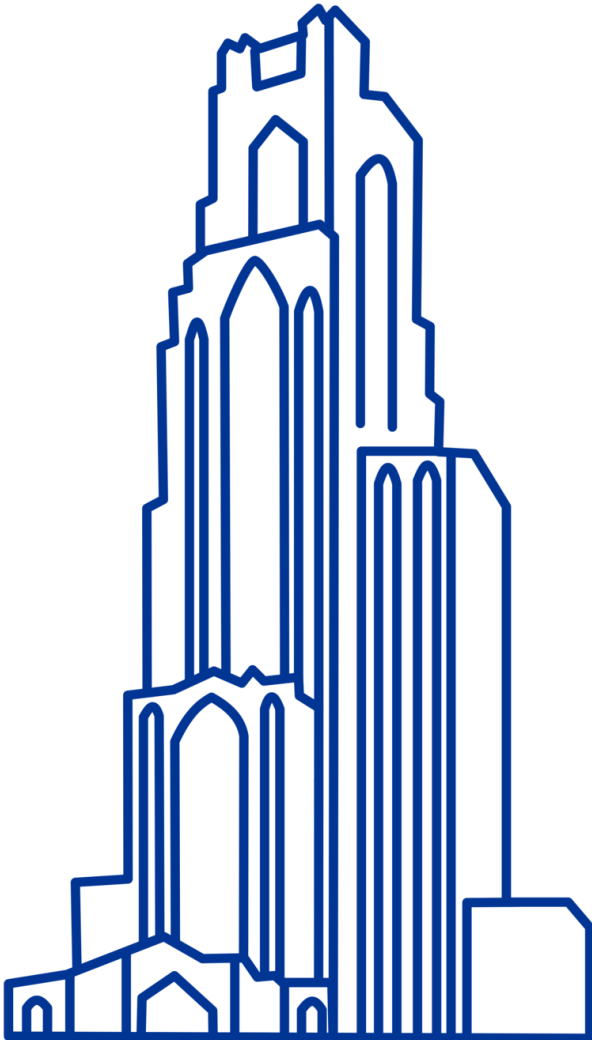


# Computational Biology

## (BIOSC 1540)

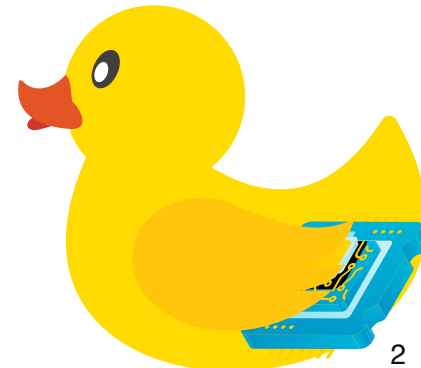
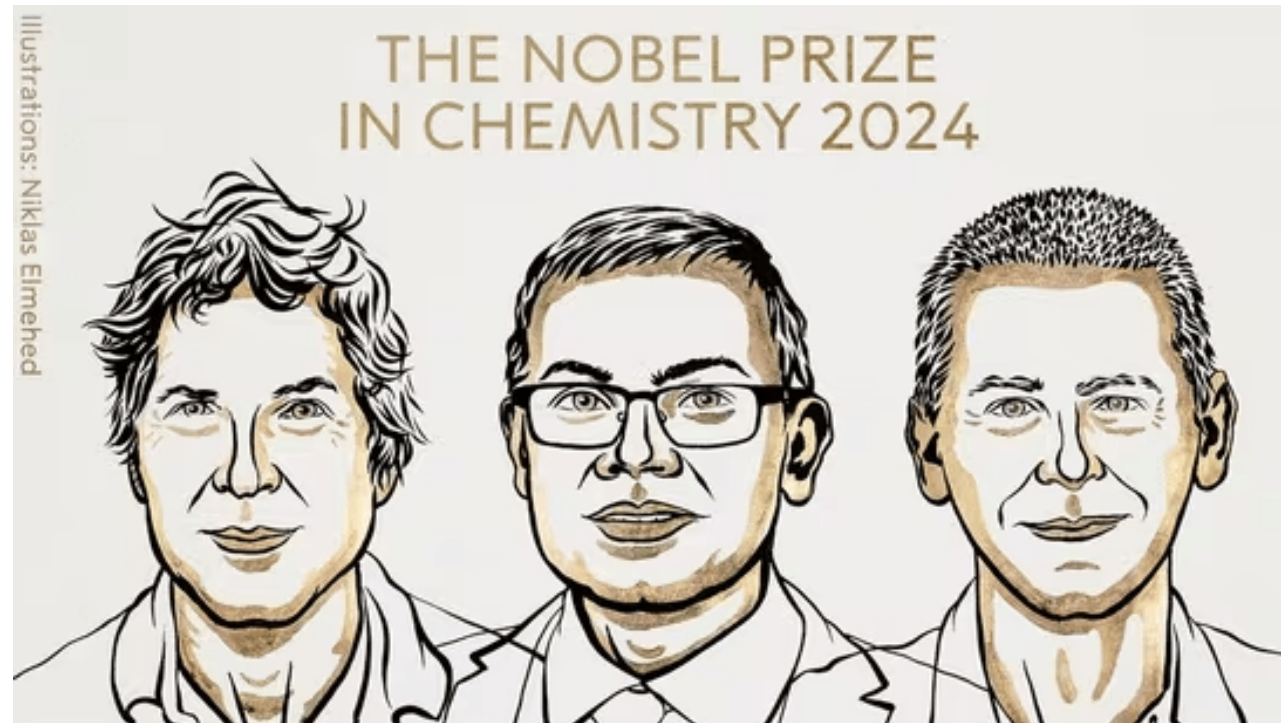
### **Lecture 12:** Protein structure prediction

Oct 10, 2024

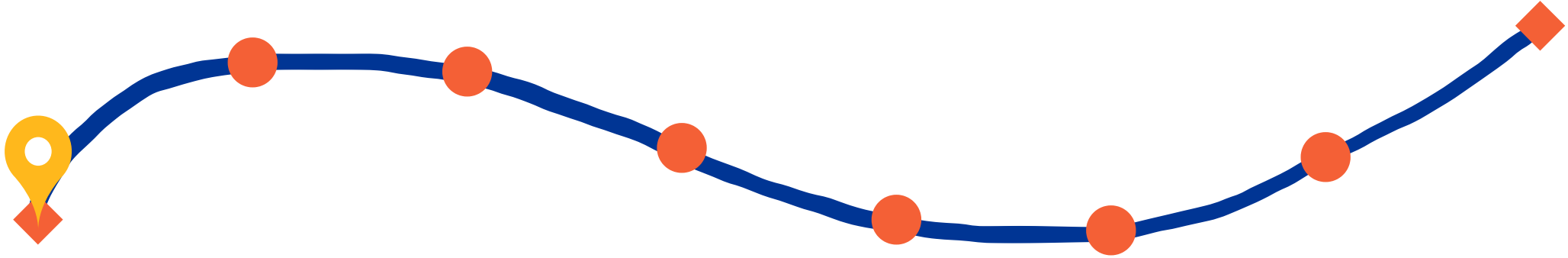


# Announcements

- No class on Tuesday (10/15)
- No office hours (mine or UTA) next week - will resume on 10/22
- Will have Programming+ recitations
- A05 will be posted tomorrow
- David Baker, John Jumper, and Demis Hassabis won the Nobel Prize in Chemistry for "computational protein design" and "protein structure prediction"



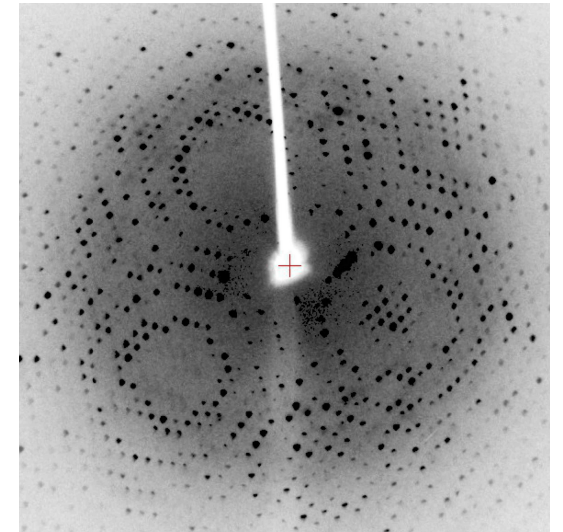
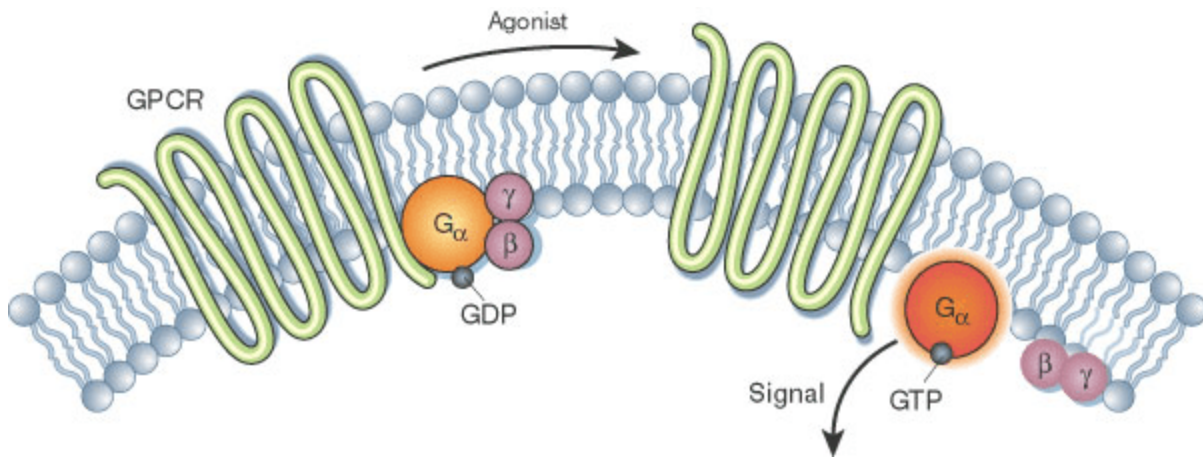
# After today, you should be able to



Why are we learning about protein  
structure prediction?

# Why predict protein structure?

Protein structure dictates interactions, signaling, and biochemical roles



Experimental methods (X-ray, Cryo-EM) provide high-resolution structures but are resource-intensive and time-consuming



# Structural insights can accelerate ... everything?

- **Drug Discovery:** Designing small-molecule inhibitors or antibodies that target specific protein conformations.
- **Biotechnology:** Engineering proteins for industrial or therapeutic applications.
- **Disease Research:** Mutations causing structural defects linked to diseases like Alzheimer's and cystic fibrosis.

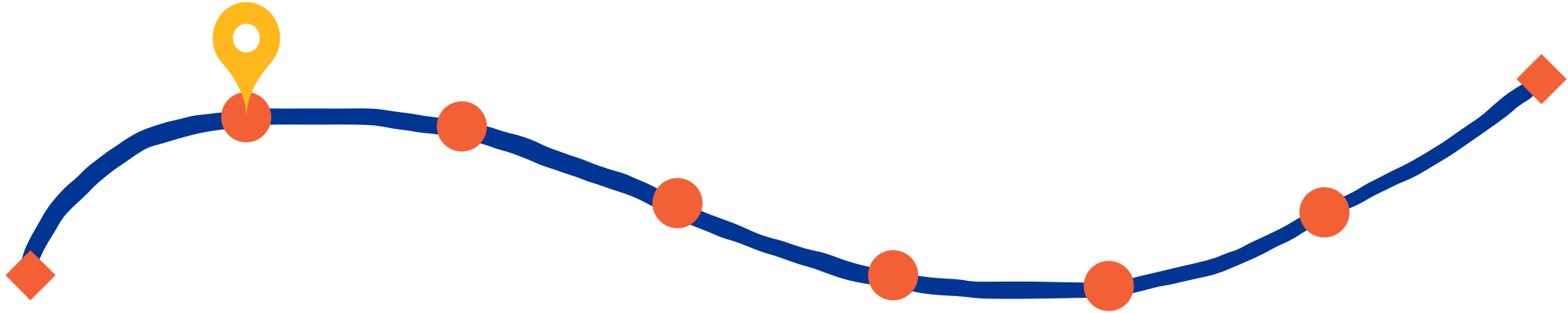
# Prediction is critical for the future of biology

Advances in predictive accuracy are  
opening new frontiers in biology

Integrating predictive models with  
experimental data is the way forward

Structure prediction complements genomics  
and transcriptomics to create a holistic  
understanding of biological function

# After today, you should be able to



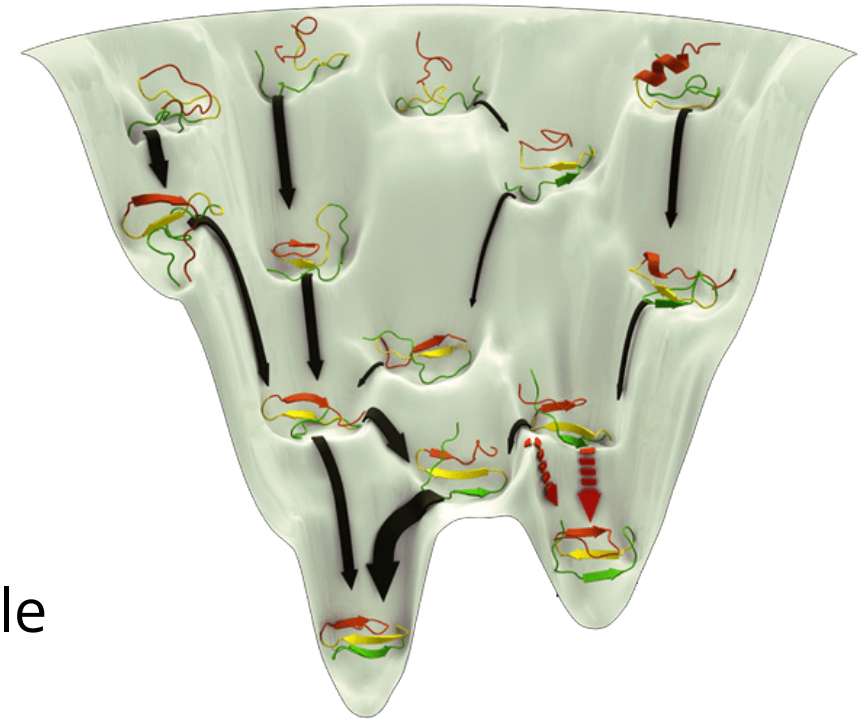
Identify what makes structure  
prediction challenging

# What makes structure prediction hard: Conformational space

Proteins can adopt a large number of possible conformations

**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  $\sim 3^{100}$  possible conformations



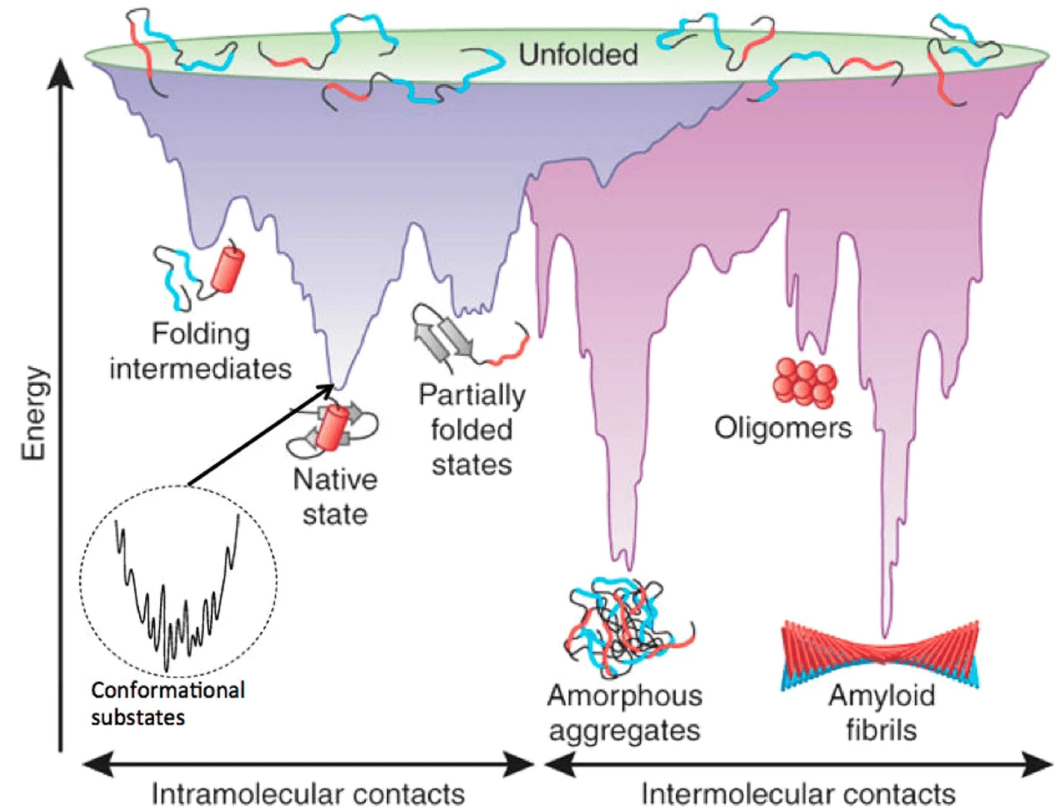
# What makes structure prediction hard: Complex energy landscape

