMBL

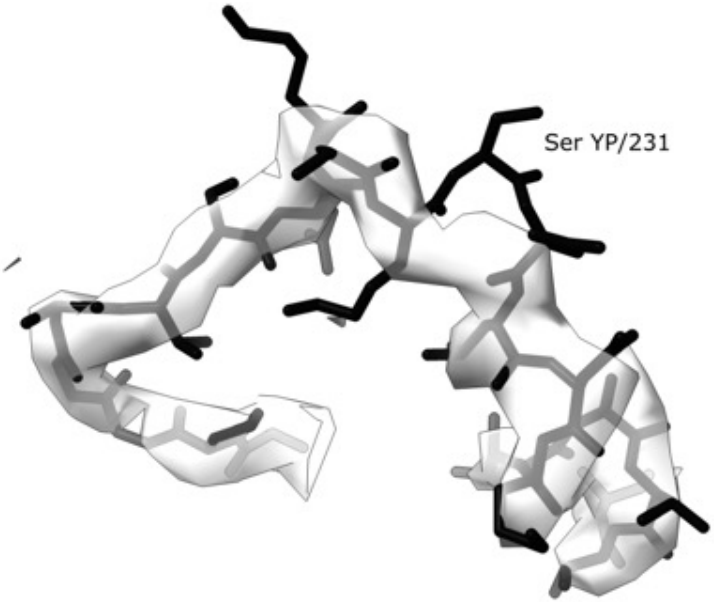
In situ photosystem megacomplex

Phycobilisome linker polypeptide



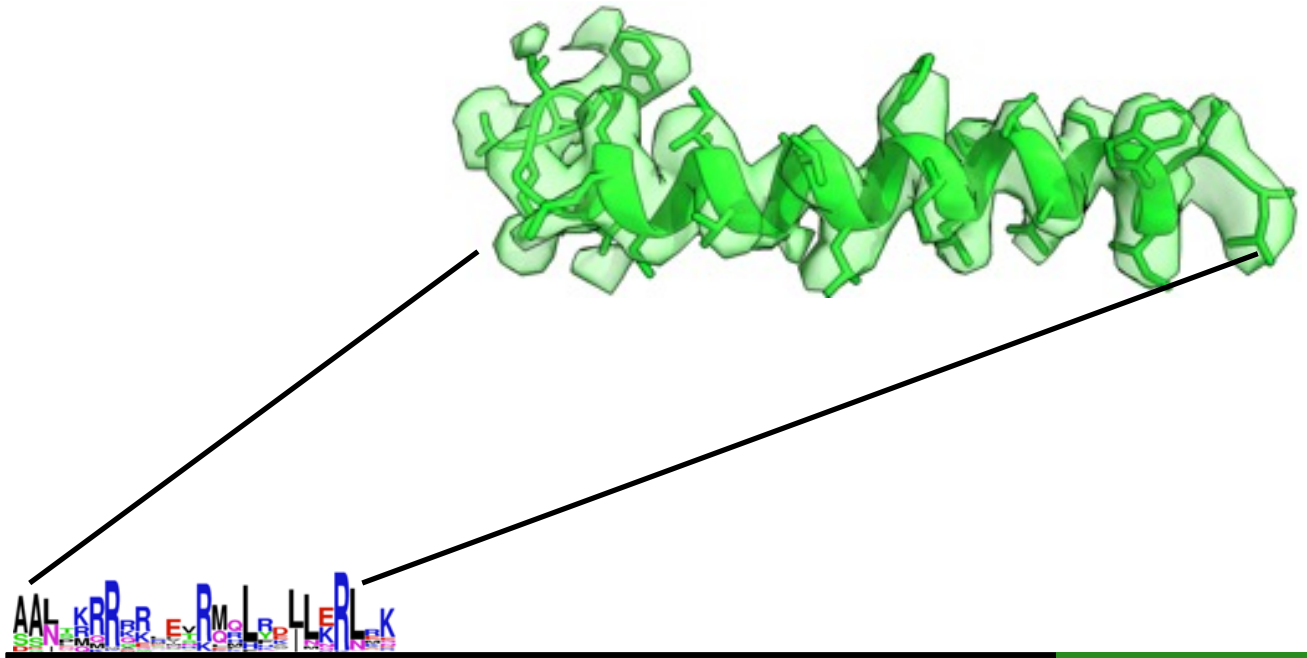
AlphaFold2 model

deposited model



EMBL

checkMySequence and protein crystal structures – model bias



target sequence

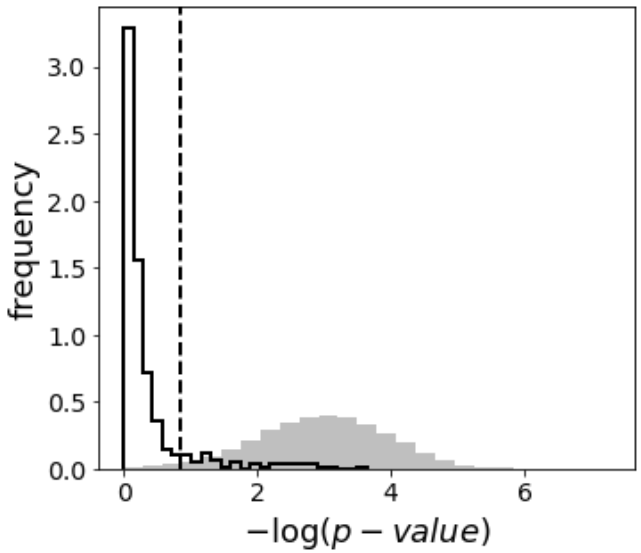
alignment p-value

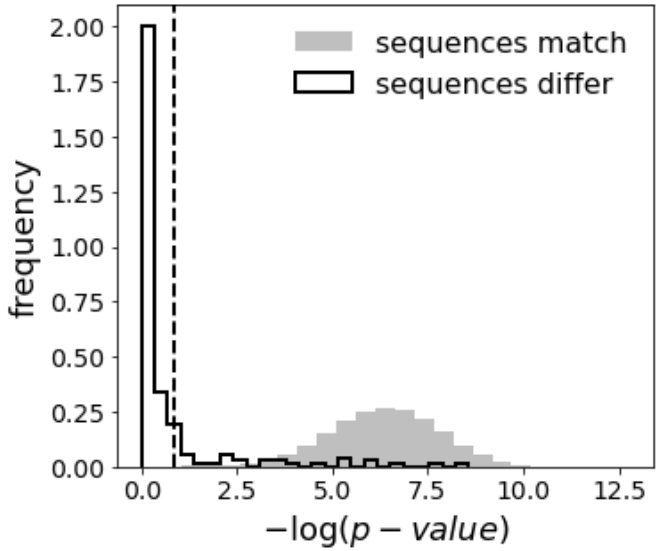
	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

Chojnowski, *in preparation*

EMBL

Sequence assignment in MX – model bias





	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

p-values for 30k protein chain fragments re-assigned to target sequence

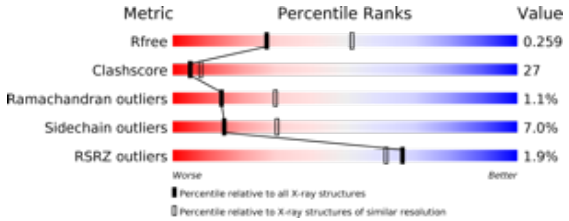
Chojnowski ActaD *2023*

EMBL

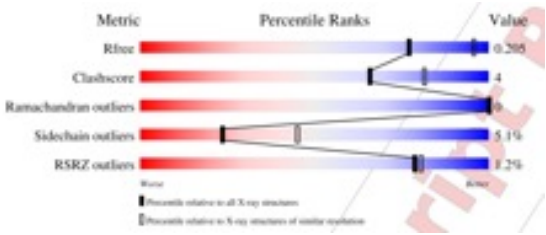
Register shifts in protein crystal structures – example 1

