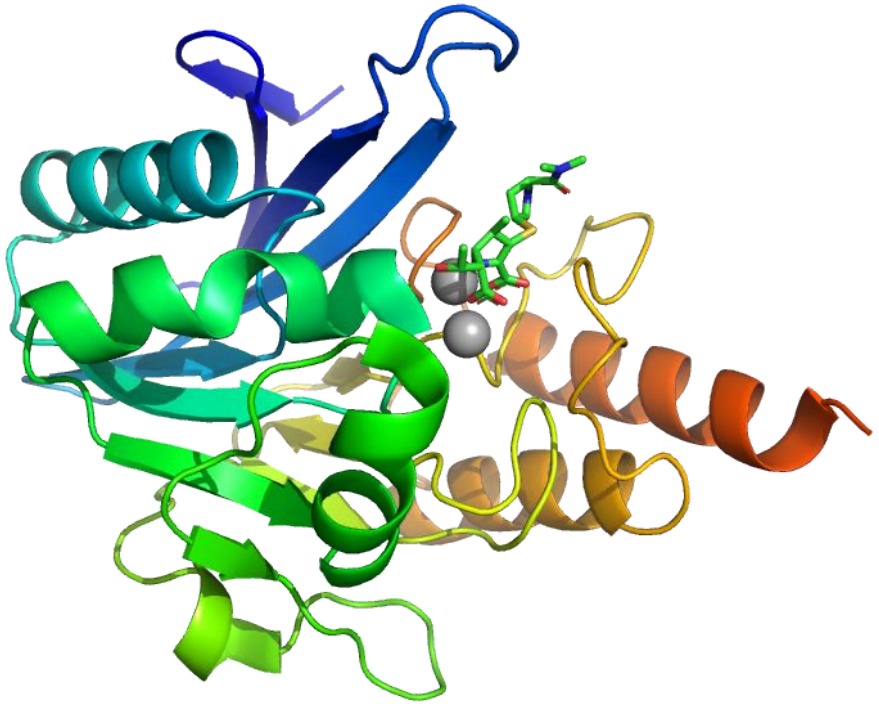
Computational Modeling of Proteins for Infectious Disease Researchers

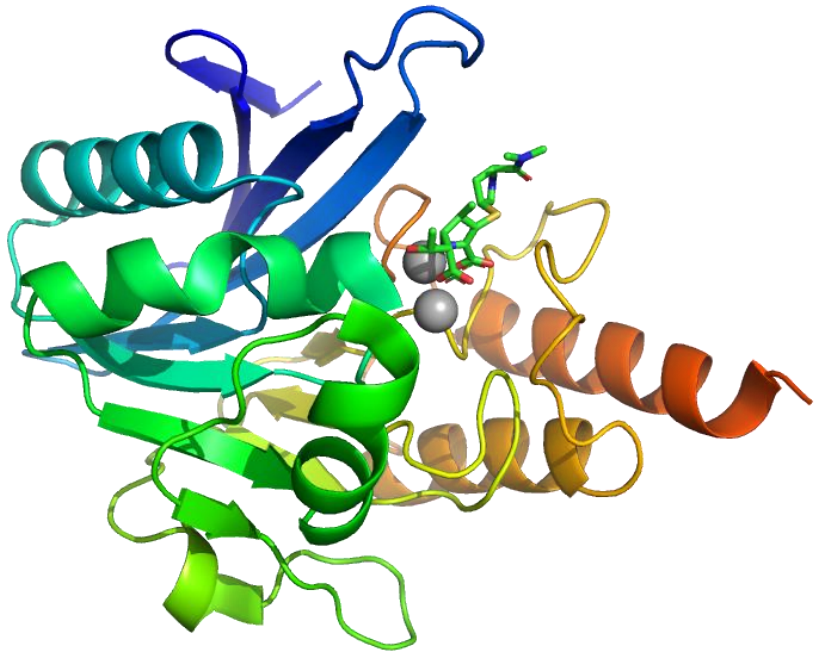


Raczynska JE, Shabalin IG, Minor W, Wlodawer A, Jaskolski M. A close look onto structural models and primary ligands of metallo-β-lactamases. Drug Resistance Updates. 2018 Sep 1;40:1-2. **PDB ID:5N0H**

Workshop

Rockville, MD July 20-24, 2026

Structure prediction and model analysis



Raczynska JE, Shabalin IG, Minor W, Wlodawer A, Jaskolski M. A close look onto structural models and primary ligands of metallo-β-lactamases. Drug Resistance Updates. 2018 Sep 1;40:1-2. **PDB ID:5N0H**

Adam Godzik

csbid

UC RIVERSIDE School of Medicine

What we are trying to achieve today

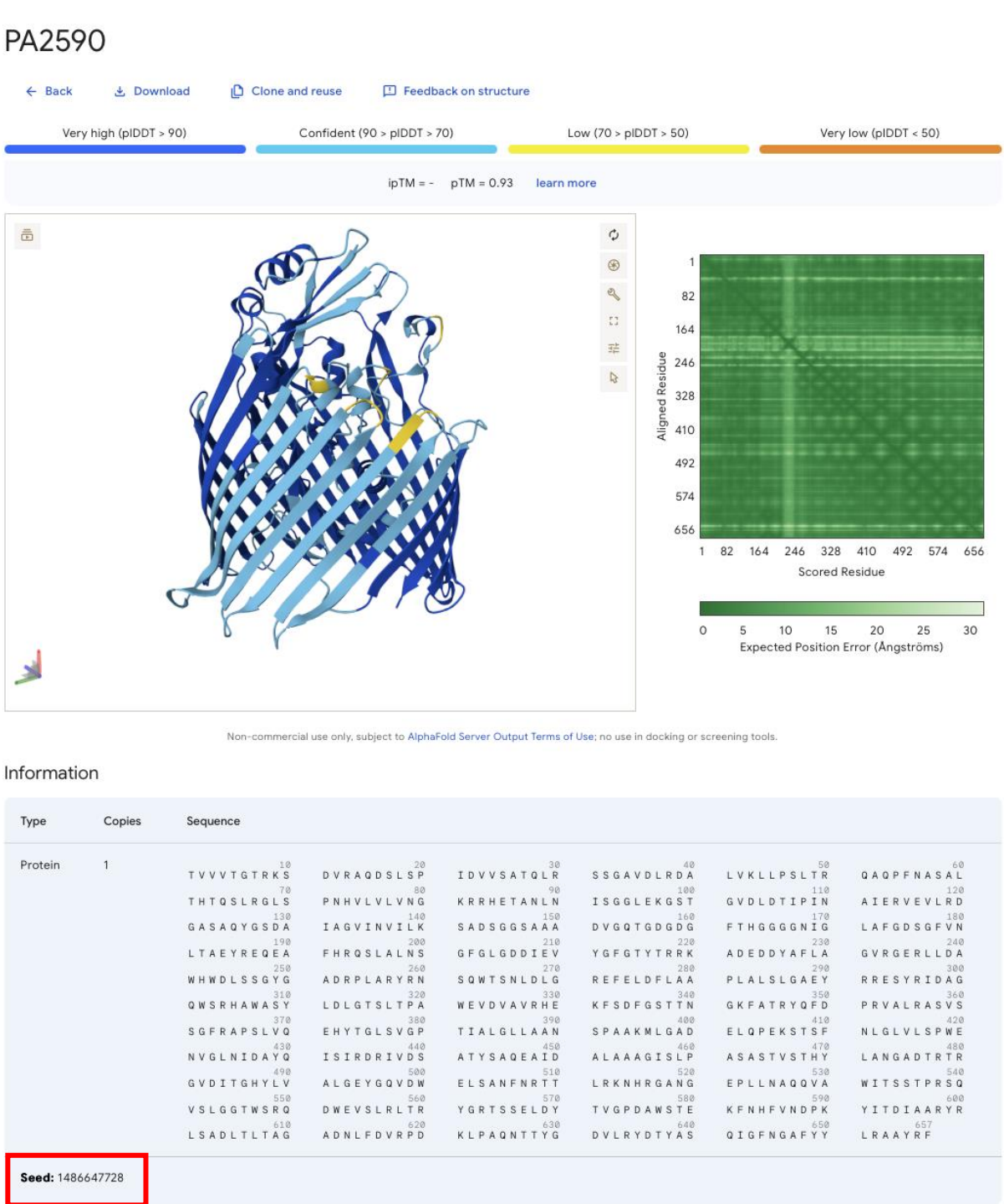
- You are interested in a specific gene/protein
- You probably known the aa sequence of this protein, maybe know something about its function
- The 3D structure of this protein could tell you more
  - Experimental structure is available in PDB (or solved by a collaborator), but you need help analyzing it
  - You learned that it can be modelled by a structure prediction algorithm. It is almost like the previous situation, but there are more caveats
- you are ultimately interested in the function of this gene/protein
  - Nothing is known (uncharacterized protein) – every piece of information is useful
  - A lot is known, but some important information is missing (effect or mechanism of a mutation, unknown variant)

Summary

- Introduction: modeling and model analysis
- Modeling: AlphaFold and beyond
  - Tools: algorithms and servers
- Model analysis: structure similarity
  - Tools: algorithms and servers
- Visualization of protein structures, similarities and differences
  - Chimera and Pymol
- Applications:
  - Fold classification
  - Phylogeny or evolution of proteins
  - Function/feature predictions

AlphaFold

- Now in a third generation (AF3), the first widely used AI protein structure prediction algorithm
- Started the ”AlphaFold revolution”
  - AlphaFold DB (uses AF2) https://alphafold.ebi.ac.uk
  - AF2 models available at UniProt
  - AF3 https://alphafoldserver.com (30/day)
  - Partly open-source
- Spawned numerous clone, variants and inspired new algorithms for protein structure prediction



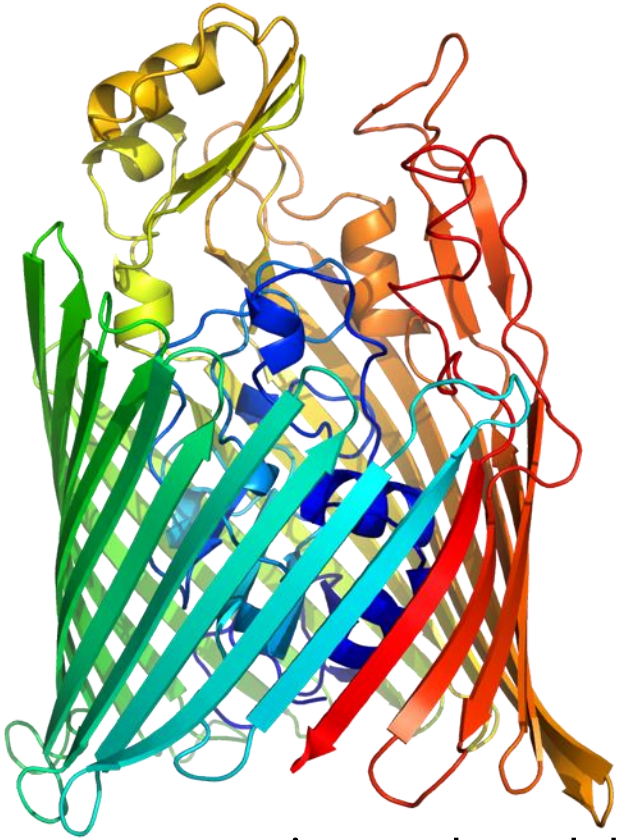
We'll get to this later →

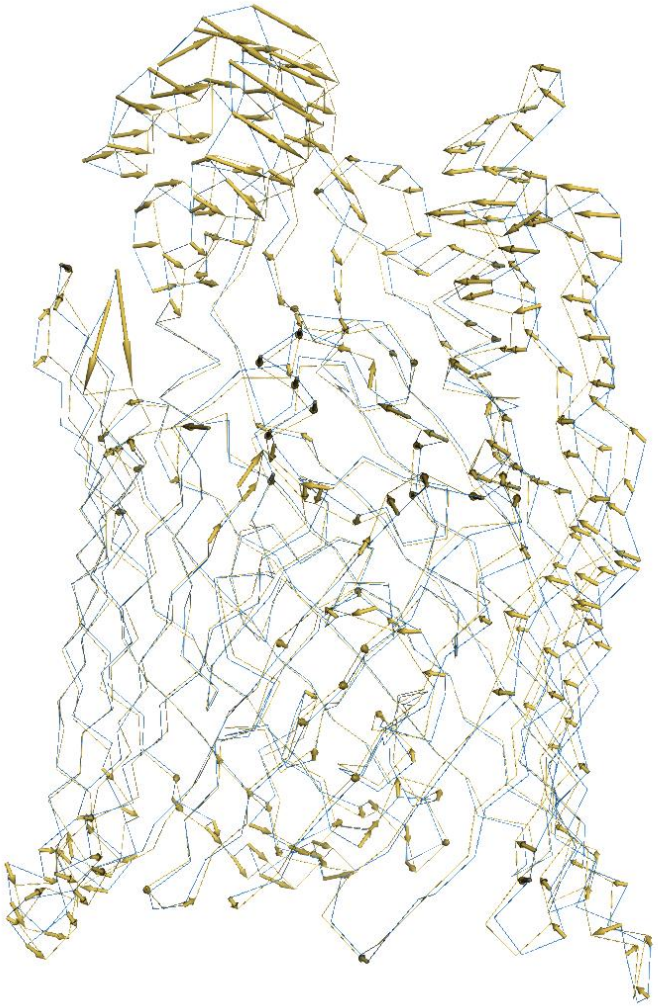
Beyond AlphaFold

- New protein structure prediction algorithms appear almost daily
- They are often validated against experimental structures or AF predictions
  - Is AF3 still a champion? CASP2026 will tell …
- Main competitors:
  - **RoseTTAFold**: David Baker lab, widely published, available as open source and on servers, but less popular as there is no equivalence to AlphaFoldDB, not distributed by UniProt, some models available at PDB
  - **ESMFold**: now in its 3 iteration (ESM3), based on protein language model
    - On-line server and database