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

Understand how the system's energy changes upon reactions or movements

Proteins fold to the lowest free-energy state, but this landscape is highly rugged

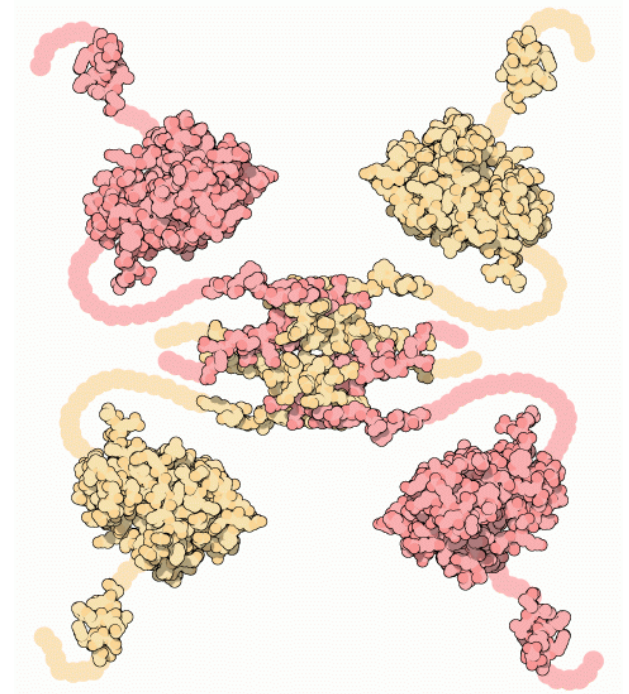
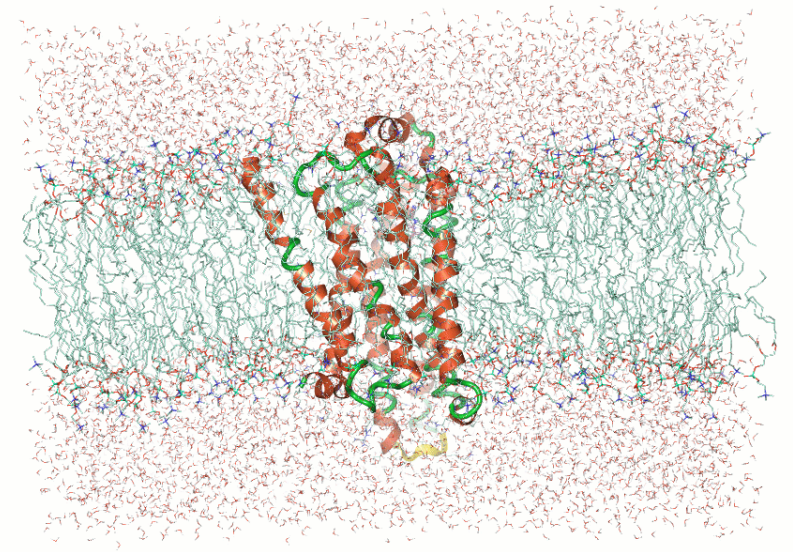
**Energy calculations** are computationally intensive and depend on accurate force fields



# What makes structure prediction hard: **Flexibility and dynamics**

**Proteins are not static;** they adopt multiple conformations (flexibility) based on their environment or interactions with other molecules

**Some proteins or regions do not adopt a fixed 3D structure** but remain disordered or flexible under physiological conditions





# What makes structure prediction hard:

## Environmental effects

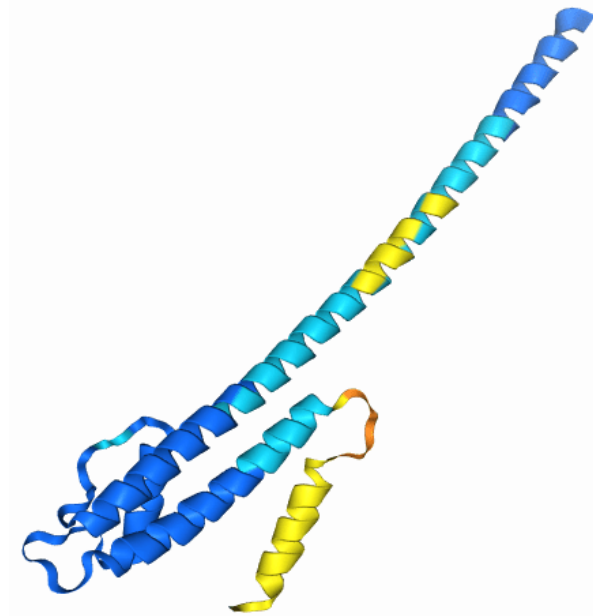
Proteins fold differently in different environments

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

7MHX



pH-gated  
K<sup>+</sup> channel



AlphaFold 3

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

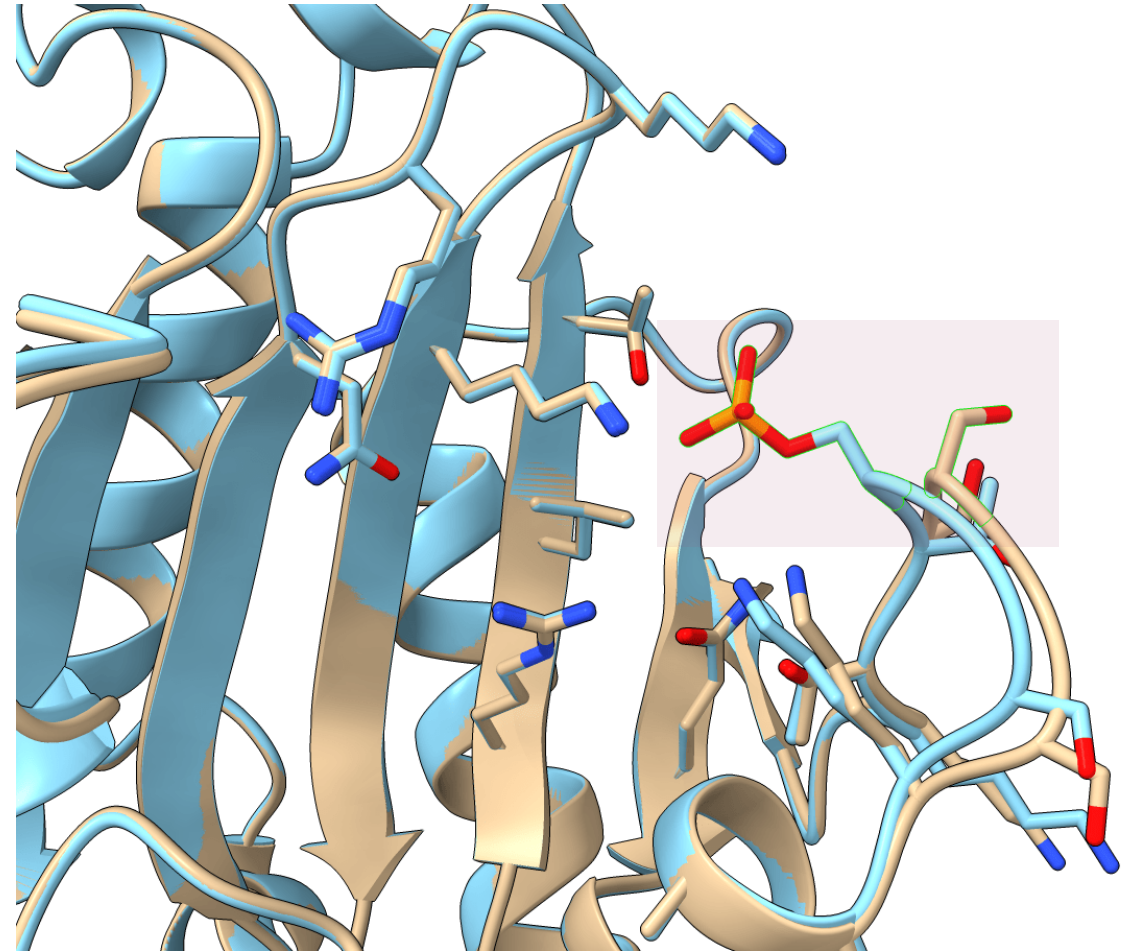
# What makes structure prediction hard: Post-translational modifications

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

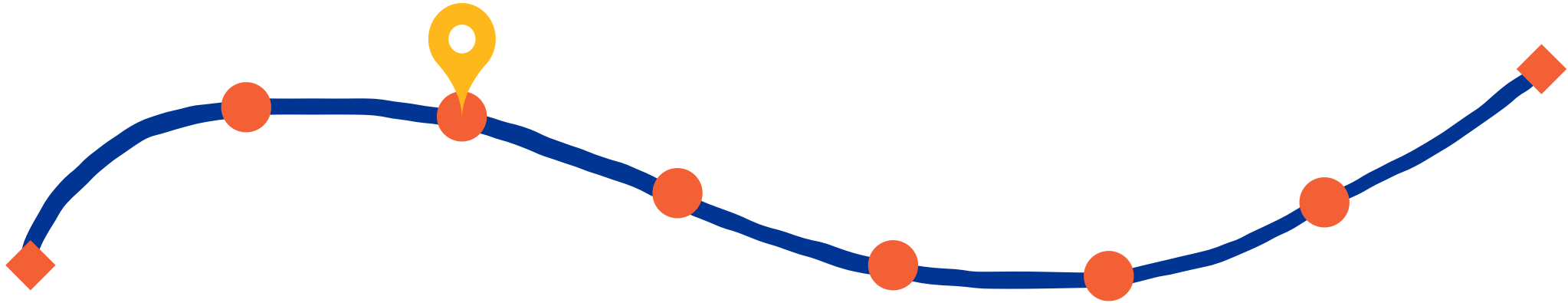


# What makes structure prediction hard: **Methods are data driven**

Our predictions **rely on similarity to known structures**, but novel sequences or folds (for which no homologous structures exist) are difficult to predict accurately

**Example:** AlphaFold has made strides, but predicting **de novo** structures remains challenging, especially for proteins with no templates

# After today, you should be able to



Explain homology modeling

# Homology modeling predicts protein structures based on evolutionary relationships

The main principle is that proteins with **similar sequences tend to fold into similar structures**

Common tools for homology modeling include **MODELLER**, **SWISS-MODEL**, and **Phyre2**

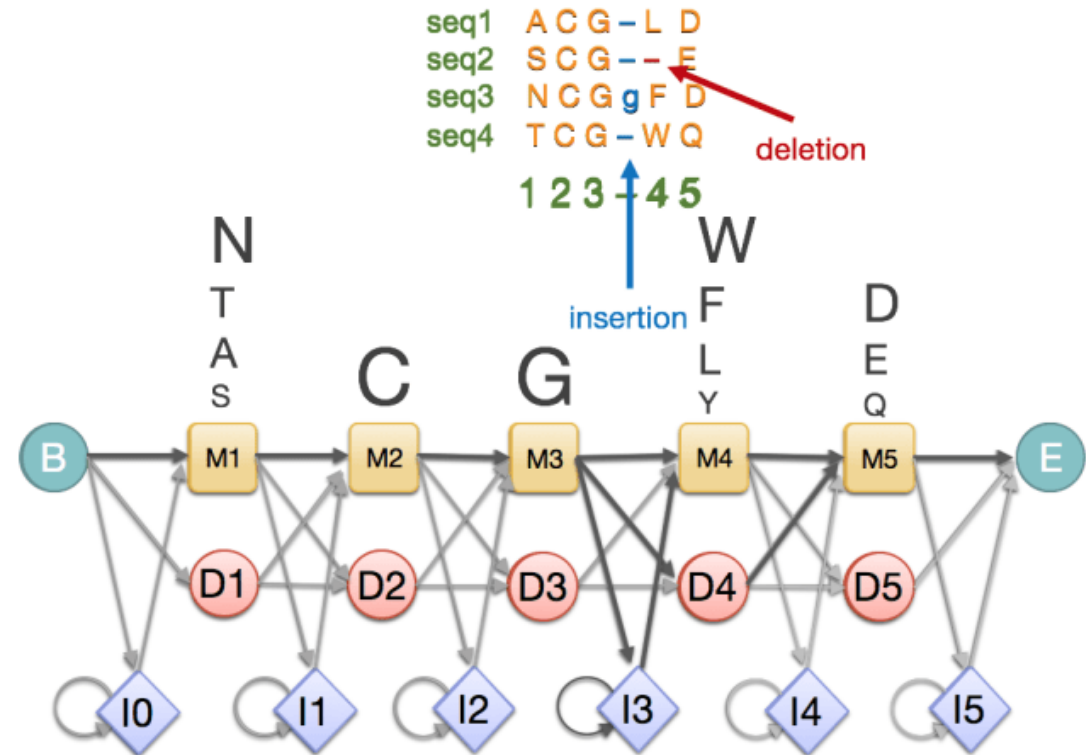
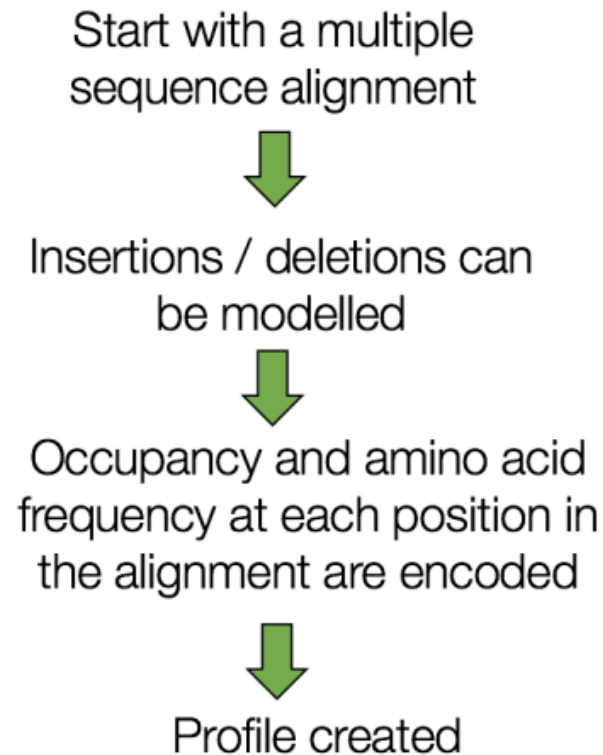
Homology modeling is the most accurate when sequence identity to other proteins is high (>30%)

The screenshot displays the SWISS-MODEL web interface. At the top, there's a navigation bar with links for Modelling, Repository, Tools, Documentation, Log in, and Create Account. Below this, the 'All Projects' section shows an 'Untitled Project' created today at 03:08. The 'Summary' tab is active, showing 50 templates and 0 models. The 'Template Results' section is visible, showing a list of templates with their sequence similarity and alignment. A 'Build Models' button is present. On the right, a 3D protein structure is shown in a cartoon representation, with a 'Clear Selection' button and a 'Cartoon' view selector. The bottom part of the interface shows a detailed sequence alignment between the target protein and several templates, with sequence identity highlighted in green.

