Introduction to AlphaFold for 3D Protein Structure Prediction on Grace

Introduction to AF2 - 05/25/2022
Devon J. Boland
Devarenne Laboratory

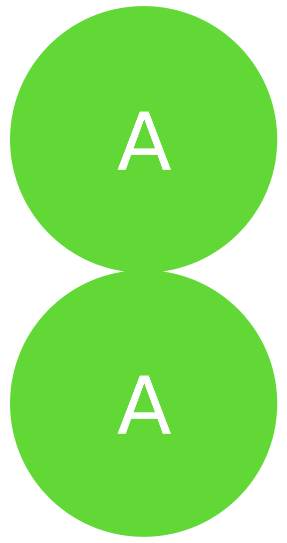
TEXAS A&M AGRILIFE RESEARCH

Basics of 3D Protein Structure

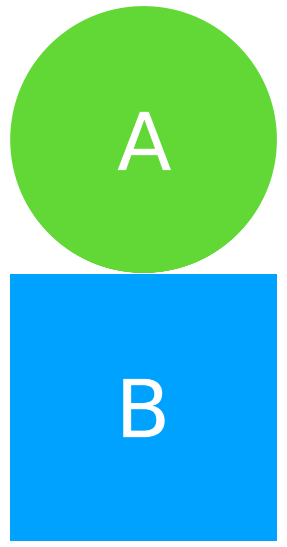
Four Tiers of Protein Structure

- **1° - Sequence of AA’s (polypeptides)**
- **2° - Interactions of the carbon backbone of 1°**
- **3° - Folding of 2° onto itself**
- **4° - Multiple 3° units (Monomers) assembling together**

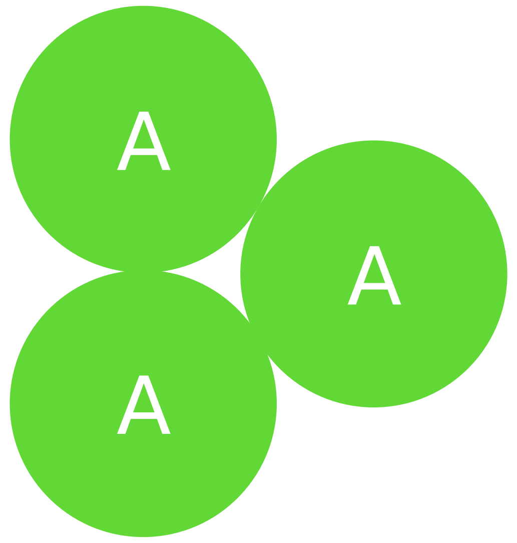
Different Types of Quaternary Structure



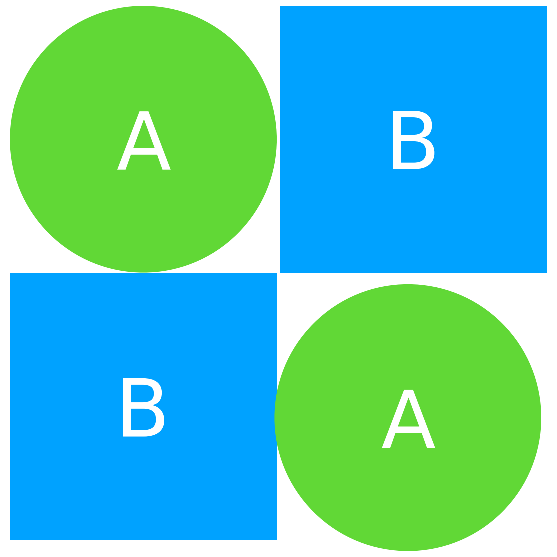
Dimer (homo)



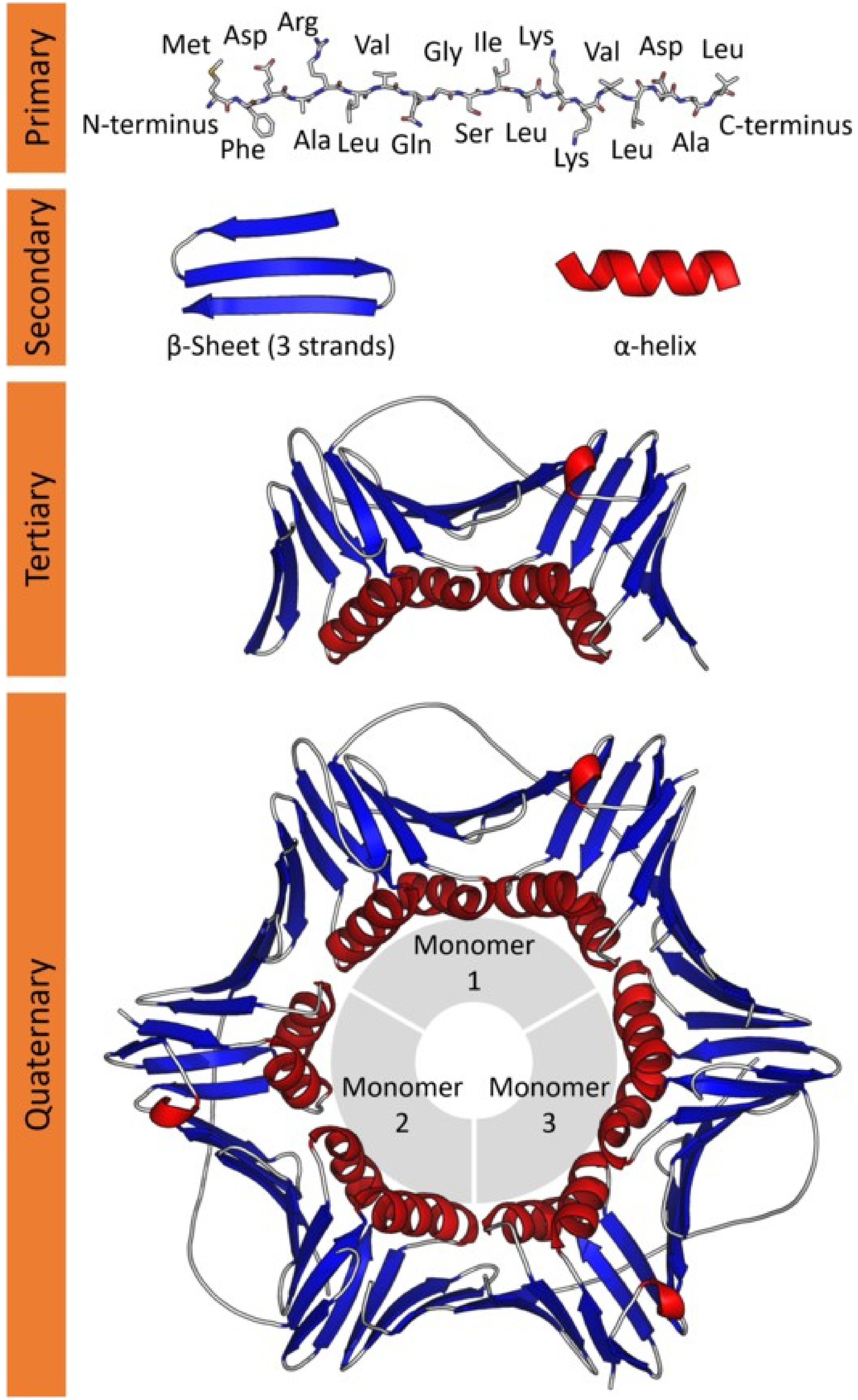
Dimer (hetero)



Trimer (homo)



Tetramer (hetero)

