SCIENCE MEETS LIFE

AlphaFold: a new age for protein folding

Jasper Zuallaert

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY
UNIVERSITEIT GENT
fwo

28/01/2022

Jasper Zuallaert

PhD in Computer Science (promotors Wesley De Neve / Yvan Saeys / Nico Callewaert)

(*deep learning for sequence analysis*)

GHENT UNIVERSITY
GLOBAL CAMPUS

Post-doc at Nico Callewaert + Lennart Martens units at Flemish Institute for Biotechnology (VIB) / UGent

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

GHENT UNIVERSITY

fwo

Introduction

AlphaFold: protein folding software by Google DeepMind

One of the most frequent software requests ever (EasyBuild)

Table of contents:

From DNA to protein

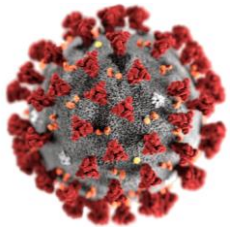
From DNA to protein

human genome

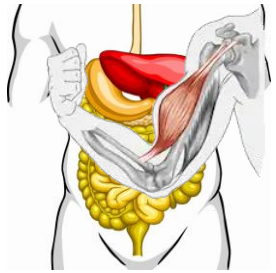
DNA

phenotypical traits
(e.g. eye color)

muscle contraction



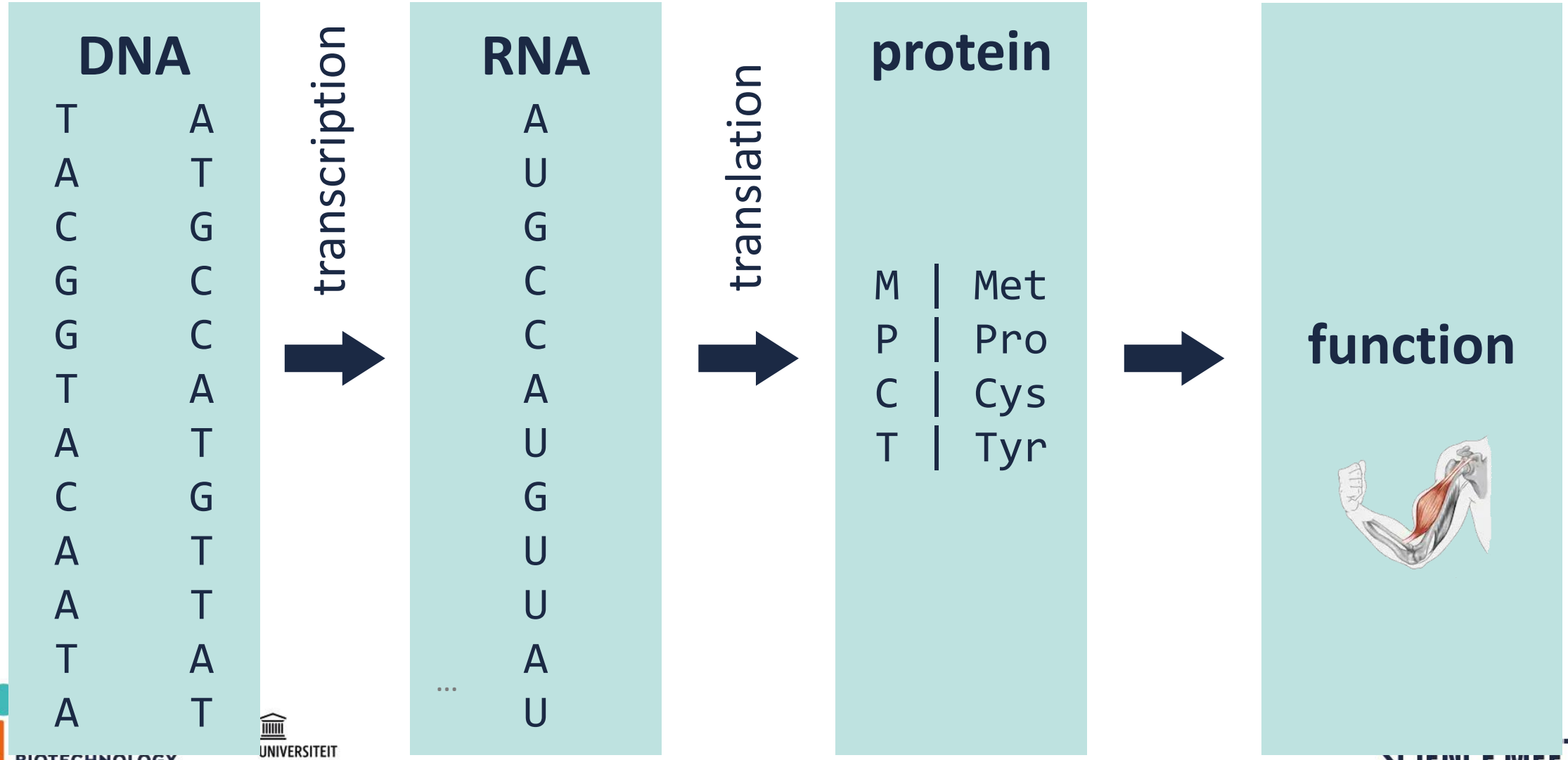
immune responses



food digestion

...

