Bioinformatics for structural biologists

Prof Dan Rigden

UNIVERSITY OF LIVERPOOL

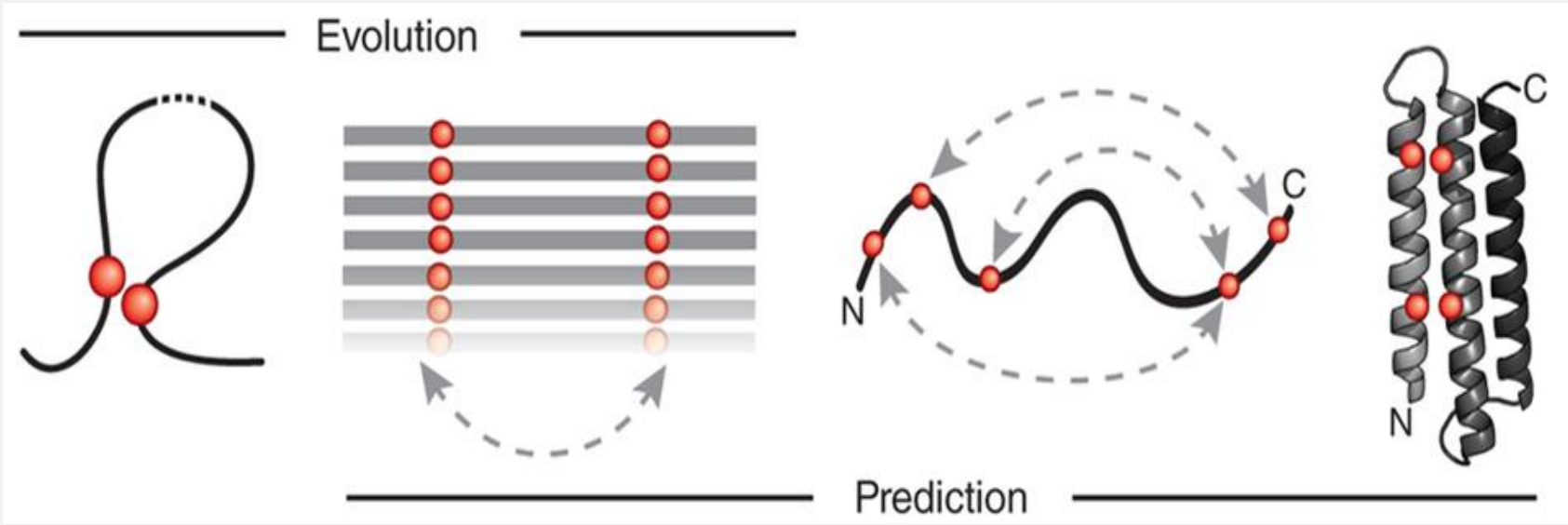
- Introduction
  - Predicting contacts from sequences using evolutionary covariance
  - Deep Learning-based structure prediction eg AlphaFold 2
- Bioinformatics throughout structure determination
  - Predicting domain structure
    - Construct design, experimental strategy
    - Protein engineering
  - ~~Predicting tertiary structure *ab initio* for MR~~
  - Quaternary structure and protein interactions
  - Finalising the structure, validation
  - Structure-based function interpretation
    - Majoring on easily available servers/predictions
    - New Deep Learning-based methods
    - Case study from Structural Genomics (if time!)
  - The sequence alignment in your paper…
- Cross-cutting messages
  - Using multiple methods for a task is good
  - CCP4 has many useful options

Introduction: Evolutionary covariance and AF2/RF

Predicting contacts and distances between residues

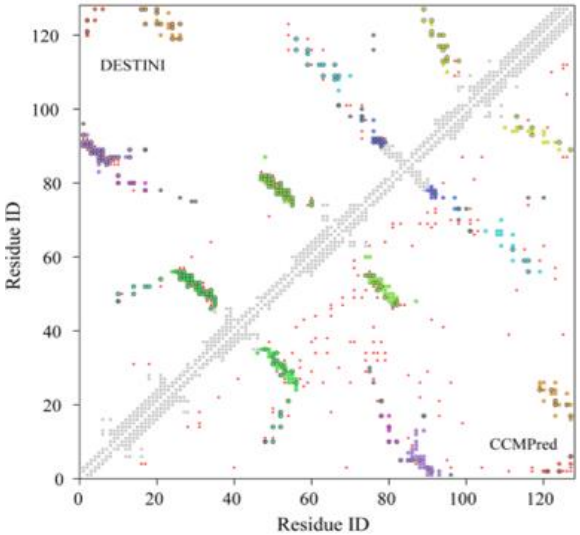
Deep Learning-based structure prediction methods

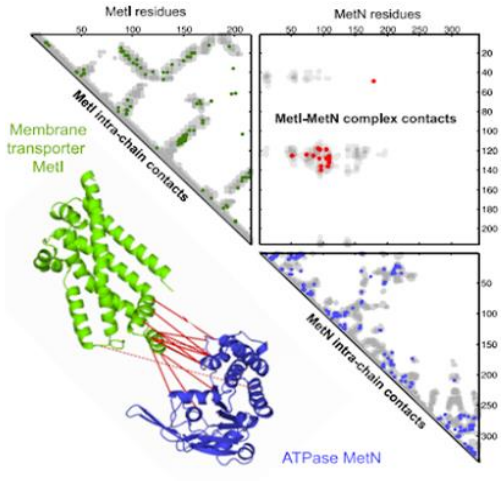
Evolutionary covariance

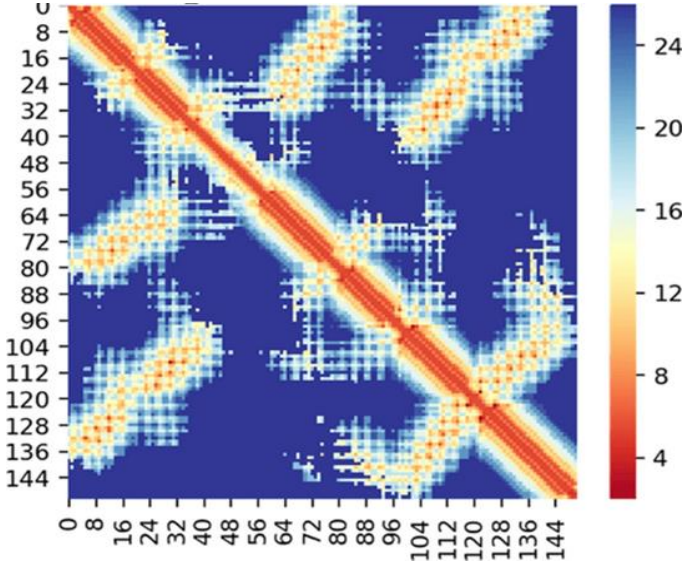


Marks *et al.* (2012) Nature Biotech 30, 1072

Predicting contact maps and distograms







Drove modelling by: EVFold, DMPfold etc

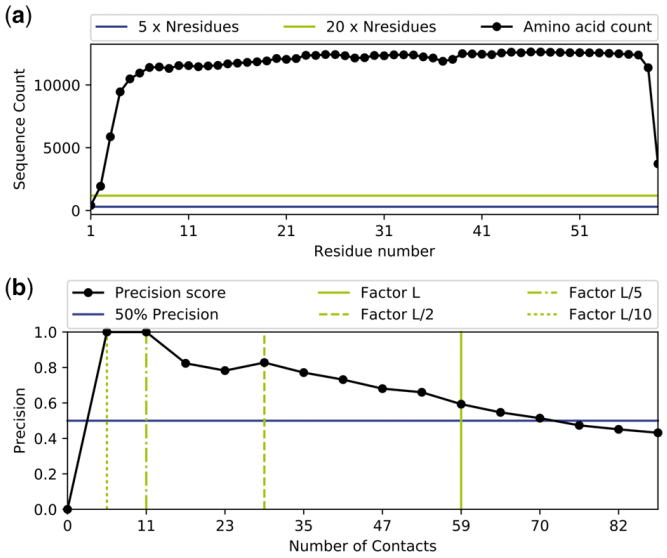
trRosetta, AlphaFold (1)

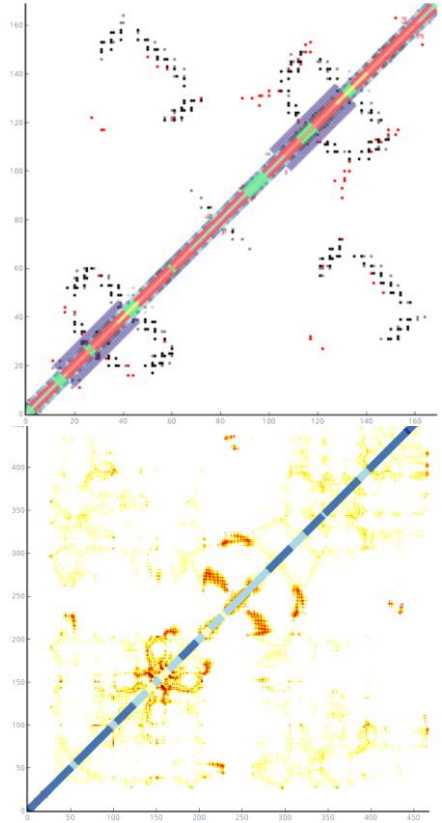
Gao et al (2019). *Sci Rep* **9,** 3514

Hopf et al. eLife (2014) 3,e03430.

Adhikari (2020) *Sci Rep* **10,** 13374

ConKit and ConPlot.org



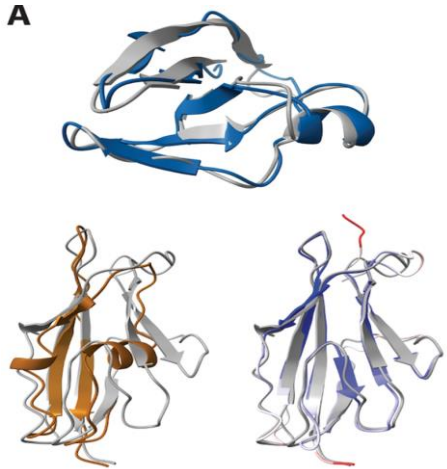


Simkovic *et al.* (2017) Bioinformatics **33** 2209

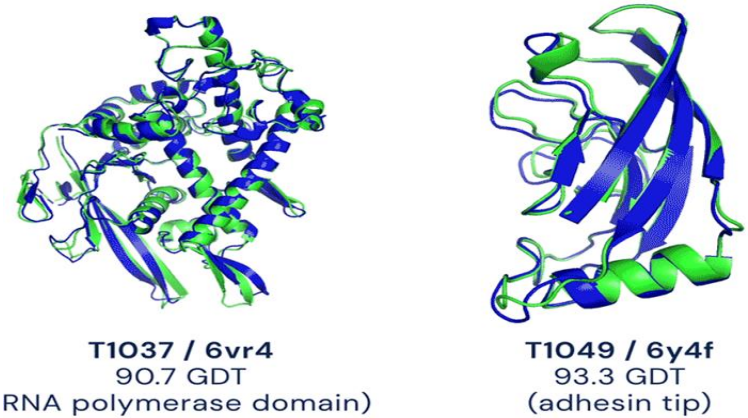
Sánchez Rodríguez *et al.* (2021) Bioinformatics btab049

Multiple methods: RoseTTAFold and AlphaFold 2

- Still use information including covariance from MSAs but networks learn to extract information without imposition of a particular model.
- End-to-end networks produce models directly rather than two separate steps



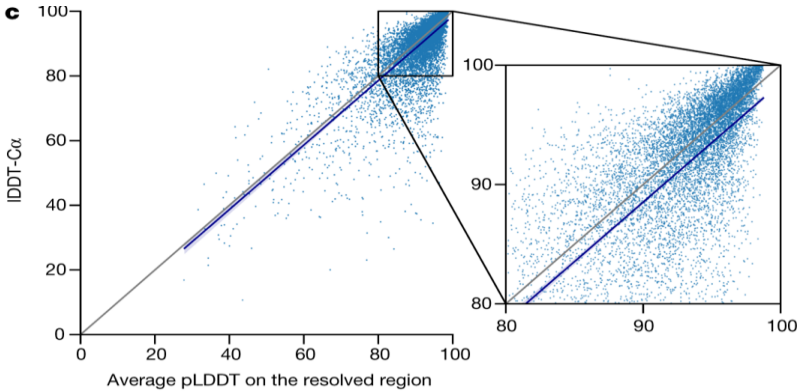
Baek et al (2021) Science 373,871



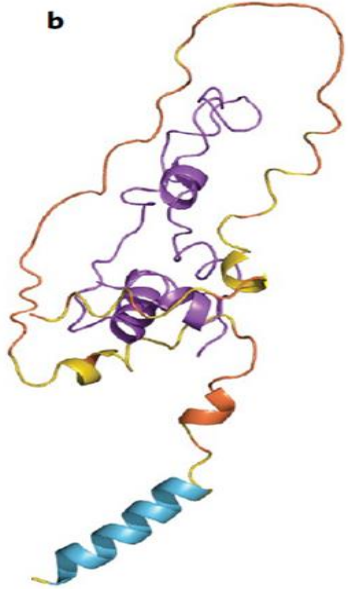
Jumper et al (2021) Nature 596,583

Multiple methods: RoseTTAFold and AlphaFold 2

- Often amazing quality but not always… Need good MSA or template!

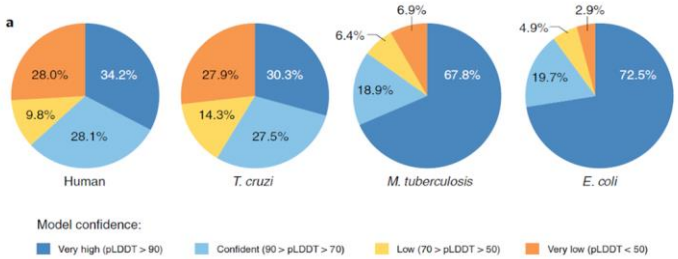


Jumper *et al.* (2021) Nature 596, 583



Phosphatase crystal structure vs AF2 model

Insulin crystal structure vs AF2 model



Thornton *et al.* (2021) Nature Medicine 27. 1666

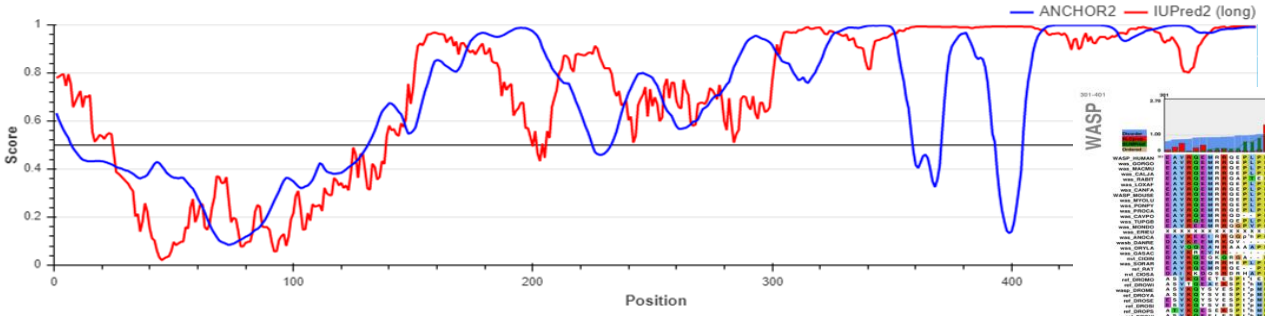
Domains, construct design and solubility

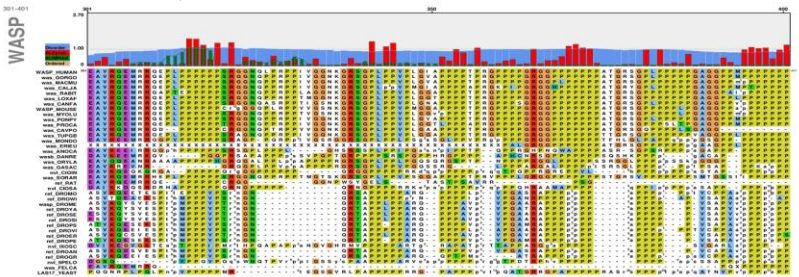
Which part to express for crystallisation?