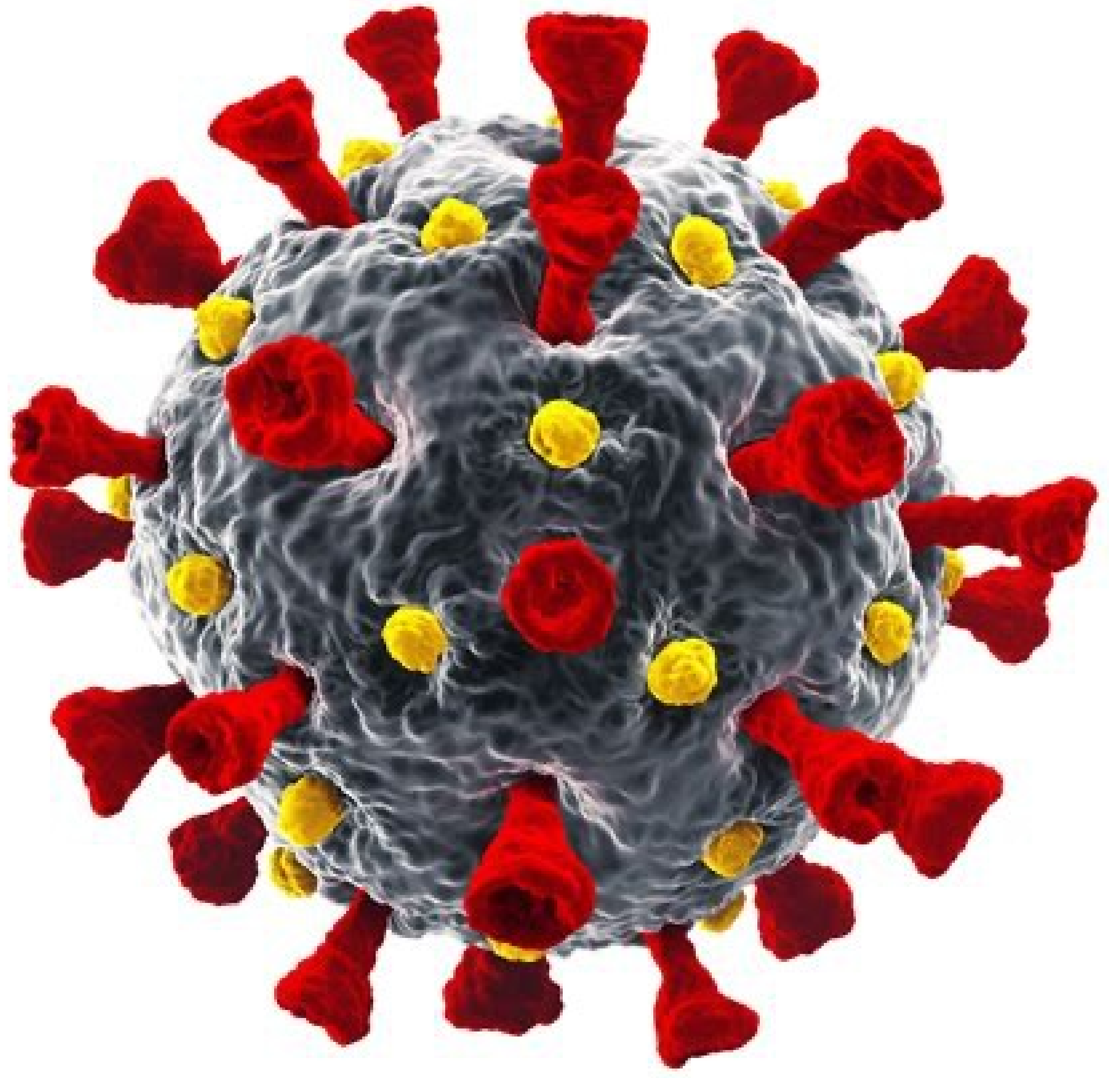


https://en.wikipedia.org/wiki/Protein_structure

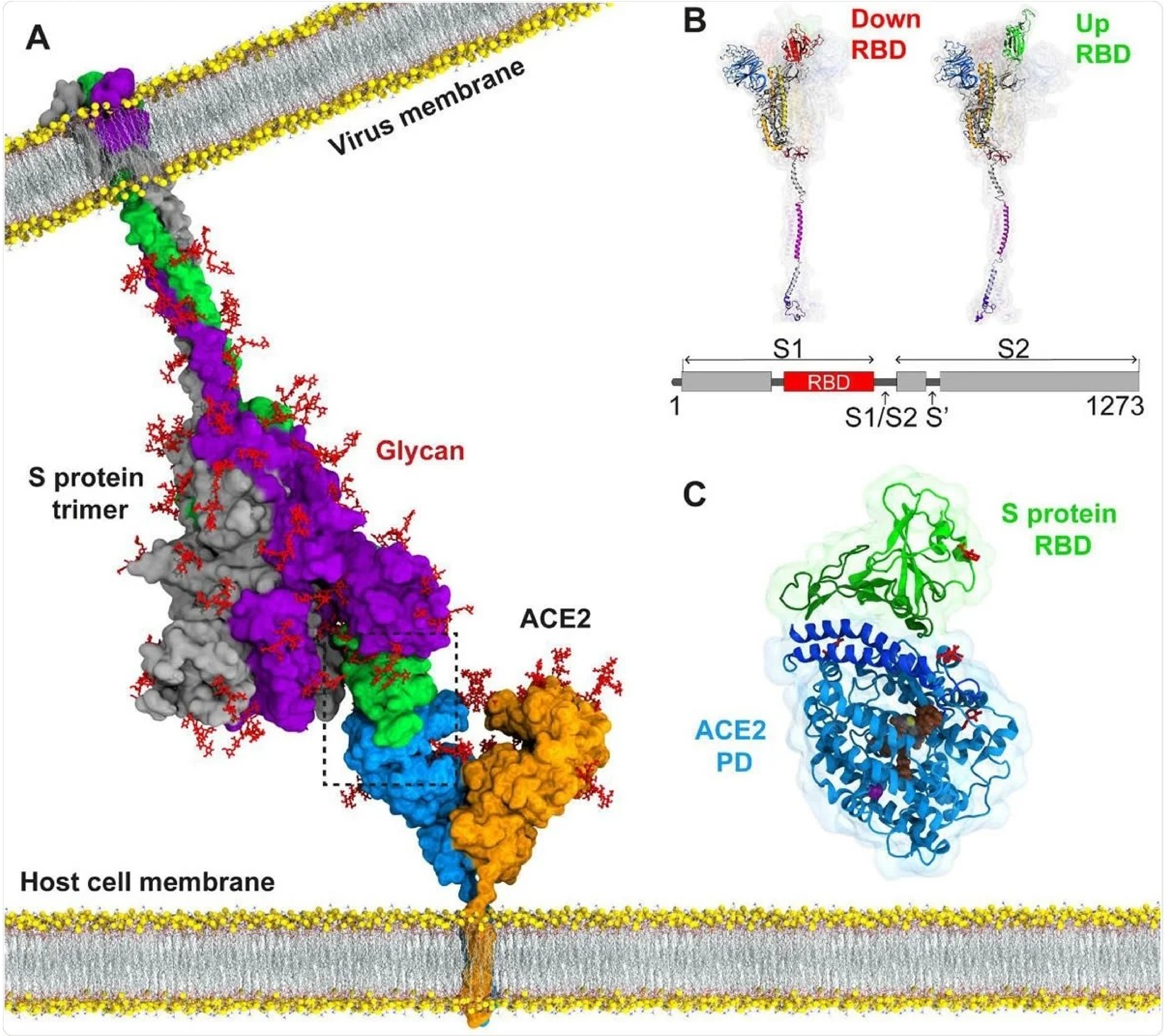
What Does Protein Structure Tell Us?

Structure = Function → Function = Structure

- **visualizing binding interface**
- **location of allosteric inhibition**
- **conformational changes**
- **molecular scaffold for molecular docking**
- **mutational design**
- **MANY OTHER FACETS…**

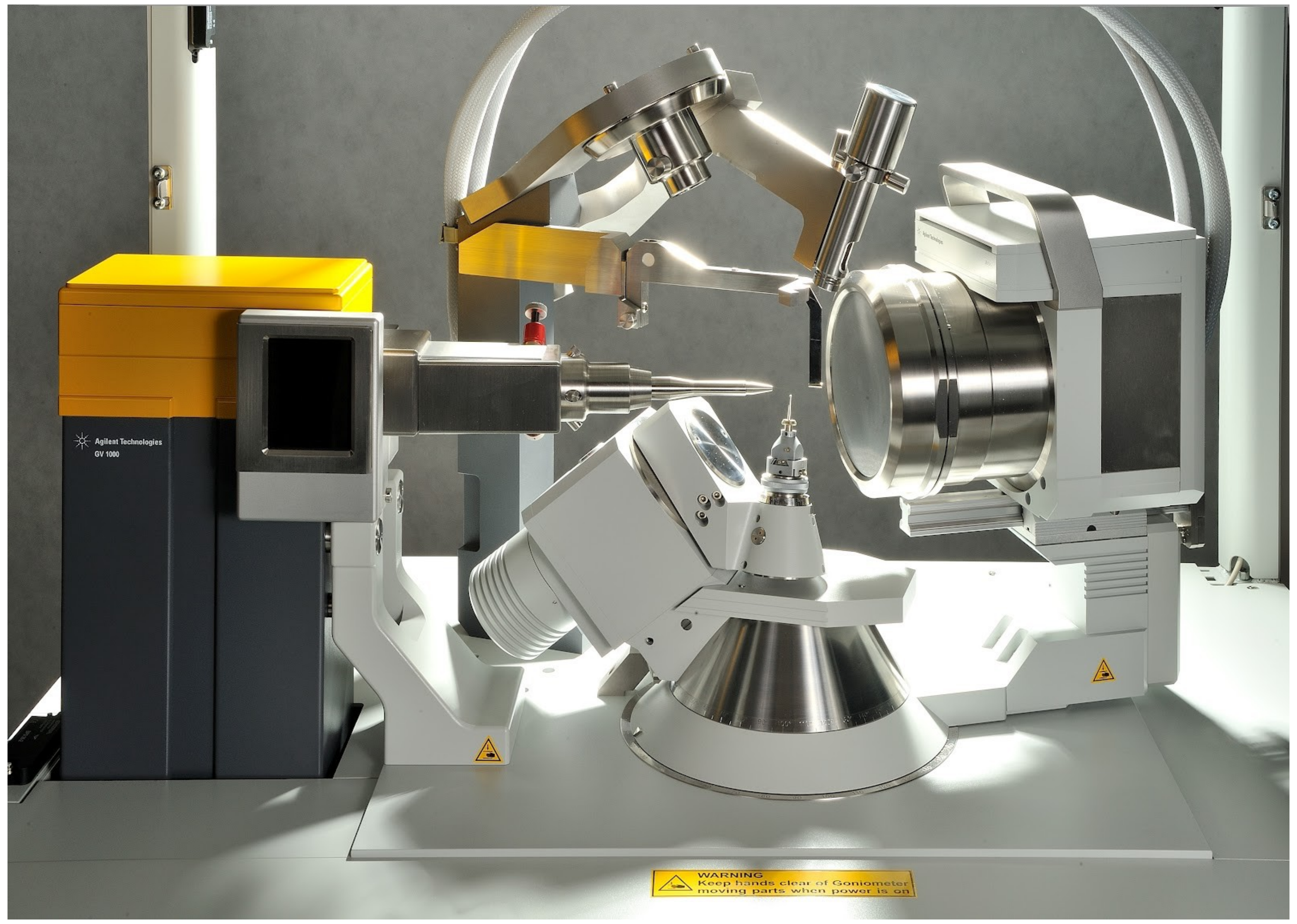


SARS-COV-2



Methods for Elucidating Protein Structure

X-Ray Crystallography



pros:

- widely used
- high-throughput
- ~2.5-1Å resolution

cons:

- crystallization process
- high variation

NMR Spectroscopy

pros:

- native state
- non-destructive
- real-time

cons:

- only works on “small” proteins
  - <100kDa
- high [protein]
- $$$ maintenance

Cryo-EM

pros:

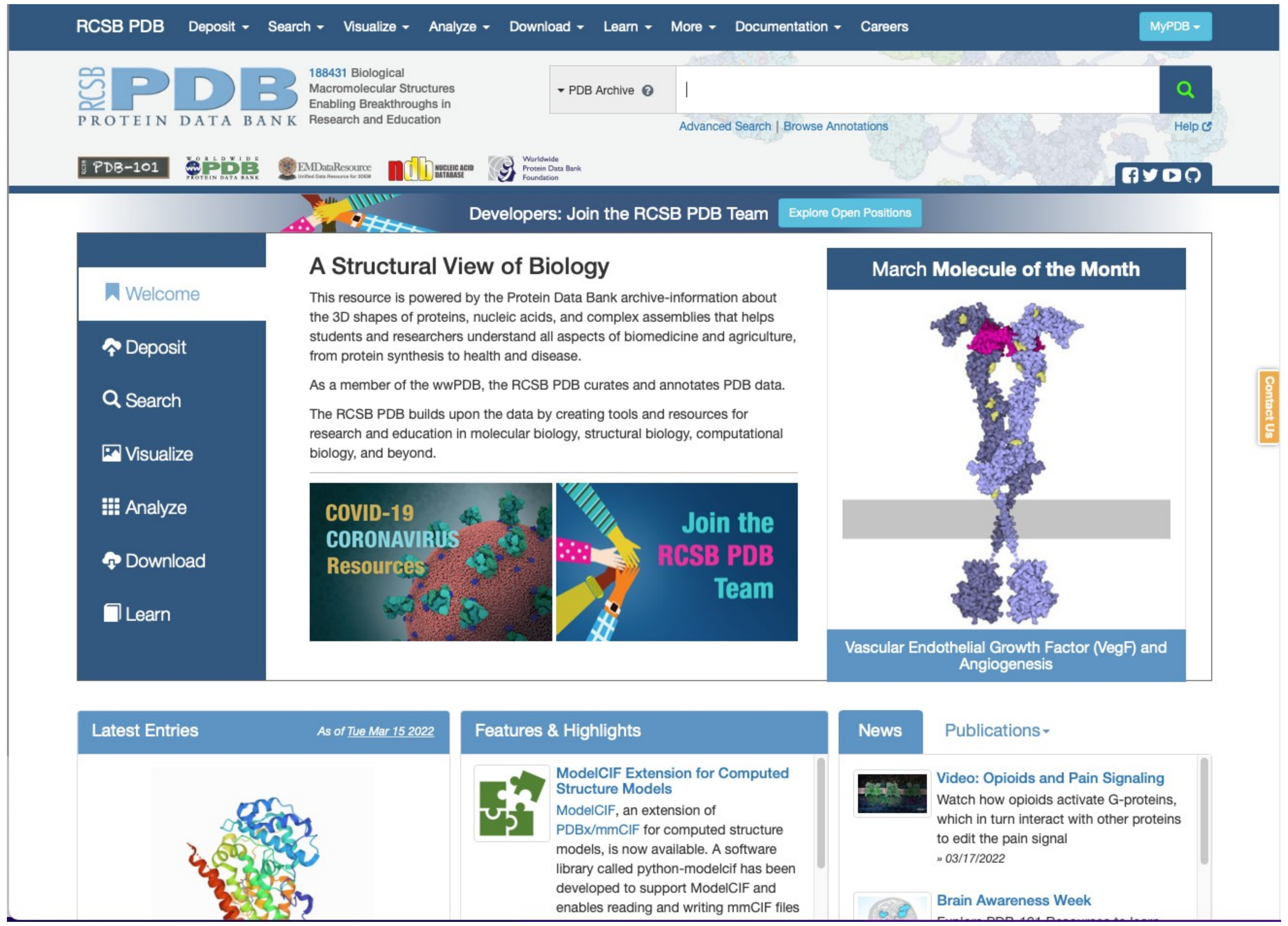
- large proteins/complexes
- near-native state

cons:

- only works on “large” proteins
  - >200kDa
- freezing samples
- computationally intensive
- $$$ maintenance

Protein Data Bank (PDB)

After we elucidate a structure where does it go?



RCSB PDB PROTEIN DATA BANK

188431 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Developers: Join the RCSB PDB Team

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

March Molecule of the Month

Vascular Endothelial Growth Factor (VegF) and Angiogenesis

Latest Entries *As of Tue Mar 15 2022*

Features & Highlights

ModelCIF Extension for Computed Structure Models

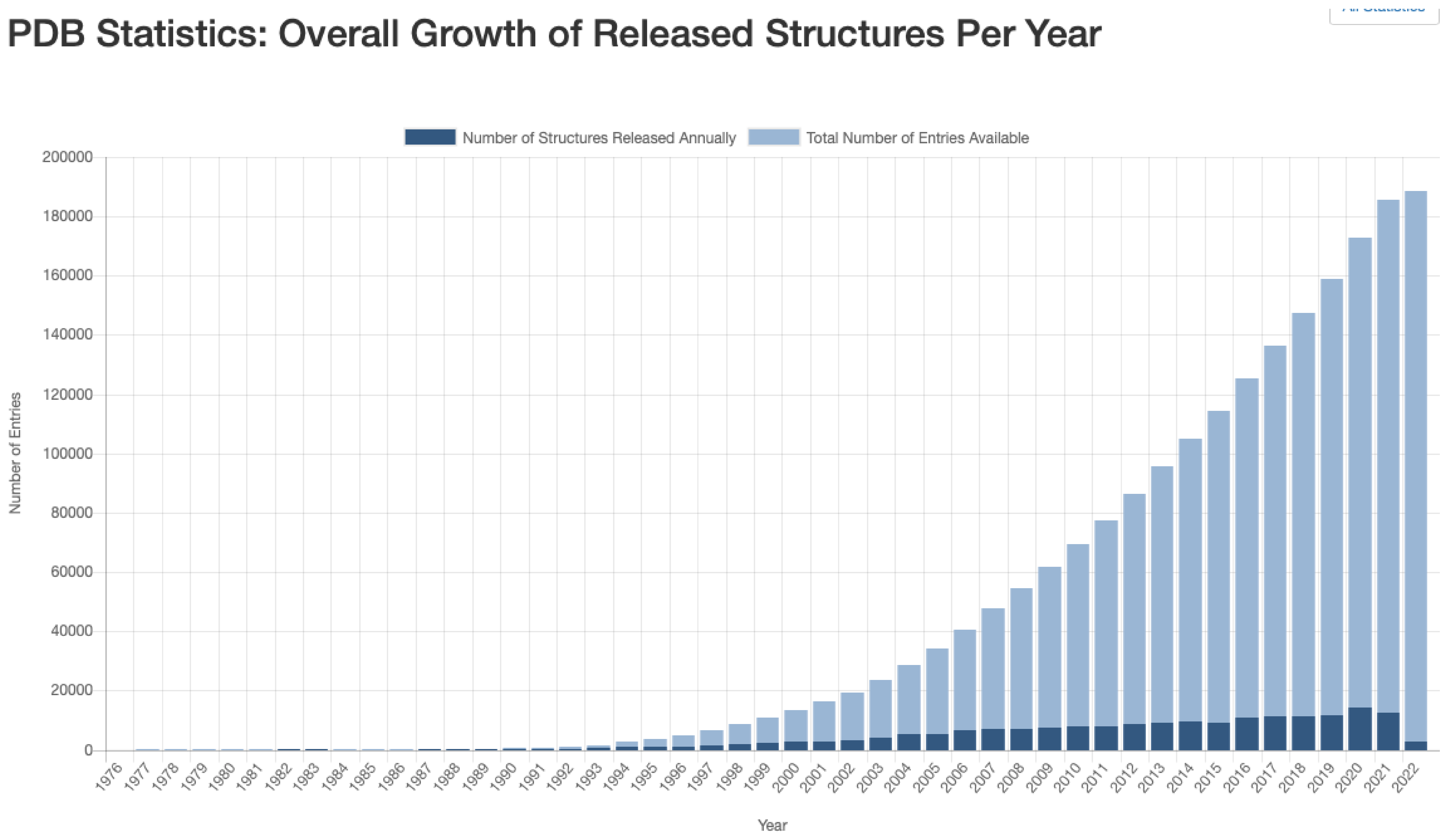
ModelCIF, an extension of PDBx/mmCIF for computed structure models, is now available. A software library called python-modelcif has been developed to support ModelCIF and enables reading and writing mmCIF files

Video: Opioids and Pain Signaling

Watch how opioids activate G-proteins, which in turn interact with other proteins to edit the pain signal

» 03/17/2022

Brain Awareness Week

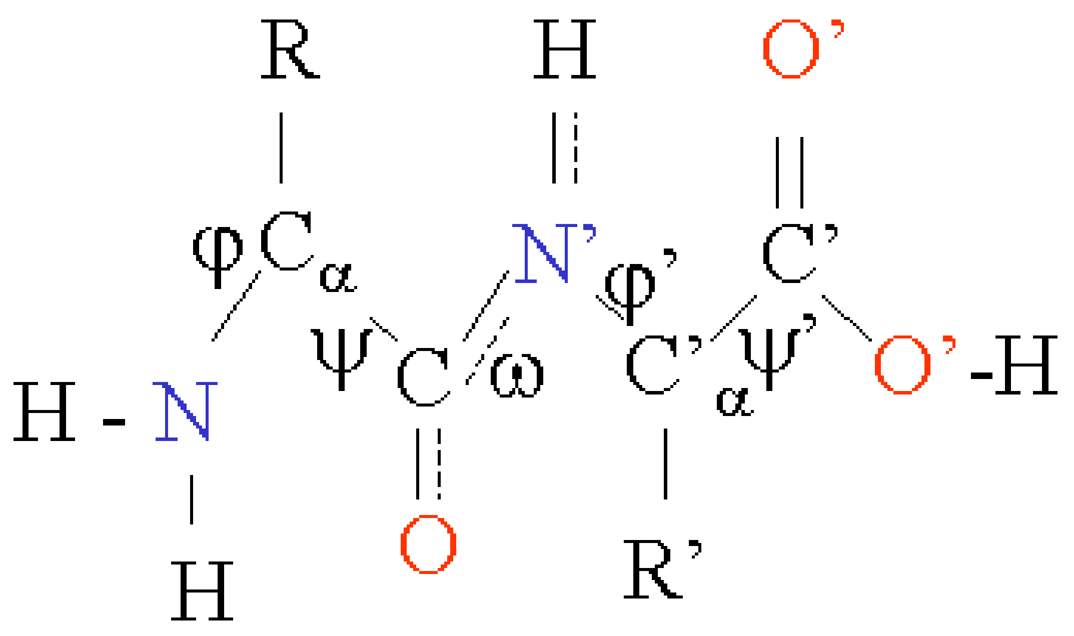


Most publishing journals require a structure be deposited to the PDB prior to publication of a study!

Enter Computers! (Not The One In Your Pocket)

Parameters that MUST be considered:

1. **Primary Structure (bond angles)**
2. **Secondary Structure (α-helix, β-sheet, loops)**
3. **Tertiary Structure (folding of secondary structure)**
4. **Quaternary Structure (Optional)**

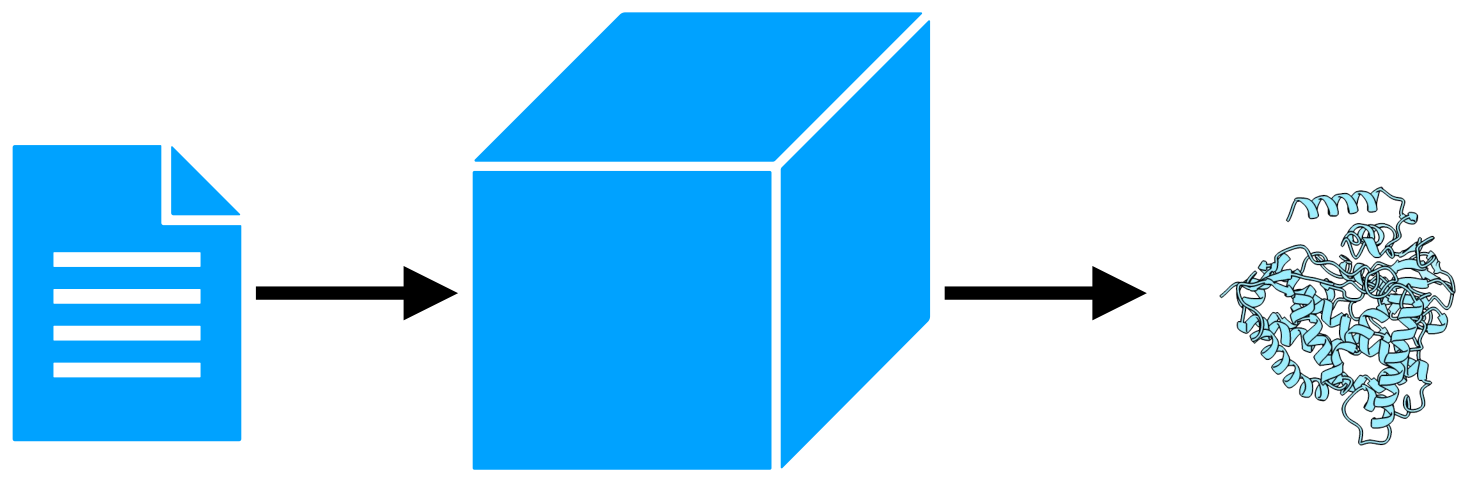


3°/4° Structure Have Additional Parameters:

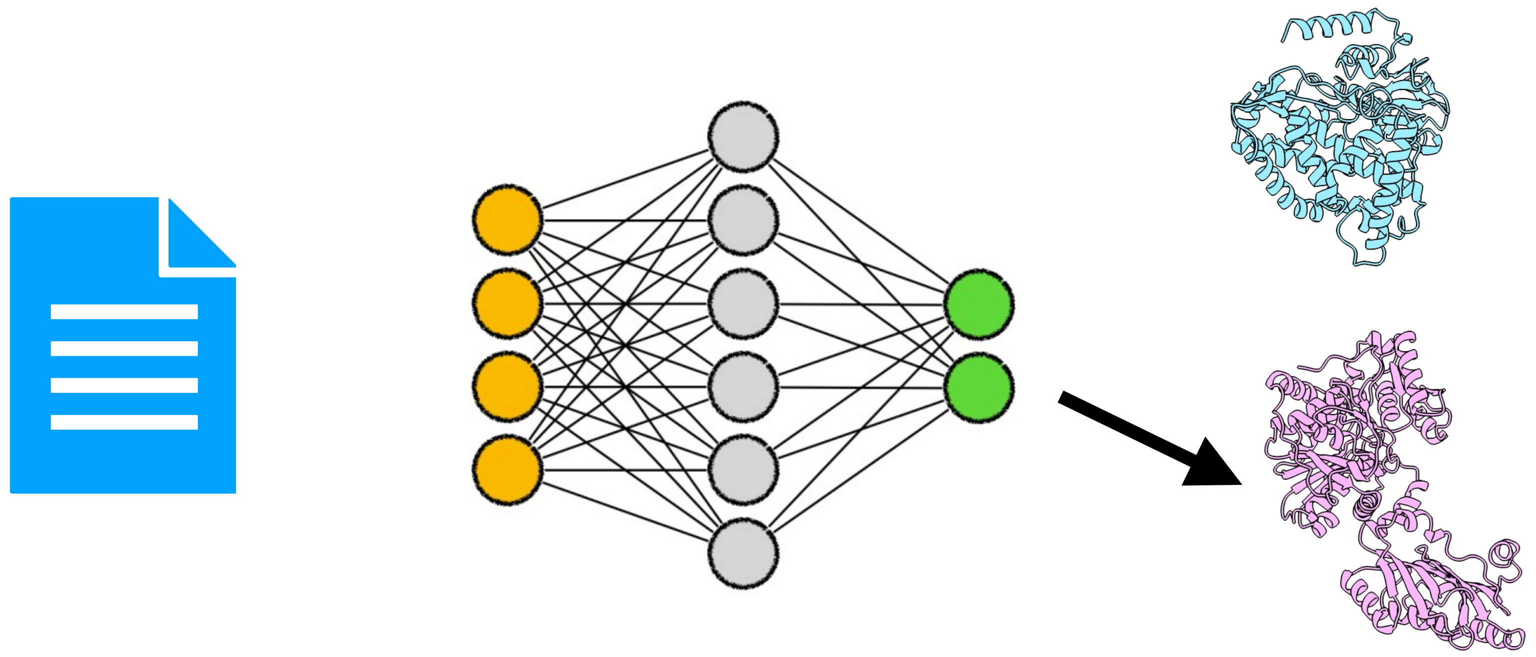
1. **Conserved protein domains**
2. **Protein families/superfamilies/clans**

Ideally, we would like to predict a protein’s 3-D structure given only its AA sequence (1° stucture)

Algorithm-Based



Machine Learning (AI)



Web-Server Based Algorithm Driven Software

I-TASSER

Protein Structure & Function Predictions

- **Widely used molecular replacement modeling system**
- **Often paired with X-ray crystallography during the refinement step**
- **Best case use when a high sequence homolog exists**

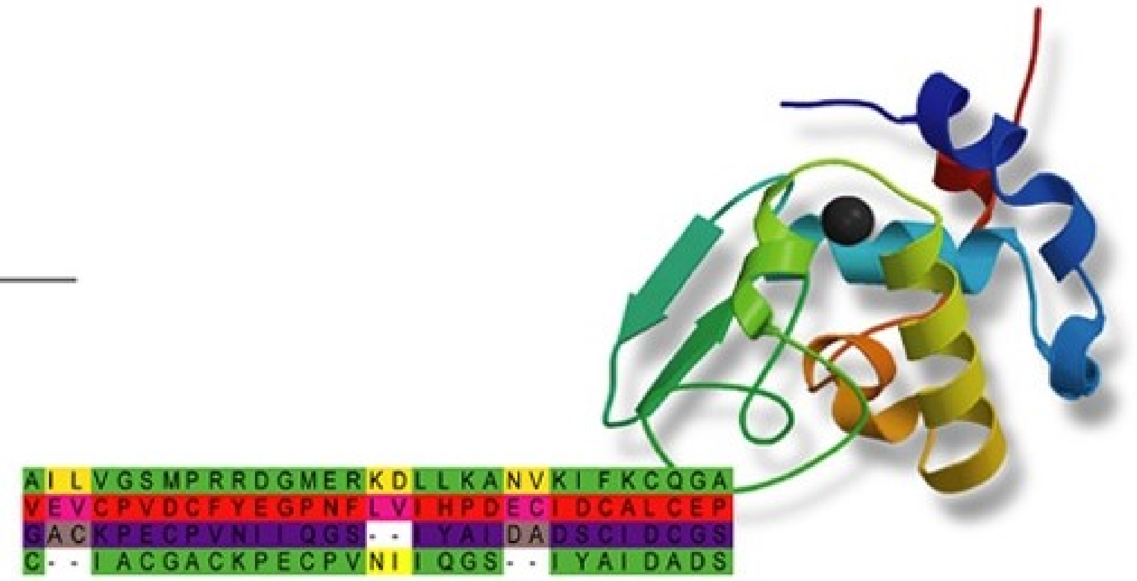
BIOZENTRUM
University of Basel
The Center for Molecular Life Sciences

SWISS-MODEL

- **Fully-integrated with ExPASy Suite**
- **Extensive manually curated databases**

Modeller

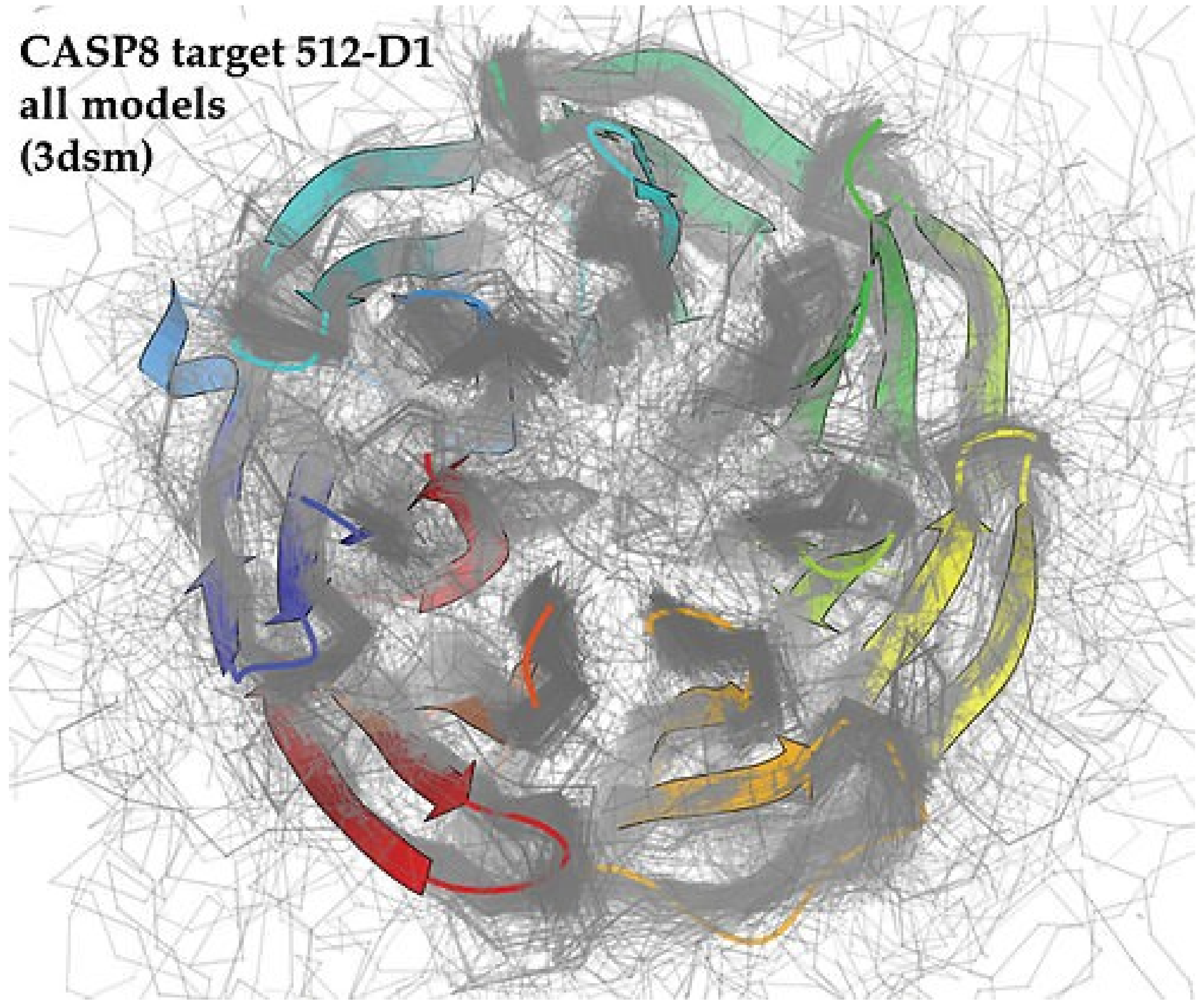
Program for Comparative Protein
Structure Modelling by Satisfaction
of Spatial Restraints



- **CASP rated highest for blind protein search/modeling**
- **Completely open-source**

