

***Dive into
computational
physical chemistry***

***Lecture 6:
Visualizing and
predicting protein
structures***

Glen Hocky
October 25, 2022



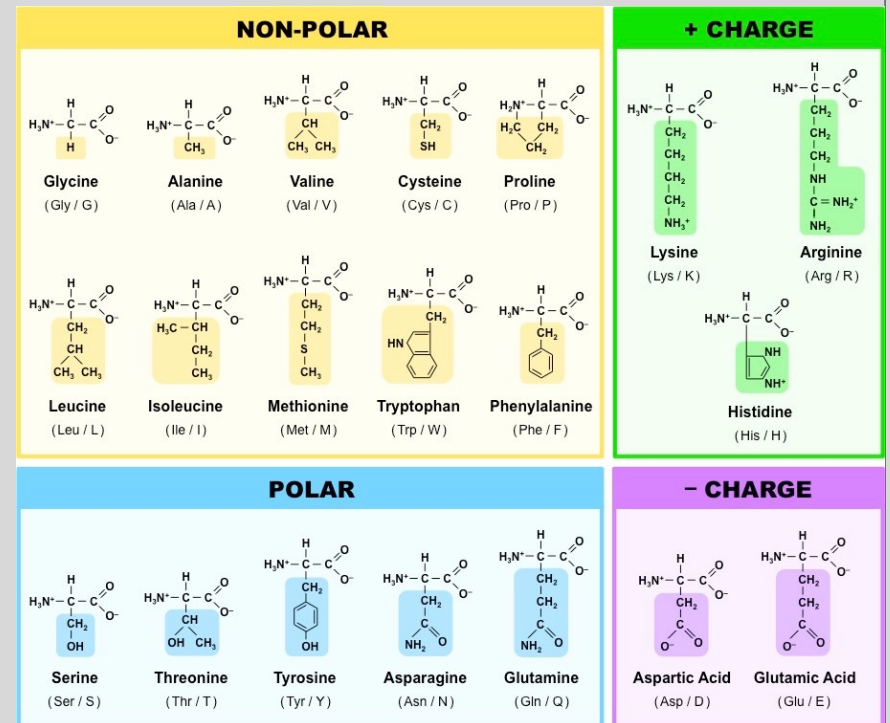
Protein structures

- Proteins do many of the functions in our bodies
- They are formed from chains of amino-acids that “fold” into three dimensional arrangements
- The three-dimensional structure determines the function, in general
- Structures are “Solved” using X-ray crystallography, electron microscopy or NMR, and deposited in the Protein Data Bank (<https://www.rcsb.org/>)



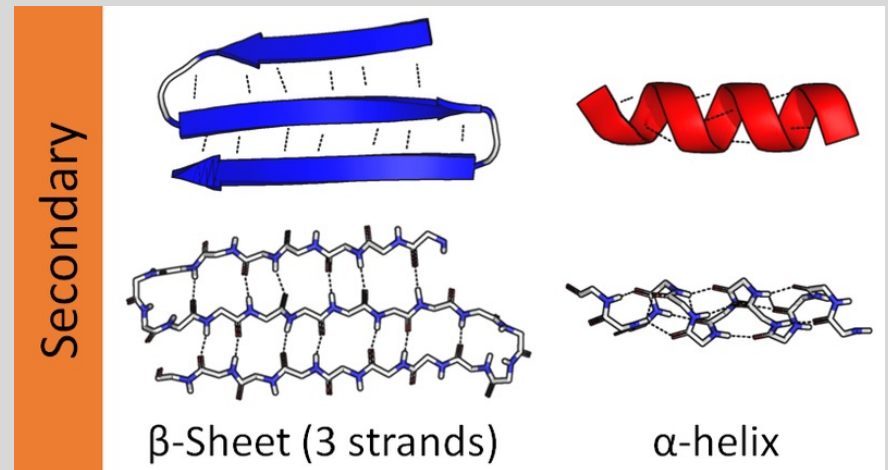
Protein primary sequence

- Proteins are typically composed of sequences of 20 common amino acids
- In PDB, you will see these given as FASTA files, e.g.
 - >1UBQ_1|Chain A|UBIQUITIN|Homo sapiens (9606)
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQ
QRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLLRGG



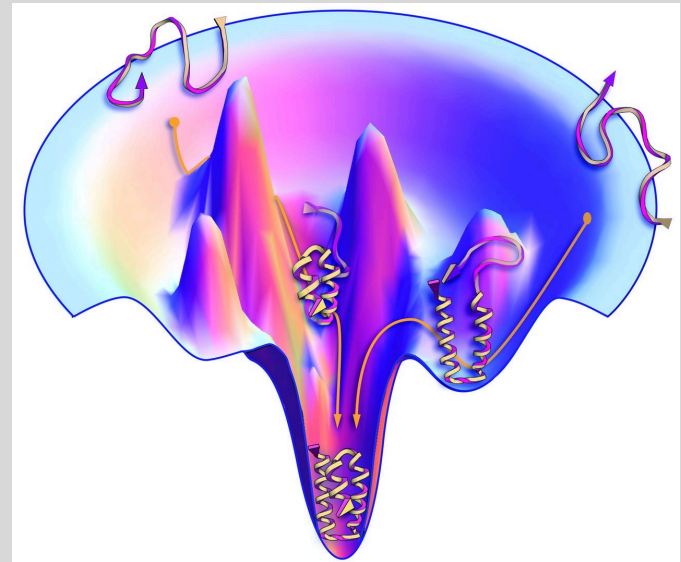
Protein secondary structure

- Proteins have locally defined structures called “secondary structure” that are repeated motifs
- Examples include sheets and helices. Programs like DSSP can assign from a PDB file
- Special cases like metal or ligand bound structures, disulfide bonds add complexity



Tertiary structure: Protein folding problem

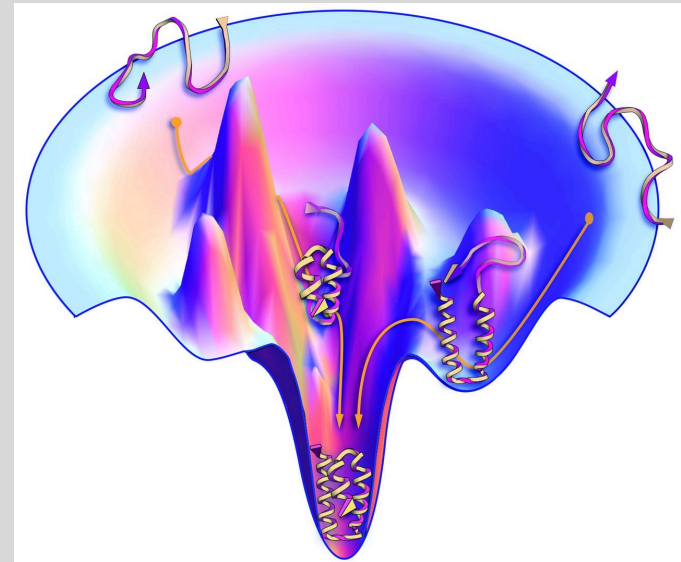
- Proteins have astronomically large number of configurations they can be in, how is the lowest free energy one found?
- Local motifs could produce single/multiple pathways or funnel picture, which makes it much less complicated
- Some protein folding facilitated by chaperones



Dill and coworkers

Tertiary structure: Protein structure prediction problem

- How can we take sequence and predict final structure
 - MD – could work, but real folding can be way too slow
 - Physical modeling/coarse-grained MD. Can try to make a lower resolution model that folds faster: this sort of works!
 - Information based modeling. Similar idea, but based on data from PDB: works better
 - Homology modeling: works great when data available
 - New frontier: deep learning



Dill and coworkers

How do we know if it's working?



