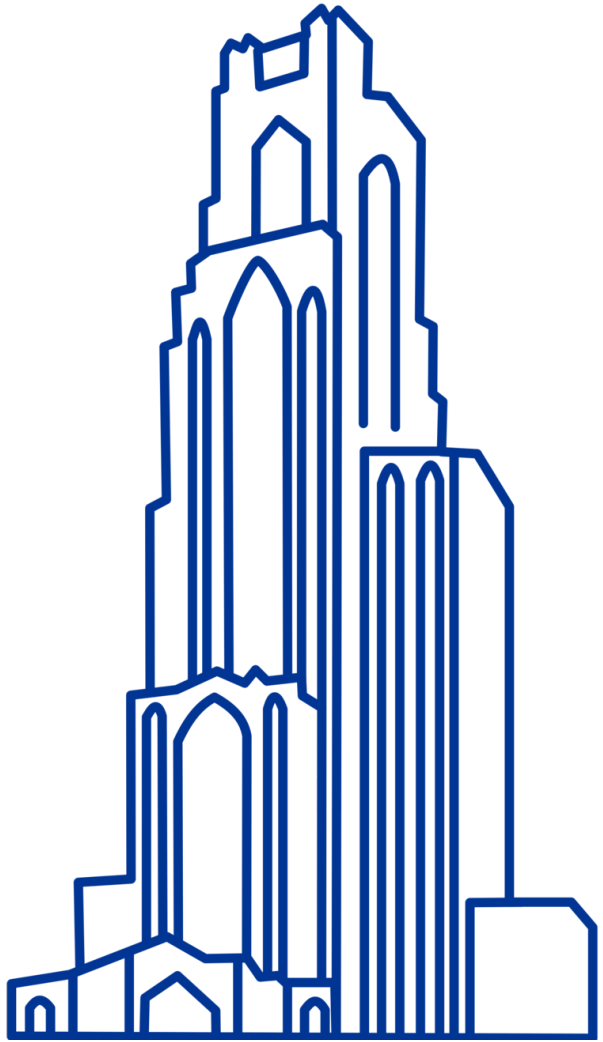


Computational Biology

(BIOSC 1540)

Lecture 11A

Protein structure prediction

Foundations

Mar 25, 2025

University of Pittsburgh

Announcements

Assignments

- P02B is due Mar 28
- P02C is due Mar 28
- P03A is due Apr 4

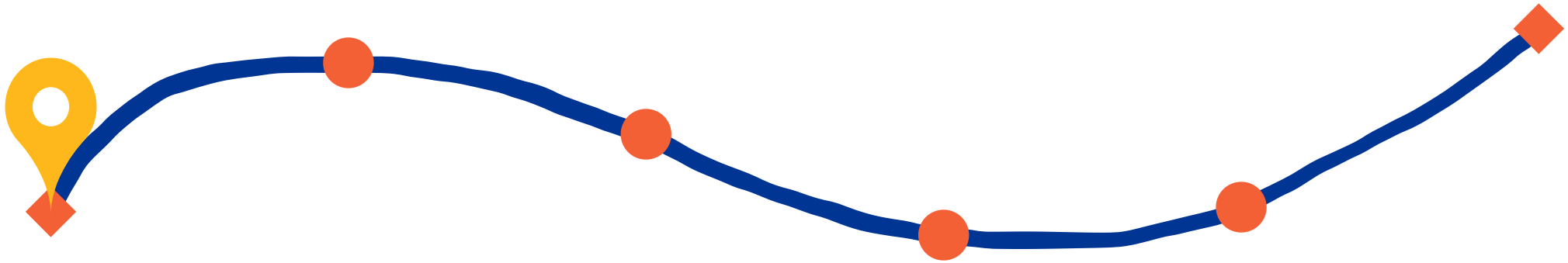
Quizzes

- Quiz 04 will be on Apr 8 and cover L09A to L12A

Final exam

- The final exam is on **Monday, Apr 28, at 4:00 pm in 244 Cathedral of Learning**

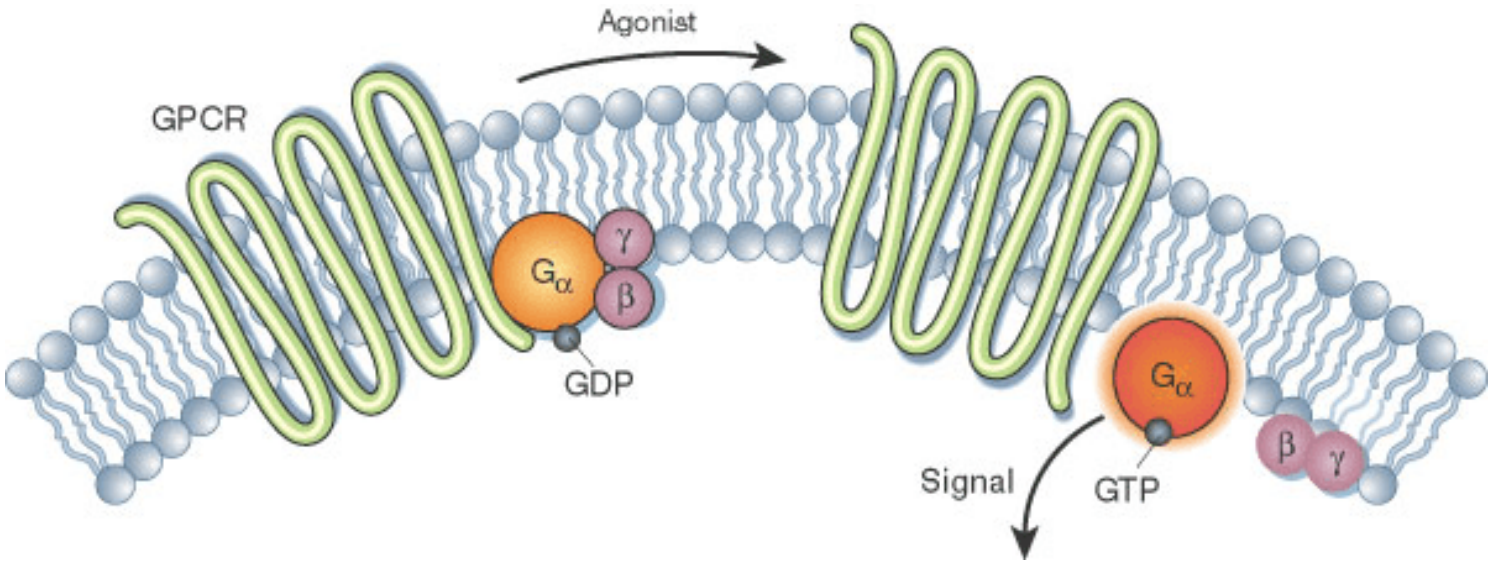
After today, you should have a better understanding of



Identify what makes structure prediction challenging

Protein structure is essential for understanding biological function

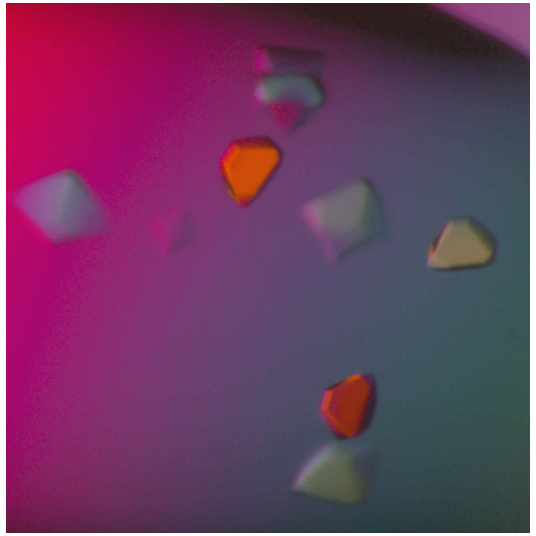
Proteins are molecular machines; their 3D shape determines how they interact with substrates, DNA, other proteins, etc.

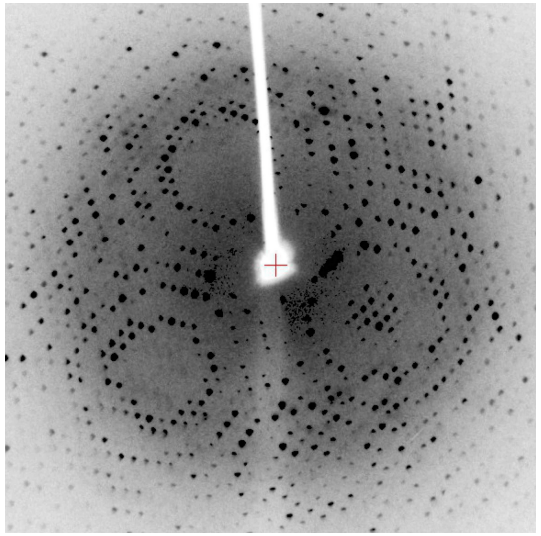


Experimental methods for structure determination are powerful but limited

X-ray crystallography, NMR, and cryo-EM provide high-resolution data.

However, these methods are time-consuming, expensive, and often fail for specific proteins.





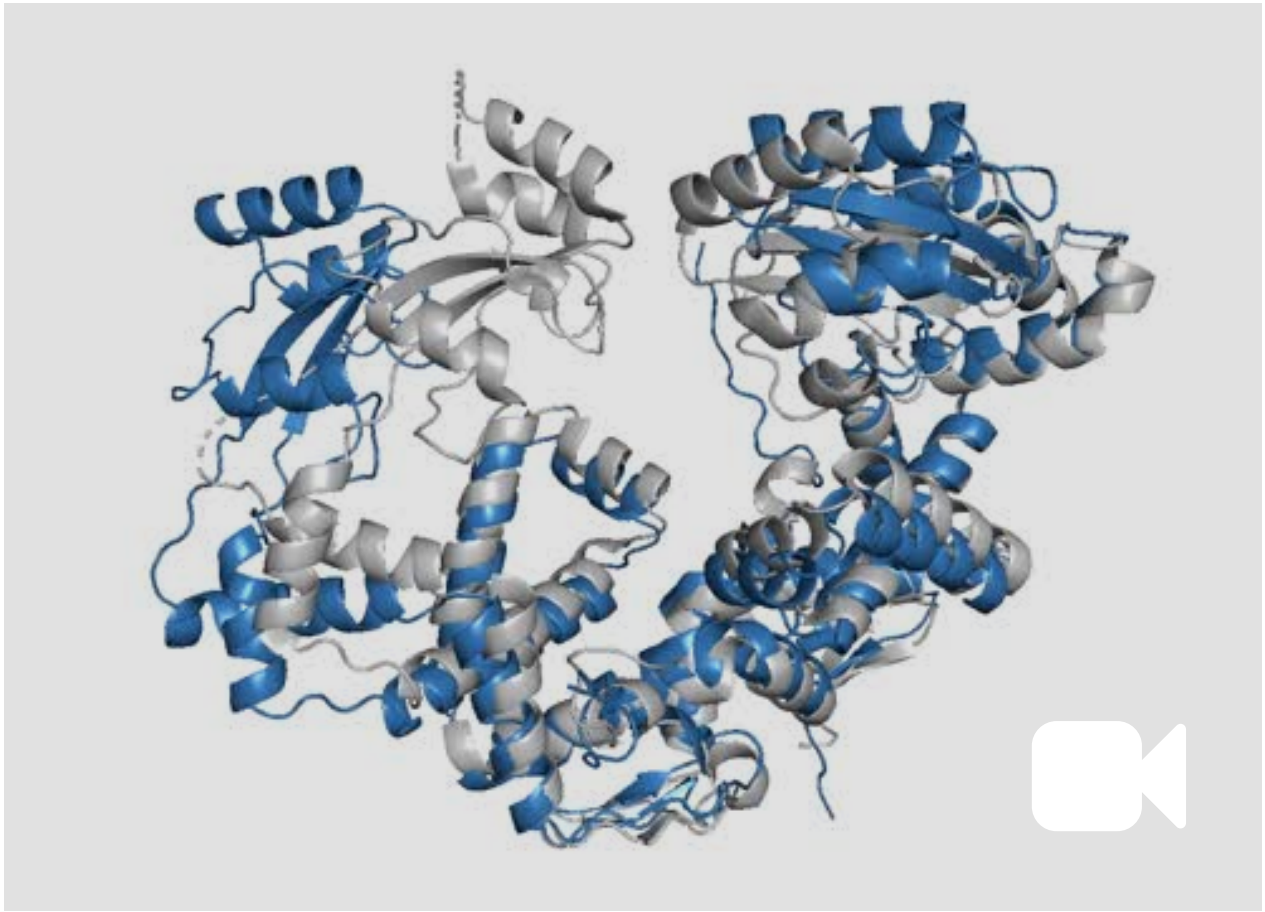
As of 2021, >200 million protein sequences exist, but <200,000 structures are known

Protein structure prediction fills a crucial gap in biological discovery

Enables structural understanding of sequences with no experimental structure. This accelerates many research fields and democratizes access to atomistic insights

PrsA1 (X-ray)

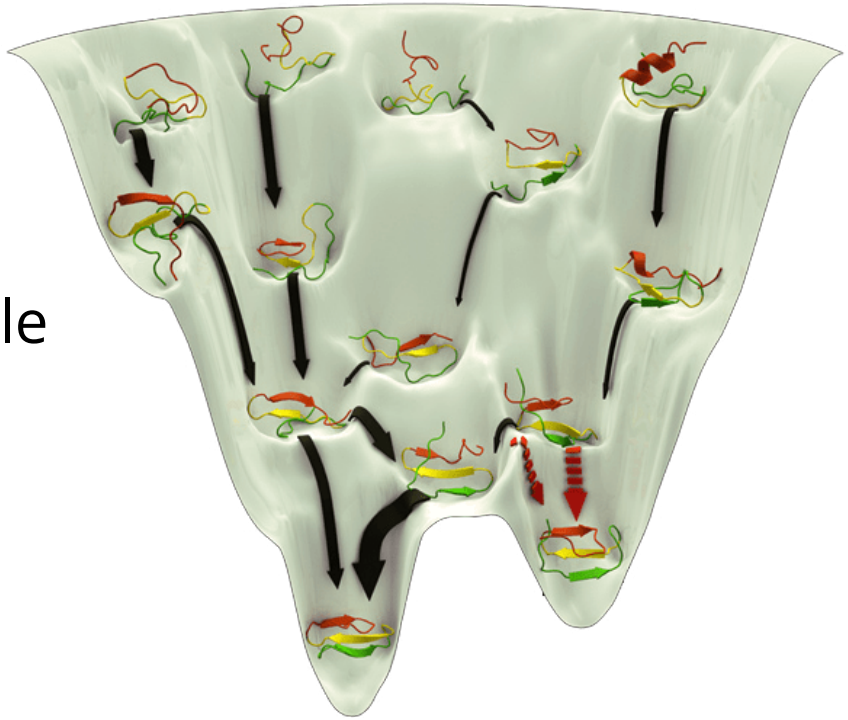
PrsA2 (AF3)



Example: Our collaborators (Dr. Cahoon) crystallized *Lm* PrsA1 in 2016, but we need a structural model of PrsA2. Instead of potentially years, AlphaFold 3 gives us a decent prediction within minutes

Protein folding is computationally hard due to the vast conformational space

Levinthal's Paradox: A protein can't sample all conformations in a biologically reasonable time, yet it folds quickly



Example: A protein with 100 amino acids, each capable of adopting about 3 torsion angles, results in ~3^{100} possible conformations

Proteins fold in milliseconds—implying nature doesn't sample all conformations.

