SCIENCE MEETS LIFE

AlphaFold: introduction + how-to

Jasper Zuallaert

Internship Major Systems Biology

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

VIB-UGENT CENTER FOR INFLAMMATION RESEARCH

VIB BIOINFORMATICS CORE

fwo

UNIVERSITEIT GENT

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

SCIENCE MEETS LIFE

The Protein folding problem

VIB

Available sequences and structures

After filtering for redundancy

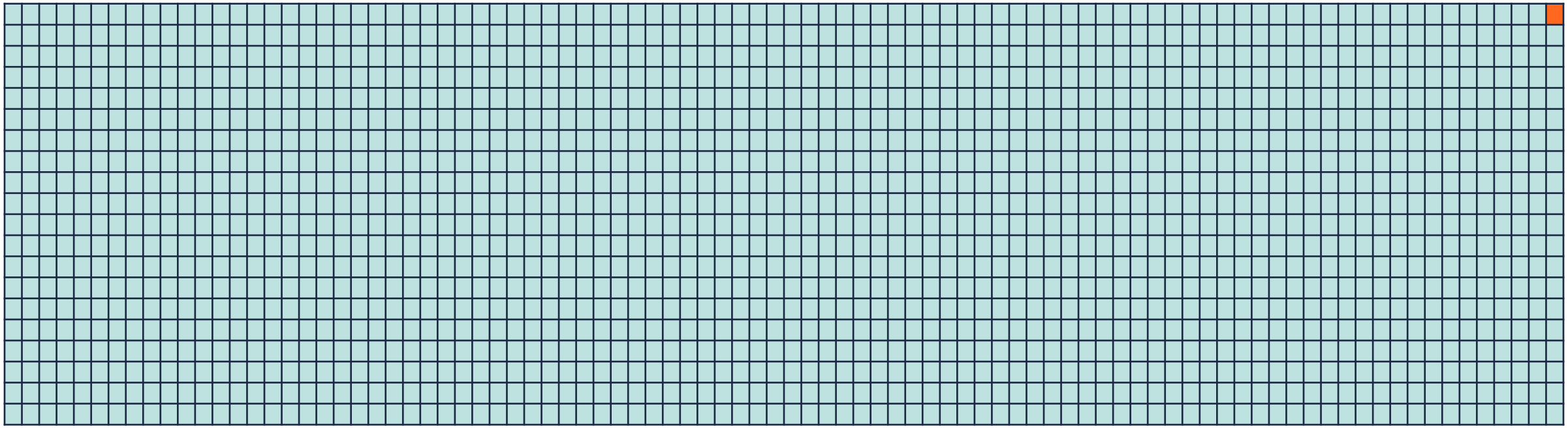
(50% similarity) --- (May 2023)

UniRef50

60 million unique sequences

Protein Data Bank (PDB)

~60 000 unique structures



□ = 60 000 proteins

Main protein structure resource

RCSB PDB
PROTEIN DATA BANK

~60 000 unique structures

Main experimental technique: X-ray crystallography

- Costly
- Labour-intensive
- Slow

(NMR)

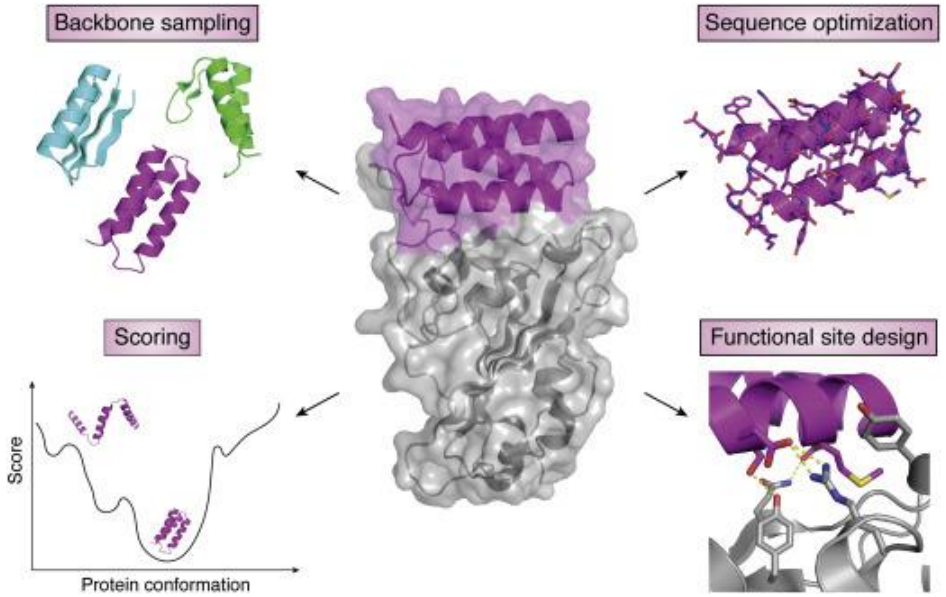
Image source: https://www.prweb.com/releases/2017/04/prweb14219956.htm

Why 3-D structure prediction? Low effort, time efficient

Examples:

- In silico drug candidate screening

- De novo protein design guidance



Source: https://www.jbc.org/article/S0021-9258(21)00336-7/fulltext

- +++

Structure prediction: homology models

Basic principle: 3-D structure more conserved than sequence

Query protein sequence: ELAIGILTVSYIPSAEKIRAPELTI

Sequence alignment:

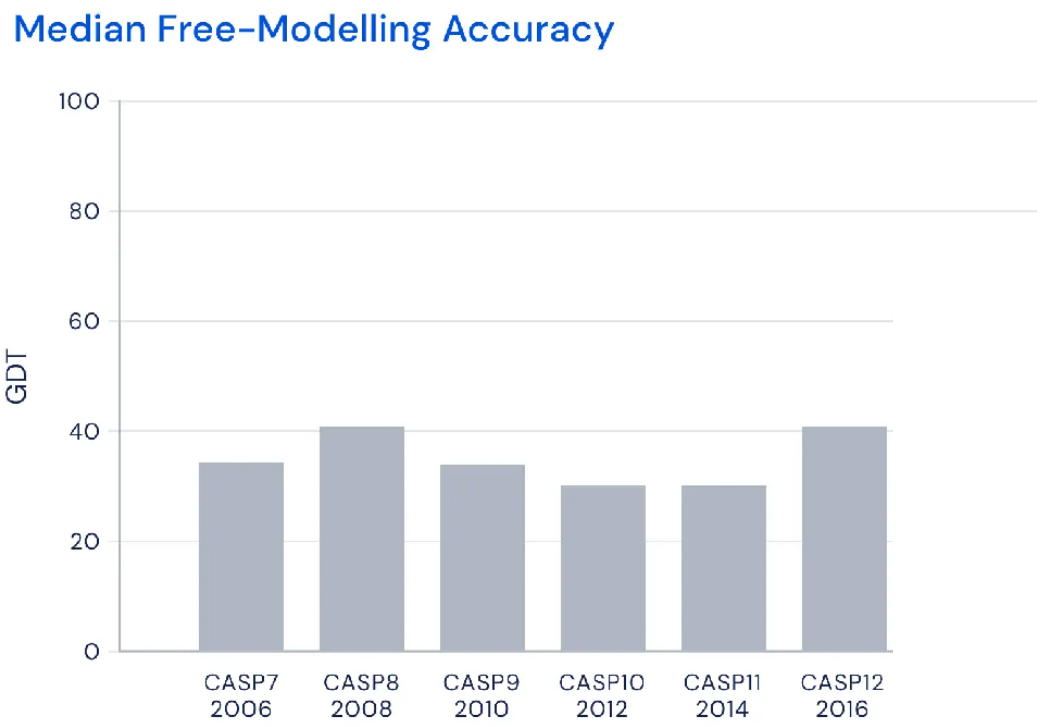
```
ELA-IGILTVSYIPSAEKIRAP--ELTI
ELAGI-ILGVSYIPSAEKI-ARACELTI
```

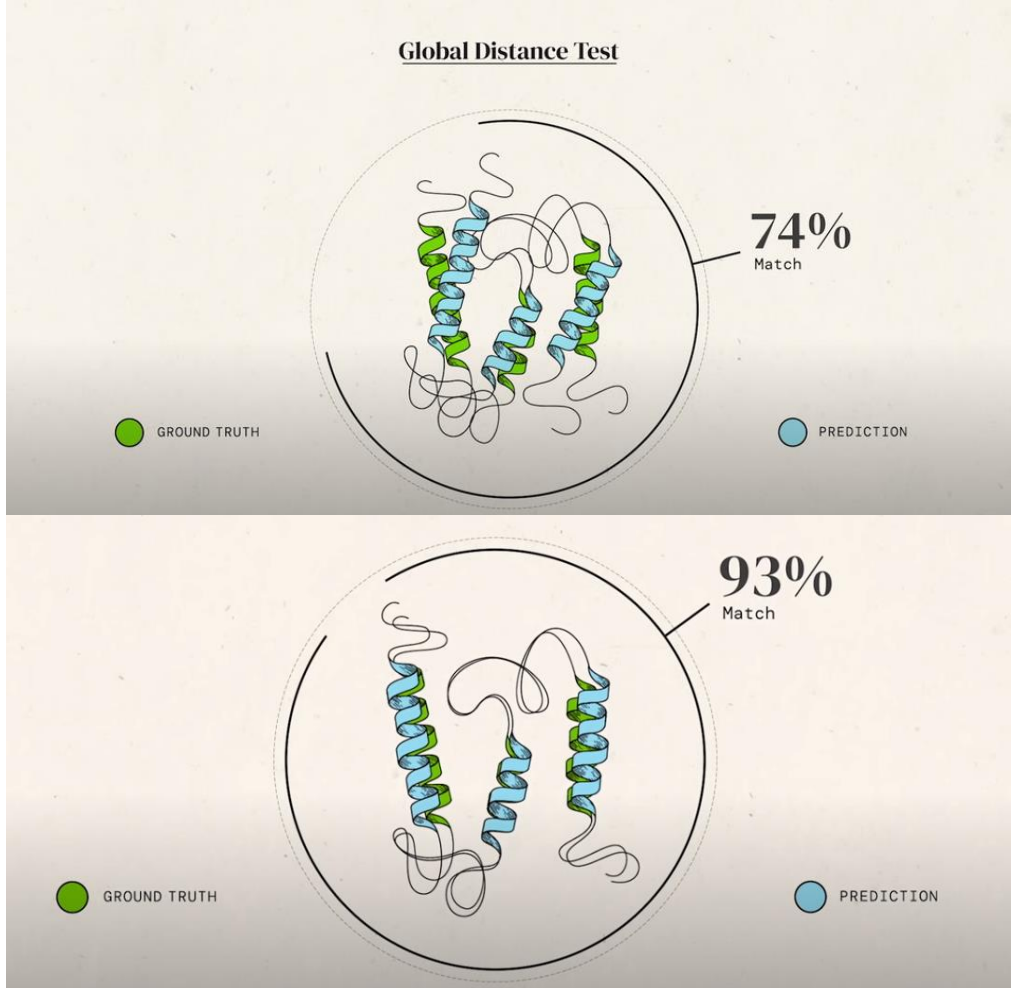
→ **structure in PDB?**

Finetuning using statistical potentials and physics-based energy calculations

Example from https://unil.ch/pmf/en/home/menuinst/technologies/homology-modeling.html

Critical Assessment of Structure Prediction





Source: https://news.machinelearning.sg/posts/alphafold2_10_things_you_want_to_know_about_biologys_imagenet_moment/

SCIENCE MEETS LIFE

AlphaFold introduction

VIB

DeepMind: a timeline

2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020

← Founded in 2010

Acquired by Google for $500M

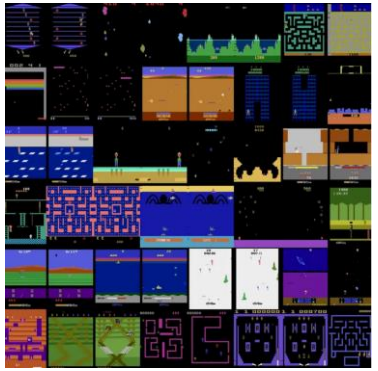
Google DeepMind

AlphaGo beats Lee Sedol (4-1)

