

# **Introduction to AlphaFold: Bioinformatics for 3D Protein Structure Prediction**

Devon J. Boland  
Norman Borlaug Endowed Research Scholar

# The Field of Bioinformatics

- “Omics”-Based
- Genome
  - Transcriptome
  - Proteome

Annotation

Gene/Protein Expression

Molecular Dynamic Simulations

Image/Microscopy Analysis

Structural Proteomics

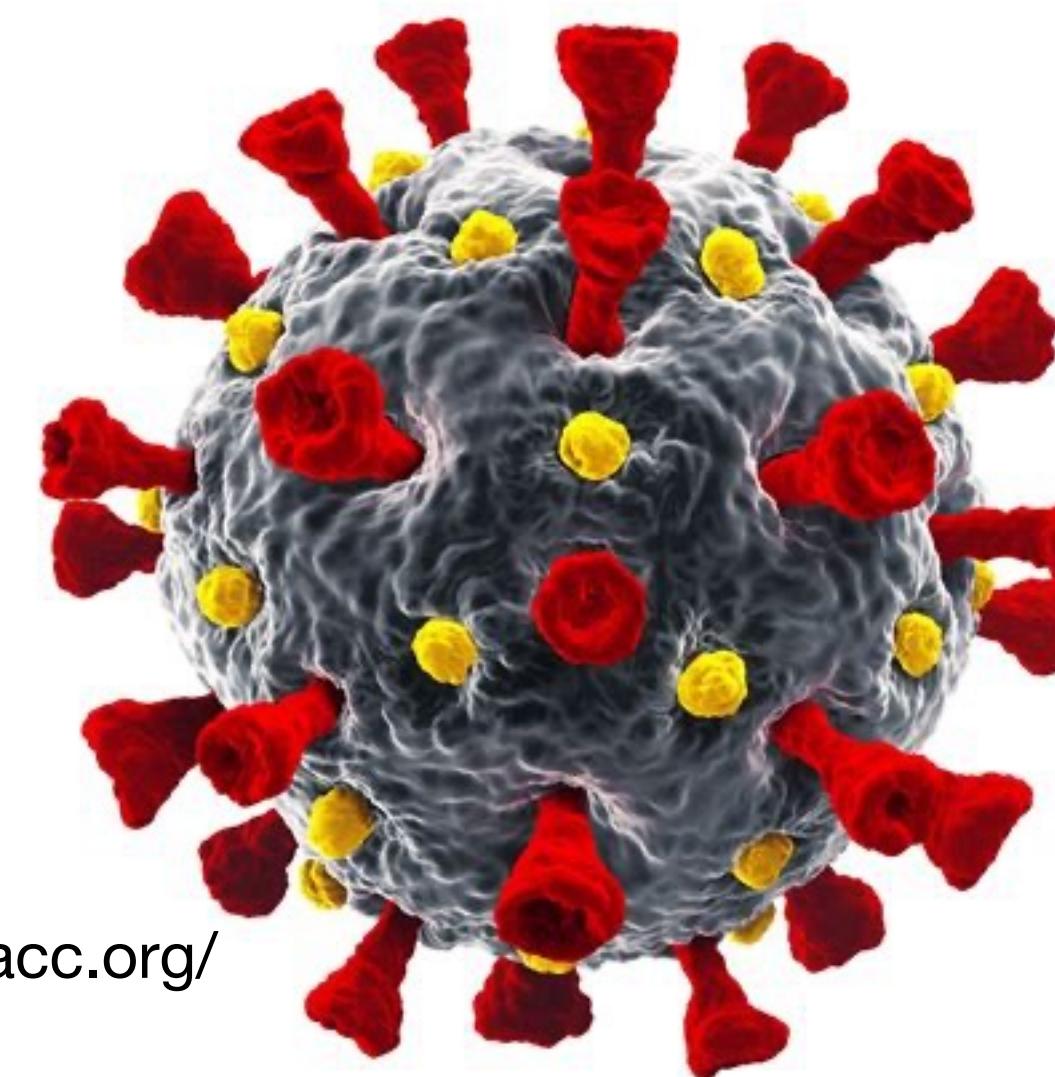


De Maio, C., et al. (2018). Text Mining Basics in Bioinformatics.

# What Does Protein Structure Tell Us?

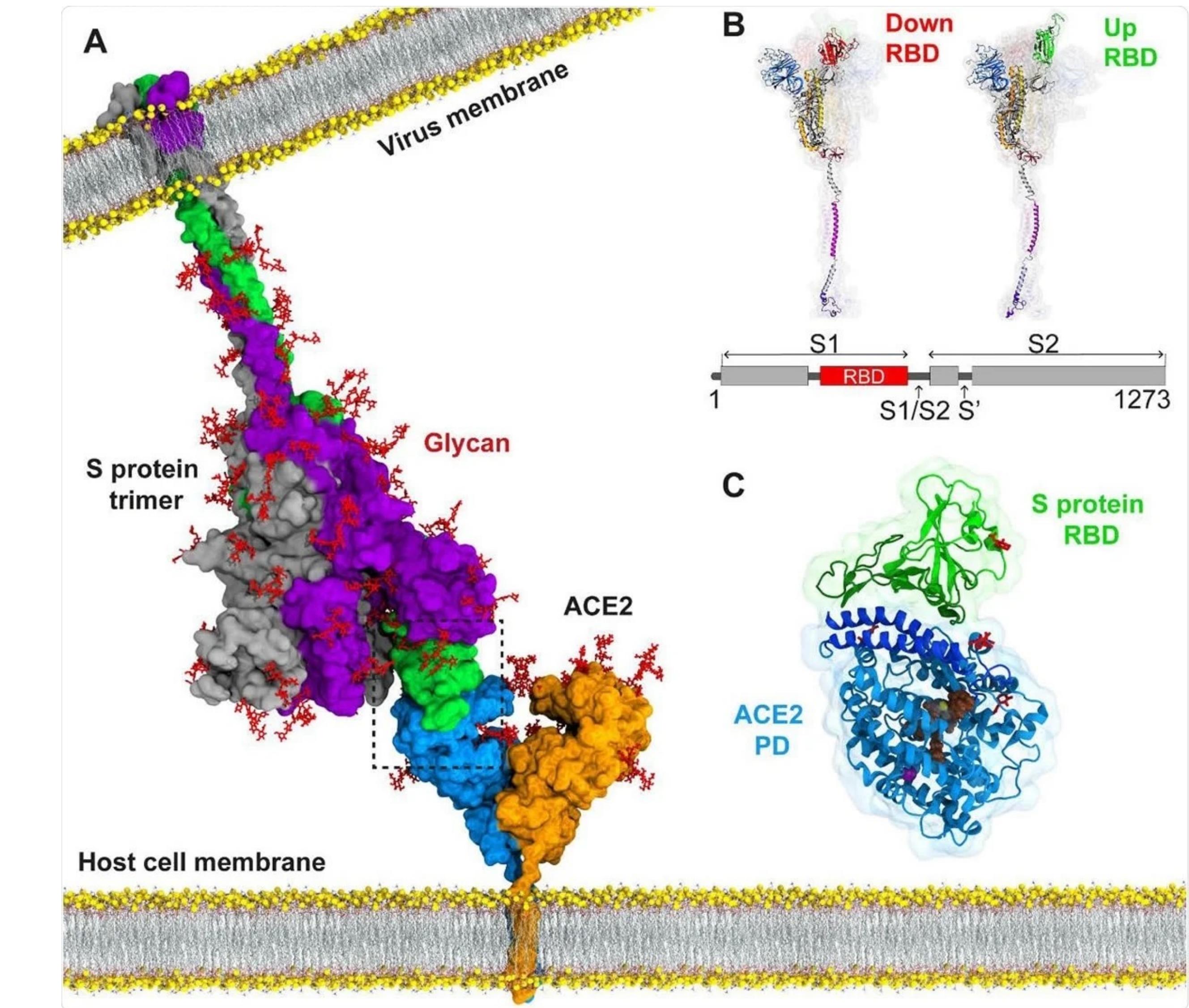
**Structure = Function → Function = Structure**

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



<https://www.acc.org/>

SARS-CoV-2

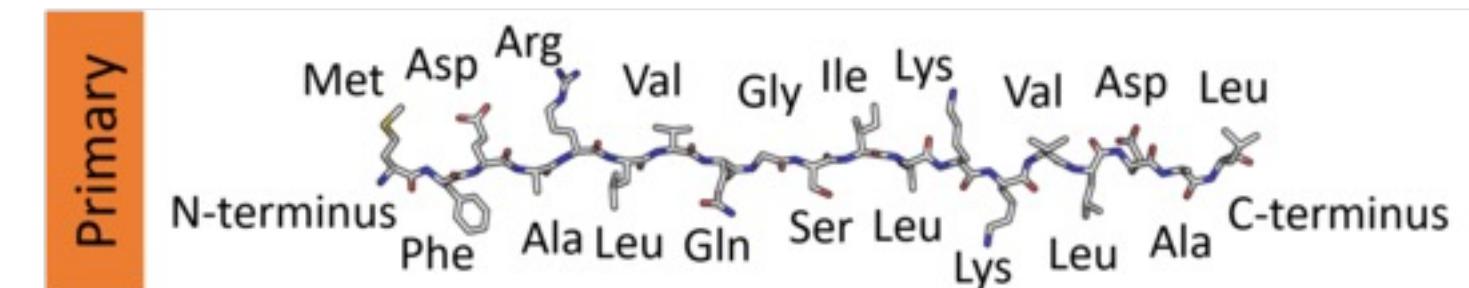


Taka, E., et al. (2021). *The Journal of Physical Chemistry B* 125(21): 5537-5548.