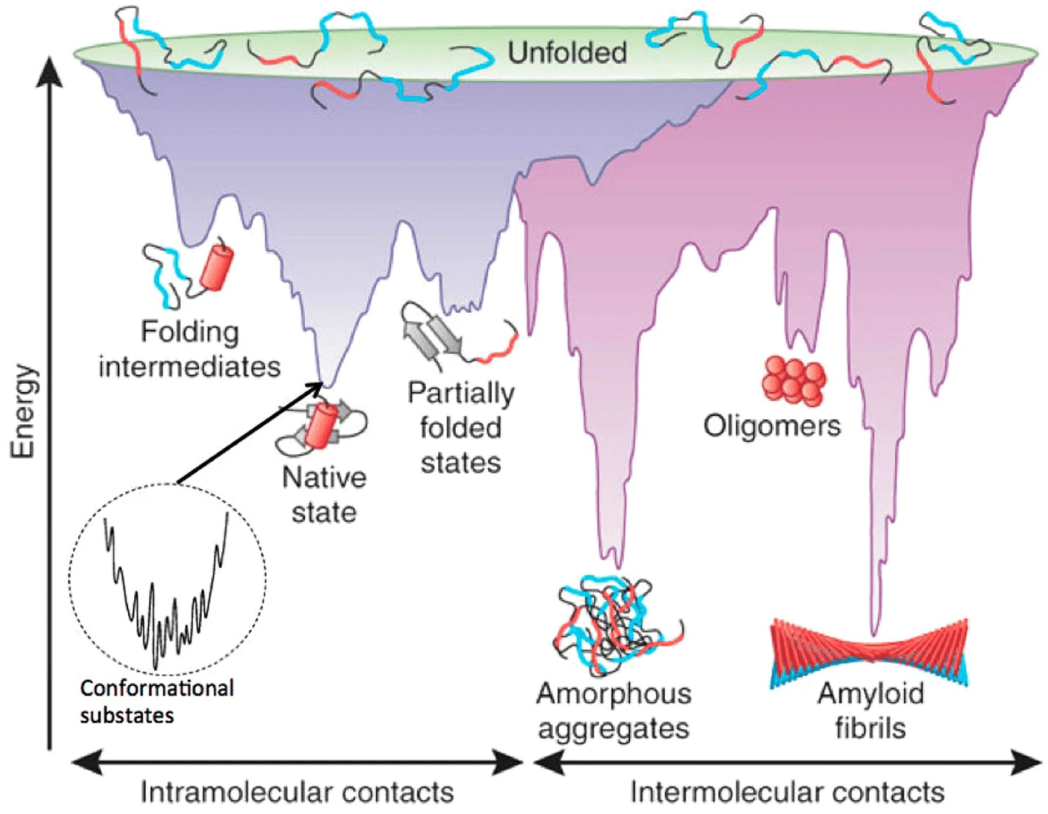
Proteins fold into their native structure by minimizing free energy

A **potential energy surface** (PES) represents the energy of a system as a function of the positions of its atoms

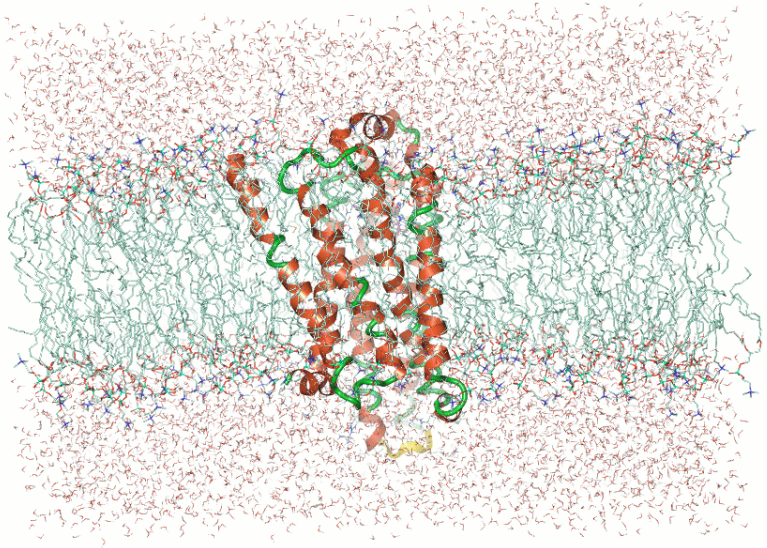
Allows us to understand how the system's energy changes upon reactions or movements

Proteins adopt conformations that minimize thermodynamic free energy

Scoring functions attempt to model this with statistical or physics-based potentials

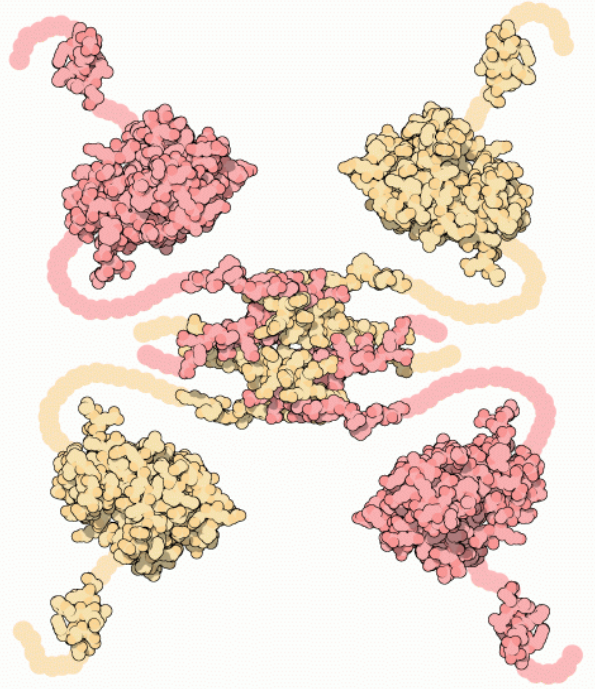


Flexible and disordered regions add complexity to structure prediction



Many proteins exist in ensembles of structures or are natively disordered.

Their function may depend on transient interactions or induced folding.



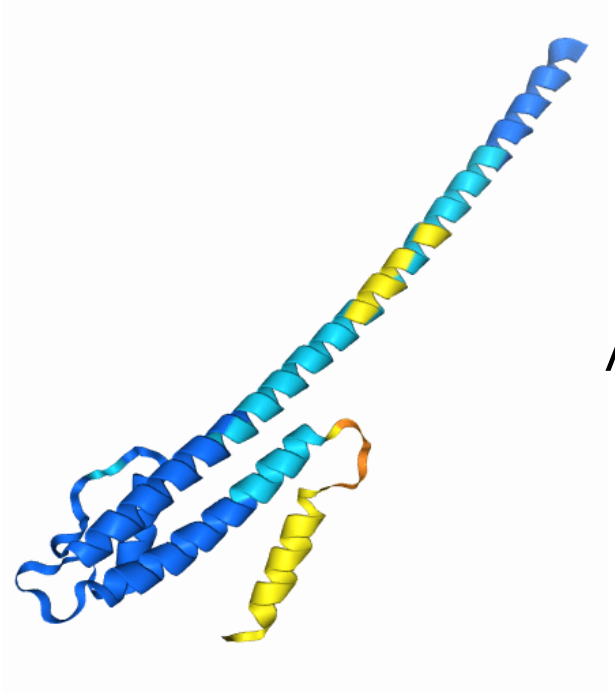
Environmental context dramatically impacts protein folding

Proteins fold differently in different environments

Predictions need to capture interactions with solvent molecules, ions, and cofactors

7MHX

pH-gated K+ channel



AlphaFold 3

Example: Predicting transmembrane protein structures, where the lipid bilayer plays a key role in folding, is particularly complex.

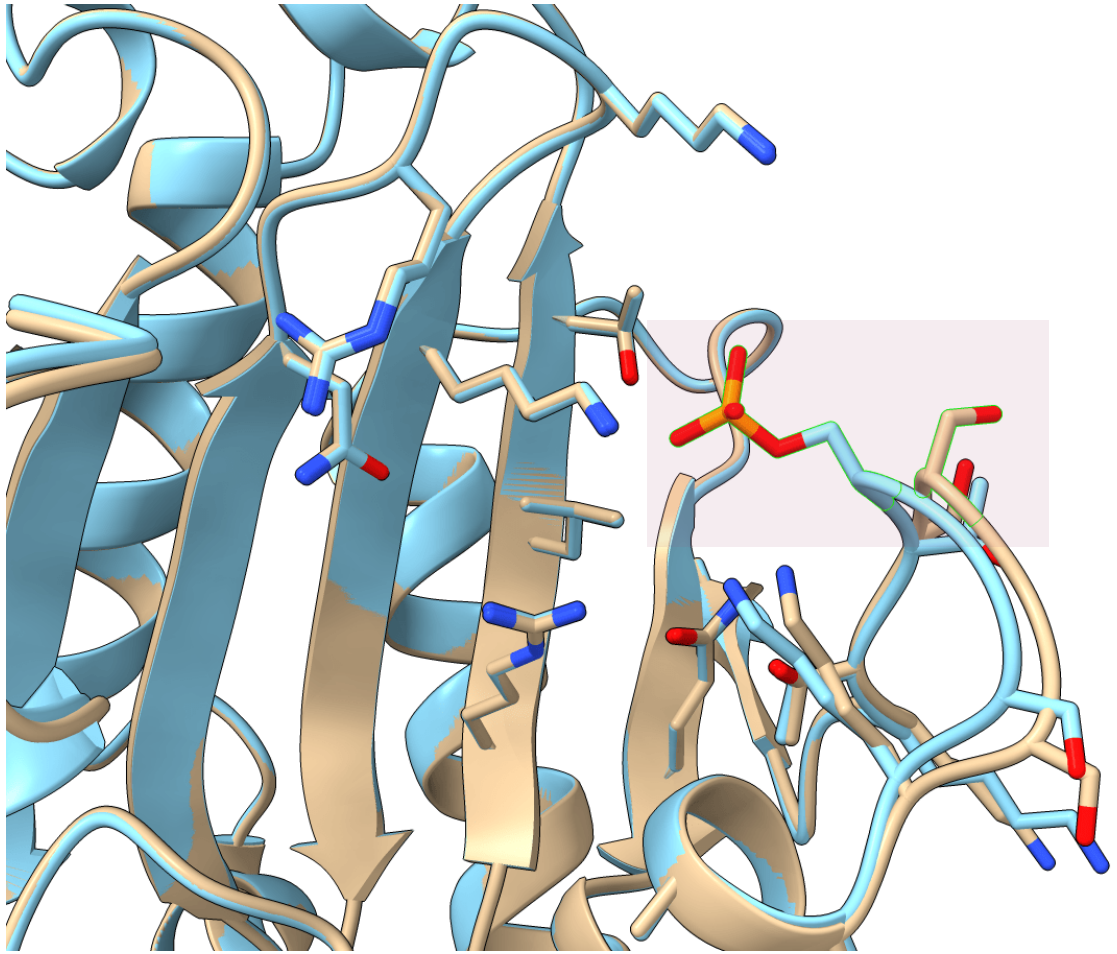
Environmental context dramatically impacts protein folding

PTMs such as phosphorylation, glycosylation, and methylation can alter protein folding and function

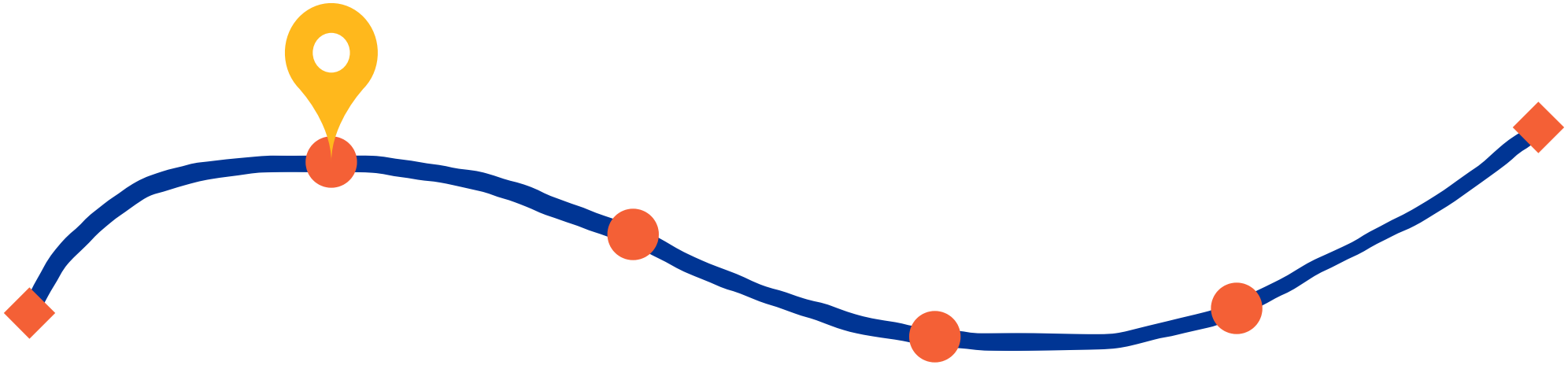
Example: eIF4E is a eukaryotic translation initiation factor involved in directing ribosomes to the cap structure of mRNAs

Ser209 is phosphorylated by MNK1

AlphaFold 3 accurately predicts these changes when they are already known



After today, you should have a better understanding of



Homology modeling

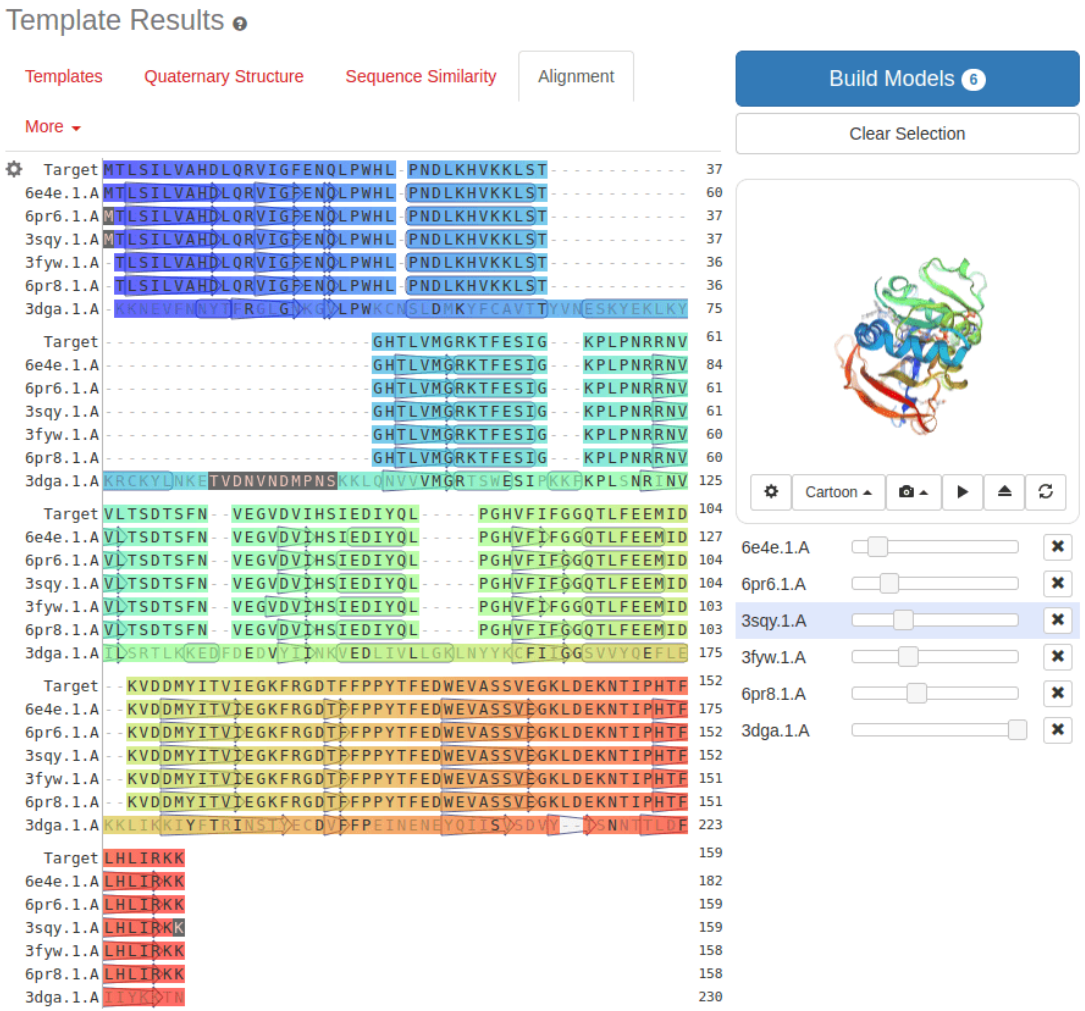
Hidden Markov Model alignments

Homology modeling predicts structure using evolutionary relationships

Based on the principle that **proteins with similar sequences tend to adopt similar structures**.

Often the first modeling strategy attempted due to simplicity and reliability.

Requires a *template*—a known structure with detectable sequence similarity.

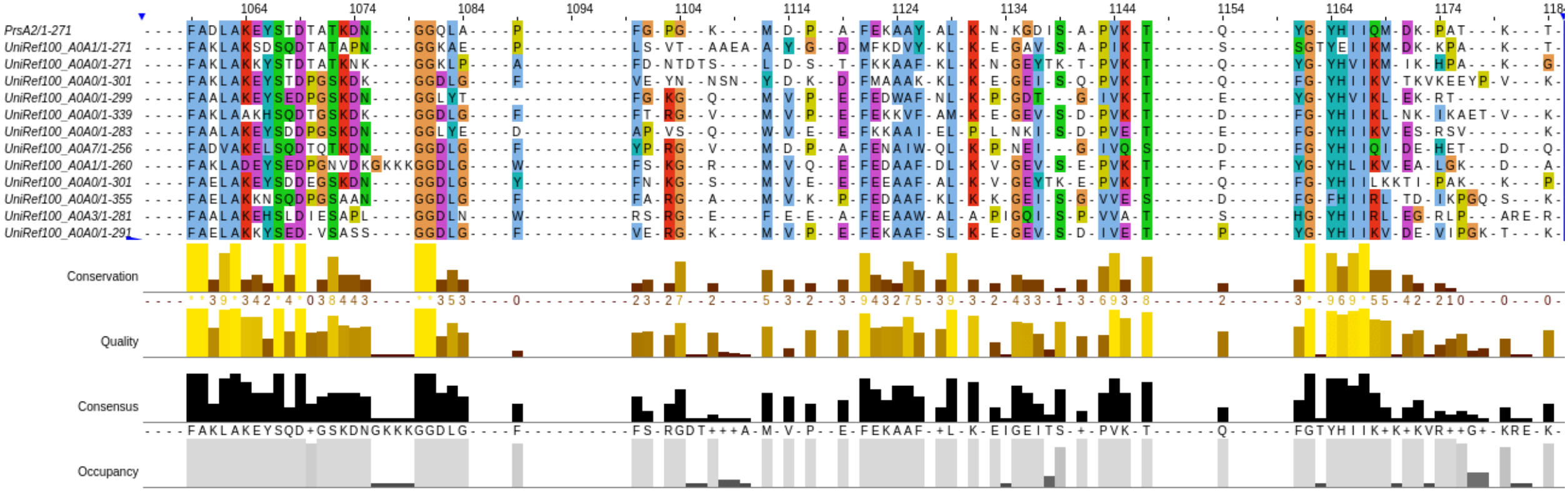


The first step in homology modeling is to search for similar sequences

You begin with a **query sequence**
(the protein you want to model)

```
> PrsA2
CGGGGDVVKTDSGDVTKDELYDAMKDKYGSEFVQQLTFEKILGDKYKVSDE
DVDKKFNEYKSQYGDQFSAVLTQSGLTEKSFKSQLKYNLLVQKATEANTDT
SDKTLKKYYETWQPDITVSHILVADENKAKEVEQKLKDGEKFADLAKEYST
DTATKDNGGQLAPFGPGKMDPAFEKAAYALKNKGDISAPVKTQYGYHIIQM
DKPATKTTFEKDKKAVKASYLESQLTTENMQKTLKKEYKDANVKVEDKDLK
DAFKDFDGSSSSDSDSSK
```

