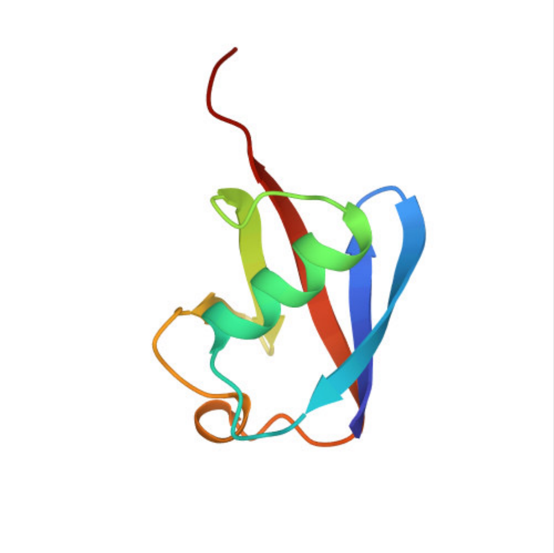
Dive into computational physical chemistry

Lecture 6: Visualizing and predicting protein structures

Glen Hocky
October 18, 2023

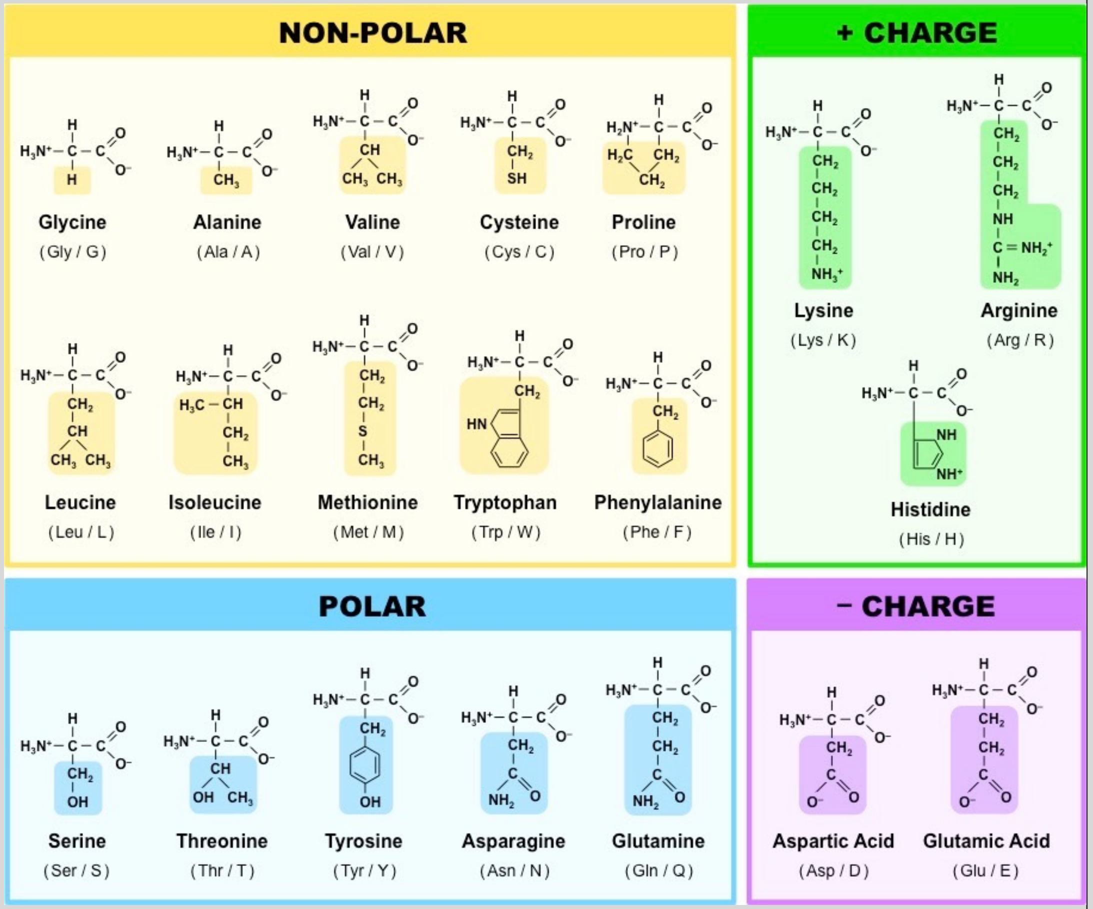
Protein structures

- 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/)