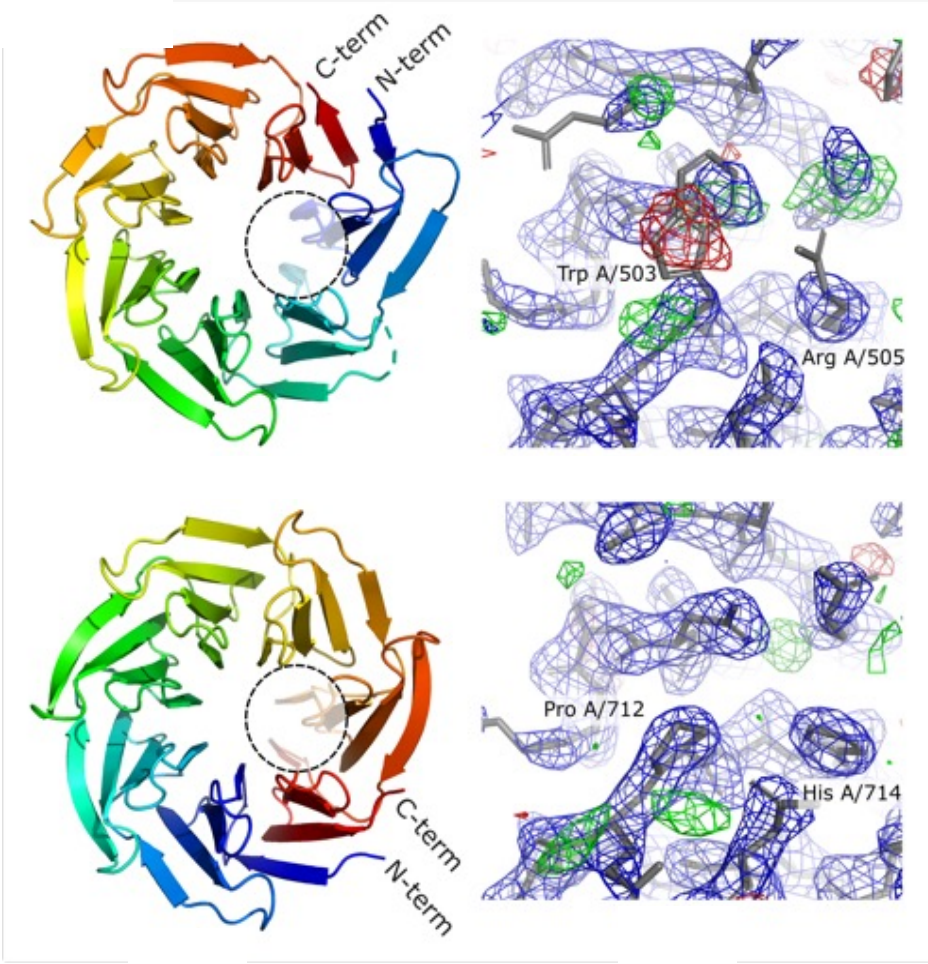
WD40-repeat domain from *T. curvata* @2.5Å

deposited
R/Rfree 22/26



corrected
R/Rfree 17/21



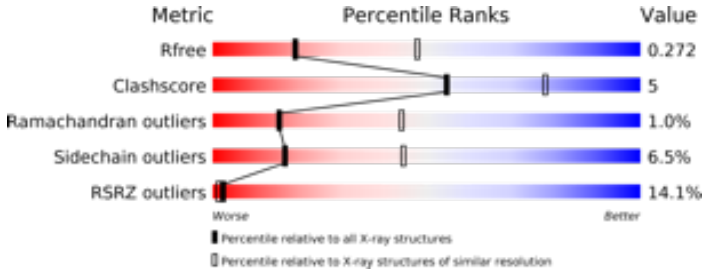


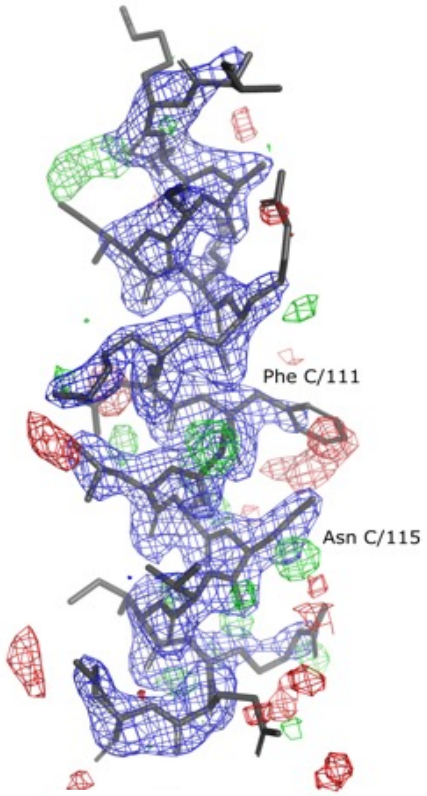
Chojnowski ActaD *2023*

EMBL

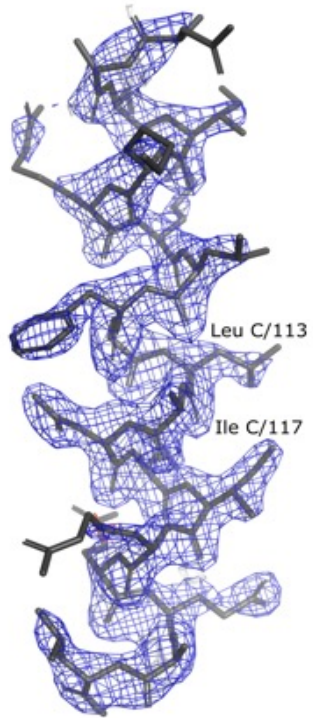
Register shifts in protein crystal structures – example 2