AlphaZero beats AlphaGo (100-0) and self-teaches chess/shogi

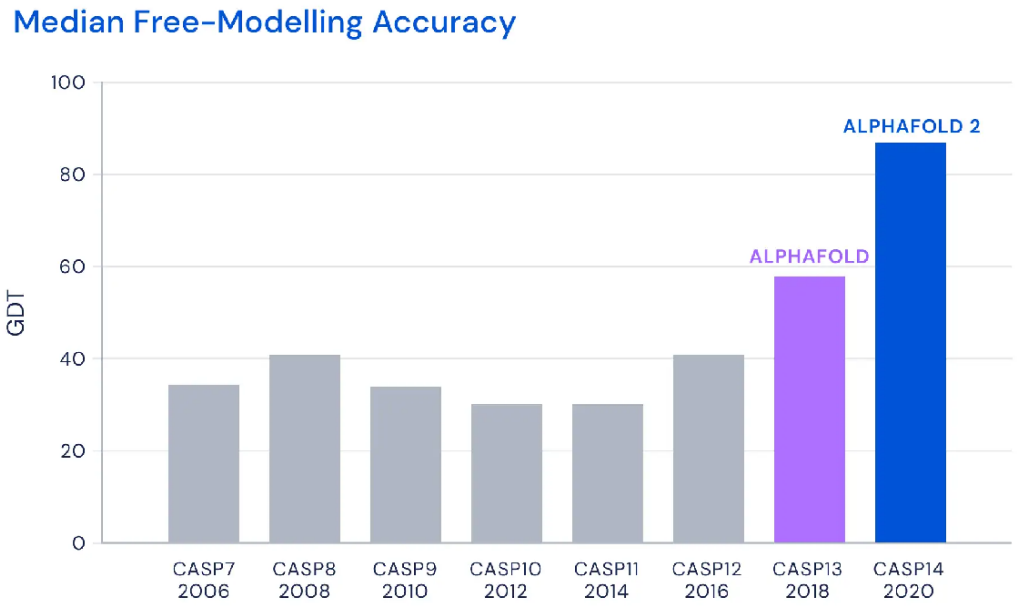
AlphaFold (1.0) wins CASP13

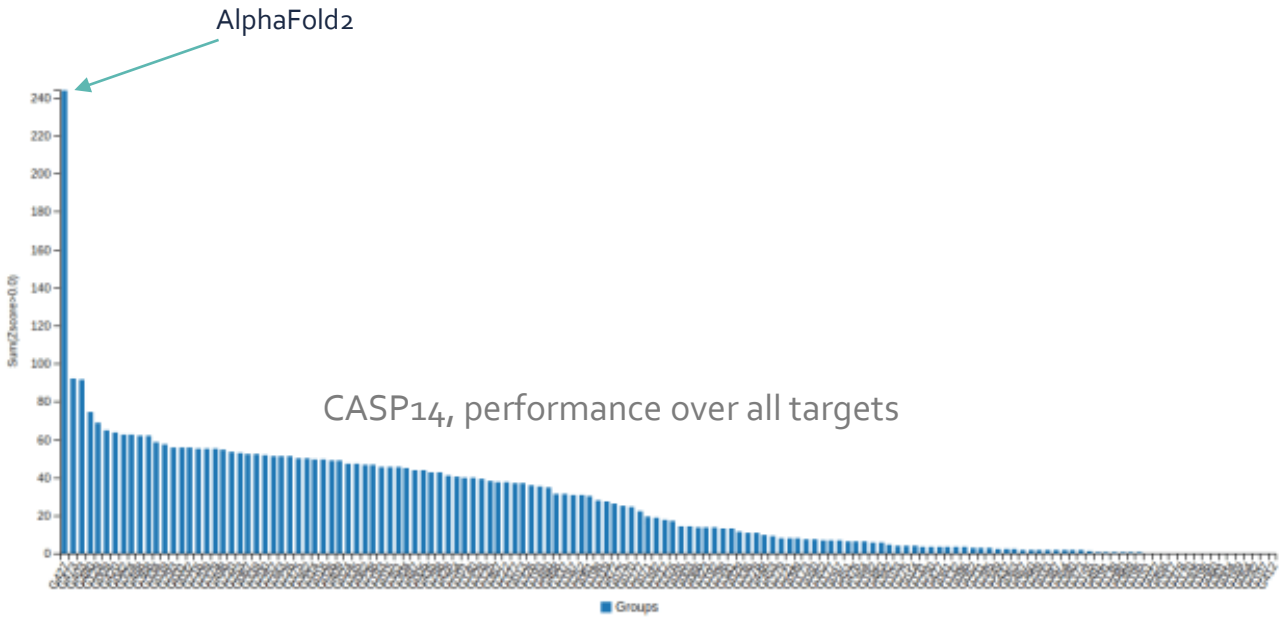
Agent57 surpasses human level performance on all Atari2600 suite games



AlphaFold (2.0) wins CASP14

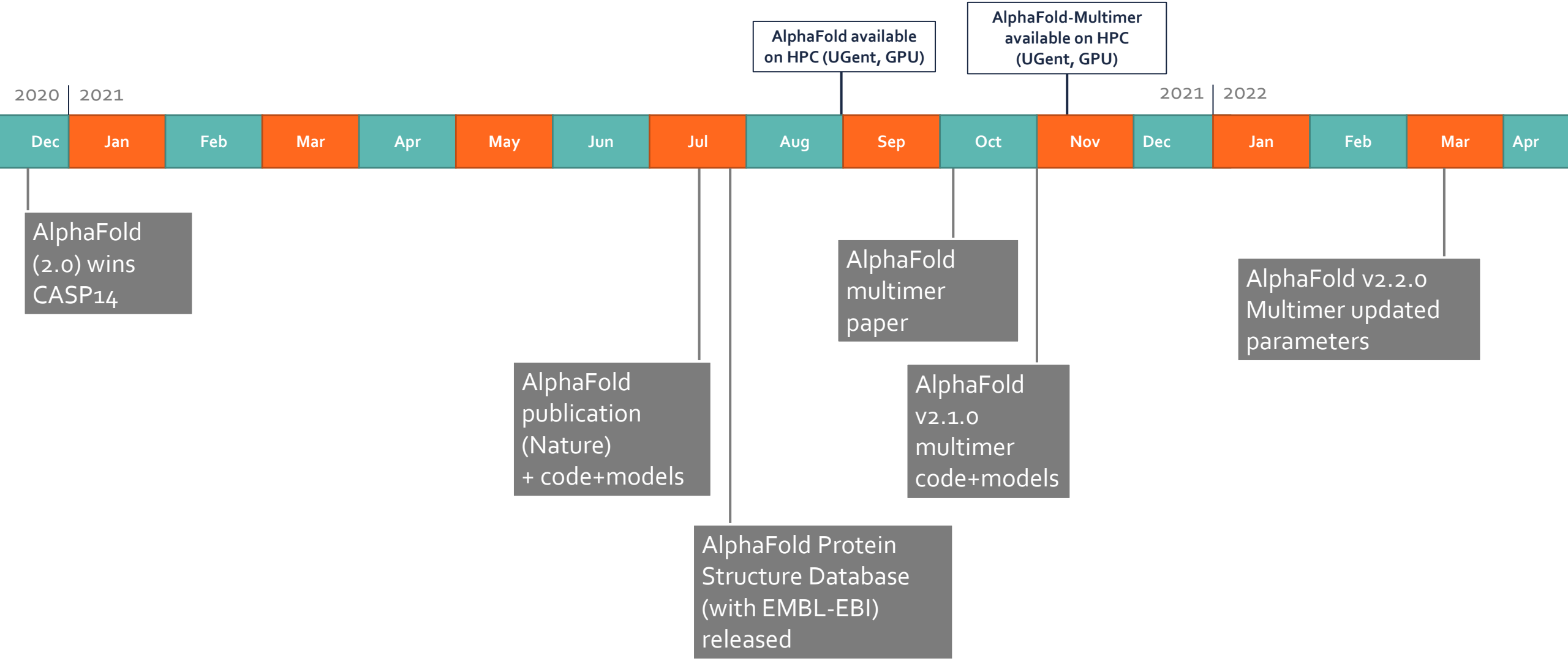
Critical Assessment of Structure Prediction (CASP14, 2020)



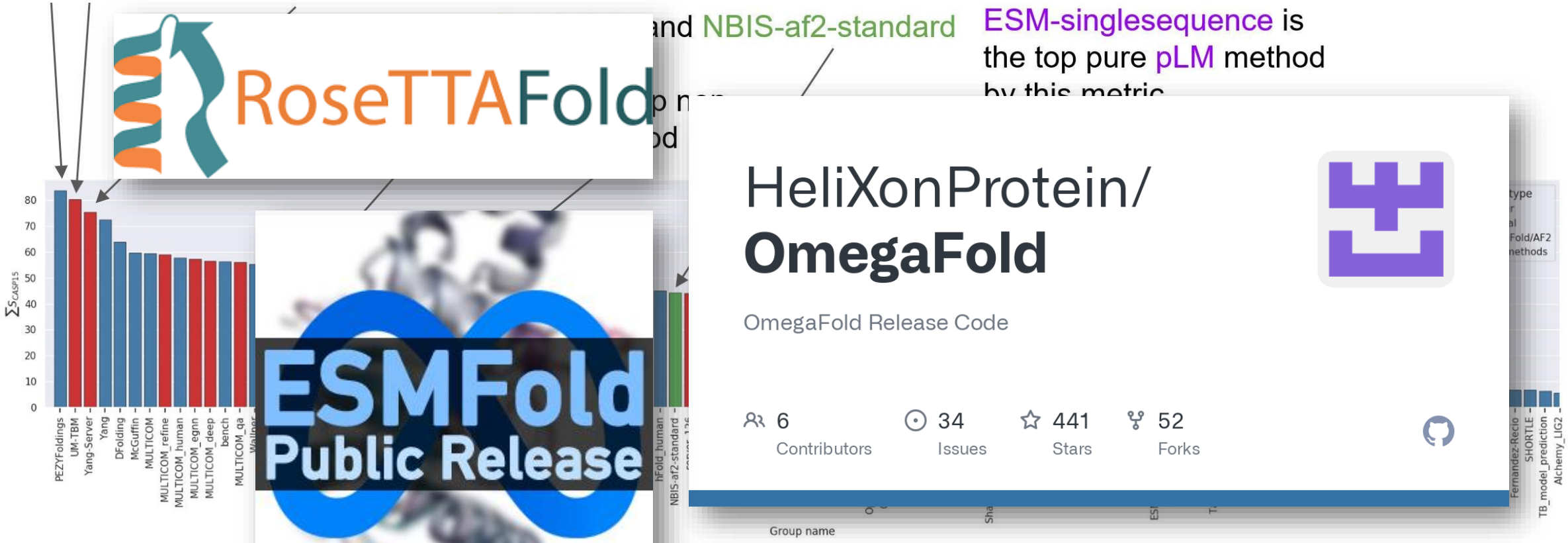


Median of 92.4 across all targets,
87.0 in free-modelling accuracy

AlphaFold timeline



CASP15: multi-run AlphaFold-centered approaches dominate



CASP15: multi-run AlphaFold-centered approaches dominate

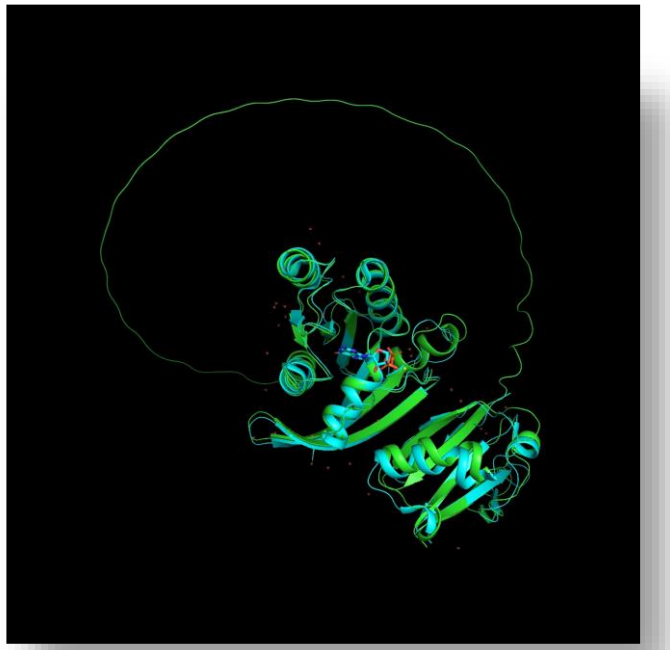
Winner of multimer competition

→ Afsample = 6000 predictions per target (+ select best)

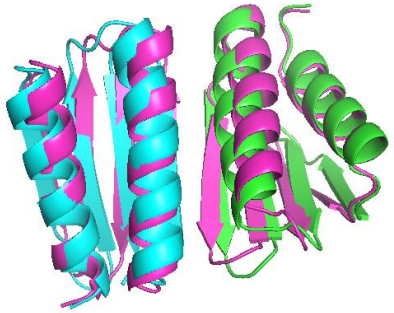
<-> default AlphaFold: 5 (v1) / 25 (v2) per target

