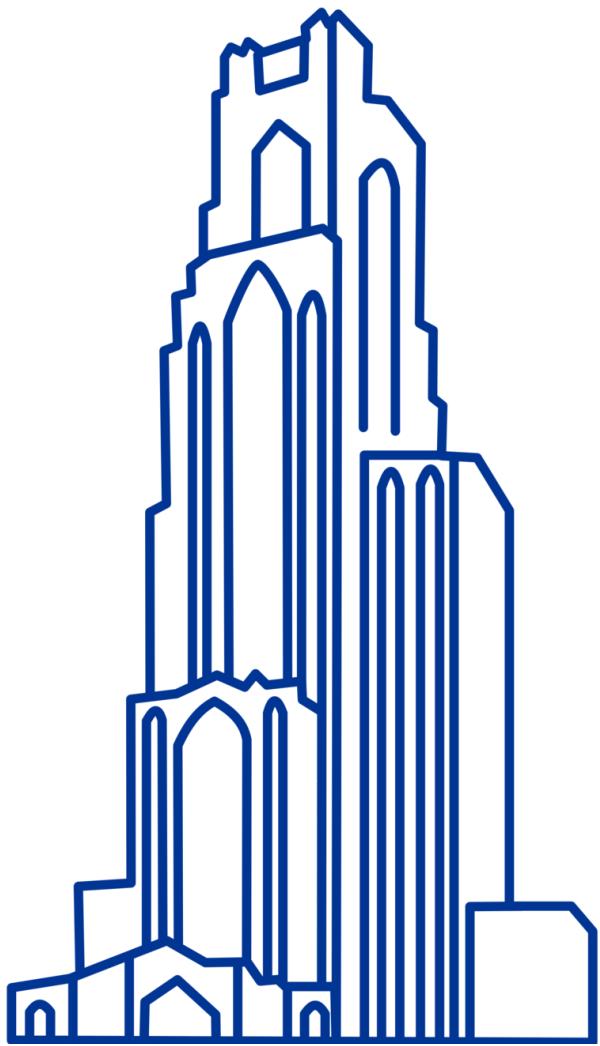


Computational Biology

(BIOSC 1540)

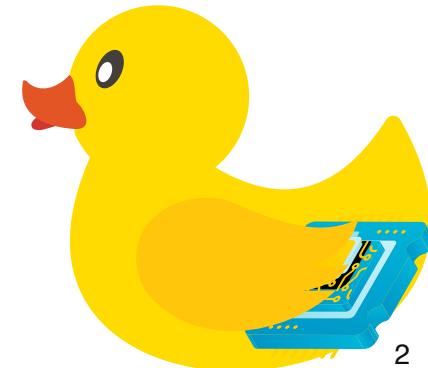


Lecture 12:
Protein structure
prediction

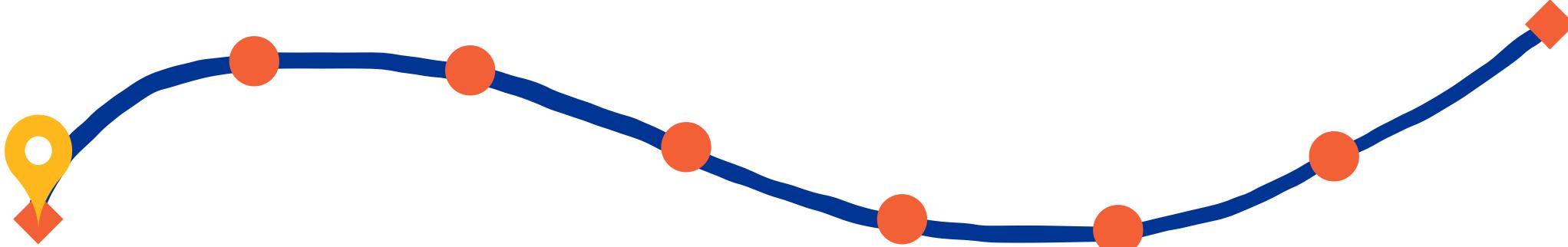
Oct 10, 2024

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"



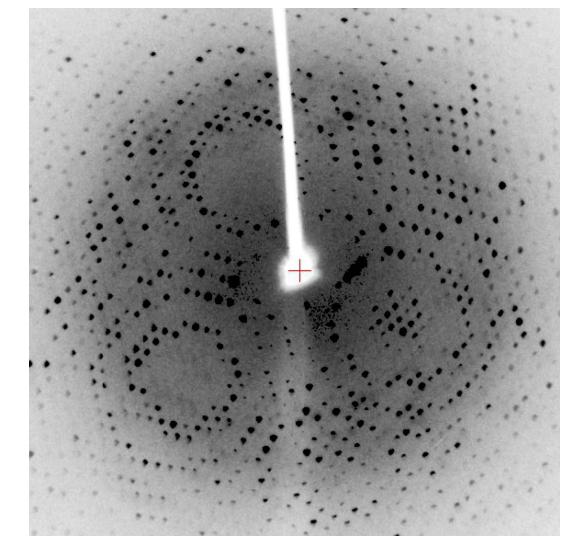
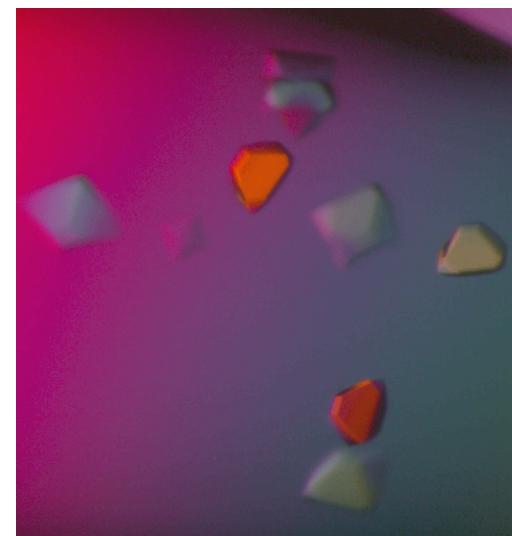
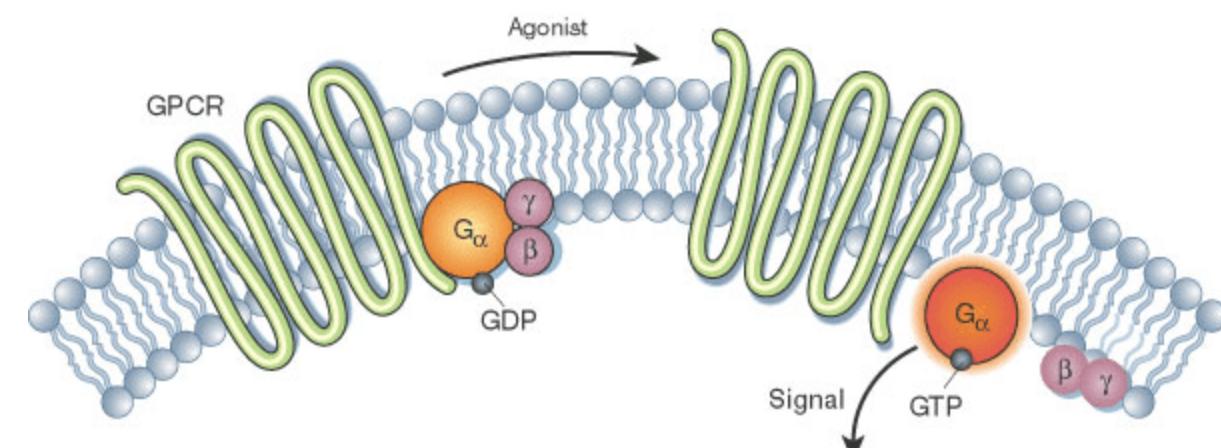
After today, you should be able to



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.

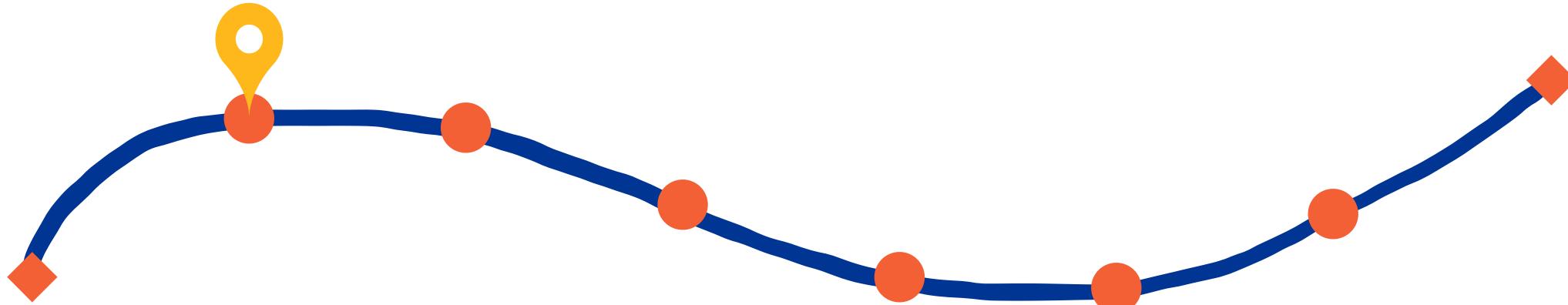
Prediction is critical for the future of biology

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



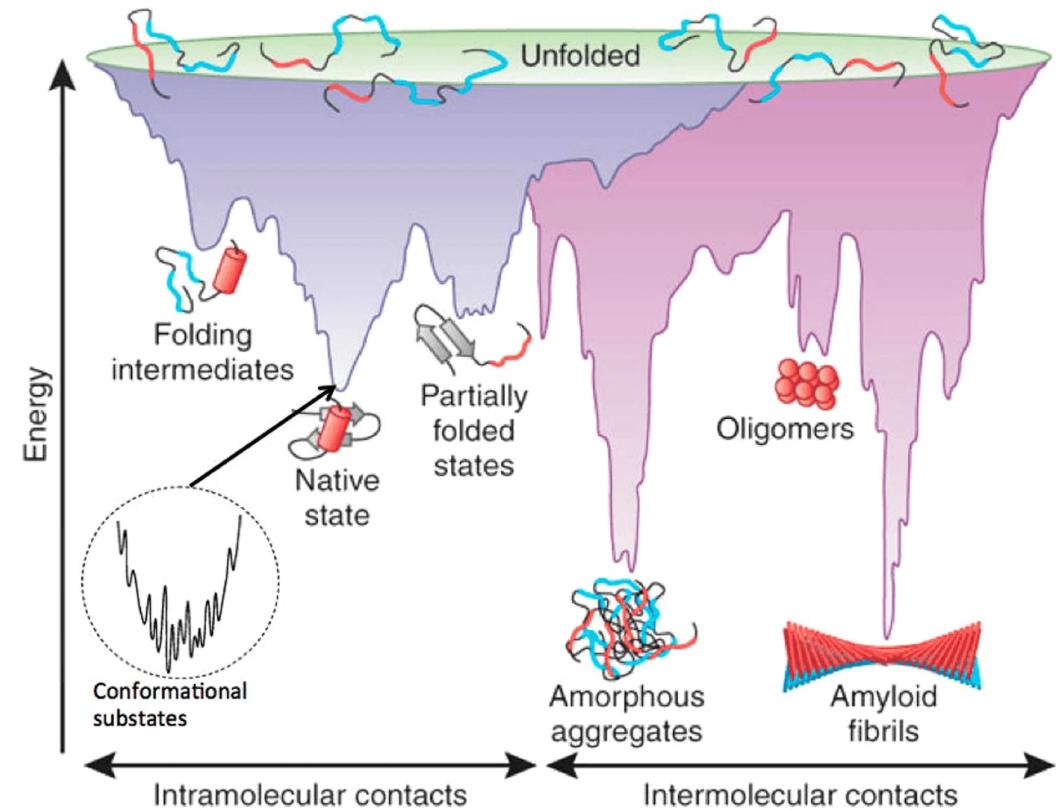
What makes structure prediction hard: Complex energy landscape

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

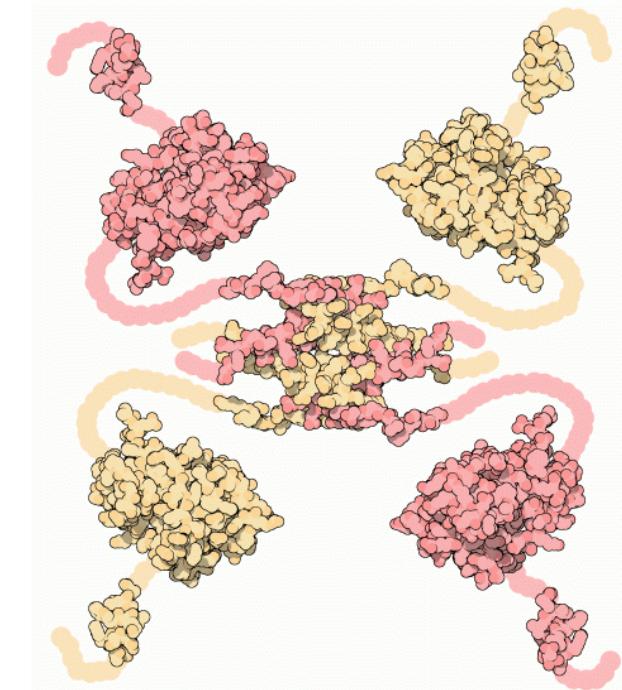
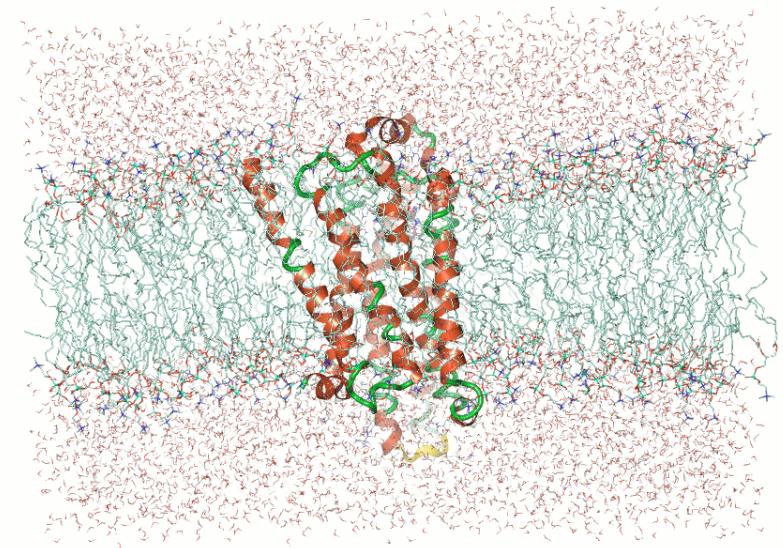
Energy calculations are computationally intensive and depend on accurate force fields