Sequence alignment maps residues from the target onto the template structure

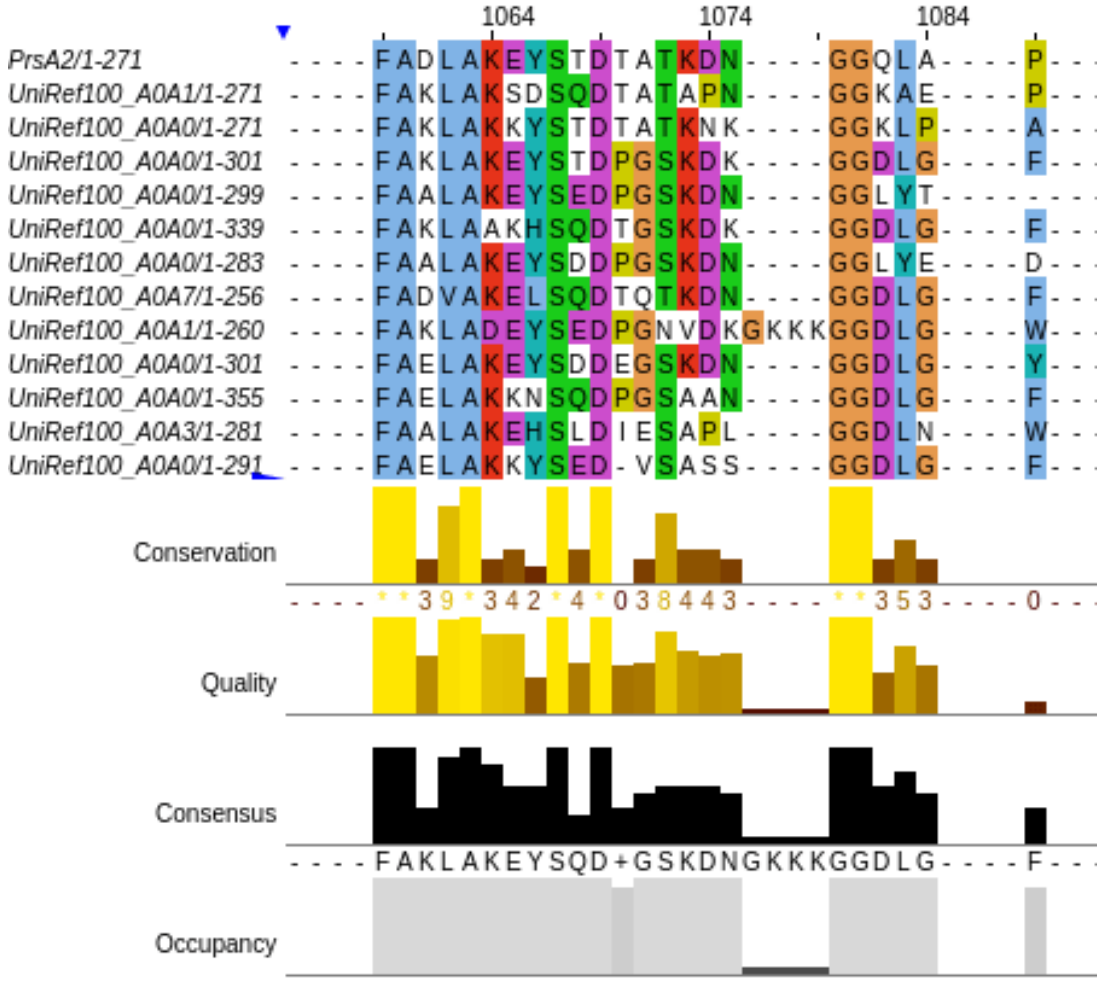


This is an MSA of *Listeria monocytogenes* PrsA2 to related proteins

Basic alignment algorithms are too simplistic for distant homolog detection

Methods like **Smith-Waterman** use direct pairwise alignment based on similarity scores (e.g., BLOSUM62).

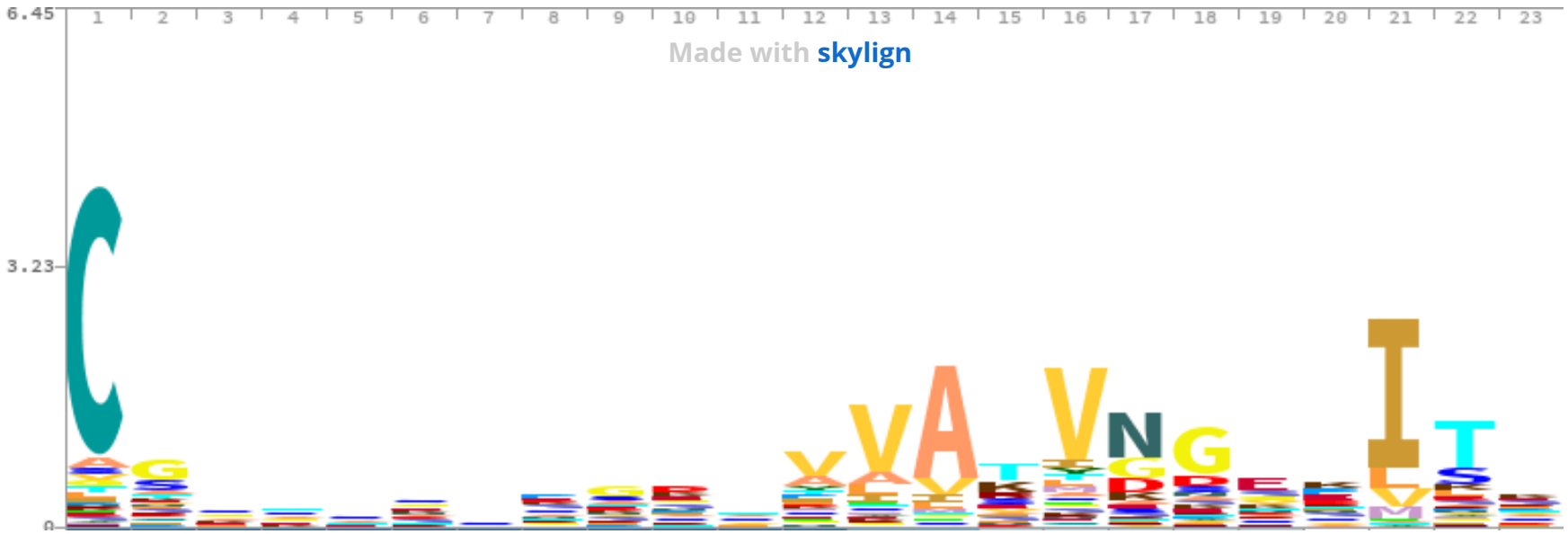
They do not consider evolutionary variation, insertions, or residue-level probabilities.



We need methods that detect **evolutionarily distant but structurally conserved** relationships.

Profile-based aligners improve sensitivity by modeling residue variability at each position

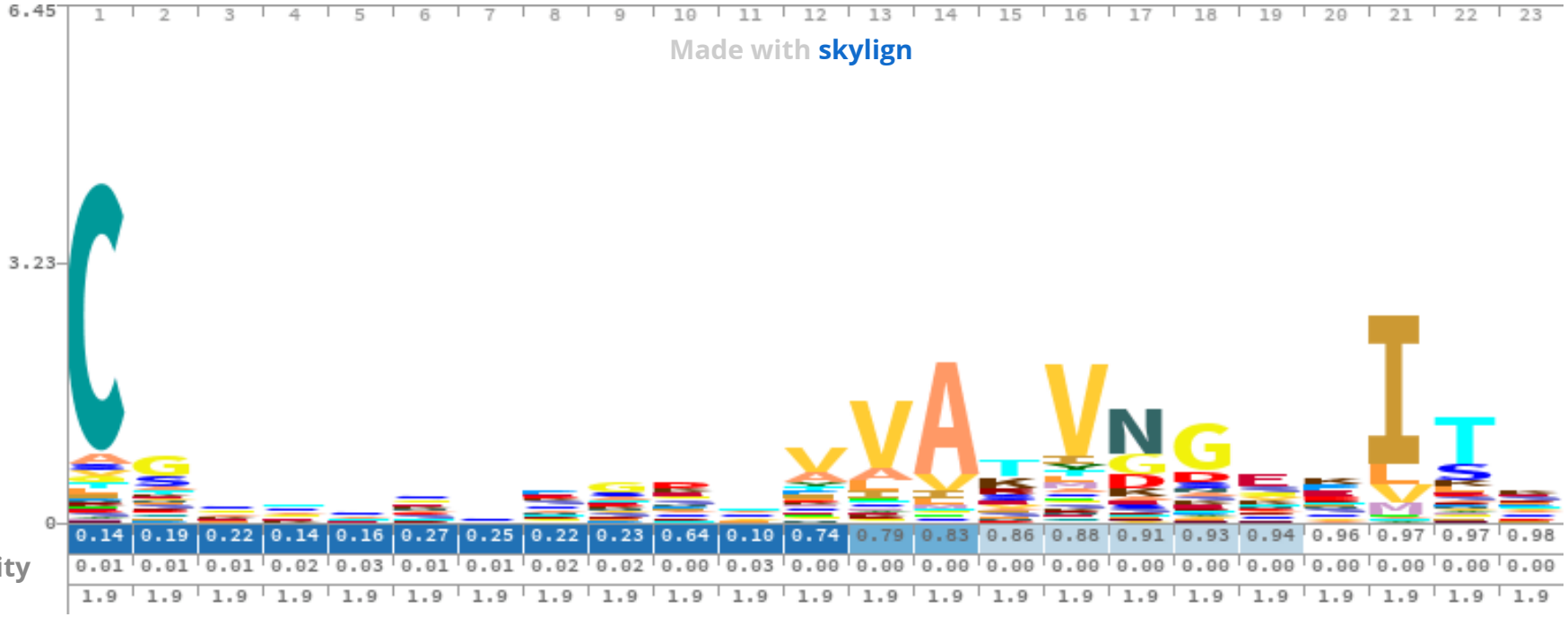
A **profile** captures how conserved each position is across an MSA



Instead of a single residue, each position becomes a **probability distribution** over all 20 amino acids.

Profile-based aligners improve sensitivity by modeling residue variability at each position

A **profile** captures how conserved each position is across an MSA



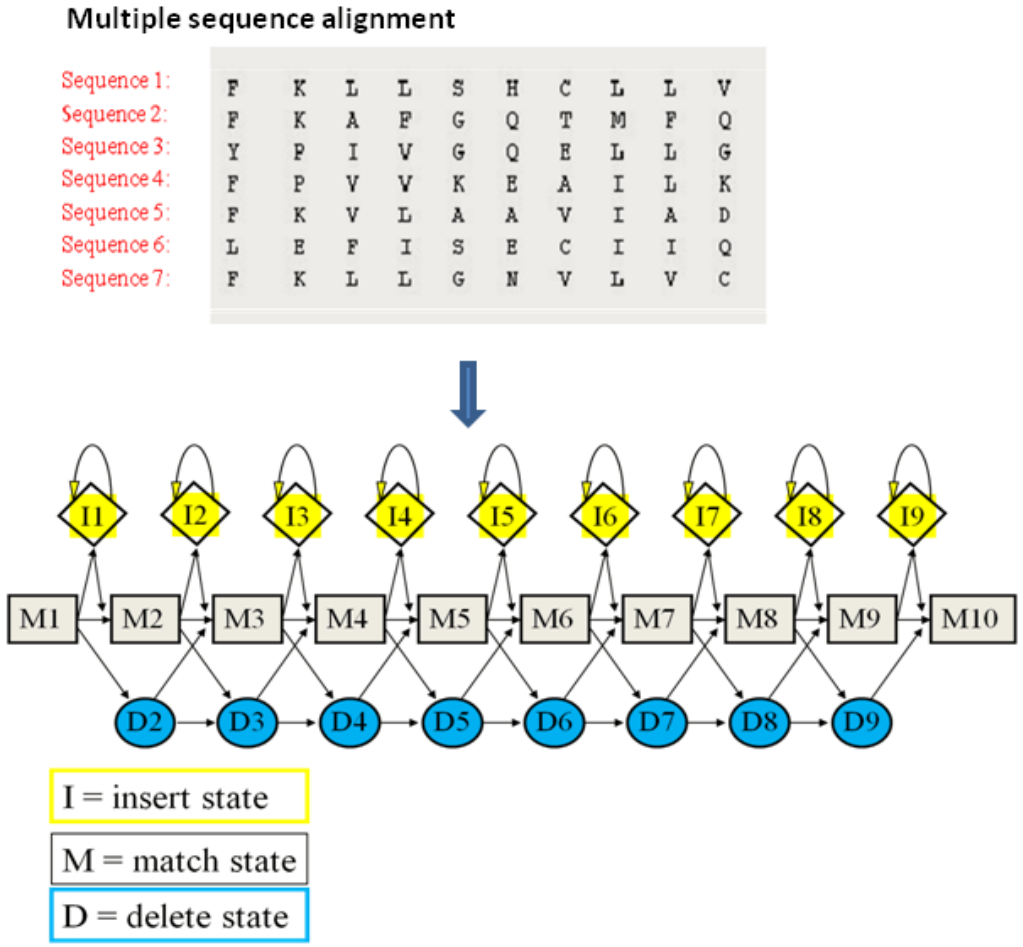
Instead of a single residue, each position becomes a **probability distribution** over all 20 amino acids.

HHblits starts by converting the query sequence or MSA into a profile HMM

A **profile HMM** models the amino acid probabilities at each position, plus insertion and deletion likelihoods.

Sequence likelihood can be computed by walking along the profile HMM

The result is a probabilistic model that captures both **conservation** and **structural variability**.



HHblits performs accurate HMM–HMM alignments on the best candidates

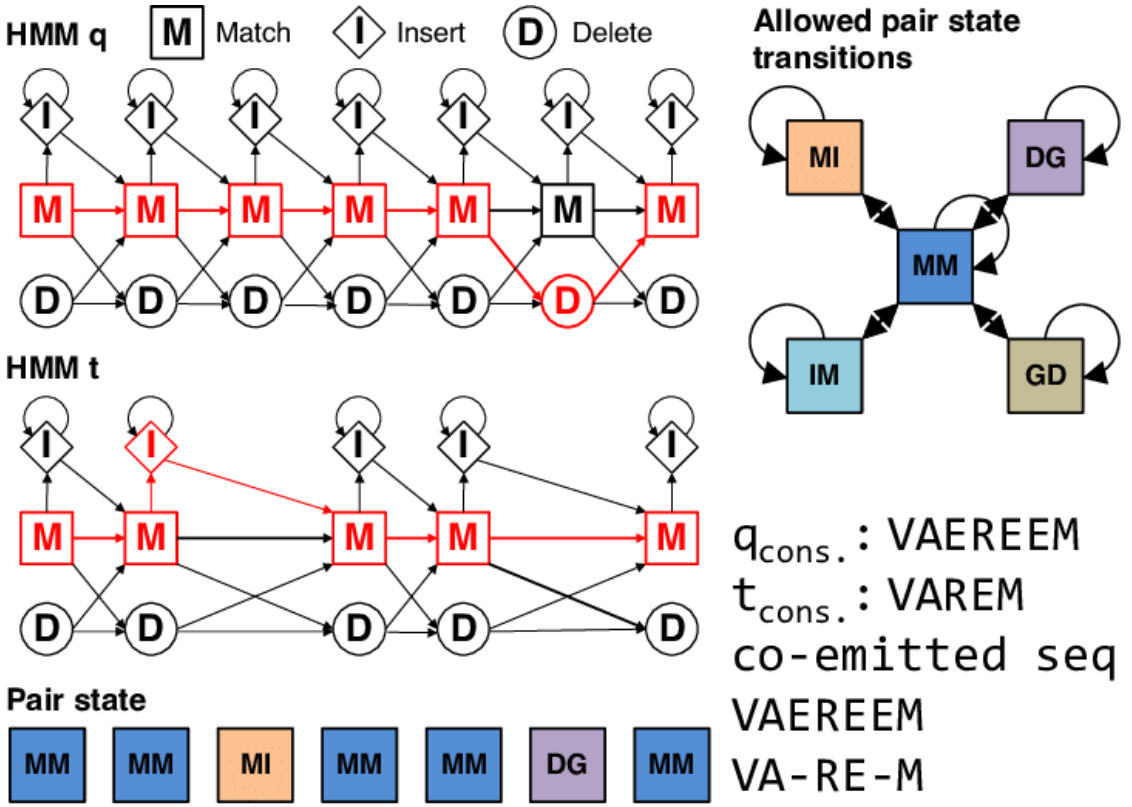
Full alignments are done using the **Viterbi algorithm** to find the best path through the HMM state space.

A **maximum accuracy (MAC)** alignment is also computed to optimize for correct residue–residue matches.

These alignments return **E-values** to estimate match confidence.

Only statistically significant hits (e.g., E < 1e-3) are retained for the next iteration.

