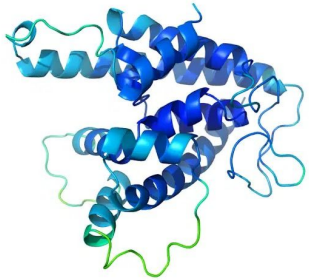
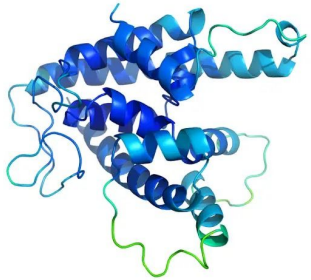


VLAAMS SUPERCOMPUTER CENTRUM
vscentrum.be

Vlaanderen
is supercomputing



Getting started with AlphaFold: installation and performance

Kenneth Hoste (HPC-UGent)

Jasper Zuallaert (VIB, UGent), Samuel Moors (HPC-VUB),

Carl Mensch (HPC-UA), Alexander Vapirev (HPC-KUL), Tim Jaenen (FWO)

1st VSC AlphaFold community meetup

30 March 2022

(source of 3D protein structure image: https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology)

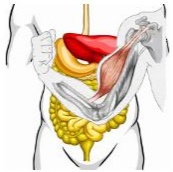
Agenda

- AlphaFold in a nutshell
- High-level overview of AlphaFold, version history, software dependencies
- Installing AlphaFold: alternatives and performance considerations
- Benchmarking AlphaFold with multiple different input sequences
- Performance aspects of running AlphaFold: CPU vs GPU, I/O, containers, …
- Beyond AlphaFold: RoseTTAFold, OpenFold, …
- VSC resources for running AlphaFold

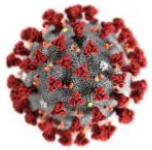
Proteins: the building blocks of life

phenotypical traits
(e.g. eye color)

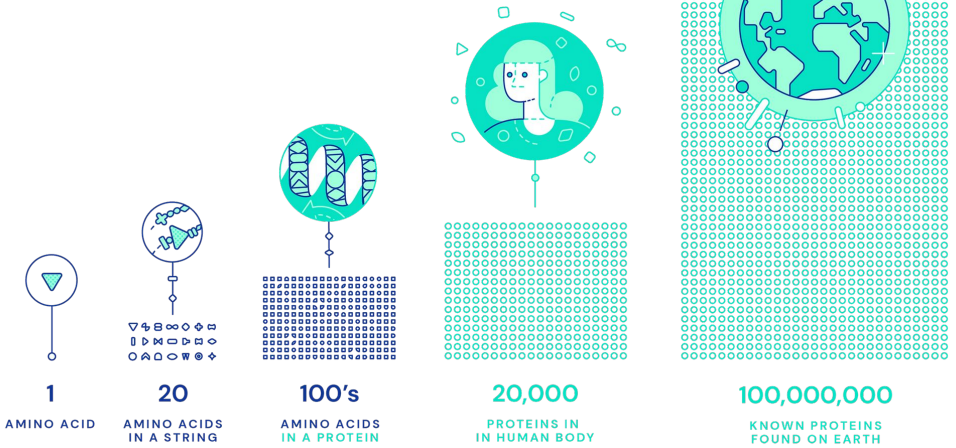
muscle contraction



food digestion



immune responses

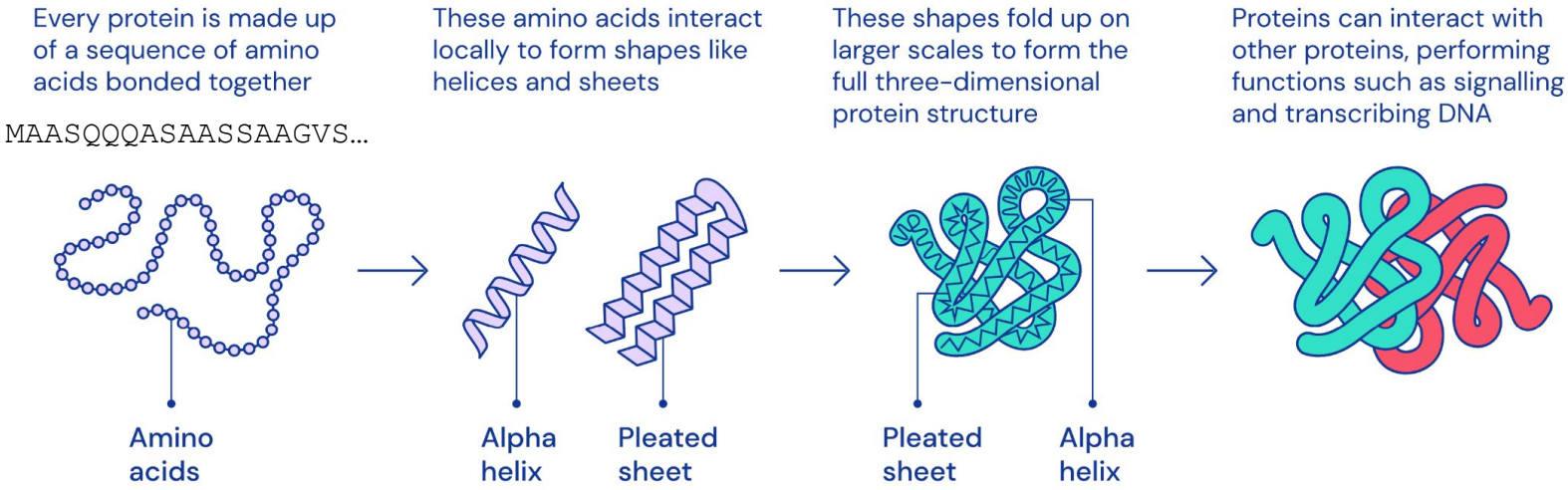


3D structure of a protein (how it "folds") determines how it works, what it does, …

~100 million known distinct proteins, but 3D structure is only known for small subset…

(source of image: https://www.deepmind.com/research/highlighted-research/alphafold)