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

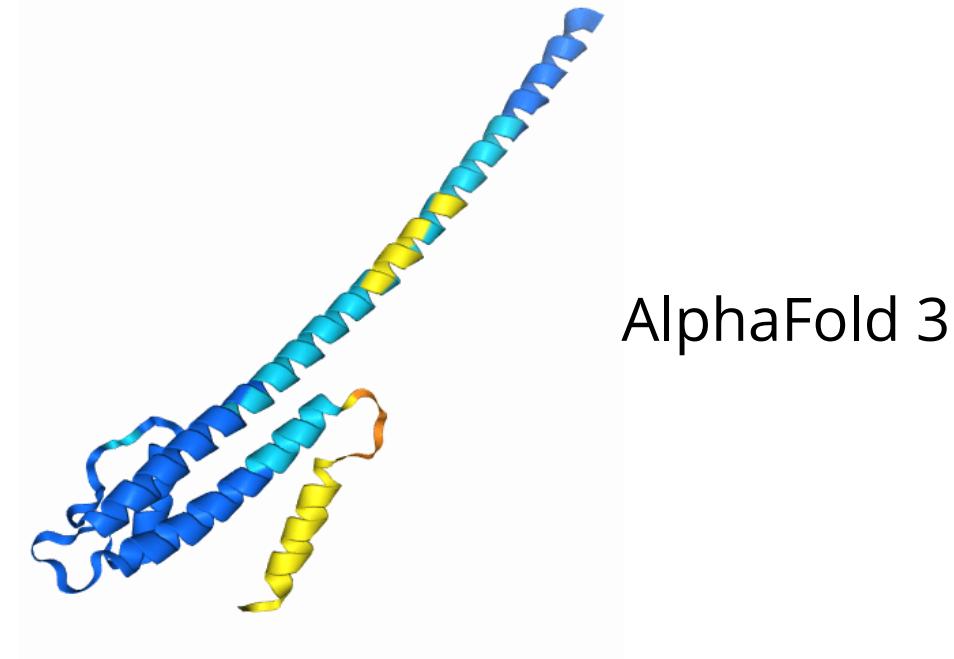
Proteins fold differently in different environments

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

7MHX



pH-gated
K⁺ channel



AlphaFold 3

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

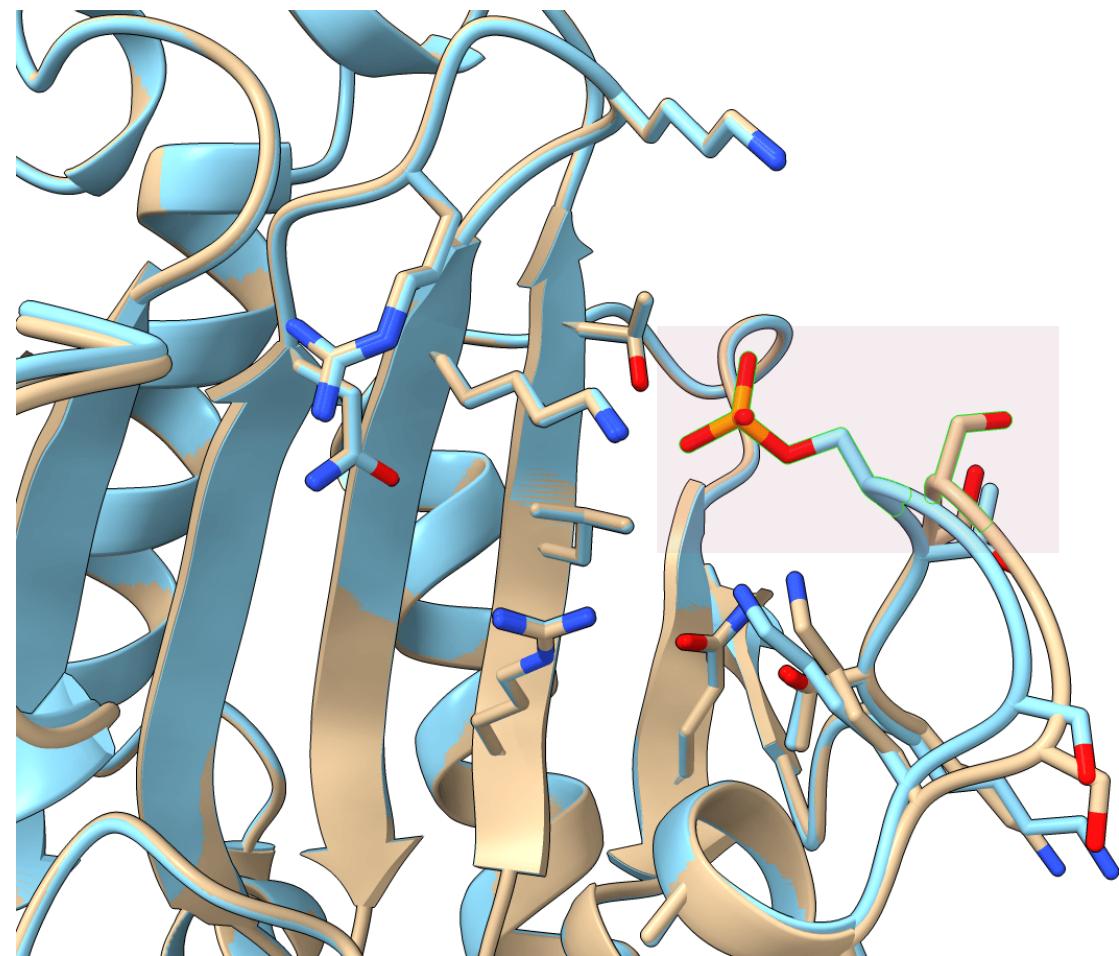
What makes structure prediction hard: Post-translational modifications

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

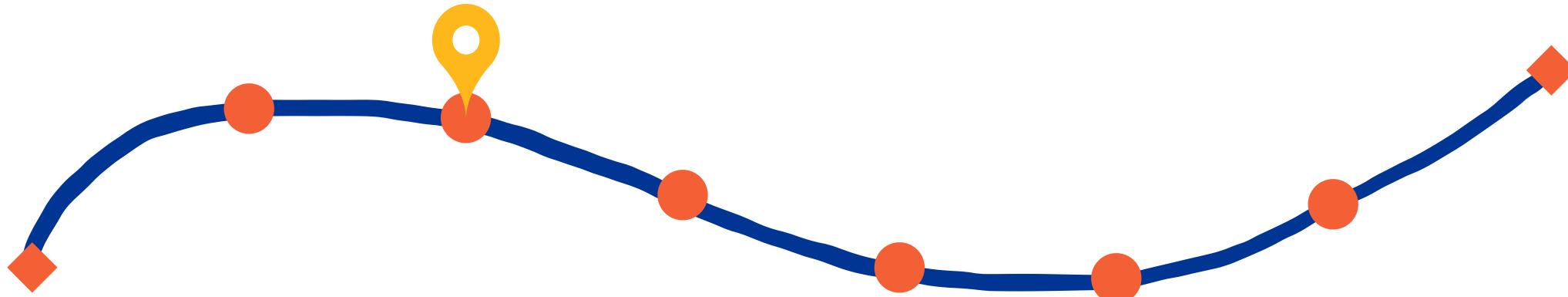


What makes structure prediction hard: Methods are data driven

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



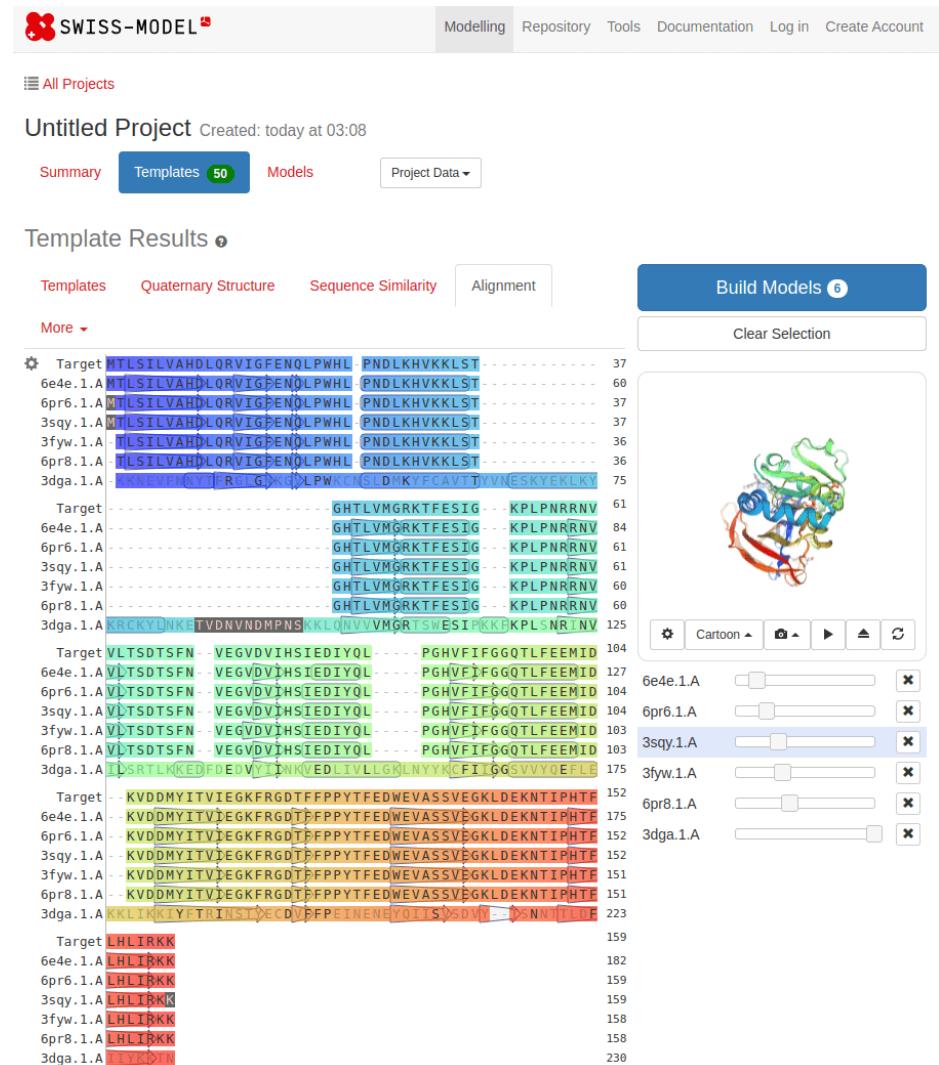
Explain homology modeling

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 shows the SWISS-MODEL web interface. At the top, there is a navigation bar with the SWISS-MODEL logo, a search bar, and links for 'Modelling', 'Repository', 'Tools', 'Documentation', 'Log in', and 'Create Account'. Below the navigation bar, the page title 'Untitled Project' is displayed, along with the creation date 'today at 03:08'. There are tabs for 'Summary', 'Templates' (50), 'Models', and 'Project Data'. The 'Templates' tab is selected, showing a table of template results. The table includes columns for Target, Sequence, and Length. The first row shows the target sequence as MTLSILVAHDLQRVIGFENOLPWHL and the template as PNDLKHKVKKLST, both with a length of 37. The table continues with other templates, including 6e4e.1.A, 6pr6.1.A, 3sqy.1.A, 3fyw.1.A, 6pr8.1.A, and 3dga.1.A. To the right of the table, there is a 3D ribbon model of a protein structure. Below the table, there are buttons for 'Build Models' (6), 'Clear Selection', and various visualization controls. The bottom of the interface shows a series of sliders and checkboxes for different models.