→ absence of evidence != evidence of absence

AlphaFold-Multimer

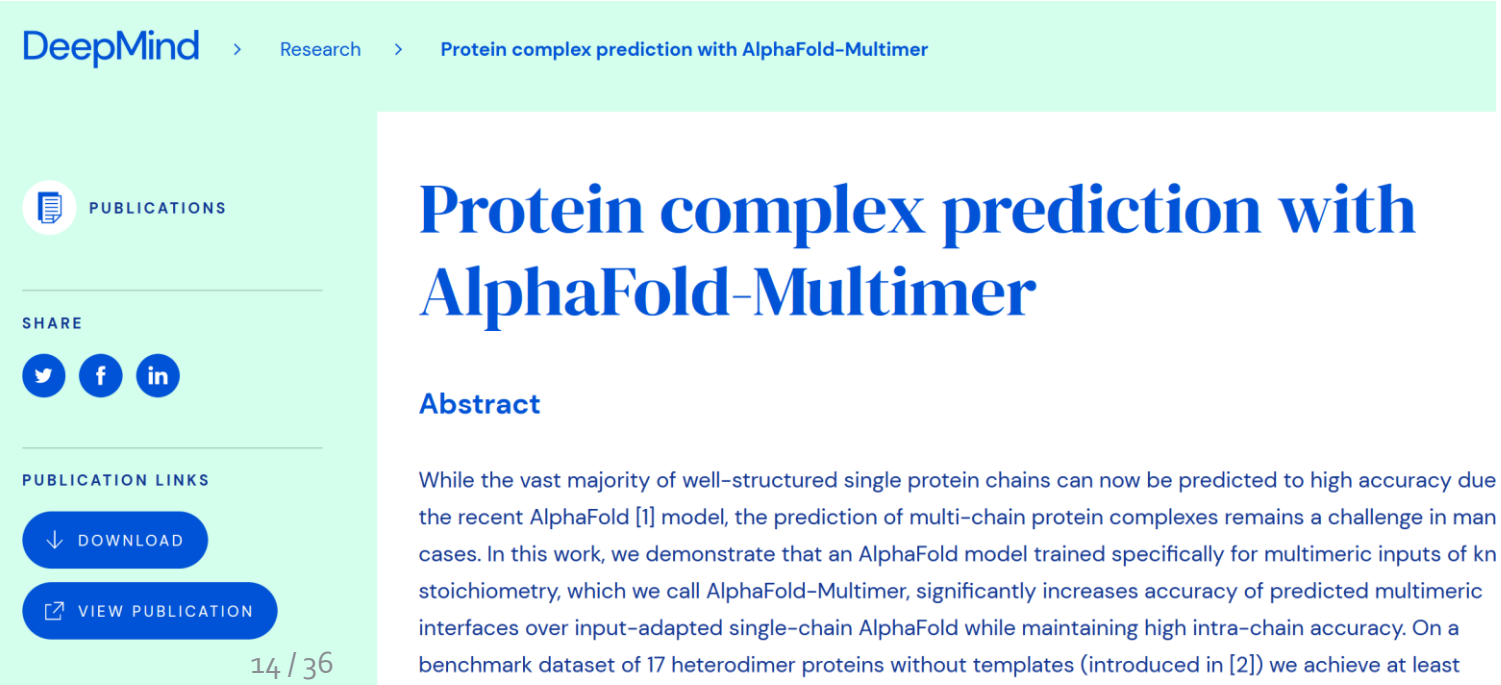


1. Single chain
with pseudolinker



2. Separate chains
programmatically

3. AlphaFold-Multimer



DeepMind > Research > Protein complex prediction with AlphaFold-Multimer

PUBLICATIONS

SHARE

PUBLICATION LINKS

DOWNLOAD

VIEW PUBLICATION

Protein complex prediction with AlphaFold-Multimer

Abstract

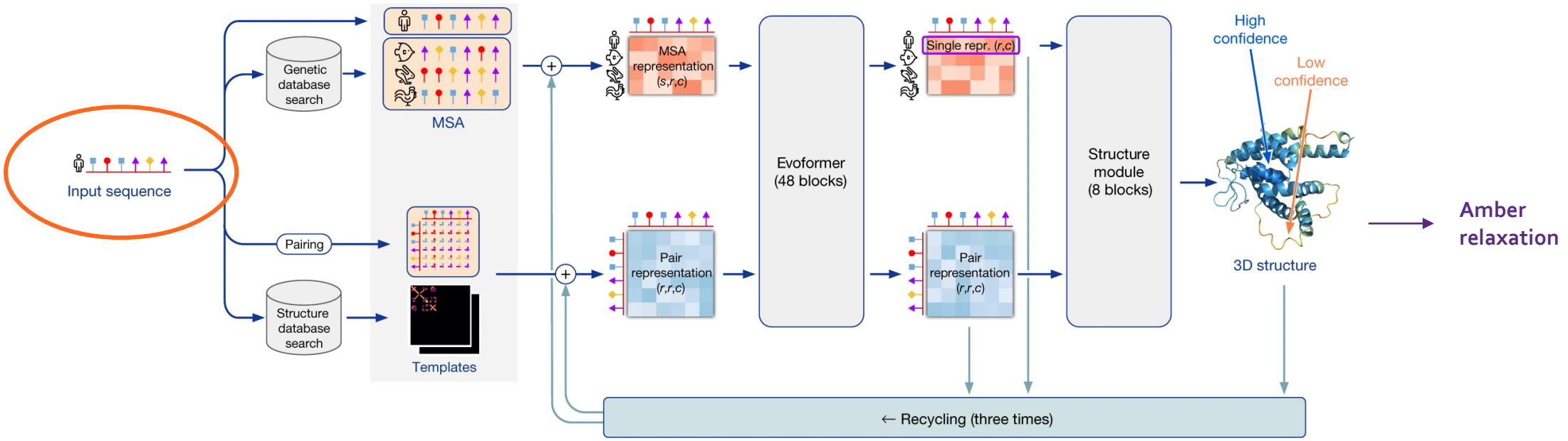
While the vast majority of well-structured single protein chains can now be predicted to high accuracy due the recent AlphaFold [1] model, the prediction of multi-chain protein complexes remains a challenge in man cases. In this work, we demonstrate that an AlphaFold model trained specifically for multimeric inputs of kn stoichiometry, which we call AlphaFold-Multimer, significantly increases accuracy of predicted multimeric interfaces over input-adapted single-chain AlphaFold while maintaining high intra-chain accuracy. On a benchmark dataset of 17 heterodimer proteins without templates (introduced in [2]) we achieve at least

SCIENCE MEETS LIFE

The AlphaFold pipeline

VIB

Stages of prediction

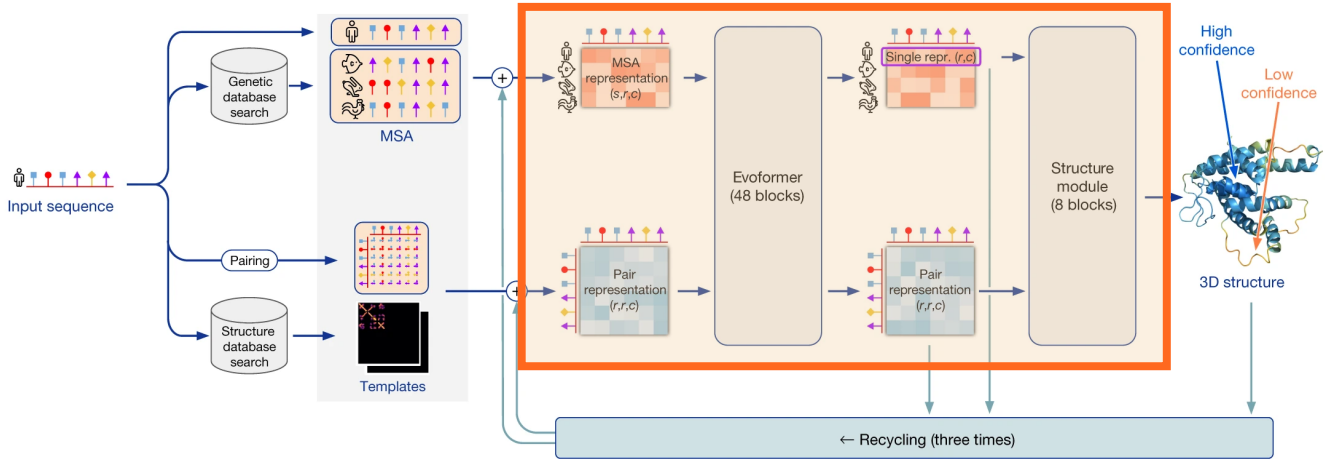


Multiple sequence alignment (CPU)
Structural templates search

Run inputs through 5 trained AlphaFold models (GPU)
→ 5 structures predicted
→ sorted on confidence

Relax all 5 structures (Amber)

AlphaFold training



5 prediction models

Trained **on**

- **PDB** protein fragments (256-384 residues)

Trained **for**

- Frame-aligned point error (FAPE)

Minimize distance between predicted structure and ground truth

+ auxiliary losses (predict confidence by estimating when mistakes are made, physical constraints, …)

SCIENCE MEETS LIFE

Access to AlphaFold predictions

- Memory (RAM)
- GPUs
- Huge databases (~2.2 TB)

VIB

AlphaFold availability

1.

Google colab

Colabfold

2.

AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

3.

GitHub

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

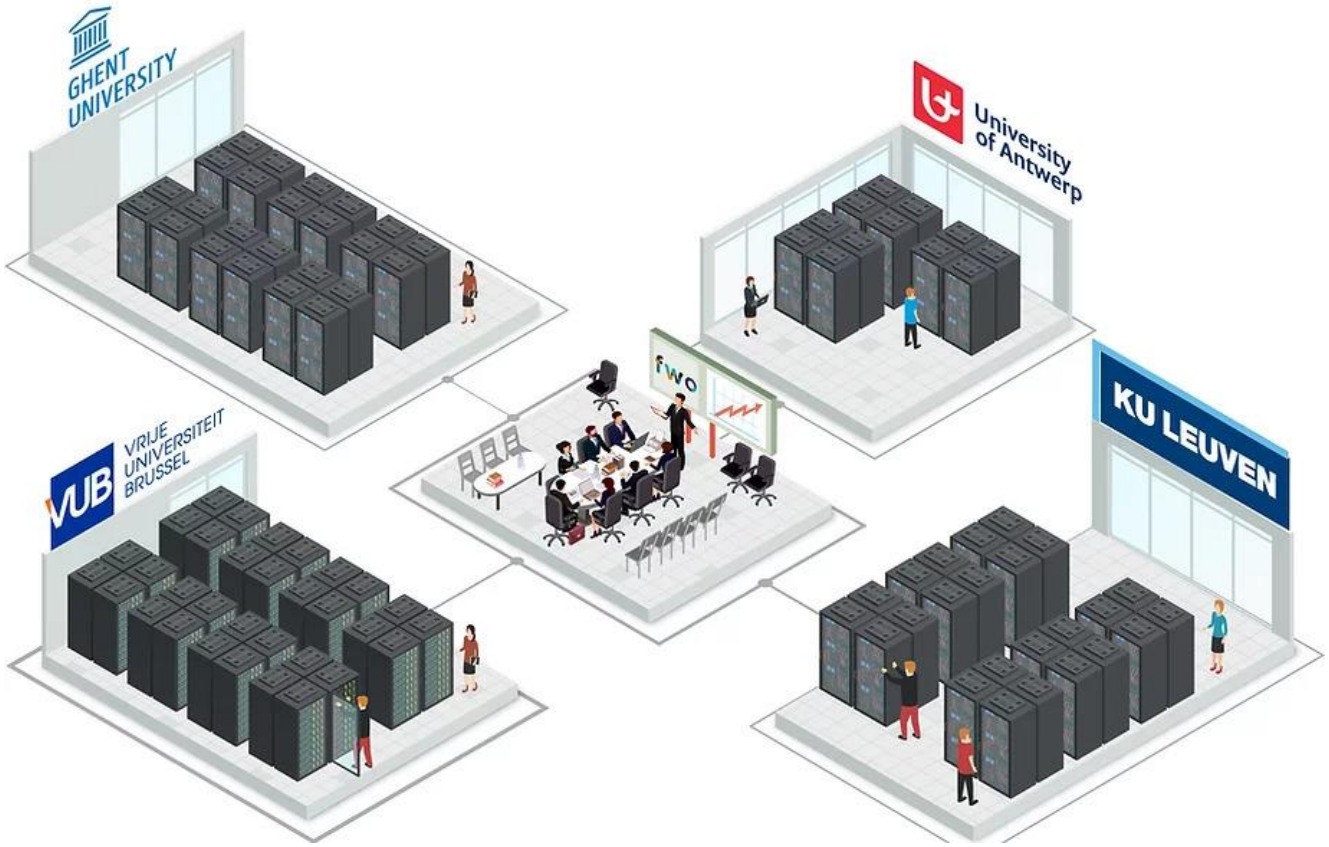
High-Performance Computing (HPC)

