Unknown sequences and hidden errors in macromolecular structure models

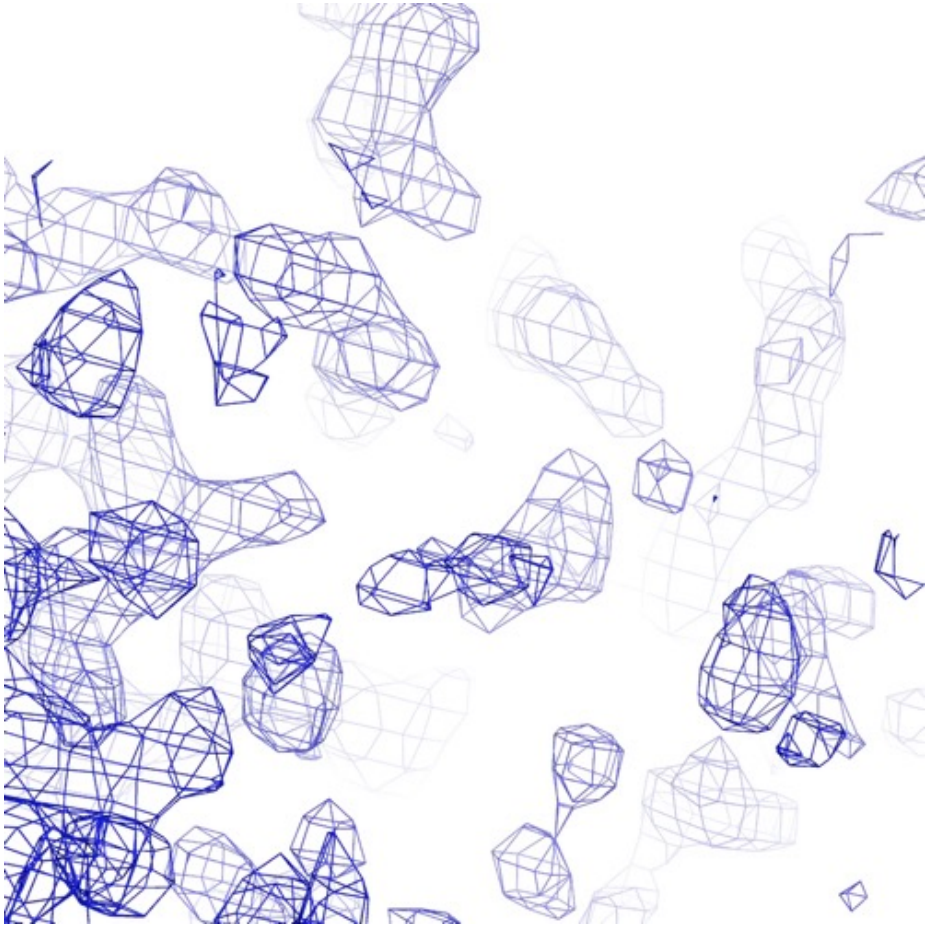
(New developments in model building and validation)

Grzegorz Chojnowski
EMBL Hamburg

EMBL

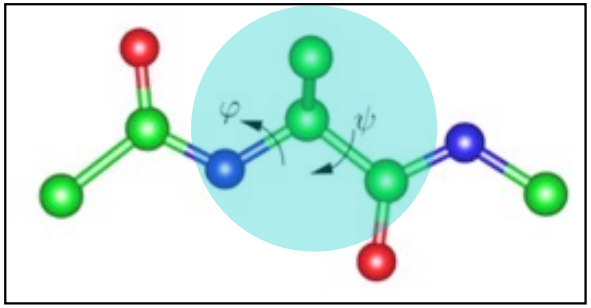
Diamond-CCP4 workshop 05.12.2022

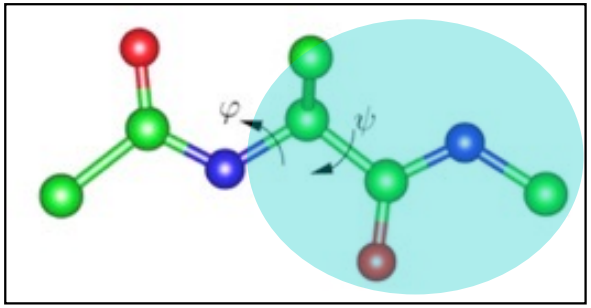
Electron density map interpretation

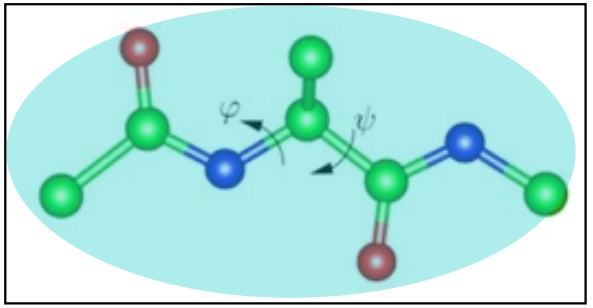


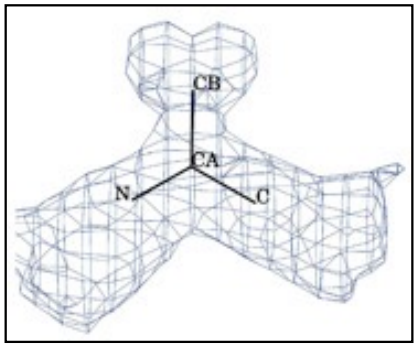
EMBL

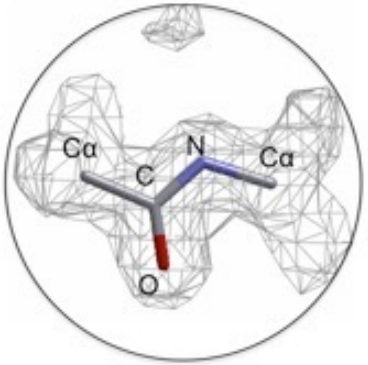
Crystal structure model building

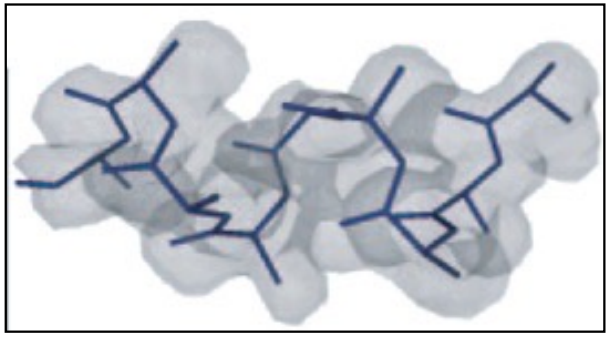












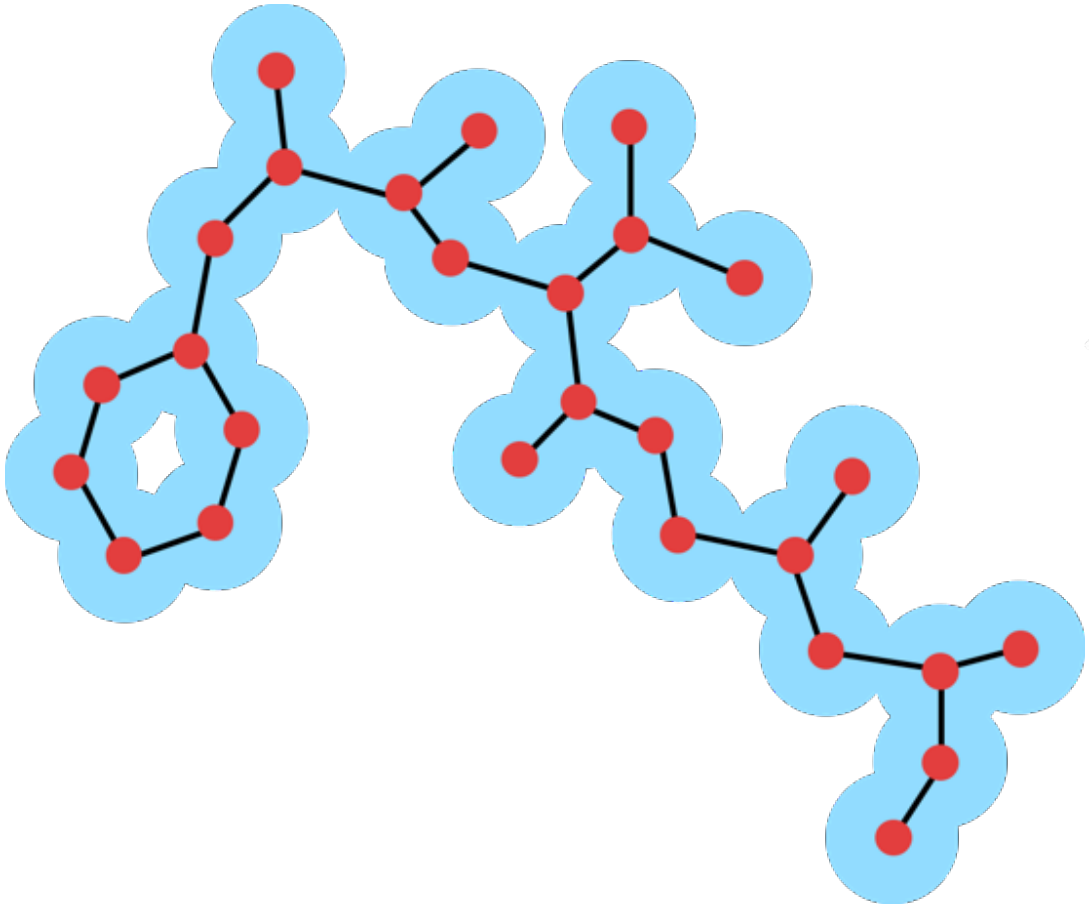
Buccaneer
ModelCraft

ARP/wARP

Phenix.autobuild

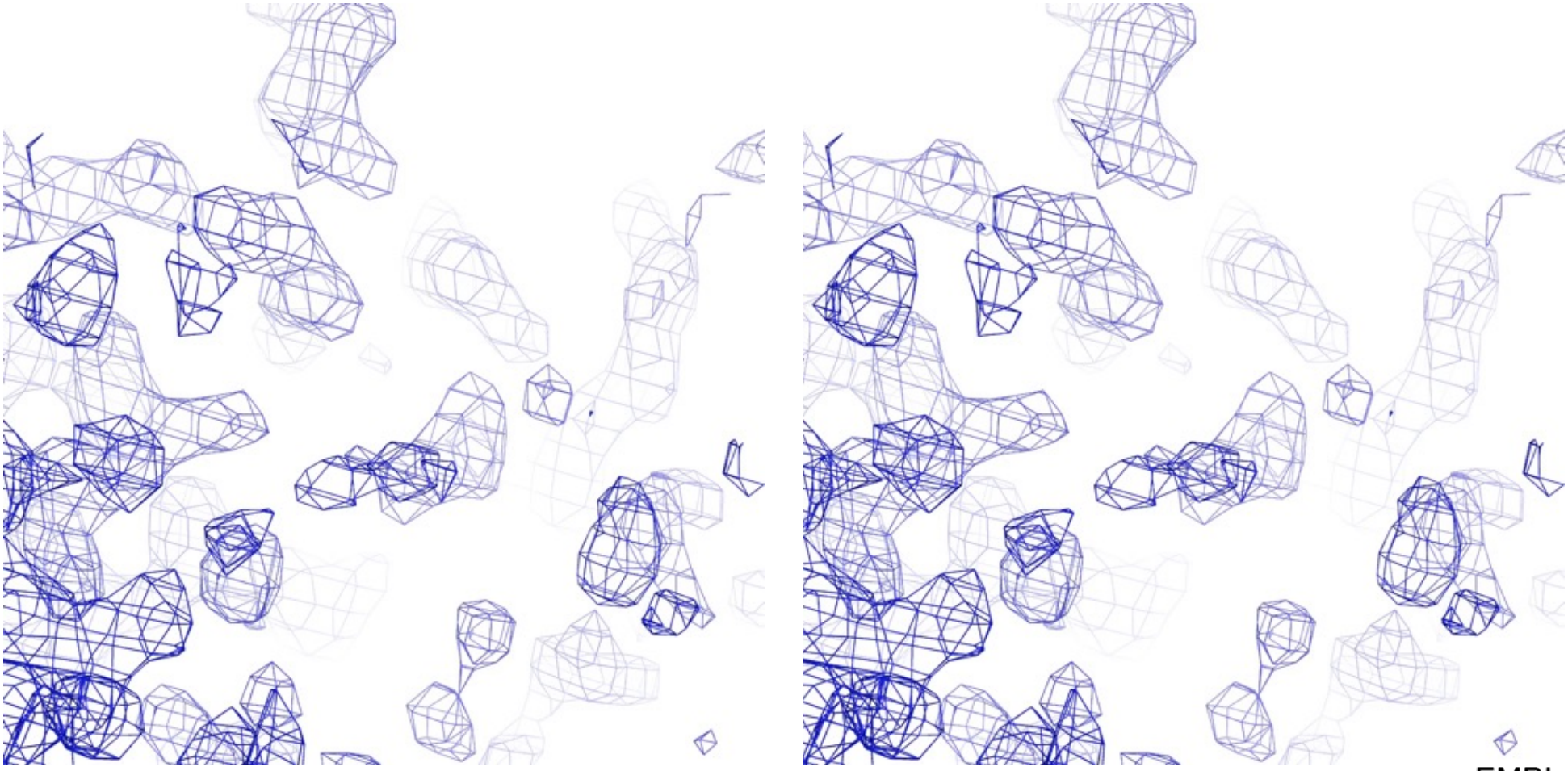
EMBL

Crystal structure model building



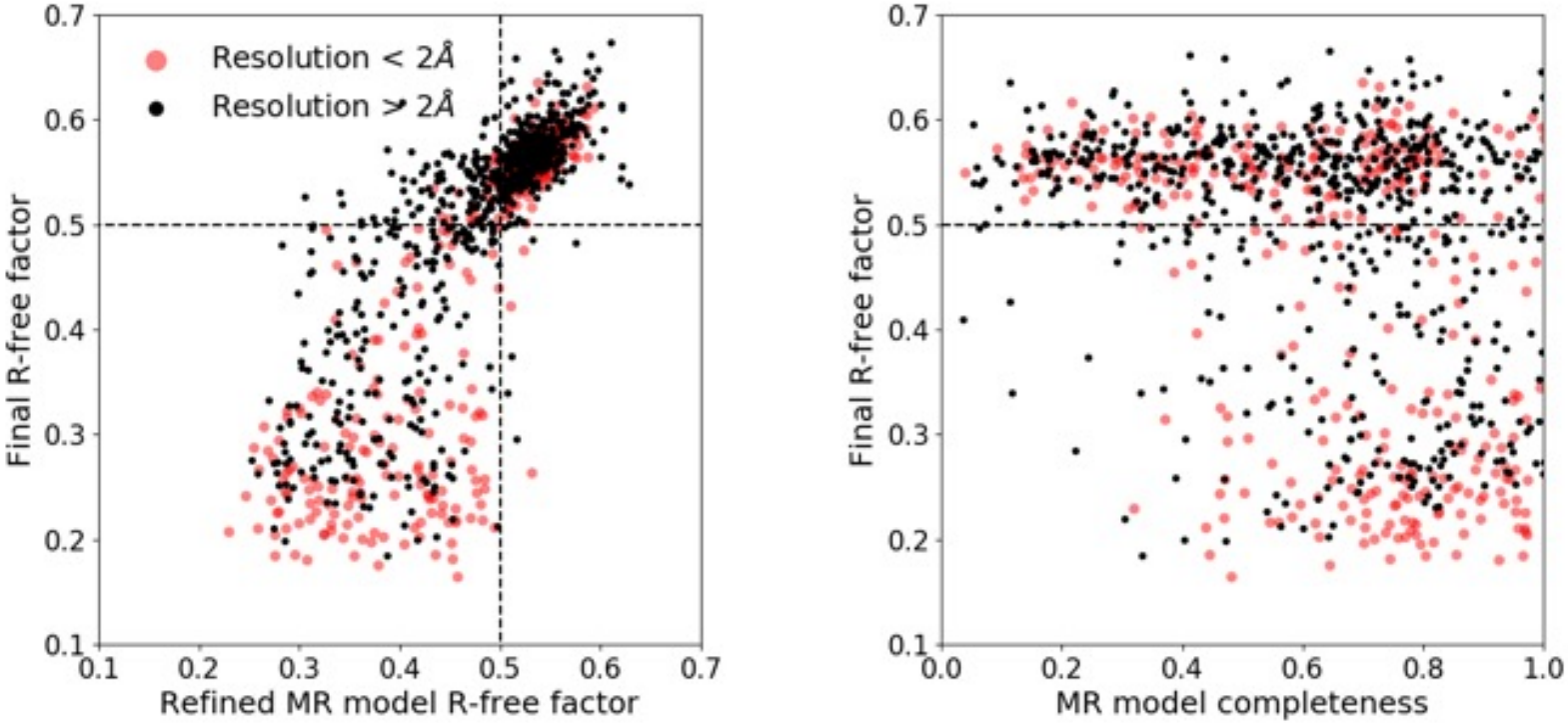
EMBL

Crystal structure model building



EMBL

Crystal structure model building

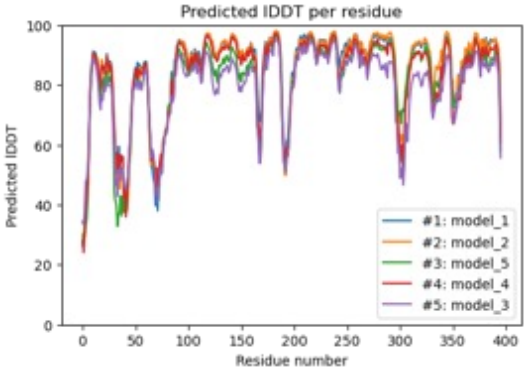


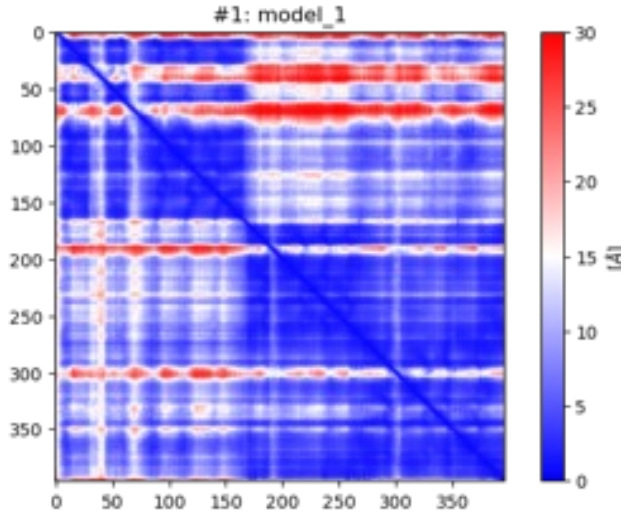
1493 MR solutions submitted to the ARP/wARP web service from automated pipelines (BALBES, MrBUMP, MORDA)

