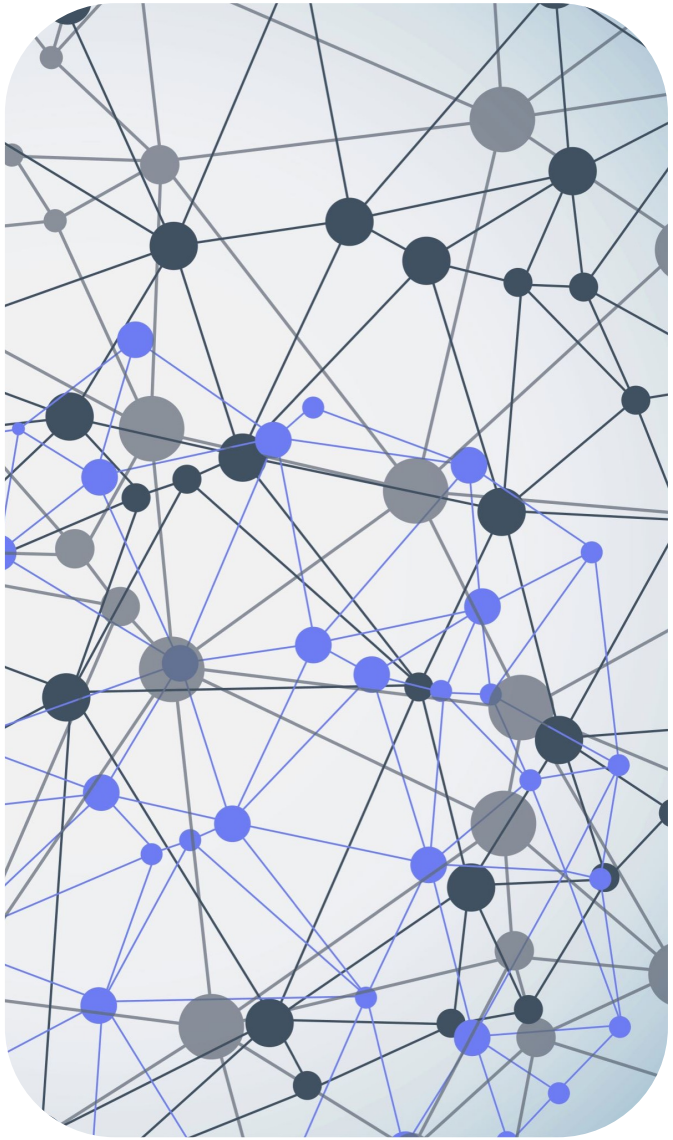
Introduction to AlphaFold2

Shirley (Xue) Li, PhD, Bioinformatician
Research Technology, TTS, Tufts University
xue.li37@tufts.edu
tts-research@tufts.edu

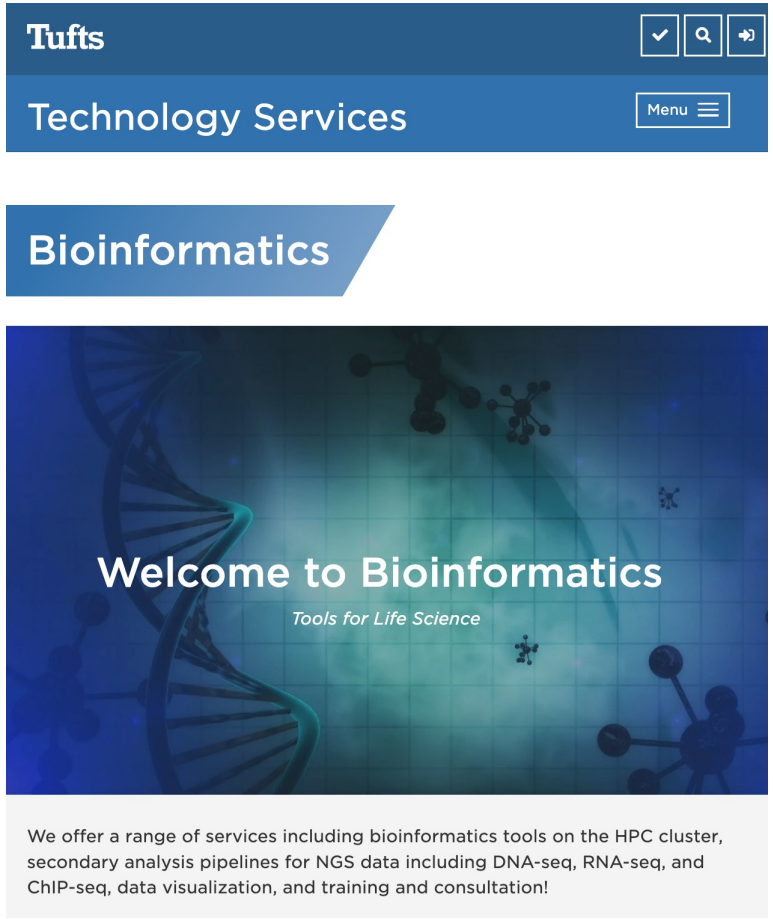


The Research Technology Team

- Consultation on Projects and Grants
- High Performance Cluster Support
- Workshops

https://it.tufts.edu/bioinformatics

https://sites.tufts.edu/datalab/workshops/



Overview

01. The importance of protein structure

Levels of protein organization
Approaches to study protein structure

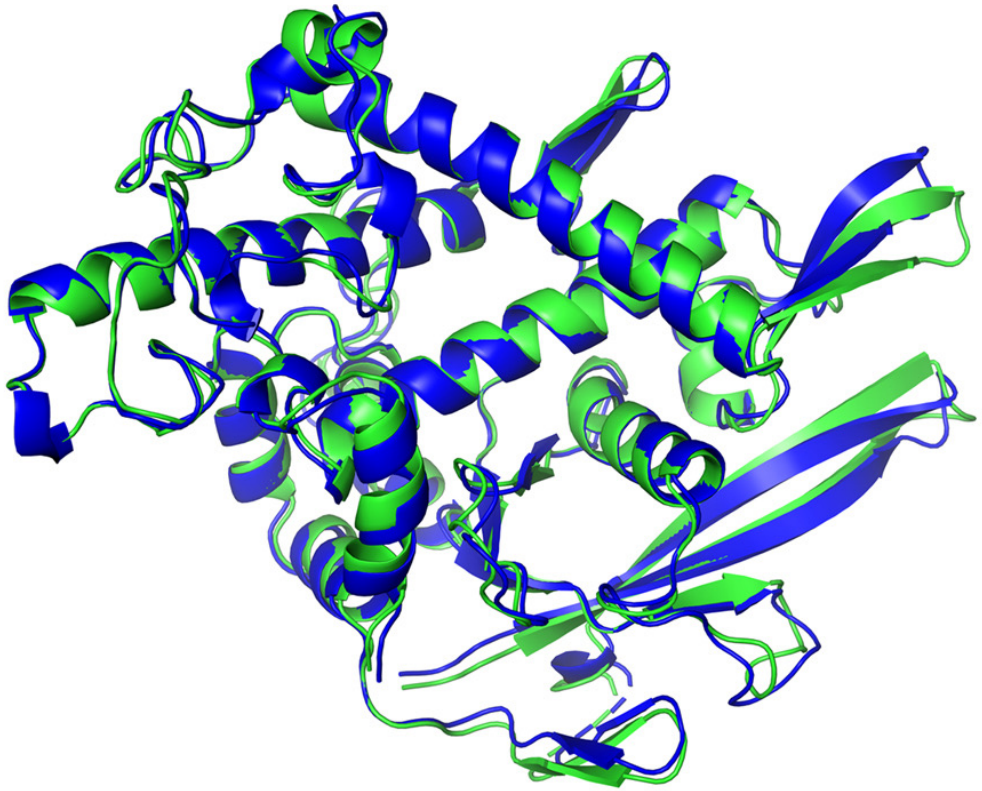
02. Introduction to AlphaFold2

AF architecture

03. Running AlphaFold2 on Tufts server

Open OnDemand
Command Line Interface

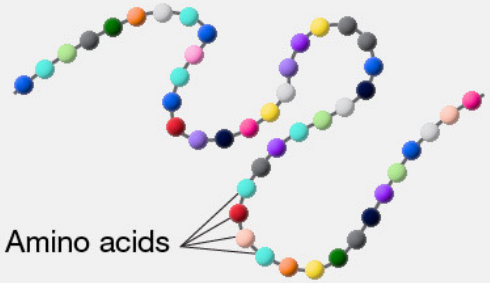
04. PyMOL: Visualizing Protein Structures



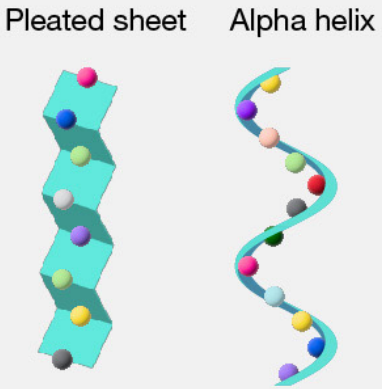
01. The importance of protein structure

Levels of protein structure

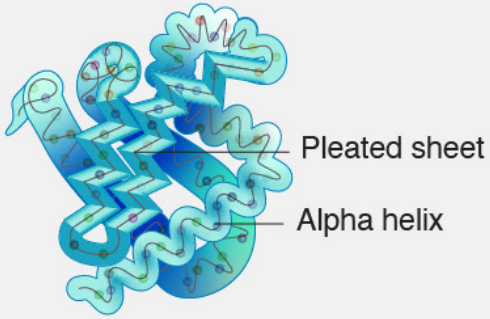
Primary protein structure
is the sequence of a chain of amino acids.



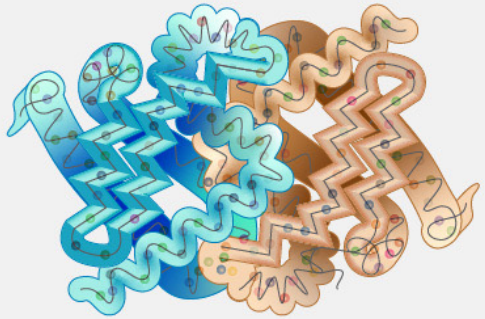
Secondary protein structure
occurs when the sequence of amino acids folds into a three-dimensional shape.



Tertiary protein structure
occurs when a mature protein folds upon itself.