All of these solutions heavily rely on homology to some characterized protein

Assessing Methods for Protein Structure Prediction



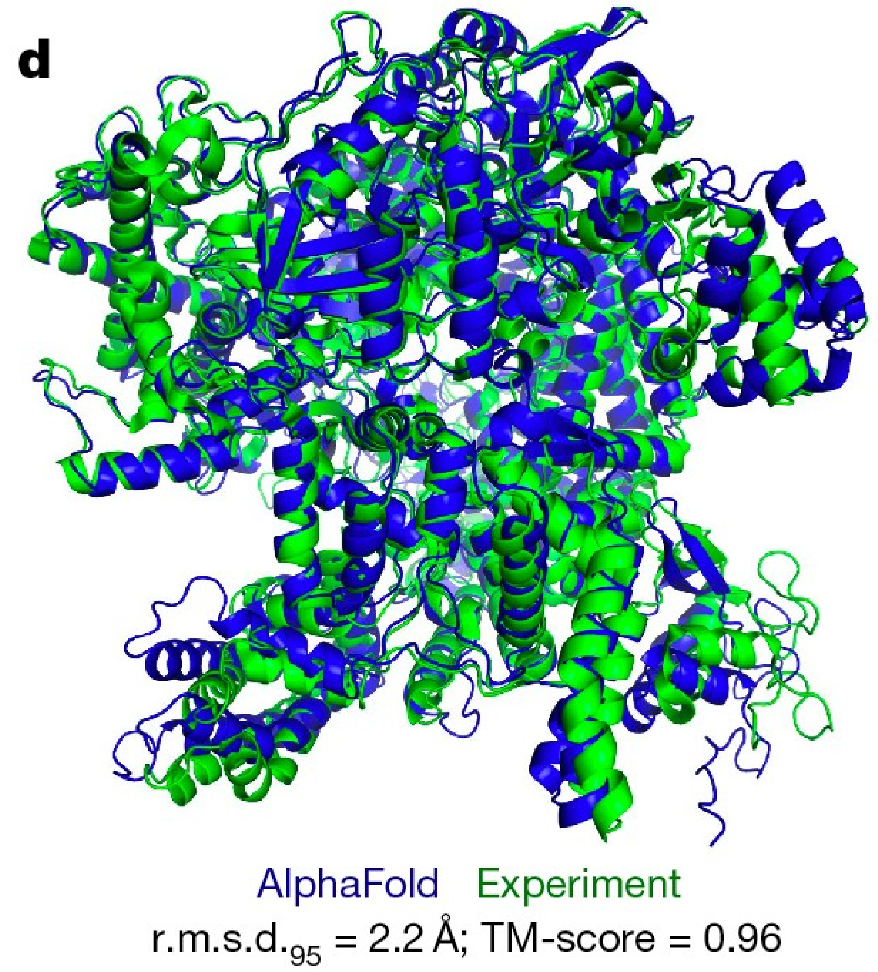
CASP evaluates current methods of predicting protein structure since 1994.

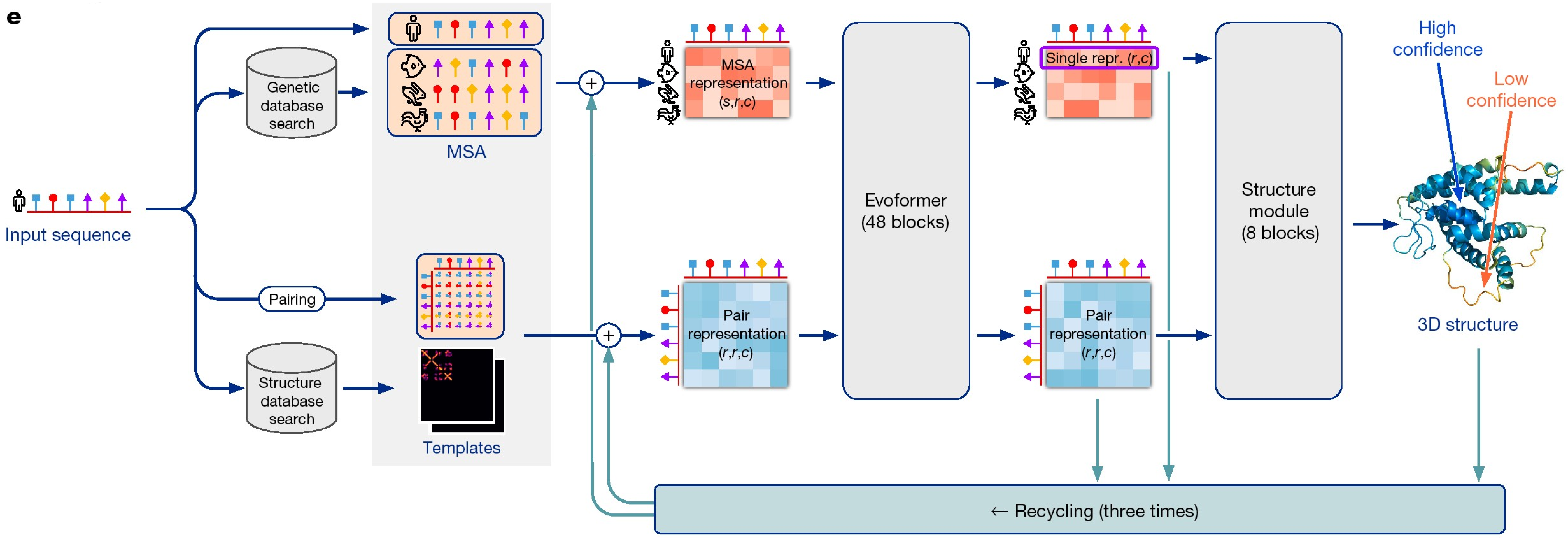
“These techniques are expensive and slow: it can take hundreds of thousands of dollars and years of trial and error for each protein. AlphaFold can find a protein’s shape in a few days.” - MIT Technology Review

- **2020 was the first year a program met a true success rate (>90%)**
  - **DeepMind’s AlphaFold2**
  - **Worked on both homologous and novel proteins**
  - **Only needed to provide the AA sequence**
  - **AlphaFold2 was released under the Apache Common Use license in 2021**

AlphaFold2: A Neural Network Trained To Predict Protein 3-D Structure

- **First program to ever meet the success rate at CASP**
- **Assessed on alignment of predicted structure to experimental**
  - **RMSD (Å)**
- **AF2 custom value to “quantify” model confidence**
  - **pLDDT**