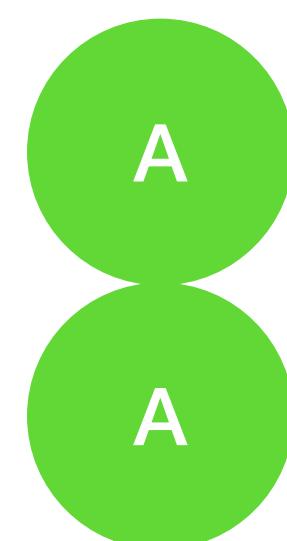
# Basics of 3D Protein Structure

## Four Tiers of Protein Structure

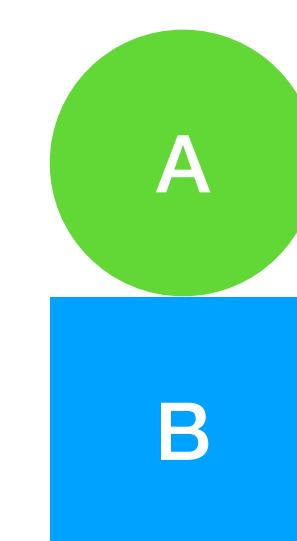
- 1° - Sequence of AA's (polypeptides)
- 2° - Interactions of the carbon backbone of 1°
- 3° - Folding of 2° onto itself
- 4° - Multiple 3° units (Monomers) assembling together



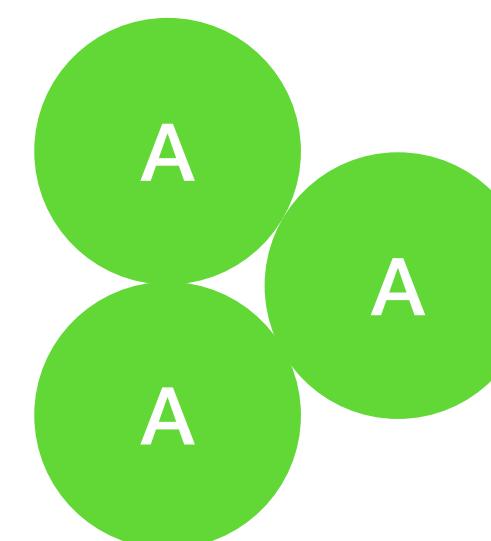
## Different Types of Quaternary Structure



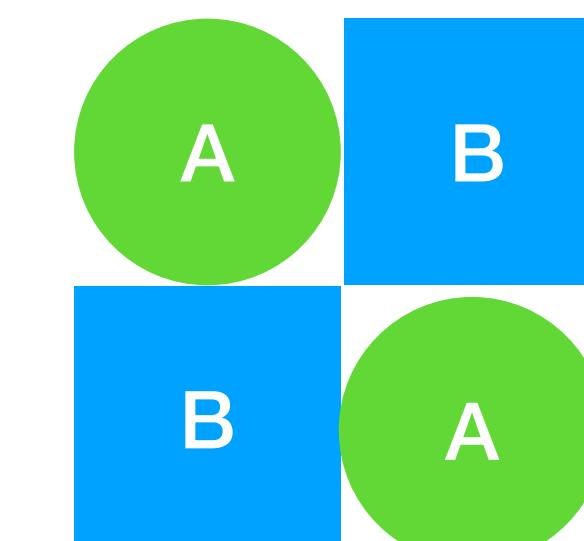
Dimer  
(homo)



Dimer  
(hetero)



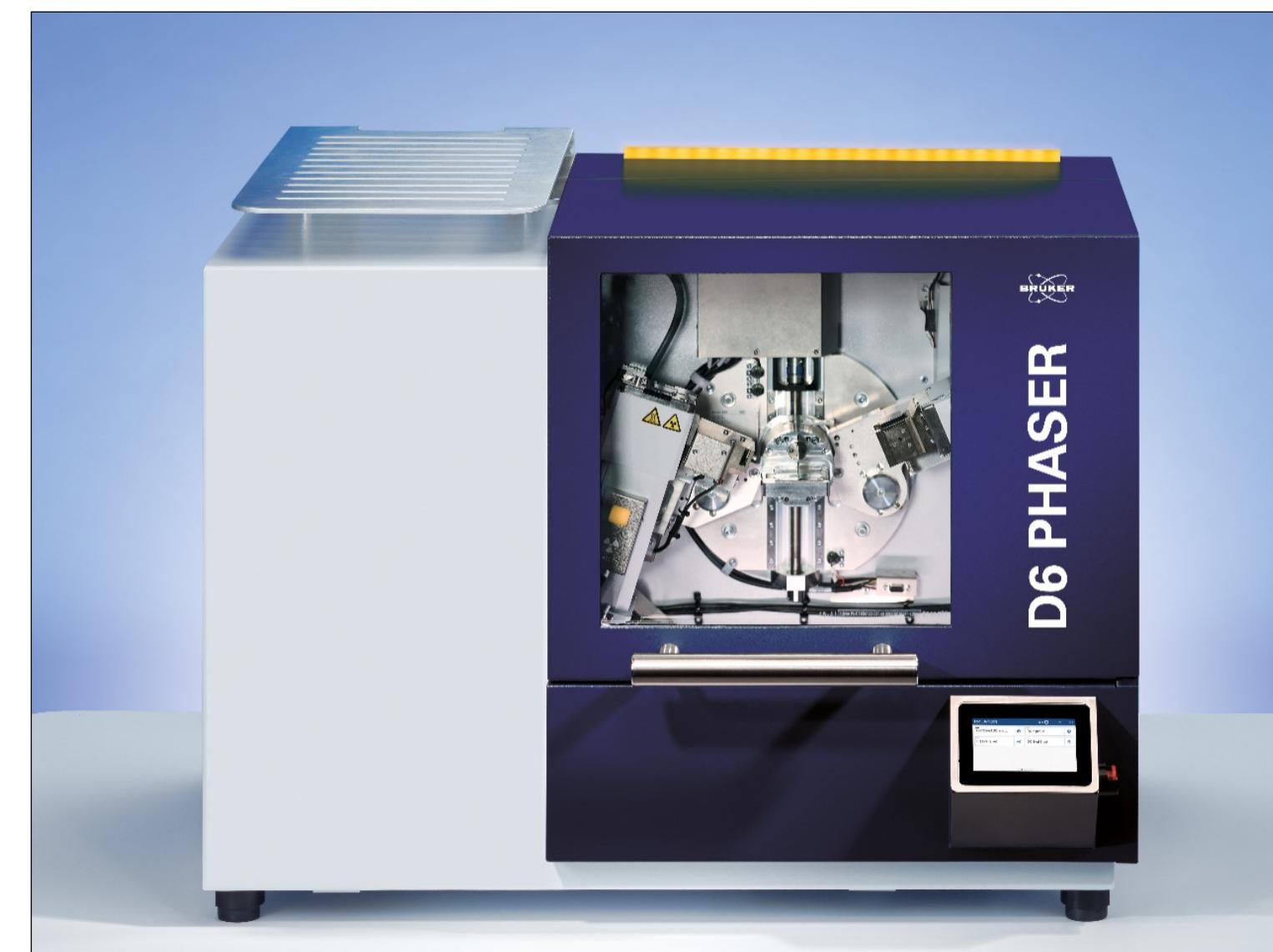
Trimer  
(homo)



Tetramer  
(hetero)

# Methods for Elucidating Protein Structure

## NMR Spectroscopy



<https://www.bruker.com/>

### pros:

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

### cons:

- crystallization process
- high variation

### pros:

- native state
- non-destructive
- real-time

### cons:

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

## Cryo-EM



<https://www.thermoscientific.com/>

### pros:

- large proteins/complexes
- near-native state

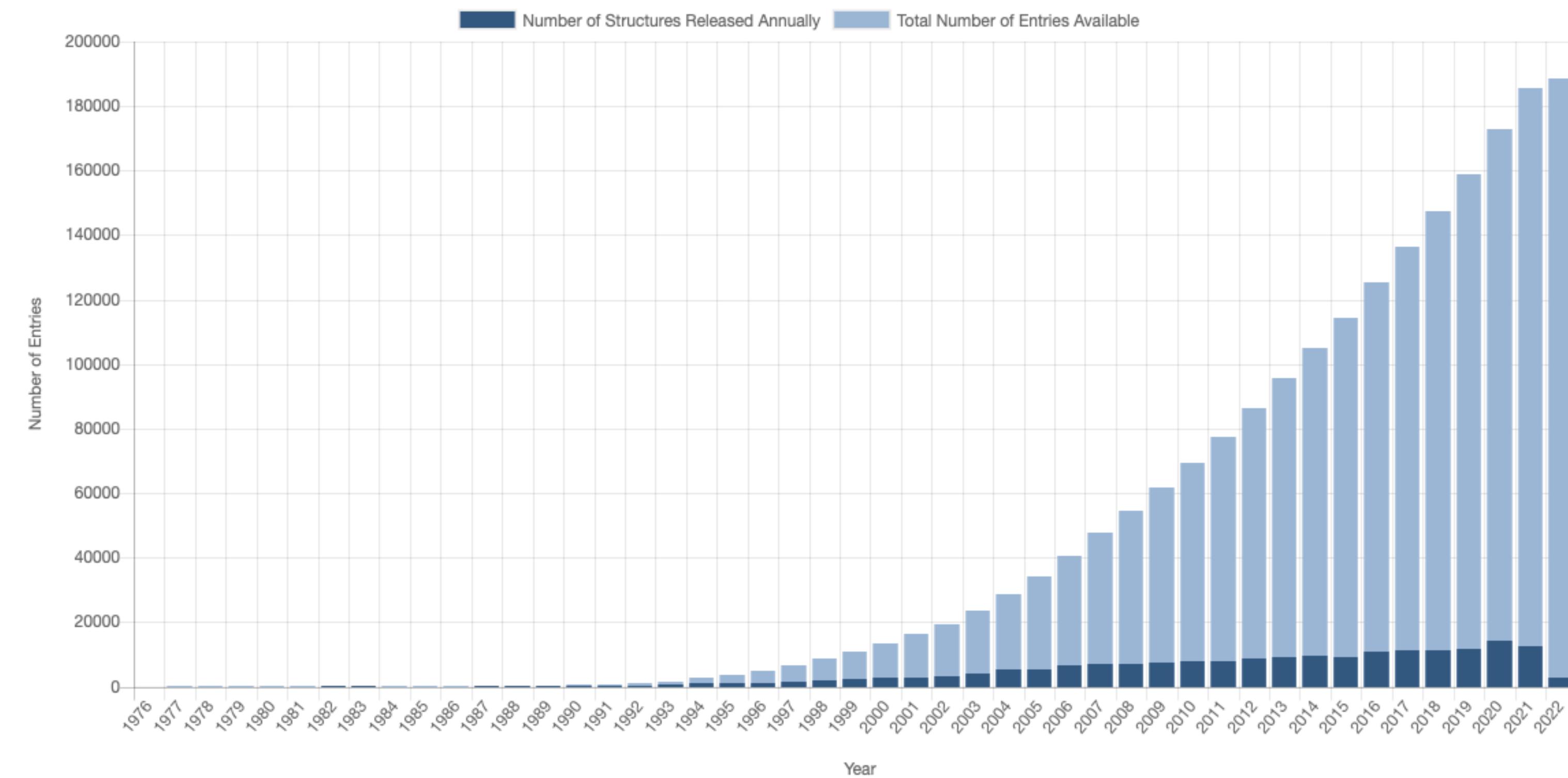
### cons:

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

# Protein Data Bank (PDB)

After we elucidate a structure where does it go?