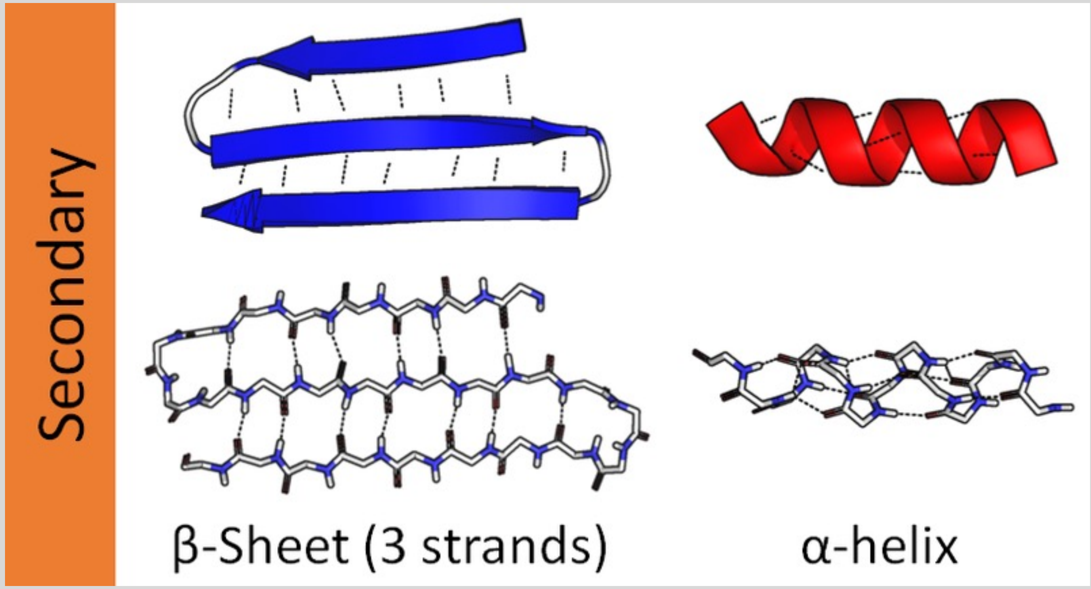
Protein primary sequence

- 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
      QRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG



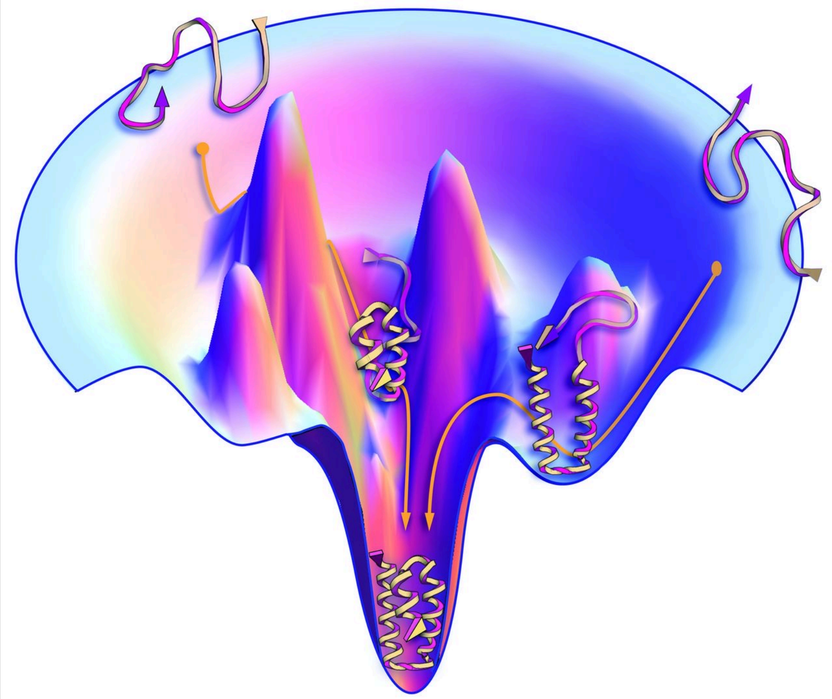
Protein secondary structure

- 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

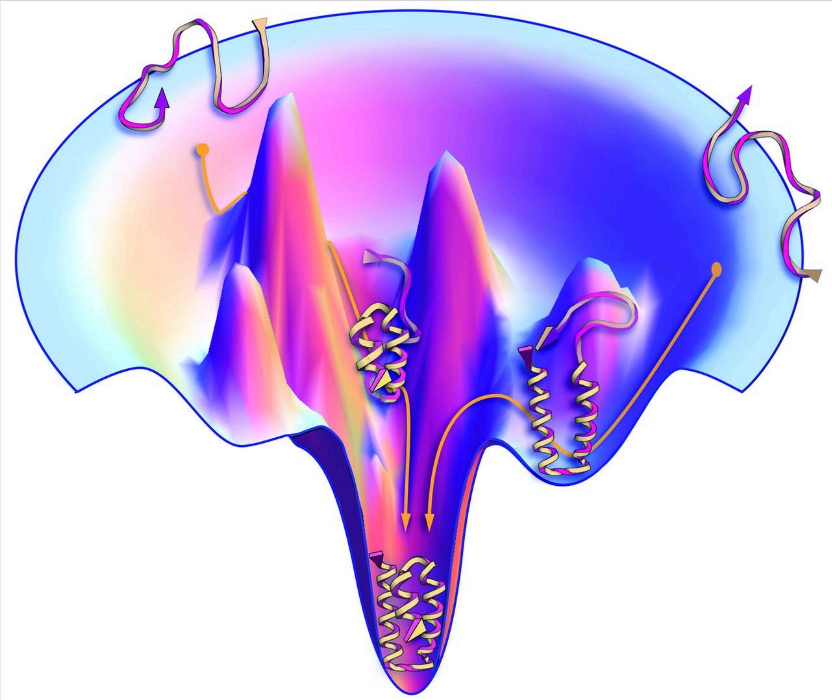
- Proteins have astronomically large number of configurations they can be in, how is the lowest free energy one found?
- Local motifs could produce single/multiple pathways or funnel picture, which makes it much less complicated
- Some protein folding facilitated by chaperones



Dill and coworkers

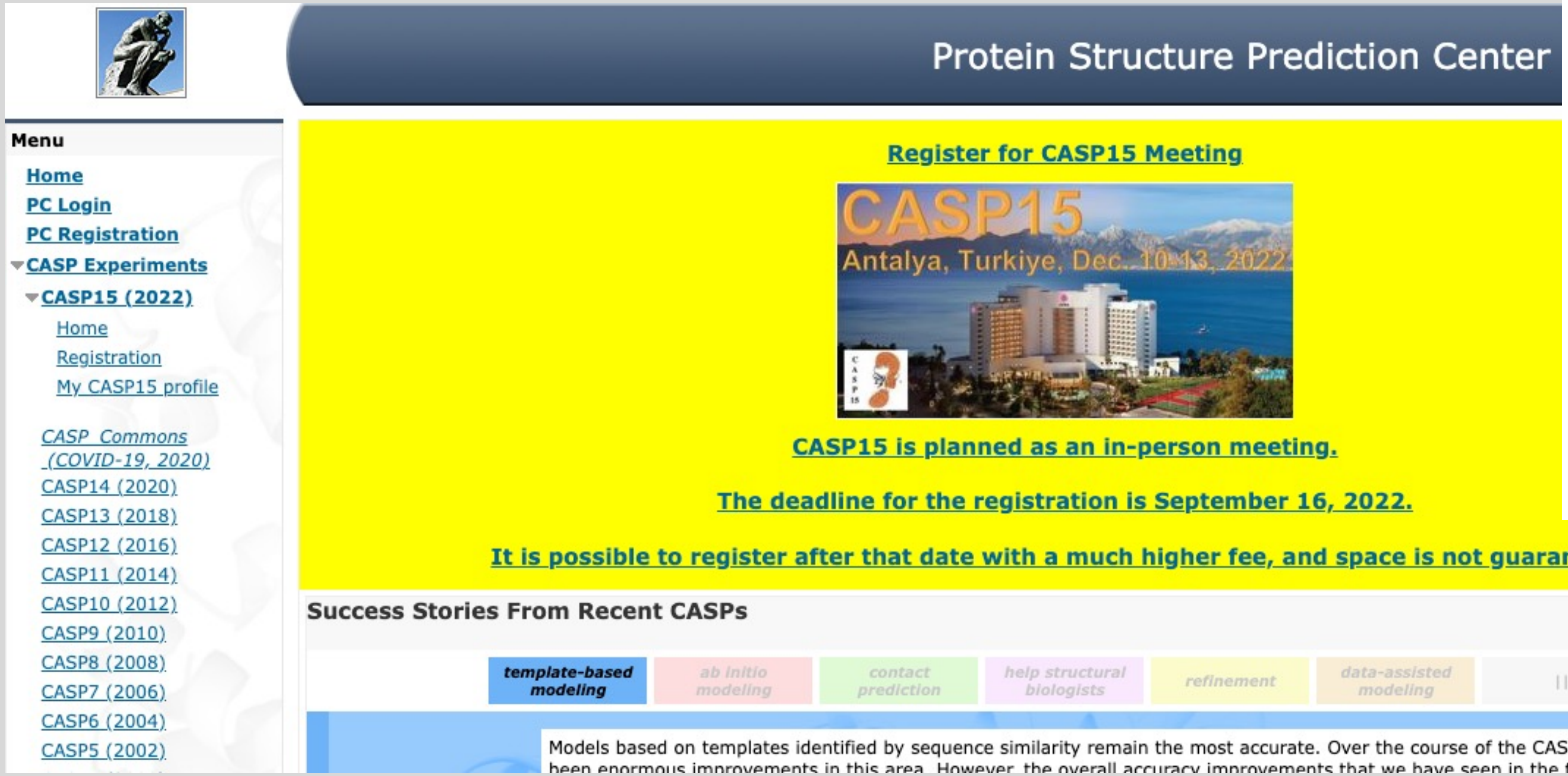
Tertiary structure: Protein structure prediction problem