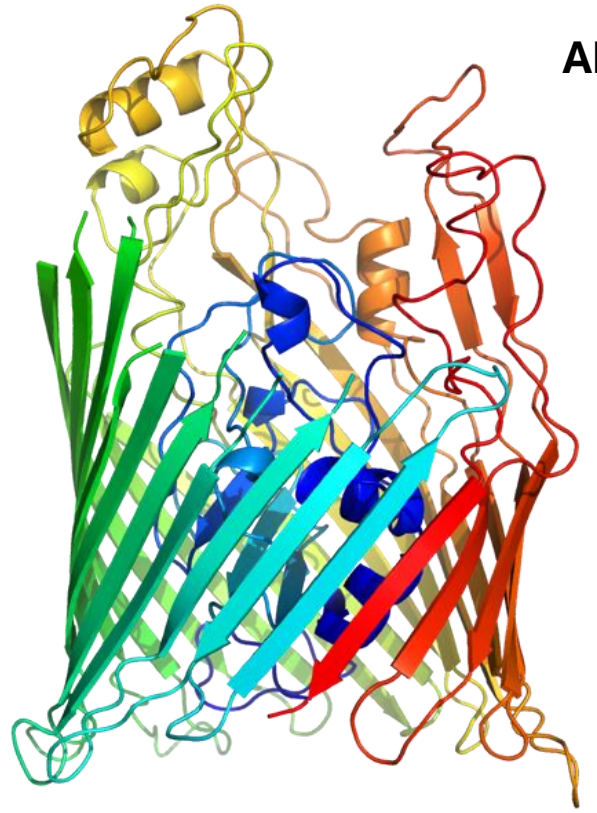
Our example: *Pseudomonas aeruginosa* PA2590

cryoEM





AF3



- Prediction vs an experimental model*
  - Similar, with important structural differences
  - 1.75Å over 657 residues

- Strong similarity in the structural core
  - Low in the loops
  - It is the loops that (mostly) define function

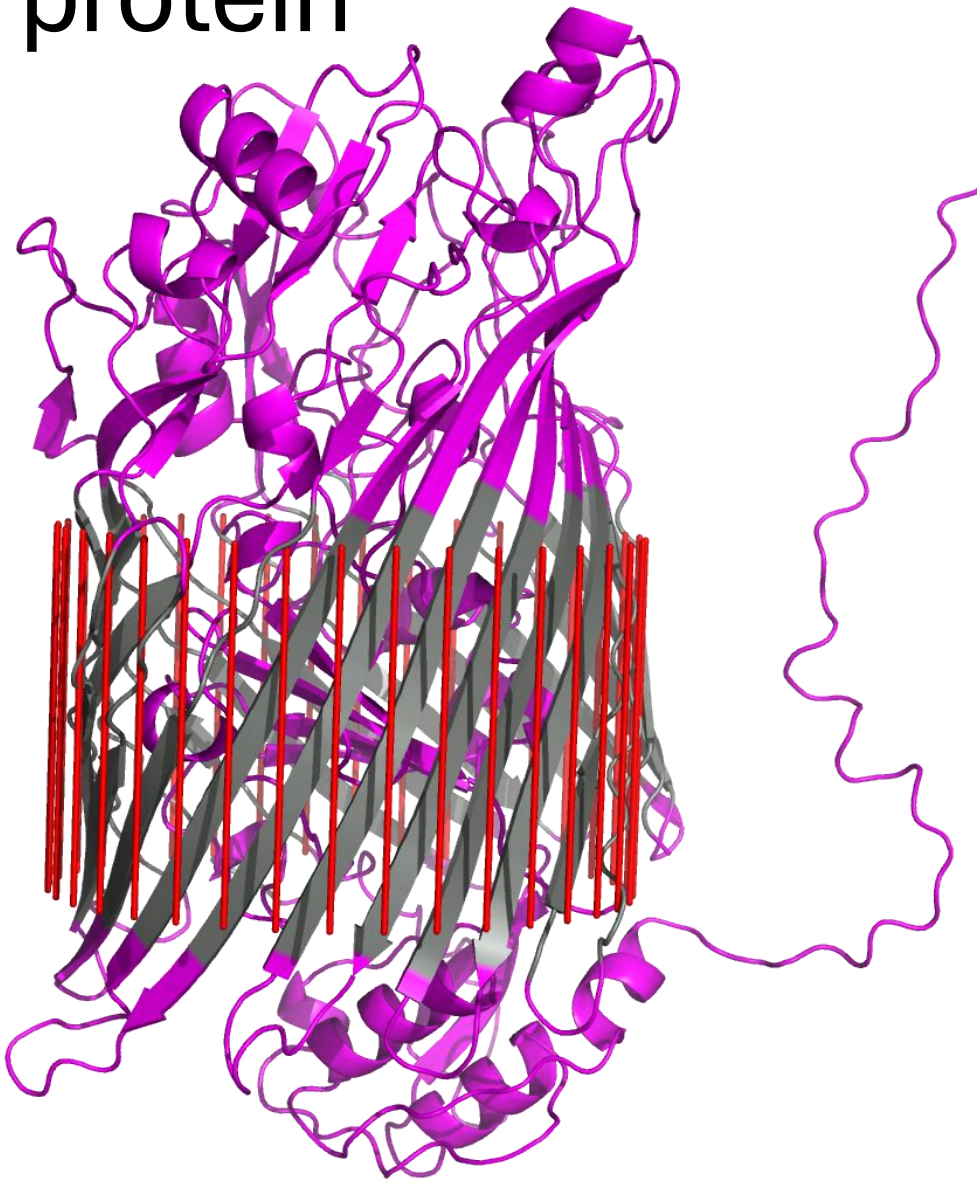
*solved by CSBID in Savchenko and Borek labs

It is a Ton-B dependent receptor, an outer membrane beta barrel protein

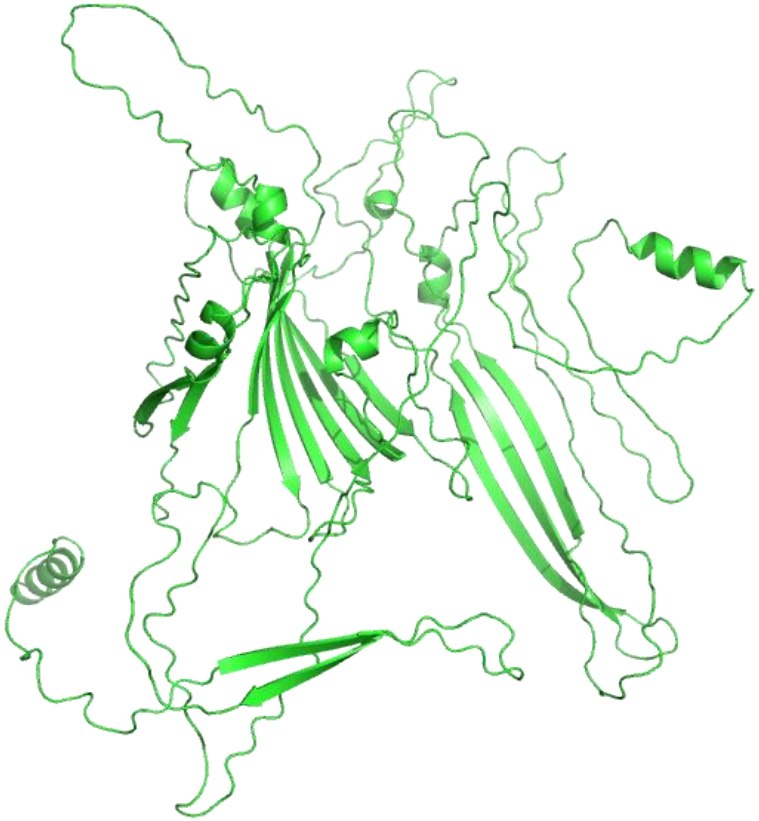
- Conserved fold, core and general function – transport across outer membrane in Gram negative bacteria
  - 8-30+ paralogs in most G- bacteria
- Variable, specific function defining regions (mostly loops)
  - Differing between homologs
- With some variations, a similar picture is seen in most protein families
  - Focus: large paralogous families – OMBBs, LRR, ankyrin repeats

Function and immune response defining extracellular loop

Elliptical cylinder fitted to the structure



Not all AI tools predict it correctly



AF2 prediction
on UniProt and
AlphaFold
database

Models from different* prediction tools are DIFFERENT from each other!

- Myth #1: AI algorithms predict a single structure and a method that is better on a benchmark would provide a better model for your protein
  - Going from the distance constraints to the 3D structure is a partly stochastic process. Repeated modeling with the same input may produce different results. The “best” model has highest probability, the difference to the second best can be very small
    - Most algorithms return several alternative models, seed for a random number generator can make the same algorithm return a different model

*and sometimes even from the same algorithm run multiple times

Differences between models from different algorithms can be interpreted as one model being better than the other.

- Myth #2: Proteins have unique structures
  - Protein structure in physiological conditions is represented by a conformation ensemble. Different algorithms (and different experimental methods and protocols) are sampling this ensemble
    - Experimental models in PDB may have significant structural differences if solved in different conditions, space groups or functional states
    - The computational models may introduce additional, non-physical biases and sample ensemble regions not sampled by real proteins
  - The next big push in protein modeling is going toward modeling ensembles.

Where is the information coming from?

- Multiple sequence alignments (MSAs) translate evolutionary correlations into distances
- **Classic**: AlphaFold 2
- **When it works**: a lot of homologs
- **When it fails**: orfans

- Middle layers in protein language embeddings encode distances between residues
- **Classic**: ESM
- **When it works**: ”typical” sequences
- **When it fails**: unusual composition

- recurrent structural motifs in proteins, put together by empirical rule with pseudo-physics scoring
- **Classic**: Rosetta
- **When it works**: ”typical” structures
- **When it fails**: novel architectures, irregular proteins

What is still missing? Physics!

- Most AI models are empirical **pattern matchers**, developed in a ML style by optimizing fit to experimental data
  - Models from such algorithms are often blind to electrostatics
- Pure physics based models (AMBER, CHARMM) are not competitive in predicting structures of full-size proteins
- Some algorithms from the beginning (Rosetta), some in later iterations (AF3) are using scoring functions with some elements of empirical potentials

Putting it all together: (partial) overview

Family	Principal evidence and inference	Central assumption	Characteristic limitations
AlphaFold2 lineage: AlphaFold2, AlphaFold-Multimer, ColabFold, OpenFold	Sequence, query-specific MSA, optional templates; learned residue-pair reasoning and coordinate generation	Correlated substitutions and templates contain sufficient geometric information, and the learned mapping transfers to the target	Shallow or biased MSAs; alternate conformations; domain orientation; complexes with weak coevolutionary evidence; confidence can remain high for a single preferred state.
Single-sequence PLM predictors: ESMFold, OmegaFold, HelixFold-Single	One sequence; embeddings from a pretrained language model replace the query-specific MSA	The training data has encoded enough evolutionary and structural regularity to infer geometry without structure of homology	Particularly useful for speed and sparse-MSA proteins, but generally less reliable for new interaction geometries, multidomain placement, conformational heterogeneity, and complexes.
All-atom co-folding and diffusion: AlphaFold 3, RoseTTAFold All-Atom, Chai-1, Boltz, Protenix, OpenFold3, HelixFold3	Proteins plus nucleic acids, ligands, ions or modifications; usually a mixture of sequence, MSA, templates and chemical features; diffusion or related all-atom generation	Joint molecular geometry can be learned from the available structural and chemical training distribution	Alternate states, induced fit, allostery, protonation and metal chemistry remain difficult.