# Hidden Markov Models (HMMs) Capture Evolutionary Patterns in Proteins

HMMs are statistical models representing sequences using probabilities for matches, insertions, and deletions

**Essentially more robust alignments**





# A Markov model predicts outcomes based on transitional probabilities

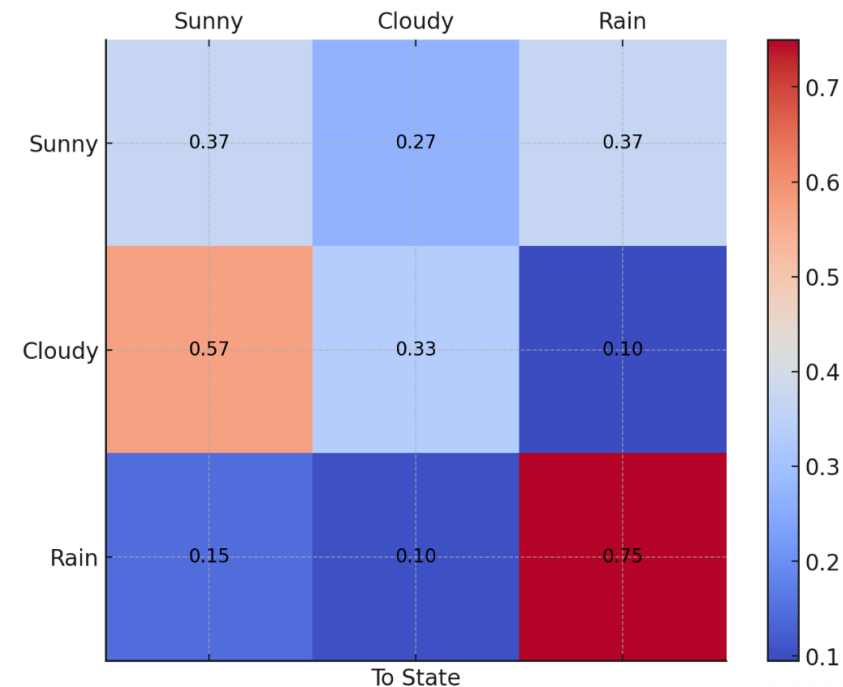
Suppose I collect weather data in Pittsburgh for the past 30 days: **Sunny**, **Cloudy**, or **Rain**

I want to figure out how to predict tomorrow's weather based on today's

**Example:** If today is cloudy, there is a 57% chance it will be Sunny tomorrow

Today's  
weather

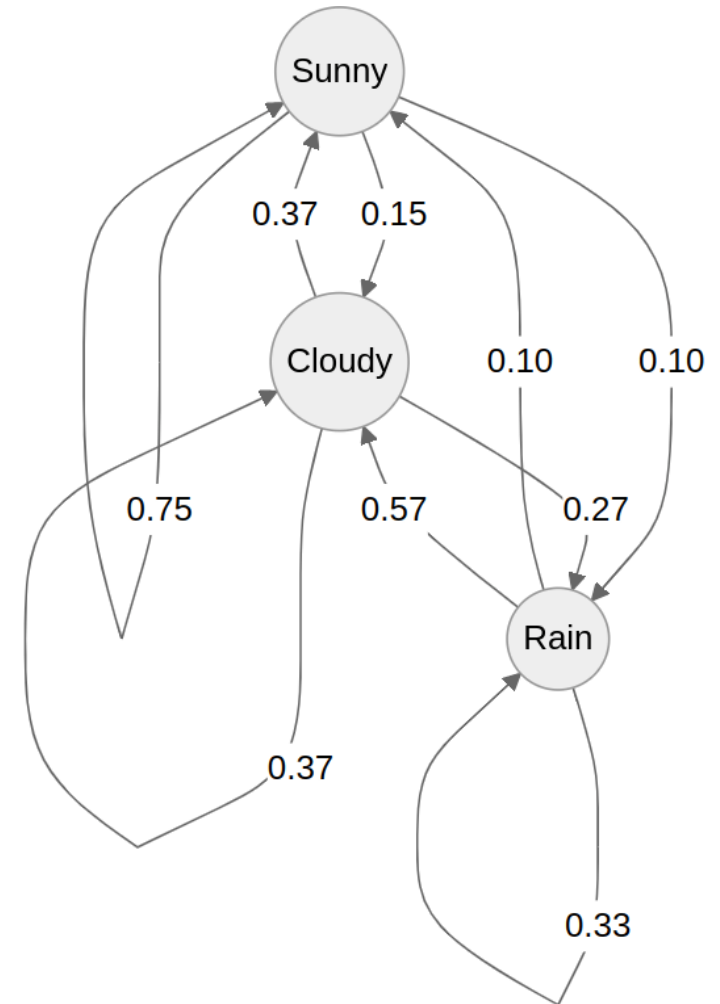
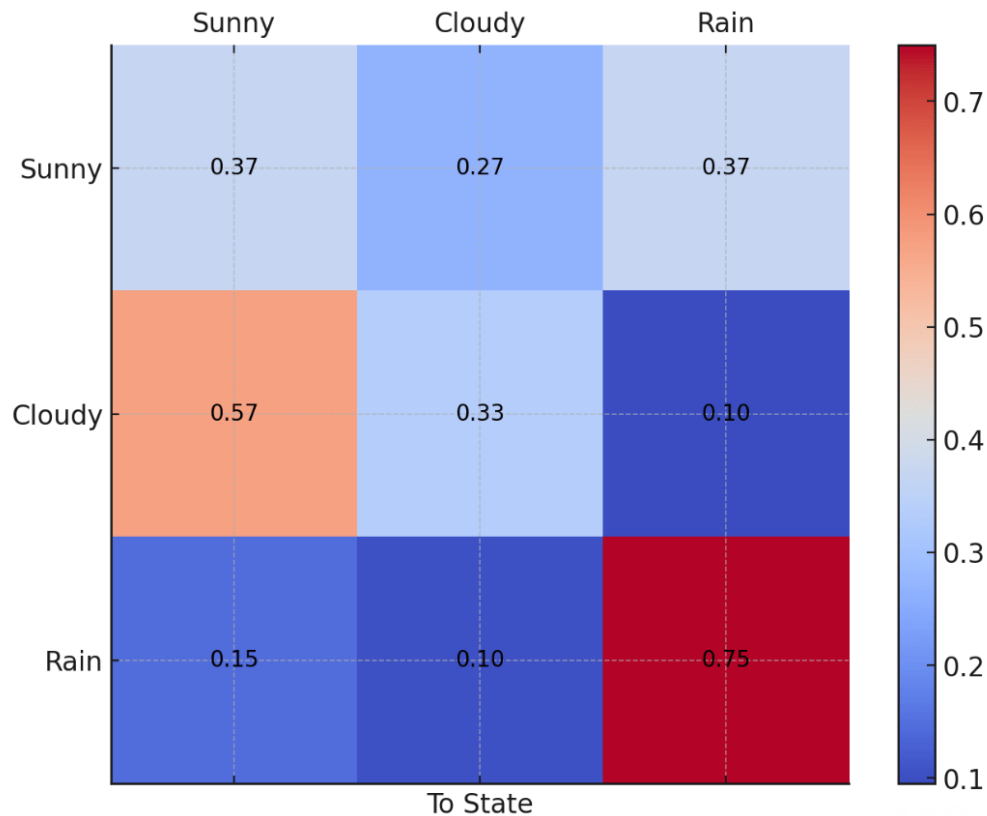
Tomorrow's  
weather



Transition  
probability

# We can represent these states and probabilities as a (cursed?) graph

Each edge represents the probability of transitioning from one state to the next



# Hidden Markov models also include additional information in "hidden states"

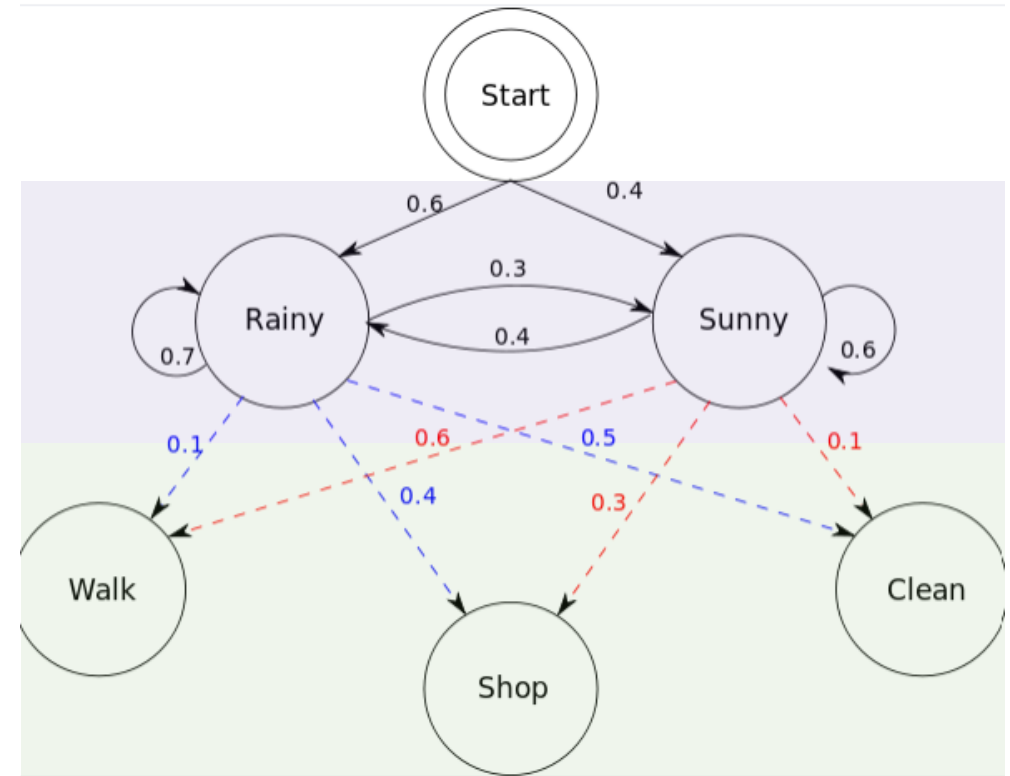
Suppose my friend lives in a remote location where it is either Rainy or Sunny

I cannot look up the weather but I have last year's weathers reports **Hidden states**

My friend can only tell me

- Walking
  - Shopping
  - Cleaning
- Observables**

We know how weather patterns transitions, but we don't have this information from our friend



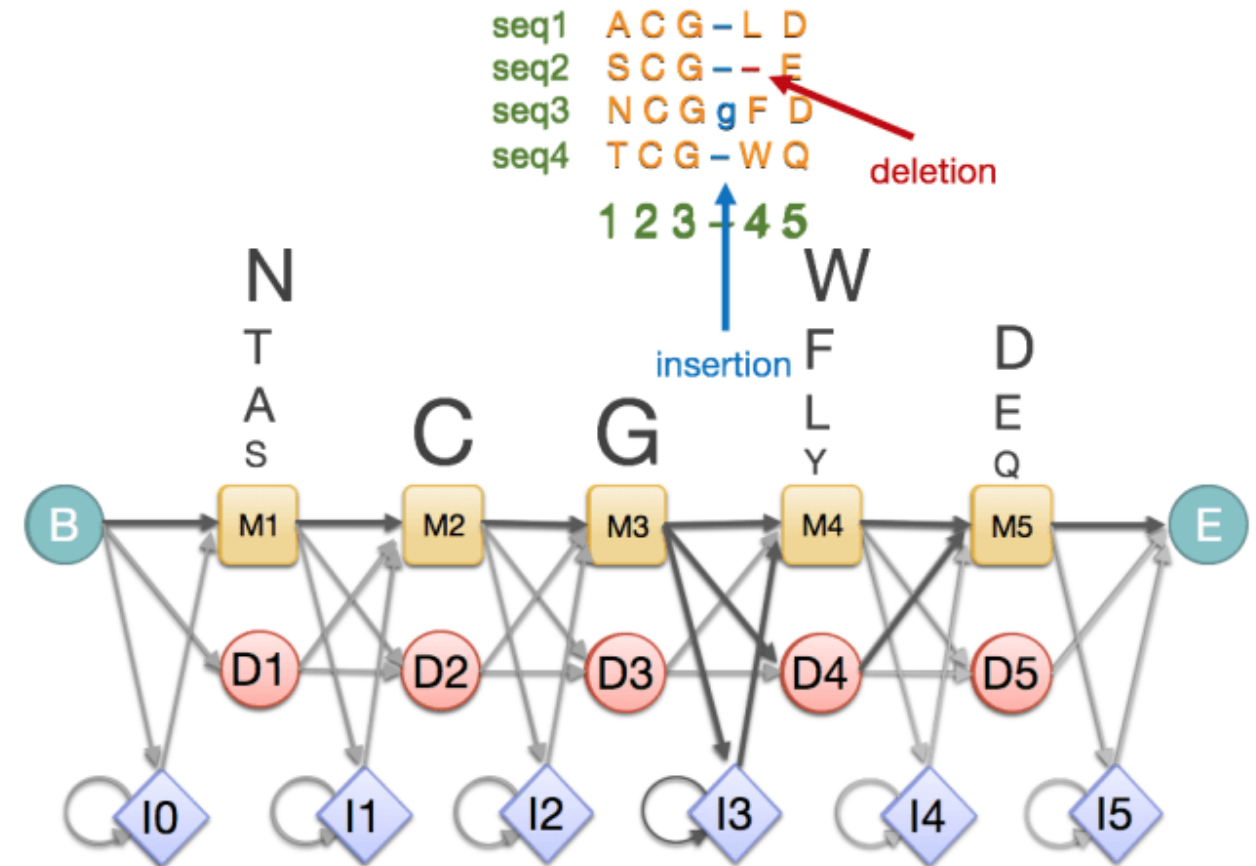
**Note:** If we had previous observable data, we could fit/learn transition probabilities of hidden states

# HMMs Model Protein Sequences as a Series of Probabilistic States

**Hidden states** represent the underlying biological events that are not directly observable

- **Match states:** conserved positions in the sequence
- **Insertion states:** positions where extra residues are added
- **Deletion states:** positions where residues are missing

**Observables** are the actual amino acids (residues) in the protein sequence that we can observe



# HMMER Uses HMMs to Search Protein Databases for Homology

HMMER is a tool that uses HMMs to search databases for sequences that match a given profile HMM

It is used to find homologous sequences, identifying evolutionary relationships across protein families



# SWISS-MODEL



```
MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKKLSTGHTL
VMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVVDVIHSIED
IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF
FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK
```

DHFR (UniProt)

The screenshot shows the SWISS-MODEL website. At the top, there is a navigation bar with links: Modelling, Repository, Tools, Documentation, Log in, and Create Account. The main header features the SWISS-MODEL logo, which consists of a red cross-like shape made of four circles, followed by the text "SWISS-MODEL" and a small red square icon. Below the header, a paragraph states: "is a fully automated protein structure homology-modelling server, accessible via the [ExPASy web server](#). The purpose of this server is to make protein modelling accessible to all life science researchers worldwide." A blue button labeled "Start Modelling" is positioned below this text. Further down, the "Repository" section is introduced with the text: "Every week we model all the sequences for thirteen core species based on the latest UniProtKB proteome. Is your protein already modelled and up to date in [SWISS-MODEL Repository](#)?" Below this is a search bar with the placeholder text "Search SWISS-MODEL Repository". At the bottom of the repository section, there is a horizontal row of twelve small, square images representing various biological specimens, including what appears to be a plant, a fish, a microorganism, and various cellular structures.