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

Start with a multiple sequence alignment

↓

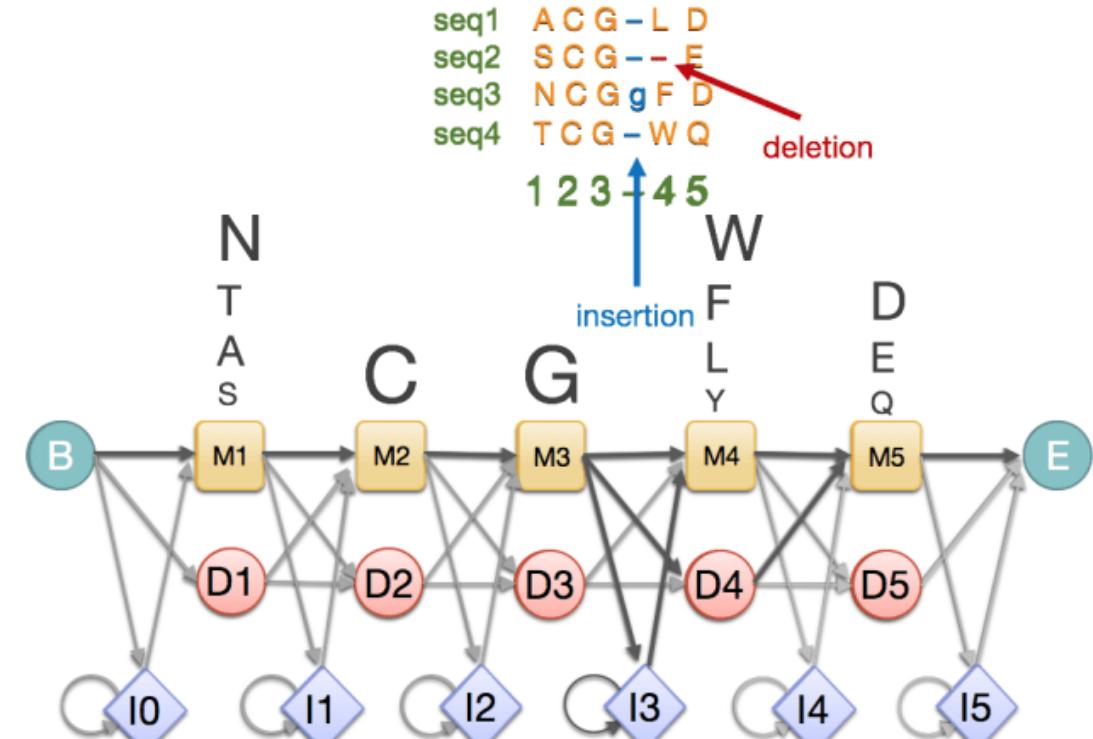
Insertions / deletions can be modelled

↓

Occupancy and amino acid frequency at each position in the alignment are encoded

↓

Profile created



A Markov model predicts outcomes based on transitional probabilities

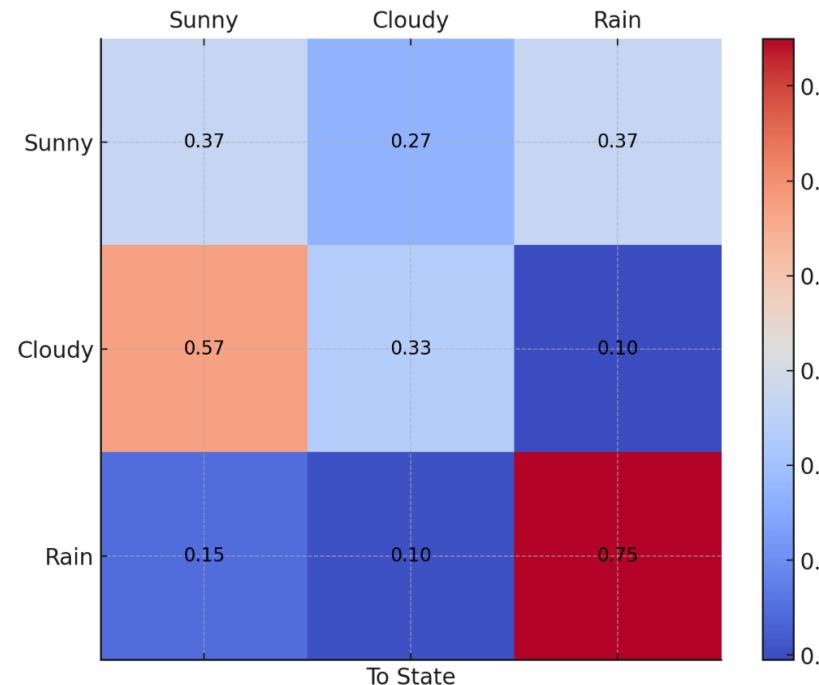
Suppose I collect weather data in Pittsburgh for the past 30 days: **Sunny**, **Cloudy**, or **Rain**

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

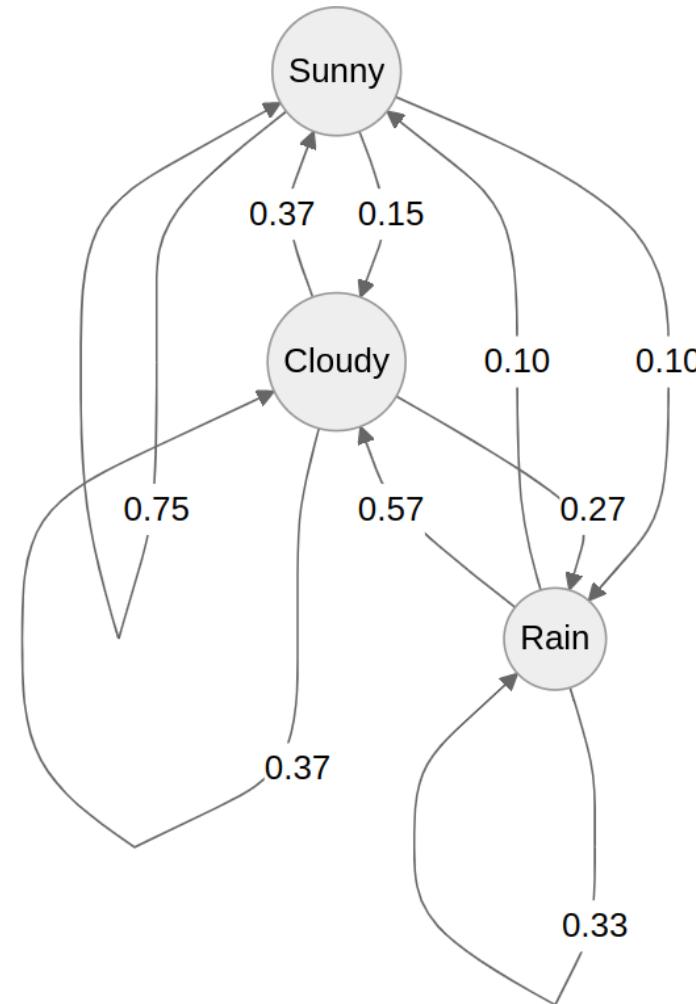
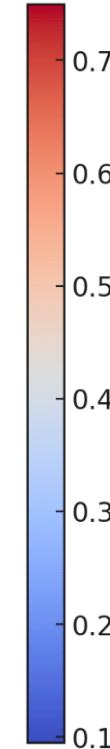
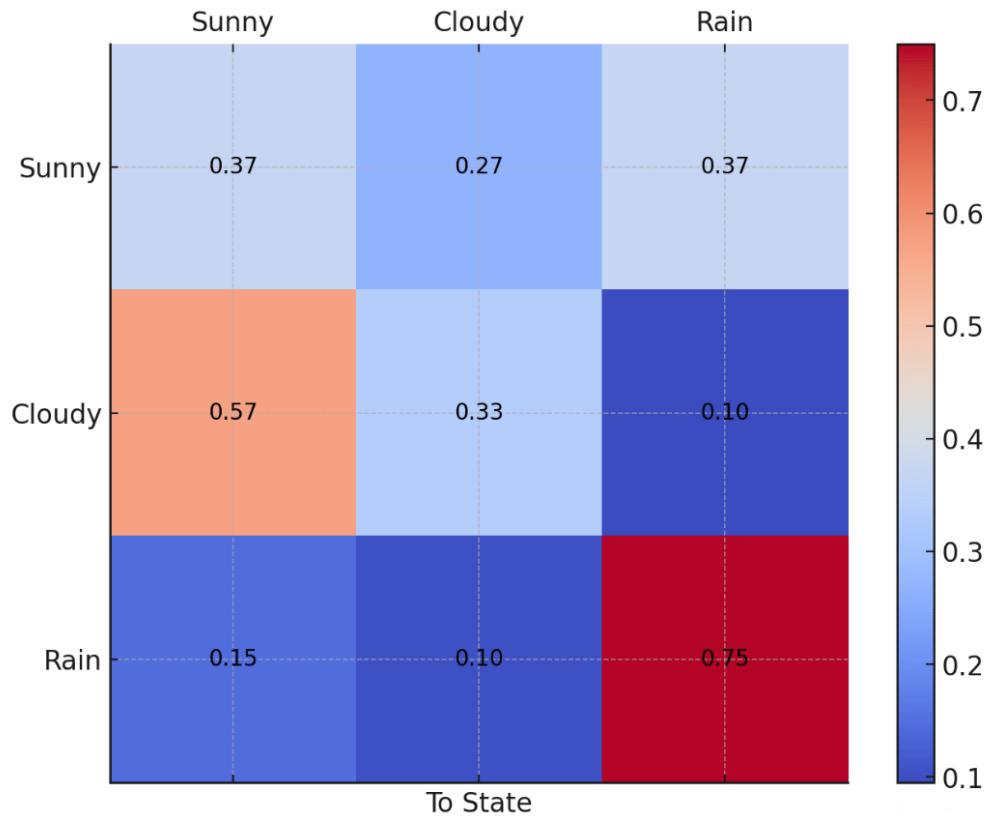
Tomorrow's weather



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

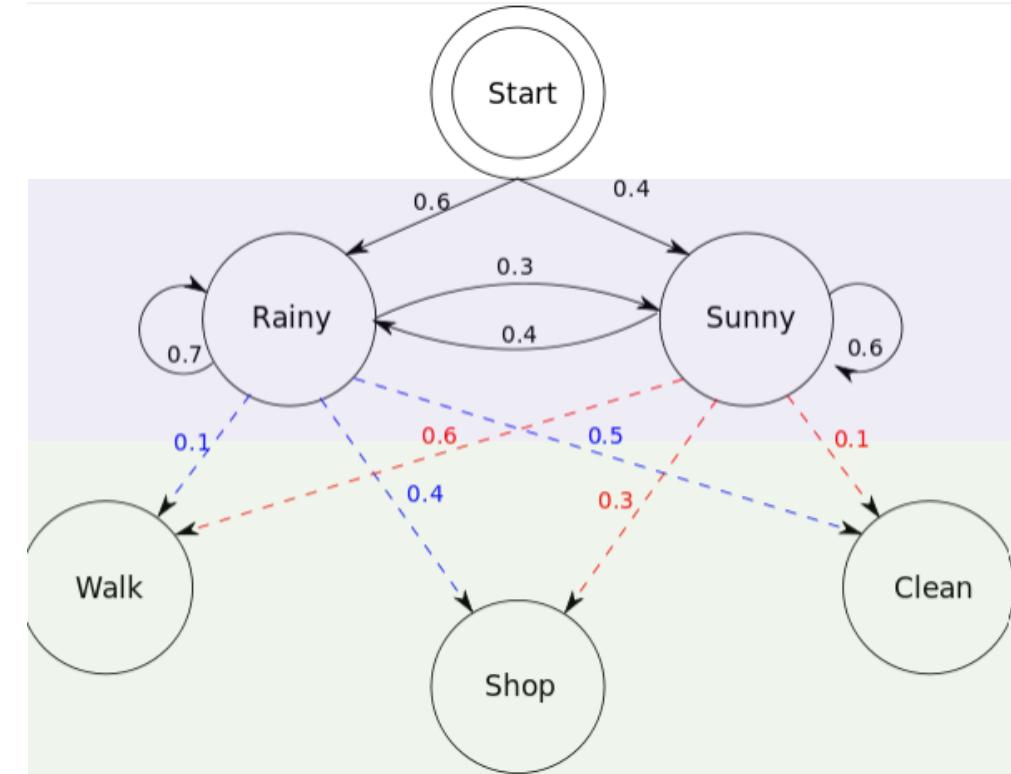
I cannot look up the weather but I have last year's weathers reports **Hidden states**

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



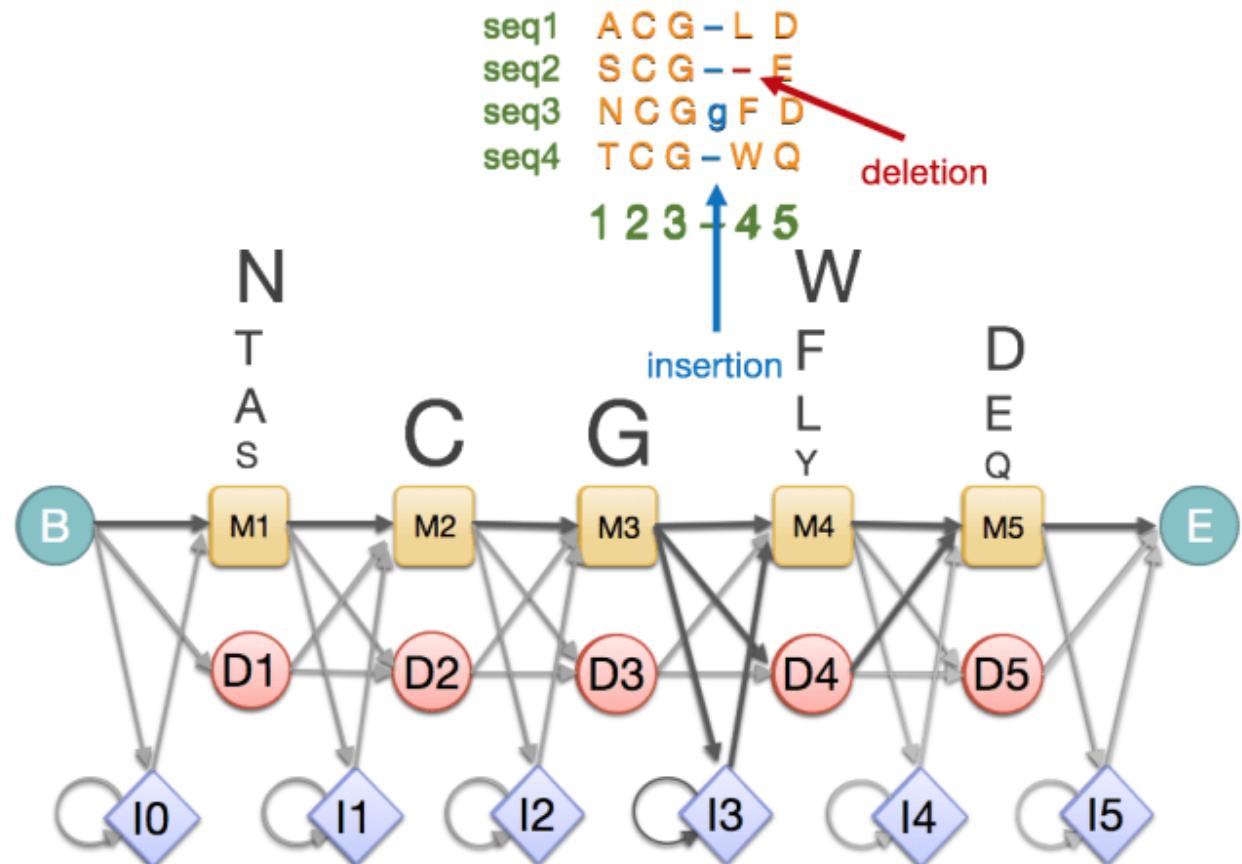
Note: If we had previous observable data, we could fit/learn transition probabilities of hidden states

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



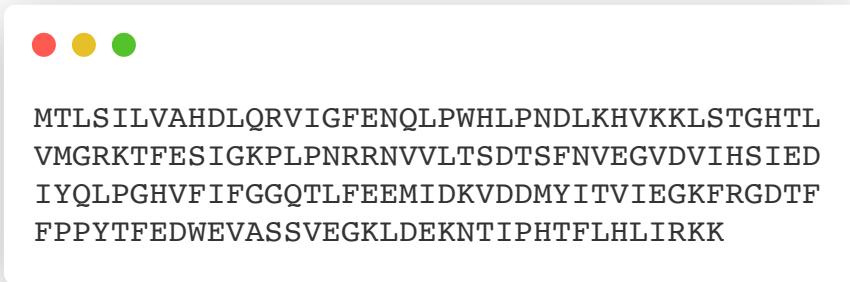
HMMER Uses HMMs to Search Protein Databases for Homology

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



```
MTLSILVAHDLQRVIGFENQLPWHLVPNLDKHKVKKLSTGHTL
VMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVDVIHSIED
IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF
FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK
```

DHFR ([UniProt](#))

ExPasy web server. The purpose of this server is to make protein modelling accessible to all life science researchers worldwide.' A blue 'Start Modelling' button is located on the right. Below this, a section titled 'Repository' is shown with a search bar and a grid of 15 small images representing different species or models." data-bbox="474 113 971 816"/>

SWISS-MODEL 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.