If necessary, can solubility etc be improved?

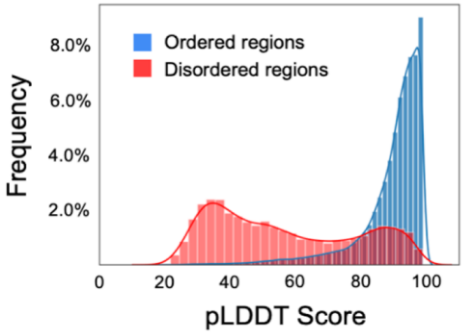
Intrinsic disorder prediction

- Not all proteins and protein regions fold into stable structured domains. ID proteins and regions will not crystallise (alone)
- There are many predictors, all performing roughly equally well
- I recommend AIUPred (fast) and MetaDisorder (slow but good)
- Can also look for short interaction motifs in ID regions (ANCHOR, SlimPred)

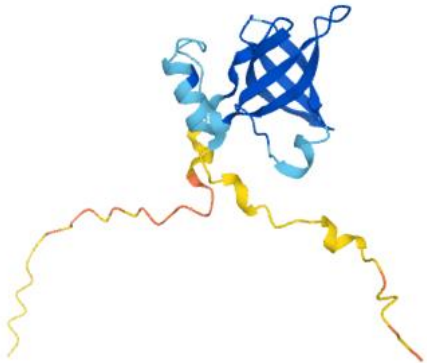


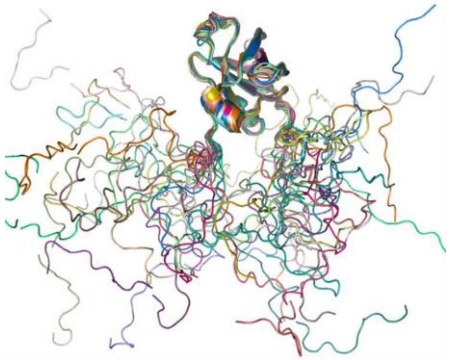


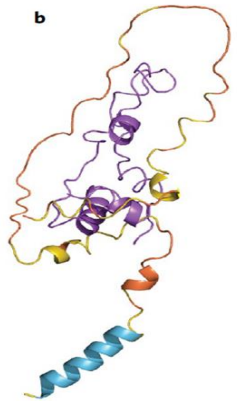
Most low-confidence AF2 regions are disordered but they may just be plain wrong



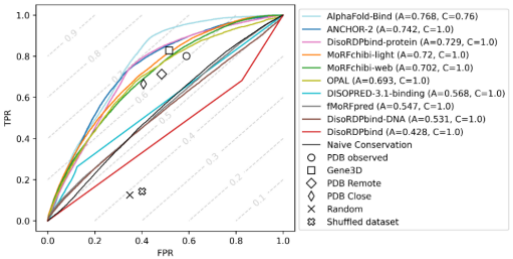
Binder et al (2022) COSB 74, 102372





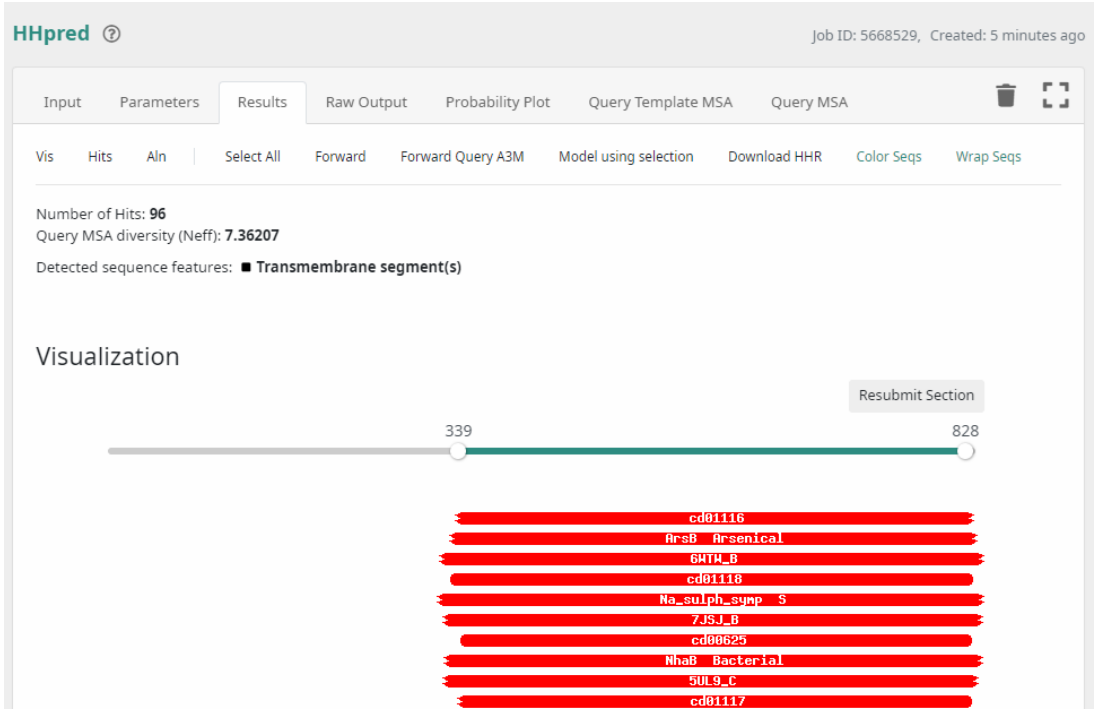


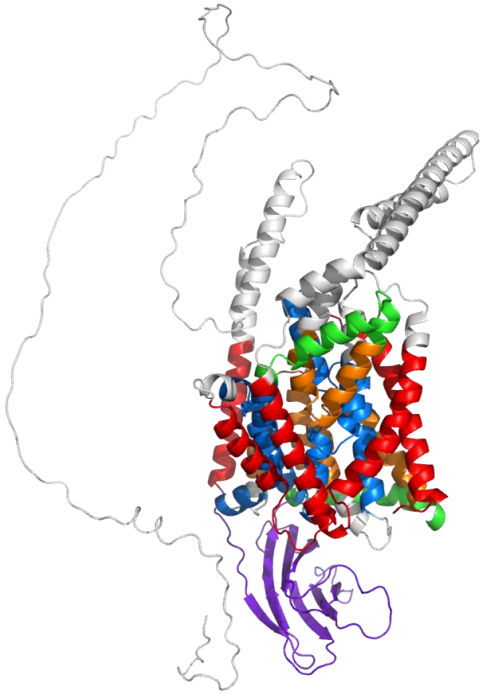
Relatively high pLDDT regions in AF2 disordered regions may predict interaction motifs



Piovesan *et al.* (2022) Protein Sci. 31, e4466

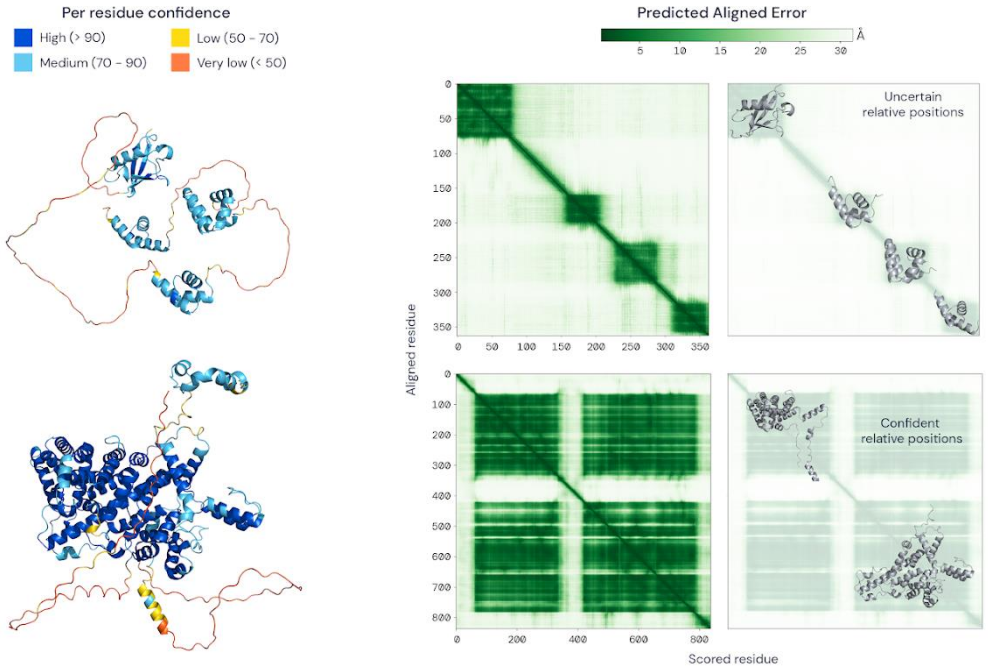
Beyond sequence matching: AF2 for domain discovery





Mesdaghi *et al.* (2023) BioSci Rep 43:BSR20230060

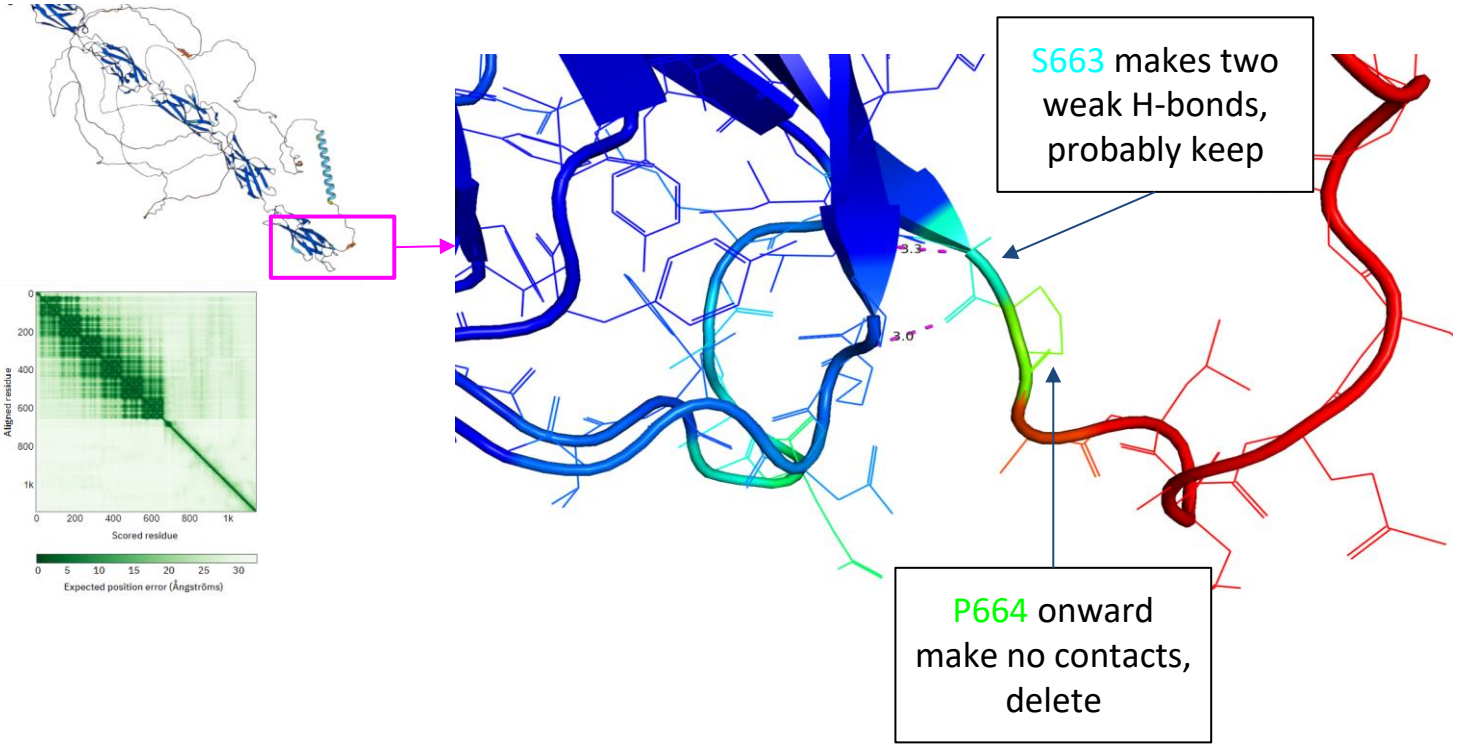
AF2 for domain boundary prediction



Slice’n’Dice uses the PAE or other kinds of clustering to split structures into domains for MR

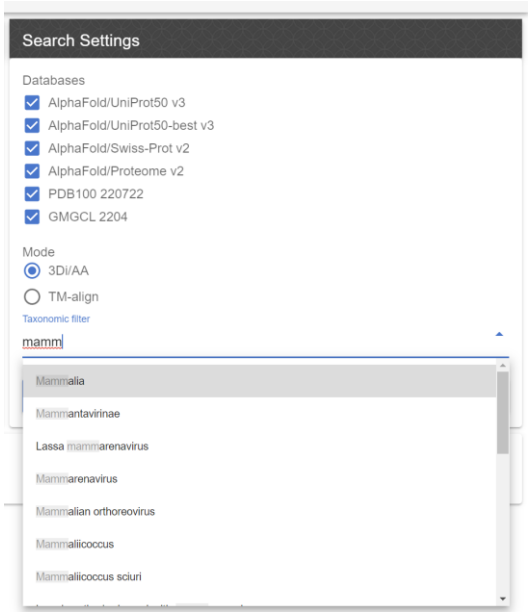
https://deepmind.com/research/publications/2021/enabling-high-accuracy-protein-structure-prediction-at-the-proteome-scale

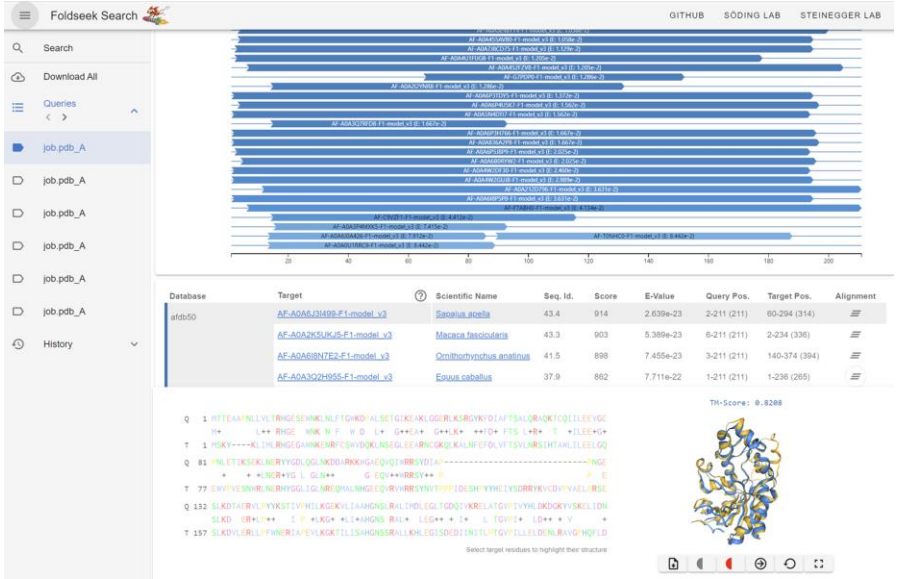
Fine details too, if reliability allows



Recognising homology using a structure

FoldSeek can search PDB, AFDB, ESMAtlas, BFVD using a structure in seconds





https://search.foldseek.com/

Van Kempen et al. (2022) Nature Biotech
https://doi.org/10.1038/s41587-023-01773-0

