Bioinformatics for structural biologists

Prof Dan Rigden

UNIVERSITY OF LIVERPOOL

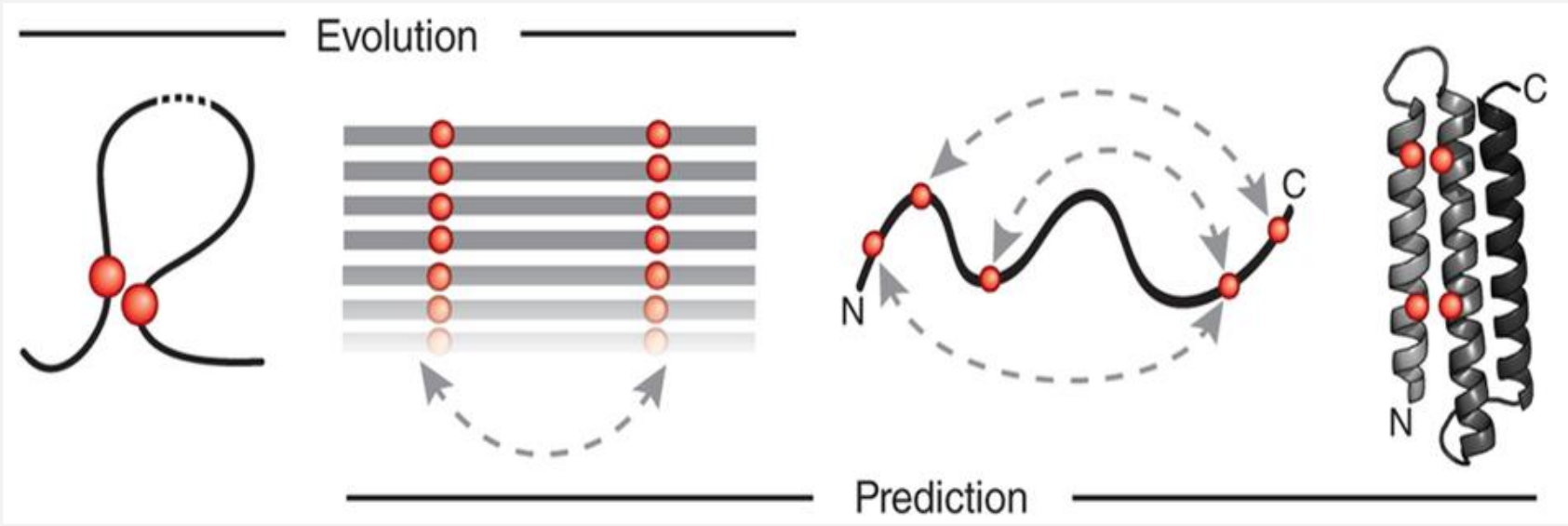
- Introduction
  - Predicting contacts from sequences using evolutionary covariance
  - Deep Learning-based structure prediction eg AlphaFold 2/3
- Bioinformatics throughout structure determination
  - Predicting domain structure
    - Construct design, experimental strategy
    - Protein engineering
  - ~~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

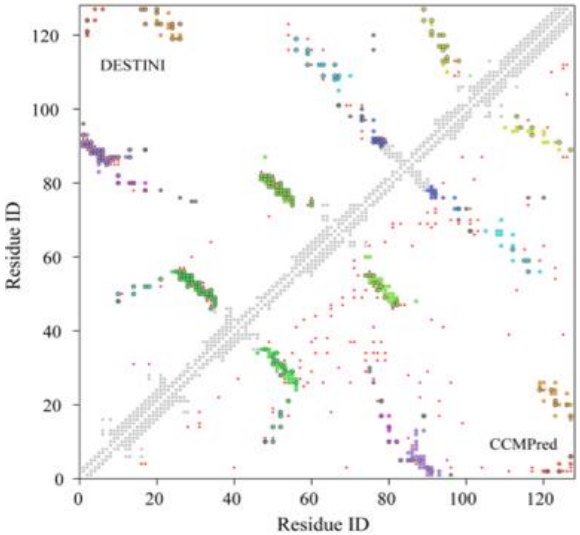
Deep Learning-based structure prediction methods

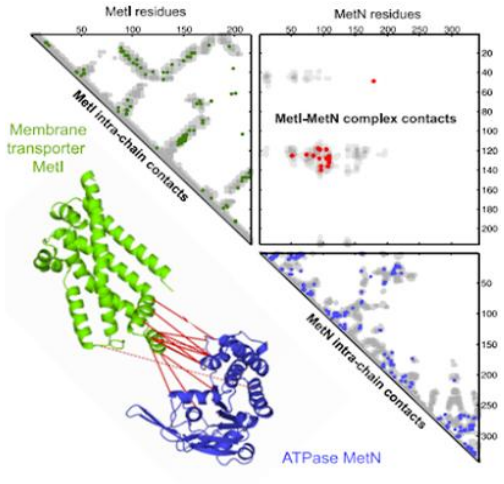
Evolutionary covariance

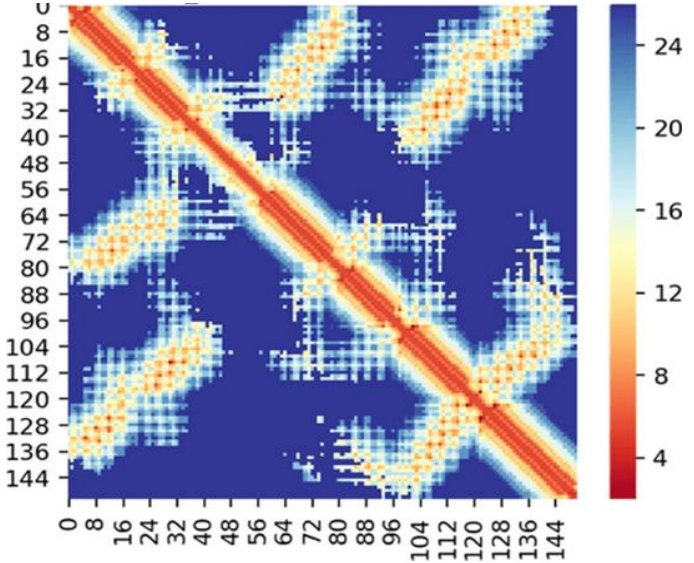


Marks *et al.* (2012) Nature Biotech 30, 1072

Predicting contact maps and distograms







Drove modelling by: EVFold, DMPfold etc

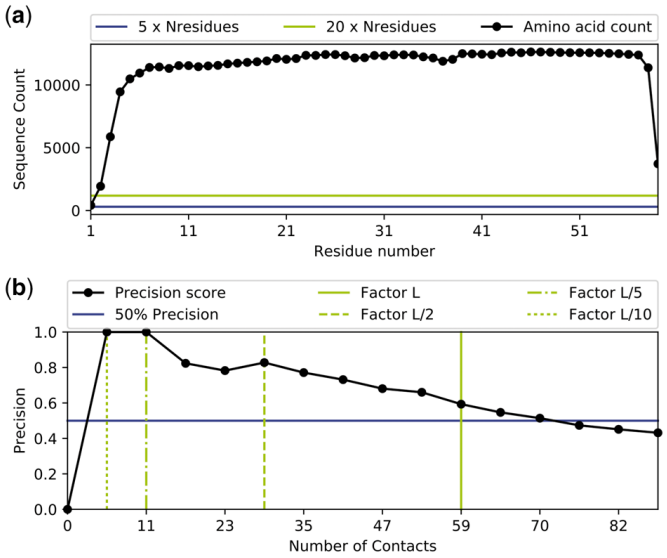
trRosetta, AlphaFold (1)

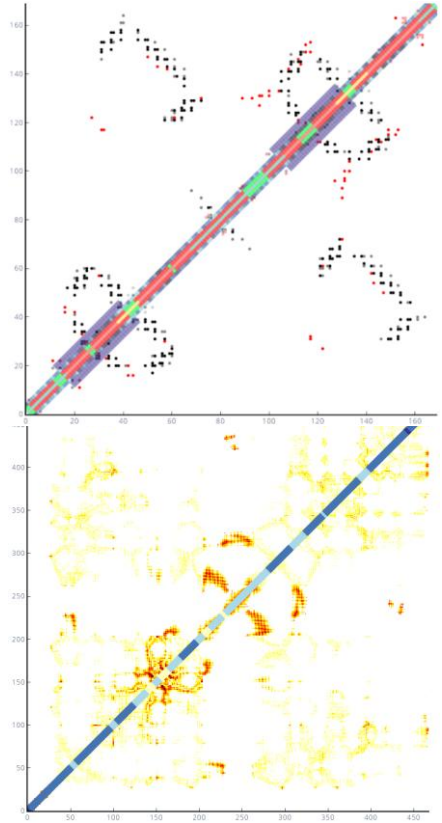
Gao et al (2019). *Sci Rep* **9,** 3514

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

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

ConKit and ConPlot.org



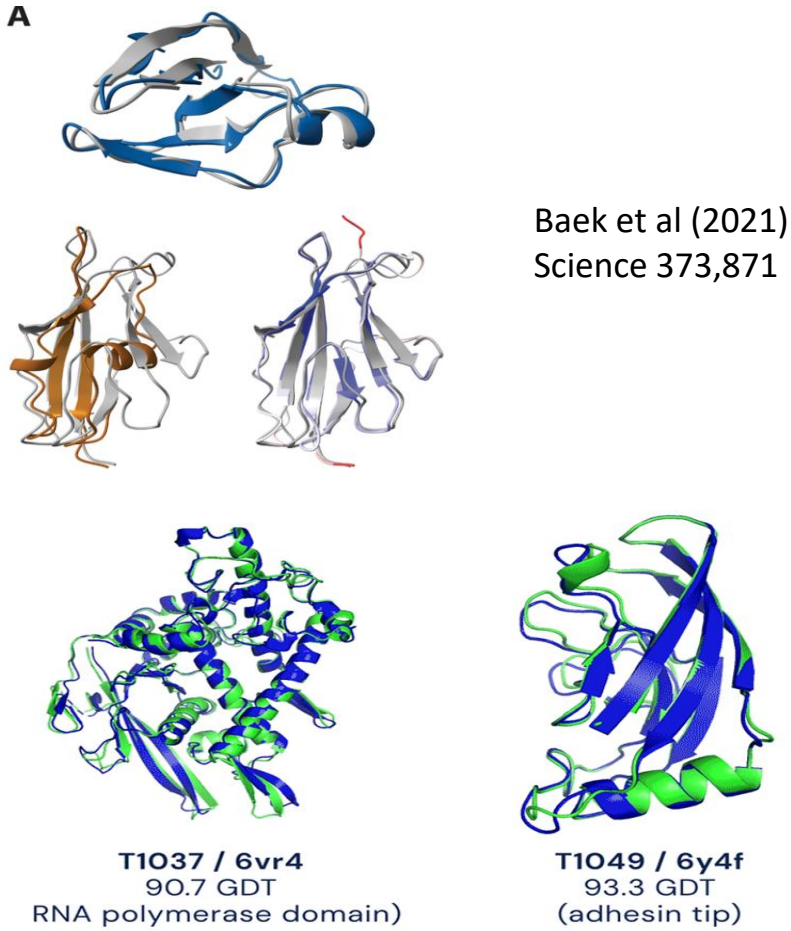


Simkovic *et al.* (2017) Bioinformatics **33** 2209

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

Multiple methods: RoseTTAFold and AlphaFold 2

- Still use information including evolutionary 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



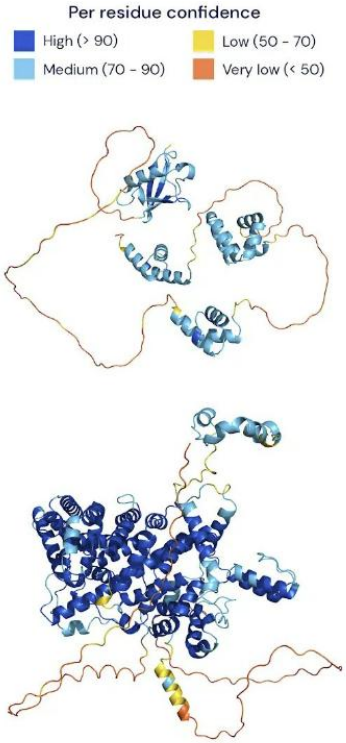
Baek et al (2021) Science 373,871

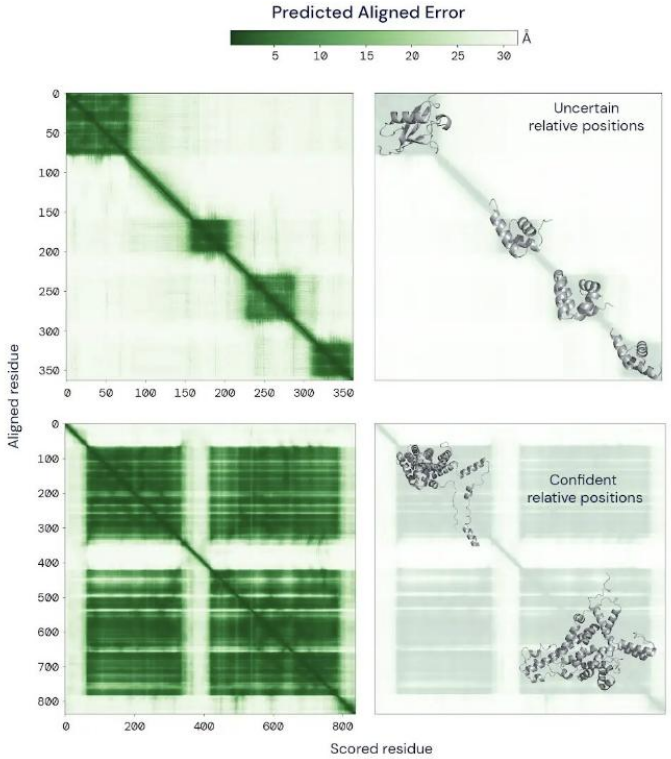
Jumper et al (2021) Nature 596,583

Three outputs from latest methods

- Coordinates *plus*

pLDDT is a measure of local confidence in the environment of a residue

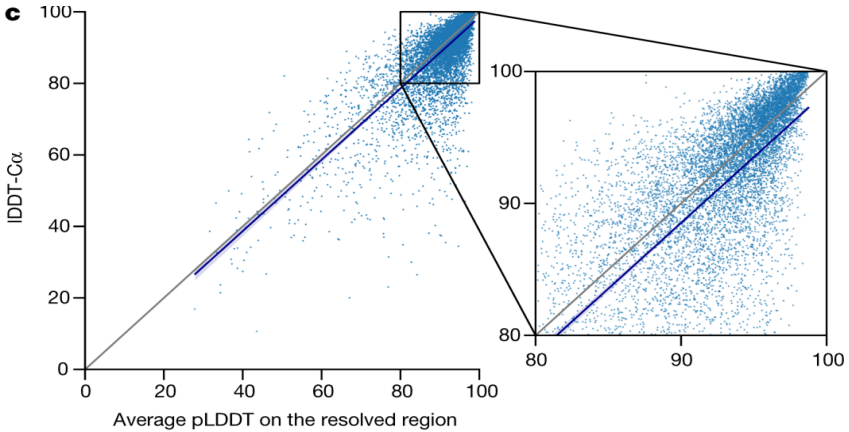




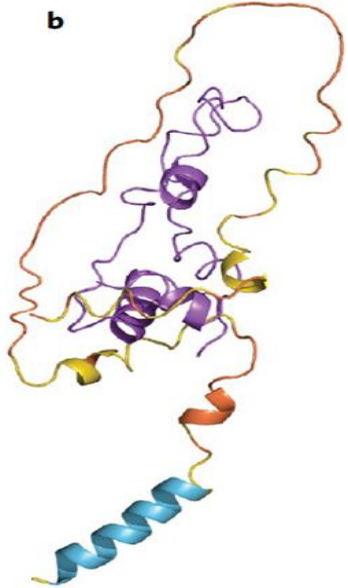
The PAE (EPE) informs on the global eg inter-domain confidence

AlphaFold 2 models

- Often amazing quality but not always… Need good MSA (evolutionary covariance) or template!
- pLDDT is a very good estimate of local quality

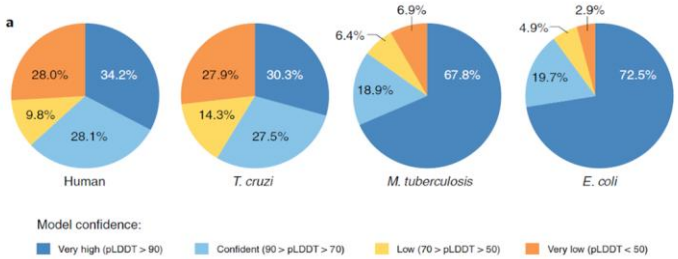


Jumper *et al.* (2021) Nature 596, 583



Phosphatase crystal structure vs AF2 model

Insulin crystal structure vs AF2 model



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

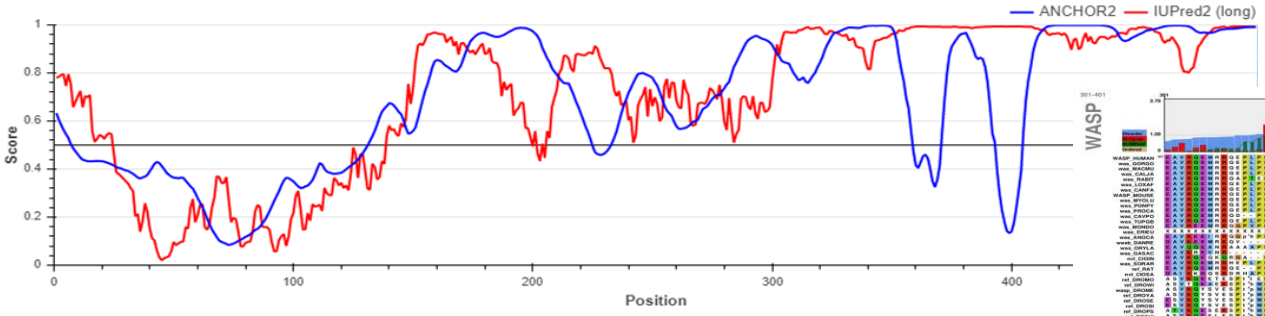
Domains, construct design and protein engineering

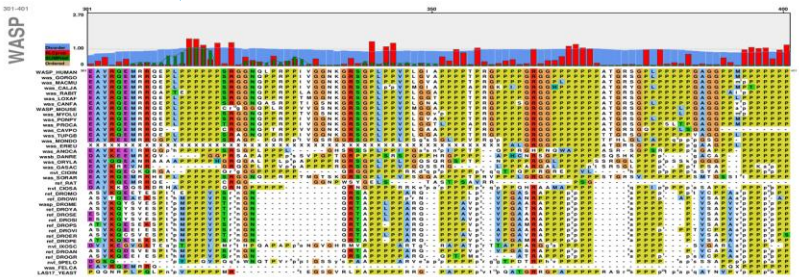
Which part to express for crystallisation?

