Running OpenFold on the VSC infrastructure

17 February 2023

Carl Mensch
NCC Belgium
VSC-UAntwerp

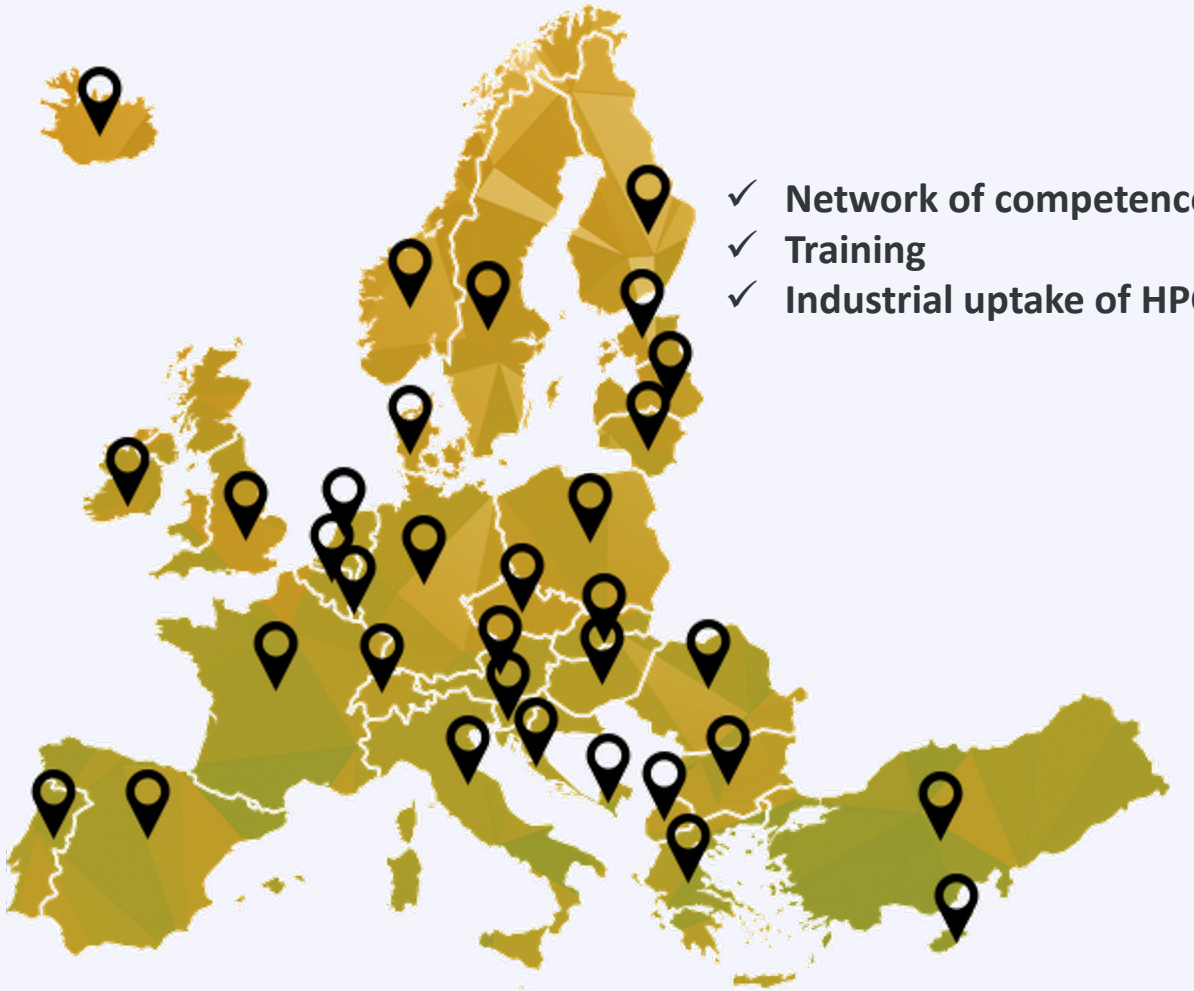
Kenneth Hoste (VSC-UGent), Samuel Moors (VSC-VUB)

VLAAMS SUPERCOMPUTER CENTRUM | *Innovative Computing for a Smarter Flanders*

EURO BELGIUM

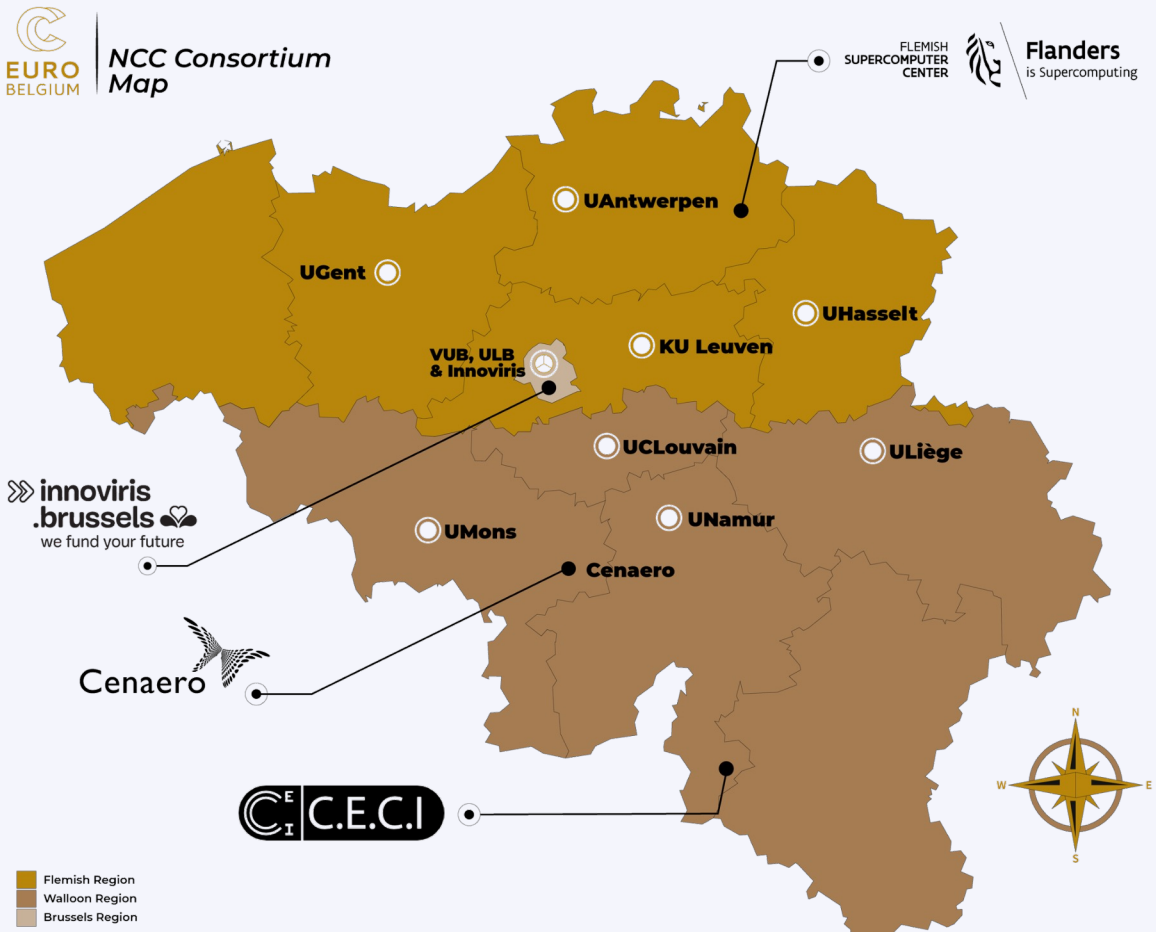
Supported by
fwo
EuroHPC Joint Undertaking

EuroCC: Supercomputing in Belgium



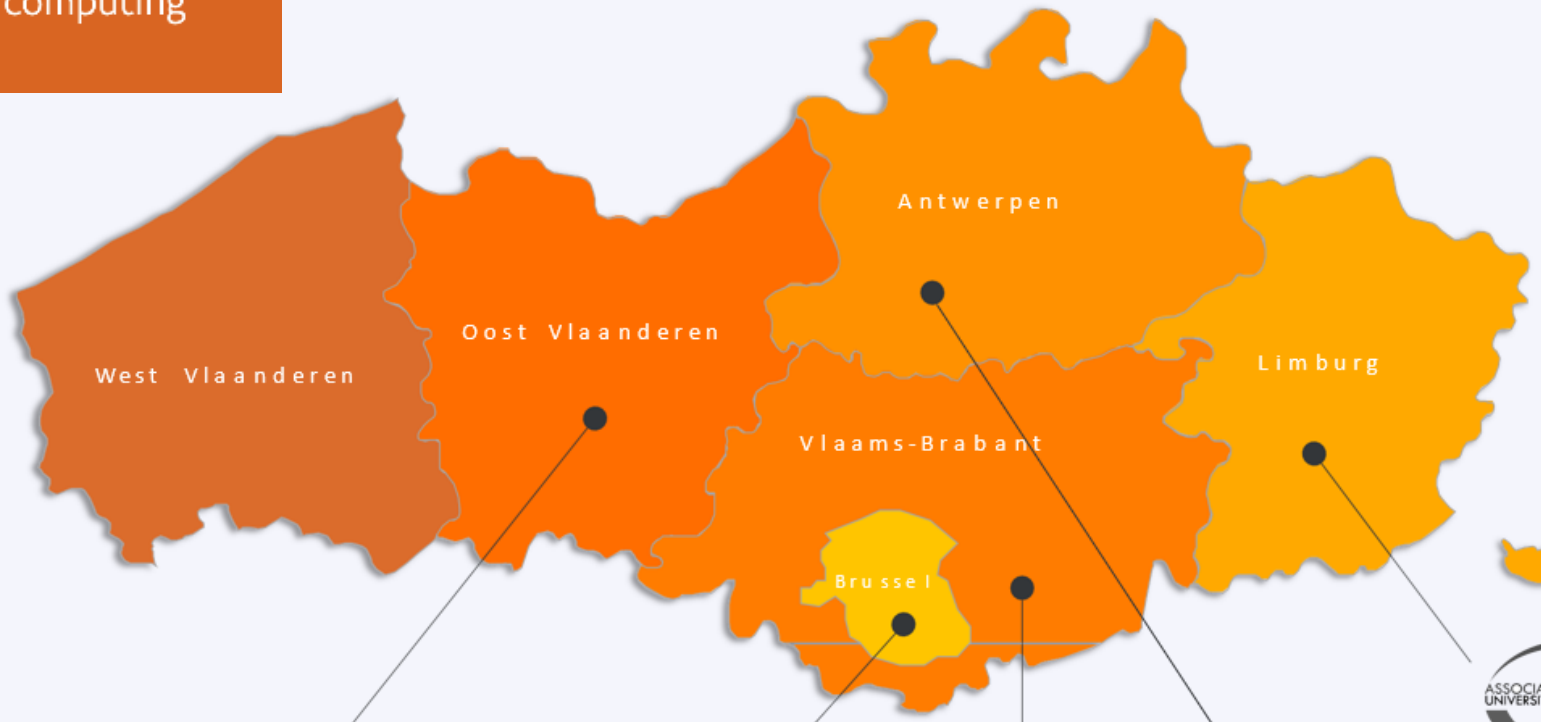
- ✓ Network of competences
- ✓ Training
- ✓ Industrial uptake of HPC

National competence centre (NCC) Belgium



EURO BELGIUM
VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen
is supercomputing