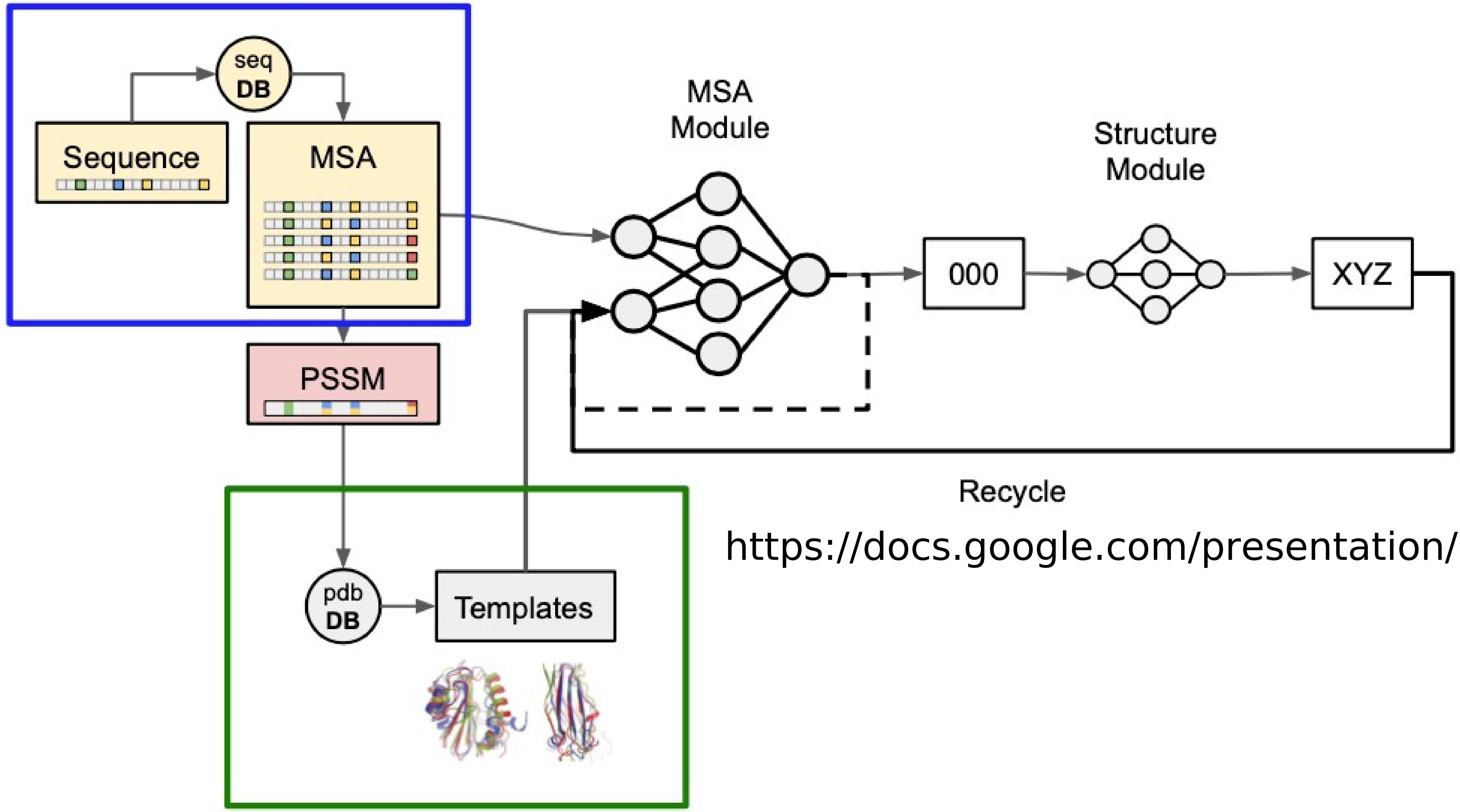



e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

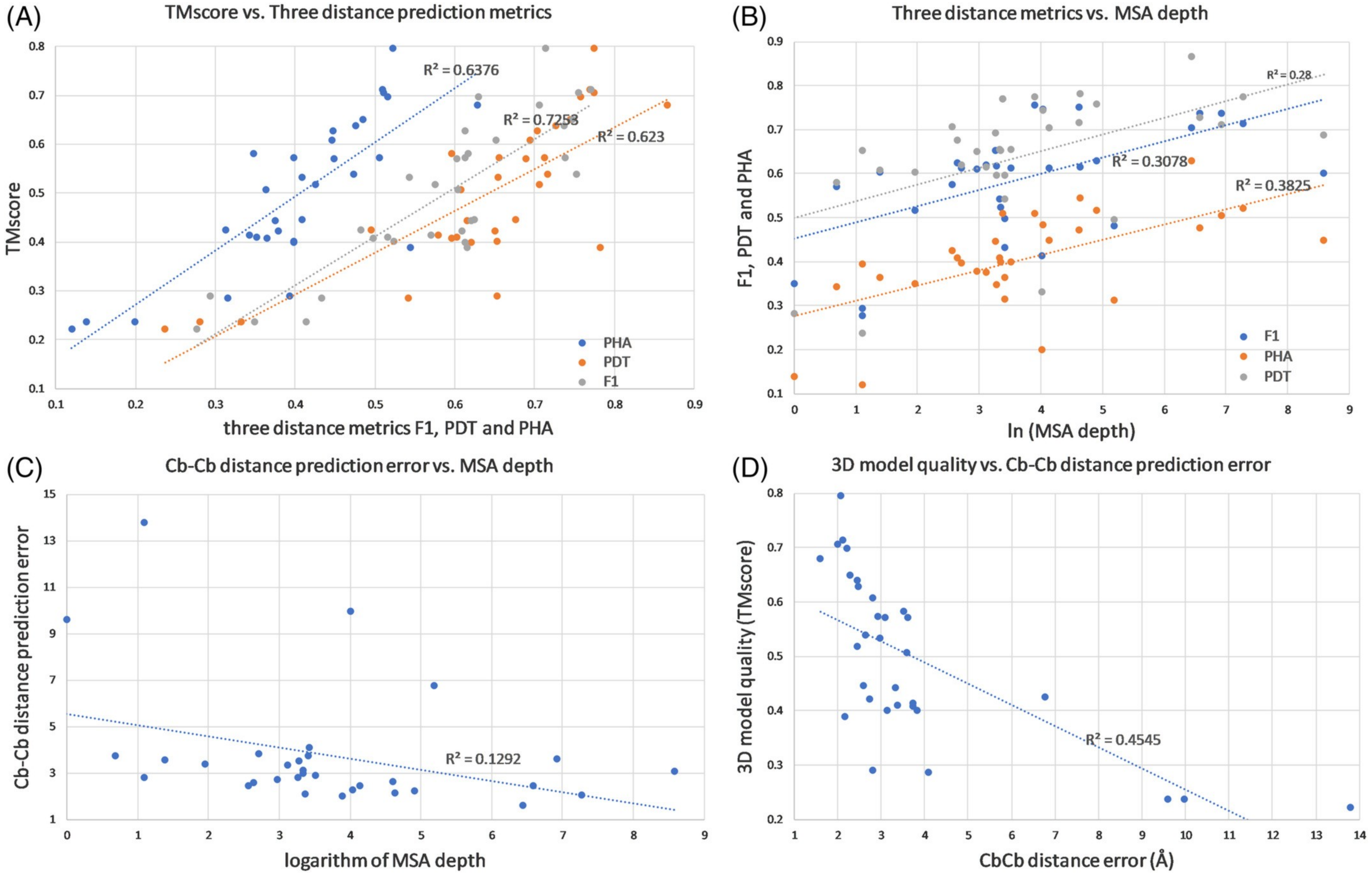
Jumper, *et. al.* 2021 Nature

MSA Depedency of AF2

Structure Prediction is only as good as the input MSA



https://docs.google.com/presentation/d/



Xu, J.*et. al*. *Proteins* 2019, *87* (12), 1069-1081.

Typically MSA depth of >30 sequences/residue produce a highly accurate and confident model

2.2 Tb of Protein Sequence Search Space

Structure Template Search Space

PDB70

PDB_mmcif

PDBseqres

Recently, it has been observed that structure templates have little to no impact on the final predicted strucutre's accuracy.

Sequencing Alignment Search Space

MGnify (bacterial proteome darkspace)

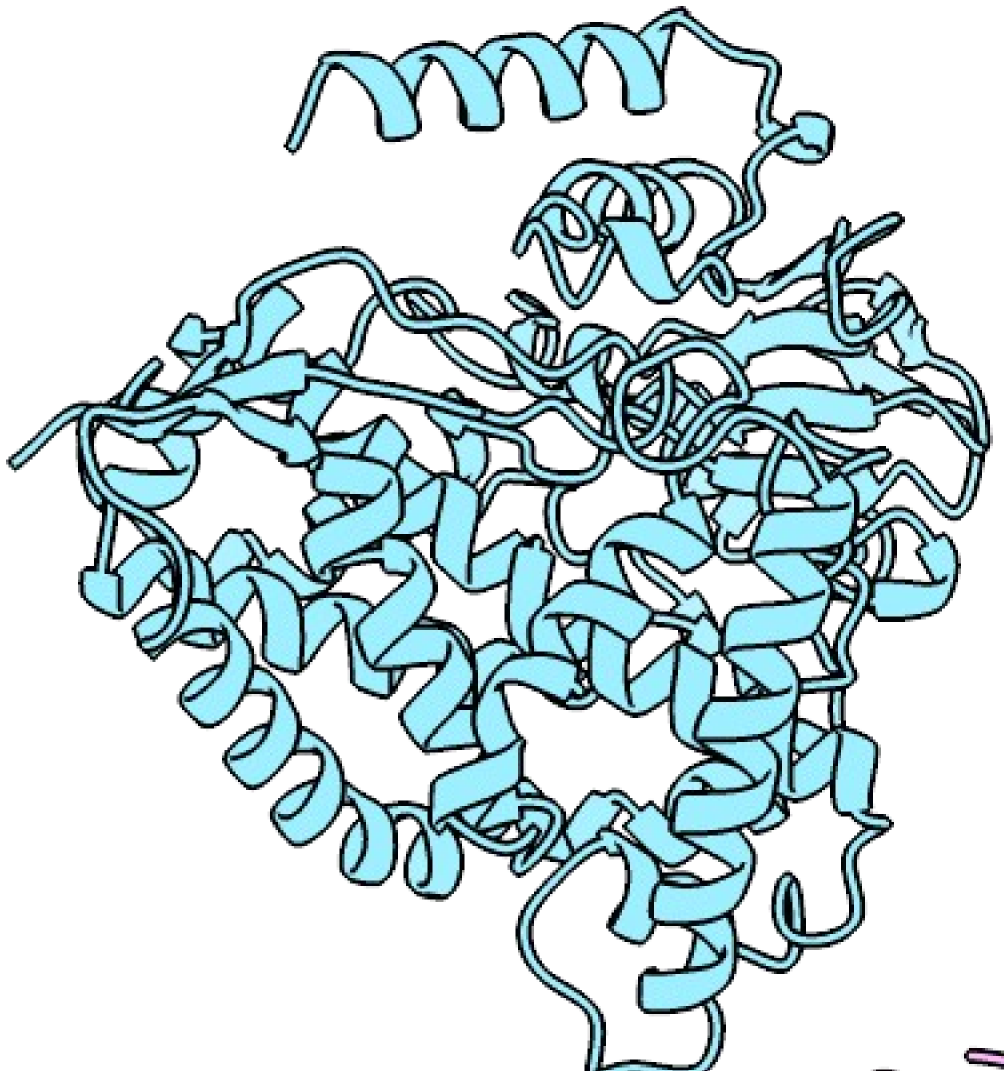
Uniref (dereplicated UniprotKB)

Uniclust30 (% clustering of UniprotKB)

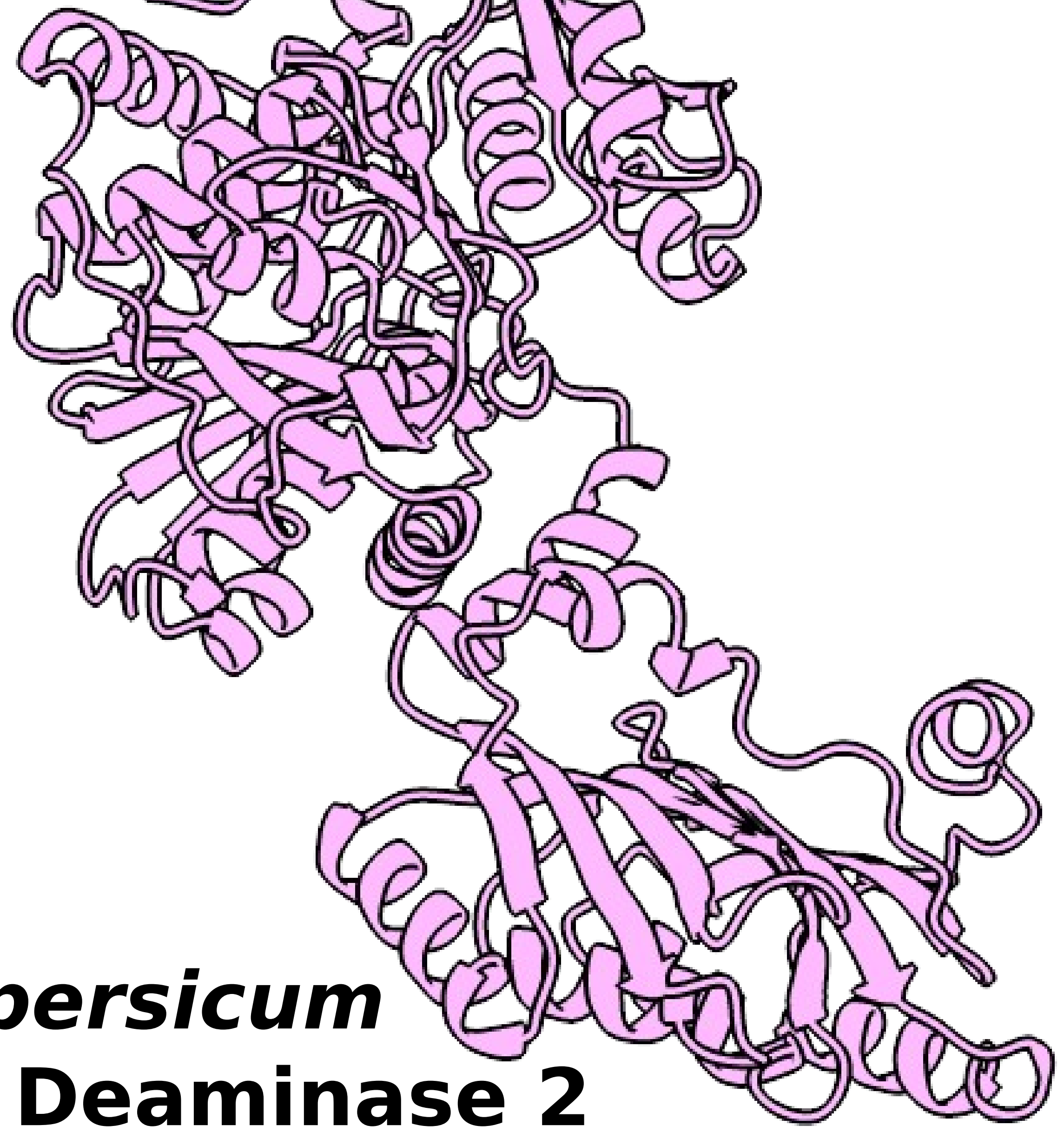
BFD (massive cluster of bacterial proteome space)

Uniprot Database)