[Start Modelling](#)

Repository

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](#)?

Search SWISS-MODEL Repository



swissmodel.expasy.org

SWISS-MODEL

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 ▾

Project Summary

Target 1 MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKKLSTGHTLVMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVVDIHSI 80
Target 1 EDIYQLPGHVIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTFFPYTFEDWEVASSVEKLDEKNTIPHTFLHLIRKK 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 <i>Staphylococcus aureus</i> MW2 bound to NADP and p218
6pr6.1	100.00	monomer	Dihydrofolate reductase <i>S. aureus</i> dihydrofolate reductase co-crystallized with para-tolyl-dihydrophthalazine inhibitor and NADP(H)
3sqy.1	100.00	monomer	Dihydrofolate reductase <i>S. aureus</i> Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines
3fyw.1	100.00	monomer	Dihydrofolate reductase <i>Staph. aureus</i> DHFR complexed with NADPH and AR-101
6pr8.1	100.00	monomer	Dihydrofolate reductase <i>S. aureus</i> dihydrofolate reductase co-crystallized with 3,5-dimethylphenyl-dihydrophthalazine inhibitor and NADP(H)

Show full template details



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

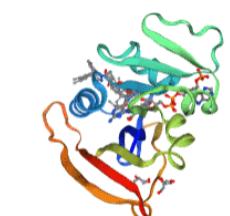
Alignment

More ▾

Sort	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
<input checked="" type="checkbox"/> 6e4e.1.A	Dihydrofolate reductase Crystal structure of dihydrofolate reductase from <i>Staphylococcus aureus</i> MW2 bound to NADP and p218	0.99	-	100.00	X-ray, 1.9Å	monomer	1 x NAP ^Q , 1 x MMV ^Q
<input checked="" type="checkbox"/> 6pr6.1.A	Dihydrofolate reductase <i>S. aureus</i> dihydrofolate reductase co-crystallized with para-tolyl-dihydrophthalazine inhibitor and NADP(H)	0.99	-	100.00	X-ray, 2.0Å	monomer	1 x NAP ^Q , 1 x OWS ^Q
<input checked="" type="checkbox"/> 3sqy.1.A	Dihydrofolate reductase <i>S. aureus</i> Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines	0.98	-	100.00	X-ray, 1.5Å	monomer	1 x NAP ^Q , 1 x Q11 ^Q
<input checked="" type="checkbox"/> 3fyw.1.A	Dihydrofolate reductase <i>Staph. aureus</i> DHFR complexed with NADPH and AR-101	0.98	-	100.00	X-ray, 2.1Å	monomer	1 x NDP ^Q , 1 x XCF ^Q
<input checked="" type="checkbox"/> 6pr8.1.A	Dihydrofolate reductase <i>S. aureus</i> dihydrofolate reductase co-crystallized with 3,5-dimethylphenyl-dihydrophthalazine inhibitor and NADP(H)	0.98	-	100.00	X-ray, 2.0Å	monomer	1 x OWJ ^Q , 1 x NAP ^Q
<input type="checkbox"/> 3sr5.1.A	Dihydrofolate reductase <i>S. aureus</i> Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines	0.98	-	100.00	X-ray, 1.7Å	monomer	1 x NAP ^Q , 1 x Q12 ^Q
<input type="checkbox"/> 2w9g.1.A	DIHYDROFOLATE REDUCTASE <i>Wild-type Staphylococcus aureus</i> DHFR in complex with NADPH and trimethoprim	0.98	-	100.00	X-ray, 2.0Å	monomer	1 x TOP ^Q , 1 x NDP ^Q

Build Models 5

Clear Selection



6e4e.1.A

6pr6.1.A

3sqy.1.A

3fyw.1.A

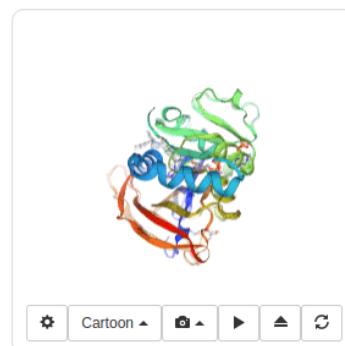
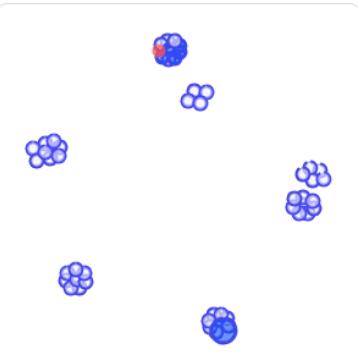
6pr8.1.A

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](#)
[Build Models 6](#)
[Clear Selection](#)


6e4e.1.A

6pr6.1.A

3sqy.1.A

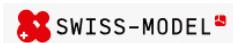
3fyw.1.A

6pr8.1.A

3dga.1.A

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](#)
[Build Models 6](#)
[Clear Selection](#)

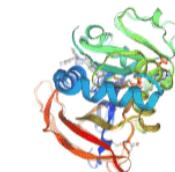
Target **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 37
 6e4e.1.A **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 60
 6pr6.1.A **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 37
 3sqy.1.A **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 37
 3fyw.1.A **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 36
 6pr8.1.A **MTLSILVAHDLQRVIGFENOLPWHL** **PNDLKHVKKKLST** 36
 3dga.1.A **KRCKYLNKE** **TVDNVNNDMPNSKKLQNVVVMGRTSWESIPKKP** **KPLPNRRNV** 75

Target **GHTLVMGRKTFESIG** **KPLPNRRNV** 61
 6e4e.1.A **GHTLVMGRKTFESIG** **KPLPNRRNV** 84
 6pr6.1.A **GHTLVMGRKTFESIG** **KPLPNRRNV** 61
 3sqy.1.A **GHTLVMGRKTFESIG** **KPLPNRRNV** 61
 3fyw.1.A **GHTLVMGRKTFESIG** **KPLPNRRNV** 60
 6pr8.1.A **GHTLVMGRKTFESIG** **KPLPNRRNV** 60
 3dga.1.A **KRCKYLNKE** **TVDNVNNDMPNSKKLQNVVVMGRTSWESIPKKP** **KPLPNRRNV** 125

Target **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 104
 6e4e.1.A **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 127
 6pr6.1.A **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 104
 3sqy.1.A **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 104
 3fyw.1.A **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 103
 6pr8.1.A **VLTSDTSFN** **VEGVVDVIHSIEDIYQL** **PGHVFIGGGQTTLFEEMID** 103
 3dga.1.A **TSRTRLKKEDFEDVY** **DNKVEDLIVLLGLNYYK** **FGGGSVVYQEFLE** 175

Target **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 152
 6e4e.1.A **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 175
 6pr6.1.A **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 152
 3sqy.1.A **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 152
 3fyw.1.A **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 151
 6pr8.1.A **KVDDMYITVIEGKFRGD** **TFPPYT** **FEDWEAVASSVEGKLDEKN** **TIPHTF** 151
 3dga.1.A **KKLIKKIYFTRINST** **ECDFPFPEINENEYQI** **TSQSDVY** **DSNNTLDE** 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



6e4e.1.A

6pr6.1.A

3sqy.1.A

3fyw.1.A

6pr8.1.A

3dga.1.A

SWISS-MODEL

SWISS-MODEL

Modelling Repository Tools Documentation Log in Create Account

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

Local Quality Estimate

Predicted Local Similarity to target

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: 100.00%

Model-Template Alignment

Model 04

Structure Assessment

Compare

159

1

180° 0° -180° 0° 180°

Ψ

General Glycine Proline Pre-Proline A

MolProbity Results

MolProbity Score: 0.93

Clash Score: 1.11 (A97 THR→ NAP)

Ramachandran Favoured: 97.45%

Ramachandran Outliers: 0.00%

Rotamer Outliers: 0.00%

C-Beta Deviations: 0

Bad Bonds: 0 / 1323

Bad Angles: 13 / 1793 (A149 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

Cis Non-Proline: 1 / 150

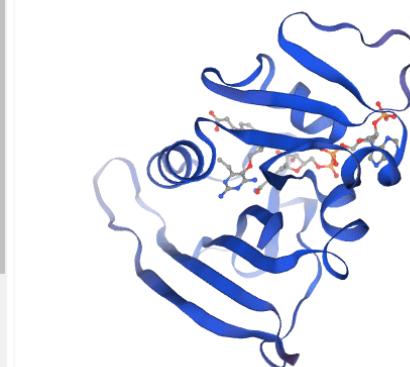
QMEANDisCo

QMEANDisCo Global: 0.91 ± 0.07

Coordinates with QMEAN local scores in the B-factor column

Cartoon

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159



What happens with a novel protein?



MGKKEVILLFLAVIFVALNTLVVAVYFRETADEQVVYGK
NNINQKLIQLKDGTYGFEPALPHVGTFKVLDNSRVPQIA
QEIIIRNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN
NLHGYEGFLFKDNVPLTYPQEFILESNLDGKVRSLQNYDY
DLDVLFGEKEEVKSEILRGLYYNTYTRAFSPYKL

Novel protein
(ChatGPT)

Templates Quaternary Structure Sequence Similarity Alignment

More ▾

Target MGKKEVILLFLAVIFVALNTLVVAVYFRETADEQVVYGKNNINQKLIQLK 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 DGTYGFEPEALPHVGTFKVLDNSRVPQIAQEIIIRNKVKRYLQEAVRIEGTY 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

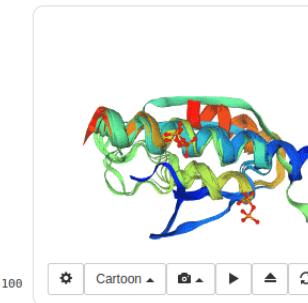
7cr6.1.B
7cr6.1.D
2vky.1.A
7cr8.1.A
7cr8.1.D
7cr8.1.B
7cr6.1.A
7cr6.1.C
7cr8.1.C
4jg4.1.A
6d1r.1.A
1d6t.1.A
1a6f.1.A

Target GKVRSLQNYDYDLDVLFGEKEEVKSEILRGLYYNTYTRAFSPYKL 190

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

Build Models 13

Clear Selection



Cartoon ▾

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

7cr8.1.B 4jg4.1.A 6d1r.1.A 1d6t.1.A 1a6f.1.A

Novel proteins are too challenging

SWISS-MODEL

Modelling Repository Tools Documentation Log in Create Account

All Projects

Novel Created: today at 03:25

Summary Templates 13 Models 13 Project Data

Model Results Order by: GMQE

Model 07

Structure Assessment

Compare 3 selected

Download files Display files

Oligo-State Monomer GMQE 0.11 QMEANDisCo Global: 0.39 ± 0.11

QMEANDisCo Local

QMEAN Z-Scores

Template 7cr6.1A CRISPR-associated endonuclease Cas1 Seq Identity 14.04% Coverage

Model-Template Alignment

Model 03

Structure Assessment

Compare 3 selected

Download files Display files

SWISS-MODEL

Modelling Repository Tools Documentation Log in Create Account

01 07 A
02 03 A
03 08 C

Remove All

Consistency with Ensemble

Consistency Residue index 0 25 50

01 02 03

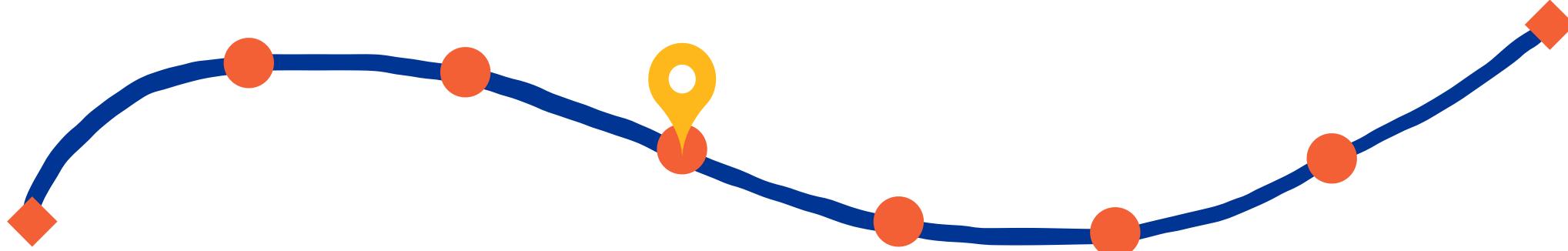
Ensemble Variance

Download set of superposed models

Superpose

Target_1 SNRVPQIAQEIIIRNKVKRYLQEAVRIEGTYPPIVDGLVNAKYTVANP
01.A SNRVPQIAQEIIIRNKVKRYLQEAVRIEGTYPPIVDGLVNAKYTVANP
02.A -----EGTYPPIVDGLVNAKYTVANP
03.C SNRVPQIAQEIIIRNKVKRYLQEAVRIEGTYPPIVDGLVNAKYTVANP