Central dogma of molecular biology



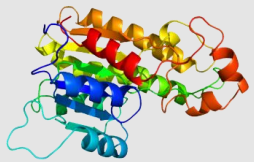
nucleotides

nucleotides

amino acids

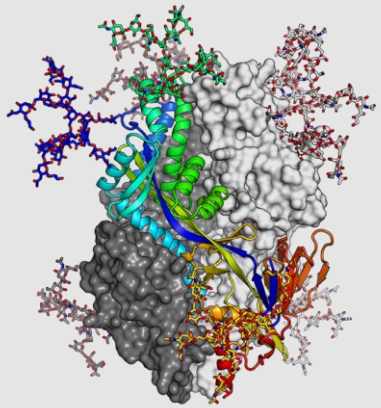
Central dogma of molecular biology

MPFWCMTRAS…



Linear protein sequence is folded into a non-linear 3-D structure

Protein structures interact with other molecules



Cell specificity

Post-translational modifications

…

nucleotides

nucleotides

protein

M | Met
P | Pro
C | Cys
T | Tyr

tory factors

t on a linear sequence

amino acids

function

Protein structure prediction

Available sequences and structures

After filtering for redundancy

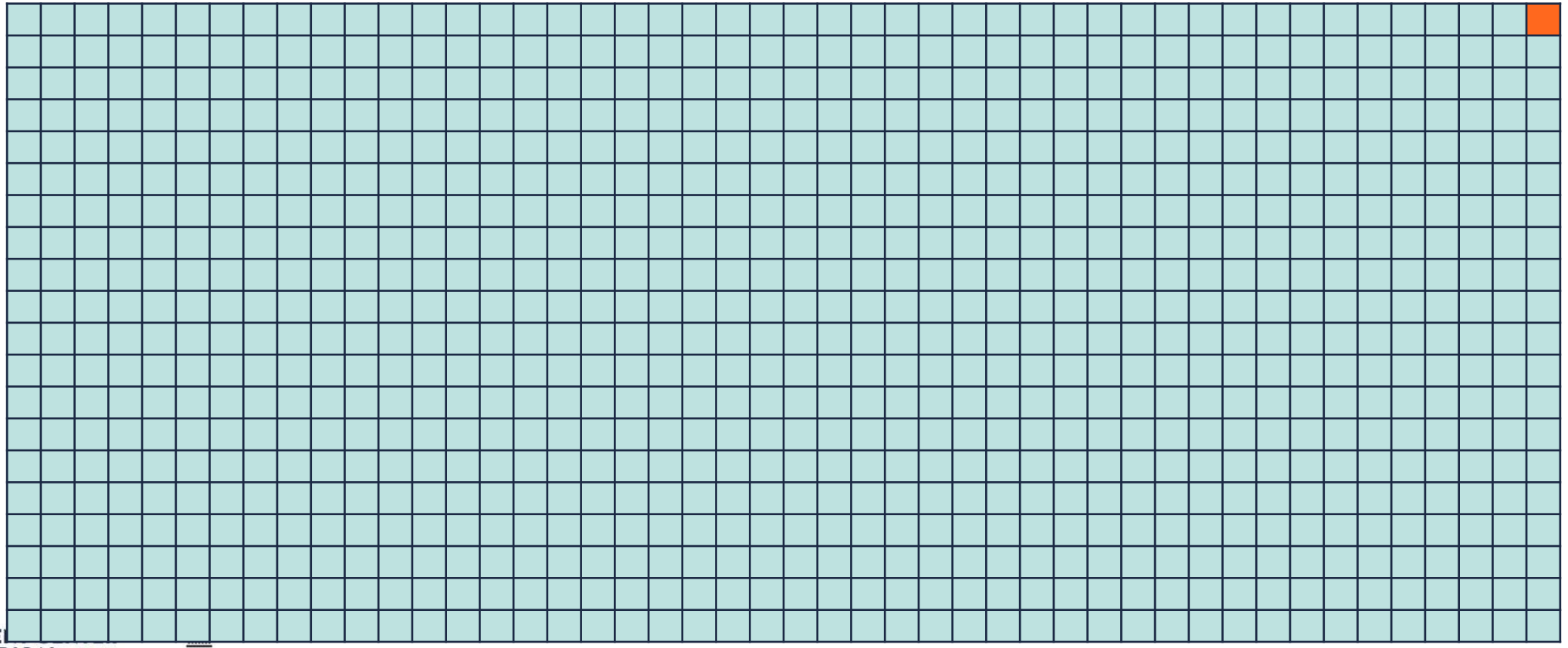
(50% similarity)

UniRef50

24 million unique sequences

Protein Data Bank (PDB)

~50 000 unique structures



□ = 50 000 proteins

VIB-UGE FOR MEDICAL BIOTECHNOLOGY
UNIVERSITEIT GENT
SCIENCE MEETS LIFE

Main protein structure resource

RCSB PDB
PROTEIN DATA BANK

~50 000 unique structures

Main experimental technique: X-ray crystallography

- Costly
- Labour-intensive
- Slow

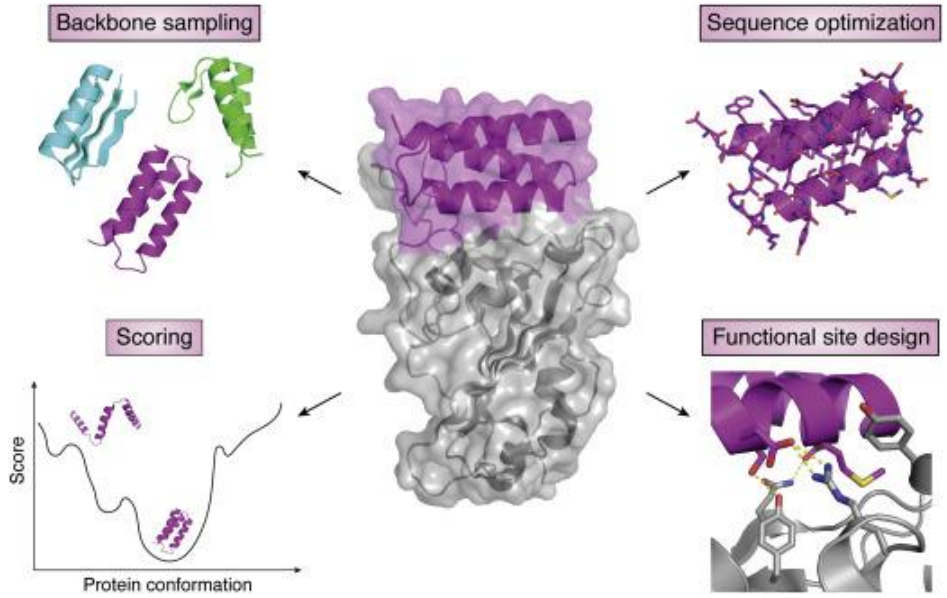
Image source: https://www.prweb.com/releases/2017/04/prweb14219956.htm

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

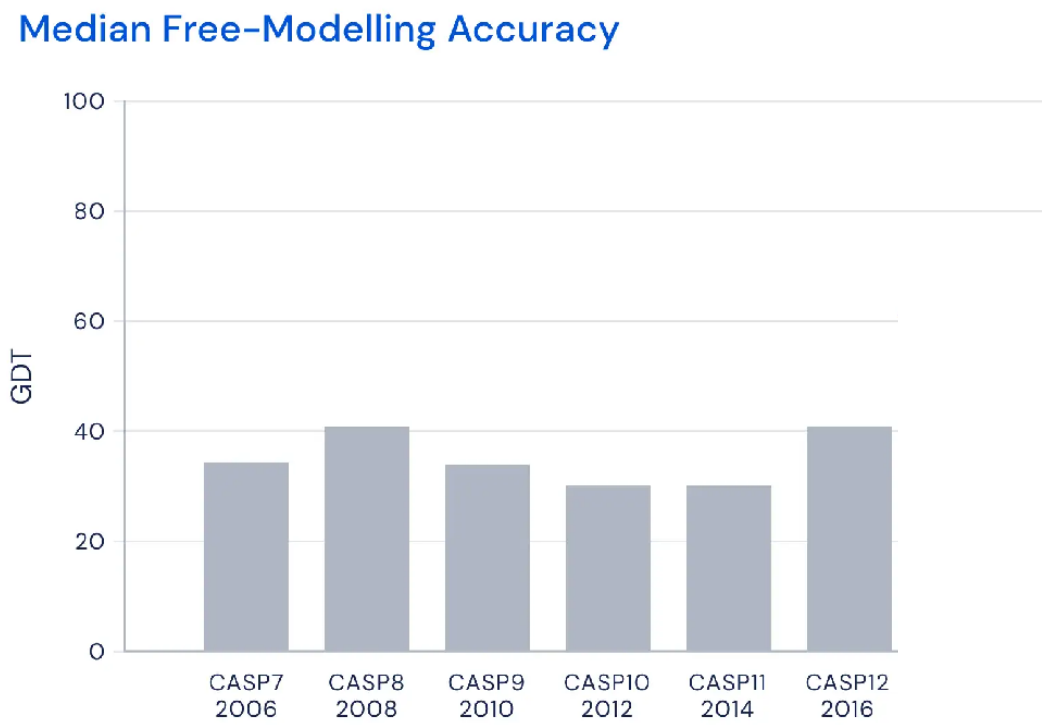
VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