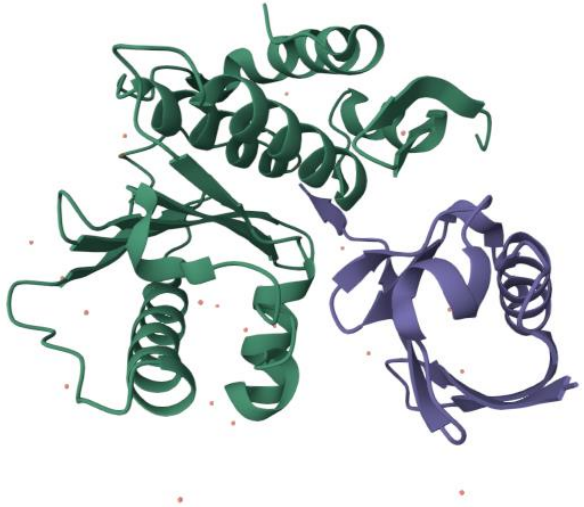
We have a structure of our protein, experimental or predicted. Now what?

- We can visualize it. Great, cool pictures, look great it papers.

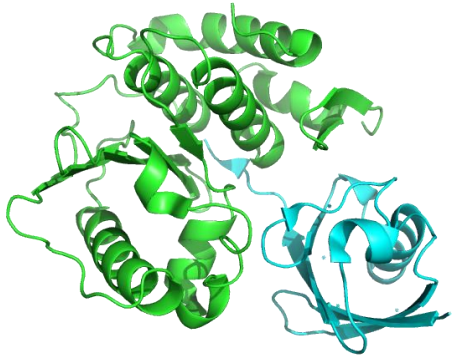
- We can classify it into a fold/superfamily/family
  - If we have enough experience, it could be done just by looking at the structure (topology of beta sheets, number of strands, helices, layers)
  - But usually, we need to compare our protein to other proteins.
    - We can compare sequences
    - We can compare structures
  - We can also analyze specific features of the structure (or sequence)
    - Patterns of charge distribution
    - Presence of cavities, clefts
    - Hydrophobic patches on the surface

Protein structure visualization

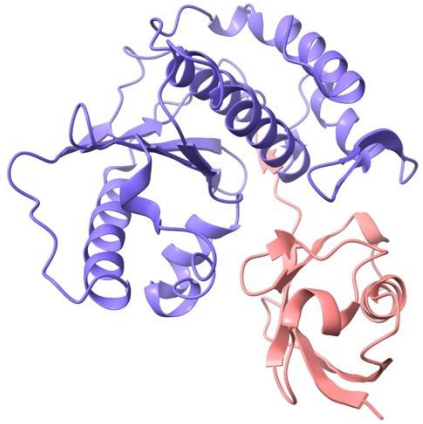
- Embedded visualization in databases (UniProt, PDB – Mol* Viewer, jMol/ jsMol)
  - Fast and easy to use
  - Embedded in web pages
  - Limits on the types of analyses
  - No publication quality figures



- Pymol
  - Distant descendant of Molscript, then MolMol
  - Used to be open source, now commercial
  - Written in python and for python
  - Naturally executes python scripts
    - Large community of users and developers
  - Phenomenal, publication quality pictures



- ChimeraX
  - Great package
  - Open source, large user and developer community
  - developed and maintained at UCSF.
  - A lot of built in functionality **without** coding



How to compare two proteins by sequence similarity

- Two aa sequences alignment optimizing similarity
  - **How to find the alignment –**
    - dynamic programming (fast)
    - BLAST heuristics (very fast)
    - hashing (really fast).
  - How to evaluate the alignment
    - Similarity matrix (Dayhoff, Blosum62)
    - PSSM (position specific similarity matrix): PSI-Blast, HMM
    - Profile-profile: HHpred, FFAS, embedding similarity
- The final alignment, which we can analyze position by position, projects our knowledge from one protein to another

>2G4D_1|Chains A |SENP1 |Homo sapiens (9606)
QDEVLSEAFRLTITRKDIQTLNHLNWLNDEIINFYMNMLMERSKEKGLPSVHAFNTFFFTKLKTAGYQAV
KRWTKKVDVFSVDILLVPIHLGVHWCLAVVDFRKKNITYYDSMGGINNEACRILLQYLKQESIDKKRKEFD
TNGWQLFSKKSQEIPQQMNGSDSGMFACKYADCITKDRPINFTQQHMPYFRKRMVWEILHRKLL

>5AEK_1|Chain A | SENP2 |HOMO SAPIENS (9606)
LELTEDMEKEISNALGHGPQDEILSSAFKLRITRGDIQTLKNYHWLNDEVINFYMNLLVERNKKQGYPALH
VFSTFFYPKLKSGGYQAVKRWTKGVNLFEQEIILVPIHRKVHWSLVVIDLRKKCLKYLDSMGQKGHRICEILL
QYLQDESKTKRNSDLNLLEWTHHSMKPHEIPQQLNGS

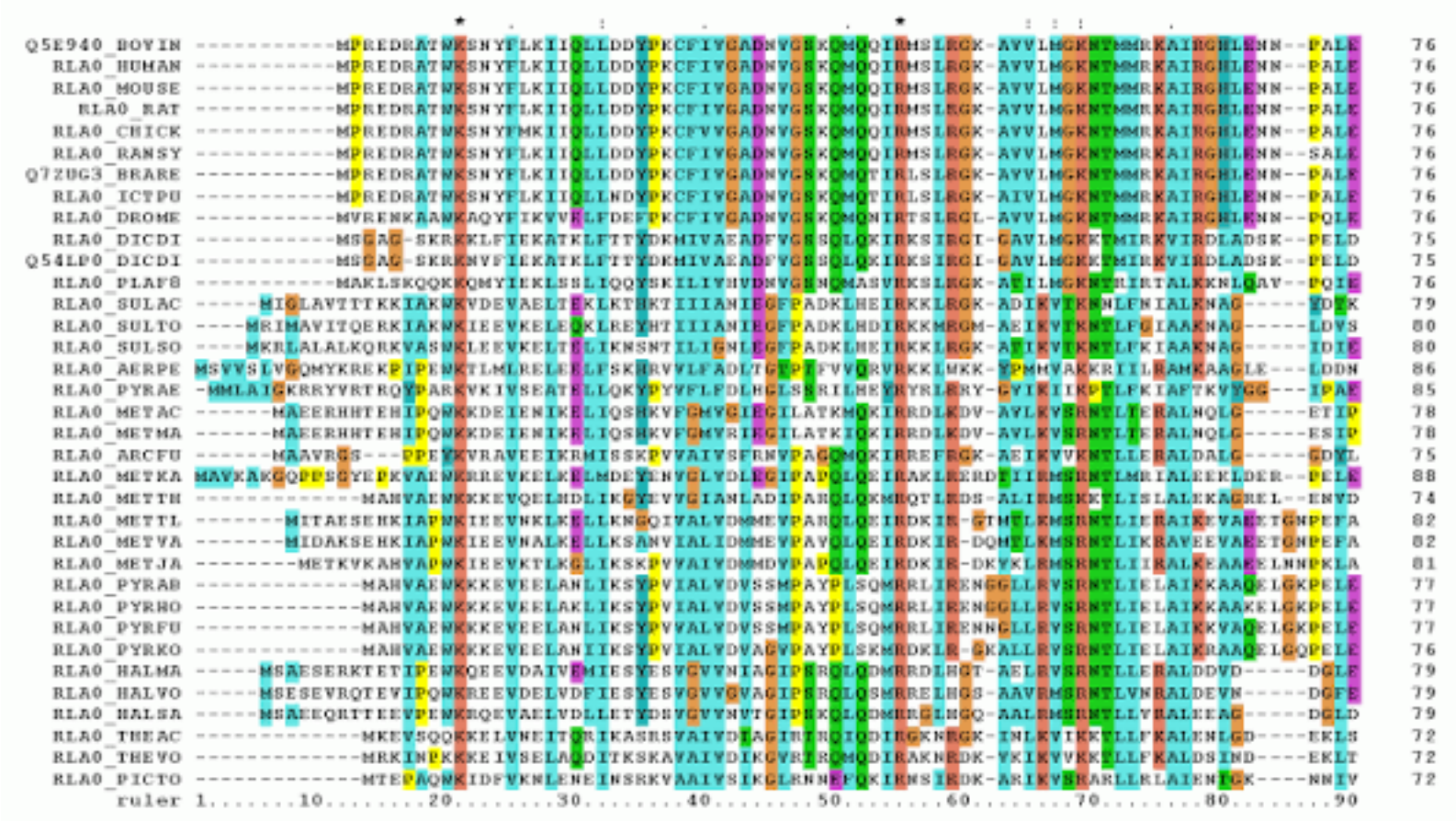
Score	Expect	Method	Identities	Positives	Gaps
202 bits(515)	1e-71	Compositional matrix adjust.	94/162(58%)	121/162(74%)	0/162(0%)

```
Query  1    QDEVLSEAFRLTITRKDIQTLNHLNWLNDEIINFYMNMLMERSKEKGLPSVHAFNTFFFT  60
            QDE+LS AF+L ITR DIQTL + +WLNDE+INFYMN+L+ER+K++G P++H F+TFF+
Sbjct  20   QDEILSSAFKLRITRGDIQTLKNYHWLNDEVINFYMNLLVERNKKQGYPALHVFSTFFYP  79

Query  61   KLKTAGYQAVKRWTKKVDVFSVDILLVPIHLGVHWCLAVVDFRKKNITYYDSMGGINNEA  120
            KLK+ GYQAVKRWTK V++F  +I+LVPIH  VHW L V+D RKK + Y DSMG   +
Sbjct  80   KLKSGGYQAVKRWTKGVNLFEQEIILVPIHRKVHWSLVVIDLRKKCLKYLDSMGQKGHRI  139

Query  121  CRILLQYLKQESIDKKRKEFDTNGWQLFSKKSQEIPQQMNGS  162
            C ILLQYL+ ES  K+  + +   W   S K  EIPQQ+NGS
Sbjct  140  CEILLQYLQDESKTKRNSDLNLLEWTHHSMKPHEIPQQLNGS  181
```

(BLAST alignment with Blosum62)

How to compare many proteins by sequence similarity

- For this we need a multiple sequence alignment (MSA)
  - Typically conservation (or lack thereof) is visualized by colors
  - Conservation at specific positions (columns) can be quantified
    - Position specific mutation matrices
    - HMMs, profiles
  - Conservation patterns between positions (columns) can be analyzed for correlations



(random figure grabbed from Wikipedia)

Protein sequence alignment and its scoring is closely related to evolutionary concepts

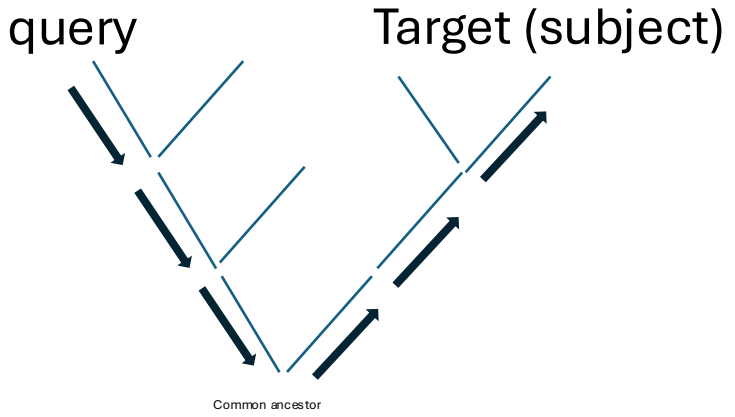
- One sequence can be “translated” to the other by a series of steps (mutations, deletions/insertions) corresponding to natural genomic processes