The alignment is represented as red path through both HMMs

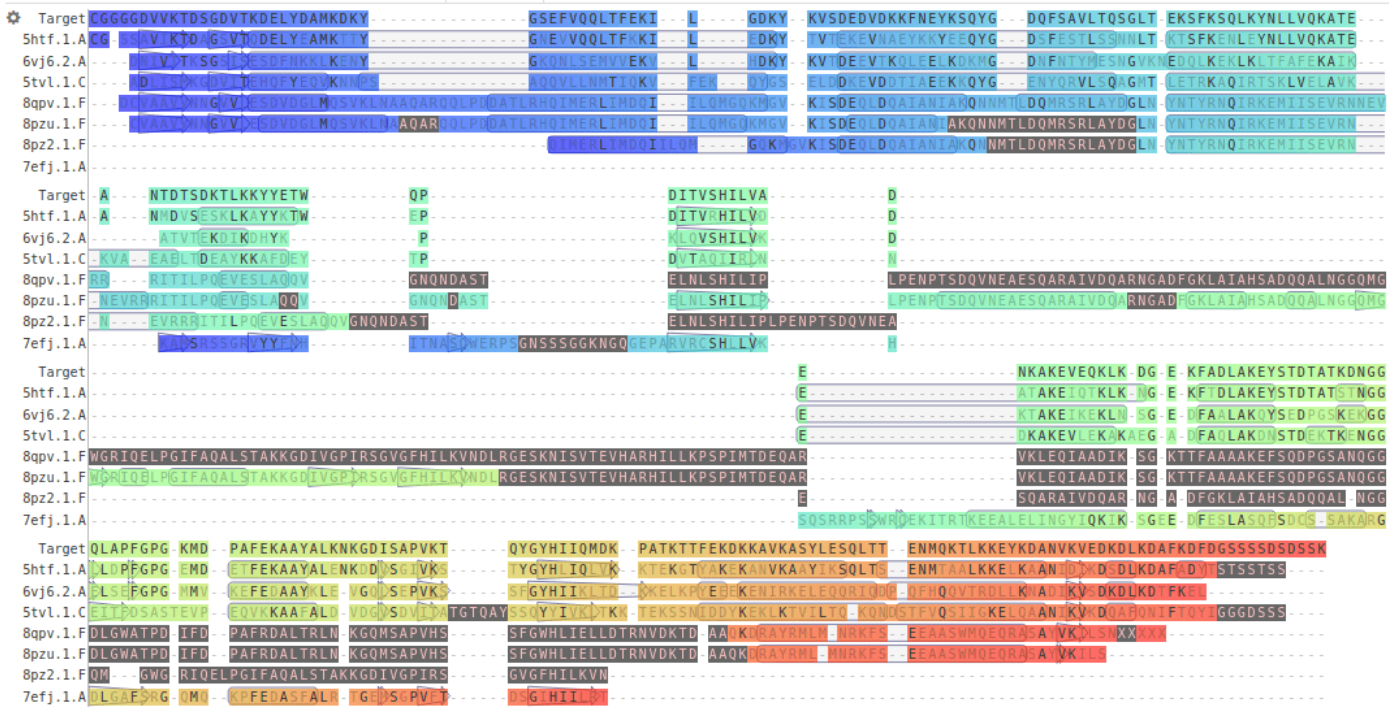


Homology modeling is most accurate when sequence identity is >30%

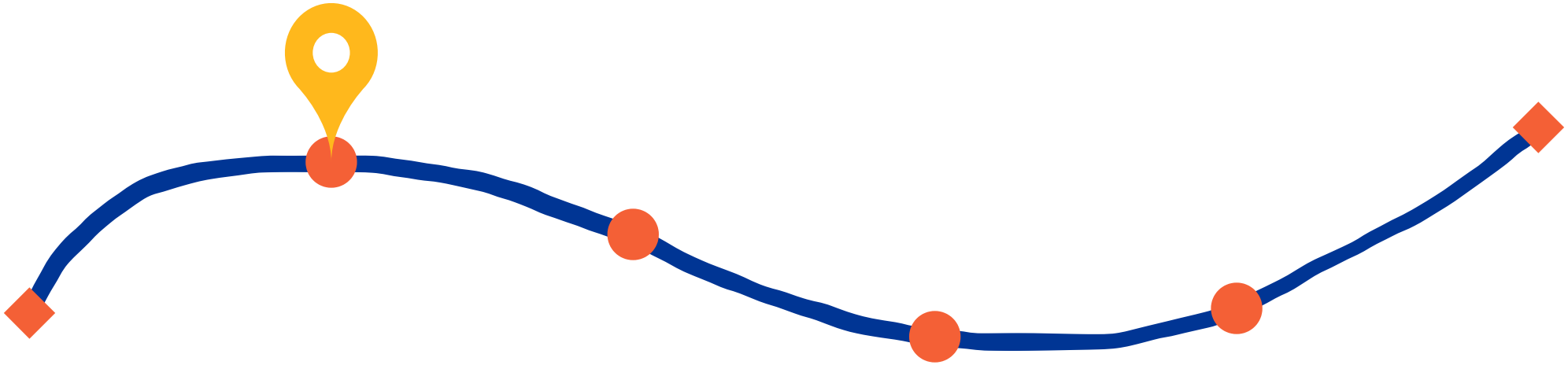
>50% identity: high-accuracy models (~1 Å RMSD) are achievable

Between 30–50%: moderate accuracy; errors appear in loops, side chains

<30% identity: The "twilight zone" where structural similarity is uncertain



After today, you should have a better understanding of



Homology modeling

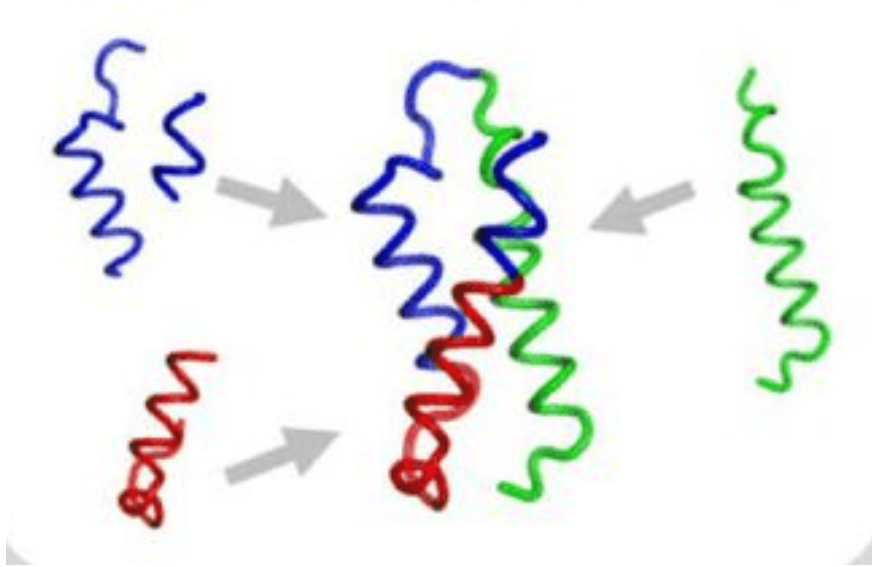
Template building

The model is built by copying the template structure and modeling variable regions

Conserved regions: Backbone atoms are copied from template directly.

Variable regions (loops, inserts): Built using fragment libraries or loop modeling algorithms.

Side chains are adjusted using rotamer libraries to fit target sequence.

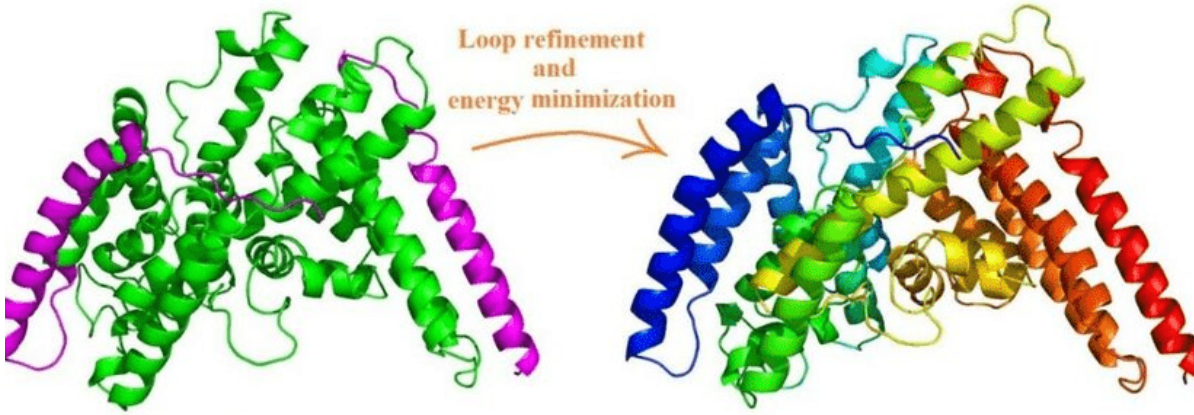


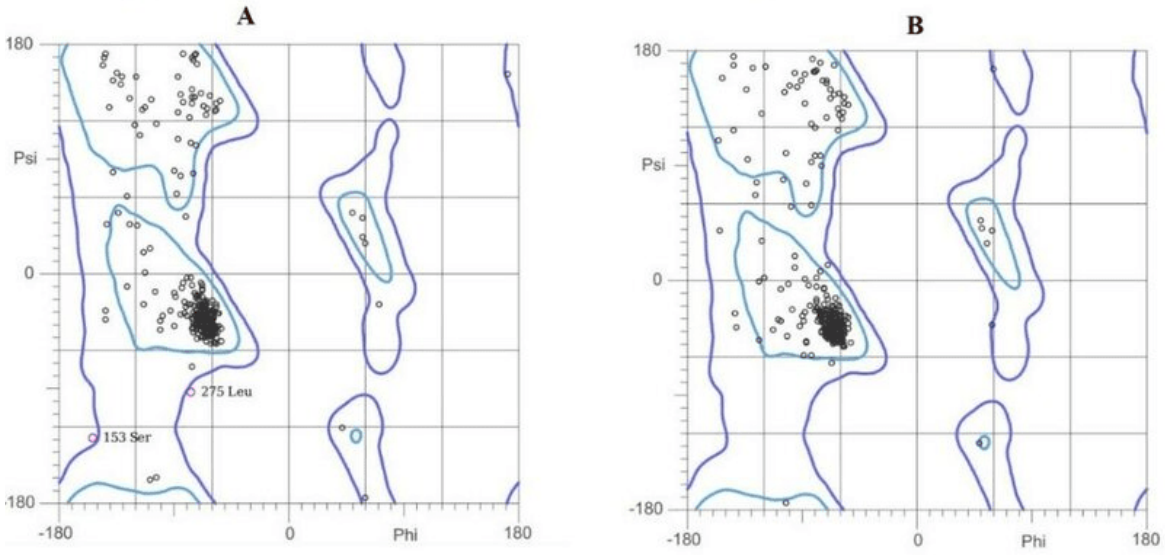
Model refinement improves geometry and resolves steric clashes

After model construction, the structure often contains **bad bond angles**, **clashes**, or **unrealistic torsions**.

Refinement includes **energy minimization** using force fields or statistical potentials.

Some tools use **molecular dynamics** or **Monte Carlo sampling**.





Model validation helps assess confidence and detect errors

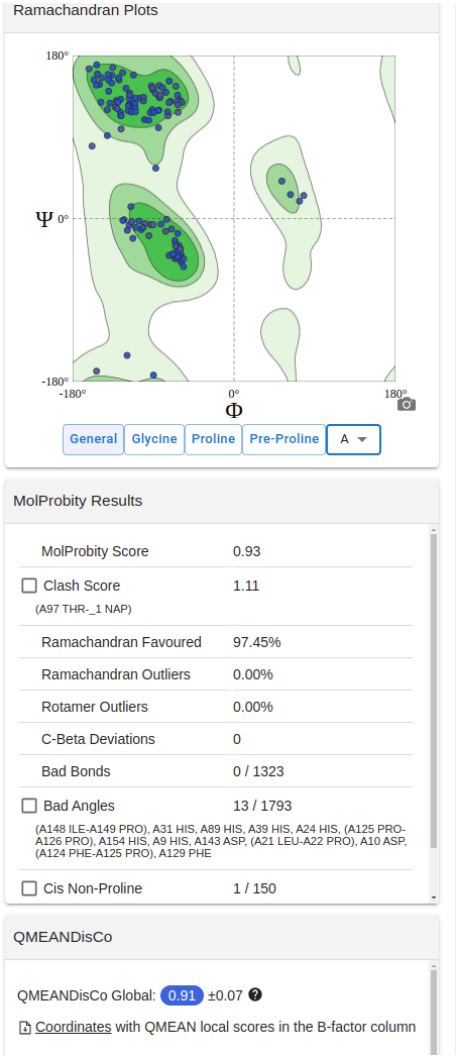
Ramachandran plots visualize backbone torsion angles.

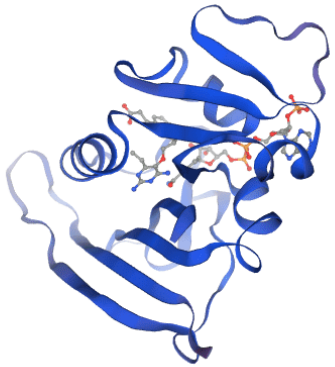
Statistical scores (e.g., DOPE, QMEAN, GA341) evaluate nativeness.

Residue-by-residue assessment helps identify weak regions (e.g., VERIFY3D, ERRAT).

Good models have:

- Most residues in favored Ramachandran regions
- Low-energy scores
- No large clashes





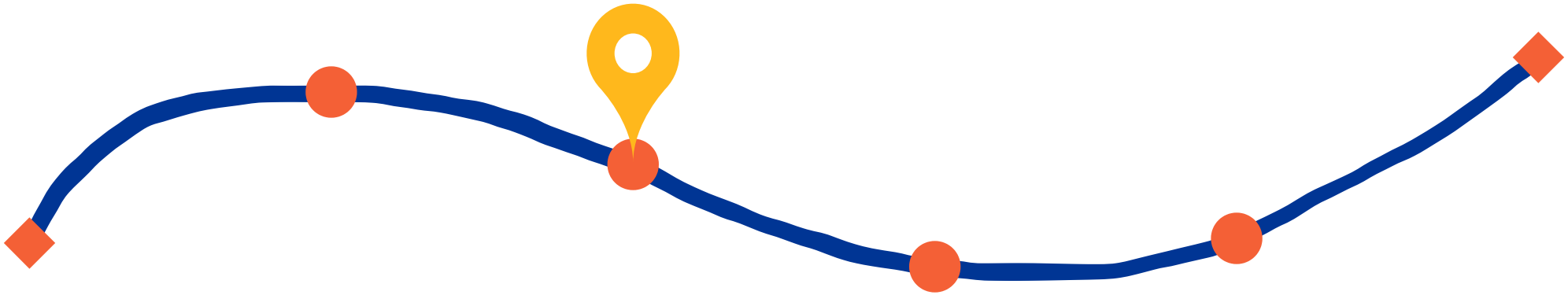
Homology modeling works best when you iterate and re-evaluate

If a model fails validation, revisit earlier steps:

- Try a different template
- Refine the alignment
- Adjust loop modeling parameters

Multiple models are often built and ranked—choose the one with the best validation metrics.

After today, you should have a better understanding of



Know when to use threading

instead of homology modeling

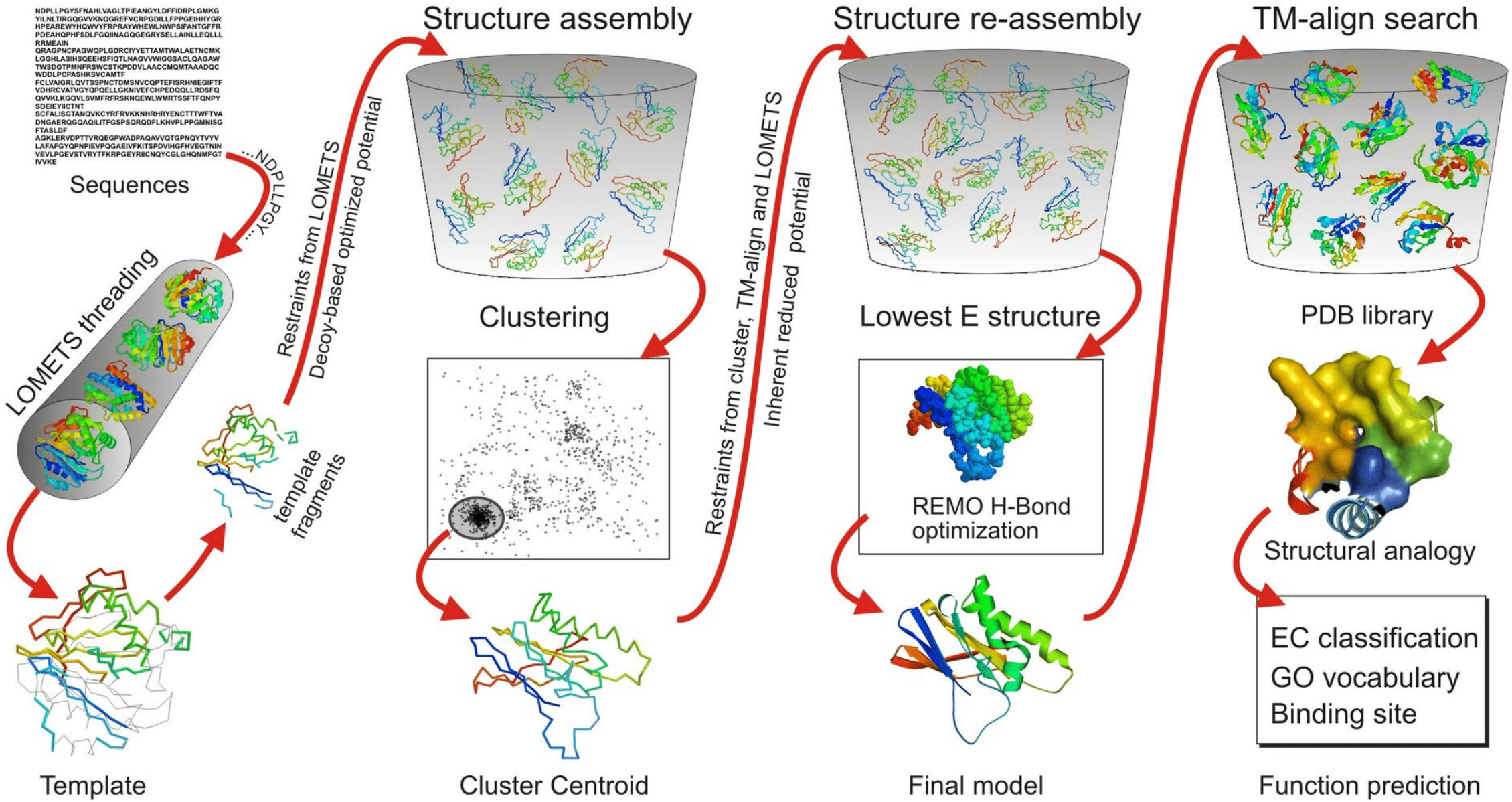
Why Use Threading?