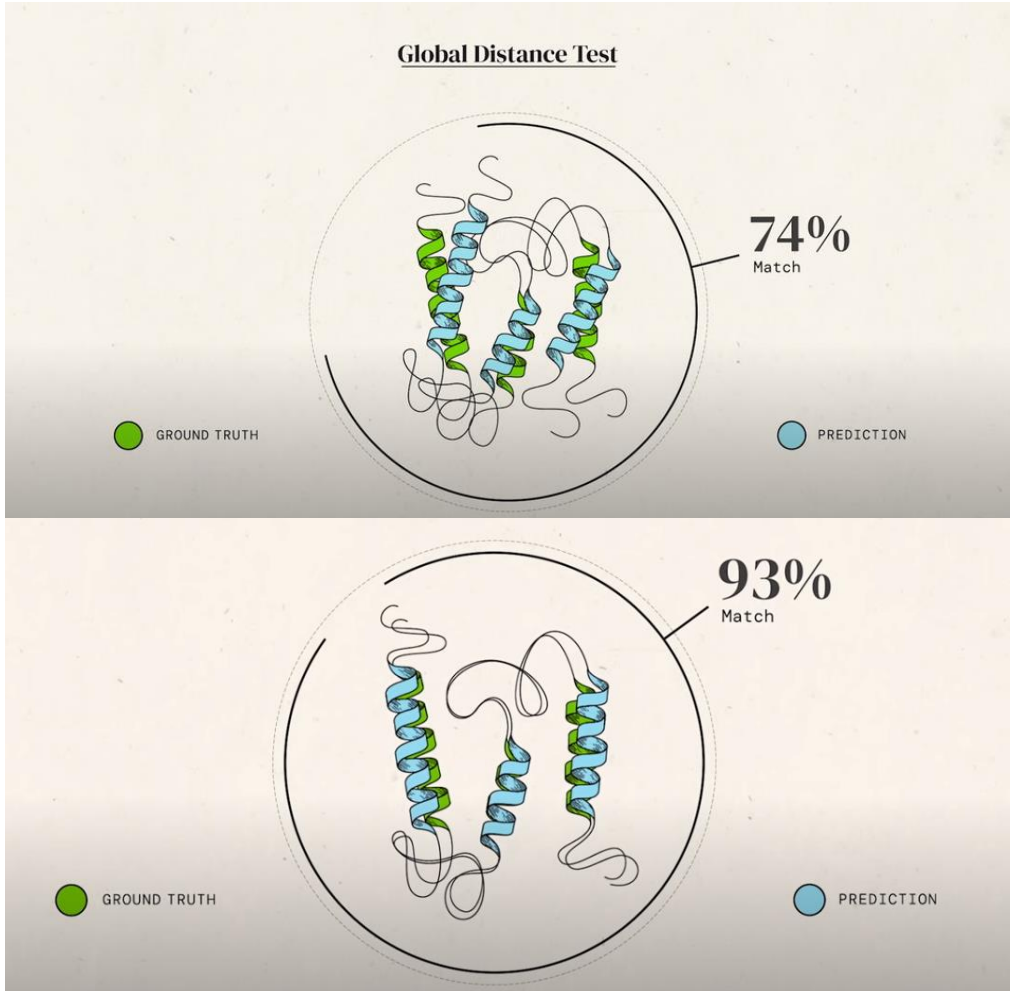
UNIVERSITEIT GENT

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

SCIENCE MEETS LIFE

Critical Assessment of Structure Prediction (CASP14, 2020)





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

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

SCIENCE MEETS LIFE

AlphaFold

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

DeepMind: a timeline

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

← Founded in 2010

Acquired by Google for $500M

Google DeepMind

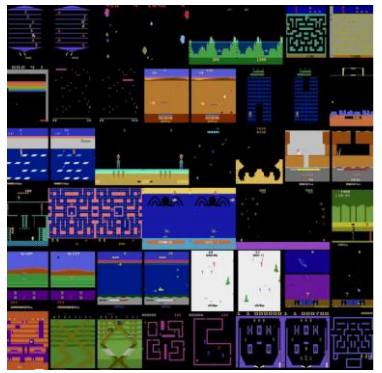
AlphaGo beats Lee Sedol (4-1)

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

AlphaFold (1.0) wins CASP13

AlphaFold (2.0) wins CASP14

Agent57 surpasses human level performance on all Atari2600 suite games



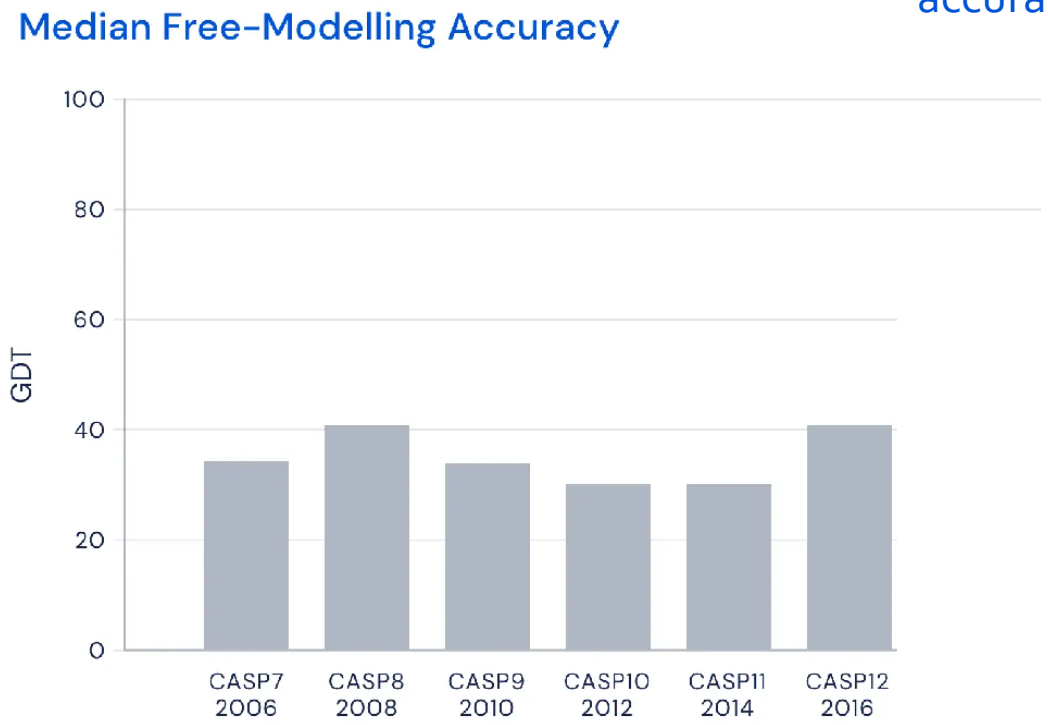
VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

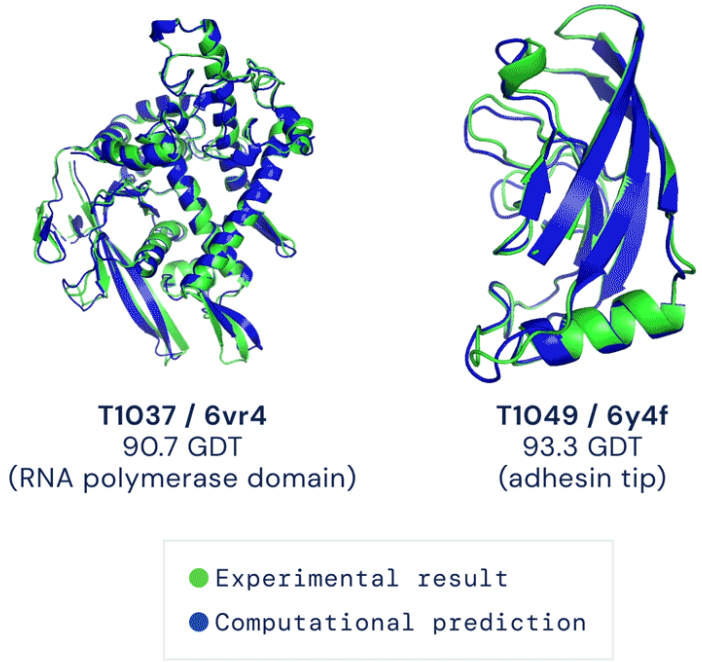
UNIVERSITEIT GENT

SCIENCE MEETS LIFE

Critical Assessment of Structure Prediction (CASP14, 2020)

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





VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

Reception of Alphafold 2

Mohammed AlQuraishi
@MoAlQuraishi

CASP14 #s just came out and they're astounding —DeepMind looks to have solved protein structure prediction. Median GDT_TS went from 68.5 (CASP13) to 92.4!!!! Cf. their 2nd best CASP13 struct scored 92.8 (out of 100). Median RMSD is 2.1Å. I think it's over predictioncenter.org/casp14/zscores...