AlphaFold in a nutshell: protein structure *prediction*

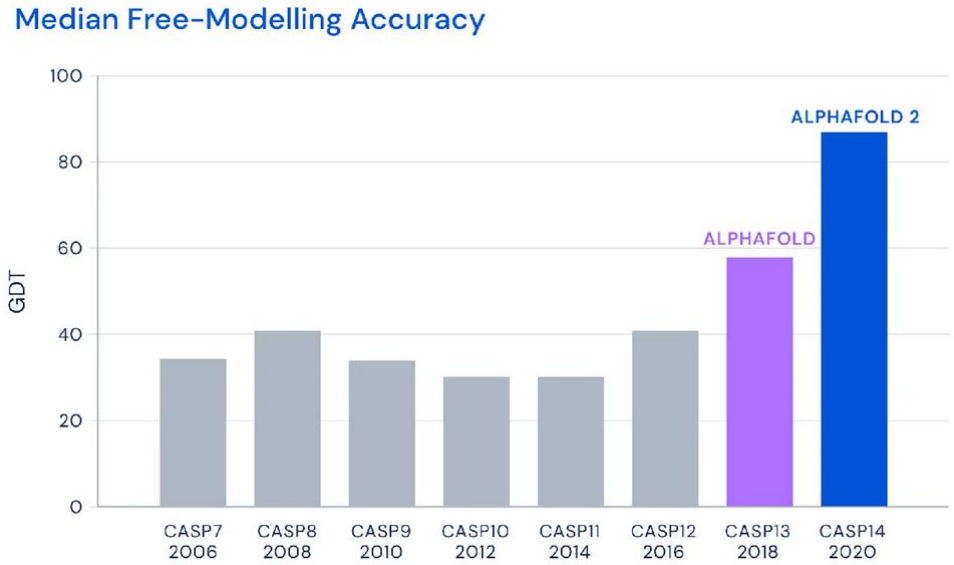


(image source: https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery)

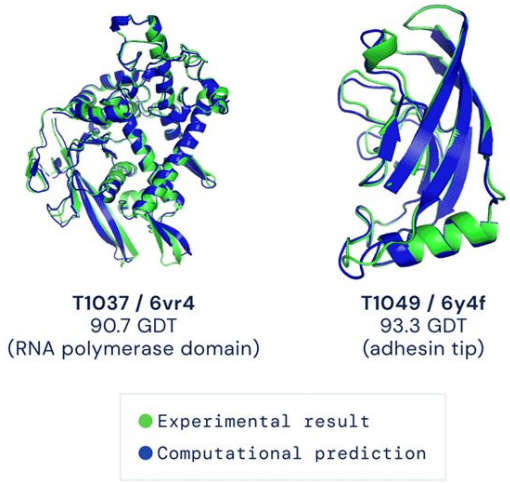
See also "*AlphaFold: a new age for protein folding*" talk
at 7th EasyBuild User Meeting: https://easybuild.io/eum22/#alphafold

Critical Assessment of protein Structure Prediction (CASP14, 2020)

In 2020, the AlphaFold 2 artificial intelligence (AI) program won the bi-annual CASP “competition”.



AlphaFold 2 achieved a GDT-TS with a median of 92.4 across all targets, and 87.0 in free-modelling accuracy.



AlphaFold: a solution to a 50-year-old grand challenge in biology

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

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*.

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

The Telegraph

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

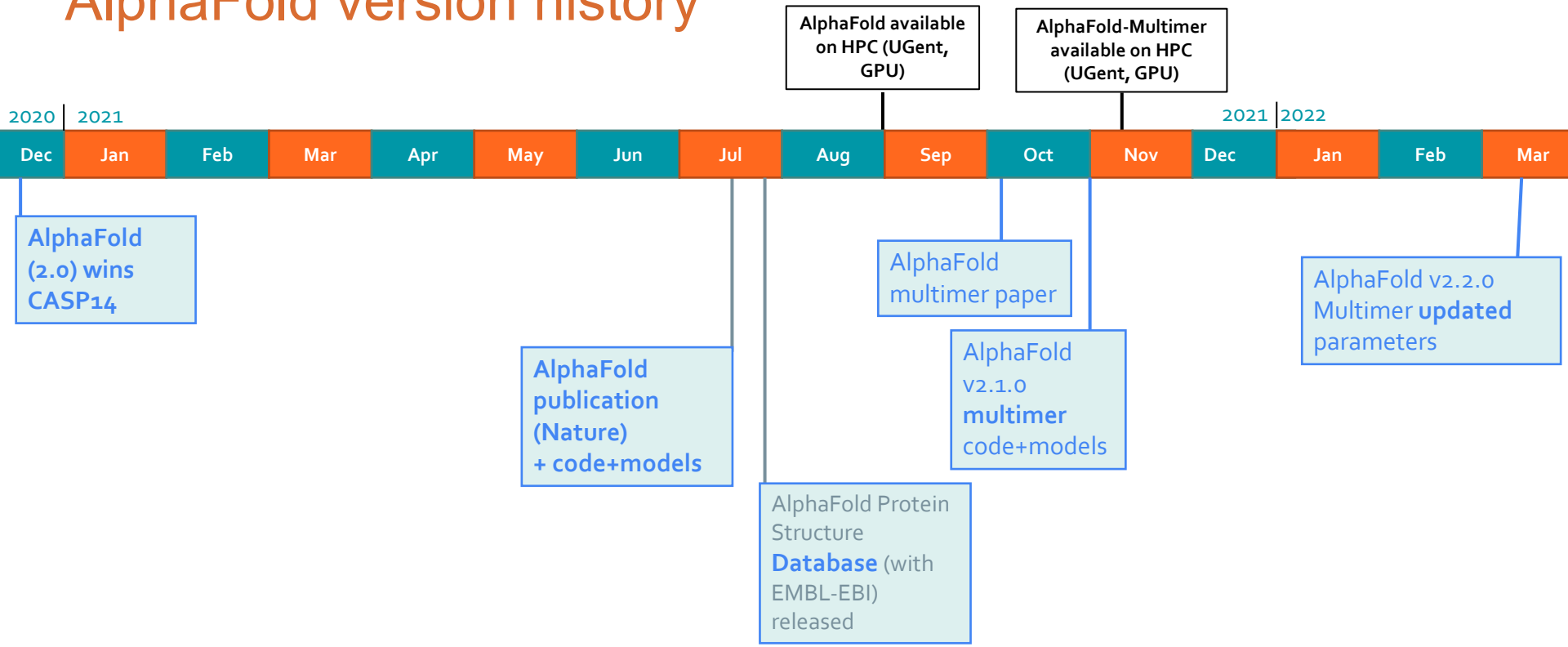
The development could 'significantly accelerate' drug development for cancer and other diseases

"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."

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."