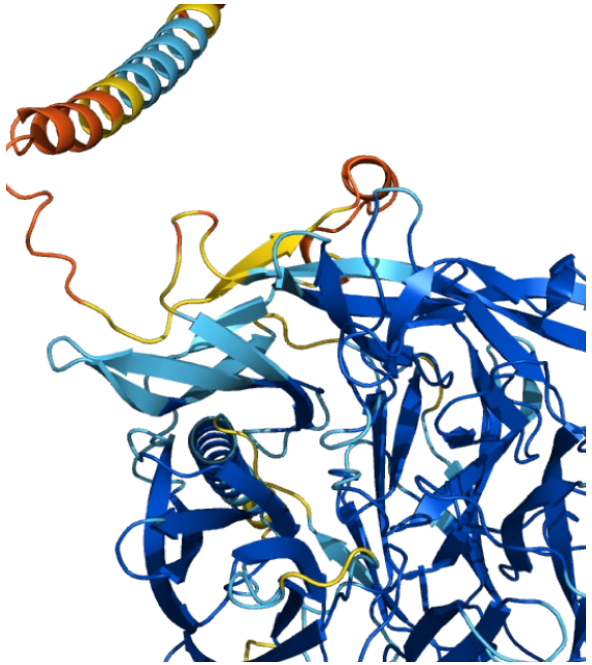
Quaternary protein structure
is a protein consisting of more than one polypeptide chain.



https://www.genome.gov/genetics-glossary/Protein

The importance of protein structure

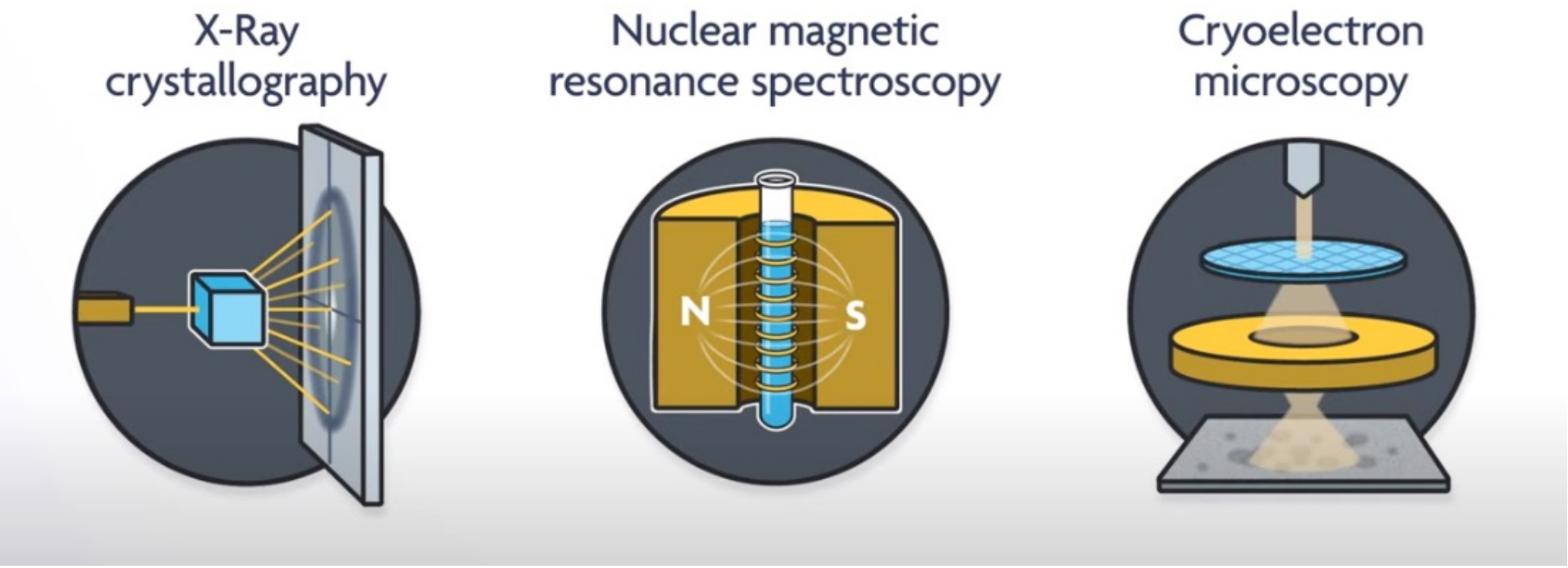
- Function Determination
- Biological Mechanisms
- Disease Understanding
- Protein Engineering
- Drug Design
- Vaccine Development
- … …



Q8I3H7: May protect the malaria parasite against attack by the immune system. Mean pLDDT 85.57.

https://alphafold.ebi.ac.uk/

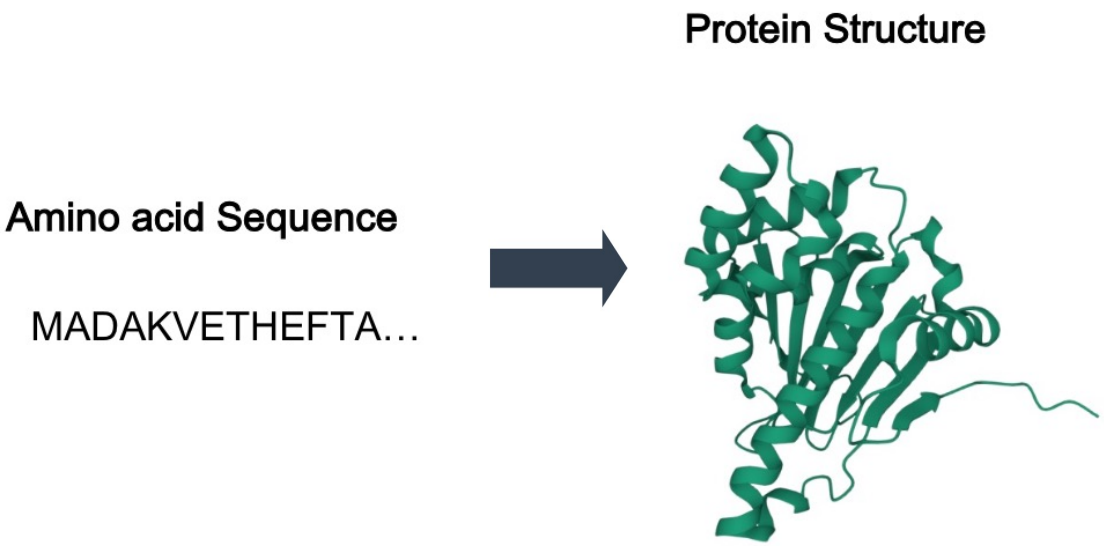
Experimental approaches to study protein structure



https://www.youtube.com/watch?v=7q8Uw3rmXyE

Computational approaches to study protein structure

- Instead of laboratory experimentation, there have been massive efforts to use a protein's sequence to determine structure.
- In 1994, the Critical Assessment of Structure Protein (CASP) was established. It's a scientific even focused on the assessment of protein structure prediction methods.



https://deepmind.google/discover/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology/

Computational approaches to study protein structure

1980s

Homology Modeling

Template-based Modeling

1990s

Rosetta

Ab inito Modeling

2000s

I-TASSER

(Iterative Threading ASSEmbly Refinement)

2010s

Threading/Fold Recognition

Machine Learning Approaches

2018

AlphaFold

(Deep Learning)

2021

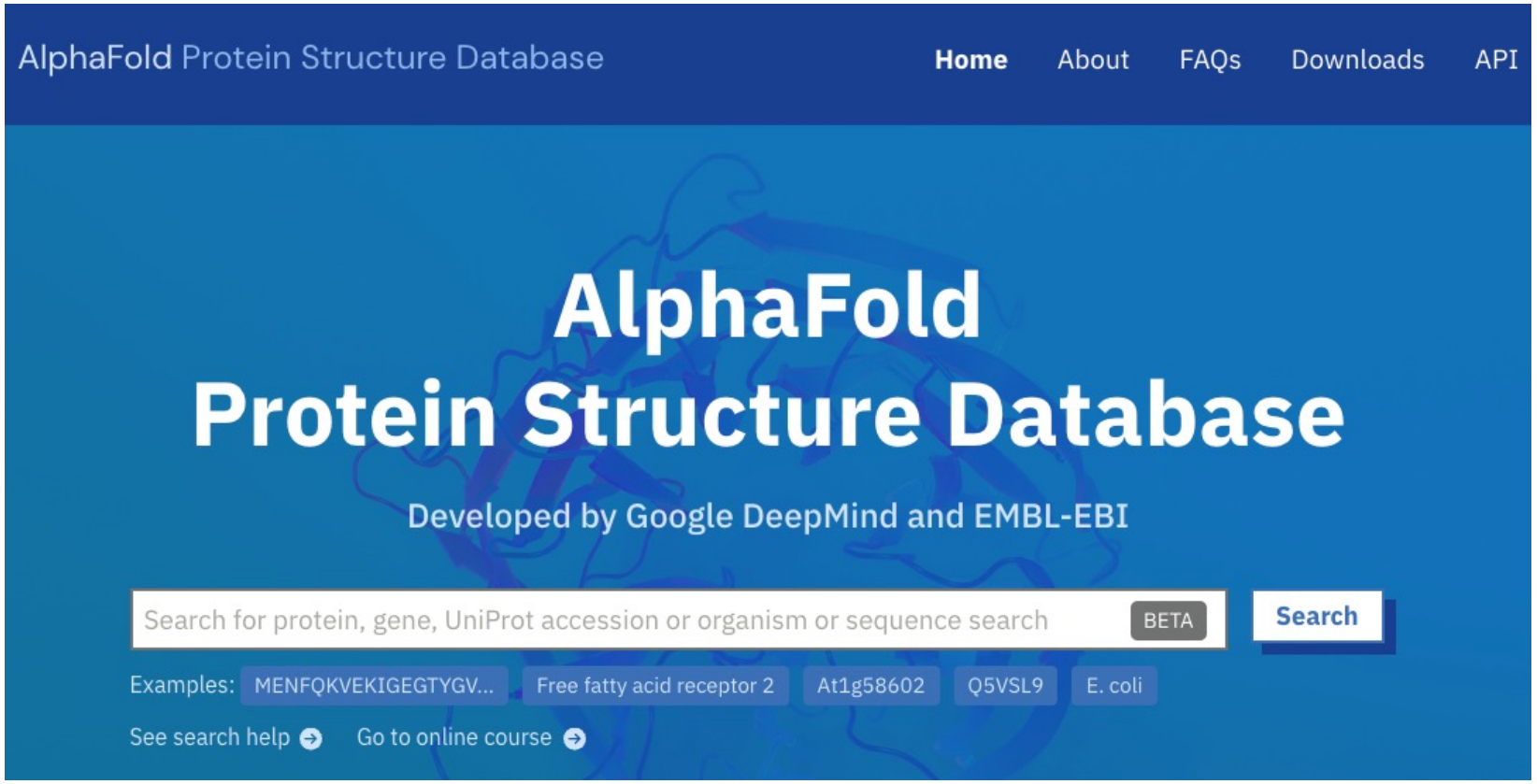
RoseTTAFold

(Deep Learning)

02. Introduction to AlphaFold2

DeepMind's AlphaFold

AlphaFold - Developed by DeepMind, it made groundbreaking progress in 2018 with AlphaFold 1 and then in 2020 with AlphaFold 2, which marked a significant leap in the field.



AlphaFold

Article | Published: 15 January 2020

Improved protein structure prediction using potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu & Demis Hassabis

Nature **577**, 706–710 (2020) | Cite this article

164k Accesses | **1704** Citations | **656** Altmetric | Metrics

AlphaFold2

Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article

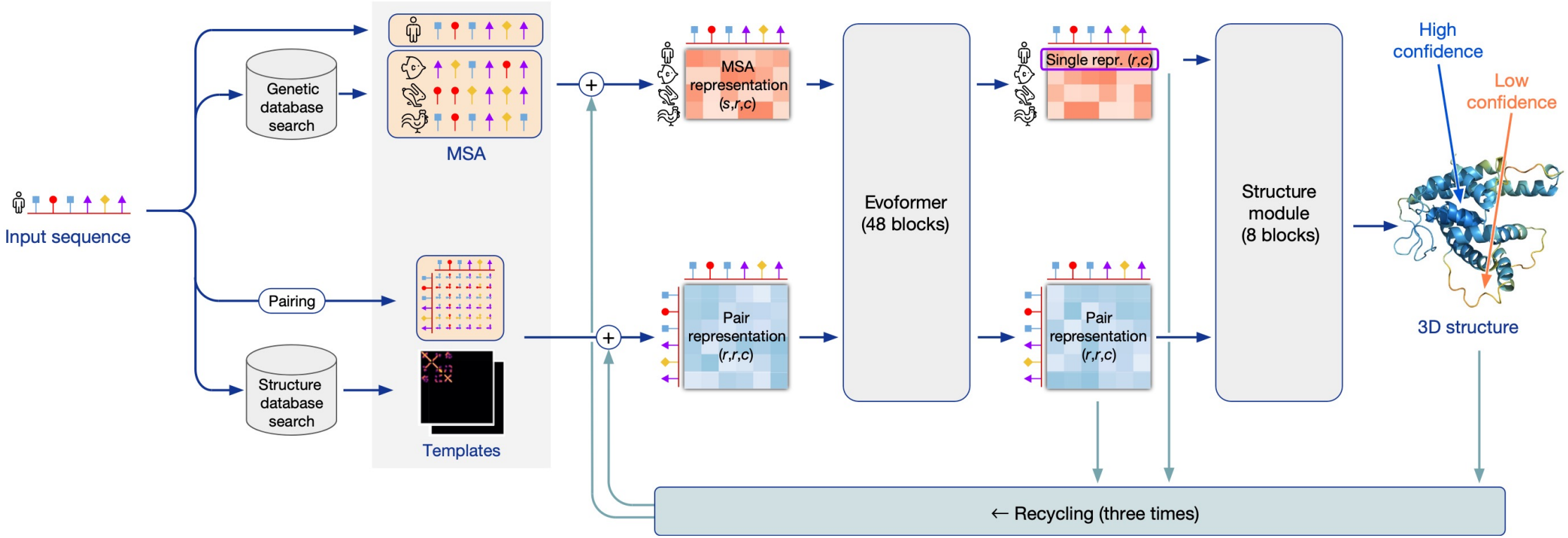
1.47m Accesses | **8815** Citations | **3517** Altmetric | Metrics

AlphaFold vs Other Computational Approaches

- Classical prediction methods require structure templates (e.g. MODELLER, I-TASSER) and they are <u>heavily dependent on sequence homology</u>.
  - These classical methods depend on the alignment of a target protein sequence with other sequences of known structure to infer the target's structure.
- AlphaFold employs deep learning, using a neural network to predict the “distance” and “angles” between residues in a protein, independent of templates.
  - This approach requires significant computational resources due to the complexity of the calculations involved.