After today, you should be able to



Know when to use threading
instead of homology modeling

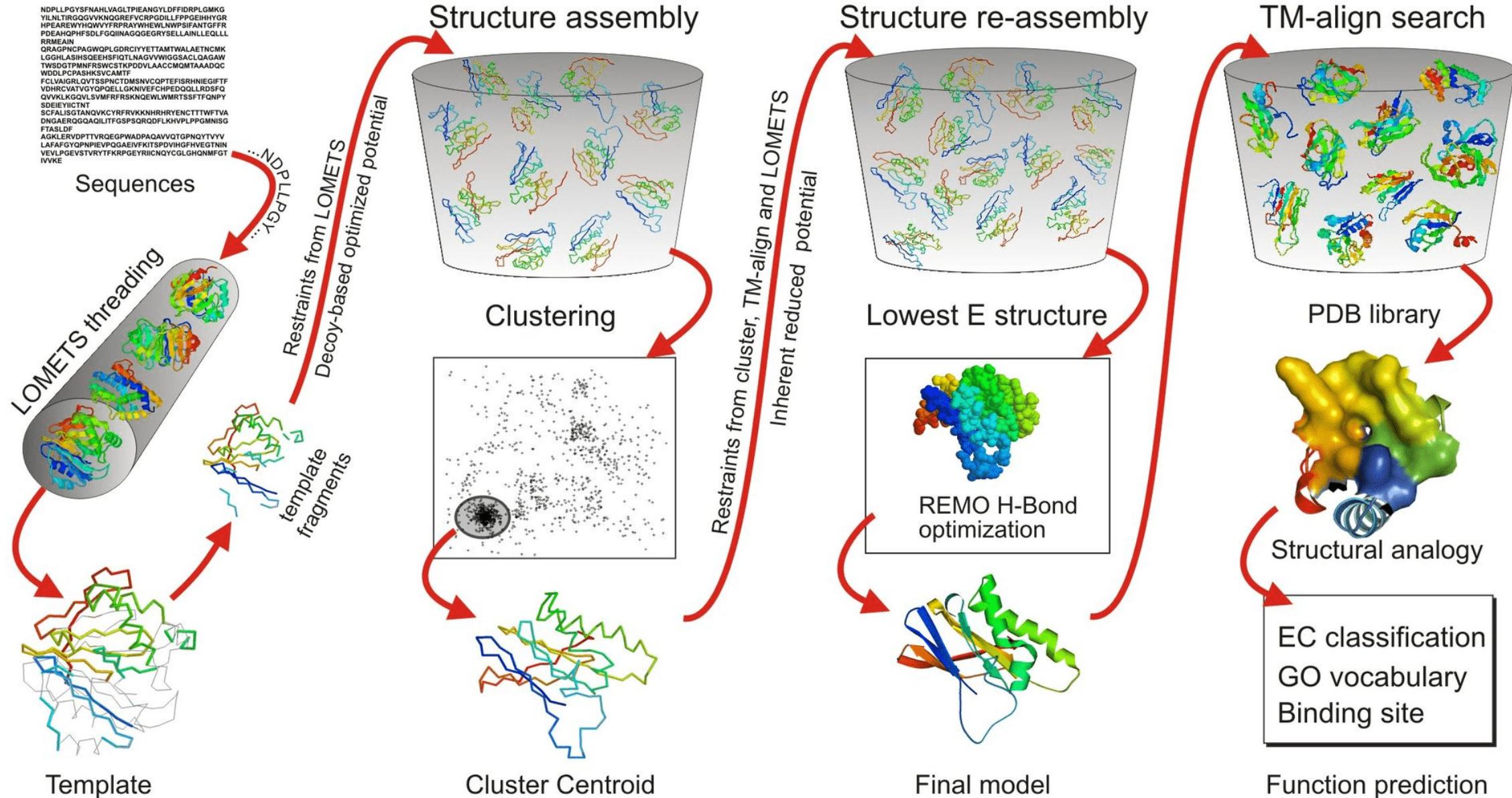
Why Use Threading?

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

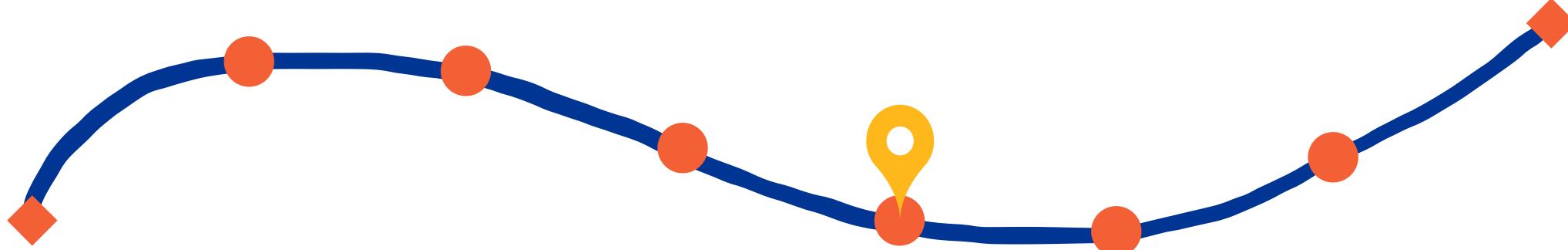
Threading matches sequences to known structural folds based on structural rather than sequence similarity

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



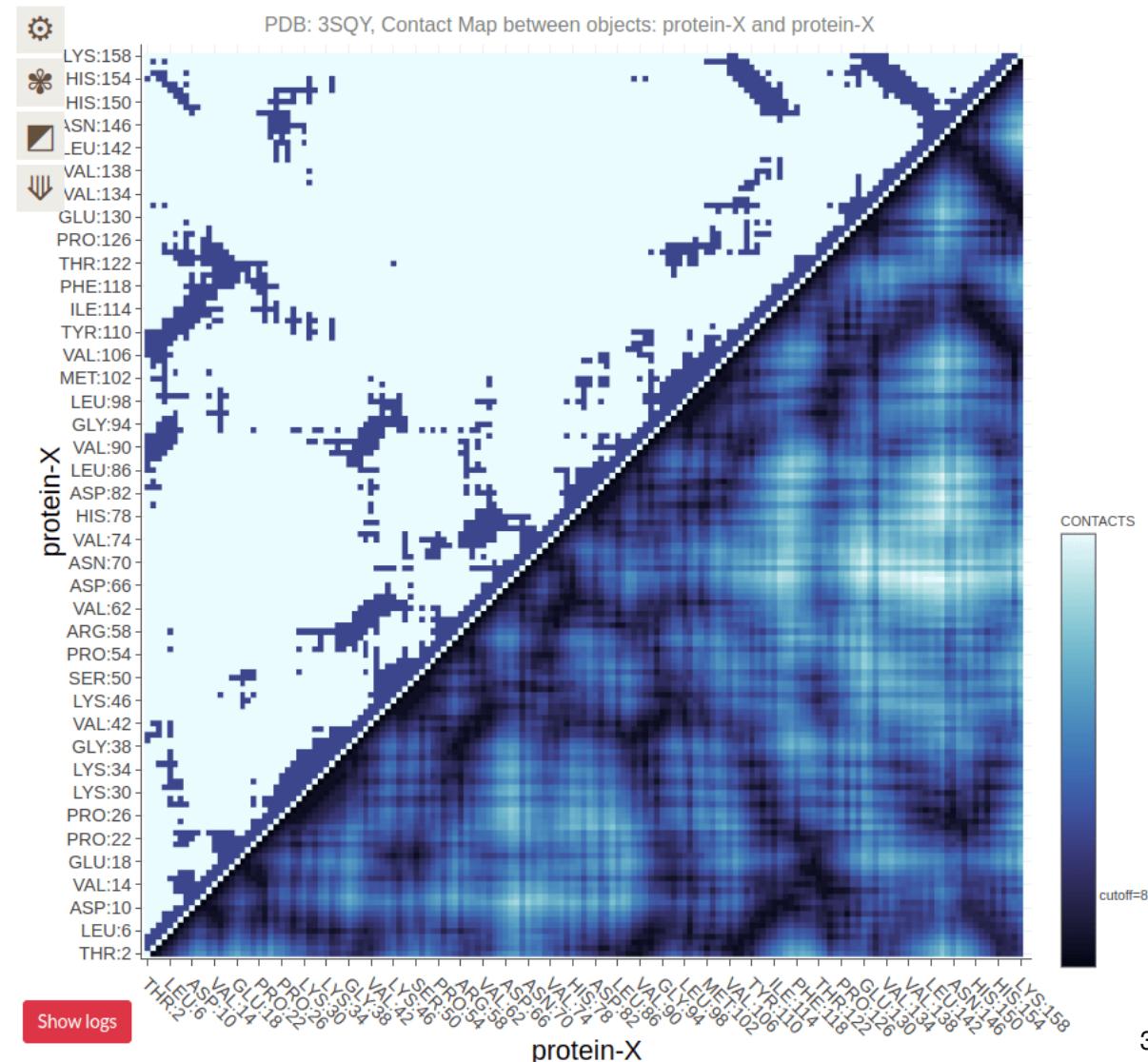
Interpret a contact map for protein
structures

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

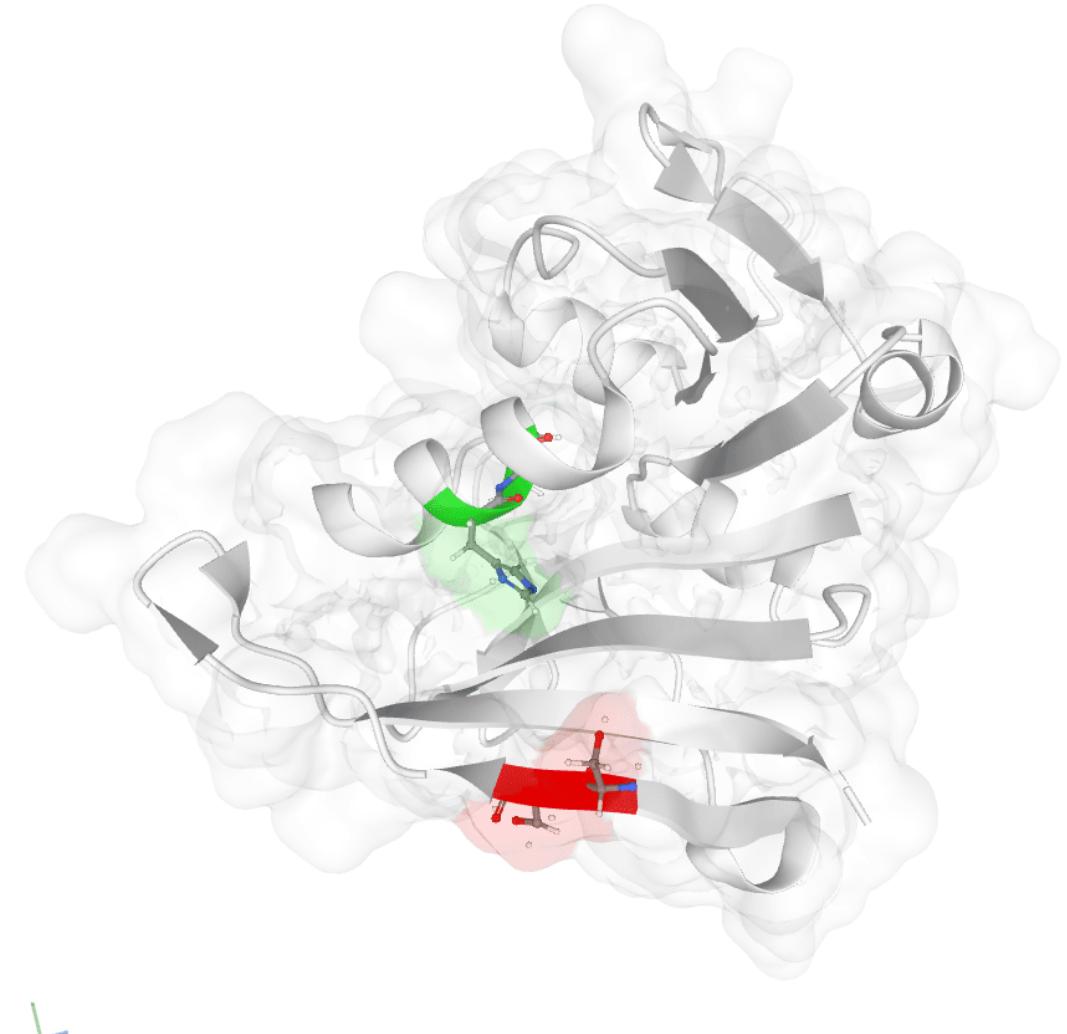
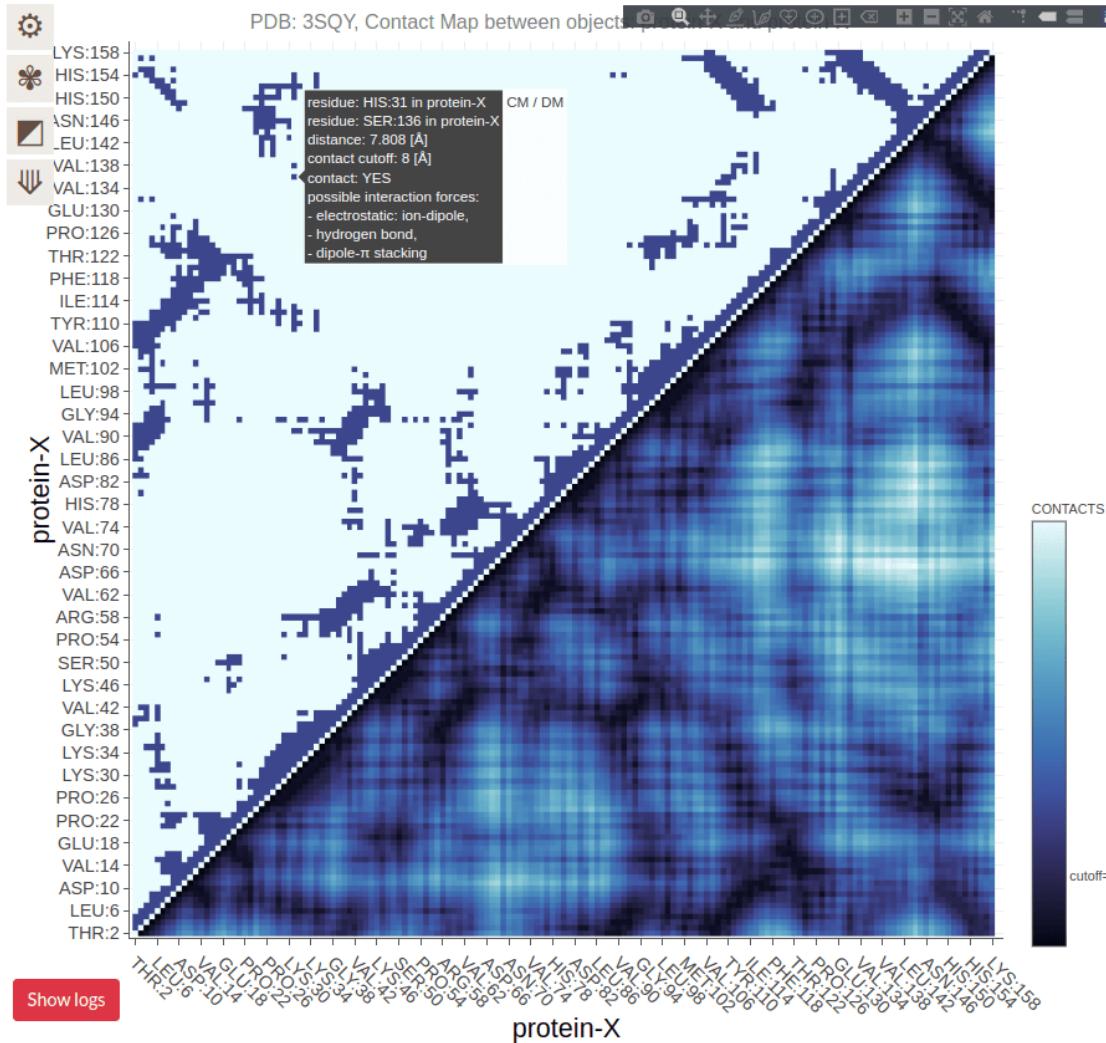
mapiya.lcbio.pl