Software and databases installed

State-of-the-art GPU/CPU/RAM hardware

Expert support

Queuing system for allocating resources



AlphaFold on HPC

Advantages:

- Set up a large batches of experiments
- Faster + GPUs have greater memory → longer sequences
- No restricted availability (←→ Google Colab)

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

Tier 2

Available GPU servers (UGent):

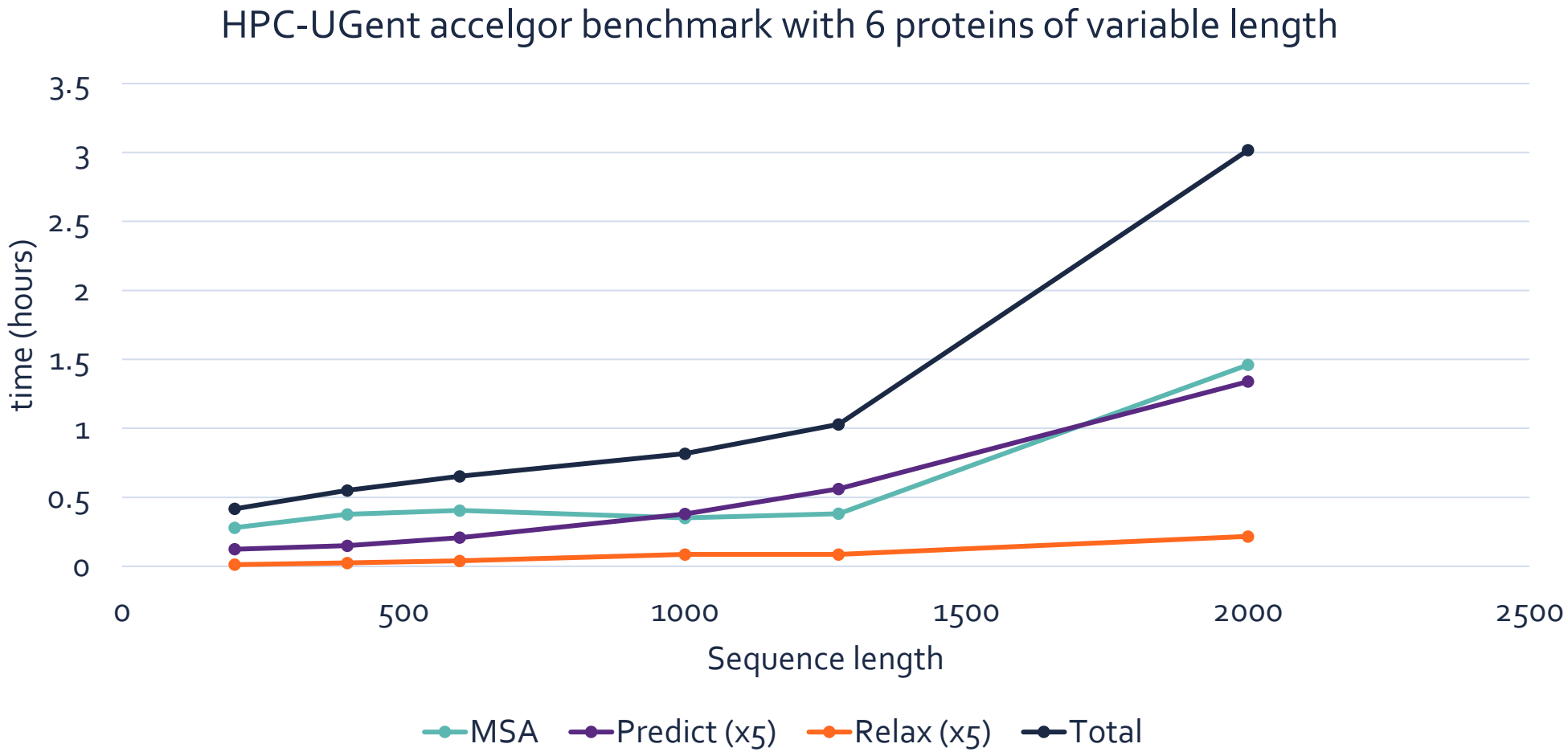
Available	joltik	10 nodes	4x V100 GPU per node (32GB)	256 GiB RAM p/n	800GB SSD p/n
Available	accelgor	9 nodes	4x A100 GPU per node (80GB)	~500 GiB RAM p/n	800GB SSD p/n

Available GPUs KUL: 20 x 4 P100 (16GB) + 2 x 4 V100 (32GB)

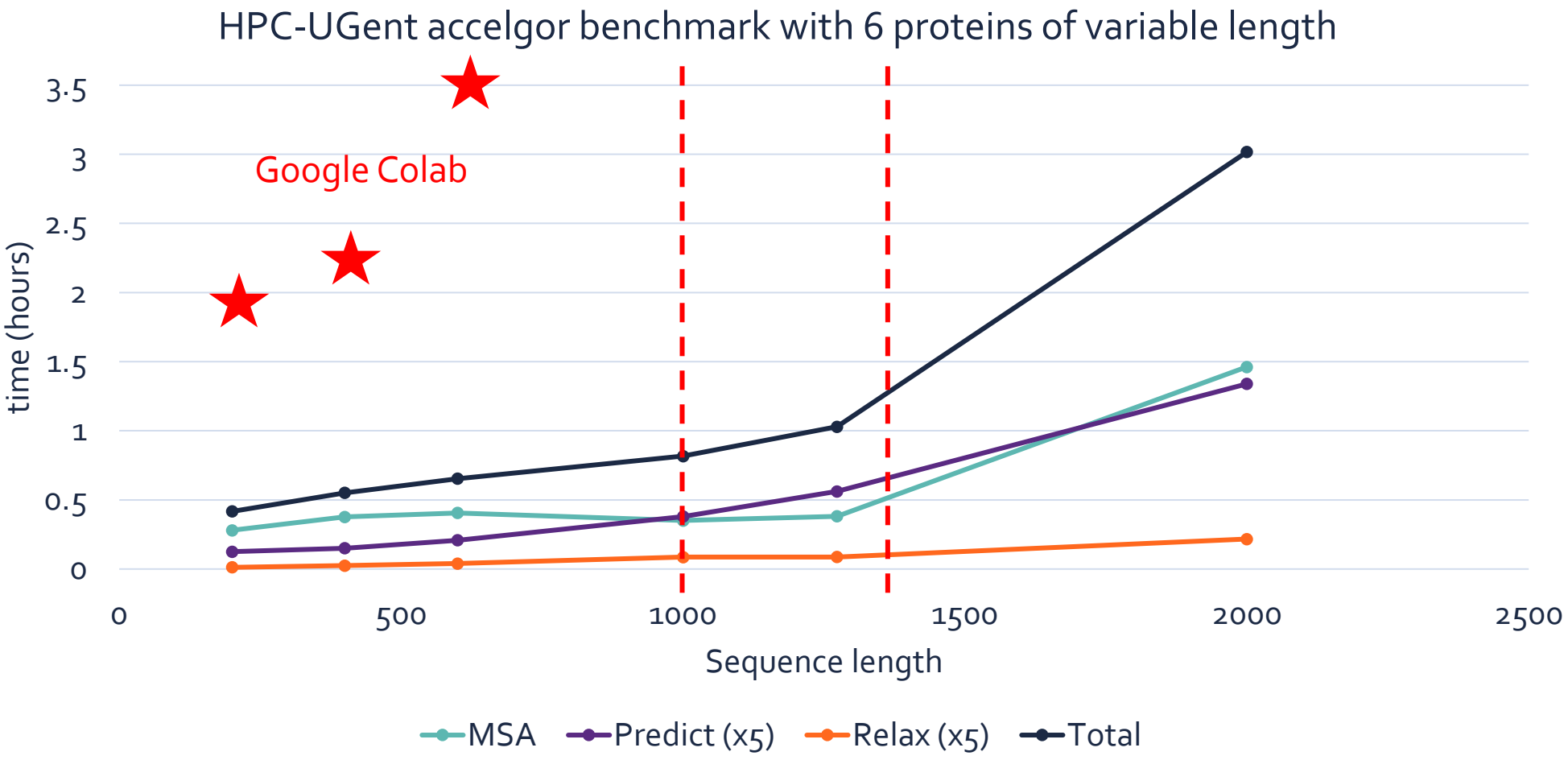
Available GPUs UA: 2 x 2 P100 (16GB)

Available GPUs VUB: 6 x 2 K20Xm (6GB) + 4 x P100 (16GB) + 6 x A100 (40GB)

Computation time on the HPC (accelgor)



Online Google Colab limits

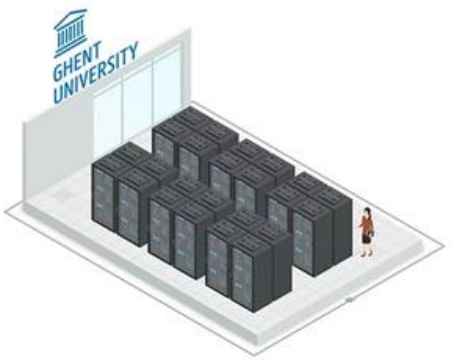


SCIENCE MEETS LIFE

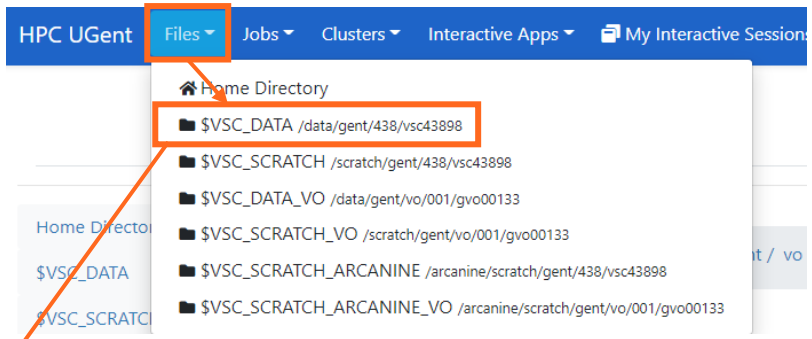
Running AlphaFold on the HPC

VIB

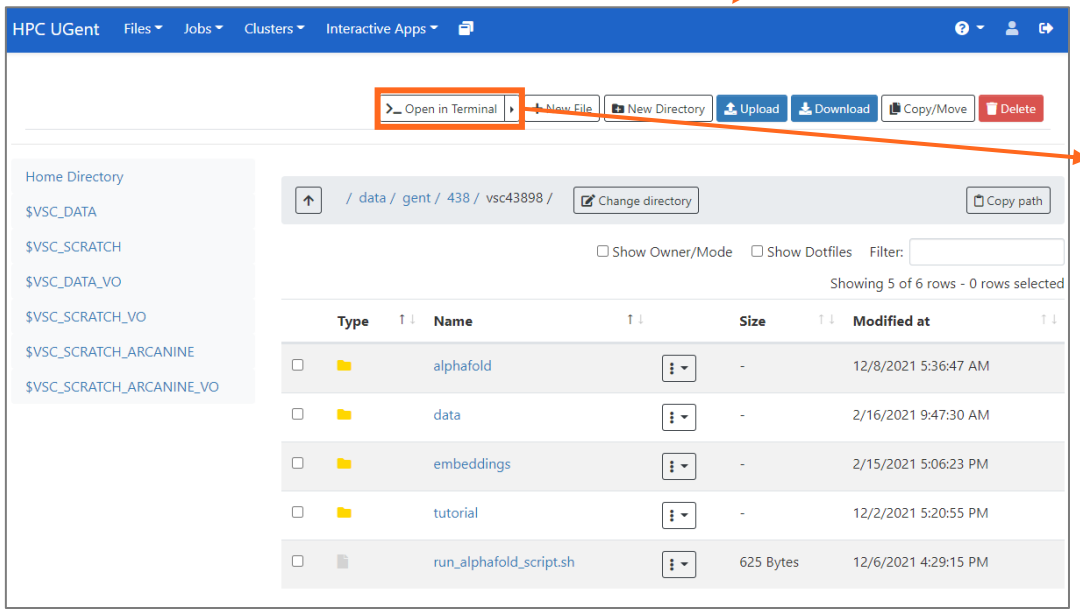
Set-up: logging in



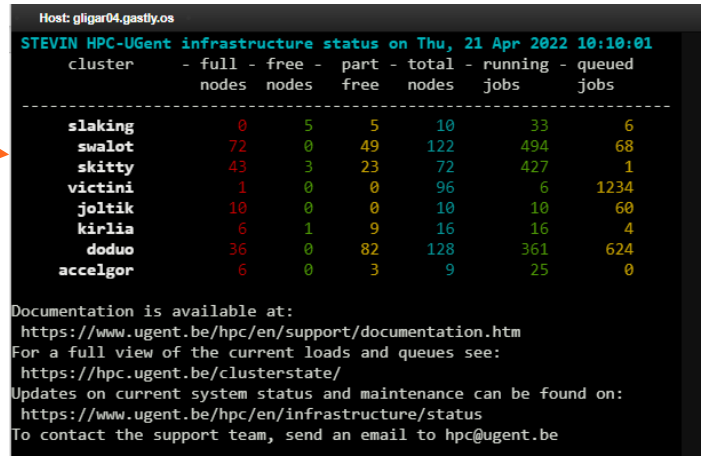
login.hpc.ugent.be



transfer files



launch scripts



Command line commands

Introduction of needed commands (full command line workflow):