## PDB Statistics: Overall Growth of Released Structures Per Year



**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:

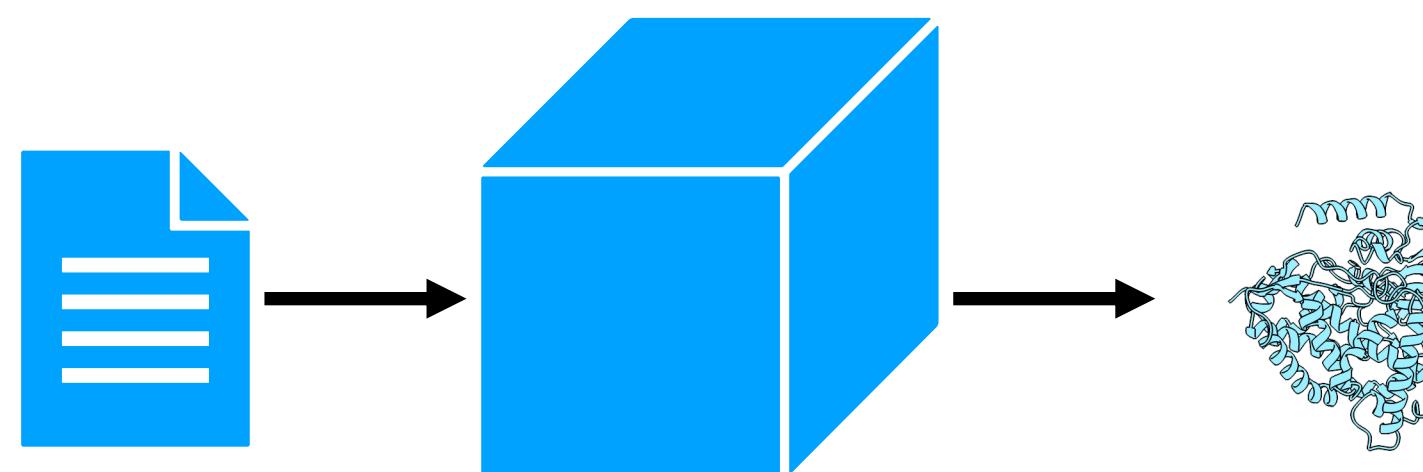
1. Primary Structure (bond angles)
2. Secondary Structure ( $\alpha$ -helix,  $\beta$ -sheet, loops)
3. Tertiary Structure (folding of secondary structure)
4. Quaternary Structure (Optional)

3°/4° Structure Have Additional Parameters:

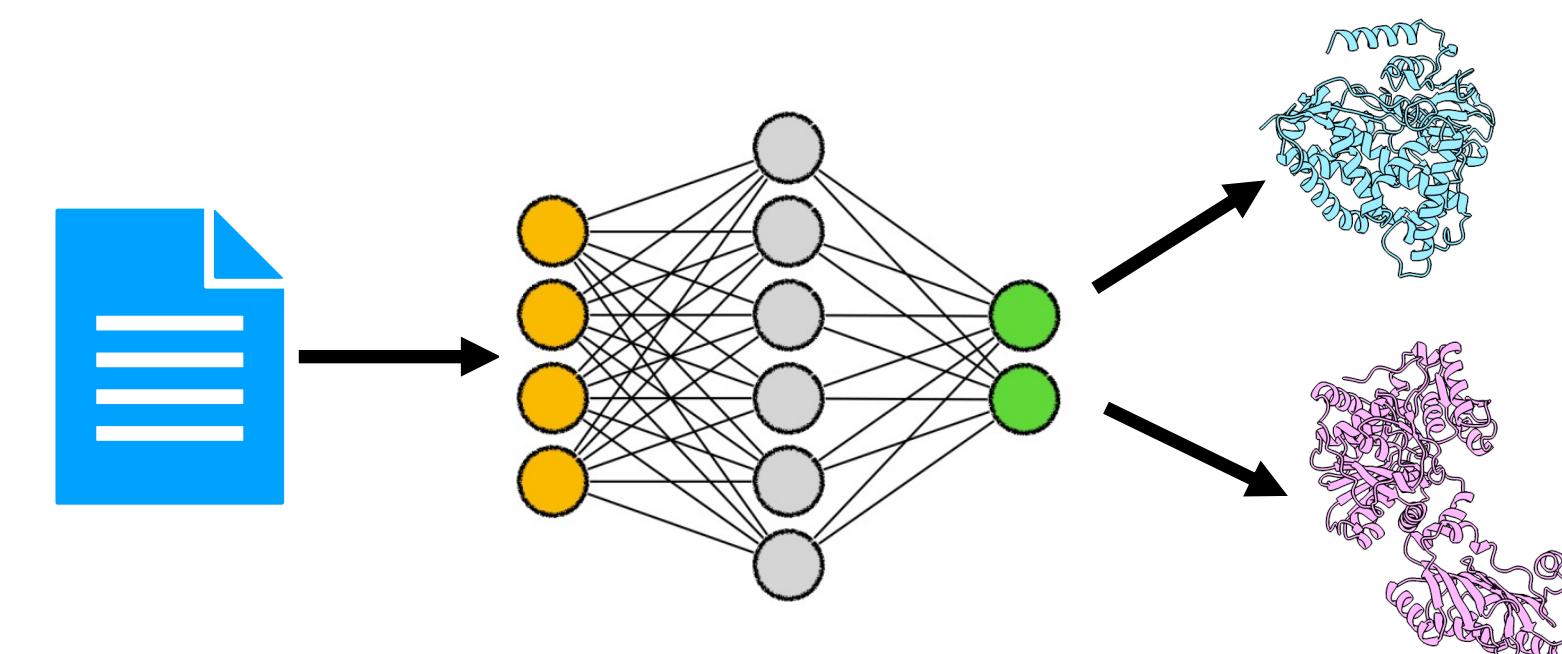
1. Conserved protein domains
2. Protein families/superfamilies/clans

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

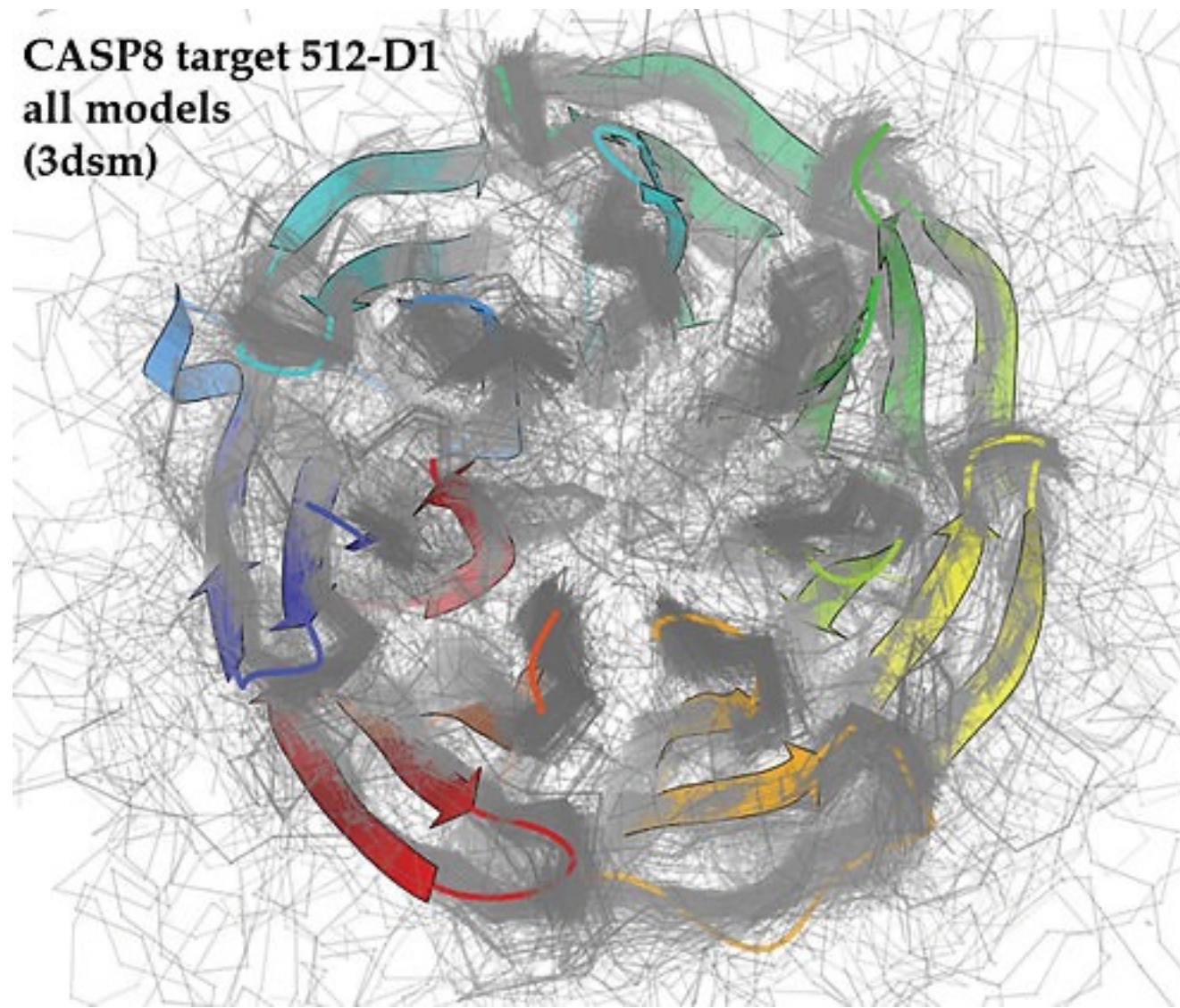
Algorithm-Based



Machine Learning (AI)