[swissmodel.expasy.org](https://swissmodel.expasy.org)



# SWISS-MODEL

All Projects

Untitled Project Created: today at 03:08

Summary Templates 50 Models Project Data

## Project Summary

Target 1 **MTLSILVAHDLQRVIGFENQLPWHLPNDLKHKVKKLSTGHTLVMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVDVIHSI** 80  
Target 1 **EDIYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTFFPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK** 159

## Template Results

A total of 685 templates were found to match the target sequence. This list was filtered by a heuristic down to 50. The top templates are:

Template	Sequence Identity	Biounit Oligo State	Description
6e4e.1	100.00	monomer	Dihydrofolate reductase Crystal structure of dihydrofolate reductase from Staphylococcus aureus MW2 bound to NADP and p218
6pr6.1	100.00	monomer	Dihydrofolate reductase S. aureus dihydrofolate reductase co-crystallized with para-tolyl-dihydrothalazine inhibitor and NADP(H)
3sqy.1	100.00	monomer	Dihydrofolate reductase S. aureus Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines
3fyw.1	100.00	monomer	Dihydrofolate reductase Staph. aureus DHFR complexed with NADPH and AR-101
6pr8.1	100.00	monomer	Dihydrofolate reductase S. aureus dihydrofoate reductase co-crystallized with 3,5-dimethylphenyl-dihydrothalazine inhibitor and NADP(H)

Show full template details

All Projects

Untitled Project Created: today at 03:08

Summary Templates 50 Models Project Data

## Template Results

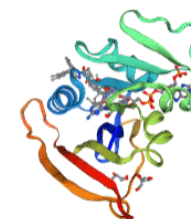
Templates Quaternary Structure Sequence Similarity Alignment

More

Sort	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
<input checked="" type="checkbox"/>	6e4e.1.A Dihydrofolate reductase Crystal structure of dihydrofolate reductase from Staphylococcus aureus MW2 bound to NADP and p218	0.99	-	100.00	X-ray, 1.9Å	monomer ✓	1 x NAP <sup>CG</sup> , 1 x MMV <sup>CG</sup>
<input checked="" type="checkbox"/>	6pr6.1.A Dihydrofolate reductase S. aureus dihydrofolate reductase co-crystallized with para-tolyl-dihydrothalazine inhibitor and NADP(H)	0.99	-	100.00	X-ray, 2.0Å	monomer ✓	1 x NAP <sup>CG</sup> , 1 x OWS <sup>CG</sup>
<input checked="" type="checkbox"/>	3sqy.1.A Dihydrofolate reductase S. aureus Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines	0.98	-	100.00	X-ray, 1.5Å	monomer ✓	1 x NAP <sup>CG</sup> , 1 x Q11 <sup>CG</sup>
<input checked="" type="checkbox"/>	3fyw.1.A Dihydrofolate reductase Staph. aureus DHFR complexed with NADPH and AR-101	0.98	-	100.00	X-ray, 2.1Å	monomer ✓	1 x NDP <sup>CG</sup> , 1 x XCF <sup>CG</sup>
<input checked="" type="checkbox"/>	6pr8.1.A Dihydrofolate reductase S. aureus dihydrofoate reductase co-crystallized with 3,5-dimethylphenyl-dihydrothalazine inhibitor and NADP(H)	0.98	-	100.00	X-ray, 2.0Å	monomer ✓	1 x OWJ <sup>CG</sup> , 1 x NAP <sup>CG</sup>
<input type="checkbox"/>	3s5.1.A Dihydrofolate reductase S. aureus Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines	0.98	-	100.00	X-ray, 1.7Å	monomer ✓	1 x NAP <sup>CG</sup> , 1 x Q12 <sup>CG</sup>
<input type="checkbox"/>	2w9g.1.A DIHYDROFOLATE REDUCTASE Wild-type Staphylococcus aureus DHFR in complex with NADPH and trimethoprim	0.98	-	100.00	X-ray, 2.0Å	monomer ✓	1 x TOP <sup>CG</sup> , 1 x NDP <sup>CG</sup>

Build Models 5

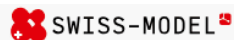
Clear Selection



Cartoon Camera Up Down Refresh

6e4e.1.A  
6pr6.1.A  
3sqy.1.A  
3fyw.1.A  
6pr8.1.A

# SWISS-MODEL



Modelling Repository Tools Documentation Log in Create Account

All Projects

Untitled Project Created: today at 03:08

Summary

Templates 50

Models

Project Data

## Template Results

Templates

Quaternary Structure

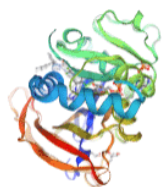
Sequence Similarity

Alignment

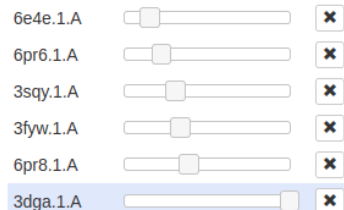
More

Build Models 6

Clear Selection



Cartoon



Template **3dga.1.A** Bifunctional dihydrofolate reductase-thymidylate synthase  
Wild-type *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase (PfDHFR-TS) complexed with RJF01302,



Modelling Repository Tools Documentation Log in Create Account

All Projects

Untitled Project Created: today at 03:08

Summary

Templates 50

Models

Project Data

## Template Results

Templates

Quaternary Structure

Sequence Similarity

Alignment

More

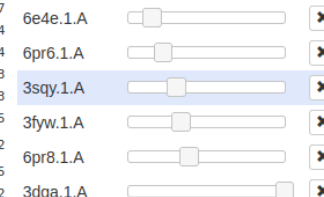
Build Models 6

Clear Selection



Cartoon

Target	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	37
6e4e.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	60
6pr6.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	37
3sqy.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	37
3fyw.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	36
6pr8.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	36
3dga.1.A	MTLSILVAHDLQRVIGFENQLPWHL	PNDLKHVKKLS	75
Target	GHTLVMGRKTFESIG	KPLNRRNV	61
6e4e.1.A	GHTLVMGRKTFESIG	KPLNRRNV	84
6pr6.1.A	GHTLVMGRKTFESIG	KPLNRRNV	61
3sqy.1.A	GHTLVMGRKTFESIG	KPLNRRNV	61
3fyw.1.A	GHTLVMGRKTFESIG	KPLNRRNV	60
6pr8.1.A	GHTLVMGRKTFESIG	KPLNRRNV	60
3dga.1.A	GHTLVMGRKTFESIG	KPLNRRNV	125
Target	VLTSDTSFN	VEGVDVIHSIEDIYQL	184
6e4e.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	127
6pr6.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	184
3sqy.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	184
3fyw.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	103
6pr8.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	103
3dga.1.A	VLTSDTSFN	VEGVDVIHSIEDIYQL	175
Target	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	152
6e4e.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	175
6pr6.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	152
3sqy.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	152
3fyw.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	151
6pr8.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	151
3dga.1.A	KVDDMYITVIEGKFRGDT	FFPPYTFEDWEVASSVEGKLDEKNTIPHT	223
Target	LHLIRKK		159
6e4e.1.A	LHLIRKK		182
6pr6.1.A	LHLIRKK		159
3sqy.1.A	LHLIRKK		159
3fyw.1.A	LHLIRKK		158
6pr8.1.A	LHLIRKK		158
3dga.1.A	LHLIRKK		230



# SWISS-MODEL

All Projects

Untitled Project Created: today at 03:08

Summary

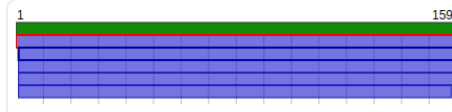
Templates 50

Models 5

Project Data

Model Results

Order by: GMQE



Model 01



Structure Assessment

Compare

Download files

Display files

Oligo-State

Monomer

GMQE

0.96

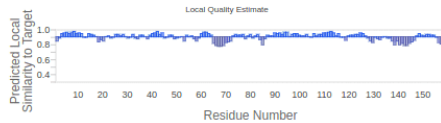
QMEANDisCo Global:

0.91 ± 0.07

Ligands

1 x MMV, 1 x NAP

QMEANDisCo Local



QMEAN Z-Scores

Template

6e4e.1.A Dihydrofolate reductase

Crystal structure of dihydrofolate reductase from Staphylococcus aureus MW2 bound to NADP and p218

Seq Identity

100.00%

Coverage

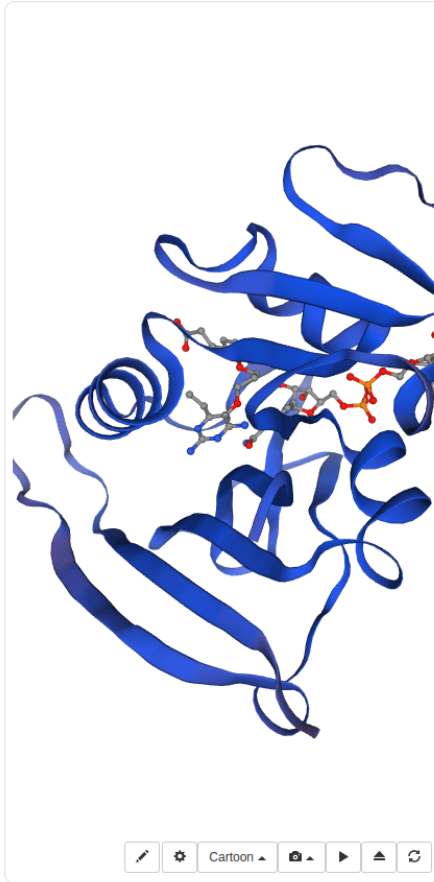
Model-Template Alignment

Model 04

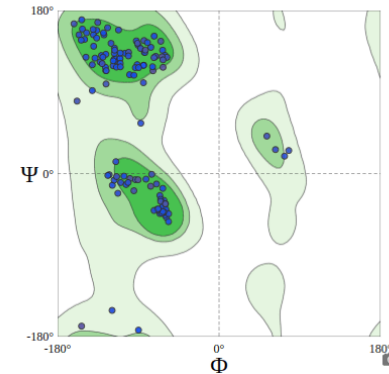


Structure Assessment

Compare



Ramachandran Plots



General Glycine Proline Pre-Proline A

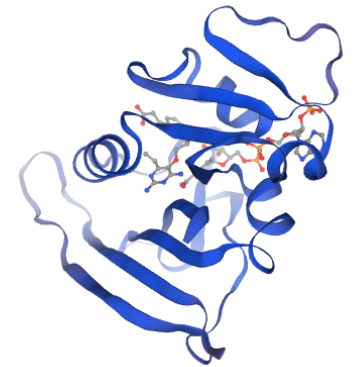
MolProbity Results

MolProbity Score	0.93
<input type="checkbox"/> Clash Score	1.11
(A97 THR_1 NAP)	
Ramachandran Favoured	97.45%
Ramachandran Outliers	0.00%
Rotamer Outliers	0.00%
C-Beta Deviations	0
Bad Bonds	0 / 1323
<input type="checkbox"/> Bad Angles	13 / 1793
(A148 ILE-A149 PRO), A31 HIS, A89 HIS, A39 HIS, A24 HIS, (A125 PRO-A126 PRO), A154 HIS, A9 HIS, A143 ASP, (A21 LEU-A22 PRO), A10 ASP, (A124 PHE-A125 PRO), A129 PHE	
<input type="checkbox"/> Cis Non-Proline	1 / 150

QMEANDisCo

QMEANDisCo Global: 0.91 ± 0.07

Coordinates with QMEAN local scores in the B-factor column



# What happens with a novel protein?



MGKKEVILLFLAVIFVALNTLVAVYFRETAEQVVYGK  
NNINQKLIQLKDGTYGFEPALPHVGTFKVLDSNRVPQIA  
QEII RNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN  
NLHGYEGFLFKDNVPLTYPQEFILSNLDGKVRSLQNYDY  
DLDVLFGEKEEVKSEILRGLYNTYTRAFSPYKL

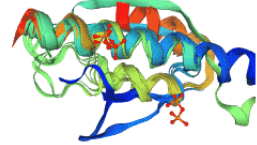
Novel protein  
(ChatGPT)

Templates   Quaternary Structure   Sequence Similarity   Alignment

More ▾

Build Models 13

Clear Selection



Target: MGKKEVILLFLAVIFVALNTLVAVYFRETAEQVVYGKNNINQKLIQLK 50

7cr6.1.B  
7cr6.1.D  
2vky.1.A  
7cr8.1.A  
7cr8.1.D  
7cr8.1.B  
7cr6.1.A  
7cr6.1.C  
7cr8.1.C  
4jg4.1.A  
6d1r.1.A  
1d6t.1.A  
1a6f.1.A

Target: DGTYGFEALPHVGTFKVLDSNRVPQIAQEII RNKVKRYLQEAVRIEGTY 100

7cr6.1.B  
7cr6.1.D  
2vky.1.A  
7cr8.1.A  
7cr8.1.D  
7cr8.1.B  
7cr6.1.A  
7cr6.1.C  
7cr8.1.C  
4jg4.1.A  
6d1r.1.A  
1d6t.1.A  
1a6f.1.A

Target: PIVDGLVNAKYTVANPNNLHGY 145

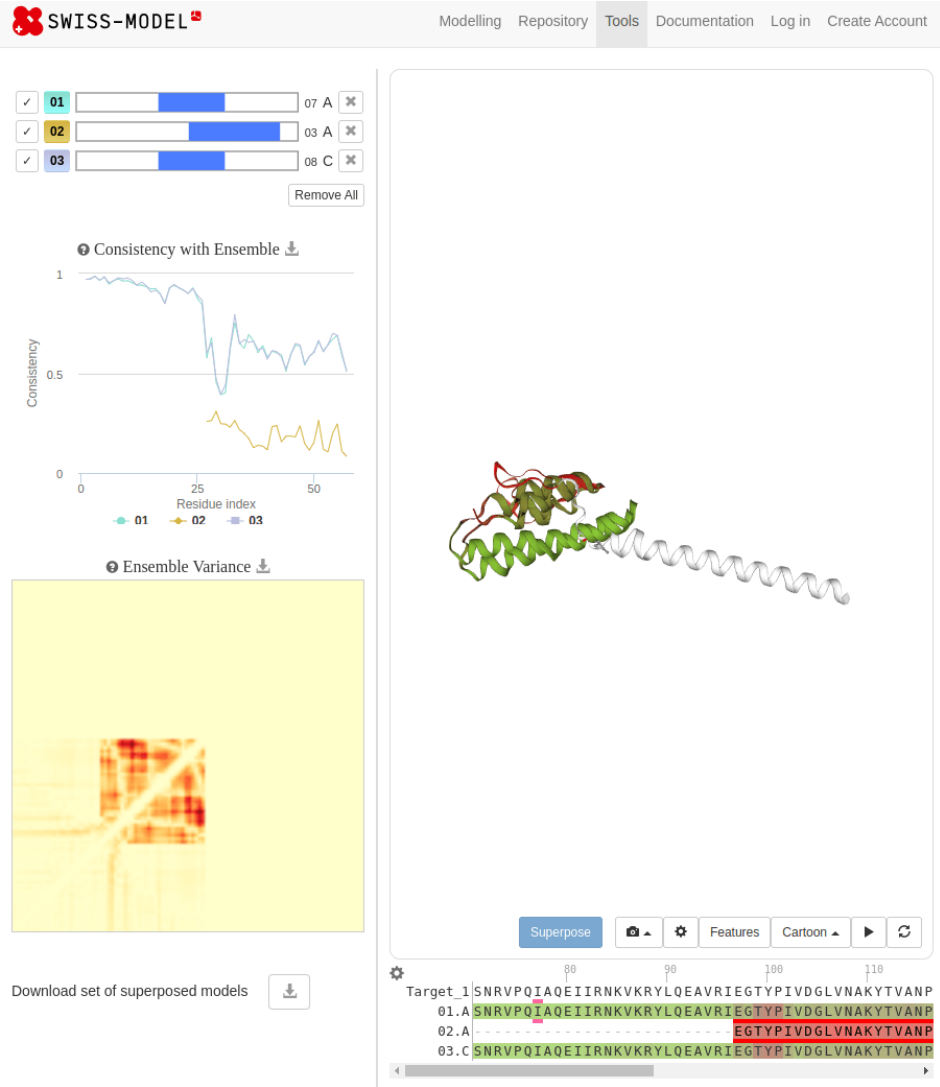
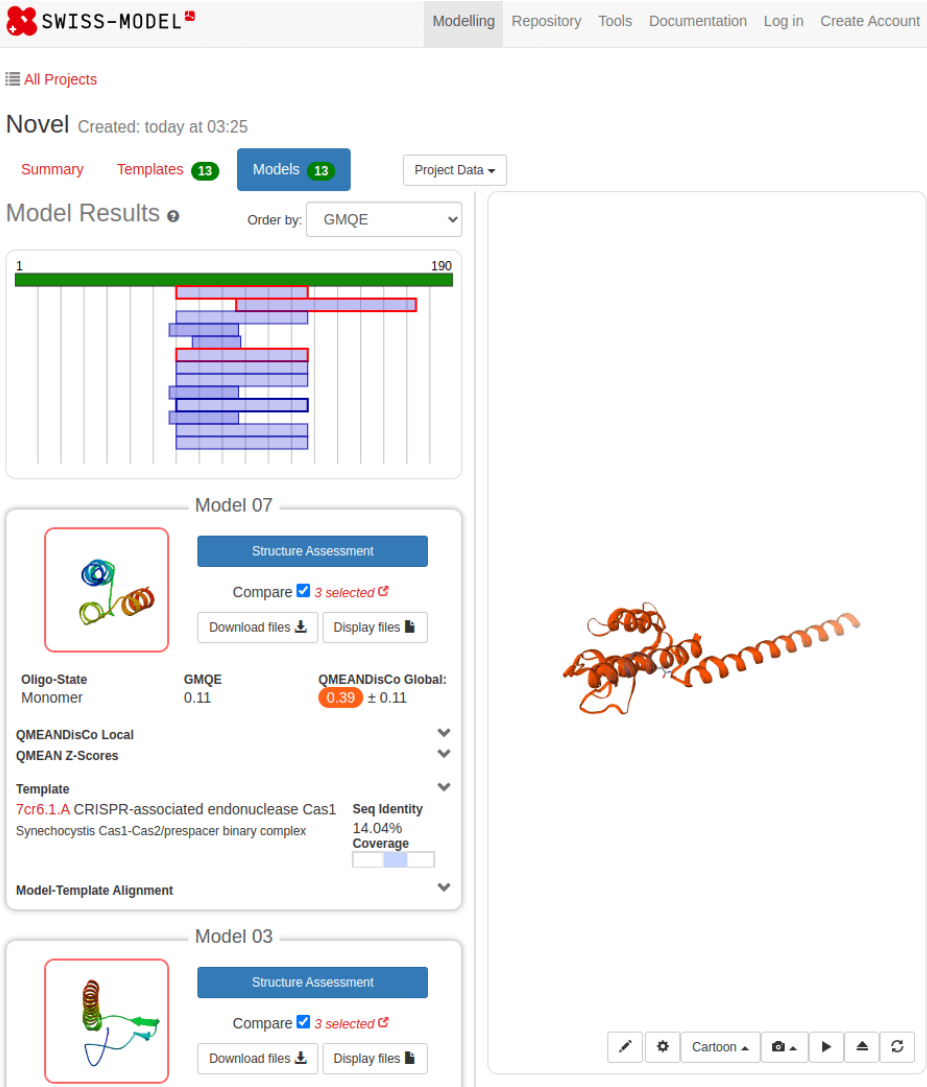
7cr6.1.B  
7cr6.1.D  
2vky.1.A  
7cr8.1.A  
7cr8.1.D  
7cr8.1.B  
7cr6.1.A  
7cr6.1.C  
7cr8.1.C  
4jg4.1.A  
6d1r.1.A  
1d6t.1.A  
1a6f.1.A