Supported by

fwo

ASSOCIATIE UNIVERSITEIT GENT

universitaire associatie BRUSSEL

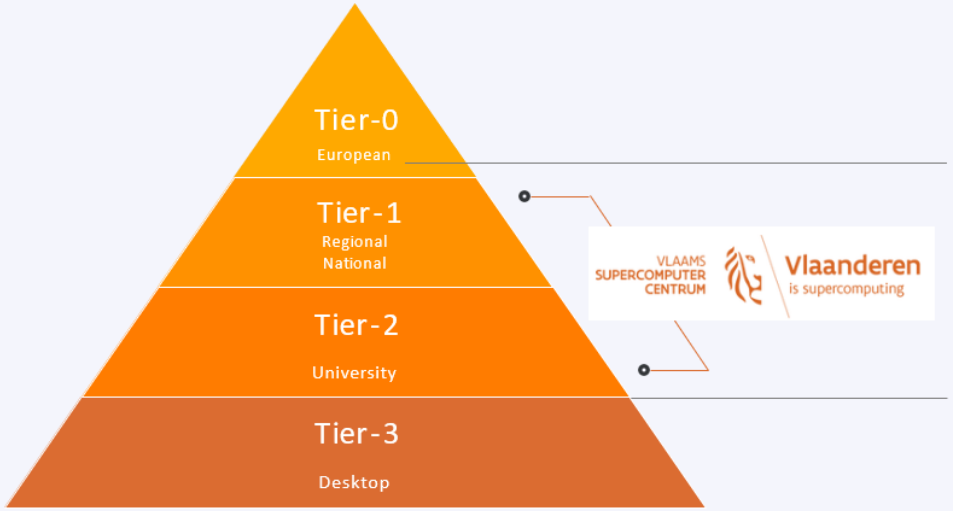
ASSOCIATIE KU LEUVEN

ASSOCIATIE UNIVERSITEIT-HOGESCHOLEN LIMBURG

auha antwerp university association

EURO BELGIUM

VLAAMS SUPERCOMPUTER CENTRUM



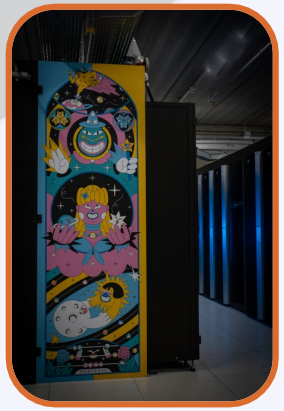
EuroHPC Joint Undertaking

LUMI

Tier-0

PRACE

Tier-1



Hortense - UGent

Tier-2

GHENT UNIVERSITY

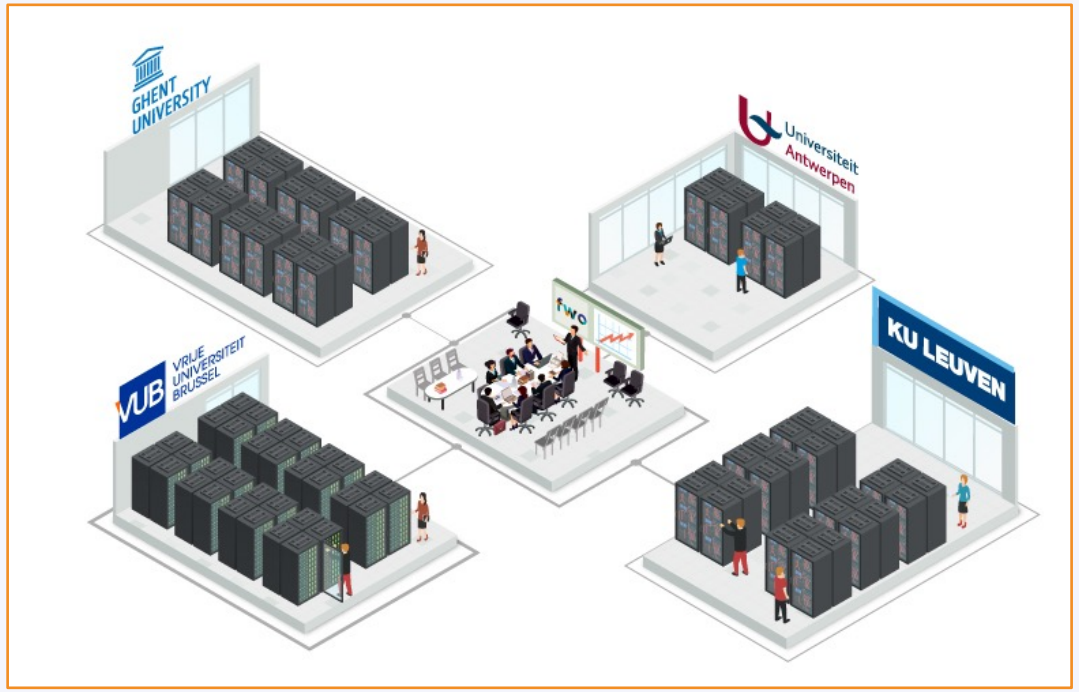
VUB VRIJE UNIVERSITEIT BRUSSEL

Universiteit Antwerpen

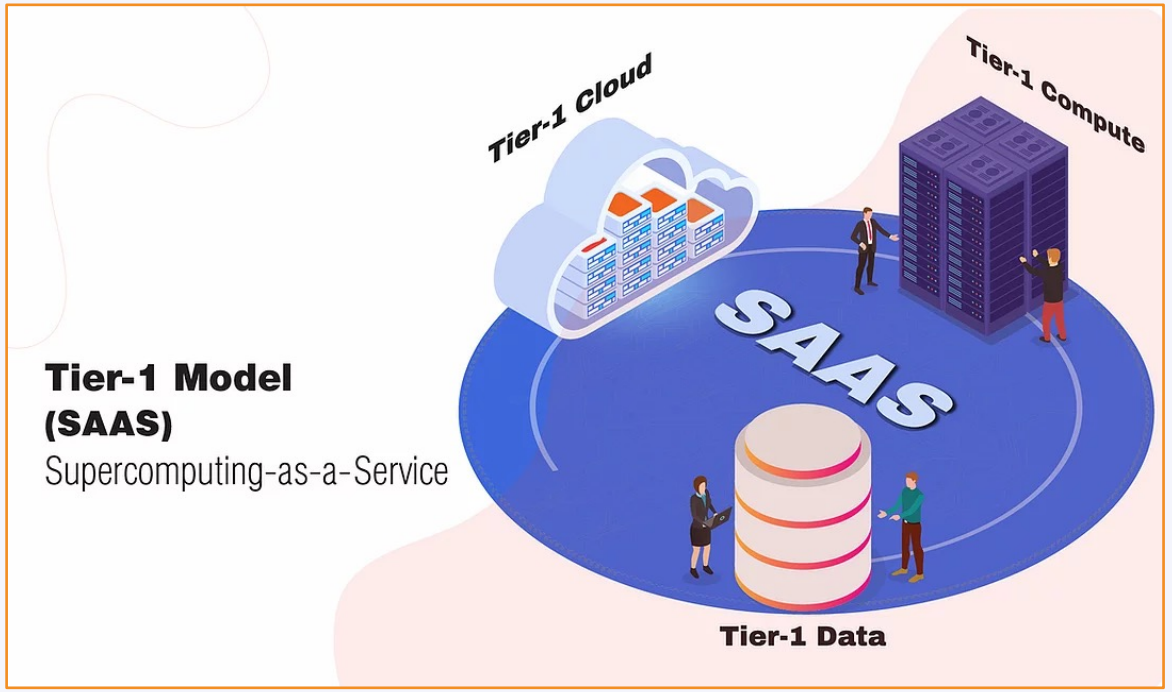
KU LEUVEN

UHASSELT

Tier-2



Tier-1



https://www.vscentrum.be/getaccess

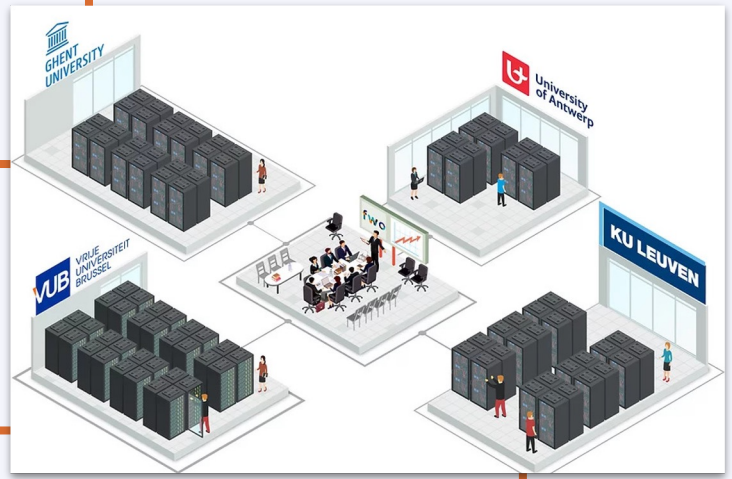
VSC Resources at the VSC hubs

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)**

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

Tier-1 Compute

infographic of

Hortense

The new Tier-1 supercomputer

MAIN PARTITION

294 WORKERNODES, EACH WITH:
2X 64-CORE AMD EPYC 7H12 CPU 2.6 GHZ (128 CORES PER NODE)
256 GIB RAM (~2GTB/CORE)
480 GB SSD LOCAL DISK

OPEN OnDemand

LARGE-MEMORY PARTITION

42 WORKERNODES, EACH WITH:
2X 64-CORE AMD EPYC 7H12 CPU 2.6 GHZ (128 CORES PER NODE)
512 GIB RAM (~4GB/CORE), NO SWAP
480 GB SSD LOCAL DISK

GPU PARTITION

20 WORKERNODES, EACH WITH:
2X 24-CORE AMD EPYC 7402 CPU 2.8 GHZ (48 CORES PER NODE)
4X NVIDIA A100-SXM4 (40 GB GPU MEMORY), NVLINK3
256 GIB RAM (~5GB/CPU CORE), NO SWAP
480 GB SSD LOCAL DISK

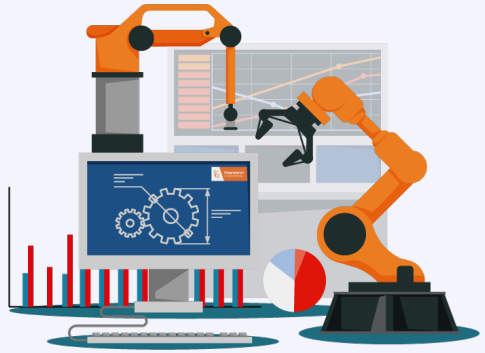
SHARED INFRASTRUCTURE

STORAGE: 3 PB SHARED SCRATCH STORAGE, BASED ON LUSTRE.ORG
INTERCONNECT INFINIBAND HDR-100
(~12.5GB/SEC), 2:1 FAT TREE TOPOLOGY
FOR THE GPU PARTITION SPECIFICALLY: DUAL HDR-100 INFINIBAND