Helicase form *H. pylori* @2.5Å





deposited



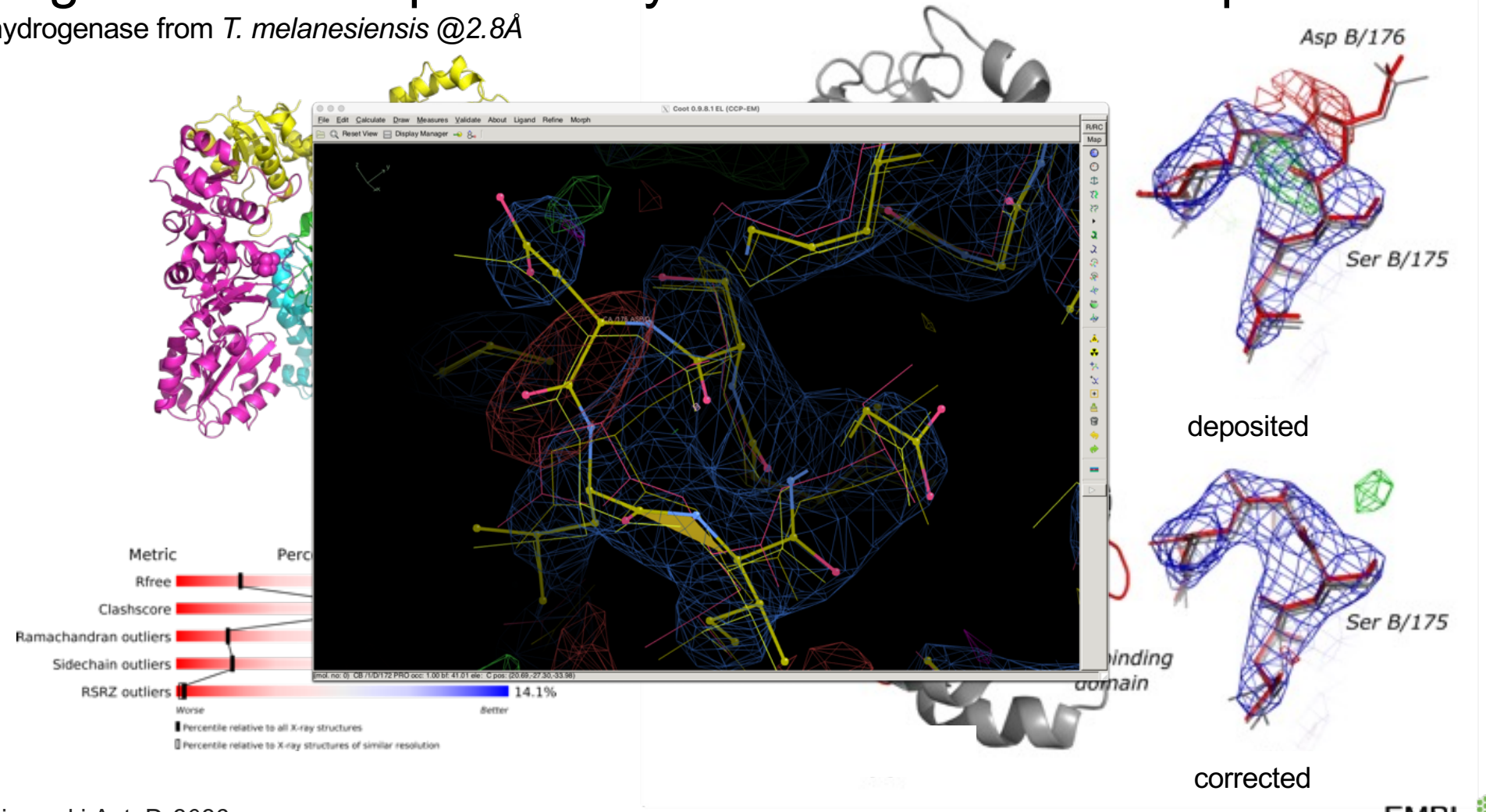
corrected

Chojnowski ActaD *2023*

EMBL

Register shifts in protein crystal structures – example 3

hydrogenase from *T. melanesiensis* @2.8Å

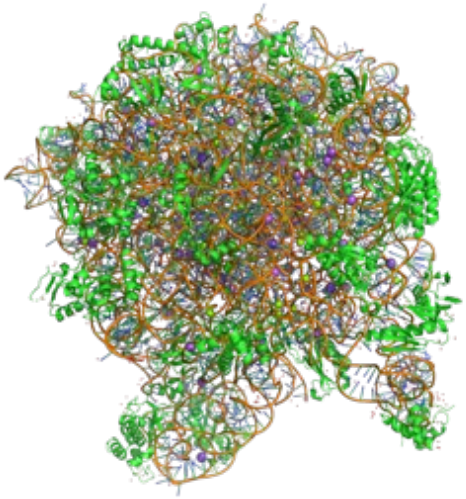


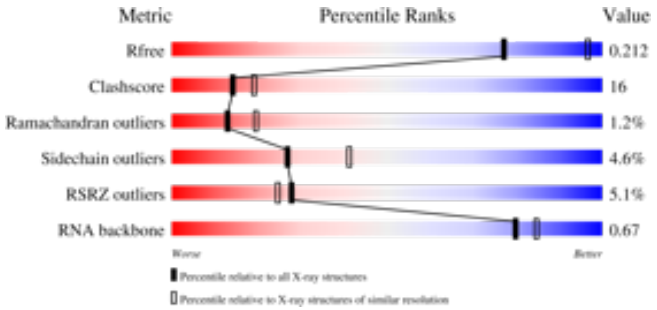
Chojnowski ActaD *2023*

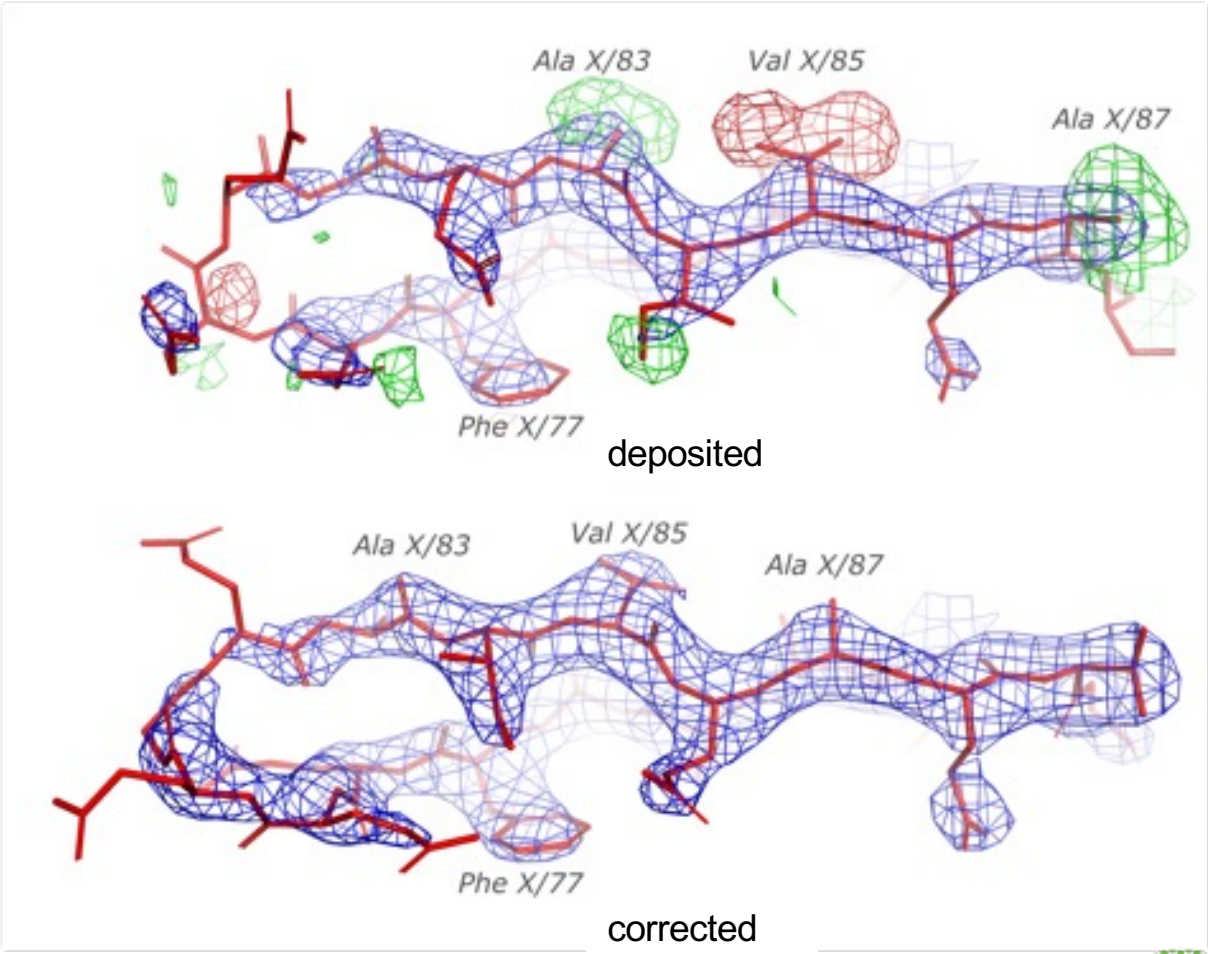
EMBL

Register shifts in protein crystal structures – example 4

Ribosomal protein L31e form *H. Marismortui @2.65Å*





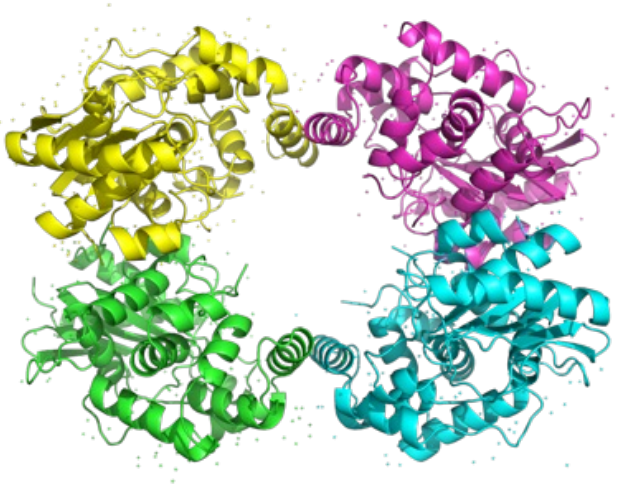


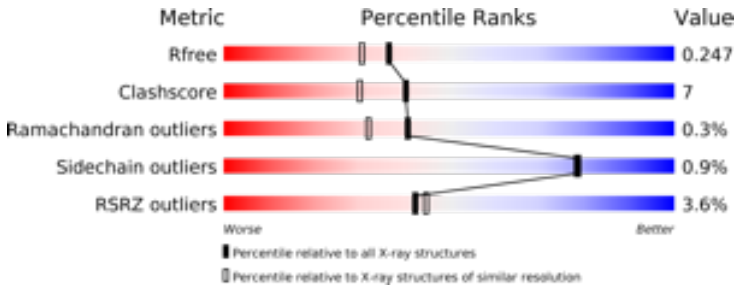
Chojnowski ActaD *2023*

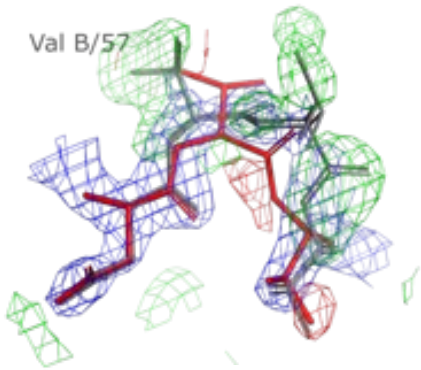
EMBL

Register shifts in protein crystal structures – example 5

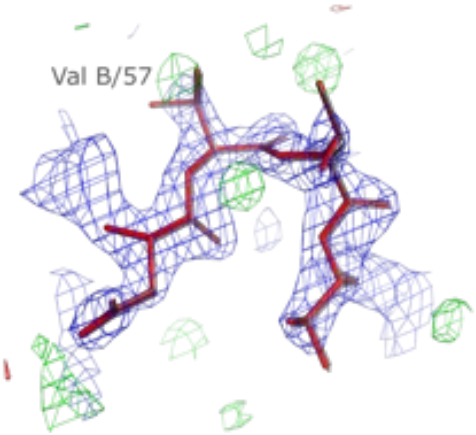
glutaminase from *G. kaustophilus @2.1Å*