# Assessing Methods for Protein Structure Prediction



<https://en.wikipedia.org/wiki/CASP>

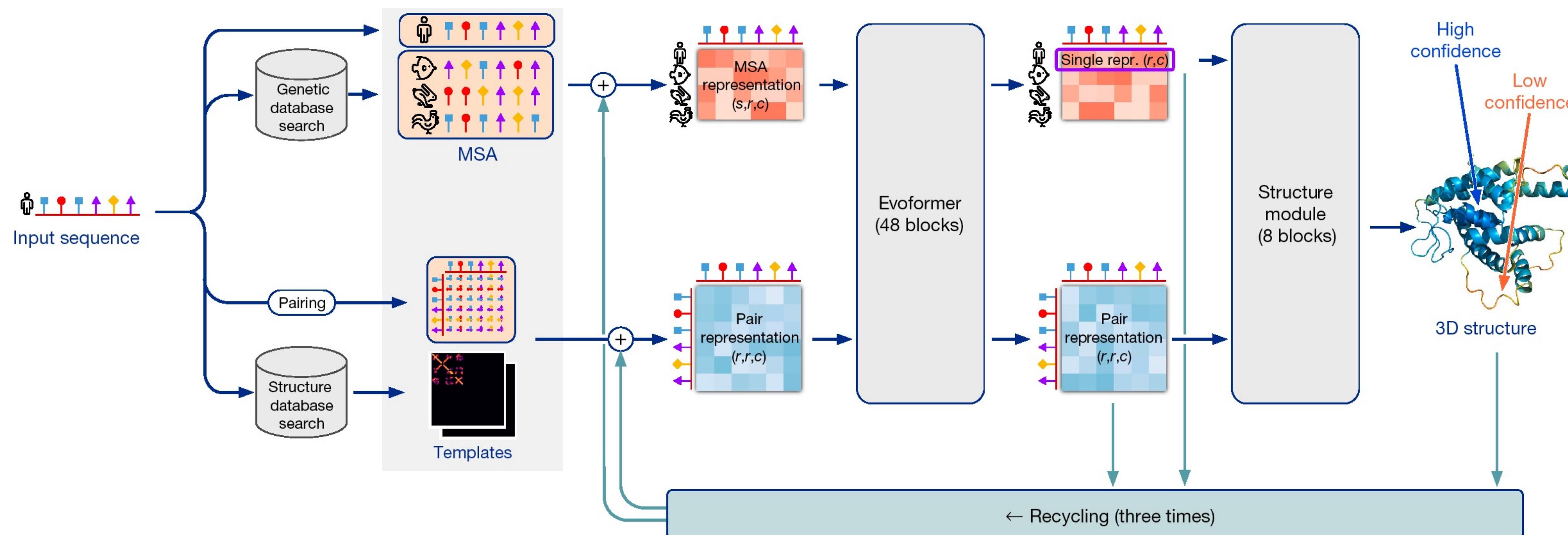
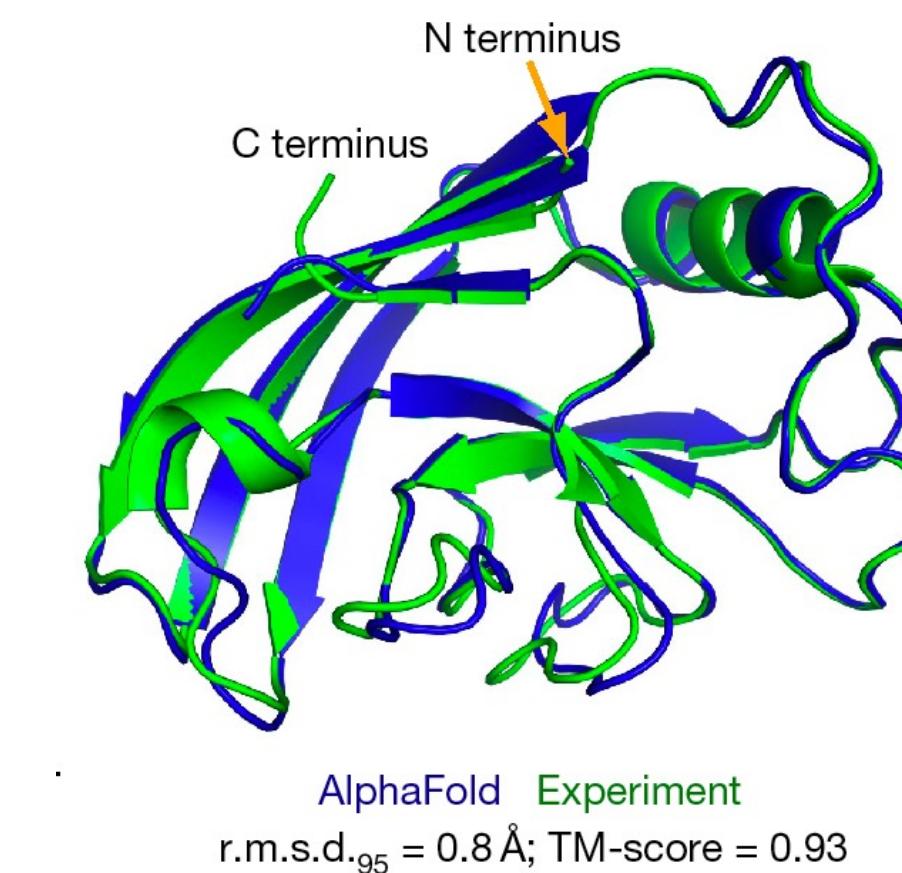
**CASP evaluates current methods of predicting protein structure since 1994.**

“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%)**
  - Google 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**

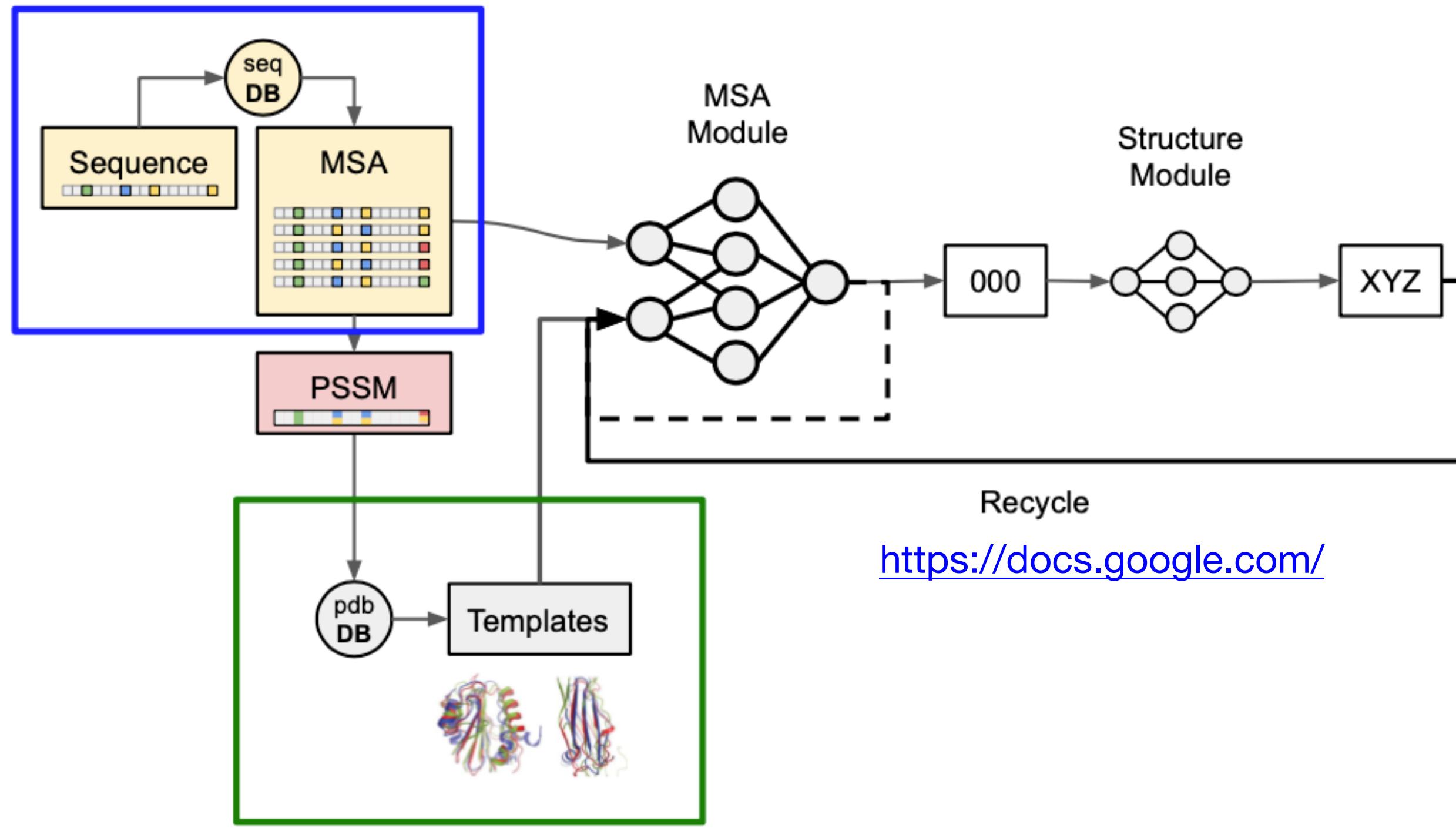
# AlphaFold2: A Neural Network Trained To Predict Protein 3-D Structure