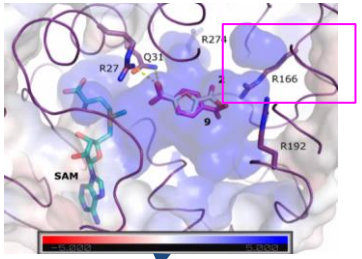
Protein engineering

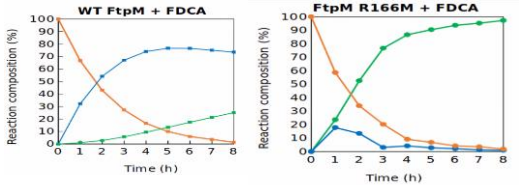
Where you have confidently interpreted a model as reliable then you can treat it almost like a crystal structure…

Rational enzyme engineering using an AF2 prediction

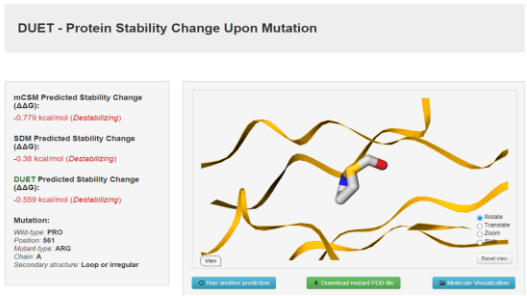
Combining AF2 predictions with other structure-based methods eg for stability prediction. AlphaMissense is another option

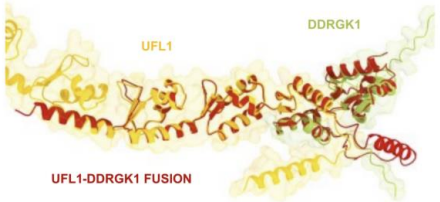
Predicting a stable fusion protein based on analysis of a complex structure prediction





Ward et al (2022) Angewandte Chemie 61,e202117324

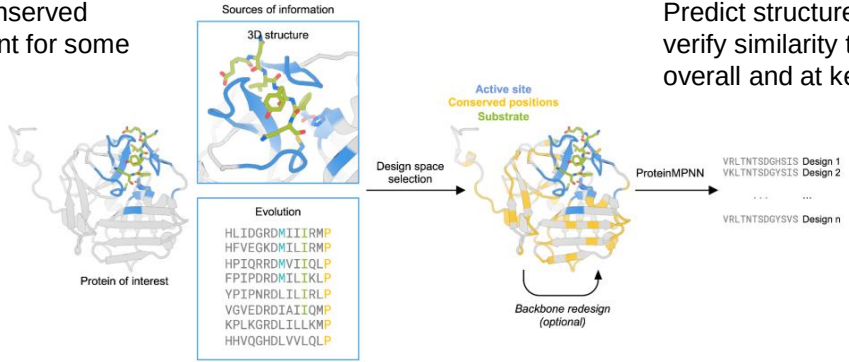




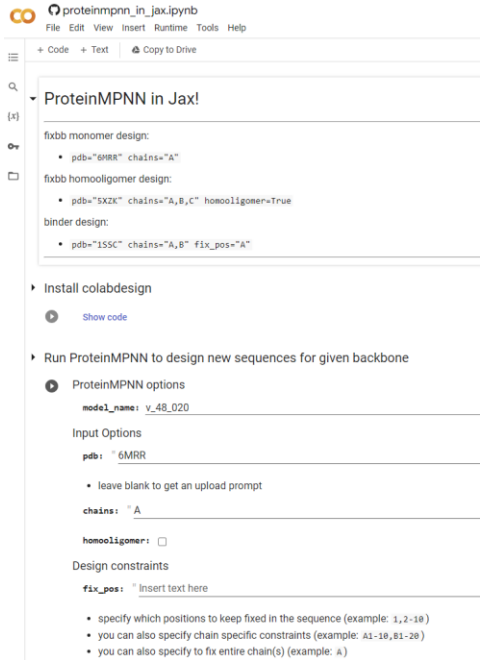
Banerjee et al (2022) BioRXiv https://doi.org/10.1101/2022.09.15.508077

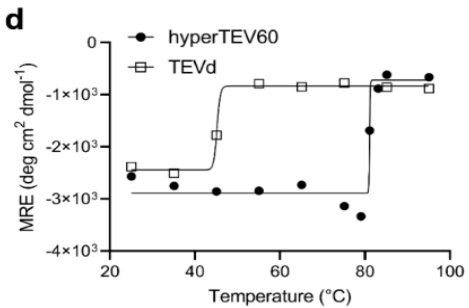
Protein engineering with deep learning

Identify and retain residues near ligand (functionally important) and conserved positions (important for some reason)

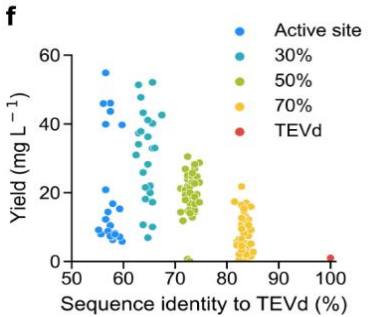


Allow ProteinMPNN to fill in the gaps and design variants. Predict structures with AF2 and verify similarity to starting point, overall and at key positions

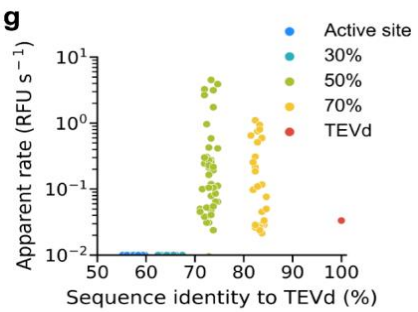




Better thermostability



Better expression



Better activity

Sumida et al Improving protein expression, stability, and function with ProteinMPNN JACS (2024) 146, 2054

Tertiary structure

AF2 and RF output as search models for MR

AF2/RF and Molecular Replacement

MR *is* a kind of structure prediction so the availability of accurate models of most proteins has impacted structure solution hugely

Similarly, accurate models can be used to interpret cryo-EM maps

See other talks for how to find and deploy AF2 models with MrParse, MrBUMP, Slice'n'Dice, ARCIMBOLDO_SHREDDER etc

Getting diversity in your models

This will be needed for hard cases and for cases where multiple conformations are accessible or sought

Ways to sample conformation more broadly

- Network dropout (eg `num_samples` and/or `is_training` on the advanced colabfold page)

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb

- Feed AF2 templates in the 'right' conformation (and maybe ignore MSA features)

de Alamo et al (2022) elife 11:e75751

- Deliberately make the input MSA more shallow

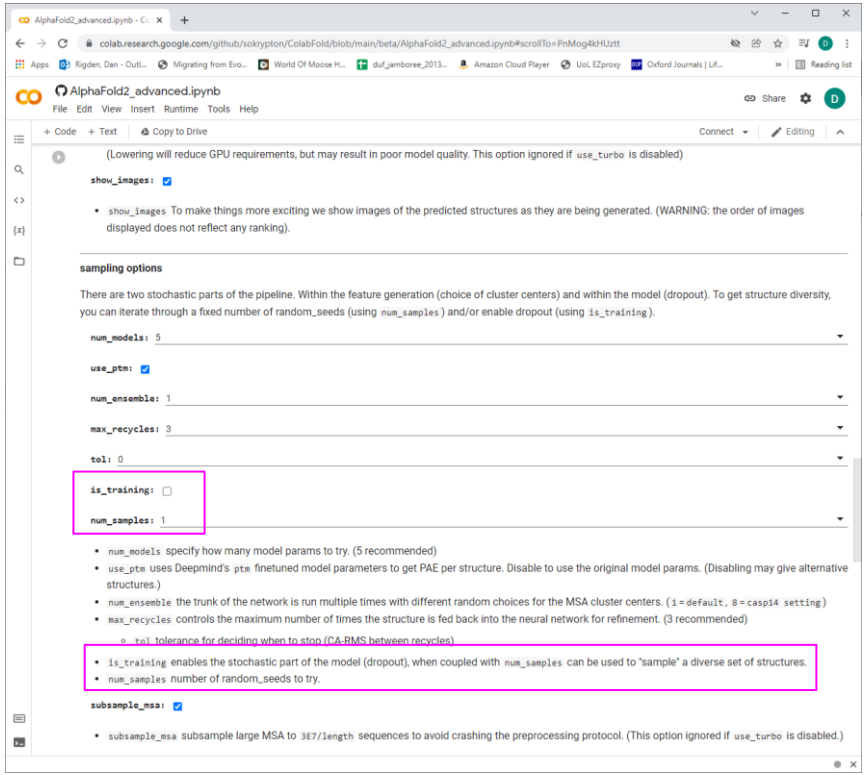
Heo & Feig (2022) PSFB DOI: 10.1002/prot.26382

- Cluster the input MSA and try individual sub-clusters

Wayment-Steele et al (2024) Nature 625, 832

- Edit the input MSA to mutate to Ala residue pairs that are driving the 'wrong' conformation

Stein and Mchaourab (2022) PLoS CB 18: e1010483.

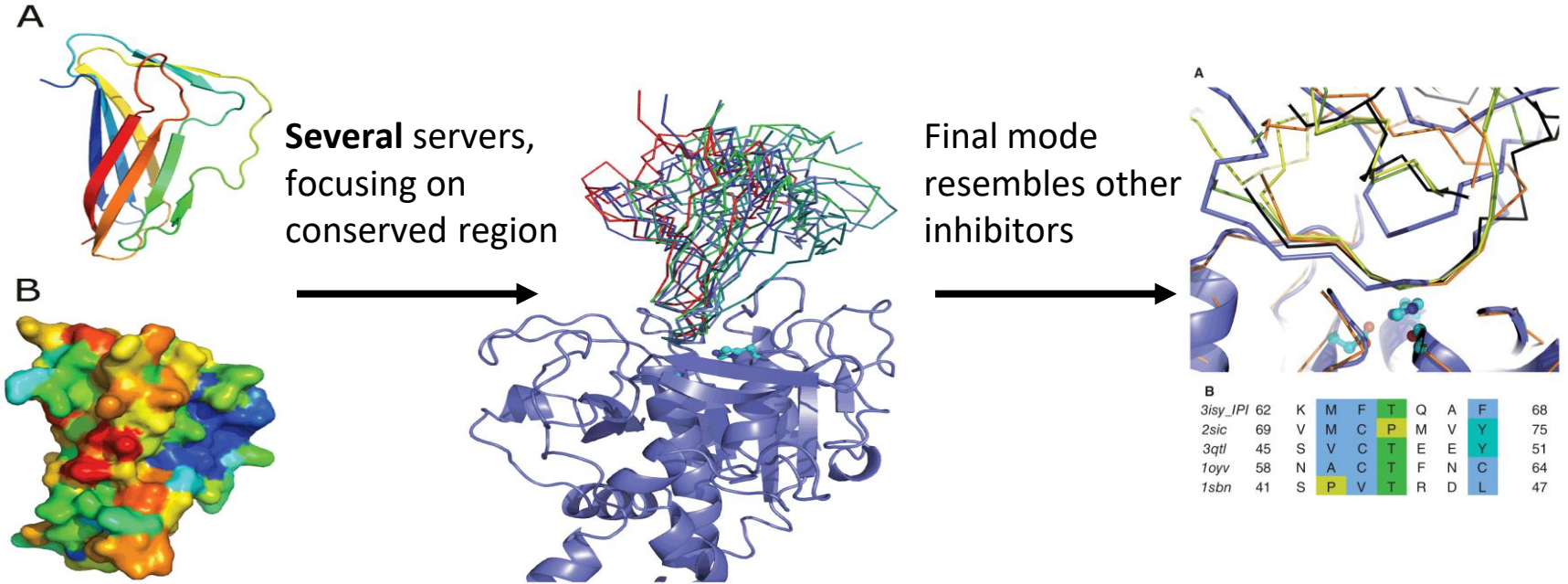


Quaternary structure and ligand interaction

Predicting protein-protein interactions

- Relevant to MR eg proteins A and B are cocrystallised but neither alone solves. An accurately predicted complex, being larger, might solve
- Many docking methods predict complexes based on steric complementarity plus other scoring functions
- Recommendable servers include
  - ClusPro, the best performing docking method
  - Haddock, which has a good server with different modes
  - Each allows inclusion of other information eg known interface residues. RF/AF2 do not (easily, yet)
  - Symmetric docking at ROSIE server. Also unavailable in RF/AF2

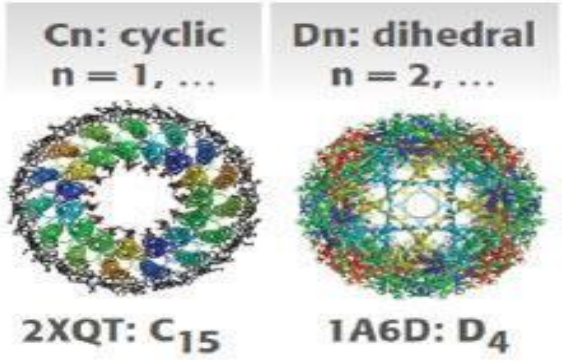
Multiple methods in bioinformatics: *B. subtilis* IPI docking to protease



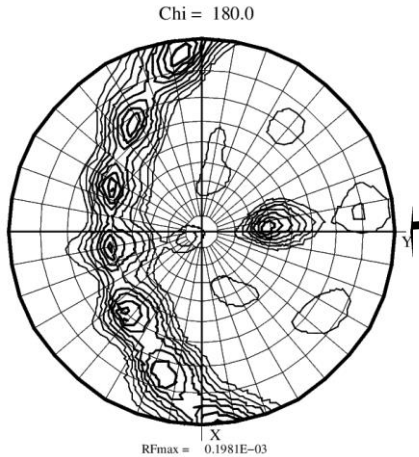
Rigden et al. (2013) F1000Research 2:154

ROSIE symmetric docking for oligomers

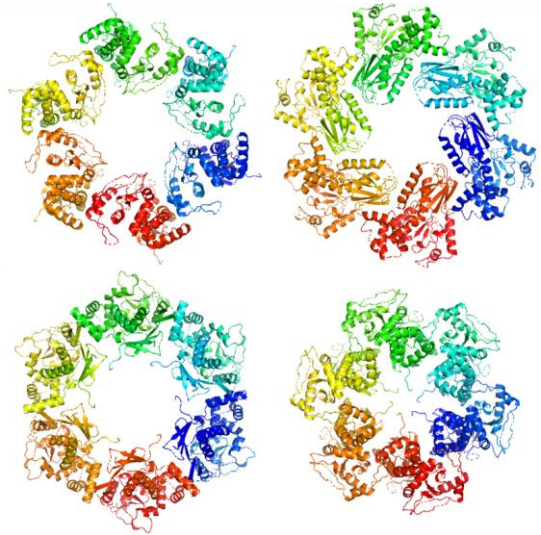
- Only cyclic (C_n) or dihedral (D_n) symmetry at server
- Clues from self-rotation function may be available
- AF2 cannot use this information!