- https://elearning.bits.vib.be/courses/alphafold/lessons/alphafold-on-the-hpc/topic/basic-unix-commands/

Needed commands for web interface workflow:

- `pwd` → print current directory
- `cd` → change directory
- `ls` → list files in directory
- `qsub` / `qdel` / `qstat` → see later
- `module` → see later

- (`mv` / `cp`)

```
(base) [vsc43898@gligar05 alphafold]$ ls
some_directory  some_text.txt
(base) [vsc43898@gligar05 alphafold]$ cp some_text.txt some_directory/
(base) [vsc43898@gligar05 alphafold]$ ls some_directory/
some_text.txt
(base) [vsc43898@gligar05 alphafold]$ mv some_text.txt new_name.txt
(base) [vsc43898@gligar05 alphafold]$ ls
new_name.txt  some_directory
(base) [vsc43898@gligar05 alphafold]$ ls some_directory/
some_text.txt
(base) [vsc43898@gligar05 alphafold]$ mv new_name.txt some_directory/
(base) [vsc43898@gligar05 alphafold]$ ls
some_directory
(base) [vsc43898@gligar05 alphafold]$ ls some_directory/
new_name.txt  some_text.txt
(base) [vsc43898@gligar05 alphafold]$
```

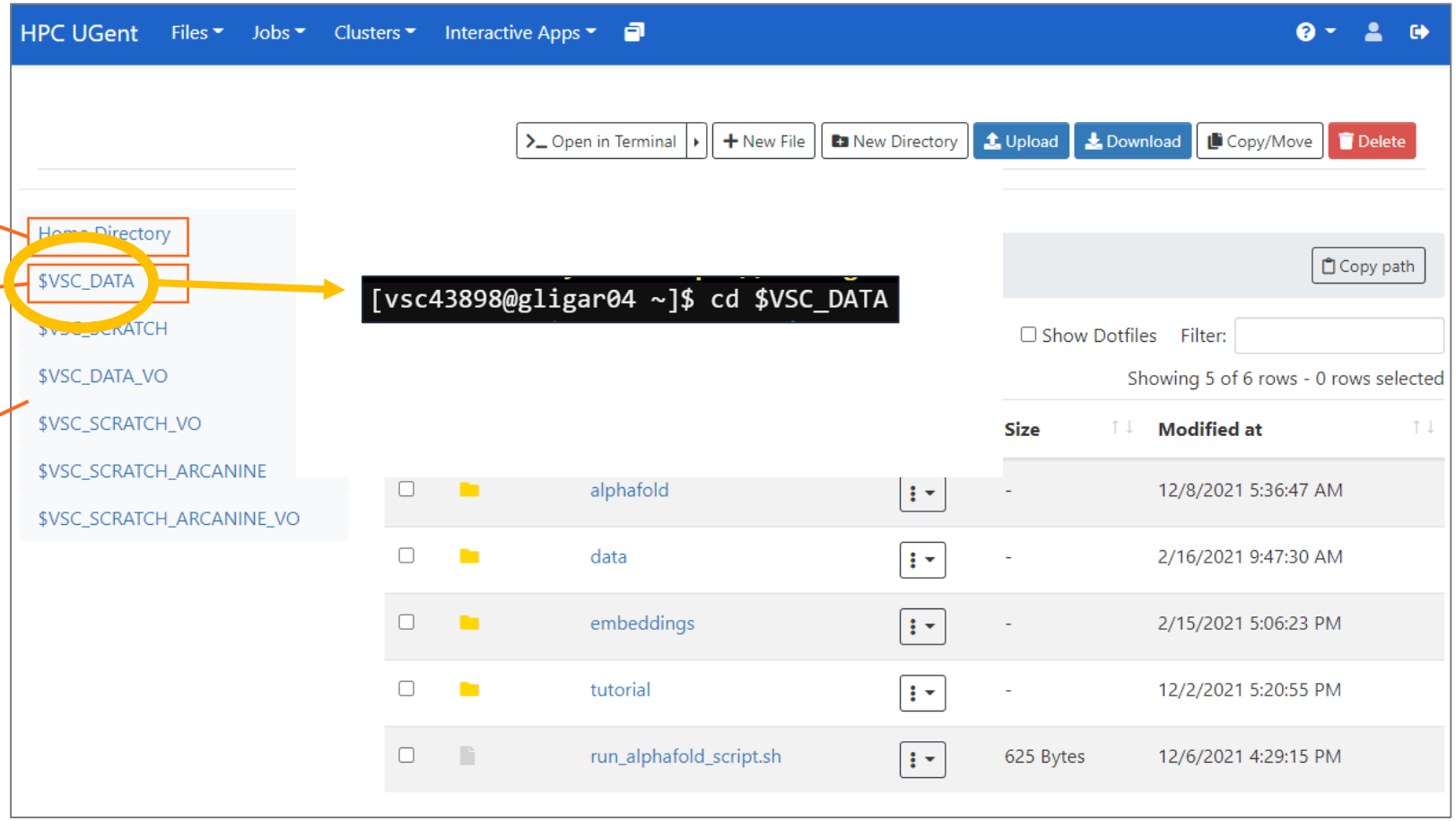
Set-up: directories and files

https://elearning.bits.vib.be/courses/alphafold/lessons/alphafold-on-the-hpc/topic/prepare-directories-and-files/

Default location when logging in (~6GB)

Extra storage capacity (~25GB)

When in a 'virtual organization' (~26TB shared)



Set-up: directories and files

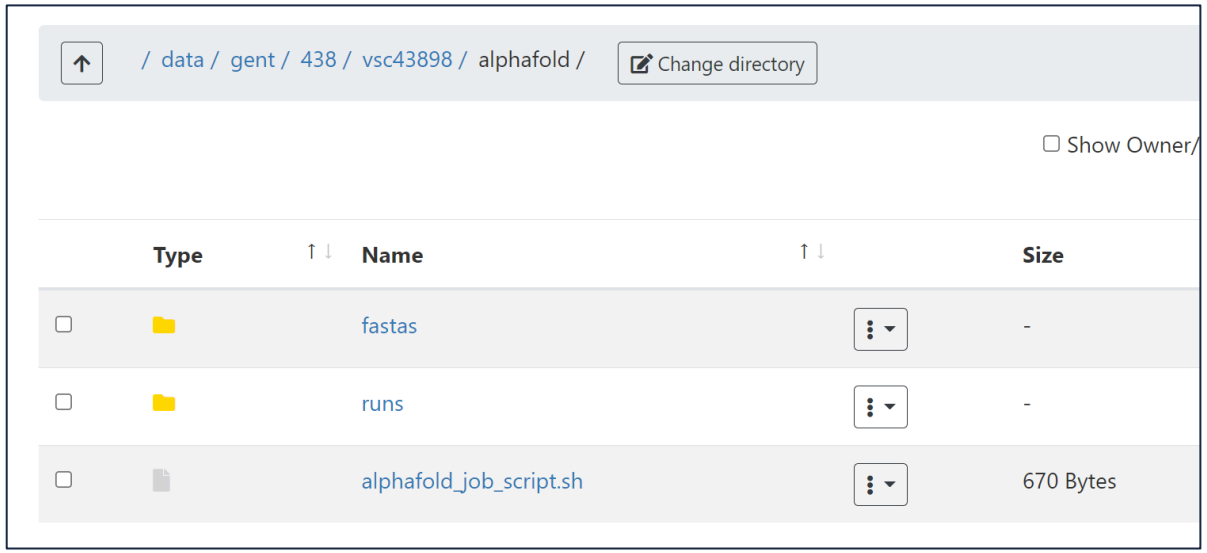
In $DATA_VSC: create directory **alphafold**

OR use *mkdir* on the command line,
OR use *New directory* on the web interface

There:

- Create directory **fastas**
- Create directory **runs**
- Download alphafold_job_script.sh

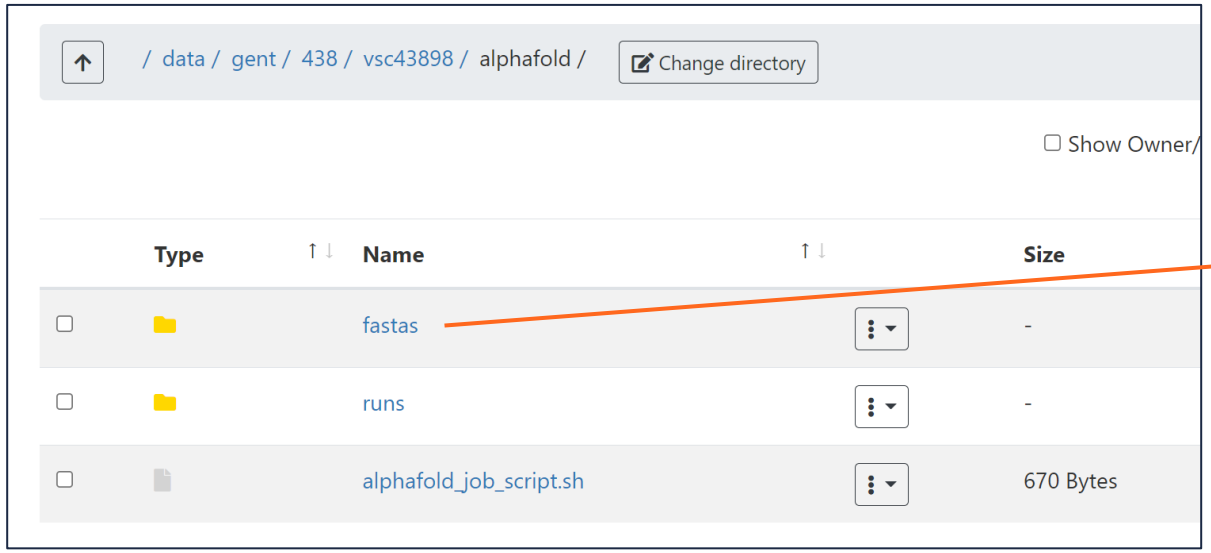
OR copy contents to a file,
OR right-click on **Raw** > Save link as,
OR use *wget* with the raw link (**Raw** > Copy link) on the command line directly



Set-up: directories and files

In **fastas**: create/upload your FASTA file

OR via command line (*nano*, *vim*, ...)
OR by uploading an existing file to the web interface
OR by creating + editing a new file via the web interface



Set-up: modify alphafold_job_script.sh

https://elearning.bits.vib.be/courses/alphafold/lessons/alphafold-on-the-hpc/topic/setting-up-the-job-script/

download link

Changes to make for individual experiments

```
#!/bin/bash
#PBS -N AlphaFold_tutorial_script
#PBS -l nodes=1:ppn=8:gpus=1
#PBS -l mem=64gb
#PBS -l walltime=24:0:0

PROTEIN=RBD

module load AlphaFold/2.1.1-fosscuda-2020b
export ALPHAFOLD_DATA_DIR=/arcanine/scratch/gent/apps/AlphaFold/20211201

WORKDIR=$VSC_DATA/alphafold/runs/$PBS_JOBID-$PROTEIN
mkdir -p $WORKDIR
cp -a $PBS_O_WORKDIR/fastas/$PROTEIN.fasta $WORKDIR/
cd $WORKDIR

echo Running $PROTEIN.fasta, output found at $WORKDIR
alphafold --fasta_paths=$PROTEIN.fasta \
          --max_template_date=2020-05-14 \
          --db_preset=full_dbs \
          --output_dir=$PWD \
          --model_preset=monomer_ptm
```

Optional: change this name to anything for experiment monitoring

Most important modification: the name of the FASTA file (located in $VSC_DATA/alphafold/fastas/)

Change this to *multimer* if you want to run protein complexes

Set-up: submitting and monitoring

```
(base) [vsc43898@gligar04 vsc43898]$ module swap cluster/joltik

We advise you to log in to a RHEL 8 login node when using the joltik cluster.

The joltik cluster is using RHEL 8 as operating system,
while the login node you are logged in to is using CentOS 7.

To avoid problems with testing installed software or submitting jobs,
it is recommended to switch to a RHEL 8 login node by running 'ssh login8'.


The following have been reloaded with a version change:
  1) cluster/victini => cluster/joltik

(base) [vsc43898@gligar04 vsc43898]$ ml

Currently Loaded Modules:
  1) cluster/joltik (S)

  Where:
   S:  Module is Sticky, requires --force to unload or purge
```

Necessary step: select the appropriate cluster (joltik in this case)

you can check it using *ml*

After setting up the cluster and the files, submit your job!