Menu

- [Home](#)
- [PC Login](#)
- [PC Registration](#)
- ▼ **CASP Experiments**
 - ▼ **CASP15 (2022)**
 - [Home](#)
 - [Registration](#)
 - [My CASP15 profile](#)
 - [CASP Commons \(COVID-19, 2020\)](#)
 - [CASP14 \(2020\)](#)
 - [CASP13 \(2018\)](#)
 - [CASP12 \(2016\)](#)
 - [CASP11 \(2014\)](#)
 - [CASP10 \(2012\)](#)
 - [CASP9 \(2010\)](#)
 - [CASP8 \(2008\)](#)
 - [CASP7 \(2006\)](#)
 - [CASP6 \(2004\)](#)
 - [CASP5 \(2002\)](#)

Protein Structure Prediction Center

Register for CASP15 Meeting



CASP15
Antalya, Turkiye, Dec. 10-13, 2022

CASP15 is planned as an in-person meeting.

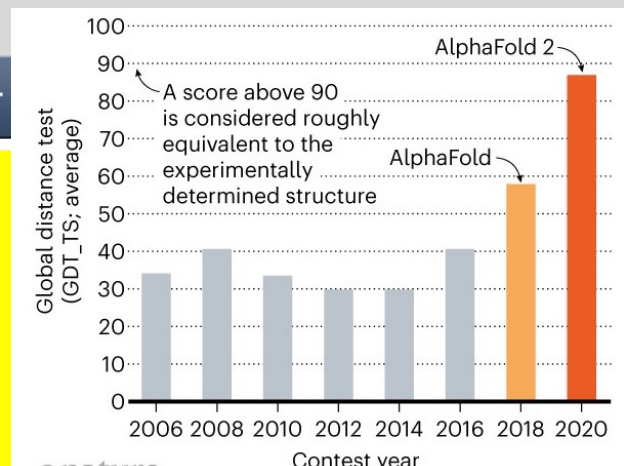
The deadline for the registration is September 16, 2022.

It is possible to register after that date with a much higher fee, and space is not guaranteed.

Success Stories From Recent CASPs

template-based modeling	ab initio modeling	contact prediction	help structural biologists	refinement	data-assisted modeling	
--------------------------------	--------------------	--------------------	----------------------------	------------	------------------------	--

Models based on templates identified by sequence similarity remain the most accurate. Over the course of the CASP experiments there have been enormous improvements in this area. However, the overall accuracy improvements that we have seen in the first 10 years of CASP