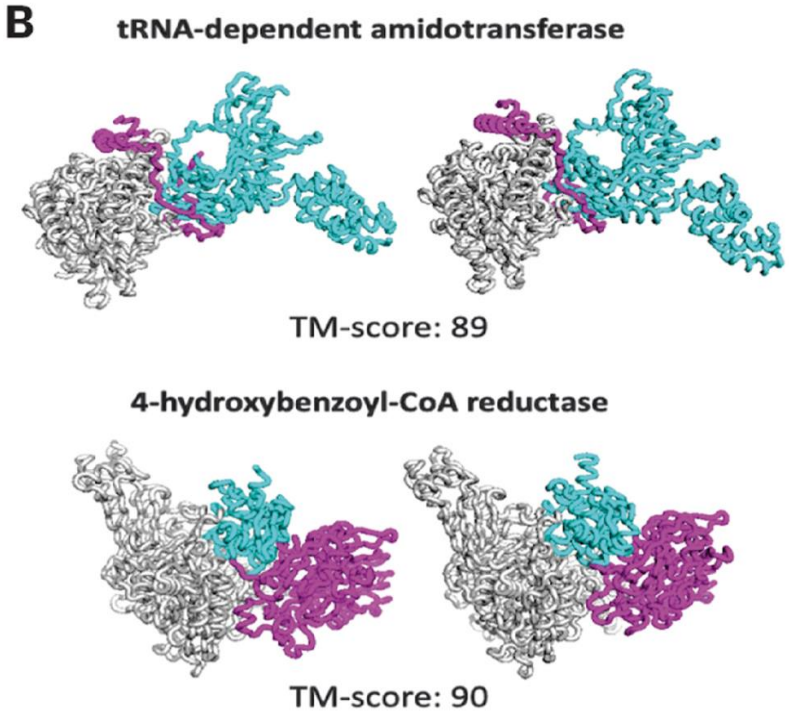
rcsb.org



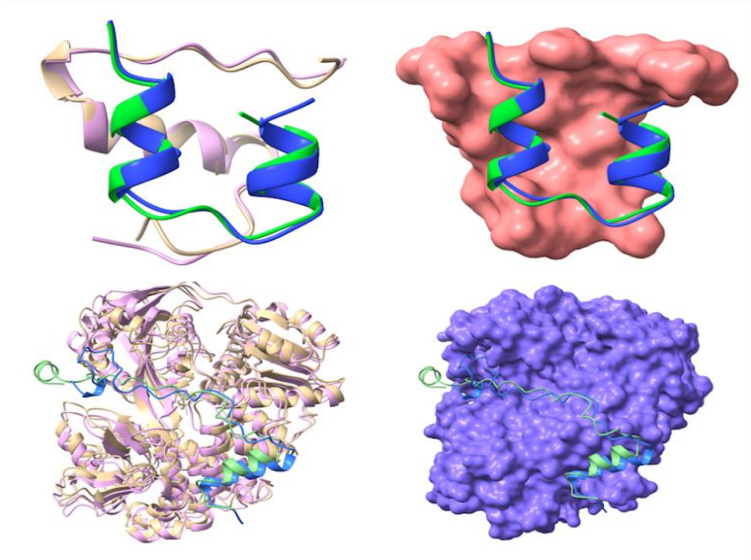
Generate hexamers



RF/AF2 to predict complexes



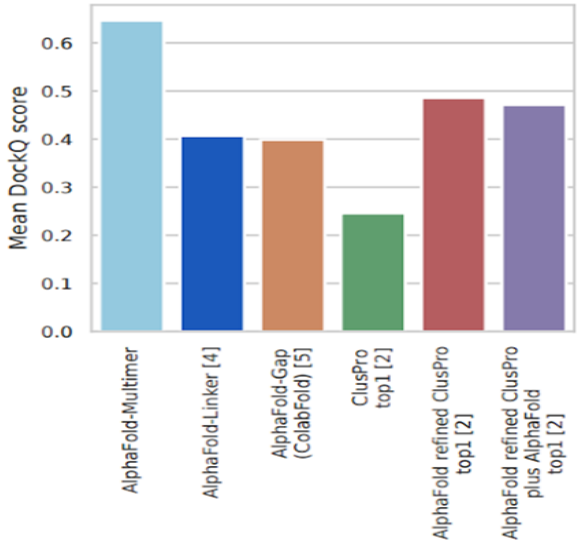
Baek et al (2021) Science 373,871

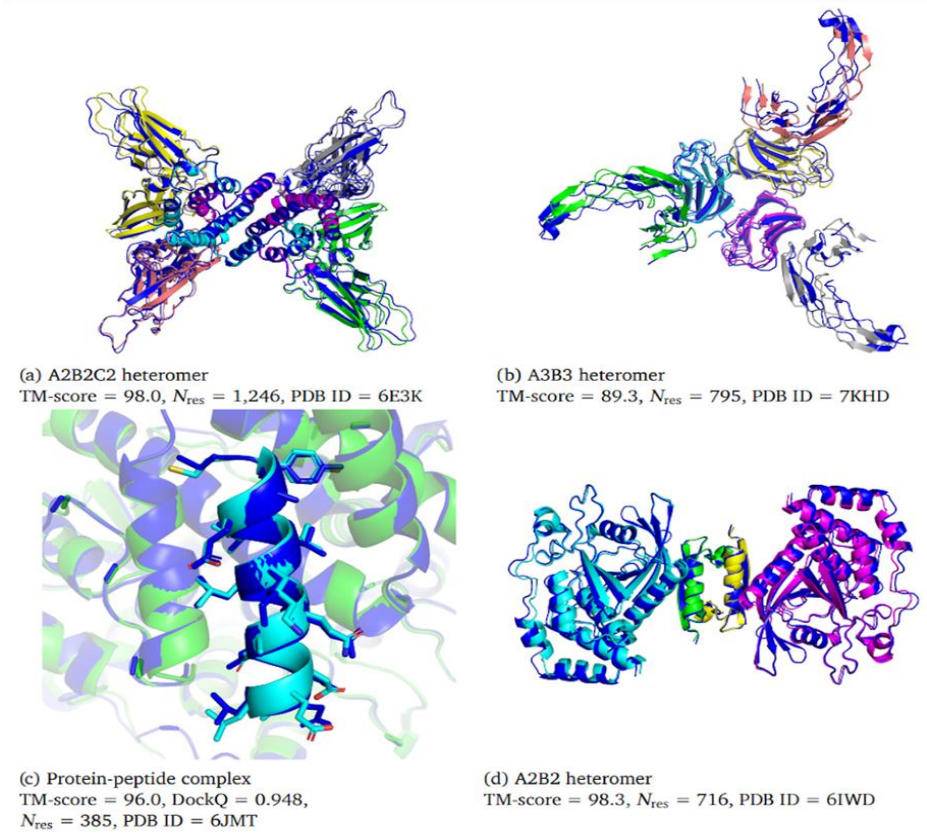


Modelling protein-peptide interactions as separate chains or linked by polyAla are complementary approaches.

Ko and Lee (2021) www.biorxiv.org/content/10.1101/2021.07.27.453972v2

New multimer-trained AF2





(a) A2B2C2 heteromer
TM-score = 98.0, N_{res} = 1,246, PDB ID = 6E3K

(b) A3B3 heteromer
TM-score = 89.3, N_{res} = 795, PDB ID = 7KHD

(c) Protein-peptide complex
TM-score = 96.0, DockQ = 0.948,
N_{res} = 385, PDB ID = 6JMT

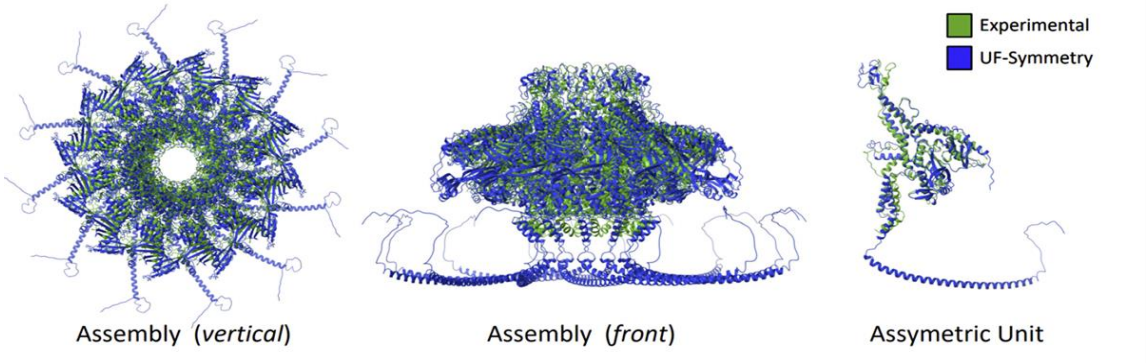
(d) A2B2 heteromer
TM-score = 98.3, N_{res} = 716, PDB ID = 6IWD

AF2 can predict oligomeric state eg asking for 5 copies may produce the natural tetramer + one left over, rather than a pentamer

Evans et al (2021) https://doi.org/10.1101/2021.10.04.463034

Uni-Fold symmetry

- Models a single chain with known symmetry to generate oligomer
- Much quicker and slightly better than other methods on oligomers



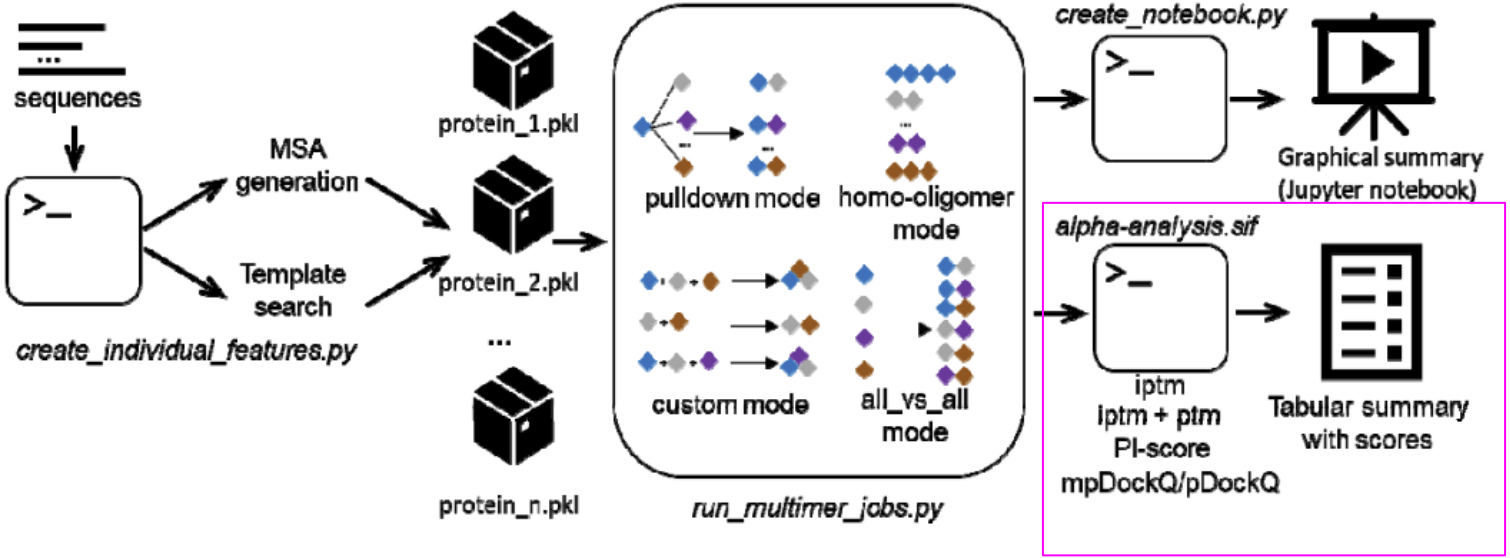
Type	Symmetry	Number of structures
Monomers	–	83,392
Multimers	Asymmetric	27,470
	C2	45,496
	C3	6,037
	C4	1,736
	C5	893
	C6	639
	Larger cyclic	587
	D2	8,571
	D3	2,577
	D4	815
	D5	302
	D6	191
	Larger dihedral	239
	Icosahedral	1,182
	Octahedral	544
	Tetrahedral	475
	Helical	581
	All	98,335
Total	–	181,727

C12 symmetry. AF2 and regular Uni-Fold fail

Li et al (2022) BioRXiv https://doi.org/10.1101/2022.08.30.505833

https://colab.research.google.com/github/dptech-corp/Uni-Fold/blob/main/notebooks/unifold.ipynb

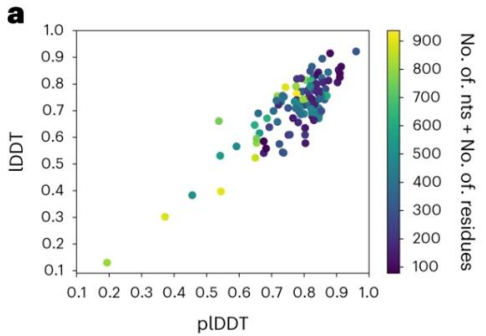
AlphaPulldown for complicated cases

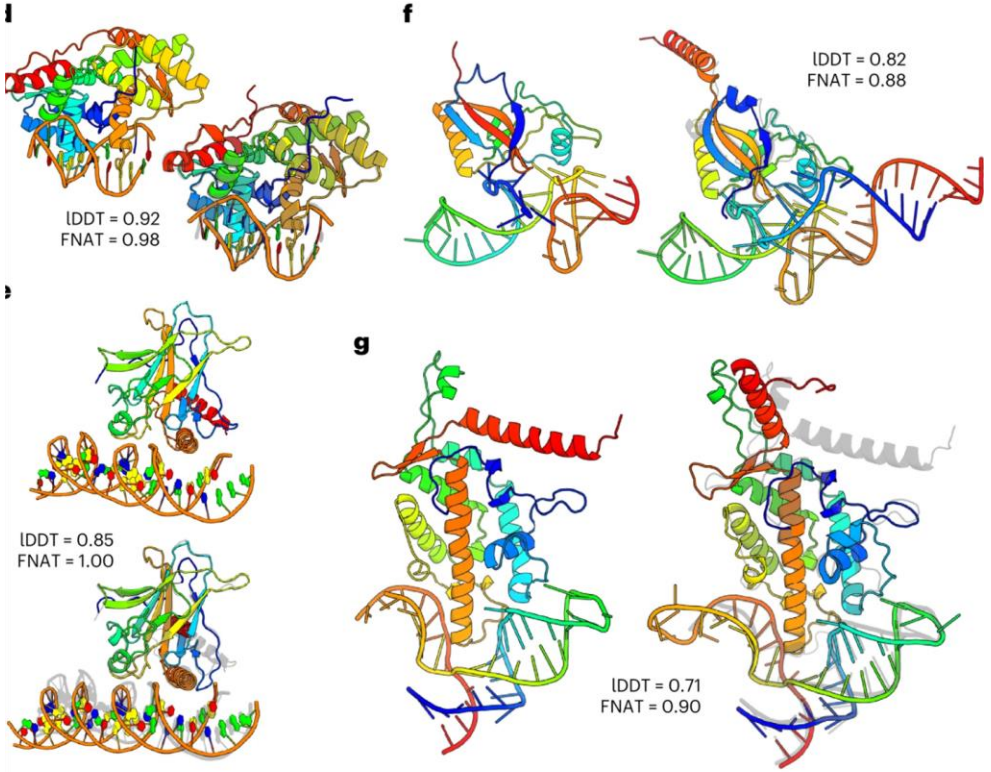


Yue al (2023) Bioinformatics 39,btac749

Protein-NA complexes with RoseTTAFoldNA

As good as AF2 on proteins; as good as DL-based RNA tools on natural RNAs; USP *was* protein-NA complexes

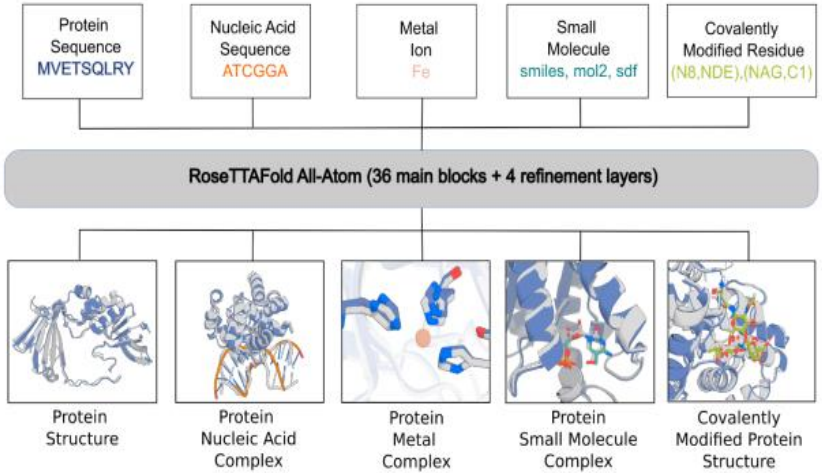




Baek et al (2023) Nature Methods https://doi.org/10.1038/s41592-023-02086-5

Generic protein-ligand prediction

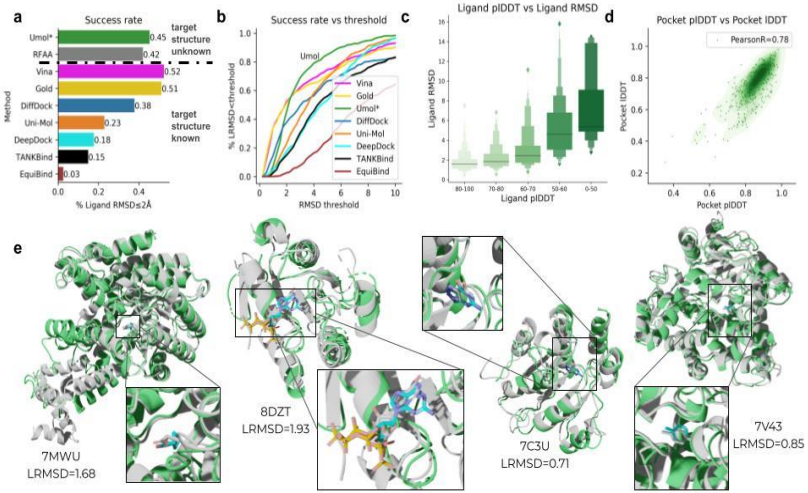
RoseTTAFold All-Atom



Code

Krishna et al (2024) Science doi: 10.1126/science.adl2528.