In cases where sequence similarity to known structures is low (< 30%), homology modeling becomes unreliable

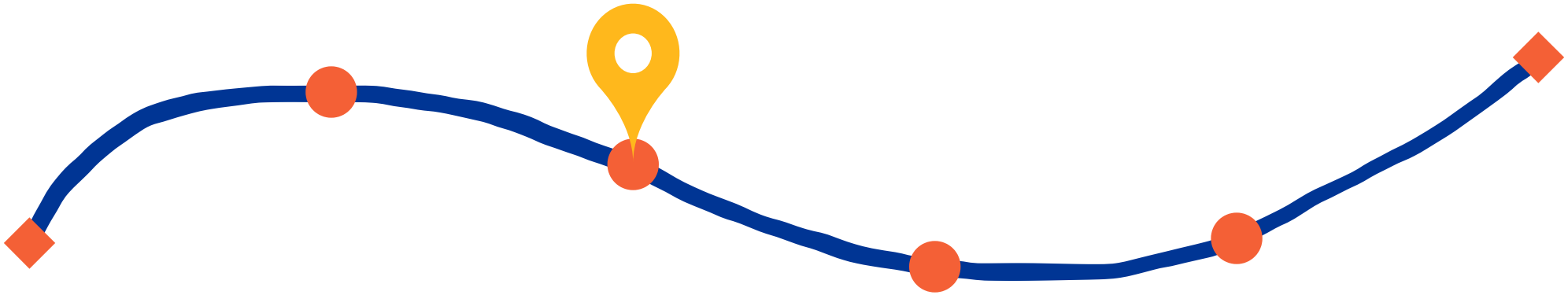
Threading **matches sequences to known structural folds** based on structural rather than sequence similarity

Phyre2, **RaptorX**, **MUSTER**, and **I-TASSER** are commonly used for threading and takes much longer than homology modeling

Identifying the Right Fold



After today, you should have a better understanding of



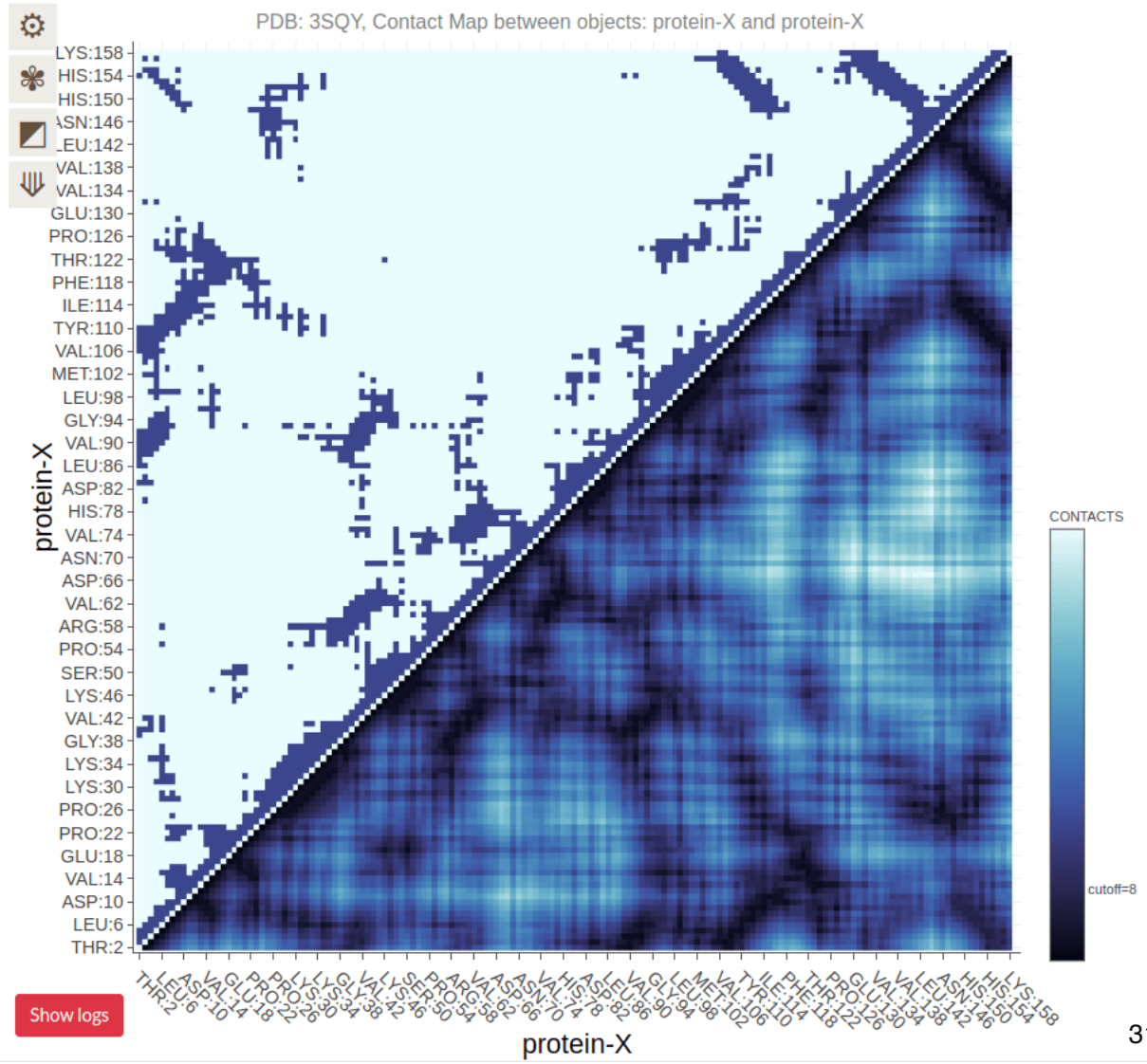
Interpret a contact map for protein structures

Contact Maps Visualize Residue Interactions in Proteins

A contact map is a 2D representation of which residues are in close proximity

Each point on the map corresponds to two residues that are close in 3D space

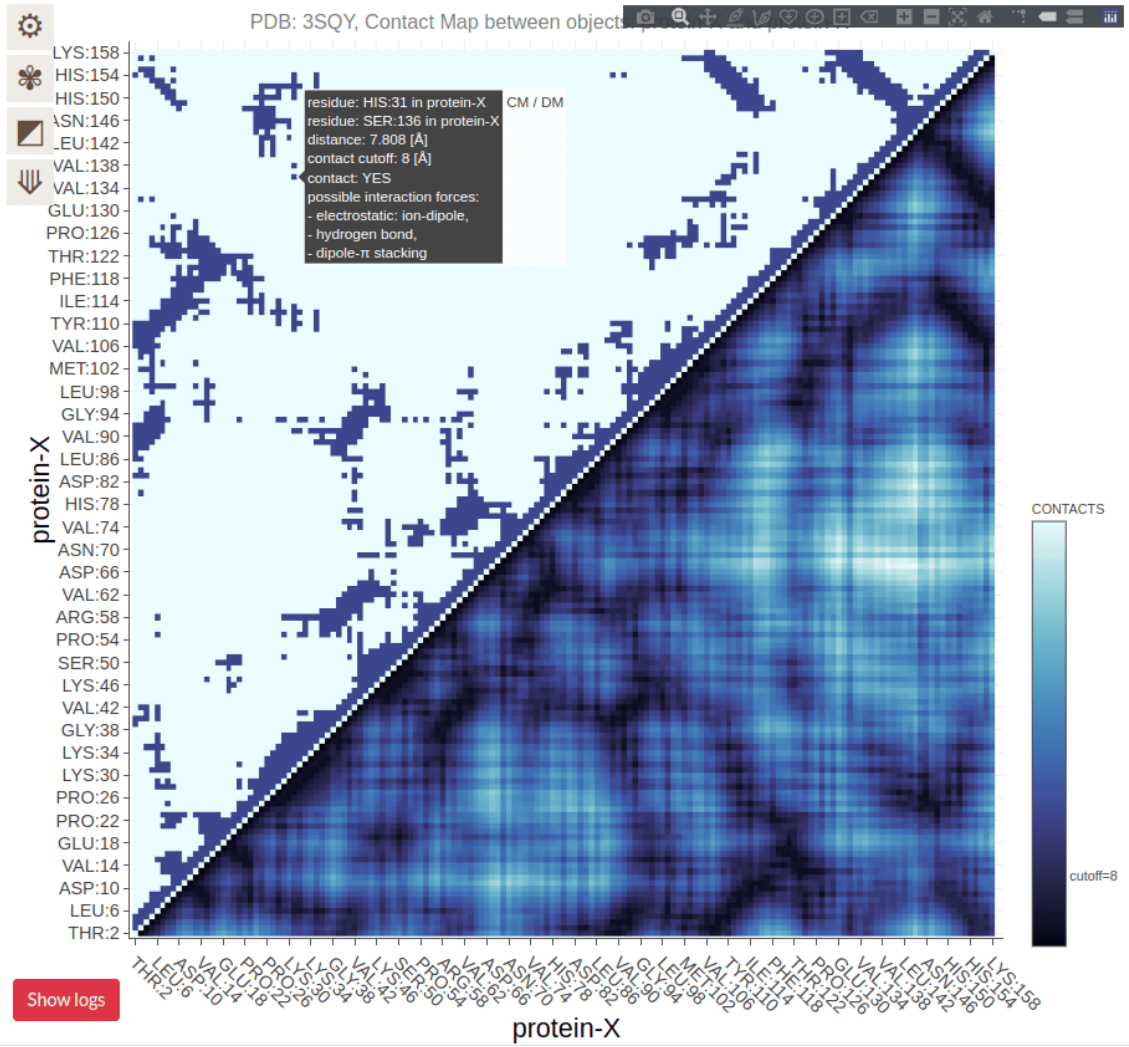
mapiya.lcbio.pl

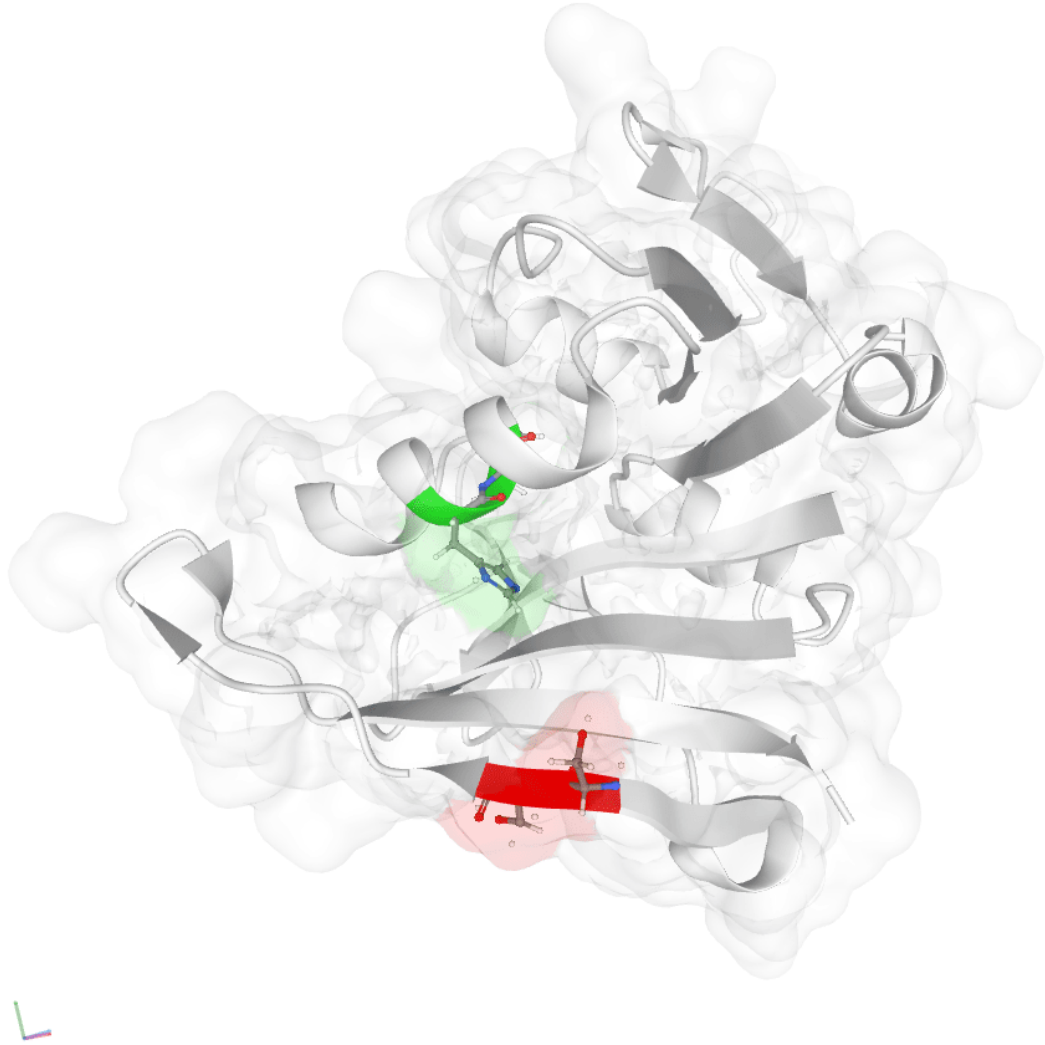


Contact Maps Represent Spatial Proximity, Not Sequence Order

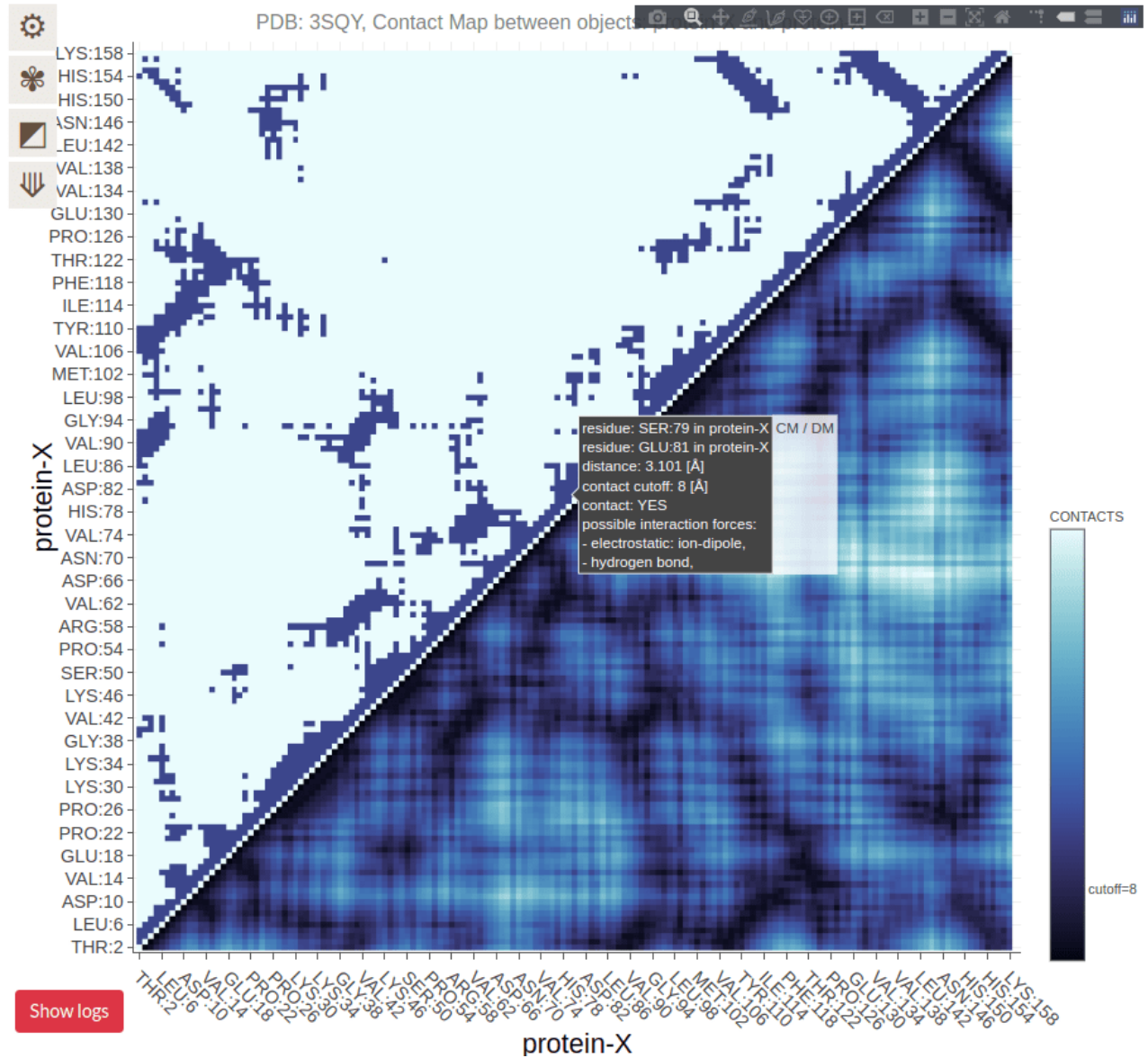
Contacts are determined by spatial proximity, typically within a certain distance threshold

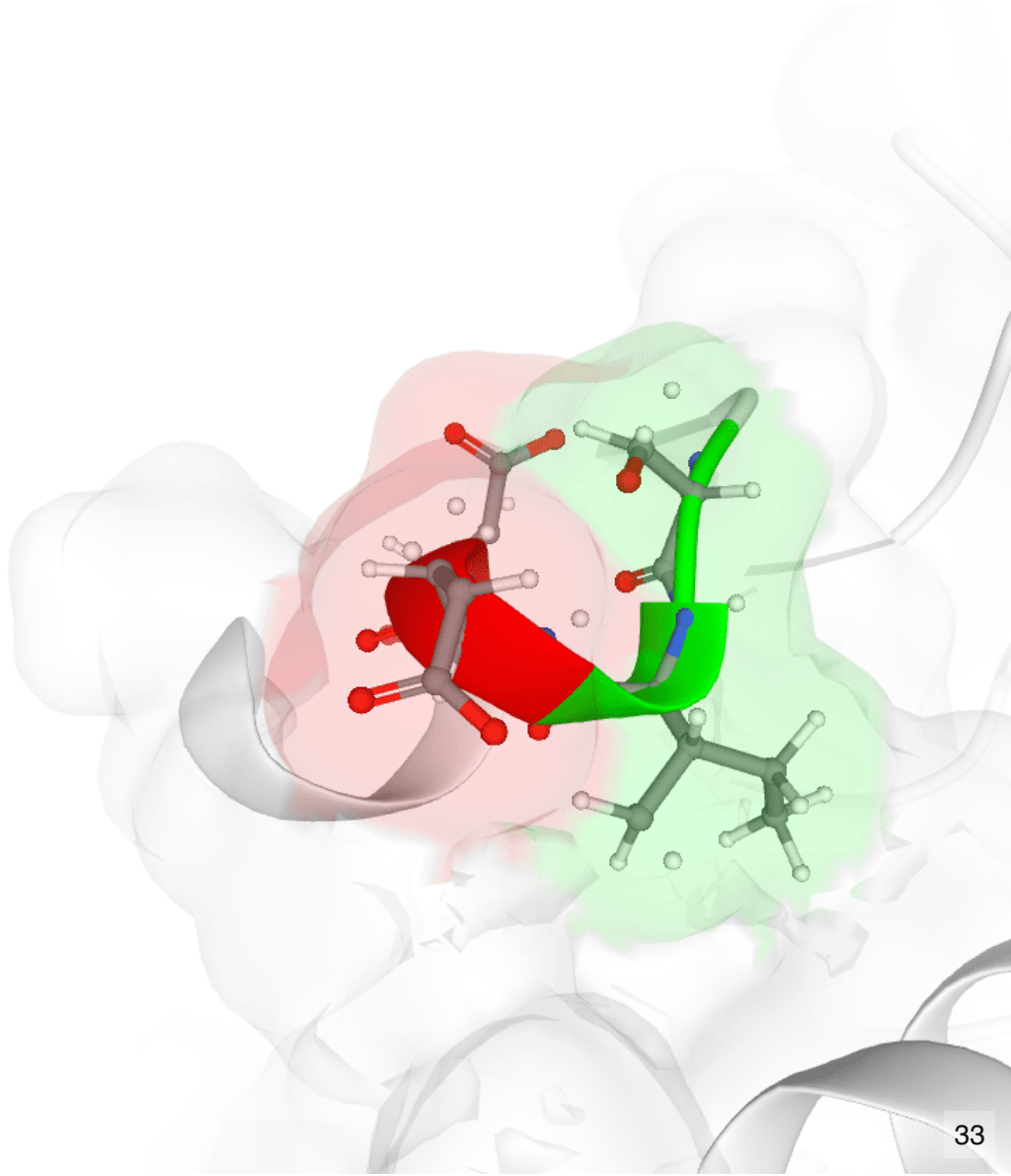
Residues far apart in the sequence can still be close in the 3D structure, reflected in the contact map