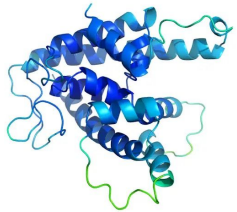
But AlphaFold had no trouble. It returned a detailed image of a three-part protein with two long helical arms in the middle. The model enabled Lupas and his colleagues to make sense of their x-ray data; within half an hour, they had fit their experimental results to AlphaFold's predicted structure. "It's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

AlphaFold version history



Kenneth Hoste (HPC-UGent), Jasper Zuallaert (VIB, UGent), Samuel Moors (HPC-VUB), Carl Mensch (HPC-UA), Alexander Vapirev (HPC-KUL), Tim Jaenen (FWO)

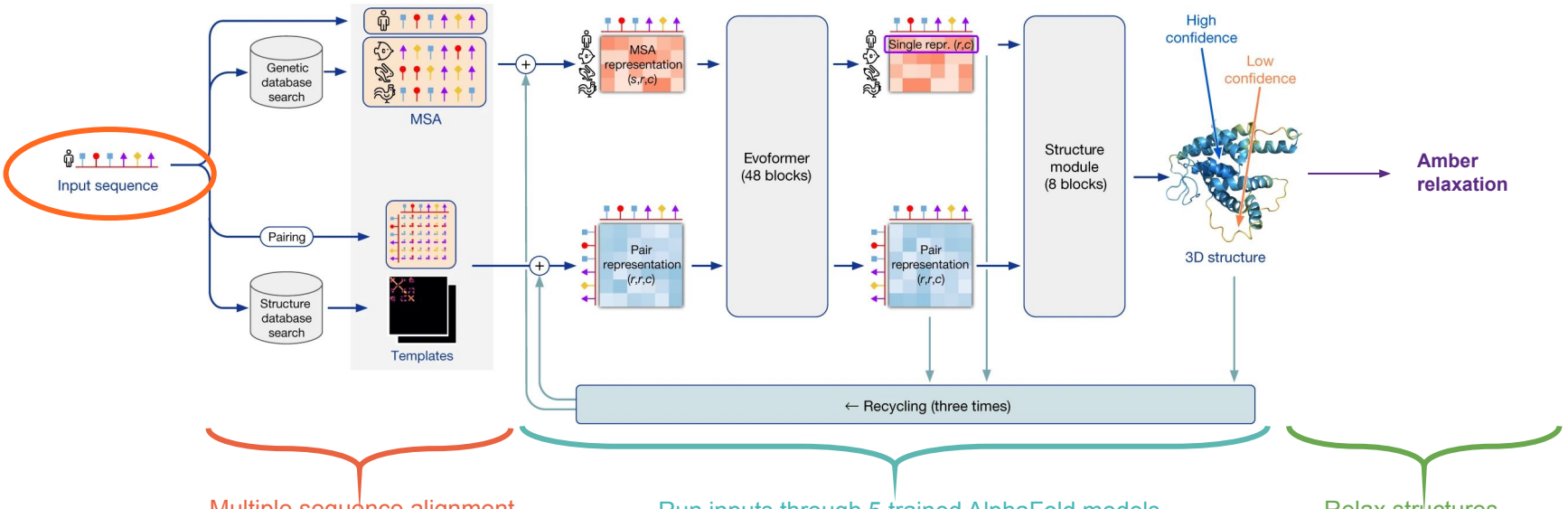
High-level overview of AlphaFold 2



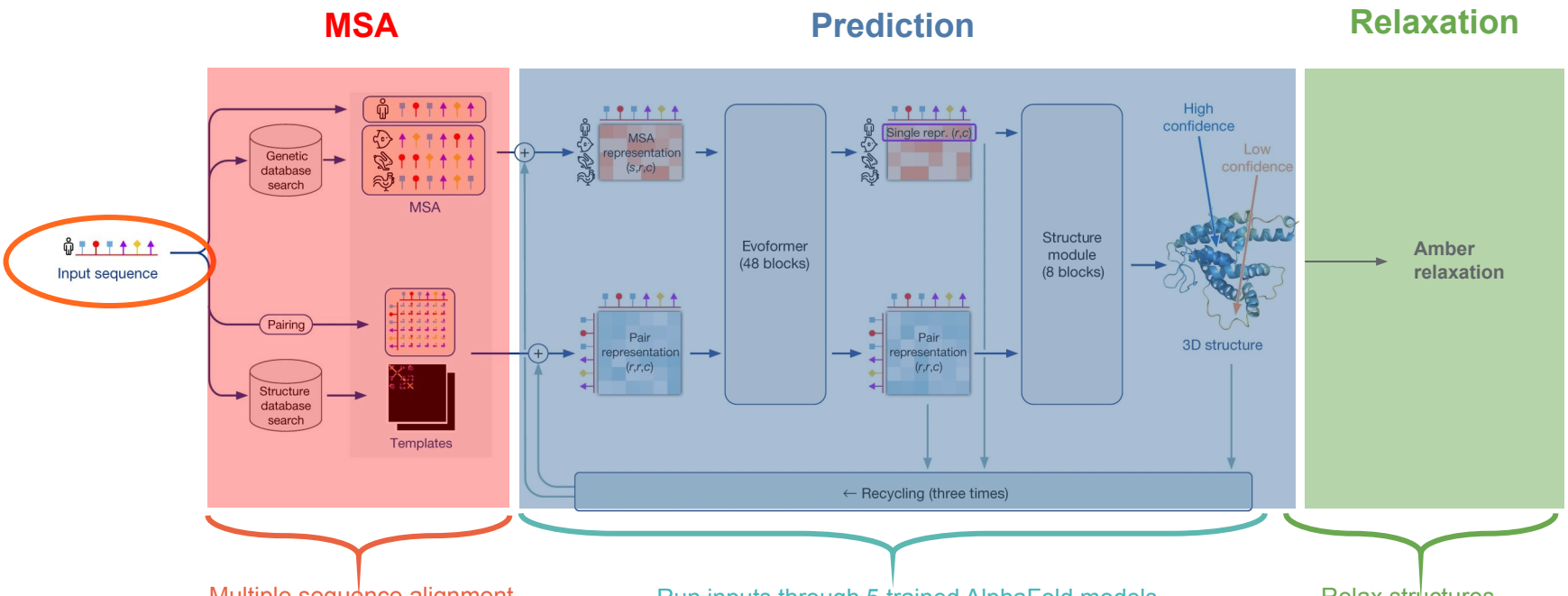
- AlphaFold 2 implements a 3-step process:
  - Multiple Sequence Alignment (MSA) to look for “similar” proteins in collection of databases
  - Structure prediction using *pre-trained* **neural networks**
  - Relaxation to minimize energy of predicted structures (optional)
- Requires **large database** of proteins with known 3D folding structure (~2.2TB on disk)
- Process is both **compute and I/O intensive**, large speedup with GPUs and SSDs
- Software is available under Apache 2 open source license: https://github.com/deepmind/alphafold
- Parameters for pre-trained models available under CC BY 4.0 license

(source of 3D protein structure image: https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology)