Umol

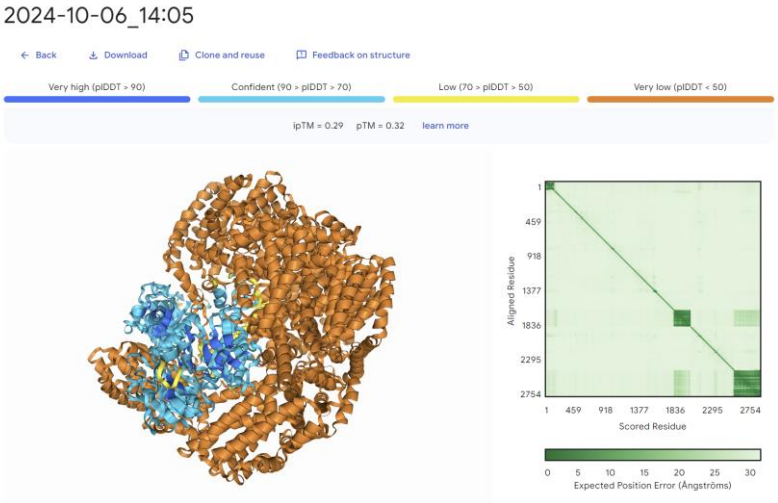


Code, Colab page!

Bryant et al (2024) Nature Communications doi: 10.1038/s41467-024-48837-6

The latest diffusion methods

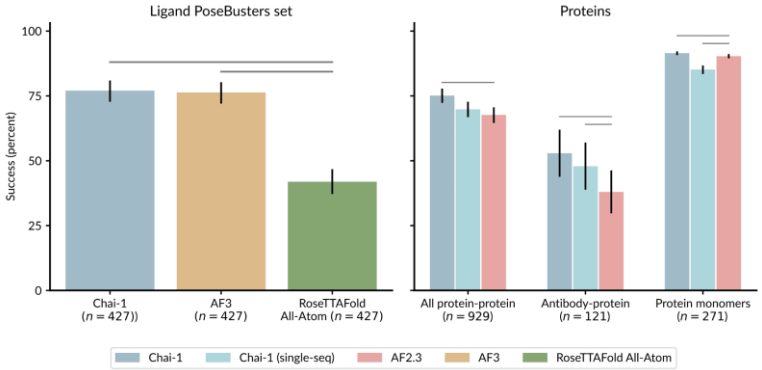
AlphaFold 3



Not a direct development of AF2. Google's server and now code. Server 20 jobs/day, limited ligand library, template always *on*

Abramson et al (2024) Nature 630, 493

Chai-1



One of five (!) AF3 copycats. Can natively use extra information. Server slow, downloads lack pLDDT. Allows commercial use (Nov 2024)

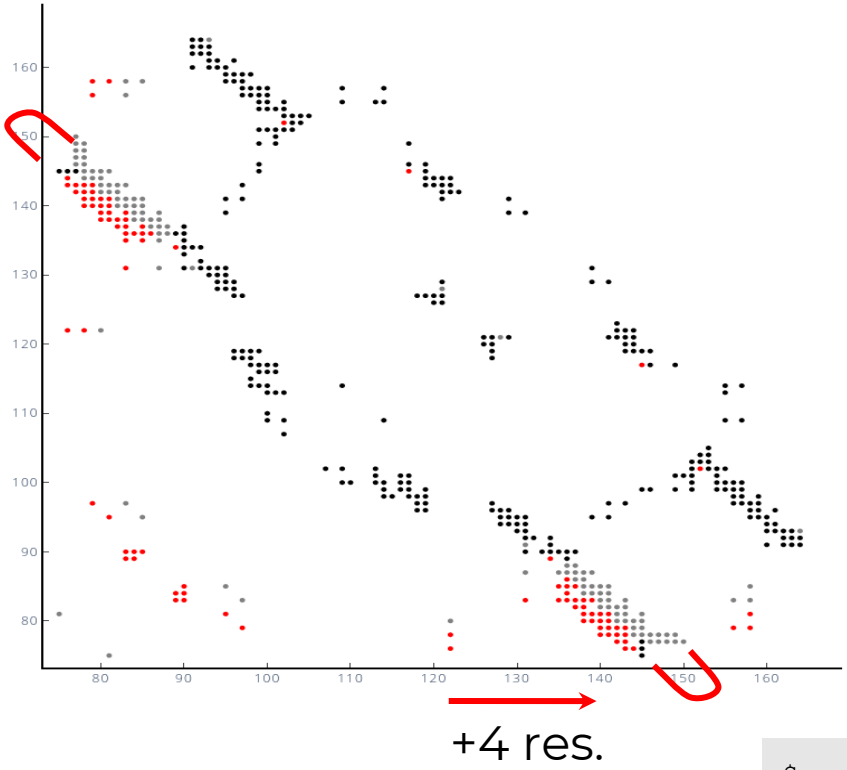
https://www.chaidiscovery.com/blog/introducing-chai-1

Finalising the structure

Does it contain any residual errors?

What is the biologically relevant quaternary structure?

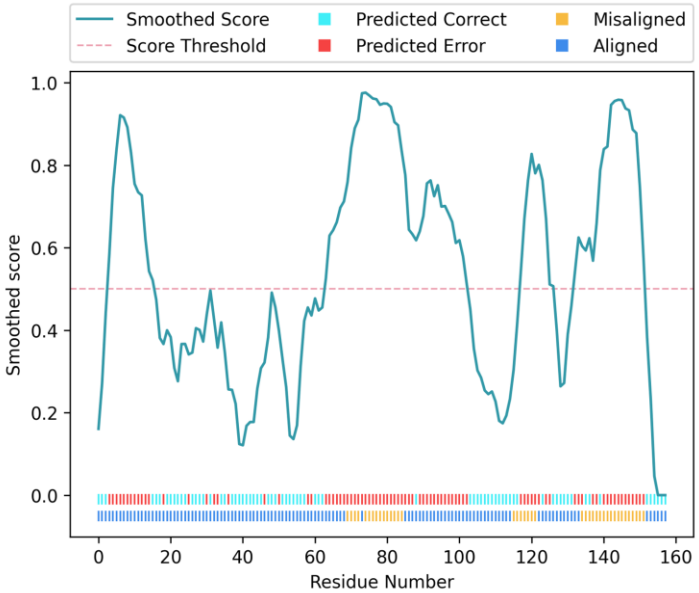
New covariance-based metrics for model validation



+4 res.

Contact map overlap picks up register errors, SVM detects errors in general

ConKit



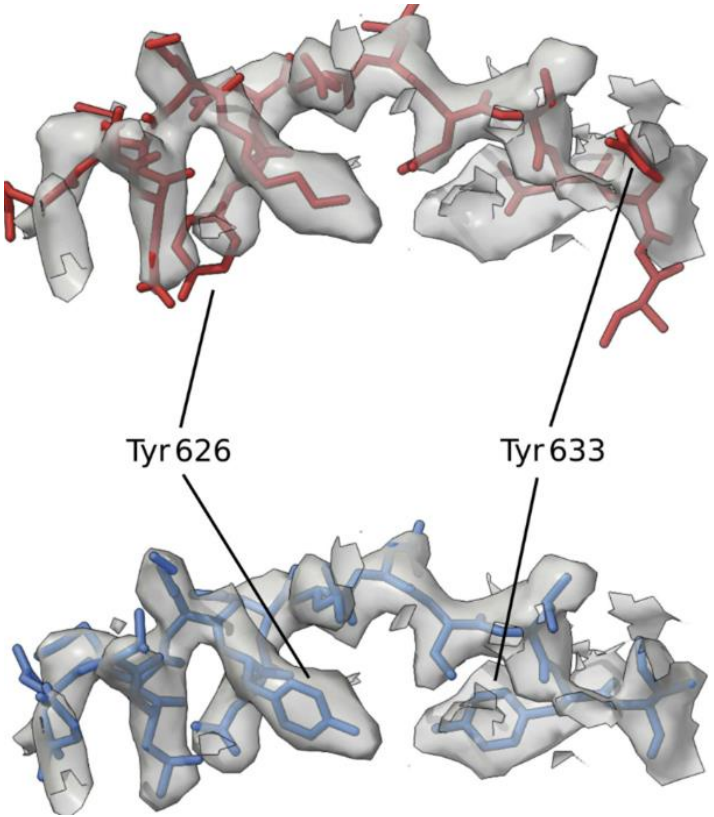

```
$ conkit-validate model.pdb prediction_af2.pkl sequence.fasta
```

Sánchez Rodríguez *et al.* (2022) Acta Cryst D, 78, 1412

New covariance-based metrics for model validation

Key points

- Is resolution- and modality-independent
- Suggests the register shift required to correct
- Limits
  - Still needs sufficient sequences for covariance signal
  - Fold-switching proteins
- 3 FP filters
- Errors validated by map, where resolution allows



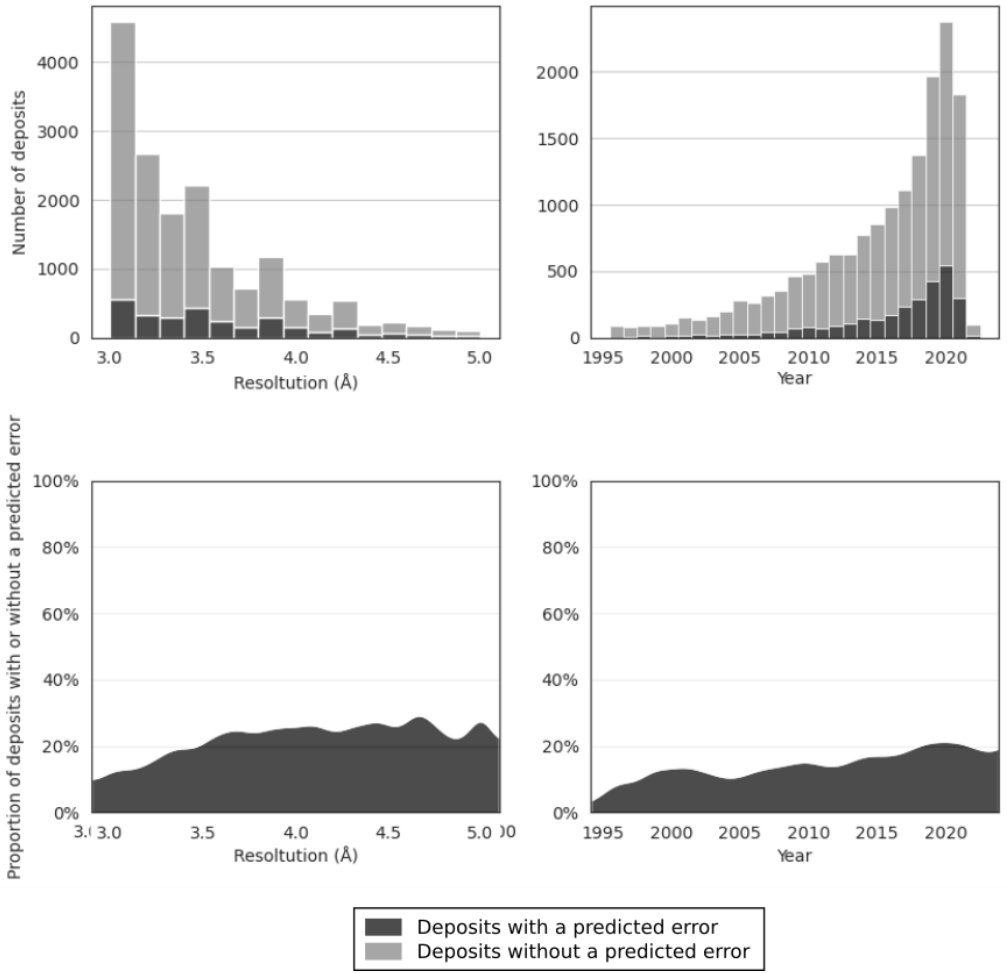
Sánchez Rodríguez *et al.* (2022) Acta Cryst D, 78, 1412

PDB-wide screen

Lower resolution structures tend to contain more errors, but dependence is quite modest

Error rate not declining recently - are experimentalists continually tempted to lower resolutions?

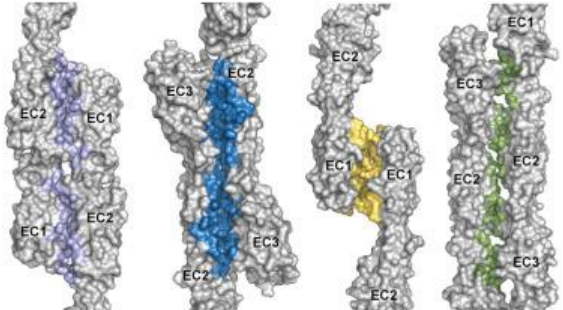
Cryo-EM structures contain more errors than crystal. Issue of local resolution, or maybe just because they are bigger?