Chojnowski et al. ActaD (2020)

EMBL

Solving crystal structures with AlphaFold2 models



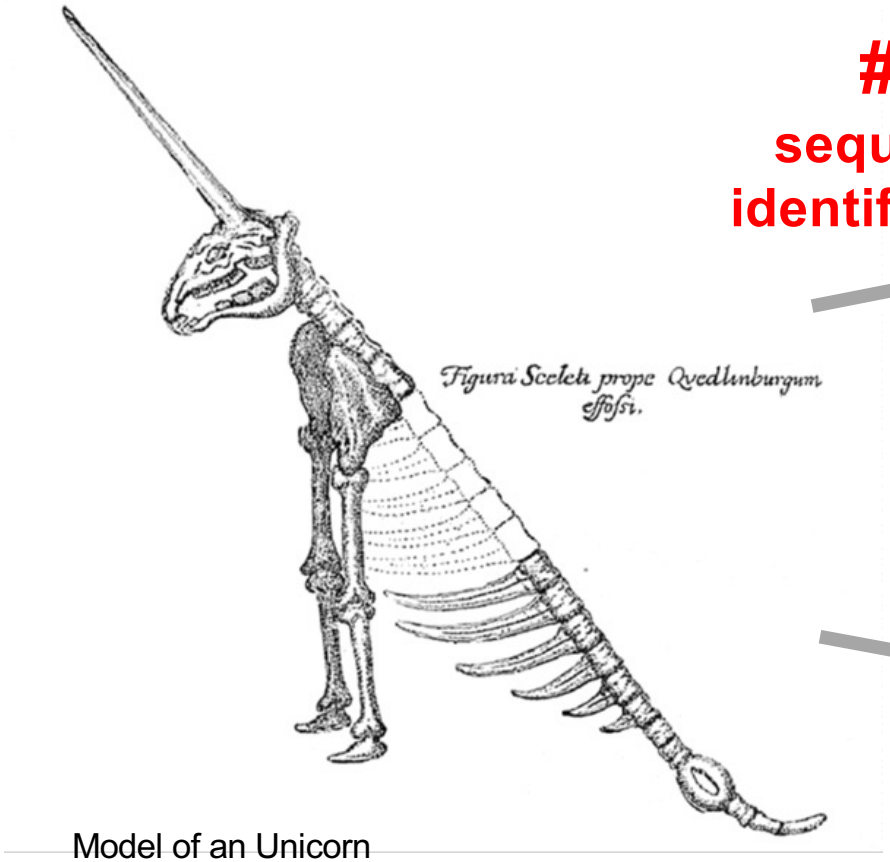


AlpaFold2 prediction

crystal structure model

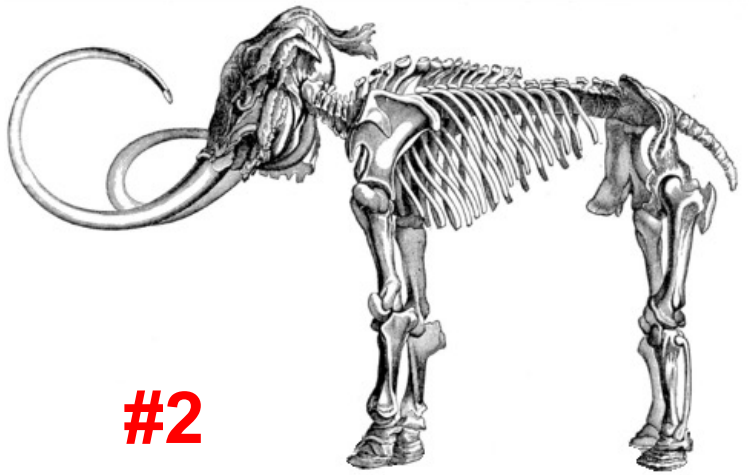
EMBL

Model building traps in MX (and cryo-EM)

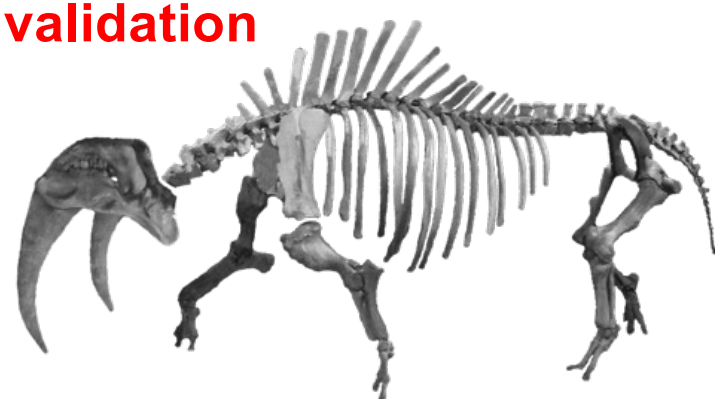


Model of an Unicorn
Gottfried Leibniz after Otto von Guericke, *Protogaea* (1719)

#1
sequence identification

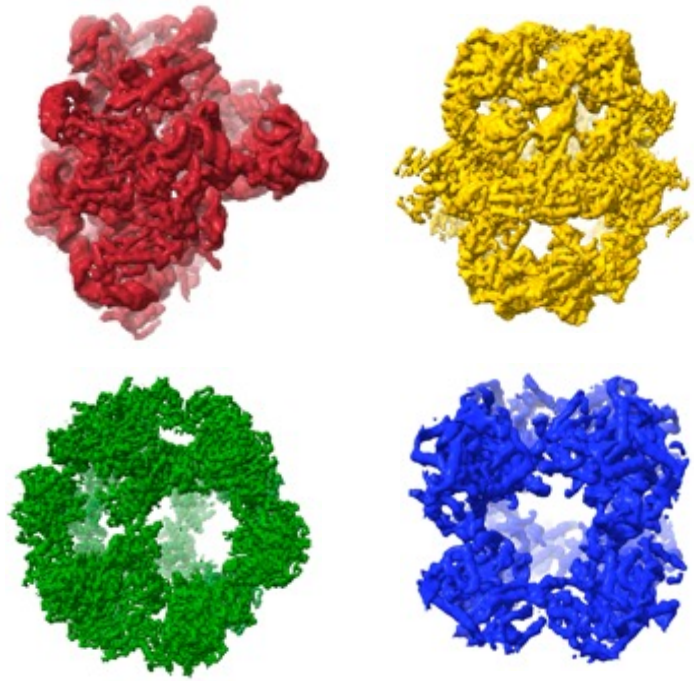


#2
sequence validation

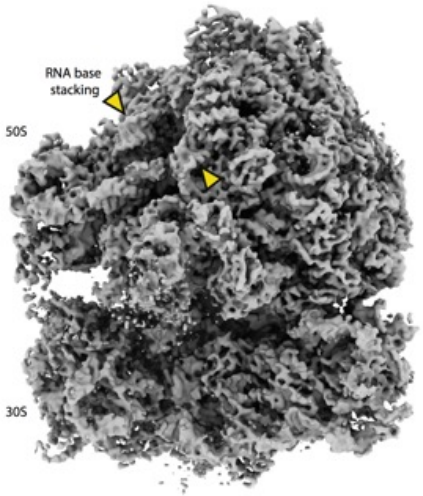


EMBL

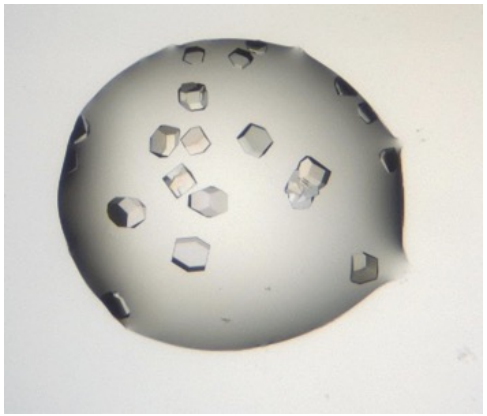
Unknown proteins in EM and MX



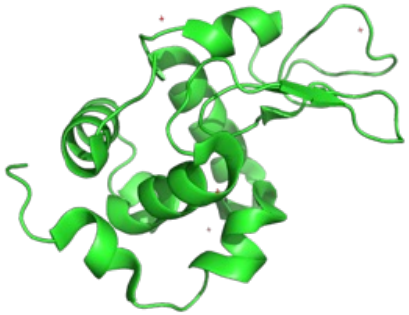
Cryo-EM and artificial intelligence visualize endogenous protein community members
Skalidis et al. Structure 2022



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

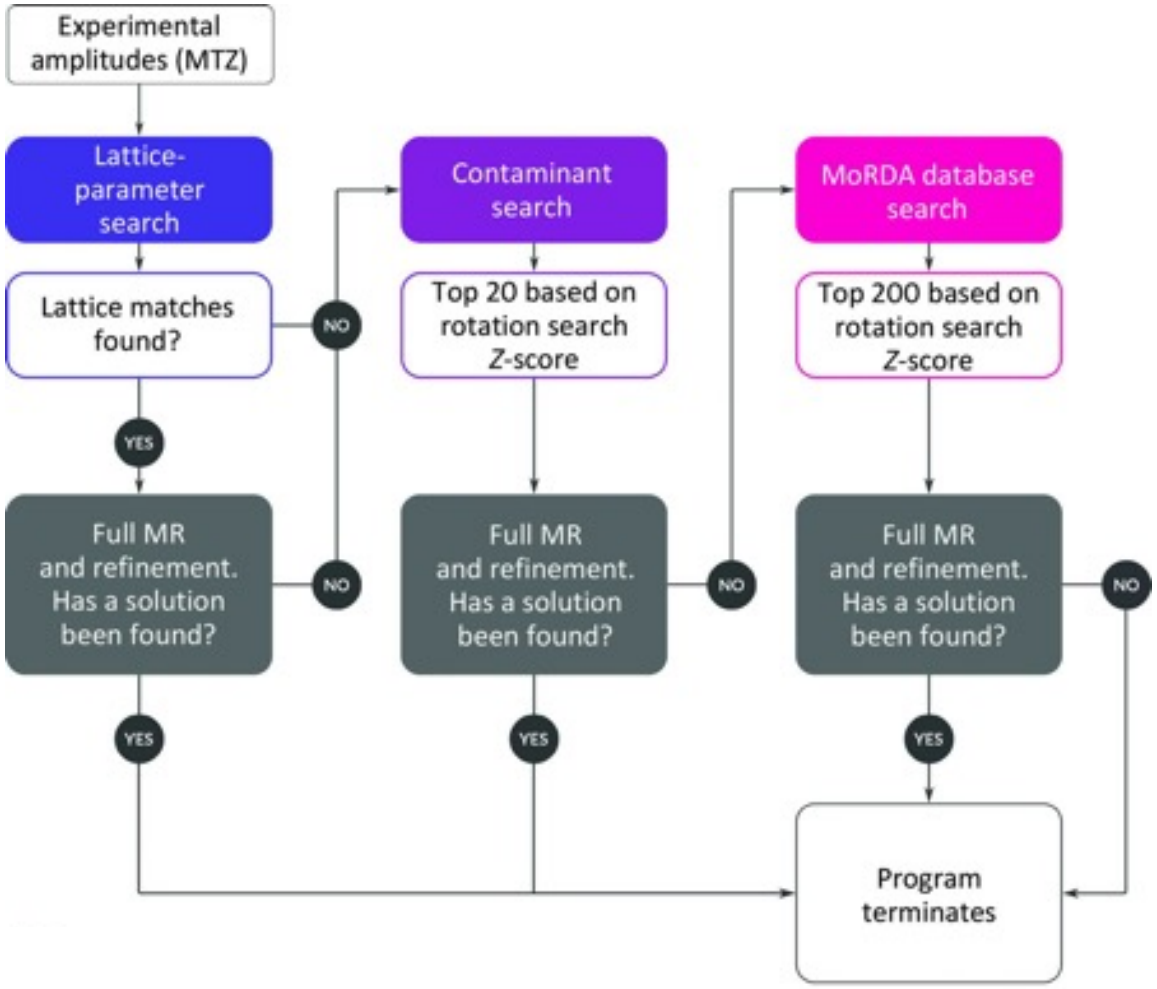


?



EMBL

Sequence-free MR with SIMBAD

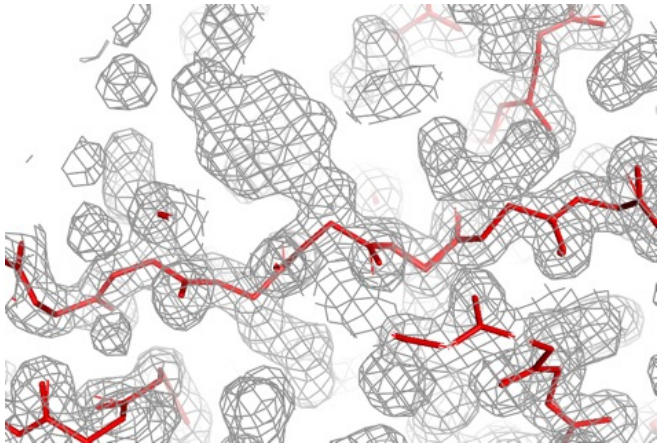


Available in CCP4

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

EMBL

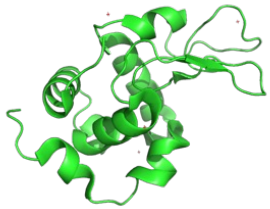
Protein sequence identification from a map