If necessary, can solubility etc be improved?

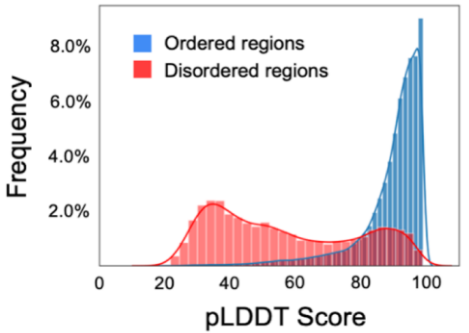
Intrinsic disorder prediction

- 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)

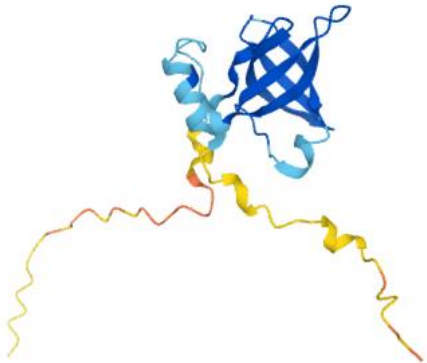


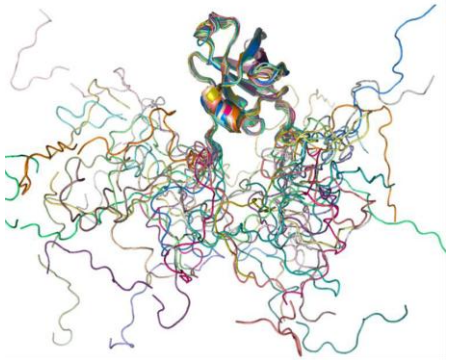


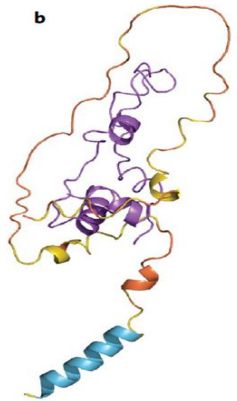
Most low-confidence AF2 regions are disordered but they may just be plain wrong



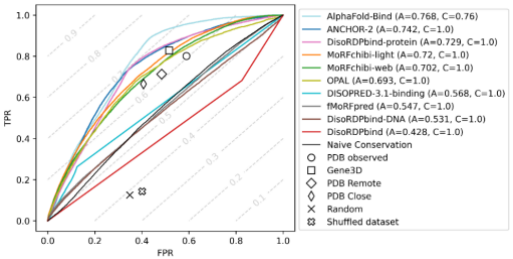
Binder et al (2022) COSB 74, 102372





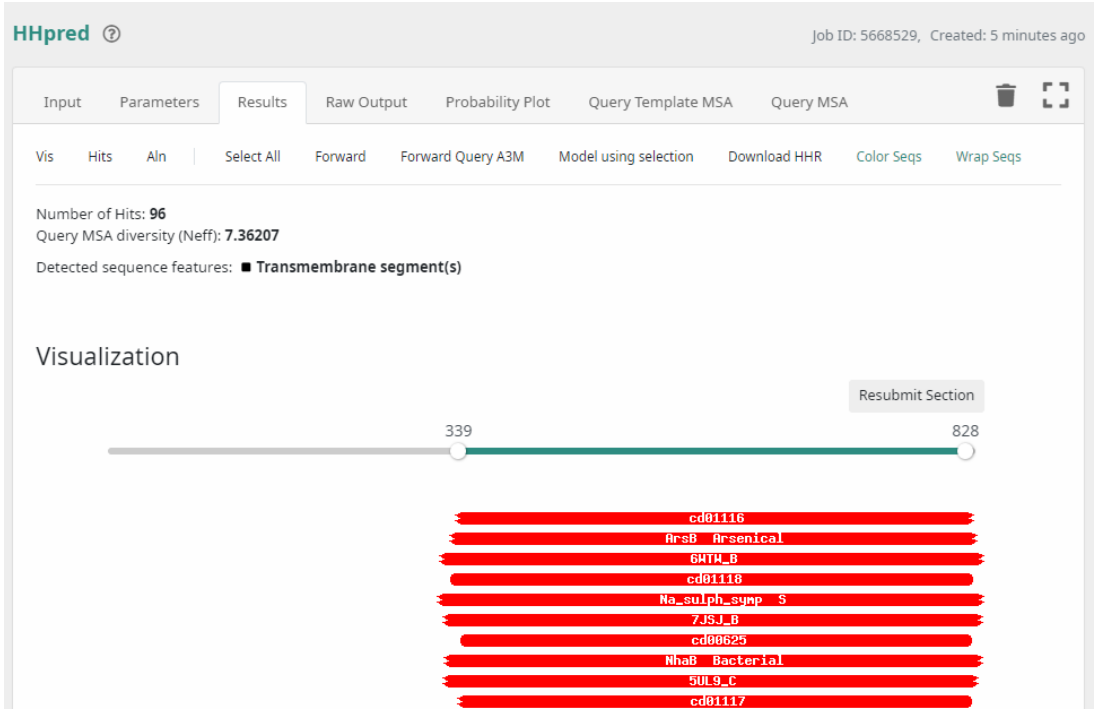


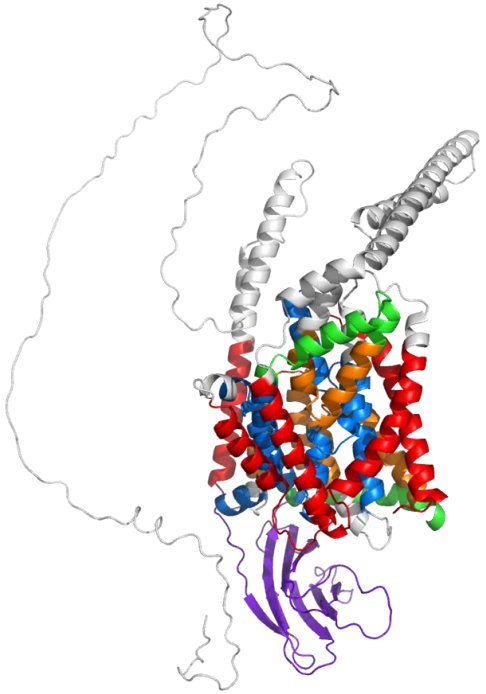
Relatively high pLDDT regions in AF2 disordered regions may predict interaction motifs



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

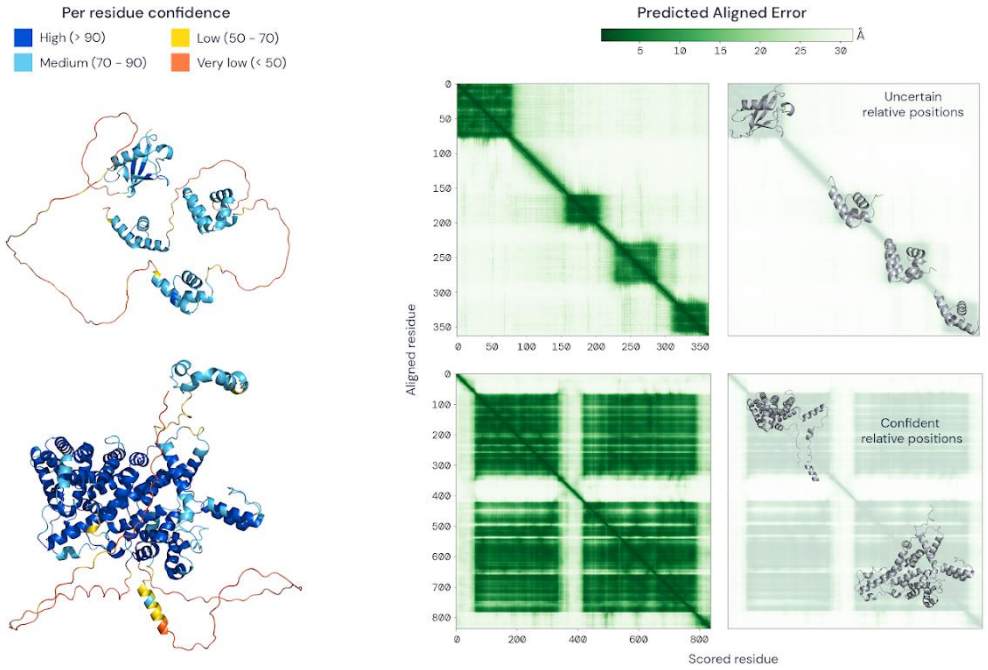
Beyond sequence matching: AF2 for domain discovery





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

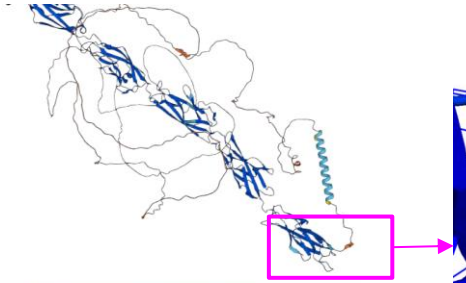
AF2 for domain boundary prediction

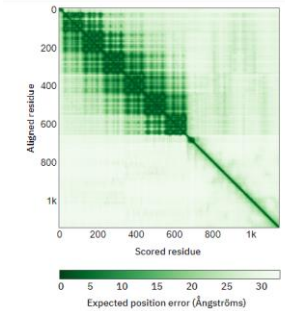


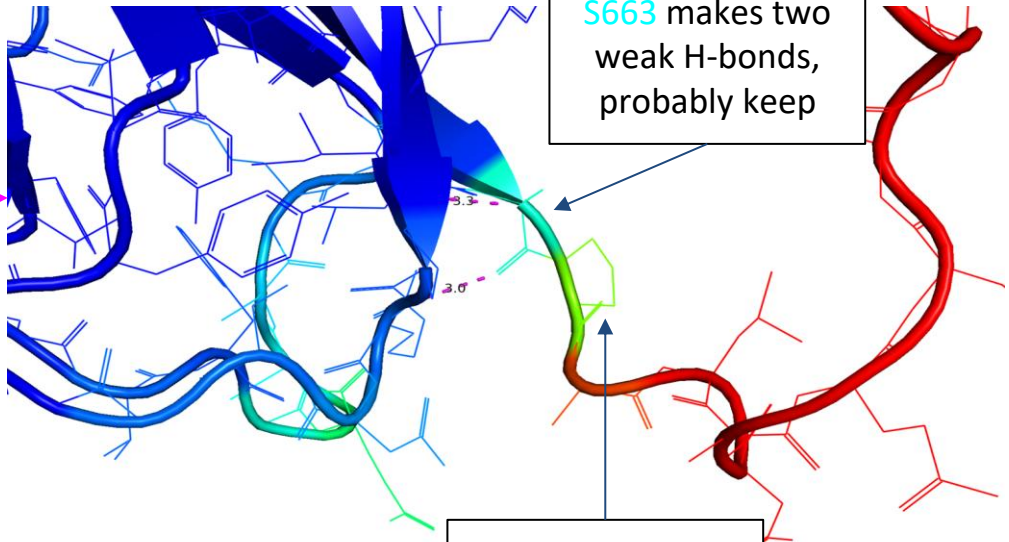
Slice’n’Dice uses the PAE or other kinds of clustering to split structures into domains for MR

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

Fine details too, if reliability allows





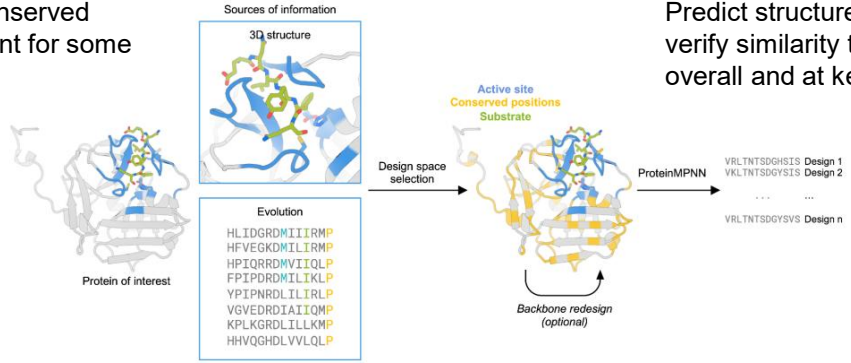


S663 makes two weak H-bonds, probably keep

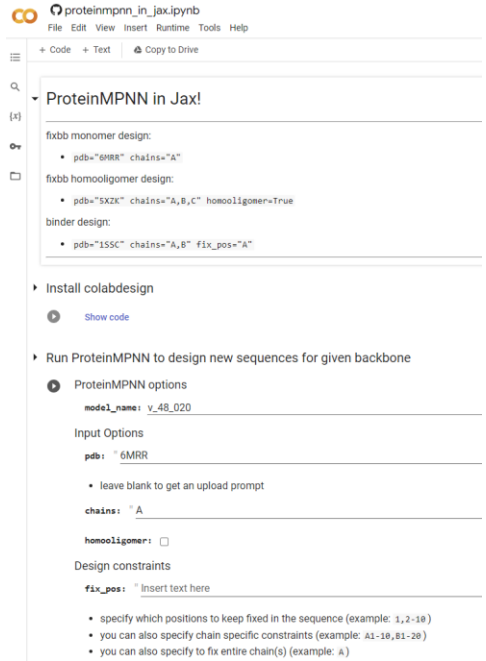
P664 onward make no contacts, delete

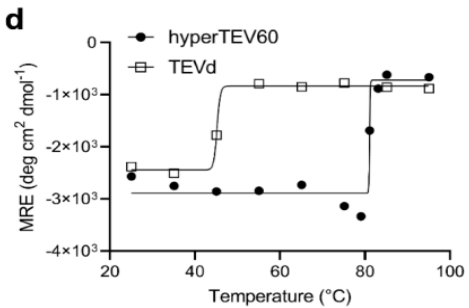
Protein engineering with ProteinMPNN for improved properties

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

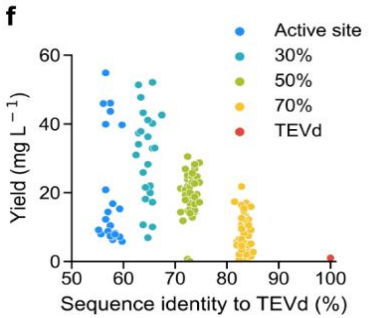


Allow ProteinMPNN to fill in the gaps and design variants. Predict structures with AF2 and verify similarity to starting point, overall and at key positions

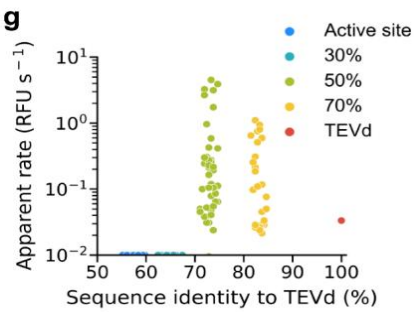




Better thermostability



Better expression

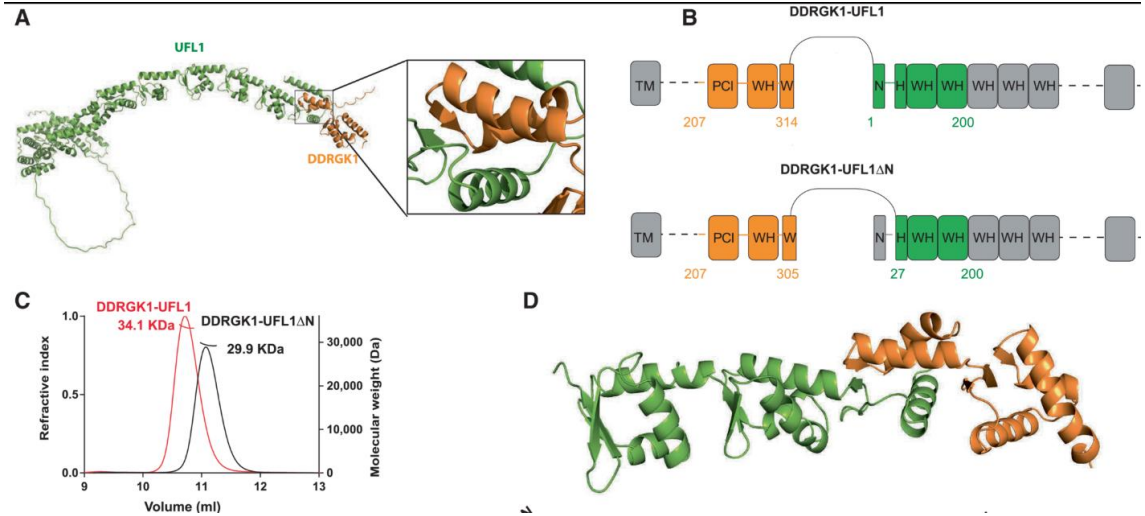


Better activity

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

Protein engineering: rational design of a fusion protein

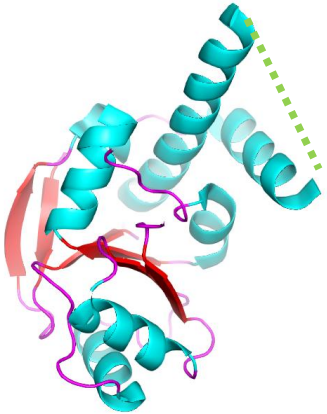
- First, AF2 used to predict the complex of URL1 and DDRGK1
- Designed linker to remove low-pLDDT (disordered) N-terminus of UFL1 and join
- AF2 confirmed viability of design and prediction matched crystal structure