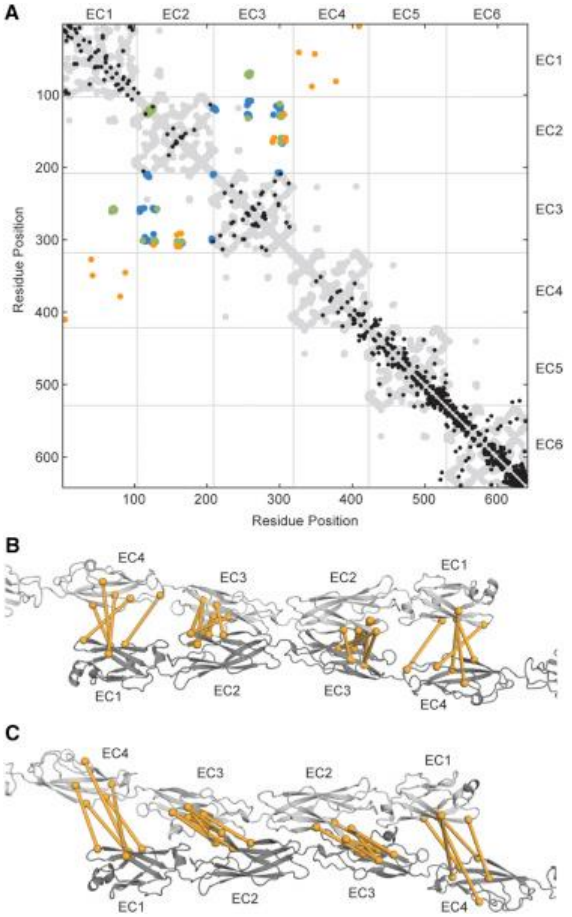


Sánchez Rodríguez *et al.* (2024) IUCrJ, 11, 938

Validating crystal structure contents

- PISA is an excellent general method, but contact predictions help in some cases
- Crystal showed various ways in which protocadherins could interact
- Predicted contacts supported only some of the modes
- PISA-cov is on the way…





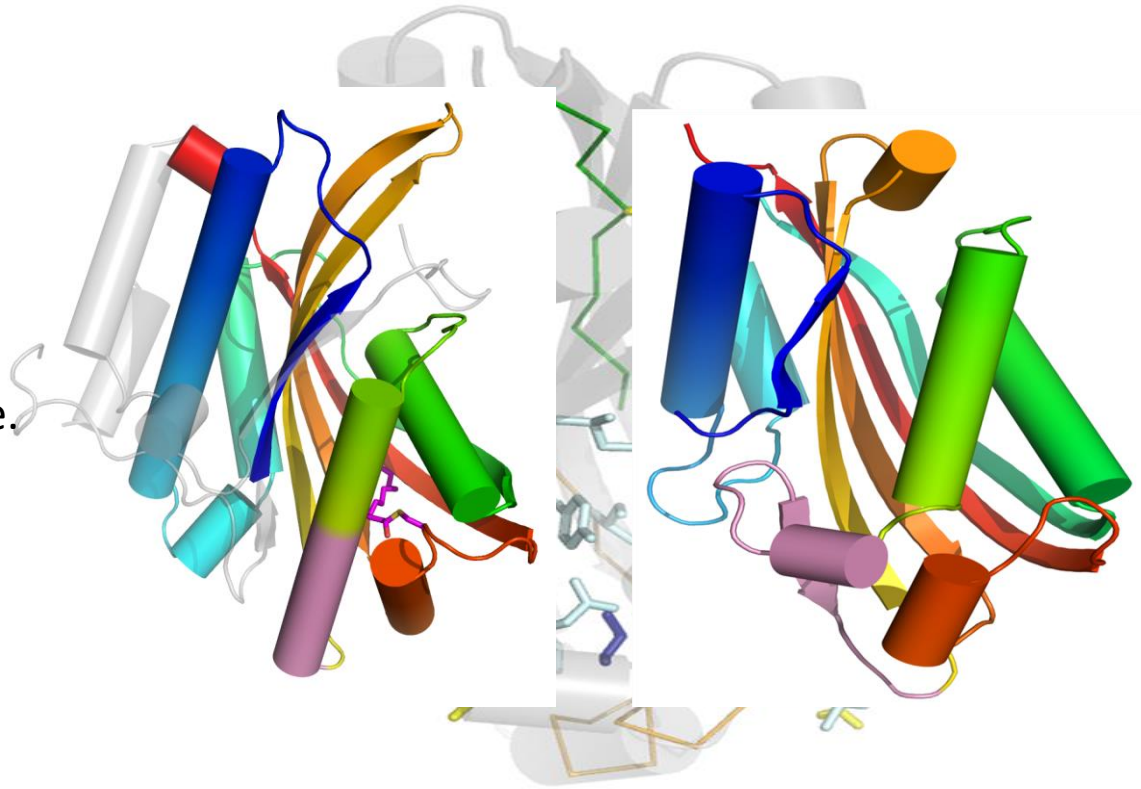
Nicoludis *et al*. (2015) Structure 23, 2087

Structure-based function interpretation

Where are the functional/catalytic sites?

Multiple methods in bioinformatics: Structure comparisons of Evf

- Reported as novel fold...
- ... but in fact related to *Bacillus* toxin structures (DALI)
- Both bind to host insect membranes
- Palmitate seen in Evf structure. Matches conserved region of toxins...
- GESAMT is an excellent CCP4 option
- FoldSeek is great for a quick search of the AFDB, ESMAtlas etc



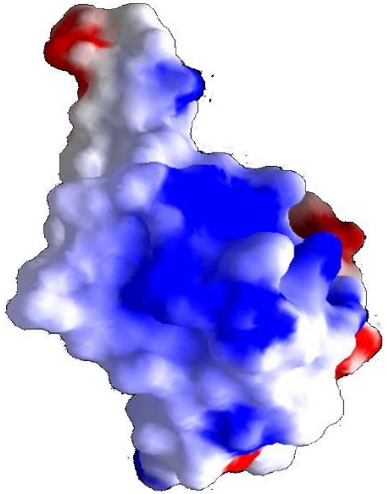
Quevillon-Cheruel et al. (2009) J. Biol. Chem. **35**,2107 Rigden (2009) FEBS Lett. **583**, 1555

Structure-based function annotation

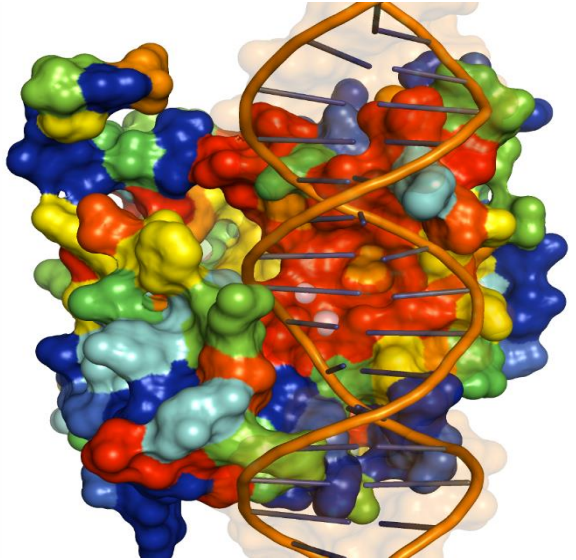
- Finding functional sites is based on their being different somehow to the rest of the protein surface. Important general methods are based on
  - Shape (castP, ProFunc, PyMOL)
  - Electrostatics (PyMOL, APBS)
  - Evolutionary conservation (Consurf)
- Less well-known but valuable characteristics are
  - Statistics of surface atom ‘triangles’ (STP)
  - Probe interaction energetics (ISMBLab)
  - Predicted pKa values (THEMATICS/POOL)
- New generation Deep Learning-based methods

Important general methods

Shape
CastP

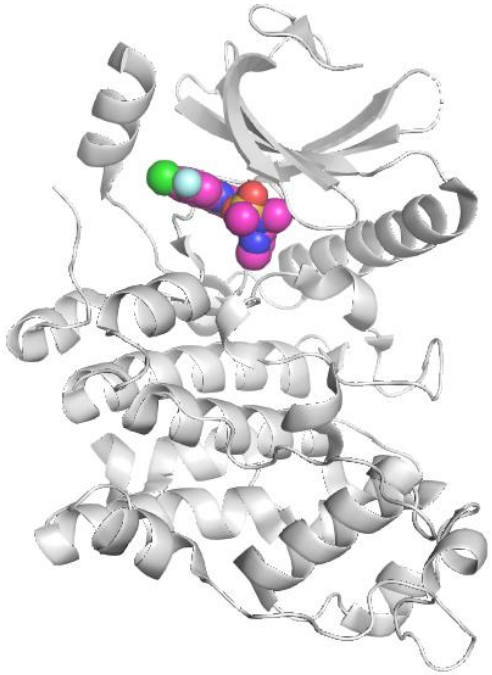


Electrostatics
APBS/PyMOL

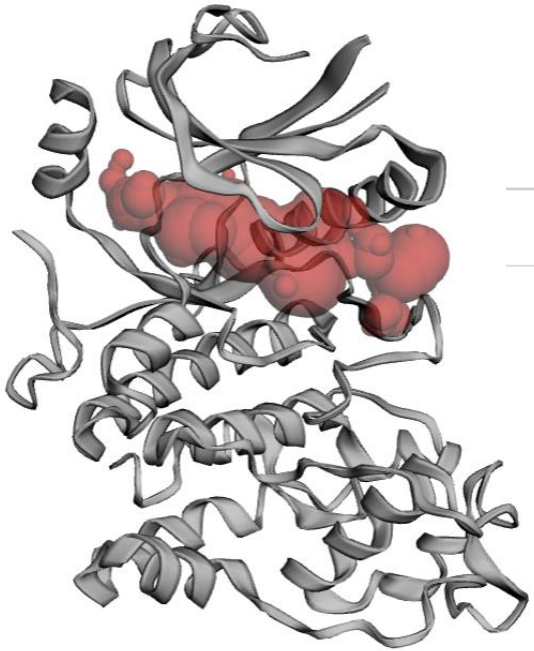


Conservation
ConSurf

Some servers require thought…



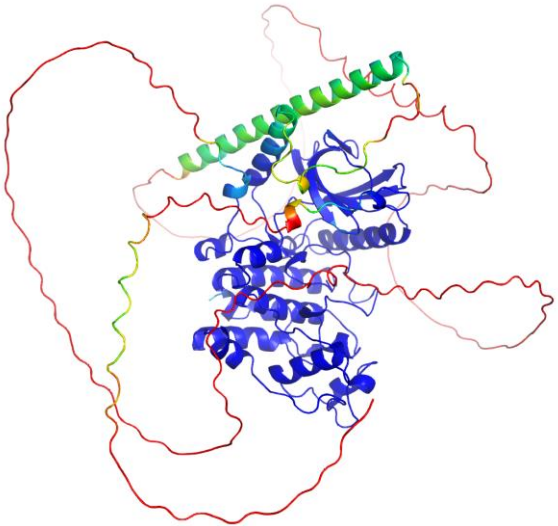
CastP



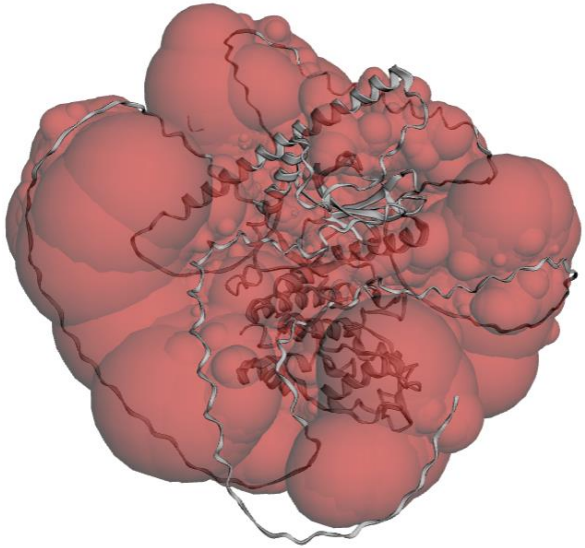
PocID	Area (SA) Å^2	Volume (SA) Å^3
1	762.617	506.394

Human SRSF protein kinase 2, PDB 7zkx

Some servers require thought…



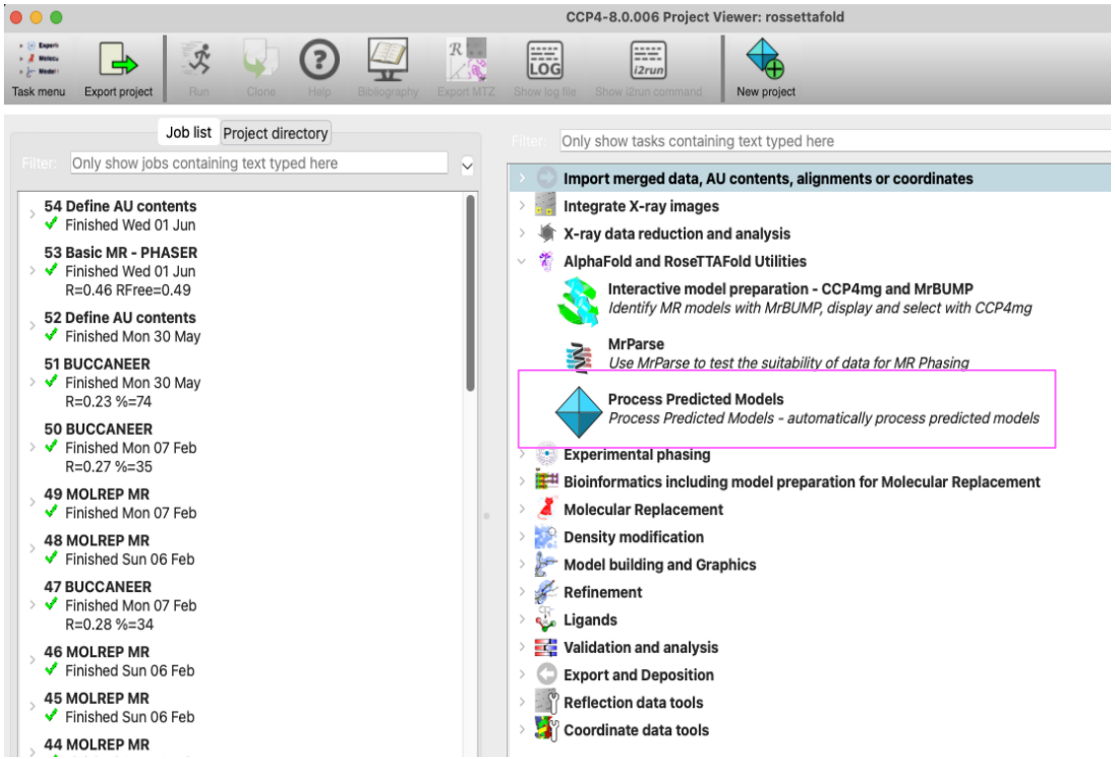
CastP



PocID	Area (SA) Å^2	Volume (SA) Å^3
1	21186.058	159841.419

Human SRSF protein kinase 2, AFDB AF-P78362-F1

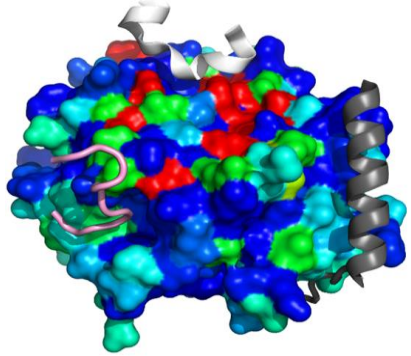
CCP4 i2 can remove the red spaghetti (and convert pLDDT to B-factor)



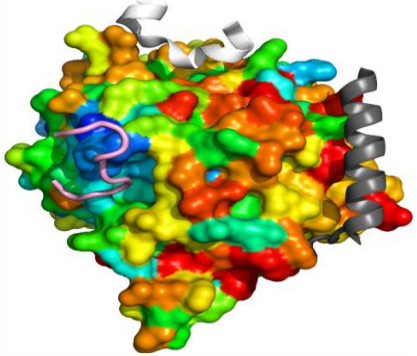
Some servers require thought…

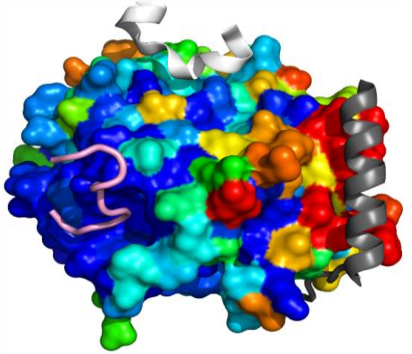
- Consurf maps sequence conservation onto a structure revealing functional sites
- Excellent, general method, but results depend on sequence set chosen for mapping: selecting all or only near relatives gives different results. Either might be more appropriate for you

Mapping 300 homologues mixes different activities so no information on binding sites