SSH CLIENT

Access Tier-1

Academia



Industry

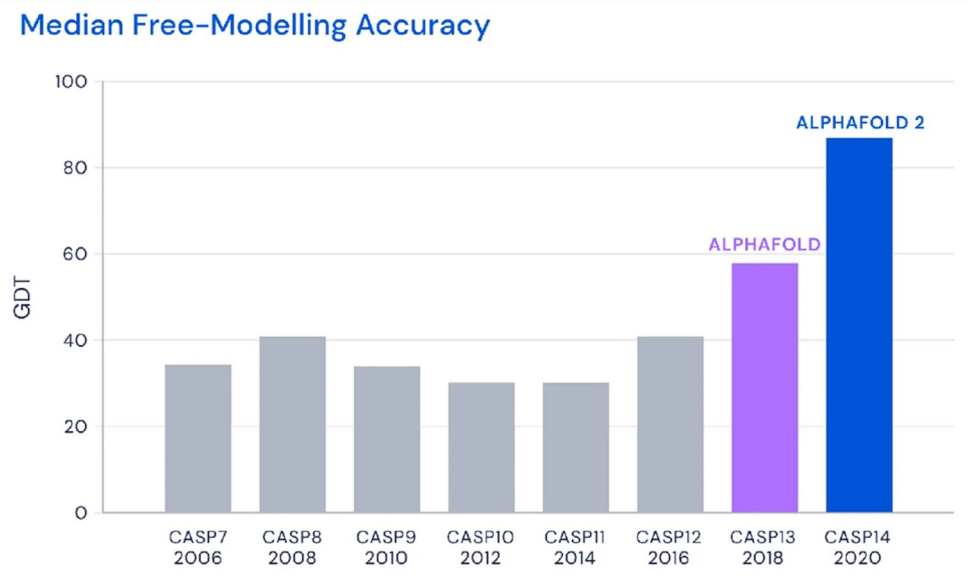
- Project based
- Three fixed submission dates per year
- Evaluation by the Tier-1 Allocation Board
- 5 non-Flemish panel members (NL, FR, UK)
- Technical evaluation
- Free

ACCESS	STORAGE	SUPPORT	TRAINING	INDUSTRY
< 500 node days **FREE** Hortense CPU 19€ Hortense GPU 94€ node/day	< 1 TB **FREE** 16€ TB/month	**FREE**	**FREE**	< 5 days **FREE** Contact us

Tier-1 Compute prices only; including VAT

2020: AlphaFold 2 grabs the headlines

In November 2020, AlphaFold was recognised as a solution to the 50-year “protein-folding problem”.



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*.

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

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

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."

AI-driven structural biology community

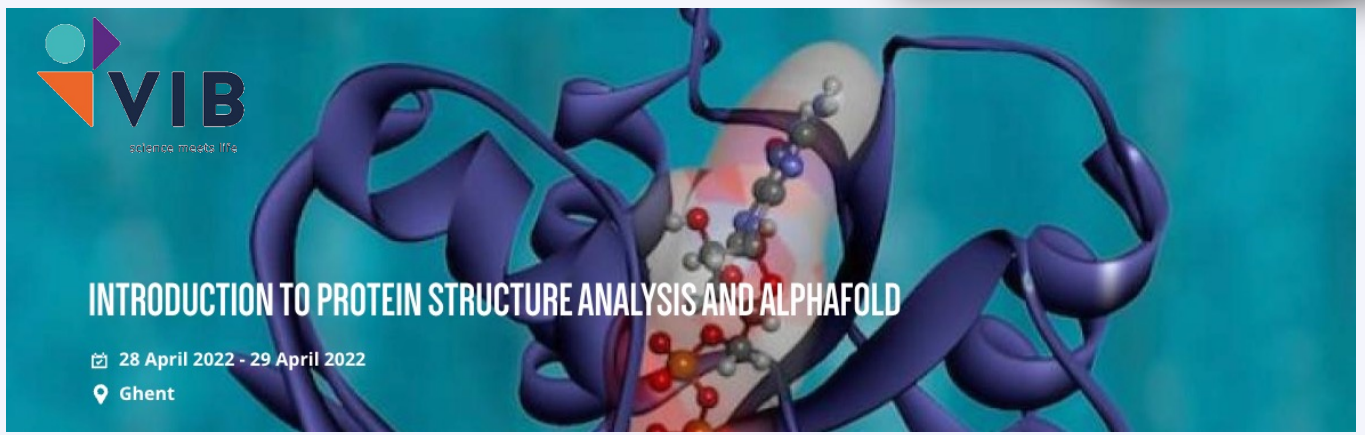
interested? Mail to carl.mensch@uantwerpen.be

EURO BELGIUM

VLAAMS SUPERCOMPUTER CENTRUM

First community meetings

YouTube



https://www.youtube.com/watch?v=jP9Qg1yBGcs

AlphaFold User Day

1. 

VSC AlphaFold Event | Sameer Velankar | EMBL-EBI

Vlaams Supercomputer Centrum • 224 weergaven • 1 maand geleden

2. 

VSC AlphaFold Event | Janani Durairaj | University of Basel

Vlaams Supercomputer Centrum • 42 weergaven • 1 maand geleden

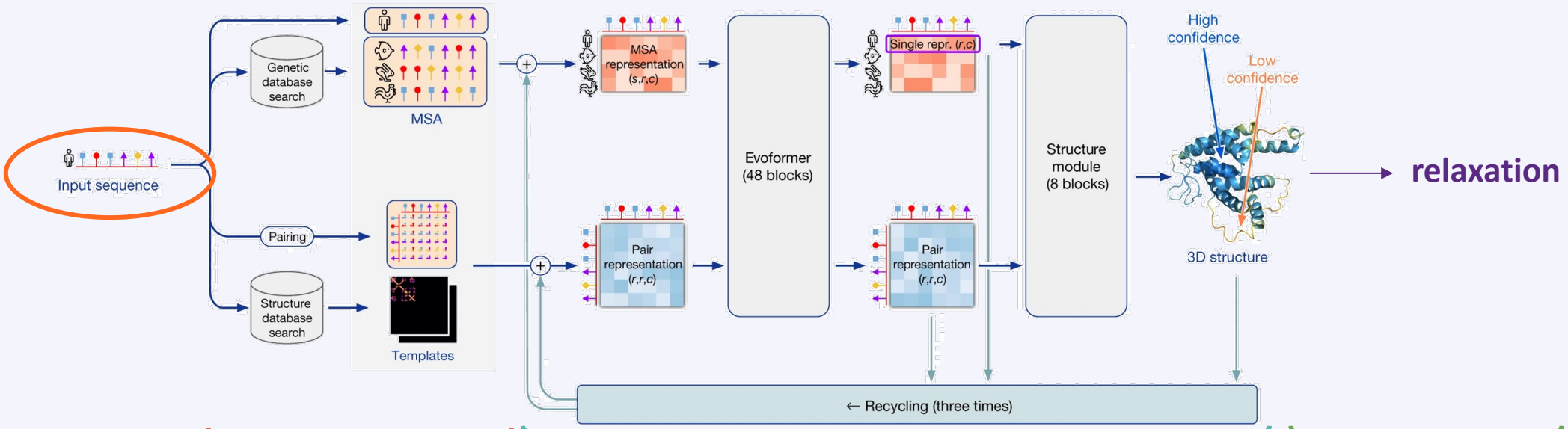
3. 

VSC AlphaFold Event | Ida de Vries | NKI

Vlaams Supercomputer Centrum • 86 weergaven • 1 maand geleden

https://www.youtube.com/playlist?list=PLtDP8nevxpVbzgt9ZiJfZhr1wGRiTAQW6

High-level overview of AlphaFold 2

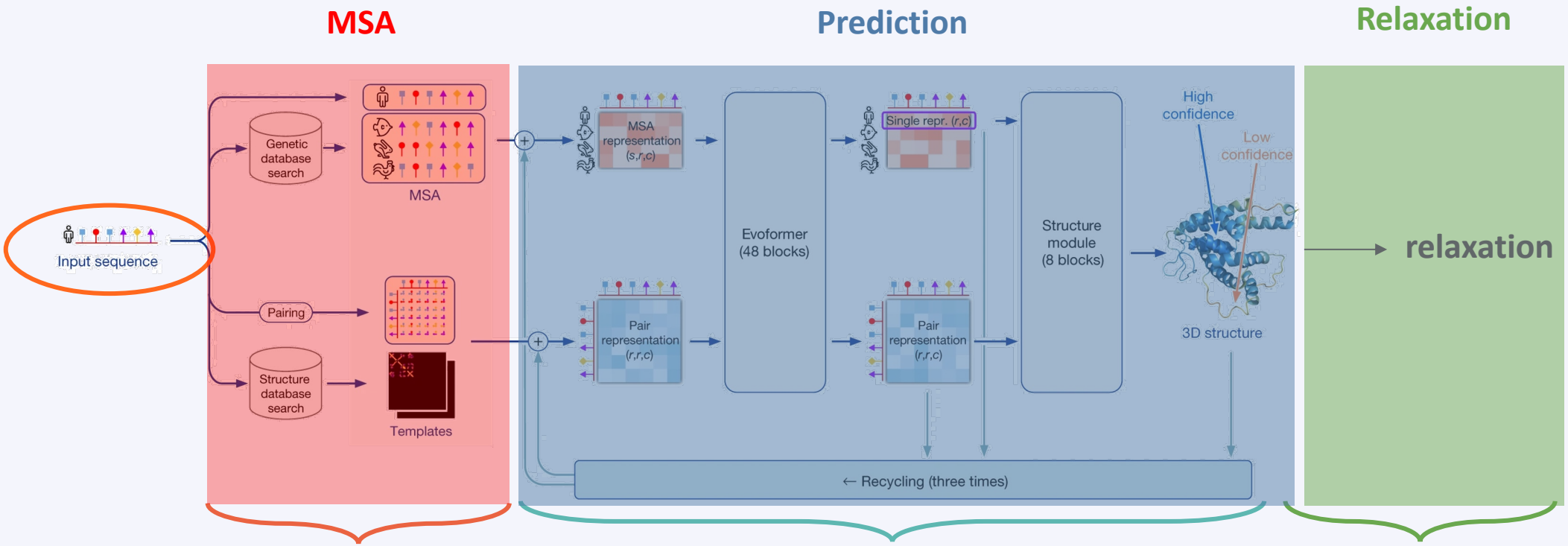


Multiple sequence alignment
Structural templates search

Run inputs through 5 trained AlphaFold models
→ 5 structures predicted
→ sorted by confidence

Relax structures (GPU)