- 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?



Register for CASP15 Meeting

CASP15 is planned as an in-person meeting.

The deadline for the registration is September 16, 2022.

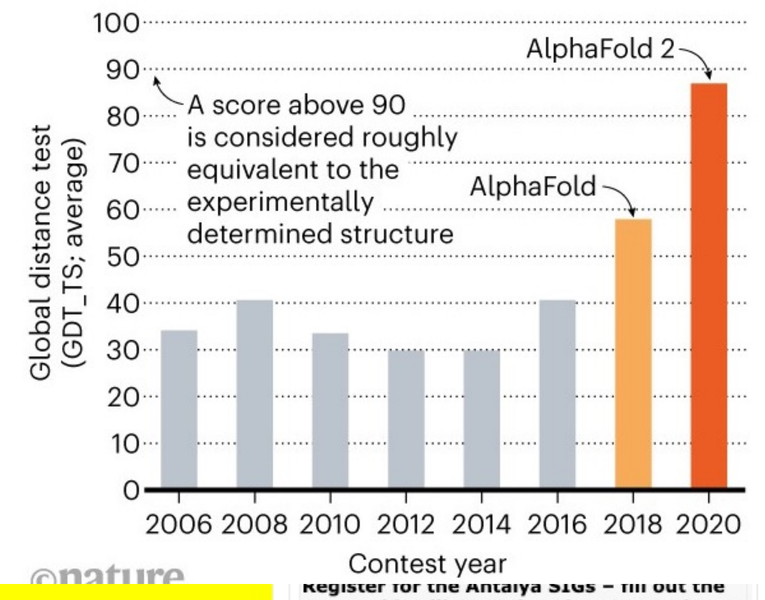
It is possible to register after that date with a much higher fee, and space is not guaranteed.