AlphaFold 2 Architecture

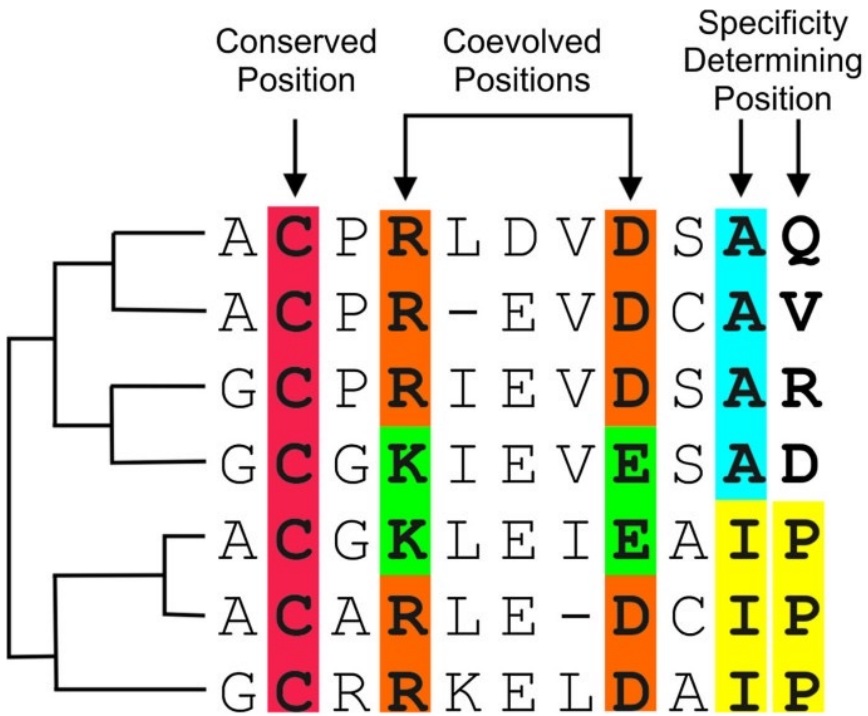
AlphaFold takes only sequence from the user



(Jumper, Evans et al. 2021)

Step 1: Database search and preprocessing

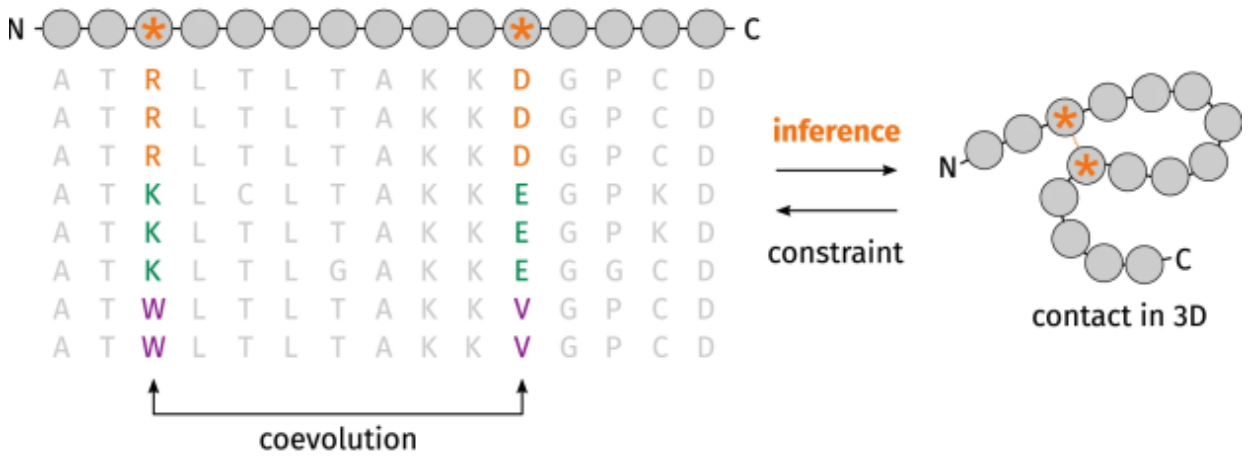
- Protein structural information can be gained by understanding multiple sequence alignments (MSA)
- When we align similar protein sequences we identify:
  - **Conserved positions**: where the letter does not change
  - **Coevolved positions**: where the letter will change with another letter
  - **Specificity determining positions**: where the letter is consistently different



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

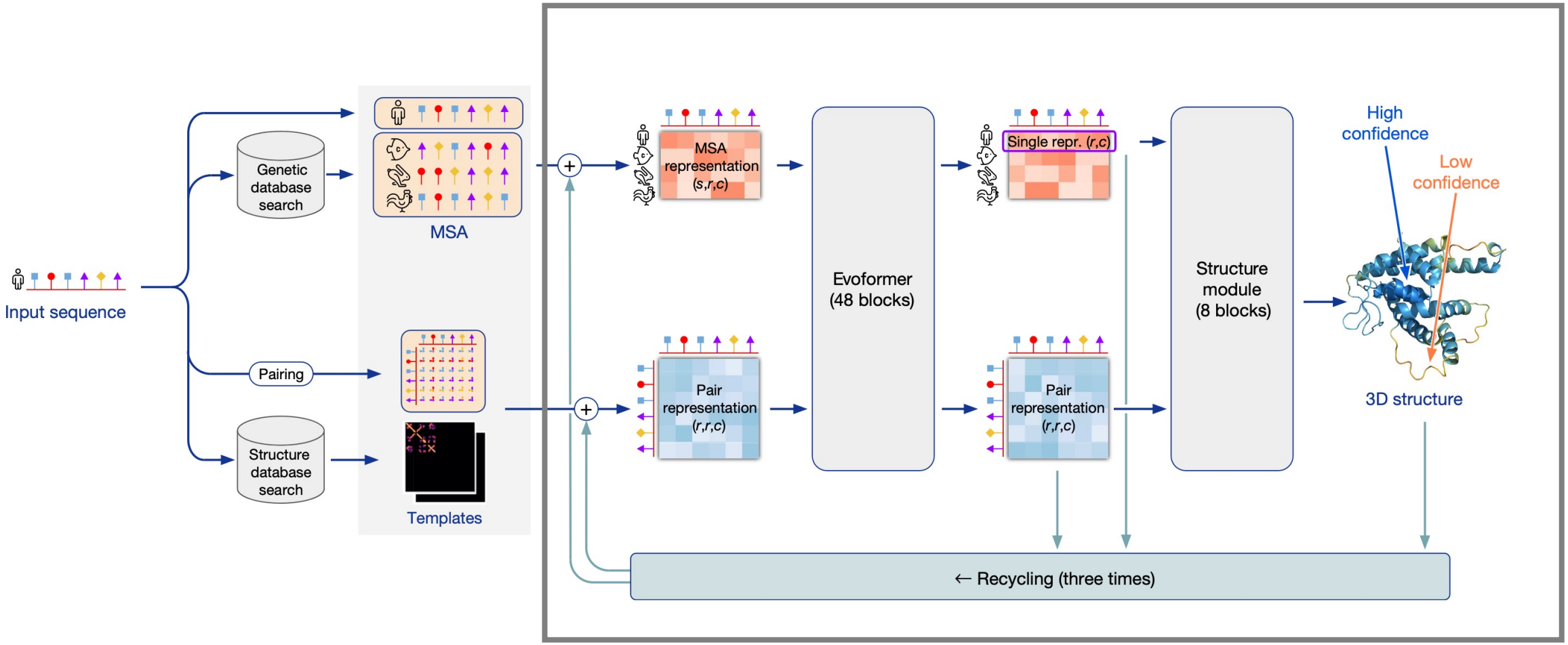
Residue Coevolution

- With an MSA we can identify residues that coevolve, or change together
- We can then reason that residues that change together must be close together in 3D space



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Step 2 & 3 : Evoformer and Structure Module



Read the paper to understand the algorithm

Article | Published: 15 January 2020

Improved protein structure predic
potentials from deep learning

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpa
Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hug
Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, Davic
Hassabis

Nature **577**, 706–710 (2020) | Cite this article

164k Accesses | **1704** Citations | **656** Altmetric | Metrics

Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article

1.47m Accesses | **8815** Citations | **3517** Altmetric | Metrics

AlphaFold represents the state of the art

- Thoroughly validated in competition, but not perfect.
- Not reliable when:
  - Too-sparse MSAs
  - Sequence are not evolutionary
    - Antibody-antigen interface
    - Point mutation studies
  - Large state-dependent structure differences

To download a copy of this slides, please go to

https://go.tufts.edu/chbe0165_af

03. Running AlphaFold on Tufts HPC

Protein Sequence Information

- Protein Sequence information
- Stored as a FASTA file. Consists of:

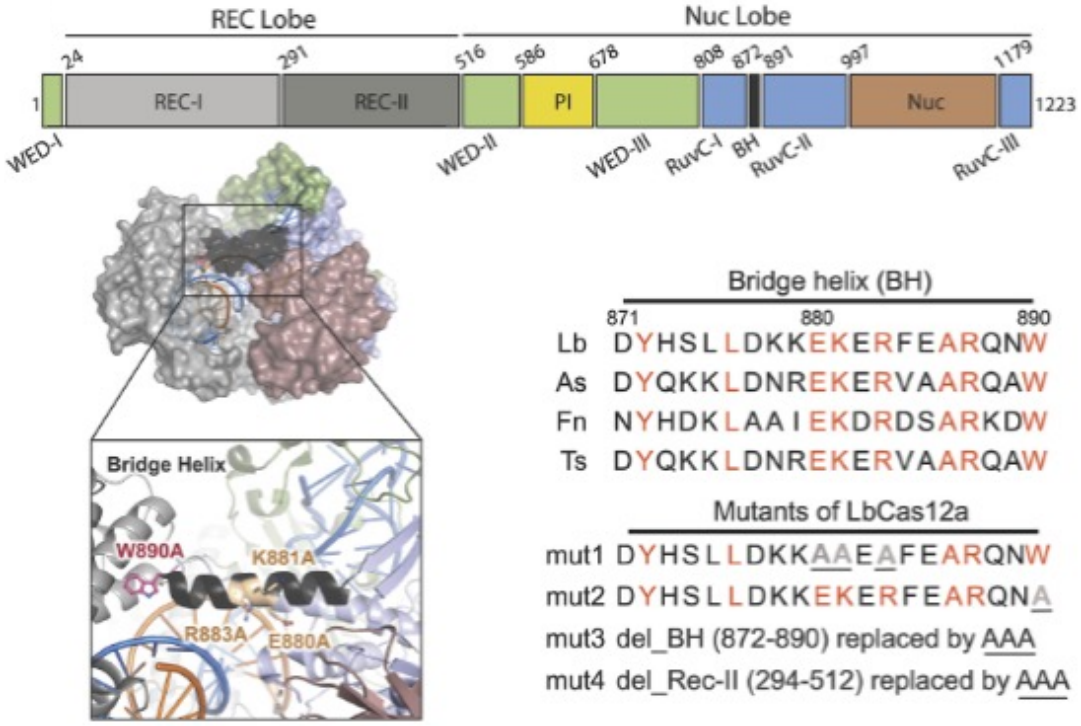
Header

Sequence

```
>sp|P46598|HSP90_CANAL Heat shock protein 90 homolog OS=Candida albicans
(strain SC5314 / ATCC MYA-2876) OX=237561 GN=HSP90 PE=1 SV=1

MADAKVETHEFTAEISQLMSLIINTVYSNKEIFLRELISNASDALDKIRYQALSDPSQLE
SEPELFIRIIPQKDQKVLEIRDSGIGMTKADLVNNLGTIAKSGTKSFMEALSAGADVSMI
GQFGVGFYSLFLVADHVQVISKHNDDEQYVWESNAGGKFTVTLDETNERLGRGTMLRLFL
KEDQLEYLEEKRIKEVVKKHSEFVAYPIQLVVTKEVEKEVPETEE
```

Today's study

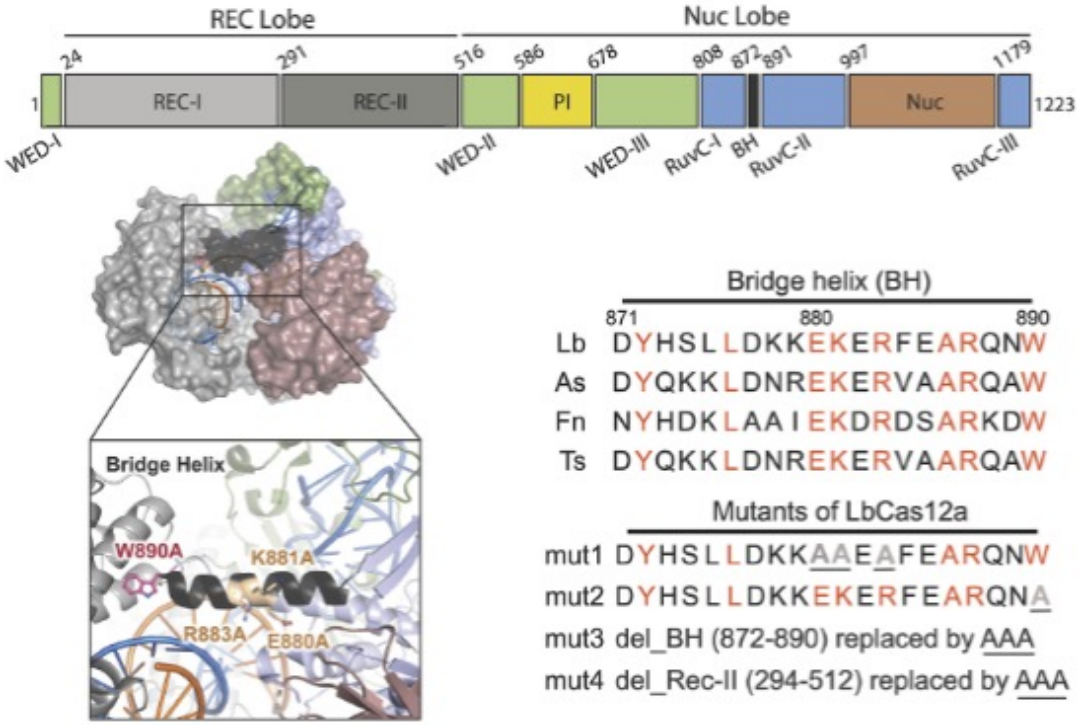
- Today we will be looking at a study by Ma et al. 2022, where they engineer Cas12a variants with reduced trans-activity while maintaining cis-activity
- They start by screening multiple mutants and identify mutant 2 as having reduced trans-activity
- Variants were then introduced in mutant 2 to create a variant with less trans-activity, and maintained cis-activity



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Today’s study

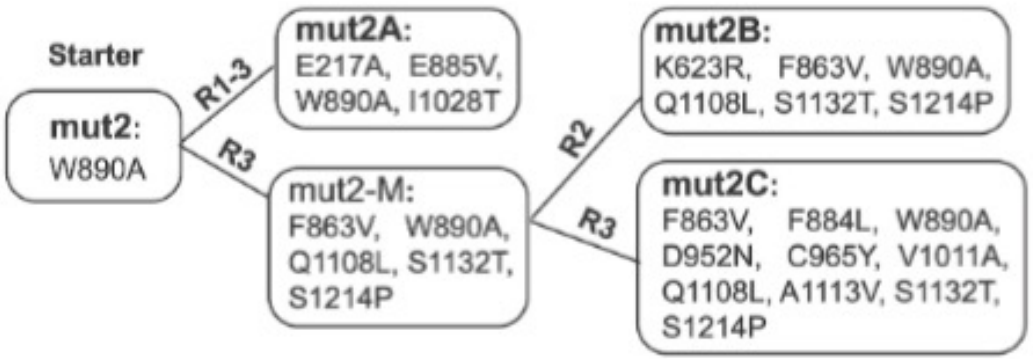
- Cas12a is used for gene editing across various organisms.
- The **cis-activity** of Cas12a refers to its ability to cleave DNA that is directly bound by the complex formed between Cas12a and its crRNA.
- The **trans-cleavage activity** of Cas12a refers to its capability to cut single-stranded DNA (ssDNA) molecules not bound by the Cas12a-crRNA complex, a process initiated upon the enzyme's activation through the recognition and cleavage of its target DNA.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Variant Structure Prediction with AlphaFold2

- Three variants were ultimately refined: mut2B-W, mut2C-W, and mut2C-WF
- We will use AlphaFold2 to predict the structure of mut2C-WF



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Variant Structure Prediction with AlphaFold2

- Three variants were ultimately refined: mut2B-W, mut2C-W, and mut2C-WF
- We will use AlphaFold2 to predict the structure of mut2C-WF

mut2C-WF

F863V, ~~F884L, W890A,~~
D952N, C965Y, V1011A,
Q1108L, A1113V, S1132T,
S1214P

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Cas12a protein (previously named Cpf1)

RCSB PDB PROTEIN DATA BANK

217,387 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

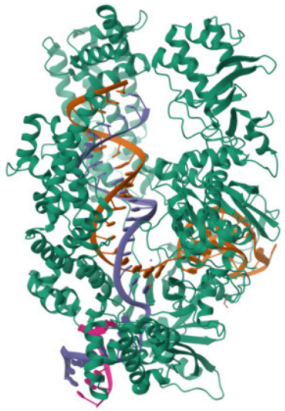
3D Structures | Enter search term(s), Entry ID(s), or sequence | Include CSM

Advanced Search | Browse Annotations | Help

PDB-101 | wwPDB | EMDataResource | NAKB | wwPDB Foundation | PDB-Dev

Structure Summary | Structure | Annotations | Experiment | Sequence | Genome | Versions

Biological Assembly 1



Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (EDO)

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Monomer - A1

Find Similar Assemblies

Display Files | Download Files | Data API

5XUS

Crystal structure of Lachnospiraceae bacterium ND2006 Cpf1 in complex with crRNA and target DNA (TTTA PAM)

PDB DOI: https://doi.org/10.2210/pdb5XUS/pdb **NAKB:** 5XUS

Classification: HYDROLASE/RNA/DNA

Organism(s): Lachnospiraceae bacterium ND2006, synthetic construct

Expression System: Escherichia coli

Mutation(s): No

Deposited: 2017-06-26 **Released:** 2017-08-09

Deposition Author(s): Yamano, T., Nishimasu, H., Ishitani, R., Nureki, O.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

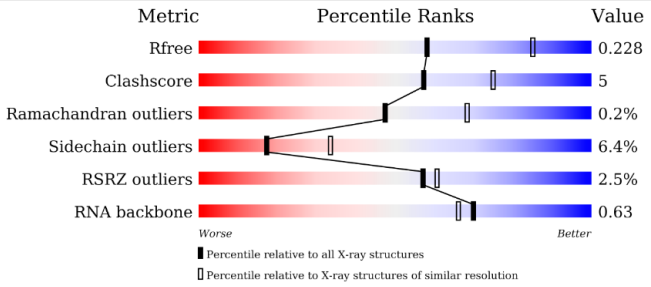
Resolution: 2.50 Å

R-Value Free: 0.228

R-Value Work: 0.178

R-Value Observed: 0.181

wwPDB Validation | 3D Report | Full Report



AA sequence of mut2C-WF

F863V, ~~F884L, W890A,~~
D952N, C965Y, V1011A,
Q1108L, A1113V, S1132T,
S1214P

```
>5XUS_1|Chain A|LbCpf1_mut2cwf|Lachnospiraceae bacterium ND2006 (1410628)
MSKLEKFTNCYSLSKTLRFKAIPVGKTQENIDNKRLLVEDEKRAEDYKGVKKLLDRYYLSFINDVLHSIKLKNLNNYISLFRKKTRTEKENKELENLEINLRKEIAKAF
KGNEGYKSLFKKDIIETILPEFLDDKDEIALVNSFNGFTTAFTGFFDNRENMFSEEAKSTSIAFRCINENLTRYISNMDIFEKVDAIFDKHEVQEIKEKILNSDYDVED
FFEGEFFNFVLTQEGIDVYNAIIGGFVTESGEKIKGLNEYINLYNQKTKQKLPKFKPLYKQVLSDRESLSFYGEGYTSDEEVLEVFRNTLNKNSEIFSSIKKLEKLFKN
FDEYSSAGIFVKNGPAISTISKDIFGEWNVIRDKWNAEYDDIHLKKKAVVTEKYEDDRRKSFKKIGSFSLEQLQEYADADLSVVEKLKEIIIQ[illegible]LF
DADFVLEKSLKKNDAVVAIMKDLLDSVKSFENYIKAFFGEGKETNRDESFYGDFVLAYDILLKVDHIYDAIRNYVTQKPYSKDKFKLYFQNPQ[illegible]AT
ILRYGSKYYLAIMDKKYAKCLQKIDKDDVNGNYEKINYKLLPGPNKMLPKVFFSKKWMAYYNPSEDIQKIYKNGTFKKGDMFNLNDCHKLIDF[illegible]YD
FNFSETEKYKDIAGFYREVEEQGYKVSFESASKKEVDKLVEEGKLYMFQIYNKDFSDKSHGTPNLHTMYFKLLFDENNHGQIRLSGGAELFMRRASLKKEELVVHPANS
PIANKNPDNPKKTTTLSYDVYKDKRFSEDQYELHIPIAINKCPKNIFKINTEVRVLLKHDDNPYVIGIDRGERNLLYIVVVDGKGNIVEQYSLNEIINNVNGIRIKTDY
HSLLDKKEKERFEARQNWTSIENIKELKAGYISQVVHKICELVEKYDAVIALEDLNSGFKNSRVKVEKQVYQKFEKMLINKLNYMVDKKSNPYATGGALKGYQITNKFE
SFKSMSTQNGFIFYIPAWLTSKIDPSTGFANLLKTKYTSIADSKKFISSFDRIMYVPEEDLFEFALDYKNFSR[illegible]IRIFRNPKKNNVFDWEEVC
LTSAYKELFNKYGINYQLGDIRVLLCEQSDKAFYSSFMALMTLMLQMRNSITGRTDVDFLISPVKNSDGIFYD[illegible]ANGAYNIARKVLWAIGQFK
KAEDEKLDKVKIAIPNKEWLEYAQTSVKH
```

F863V

D952N

Running Alphafold2

Hardware Requirements

GPU: It requires NVIDIA GPUs with CUDA support, and for optimal performance, it's recommended to use a high-performance GPU such as the NVIDIA A100, V100, or at least a T4 or RTX 2080 Ti for smaller proteins.