- Similarity score can be interpreted as a distance between two proteins on a phylogenetic tree while traveling from one protein to another through a common ancestor

- We may not be able to recover it, but there is a unique, correct solution to the sequence alignment between two homologous proteins. All we need are sequences and a time machine (or a good algorithm)

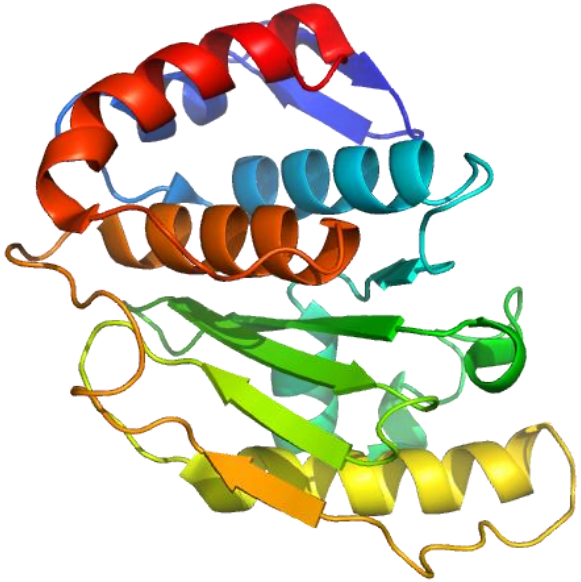
Score	Expect	Method	Identities	Positives	Gaps
202 bits(515)	1e-71	Compositional matrix adjust.	94/162(58%)	121/162(74%)	0/162(0%)

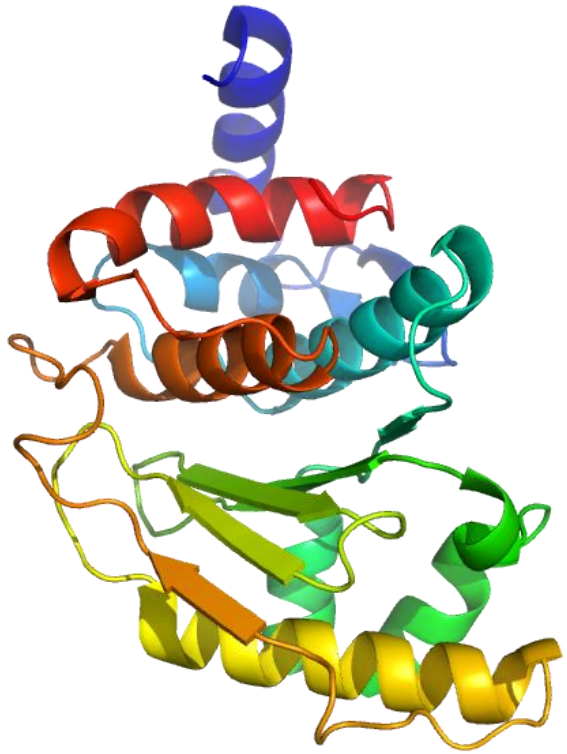
```
Query  1    QDEVLSEAFRLTITRKDIQTLNHLNWLNDEIINFYMNMLMERSKEKGLPSVHAFNTFFFT  60
            QDE+LS AF+L ITR DIQTL + +WLNDE+INFYMN+L+ER+K++G P++H F+TFF+
Sbjct  20   QDEILSSAFKLRITRGDIQTLKNYHWLNDEVINFYMNLLVERNKKQGYPALHVFSTFFYP  79

Query  61   KLKTAGYQAVKRWTKKVDVFSVDILLVPIHLGVHWCLAVVDFRKKNITYYDSMGGINNEA  120
            KLK+ GYQAVKRWTK V++F  +I+LVPIH  VHW L V+D RKK + Y DSMG   +
Sbjct  80   KLKSGGYQAVKRWTKGVNLFEQEIILVPIHRKVHWSLVVIDLRKKCLKYLDSMGQKGHRI  139

Query  121  CRILLQYLKQESIDKKRKEFDTNGWQLFSKKSQEIPQQMNGS  162
            C ILLQYL+ ES  K+  + +   W   S K  EIPQQ+NGS
Sbjct  140  CEILLQYLQDESKTKRNSDLNLLEWTHHSMKPHEIPQQLNGS  181
```



We can also compare 3D structures of two proteins by displaying them next to each other

- And are indeed very similar, especially in the mixed beta sheet between two helical sections





- This is a typical cysteine proteinase fold from $\alpha+\beta$ class (SCOP classification)

There are several classifications of protein structures

- SCOP 2.0*
  - Last update -Jun 2022)
  -
  - Class (7, but really 4)
  - Fold (1,547)
  - Super-family (2,778)
  - Family (5828)
- https://scop.mrc-lmb.cam.ac.uk/

- CATH 4.4 **
  - Last update (Jan 2024, automated updates)
  - Class (4)
  - Architecture (40)
  - Topology (fold) (1375)
  - Homologous super-family (6,631)
  - Sequence family (16,933)
- https://www.cathdb.info/

- ECOD
  - Mar 25 2026
  - Architecture (20)
  - X-group (2,459)
  - H-group (3,715)
  - F-group (16,539)
  - http://prodata.swmed.edu/ecod/

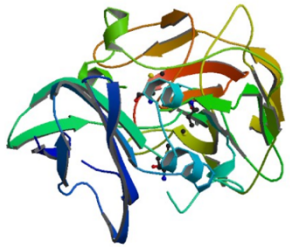
*oldest and most established. Its classifications became a natural language of structural biology

**now expanded into AI prediction databases (TED)

PDB provides classification of experimentally solved proteins

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



View in 3D: NGL or JSmol (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Protein Symmetry: Asymmetric (View in 3D)

Protein Stoichiometry: Monomer

Biological assembly 1 assigned by authors and generated by PISA (software)

Macromolecule Content

- Unique protein chains: 1

Display Files | Download Files

3UP6

Crystal structure of a putative cell adhesion protein (BACOVA_04078) from Bacteroides ovatus ATCC 8483 at 2.80 A resolution

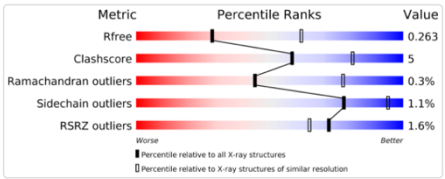
DOI: 10.2210/pdb3up6/pdb

Classification: CELL ADHESION
Deposited: 2011-11-17 **Released**: 2011-12-21
Deposition author(s): Joint Center for Structural Genomics
Organism: Bacteroides ovatus
Expression System: Escherichia Coli

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.8 Å
R-Value Free: 0.250
R-Value Work: 0.203

wwPDB Validation | 3D Report | Full Report



Literature | Download Primary Citation

Crystal structure of a hypothetical protein BACOVA_04078 [Bacteroides ovatus ATCC 8483] (ZP_02067074.1, SP17169A, JCSG 417104) from Bacteroides ovatus ATCC 8483 at 2.80 A resolution

JOINT CENTER FOR STRUCTURAL GENOMICS

To be published

Display Files | Download Files | Data API

3UP6

Crystal structure of a putative cell adhesion protein (BACOVA_04078) from Bacteroides ovatus ATCC 8483 at 2.80 A resolution

External Resource: Annotation

- Domain Annotation: SCOP2 Classification
- Domain Annotation: ECOD Classification
- Domain Annotation: CATH

Domain Annotation: SCOP2 Classification — SCOP2 Database Homepage

Chains	Type	Family Name	Domain Identifier	Family Identifier	Provenance Source (Version)
B	SCOP2 Family	Major fimbrial protein FimA-like	8048038	4005929	SCOP2 (2022-06-29)
B	SCOP2 Superfamily	FimA/Mfa2-like	8102572	3002736	SCOP2 (2022-06-29)
B	SCOP2 Superfamily	FimA/Mfa2-like	8102566	3002736	SCOP2 (2022-06-29)

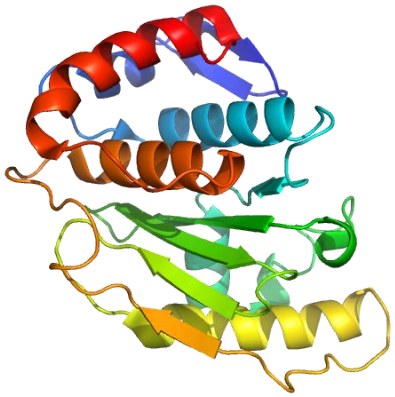
Domain Annotation: ECOD Classification — ECOD Database Homepage

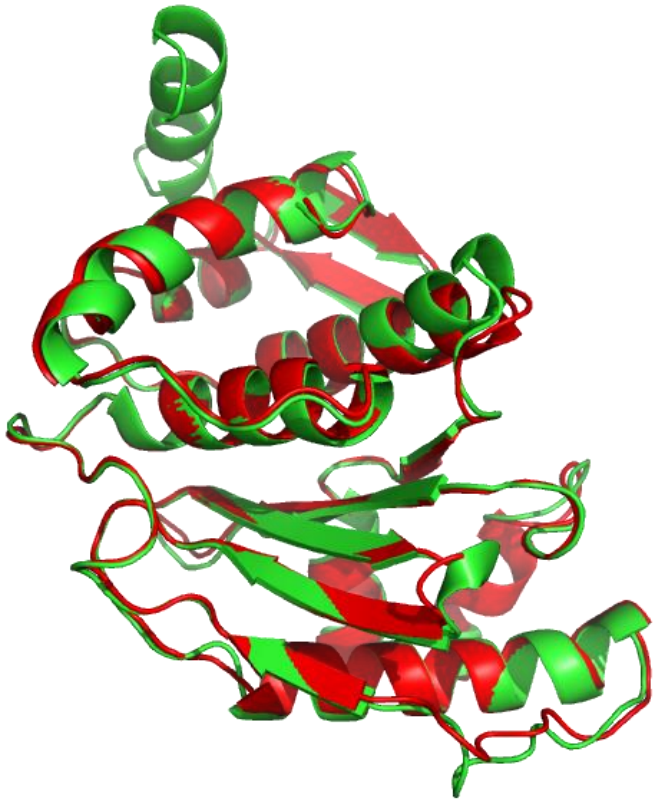
Chains	Family Name	Domain Identifier	Architecture	Possible Homology	Homology	Topology	Family	Provenance Source (Version)
A	P_gingi_FimA_like_1	e3up6A1	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Prealbumin-like	F: P_gingi_FimA_like_1	ECOD (1.6)
A	P_gingi_FimA_like	e3up6A2	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Prealbumin-like	F: P_gingi_FimA_like	ECOD (1.6)
B	P_gingi_FimA_like_1	e3up6B2	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Prealbumin-like	F: P_gingi_FimA_like_1	ECOD (1.6)
B	P_gingi_FimA_like	e3up6B1	A: beta sandwiches	X: Immunoglobulin-like beta-sandwich	H: Immunoglobulin-related	T: Prealbumin-like	F: P_gingi_FimA_like	ECOD (1.6)

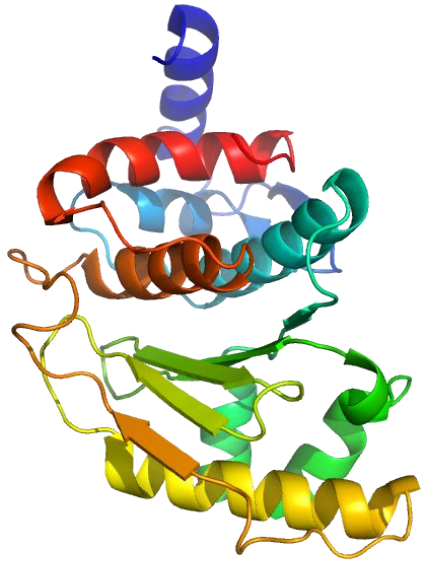
Domain Annotation: CATH — CATH Database Homepage

Chain	Domain	Class	Architecture	Topology	Homology	Provenance Source (Version)
A	2.60.40.2580	Mainly Beta	Sandwich	Immunoglobulin-like		CATH (4.3.0)
A	2.60.40.3380	Mainly Beta	Sandwich	Immunoglobulin-like		CATH (4.3.0)
B	2.60.40.2580	Mainly Beta	Sandwich	Immunoglobulin-like		CATH (4.3.0)
B	2.60.40.3380	Mainly Beta	Sandwich	Immunoglobulin-like		CATH (4.3.0)