!! During this session only: priority access to joltik via

```
qsub alphafold_job_script.sh --pass=reservation=alphafold
```

```
(base) [vsc43898@gligar05 alphafold]$ qsub alphafold_job_script.sh
40169067
(base) [vsc43898@gligar05 alphafold]$ qstat
Job ID     Name             User            Time Use S Queue
---------- ---------------- --------------- -------- - -------
40169067   AlphaFold_tu...  vsc43898        0:00:00  Q joltik
```

Check status

```
Job ID     Name             User            Time Use S Queue
---------- ---------------- --------------- -------- - -------
40169067   AlphaFold_tu...  vsc43898        0:14:11  R joltik
```

To cancel runs: use *qdel*

```
(base) [vsc43898@gligar05 alphafold]$ qdel 40168964
```

Monitoring + outputs

/ data / gent / 438 / vsc43898 / alphafold / Change directory

Type	Name
	fastas
	runs
	alphafold_job_script.sh
	AlphaFold_tutorial_script.e40191367
	AlphaFold_tutorial_script.o40191367

error file: tells when something goes wrong

output file: tells where output files are located

Type	Name
	40191367-RBD

Type	Name
	RBD
	RBD.fasta

Type	Name
	msas
	features.pkl
	ranked_0.pdb
	ranked_1.pdb
	ranked_2.pdb
	ranked_3.pdb
	ranked_4.pdb
	ranking_debug.json
	relaxed_model_1_ptm.pdb
	relaxed_model_2_ptm.pdb
	relaxed_model_3_ptm.pdb
	relaxed_model_4_ptm.pdb
	relaxed_model_5_ptm.pdb
	result_model_1_ptm.pkl
	result_model_2_ptm.pkl
	result_model_3_ptm.pkl
	result_model_4_ptm.pkl
	result_model_5_ptm.pkl
	timings.json
	unrelaxed_model_1_ptm.pdb
	unrelaxed_model_2_ptm.pdb
	unrelaxed_model_3_ptm.pdb
	unrelaxed_model_4_ptm.pdb
	unrelaxed_model_5_ptm.pdb

Troubleshooting:

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

SCIENCE MEETS LIFE

Exercise:

run a prediction on the HPC

https://elearning.bits.vib.be/courses/alphafold/lessons/vib-training-session-alphafold/topic/first-experiment-on-the-hpc/

VIB

SCIENCE MEETS LIFE

AlphaFold outputs

VIB

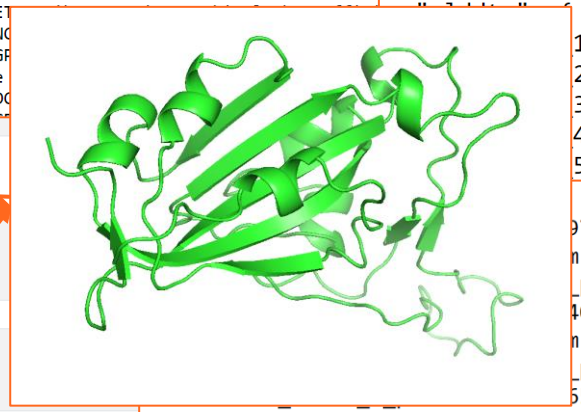
AlphaFold output

Type	Name	
	msas	
	features.pkl	
	ranked_0.pdb ... ranked_4.pdb	
	ranking_debug.json	
	relaxed_model_1_ptm.pdb ... relaxed_model_5_ptm.pdb	
	result_model_1_ptm.pkl ... result_model_5_ptm.pkl	
	timings.json	826 Bytes
	unrelaxed_model_1_ptm.pdb ... unrelaxed_model_5_ptm.pdb	124 KB 124 KB

```
>SARS-CoV-2 spike RBD
ITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNL
>tr|A0A088DJY6|A0A088DJY6_9BETC Spike protein OS=Bat Hp-betacoronavirus/Zhejiang2013 OX=1541205 PE=4 SV=1
------FQSLINVTeaTIPSPAFWRRHYVRNCNYDISVFTDNADVYSLQCYGVAPSSLADMCWEEAHIDYMKISEKDIFSFKPSGAGDFAKYNYKLPSDFMGCTVVFTNqeltcNA----TSGQLchvyt
MAITLKPTRTSATVCGYkQK
>tr|F1BYL9|F1BYL9_9BETC Putative spike glycoprotein OS=Zaria bat coronavirus OX=989337 GN=S PE=4 SV=1
------AERLLNASidQIPDPAFWKRHVIRNCKFNFSHIMALSHVYDMQCYGIDASKLPSTCWNEVYADVFRLAQDDFYSFKPSASGDLATYNYKLPSDFLGCTLILTNpnlycGN----STT--cgipg
LALTLKPATVASTVCDVtAQ
>tr|E0ZN60|E0ZN60_BCHK9 Spike glycoprotein OS=Bat coronavirus HKU9-10-2 OX=875613 GN=S PE=4 SV=1
------ICSLPYSDilKPPQPIVWKRHTVTNCSFDFDAIVNRLPTFQLKCFGVSPAKLAQMCYSSVTLDIFRANTTHLANMLGKVPDVFSKYNYALPPDFYGCVHSYYIND----TSPMYAIAQQW----
>tr|F1DAZ9|F1DAZ9_9BETC Spike protein OS=Rousettus bat coronavirus/Kenya/KY06/2006 OX=983925 GN=S PE=4 SV=1
------YCRPPYNAllDPPQPVVWRRFMLYDCAFDFSVVIDNLPTHQLQCYGISPRRLASMCYSSVTIDVMRINATHLNNLLNRVPDSFSLYNYAVPDDFYGCLHAFYLNS----T-TAYAVANQF----
>sp|A3EXG6|SPIKE_BCHK9 Spike glycoprotein OS=Bat coronavirus HKU9 OX=694006 GN=S PE=1 SV=1
------YCTPPYSVlqDPPQPVVWRRYMLYDCVFDFTVVVDSLPTHQLQCYGVSPRRLASMCYGSVTLDVMRINETHLNNLFNRVPDTFSLYNYALPDNFYGCLHAFYLNS----T-APYAVANRF----
>tr|A3EXJ0|A3EXJ0_BCHK9 Spike glycoprotein OS=Bat coronavirus HKU9-4 OX=424370 GN=S PE=4 SV=1
------TCDIPYAAlqTPPQPIAWRRYAVSKCGFDFEAVINRLPTFELKCFGVSPARLASMCYGKVTIDVFRINVTHLANLIAGVPDAFSKYNYALPRDFYGCVHAFYVNM----S-SDYIIADSW----
>tr|A0A2P1M5J5|A0A2P1M5J5_BCHK9 Spike glycoprotein OS=Bat coronavirus HKU9-2 OX=694006 GN=S PE=4 SV=1
------VCQPPYAAleNPPQPVVWRRYLVRDCAFDFATVINNLPTYQLHCYGVSPSRLASMCYNTITIDVMRINTTHLNNLLQKVPDAFSLYNYAIPSDFYGCIHAYYLNV----T-DTYAIATQR----
>tr|A0A0P0INJ4|A0A0P0INJ4_9BETC Spike glycoprotein OS=SARS-like coronavirus bat CoV BB9904/BGR/2008 OX=1737344 PE=4 SV=1
------FDQVFNASSFPSVYAWERVRITDCVANYAVLYNSsVSFSTFQCYGVSPTKLNDLCFSSVYA
TQSSGISFQPYRVVVLSFELLNAPATVCGPKQ
>tr|A0A023PTS3|A0A023PTS3_9BET
------FGEVFNATTFPSVYAWERKRISNC
YTTNGIGYQPYRVVVLSFELLNAPATVCGF
>tr|D5HJQ1|D5HJQ1_BCHK3 Spike
------FDKVFNATRFPNVYAWERTKISDC
```

Actual input given to the predictor (after MSA)

.pkl = specific python data format



```
1_ptm": 88.7186512177675,
2_ptm": 89.14702063046553,
3_ptm": 71.71698708689162,
4_ptm": 84.52844698435865,
5_ptm": 70.38715136492051
97,
m": 6.088567495346069,
_ptm": 243.2005693912506,
4076690674,
m": 2.400113821029663,
_ptm": 222.6782784461975,
5502456665,
"relax_model_4_ptm": 9.147892951965332,
"process_features_model_5_ptm": 2.1541543006896973,
"predict_and_compile_model_5_ptm": 187.0064091682434,
"relax_model_5_ptm": 7.8038649559021
}
```

Open with PyMOL or any other protein structure viewer, or online at: https://www.ncbi.nlm.nih.gov/Structure/icn3d/

AlphaFold output folder

Type	Name	Size
	msas	-
	features.pkl	4.71 MB
	ranked_0.pdb	243 KB
	...	
	ranked_4.pdb	243 KB
	ranking_debug.json	369 Bytes
	relaxed_model_1_ptm.pdb	243 KB
	...	
	relaxed_model_5_ptm.pdb	243 KB
	result_model_1_ptm.pkl	28.3 MB
	...	
	result_model_5_ptm.pkl	28.4 MB
	timings.json	826 Bytes
	unrelaxed_model_1_ptm.pdb	124 KB
	...	
	unrelaxed_model_5_ptm.pdb	124 KB

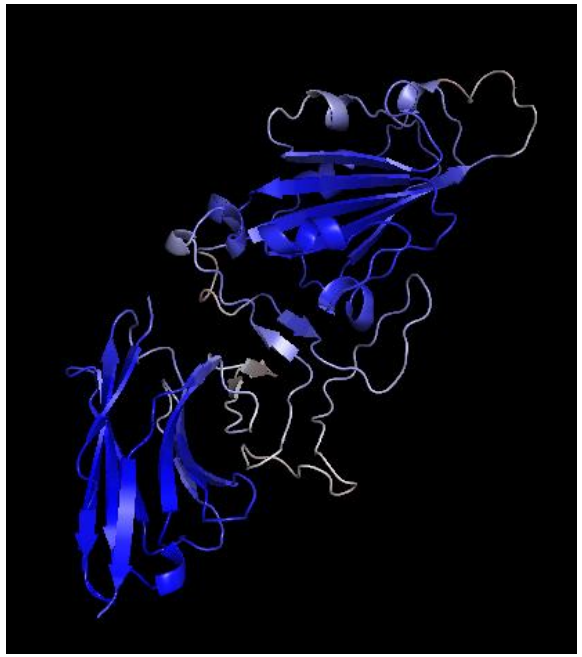
SCIENCE MEETS LIFE

Evaluating & interpreting AlphaFold predictions

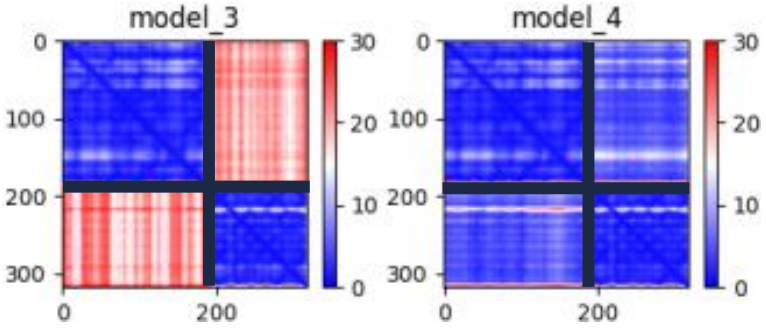
VIB

Metrics and scores to evaluate model accuracy

Local confidence
pLDDT (predicted local distance difference test)



Global confidence
PAE (predicted aligned error)



Individual scores:

- **Average PLDDT**
- **PTM** (predicted template modelling score)
- **iPTM** (interface PTM)

pTM-score: intrinsic global accuracy estimate

- **TM-score** indicates the similarity between two structures by a score **between 0 and 1,** where 1 indicates a perfect match between two structures

- The TM-score is intended as a more accurate measure of the global similarity of full-length protein structures than the often used RMSD measure (independent of protein length, less sensitive to local dissimilarity).