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

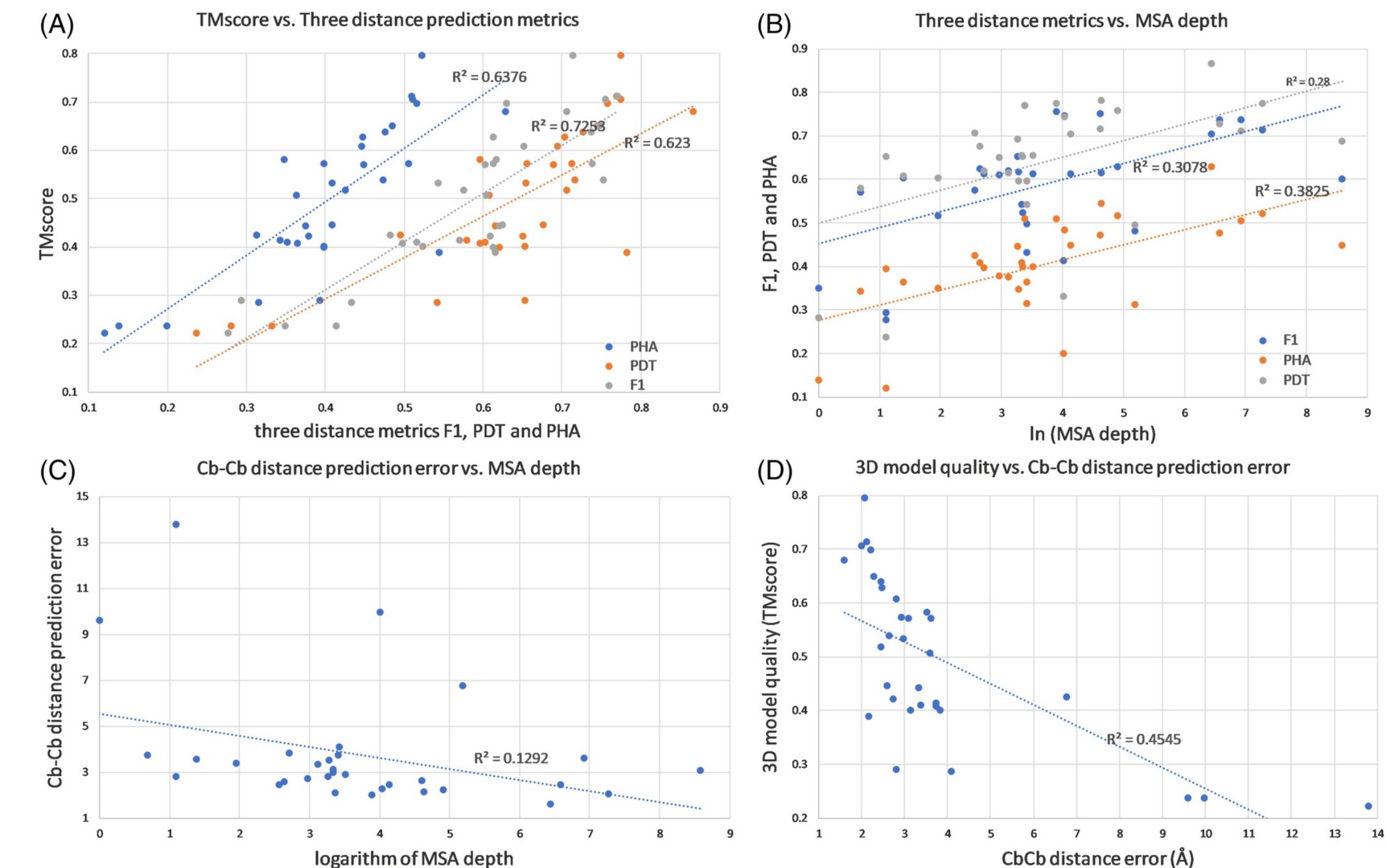


# MSA Depedency of AF2

**Structure Prediction is only as good as the input MSA**



Recycle  
<https://docs.google.com/>

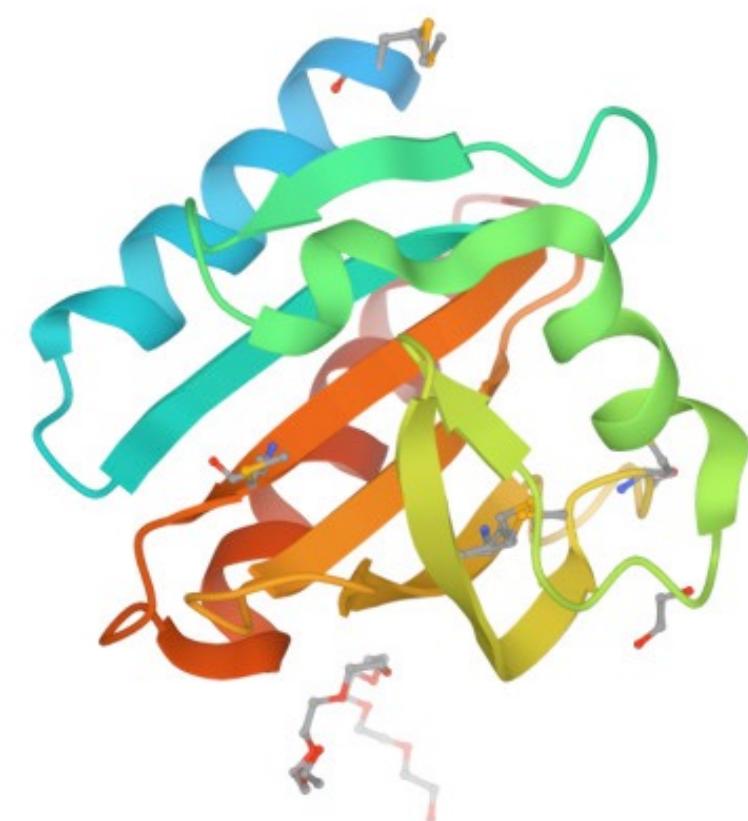
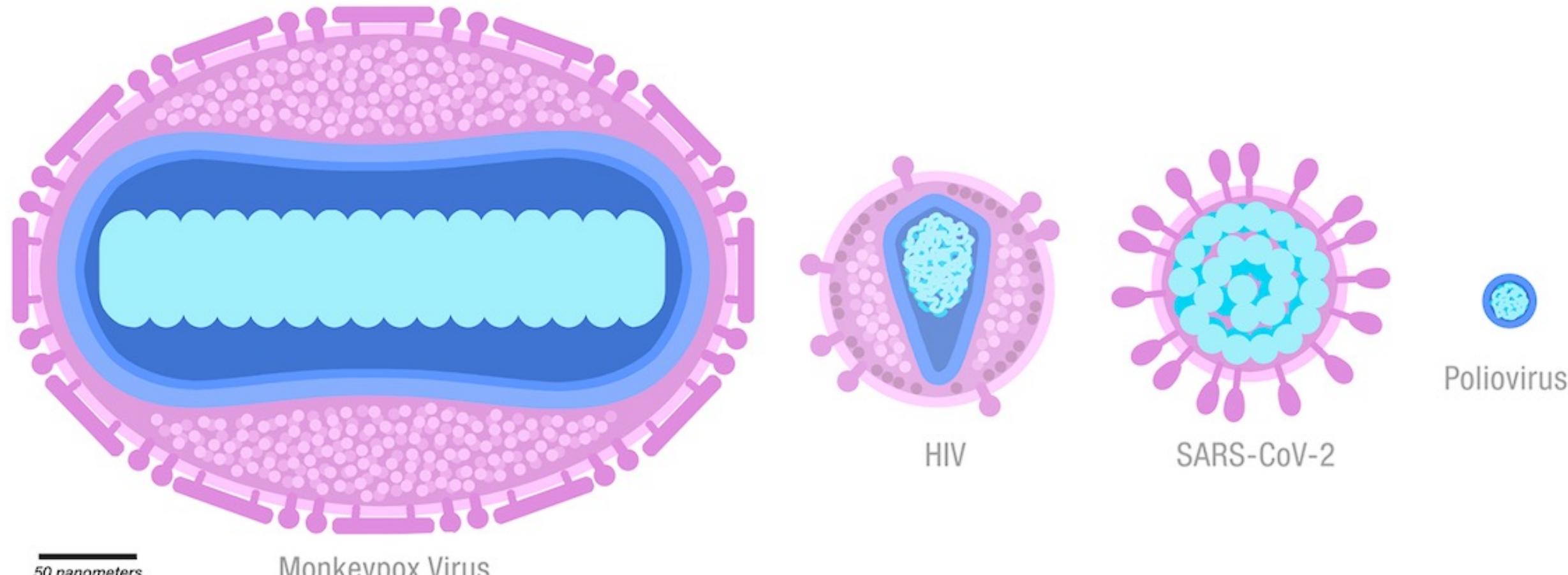


Xu, J. et. al. *Proteins* 2019, 87 (12), 1069-1081.