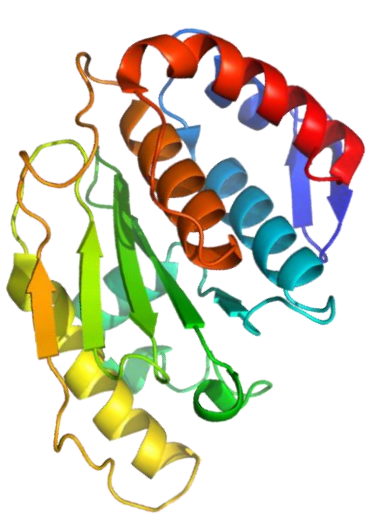
Their similarity can be visualized as the structural superposition

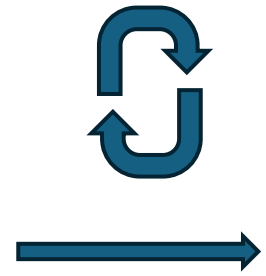




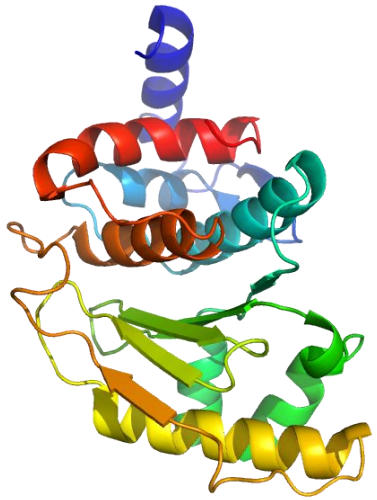


This is an easy calculation





6 degrees of freedom
- **3 d transcription vector**
- **3 Euler angles**

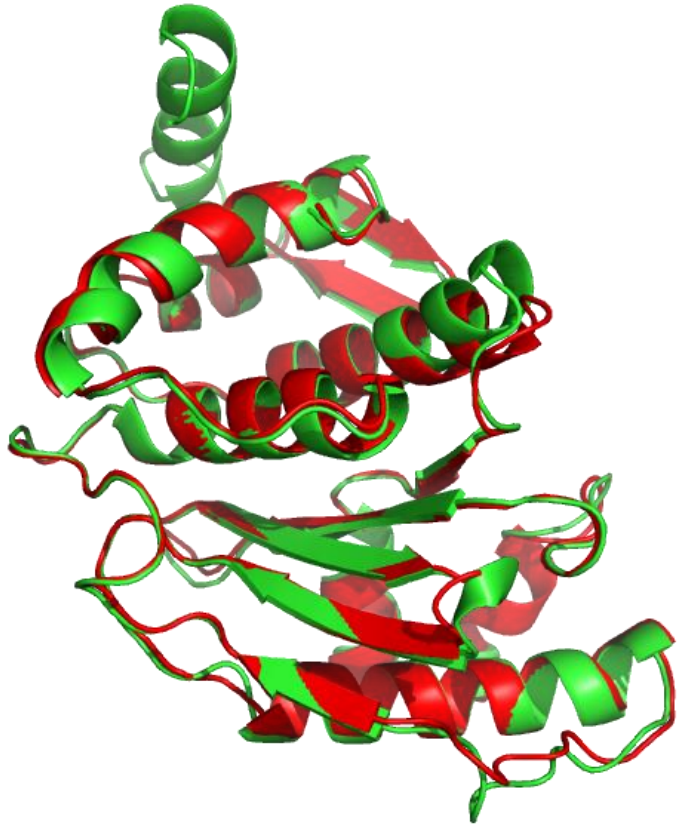


How to quantify the structural similarity?

- 3D approaches: for a given alignment, optimal superposition can be easily calculated, then:
  - RMSD – root mean square after optimal superposition
    - 205 positions with an RMSD of 0.92Å, values below 3Å are considered significant
  - TM-score - (Template Modeling score)

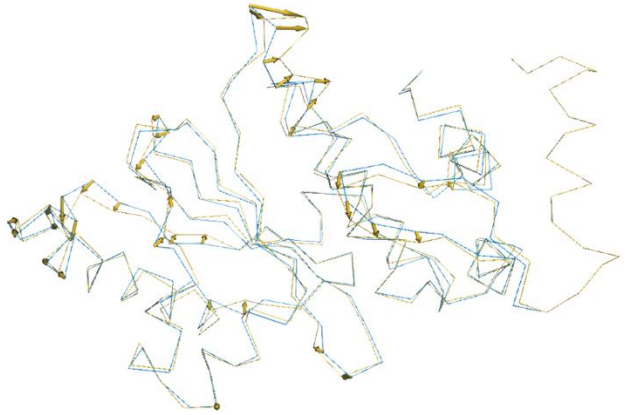
$$\text{TM-score} = \max\left[\frac{1}{L_{\text{target}}}\sum_{i}^{L_{\text{common}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right]$$

    - TM-score falls between 1 (identical), >0.5 (similar fold) to <0.2 (no similarity), for our pair T-score is 0.97
  - distance (or contact) map overlap (rarely used)



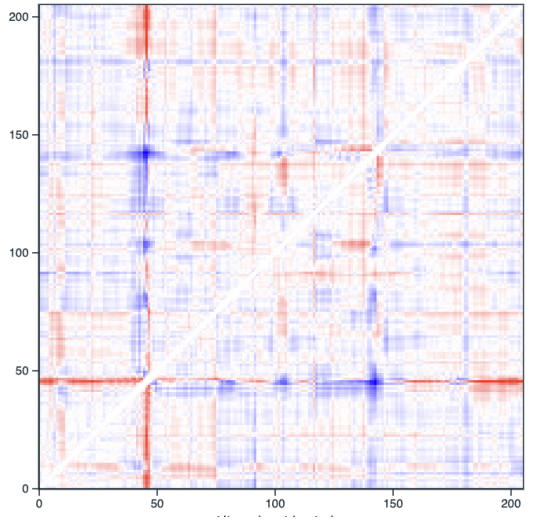
Each tells us something, but …

- RMSD
  - Strongly depends on the length of the alignment
  - Measures global similarity, sacrifices local one
- it is very intuitive and widely used, great visualization



- TM-score
  - Not very intuitive, no useful visualization
  - Initially designed to evaluate quality of models
  - Measures global similarity , sacrifices local one
- (partly) addresses the score vs. length issue

- Distance/contact map overlap
  - Not very popular
  - But (partly) addresses the score vs. length issue and local vs. global similarity
  - Can identify local conservation patterns



Can the optimal the alignment optimizing structural similarity be found?

- For a full 3D alignment, no
  - Combination of the alignment and superposition is an NP-complete problem (hard)[1]
  - A solution with the accuracy (X-e), where X is the optimal solution and e is arbitrarily small can be found in polynomial time
- All existing programs use simplified assumptions (heuristics) to find **approximate** solutions. They differ between programs!
  - They also use different similarity measures, gap penalties and other parameters that further make their results different

[1] Goldman, D., Istrail, S. and Papadimitriou, C.H., 1999, October. Algorithmic aspects of protein structure similarity. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)* (pp. 512-521). IEEE.

[2] Kolodny, R. and Linial, N., 2004. Approximate protein structural alignment in polynomial time. *Proceedings of the National Academy of Sciences*, *101*(33), pp.12201-12206.

Heuristics give reasonable results, but don't expect them to be identical

- **DALI** (Distance-matrix Alignment) – similarity of distance matrices
- **TM-align** based on TM-score (Template Modeling score)
- **FATCAT** - (Flexible structure alignment by chaining aligned fragment pairs) – internal score, including penalties for hinge movements
- **MatchMaker** (ChimeraX) – hybrid method based on a CE algorithm + later iteration
- **Align** (Pymol) – sequence alignment iteratively improved to optimize RMSD

- And 100+ other algorithms ….

There is another solution - FoldSeek

- Instead of comparing 3D coordinates directly, 3D structure is translated into a “structural alphabet” (3Di)
- A modified MMseqs2 algorithm is used to compare now linear sequences (in 3Di alphabet)
- A result is 50-1000 times faster than Dali or other 3D based programs
  - This approach was used before, but the additional speed-up of FoldSeek made is uniquely useful with the availability of massive databases of predicted structures
    - SSAP (**Taylor** & **Orengo**, 1989b) used scoring between “structural environments”
    - *PDBeFold* used secondary structure matching

Each is based on certain assumptions (heuristic) to provide an estimate to the undoable (NP-hard) problem

- **DALI**
  - Contact map is split into 6x6 submatrices and similar contact patterns are recognized
  - Iteratively expands submatrices to build a whole protein alignments.
  - Unique features/ advantages – very stable significance analysis
  - Sensitive to noise/errors in sidechain positions
- **TM-align**
  - Calculates structural similarity between short (5aa) structural fragments and connects them by iteratively extending them optimizing TM-score
  - Unique features/ advantages – handles proteins of different lengths, easy interpretation of the RM-score
  - If the initial seed is wrong, it may not recover
- **FATCAT**
  - Identifies short aligned structural fragments and chains them using dynamic programming with two types of gaps – inserts/deletions and hinge motions
  - Unique features/ advantages - can account for protein flexibility
  - May overfit structures using hinge movements
- **FoldSeek**
  - Translates 3D structure into one dimensional “structural alphabet
  - Unique features/ advantages - **speed**
  - Is not sensitive to small hinge motions in the structures

Structural alignments obtained using different algorithms are different from each other

- Myth #3. There is a single unique structural alignment between two protein structures.
  - Alignments differ because of different heuristics used by different algorithms
  - Alignments differ because of different measures of structural similarity used by different algorithms
    - Different algorithms use different internal similarity measures, but usually present their results as RMSD or TM-score

There is an old paper on this …

Table 1. *Comparison of several published structural alignments of azurin (1azcA) and plastocyanin (1plc) structures*[a]

Alignment source	RMSD	No. contacts	No. identities	Alignment length	Structural information
Contact overlap	5.1	108	15	89	Interactions
Adman (1985)	6.7	76	18	94	Visual
Taylor and Orengo (1989)	3.9	96	13	86	Environment
Chothia and Lesk (1982)	2.9	72	11	70	Core packing
DALI (1995)	4.5	68	13	93	Cα distances
Cα distance	2.4	72	12	80	Cα distances
Gerstein et al. (1994)	5.7	75	14	96	Core packing
Pascarella and Argos (1992)	2.5	75	16	73	Hybrid
May and Johnson (1994)	4.7	73	12	89	Cα RMSD
GAFIT	1.5	65	11	60	Cα RMSD
MNYFIT	1.6	13	4	51	Cα RMSD
Sequence	8.4	52	24	98	Gonnet

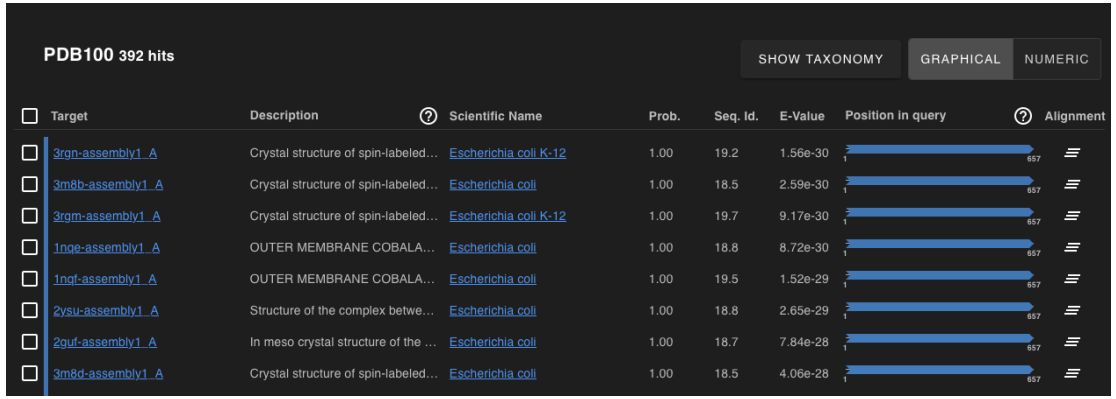
[a] The entire length of azurin (plastocyanin) is 129 (99) amino acids and there are 239 (173) pair interactions in the 1azcA (1plc) structure.