High-level overview of AlphaFold 2

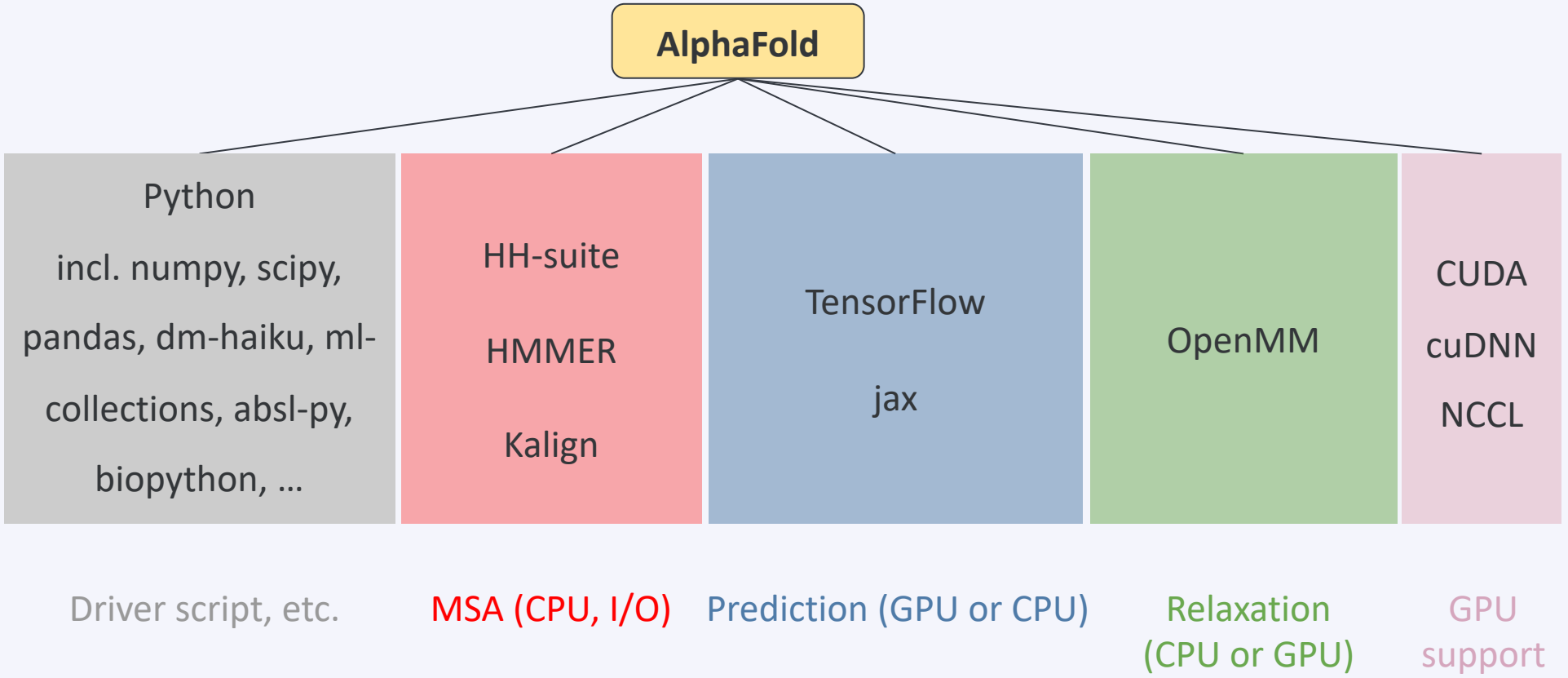


Multiple sequence alignment
Structural templates search

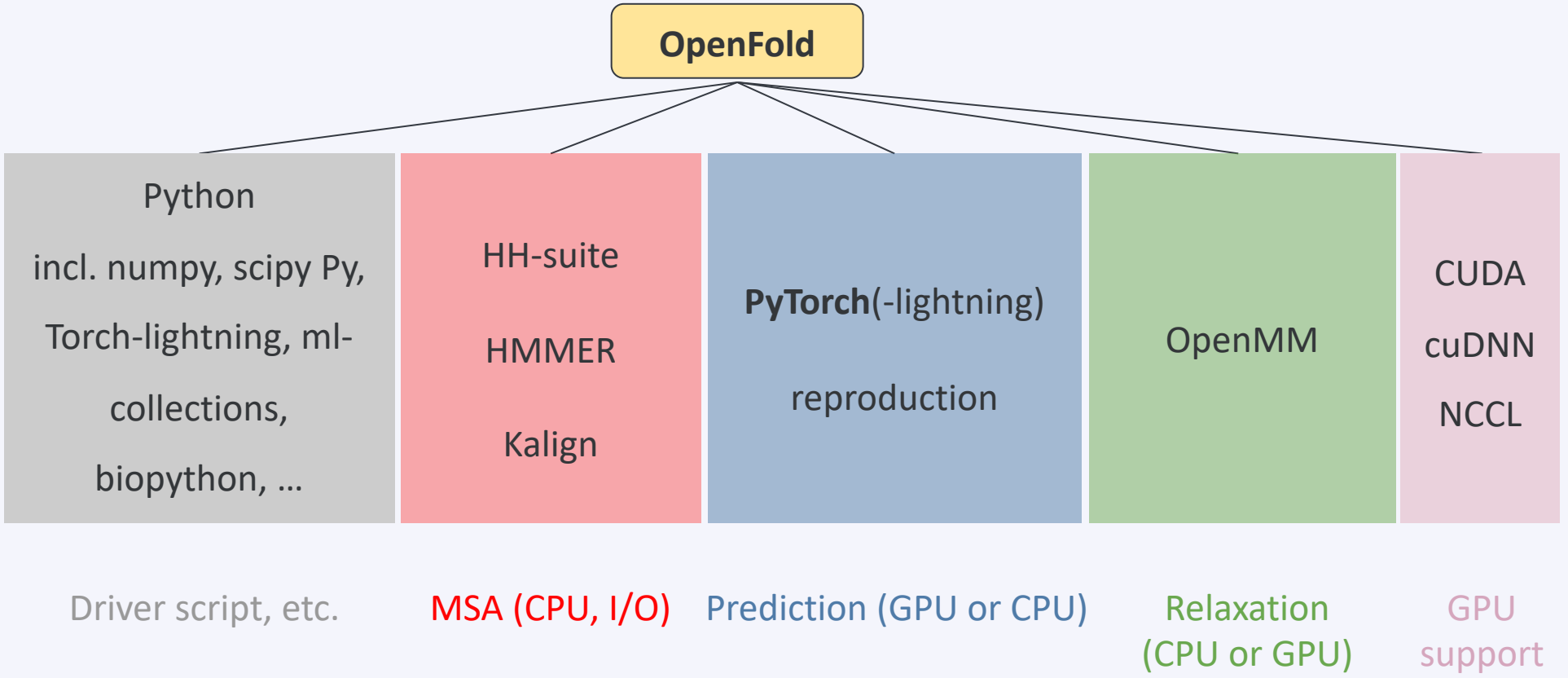
Run inputs through 5 trained AlphaFold models
→ 5 structures predicted
→ sorted by confidence

Relax structures (GPU)

Dependencies AlphaFold



Dependencies OpenFold



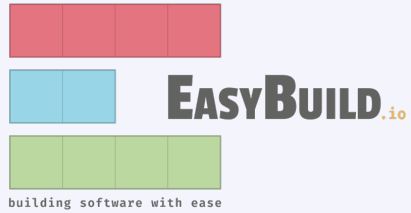
Installation

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

- Docker file provided: 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

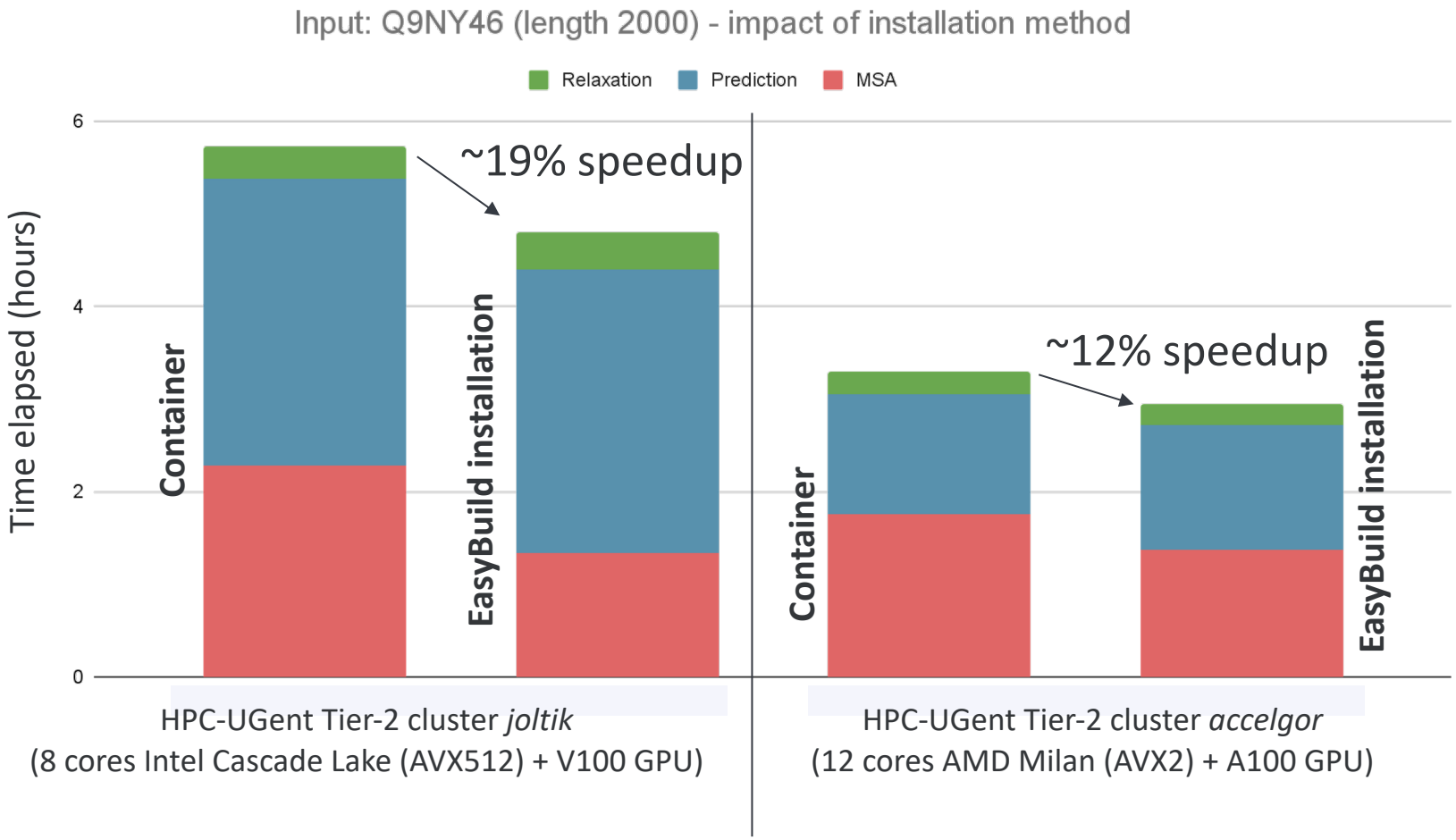
Option 2: install package and all required dependencies **from source code**

