Computational efforts against COVID19: Folding@Home at the WLCG

Antonio Pérez-Calero Yzquierdo

Port d'Informació Científica and CIEMAT

IFAE Pizza Seminar - June 3rd, 2020

PIC
port d'informació
científica

Outline

- Protein structures
- SARS-CoV-2 infection mechanism and the role of certain proteins
- The Folding@Home project
- CERN & WLCG Computing contribution to Folding@Home against COVID19
- Folding@Home against COVID19 highlights

PIC
port d'informació
científica

DISCLAIMER

- I am a **physicist**, PhD in experimental **Particle Physics**, LHCb and CMS experiments at the LHC at CERN. I have been working in **Computing for Particle Physics** (CMS) at **PIC** for the last years of my career

So **why** would I give a talk on COVID19 research, Folding@Home or molecular biology at all?

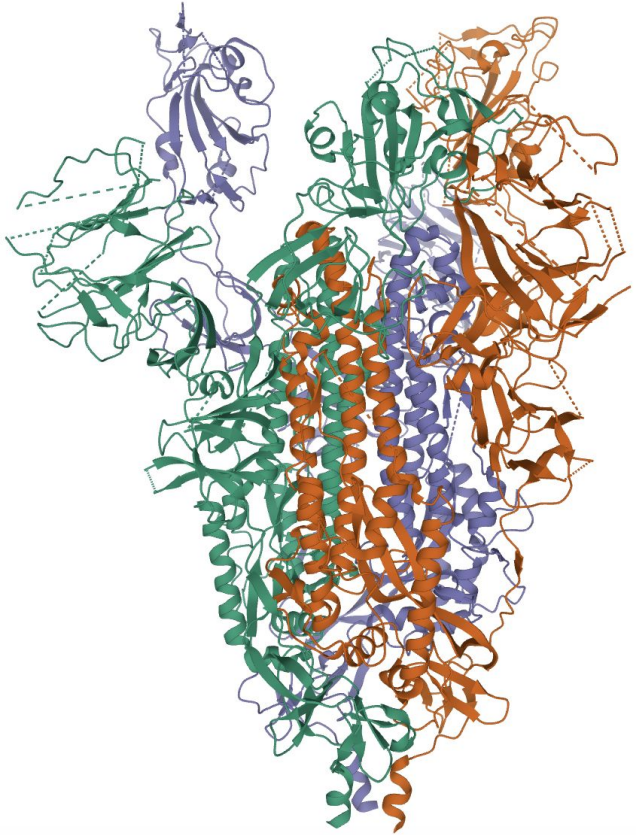
- My involvement in this topic originated from my experience in **Distributed Computing,** my contribution in this area being the adaptation of the **CMS Workload Management** infrastructure to run **Folding@Home** tasks, instead of CMS data, in the LHC **Computing Grid**

I will **humbly** share with you what I have learned about COVID19 while trying to be of some use and contribute in the fight against the pandemic

Apologies in advance for over-simplified & erroneous remarks, please bear with me!

PIC
port d'informació
científica

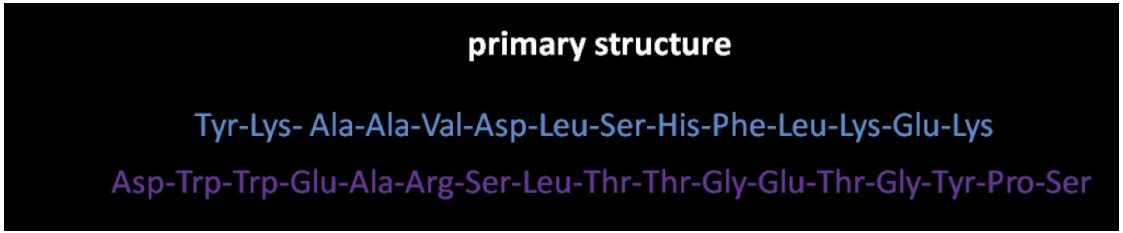
Protein structures



Proteins and their structure

- Proteins compose **structural** and **motor** elements in the cell, and they serve as the **catalysts** for virtually every biochemical reaction that occurs in living things
  - This incredible **array of functions derives from a simple code** that specifies a hugely diverse set of structures.
  - Proteins interactions depend on their ability to **bind to other molecules**. Their **dynamic tridimensional structures** are crucial in determining their **biological functions**.