From **A. Godzik "The structural alignment between two proteins: Is there a unique answer ?"** ***Protein Science*** **5:1325-1338 (1996).**

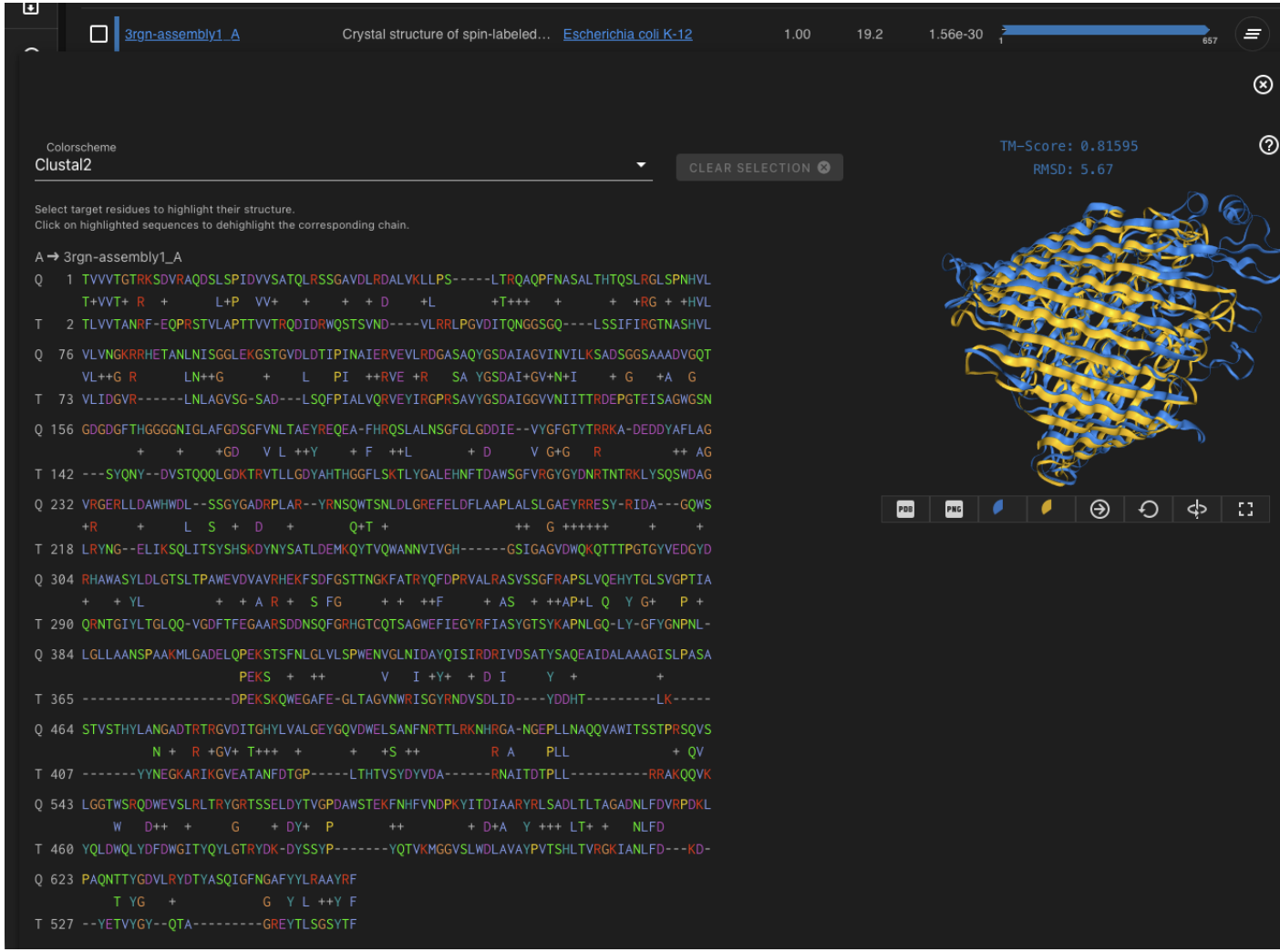
Let's try some examples

- Imagine we don’t know anything about PA2590. Let's search for similar structures in PDB

- FoldSeek
  - Time - instantaneous



Top hit: 3RGN, 5.67Å RMSD



Let's try some examples

- Imagine we don’t know anything about PA2590. Let's search for similar structures in PDB

- Dali
  - Time ~10 minutes

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

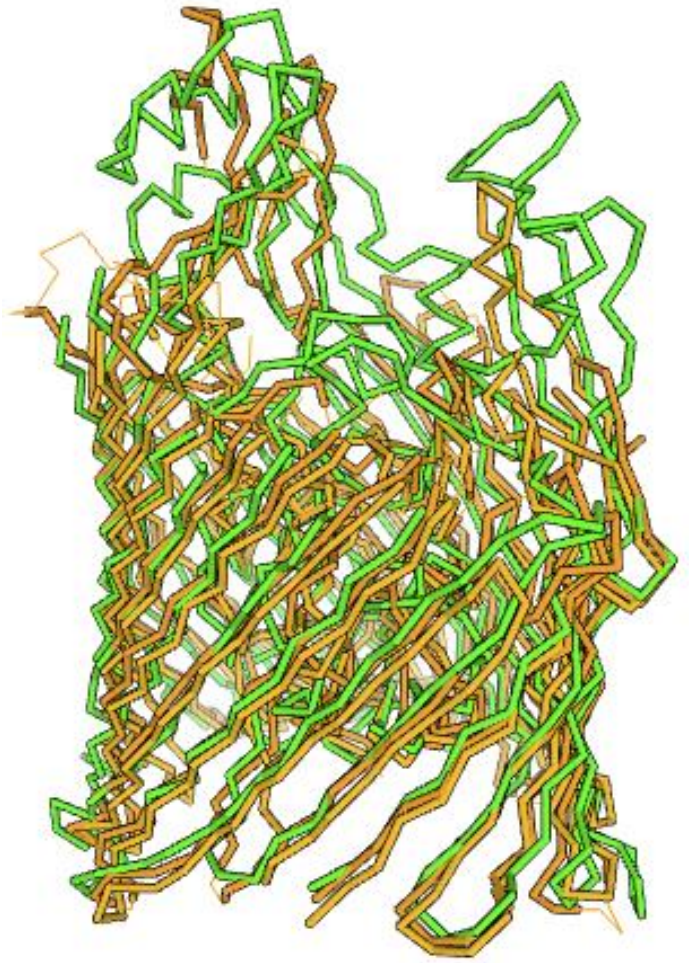
Structural Alignment | Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Pfam | Reset Selection

Summary

The list is truncated at 100 hits.

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	2hdi-A	37.1	2.3	544	598	22	PDB	MOLECULE: COLICIN I RECEPTOR;
2:	1nqf-A	36.8	2.5	527	544	21	PDB	MOLECULE: VITAMIN B12 RECEPTOR;
3:	1nqe-A	36.8	2.5	527	549	21	PDB	MOLECULE: VITAMIN B12 RECEPTOR;
4:	3rgm-A	36.7	2.5	526	548	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;
5:	3rgn-A	36.7	2.5	527	549	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;
6:	1nqh-A	36.6	2.5	527	584	21	PDB	MOLECULE: VITAMIN B12 RECEPTOR;
7:	3m8b-A	36.6	2.5	526	549	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;
8:	9nn6-A	36.2	2.2	556	632	21	PDB	MOLECULE: COLICIN I RECEPTOR;
9:	2guf-A	36.0	2.5	517	551	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;
10:	2ysu-A	35.4	2.4	527	576	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;
11:	1ujw-A	35.4	2.4	527	576	21	PDB	MOLECULE: VITAMIN B12 RECEPTOR;
12:	2hdf-A	35.1	3.2	528	584	21	PDB	MOLECULE: COLICIN I RECEPTOR;
13:	1nqg-A	35.1	2.6	527	576	21	PDB	MOLECULE: VITAMIN B12 RECEPTOR;
14:	3m8d-A	35.1	2.5	527	582	21	PDB	MOLECULE: VITAMIN B12 TRANSPORTER BTUB;

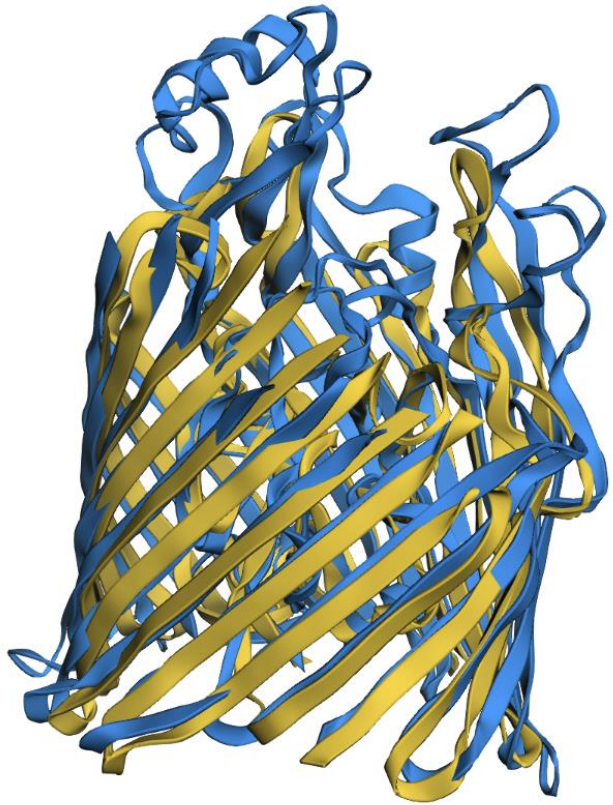
Top hit: 2HDI, 2.3Å RMSD



Multiple structure alignment of 3 top hits

Let's compare the top hit using different alignment programs

- FATCAT
  - 2.69Å RMSD, p-value 0.0



- DALI
  - 2.5Å RMSD, Z-score 36.7
  - No working visualization on the website

Why structure classification or similarity searches are interesting?

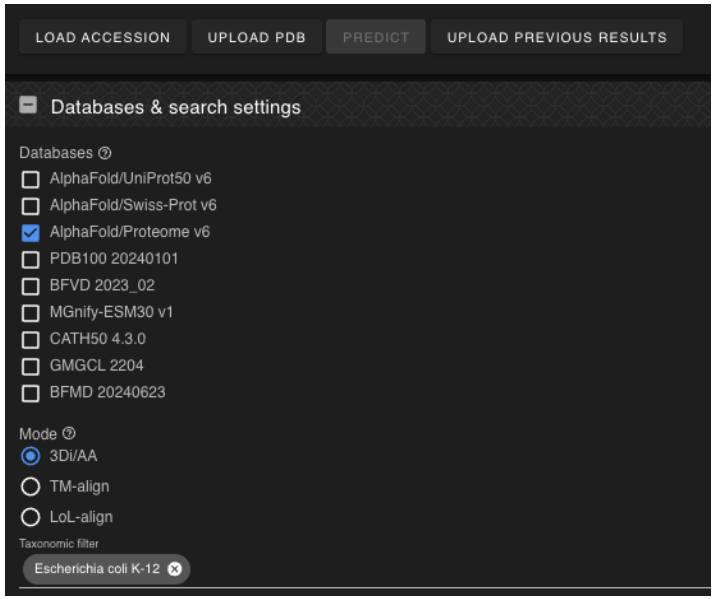
- Sequences
  - ① Identical
  - ② Very similar, visual inspection
  - ③ Similar (30% seq identity (BLAST)
  - ④ Twilight zone, PSI-BLAST, HMM
  - ⑤ Profile-profile, sometimes undetectable
  - ⑥ undetectable