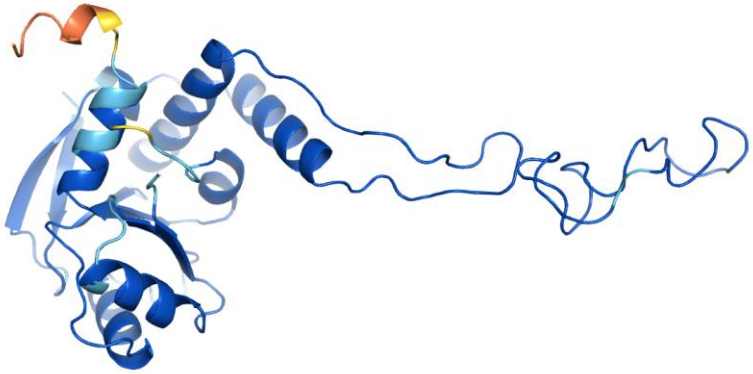
Banerjee et al (2023) EMBO reports 24: e56920

Protein engineering: rational removal of disordered loop

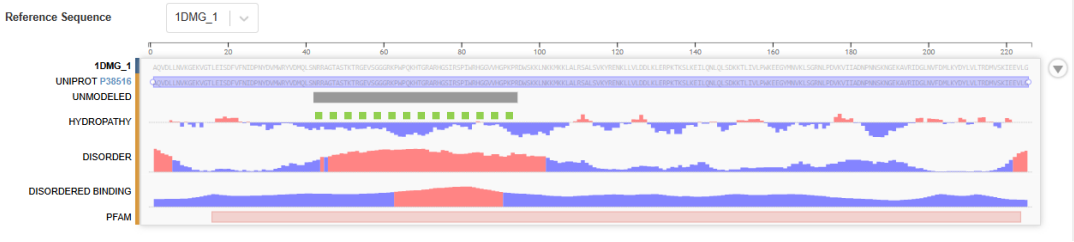
Ribosomal protein L4 has a large disordered loop that may hamper study



PDB:1dmg



AFDB: P38516



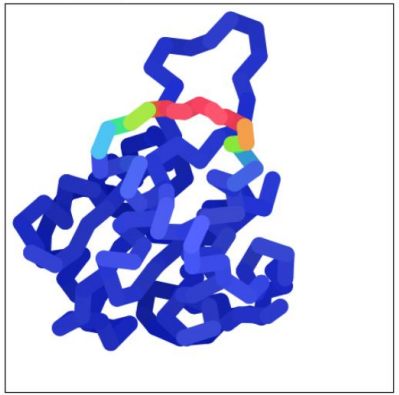
Protein engineering: rational removal of disordered loop

> run **RFdiffusion** to generate a backbone

```
name:        remove_disordered_loop
contigs:     'A2-42/15/A105-223'
pdb:         '1DMG'
iterations:  50
hotspot:     Insert text here
num_designs: 4
visual:      image
```

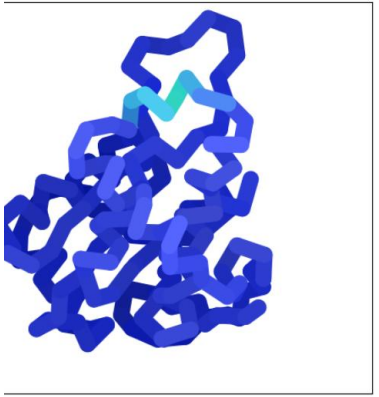
```
mode: fixed
output: outputs/remove_disordered_loop
contigs: ['A2-42/15-15/A105-223']
./RFdiffusion/run_inference.py inference.output_prefix=outputs/re
```

running



```
fixed
t: outputs/remove_disordered_loop
gs: ['A2-42/15-15/A105-223']
iffusion/run_inference.py inference.output_prefix=outputs/
```

running



> run **ProteinMPNN** to generate a sequence and **AlphaFold** to validate

ProteinMPNN Settings

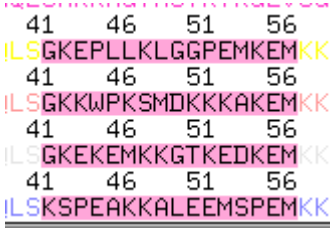
```
num_seqs:          8
mpnn_sampling_temp: 0.1
rm_aa:             C
use_solubleMPNN:   ☐
```

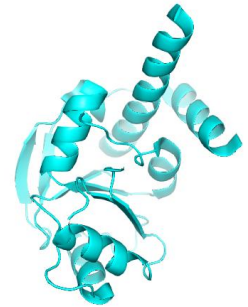
- `mpnn_sampling_temp` - control diversity of sampled sequences. (higher = more diverse).
- `rm_aa='C'` - do not use [C]ysteines.
- `use_solubleMPNN` - use weights trained only on soluble proteins. See preprint.

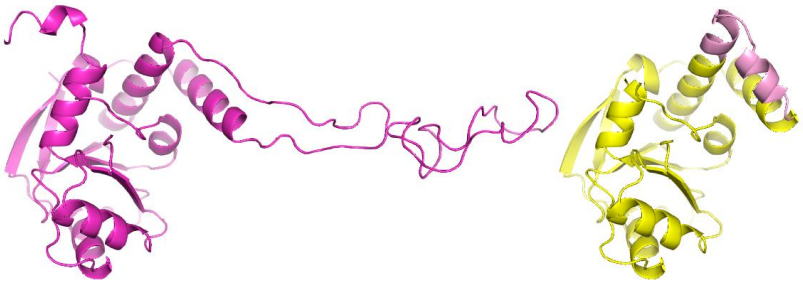
https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb

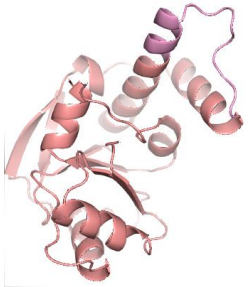
Protein engineering: rational removal of disordered loop

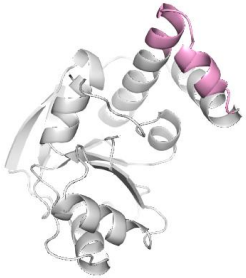
Designed connections, mainly helical, have different sequences

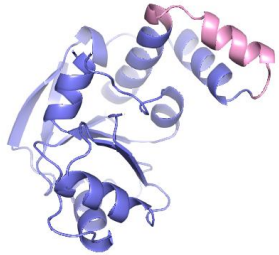








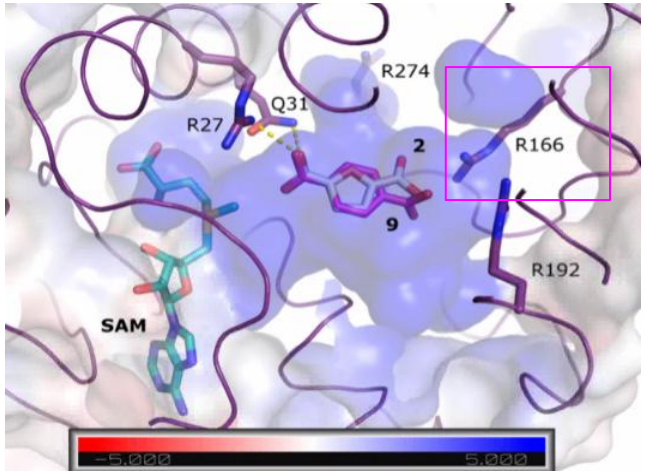


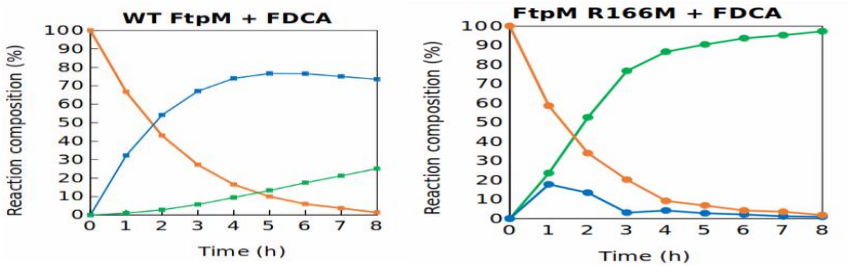


Protein engineering: AF model guides mutation

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

Methyltransferase R166M mutation improves binding of monomethyl intermediate, expediting formation of bismethylated derivative





Ward et al (2023) ChemSusChem 16, e202300516

Tertiary structure

AF2/3 and RF output as search models for MR

AlphaFold etc 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

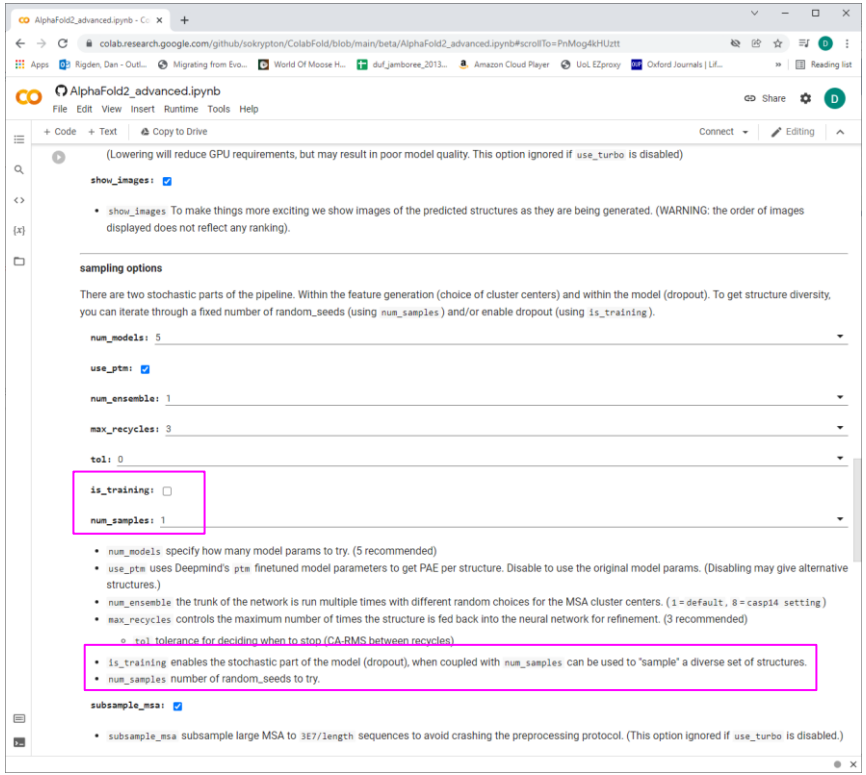
Heo & Feig (2022) PSFB DOI: 10.1002/prot.26382

- 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.

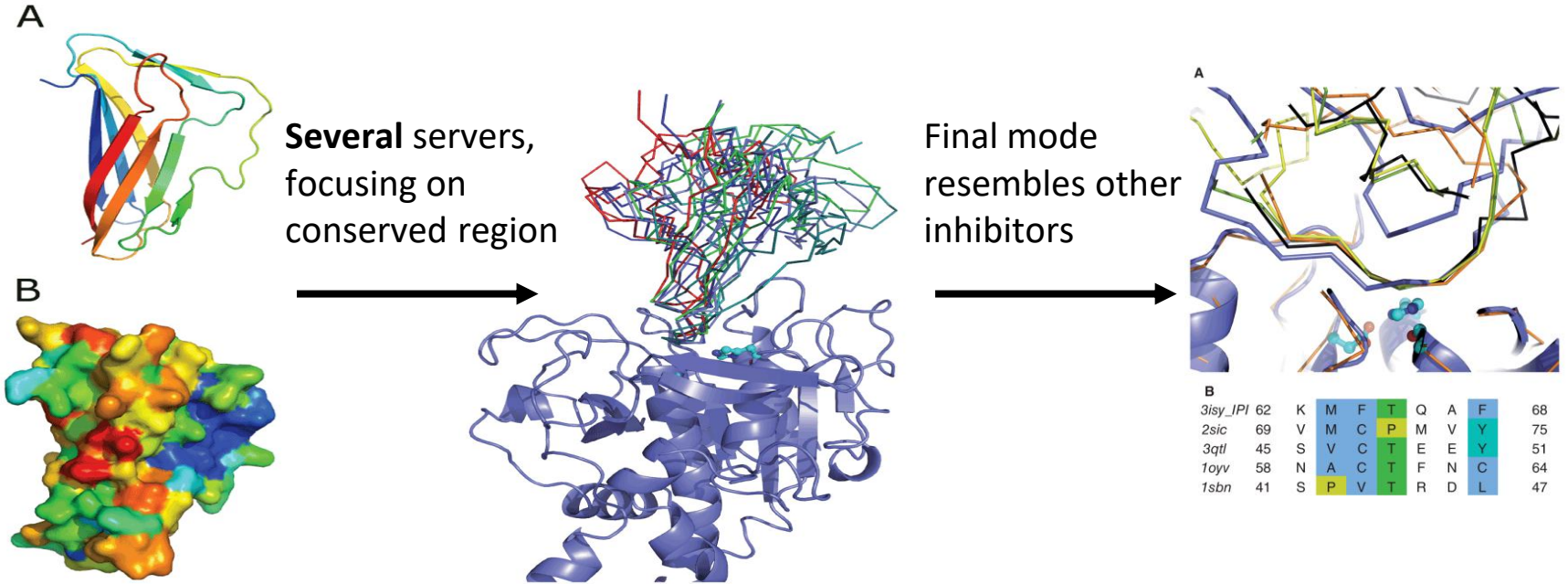


Quaternary structure and ligand interaction

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

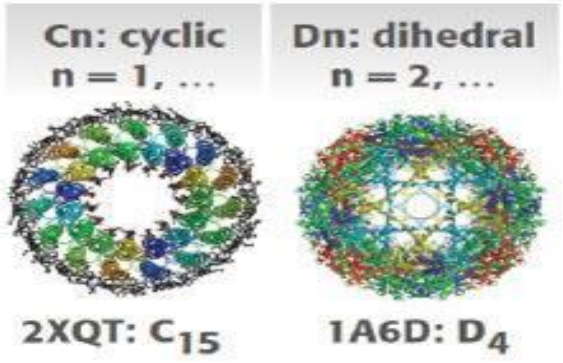
Multiple methods in bioinformatics: *B. subtilis* IPI docking to protease



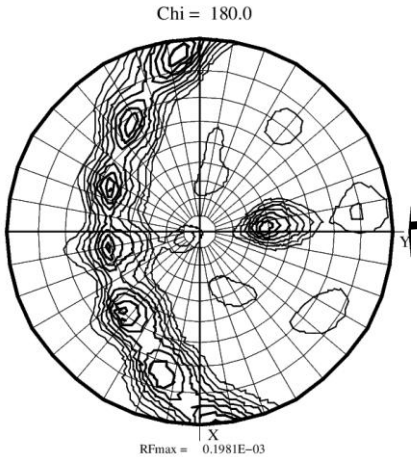
Rigden et al. (2013) F1000Research 2:154

ROSIE symmetric docking for oligomers

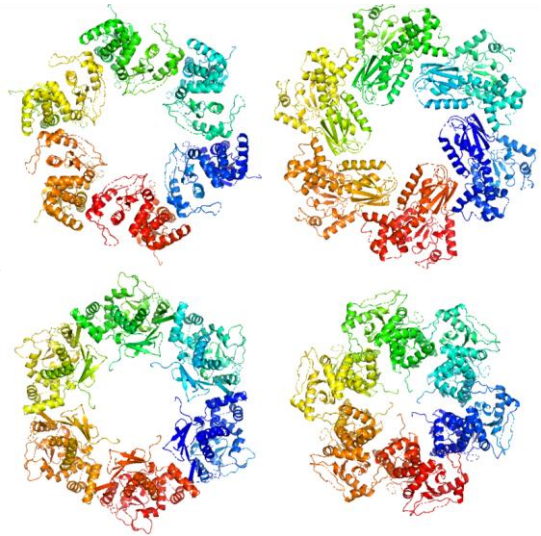
- Only cyclic (C_n) or dihedral (D_n) symmetry at server
- Clues from self-rotation function may be available
- AF2 cannot use this information!