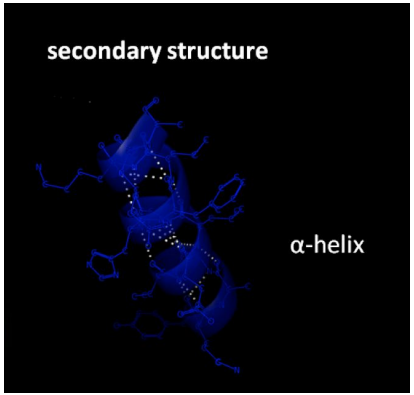
- Going from amino acids to the final 3D structure is a really **complex problem**, as it evolves to its most **energetically favorable conformation**
  - **Primary structure:** linear chain of amino acids, as encoded in the DNA and built by ribosomes (**polypeptides**)

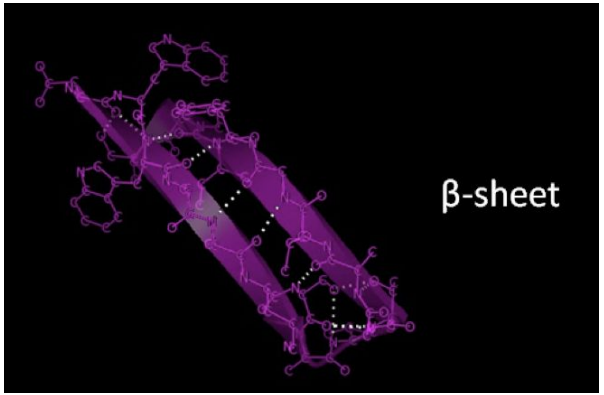


Proteins and their structure

- Proteins compose **structural** and **motor** elements in the cell, and they serve as the **catalysts** for virtually every biochemical reaction that occurs in living things
  - This incredible **array of functions derives from a simple code** that specifies a hugely diverse set of structures.
  - Proteins interactions depend on their ability to **bind to other molecules**. Their **dynamic tridimensional structures** are crucial in determining their **biological functions**.

- Going from amino acids to the final 3D structure is a really **complex problem**, as it evolves to its most **energetically favorable conformation**
  - **Secondary structure**: three dimensional form of local segments of proteins, commonly into **helices** and **sheets**, determined by pattern of bonds (e.g. hydrogen-bonds) between non-consecutive amino-acids

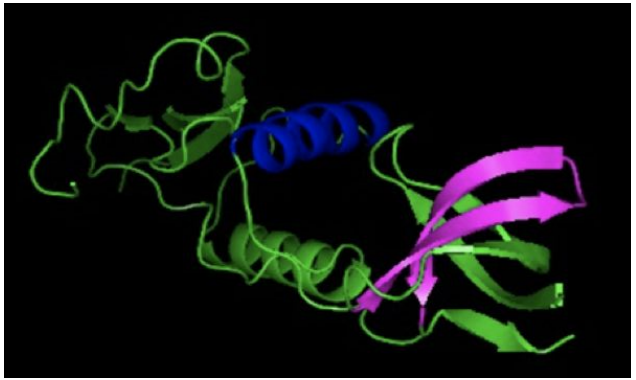




Proteins and their structure

- Proteins compose **structural** and **motor** elements in the cell, and they serve as the **catalysts** for virtually every biochemical reaction that occurs in living things
  - This incredible **array of functions derives from a simple code** that specifies a hugely diverse set of structures.
  - Proteins interactions depend on their ability to **bind to other molecules**. Their **dynamic tridimensional structures** are crucial in determining their **biological functions**.