CPU: A modern multi-core CPU (e.g., 8 cores or more) is important for efficient data processing.

Memory (RAM): The amount of system memory required can vary. For predicting structures of individual proteins (monomers), at least 16 GB of RAM is recommended, but 32 GB or more may be required for larger proteins or for multimer predictions.

Computational Time

The time it takes to run a prediction can vary from a few hours to several days, depending on:

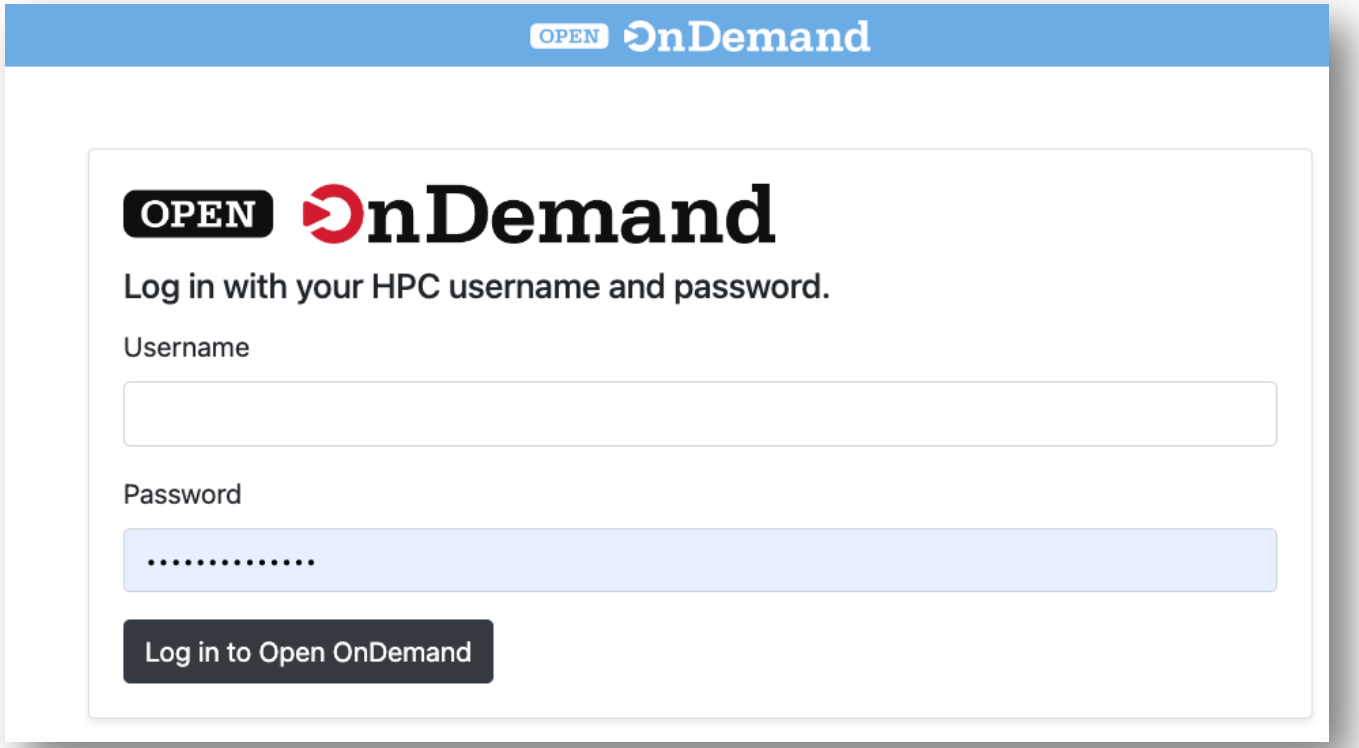
- The complexity of the protein or protein complex.
- The model_preset used (monomer vs. multimer).
- The performance of the hardware, especially the GPU.

Accessing AlphaFold2 on Tufts HPC

- Command Line Interface (CLI)

```
xli37@login-prod-01:~>module load alphafold/2.3.2
xli37@login-prod-01:~>
```

- Open OnDemand



Run AlphaFold2 on Tufts HPC

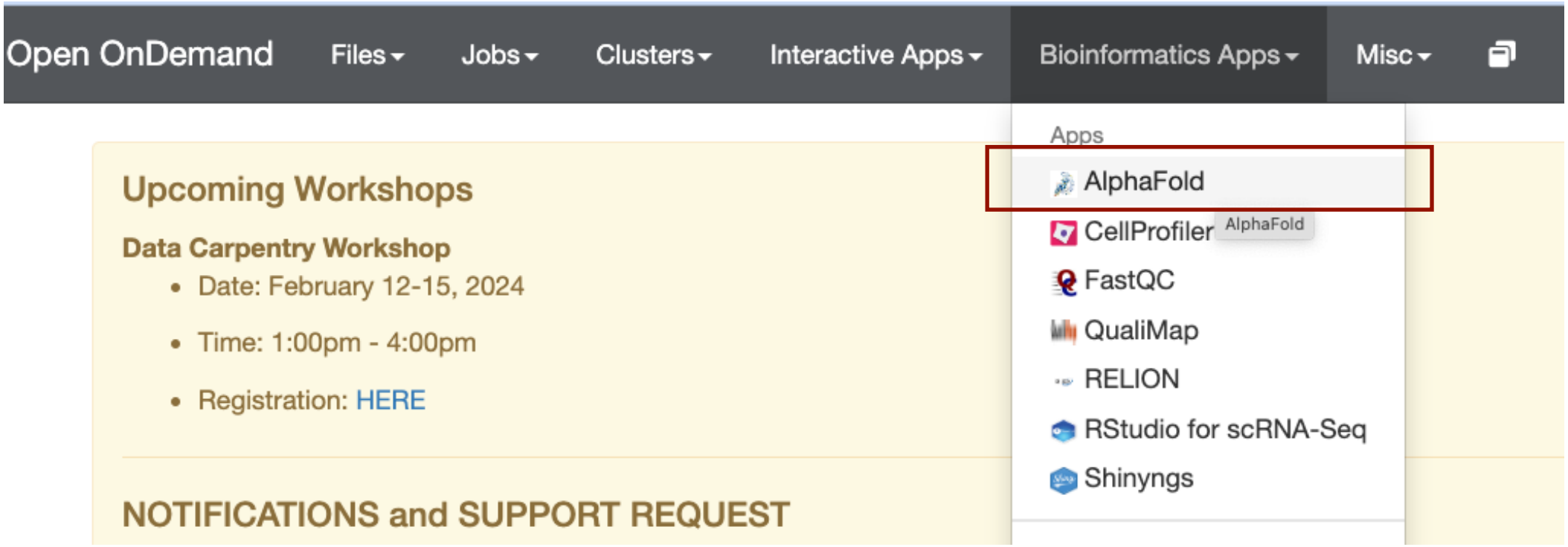
Example script is provided

/cluster/tufts/bio/tools/training/cas12a_af2_sp24/script/runaf.sh

```
#!/bin/bash
#SBATCH -p gpu
#SBATCH -n 8
#SBATCH --mem=64g
#SBATCH --time=2-0
#SBATCH -o output.%j
#SBATCH -e error.%j
#SBATCH -N 1
#SBATCH --gres=gpu:a100:1

# Load the AlphaFold2 and NVIDIA modules
module load alphafold/2.3.2
nvidia-smi

# Make the results directories
mkdir /cluster/home/xli37/cas12a_af2_sp24/out/

# Specify where your output directories and raw data are
outputpath1=/cluster/home/xli37/cas12a_af2_sp24/out/
fastapath=/cluster/home/xli37/cas12a_af2_sp24/5XUS_mut2cwf_modified.fasta

# Date to specify if you want to avoid using template
maxtemplatedate1=2020-01-01

run_alphafold.sh --output_dir=$outputpath1 \
        --fasta_paths=$fastapath \
        --max_template_date=$maxtemplatedate1 \
        --model_preset=multimer \
        --models_to_relax=best \
        --data_dir=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/ \
        --uniref90_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/uniref90/uniref90.fasta \
        --mgnify_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/mgnify/mgy_clusters_2022_05.fa \
        --pdb_seqres_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/pdb_seqres/pdb_seqres.txt \
        --template_mmcif_dir=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/pdb_mmcif/mmcif_files \
        --max_template_date=2022-01-01 \
        --obsolete_pdbs_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/pdb_mmcif/obsolete.dat \
        --use_gpu_relax=True \
        --bfd_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
        --uniref30_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/uniref30/UniRef30_2021_03 \
        --uniprot_database_path=/cluster/tufts/biocontainers/datasets/alphafold/db_20231031/uniprot/uniprot.fasta
```

Running AlphaFold2 with Open OnDemand

https://ondemand.pax.tufts.edu



AlphaFold

This app will launch AlphaFold. More information about AlphaFold can be found here (https://github.com/deepmind/alphafold).

Number of hours

24

Number of cores

8

Numbers can be changed based on the size of your protein

Amount of memory

32GB

Select preempt or normal gpu partition

gpu

NOTE: jobs submitted to the preempt partition may get automatically killed to allow higher priority jobs to run

NOTE: jobs submitted to the preempt partition may get automatically killed to allow higher priority jobs to run

Select the GPU type

a100

Software Version

2.3.2

Database

20231031

Working Directory

Change it to your own working directory

/cluster/home/tutln02/cas12a_af2_sp24/

Select your project directory; defaults to $HOME

Output directory Name

/cluster/home/tutln02/cas12a_af2_sp24/

Change it to your own output directory

Where the results will be going to (relative to the working directory field above). Example: alphafold.out

fasta_paths

Input file. Fasta format.

/cluster/home/tutln02/cas12a_af2_sp24/5XUS_mut2cwf_modified.fasta

The fasta files containing amino acid sequence(s) to fold. If there are more multiple files, please separate them using comma(e.g. seq1.fasta,seq2.fasta)

model_preset

Let's use multimer for now

multimer

Select to run the monomer or multimer model for sequences.

models_to_relax

best

After generating the predicted model, AlphaFold runs a relaxation step to improve local geometry. By default, only the best model (by pLDDT) is relaxed (--models_to_relax=best), but also all of the models (--models_to_relax=all) or none of the models (--models_to_relax=none) can be relaxed.

num_multimer_predictions_per_model

1

How many predictions (each with a different random seed) will be generated per model. E.g. if this is 2 and there are 5 models then there will be 10 predictions per input. Note: this FLAG only applies if model_preset=multimer. (default: 5).

max_template_date

2020-01-01

Maximum template release date to c
DD).Important if folding historical tes

This parameter is crucial for benchmarking and studies, ensuring predictions replicate original conditions without using future knowledge unavailable at the study time.
It can be any past date.

This date acts as a cutoff, meaning that only protein templates solved on or before this date will be considered during the structure prediction process.

Extra parameters

Extra parameters to use. Multiple space-separated parameters can be used.

Launch

* The AlphaFold session data for this session can be accessed under the data root directory.