Success Stories From Recent CASPs

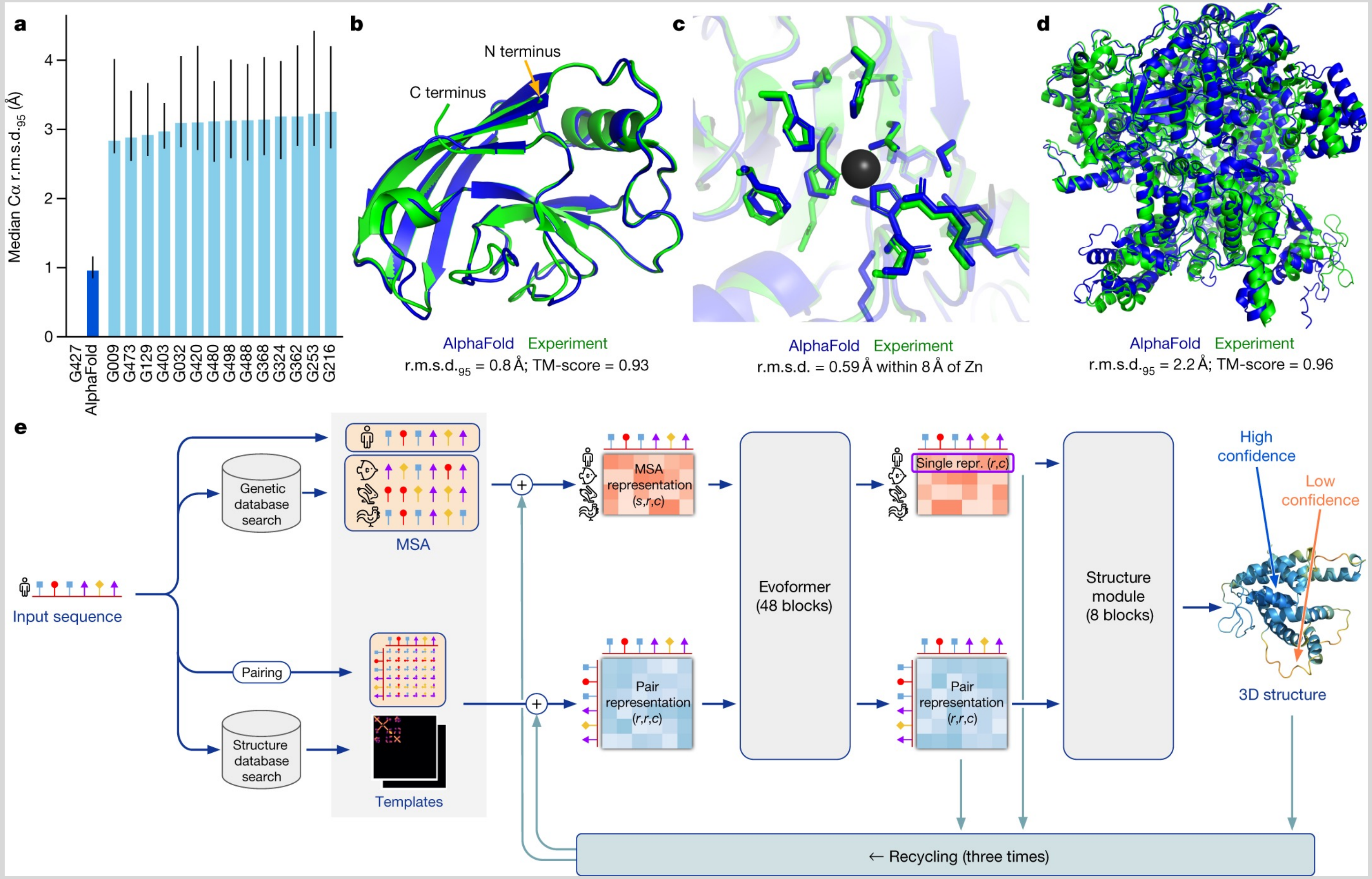
Models based on templates identified by sequence similarity remain the most accurate. Over the course of the CASP experiments there have been enormous improvements in this area. However, the overall accuracy improvements that we have seen in the first 10 years of CASP

Register for the Antalya SIGs – fill out the Survey (deadline Wednesday, September 7, 2022)

Please let us know what you think about the SIGs: https://forms.gle/xz5f9tgtk5bWc9j66 Register for the Antalya CASP15 meeting (the deadline is September 16, 2022) using the following link at the ...



Alphafold2 general architecture

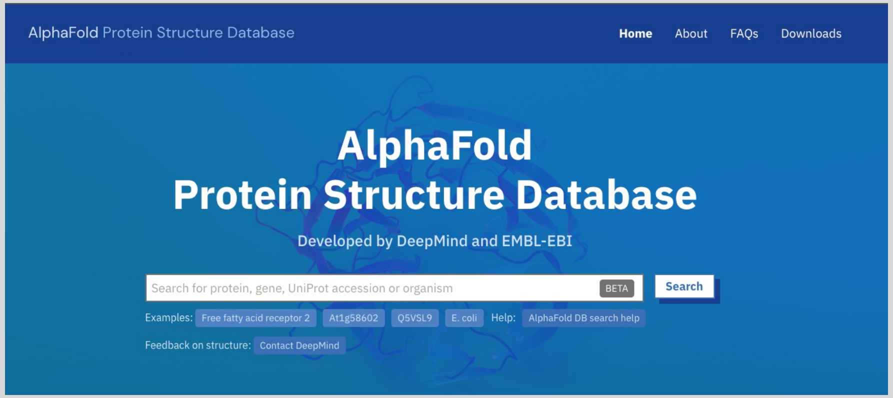


a
Median Cα r.m.s.d.$_{95}$ (Å)
4
3
2
1
0
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G368
G488
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