main-chain model and a map

sequence profile

GESAMT, FATCAT, DALI

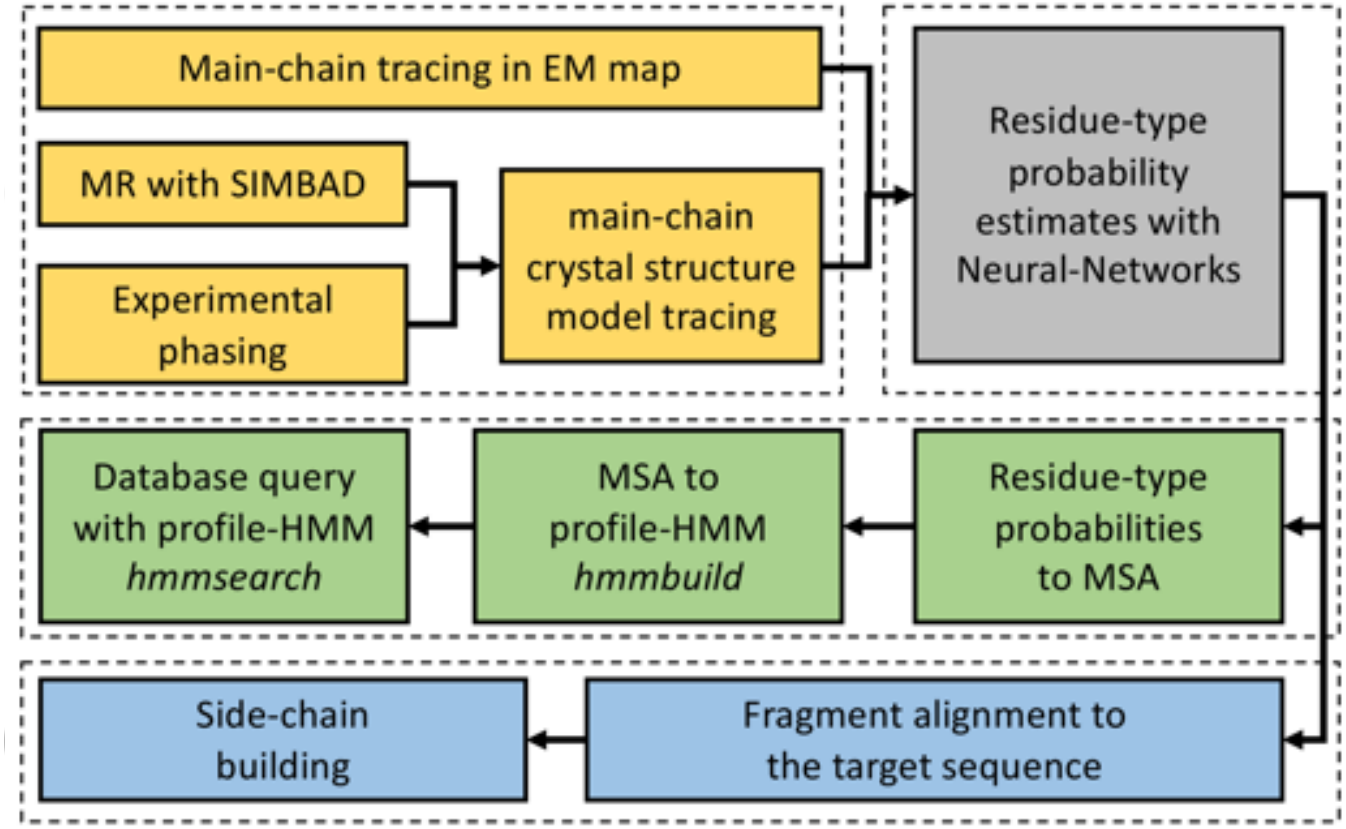


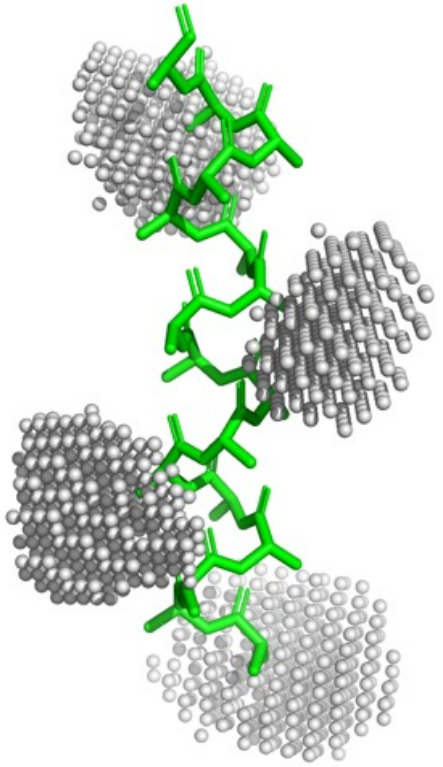
protein structure database

sequence database

EMBL

Protein sequence identification with findMySequence





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

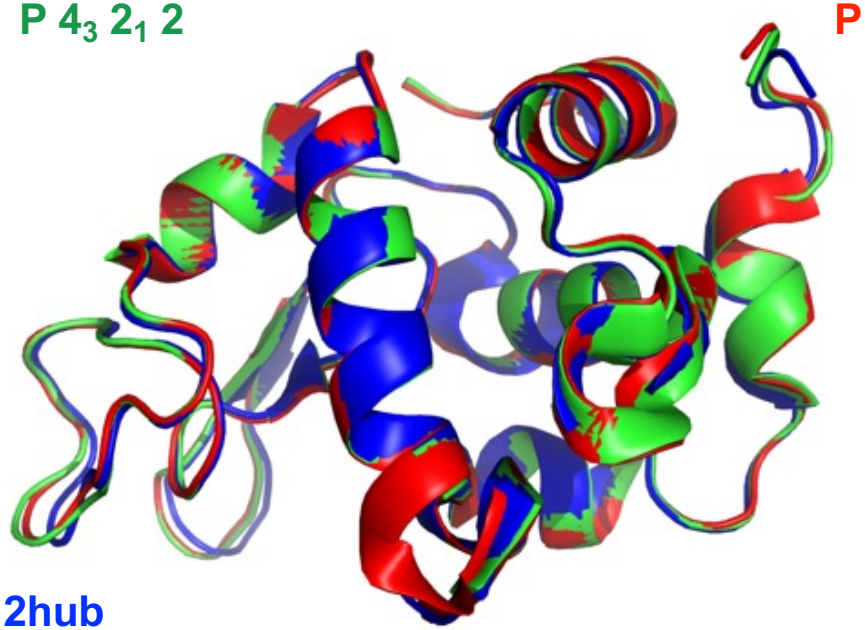
Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX benchmarks: three hen egg-white lysozyme targets

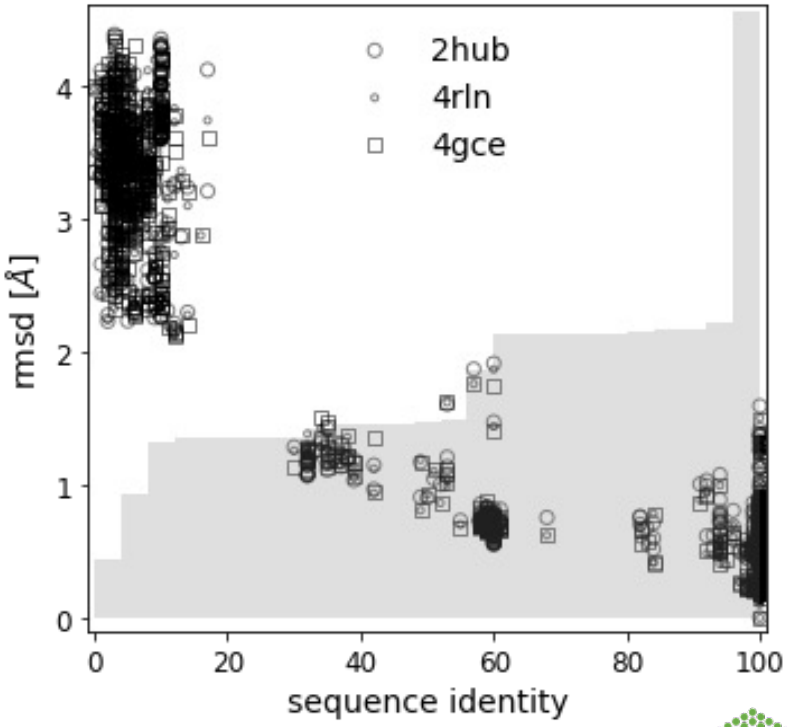
4gce
2.9Å
P 4_3 2_1 2

4rln
2.2Å
P 4_3 2_1 2



2hub
1.2Å
P 4_3 2_1 2

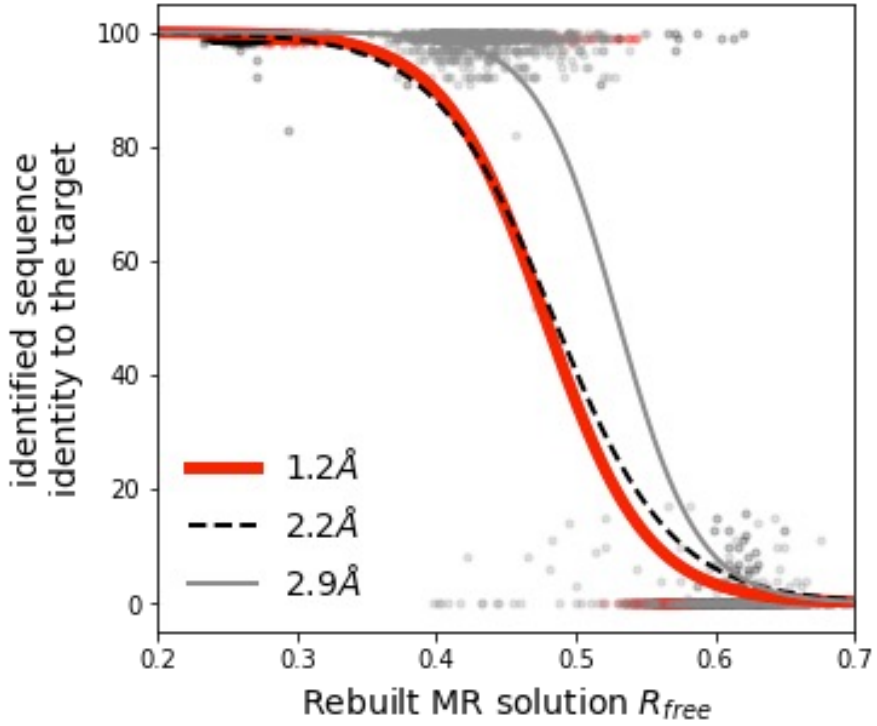
1,500 artificial
MR search models

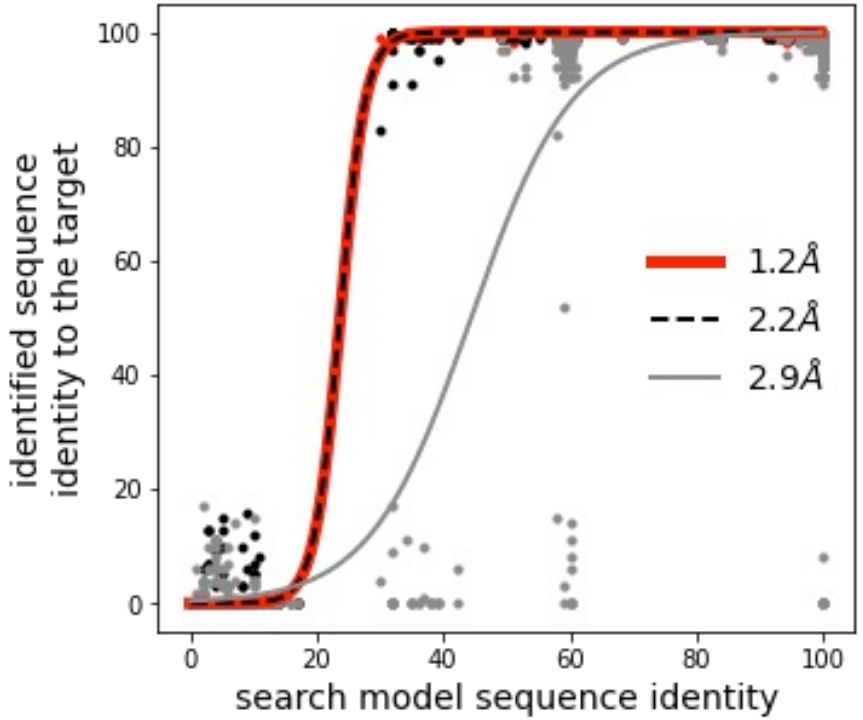


Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX benchmarks: three hen egg-white lysozyme targets

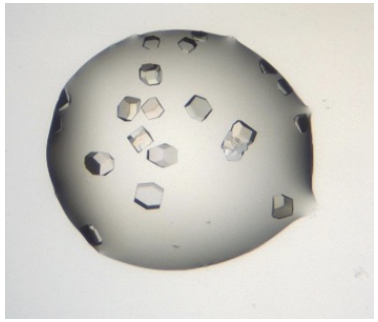


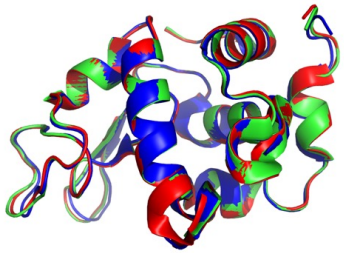


Chojnowski et al IUCrJ 9.1 (2022)

EMBL

MX use-case: *Bothrops atrox* venom proteins



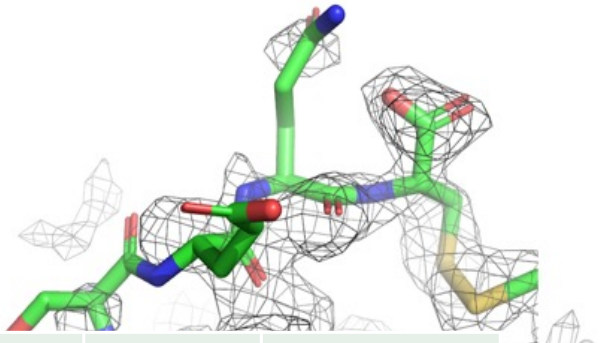


Diego A. Leonardo
Wolfram Seifert-Davila
Dan E. Vivas-Ruiz

EMBL

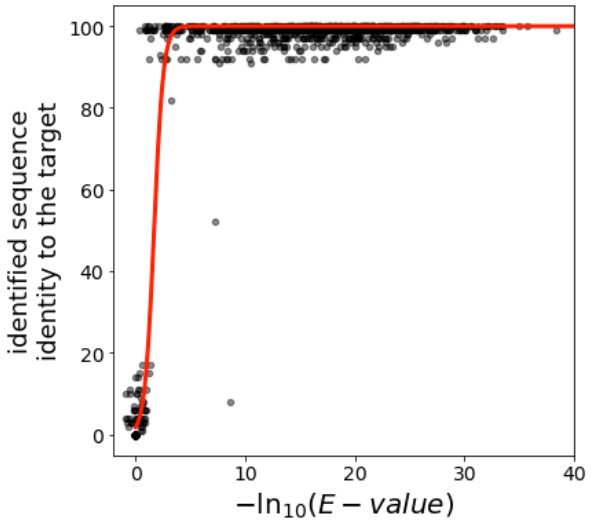
MX use-case: *Bothrops atrox* venom proteins

initial MR solution

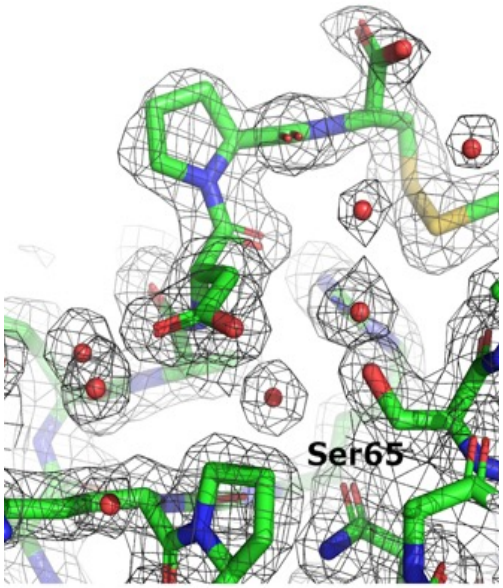


uniprot id	hmmsearch E-value	sequence identity to the top hit [%]
A0A1L8D5Z7	1.40e-28	100
A0A1L8D611	2.10e-23	86
A0A1L8D605	1.00e-17	61
A0A1L8D602	9.30e-16	56
A0A1L8D6C4	3.00e-15	52
A0A1L8D5Z4	4.10e-12	50

model rebuilt
wout sequence



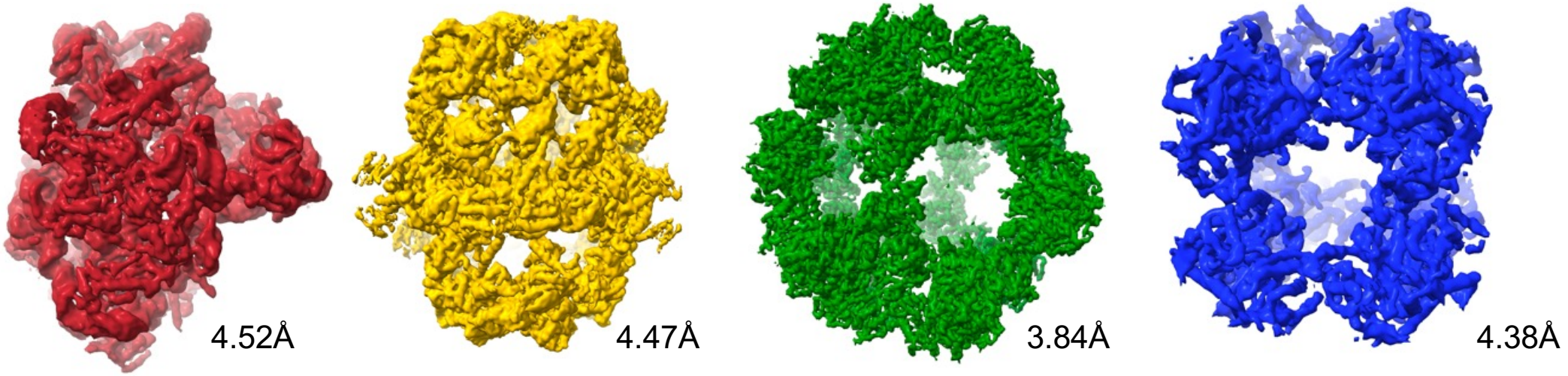
final model



Chojnowski et al IUCrJ 9.1 (2022)