Output

```
4.8M Dec  6 17:26 features.pkl
4.0K Dec  6 17:26 msas
125K Dec  6 17:29 unrelaxed_model_1_ptm.pdb
125K Dec  6 17:32 unrelaxed_model_2_ptm.pdb
125K Dec  6 17:34 unrelaxed_model_3_ptm.pdb
125K Dec  6 17:35 unrelaxed_model_4_ptm.pdb
125K Dec  6 17:37 unrelaxed_model_5_ptm.pdb
243K Dec  6 17:29 relaxed_model_1_ptm.pdb
243K Dec  6 17:32 relaxed_model_2_ptm.pdb
243K Dec  6 17:34 relaxed_model_3_ptm.pdb
243K Dec  6 17:36 relaxed_model_4_ptm.pdb
243K Dec  6 17:37 relaxed_model_5_ptm.pdb
243K Dec  6 17:37 ranked_0.pdb
243K Dec  6 17:37 ranked_1.pdb
243K Dec  6 17:37 ranked_2.pdb
243K Dec  6 17:37 ranked_3.pdb
243K Dec  6 17:37 ranked_4.pdb
 29M Dec  6 17:29 result_model_1_ptm.pkl
 29M Dec  6 17:32 result_model_2_ptm.pkl
 29M Dec  6 17:34 result_model_3_ptm.pkl
 29M Dec  6 17:35 result_model_4_ptm.pkl
 29M Dec  6 17:37 result_model_5_ptm.pkl
 829 Dec  6 17:37 timings.json
 370 Dec  6 17:37 ranking_debug.json
```

the MSA information, processed to be given to AlphaFold as input (+ more)

subdirectory with MSA information, in human-readable format

the (not yet relaxed) predictions from the five AlphaFold models.

the relaxed versions of the predicted structures

the relaxed versions of the predicted structures, but now ranked based on pLDDT (with highest pLDDT for ranked_0.pdb)

extra outputs in .pkl format. Contains a lot of information, including pLDDT and PTM values

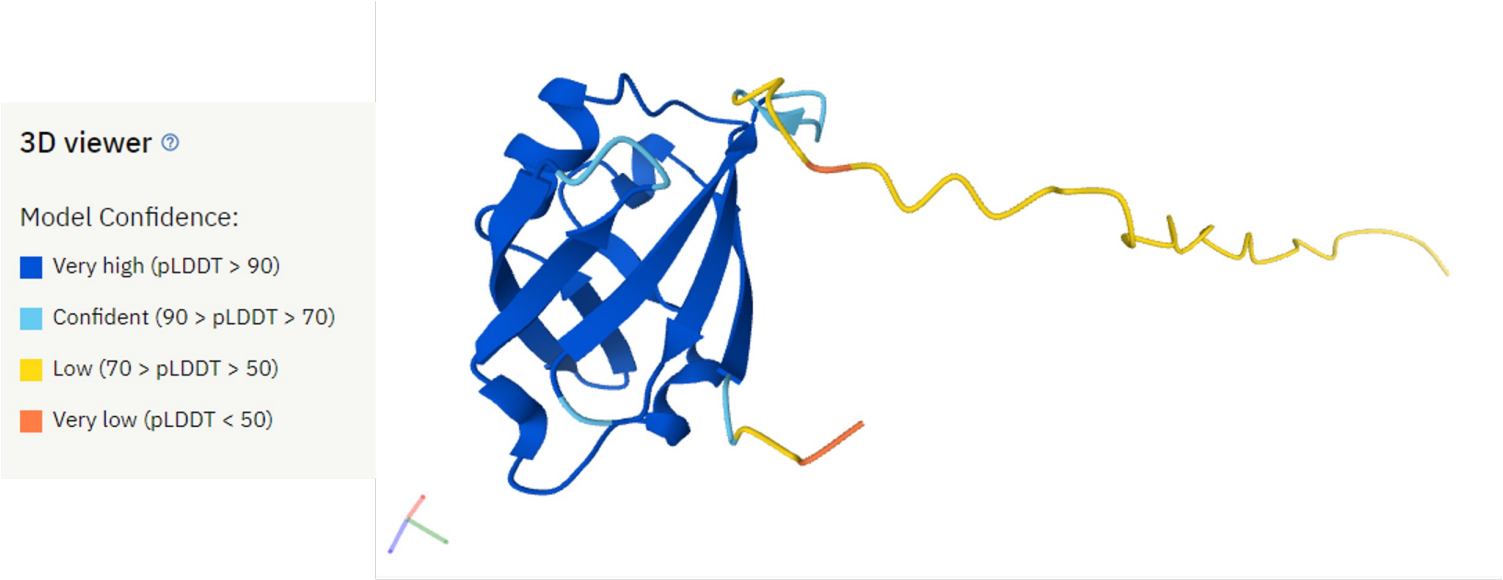
information on how long the different parts of the AlphaFold run took, in seconds

information on the pLDDT of each model, and how they were ranked

https://elearning.vib.be/courses/alphafold/lessons/alphafold-on-the-hpc/topic/alphafold-outputs/

AlphaFold2 Accuracy

Predicted Local Distance Difference Test



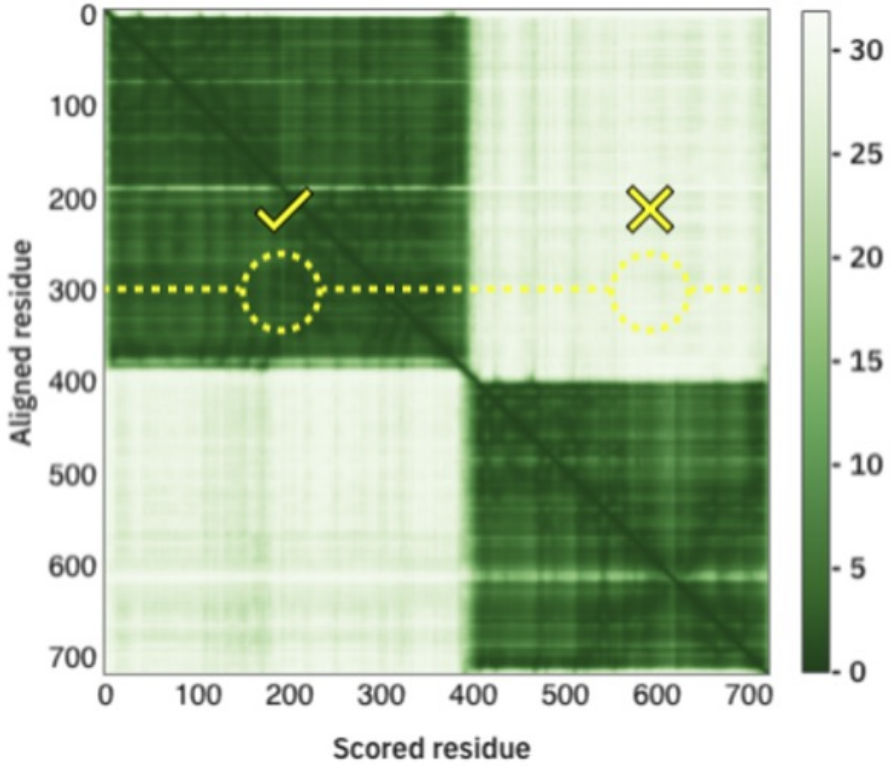
- The Predicted Local Distance Difference Test (pLDDT) is a per-residue confidence metric ranging from 0-100 (100 being the highest confidence)
- Regions below 50 could indicate disordered regions

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

AlphaFold2 Accuracy

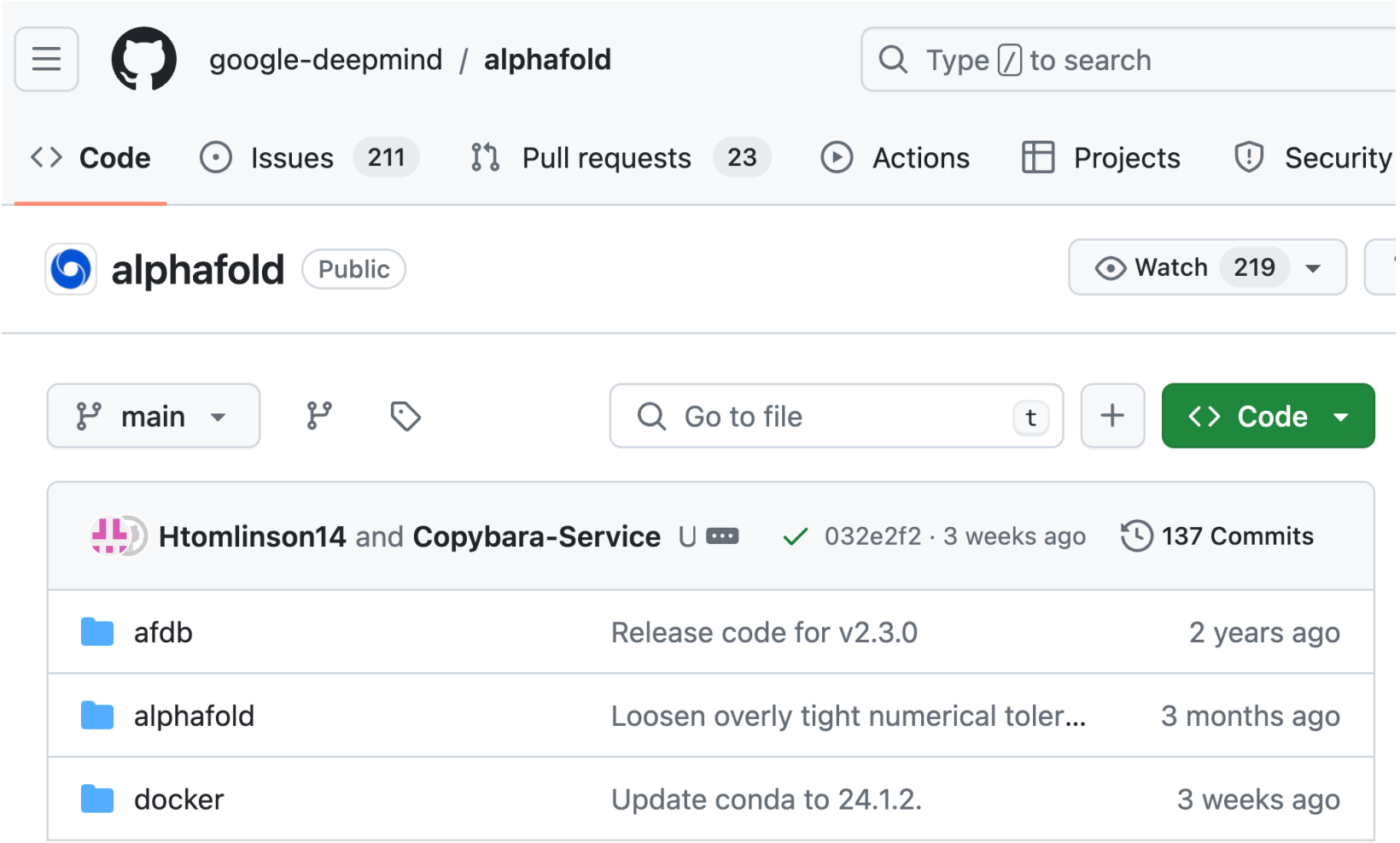
Predicted Alignment Error

- The Predicted Alignment Error (PAE) gives us an expected distance error based on each residue.
- If we are more confident that the distance between two residues is accurate, then the PAE is lower (darker green). If we are less confident that the distance between two residues is accurate, the PAE is higher (lighter green)