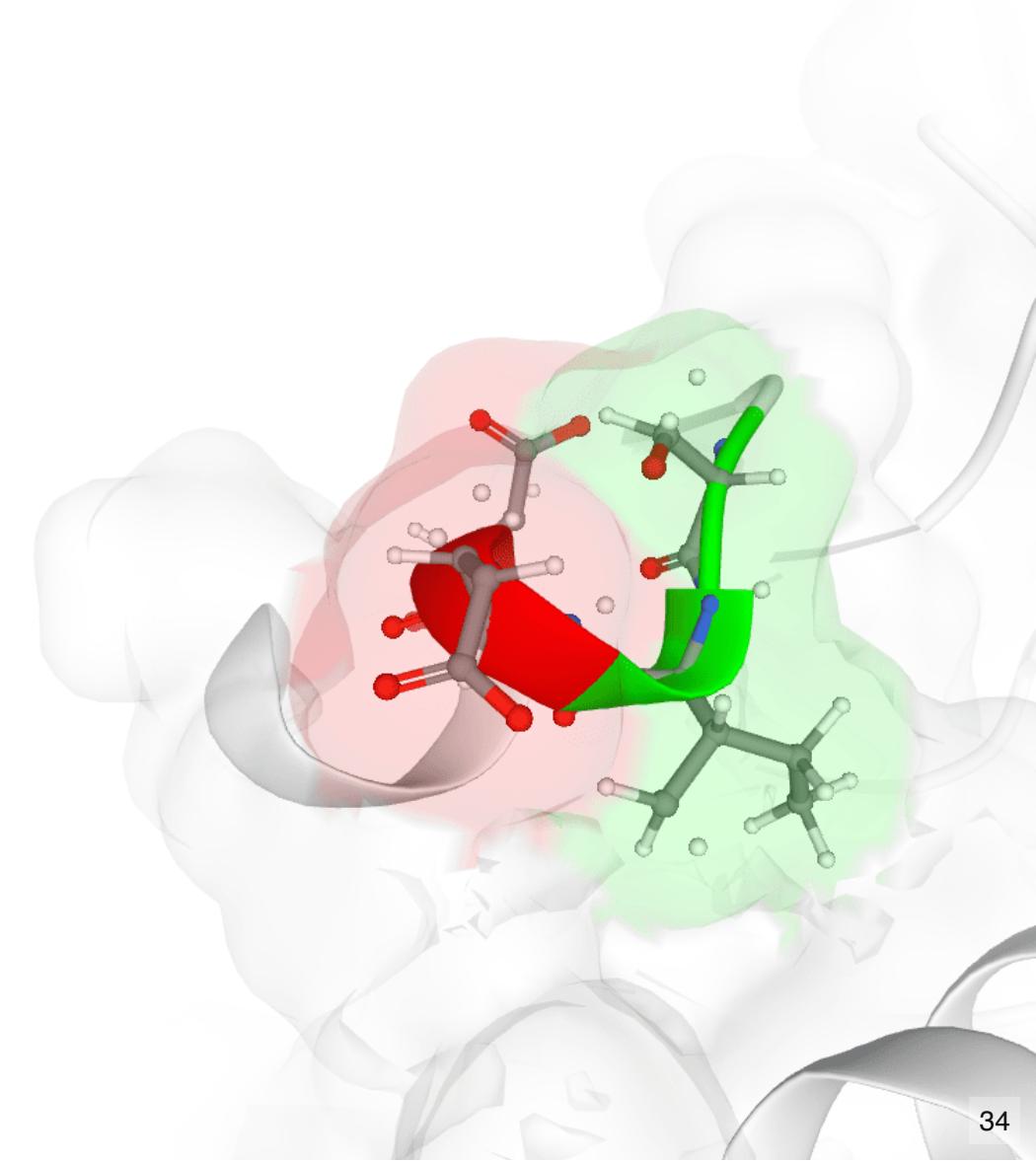
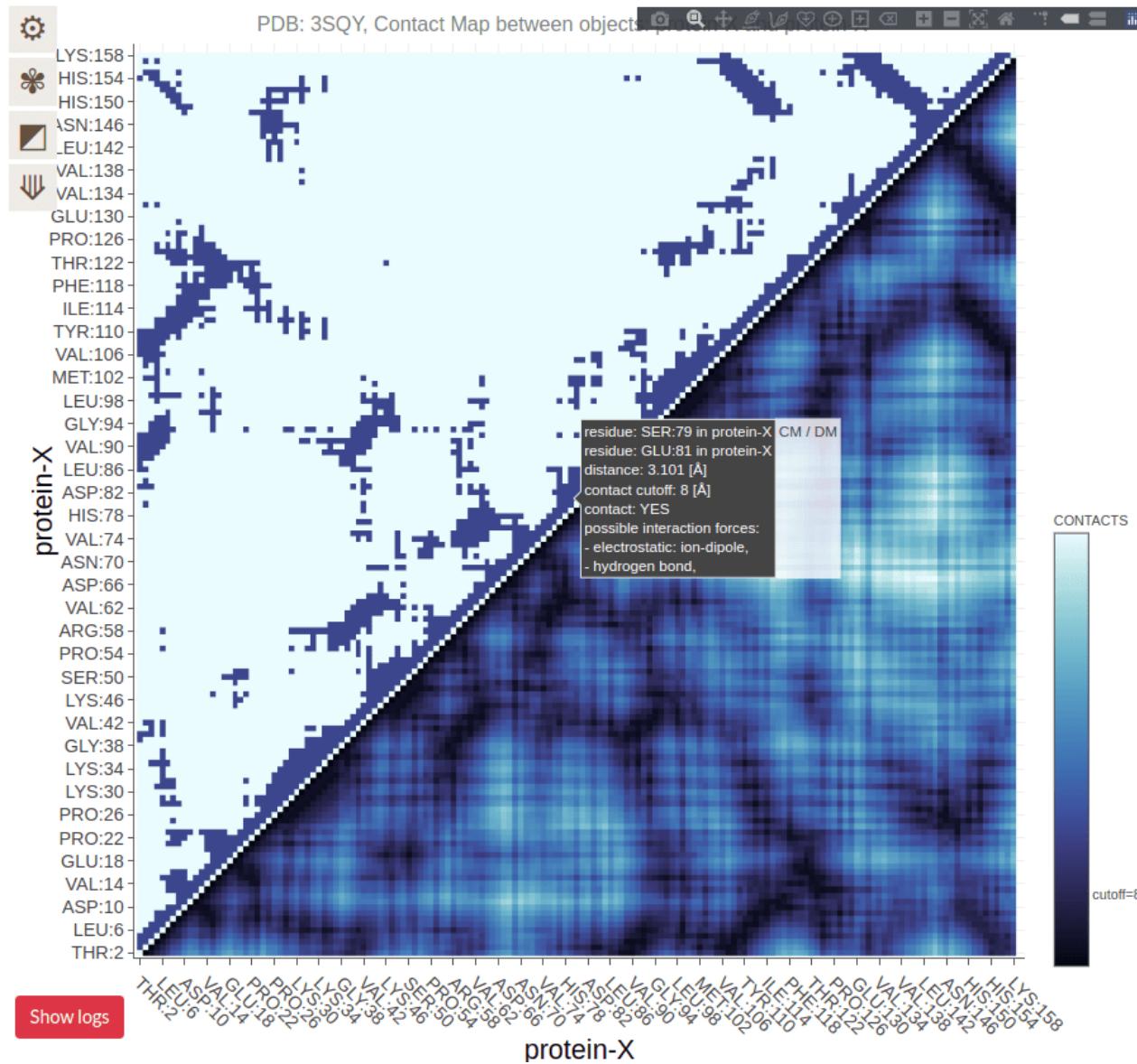
Contact Maps Represent Spatial Proximity, Not Sequence Order

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

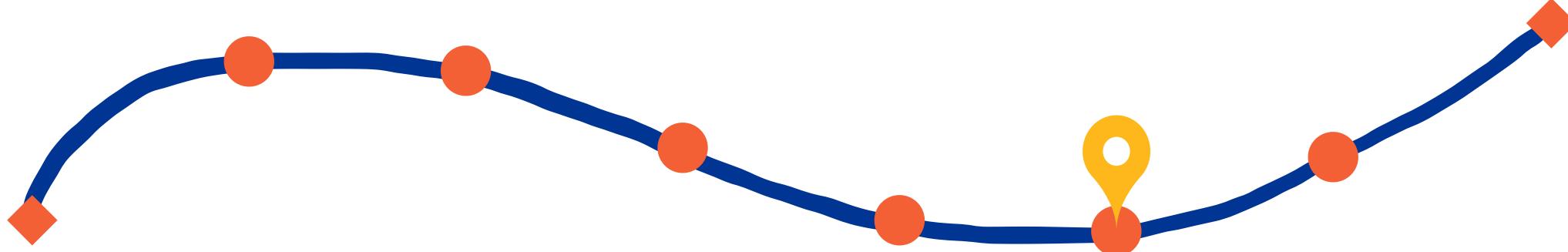
Residues far apart in the sequence can still be close in the 3D structure, reflected in the contact map



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)

N	●	●	*	●	●	●	●	●	●	●	●	●	c		
	A	T	R	L	T	L	T	A	K	K	D	G	P	C	D
	A	T	R	L	T	L	T	A	K	K	D	G	P	C	D
	A	T	R	L	T	L	T	A	K	K	D	G	P	C	D
	A	T	K	L	C	L	T	A	K	K	E	G	P	K	D
	A	T	K	L	T	L	T	A	K	K	E	G	P	K	D
	A	T	K	L	T	L	G	A	K	K	E	G	G	C	D
	A	T	W	L	T	L	T	A	K	K	V	G	P	C	D
	A	T	W	L	T	L	T	A	K	K	V	G	P	C	D

Asp (negative)

Glu (negative)

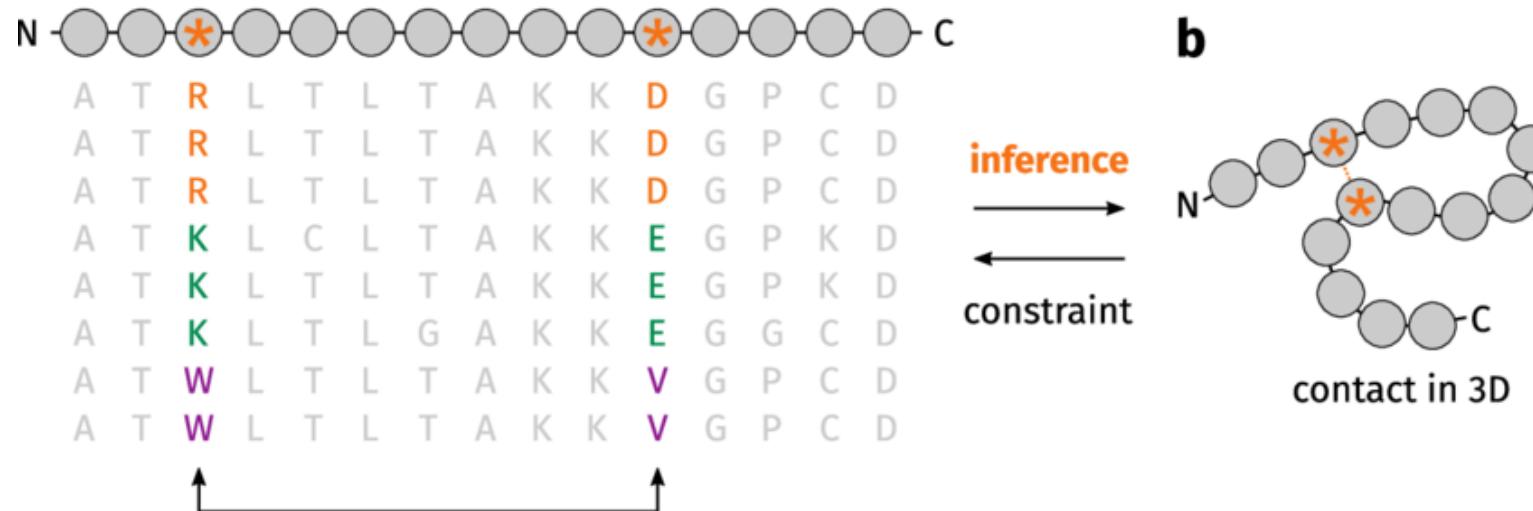
Val (hydrophobic)

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

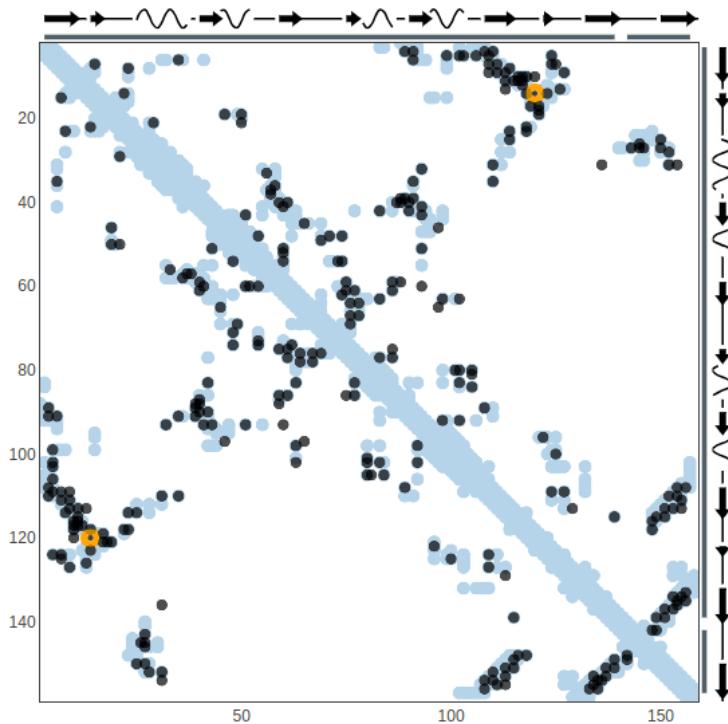
Evolutionary information from MSAs guides predictions for residue-residue contacts



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 _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



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.