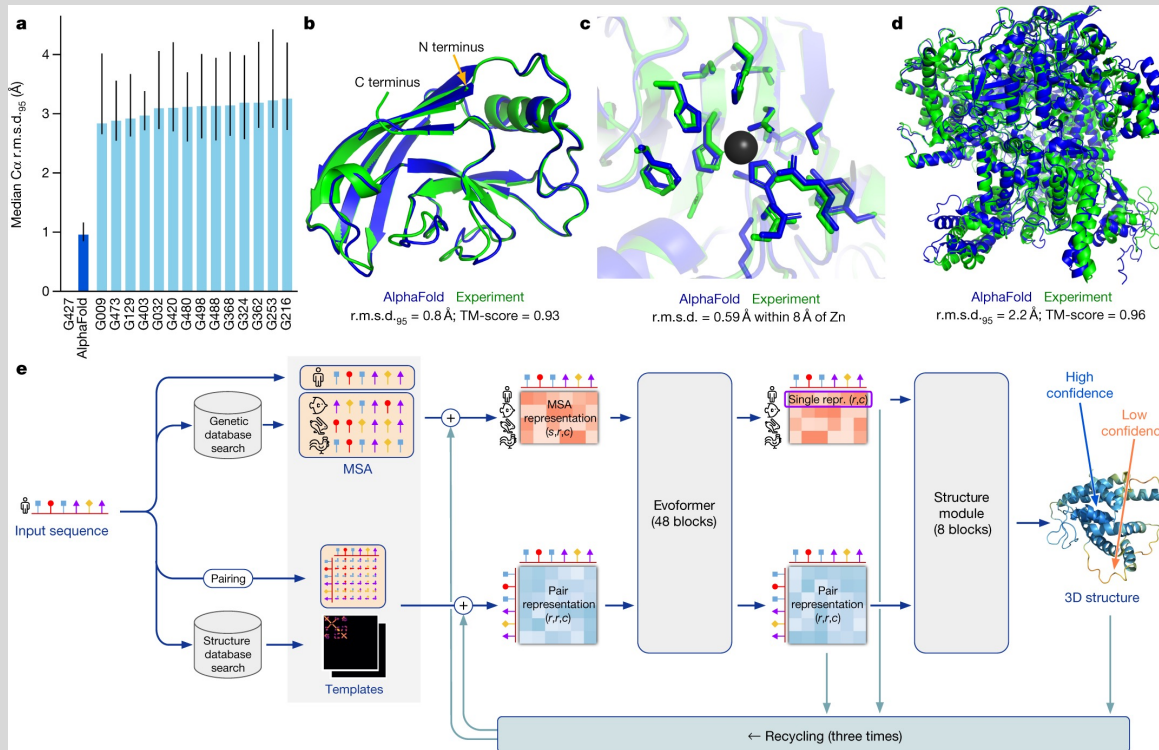


register for the Antalya 2022 - fill out the Survey (deadline Wednesday, September 7, 2022)

Please let us know what you think about the SIGs: <https://forms.gle/xz5f9tqtk5bWc9j66>

Register for the Antalya CASP15 meeting (the deadline is September 16, 2022) using the following link at the ...

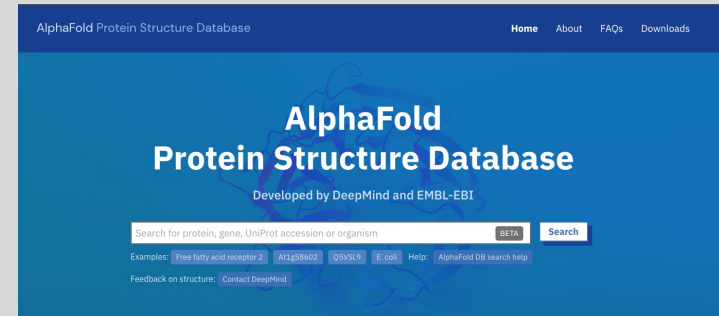
AlphaFold2 general architecture



- Generate and use “Multiple Sequence Alignment” to help with pair contacts (see e.g. Ranganathan, Nature 2005)
- Use templates (doesn’t add much)
- Use MD refinement (doesn’t add much)
- Simultaneously predict confidence (pLDDT score) of each residue

Jumper et al. Nature, 2021; See also Baek et al, Science 2021, RoseTTAFold

Alphafold2 database



3D viewer

Model Confidence:

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

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-P0CG48-F1 Chain 1: Polyubiquitin...










```
MQIFVKTLTGKITLEVEPSDTIENVKAKIQDKEGIPPDQRLIFAGKQLEDGRITLSDYNIQKESTLHLVLRGGQIIFVKTLTGKITLEVEPSDTIENVKAKIQDKEGIPPDQRLIFAGK
QLEDGRITLSDYNIQKESTLHLVLRGGQIIFVKTLTGKITLEVEPSDTIENVKAKIQDKEGIPPDQRLIFAGKQLEDGRITLSDYNIQKESTLHLVLRGGQIIFVKTLTGKITLEVEPS
DTIENVKAKIQDKEGIPPDQRLIFAGKQLEDGRITLSDYNIQKESTLHLVLRGGQIIFVKTLTGKITLEVEPSDTIENVKAKIQDKEGIPPDQRLIFAGKQLEDGRITLSDYNIQKESTLH
```