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

Github page for AlphaFold

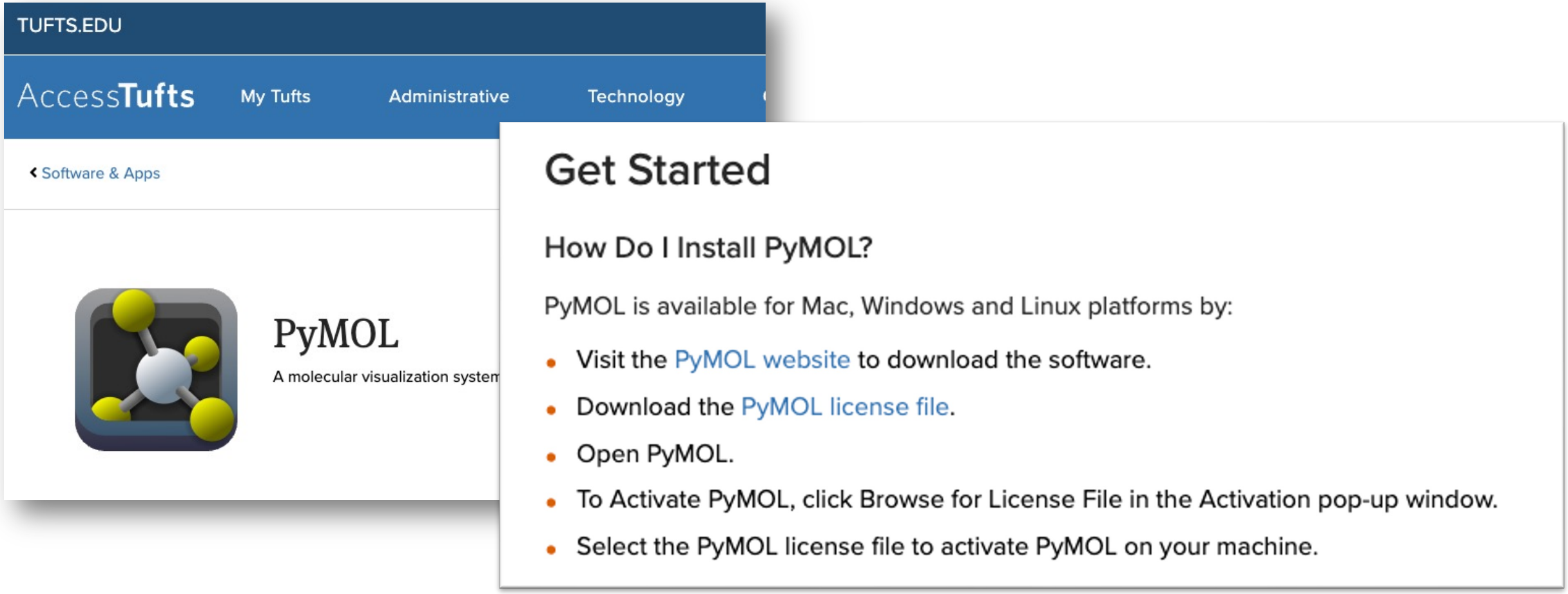


https://github.com/google-deepmind/alphafold/?tab=readme-ov-file#running-alphafold

04. PyMOL: Visualizing Protein Structures

Pymol is accessible for free with Tufts credentials

https://access.tufts.edu/pymol



PyMOL

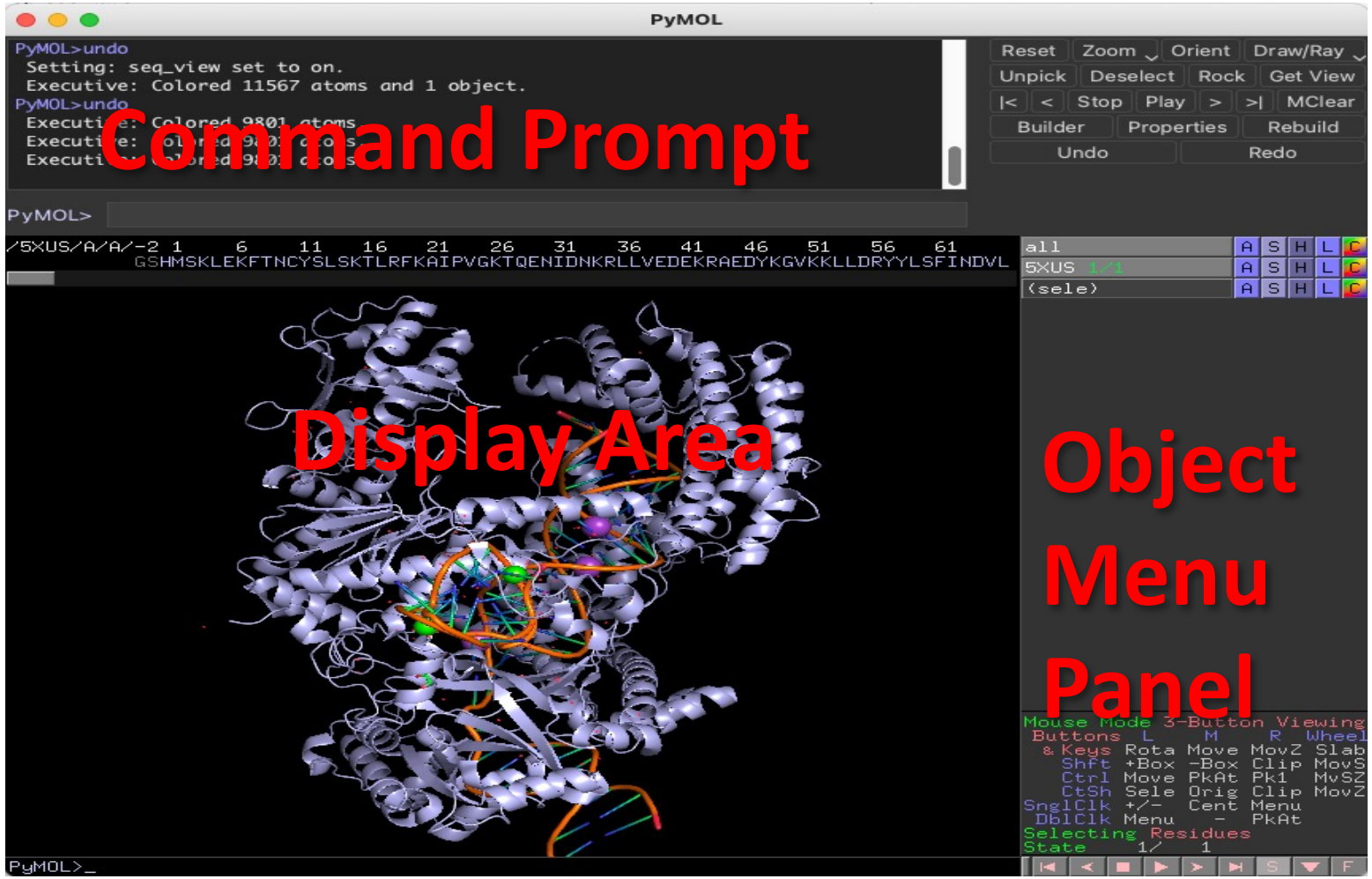
https://www.rcac.purdue.edu/files/training/AlphaFold_Protein_Structure_Prediction.pdf

Molecular visualization software

- Given atomic coordinate or volumetric data
- X-ray, NMR, EM, AlphaFold, etc.
- Generates an interactive visualization
- Can render and save publication-quality images and videos.

PyMOL

https://www.rcac.purdue.edu/files/training/AlphaFold_Protein_Structure_Prediction.pdf



Misc Controls

Misc Controls

Pymol Reference Card

https://pymolwiki.org/images/7/77/PymolRef.pdf

Pymol Reference Card

Modes

Pymol supports two modes of input: point and click mode, and command line mode. The point and click allows you to quickly rotate the molecule(s) zoom in and out and change the clipping planes. The command line mode where commands are entered into the external GUI window supports all of the commands in the point and click mode, but is more flexible and possibly useful for complex selection and command issuing. Commands entered on the command line are executed when you press the return key.

command help `help` *`keyword`*

Pymol Reference Card

https://pymolwiki.org/images/7/77/PymolRef.pdf

Pymol Reference Card

Modes

Pymol supports two modes of input: point and click mode, and command line mode. The point and click allows you to quickly rotate the molecule(s) zoom in and out and change the clipping planes. The command line mode where commands are entered into the external GUI window supports all of the commands in the point and click mode, but is more flexible and possibly useful for complex selection and command issuing. Commands entered on the command line are executed when you press the return key.

command help	`help` *keyword*

Loading Files

file loading	`load data/test/pept.pdb`
loading from terminal	`pymol data/test/pept.pdb`
toggle between text and graphics	*Esc*
toggle Y axis rocking	`rock`
stereo view	`stereo on/off`
stereo type	`stereo crosseye / walleye / quadbuffer`
undo action	`undo`
reset view	`reset`
reinitialize Pymol	`reinitialize`
quit (force, even if unsaved)	`quit`

Mouse Control

	L Rota	M Move	R MovZ	Wheel Slab
Shift	+Box	-Box	Clip	MovS
Ctrl	+/-	PkAt	Pk1	—
CtSh	Sele	Cent	Menu	—
DblClk	Menu	Cent	PkAt	—

set the center of rotation	`origin` *selection*

Atom Selection

object-name/segi-id/chain-id/resi-id/name-id

molecular system selection	`/pept`
molecule selection	`/pept/lig`
chain selection	`/pept/lig/a`
residue selection	`/pept/lig/a/10`
atom	`/pept/lig/a/10/ca`
ranges	`lig/a/10-12/ca`
ranges	`a/6+8/c+o`
missing selections	`/pept//a`
naming a selection	`select bb, name c+o+n+ca`
count atoms in a selection	`count_atoms bb`
remove atoms from a selection	`remove resi 5`
general	`all, none, hydro, hetatm, visible, present`
atoms not in a selection	`select sidechains, ! bb`
atoms with a vdW gap < 3 Å	`resi 6 around 3`
atom centers with a gap < 1.0 Å	`all near 1 of resi 6`
atom centers within < 4.0 Å	`all within 4 of resi 6`

Basic Commands

Some commands used with atoms selections. If you are unsure about the selection, click on the molecule part that you want in the viewing window and then look at the output line to see the selection.

fill viewer with selection	`zoom /pept//a`
center a selection	`center /pept//a`
colour a selection	`colour pink, /pept//a`
force Pymol to reapply colours	`recolor`
set background colour	`bg_color white`
vdW representation of selection	`show spheres, 156/ca`
stick representation of selection	`show sticks, a//`
line representation of selection	`show lines, /pept`
ribbon representation of selection	`show ribbon, /pept`
dot representation of selection	`show dots, /pept`
mesh representation of selection	`show mesh, /pept`
surface representation of selection	`show surface, /pept`
nonbonded representation of selection	`show nonbonded, /pept`
nonbonded sphere representation of selection	`show nb_spheres, /pept`
cartoon representation of selection	`show cartoon, a//`
clear all	`hide all`
rotate a selection	`rotate` *axis, angle, selection*
translate a selection	`translate [x,y,z],` *selection*

Cartoon Settings

Setting the value at the end to 0 forces the secondary structure to go though the CA position.

cylindrical helices	`set cartoon_cylindrical_helices,1`
fancy helices [tubular edge]	`set cartoon_fancy_helices,1`
flat sheets	`set cartoon_flat_sheets,1`
smooth loops	`set cartoon_smooth_loops,1`
find rings for cartoon	`set cartoon_ring_finder,[1,2,3,4]`
ring mode	`set cartoon_ring_mode,[1,2,3]`
nucleic acid mode	`set nucleic_acid_mode,[0,1,2,3,4]`
cartoon sidechains	`set cartoon_side_chain_helper; rebuild`
primary colour	`set cartoon_color,blue`
secondary colour	`set cartoon_highlight_color,grey`
limit colour to ss	`set cartoon_discrete_colors,on`
cartoon transparency	`set cartoon_transparency,0.5`
cartoon loop	`cartoon loop, a//`
cartoon loop	`cartoon loop, a//`
cartoon rectangular	`cartoon rect, a//`
cartoon oval	`cartoon oval, a//`
cartoon tubular	`cartoon tube, a//`
cartoon arrow	`cartoon arrow, a//`
cartoon dumbell	`cartoon dumbell, a//`
b-factor sausage	`cartoon putty, a//`