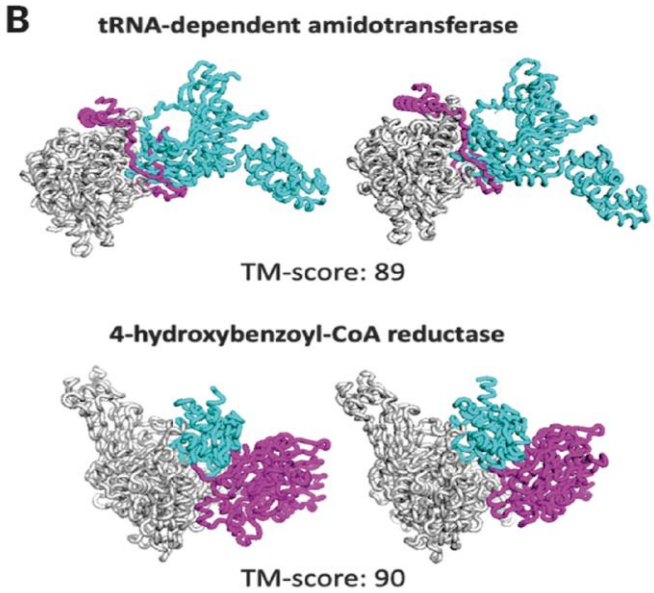
rcsb.org

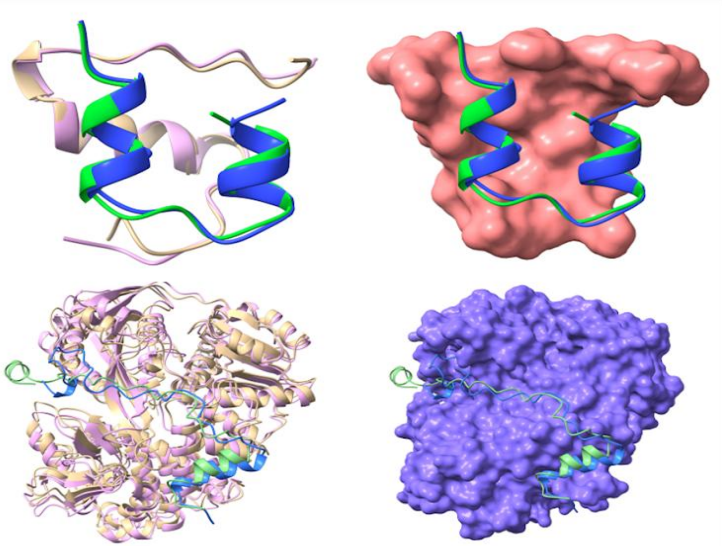


Generate hexamers



RF/AF2 cofolding to predict complexes





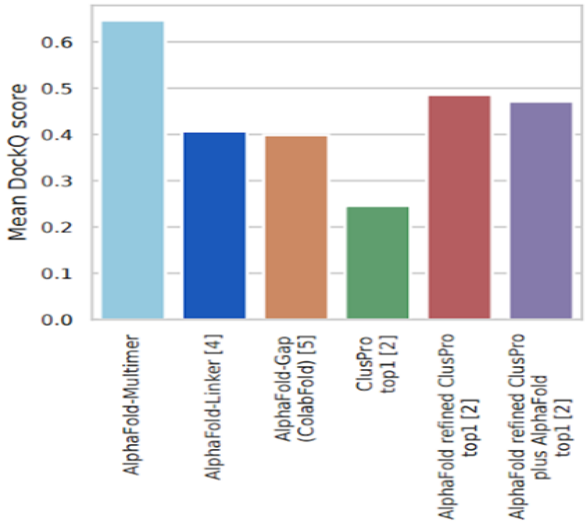
As with single chain folding, cofolding of multiple chains works best with good evolutionary covariance information

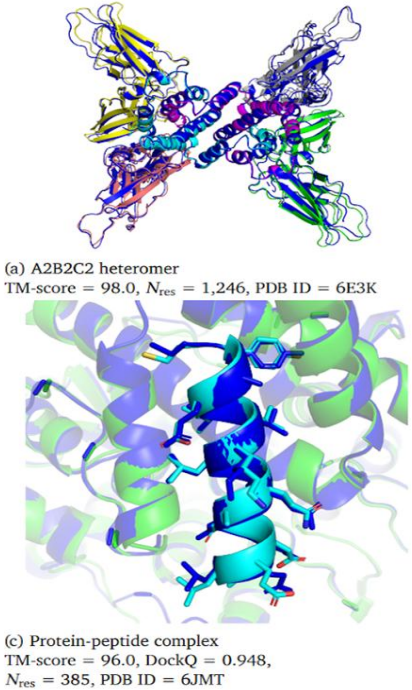
Baek et al (2021) Science 373,871

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

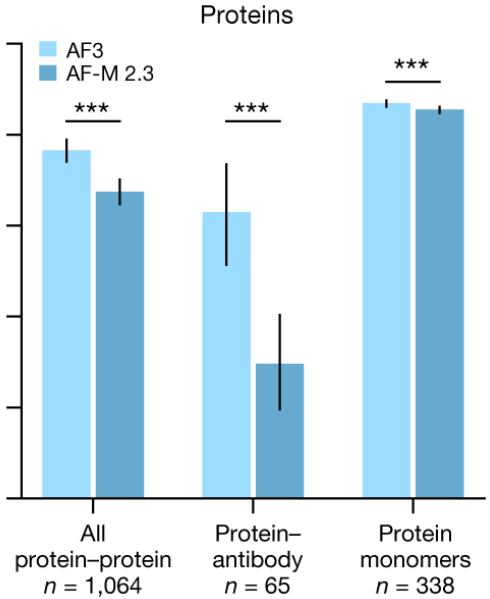
AFM and AF3





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

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



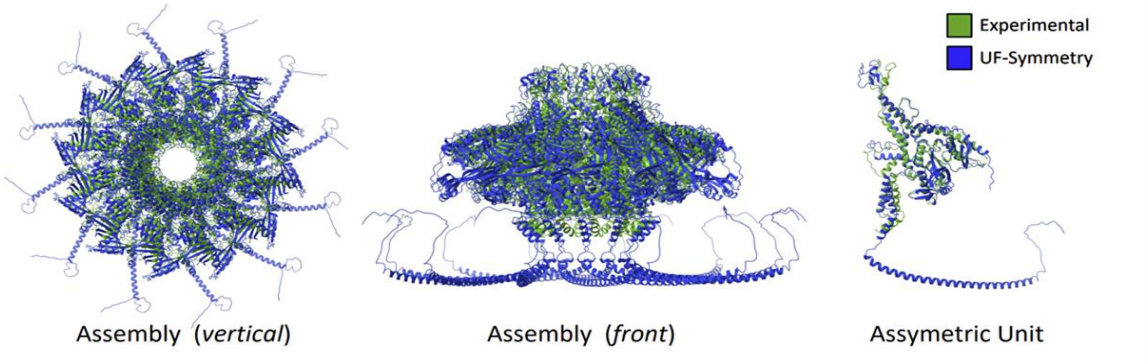
AF2 can predict oligomeric state eg asking for 5 copies may produce the natural tetramer + one left over, rather than a pentamer

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

Abramson et al (2024) Nature 630, 493

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

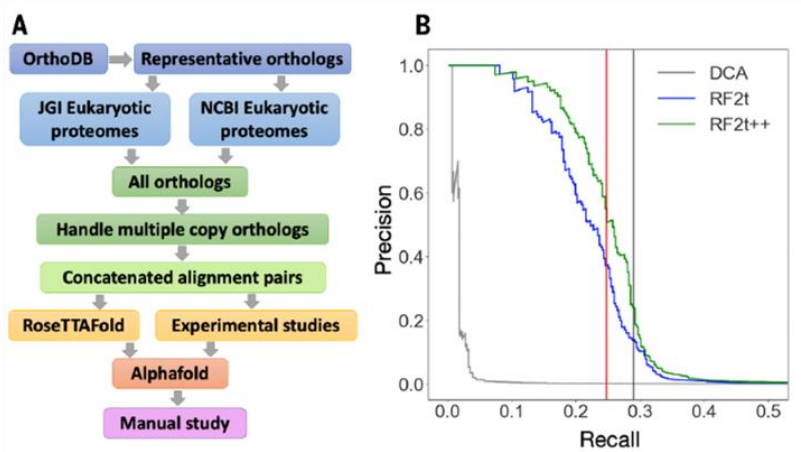
Li et al (2022) BioRXiv https://doi.org/10.1101/2022.08.30.505833

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

RF and AF2 predict *if* two proteins interact

Several groups use intermolecular covariance, as inferred by RF and/or AF2 to

- Model known complexes
- Show some annotated interactions likely indirect
- Predict new interactions/ complexes



Humphreys et al (2021) Science 374, 1340

Human PPI predictions

Final predictions

Show 10 entries Search:

	Pair	Gene1	Gene2	Confidence level	RF Prob	AF Prob	CF Prob	AFMM Prob	PDB	UniProt	BioGrid	STRING Physical	STRING Gene Score
⊖	Q7Z6A9 Q92956	BTLA	TNFRSF14	high confidence	0.998	0.9985	0.999	1.0	exact	N	N	Y	999
	Q7Z6A9 — **Protein Name:** B- and T-lymphocyte attenuator; **Knownness score:** 1.0; **Locality:** Cell membrane,Membrane; **Other information:** UniProt AFDB												
	Q92956 — **Protein Name:** Tumor necrosis factor receptor superfamily member 14; **Knowness score:** 8.9; **Locality:** Cell membrane,Membrane; **Other information:** UniProt AFDB												
⊕	Q14978 Q9UNY5	ZNF263	ZNF232	high confidence	0.13	0.9985	1.0	0.9995	homolog	N	N	N	452
⊕	P60059 Q9H9S3	SEC61G	SEC61A2	high confidence	1.0	0.995	0.995	0.999	homolog	N	N	Y	993
⊕	P23396 P63244	RPS3	RACK1	high confidence	0.9995	0.9453	0.969	0.9966	exact	N	Y	Y	999
⊕	P49754 Q9H270	VPS41	VPS11	high confidence	0.9844	0.9937	0.993	0.9756	homolog	Y	Y	Y	999
⊕	P63272 Q8WVC0	SUPT4H1	LEO1	high confidence	0.465	0.02788	0.0313	0.7593	none	N	N	Y	997
⊕	O14625 P49682	CXCL11	CXCR3	high confidence	0.913	0.7007	0.8677	0.8115	exact	N	N	N	0
⊕	O15323 O9NS84	KRT31	KRT82	high confidence	0.998	0.994	0.999	0.9976	homolog	Y	N	N	490
⊕	P18846 P47928	ATF1	ID4	high confidence	0.9995	0.003252	0.02792	0.995	none	N	N	N	0
⊕	O8N0V3 O9Y291	RBFA	MRPS33	high confidence	0.5513	0.0	0.0134	0.7046	none	N	N	Y	832

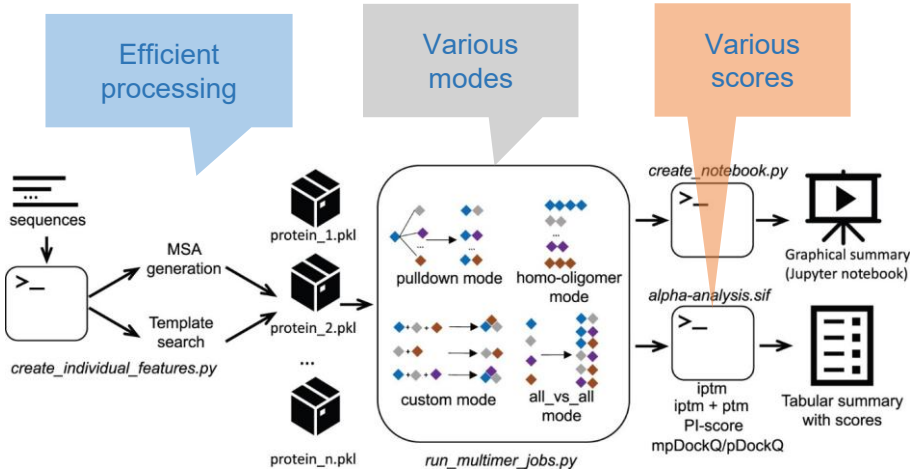
Showing 1 to 10 of 29,246 entries Previous 1 2 3 4 5 ... 2,925 Next

http://prodata.swmed.edu/humanPPI

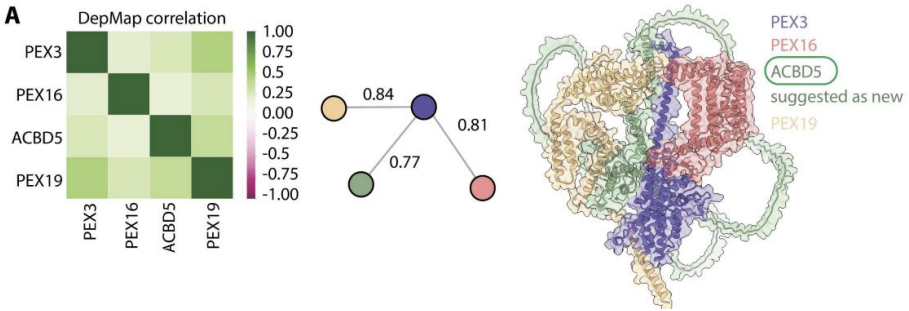
Zhang et al (2021) Science 10.1126/science.adt1630

Fishing for novel interactions

- AlphaPulldown is a great way to run 1-to-many or all-against-all screens with AF2 or AF3



- Screens can work with suspected or plausible interactions
- Works for motifs in disordered regions too

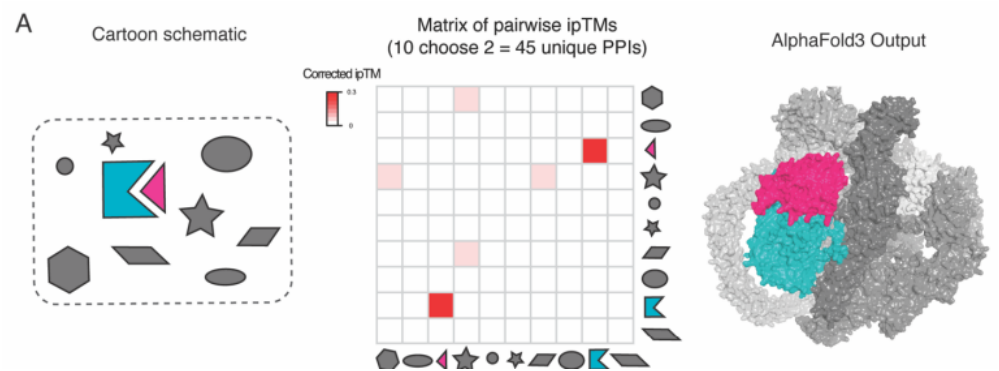


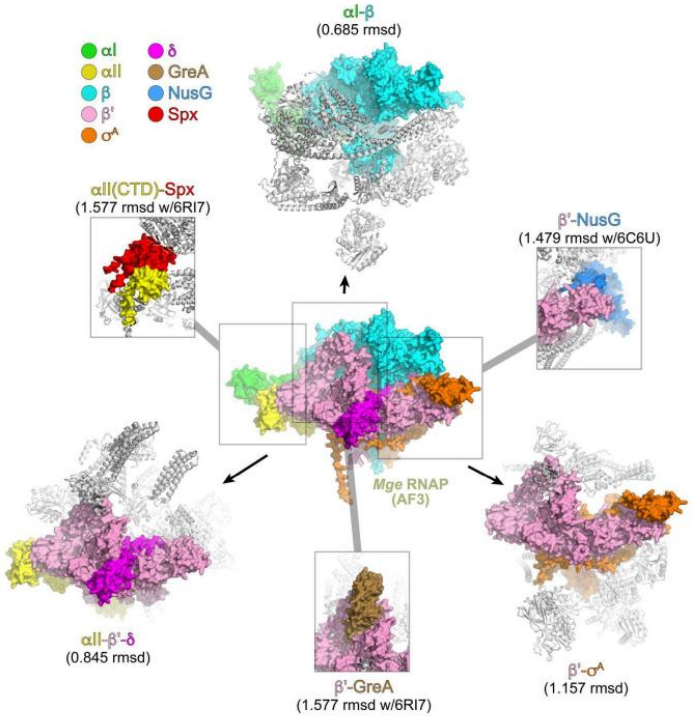
Cluster genetic co-dependencies from DepMap and model complexes with AF3

Yu et al (2022) Bioinformatics 39, btac749
Molodenskiy et al (2025) Bioinformatics 41, btaf115

Uckelmann et al (2025) bioRXiv doi: 10.1101/2025.09.09.675133

Testing a whole proteome with pooled-AF3





- *"We demonstrate the utility of this strategy by predicting all pairwise PPIs (113,050) in the M. genitalium genome using only 2,027 jobs, which required just 68 person-days using our personal AlphaFold3 accounts"*
- Corrected iPTMs for size and noted that competition-based iPTMs maybe be better than the regular ones

Todor et al (2025) bioRXiv doi: 10.1101/2025.07.01.662654

AF3 lookalikes

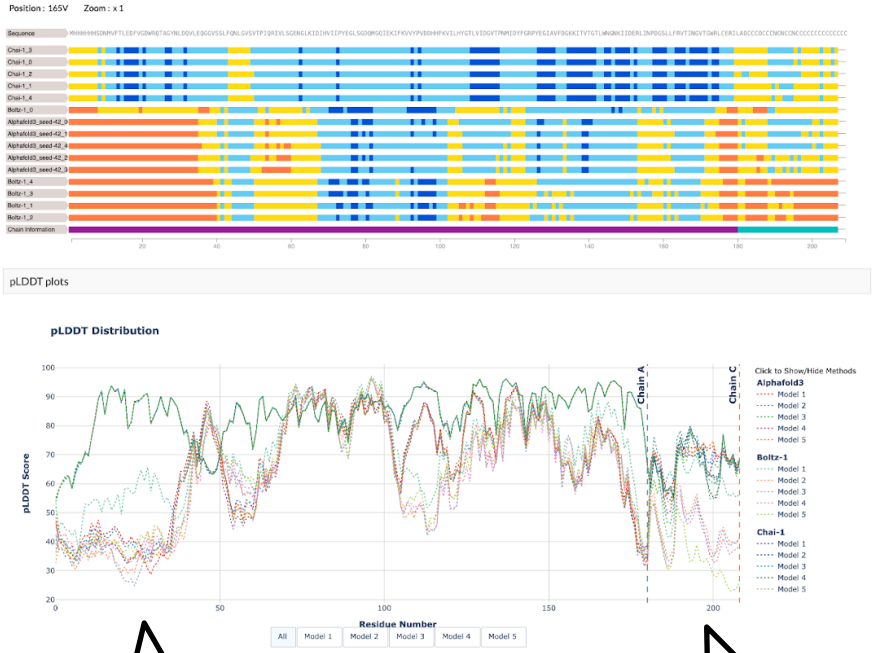
- AF3 license prohibits commercial use and use of results for further ML training
- A variety of rivals have less restrictive (not necessarily unrestricted!) licenses
  - Boltz from MIT with industry collaborators
  - Chai from Chai Discovery company
  - Protenix. ByteDance Seed-AI for Science
  - OpenFold 3. OpenFold consortium of academic and industry groups
  - RosettaFold 3. Baker group at UWash
- The field is fast-moving. In general terms (AFAIK!), none currently improves on AF3 on basic performance, but they have some specific niche advantages
- neurosnap.ai or app.tamarind.bio ($, free trial) are ways to access several