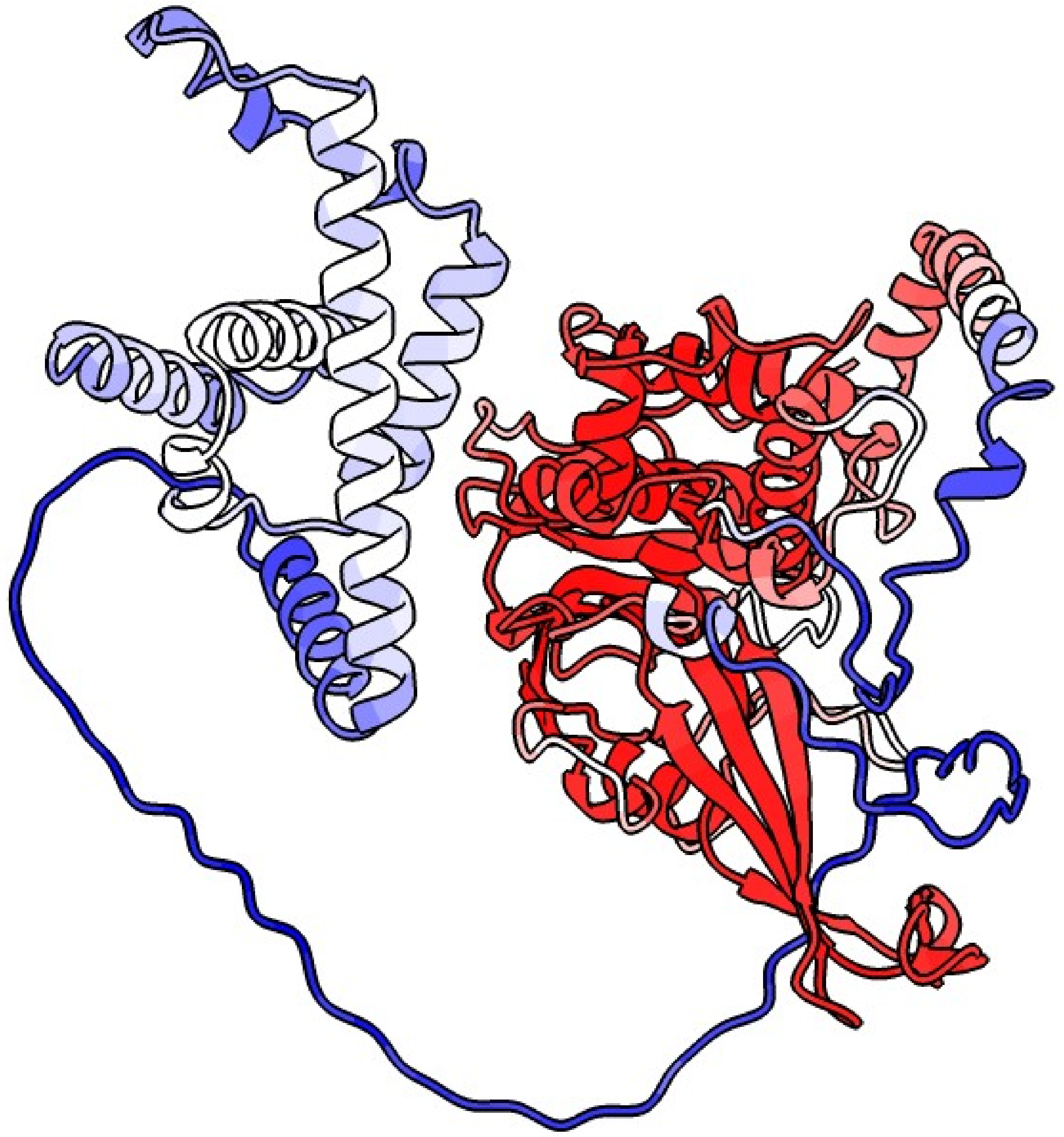
Should/Could/Would….. Just Do It!



CYP153A - Bacterial cytochrome p450
(collaboration)



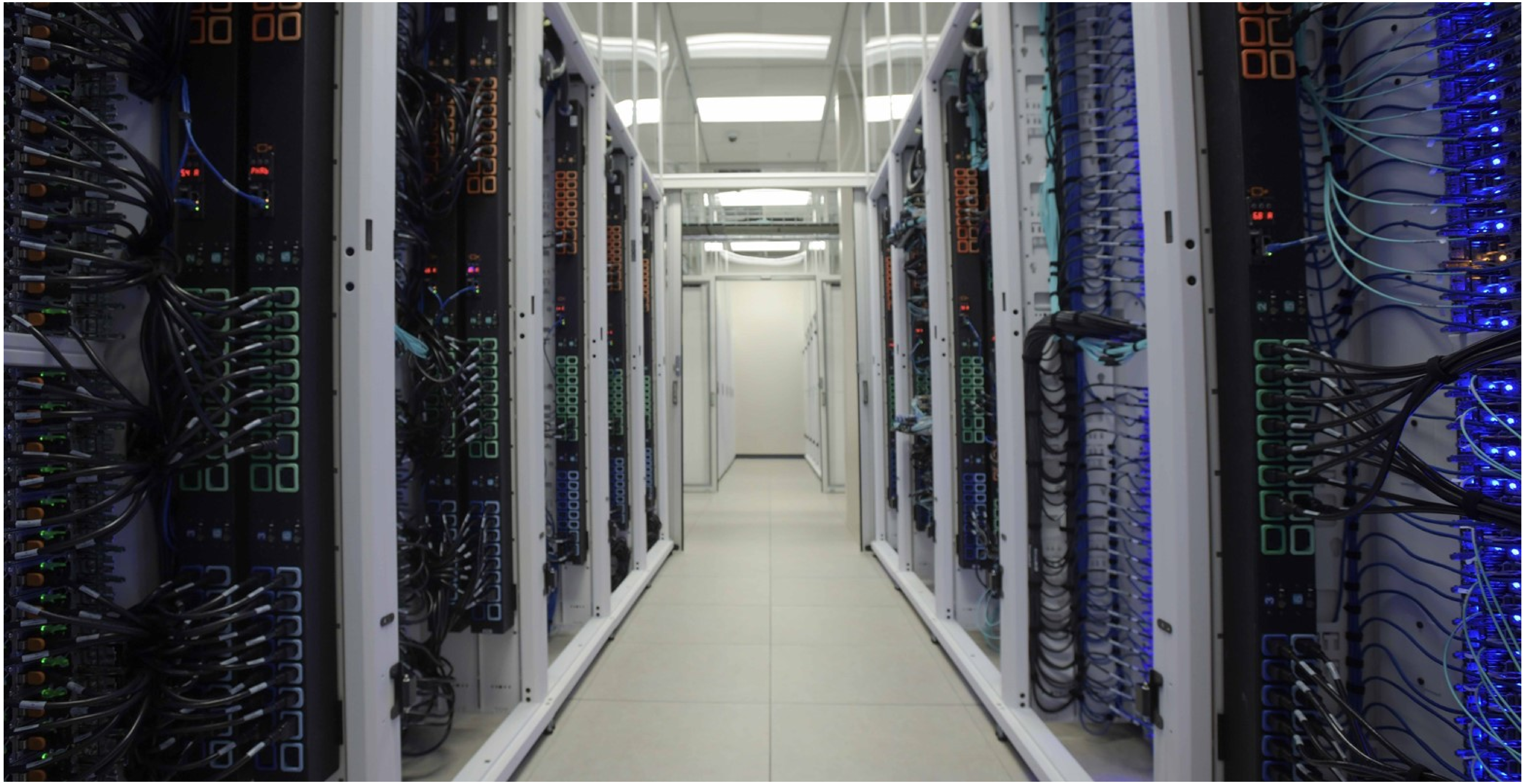
S. lycopersicum
Threonine Deaminase 2



Flagellum-22 Up-regulated
S. lycopersicum Protease

- No “cost” to you
- Best accuracy with homolog. seqns.
- Running the same protein yeilds differences (Å)
  - CollabFold number of recycling steps

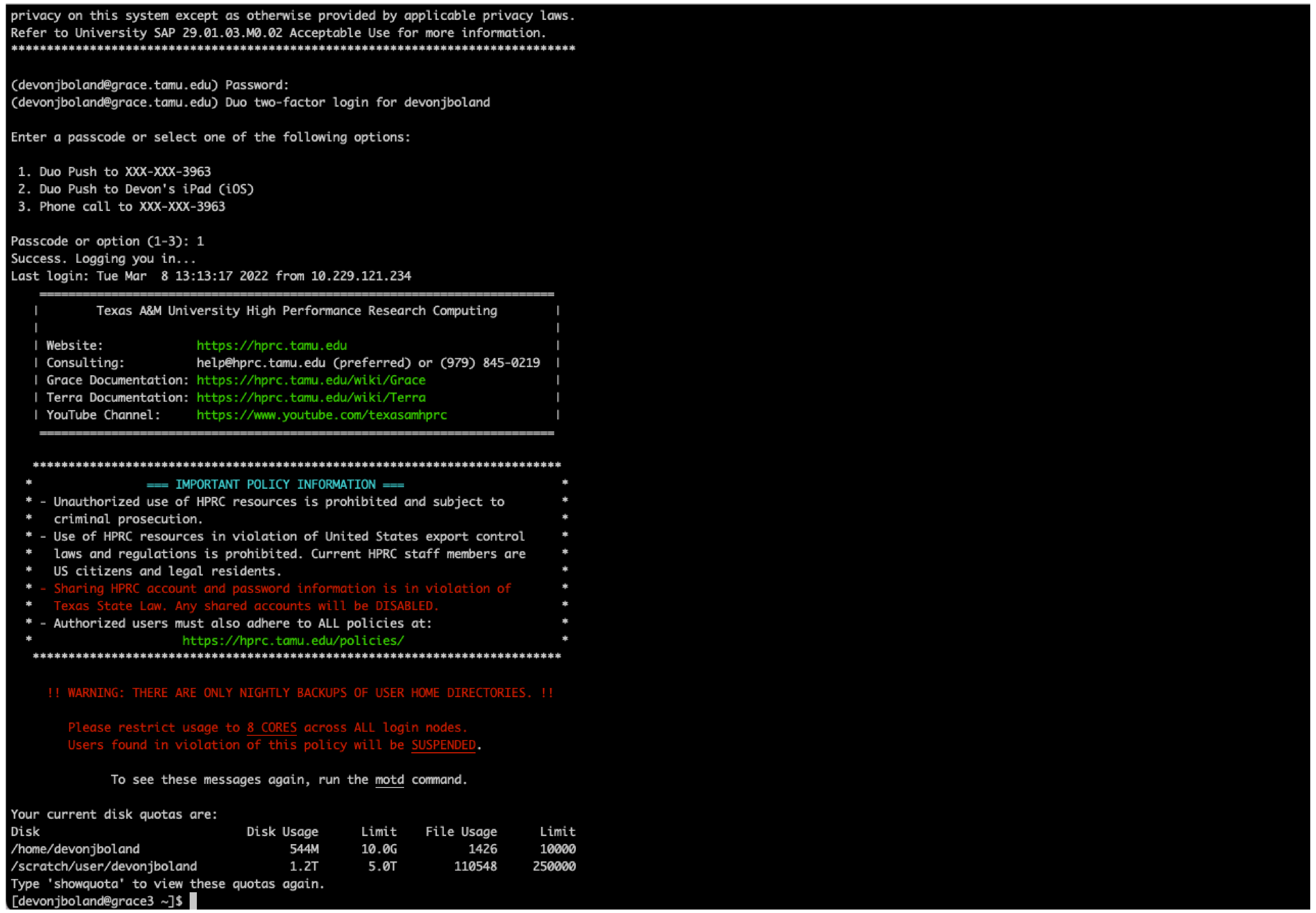
Grace Cluster



“Grace is a 925-node Intel cluster from Dell with an InfiniBand HDR-100 interconnect, A100 GPUs, RTX 6000 GPUs and T4 GPUs. All nodes are based on the Intel Cascade Lake processor.” - HPRC