Target: GKVRSLQNYDYDLDVLFGEKEEVKSEILRGLYNTYTRAFSPYKL 190

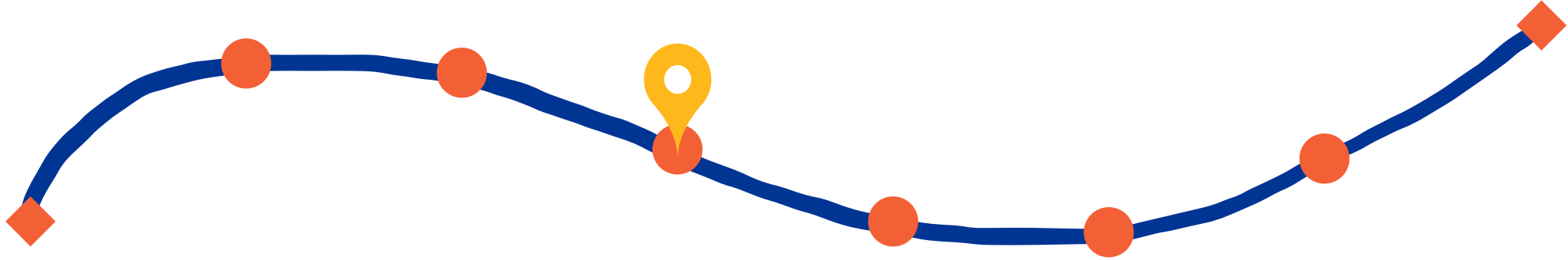
7cr6.1.B  
7cr6.1.D  
2vky.1.A  
7cr8.1.A  
7cr8.1.D  
7cr8.1.B

7cr6.1.B  
7cr6.1.D  
2vky.1.A  
7cr8.1.A  
7cr8.1.D  
7cr8.1.B  
7cr6.1.A  
7cr6.1.C  
7cr8.1.C  
4jg4.1.A  
6d1r.1.A  
1d6t.1.A  
1a6f.1.A

# Novel proteins are too challenging



# After today, you should be able to



Know when to use threading  
instead of homology modeling



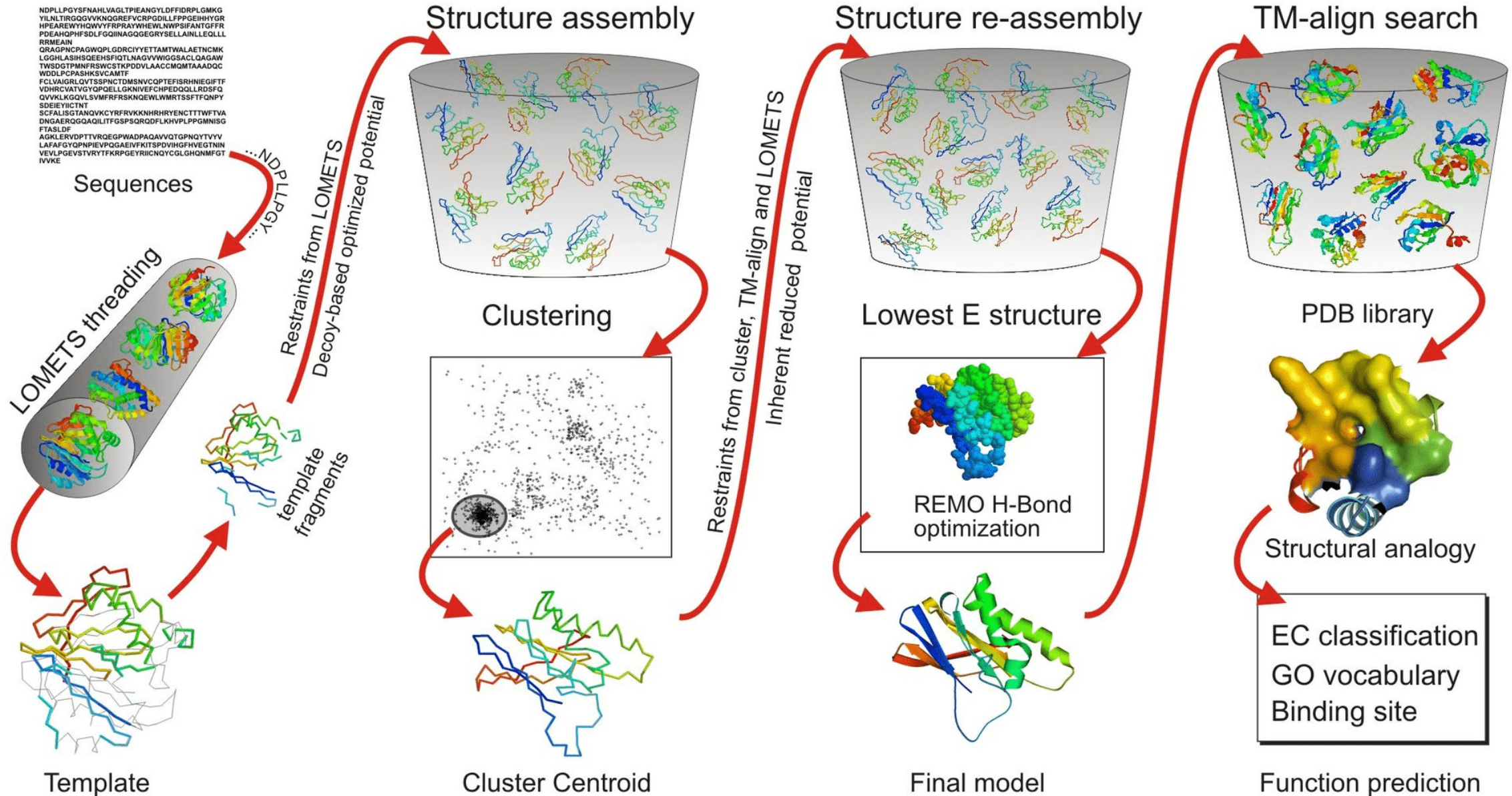
# Why Use Threading?

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

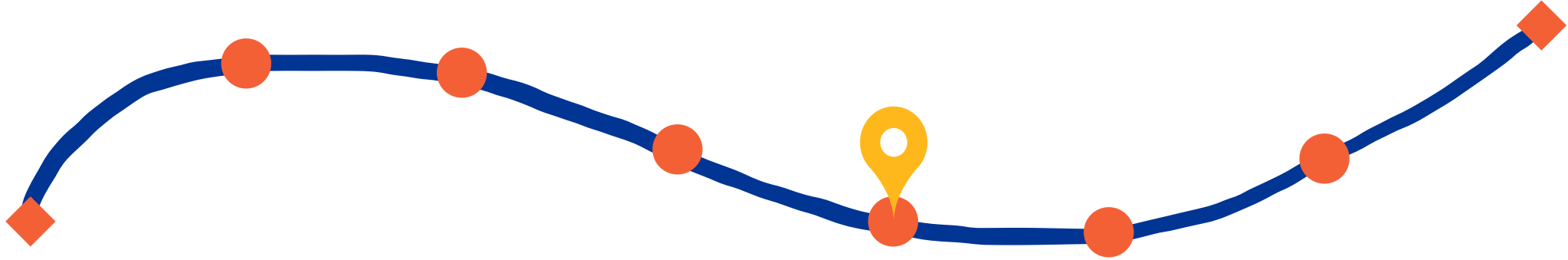
Threading matches sequences to known structural folds based on structural rather than sequence similarity

**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 be able to



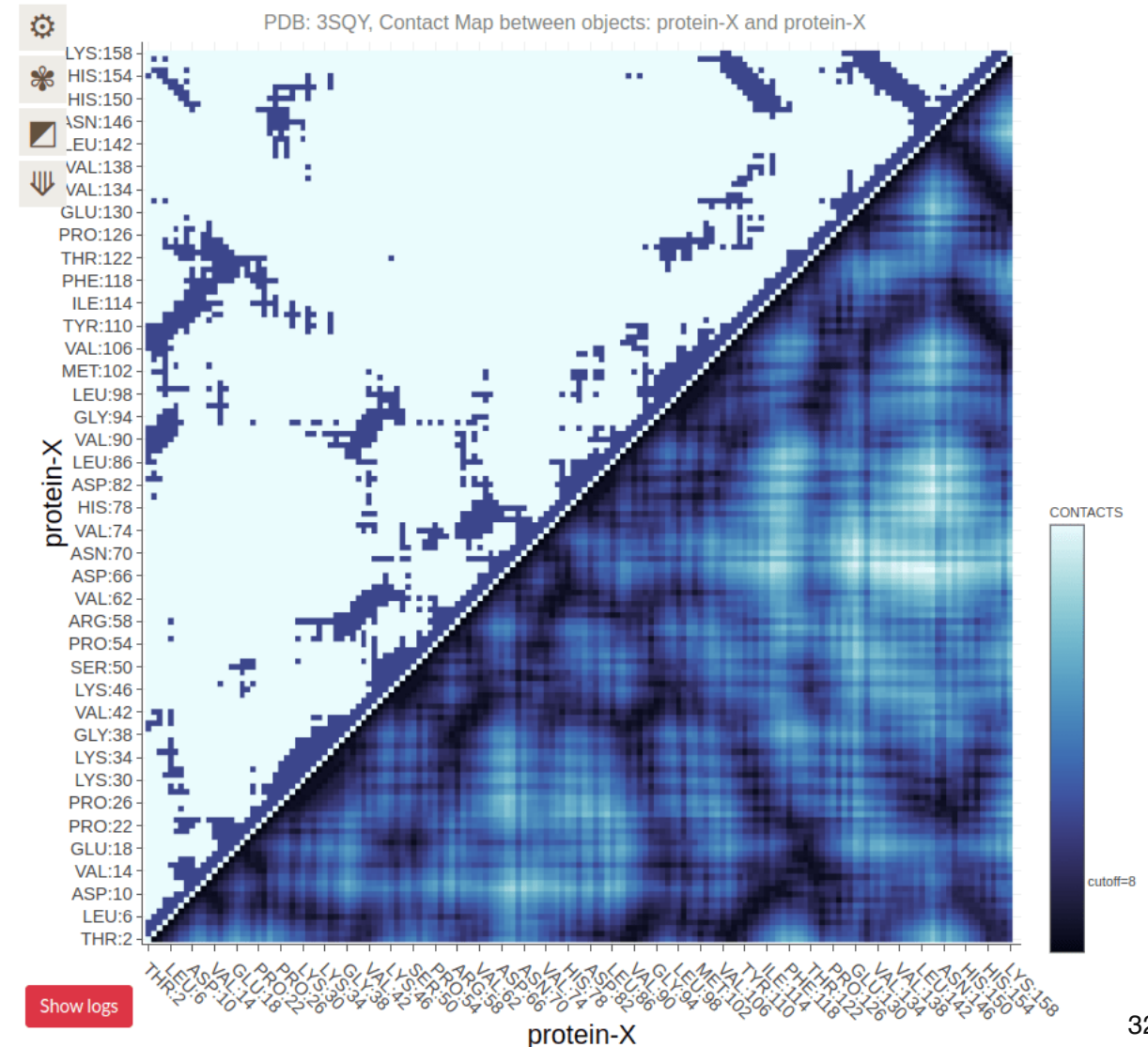
Interpret a contact map for protein  
structures

# 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](http://mapiya.lcbio.pl)

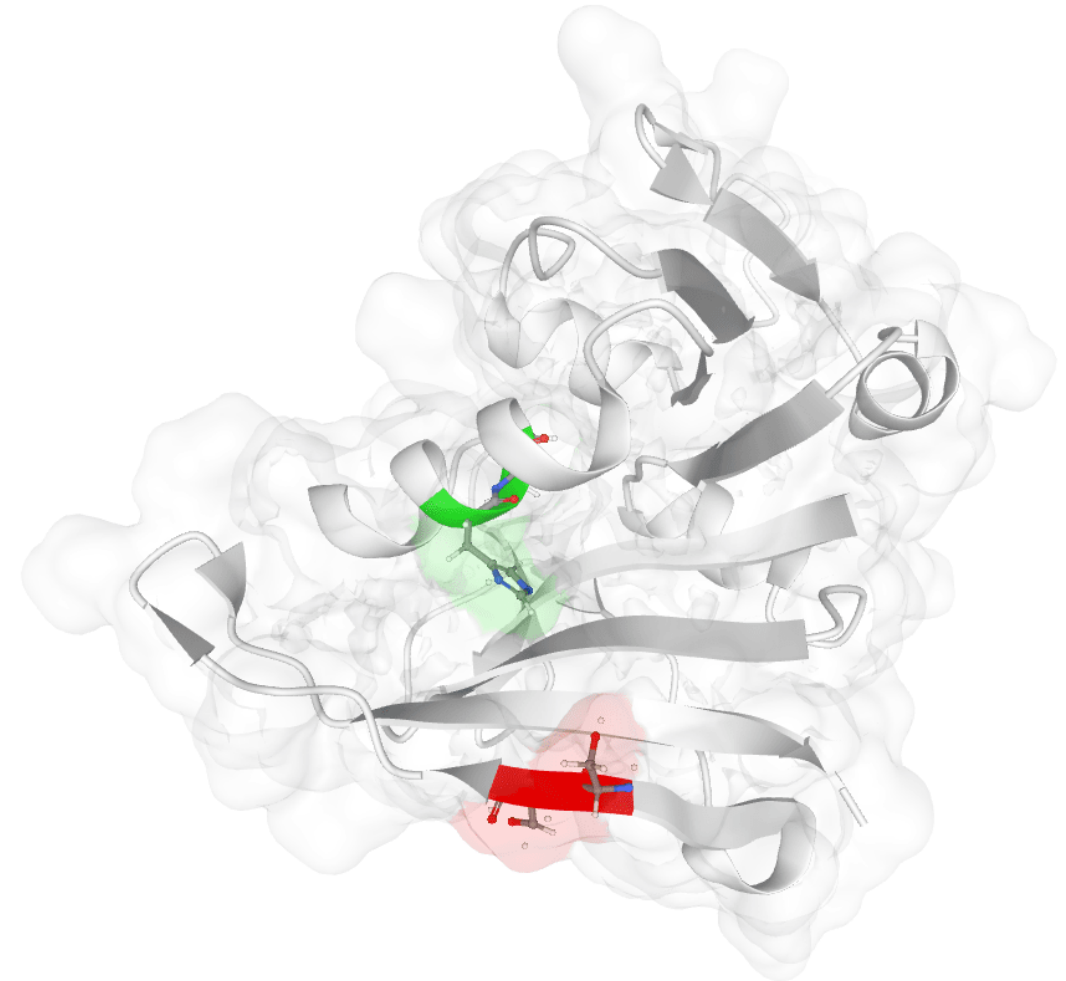
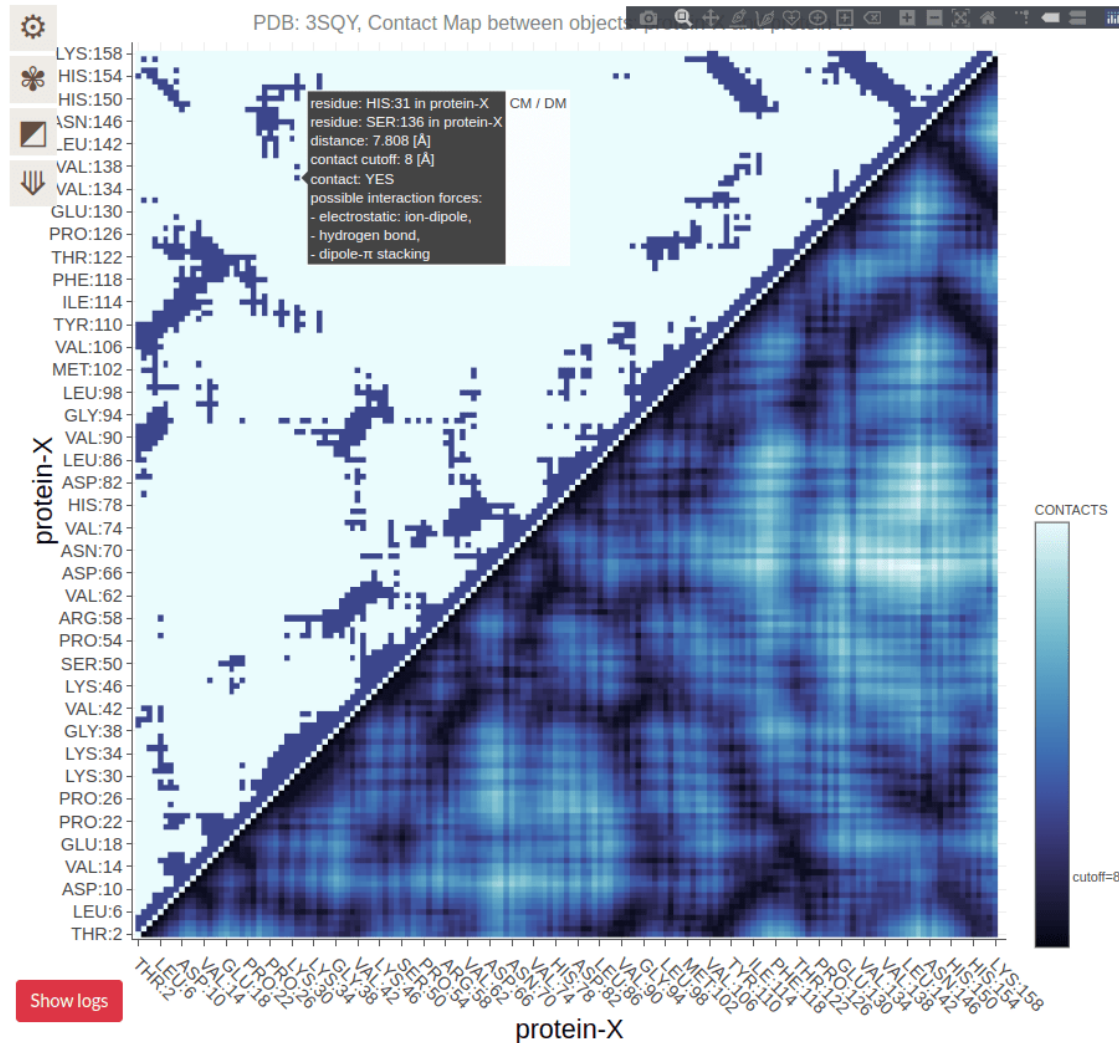