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

Installation

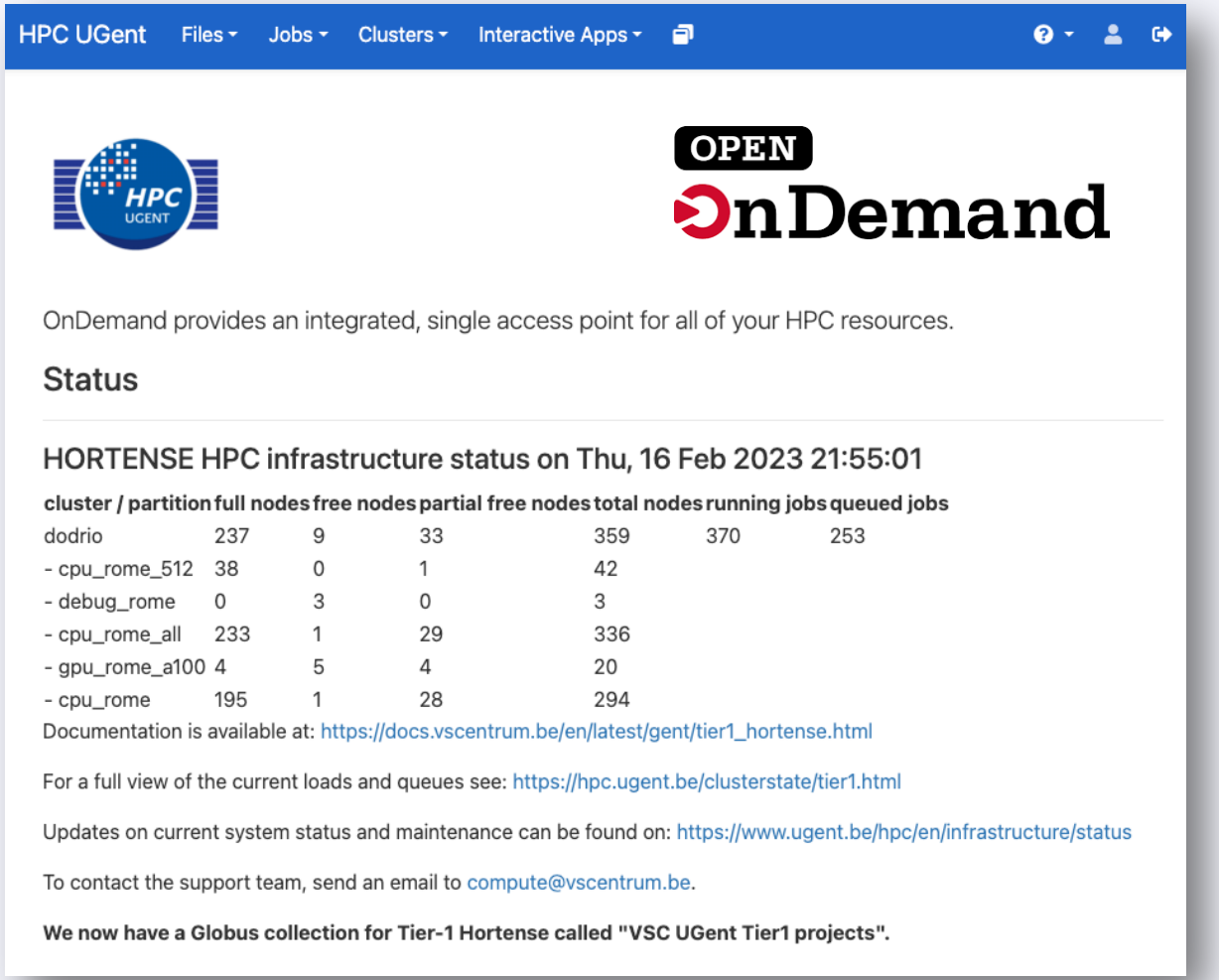


CONNECTING TO HORTENSE

Option 1: Use ssh to connect (requires private key)

```
→  ~ ssh [user-id]@tier1.hpc.ugent.be
```

Option 2: Open On-Demand web portal https://tier1.hpc.ugent.be.



Connecting

```
➜  ~ ssh hortense


 HORTENSE HPC infrastructure status on Thu, 16 Feb 2023 11:25:02
      cluster     - full - free -  part - total - running - queued
      partition     nodes  nodes   free   nodes   jobs      jobs
 ---------------------------------------------------------------------
 dodrio               227     19     36     359       358      233
  - cpu_rome_512       39      0      0      42
  - gpu_rome_a100       5      5      3      20
  - debug_rome          0      3      0       3
  - cpu_rome          183     11     33     294
  - cpu_rome_all      222     11     33     336


Documentation is available at:
 https://docs.vscentrum.be/en/latest/gent/tier1_hortense.html
For a full view of the current loads and queues see:
 https://hpc.ugent.be/clusterstate/tier1.html
Updates on current system status and maintenance can be found on:
 https://www.ugent.be/hpc/en/infrastructure/status
To contact the support team, send an email to compute@vscentrum.be


We now have a Globus collection for Tier-1 Hortense called "VSC UGent Tier1 projects".


Last login: Wed Feb 15 23:17:21 2023 from 10.141.47.148
[vsc20133@login56 ~]$
```

Connecting

```
➜  ~ ssh hortense

 HORTENSE HPC infrastructure status on Thu, 16 Feb 2023 11:25:02
      cluster     - full - free -  part - total - running - queued
      partition     nodes  nodes   free   nodes   jobs      jobs
 ---------------------------------------------------------------------
 dodrio               227     19     36     359       358      233
  - cpu_rome_512       39      0      0      42
  - gpu_rome_a100       5      5      3      20
  - debug_rome          0      3      0       3
  - cpu_rome          183     11     33     294
  - cpu_rome_all      222     11     33     336


Documentation is available at:
 https://docs.vscentrum.be/en/latest/gent/tier1_hortense.html
For a full view of the current loads and queues see:
 https://hpc.ugent.be/clusterstate/tier1.html
Updates on current system status and maintenance can be found on:
 https://www.ugent.be/hpc/en/infrastructure/status
To contact the support team, send an email to compute@vscentrum.be


We now have a Globus collection for Tier-1 Hortense called "VSC UGent Tier1 projects".


Last login: Wed Feb 15 23:17:21 2023 from 10.141.47.148
[vsc20133@login56 ~]$
```

```
$VSC_ARCH_LOCAL                 $VSC_SCRATCH_ARCANINE
$VSC_ARCH_SUFFIX                $VSC_SCRATCH_ARCANINE_VO
$VSC_DATA                       $VSC_SCRATCH_ARCANINE_VO_USER
$VSC_DATA_VO                    $VSC_SCRATCH_CLUSTER
$VSC_DATA_VO_USER               $VSC_SCRATCH_KYUKON
$VSC_DEFAULT_CLUSTER_MODULE     $VSC_SCRATCH_KYUKON_VO
$VSC_HOME                       $VSC_SCRATCH_KYUKON_VO_USER
$VSC_INSTITUTE                  $VSC_SCRATCH_NODE
$VSC_INSTITUTE_CLUSTER          $VSC_SCRATCH_PROJECTS_BASE
$VSC_INSTITUTE_LOCAL            $VSC_SCRATCH_SITE
$VSC_OS_LOCAL                   $VSC_SCRATCH_VO
$VSCPROFILELOADED               $VSC_SCRATCH_VO_USER
$VSC_SCRATCH                    $VSC_VO
[vsc20133@login56 ~]$ ls $VSC
```

Connecting

```
➜  ~ ssh hortense

 HORTENSE HPC infrastructure status on Thu, 16 Feb 2023 11:25:02
      cluster     - full - free -  part - total - running - queued
      partition     nodes  nodes   free   nodes   jobs      jobs
 ---------------------------------------------------------------------
 dodrio               227     19     36     359       358      233
  - cpu_rome_512       39      0      0      42
  - gpu_rome_a100       5      5      3      20
  - debug_rome          0      3      0       3
  - cpu_rome          183     11     33     294
  - cpu_rome_all      222     11     33     336


Documentation is available at:
 https://docs.vscentrum.be/en/latest/gent/tier1_hortense.html
For a full view of the current loads and queues see:
 https://hpc.ugent.be/clusterstate/tier1.html
Updates on current system status and maintenance can be found on:
 https://www.ugent.be/hpc/en/infrastructure/status
To contact the support team, send an email to compute@vscentrum.be


We now have a Globus collection for Tier-1 Hortense called "VSC UGent Tier1 projects".


Last login: Wed Feb 15 23:17:21 2023 from 10.141.47.148
[vsc20133@login56 ~]$
```

```
$VSC_ARCH_LOCAL                 $VSC_SCRATCH_ARCANINE
$VSC_ARCH_SUFFIX                $VSC_SCRATCH_ARCANINE_VO
$VSC_DATA                       $VSC_SCRATCH_ARCANINE_VO_USER
$VSC_DATA_VO                    $VSC_SCRATCH_CLUSTER
$VSC_DATA_VO_USER               $VSC_SCRATCH_KYUKON
$VSC_DEFAULT_CLUSTER_MODULE     $VSC_SCRATCH_KYUKON_VO
$VSC_HOME                       $VSC_SCRATCH_KYUKON_VO_USER
$VSC_INSTITUTE                  $VSC_SCRATCH_NODE
$VSC_INSTITUTE_CLUSTER          $VSC_SCRATCH_PROJECTS_BASE
$VSC_INSTITUTE_LOCAL            $VSC_SCRATCH_SITE
$VSC_OS_LOCAL                   $VSC_SCRATCH_VO
$VSCPROFILELOADED               $VSC_SCRATCH_VO_USER
$VSC_SCRATCH                    $VSC_VO
[vsc20133@login56 ~]$ ls $VSC
```

Connecting

Home directory $VSC_HOME

•The data stored here should be relatively small, and not generating very intense I/O during jobs.

• Store all kinds of configuration files

• Readable and writable on all VSC sites.

Data directory $VSC_DATA

•A bigger 'workspace', for program code, datasets or results that must be stored for a longer period of time.

• Depending on the cluster, performance may not be very high.

• Readable and writable on all VSC sites.

Scratch directories $VSC_SCRATCH_XXX

•For temporary or transient data; there is typically no backup

•These file systems are not exported to other VSC sites.

```
$VSC_ARCH_LOCAL                 $VSC_SCRATCH_ARCANINE
$VSC_ARCH_SUFFIX                $VSC_SCRATCH_ARCANINE_VO
$VSC_DATA                       $VSC_SCRATCH_ARCANINE_VO_USER
$VSC_DATA_VO                    $VSC_SCRATCH_CLUSTER
$VSC_DATA_VO_USER               $VSC_SCRATCH_KYUKON
$VSC_DEFAULT_CLUSTER_MODULE     $VSC_SCRATCH_KYUKON_VO
$VSC_HOME                       $VSC_SCRATCH_KYUKON_VO_USER
$VSC_INSTITUTE                  $VSC_SCRATCH_NODE
$VSC_INSTITUTE_CLUSTER          $VSC_SCRATCH_PROJECTS_BASE
$VSC_INSTITUTE_LOCAL            $VSC_SCRATCH_SITE
$VSC_OS_LOCAL                   $VSC_SCRATCH_VO
$VSCPROFILELOADED               $VSC_SCRATCH_VO_USER
$VSC_SCRATCH                    $VSC_VO
[vsc20133@login56 ~]$ ls $VSC
```

Connecting

```
➜  ~ ssh hortense

 HORTENSE HPC infrastructure status on Thu, 16 Feb 2023 11:25:02
      cluster     - full - free -  part - total - running - queued
      partition     nodes  nodes   free   nodes   jobs      jobs
 ---------------------------------------------------------------------
 dodrio               227     19     36     359       358      233
  - cpu_rome_512       39      0      0      42
  - gpu_rome_a100       5      5      3      20
  - debug_rome          0      3      0       3
  - cpu_rome          183     11     33     294
  - cpu_rome_all      222     11     33     336


Documentation is available at:
 https://docs.vscentrum.be/en/latest/gent/tier1_hortense.html
For a full view of the current loads and queues see:
 https://hpc.ugent.be/clusterstate/tier1.html
Updates on current system status and maintenance can be found on:
 https://www.ugent.be/hpc/en/infrastructure/status
To contact the support team, send an email to compute@vscentrum.be


We now have a Globus collection for Tier-1 Hortense called "VSC UGent Tier1 projects".


Last login: Wed Feb 15 23:17:21 2023 from 10.141.47.148
[vsc20133@login56 ~]$
```

```
$VSC_ARCH_LOCAL                 $VSC_SCRATCH_ARCANINE
$VSC_ARCH_SUFFIX                $VSC_SCRATCH_ARCANINE_VO
$VSC_DATA                       $VSC_SCRATCH_ARCANINE_VO_USER
$VSC_DATA_VO                    $VSC_SCRATCH_CLUSTER
$VSC_DATA_VO_USER               $VSC_SCRATCH_KYUKON
$VSC_DEFAULT_CLUSTER_MODULE     $VSC_SCRATCH_KYUKON_VO
$VSC_HOME                       $VSC_SCRATCH_KYUKON_VO_USER
$VSC_INSTITUTE                  $VSC_SCRATCH_NODE
$VSC_INSTITUTE_CLUSTER          $VSC_SCRATCH_PROJECTS_BASE
$VSC_INSTITUTE_LOCAL            $VSC_SCRATCH_SITE
$VSC_OS_LOCAL                   $VSC_SCRATCH_VO
$VSCPROFILELOADED               $VSC_SCRATCH_VO_USER
$VSC_SCRATCH                    $VSC_VO
[vsc20133@login56 ~]$ ls $VSC
```