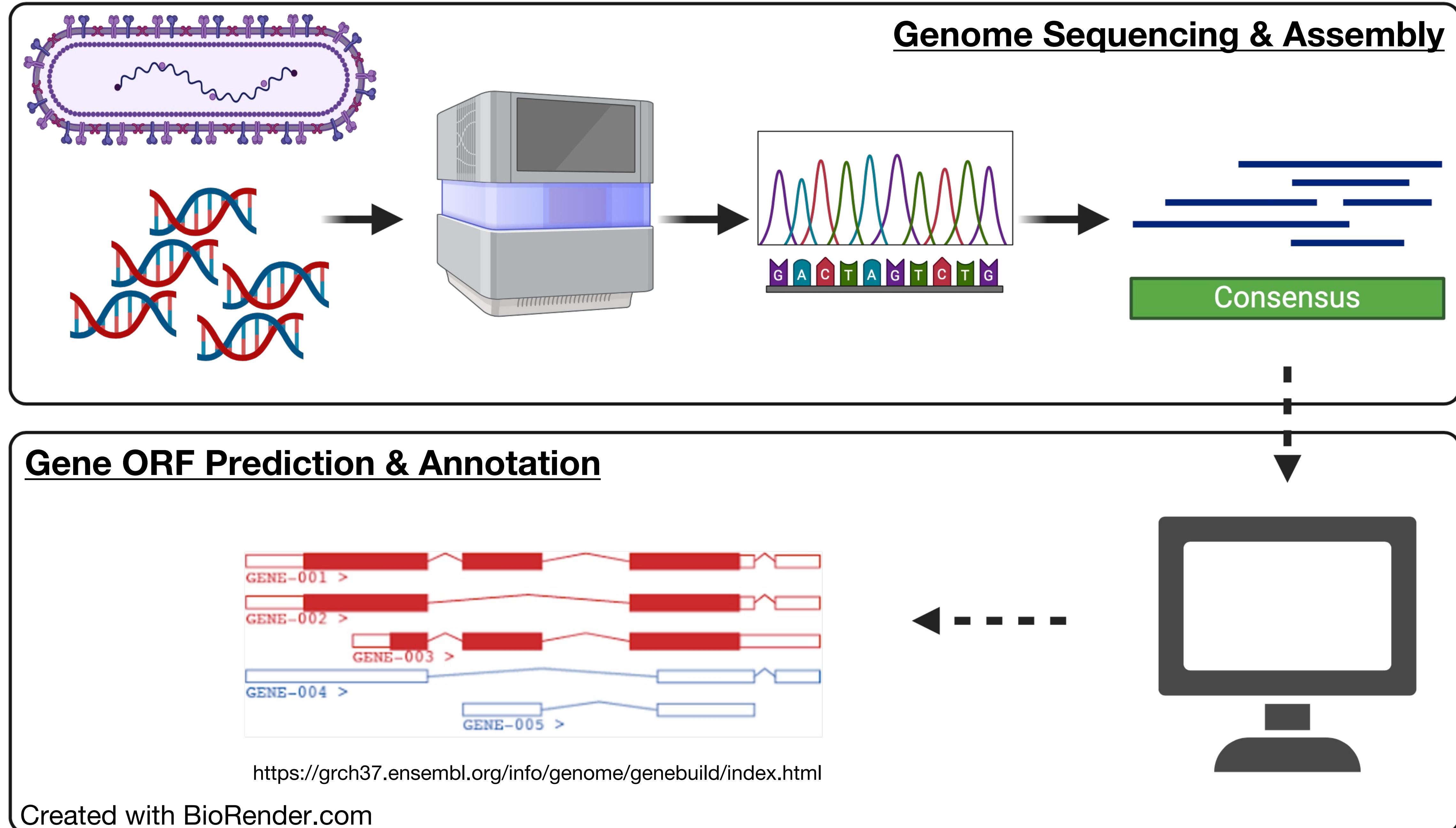
**Typically MSA depth of >100 sequences/residue produce a highly accurate and confident model**

# MPOX (Monkey Pox)

- MPOX first isolated 1958
  - Copenhagen, Denmark
  - Started in children from a playgroup
- Easily diagnosed by lesions that develop in 2-27 days
- Can spread to genitalia
- Can cause:
  - pneumonia, sepsis, and stillbirth (pregnancy)
- Genome ~197kb
  - Contains 190 protein-encoding genes
  - We will predict structures for 60 of these proteins
- To date only two proteins have been elucidated
  - MPXV polymerase holoenzyme in replicating state
  - Profilin-like Protein

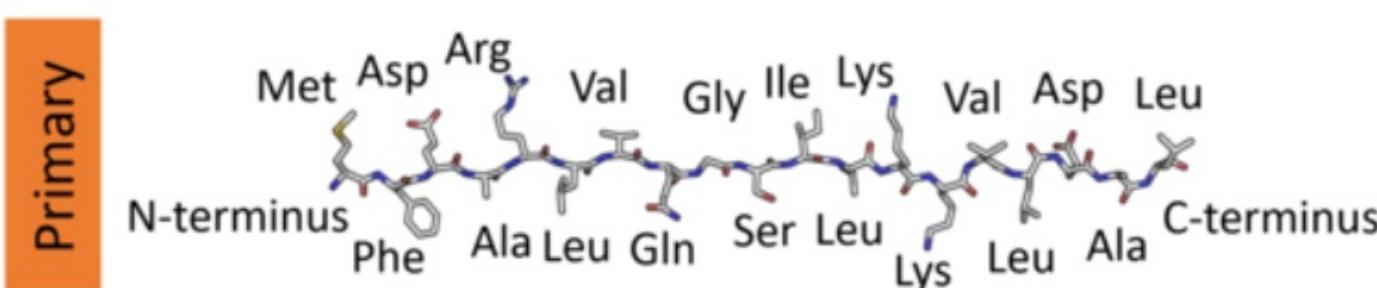


# Assembly of the MPOX-22 Global Outbreak Genome



# Pearson Fasta Format

## *In vivo* “Biological” Representation



[https://en.wikipedia.org/wiki/Protein\\_structure](https://en.wikipedia.org/wiki/Protein_structure)

## *In silico* “Computer” Representation

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken  
MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAEQDMINEVDADGN  
FPEFLTMMARKMKDTDSEEEIREAFRVDKGNGYISAAELRHVMTNLGEKLTDEEVDEM  
DIDGDGQVNYEEFVQMMTAK\*

- They are separated by the ‘>’ sign
  - Followed by a descriptor, known as a ‘header’

How would you represent a genome on a computer?  
What about a transcriptome? or proteome?

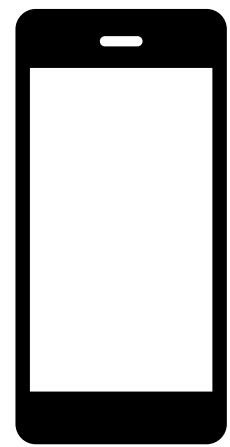
>lcl|ON563414.3\_prot\_URK20440.1\_1  
MKQYIVLACMCLVAAAMPTSLQQSSSCTEEENKHHMGIDVIKVTQDKCQSVTEVTETEDD

- Each sequence is represented as two lines

- The next line contains the amino acid sequence

# Running AlphaFold2 In A Timely Manner

>50 years

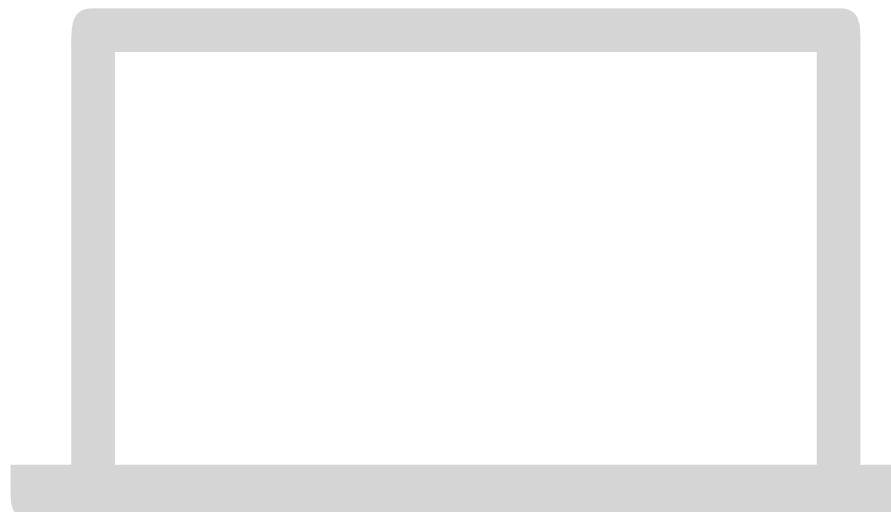


4-8 CPU core

CUSTOM Graphics  
Chip

8-16Gb RAM

Months - Years

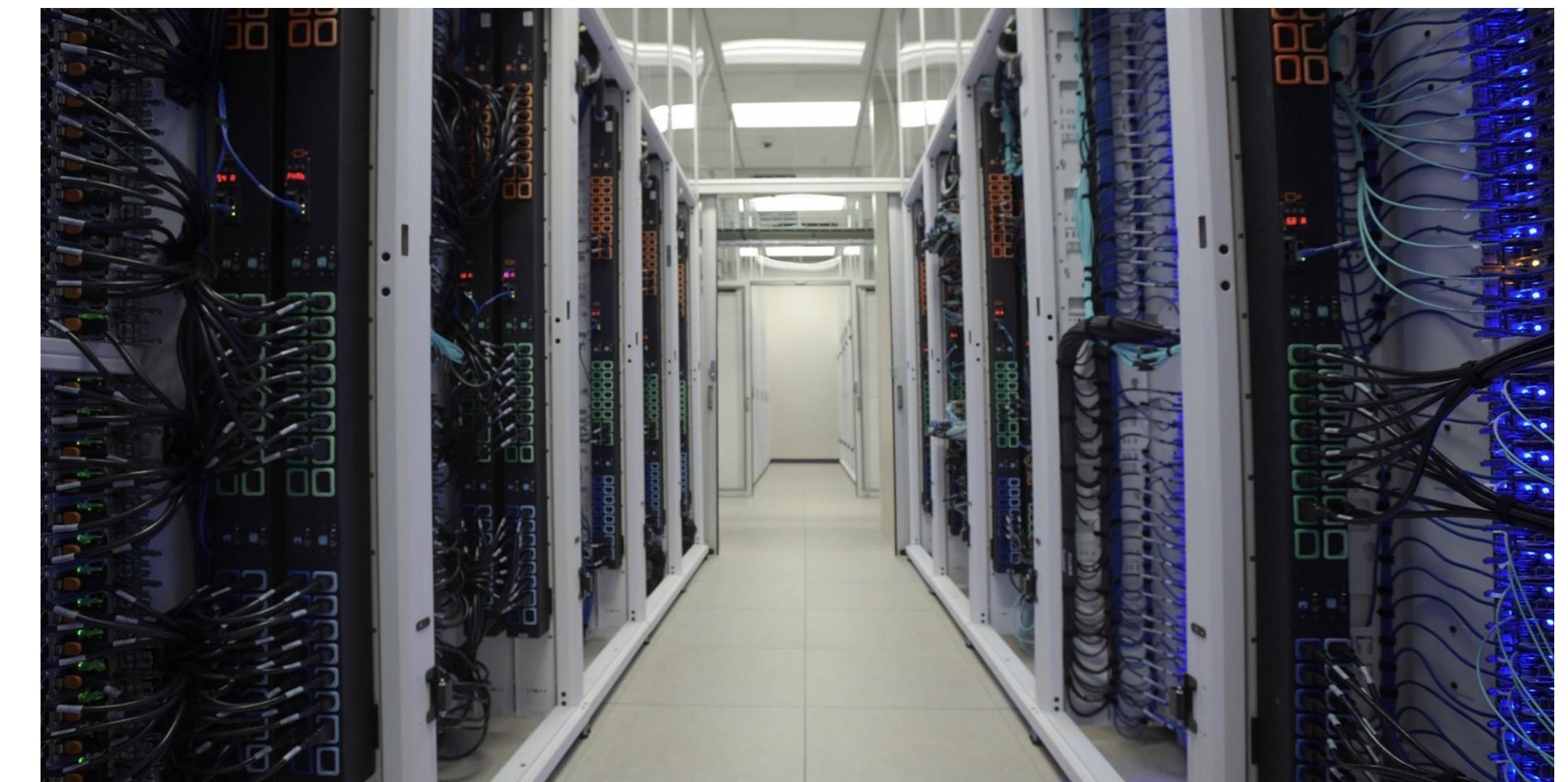


8-10 CPU core

NVIDIA/AMD Entry GPU

16-32Gb RAM

Hours-Days



<https://hprc.tamu.edu/resources/>

925 nodes (40cpu per node)