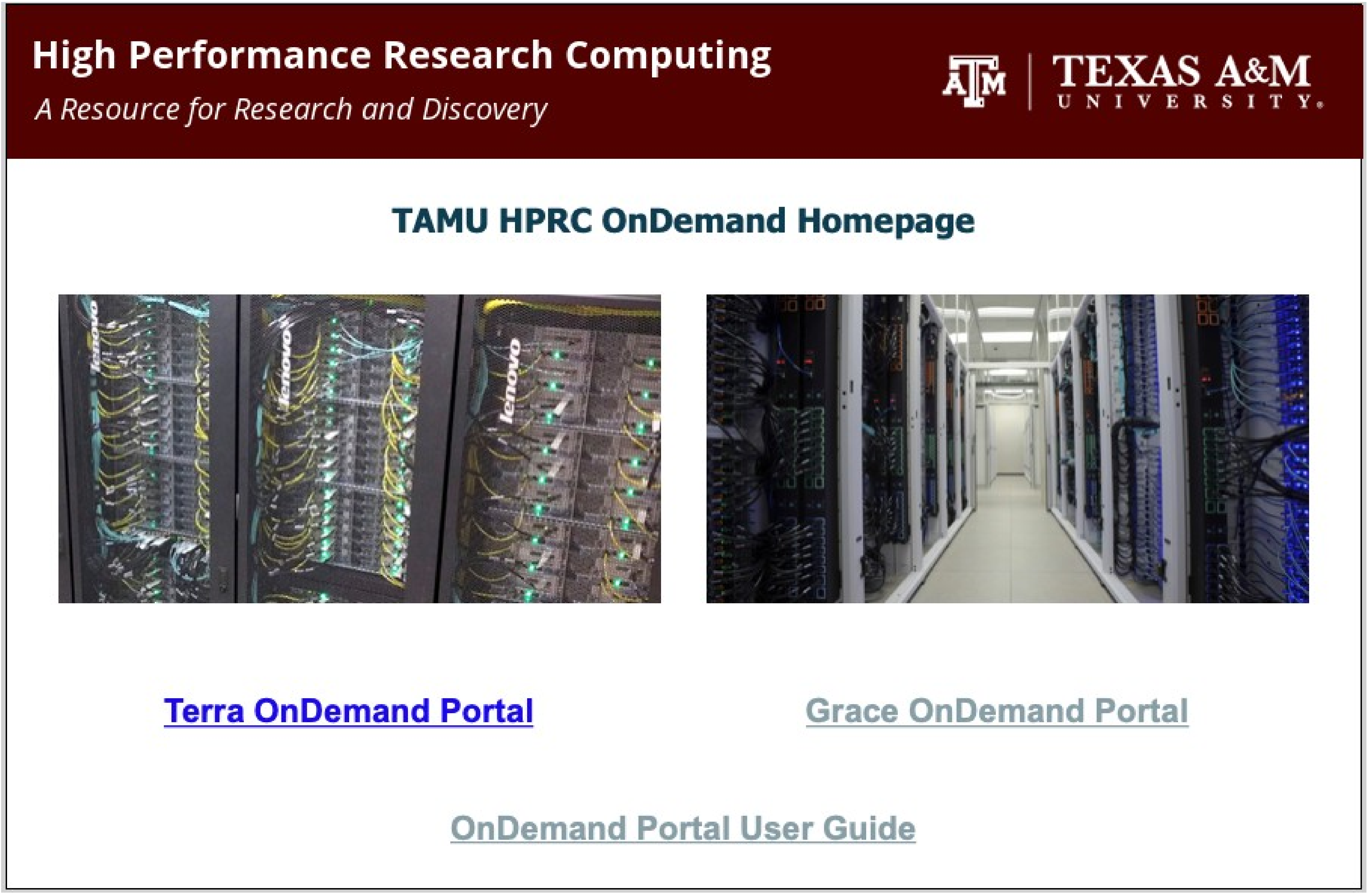
Using AlphaFold2

Terminal via SSH (Local Client)



```
privacy on this system except as otherwise provided by applicable privacy laws.
Refer to University SAP 29.01.03.M0.02 Acceptable Use for more information.
*******************************************************************************

(devonjboland@grace.tamu.edu) Password:
(devonjboland@grace.tamu.edu) Duo two-factor login for devonjboland

Enter a passcode or select one of the following options:

 1. Duo Push to XXX-XXX-3963
 2. Duo Push to Devon's iPad (iOS)
 3. Phone call to XXX-XXX-3963

Passcode or option (1-3): 1
Success. Logging you in...
Last login: Tue Mar  8 13:13:17 2022 from 10.229.121.234
    ========================================================================
   |        Texas A&M University High Performance Research Computing        |
   |                                                                        |
   | Website:              https://hprc.tamu.edu                            |
   | Consulting:           help@hprc.tamu.edu (preferred) or (979) 845-0219 |
   | Grace Documentation:  https://hprc.tamu.edu/wiki/Grace                 |
   | Terra Documentation:  https://hprc.tamu.edu/wiki/Terra                 |
   | YouTube Channel:      https://www.youtube.com/texasamhprc              |
    ========================================================================

   **************************************************************************
  *                 === IMPORTANT POLICY INFORMATION ===                      *
  * - Unauthorized use of HPRC resources is prohibited and subject to         *
  *   criminal prosecution.                                                   *
  * - Use of HPRC resources in violation of United States export control     *
  *   laws and regulations is prohibited. Current HPRC staff members are     *
  *   US citizens and legal residents.                                        *
  * - Sharing HPRC account and password information is in violation of       *
  *   Texas State Law. Any shared accounts will be DISABLED.                  *
  * - Authorized users must also adhere to ALL policies at:                   *
  *                  https://hprc.tamu.edu/policies/                          *
   **************************************************************************

     !! WARNING: THERE ARE ONLY NIGHTLY BACKUPS OF USER HOME DIRECTORIES. !!

        Please restrict usage to 8 CORES across ALL login nodes.
        Users found in violation of this policy will be SUSPENDED.

              To see these messages again, run the motd command.

Your current disk quotas are:
Disk                            Disk Usage      Limit    File Usage      Limit
/home/devonjboland                    544M      10.0G          1426      10000
/scratch/user/devonjboland            1.2T       5.0T        110548     250000
Type 'showquota' to view these quotas again.
[devonjboland@grace3 ~]$ 
```

OnDemand Portal (Web Browser)



High Performance Research Computing

A Resource for Research and Discovery

TEXAS A&M UNIVERSITY

TAMU HPRC OnDemand Homepage

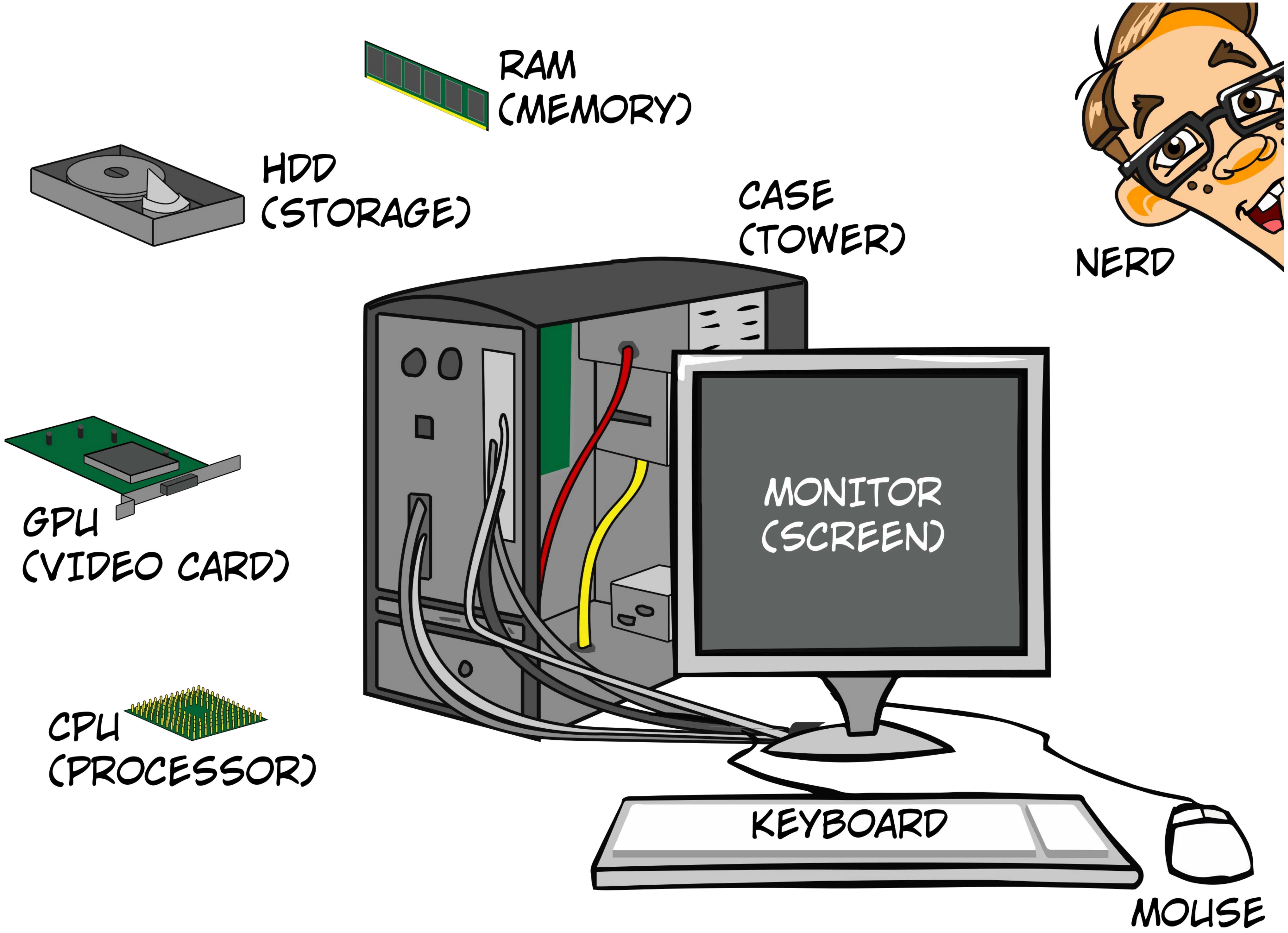
[Terra OnDemand Portal]

[Grace OnDemand Portal]

[OnDemand Portal User Guide]

Google Colab Notebook (Colab Fold)

AF2 is Not a Well Optimized Program



Me...

Summary

- AlphaFold2 was the first program to be considered a success for protein structure prediction
- It works on proteins with structural homologs, and those without
- Its major limitation is still requiring homologous sequences for accurate structure prediction
- Its dependent on two parts:
  - Finding homologous sequences
  - Finding homologous structures
- The confidence of the model is expressed as pLDDT
- When in doubt, run AF2 on your protein, worst that happens is you waste your time

Any Questions!

devonjboland@tamu.edu