1:13 PM · Nov 30, 2020 · Twitter for iPhone

619 Retweets **293** Quote Tweets **2,125** Likes

The Telegraph

'Once in a generation advance' as Google AI researchers crack 50-year-old biological challenge

The development could 'significantly accelerate' drug development for cancer and other diseases

"This is a big deal," says John Moult, a computational biologist at the University of Maryland in College Park, who co-founded CASP in 1994 to improve computational methods for accurately predicting protein structures. "In some sense the problem is solved."

Dr. Mohammed AlQuraishi at Columbia University, who also participated in CASP, lauded the AI as transformational. "It's a breakthrough of the first order, certainly one of the most significant scientific results of my lifetime," he said to *Nature*.

All of the groups in this year's competition improved, Moult says. But with AlphaFold, Lupas says, "The game has changed." The organizers even worried DeepMind may have been cheating somehow. So Lupas set a special challenge: a membrane protein from a species of archaea, an ancient group of microbes. For 10 years, his research team tried every trick in the book to get an x-ray crystal structure of the protein. "We couldn't solve it."

But AlphaFold had no trouble. It returned a detailed image of a three-part protein with two long helical arms in the middle. The model enabled Lupas and his colleagues to make sense of their x-ray data; within half an hour, they had fit their experimental results to AlphaFold's predicted structure. "It's almost perfect," Lupas says. "They could not possibly have cheated on this. I don't know how they do it."

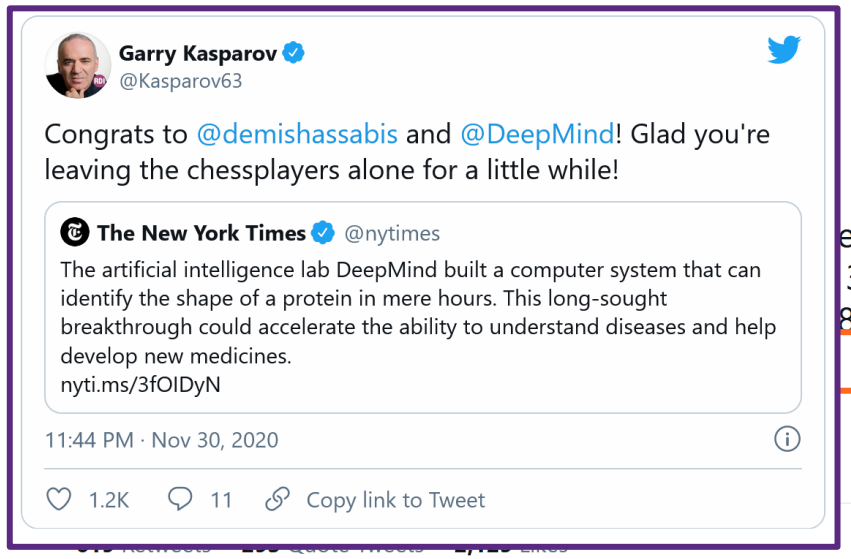
Eric Topol
@EricTopol

A "gargantuan" leap today for #AI life science. @DeepMind @GoogleAI #AlphaFold2 prediction of #3D protein structure from amino acids nature.com/articles/d4158... by @ewencallaway @NatureNews twitter.com/demishassabis/... @demishassabis

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

SCIENCE MEETS LIFE

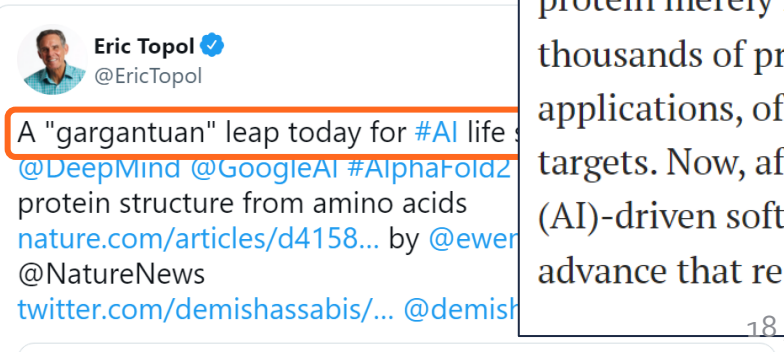
Reception of Alphafold 2

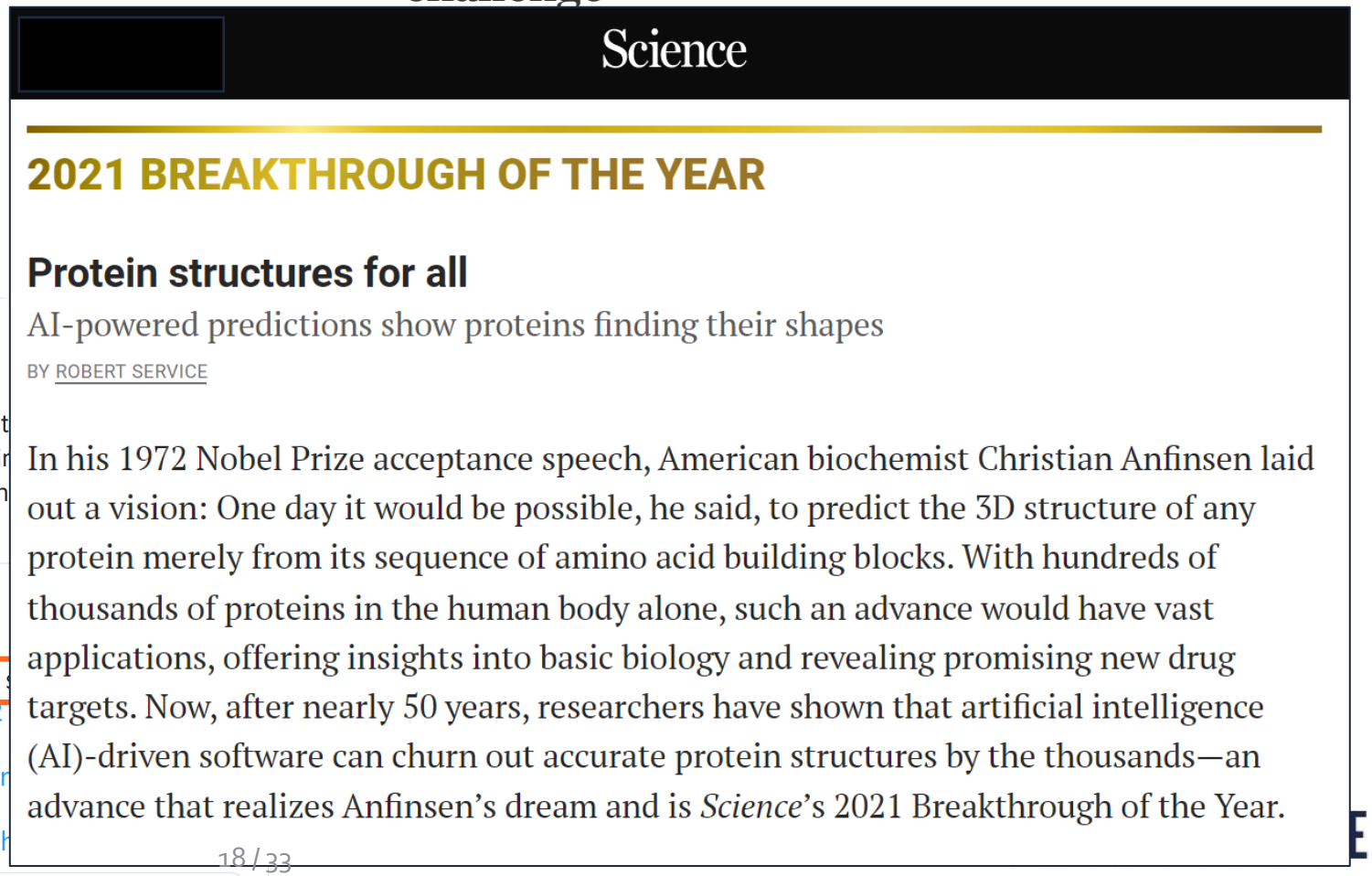


The Telegraph

'Once in a generation advance' as Google AI researchers crack 50-year-old biological challenge

Dr. Mohammed AlQuraishi at Columbia University, who also participat
CASP, lauded the AI as transformational. "It's a breakthrough of the fi
certainly one of the most significant scientific results of my lifetime," h
Nature.