Set up your files and job script in the $VSC_DATA and $VSC_SCRATCH_PROJECTS_BASE
scp or Globus can be use to transfer files between your system and the clusters

Track quota:

```
>> my_dodrio_quota -p [project]
```

Connecting

```
→  ~ ssh hortense

HORTENSE HPC infrastructure status on Thu, 16 Feb 2023 11:25:02
     cluster     - full - free -  part - total - running - queued
     partition     nodes  nodes   free   nodes   jobs      jobs
-----------------------------------------------------------------------
dodrio               227     19     36     359       358      233
 - cpu_rome_512       39      0      0      42
 - gpu_rome_a100       5      5      3      20
 - debug_rome          0      3      0       3
 - cpu_rome          183     11     33     294
 - cpu_rome_all      222     11     33     336

Documentation is available at:
 https://docs.vscentrum.be/en/latest/gent/tier1_hortense.html
For a full view of the current loads and queues see:
 https://hpc.ugent.be/clusterstate/tier1.html
Updates on current system status and maintenance can be found on:
 https://www.ugent.be/hpc/en/infrastructure/status
To contact the support team, send an email to compute@vscentrum.be


We now have a Globus collection for Tier-1 Hortense called "VSC UGent Tier1 projects".


Last login: Wed Feb 15 23:17:21 2023 from 10.141.47.148
[vsc20133@login56 ~]$
```

Select the partition you need.

```
[vsc20133@login56 ~]$ module swap cluster/dodrio/gpu_rome


The following have been reloaded with a version change:
  1) cluster/dodrio/cpu_rome => cluster/dodrio/gpu_rome_a100
```

JOB SCRIPT

```
#!/bin/bash
#PBS -N OpenFold
#PBS -l nodes=1:ppn=12,gpus=1
#PBS -l walltime=12:00:00
#PBS -o OF-${PBS_JOBID}.o.txt
#PBS -e OF-${PBS_JOBID}.e.txt
#PBS -A [accounting project name]
#PBS -m abe


# Load modules:
module load OpenFold/1.0.1-foss-2021a-CUDA-11.3.1


cd ${PBS_O_WORKDIR}


# Databases location:
# VUB data=/databases/bio/alphafold-2.1.1
# GHENT TIER-2 data=/arcanine/scratch/gent/apps/AlphaFold/20220701

data=/readonly/dodrio/apps/AlphaFold/20220701


# Run the run_pretrained_openfold script:
run_pretrained_openfold.py \
    fasta_dir \
    $data/pdb_mmcif/mmcif_files/ \
    --uniref90_database_path $data/uniref90/uniref90.fasta \
    --mgnify_database_path $data/mgnify/mgy_clusters_2018_12.fa \
    --pdb70_database_path $data/pdb70/pdb70 \
    --uniclust30_database_path $data/uniclust30/uniclust30_2018_08/uniclust30_2018_08 \
    --output_dir ./OUTPUT \
    --bfd_database_path $data/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
    --model_device "cuda:0" \
    --jackhmmer_binary_path $EBROOTHMMER/bin/jackhmmer \
    --hhblits_binary_path $EBROOTHHMINSUITE/bin/hhblits \
    --hhsearch_binary_path $EBROOTHHMINSUITE/bin/hhsearch \
    --kalign_binary_path $EBROOTKALIGN/bin/kalign \
    --cpus=12 \
    --config_preset "model_1_ptm" \
    --openfold_checkpoint_path /readonly/dodrio/apps/OpenFold/20220719/openfold_params/finetuning_ptm_2.pt
```

Torque resource configuration

Load module including dependencies

Databases location

Run the script

RESOURCES

```
#!/bin/bash
#PBS -N OpenFold
#PBS -l nodes=1:ppn=12,gpus=1
#PBS -l walltime=24:00:00
#PBS -o OF-${PBS_JOBID}.o.txt
#PBS -e OF-${PBS_JOBID}.e.txt
#PBS -A [accounting project name]
#PBS -m abe
```

Choose a job name

Number of cores should correspond to the 1 GPU. Hortense has 48 CPU cores and 4 GPUs, so set ppn=12 and gpus=1

Request walltime depending on the size of the protein and number of sequences you are running

Write ouput and error to these files.

Specify your Tier-1 project.

Stay informed on various job steps (queueing, start, end) via e-mail

Optional – memory issues

```
#PBS -l mem=64GB
```

Hortense GPU partition nodes have 256 GiB RAM (~5GB/CPU core) each. Increase this if you run into memory issues (for large MSAs)

DATABASES

The location of the shared available databases depends on the cluster.

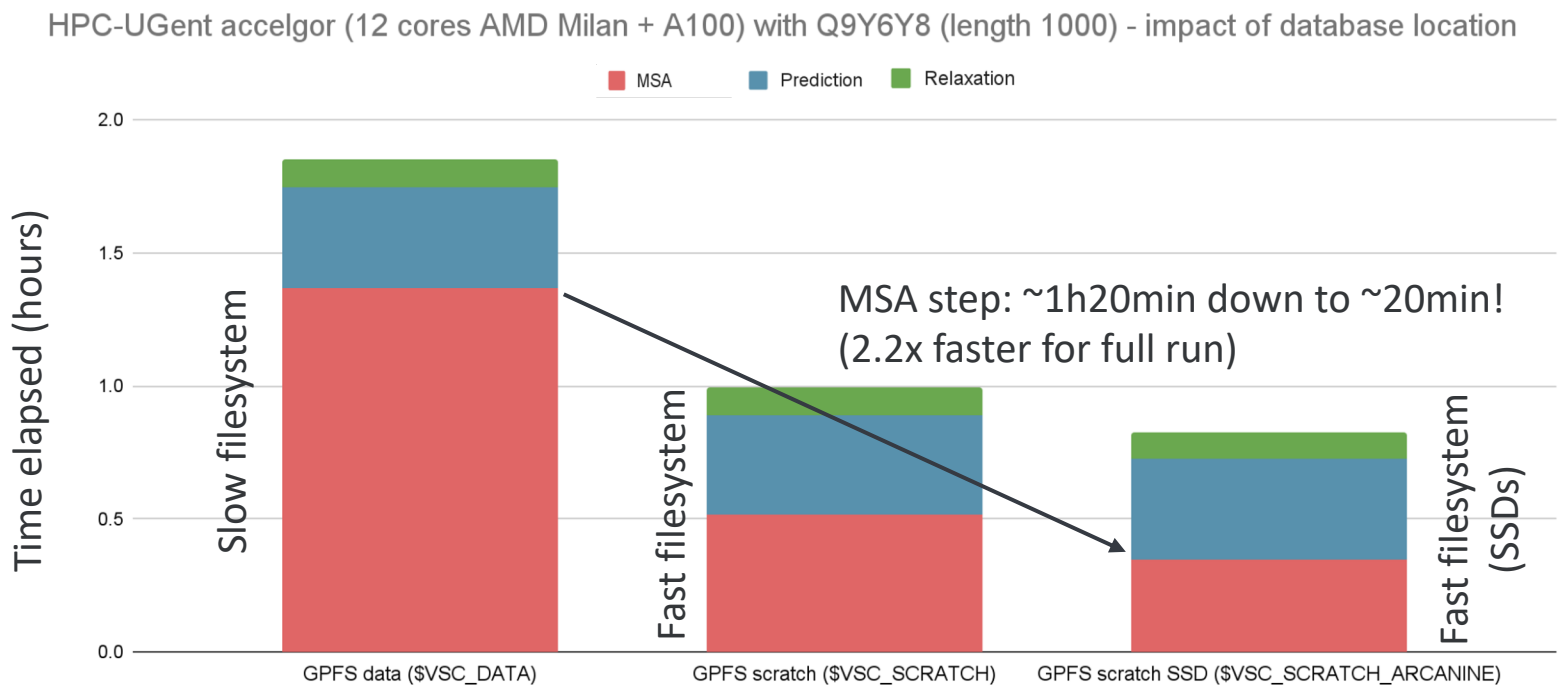
```
# Databases location:
# VUB data=/databases/bio/alphafold-2.1.1
# Ghent TIER-2 data=/arcanine/scratch/gent/apps/AlphaFold/20220701

data=/readonly/dodrio/apps/AlphaFold/20220701
```

Tier-2 VUB (Hydra-cluster)

Tier-2 in Ghent (accelgor, joltik)

Tier-1 Ghent

DATABASES



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.

RUN OPENFOLD

```
# Run the run_pretrained_openfold script:
run_pretrained_openfold.py \
    fasta_dir \
    $data/pdb_mmcif/mmcif_files/ \
    --uniref90_database_path $data/uniref90/uniref90.fasta \
    --mgnify_database_path $data/mgnify/mgy_clusters_2018_12.fa \
    --pdb70_database_path $data/pdb70/pdb70 \
    --uniclust30_database_path $data/uniclust30/uniclust30_2018_08/uniclust30_2018_08 \
    --bfd_database_path $data/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
    --output_dir ./OUTPUT \
    --model_device "cuda:0" \
    --jackhmmer_binary_path $EBROOTHMMER/bin/jackhmmer \
    --hhblits_binary_path $EBROOTHHMINSUITE/bin/hhblits \
    --hhsearch_binary_path $EBROOTHHMINSUITE/bin/hhsearch \
    --kalign_binary_path $EBROOTKALIGN/bin/kalign \
    --cpus=12 \
    --config_preset "model_1_ptm" \
    --openfold_checkpoint_path /readonly/dodrio/apps/OpenFold/20220719/openfold_params/finetuning_ptm_2.pt
```

RUN OPENFOLD

```
# Run the run_pretrained_openfold script:
run_pretrained_openfold.py \
    fasta_dir \
    $data/pdb_mmcif/mmcif_files/ \
    --uniref90_database_path $data/uniref90/uniref90.fasta \
    --mgnify_database_path $data/mgnify/mgy_clusters_2018_12.fa \
    --pdb70_database_path $data/pdb70/pdb70 \
    --uniclust30_database_path $data/uniclust30/uniclust30_2018_08/uniclust30_2018_08 \
    --bfd_database_path $data/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
    --output_dir ./OUTPUT \
    --model_device "cuda:0" \
    --jackhmmer_binary_path $EBROOTHMMER/bin/jackhmmer \
    --hhblits_binary_path $EBROOTHHMINSUITE/bin/hhblits \
    --hhsearch_binary_path $EBROOTHHMINSUITE/bin/hhsearch \
    --kalign_binary_path $EBROOTKALIGN/bin/kalign \
    --cpus=12 \
    --config_preset "model_1_ptm" \
    --openfold_checkpoint_path /readonly/dodrio/apps/OpenFold/20220719/openfold_params/finetuning_ptm_2.pt
```

Prepare a directory with your fasta sequences.