- 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



AlphaFold Protein Structure Database

Home About FAQs Downloads

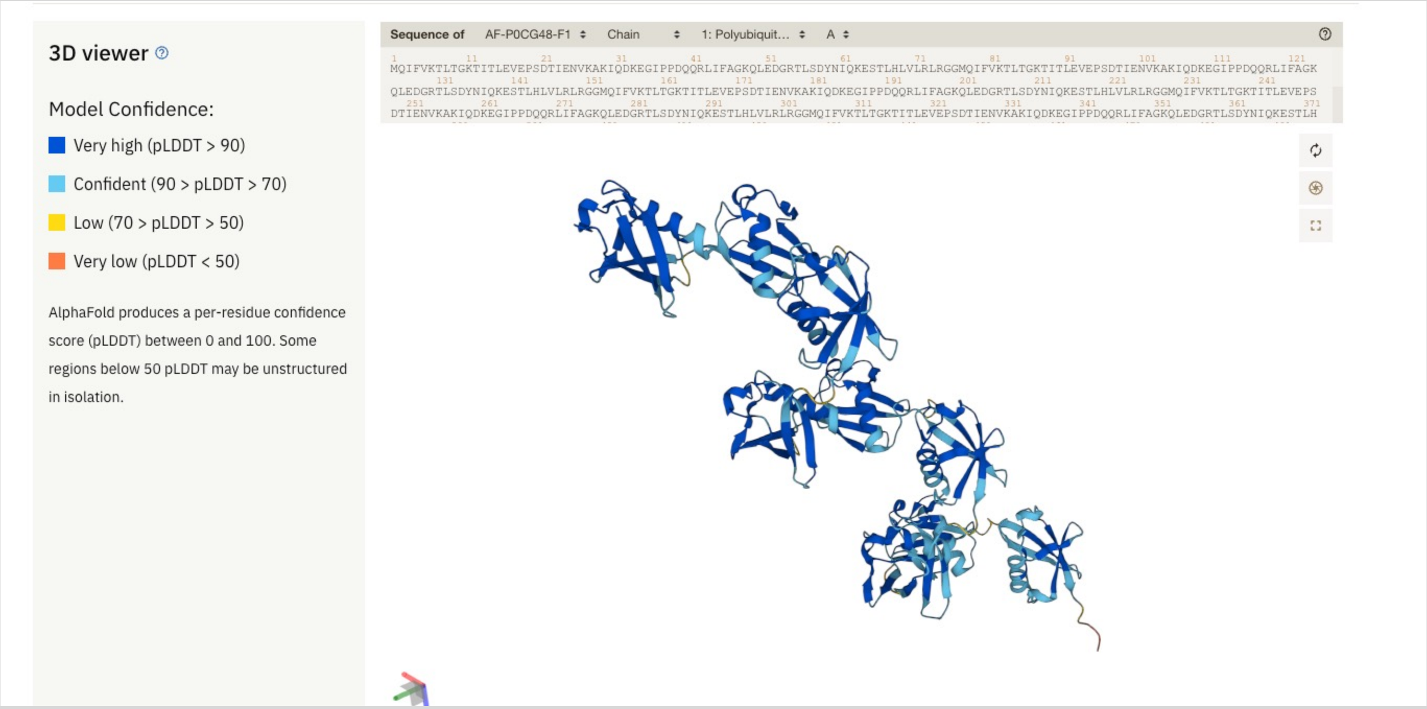
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA Search

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Feedback on structure: Contact DeepMind



3D viewer

Model Confidence:

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

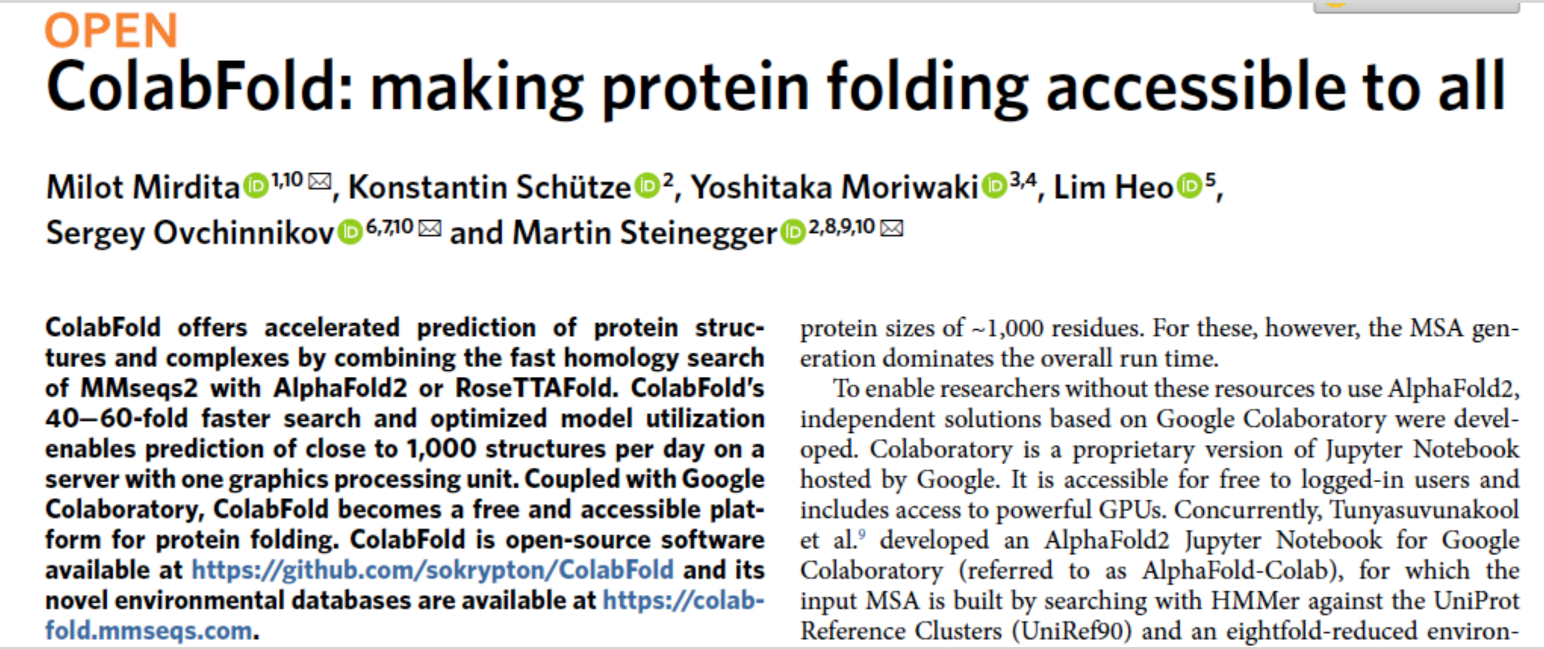
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-P0CG48-F1 Chain 1: Polyubiquit... A

```
1         11        21        31        41        51        61        71        81        91        101       111       121
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGK
      131       141       151       161       171       181       191       201       211       221       231       241
QLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPS
  251       261       271       281       291       301       311       321       331       341       351       361       371
DTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLH
```

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://colabfold.mmseqs.com.