NVIDIA A100/AMD Ryzen GPUs

>3TB RAM

# Architecture of HPC Clusters

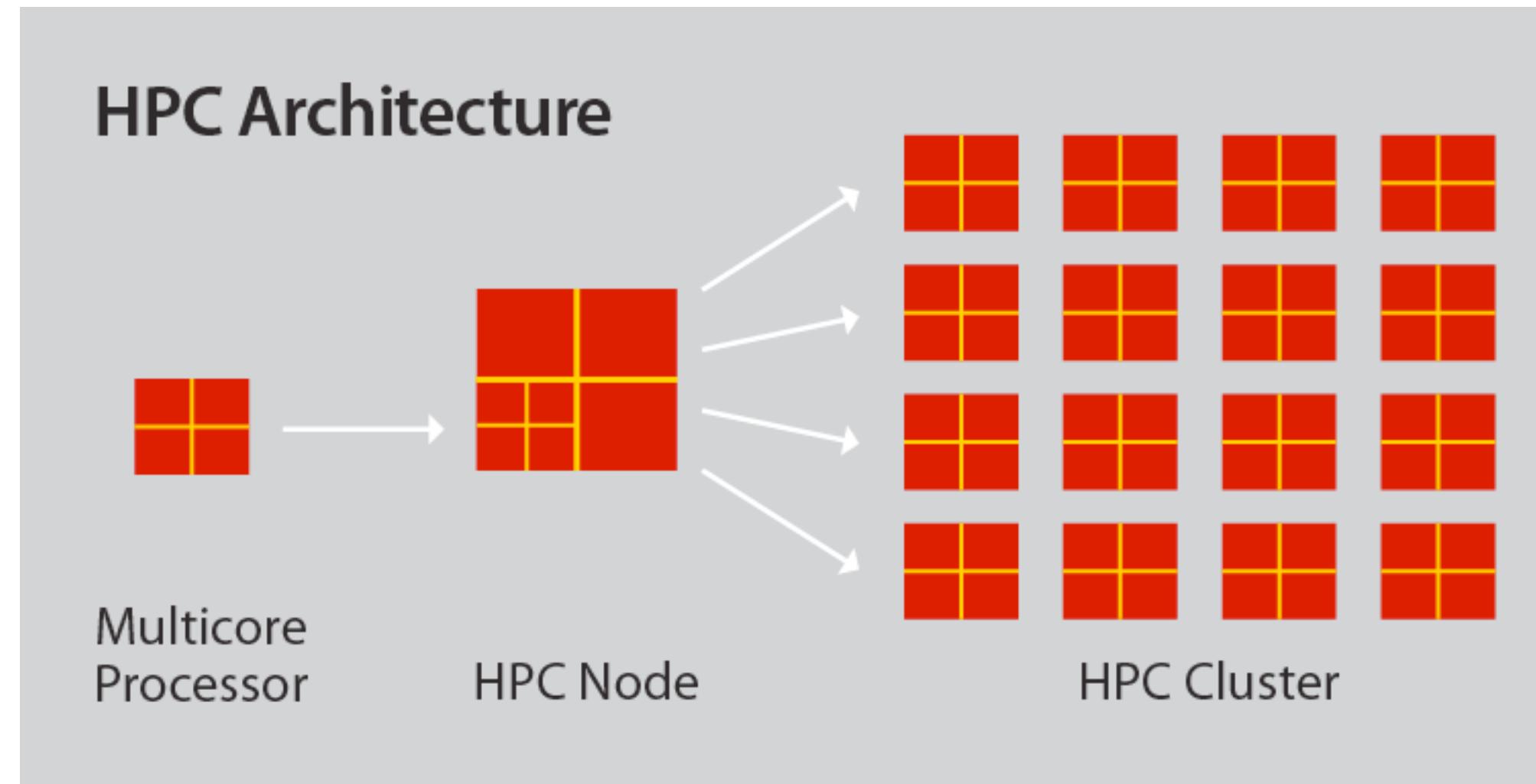
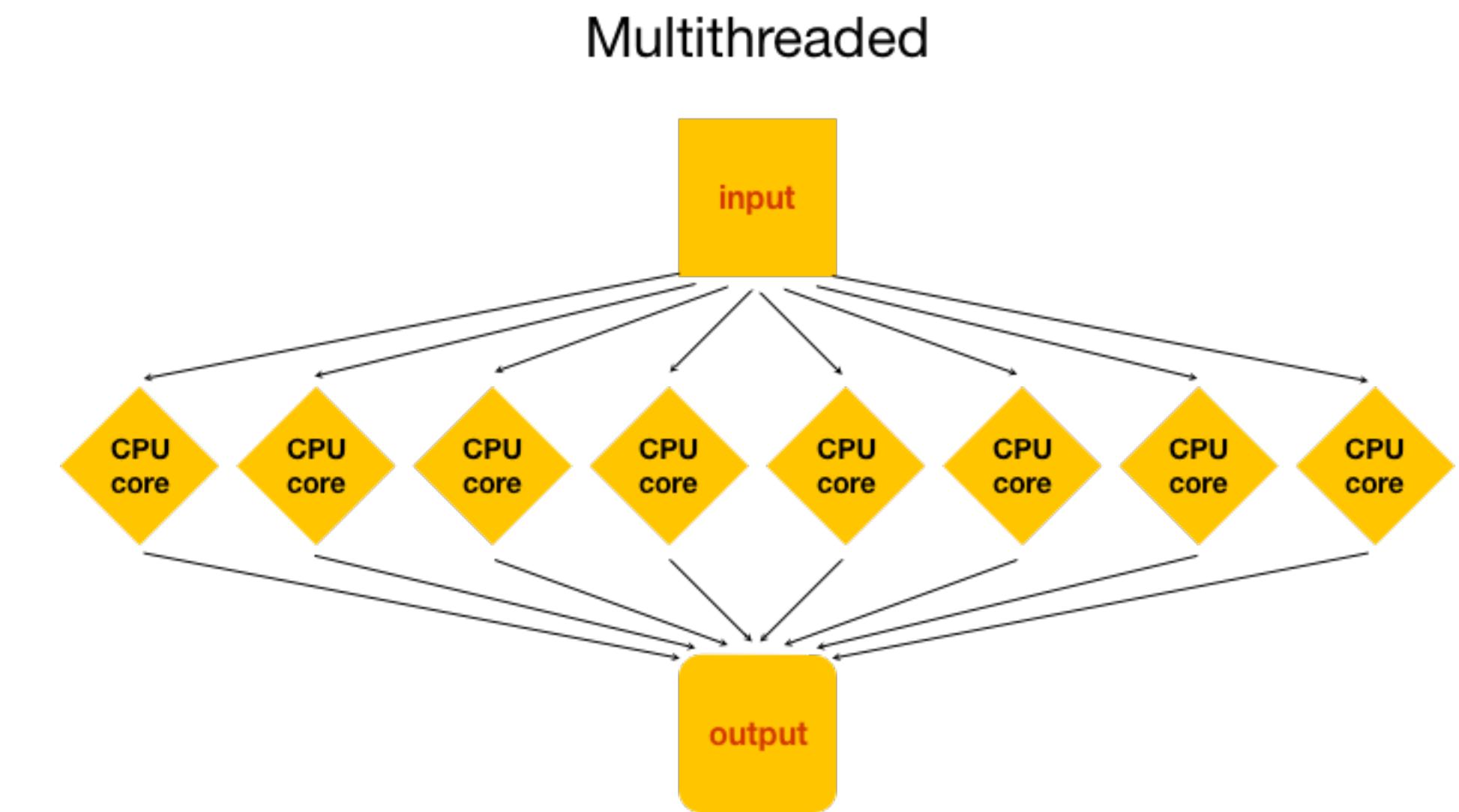
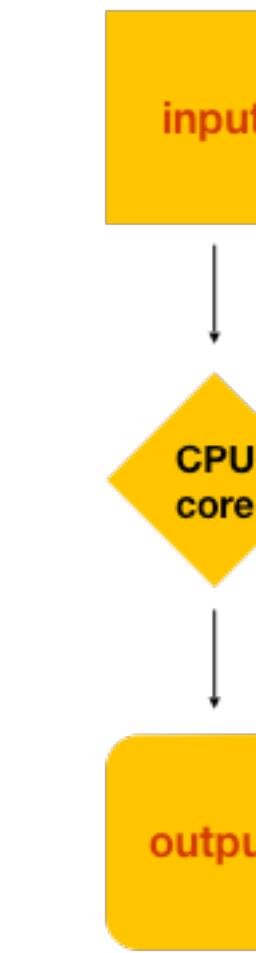
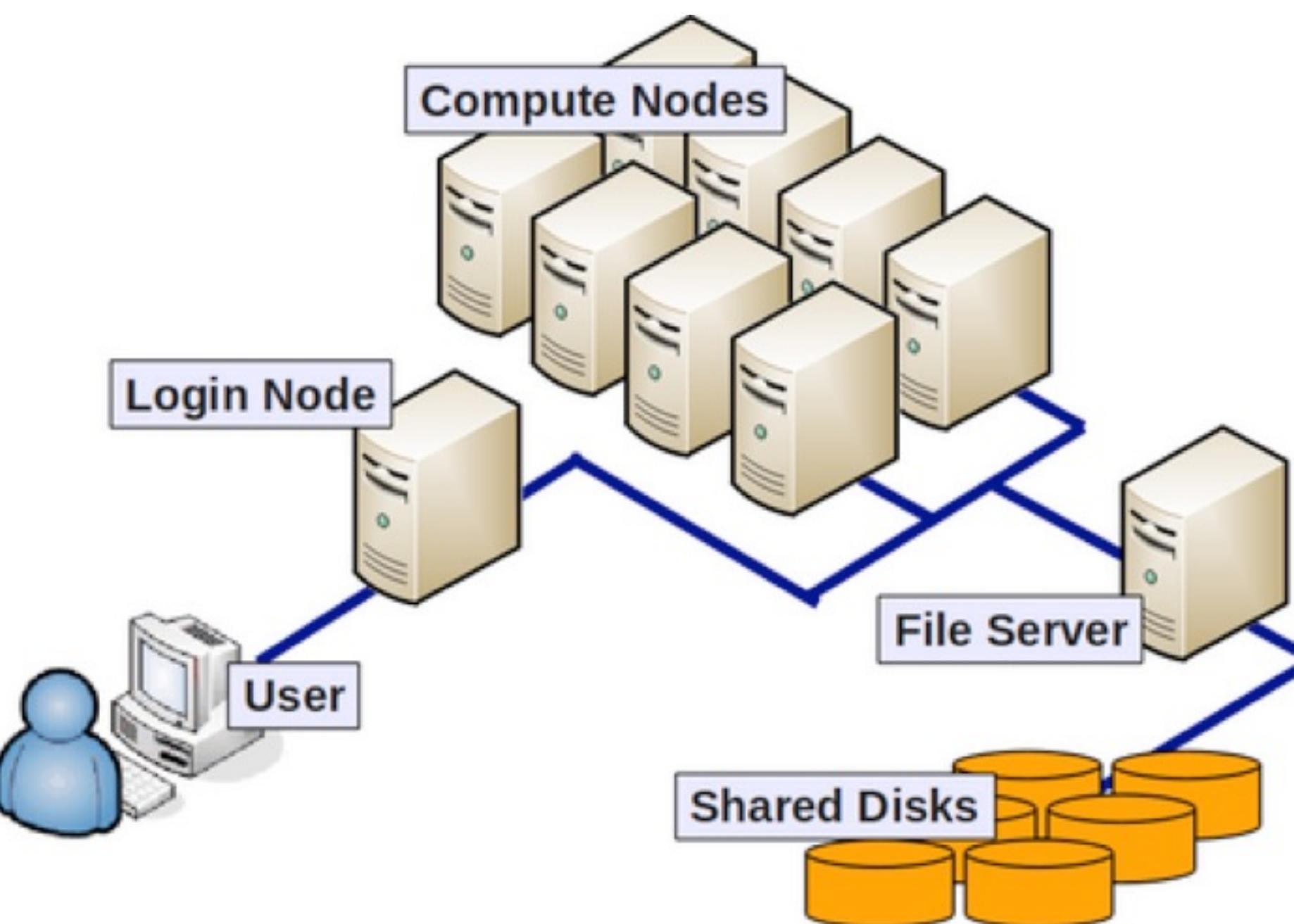