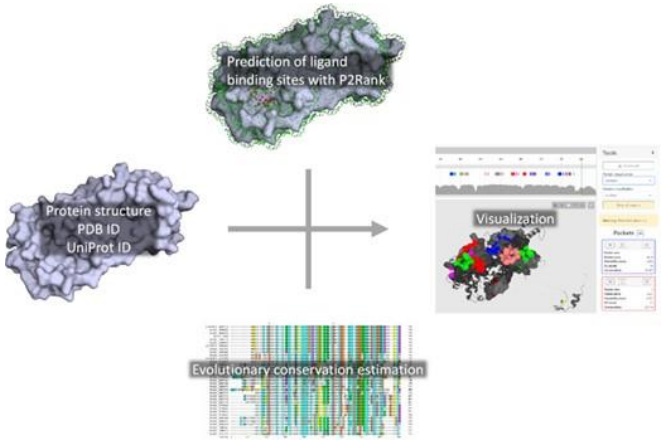


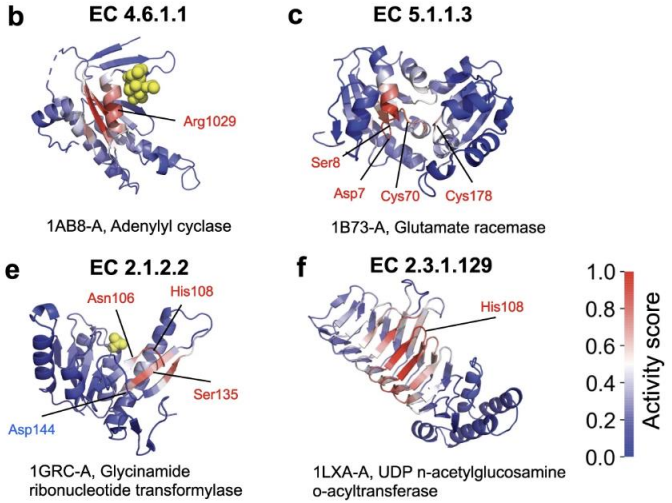
But restricting to a single protein family shows only ‘pink’ site is function in both Diptera and Lepidoptera





Advanced methods use Machine Learning and/or multiple signals





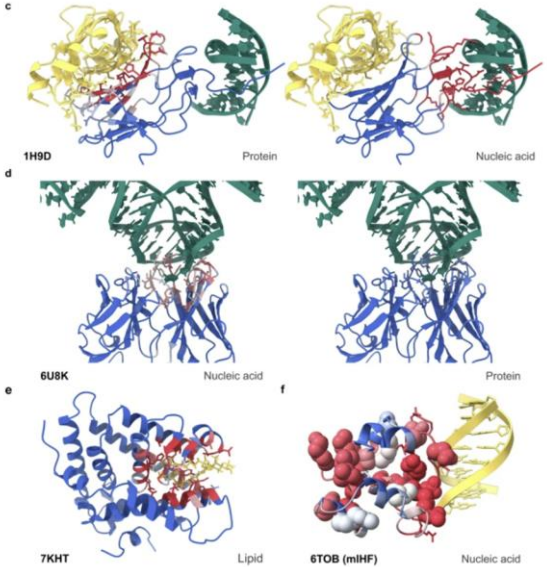
PrankWeb uses structural and physicochemical properties then displays pockets with conservation analysis https://prankweb.cz

Combines LM and contact map features in CNN to predict GO terms and sites

https://beta.deepfri.flatironinstitute.org/

My favourite Deep Learning-based methods

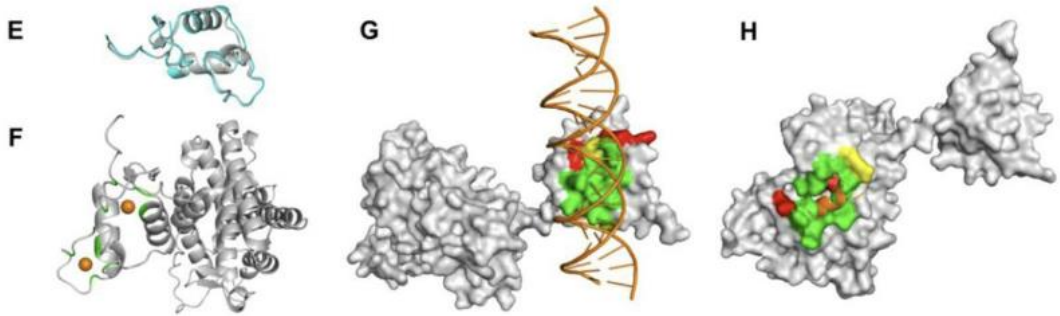
PESTO



Geometry and contacts of input structure analysed. Interface propensities displayed

Krapp et al (2023) Nature Comms 14, 2175
https://pesto.epfl.ch/

GPsite

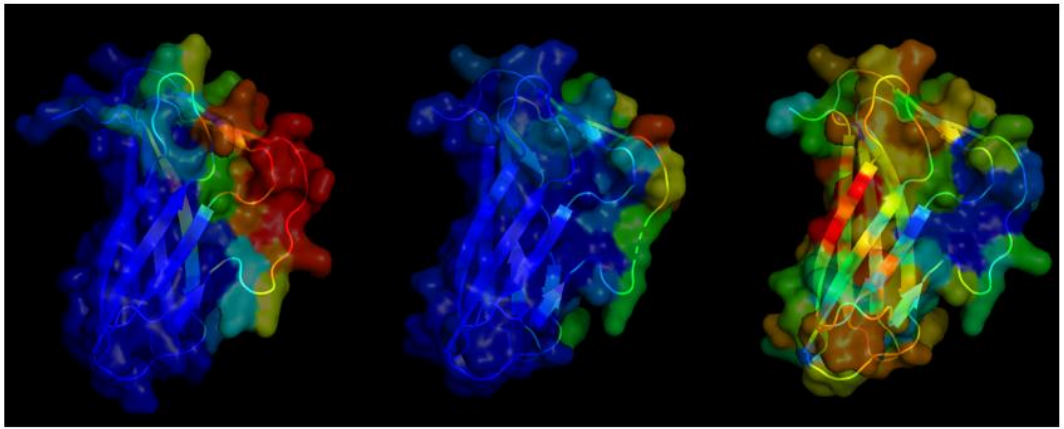


True Positive False Positive False Negative

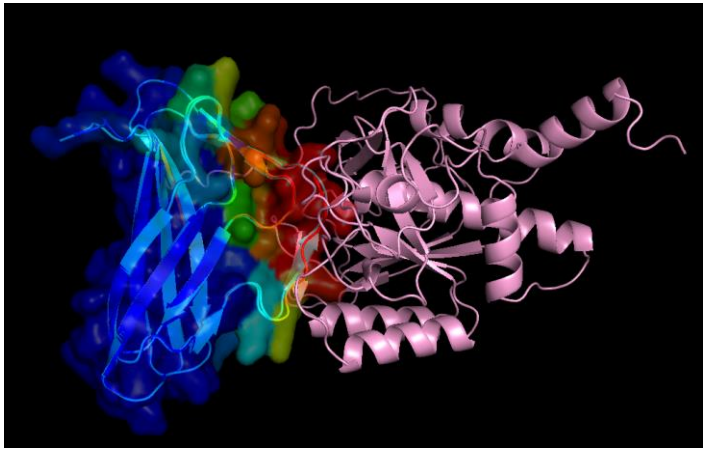
Sequence modelled with ESMFold. Various binding sites and GO terms predicted

Yuan et al (2024) eLife https://doi.org/10.7554/eLife.93695.1
https://bio-web1.nscc-gz.cn/app/GPSite

Multiple methods in bioinformatics: Structure-based function methods



GPsite PESTO Consurf

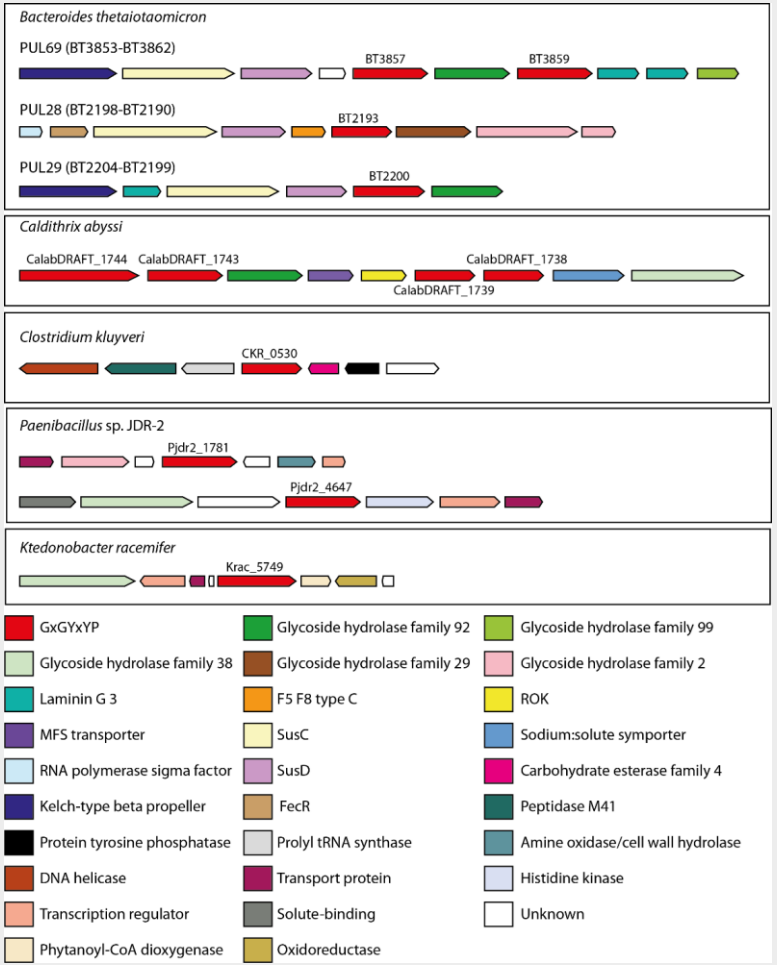


All methods pick out the very probable binding site of this protease inhibitor. DL methods give a somewhat cleaner signal than general conservation

Case study from Structural Genomics

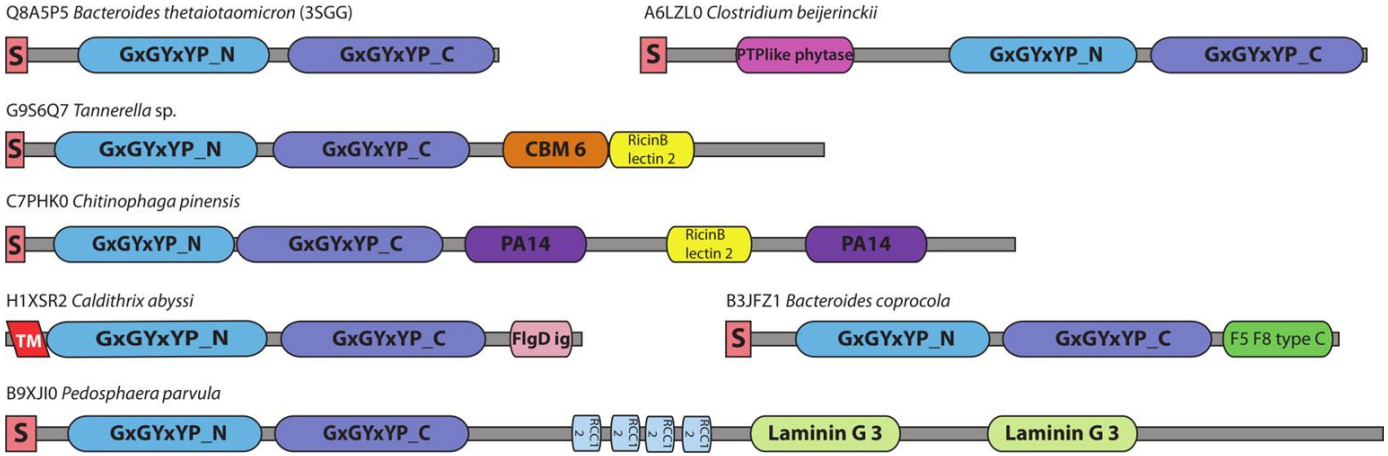
GxGYxYP proteins

- Named for a conserved sequence motif. Molecular function unknown
- Over-represented in gut bacteria
- Found in Polysaccharide Utilization Loci in *Bacteroides thetaiotaomicron*
- *Q: What does the protein do?*



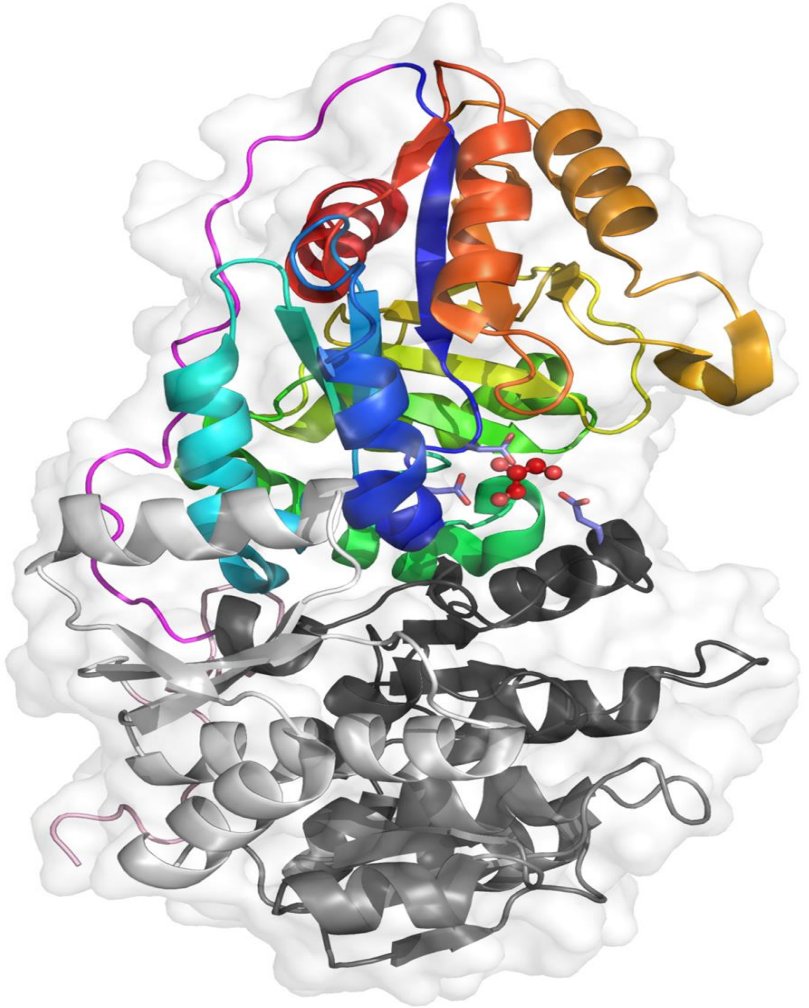
GxGYxYP proteins

- Domain architectures also predict carbohydrate connection



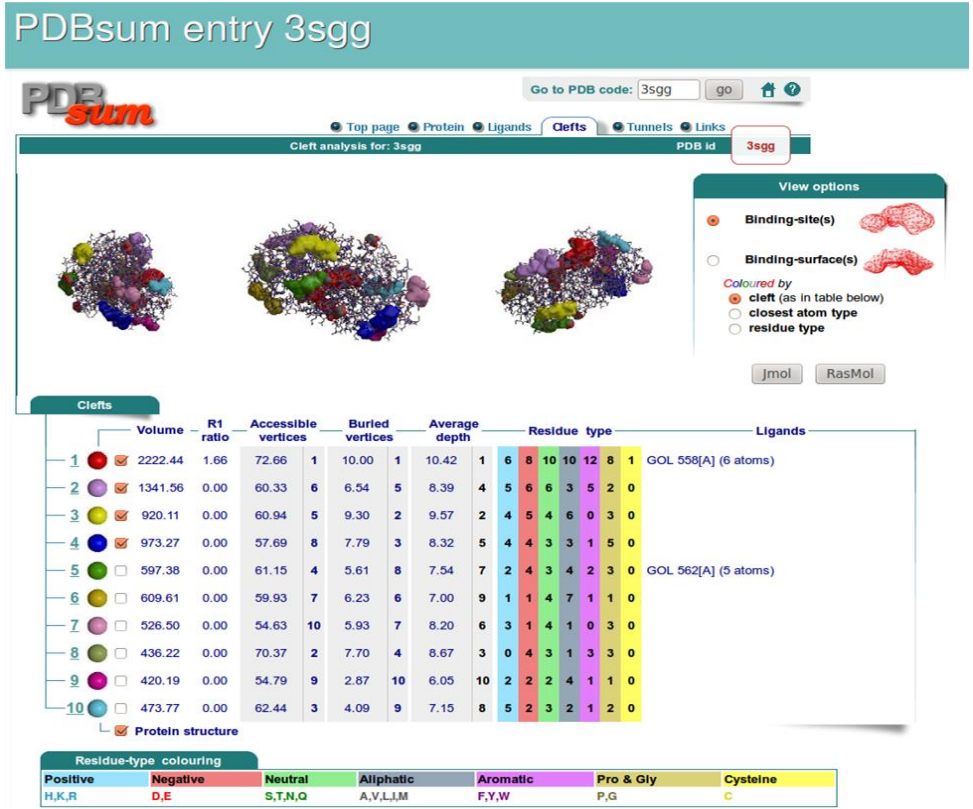
GxGYxYP proteins

- Overall folds don't help much
- TIM barrel + 3 x novel α+β unit
- TIM barrel DALI Z >13 for
  - Allantoinase
  - Polysaccharide deacetylase
  - Glycosyltransferase