deposited



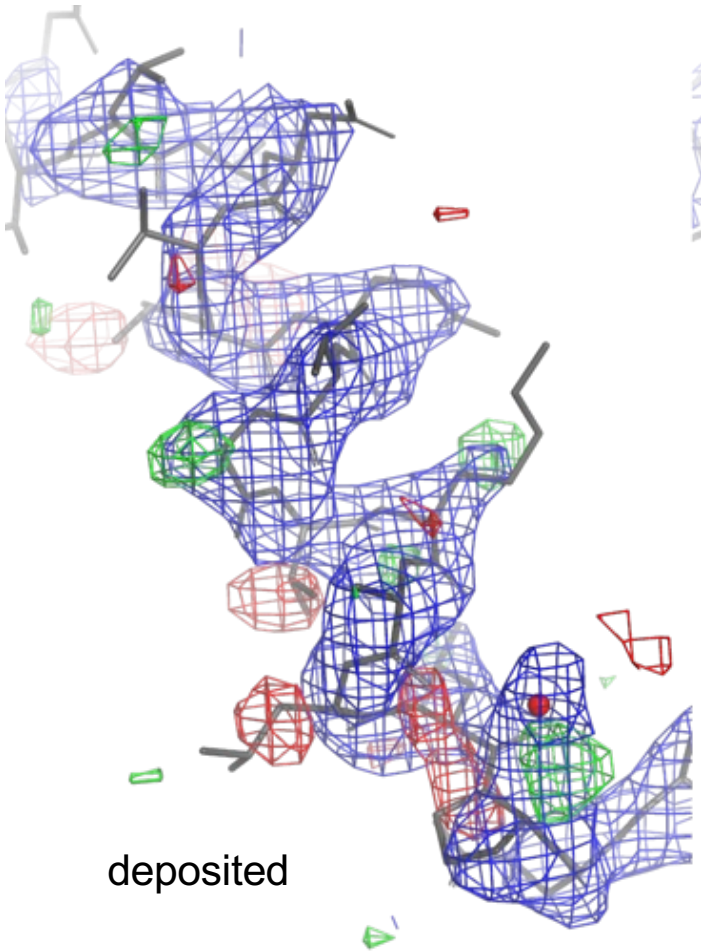
corrected

Chojnowski ActaD *2023*

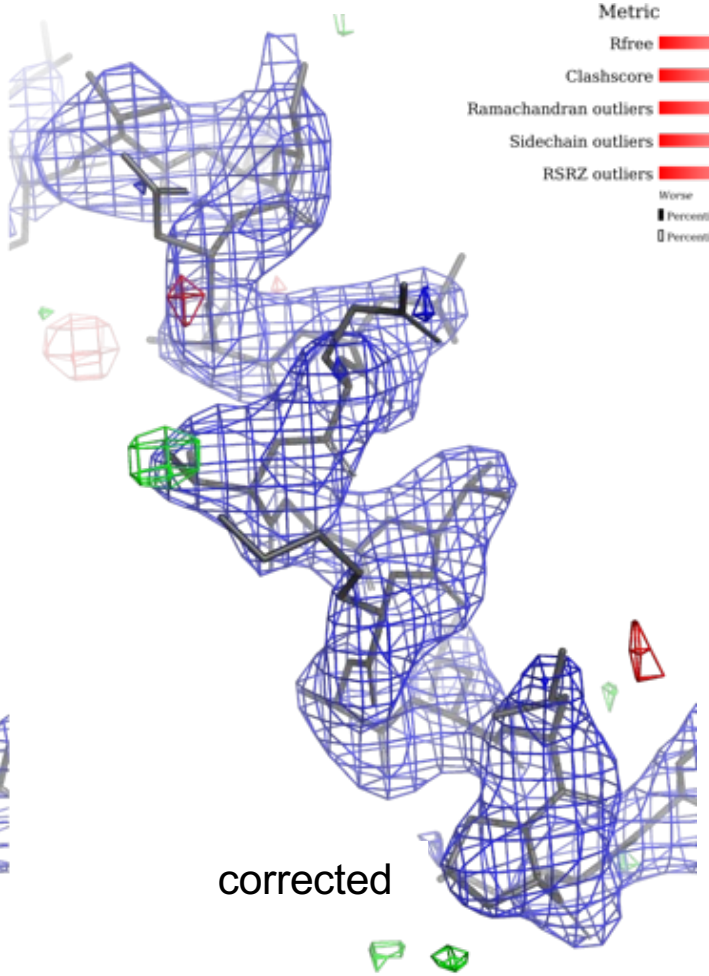
EMBL

Example from last month's PDB release…

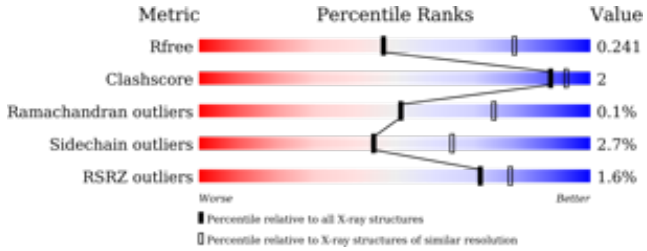
DNA binding protein at 2.75Å



deposited

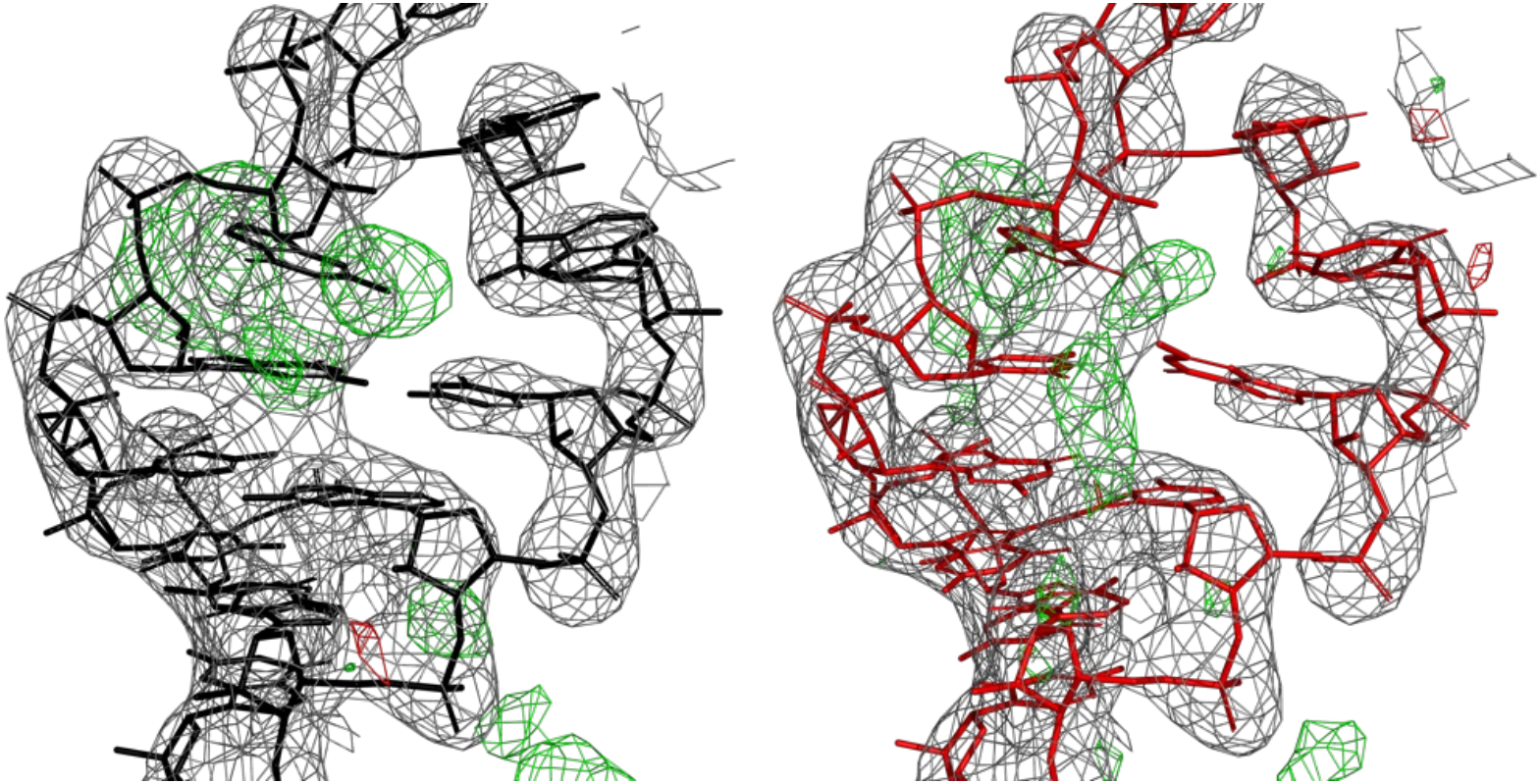


corrected



EMBL

Nucleic acid model building can be challenging

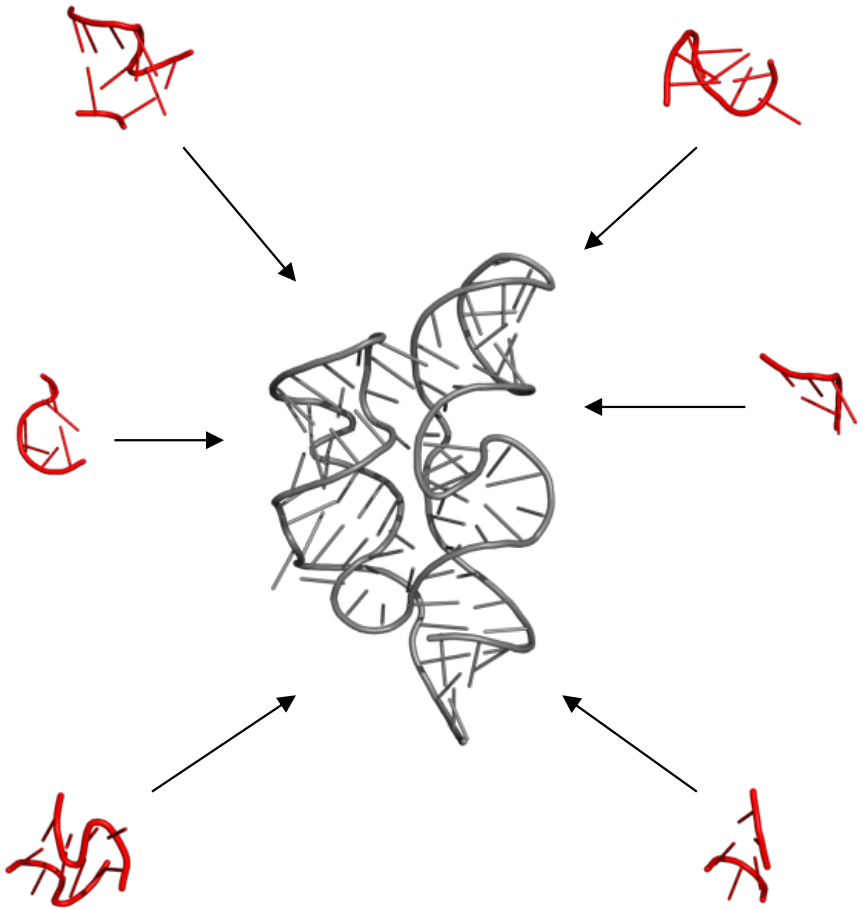


Different types of purines and pyrimidines are (usually) indistinguishable in MX or EM maps

Chojnowski NAR, 2023

EMBL

Base-pairs from backbone geometry



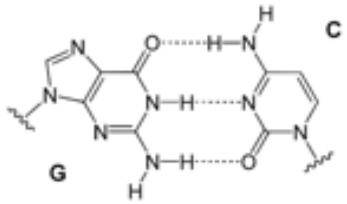
input RNA model backbone is “covered” with motifs with known secondary structure

- base pairs assignment
- sequence validation
- refinement restraints

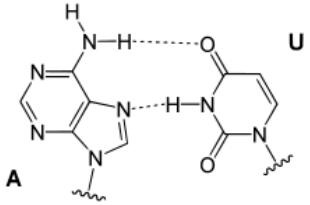
Chojnowski et al. NAR 42 D123-D131 2014
Chojnowski at al. ActaD 71 697-705 2015