Multiple methods in bioinformatics: ABCFold

ABCFold

Structure predictions for: AlphaFold3, Boltz-1 and Chai-1

Model Name	Model Source	Average pLDDT ▼	H-score	pTM score	ipTM score	Residue Clashes	Atom Clashes	Model visualisations
Chai-1_3	Chai-1	83.69	77	0.82	0.45	0	0	Click for PAE Plot
Chai-1_0	Chai-1	83.68	77	0.82	0.45	0	0	Click for PAE Plot
Chai-1_2	Chai-1	83.65	77	0.82	0.47	0	0	Click for PAE Plot
Chai-1_1	Chai-1	83.62	77	0.82	0.42	0	0	Click for PAE Plot
Chai-1_4	Chai-1	83.60	77	0.82	0.4	0	0	Click for PAE Plot
Boltz-1_0	Boltz-1	70.90	64	0.72	0.83	0	0	Click for PAE Plot
Alphafold3_seed-42_0	AlphaFold3	67.08	62	0.6	0.85	0	0	Click for PAE Plot
Alphafold3_seed-42_1	AlphaFold3	66.14	61	0.59	0.84	0	0	Click for PAE Plot
Alphafold3_seed-42_4	AlphaFold3	65.41	61	0.58	0.84	0	0	Click for PAE Plot
Alphafold3_seed-42_2	AlphaFold3	65.35	61	0.57	0.82	0	0	Click for PAE Plot
Alphafold3_seed-42_3	AlphaFold3	64.78	61	0.57	0.83	0	0	Click for PAE Plot
Boltz-1_4	Boltz-1	64.07	60	0.55	0.52	0	0	Click for PAE Plot
Boltz-1_3	Boltz-1	62.70	59	0.51	0.59	0	0	Click for PAE Plot
Boltz-1_1	Boltz-1	62.62	59	0.54	0.64	0	0	Click for PAE Plot
Boltz-1_2	Boltz-1	62.20	58	0.52	0.64	0	0	Click for PAE Plot



Sort the models by average pLDDT, pTM, ipTM, number of clashes

Elliott et al (2025) Bioinformatics Advances, 5, vbaf153

In this example, only Boltz-1 is confident in the N-terminus

In this example, only Boltz-1 and AF3 are moderately confident in ligand placement

ABCFold

- Need AF3 installation; Boltz-1 and Chai-1 installed during run
- Local, API or bespoke MSAs
- Helps deal with templates too
- Consistent display of models, clashes, PAEs
- https://github.com/rigdenlab/ABCFold

Elliott et al (2025) Bioinformatics Advances, 5, vbaf153

ABCFold – ligand identification

WWW.JISCMAIL.AC.UK (26 Matches)

Subject

Maps look different from auto-mtz vs EDS vs FFT in Coot or CCP4MG.

I'd like help in interpreting some mystery density in a structure.
ligands. The apo structure is already released, as pdb code 2B23. The
question is whether there is a mystery molecule in the pocket of the apo
receptor. If you superimpose 3ERD, you can see where the ligand binds. The
If you look at the 2mFo-DFc map from EDS in Coot or CCP4mg, you see mystery
density in the pocket. If I use the MTZ, you see density in Coot, but not
CCP4MG. I then downloaded the structure factors from the PDB and made an
MTZ. The map in CCP4MG shows some density, but much less that with the map
from EDS. When I used FFT to make a 2mF1-1nF2 map, there is no mystery
density in either CCP4MG or coot.
I was ready to submit the manuscript with a picture of the mystery density,
but now I'm not sure if that is appropriate. Any suggestions, as far as how
to interpret this mystery density would be greatly appreciated.

Re: Maps look different from auto-mtz vs EDS vs FFT in Coot or CCP4MG.

> I'd like help in interpreting some mystery density in a structure.
>
> different ligands. The apo structure is already released, as pdb
> code 2B23. The question is whether there is a mystery molecule in
> the pocket of the apo receptor. If you superimpose 3ERD, you can
> If you look at the 2mFo-DFc map from EDS in Coot or CCP4mg, you see
> mystery density in the pocket. If I use the MTZ, you see density
> in Coot, but not CCP4MG. I then downloaded the structure factors
> from the PDB and made an MTZ. The map in CCP4MG shows some density,
> but much less that with the map from EDS. When I used FFT to make a
> 2mF1-1nF2 map, there is no mystery density in either CCP4MG or coot.
>
> I was ready to submit the manuscript with a picture of the mystery
> density, but now I'm not sure if that is appropriate. Any
> suggestions, as far as how to interpret this mystery density would
> be greatly appreciated.

ACA 2010, structural enzymology - mechanistic: call for abstracts

intermediates provide valuable mechanistic insight, but for which the
crystallographers often find themselves interpreting □mystery density□
within the data. Thus, the talks and posters in this session are intended

March 31, ACA abstract deadline

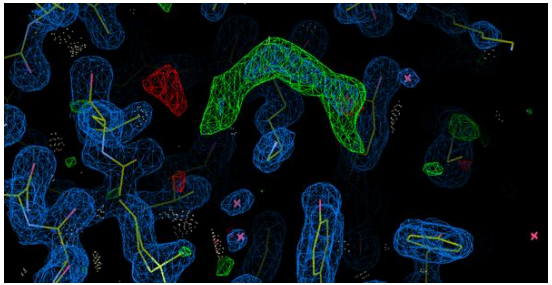
intermediates provide valuable mechanistic insight, but for which the
crystallographers often find themselves interpreting □mystery density□
within the data. Thus, the talks and posters in this session are intended

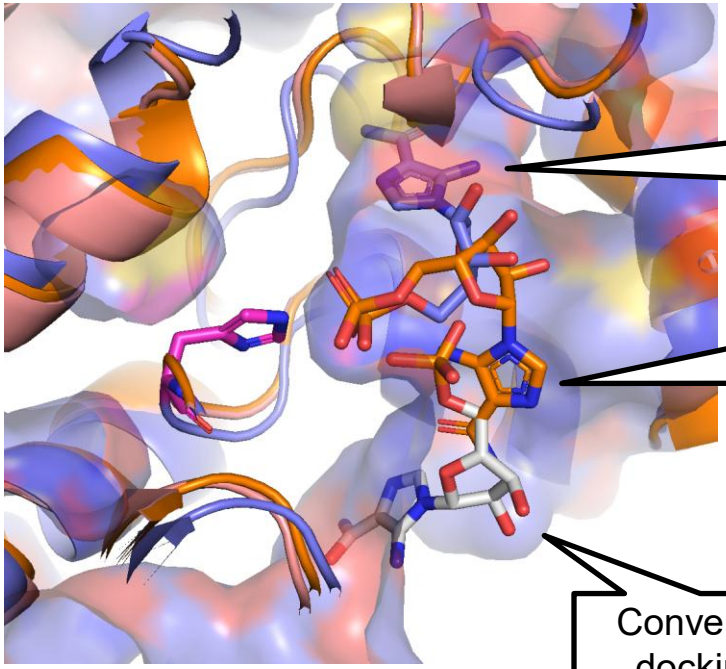
Re: Weird blob appears

were the dimensions consistent with three sugars? were there any collisions
with backbone or side chains inside/near the mystery density?
It really looks like maltotriose... is the protein a sugar binder by any
> > 1- It is in the symmetry axis so maybe be a little noisy there
> > 2-If you are in doubt about it being real or not check the density and
> how it fits into your protein and (symm. related neighbours of course). If

Mystery density

From: Martin Montgomery <[log in to unmask]>
Subject: Mystery density
Content-Type: multipart/alternative;
we have a puzzling density and hope someone has an idea of what it may be.





Boltz-1 packs the base in a pocket

AF3 pose more plausible but base not well-packed

Conventional docking on AFDB model clearly wrong

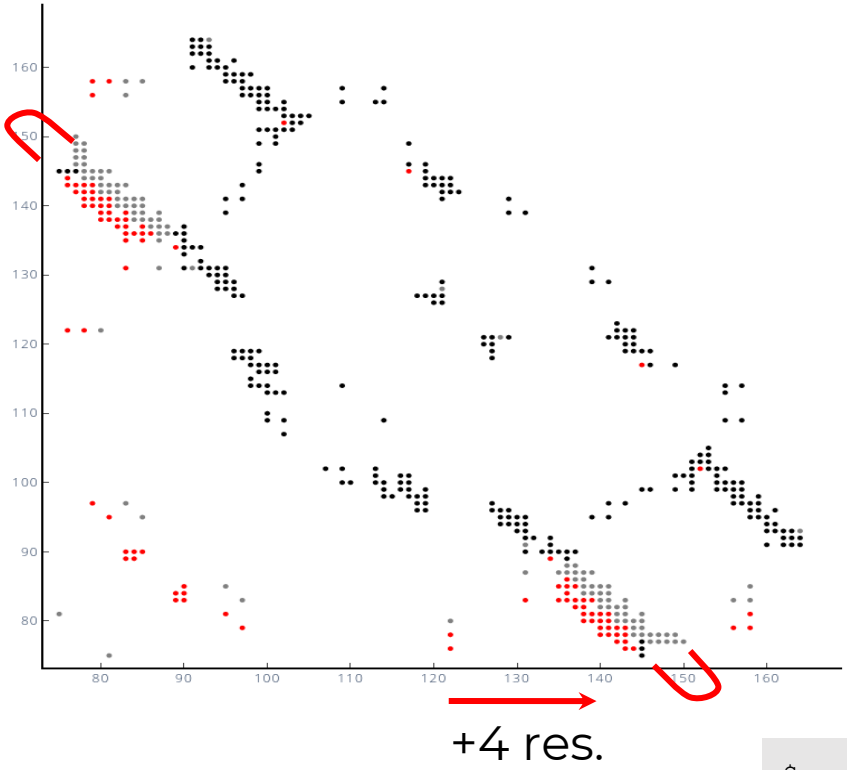
Note differences in protein!

Finalising the structure

Does it contain any residual errors?

What is the biologically relevant quaternary structure?

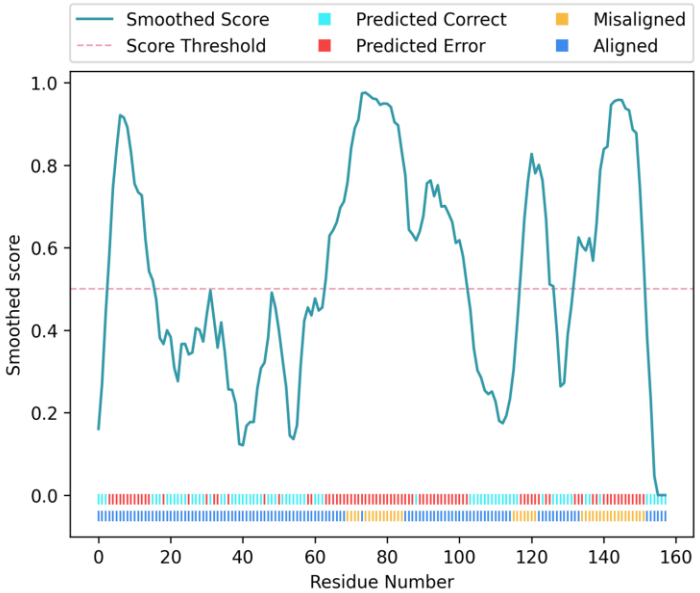
New covariance-based metrics for model validation



+4 res.

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

ConKit

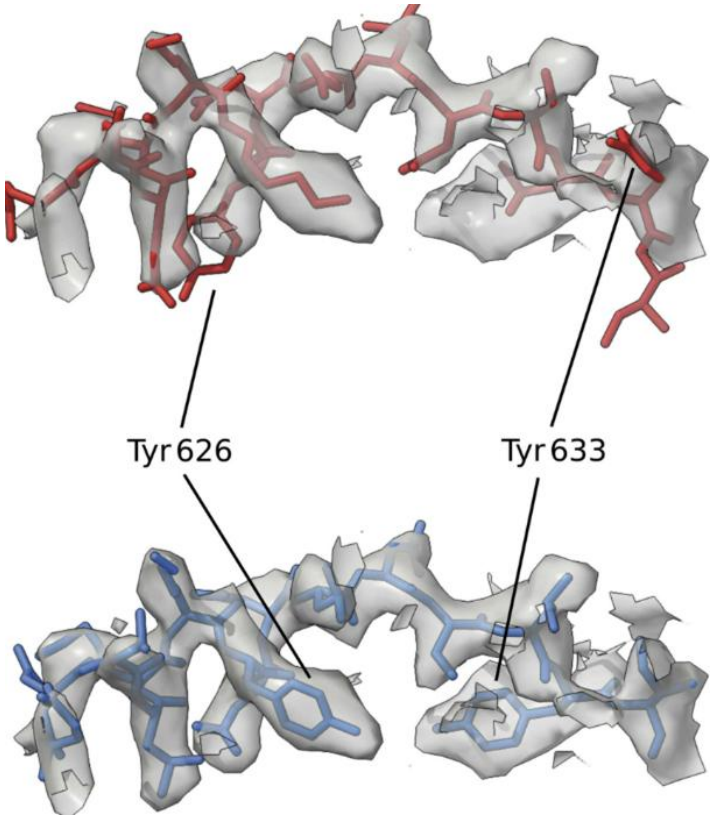



```
$ conkit-validate model.pdb prediction_af2.pkl sequence.fasta
```

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



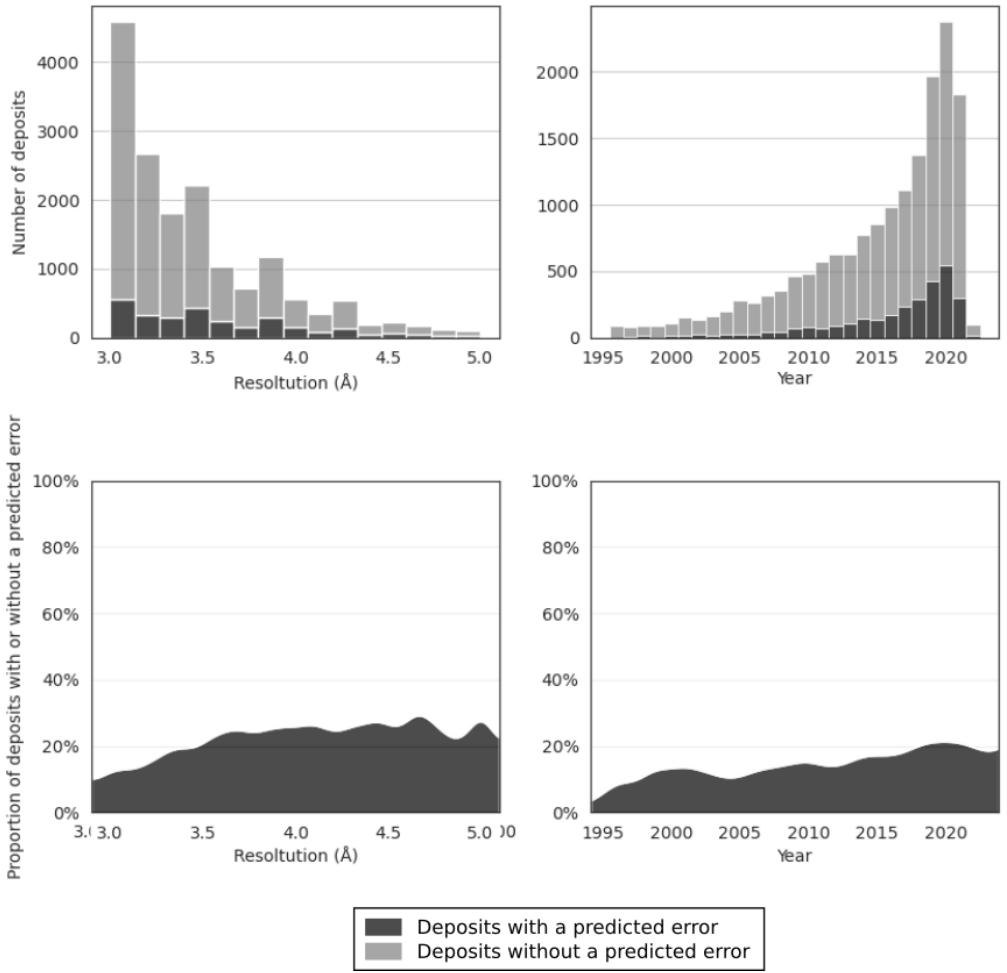
Sánchez Rodríguez *et al.* (2022) Acta Cryst D, 78, 1412

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?

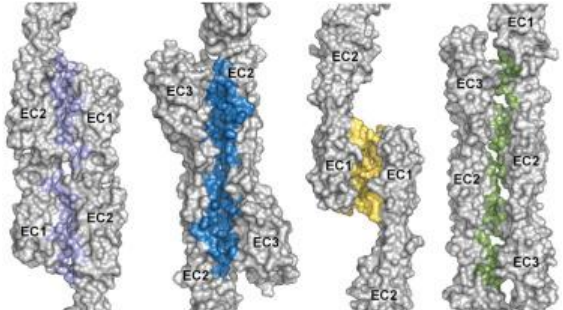
Cryo-EM structures contain more errors than crystal. Issue of local resolution, or maybe just because they are bigger?