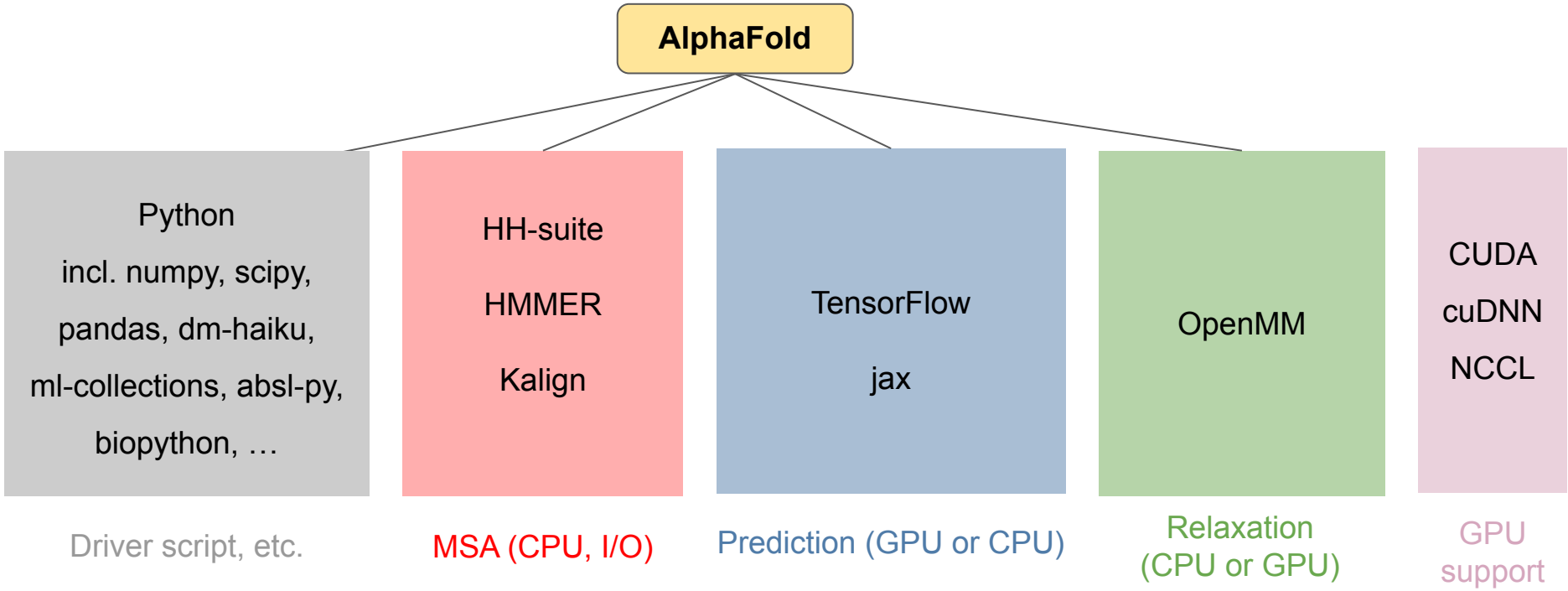
High-level overview of AlphaFold 2



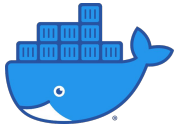
High-level overview of AlphaFold 2



AlphaFold software dependencies



Getting AlphaFold installed



Option 1: using a (pre-built) **container**

- Docker file provided by DeepMind: https://github.com/deepmind/alphafold#first-time-setup
- Pre-built container images available on Docker Hub: https://hub.docker.com/r/catgumag/alphafold
- Can also be run via Singularity/Apptainer (HPC-friendly): https://github.com/hyoo/alphafold_singularity

```
singularity exec --nv alphafold.sif  bash -c "cd /opt/alphafold/; ./run.sh …"
```

- Pros: quick to “install” and get started (~15min)
- Cons:
  - **Can be significantly slower** due to use of *generic* binaries (not optimized for “your” hardware)
  - When using pre-built container images, there may be trust issues

Getting AlphaFold installed

Option 2: install AlphaFold + all required dependencies **from source code** (where possible)

- Time-consuming and tedious process, especially if done manually and if you're not used to doing this…
- **Made easy via EasyBuild + environment modules**, no admin privileges required:

https://easybuild.io

```
pip3 install easybuild
eb AlphaFold-2.1.2-foss-2021a-CUDA-11.3.1.eb --robot
module load AlphaFold/2.1.2-foss-2021a-CUDA-11.3.1
alphafold ...
```

- EasyBuild was created by HPC-UGent and VSC, now a worldwide community of "installation experts"
- Pros: resulting AlphaFold installation will be **significantly faster** (optimized for "your" hardware)
- Cons: installation process takes a while when done from scratch (hours)

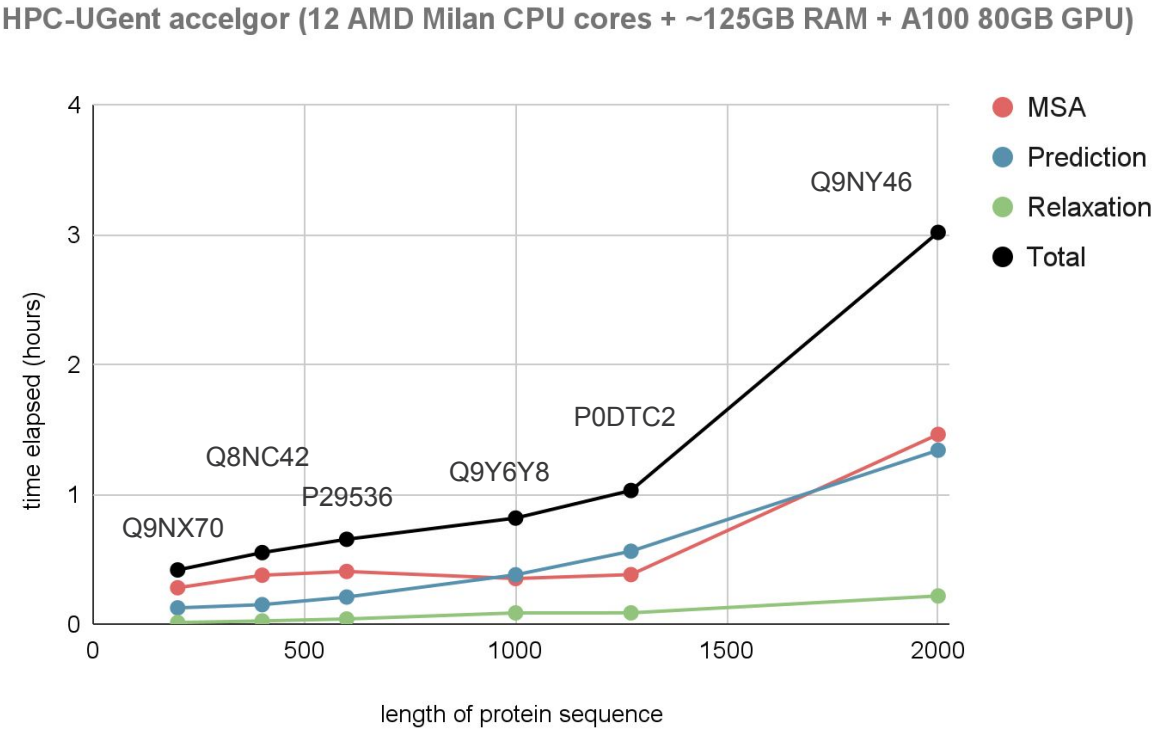
Benchmarking AlphaFold

- Time-to-result (performance) of **AlphaFold v2.1.2** was evaluated
- Using set of input sequences of different length:

Q9NX70 (200 amino acids)	Q8NC42 (400 amino acids)
P29536 (600 amino acids)	Q9Y6Y8 (1000 amino acids)
P0DTC2 - SARS-CoV-2 (1273 amino acids)	Q9NY46 (2000 amino acids)

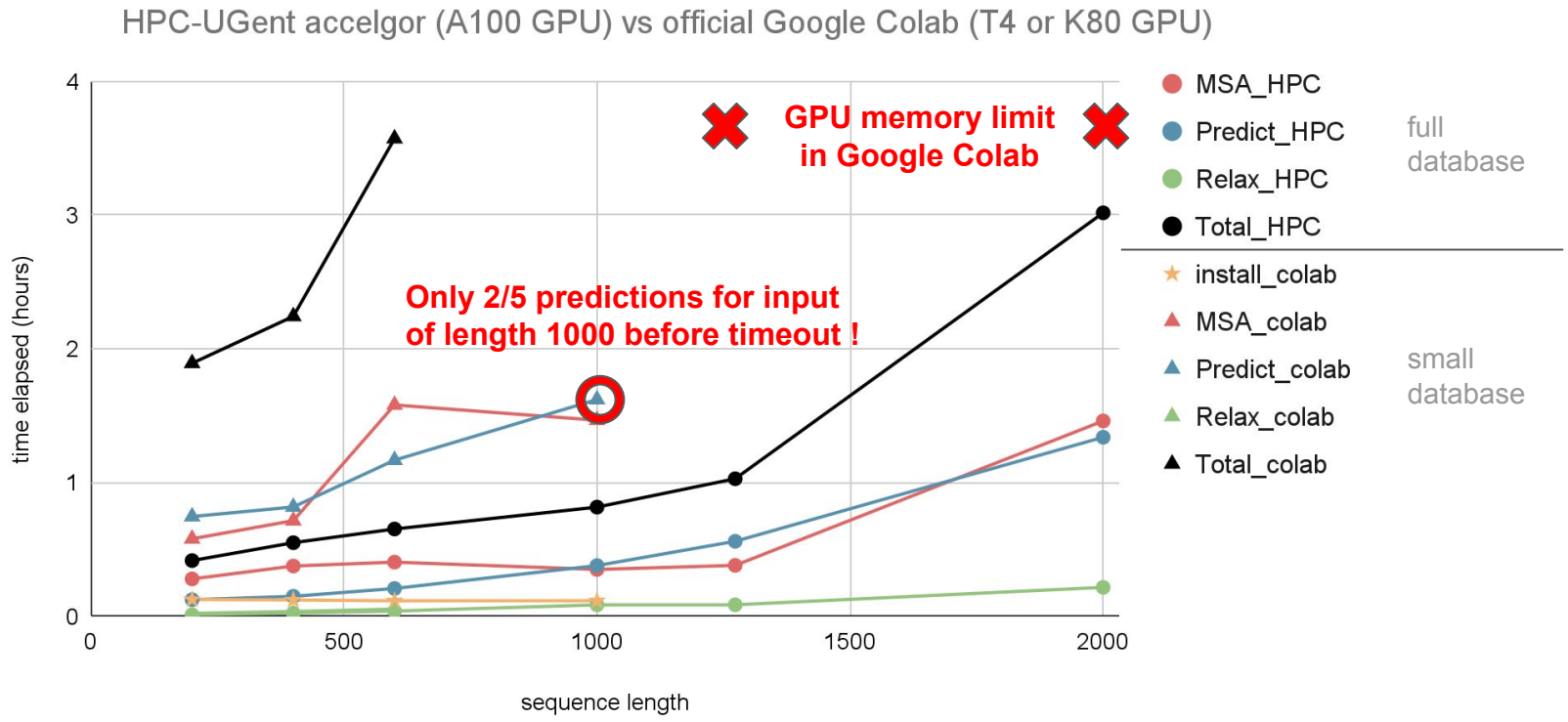
- Selected sequences + scripts are available at https://github.com/vscentrum/vsc-alphafold
- Impact of different aspects regarding running AlphaFold was assessed…

Benchmarking AlphaFold: length of protein sequence

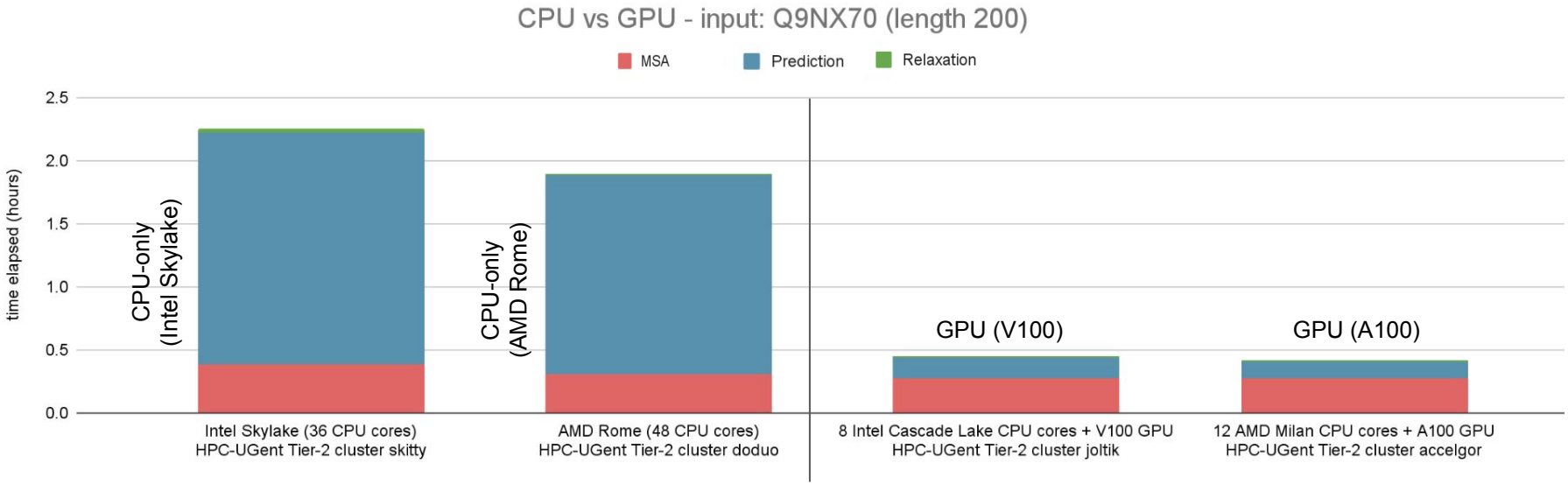


Kenneth Hoste (HPC-UGent), Jasper Zuallaert (VIB, UGent), Samuel Moors (HPC-VUB), Carl Mensch (HPC-UA), Alexander Vapirev (HPC-KUL), Tim Jaenen (FWO)