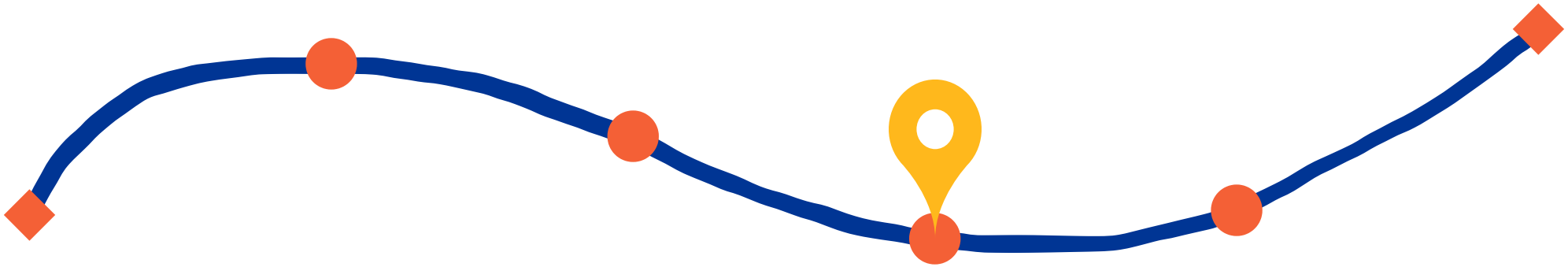


Residues on the diagonal are adjacent in sequence (and spatially)



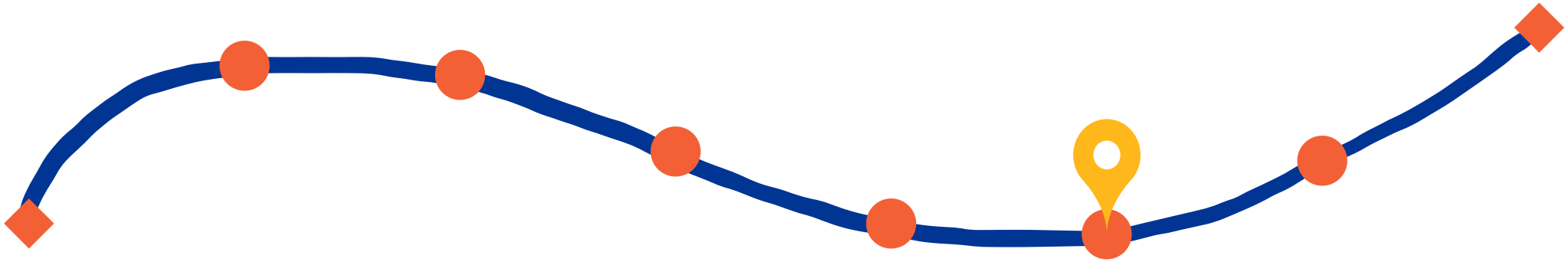


After today, you should have a better understanding of



Comprehend how coevolution provides structural insights

After today, you should be able to



Comprehend how coevolution provides structural insights

The Rise of Machine Learning in Structural Biology

Traditional methods like **homology modeling** and threading rely on **templates and known structures**

ML predicts 3D structures **only from sequence data**

AlphaFold (DeepMind) and **RosettaFold** (Baker Lab) lead the charge in this area

What is AlphaFold?

Developed by DeepMind, **AlphaFold** predicts protein structures with atomic accuracy by using deep learning models trained on large structural datasets

Breakthroughs

- AlphaFold 2 achieved near-experimental level accuracy in the 2020 **CASP14** competition (Critical Assessment of protein Structure Prediction)
- **AlphaFold 3** (2024) predicts proteins, DNA, RNA, ligands, and post-translational modifications

Coevolving residues mutate in a correlated manner

Mutations in one residue often result in **compensatory mutations** in its interacting partner

This is observed across species through **analysis of homologous protein sequences**

Correlated mutations indicate **functionally significant** residue pairs

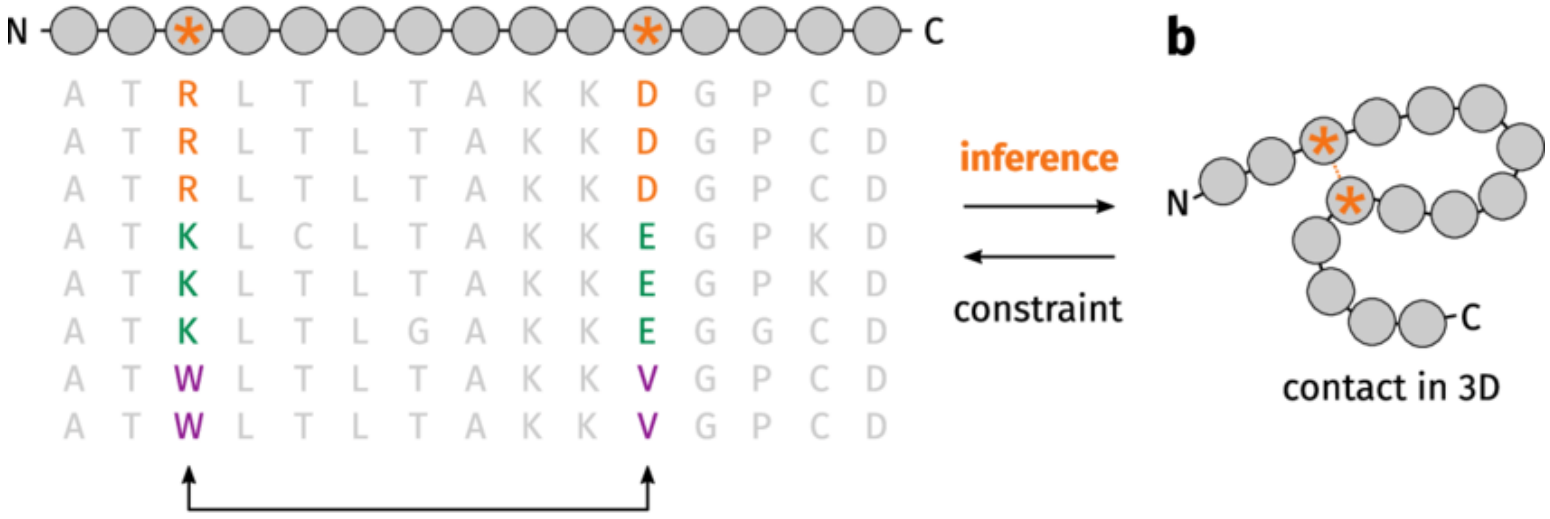
Evolution

Evolutionary Analysis Reveals Structural Insights

Coevolution analysis helps predict which residues are close in the 3D structure

Residues showing correlated mutations are likely to be spatially close in the folded protein

This is particularly useful when no experimental structure is available



Multiple Sequence Alignments Enable Coevolution Detection

Coevolution is detected using large MSAs from homologous proteins

The more diverse the sequences in the MSA, the better the resolution of coevolving residues

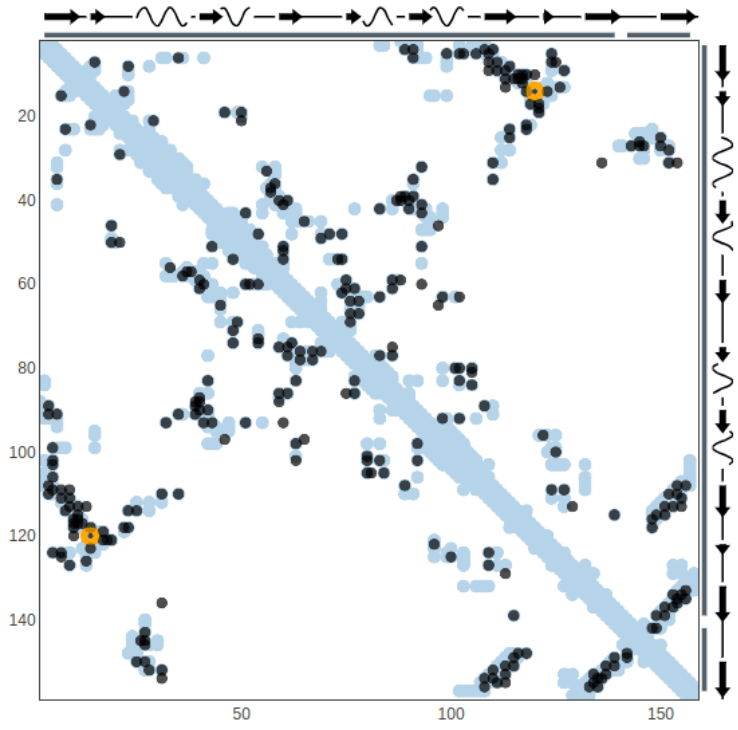
Evolutionary information from MSAs guides predictions for residue-residue contacts

evcouplings.org

Coevolution example: DHFR

Residues with a high Score (i.e., coevolve) are near each other in the protein's structure (i.e., small distance)

Val14 and Gly120 coevolved



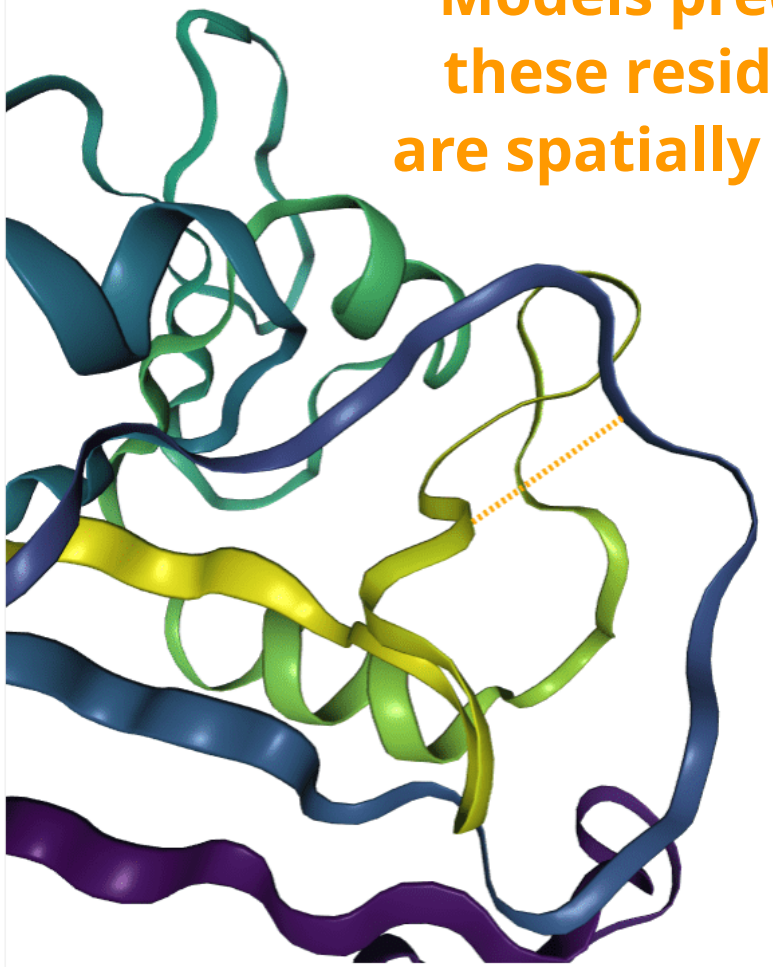
evcouplings.org

Structures | Legend 153 ECs

	i	A_i	j	A_j	Score	Prob	Dist	PPV
1	59	R	75	D	28.14	1.00	2.58	1.00
2	14	V	120	G	25.06	1.00	3.49	1.00
3	77	I	86	L	21.54	1.00	4.85	1.00
4	113	V	153	L	21.54	1.00	3.66	1.00
5	134	V	153	L	21.45	1.00	4.42	1.00
6	11	L	115	E	20.84	1.00	3.27	1.00
7	37	T	57	N	20.63	1.00	3.56	1.00
8	5	I	106	V	20.23	1.00	3.68	1.00
9	80	I	102	M	20.18	1.00	3.77	1.00
10	108	D	154	H	19.81	1.00	2.62	1.00
11	115	E	151	T	17.78	1.00	2.62	1.00
12	39	H	91	F	17.39	1.00	3.73	1.00
13	135	A	156	I	17.11	1.00	3.59	1.00
14	4	S	110	Y	17.04	1.00	3.79	1.00
15	139	E	151	T	16.45	1.00	2.65	1.00
16	137	S	151	T	16.01	1.00	3.30	1.00
17	32	V	93	F	15.74	1.00	3.35	1.00
18	142	L	149	P	15.33	1.00	3.71	1.00
19	14	V	123	F	15.32	1.00	3.49	1.00
20	109	M	127	Y	15.06	1.00	4.03	1.00
21	40	T	59	R	14.60	1.00	3.32	1.00
22	92	I	102	M	14.51	1.00	3.95	1.00
23	4	S	89	H	14.19	1.00	2.63	1.00
24	39	H	89	H	14.19	1.00	3.23	1.00
25	10	D	118	F	13.94	1.00	2.75	1.00
26	11	L	117	K	13.87	1.00	3.74	1.00
27	4	S	108	D	13.85	1.00	2.45	1.00
28	108	D	156	I	13.82	1.00	3.93	1.00
29	31	H	110	Y	13.66	1.00	4.48	1.00
30	137	S	153	L	13.61	1.00	3.27	1.00
31	110	Y	154	H	13.29	1.00	3.32	1.00
32	12	Q	117	K	13.04	1.00	3.48	1.00
33	35	L	110	Y	12.87	1.00	3.42	1.00
34	9	H	113	V	12.78	1.00	3.27	1.00
35	136	S	154	H	12.54	1.00	2.93	1.00
36	75	D	86	L	12.45	1.00	6.11	0.97
37	7	V	111	I	12.42	1.00	3.20	0.97
38	27	N	150	H	12.17	1.00	2.87	0.97

Legend 153 ECs

Models predict these residues are spatially close



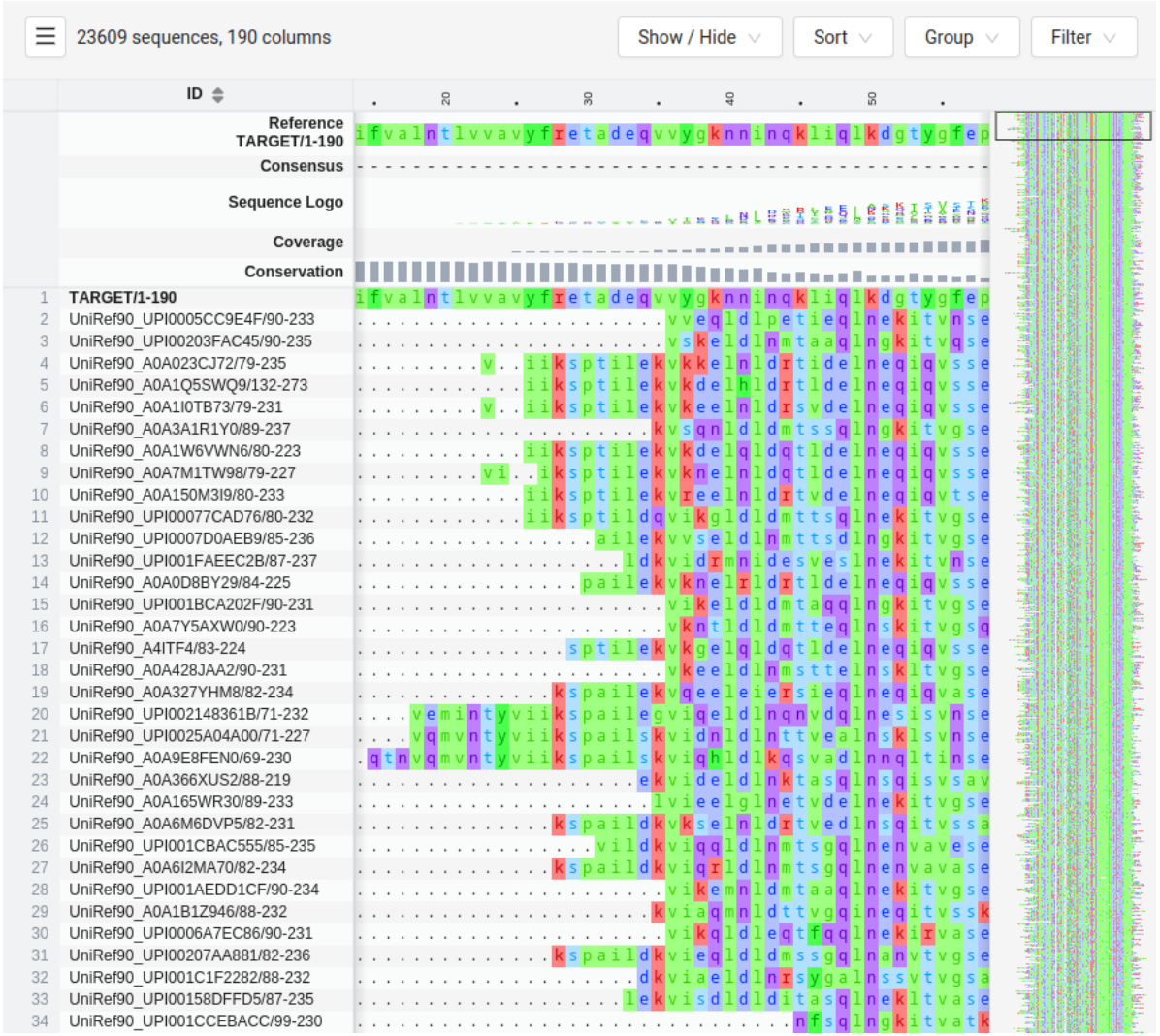
Structures

Coevolutionary signals can be noisy

Not all correlated mutations are due to direct physical interactions; some may be indirect

Noise in the data can come from random mutations or insufficient evolutionary diversity.