EMBL

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



pre 60S

omokage search
AF2
findMySequence

Fatty Acid Synthase

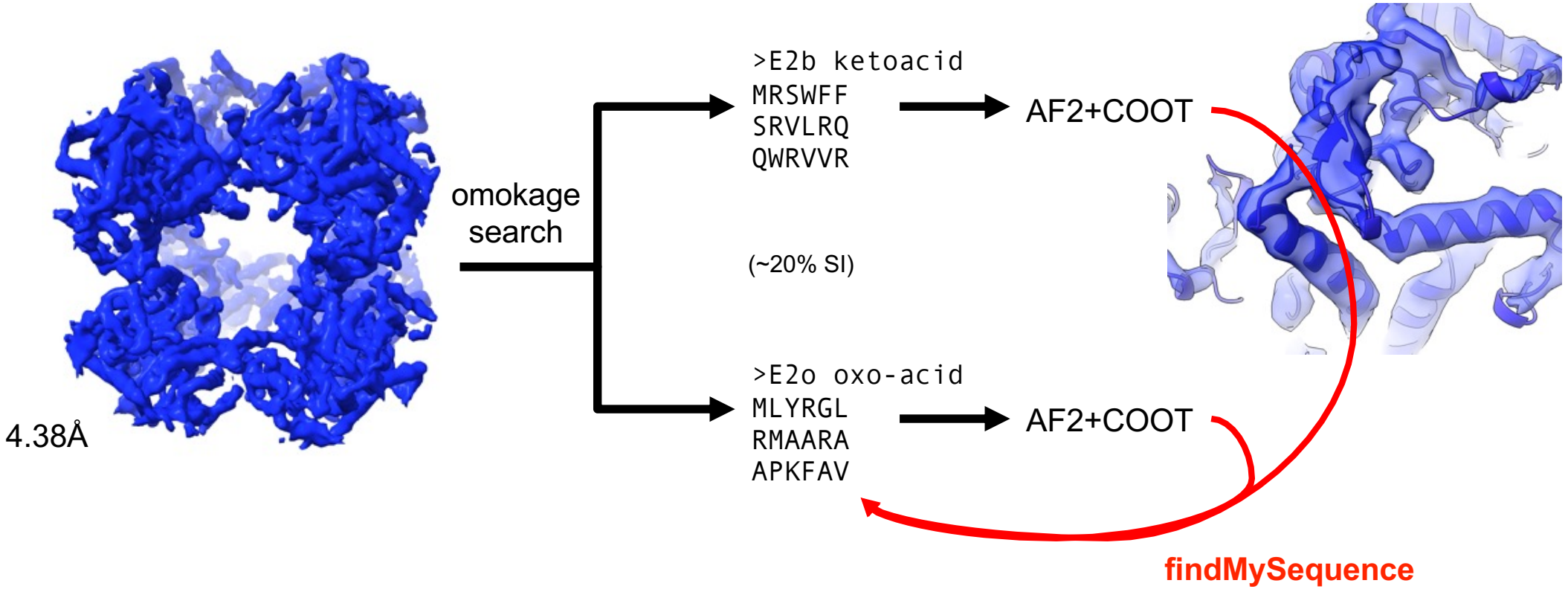
de novo
findMySequence

pyruvate dehydrogenase

with Panos Kastritis and Ioannis Skalidis
Skalidis et al Structure 2022

EMBL

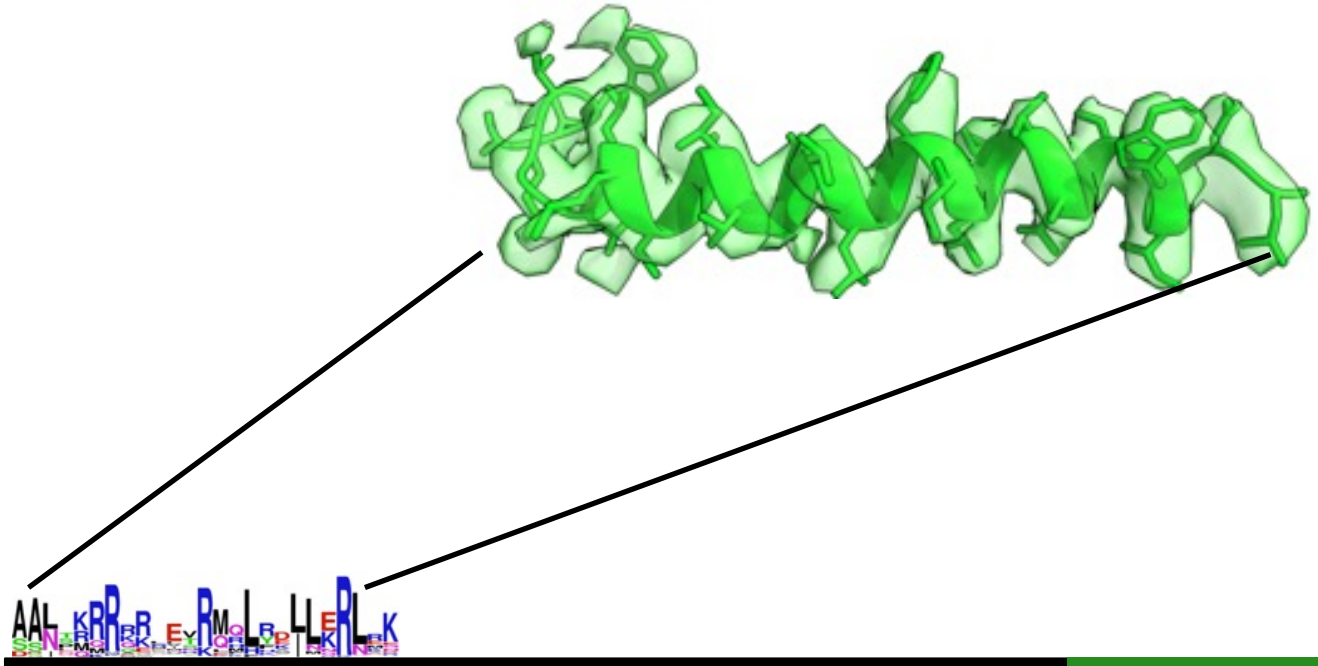
two dehydrogenases in *C. thermophilum* native cell extracts



Chojnowski et al IUCrJ 2022
Skalidis et al Structure 2022

EMBL

Sequence re-assignment as a model validation



target sequence

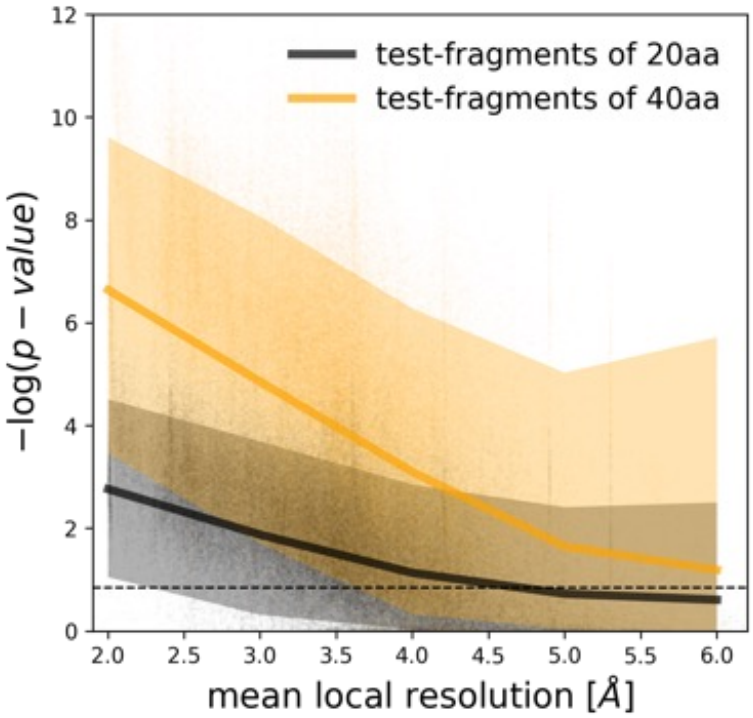
alignment p-value

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

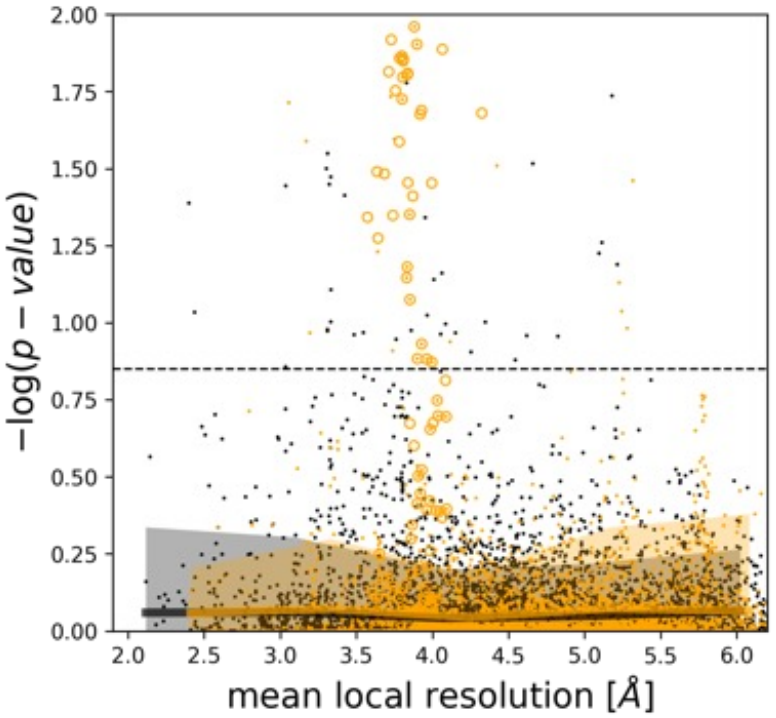
EMBL

Is the sequence assignment p-value a reliable score?

sequence OK



sequences differ



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

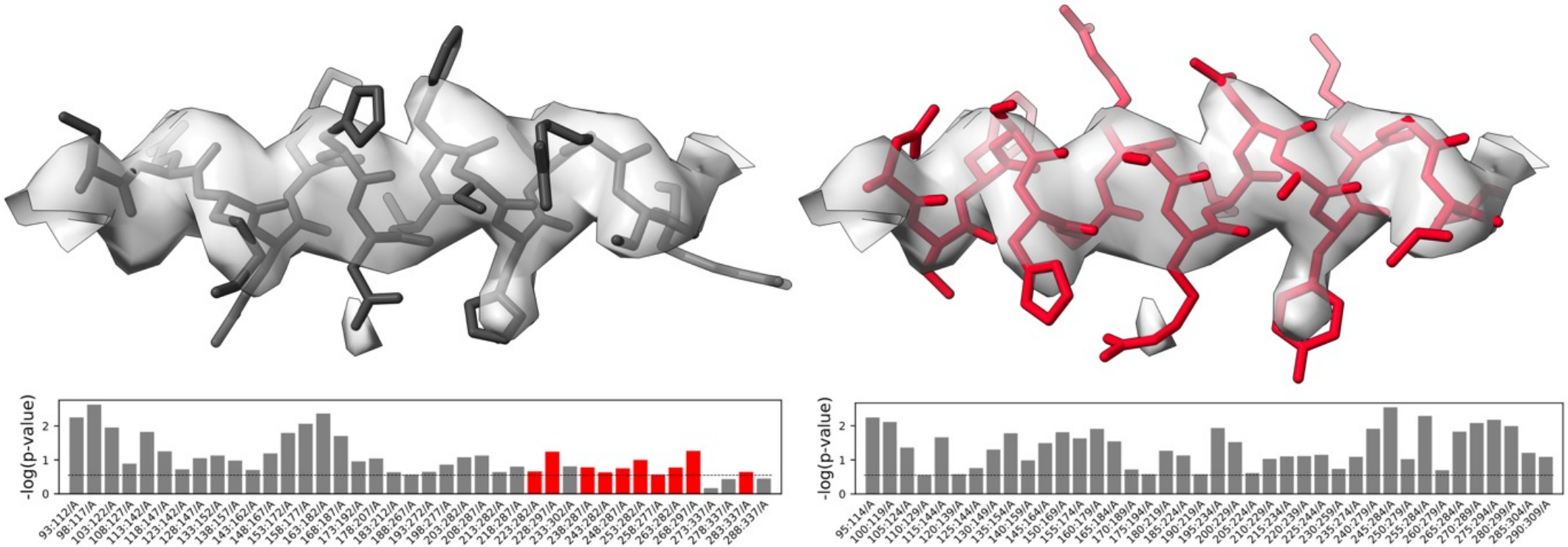
p-values for 30k protein chain fragments re-assigned to target sequence

Chojnowski ActaD 2022

EMBL

hidden errors in EM models: finding a better hypothesis

AlphaFold2 model plddt>90



cytoplasmic domain of a cation channel at 3.8Å resolution

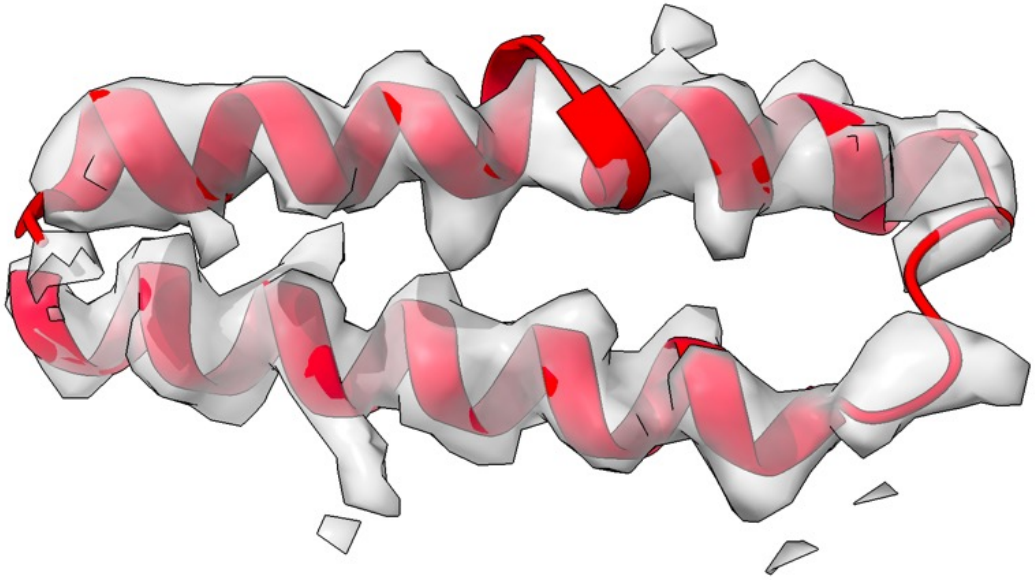
Chojnowski ActaD 2022

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
```

checkMySequence for 6,000 cryo-EM models at 5Å or better

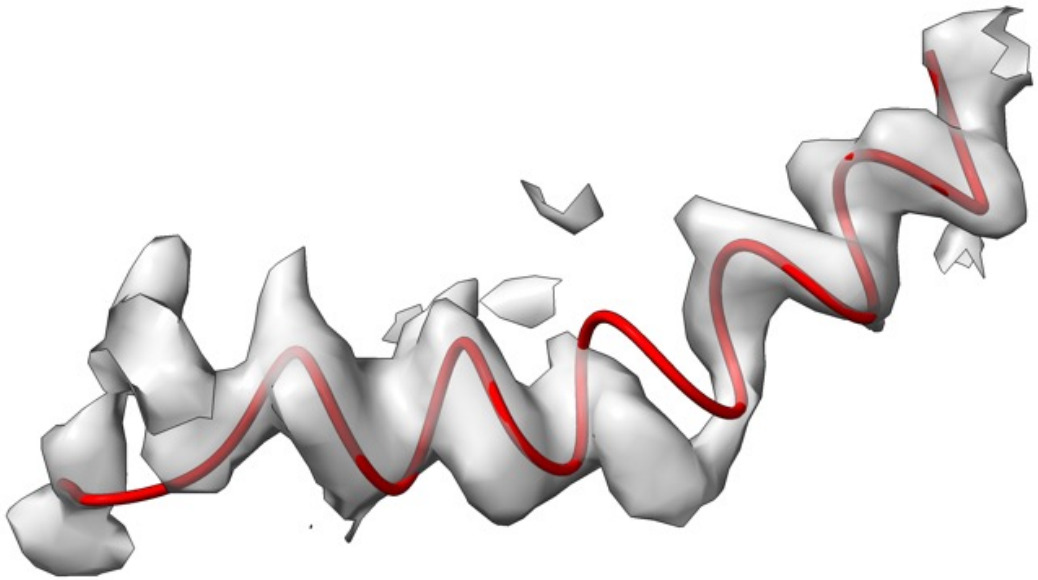


```
- Fragment y/1-20 is shifted by -1 residue [-log(p-value)=2.95]
  model seq 1-20
      MKAKELREKSVEELNTELLNllreqfn
    new seq 2-21
      mKAKELREKSVEELNTELLNLlreqfn
```

70S ribosomal protein 2.9 Å

EMBL

checkMySequence for 6,000 cryo-EM models at 5Å or better

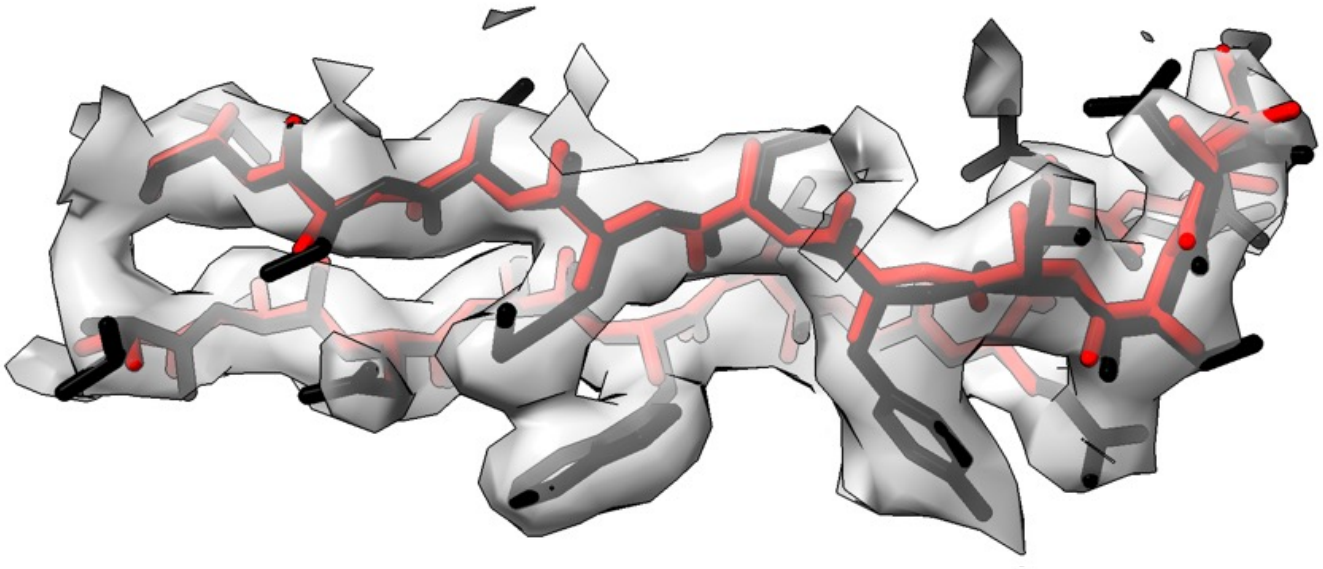