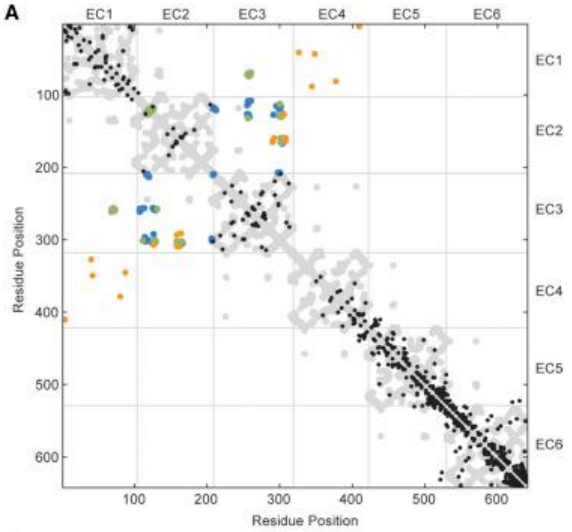


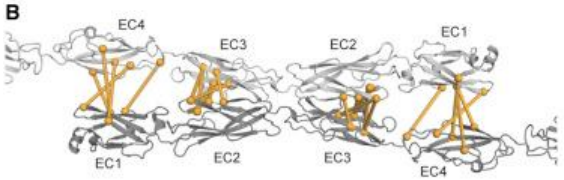
Sánchez Rodríguez *et al.* (2024) IUCrJ, 11, 938

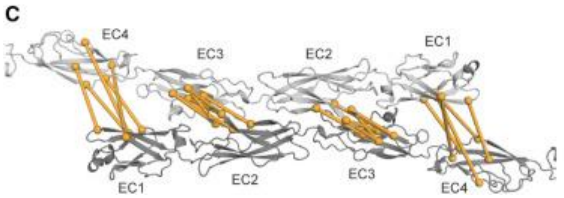
Validating crystal structure contents

- PISA is an excellent general method, but contact predictions help in some cases
- Crystal showed various ways in which protocadherins could interact
- Predicted contacts from evolutionary covariance supported only some of the modes
- Suggest https://evcouplings.org/ then ConPlot









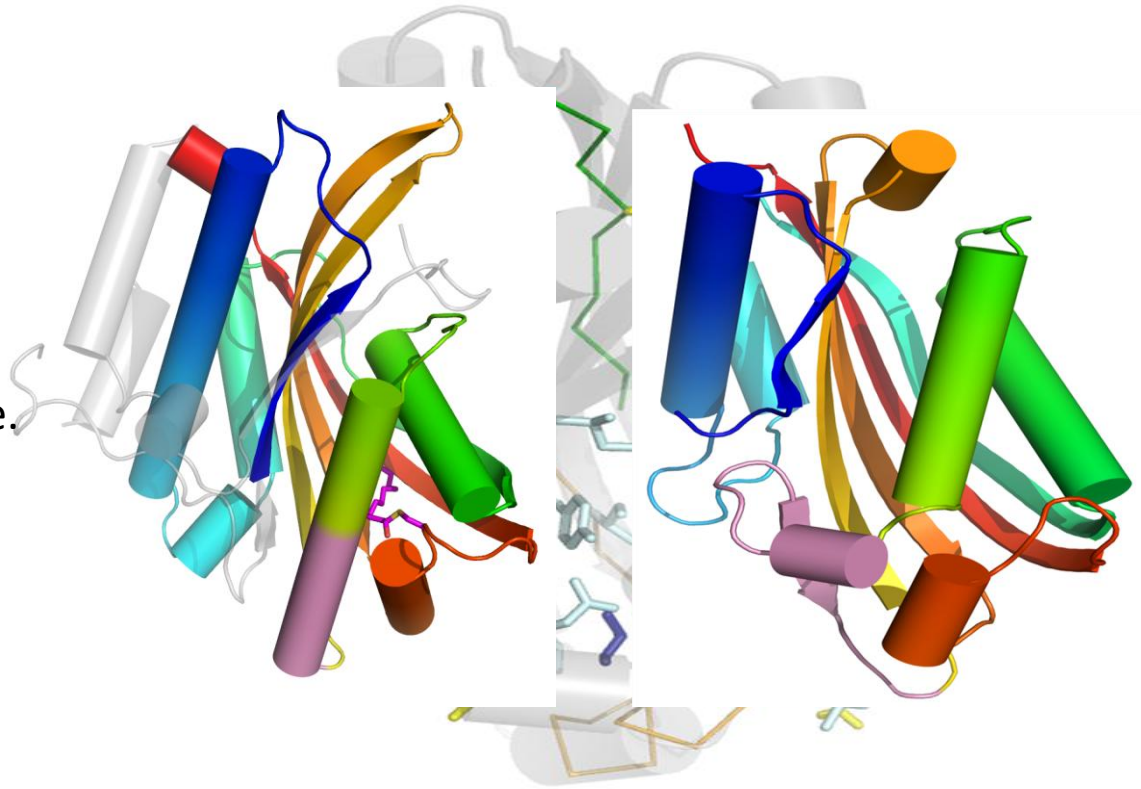
Nicoludis *et al*. (2015) Structure 23, 2087

Structure-based function interpretation

Where are the functional/catalytic sites?

Multiple methods in bioinformatics: Structure comparisons of Evf

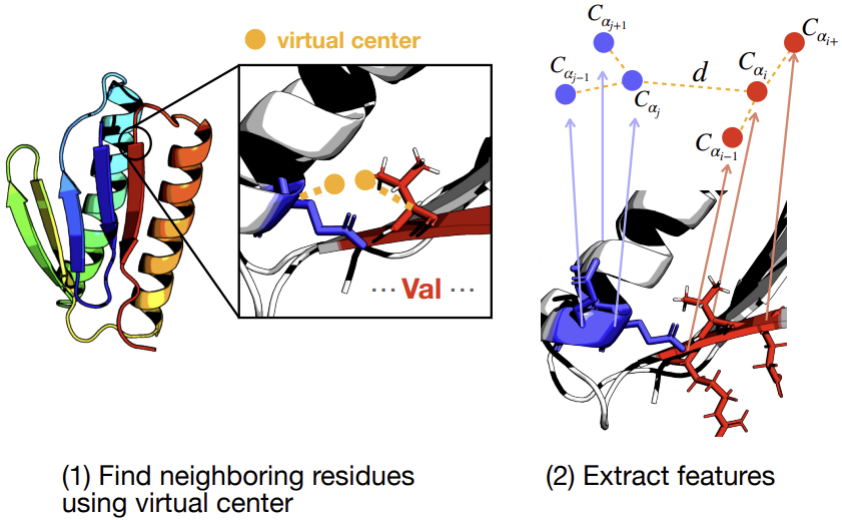
- Reported as novel fold...
- ... but in fact related to *Bacillus* toxin structures (DALI)
- Both bind to host insect membranes
- Palmitate seen in Evf structure. Matches conserved region of toxins...
- GESAMT is an excellent CCP4 option
- FoldSeek is great for a quick search of the AFDB, ESMAtlas etc



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

FoldSeek to search AFDB etc

- Almost as sensitive as DALI
- Very fast
- Scans UniRef50 of AFDB v3
- Also ESMAtlas, BFMD etc
- Works by representing structural neighbourhoods as strings and string matching
- Phylogenetic filtering



(1) Find neighboring residues using virtual center

(2) Extract features

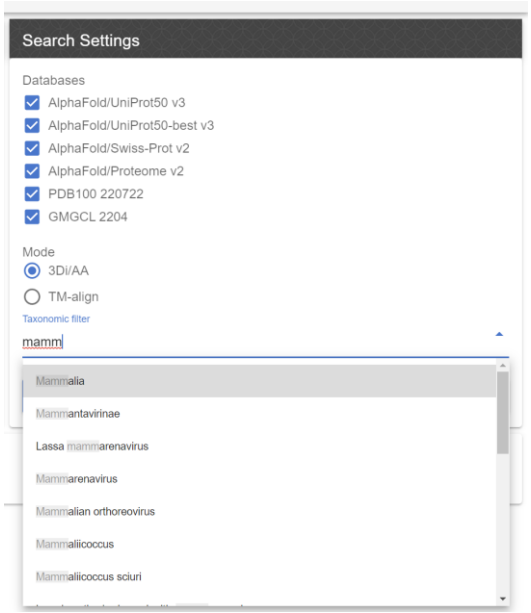
"The 20 states of the 3D-interactions (3Di) alphabet describe for each residue i the geometric conformation with its spatially closest residue j."

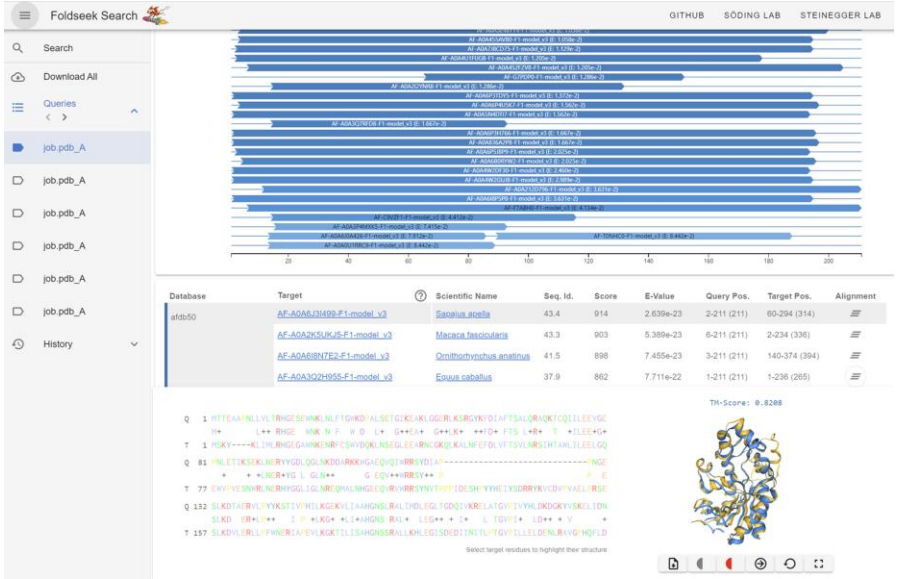
https://search.foldseek.com/

van Kempen et al (2023) Nature Biotech https://doi.org/10.1038/s41587-023-01773-0

Foldseek server

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





https://search.foldseek.com/

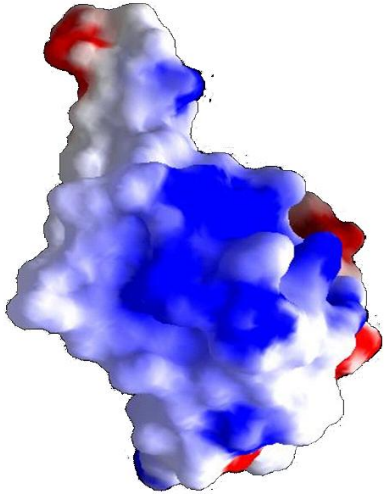
Van Kempen et al. (2022) Nature Biotech
https://doi.org/10.1038/s41587-023-01773-0

Structure-based function annotation

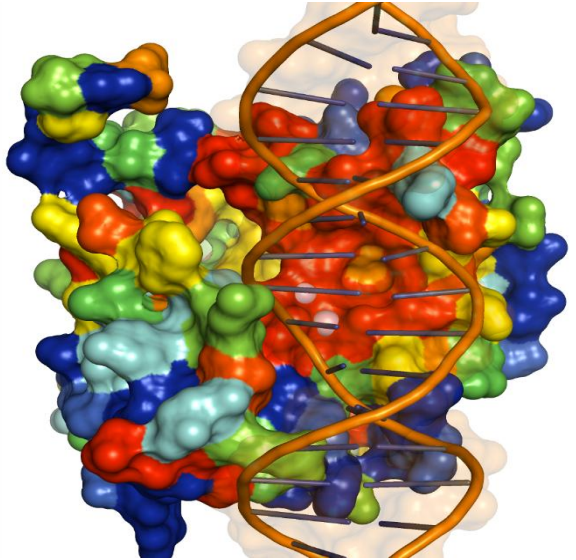
- Finding functional sites is based on their being different somehow to the rest of the protein surface. Important general methods are based on
  - Shape (castP, ProFunc, PyMOL)
  - Electrostatics (PyMOL, APBS)
  - Evolutionary conservation (Consurf)
- Less well-known but valuable characteristics are
  - 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

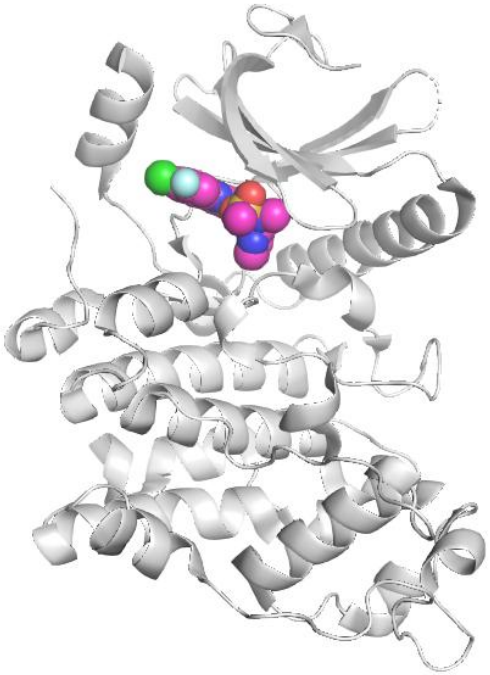


Electrostatics
APBS/PyMOL

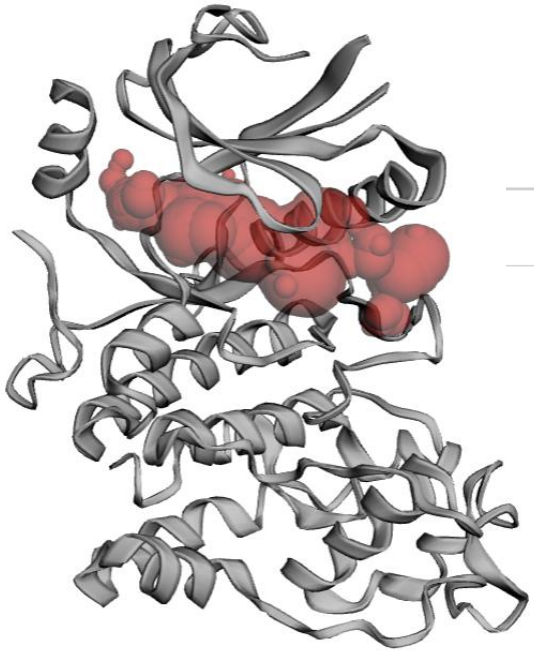


Conservation
ConSurf

Some servers require thought…



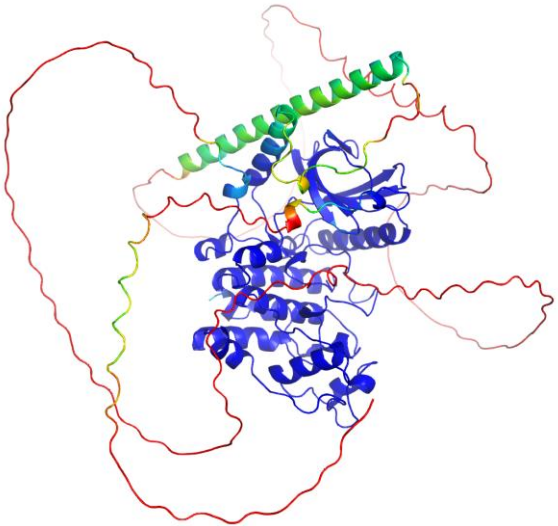
CastP



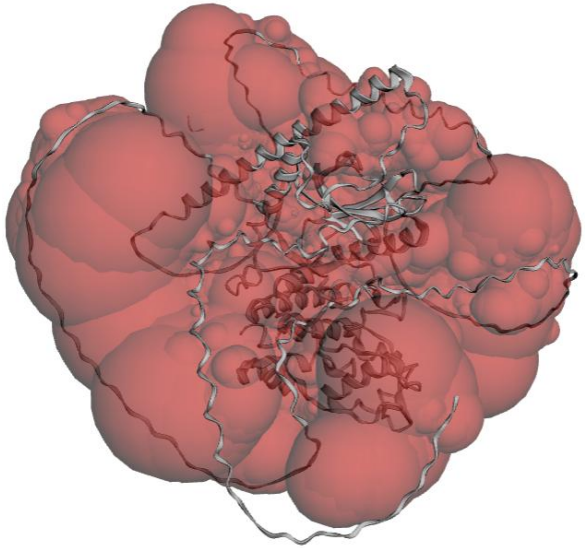
PocID	Area (SA) Å^2	Volume (SA) Å^3
1	762.617	506.394