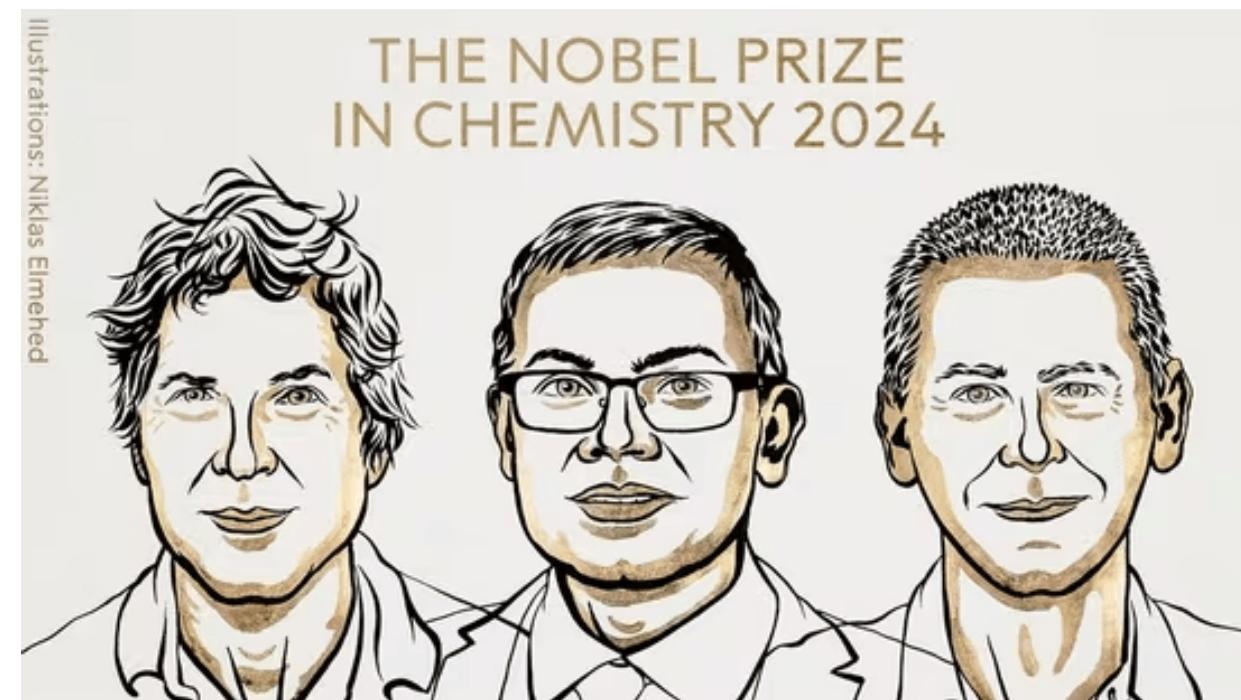
Large and diverse sequence data sets are needed for reliable coevolution predictions.



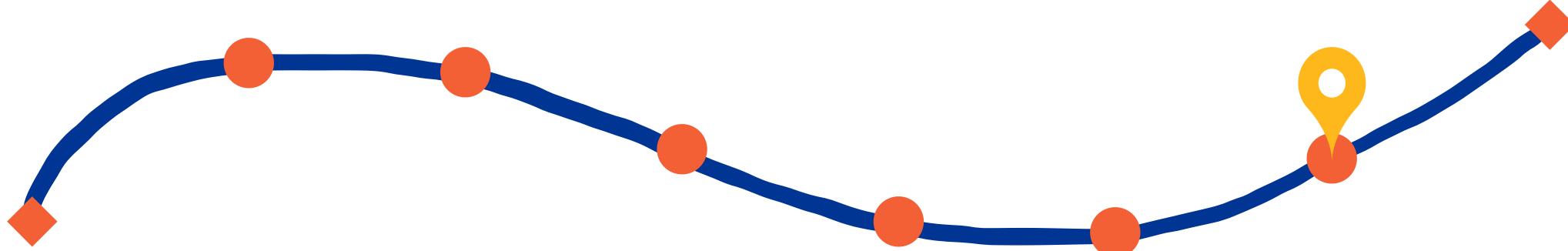
Machine learning leverages coevolution for high-accuracy predictions

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

Predict

Input sequence

Multiple Sequence Alignment

Page 10

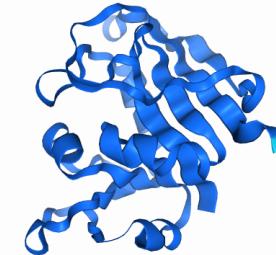
MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKK
LSTGHTLVMGRKTFSIGKPLPNRRNVVLTSDTS
FNVEGVDVIIHSIEDIYQLPGHVFIFGGQTLFEEM
IDKVDDMYITVIEGKFRGDTFFPPYTfedWEVAS
SVEGKLDEKNTIPHTFLHLIRKK

ML models

DHFR

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

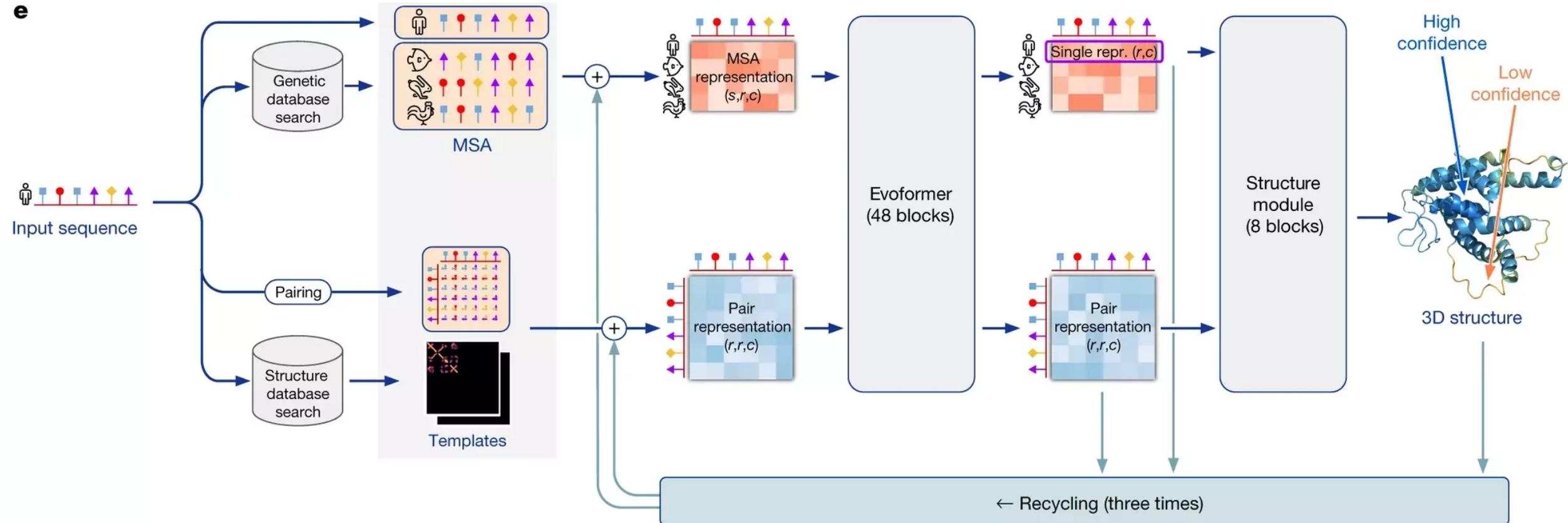
ipTM = - pTM = 0.95 [learn more](#)



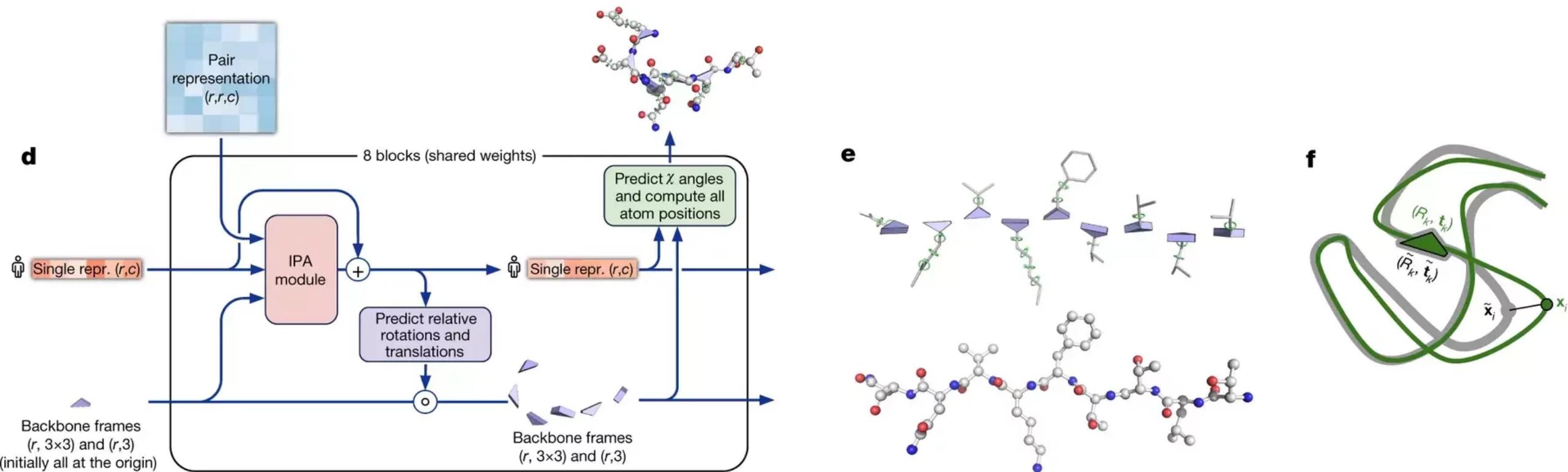
Atomistic structure

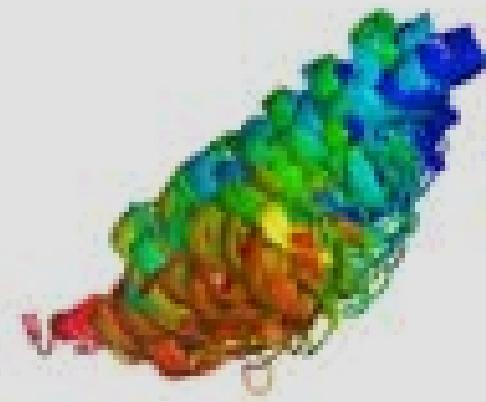
AlphaFold 2 pipeline: Evoformer

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



Contact maps are converted into dihedral angles





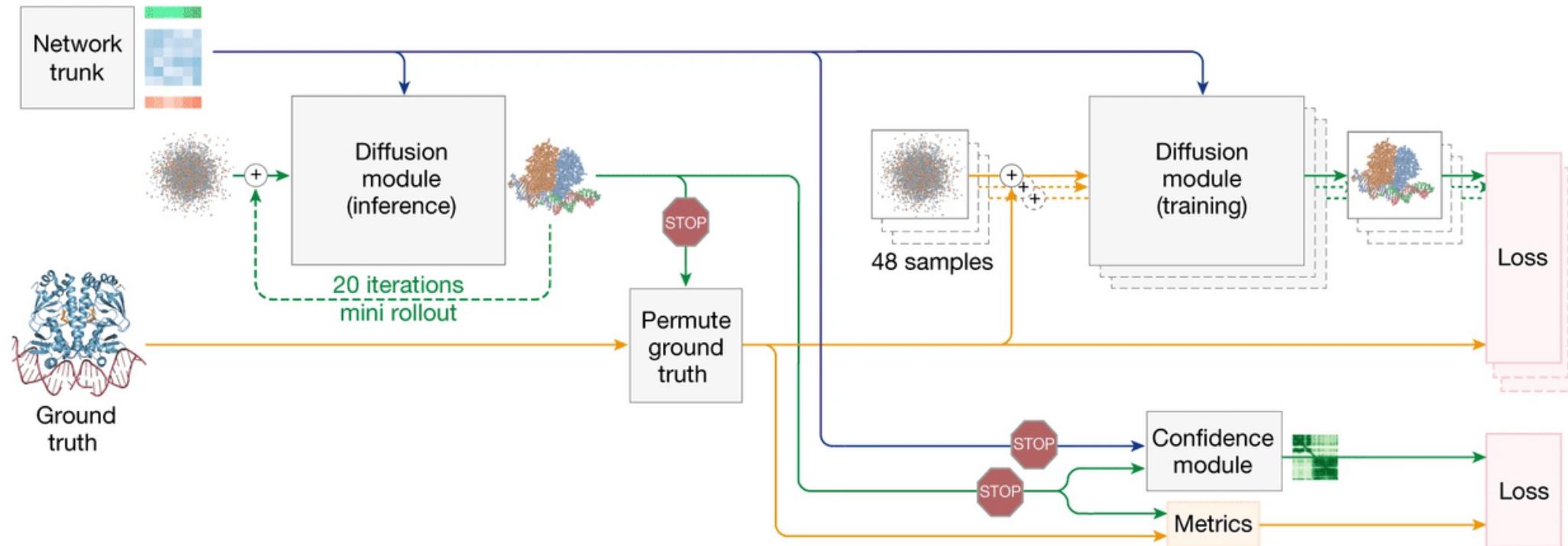
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction



What is new in AlphaFold 3?