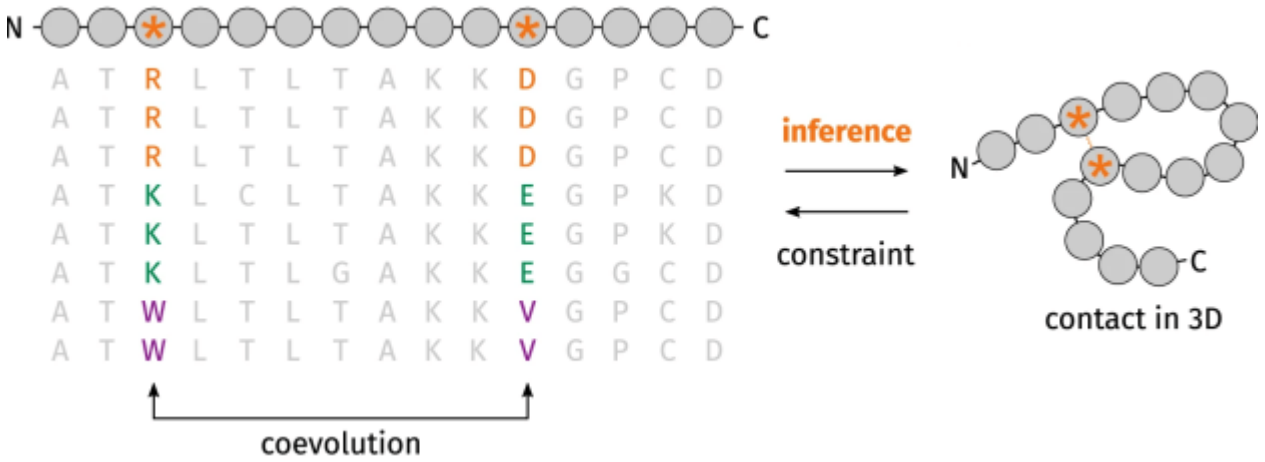


VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

Multiple Sequence Alignment

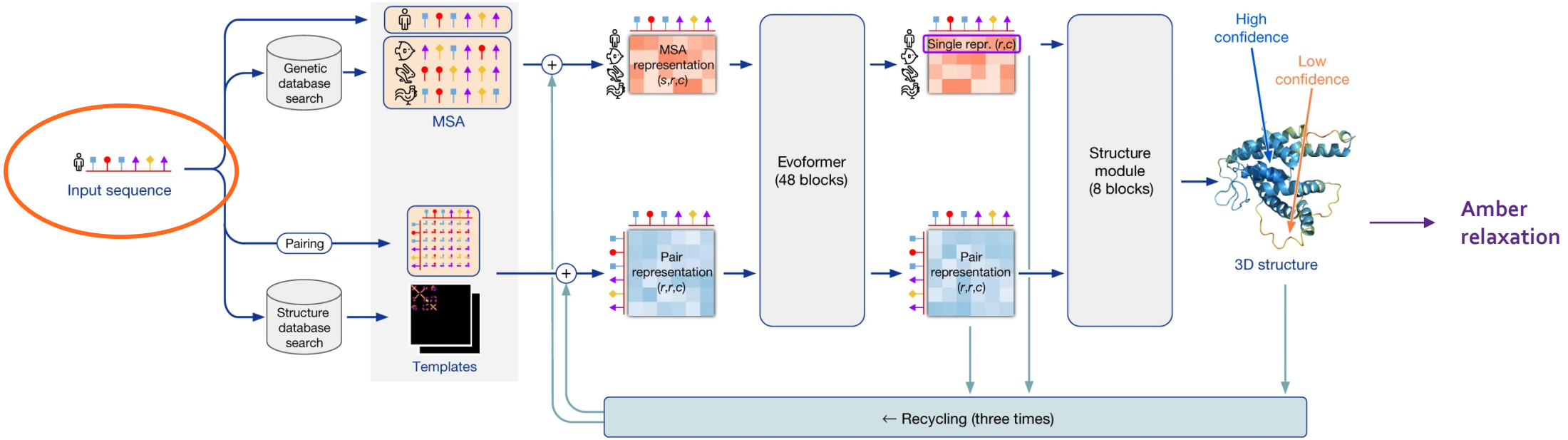
AlphaFold input = query protein sequence

First step: calculate multiple sequence alignment



Source: https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

AlphaFold stages

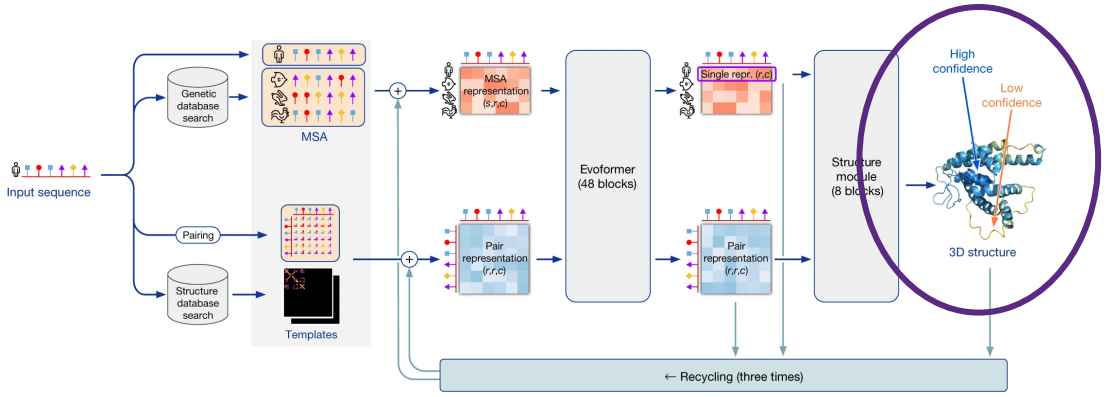


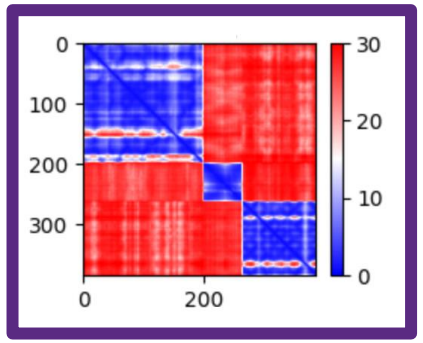
Multiple sequence alignment (CPU)
Structural templates search

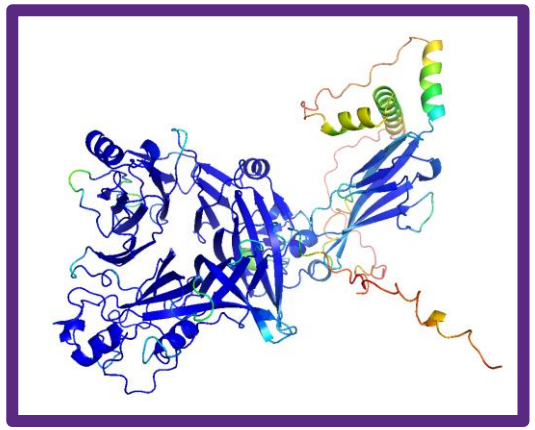
Run inputs through 5 trained AlphaFold models (GPU)
→ 5 structures predicted
→ sorted on confidence

Relax all 5 structures (Amber)