Human SRSF protein kinase 2, PDB 7zkx

Some servers require thought…



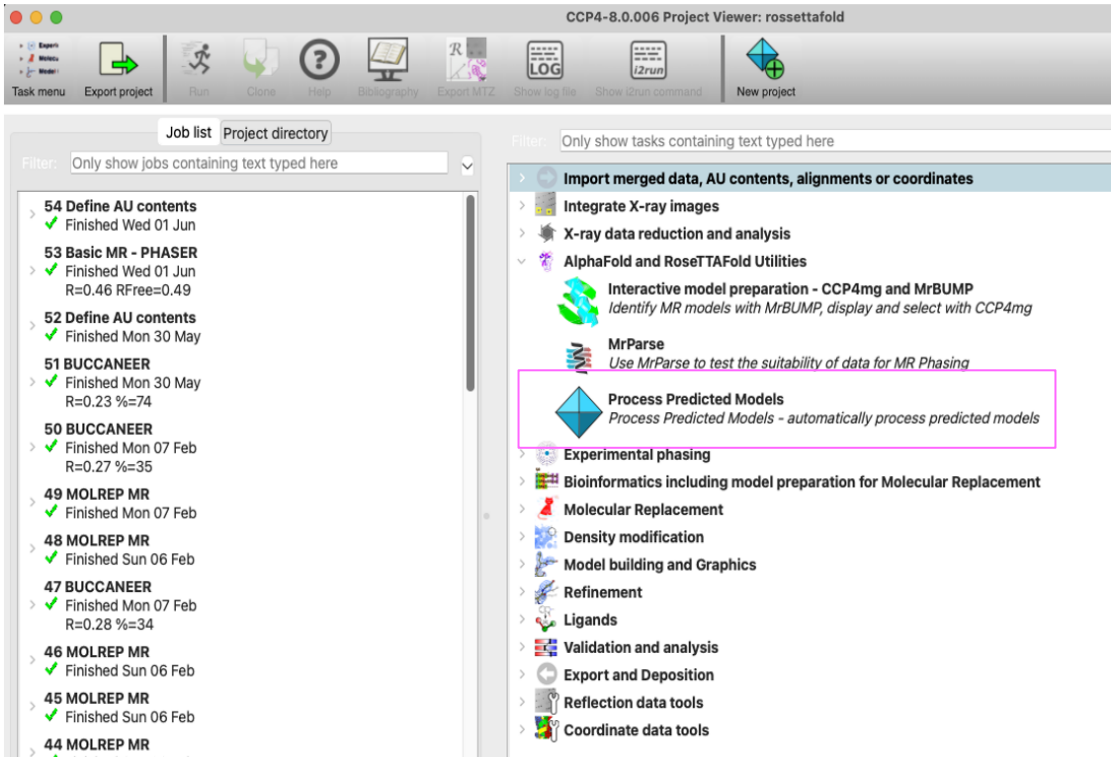
CastP



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

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

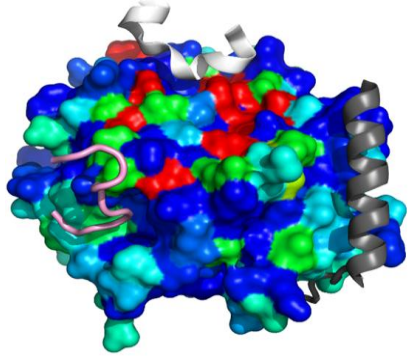
CCP4 i2 can remove the red spaghetti (and convert pLDDT to B-factor)



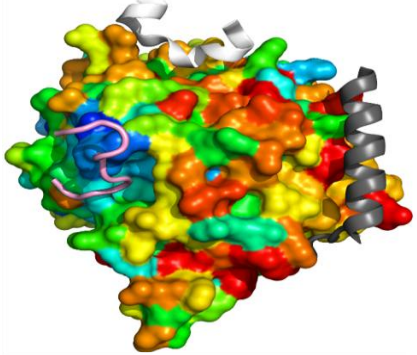
Some servers require thought…

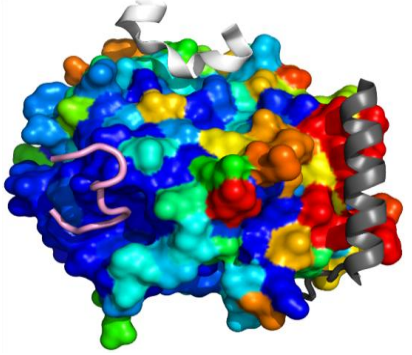
- Consurf maps sequence conservation onto a structure revealing functional sites
- Excellent, general method, but results depend on sequence set chosen for mapping: selecting all or only near relatives gives different results. Either might be more appropriate for you

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



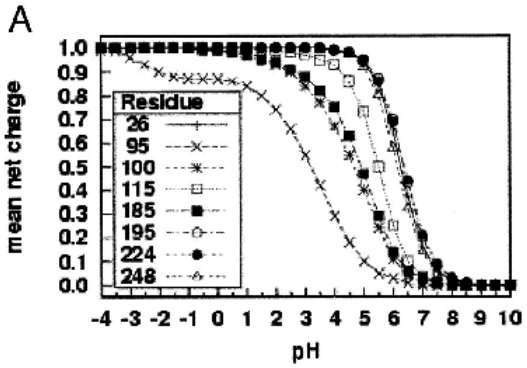
But restricting to a single protein family shows only ‘pink’ site is function in both Diptera and Lepidoptera



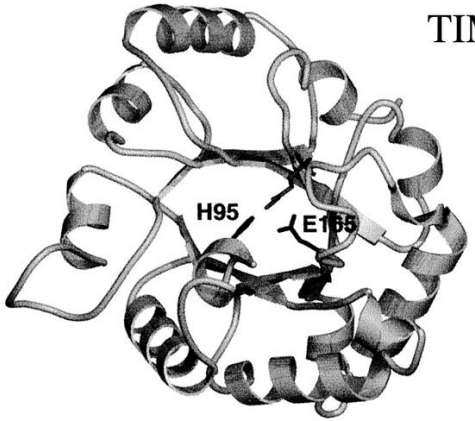


Theoretical microscopic titration

- Computer analysis of a reliable protein structure can predict pKa values for acids and bases. Residues with perturbed pKa values are possible catalytic residues, especially if clustered.



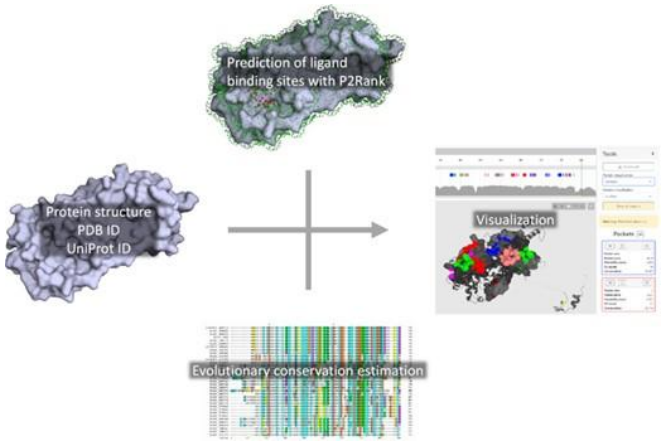
pKa of His95 is atypical compared to other His residues in enzyme

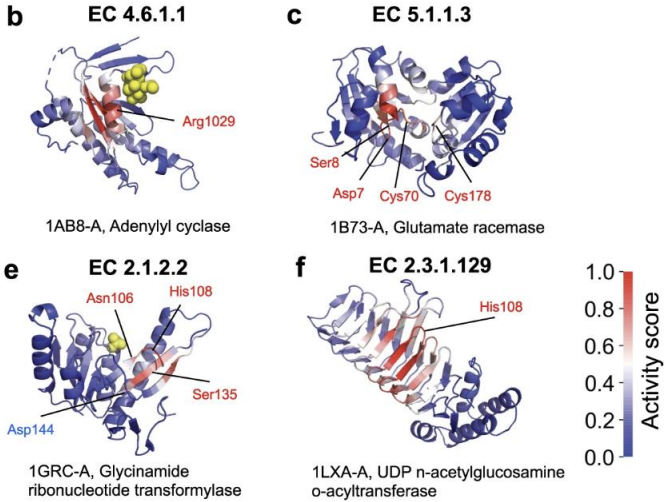


His95 and other residues with atypical pKa cluster at catalytic site

www.pool.neu.edu/wPOOL/index2.jsp

Advanced methods use Machine Learning and/or multiple signals





PrankWeb uses structural and physicochemical properties then displays pockets with conservation analysis. Now with onward Autodock Vina

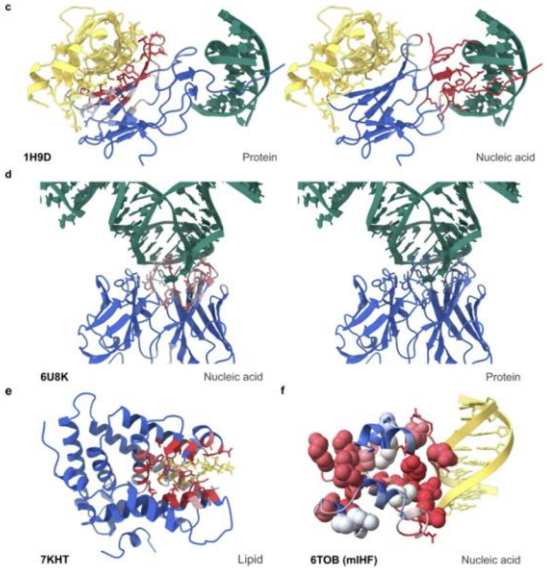
https://prankweb.cz

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

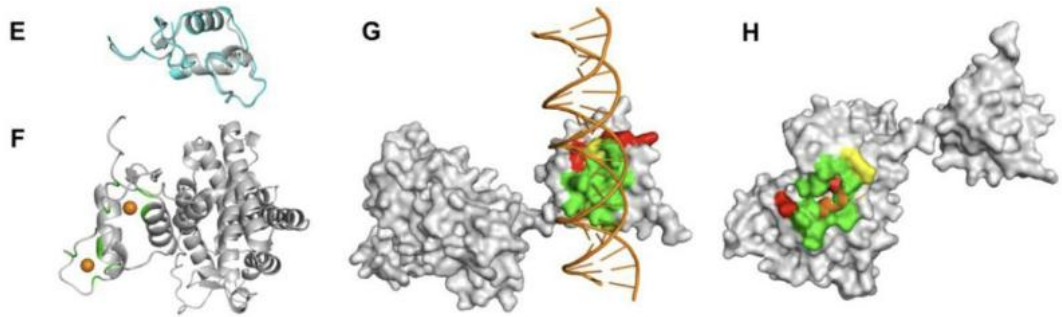
PESTO



Geometry and contacts of input structure analysed. Interface propensities displayed

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

GPsite

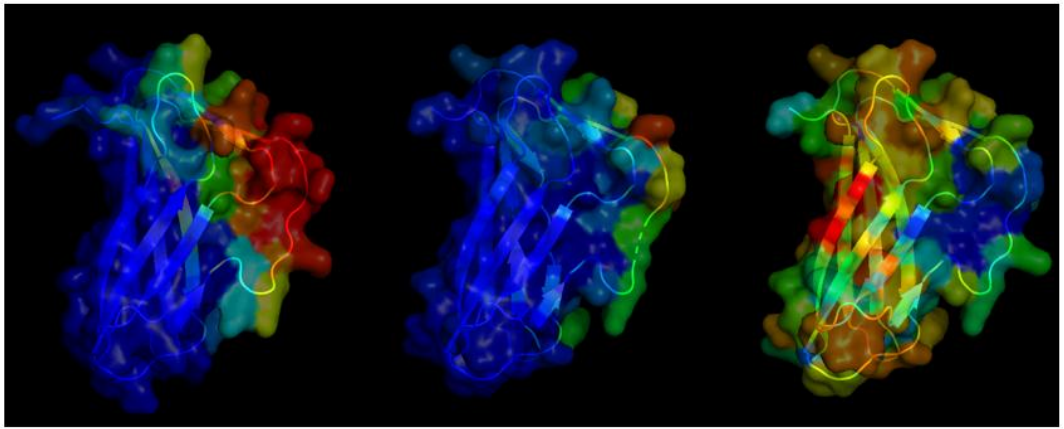


True Positive False Positive False Negative

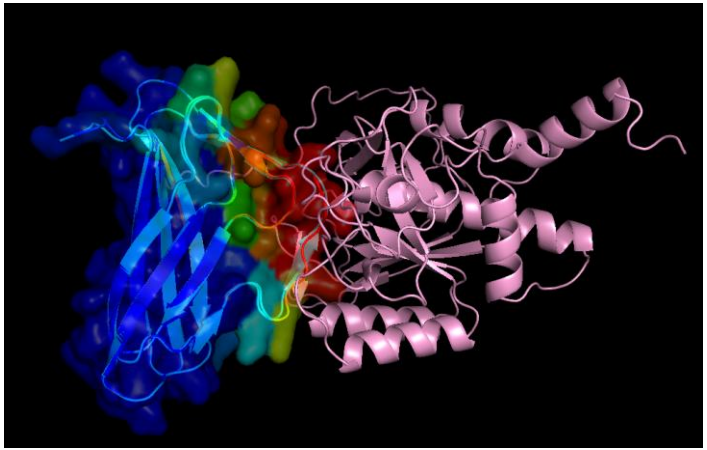
Sequence modelled with ESMFold. Various binding sites and GO terms predicted

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

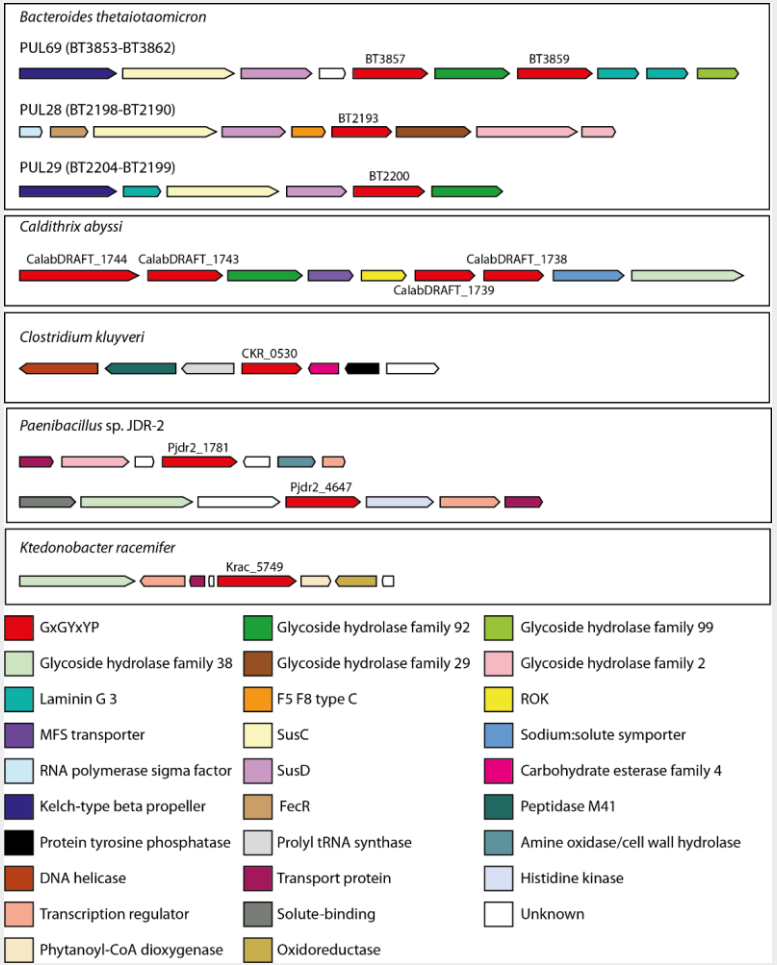


All methods pick out the very probable binding site of this protease inhibitor. DL methods give a somewhat cleaner signal than general conservation