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

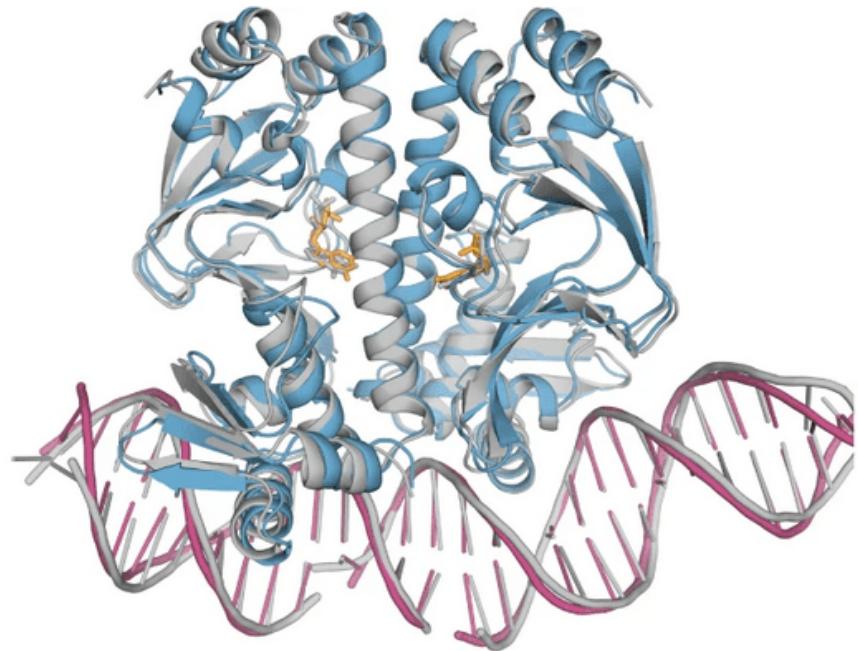
Diffusion models essentially learn to **unscramble atoms into a structure**



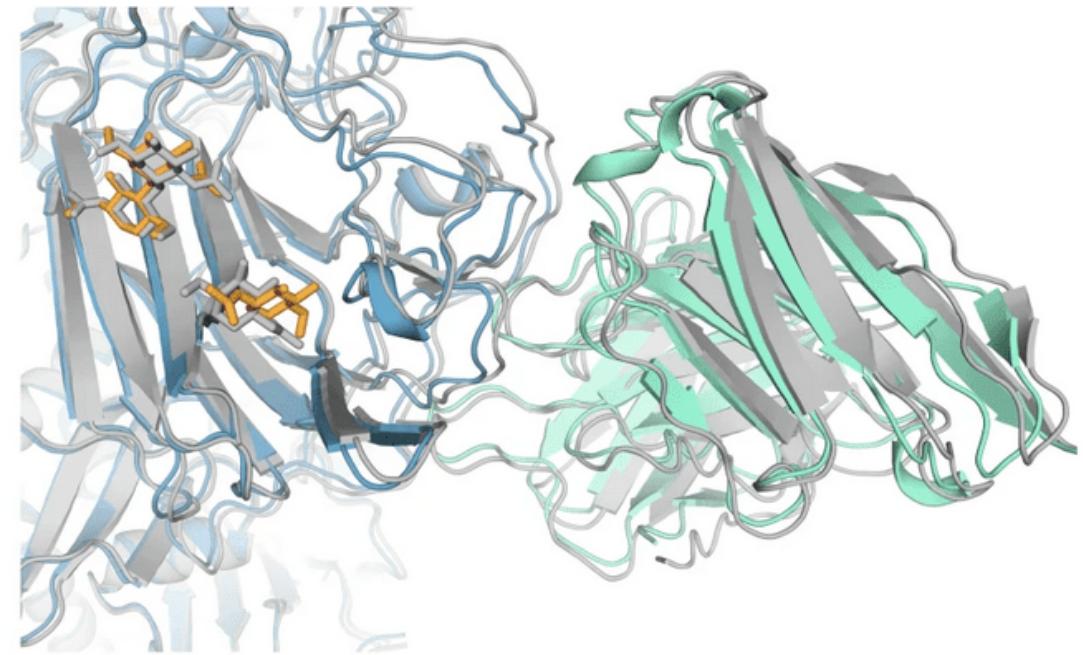
AlphaFold 3 is supercharged for any biomolecule

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

a



b



AlphaFold 3

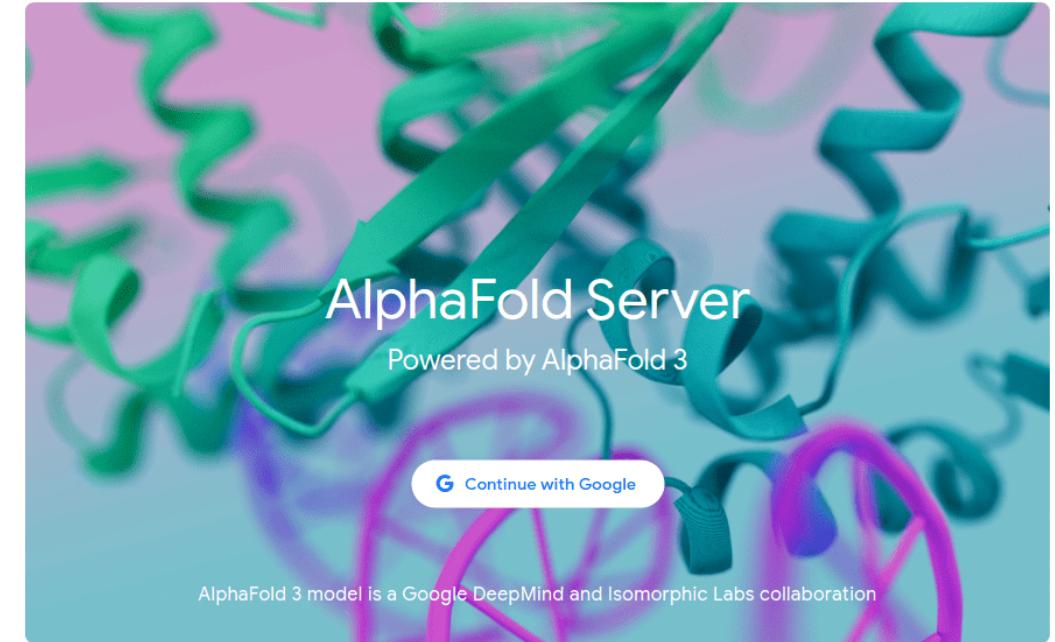
≡ AlphaFold Server BETA

MTLSILVAHDLQRVIGFENQLPWHPNDLKHKVKKLSTGHTL
VMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVDVIHSIED
IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF
FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK

DHFR ([UniProt](#))

MGKKEVILLFLAVIFVALNTLVVAVYFRETADEQVVYGK
NNINQKLIQLKDGTGFEPALPHVGTFKVLDNSRVPQIA
QEIIIRNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN
NLHGYEGFLFKDNVPLTYPQEFLSNDGKVRSLQNYDY
DLDVLFGEKEEVKSEILRGLYYNTYTRAFSPYKL

Novel protein
(ChatGPT)



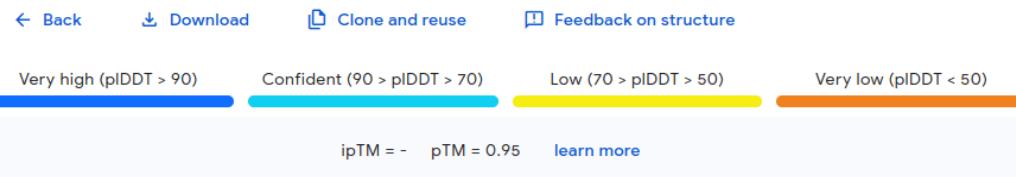
How does AlphaFold Server work?

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



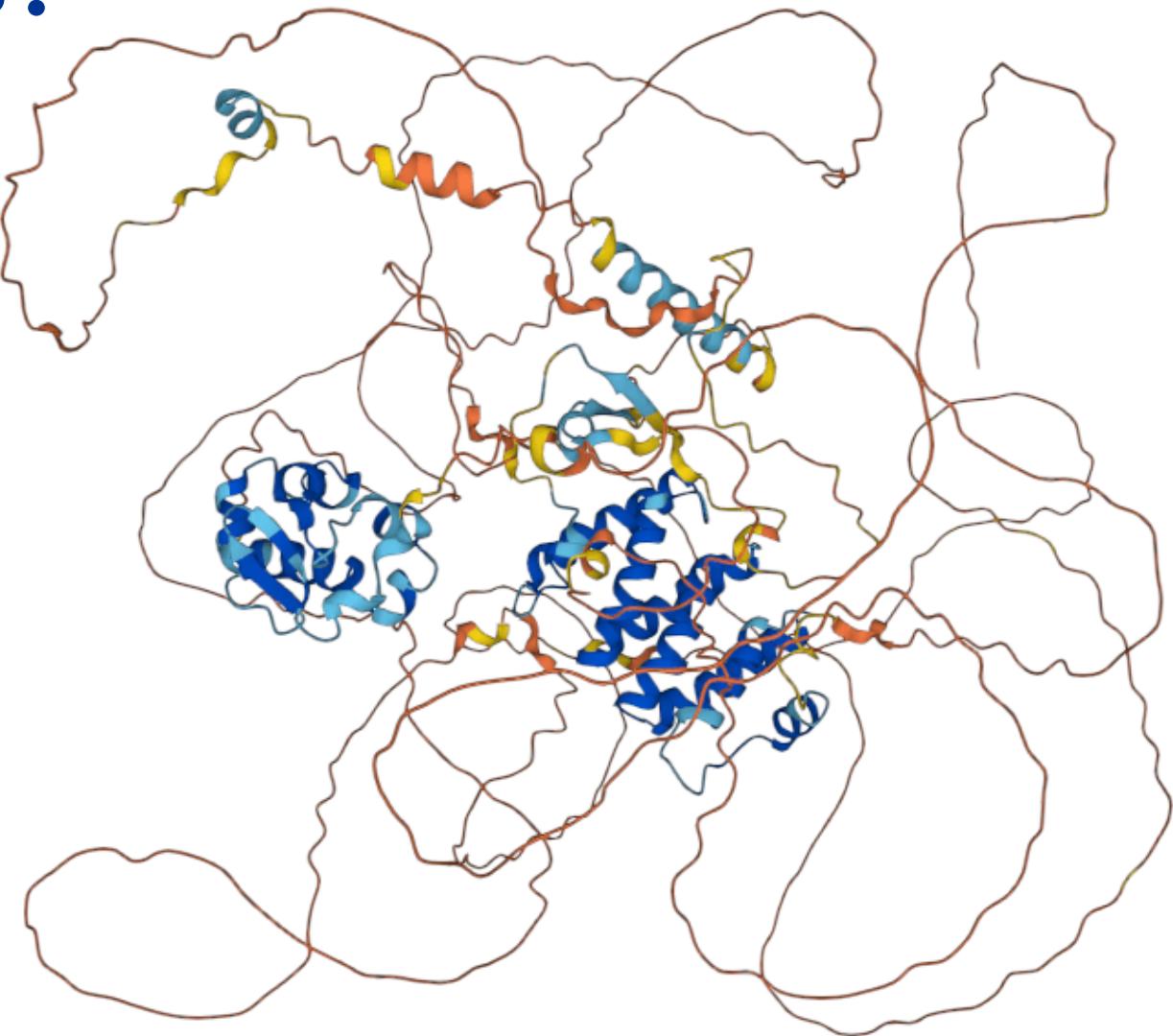
Caveat: Proteins are dynamic

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



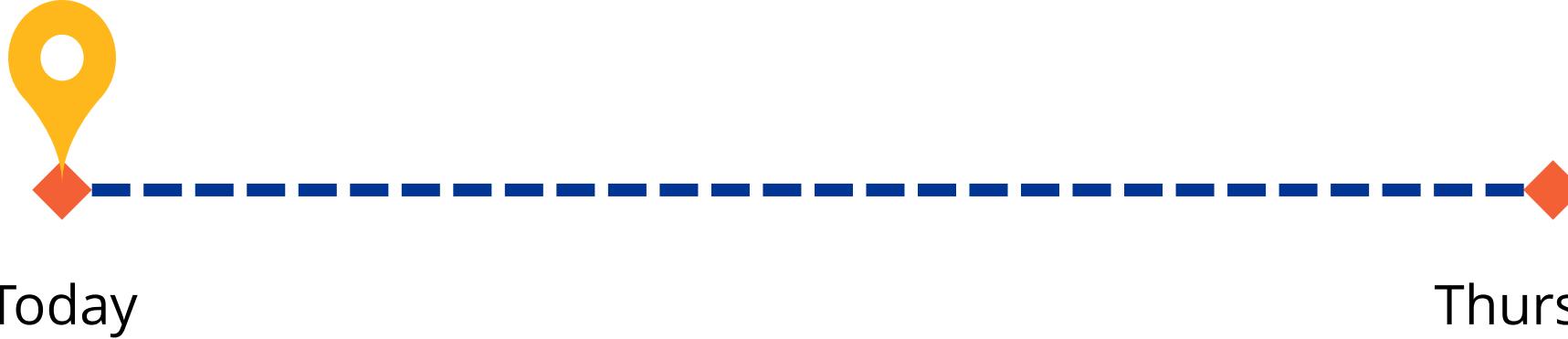
Before the next class, you should

Lecture 12:

Protein structure prediction

Lecture 13:

Molecular simulation principles



- Work on A05
- Review material