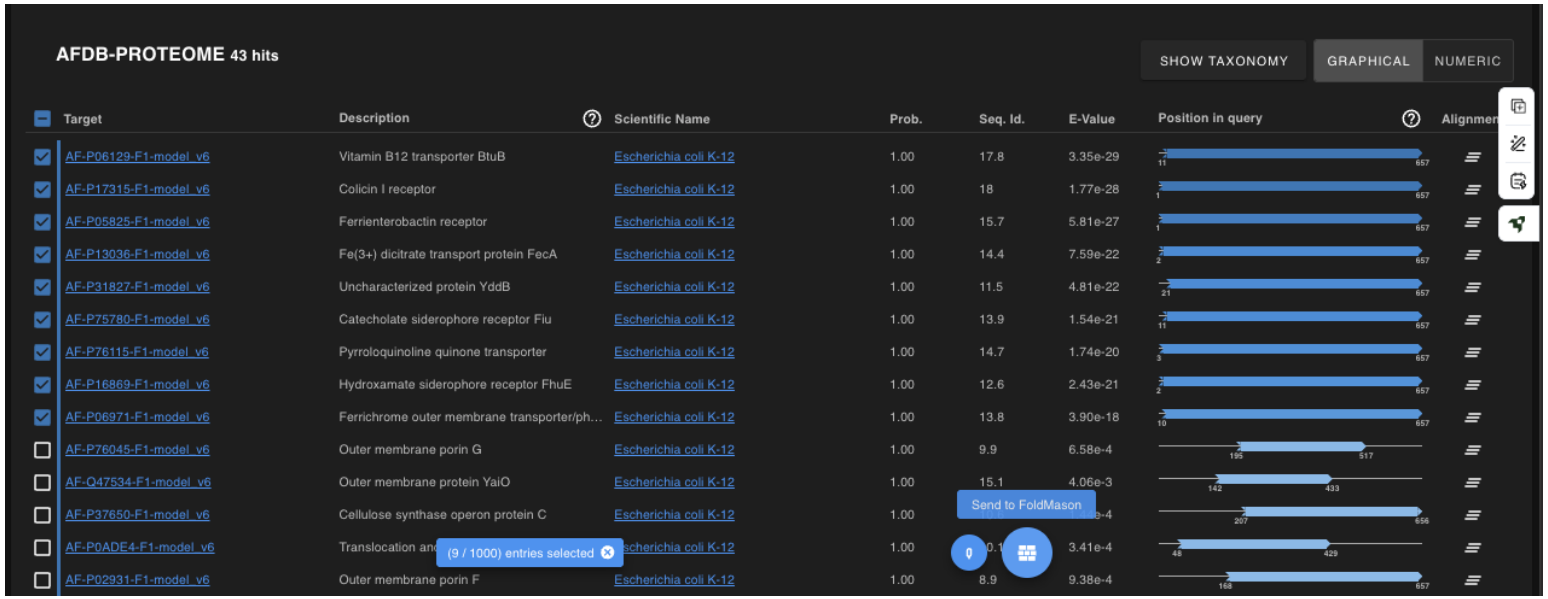
- **"predictions" from molecular taxonomy**
  - ① Individual proteins
  - ② Very close homologs
    - **one-to-one orthologs**
    - **conserved function**
  - ③ Close homologs (modeling families)
    - **orthologs**
    - **similar, but different function**
  - ④ Families
    - **broad functional similarity**
  - ⑤ Superfamilies
    - **similar functional class**
  - ⑥ Folds
    - ????

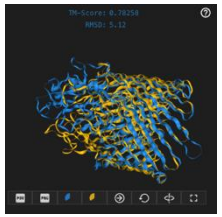
- Structures
  - ① Differences between functional states
  - ② Very similar, <1Å RMSD
  - ③ Similar, 1-2Å RMS
  - ④ Most of the structure is similar, 2-3Å
  - ⑤ Parts of the structures are similar ~3Å
  - ⑥ Simplified diagrams, sometimes impossible to align

Because of its speed, FoldSeek opens new possibilities

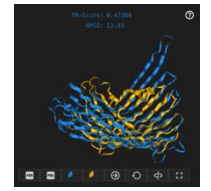
- Genome wide searches for paralogs
  - How many PA2590-homologs are in *E.coli*?
- Links to multiple structure alignments







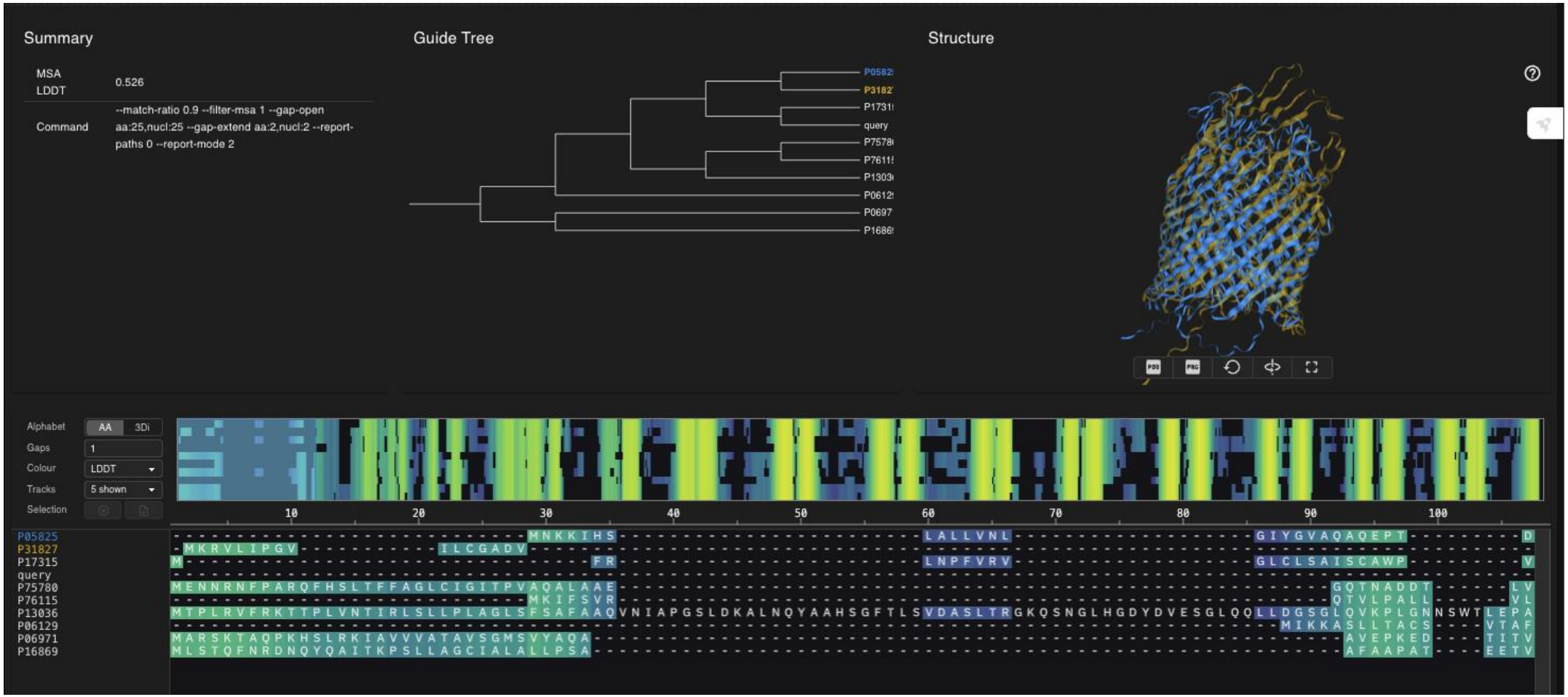
Full length



Partial
or bad models

FoldMason

FoldMason


Summary
MSA
LDDT
0.526
Command
--match-ratio 0.9 --filter-msa 1 --gap-open aa:25,nucl:25 --gap-extend aa:2,nucl:2 --report-paths 0 --report-mode 2
Guide Tree
query
Structure
Alphabet
AA
3Di
Gaps
1
Colour
LDDT
Tracks
5 shown
Selection
10
20
30
40
50
60
70
80
90
100
P05825
P31827
P17315
query
P75780
P76115
P13036
P06129
P06971
P16869

How to make use from the massive structural coverage of protein families?

- Structural phylogeny
  - Treat structural similarity as an ultimate measure of the phylogenetic distance
  - Do (if previously impossible) or redo (if previously didn't make sense) phylogenetic trees and evolutionary history of families
    - Structural comparison as a ground truth

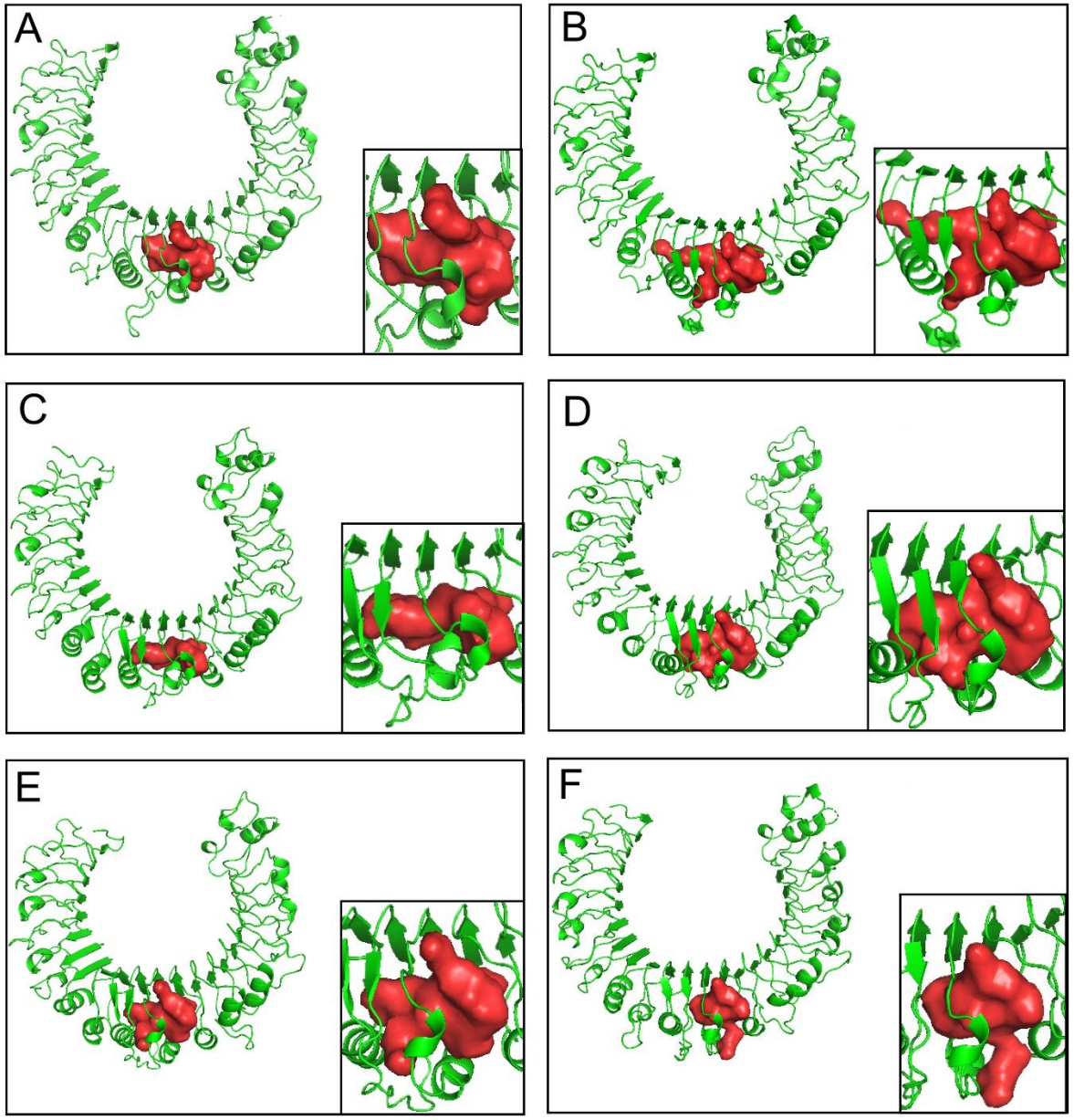
- **The good**: can extend phylogeny beyond the limits of sequence based measures
- **The bad**: structural similarity is not a good measure of evolutionary distances
  - Convergent evolution
  - Uneven pace of structural changes in evolution

- phylogenetic mapping of structural features
  - Core structural features are usually ancestral: fold defining
  - Small structural features can arise independently, recur during evolution responding to evolutionary pressure
  - And yet they often define function, or at least the features we would like to control
    - Binding cavities
    - Long loops
    - Inserted domains

- **The good**: detailed function prediction
- **The bad**: need examples or better physics based methods to predict detailed functions