Case study from Structural Genomics

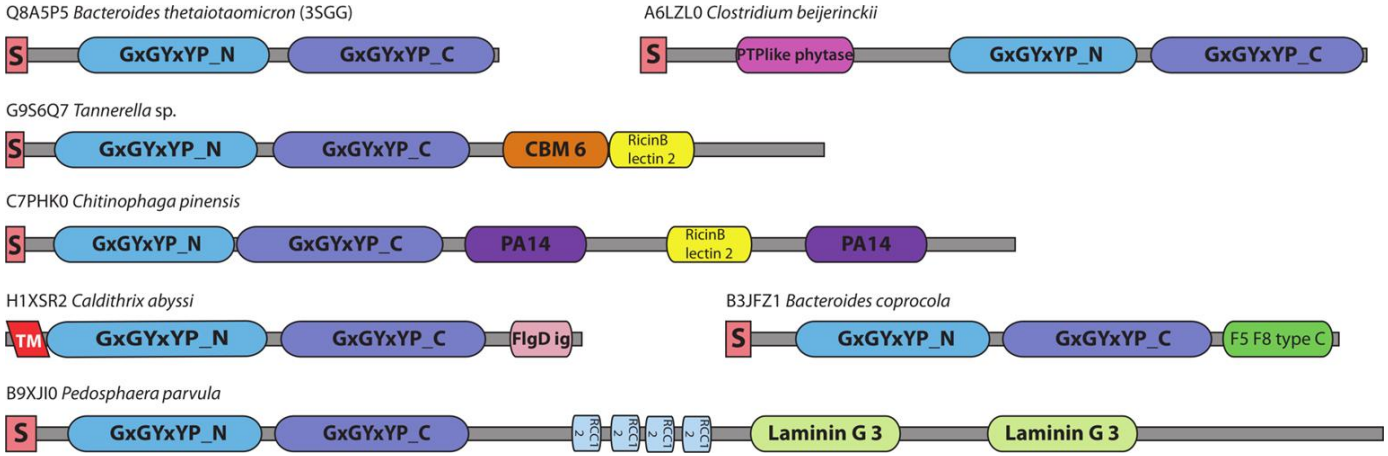
GxGYxYP proteins

- 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?*



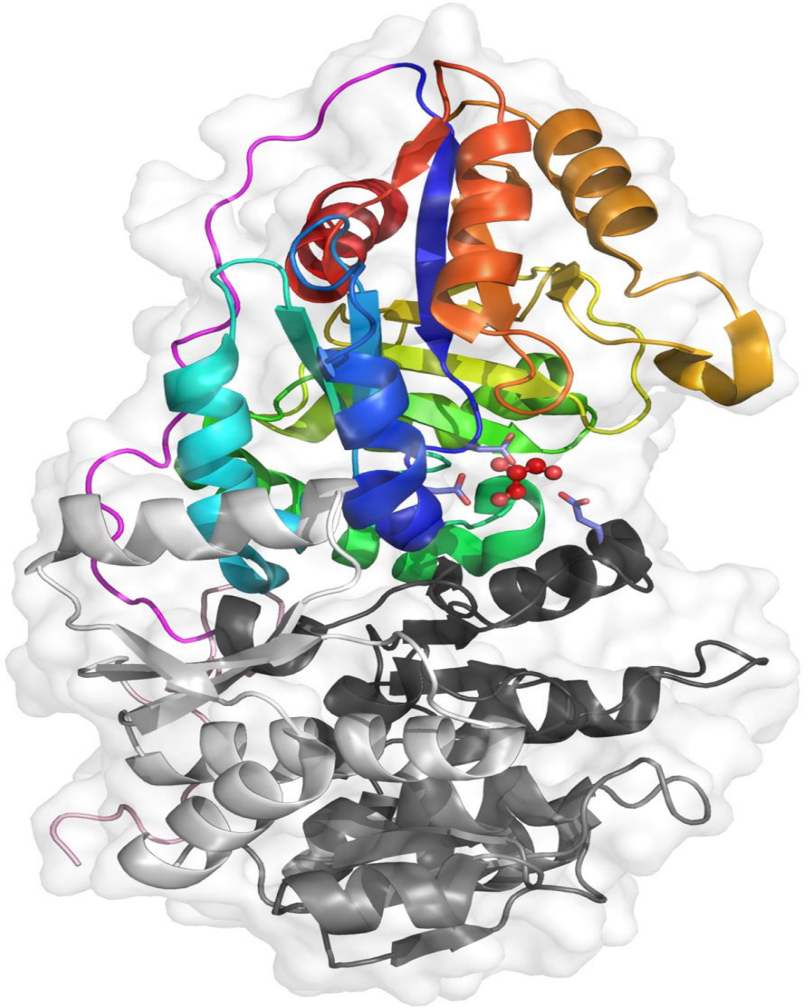
GxGYxYP proteins

- Domain architectures also predict carbohydrate connection



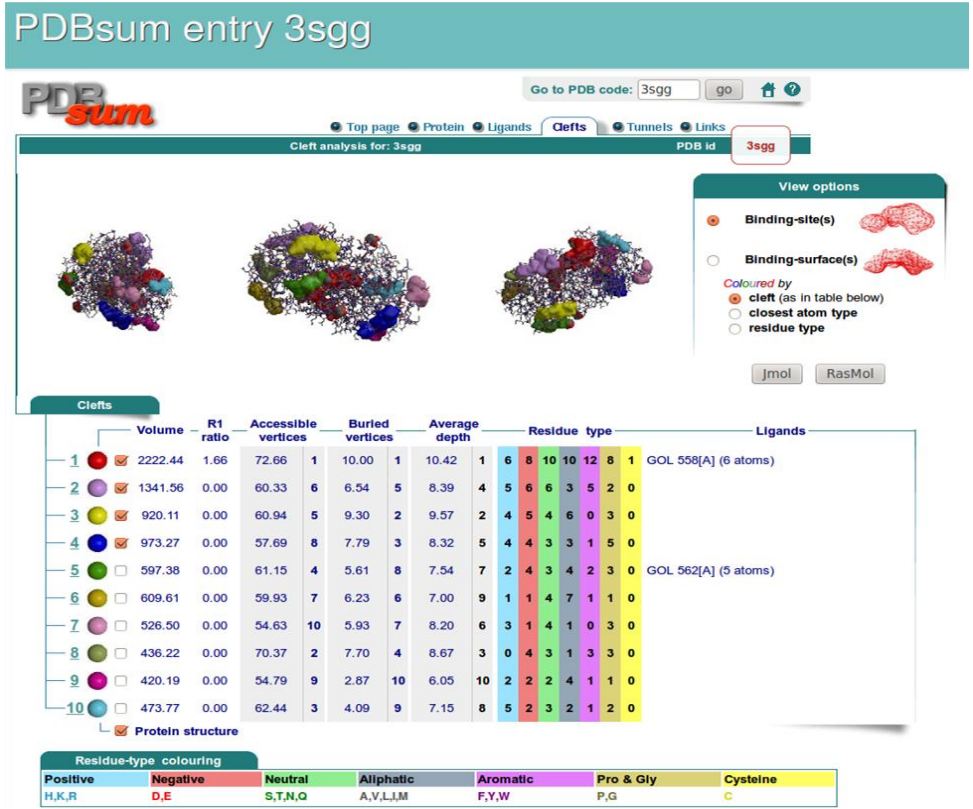
GxGYxYP proteins

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



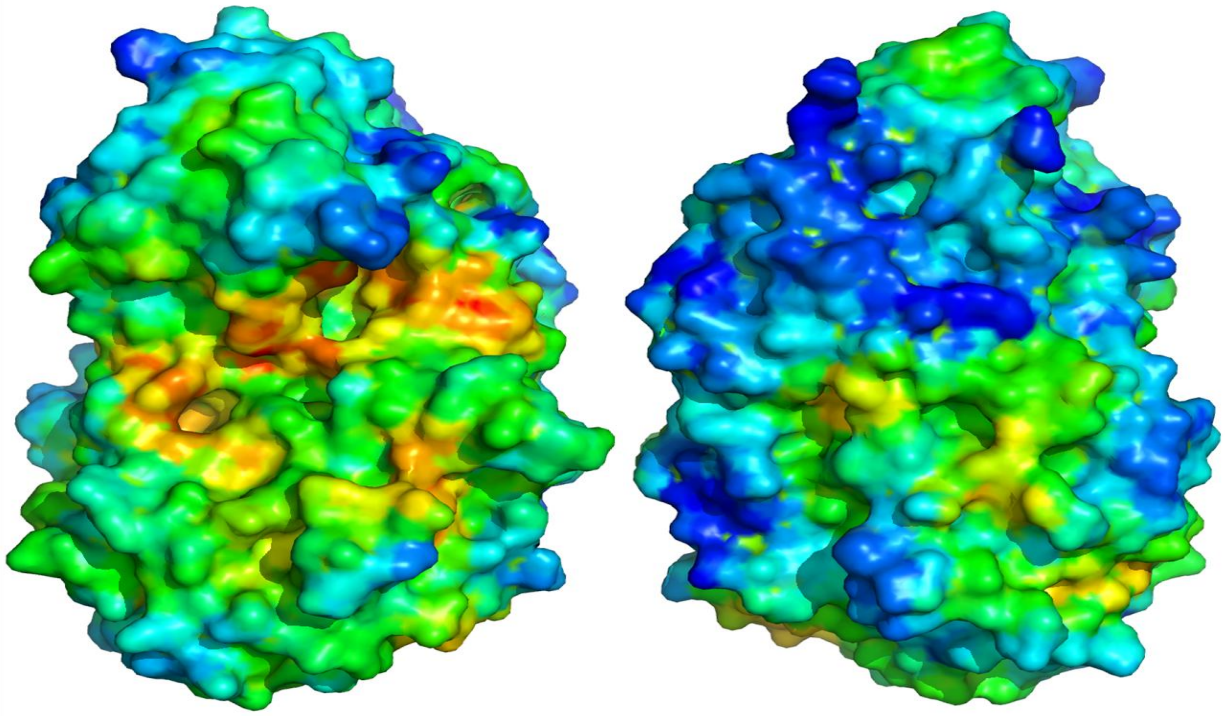
GxGYxYP proteins

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



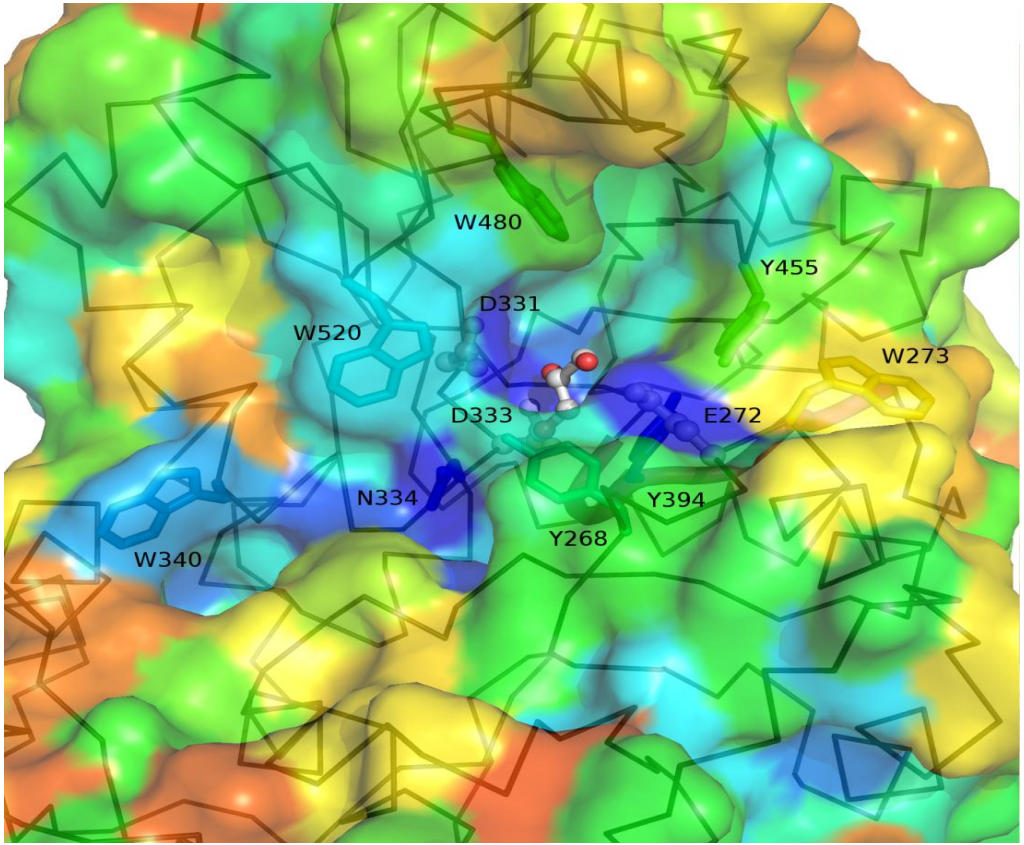
GxGYxYP proteins

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



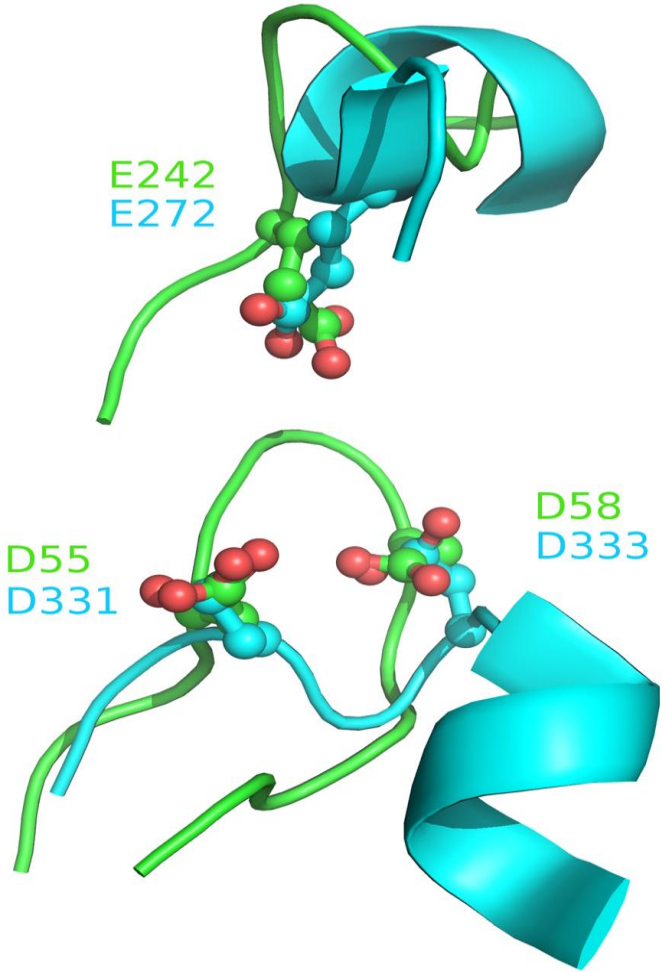
GxGYxYP proteins

- 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

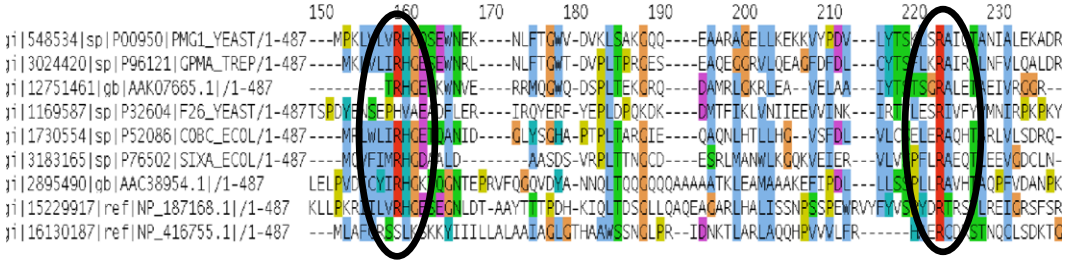
…and finally, you're putting a manuscript together

Calculating and presenting sequence alignments

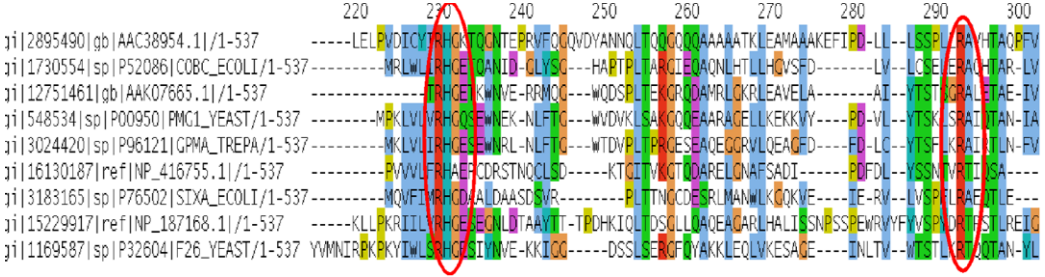
Your sequence alignment

- 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…

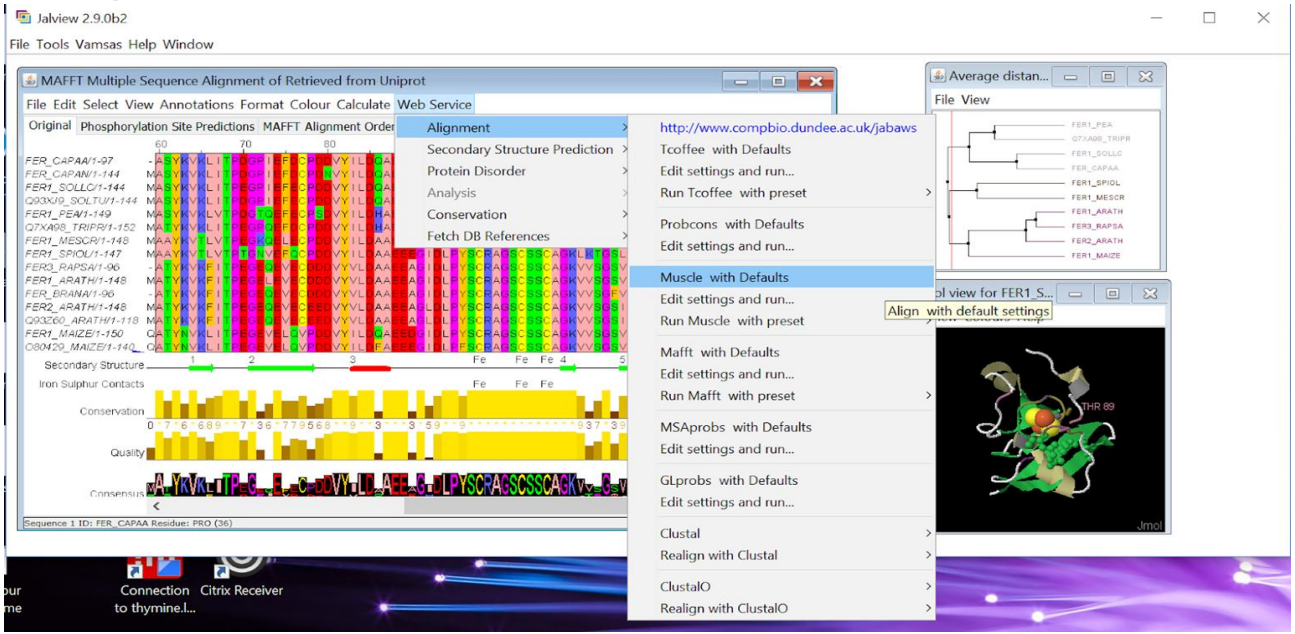


… but MUSCLE gets it



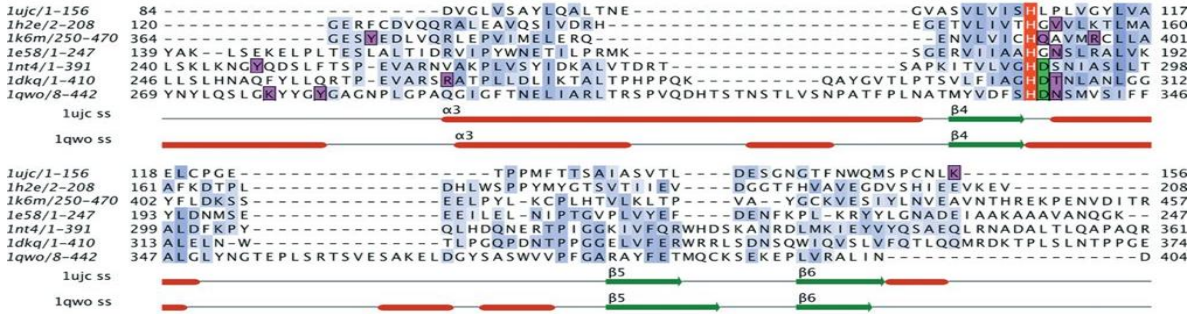
Multiple methods in bioinformatics : Jalview.org

- 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