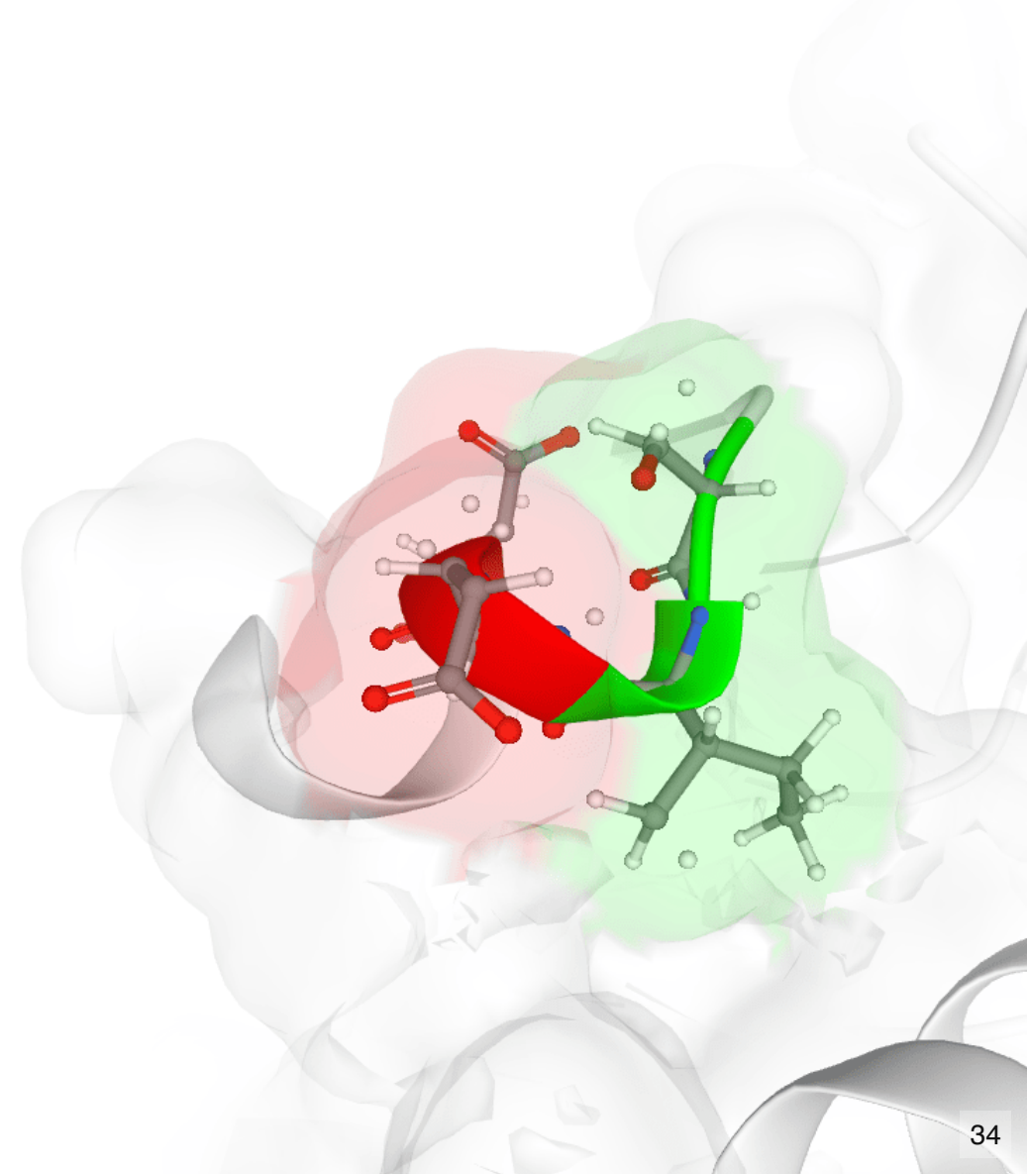
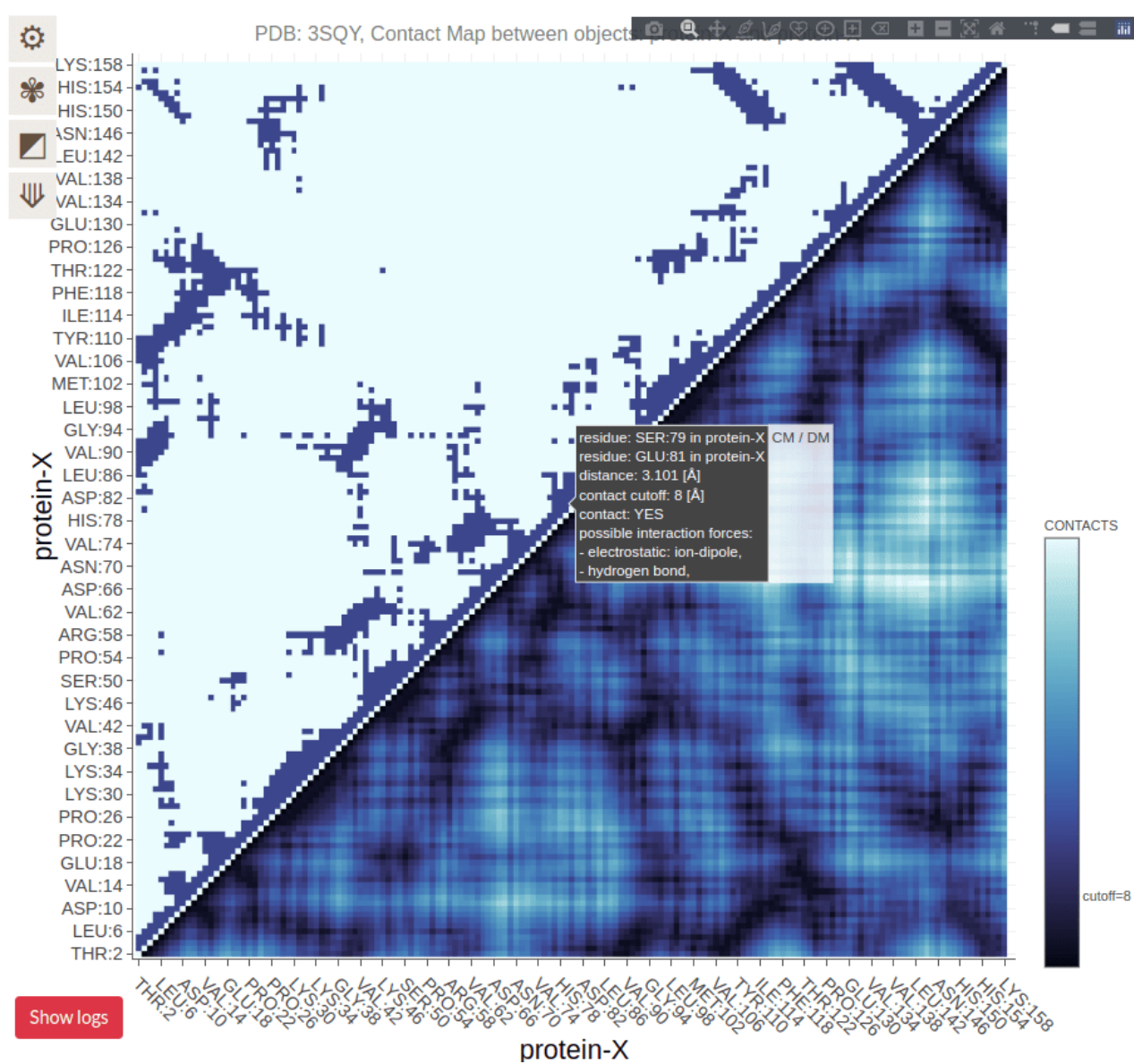
# Contact Maps Represent Spatial Proximity, Not Sequence Order

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

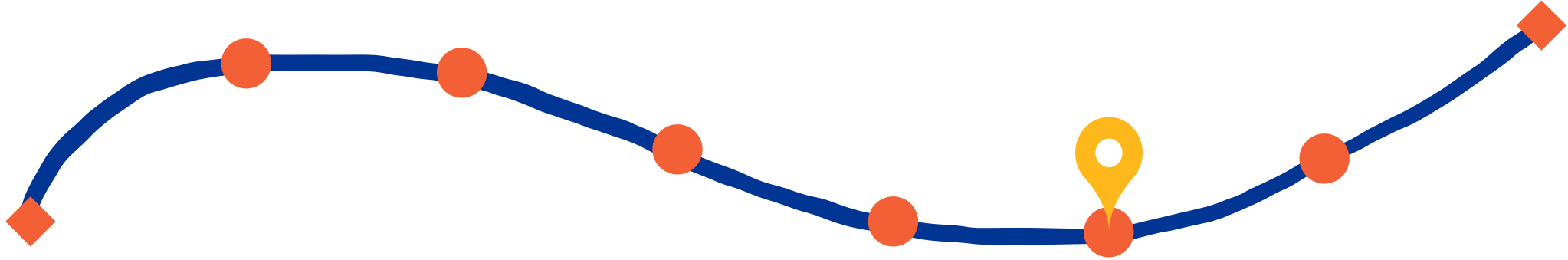
Residues far apart in the sequence can still be close in the 3D structure, reflected in the contact map



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



# After today, you should 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



Arg (positive)  
Lys (positive)  
Trp (hydrophobic)



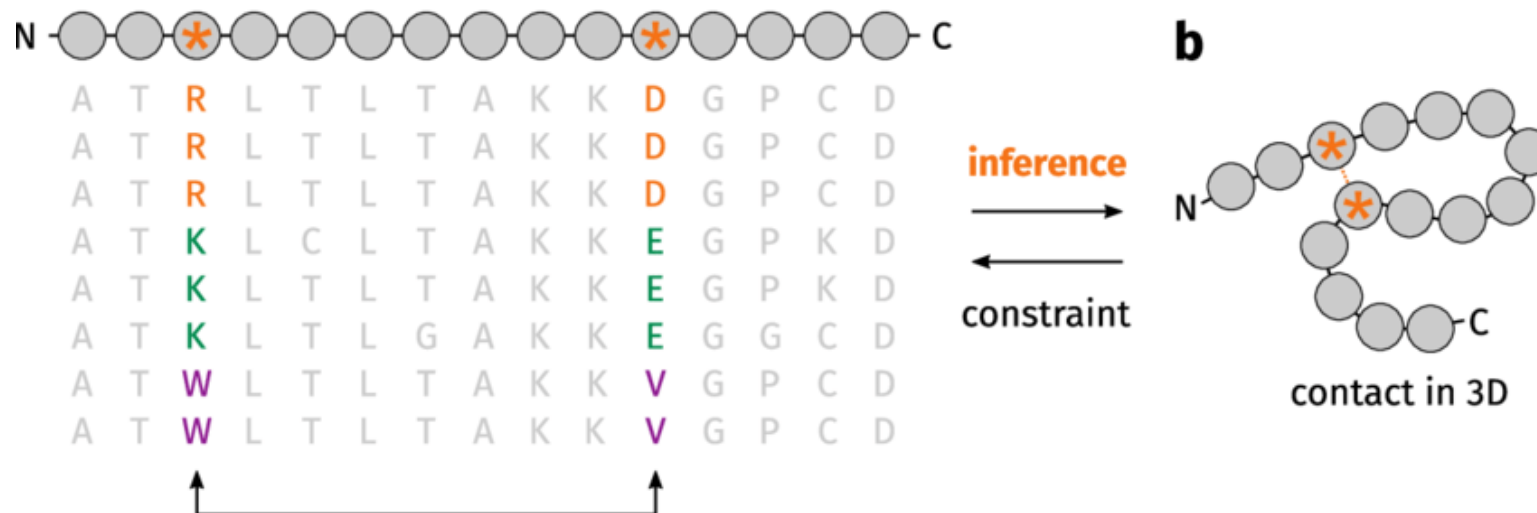
Asp (negative)  
Glu (negative)  
Val (hydrophobic)

# 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

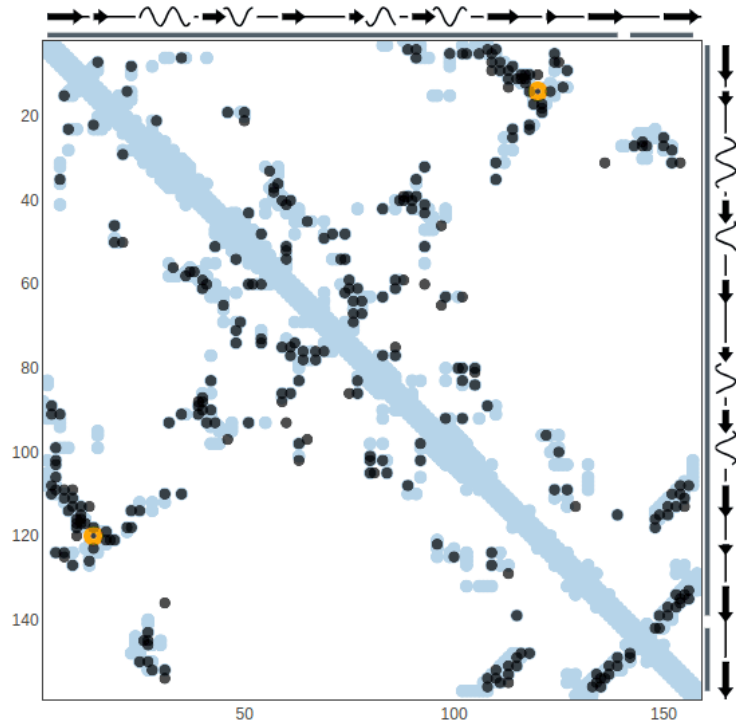
Evolutionary information from MSAs guides predictions for residue-residue contacts