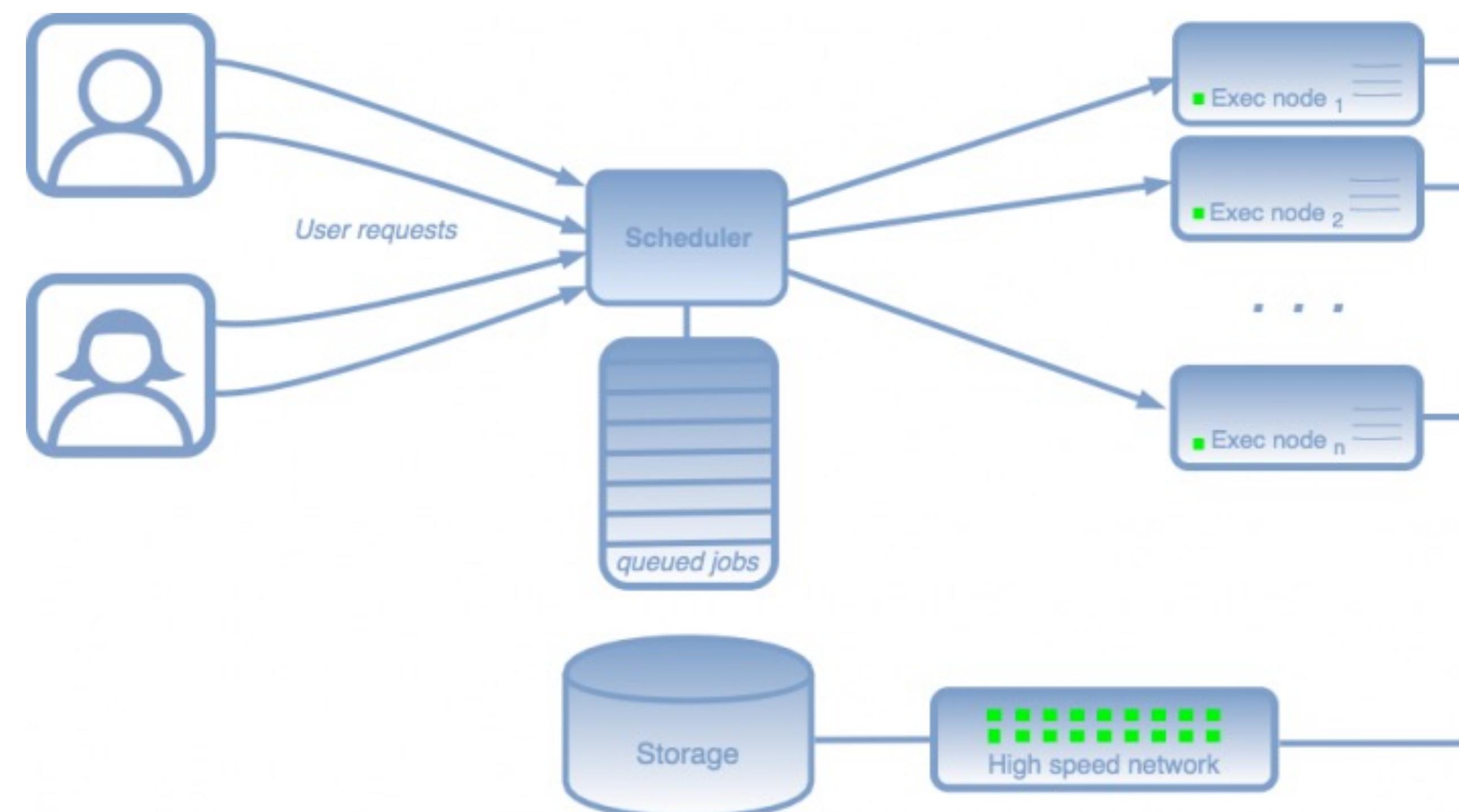
Photo: Illinois State University ([ISU.EDU](http://ISU.EDU))



Photos: GitHub HPCTraining

# Process of Running A Job On Grace

- Unlike to your personal computer, Grace will not do things automatically
- We will build a list of commands which will include running AlphaFold2
  - This is called “scripting/building” a **job file**
- Our job file will contain “directives” and “commands”
  - **Directives** - tell Grace what resources to use (CPU/GPU/Time/RAM)
  - **Commands** - will execute programs or lines of code



# The AlphaFold2 Job Script

```
#!/bin/bash
#SBATCH --job-name=LASTNAME_FIRSTNAME_PROTEIN_HEADER # Edit this line with your name and protein fasta header
#SBATCH --time=1-00:00:00 # Leave this and all other lines starting with #SBATCH alone
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=24
#SBATCH --mem=180G
#SBATCH --output=stdout.%j.txt
#SBATCH --error=stderr.%j.txt
```

**Directives** - tells the computer what hardware to use

**Modules** - tells the computer what software to use

```
##### Load Program Modules #####
# This will load the modules (programs) needed to run AF2 and analyze the output
module purge
module load GCC/10.2.0 CUDA/11.1.1 OpenMPI/4.0.5 AlphaPickle/1.4.1
```

**Variables** - our input for the program

```
##### VARIABLES #####
#####
# INPUTS #####
# Edit this variable to point towards your fasta file
protein_fasta='/Path/to/your/fasta_file.fasta'
```

**Parameters** - “arguments” the program uses, typically specific to your data

```
##### PARAMETERS #####
# Leave these variables alone
DOWNLOAD_DIR='/scratch/data/bio/alphafold/2.3.0'
max_template_date='2023-1-1'
model_preset='monomer' # monomer, monomer_casp14, monomer_ptm, multimer
db_preset='full_dbs' # full_dbs, reduced_dbs
```

**Variables** - specific to the program you are using

```
##### OUTPUTS #####
protein_basename=$(basename ${protein_fasta%.*})
output_dir="out_${protein_basename}_${model_preset}"
pickle_out_dir=$protein_basename
```

**Commands** - specific to the program, and tell the computer what to do

```
##### COMMANDS #####
singularity exec --nv /sw/hpcr/sw/bio/containers/alphafold/alphafold_2.3.0.sif python /app/alphafold/run_alphafold.py \
--data_dir=$DOWNLOAD_DIR \
--uniref90_database_path=$DOWNLOAD_DIR/uniref90/uniref90.fasta \
--mgnify_database_path=$DOWNLOAD_DIR/mgnify/mgy_clusters_2022_05.fa \
--bfd_database_path=$DOWNLOAD_DIR/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
--uniref30_database_path=$DOWNLOAD_DIR/uniref30/UniRef30_2021_03 \
--pdb_seqres_database_path=$DOWNLOAD_DIR/pdb_seqres/pdb_seqres.txt \
--template_mmcif_dir=$DOWNLOAD_DIR/pdb_mmcif/mmcif_files \
--obsolete_pdbs_path=$DOWNLOAD_DIR/pdb_mmcif/obsolete.dat \
--uniprot_database_path=$DOWNLOAD_DIR/uniprot/uniprot.fasta \
--model_preset=$model_preset \
--max_template_date=$max_template_date \
--db_preset=$db_preset \
--output_dir=$output_dir \
--fasta_paths=$protein_fasta
# graph pLDDT and PAE .pkl files
# This line command will run a python script to plot our pLDDT and pAE plots
# if you would like more information on how this works please email Devon
run_AlphaPickle.py -od $output_dir/$pickle_out_dir
```

# 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 and pAE
- When in doubt, run AF2 on your protein, worst that happens is you waste your time

## Next Time

We will analyze the predicted structures