AlphaFold training







Any neural network is trained using a loss function

Main: **Frame Aligned Point Error**

+ auxiliary losses:
- → FAPE + torsion loss at intermediate structures
- → distogram prediction loss
- → Masked MSA prediction loss
- → Structural violation loss
- → ...
- → Predict local distance difference test (pLDDT)
- → Predicted Alignment Error (PAE)

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

AlphaFold availability

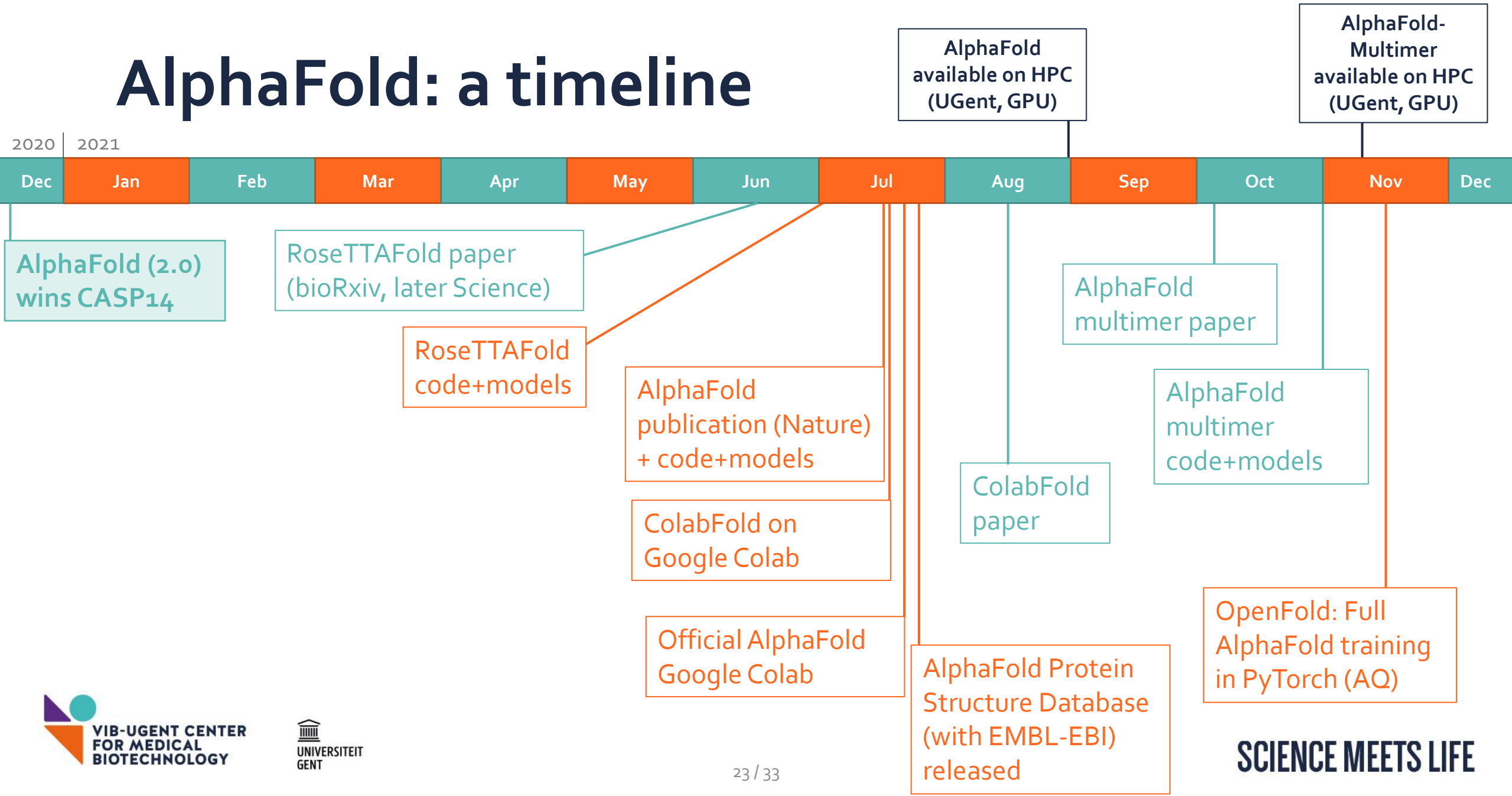
GitHub

Colabfold

Google colab

AlphaFold
Protein Structure Database
Developed by DeepMind and EMBL-EBI

AlphaFold: a timeline



VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

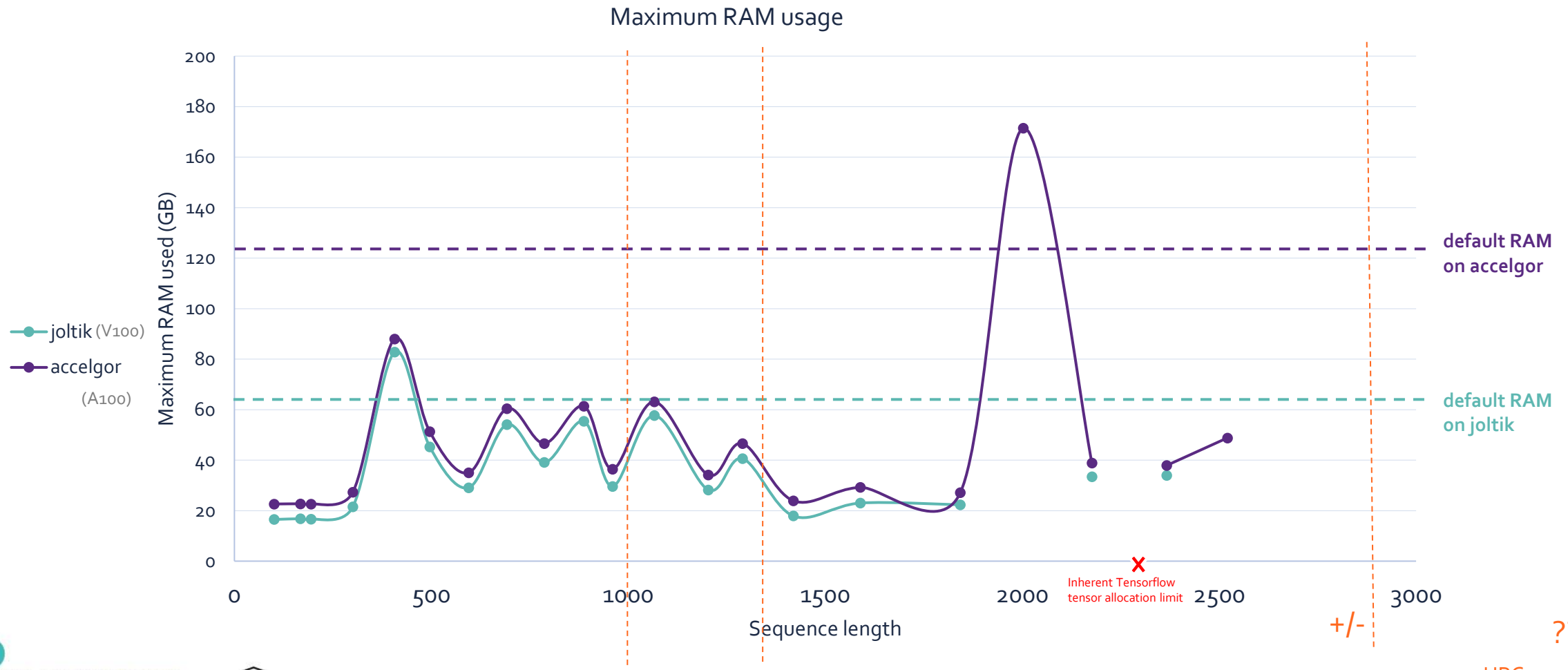
SCIENCE MEETS LIFE

AlphaFold on the HPC

AlphaFold resource usage

1. Sequence database search
   - → RAM-intensive
   - → High disk reading speed desirable
   - → Datasets ~2.2 TB storage
2. Deep learning prediction models
   - → GPU memory-intensive