Benchmarking AlphaFold: HPC cluster vs Google Colab

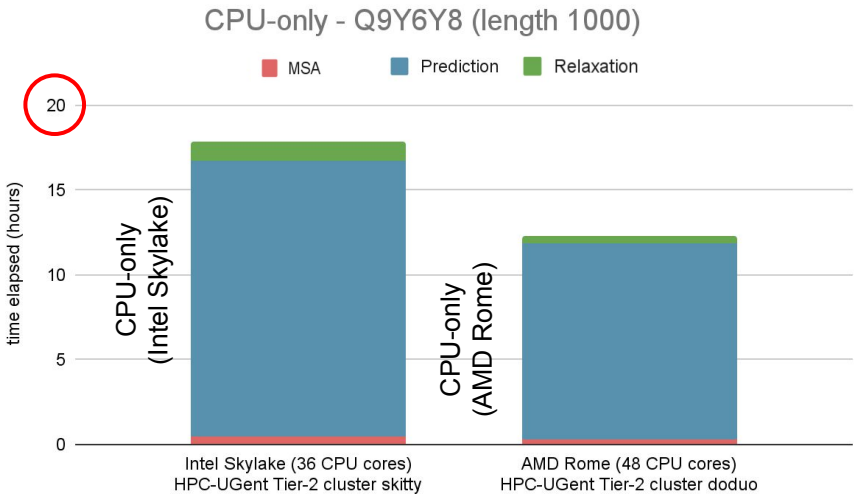


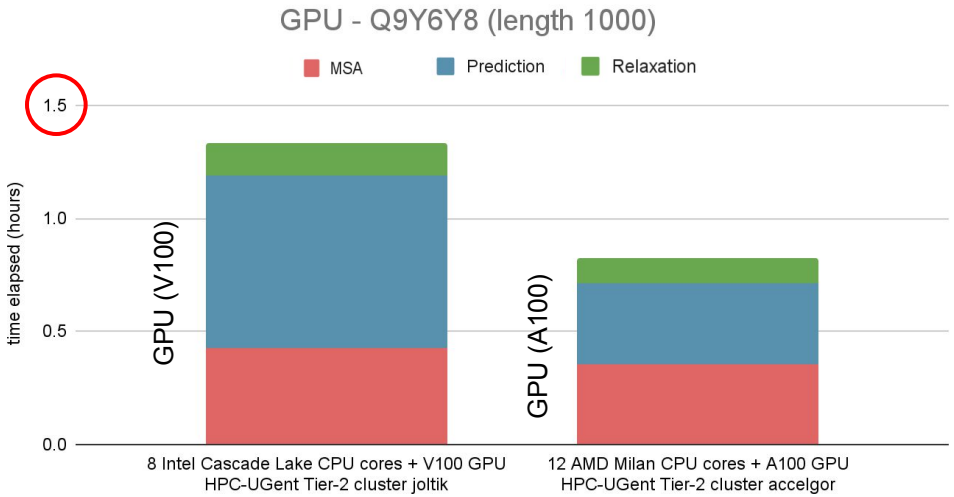
Benchmarking AlphaFold: CPU versus GPU



~5x faster using a (powerful) GPU compared to CPU-only
for short input sequences (length 200)

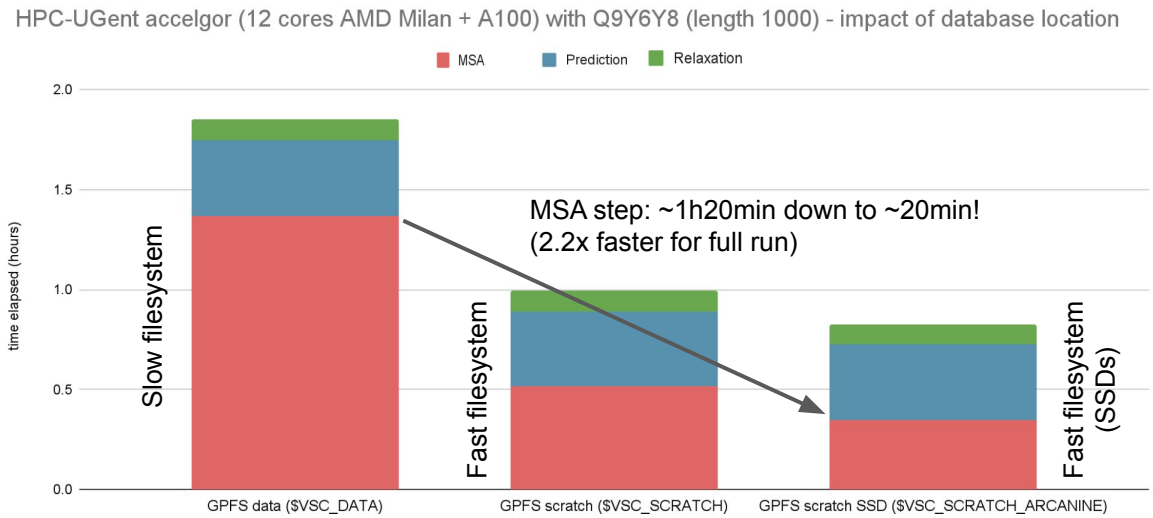
Benchmarking AlphaFold: CPU versus GPU





15-20x faster using a (powerful) GPU compared to CPU-only
for longer input sequences (length 1000)