```
- Fragment u/43-62 is shifted by -3 residues [-log(p-value)=2.40]
  model seq 43-62
      ekpttERKRAKASAVKRHAKKLARErnarrt
    new seq 46-65
      ekptterkRAKASAVKRHAKKLARENARrt
```

70S ribosomal protein at 3 Å

EMBL

checkMySequence for 6,000 cryo-EM models at 5Å or better



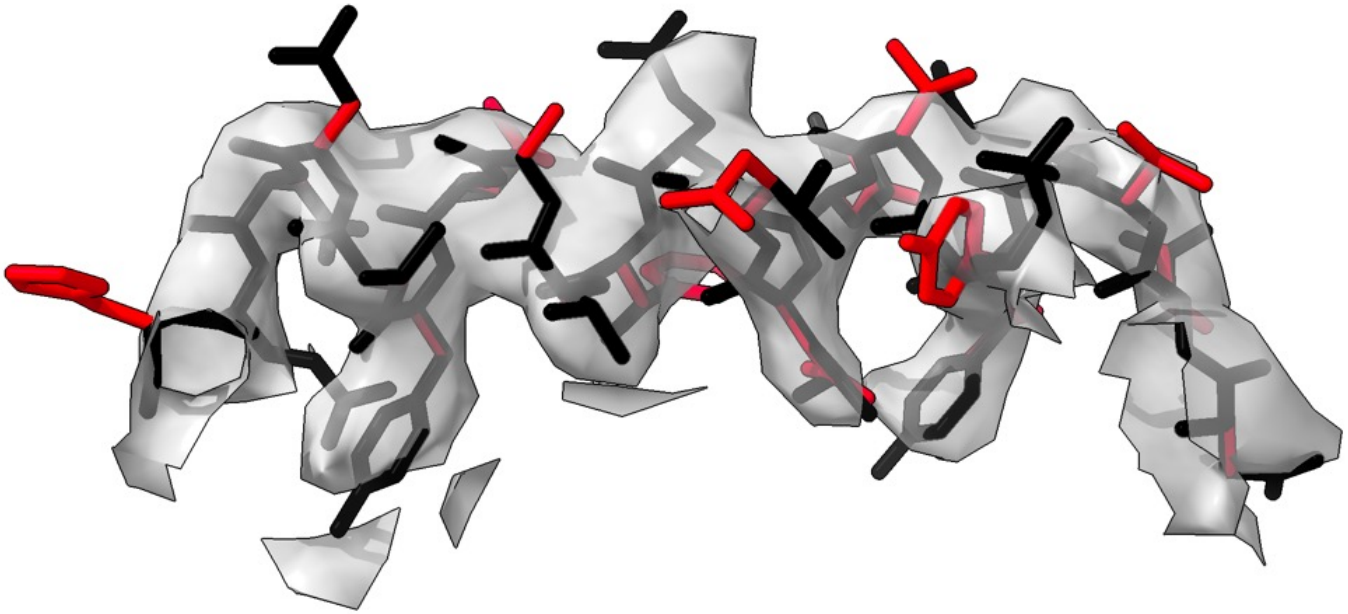
deposited model
rebuilt with findMySequence

```
- Fragment C/161-200 is shifted by 2 residue [-log(p-value)=3.11]
  model seq 161-200
      lnnfypkDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTkdeyer
    new seq 159-198
      lnnfyPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTltkdeyer
```

ferroportin at 3 Å

EMBL

checkMySequence for 6,000 cryo-EM models at 5Å or better



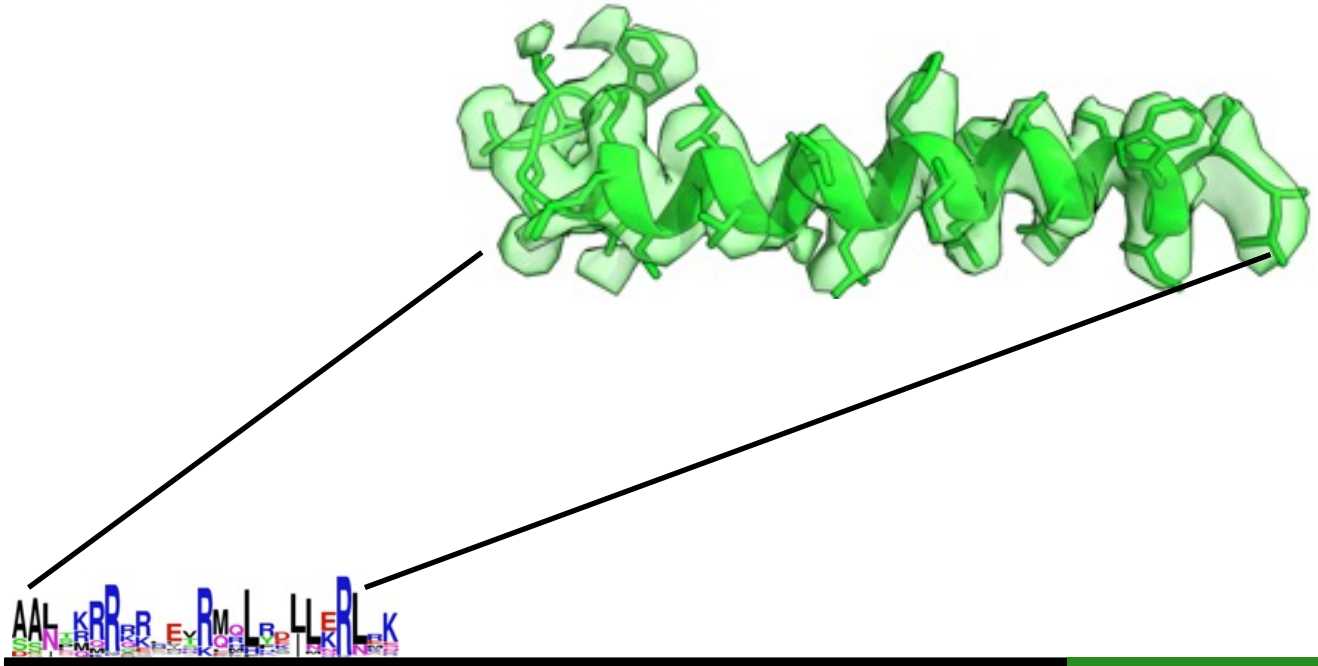
deposited model
rebuilt with findMySequence