- pTM score is the **predicted TM-score** between the structural prediction and the (unknown) true structure

- pTM for complexes == 0.2 * PTM + 0.8 * iPTM

- **Alphafold predictions are ranked based on the pTM score**

- References:
  TM-score: Zhang Y and Skolnick J (2004). "Scoring function for automated assessment of protein structure template quality". *Proteins*
  pTM score: Jumper et al., (2021). "Highly accurate protein structure prediction with AlphaFold". *Nature*

pLDDT reflects local model accuracy

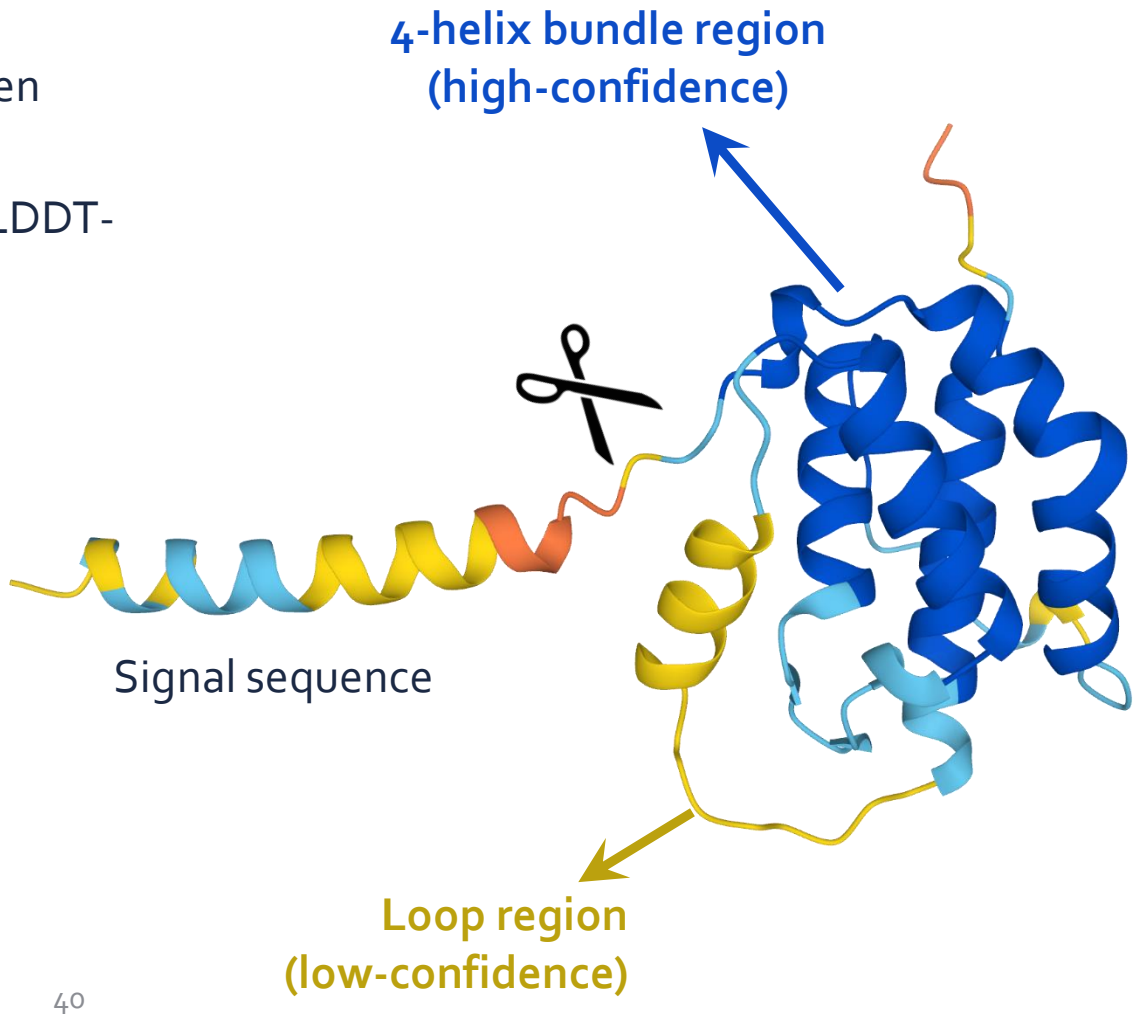
- pLDDT: **p**redicted **l**ocal **d**istance **d**ifference **t**est
- per residue confidence metric
- Estimates agreement of local environment between true structure and prediction
- structural models are coloured according to the pLDDT-metric (relative scale 1- 100):

Model Confidence:

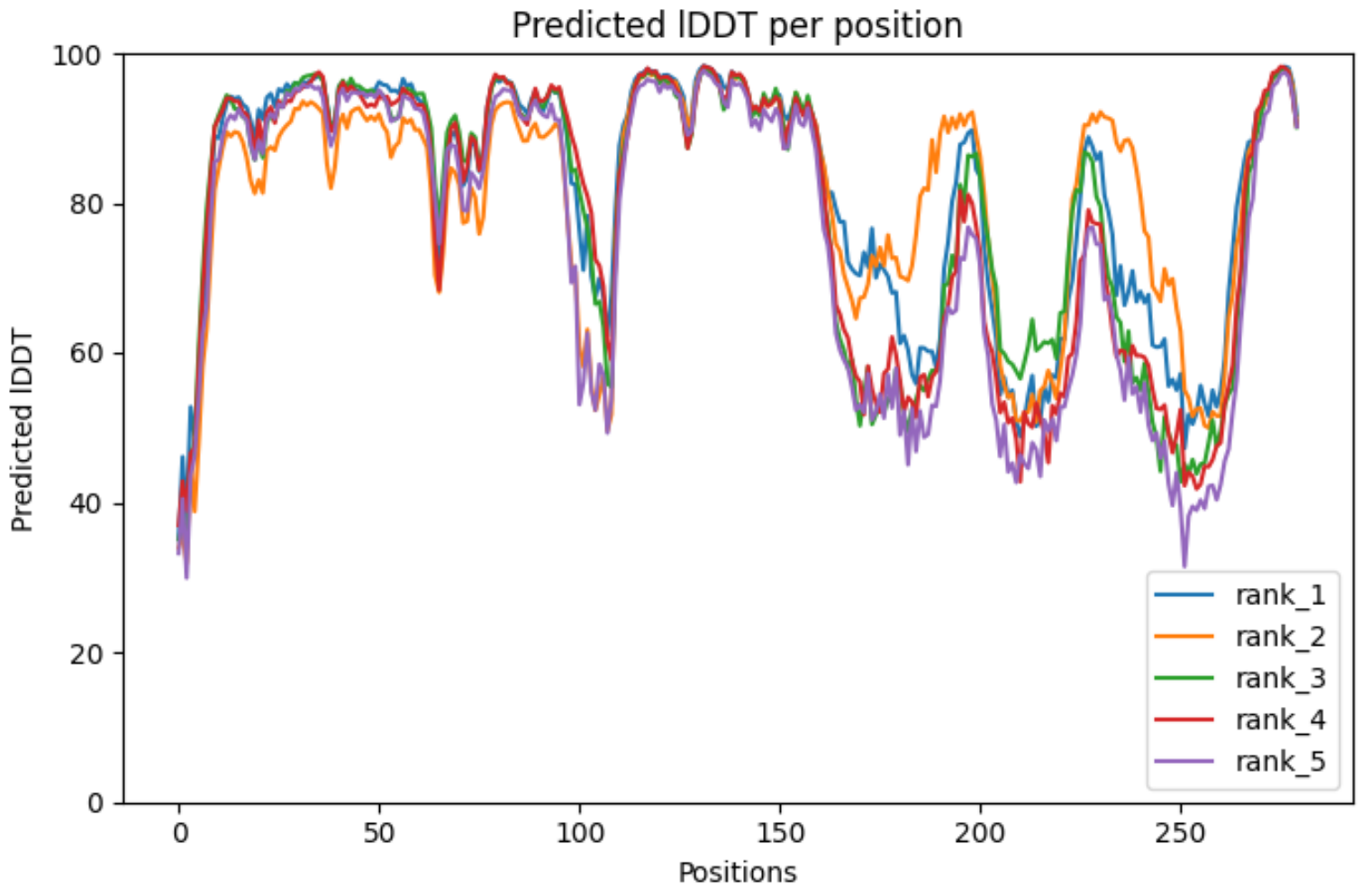
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