EMBL

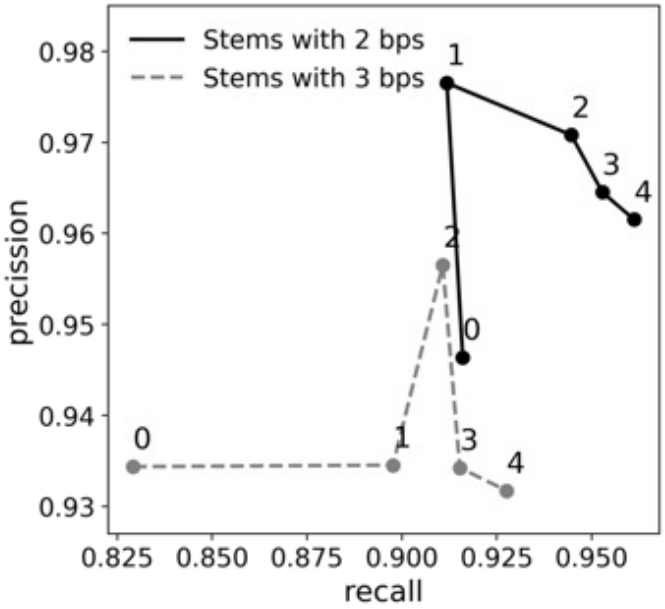
Base-pairs from backbone geometry

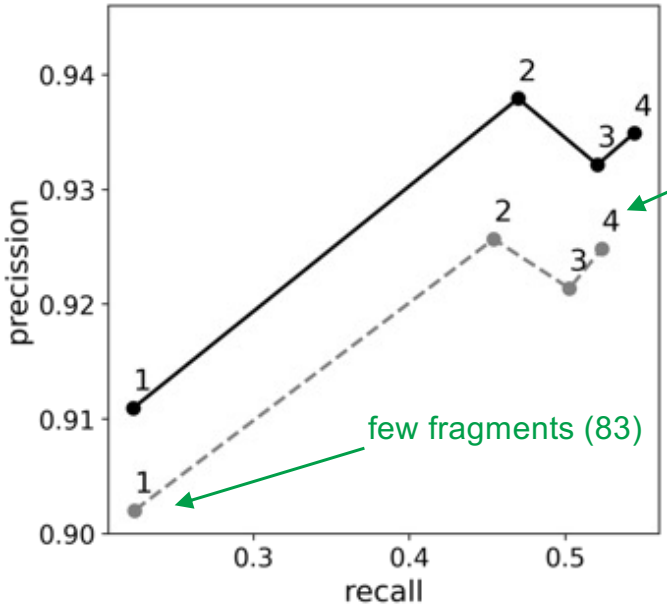


Watson-Crick (canonical) base-pairs



non-canonical base-pairs

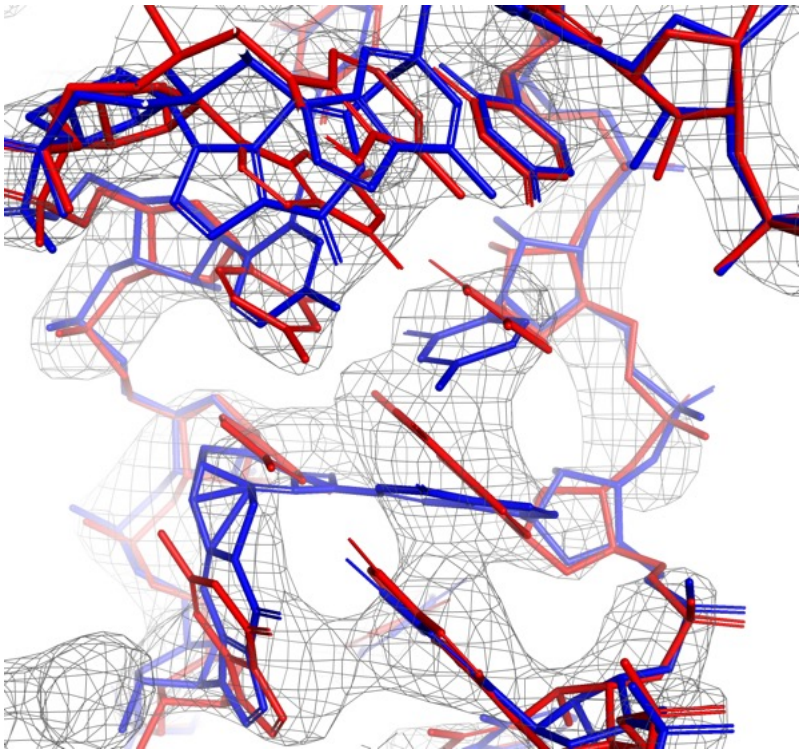




Chojnowski NAR, 2023

EMBL

doubleHelix - base-pairs from backbone geometry



ARP/wARP model of 23S fragment
in cryo-EM map @ 3.2Å resolution

Standard tools

phenix.secondary_structure_restraints
LibG

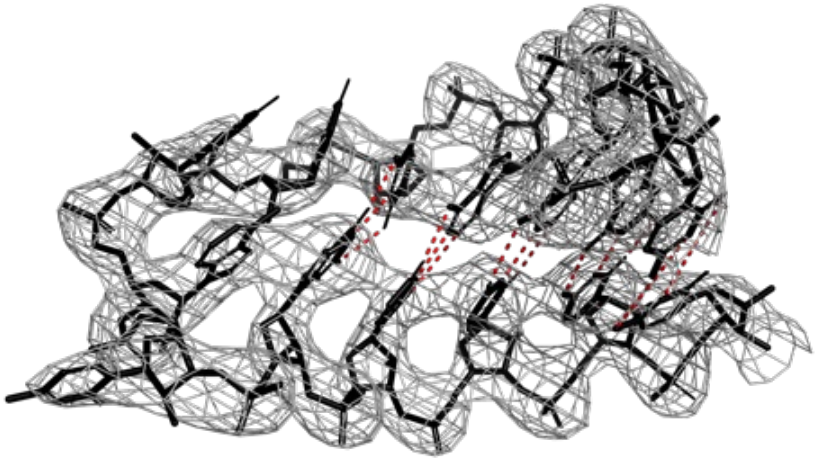
cannot restrain RNA base-pairs if

- geometry is very bad
- sequence is wrong

Chojnowski NAR, 2023

EMBL

checkMySequence and NA crystal structures