```
- Fragment C/145-204 is shifted by 10 residue [-log(p-value)=1.94]
  model seq 145-204
      gfaelarryahnlanARFLWRNRVGAEAVEVRINHIRQGEVARAWRFDALAIGLRDFKADAELDALAELIASGLSgsghvl
    new seq 135-194
      gfaelARRYAHNLANARFLWRNRVGAEAVEVRINHIRQGEVARAWRFDALAIGLRDFKADAELDAlaeliasglsgsghvl
```

CRISPR-associated protein at 3.8 Å

EMBL

checkMySequence and protein crystal structures – model bias



target sequence

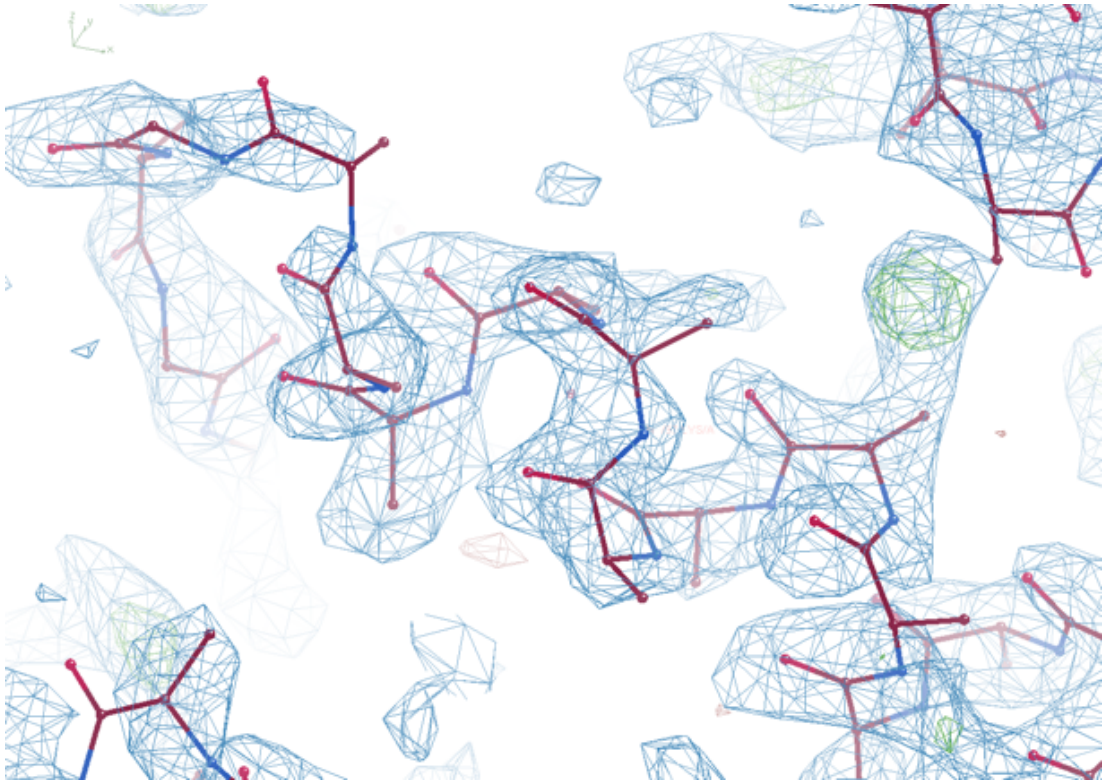
alignment p-value

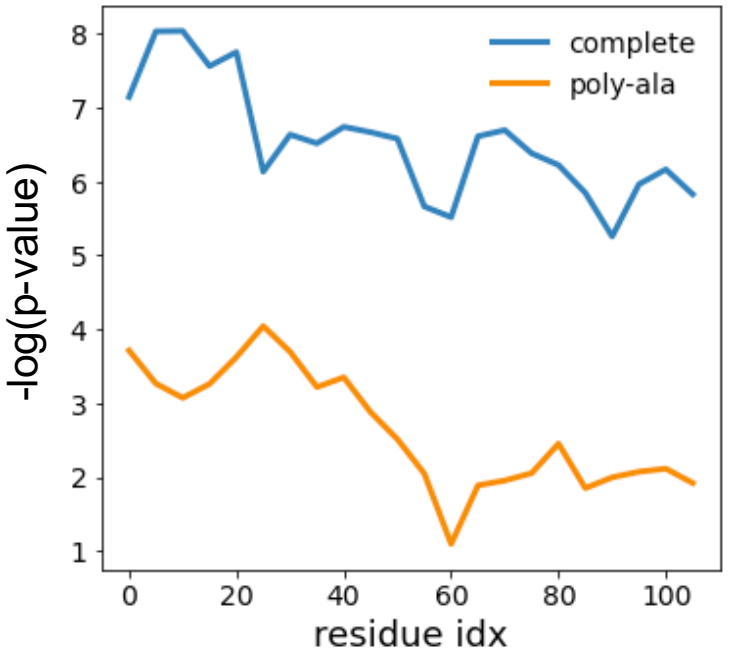
	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

Chojnowski, *in preparation*

EMBL

Sequence assignment in MX – model bias

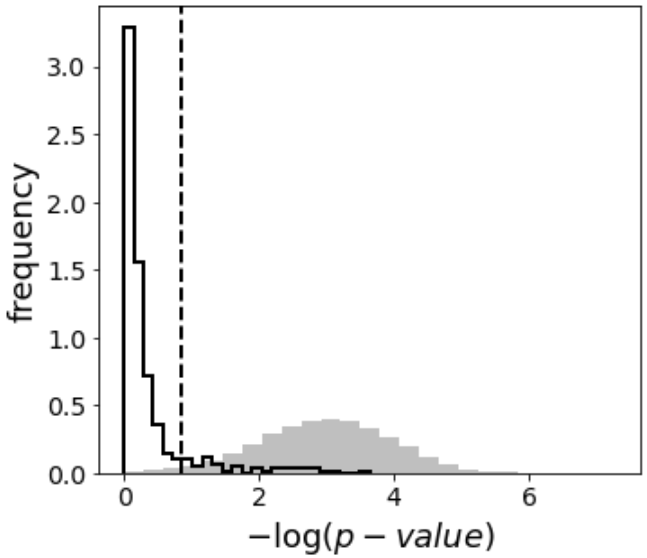


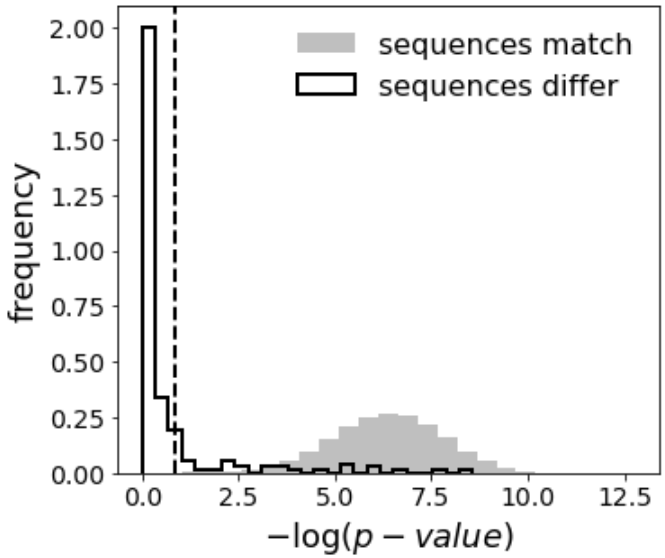


hen egg-white lysozyme @2.9Å (poly-ala + Refmac5, 4gce)

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

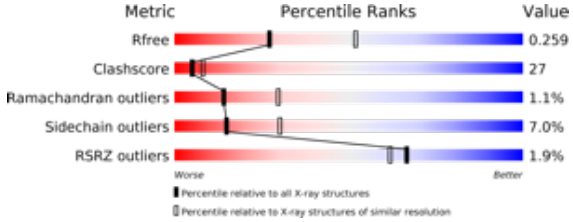
Chojnowski, *in preparation*

EMBL

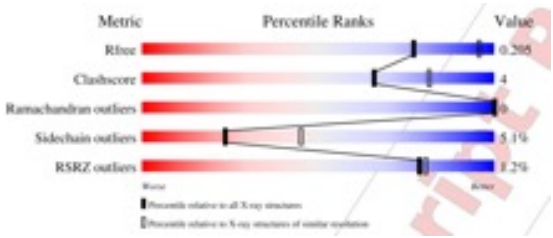
Register shifts in protein crystal structures – example 1

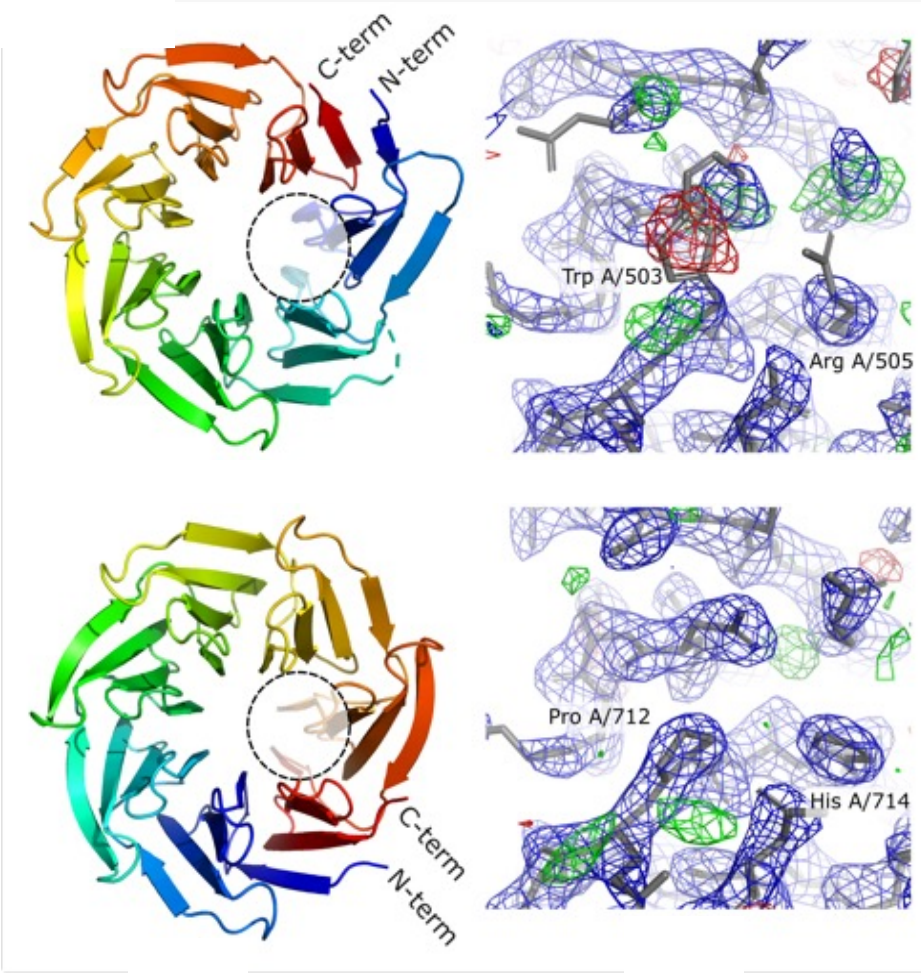
WD40-repeat domain from *T. curvata @2.5Å*

deposited
R/Rfree 22/26



corrected
R/Rfree 17/21



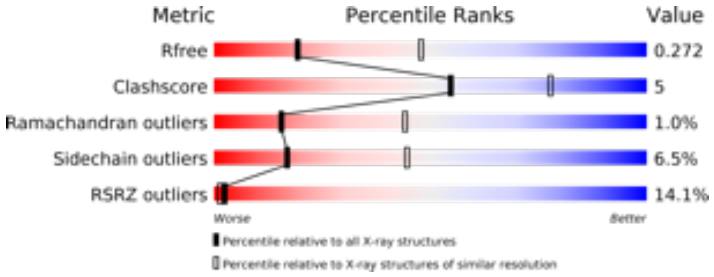


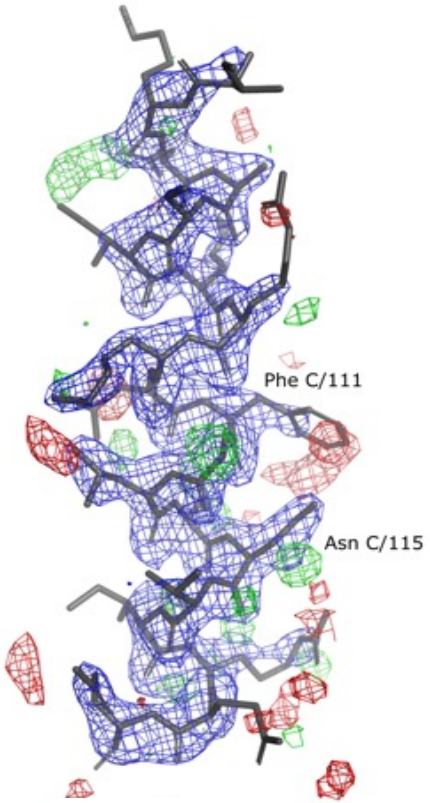
Chojnowski, *in preparation*

EMBL

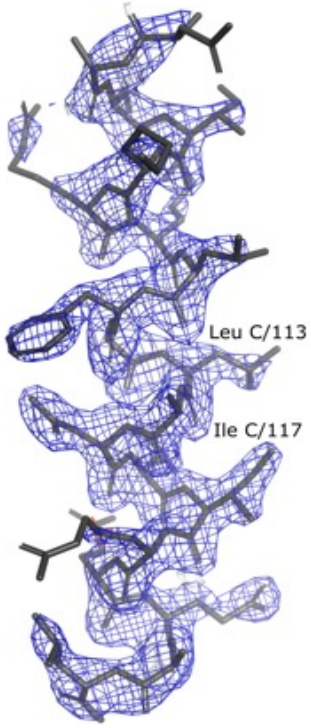
Register shifts in protein crystal structures – example 2

Helicase form *H. pylori @2.5Å*





deposited



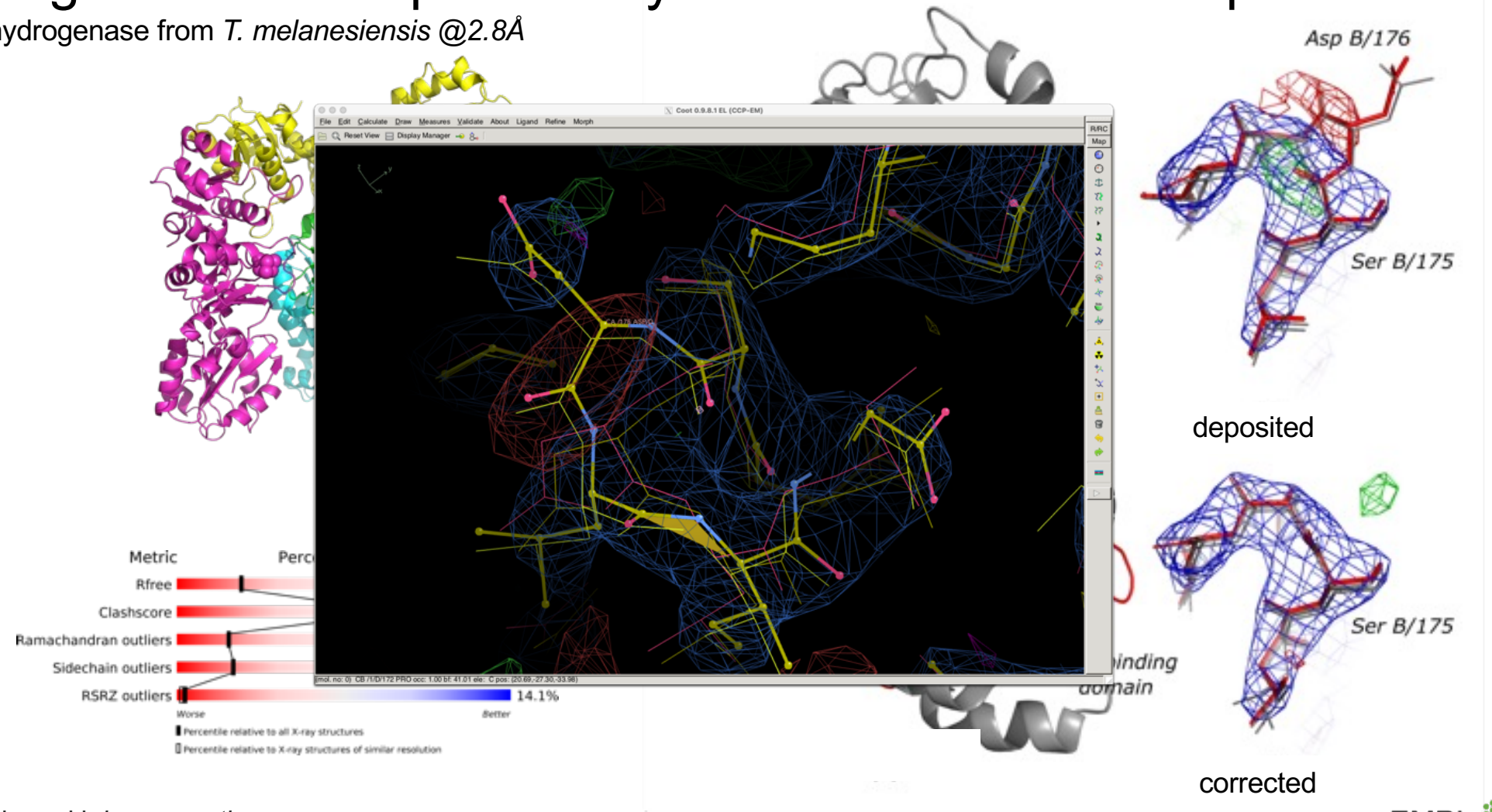
corrected

Chojnowski, *in preparation*

EMBL

Register shifts in protein crystal structures – example 3

hydrogenase from *T. melanesiensis* @2.8Å

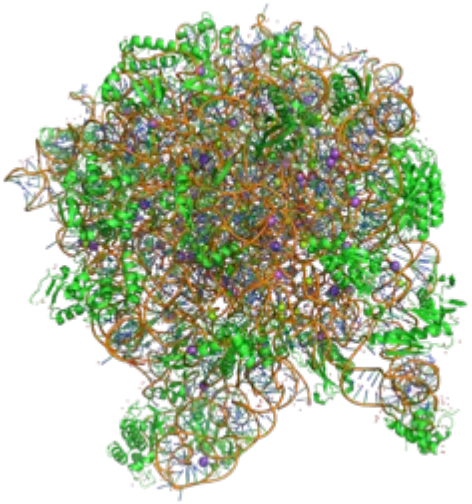


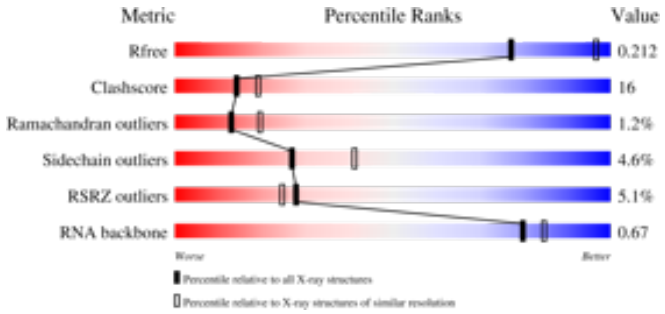
Chojnowski, *in preparation*

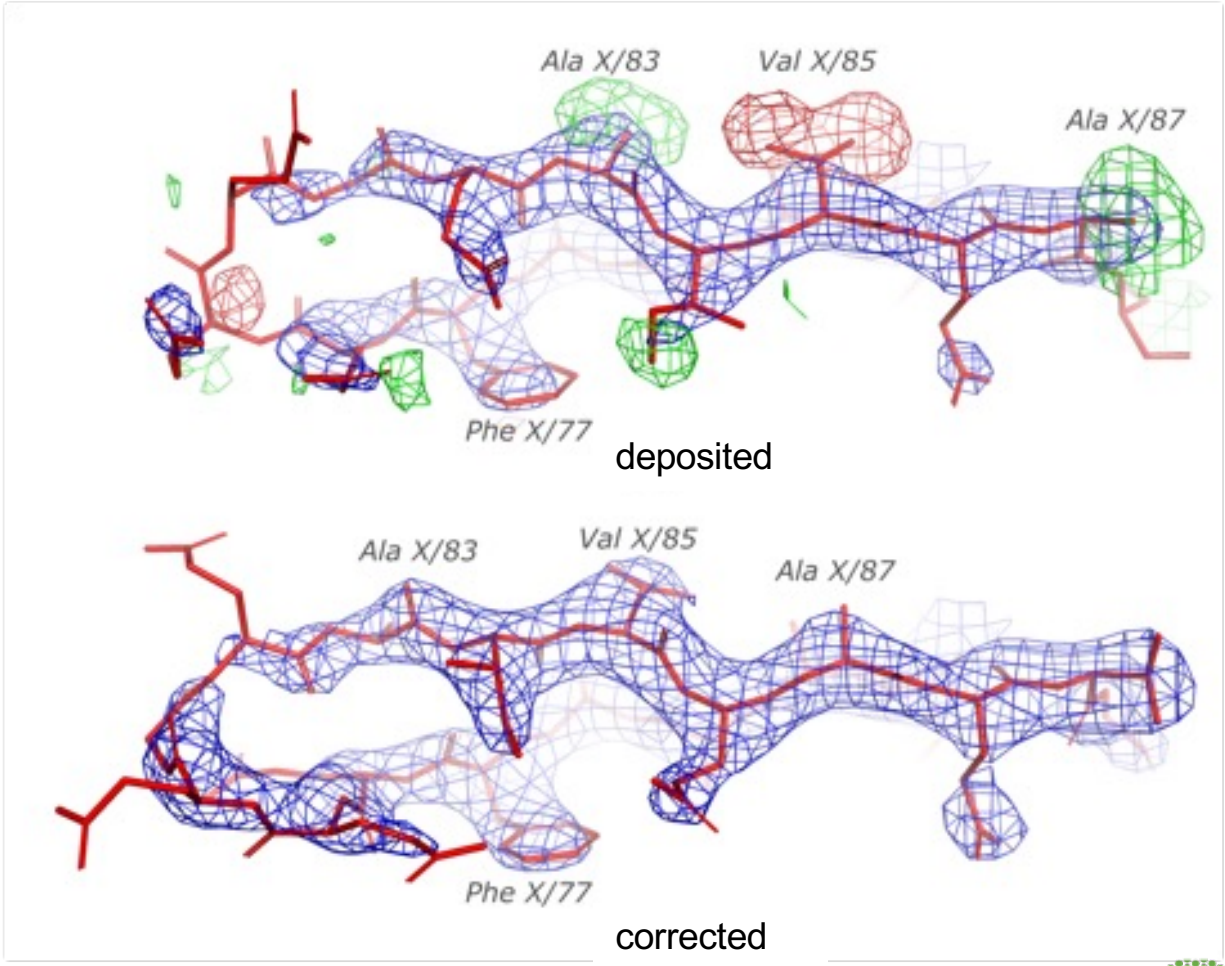
EMBL

Register shifts in protein crystal structures – example 4

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