Memory limits and time elapsed



Google Colab single GPU limits (12 or 16 GB)

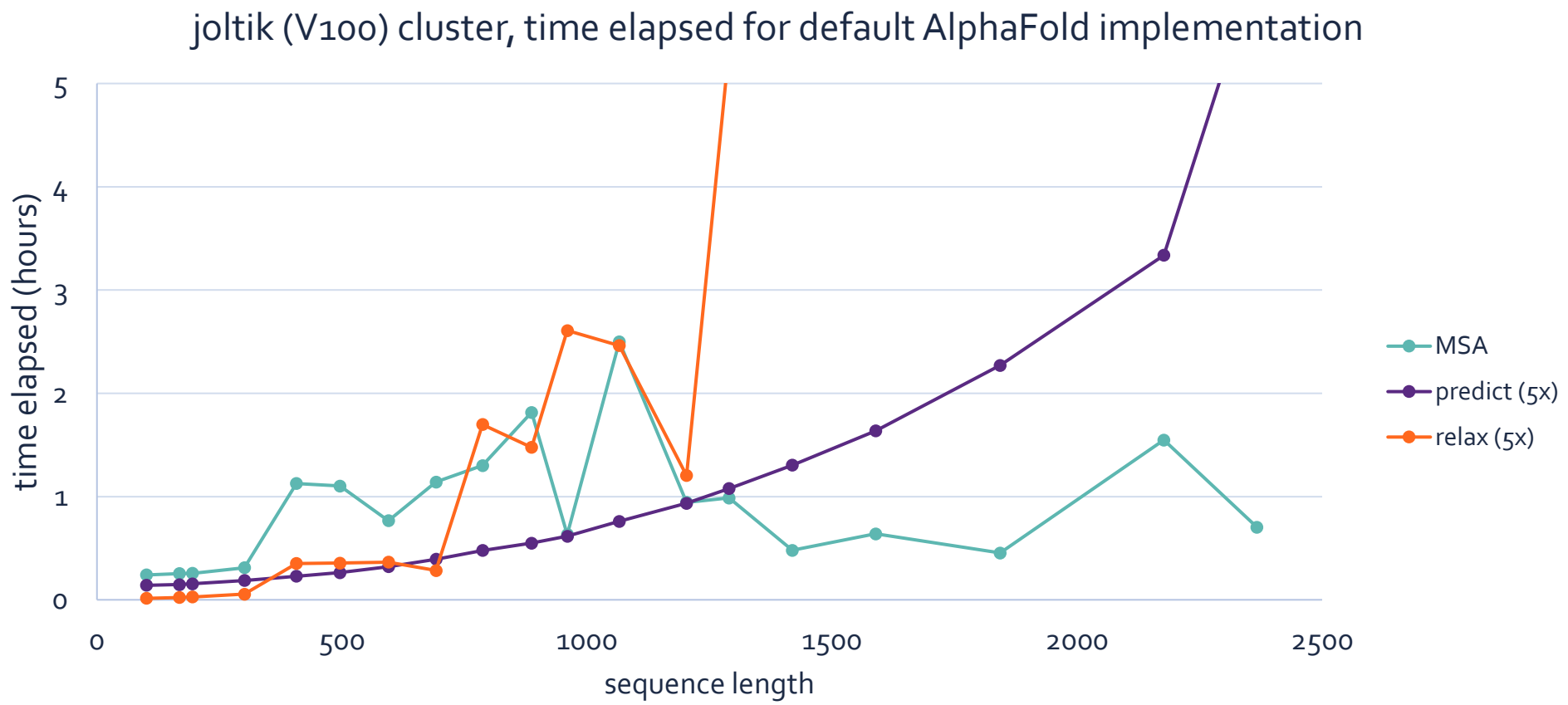
HPC joltik V100 single GPU limit (32 GB)

HPC accelgor A100 single GPU limit (80 GB)

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

Memory limits **and time elapsed**

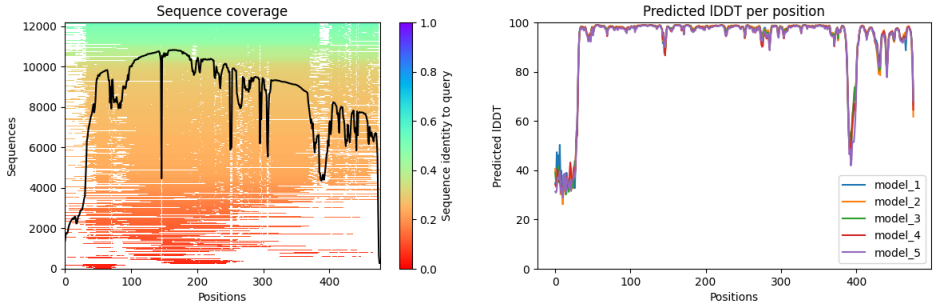


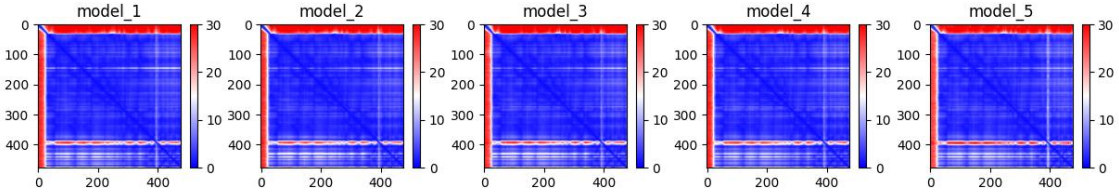
VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

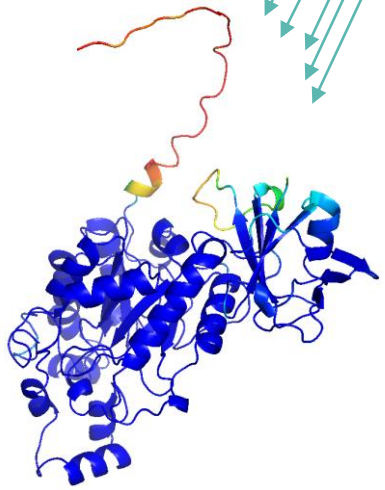
AlphaFold on HPC





4. profit!

- BCX803421_15000808.fasta
- BCX803421_15000808_coverage_lDDT.png
- BCX803421_15000808_PAE.png
- BCX803421_15000808_relaxed_model_1.pdb
- BCX803421_15000808_seq0.a3m
- BCX803421_15000808_unrelaxed_model_1.pdb
- BCX803421_15000808_unrelaxed_model_2.pdb
- BCX803421_15000808_unrelaxed_model_3.pdb
- BCX803421_15000808_unrelaxed_model_4.pdb
- BCX803421_15000808_unrelaxed_model_5.pdb
- info.log

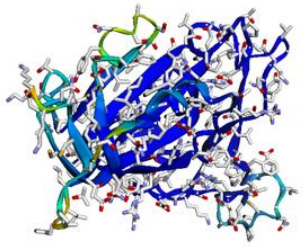


VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

SCIENCE MEETS LIFE

Example comparison of time required



Target: SARS-CoV-2 spike RBD

```
>rbd
ITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLC
FTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNY
LYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVV
VLSFELLHAPATVCGPKK
```