	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

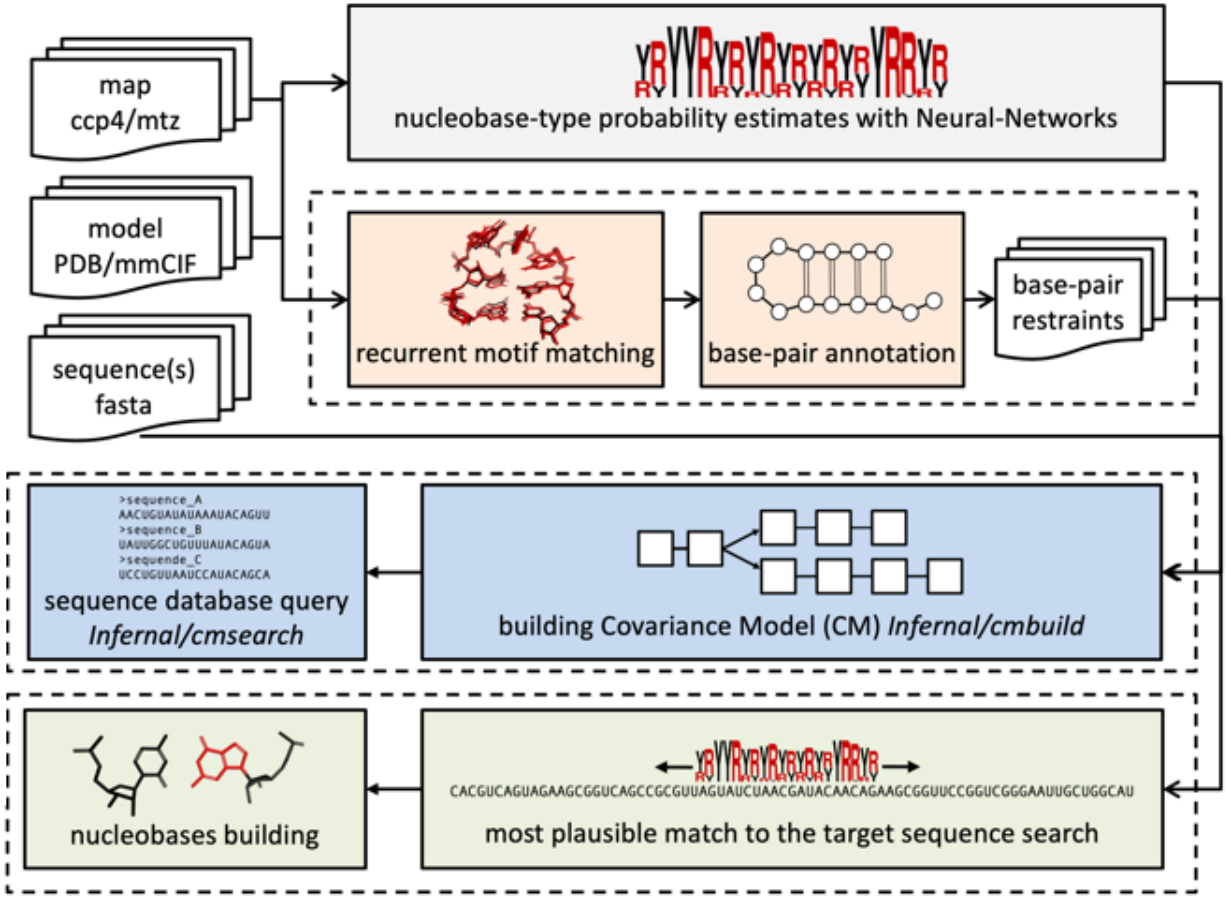
target sequence

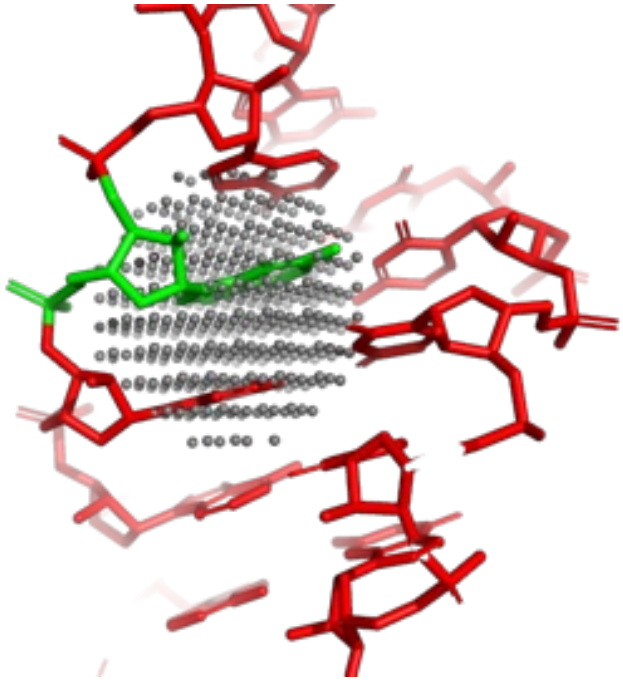
alignment p-value

Chojnowski NAR, 2023

EMBL

Nucleic acid sequence assignment and validation in EM/MX



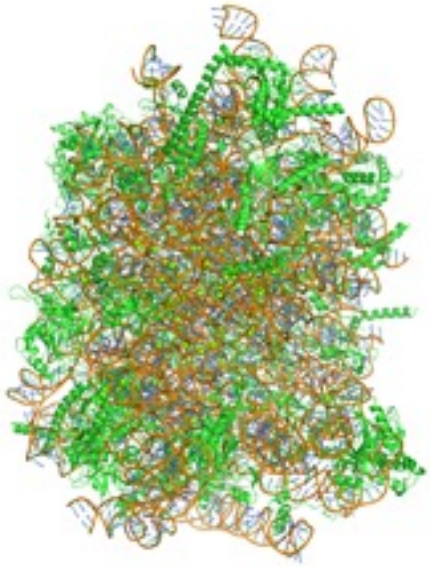


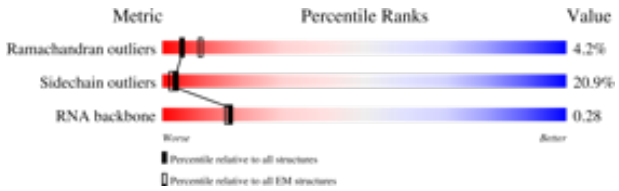
Chojnowski NAR, 2023

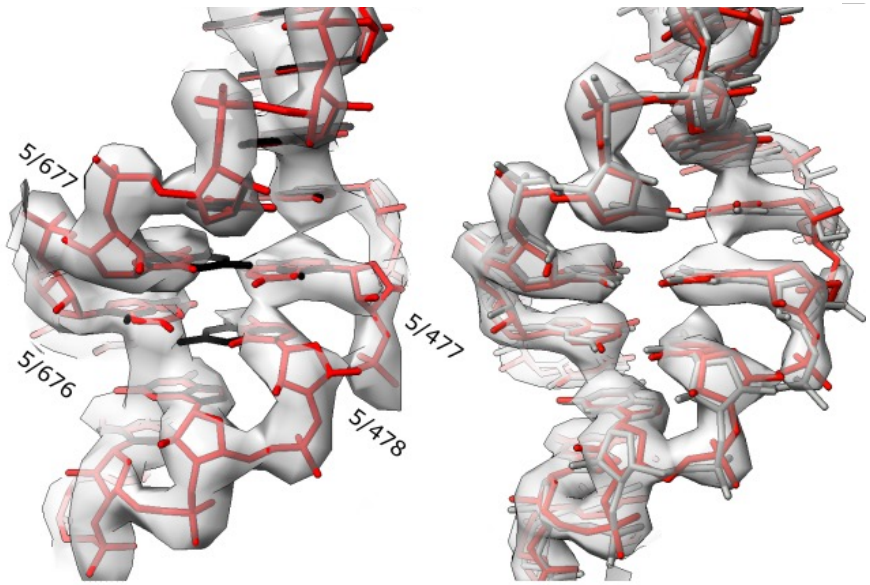
EMBL

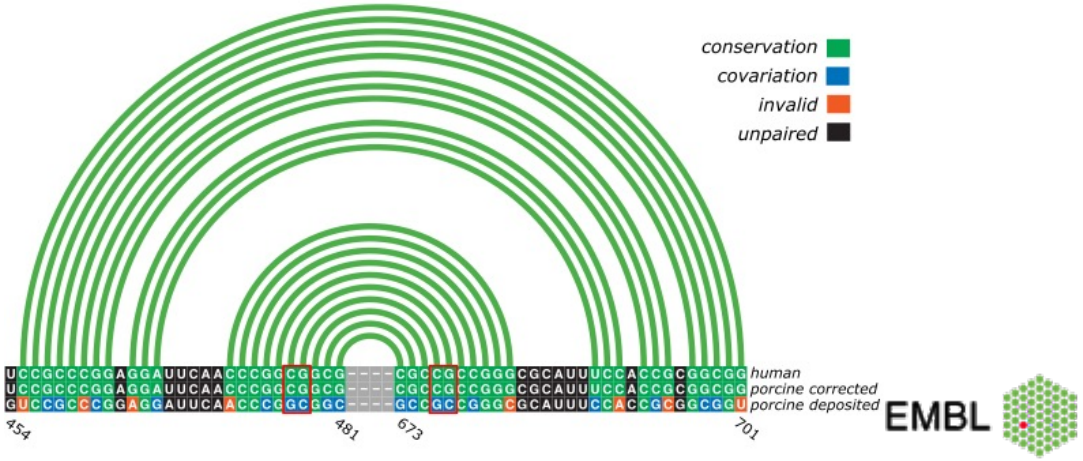
Sequence assignment issues in rRNA

EM model of mammalian ribosome @3.4Å (deposited 2014)





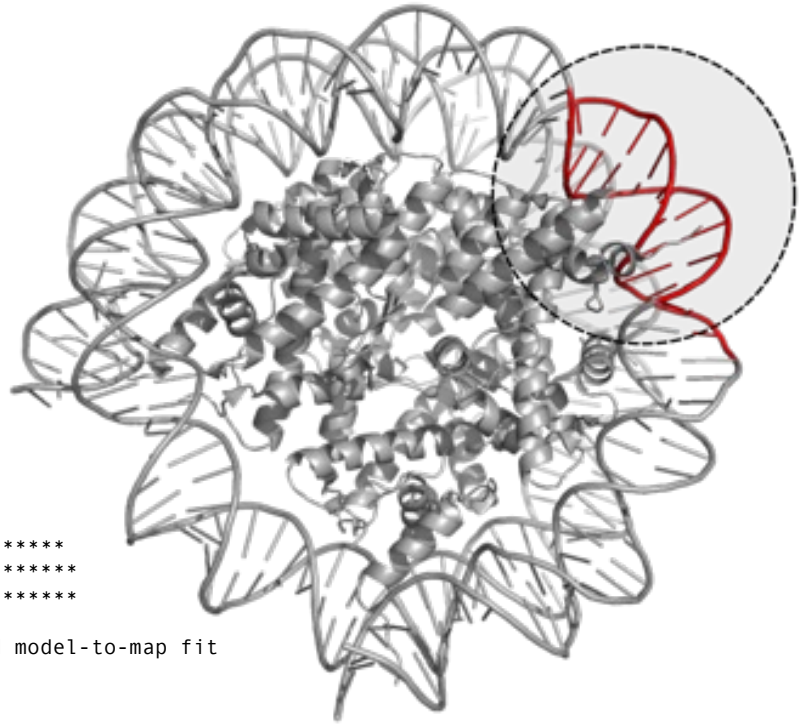


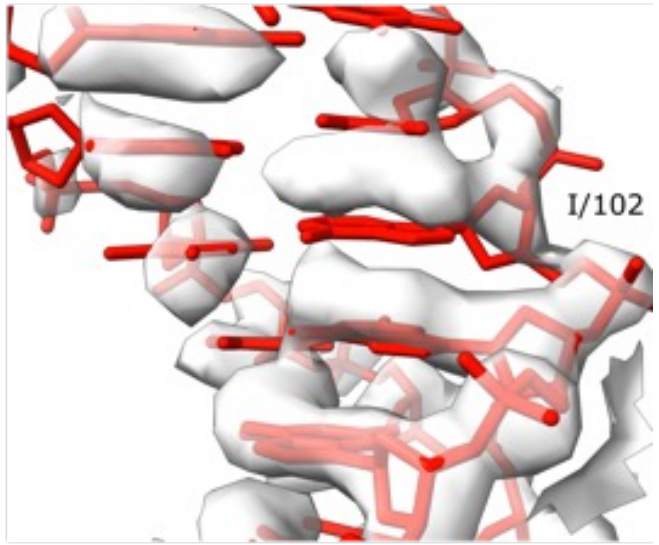


Chojnowski NAR, 2023

EMBL