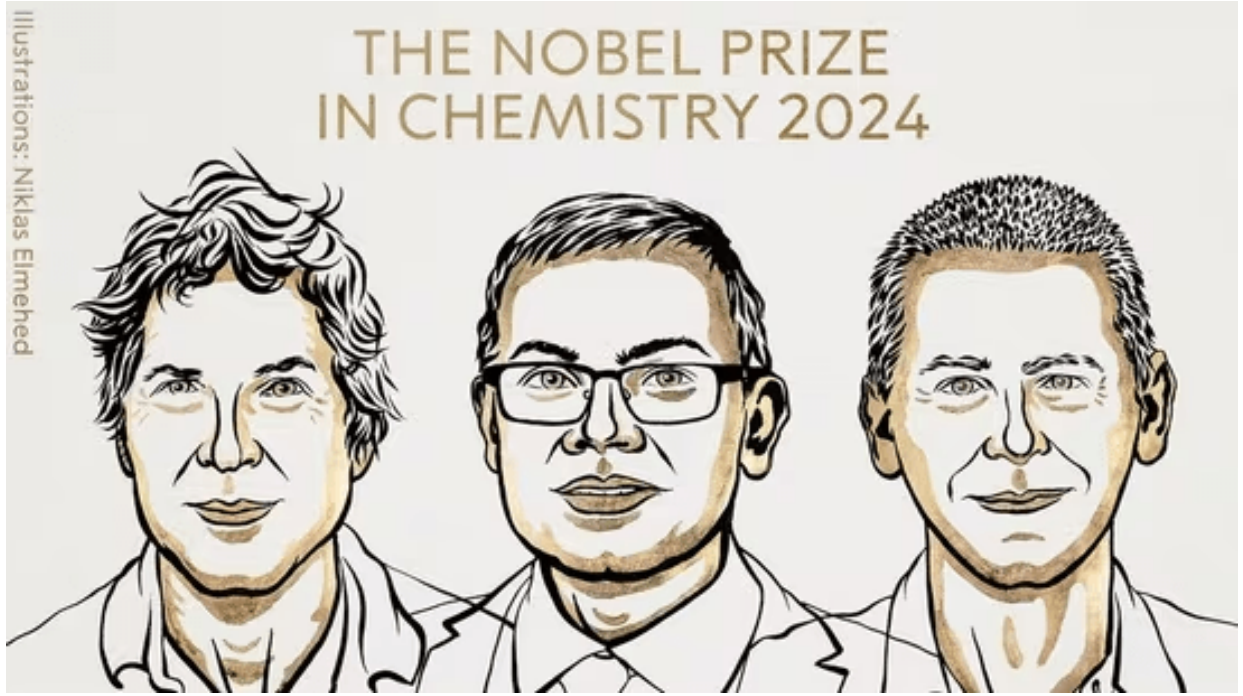
Large and diverse sequence data sets are needed for reliable coevolution predictions.



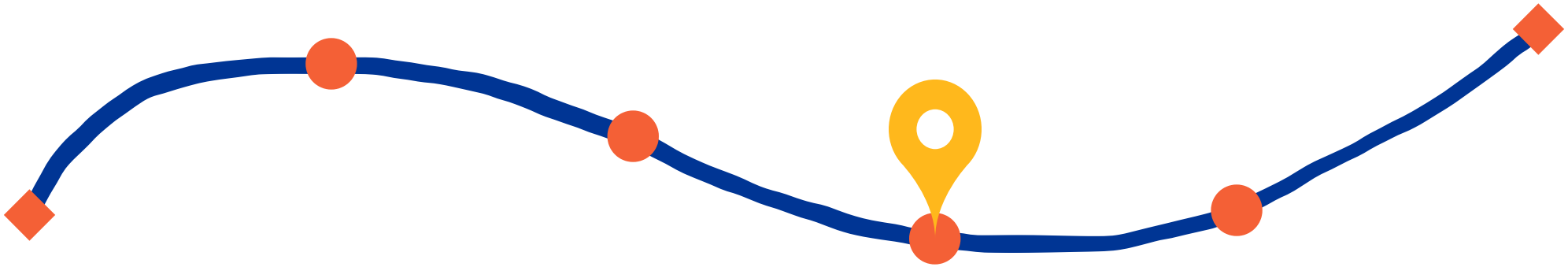
Machine learning leverages coevolution for high-accuracy predictions

AlphaFold and RosettaFold utilize coevolutionary data from MSAs to predict residue interactions

These models incorporate evolutionary information along with structural features, leading to highly accurate predictions



After today, you should have a better understanding of



Explain why ML models are dominate

protein structure prediction

AlphaFold pipeline, simplified

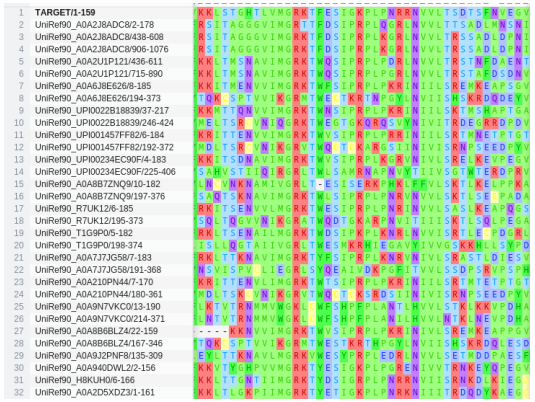
Given the following data

Predict

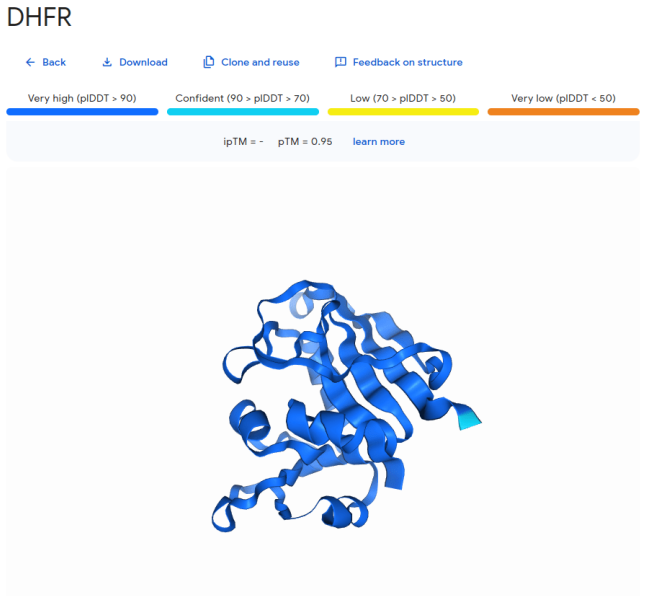
Input sequence

```
MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKK
LSTGHTLVMGRKTFESIGKPLPNRRNVVLTSDTS
FNVEGVDVIHSIEDIYQLPGHVFIFGGQTLFEEM
IDKVDDMYITVIEGKFRGDTFFPPYTFEDWEVAS
SVEGKLDEKNTIPHTFLHLIRKK
```

Multiple Sequence Alignment



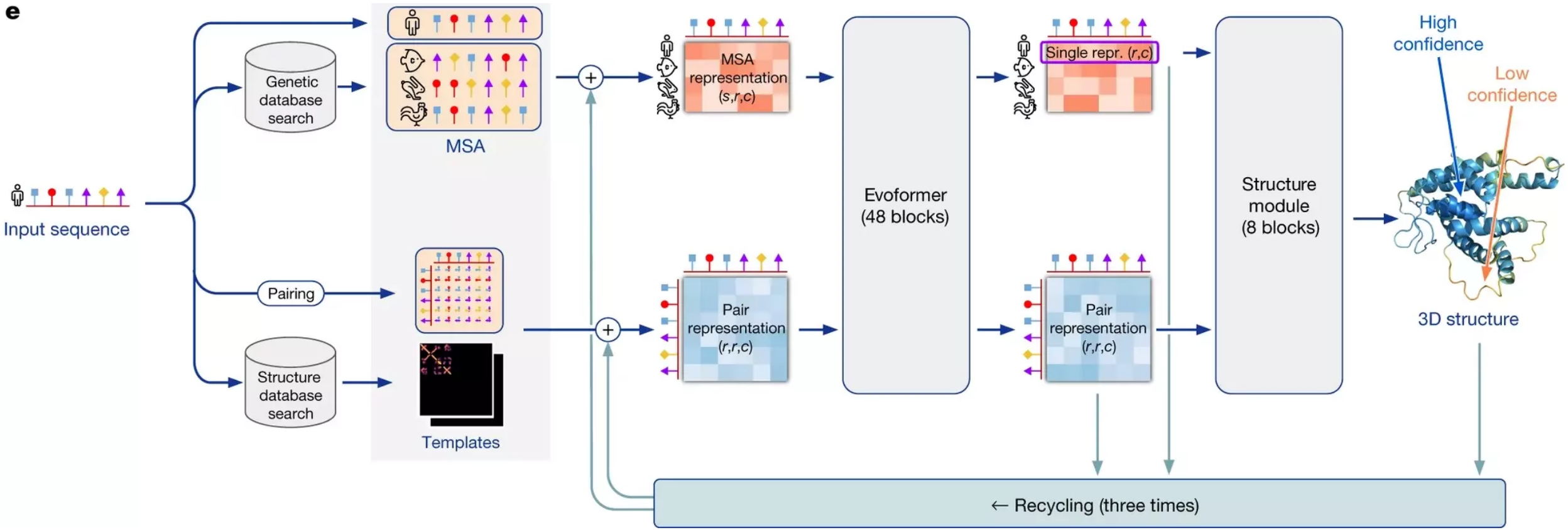
ML models



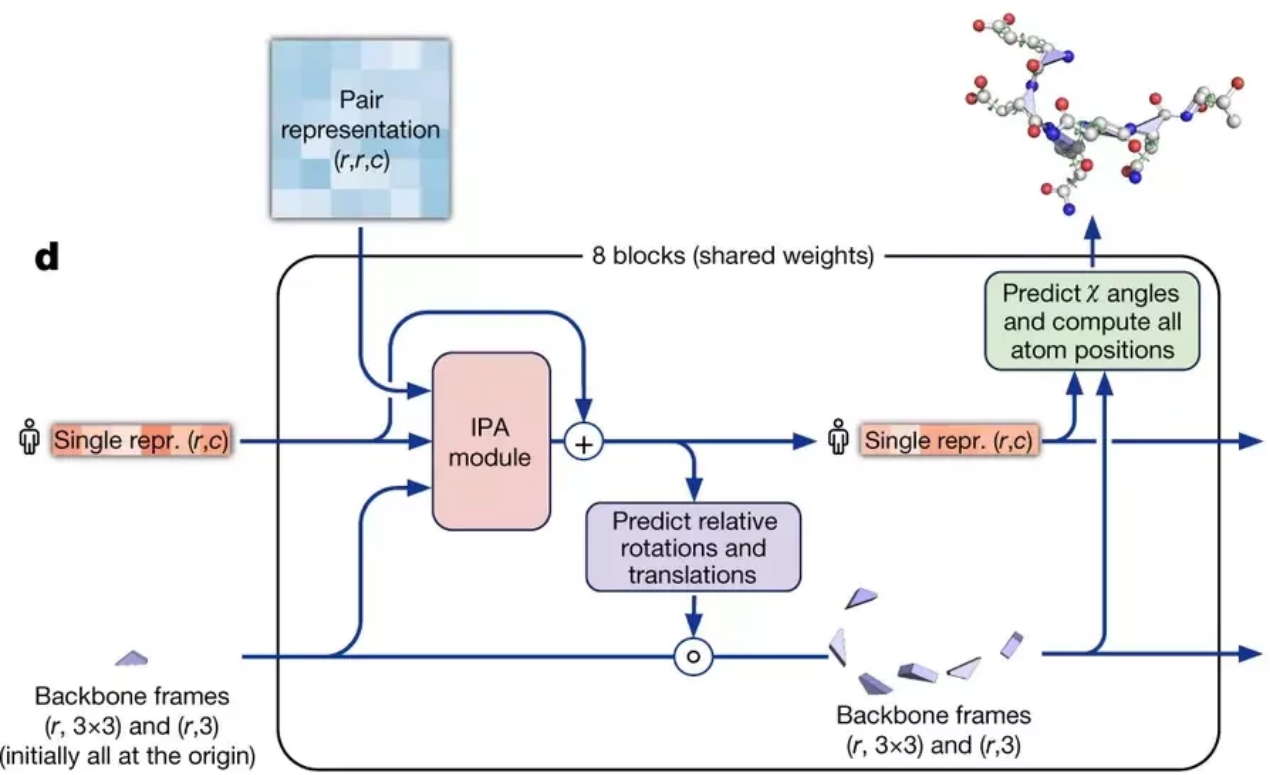
Atomistic structure

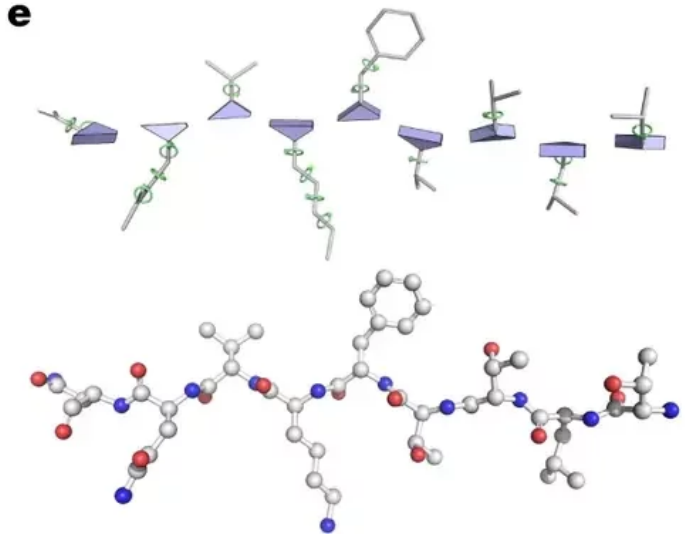
AlphaFold 2 pipeline: **Evoformer**

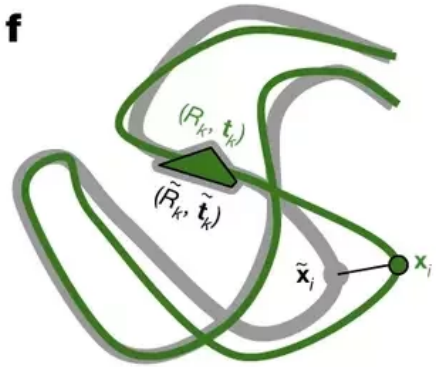
Using MSAs and contact maps, DeepMind trained a model to predict protein structures



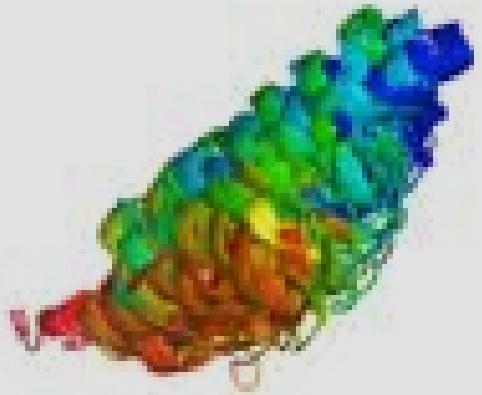
Contact maps are converted into dihedral angles







AF2 iterations

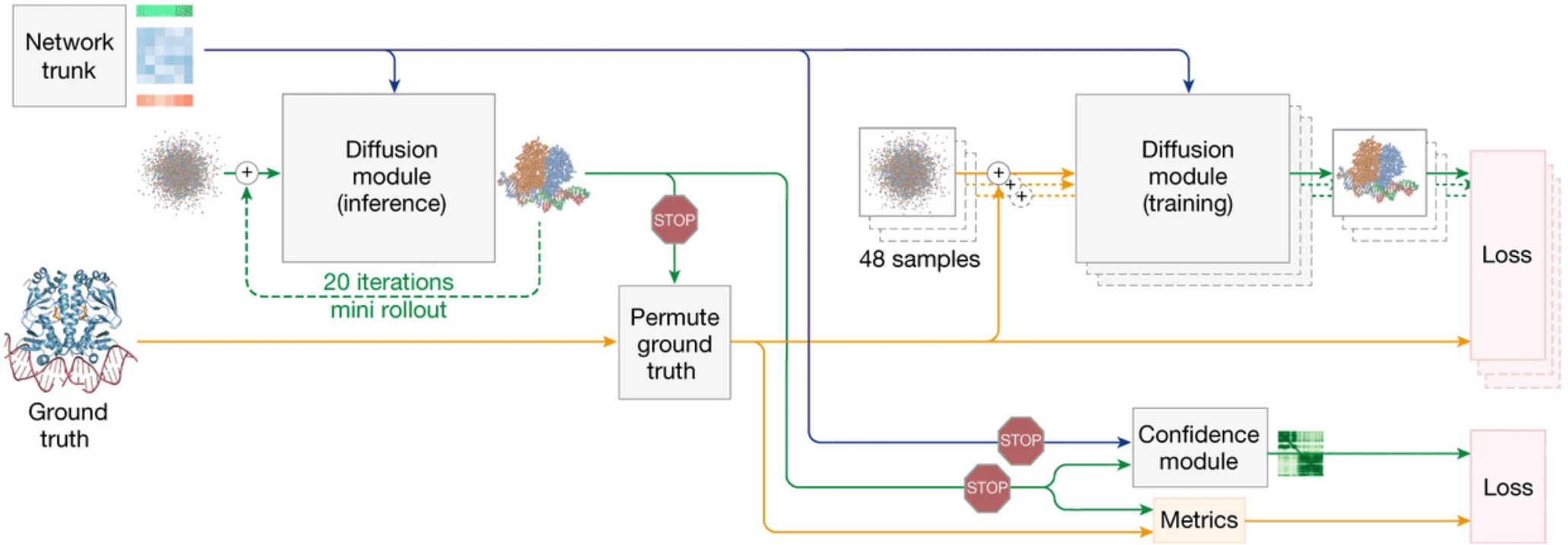


Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

What is new in AlphaFold 3?