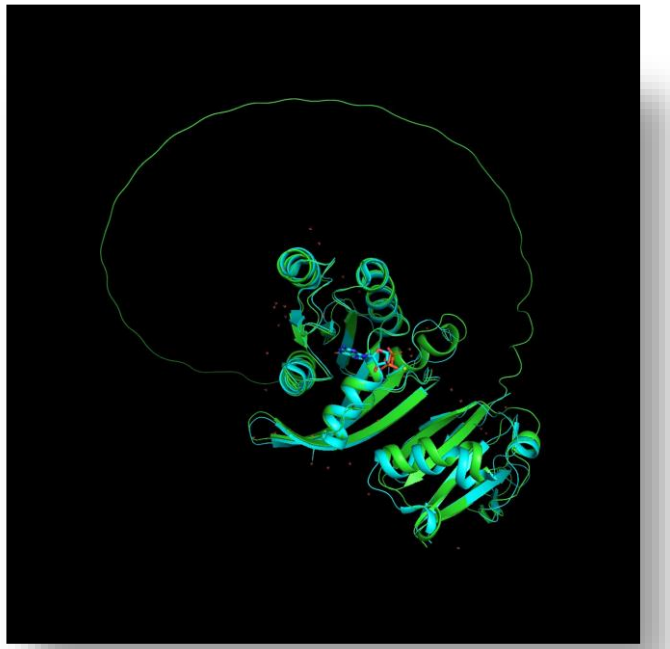
(198 residues)

accelgor (A100)

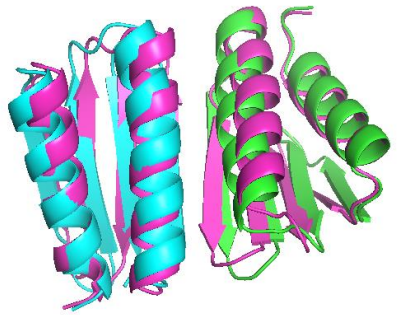
	Installation	MSA	Predict (5x)	Relax (1x/5x)	Total time elapsed
Colab: Official AlphaFold (v2.1)	9 min	38 min	48 min (relax 1x)		**1 hr 35 min**
HPC: Official AlphaFold (v2.0)	-	15 min	7 min	<1 min (5x)	**23 min**
HPC: *adapted script* (ColabFold modified version)	-	14 min	4 min	<1 min (1x)	**19 min**

What’s next?

Protein complex prediction



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

31 / 33

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

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

OpenFold

Mohammed AlQuraishi
@MoAlQuraishi

An announcement I've been aching to make! After much sweat, we've built a trainable version of AlphaFold2, implemented in PyTorch, which we're calling OpenFold.

GitHub: github.com/aqlaboratory/o...

Colab: colab.research.google.com/github/aqlabor...

Why a trainable version of AlphaFold2 you ask? ⬇️

4:57 PM · Nov 12, 2021 · Twitter Web App

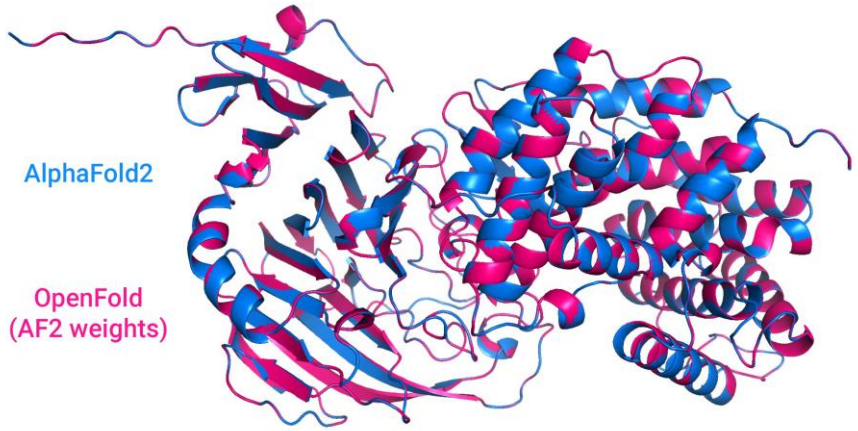
510 Retweets **46** Quote Tweets **2,052** Likes

Mohammed AlQuraishi @MoAlQuraishi · Nov 12, 2021

Replying to @MoAlQuraishi

As we saw with the recent AlphaFold-Multimer, some applications can benefit from training new AF2 variants and possibly integrating AF2 within larger models. DeepMind's JAX version, while excellent, is missing training code. PyTorch is also more widely used, hence OpenFold.

1 6 74



Single sequence prediction

bioRxiv, August 2021

Single-sequence protein structure prediction using language models from deep learning

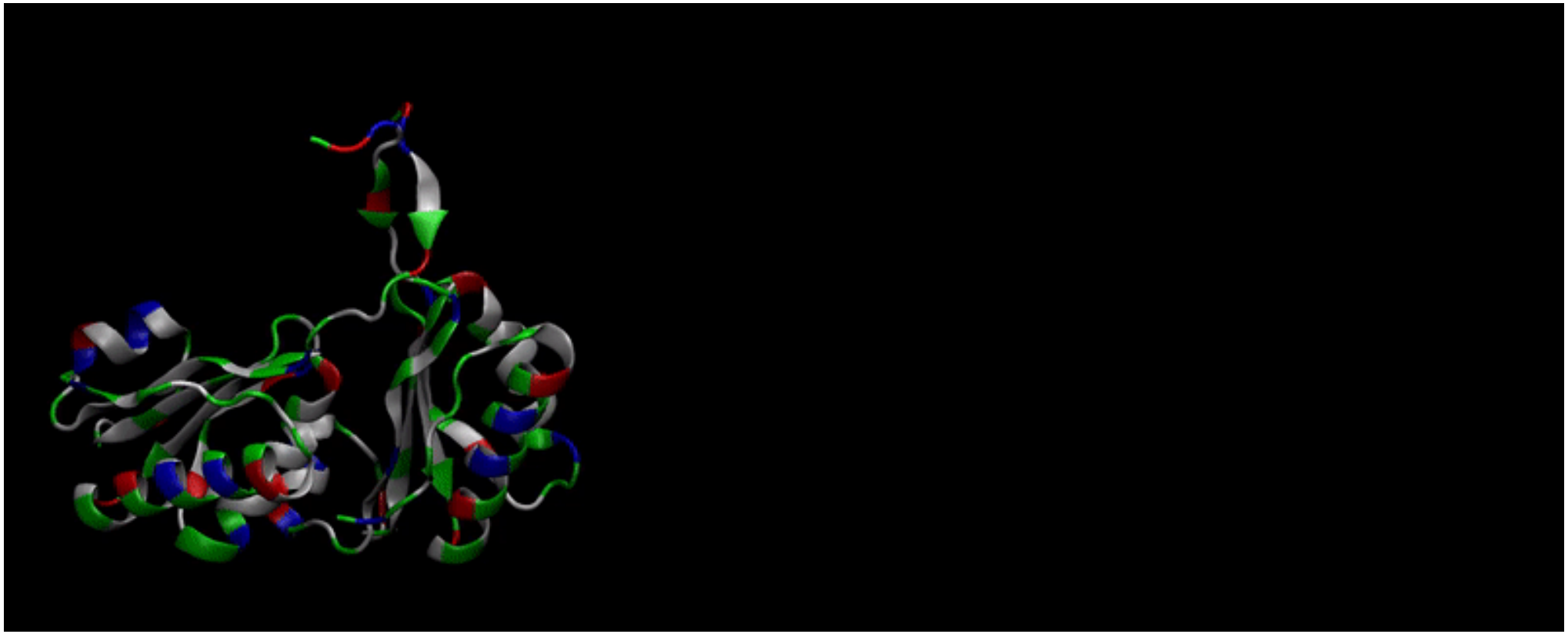
Ratul Chowdhury[a,*], Nazim Bouatta[a,*], Surojit Biswas[b,c,*], Charlotte Rochereau[d], George M. Church[a,b], Peter K. Sorger[a,†], and Mohammed AlQuraishi[a,e,†]

bioRxiv, January 2022

Single-sequence protein structure prediction using supervised transformer protein language models

Wenkai Wang[1], Zhenling Peng[2], Jianyi Yang[1,*]
[1]School of Mathematical Sciences, Nankai University, Tianjin 300071, China.
[2]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, 266237, China.

Molecular Dynamics



Source: https://2019.igem.org/Team:UANL/Simulations

Thank you for your attention.

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

Q&A

jasper.zuallaert@ugent.be