# Coevolution example: DHFR

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

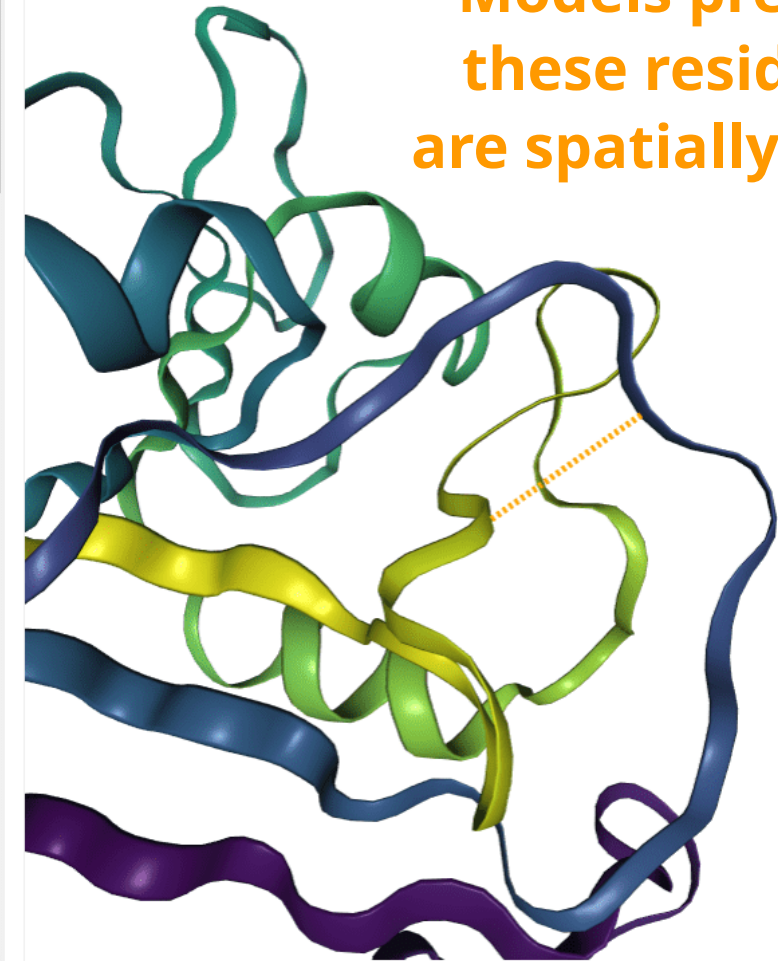
**Val14 and Gly120 coevolved**



evcouplings.org

	i	A <sub>i</sub>	j	A <sub>j</sub>	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

**Models predict these residues are spatially close**





# 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.

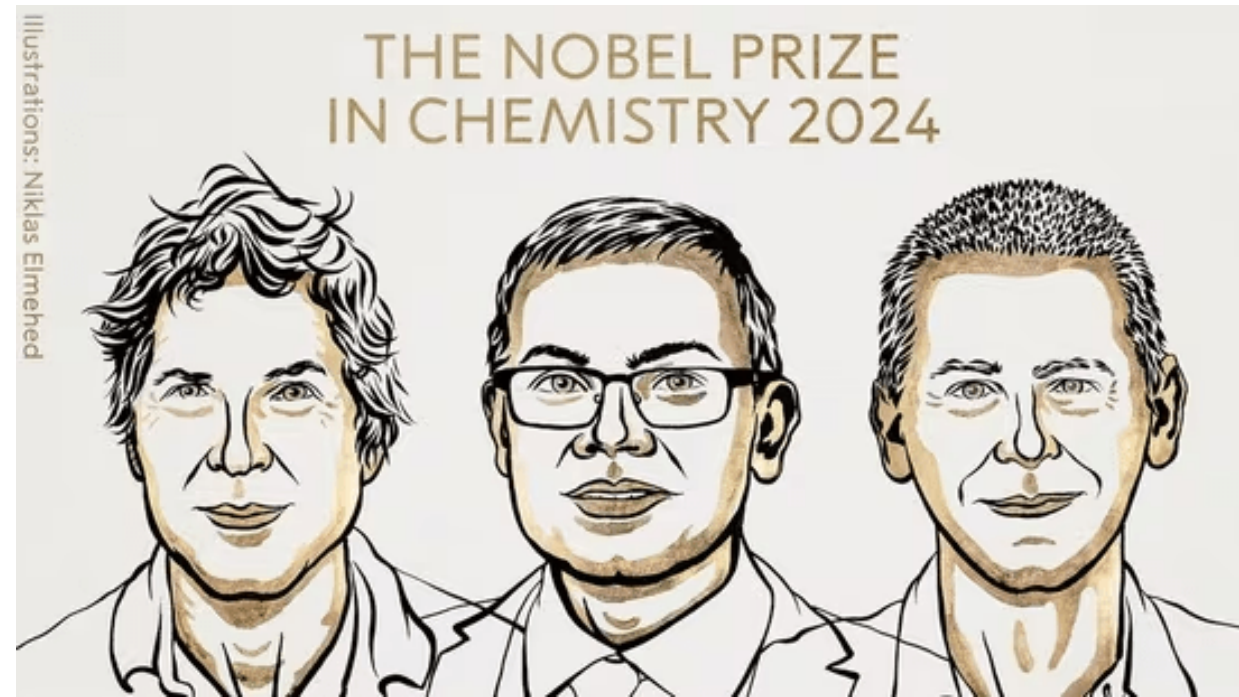




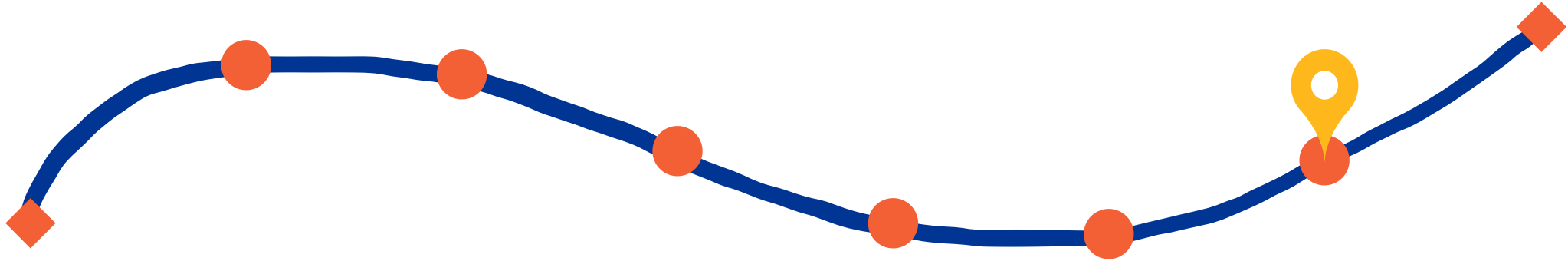
# Machine learning leverages coevolution for high-accuracy predictions

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 be able to



Explain why ML models are dominate  
protein structure prediction

# AlphaFold pipeline, simplified

Given the following data

Input  
sequence



```
MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKK  
LSTGHTLVMGRKTFESIGKPLPNRRNVVLTSDTS  
FNVEGVVDVIHSIEDIYQLPGHVFIFGGQTLFEEM  
IDKVDDMYITVIEGKFRGDTFFPPYTFEDWEVAS  
SVEGKLDEKNTIPHTFLHLIRKK
```

Multiple  
Sequence  
Alignment

1	TARGET1-159	KKLSTGRLLVMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGV
2	UniRef90_A0A238ADC8/2-178	RSITAGGGVINGRTTFDSIPRPLQGRINNVLTSSADLMQW
3	UniRef90_A0A238ADC8/38-508	RSITAGGGVINGRTTFDSIPRPLQGRINNVLTSSADLMQW
4	UniRef90_A0A238ADC8/906-1076	RSITAGGGVINGRTTFDSIPRPLQGRINNVLTSSADLMQW
5	UniRef90_A0A2U1P121/436-611	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
6	UniRef90_A0A2U1P121/715-890	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
7	UniRef90_A0A2JNE268-185	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
8	UniRef90_A0A2JNE268/184-373	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
9	UniRef90_UPI0022818839/37-217	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
10	UniRef90_UPI0022818839/246-424	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
11	UniRef90_UPI001457FF826-184	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
12	UniRef90_UPI001457FF82192-372	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
13	UniRef90_UPI00234EC90F/4-183	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
14	UniRef90_UPI00234EC90F/225-406	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
15	UniRef90_A0A8B72N/910-182	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
16	UniRef90_A0A8B72N/917-376	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
17	UniRef90_R7UK12/6-185	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
18	UniRef90_R7UK12/195-373	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
19	UniRef90_T1G9P05-182	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
20	UniRef90_T1G9P05-188-374	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
21	UniRef90_A0A7J7G58/7-183	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
22	UniRef90_A0A7J7G58/191-368	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
23	UniRef90_A0A210P44/7-170	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
24	UniRef90_A0A210P44/180-361	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
25	UniRef90_A0A9N7VKK02/13-190	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
26	UniRef90_A0A9N7VKK02/214-371	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
27	UniRef90_A0A8B6L24/22-159	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
28	UniRef90_A0A8B6L24/167-346	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
29	UniRef90_A0A932P4F/3125-309	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
30	UniRef90_A0A940DWL2/2-156	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
31	UniRef90_H8KUH06-166	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW
32	UniRef90_A0A2D5XDZ31-161	KKLTKNAVINGRKTWSIPRPLPDRINNVLTSSADLMQW



ML  
models

Predict

DHFR

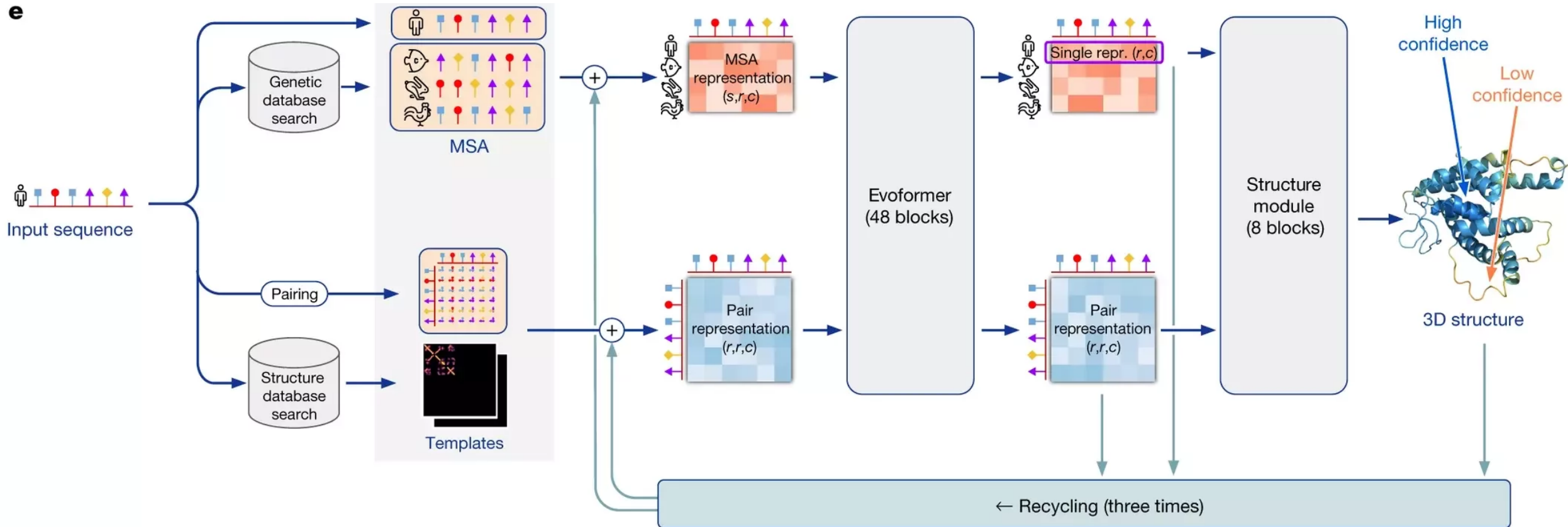
[← Back](#) [Download](#) [Clone and reuse](#) [Feedback on structure](#)



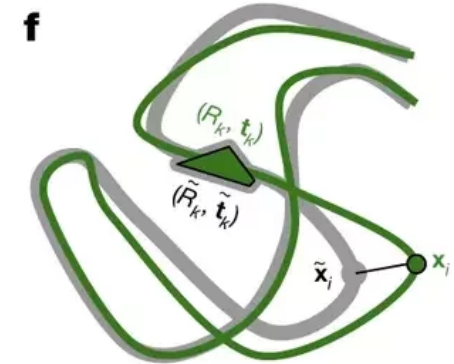
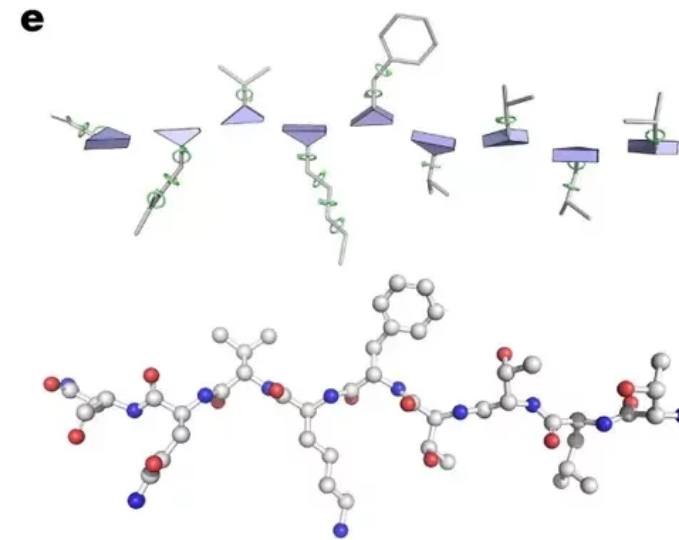
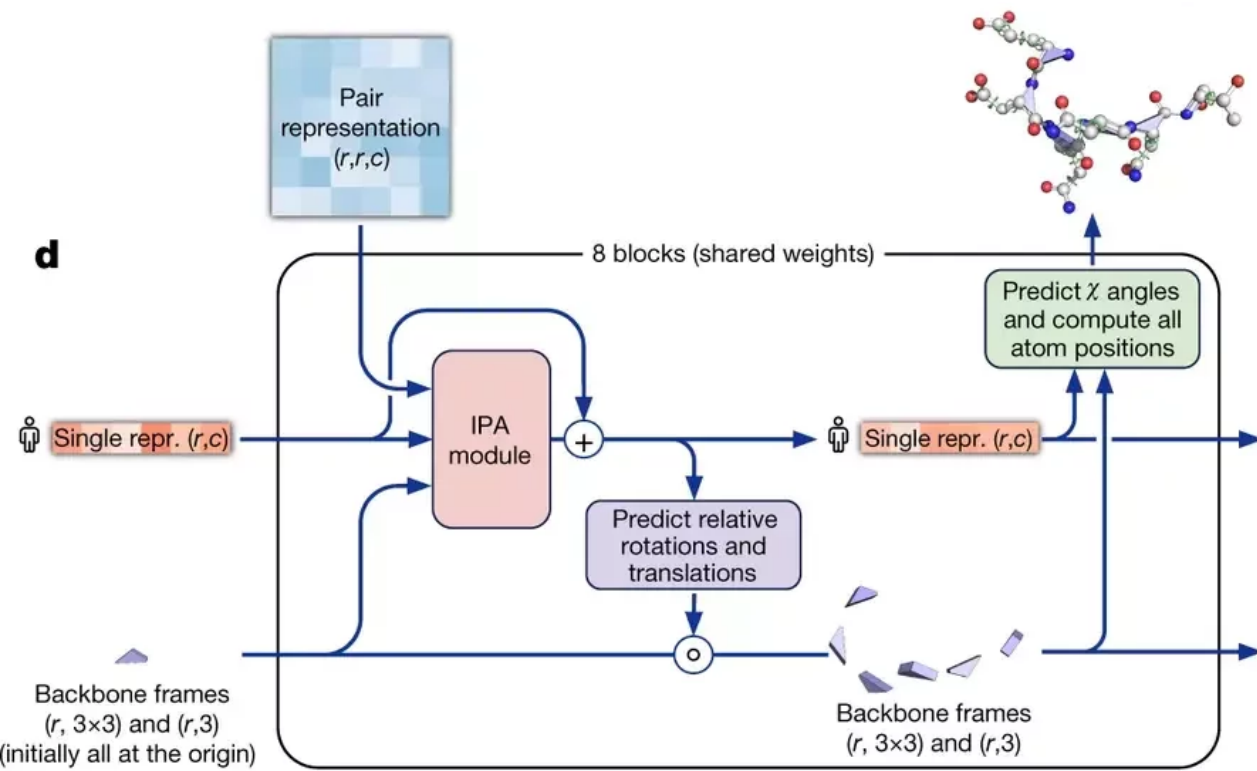
Atomistic  
structure