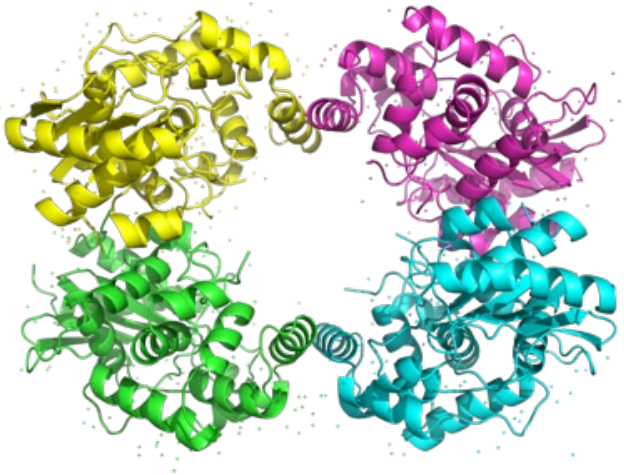


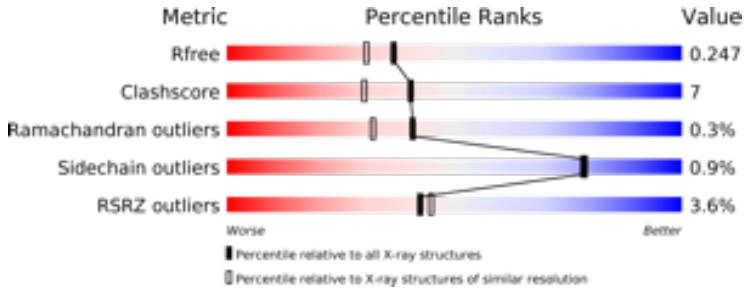
Chojnowski, *in preparation*

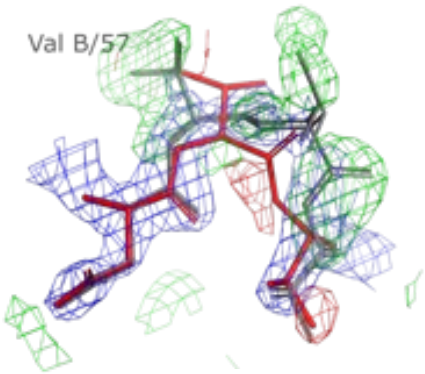
EMBL

Register shifts in protein crystal structures – example 5

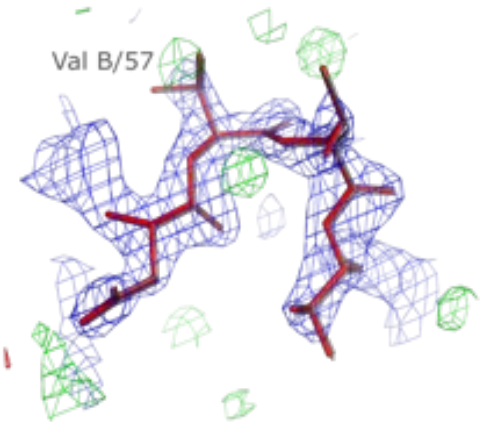
glutaminase from *G. kaustophilus @2.1Å*







deposited



corrected

Chojnowski, *in preparation*

EMBL

Register shifts in protein crystal structures

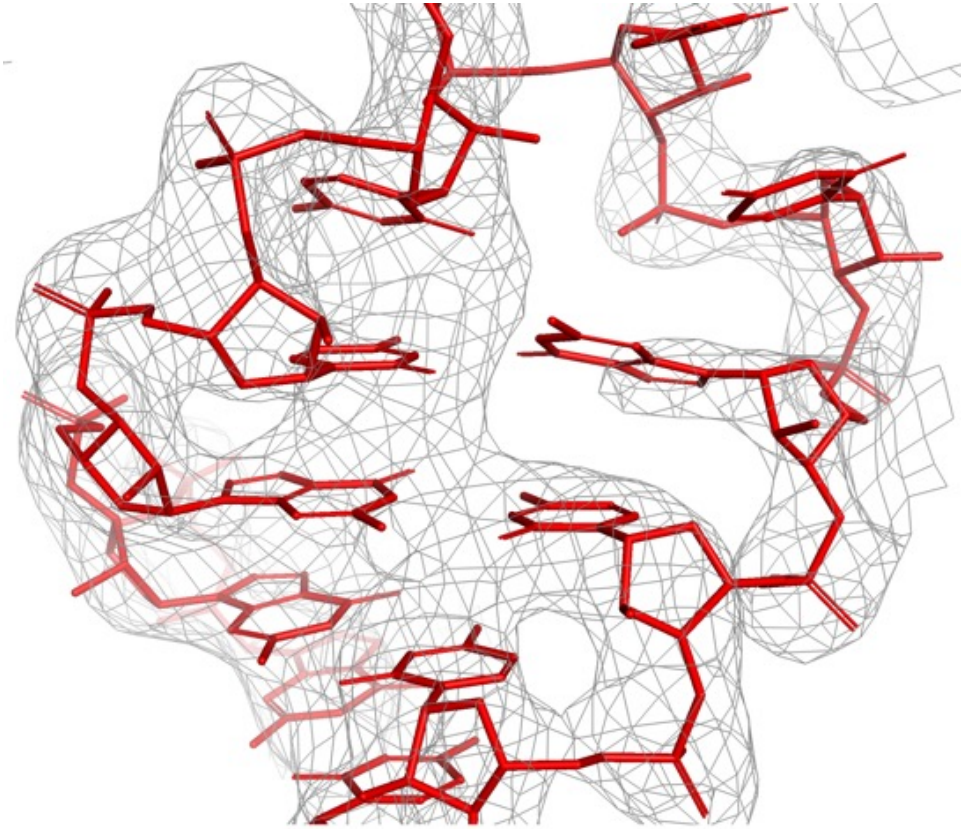
Statistics from a scan of 10,000 deposited protein crystal structures at 2-3Å resolution

- 27 models with indexing issues (residue numbering ignores gaps)
- 89 structures with sequence mismatches
- 115 with poorly resolved chains (at least 1)
- 70 plausible register shifts

Chojnowski, *in preparation*

EMBL

Nucleic acid sequence assignment in EM/MX

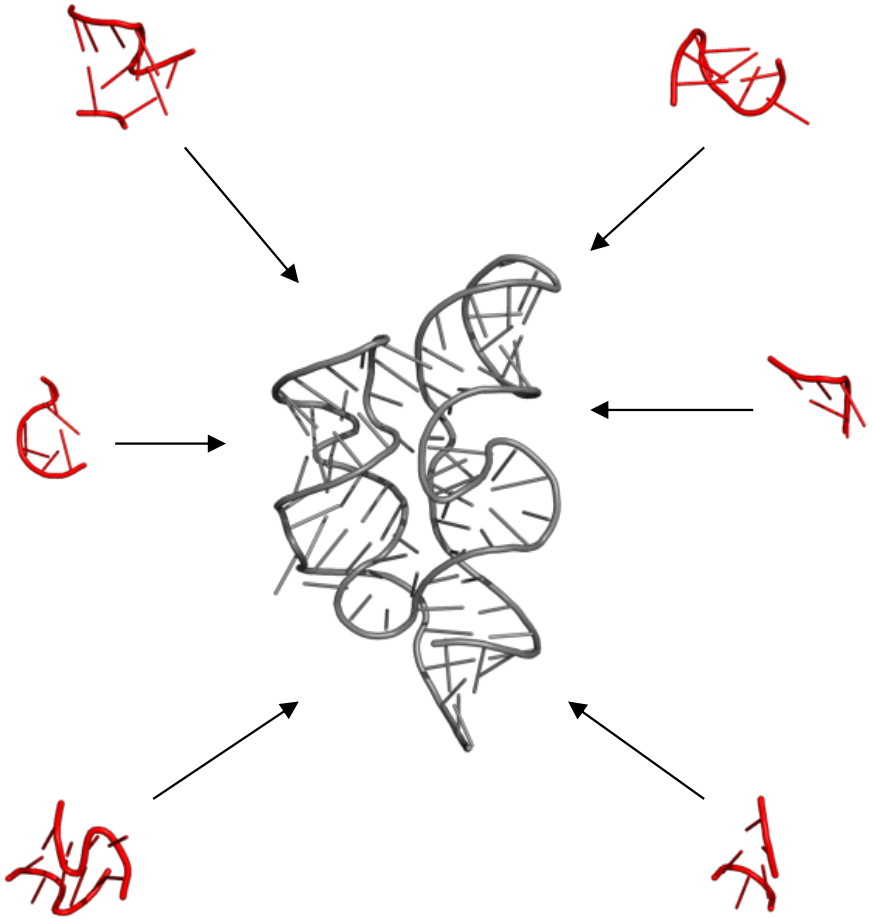


Different types of purines and pyrimidines are (usually) indistinguishable in MX or EM maps

Chojnowski, *in preparation*

EMBL

Base-pairs from backbone geometry



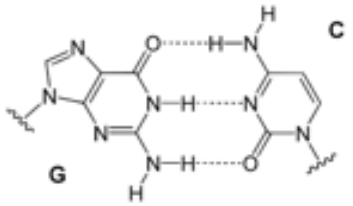
input RNA model backbone is “covered” with motifs with known secondary structure

- base pairs assignment
- sequence validation
- refinement restraints

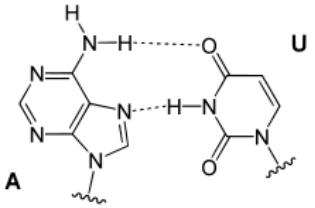
Chojnowski et al. NAR 42 D123-D131 2014
Chojnowski at al. ActaD 71 697-705 2015

EMBL

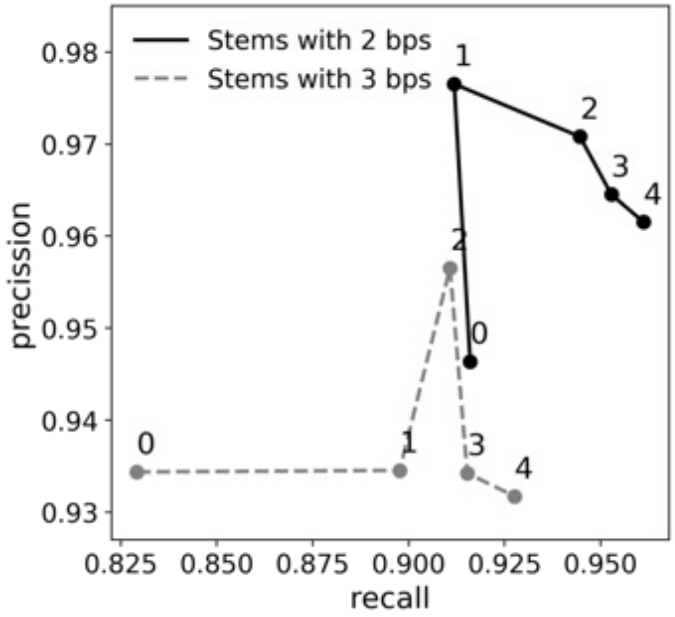
Base-pairs from backbone geometry

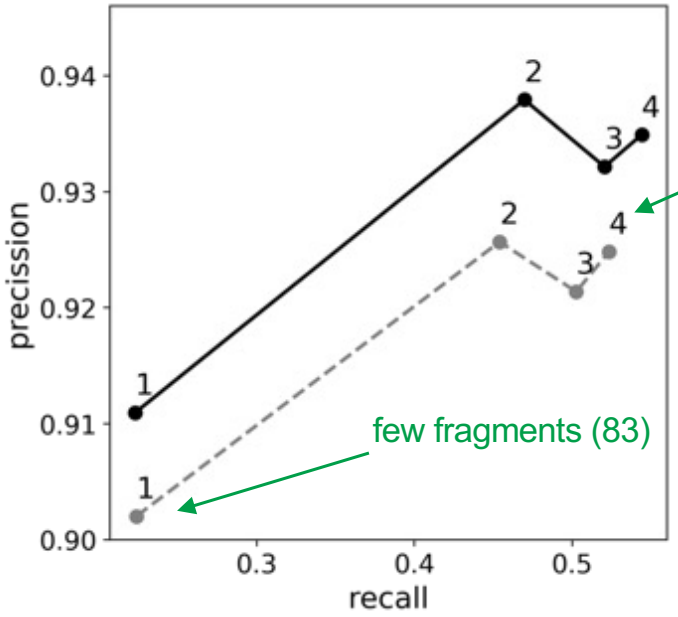


Watson-Crick (canonical) base-pairs



non-canonical base-pairs

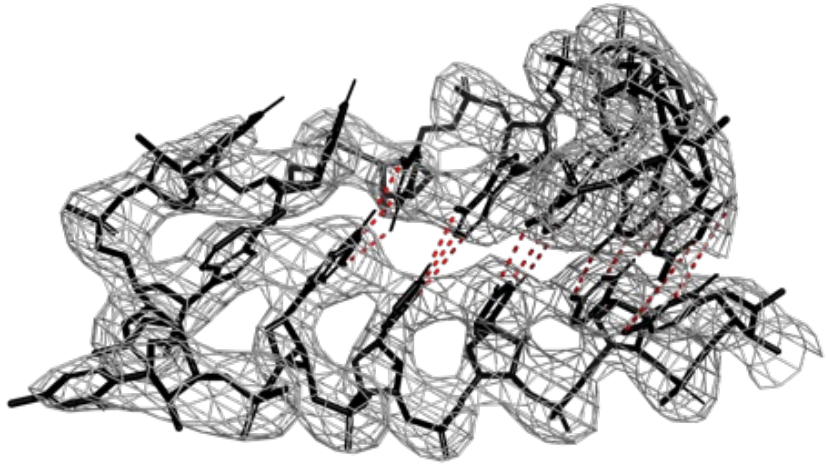




Chojnowski, *in preparation*

EMBL

checkMySequence and NA crystal structures

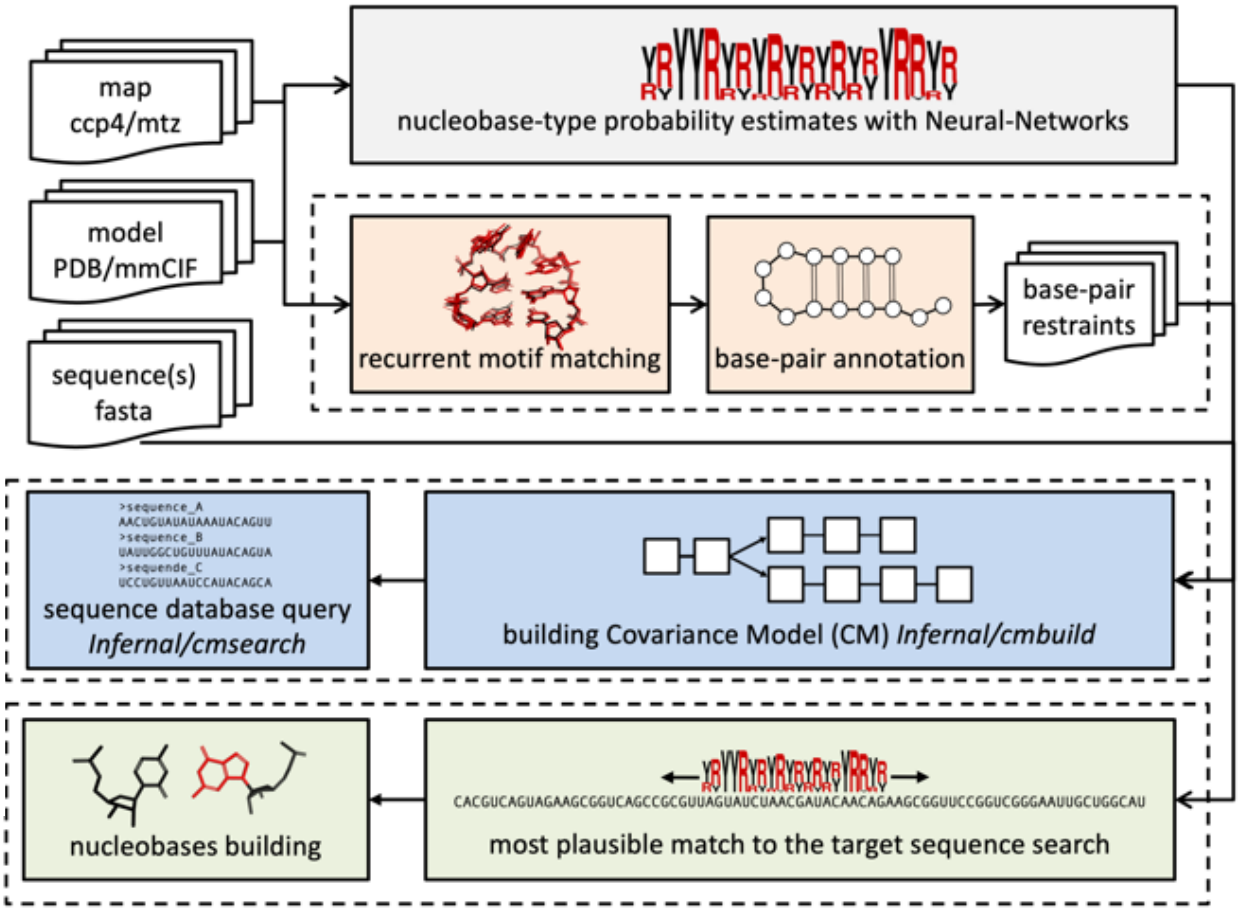


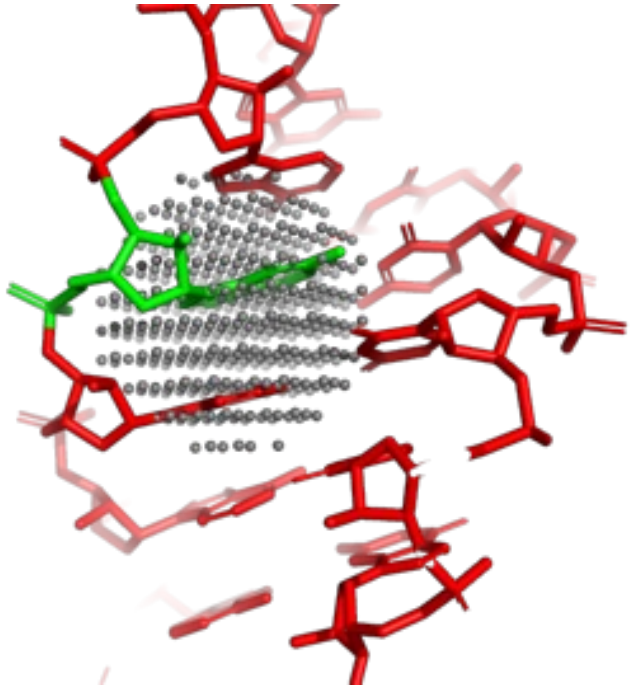
	p-value small	p-value ~ 1
sequences OK	✓	?
sequences differ	⚠	?

Chojnowski, *in preparation*

EMBL

Nucleic acid sequence assignment and validation in EM/MX