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

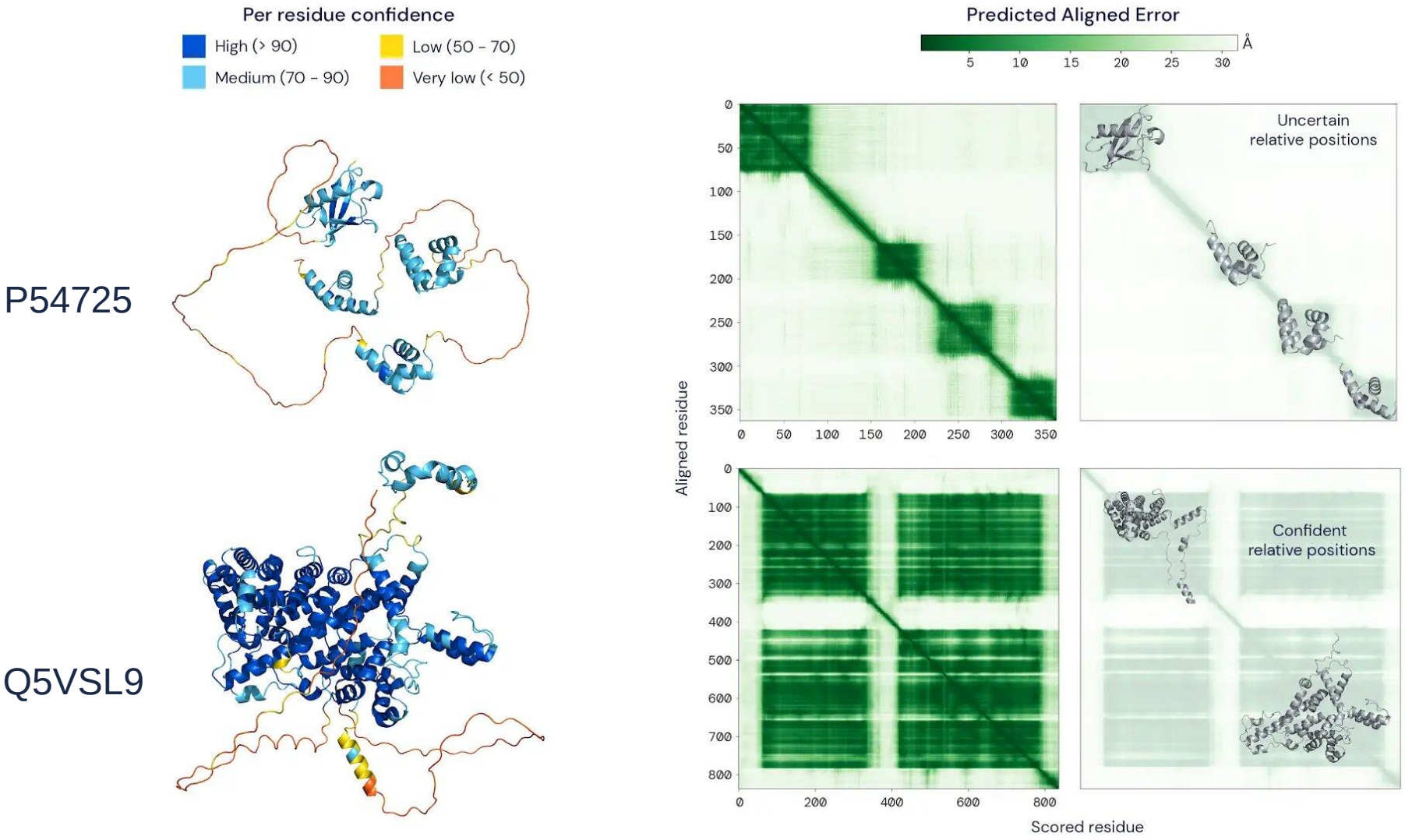
Structural model for Human TSLP



pLDDT reflects local model accuracy



PAE evaluates interdomain accuracy



https://www.deepmind.com/

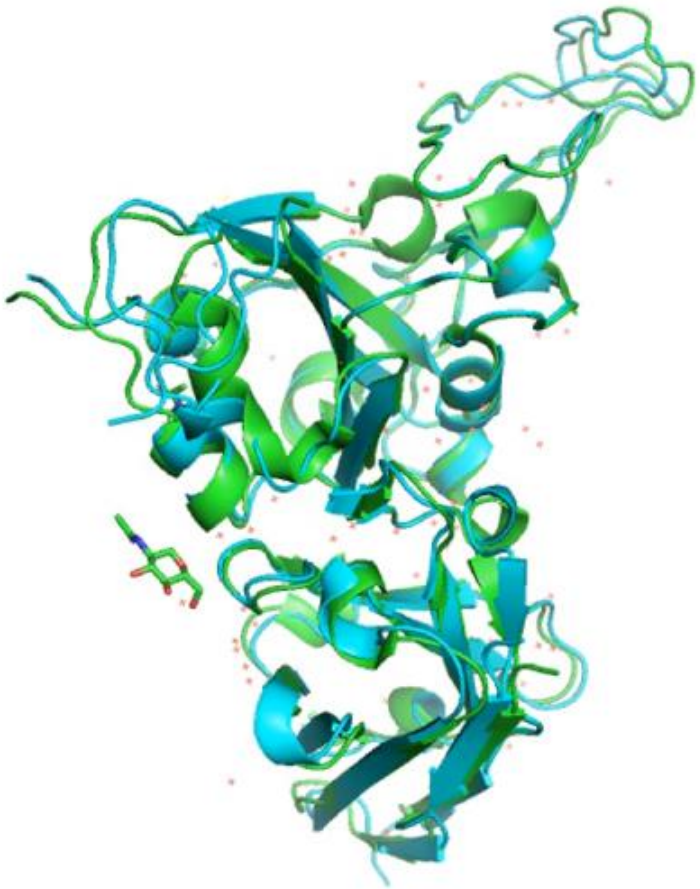
PAE-plot evaluates interdomain accuracy

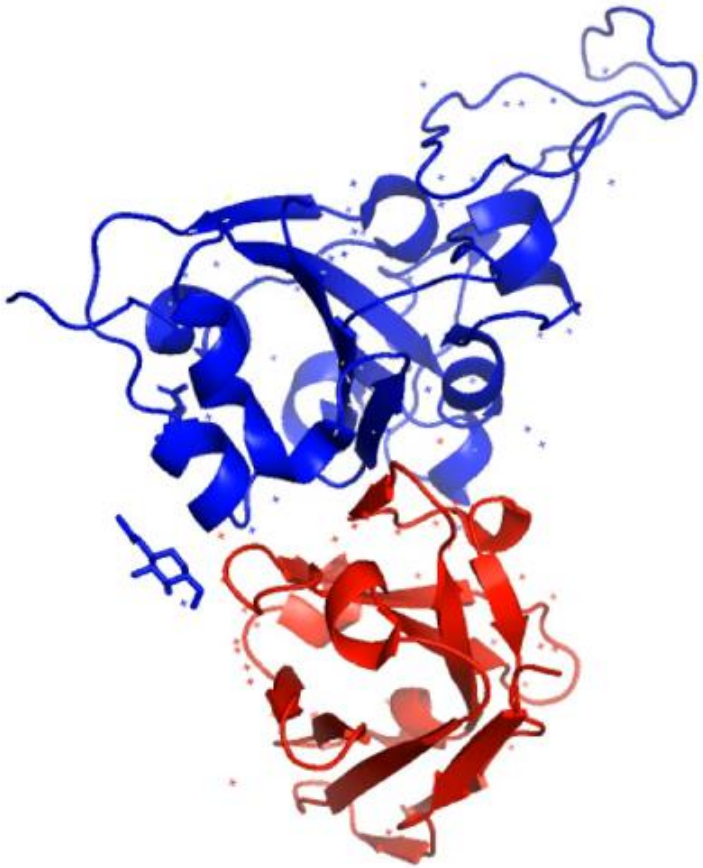
Example:

https://alphafold.ebi.ac.uk/entry/Q13049

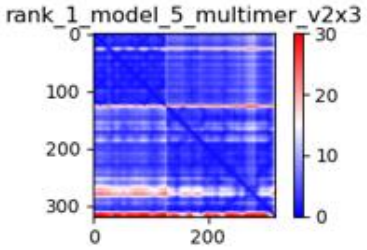
(see *Predicted aligned error tutorial* (scroll down))

Example: VHH72 + SARS-CoV-1 (PDB:6WAQ)





PTM=0.87

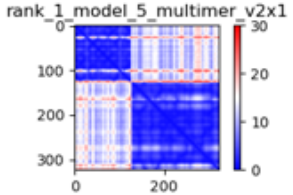


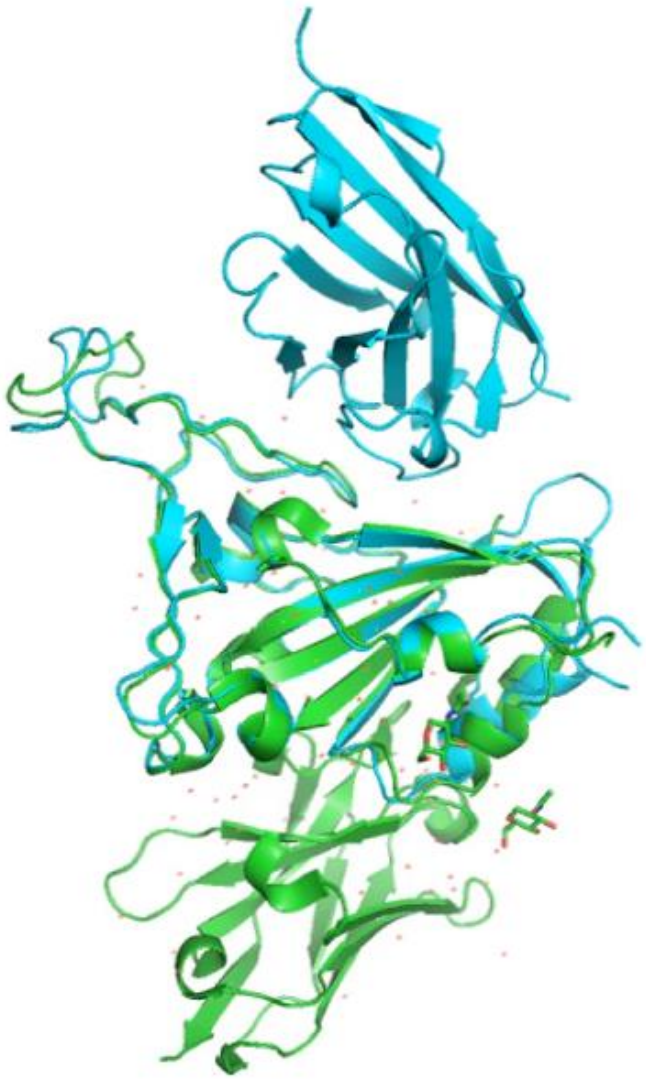
Green: PDB structure
Cyan: predicted by AlphaFold

Blue: SARS-CoV-1 RBD
Red: VHH72

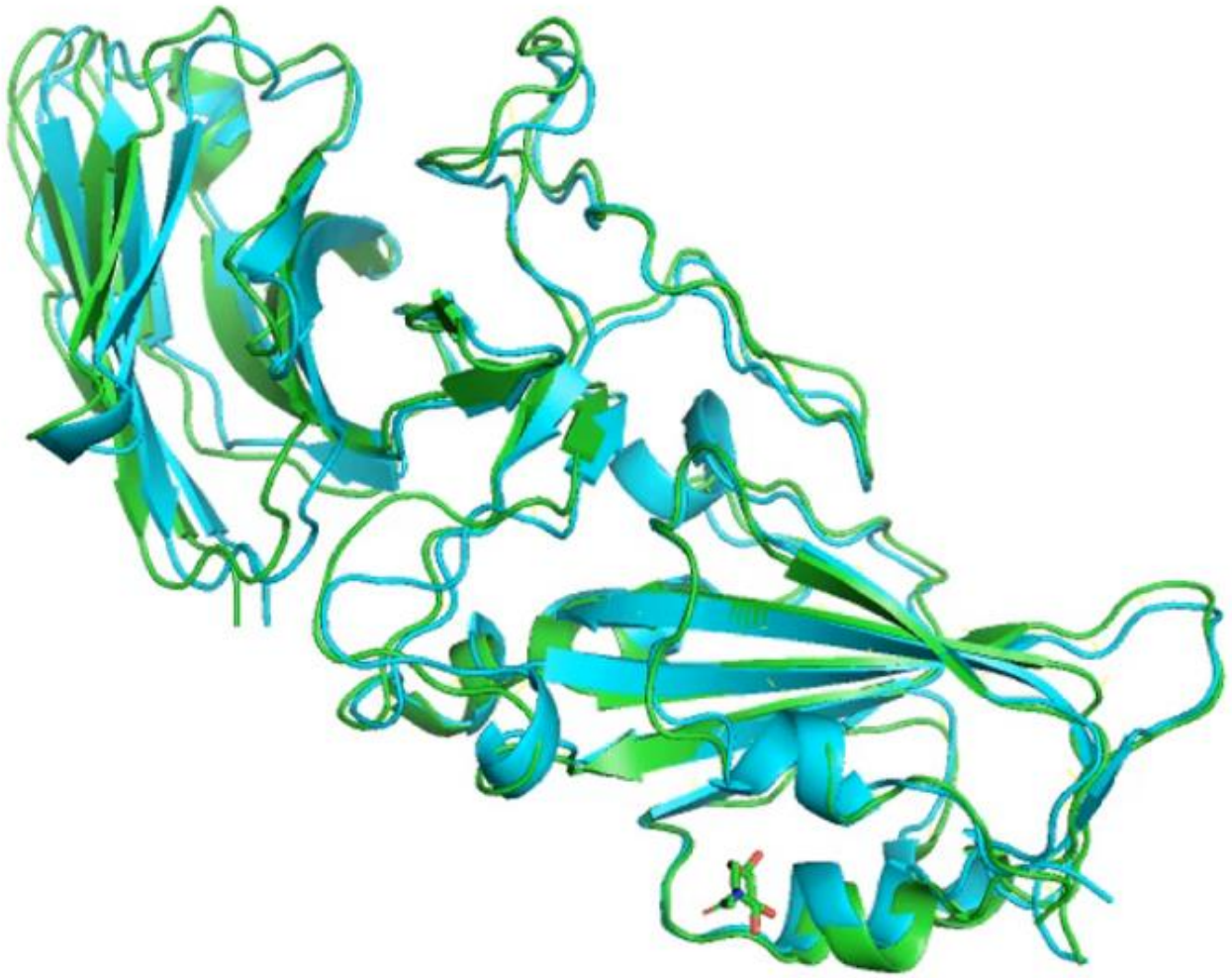
Example: SARS-CoV-2 + VHH72

PTM=0.53

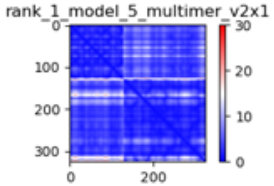




Example: SARS-CoV-2 + VHH E (PDB:7B14)



PTM=0.89



SCIENCE MEETS LIFE

Pitfalls of AlphaFold

VIB

AlphaFold: uncertainties and potential pitfalls

1. Low accuracy of regions where **few sequences are available** for alignment
2. Low accuracy of **intrinsically unstructured regions**
3. Limited information on **structural dynamics**
4. Variable accuracy of **multimeric assemblies**
5. No information on **co-factors/ligands/post-translational modifications**
6. No post-translational processing, no compartmentalization of transmembrane assemblies

SCIENCE MEETS LIFE

Exercise:

Generate PAE/PLDDT using python script

!! Recent problem on joltik: if you have used “module load” after swapping to joltik, you need to open a new terminal session to run the python scripts in this exercise.

(otherwise you get the *Illegal instruction (core dumped)* error)

VIB

https://elearning.bits.vib.be/courses/alphafold/lessons/vib-training-session-alphafold/topic/generate-pae-plddt-plots/

SCIENCE MEETS LIFE

Acknowledgements

Jasper Zuallaert

Kenneth Verstraete

Alexander Botzki

Janick Mathys

Kenneth Hoste

HPC UGENT

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

fwo

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

VIB-UGENT CENTER FOR INFLAMMATION RESEARCH

VIB BIOINFORMATICS CORE

UNIVERSITEIT GENT