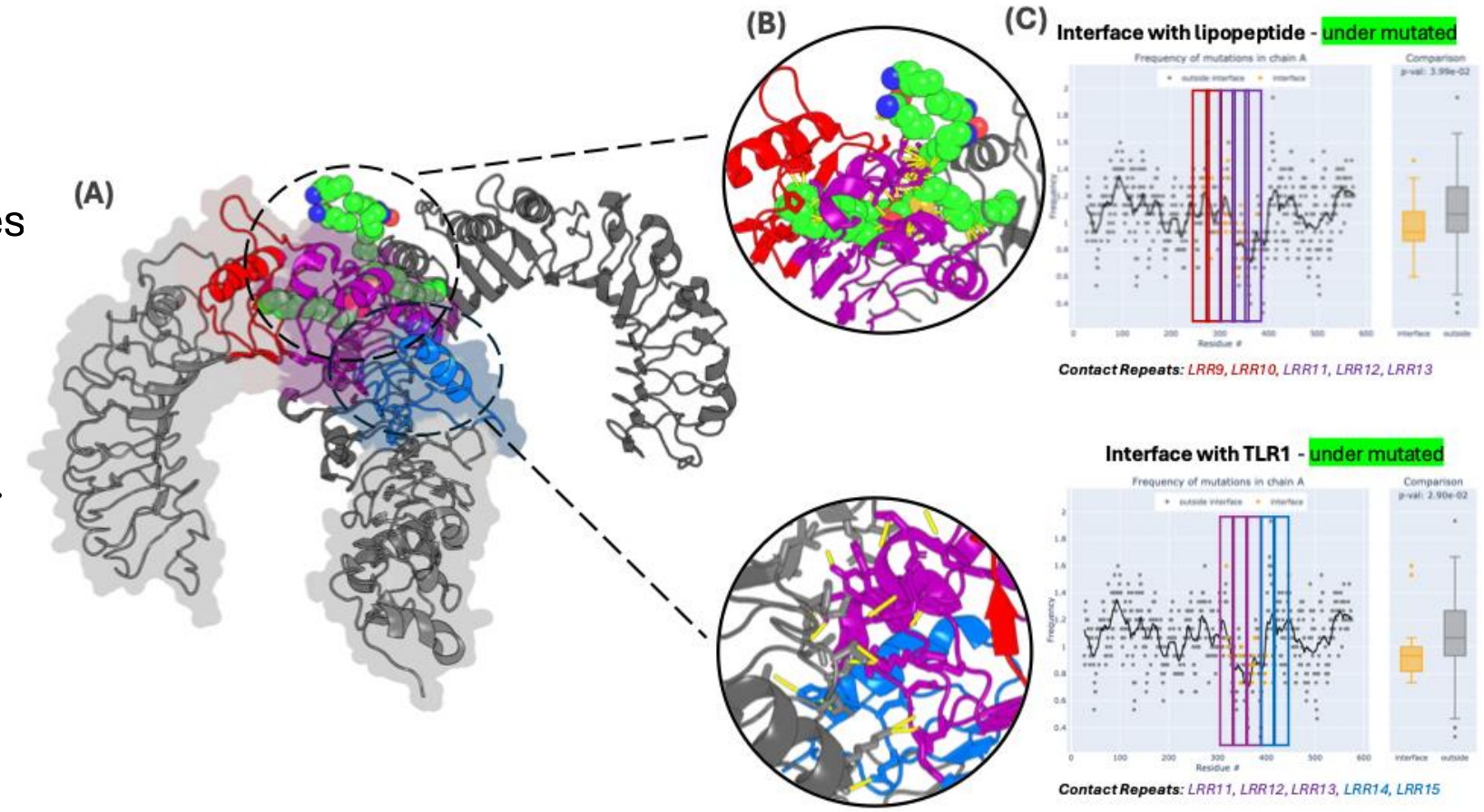
Examples from our recent papers

- Toll-like receptor 2 (TLR2)
  - TLR with a ligand binding cavity
  - Conserved in (almost) all vertebrates
    - >250 models analyzed, previously 6 structures were available
    - The cavity is apparently lost in some species of jawless fish
    - More variable in paralogs
  - Hypothesized to predate vertebrates as ligand binding cavities were found in some invertebrates, but detailed analysis suggests relative dearth and independent origin of such cavities in invertebrates

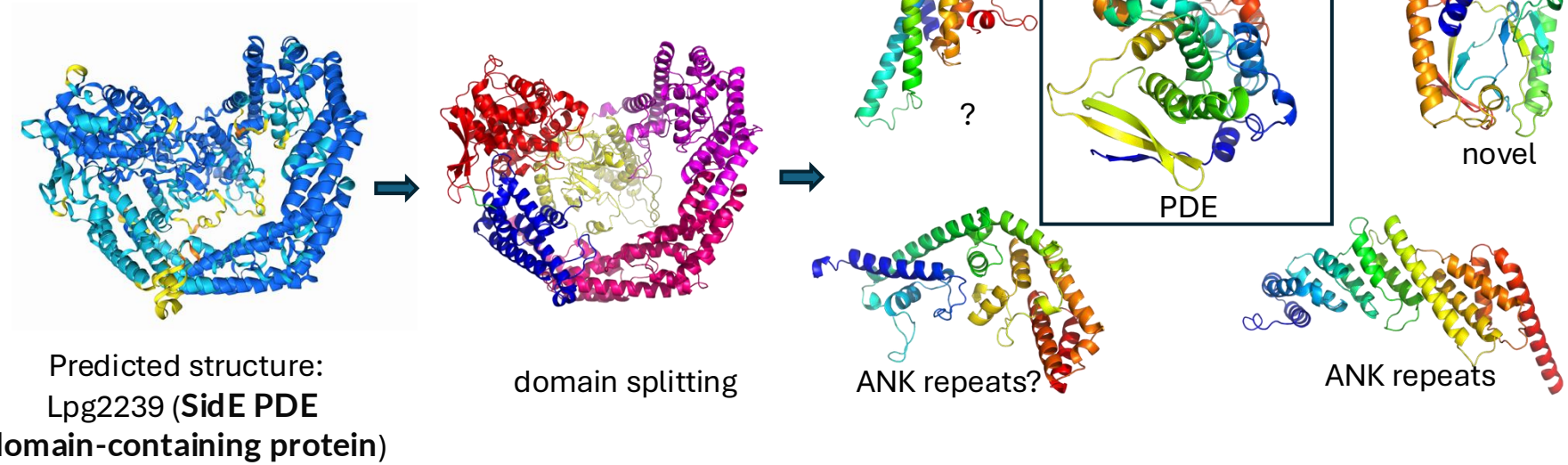


Another aspect of the same story

- TLR2 and other TLRs
  - High level of polymorphism in humans around the binding sites
    - Not in the residues lining the cavity, but in the loop binding the peptide head of the lipoprotein
    - In TLR2, but also in other TLRs.
  - Signs of subpopulations adapting to their specific pathogens?
    - Observed differences in pathogen clearing efficiency depending on the genetic background



Real life example of the novel structure analysis



Predicted structure:
Lpg2239 (**SidE PDE domain-containing protein**)

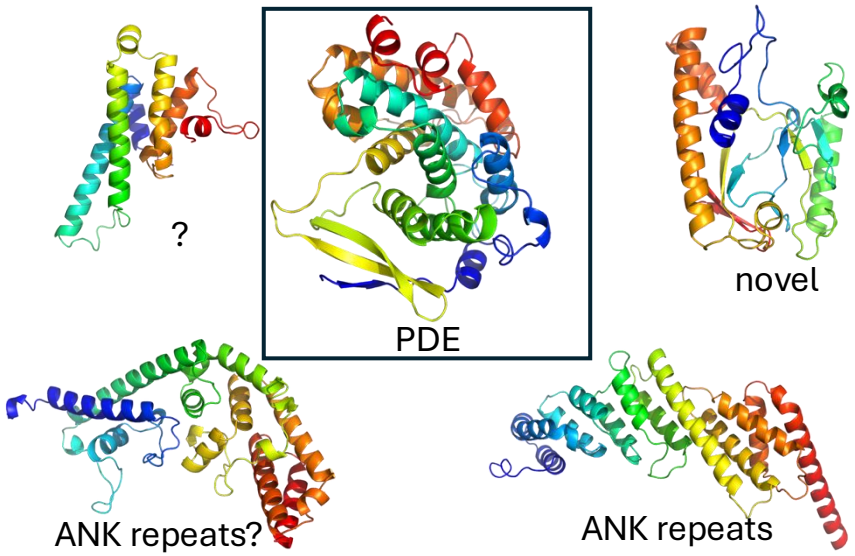
domain splitting

UniProt, InterPro, CDD

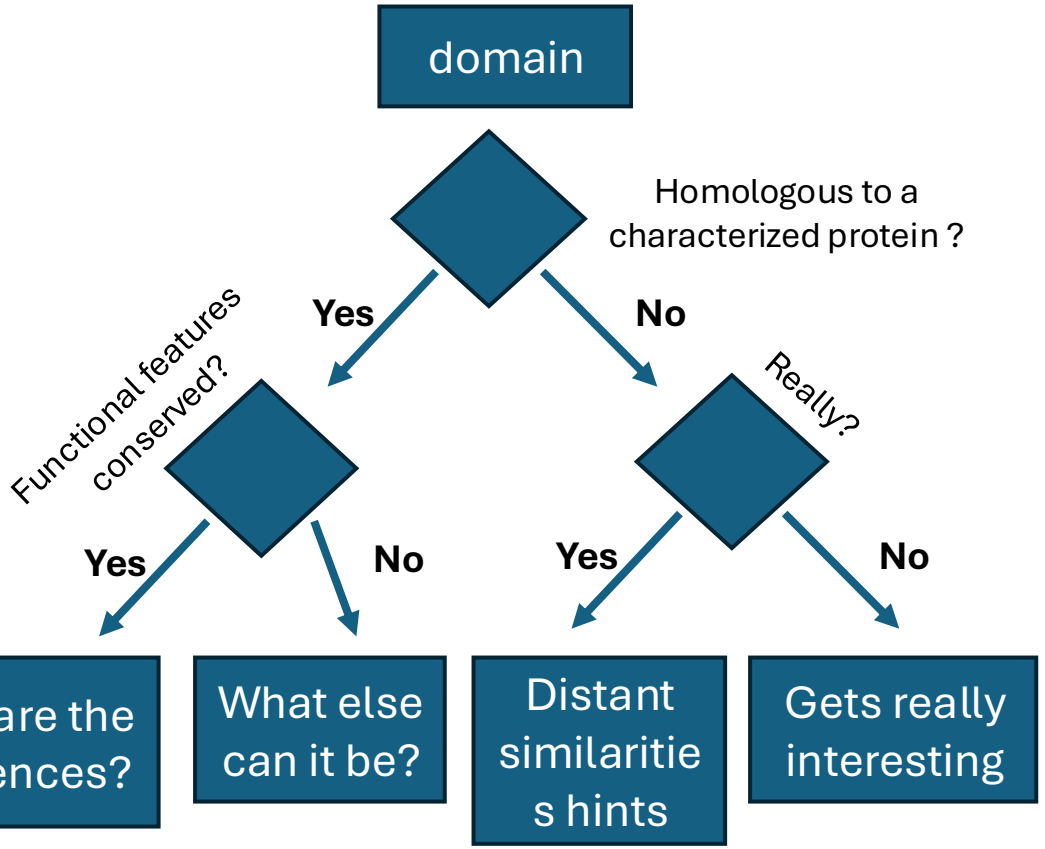
After AF model structure analysis

Patel, Deepak T., et al. "Global atlas of predicted functional domains in *Legionella pneumophila* Dot/Icm translocated effectors." *Molecular systems biology* 21.1 (2025): 59-89.

(predicted) structure analysis workflow



After AF model structure analysis



Before the hands-on session

- if you haven’t done so already, register at https://alphafoldserver.com
- Open free account on https://neurosnap.ai, https://www.tamarind.bio. Use their free credits to run other fold prediction programs. If you know about any other platform like these, use it (and let me know)

In the practical session

- Pick up a protein you are interested in
- Try different algorithms to predict its structure and compare the models
  - Easiest – AF2 from AlphaFoldDB or Uniprot to AF3 from the server to ESM from the ESM atlas
- Use Chimera or Pymol to look and visually compare different models
- Use Dali or FATCAT to quantify differences between the models
- Use FoldSeek to find homologs of your favorite protein in a selected genome
- Use FoldMason for multiple structure alignment

Acknowledgments

- Center for Structural Biology of Infectious Diseases
- University of California, Riverside School of Medicine, Biosciences Division, computational biology group

This project is supported in part by the NIH NIAID contract #75N93022C00035