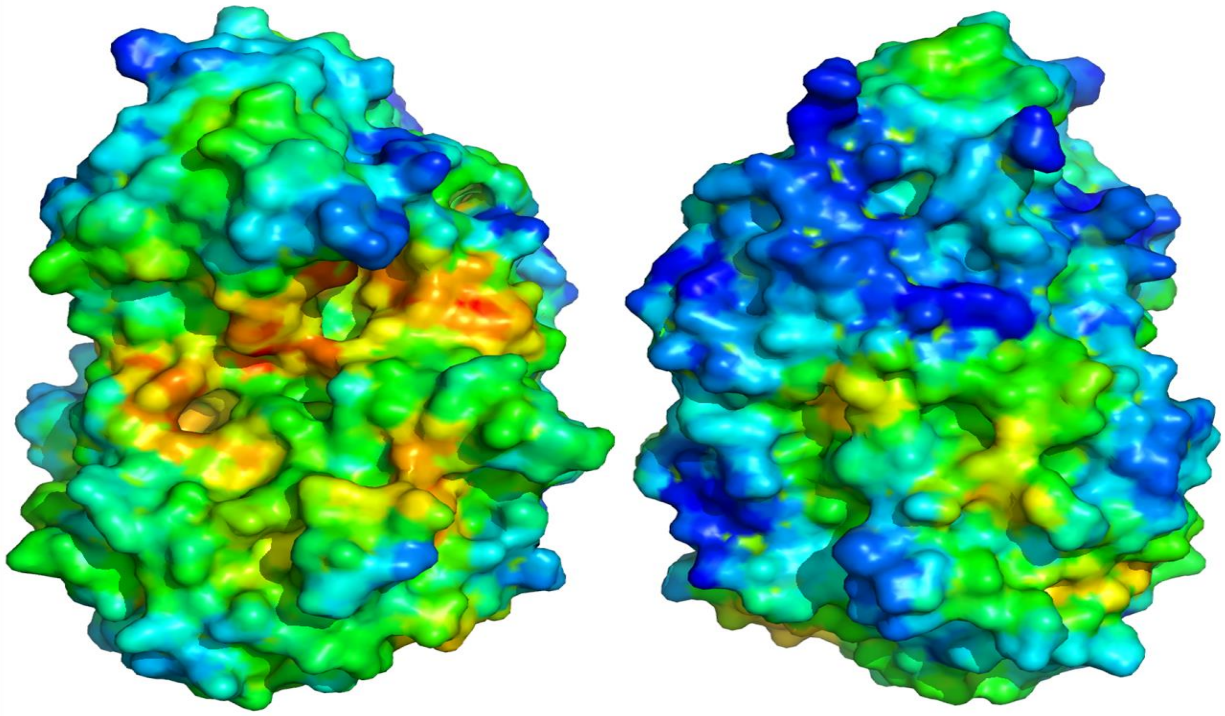
GxGYxYP proteins

- Largest cavity lies between domains
- Glycerol from crystallisation solution in it



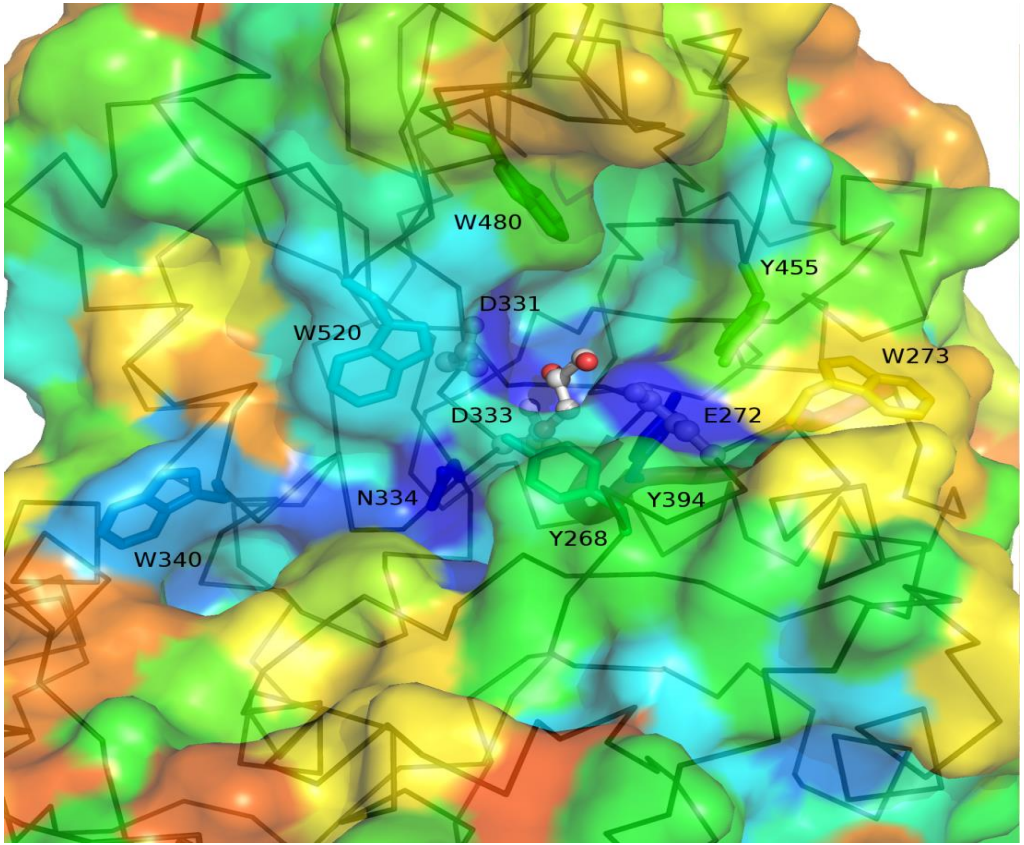
GxGYxYP proteins

- Largest cavity lies between domains
- Glycerol from crystallisation solution in it
- Picked out by non-geometry based STP (surface triplet propensities)



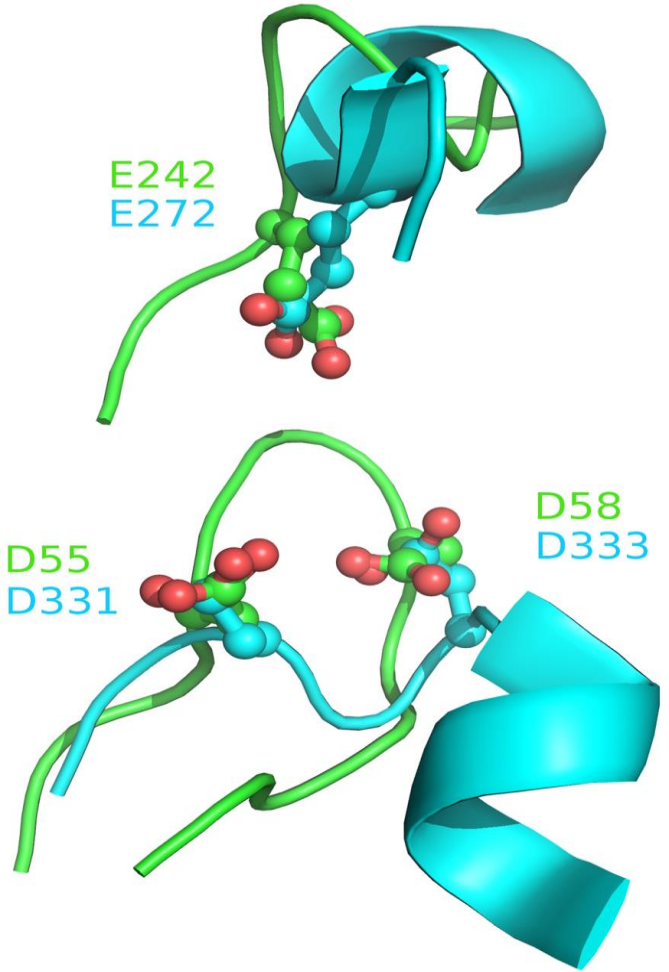
GxGYxYP proteins

- The patch is conserved
- Contains multiple aromatic residues, often surface lying in carbohydrate binding sites…



GxGYxYP proteins

- The patch is conserved
- Contains multiple aromatic residues, often surface lying in carbohydrate binding sites…
- … and acidic residues resembling known glycosidase site… (SPRITE)
- … and with perturbed pKa values (THEMATICS)



cellulase
3sgg

Multiple methods in bioinformatics

GxGYxYP conclusion

- GxGYxYP is a novel Glycoside Hydrolase family
  - Genome context
  - Domain composition
  - Cavity
  - Bound glycerol
  - STP
  - Conservation
  - Match to known GH catalytic site
  - pKa perturbation

Structure-based methods

Rigden et al. (2014) BMC Bioinformatics 15:196

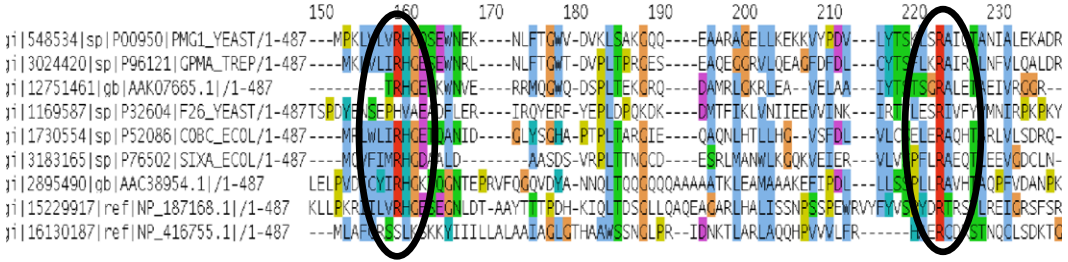
…and finally, you're putting a manuscript together

Calculating and presenting sequence alignments

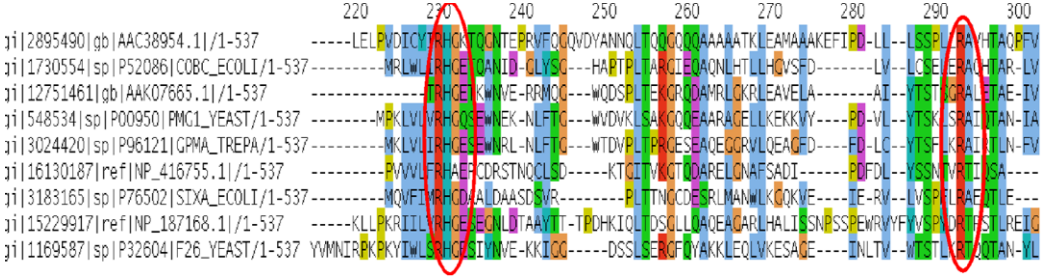
Your sequence alignment

- Don’t use ClustalW! It’s 24 years old! Modern methods like MUSCLE, Probcons and MAFFT are much better

ClustalW misses relatively obvious RHG motif in some of diverse sequence set…

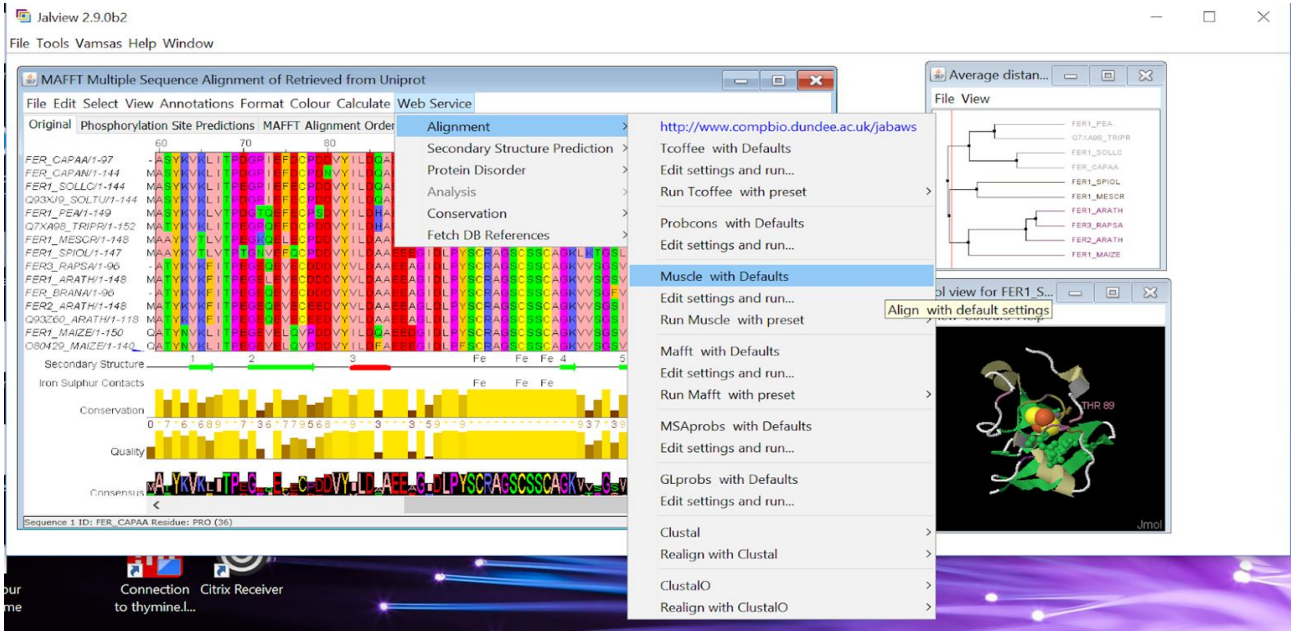


… but MUSCLE gets it



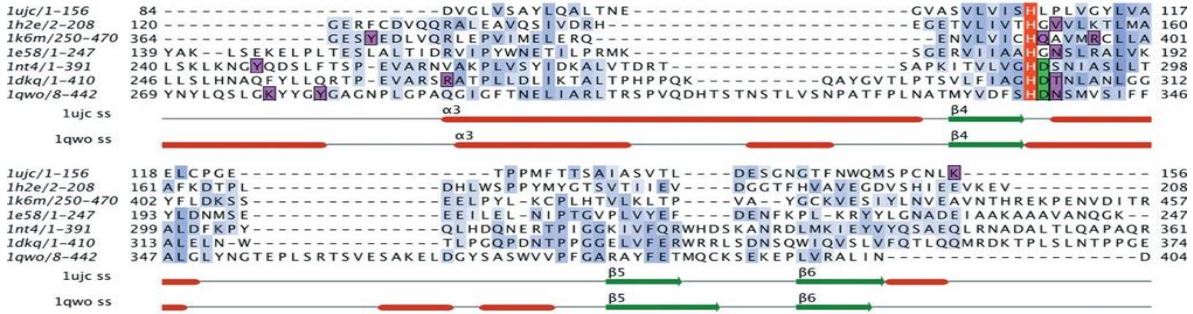
Multiple methods in bioinformatics : Jalview.org

- All these alignment methods and more are available through Jalview on Dundee servers



Jalview

- Also helps you produce figures like this…



- … rather than like this

Don’t forget to cite it (and all your bioinformatics)!

Waterhouse et al (2009), *Bioinformatics* **25**, 1189

Questions?

drigden@liv.ac.uk