recent EM use case: nucleosome



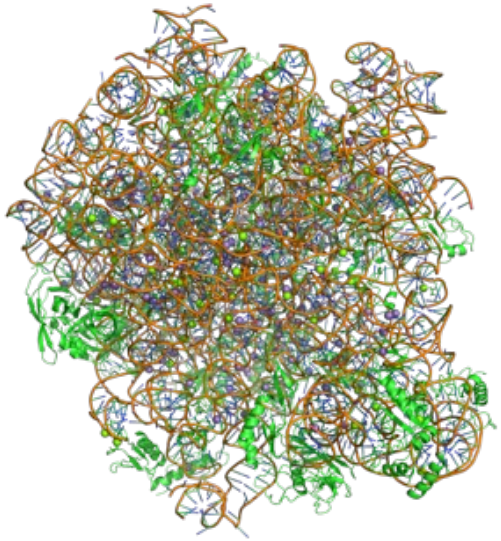


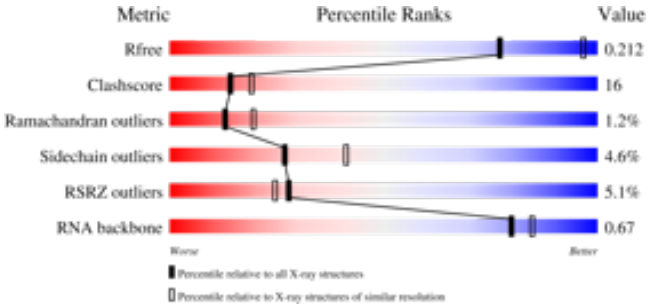

```
*********************************************************
*********************************   SUMMARY   ***********
*********************************************************

 ==> Unidentified chains; check input sequences and model-to-map fit

I/2:144

 ==> Sequence register shifts


 - nucleic-acid chain fragment J/182-211 may be shifted by -1 residue [p-value=9.28e-02]
   model seq 182-211
       actagGGAGTAATCCCCTTGGCGGTTAAAACGCGGgggacag
     new seq 183-212
       actaggGAGTAATCCCCTTGGCGGTTAAAACGCGGgggacag


Time elapsed 0:00:19
```