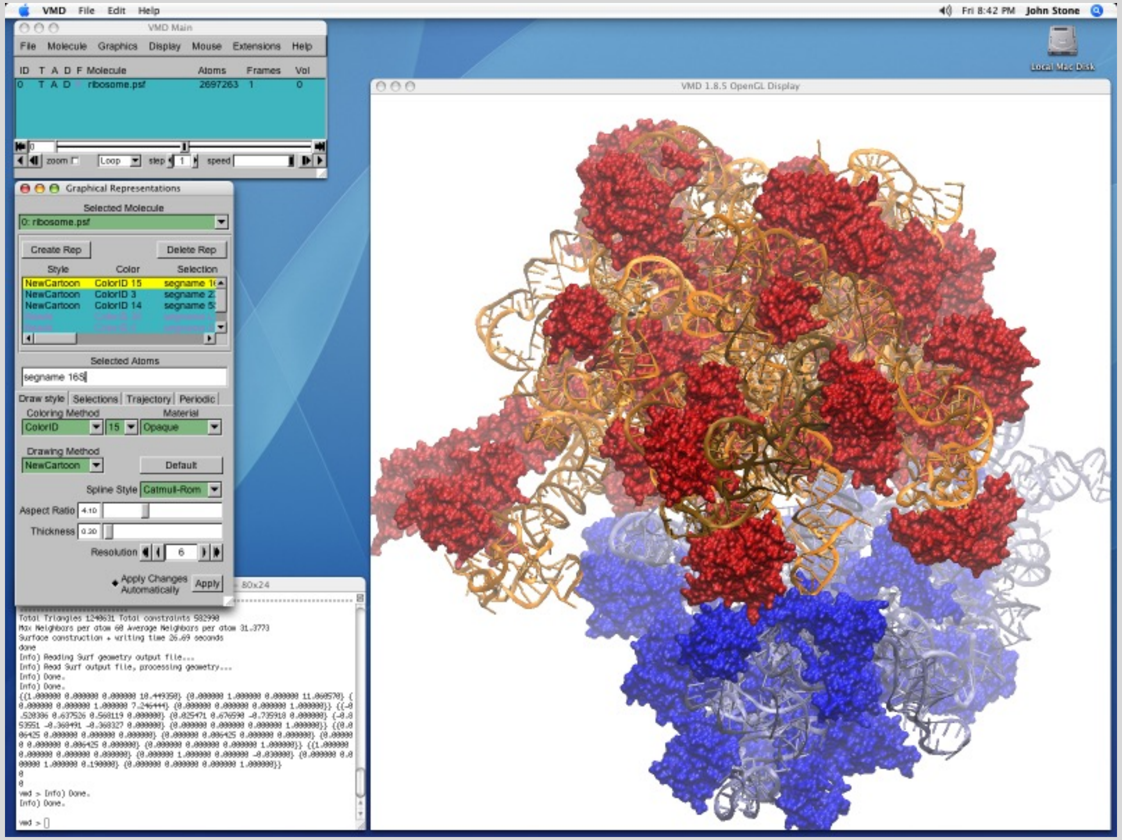
protein sizes of ~1,000 residues. For these, however, the MSA generation dominates the overall run time.

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*, ...



ID	T	A	D	F	Molecule	Atoms	Frames	Vol
0	T	A	D		ribosome.psf	2697263	1	0

Selected Molecule: 0: ribosome.psf

Style	Color	Selection
NewCartoon	ColorID 15	segname 1
NewCartoon	ColorID 3	segname 2
NewCartoon	ColorID 14	segname 5

Selected Atoms: segname 16S

Coloring Method: ColorID 15

Material: Opaque

Drawing Method: NewCartoon

Spline Style: Catmull-Rom

Aspect Ratio: 4.10

Thickness: 0.20

Resolution: 6

◆ Apply Changes Automatically

```
Total Triangles 1240631 Total constraints 582990
Max Neighbors per atom 60 Average Neighbors per atom 31.3773
Surface construction + writing time 26.69 seconds
done
Info) Reading Surf geometry output file...
Info) Read Surf output file, processing geometry...
Info) Done.
Info) Done.
{{1.000000 0.000000 0.000000 10.449350} {0.000000 1.000000 0.000000 11.060570} {
0.000000 0.000000 1.000000 7.246444} {0.000000 0.000000 0.000000 1.000000}} {{-0
.520306 0.637526 0.568119 0.000000} {0.025471 0.676590 -0.735918 0.000000} {-0.8
53551 -0.360491 -0.368327 0.000000} {0.000000 0.000000 0.000000 1.000000}} {{0.0
06425 0.000000 0.000000 0.000000} {0.000000 0.006425 0.000000 0.000000} {0.00000
0 0.000000 0.006425 0.000000} {0.000000 0.000000 0.000000 1.000000}} {{1.000000
0.000000 0.000000 0.000000} {0.000000 1.000000 0.000000 -0.030000} {0.000000 0.0
00000 1.000000 0.190000} {0.000000 0.000000 0.000000 1.000000}}
0
0
vmd > Info) Done.
Info) Done.

vmd > 
```

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-2023/blob/main/Week7/Assignment.md