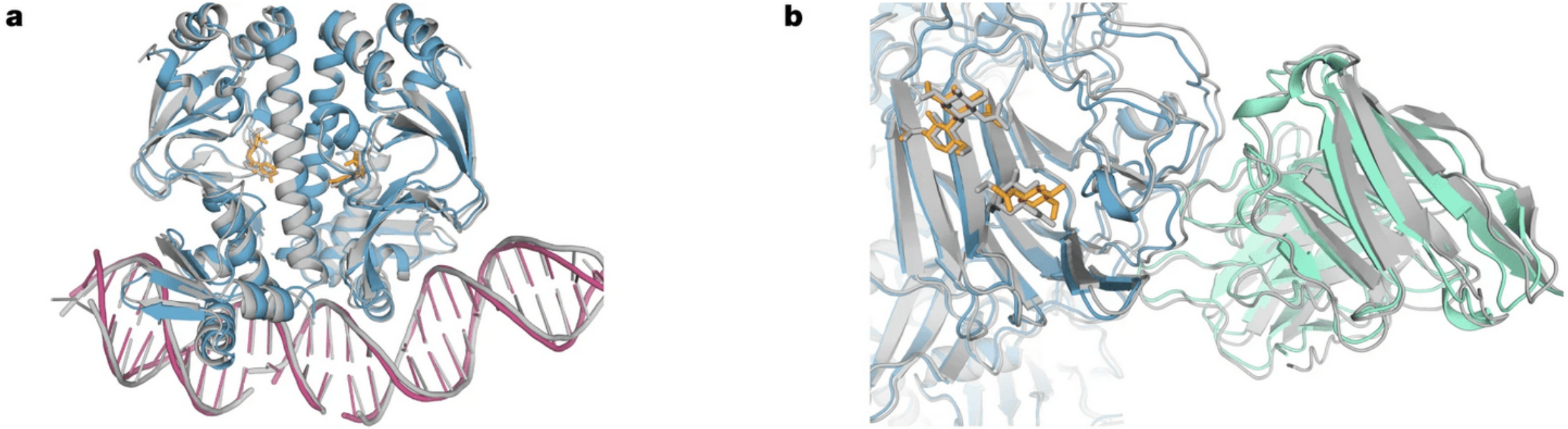
Biggest change is the use of a **diffusion model**

Diffusion models essentially learn to **unscramble atoms into a structure**



AlphaFold 3 is supercharged for any biomolecule

Proteins, DNA, RNA, ligands, PTMs, protein-proteins, etc.



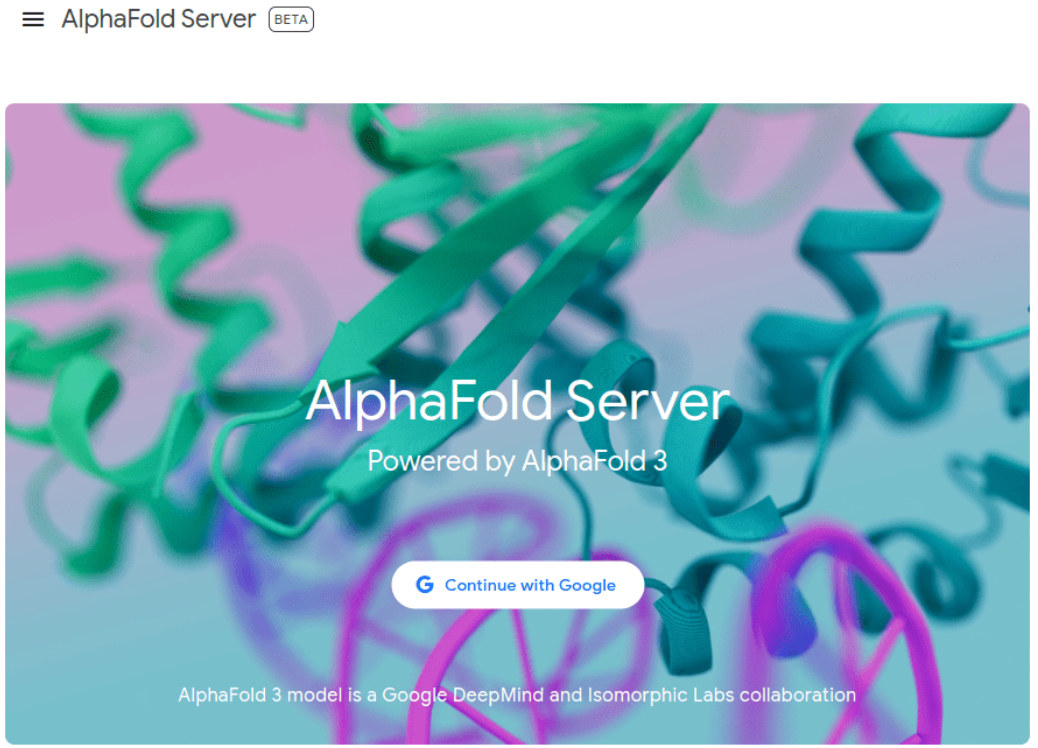
AlphaFold 3

```
MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKKLSTGHTL
VMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVDVIHSIED
IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF
FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK
```

DHFR (UniProt)

```
MGKKEVILLFLAVIFVALNTLVVAVYFRETADEQVVYGK
NNINQKLIQLKDGTYGFEPALPHVGTFKVLDSNRVPQIA
QEIIRNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN
NLHGYEGFLFKDNVPLTYPQEFILSNLDGKVRSLQNYDY
DLDVLFGEKEEVKSEILRGLYYNTYTRAFSPYKL
```

Novel protein
(ChatGPT)



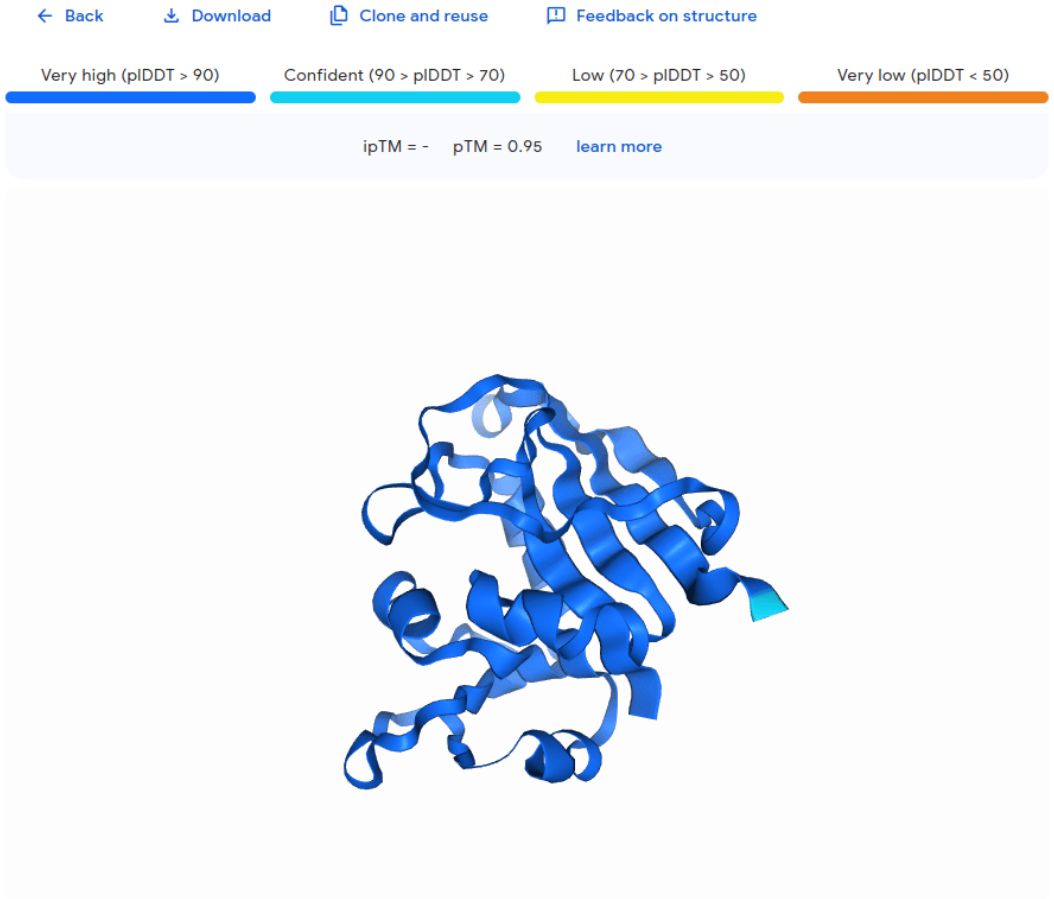
How does AlphaFold Server work?

AlphaFold Server is a web-service that can generate highly accurate biomolecular structure predictions containing proteins, DNA, RNA, ligands, ions, and also model chemical modifications for proteins and nucleic acids in one platform. It's powered by the newest AlphaFold 3 model.

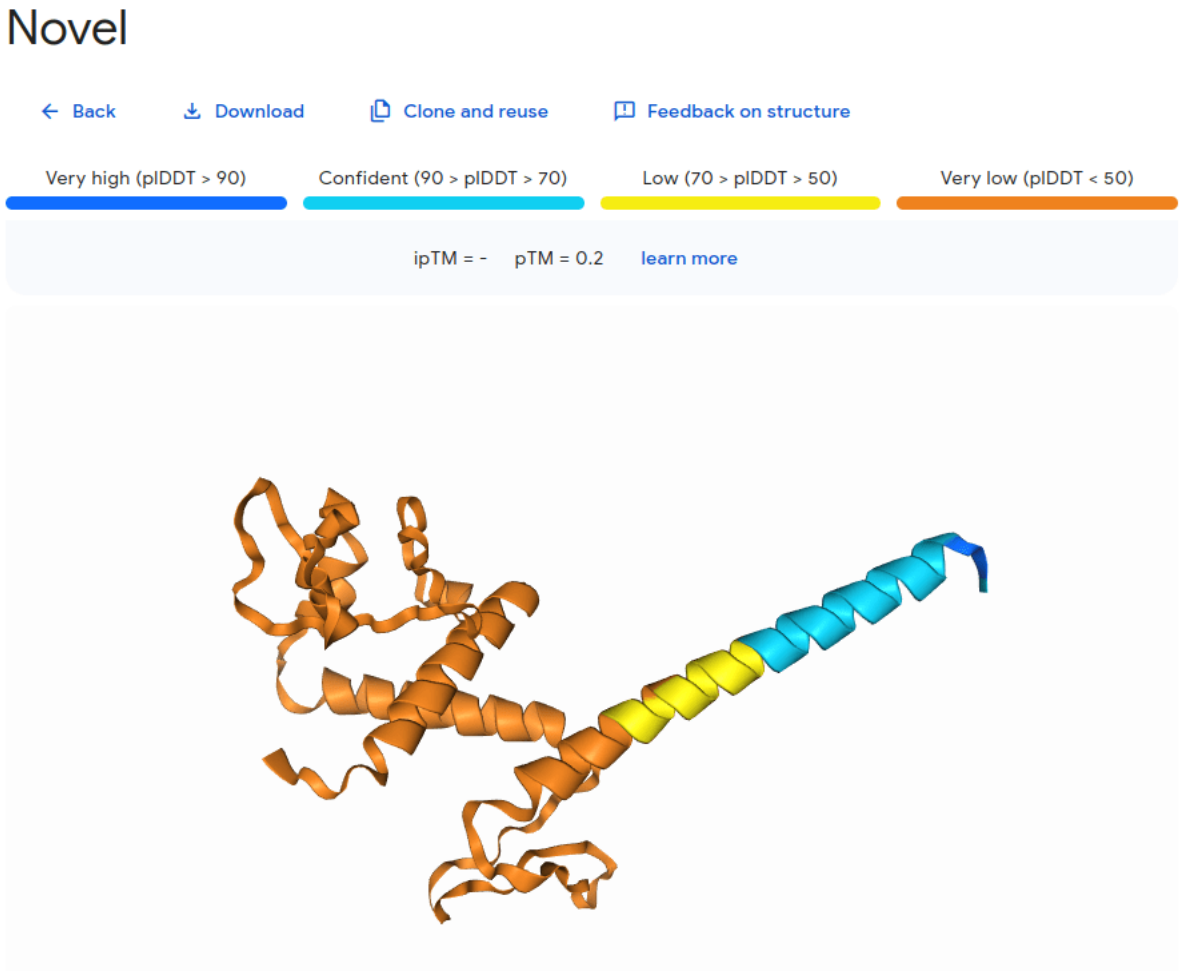
alphafoldserver.com

AlphaFold 3 is a breakthrough, not the final solution

DHFR



Novel



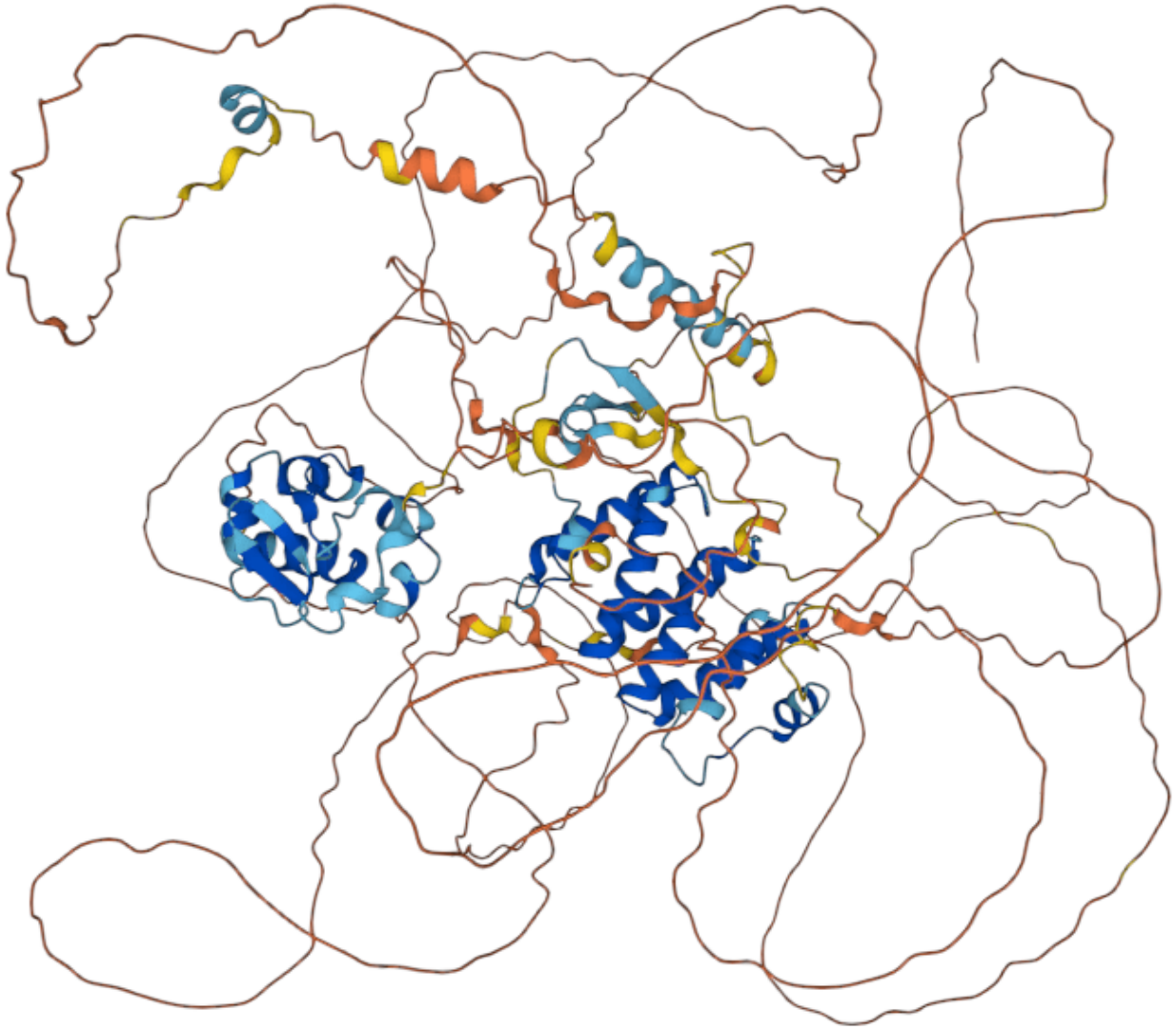
Caveat: Proteins are **dynamic**

https://www.youtube.com/embed/AjcUmxT-QEA?si=qupqTpuV5IvOB_ut&start=43&enablejsapi=1

What about intrinsically disordered proteins?

At least 40% of proteins have disordered regions

AlphaFold (and all other methods) struggle with disordered regions



LARP1

Before the next class, you should

Lecture 11A:

Protein structure prediction - Foundations

Lecture 11B:

Protein structure prediction - Methodology

Today

Thursday

- Work on P02B, P02C, and P03A