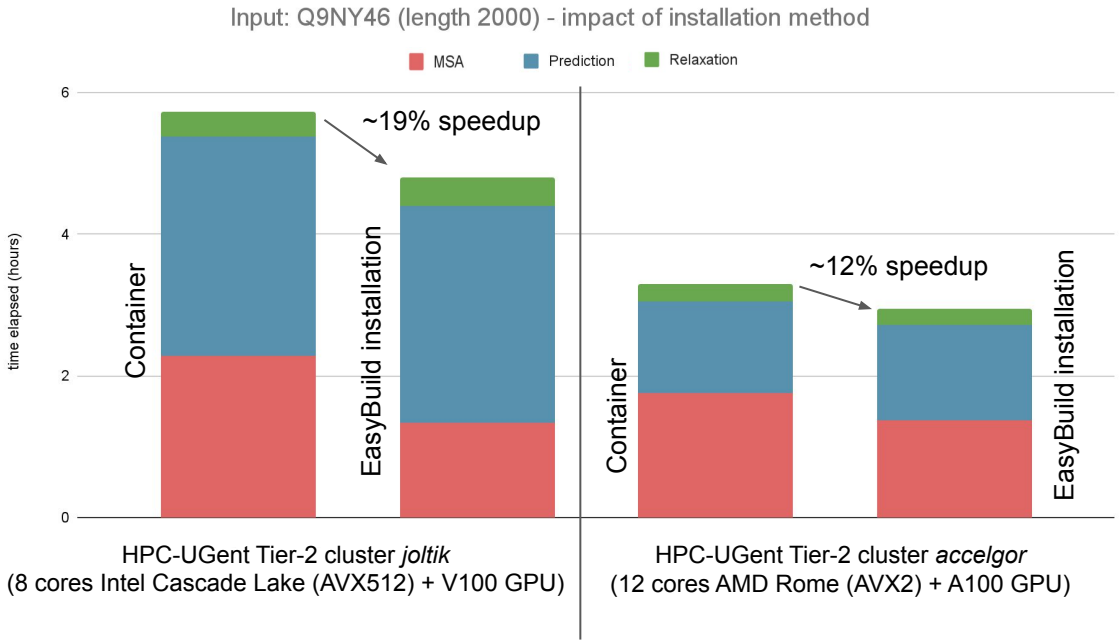
Benchmarking AlphaFold: location of databases (I/O)



Location of database has *big* impact on performance of MSA step!

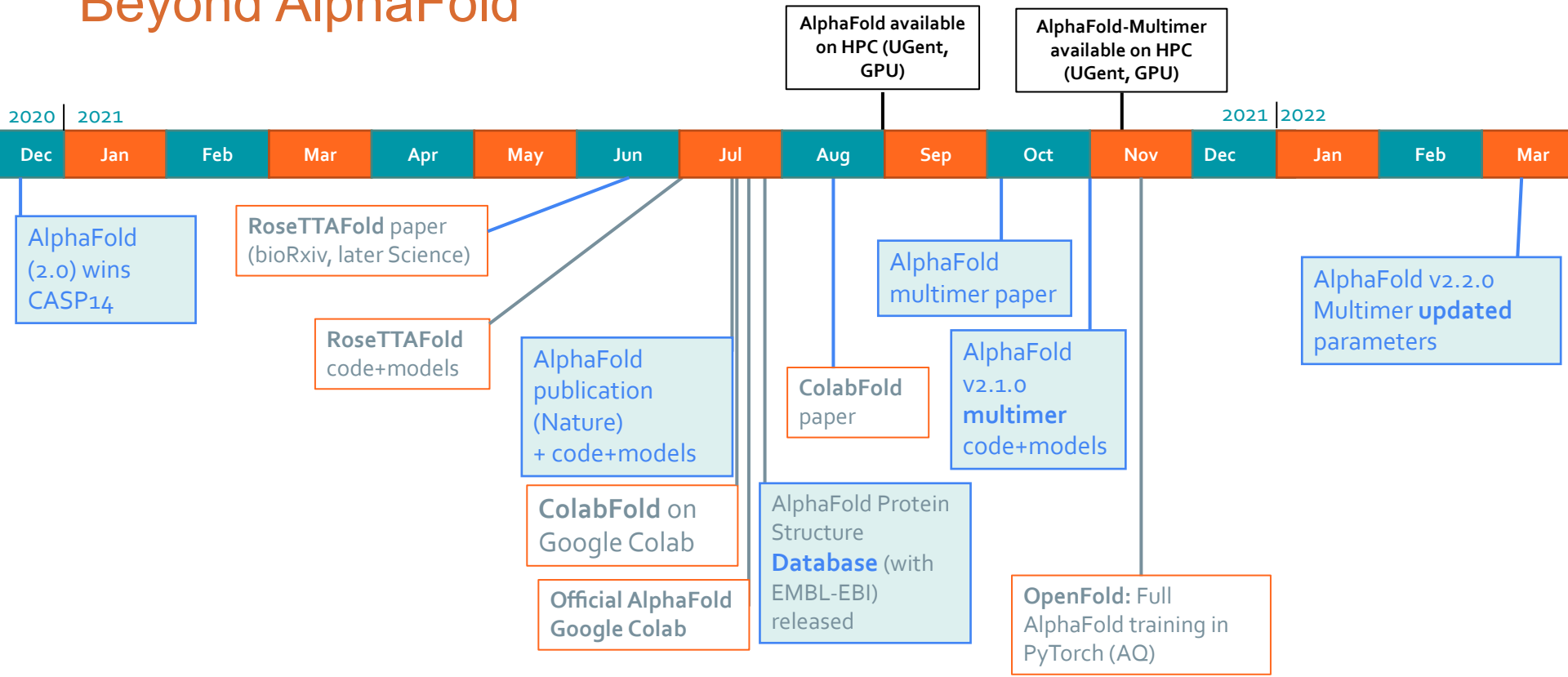
MSA step is **~4x faster** on shared scratch SSD filesystem compared to shared data filesystem,
50% faster when using scratch SSD filesystem compared to standard scratch filesystem.

Benchmarking AlphaFold: container vs EasyBuild installation



~10-20% faster runtime for long input sequences when using installation that is optimized for hardware on which it is used (mostly in CPU-only MSA step)

Beyond AlphaFold



Beyond AlphaFold - ColabFold

Adds more customizability to AlphaFold runs (MSA search, # of recycles, …)

Runs in Google Colab

Faster MSA search via MMSeqs2, either via

- dedicated webserver
- local installation of databases

Input: P29536 (length 600)		
	MSA search	Predict (5 models)
Official AlphaFold Google Colab	~ 1 hr 35 min	~ 1 hr 10 min
ColabFold	< 1 min	~ 1 hr 24 min
HPC-UGent Tier-2 accelgor	~ 20 min.	~ 12 min.