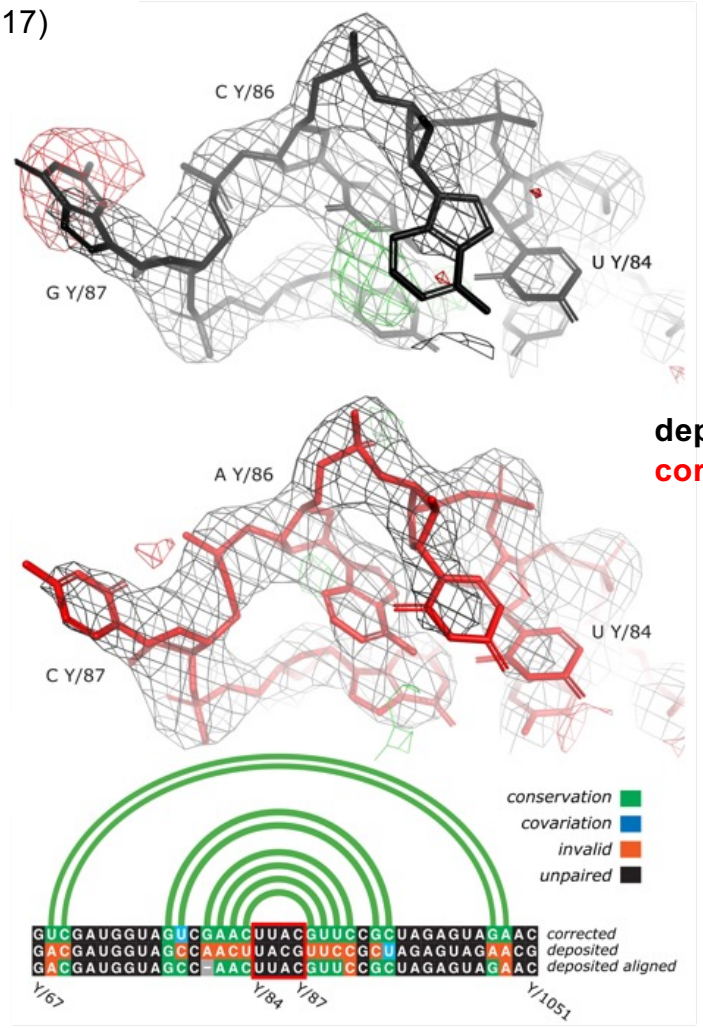
EMBL

Sequence assignment issues in rRNA

Crystal structure model of bacterial ribosome @3.5Å (deposited 2017)







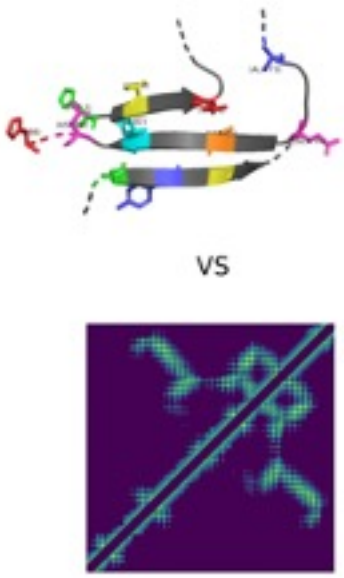
deposited
corrected

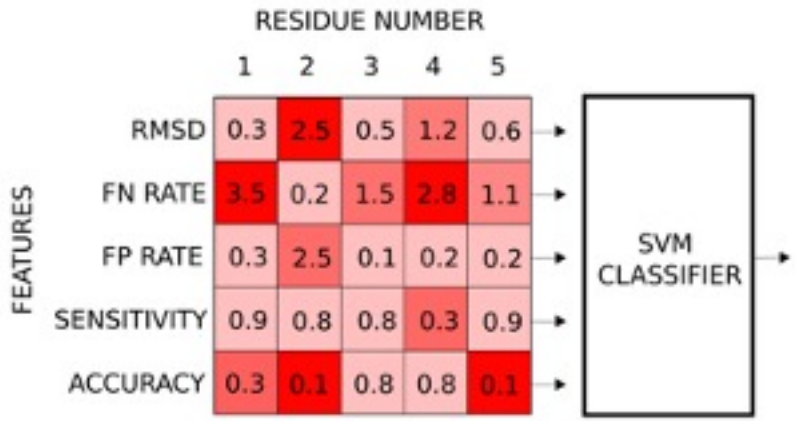
Chojnowski NAR, 2023

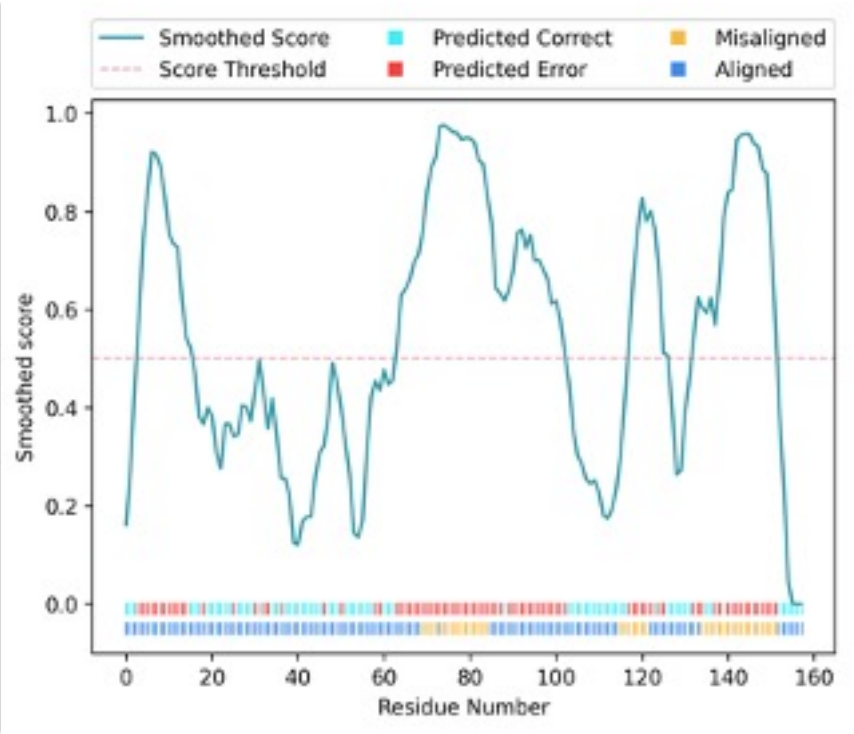
EMBL

Predicted inter-residue distances for model validation

www.conkit.org







Sanchez Rodriguez, F., Chojnowski, G., Keegan, R. M. & Rigden, D. J. (2022). Acta Cryst. D78, 1412-1427.

Acknowledgements

University of Liverpool
Daniel Rigden
Adam Simpkin
Filomeno Sánchez Rodríguez

CCP4 Core Team
Ronan Keegan
Charles Ballard
Eugene Krissinel
Ville Uski
Kyle Stevenson
Maria Fando

CCPEM team
Agnel Joseph
Tom Burnley
Colin Palmer
Matt Iadanza

Martin Luther University
Panos Kastritis
Ioannis Skalidis

EMBL Hamburg
Matthias Wilmanns
Kate Beckham
Jan Kosiński
Christina Ritter
Edukondalu Mullapudi
Isabel Bento
Alice Bochel
Wolfram Seifert-Davila

São Carlos Institute of Physics
Diego A. Leonardo
Laboratorio de Biolgia Molecular, Peru
Dan E. Vivas-Ruiz

findMySequence
protein sequence identification

checkMySequence
sequence validation

doubleHelix
NA sequence identification