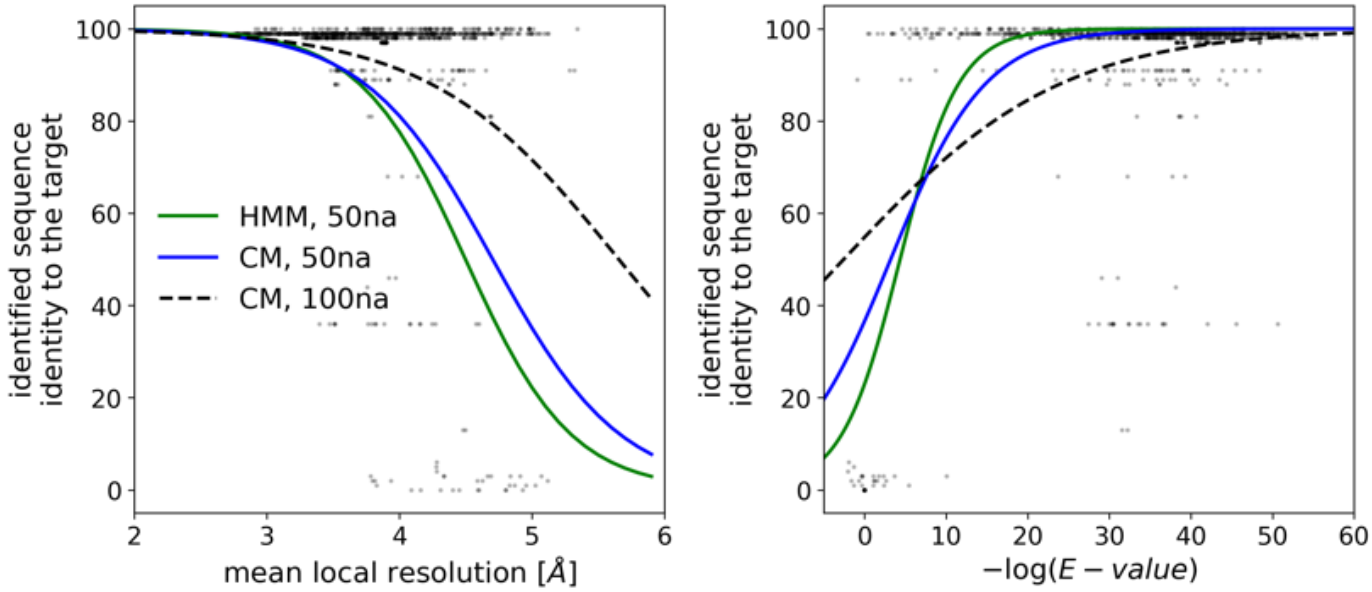




Chojnowski, *in preparation*

EMBL

Nucleic acid sequence identification in EM strcutures

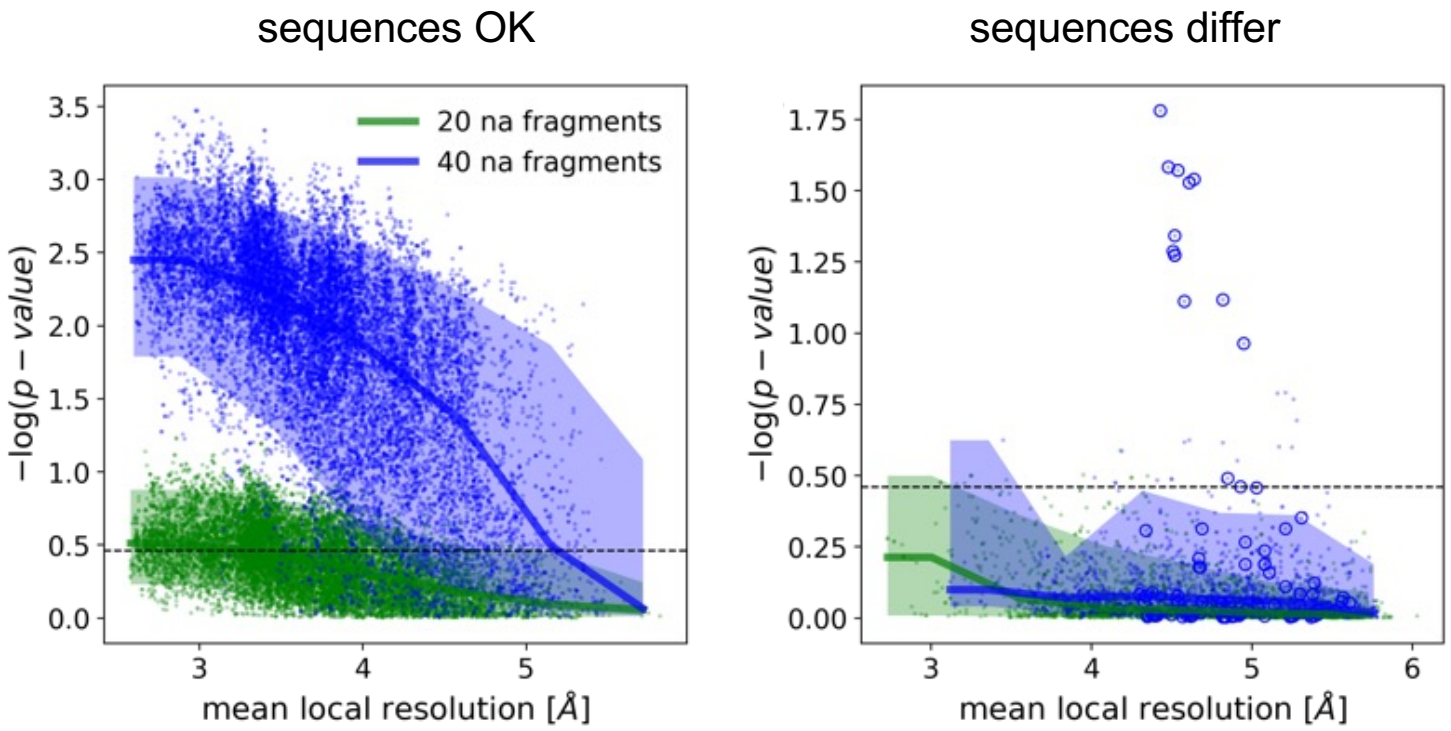


Sequence identification of short RNA fragments from
17 EM structures of ribosomes at 3.5Å or better

Chojnowski, *in preparation*

EMBL

Nucleic acid sequence assignment in EM structures



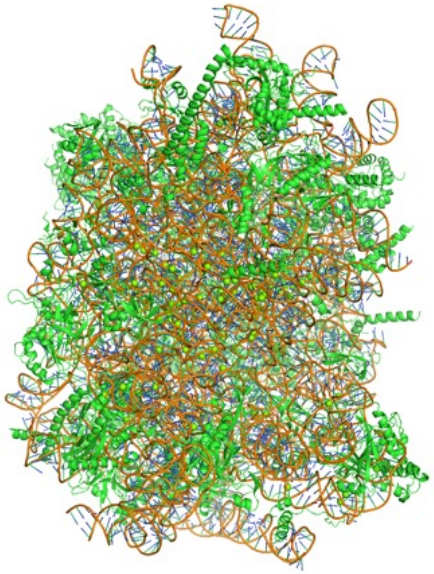
p-values for 20k RNA chain fragments re-assigned to target sequence

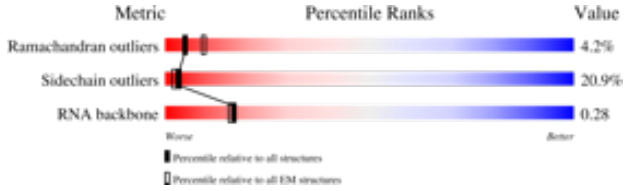
Chojnowski, *in preparation*

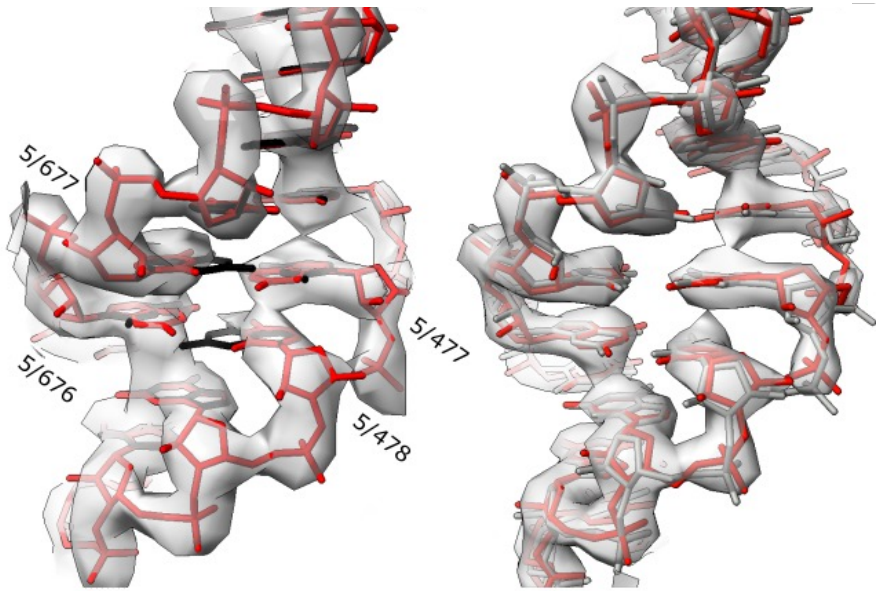
EMBL

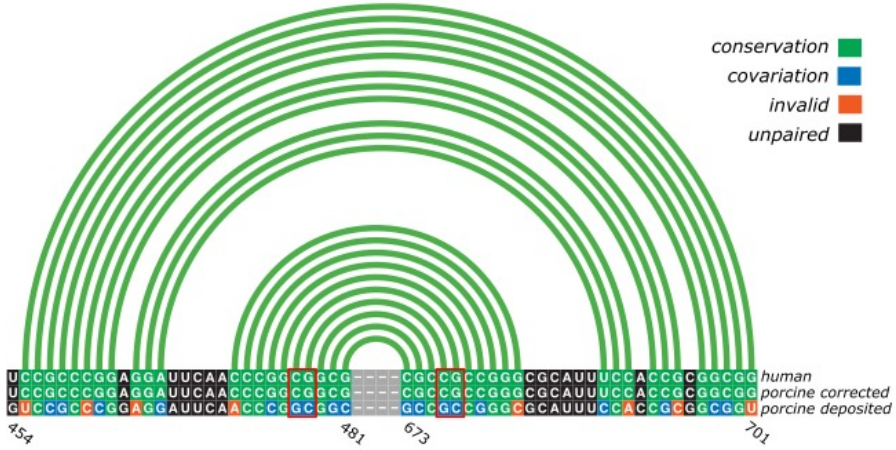
Sequence assignment issues in rRNA

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







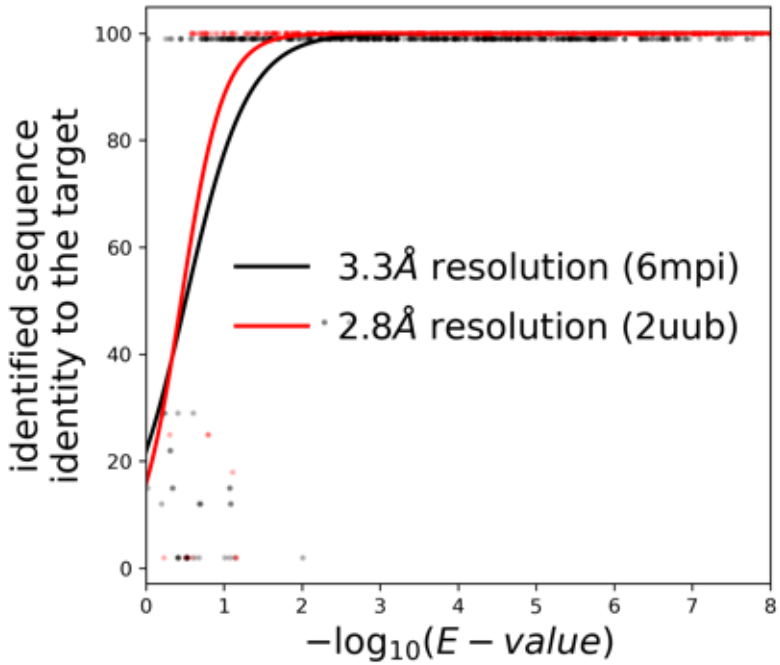


Chojnowski, *in preparation*

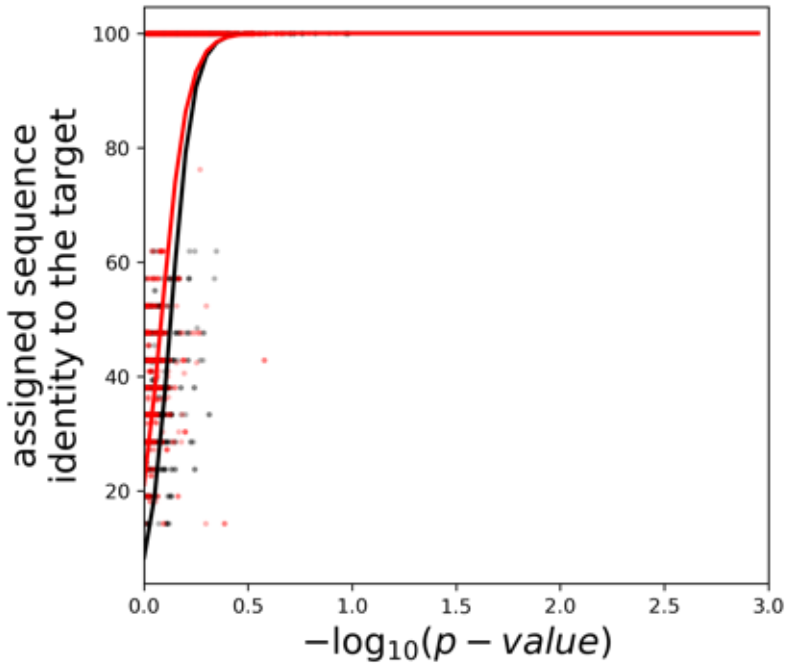
EMBL

Nucleic acid sequence assignment and identification in MX

sequence identification



sequence assignment

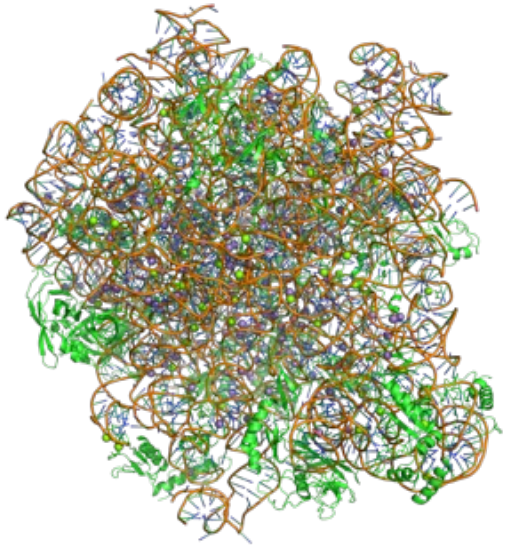


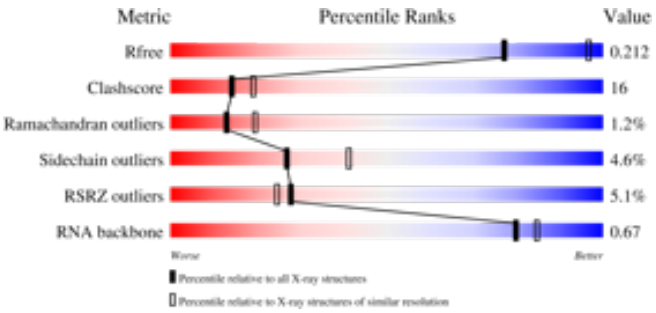
Chojnowski, *in preparation*

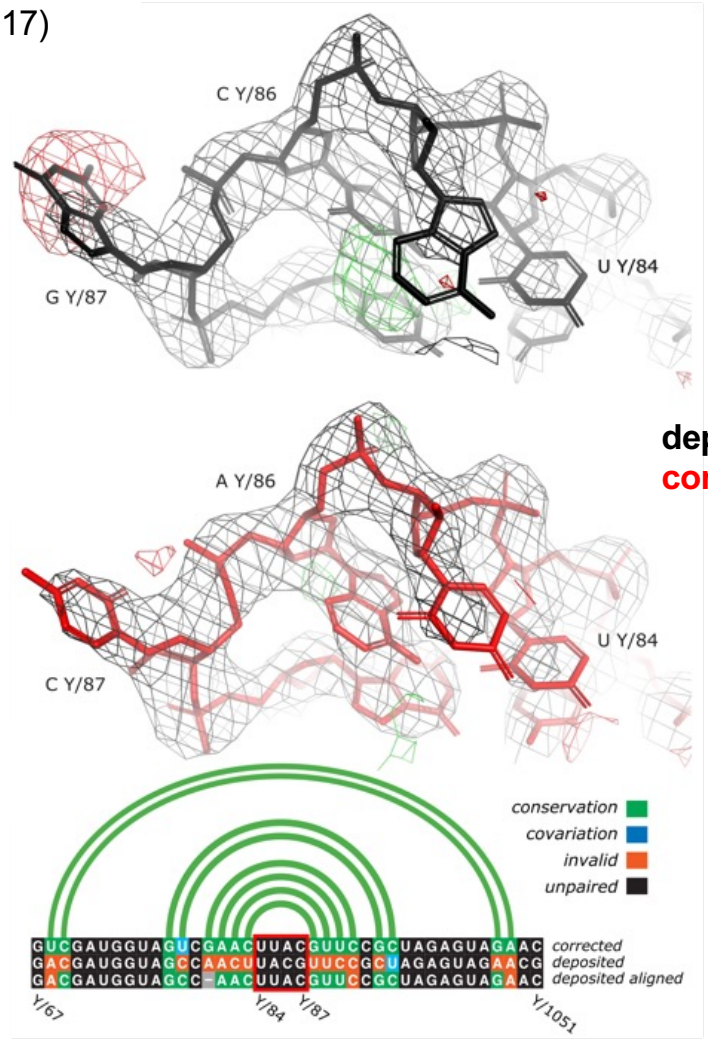
EMBL

Sequence assignment issues in rRNA

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





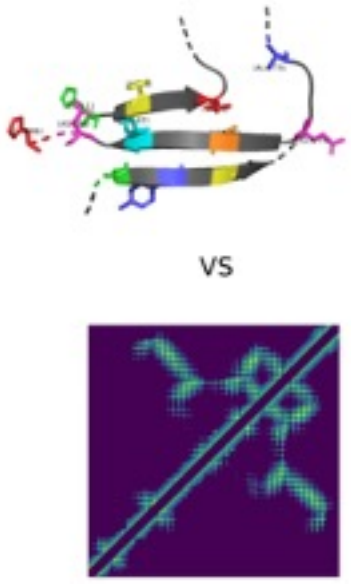


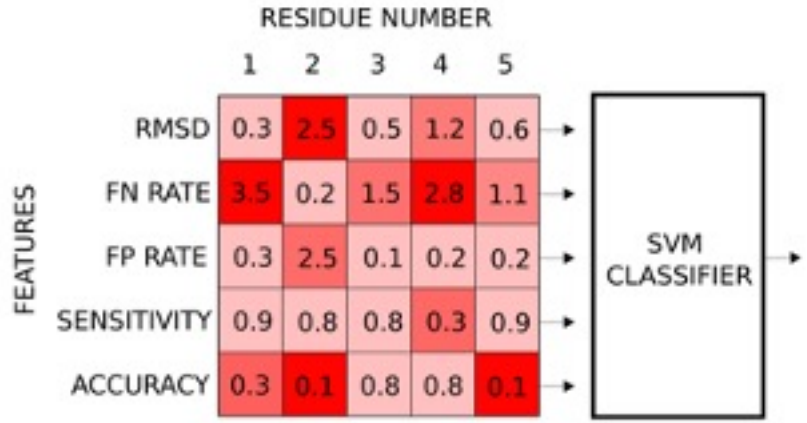
Chojnowski, *in preparation*

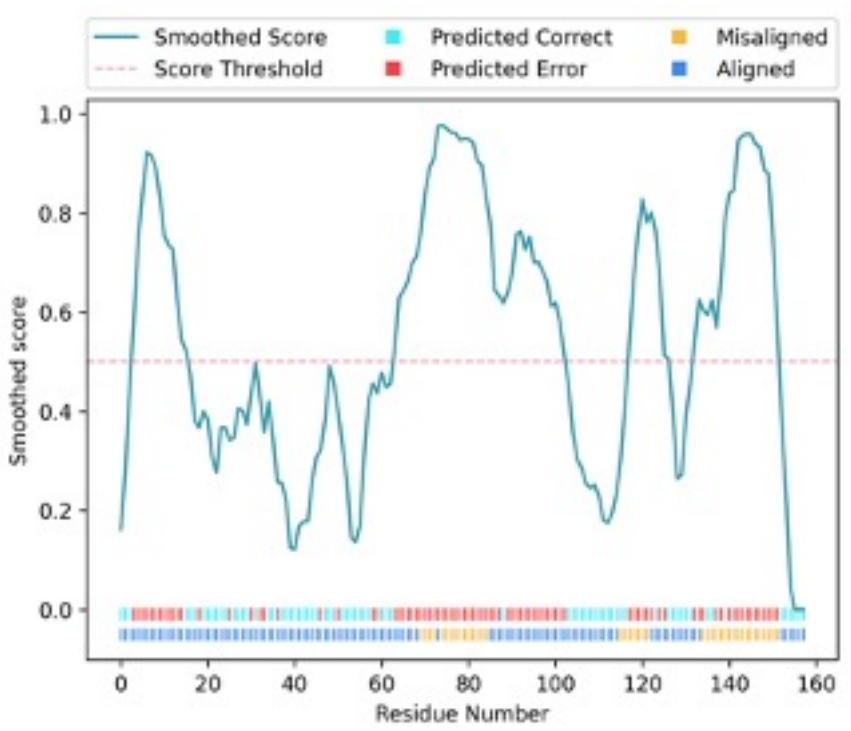
EMBL

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 Krissignel

São Carlos Institute of Physics
Diego A. Leonardo

EMBL Heidelberg
Wolfram Seifert-Davila

Laboratorio de Biolgia Molecular, Peru
Dan E. Vivas-Ruiz

Martin Luther University
Panos Kastritis
Ioannis Skalidis

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

CCPEM and **CCP4** workshop students!

findMySequence

GitLab IUCrJ CCP4

checkMySequence

GitLab Acta Cryst D

doubleHelix
available soon
at gitlab.com/gchojnowski

EMBL