https://github.com/sokrypton/ColabFold

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

Beyond AlphaFold - RoseTTAFold

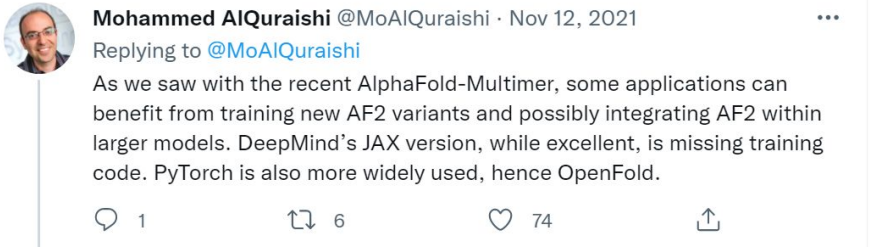
https://github.com/RosettaCommons/RoseTTAFold

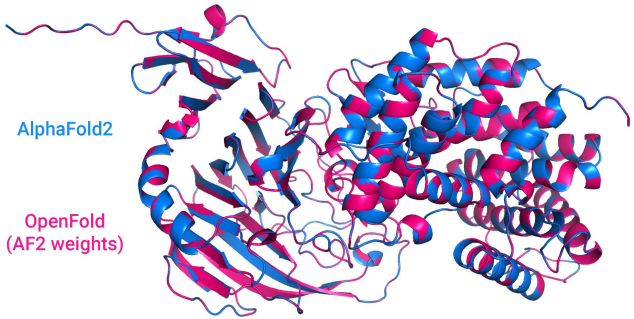
- Analogous, concurring deep learning approach for protein folding
- Baker lab (University of Washington), long-standing protein folding history
- Runners-up of CASP14, now updated with AlphaFold2-inspired ideas

M. Baek, et al., Accurate prediction of protein structures and interactions using a three-track neural network, Science (2021)

Beyond AlphaFold - OpenFold

https://github.com/aqlaboratory/openfold





OpenFold uses PyTorch rather than TensorFlow, allows training own models rather than using pre-trained models (very compute-intensive!)

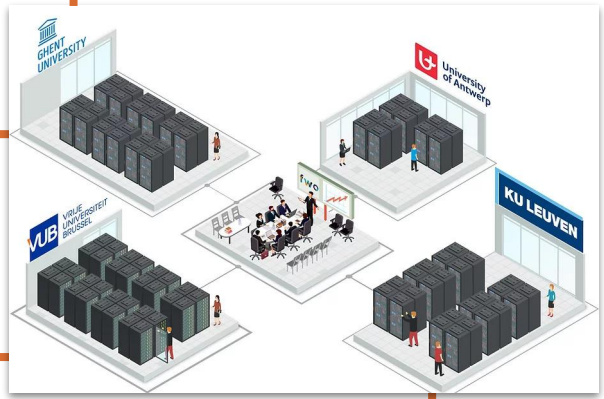
HPC resources available at VSC hubs

UGent Tier-2 clusters 'Joltik + 'Accelgor'

10 nodes: 32 CPU cores Intel Cascade Lake, ~250GB RAM, 4x NVIDIA V100 (32GB)

9 nodes: 48 CPU cores AMD Milan, ~500GB RAM, 4x NVIDIA A100 (80GB)

Fast shared scratch powered by SSDs



VUB Tier-2 cluster 'Hydra'

4 nodes: 24 CPU cores Intel Broadwell, ~250GB RAM, 2x NVIDIA P100 GPUs (16GB)

6 nodes: 32 CPU cores AMD Rome, ~250GB RAM, 2x NVIDIA A100 GPUs (40GB)

KU Leuven Tier-2 cluster 'Genius'

20 nodes: 36 CPU cores Intel Skylake, 192GB RAM, 4x NVIDIA P100 GPUs (16GB)

2 nodes, 36 CPU cores Intel Cascade Lake, 768GB RAM, 8x NVIDIA V100 GPUs (32GB)

Tier-1 cluster 'Hortense' (phase 1)

20 nodes: 48 CPU cores AMD Rome, ~256GB RAM, 4x NVIDIA A100 (40GB)

Beyond the VSC: EuroHPC Tier-0 supercomputer LUMI

https://lumi-supercomputer.eu

LUMI

LUMI, the Queen of the North

LUMI is a Tier-0 ***GPU-accelerated supercomputer*** *that enables the convergence of* ***high-performance computing***, ***artificial intelligence***, *and* ***high-performance data analytics.***

- Supplementary CPU partition
- ~200,000 AMD EPYC CPU cores

Possibility for combining different resources within a single run. HPE Slingshot technology.

30 PB encrypted object storage (Ceph) for storing, sharing and staging data

LUMI-G: GPU Partition

LUMI-C: x86 Partition

LUMI-D: Data Analytics Partition

LUMI-K: Container Cloud Service

High-speed interconnect

LUMI-F: Accelerated Storage

LUMI-Q: Emerging tech

LUMI-O: Object Storage Service

LUMI-P: Lustre Storage

Tier-0 GPU partition: over 550 Pflop/s powered by AMD Instinct GPUs

Interactive partition with 32 TB of memory and graphics GPUs for data analytics and visualization

7 PB Flash-based storage layer with extreme I/O bandwidth of 2 TB/s and IOPS capability. Cray ClusterStor E1000.

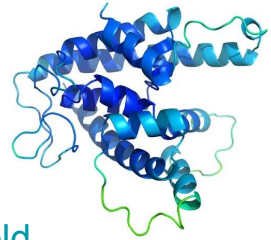
80 PB parallel file system

www.lumi-supercomputer.eu #lumisupercomputer #lumieurohpc

VSC AlphaFold community sessions

- Goal: bring different parties (academia + industry) together who are interested in AlphaFold
- To exchange ideas, experiences, and expertise on using AlphaFold (and beyond)
- Multiple sessions per year, focus on a specific topic
- Next session: **Introduction to Protein Structure Analysis and AlphaFold** (28-29 April 2022)
  2-day training session, incl. hands-on exercises with AlphaFold on VSC HPC infrastructure
  https://training.vib.be/all-trainings/introduction-protein-structure-analysis-and-alphafold
  Contact FWO (tim.jaenen@fwo.be) if you would like to attend the second session free of cost!

Questions are welcome!



Useful resources:

- AlphaFold website: https://www.deepmind.com/research/highlighted-research/alphafold
- Benchmark input sequences and scripts: https://github.com/vscentrum/vsc-alphafold
- Jasper's AlphaFold talk at EasyBuild User Meeting 2022: https://easybuild.io/eum22/#alphafold
- AlphaFold course by VIB on VSC resources: https://elearning.bits.vib.be/courses/alphafold
- **Flemish Supercomputing Centre (VSC)**
  - Website: https://www.vscentrum.be
  - Overview of HPC resources at VSC hubs: https://docs.vscentrum.be/en/latest/hardware.html
  - Contact: compute@vscentrum.be