Databases

output – if it contains ./OUTPUT/alignments, the MSA step is skipped

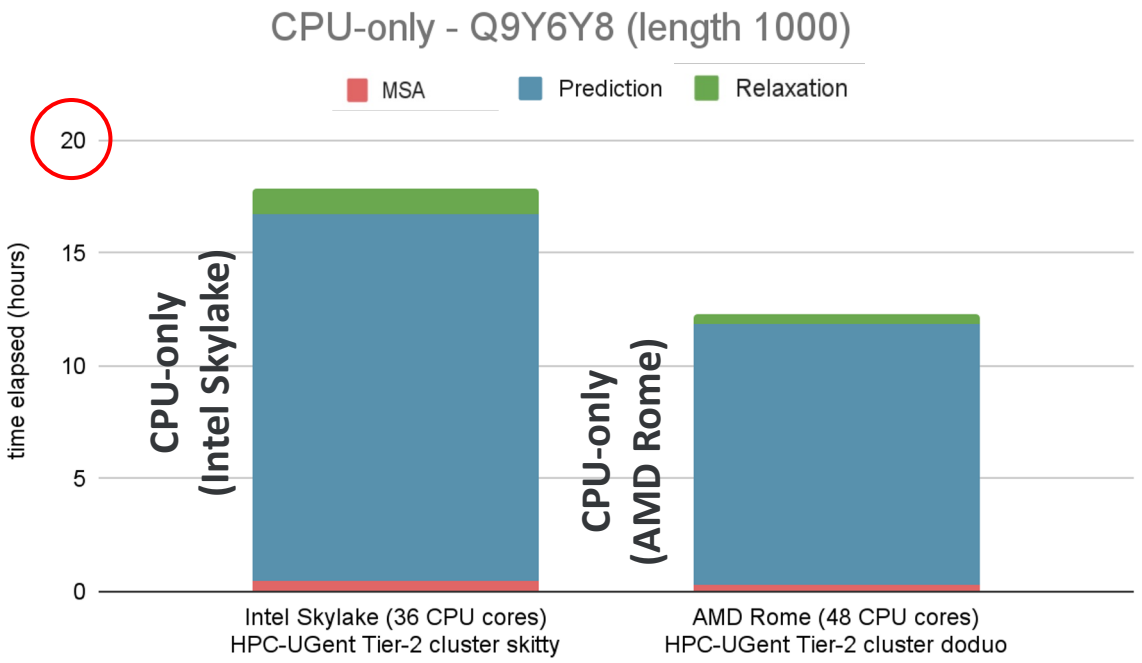
Run on GPU

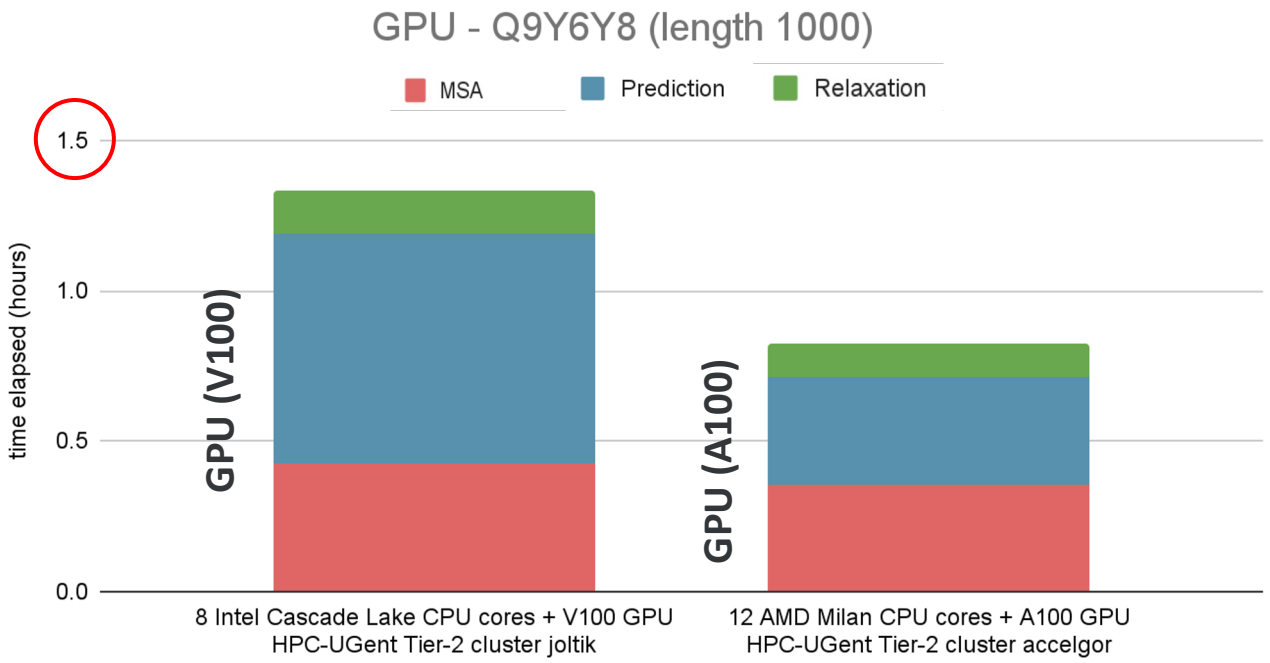
Binaries

The number of CPUs might impact the performance

Depends on cluster

CPU vs GPU AlphaFold





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

PBS AND SLURM

On Tier-1 Torque is recommended for submitting and managing your jobs. It serves as frontend for SLURM, the resource and job scheduler.

```
>> qsub --dryrun submit.pbs
```

With `qsub --dryrun` the SLURM set-up can be verified

PBS

```
#!/bin/bash
#PBS -N OpenFold
#PBS -l nodes=1:ppn=12,gpus=1
#PBS -l walltime=12:00:00
#PBS -o OF-${PBS_JOBID}.o.txt
#PBS -e OF-${PBS_JOBID}.e.txt
#PBS -A [accounting project name]
#PBS -m abe


...
```

SLURM

```
#!/bin/bash
#SBATCH --account="---"
#SBATCH --chdir=”---"
#SBATCH --cpus-per-gpu="12"
#SBATCH –error=---/OF-%A.e.txt"
#SBATCH --export="NONE"
#SBATCH --get-user-env="60L"
#SBATCH --gres="gpu:4"
#SBATCH --job-name="---"
#SBATCH --mail-type="BEGIN,END,FAIL,TIME_LIMIT"
#SBATCH --nodes="1"
#SBATCH --ntasks-per-node="12"
#SBATCH --ntasks="12"
#SBATCH --output="---/OF-%A.o.txt"
#SBATCH --time="1-00:00:00"
```

SUBMIT AND MONITOR

```
#!/bin/bash
#PBS -N OpenFold
#PBS -l nodes=1:ppn=12,gpus=1
#PBS -l walltime=12:00:00
#PBS -o OF-${PBS_JOBID}.o.txt
#PBS -e OF-${PBS_JOBID}.e.txt
#PBS -A [accounting project name]
#PBS -m abe


# Load modules:
module load OpenFold/1.0.1-foss-2021a-CUDA-11.3.1


cd ${PBS_O_WORKDIR}

...
```

The job can be submitted with qsub (or `sbatch` if using SLURM)

```
>> qsub submit.pbs
```

The job can be monitored with `qstat` or `squeue`

```
>> qstat

>> squeue
```

BENCHMARKING OPENFOLD

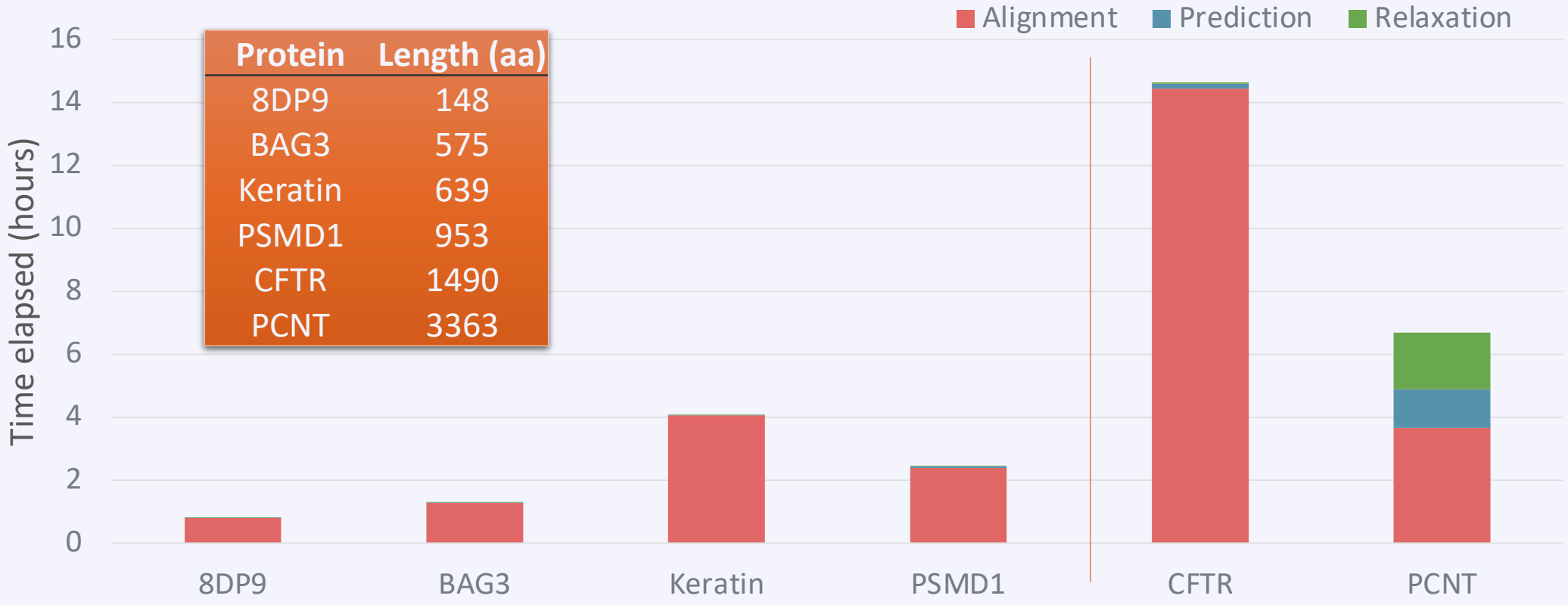
- Time-to-result performance of **OpenFold v1.0.1** was evaluated
- Using a set of protein sequences of different length

Protein	Length (aa)
8DP9	148
BAG3	575
Keratin	639
PSMD1	953
CFTR	1490
PCNT	3363

- Run on Tier-1 Hortense and Tier-2 Accelgor with A100 GPUs.
- Run by Moisés on local workstation with NVIDIA RTX A5000 GPU (24Gb memory)