Tunyasuvunakool et al. Nature, 2021

ColabFold

OPEN

ColabFold: making protein folding accessible to all

Milot Mirdita ^{1,10} , Konstantin Schütze ², Yoshitaka Moriwaki ^{3,4}, Lim Heo ⁵,
Sergey Ovchinnikov ^{6,7,10}  and Martin Steinegger ^{2,8,9,10} 

ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold. ColabFold's 40–60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit. Coupled with Google Colaboratory, ColabFold becomes a free and accessible platform for protein folding. ColabFold is open-source software available at <https://github.com/sokrypton/ColabFold> and its novel environmental databases are available at <https://colab-fold.mmseqs.com>.

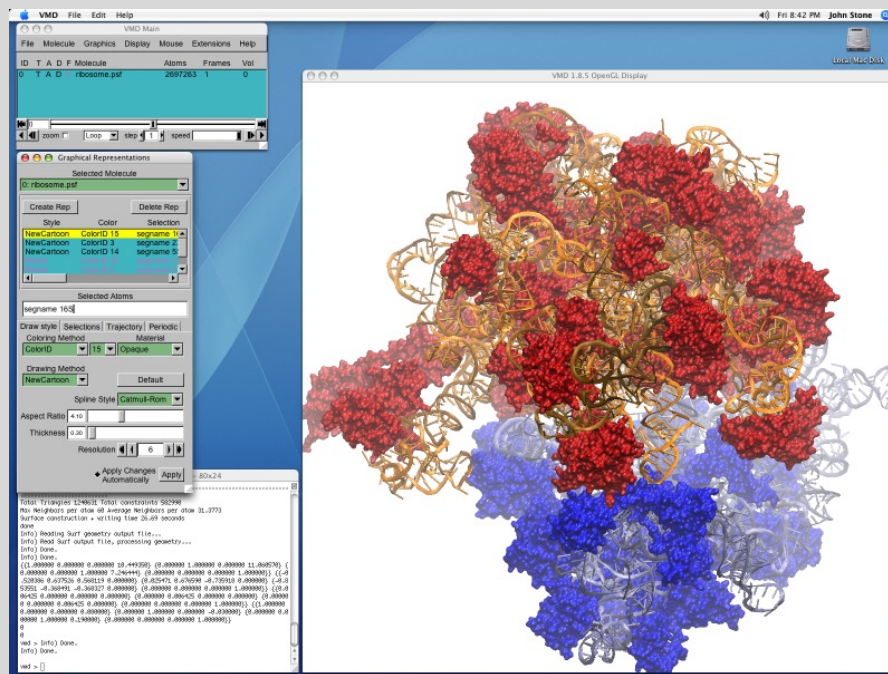
protein sizes of ~1,000 residues. For these, however, the MSA generation dominates the overall run time.

To enable researchers without these resources to use AlphaFold2, independent solutions based on Google Colaboratory were developed. Colaboratory is a proprietary version of Jupyter Notebook hosted by Google. It is accessible for free to logged-in users and includes access to powerful GPUs. Concurrently, Tunyasuvunakool et al.⁹ developed an AlphaFold2 Jupyter Notebook for Google Colaboratory (referred to as AlphaFold-Colab), for which the input MSA is built by searching with HMMer against the UniProt Reference Clusters (UniRef90) and an eightfold-reduced environ-

<https://github.com/sokrypton/ColabFold>

Protein structure visualization

Many options with different features: **VMD**, PyMol, Chimera, ChimeraX, Mol*, ...



Today

1. Two versions of alphafold on an assigned protein
2. Learn to visualize and analyze structures using VMD

<https://github.com/hockyg/comp-lab-class/blob/main/Week7/Assignment.md>