# AlphaFold 2 pipeline: Evoformer

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



# Contact maps are converted into dihedral angles





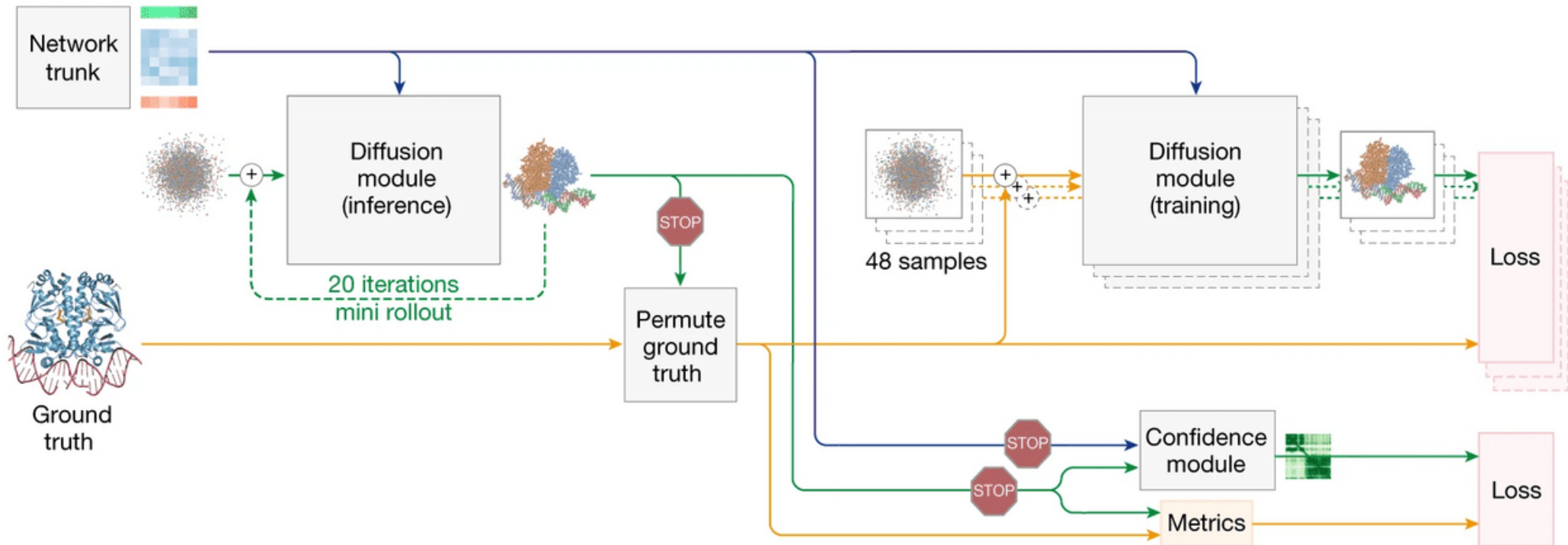
Recycling iteration 0, block 01  
Secondary structure assigned from the final prediction



# 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**

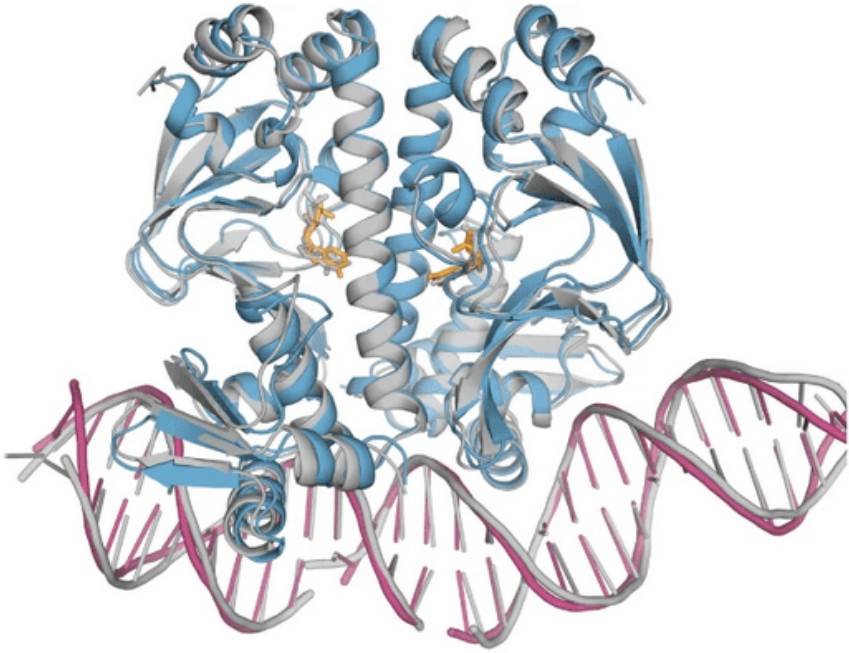




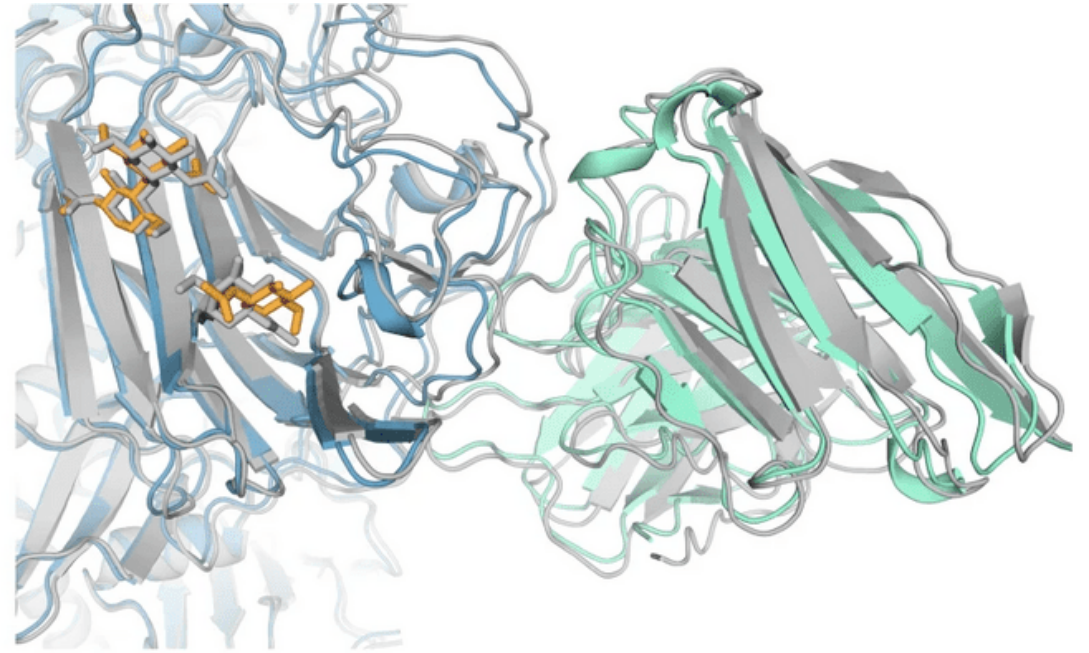
# AlphaFold 3 is supercharged for any biomolecule

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

**a**



**b**



# AlphaFold 3

AlphaFold Server BETA



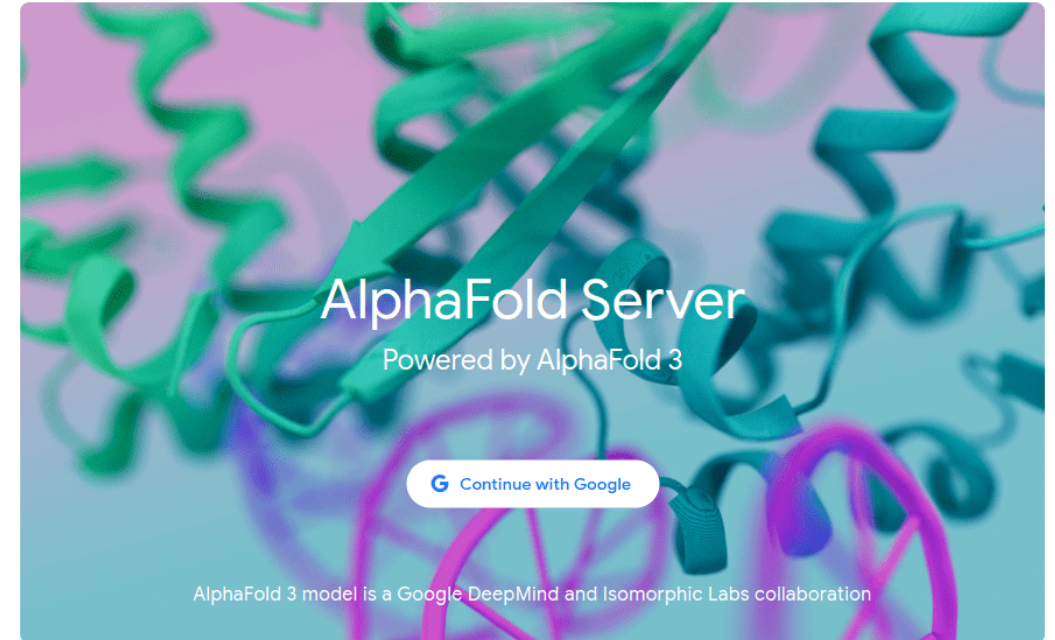
```
MTLSILVAHDLQRVIGFENQLPWHL PNDLKHVKKLSTGHTL
VMGRKTFESIGKPLPNRRNVVLTS DTSFNVEGV DVIHSIED
IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF
FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK
```

DHFR (UniProt)



```
MGKKEVILLFLAVIFVALNTLVVAVYFRETAEQVVYGK
NNINQKLIQLKDGT YGFEPALPHVGTFKVLDSNRVPQIA
QEII RNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN
NLHGYEGFLFKDNVPLTYPQEFILSNLDGKVRSLQNYDY
DL DVLFG EKEEVKSEILRGLYYNTYTRAFSPYKL
```

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](https://alphafoldserver.com)

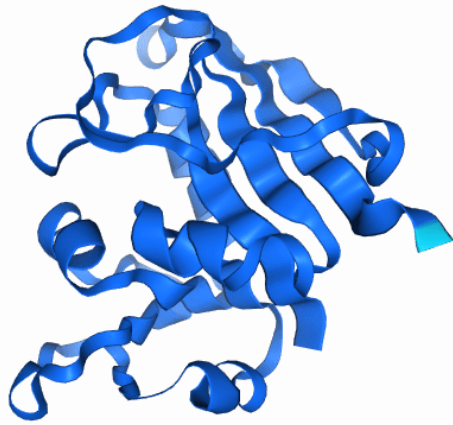
# AlphaFold 3 is a breakthrough, not the final solution

## DHFR

[← Back](#) [Download](#) [Clone and reuse](#) [Feedback on structure](#)

Very high (pLDDT > 90) Confident (90 > pLDDT > 70) Low (70 > pLDDT > 50) Very low (pLDDT < 50)

ipTM = - pTM = 0.95 [learn more](#)

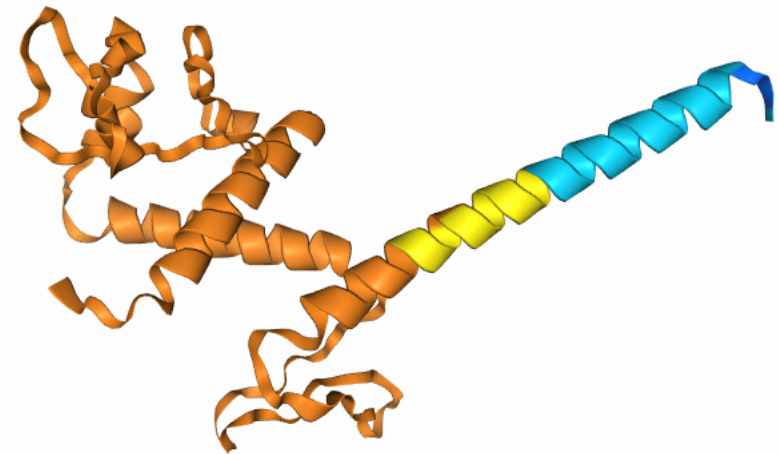


## Novel

[← Back](#) [Download](#) [Clone and reuse](#) [Feedback on structure](#)

Very high (pLDDT > 90) Confident (90 > pLDDT > 70) Low (70 > pLDDT > 50) Very low (pLDDT < 50)

ipTM = - pTM = 0.2 [learn more](#)



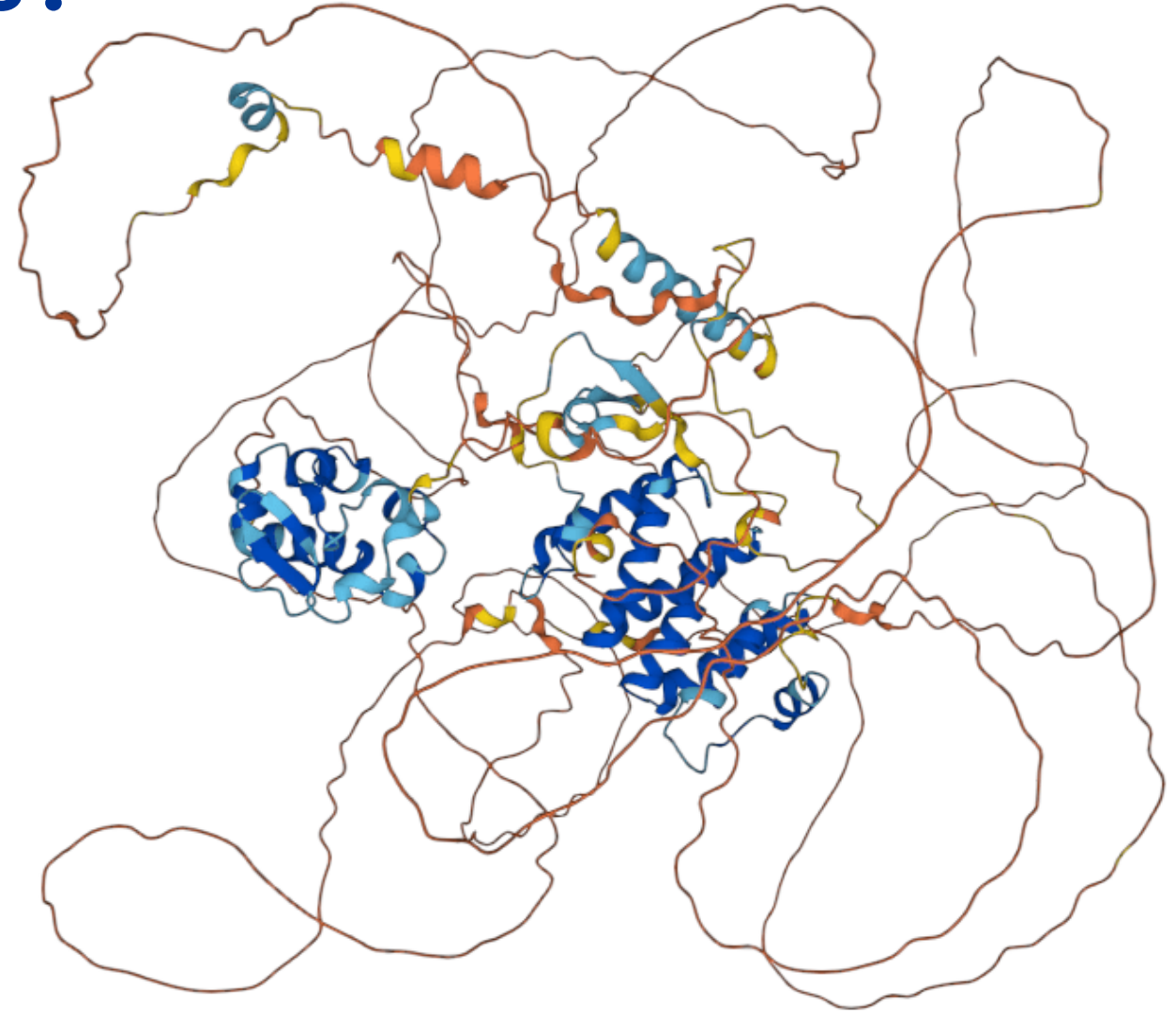
# Caveat: Proteins are **dynamic**

[https://www.youtube.com/embed/AjcUmxT-QEA?si=qupqTpuV5lvOB\\_ut&start=43&enablejsapi=1](https://www.youtube.com/embed/AjcUmxT-QEA?si=qupqTpuV5lvOB_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 12:

Protein structure prediction

## Lecture 13:

Molecular simulation principles



Today



Thursday

- Work on A05
- Review material