BENCHMARKING OPENFOLD

Protein	Length (aa)
8DP9	148
BAG3	575
Keratin	639
PSMD1	953
CFTR	1490
PCNT	3363

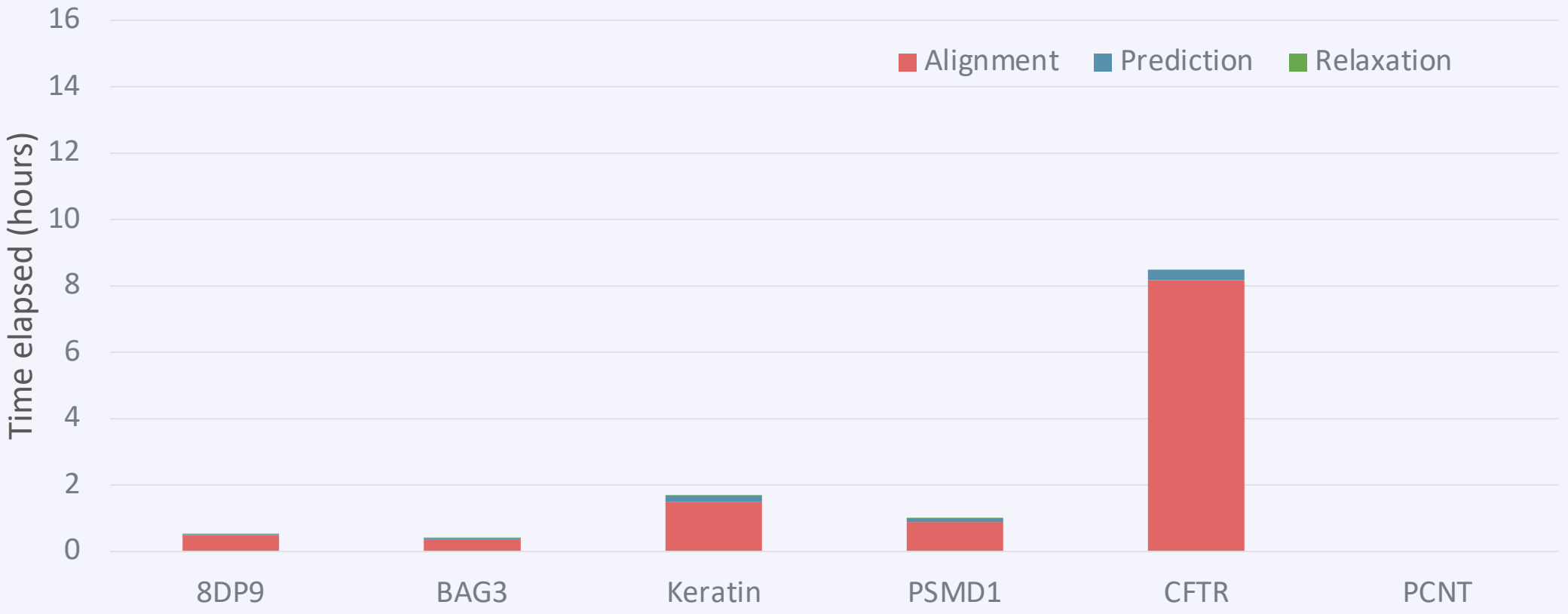


HPC-UGent Tier-1 cluster *Hortense*
(12 cores AMD Rome + A100 GPU (40 GB)

HPC-UGent Tier-2 cluster *Accelgor*
(12 cores AMD Rome + A100 GPU (80 GB)

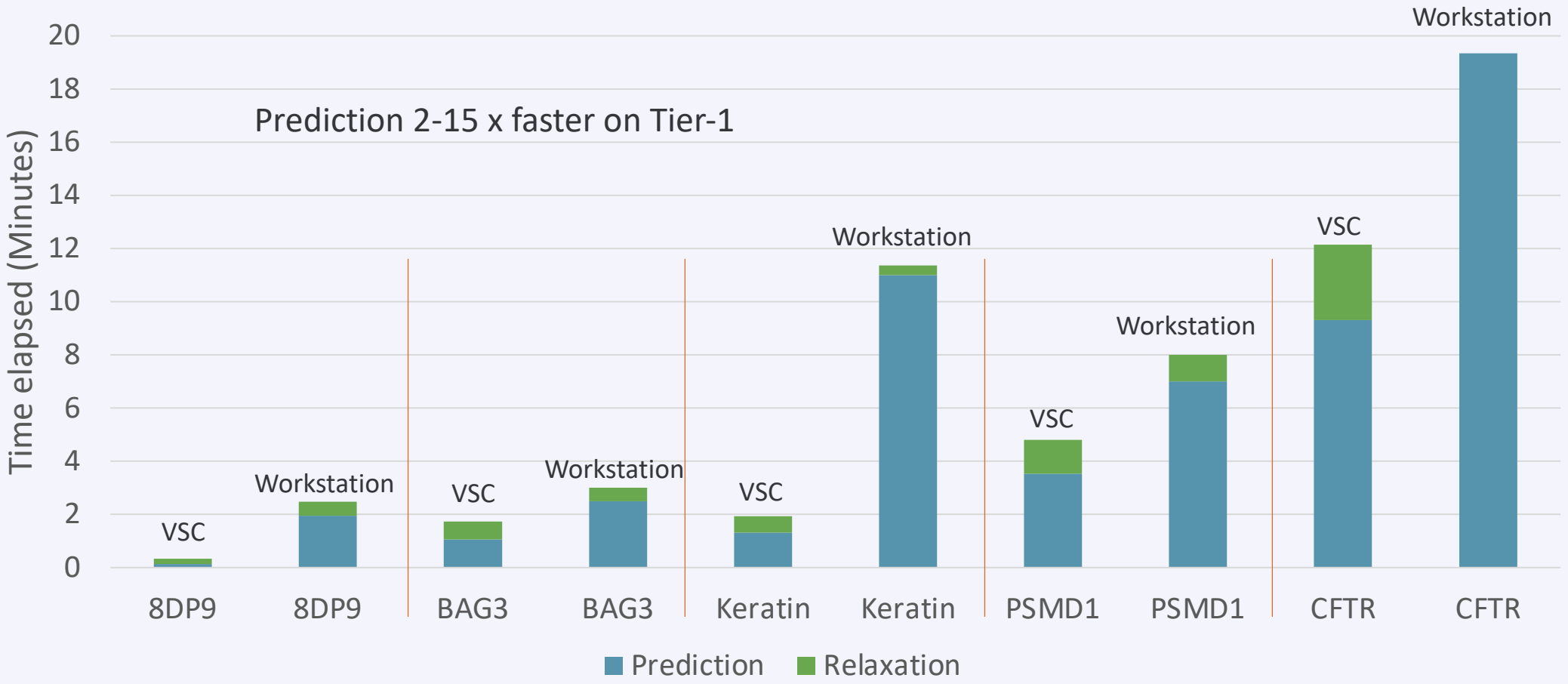
- MSA determines the total runtime the most
- MSA time depends on sequence and not merely the sequence length
- Enough memory and GPU memory needed for some MSAs

BENCHMARKING OPENFOLD



- Puxano workstation with NVIDIA RTX A5000 GPU (24Gb memory)
- Same trends
- MSA faster on local machine vs shared file system VSC

BENCHMARKING OPENFOLD



What's next?

Improve MSA performance

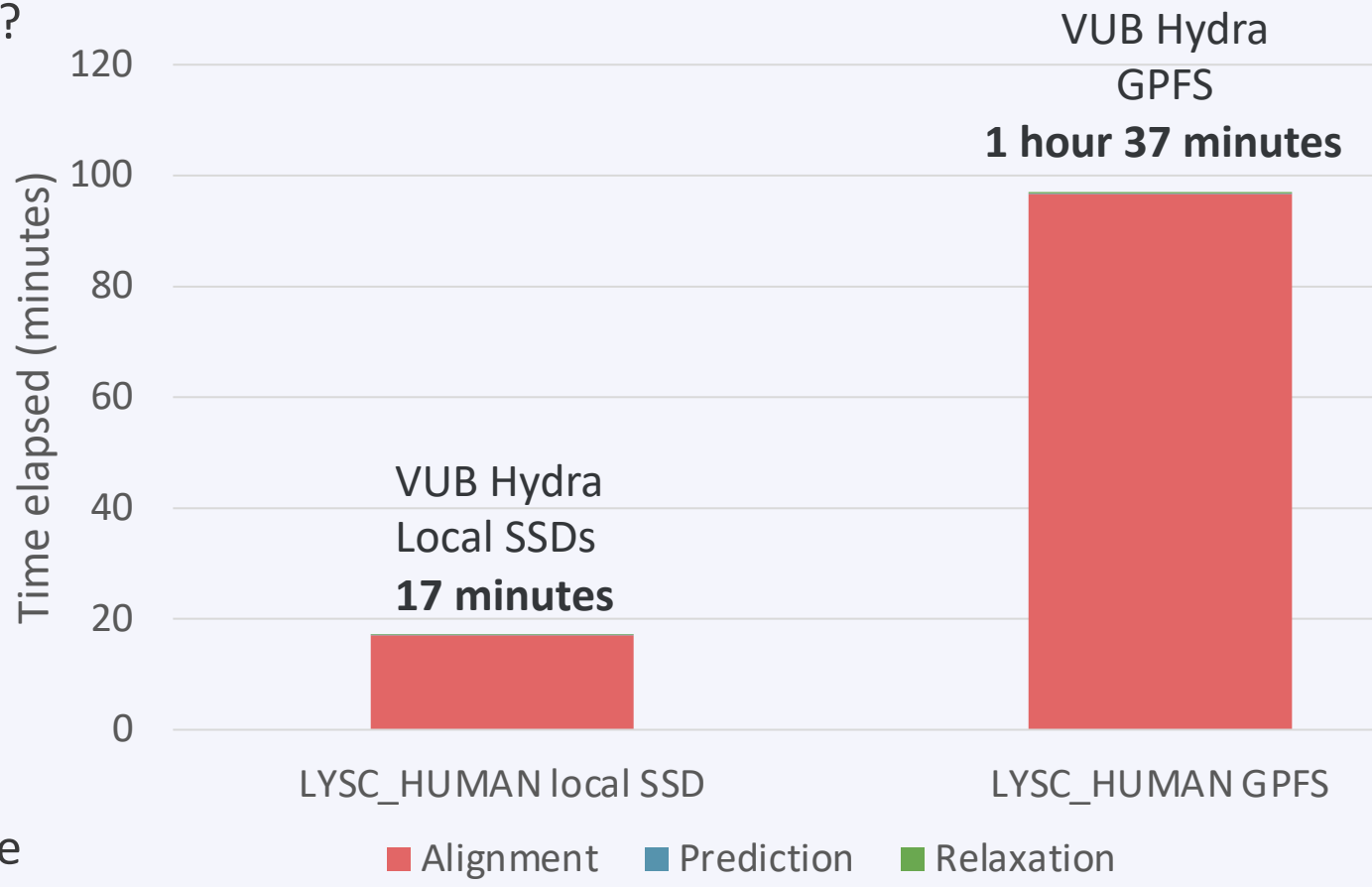
- Optimal use of dependencies in MSA?
- Options or configurations
- Local SSD (VUB test phase)
- CPU only MSA?

Other

- Multimer
- Training

AI-driven structural biology community

- Open for ideas
- Contact carl.mensch@uantwerpen.be



Useful resources

- OpenFold website: https://openfold.io/
- 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
- Kenneth's AlphaFold talk at the first AI-driven structural biology community meetup

  https://www.youtube.com/watch?v=jP9Qg1yBGcs&t=834s
- AlphaFold course by VIB on VSC resources: https://elearning.bits.vib.be/courses/alphafold

Flemish Supercomputing Centre (VSC) and NCC Belgium

- VSC-website: https://www.vscentrum.be, and NCC Belgium: https://www.enccb.be/
- Overview of HPC resources at VSC hubs: https://docs.vscentrum.be/en/latest/hardware.html
- Contact: compute@vscentrum.be

Thanks

Wouter Van Putte
Moisés Maestro López

Jasper Zuallaert
Alexander Botzki

Kenneth Hoste
Samuel Moors
Alexander Vapirev

Tim Jaenen