Image Output

low resolution	`ray`
high resolution	`ray 2000,2000`
ultra-high resolution	`ray 5000,5000`
change the default size [pts]	`viewport 640,480`
image shadow control	`set ray_shadow,0`
image fog control	`set ray_trace_fog,0`
image depth cue control	`set depth_cue,0`
image antialiasing control	`set antialias,1`
export image as .png	`png` *image*`.png`

Hydrogen Bonding

Draw bonds between atoms and label the residues that are involved.

draw a line between atoms	`distance 542/oe1,538/ne`
set the line dash gap	`set dash_gap,0.09`
set the line dash width	`set dash_width,3.0`
set the line dash radius	`set dash_radius,0.0`
set the line dash length	`set dash_length,0.15`
set round dash ends	`set dash_round_ends,on`
hide a label	`hide labels, dist01`
label a reside	`label (542/oe1), "%s" %("E542")`
set label font	`set label_font_id,4`
set label colour	`set label_color,white`

Electrostatics

There are a number of ways to apply electrostatics in Pymol. The user can use GRASP to generate a map and then import it. Alternatively the user can use the APBS Pymol plugin. Pymol also has a built in function that is quick and dirty.

generate electrostatic surface `action > generate>vacuum electrostatics > protein contact potential`

Pymol Movies (mac)

move the camera	`move x,10`
turn the camera	`turn x,90`
play the movie	`mplay`
stop the movie	`mstop`
writeout png files	`mpng` *prefix* `[, first [, last]]`
show a particular frame	`frame` *number*
move forward on frame	`forward`
move back one frame	`backwards`
go to the start of the movie	`rewind`
go to the middle of the movie	`middle`
go to the movie end	`ending`
determine the current frame	`get_frame`
clear the movie cache	`mclear`
execute a command in a frame	`mdo 1, turn x,5; turn y,5;`
dump current movie commands	`mdump`
reset the number of frames per second	`meter_reset`

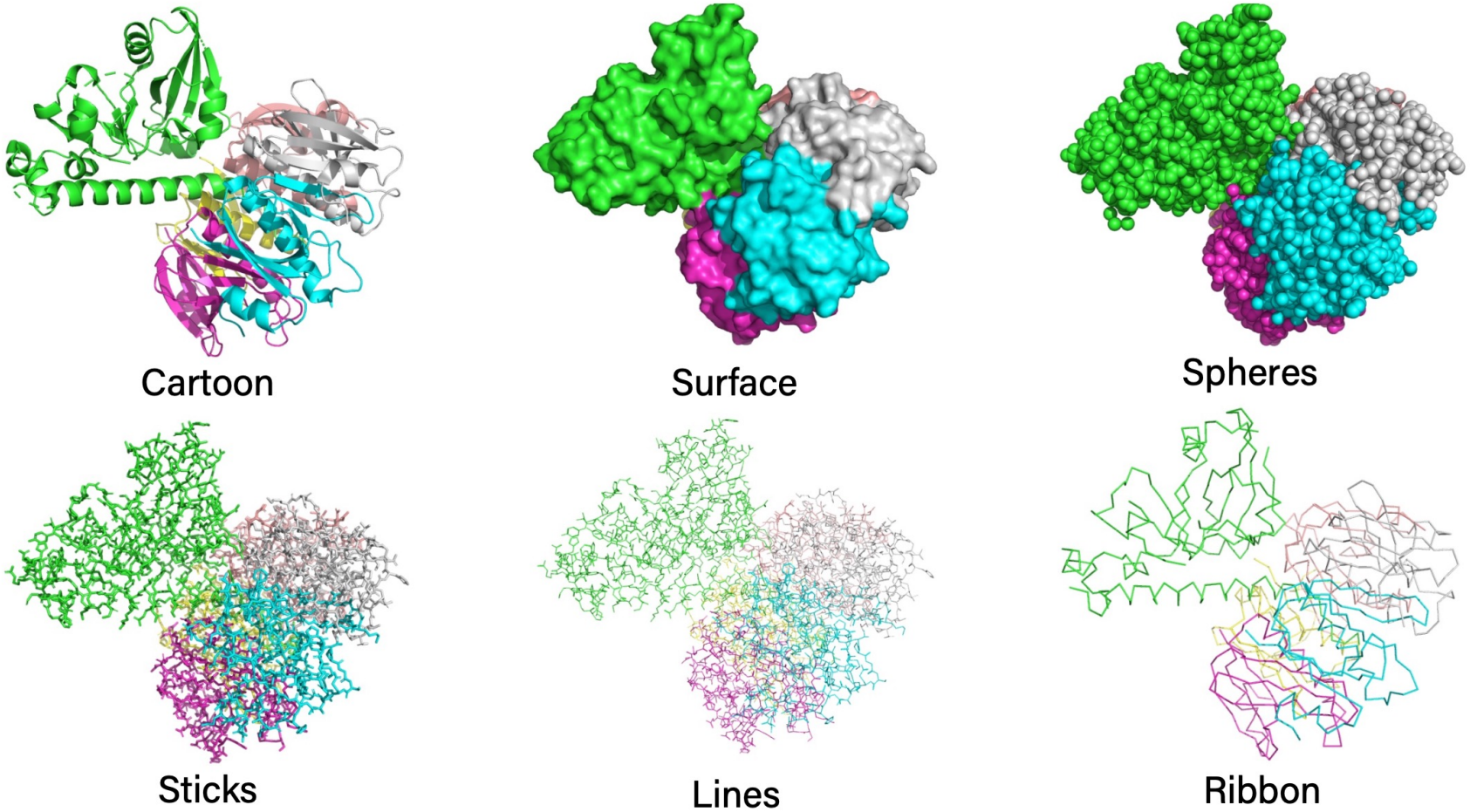
Loading Data

https://www.rcac.purdue.edu/files/training/AlphaFold_Protein_Structure_Prediction.pdf

PyMOL handles PDB, mmCIF, MRC, SITUS, etc

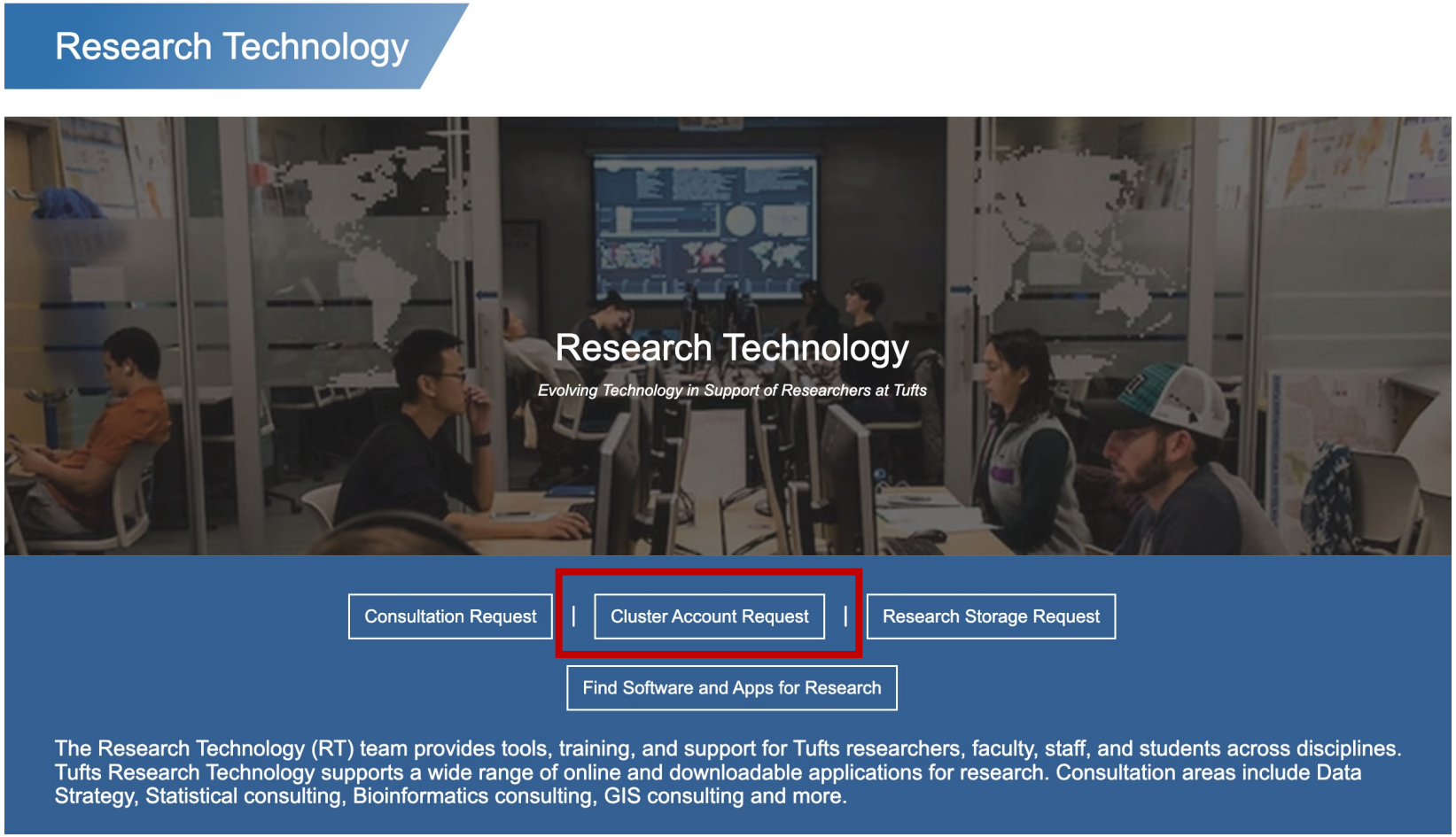
- Can open files on your computer
  - File → Open
  - `load <path to file>`
- Can download directly from PDB
  - File → Get PDB
  - `fetch <PDB code>`

Representations for Atomic Coordinate Data



https://www.rcac.purdue.edu/files/training/AlphaFold_Protein_Structure_Prediction.pdf

For those without access to an HPC account



https://it.tufts.edu/researchtechnology.tufts.edu

Hands-on tutorial
2024 Spring
Latest version

https://go.tufts.edu/chbe0165_af

Hands-on session 1: Run AlphaFold2 on Tufts HPC with Open OnDemand App

https://github.com/tuftsdatalab/tuftsWorkshops/blob/main/docs/2024_workshops/cas12aAlphaFold2_sp24/02_Run_AlphaFold2_OpenOndemandApp.md

Hands-on session 2: Visualize alphafold2 predicted structure with PYMOL

https://github.com/tuftsdatalab/tuftsWorkshops/blob/main/docs/2024_workshops/cas12aAlphaFold2_sp24/03_Vizualize_predicted_structure_with_PYMOL.md

Hands-on tutorial, 2023 Spring:
Content developed by Jason Larid, former bioinformatics scientist.

https://github.com/tuftsdatalab/tuftsWorkshops/tree/main/docs/2023_workshops/cas12aAlphaFold2

References

https://www.sciencedirect.com/science/article/pii/S2319417019305050

https://www.yourgenome.org/facts/what-is-crispr-cas9/

https://www.nature.com/articles/emm2016111

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-235

https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

https://www.nature.com/articles/s41586-021-03819-2

https://www.uniprot.org/help/uniref

https://www.rcsb.org/

https://alphafold.ebi.ac.uk/faq

https://alphafold.com/entry/Q9FX77

https://www.rcsb.org/3d-view/5XUS/1

https://pymol.org/2/

https://github.com/google-deepmind/alphafold/tree/main

https://hpc.nih.gov/apps/alphafold2.html