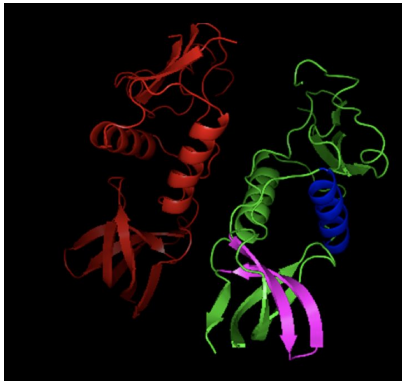
- Going from amino acids to the final 3D structure is a really **complex problem**, as it evolves to its most **energetically favorable conformation**
  - **Tertiary structure**: final tridimensional shape of a **single polypeptide** chain

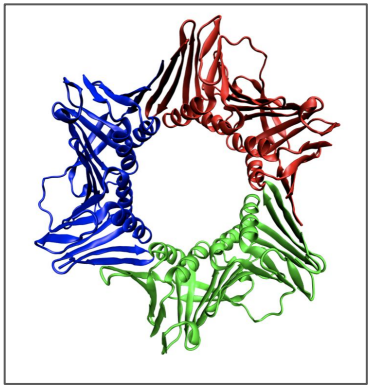


Proteins and their structure

- Proteins compose **structural** and **motor** elements in the cell, and they serve as the **catalysts** for virtually every biochemical reaction that occurs in living things
  - This incredible **array of functions derives from a simple code** that specifies a hugely diverse set of structures.
  - Proteins interactions depend on their ability to **bind to other molecules**. Their **dynamic tridimensional structures** are crucial in determining their **biological functions**.

- Going from amino acids to the final 3D structure is a really **complex problem**, as it evolves to its most **energetically favorable conformation**
  - **Quaternary structure**: final structure of a protein containing **multiple polypeptide chains**





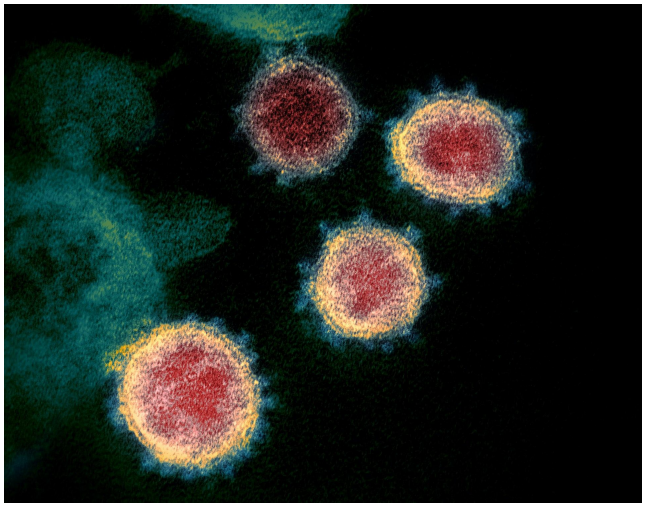
Proteins and their structure

- Proteins compose **structural** and **motor** elements in the cell, and they serve as the **catalysts** for virtually every biochemical reaction that occurs in living things
  - This incredible **array of functions derives from a simple code** that specifies a hugely diverse set of structures.
  - Proteins interactions depend on their ability to **bind to other molecules**. Their **dynamic tridimensional structures** are crucial in determining their **biological functions**.

- Going from amino acids to the final 3D structure is a really **complex problem**, as it evolves to its most **energetically favorable conformation**
  - **Primary structure**: linear chain of amino acids, as encoded in the DNA and built by ribosomes (**polypeptides**)
  - **Secondary structure**: three dimensional form of local segments of proteins, commonly into **helices** and **sheets**, determined by pattern of bonds (e.g. hydrogen-bonds) between non-consecutive amino-acids
  - **Tertiary structure**: tridimensional shape of a **single polypeptide** chain
  - **Quaternary structure**: final structure of a protein containing **multiple polypeptide chains**

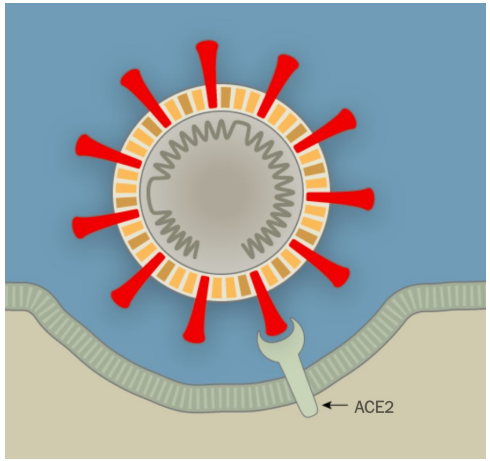
PIC
port d'informació
científica

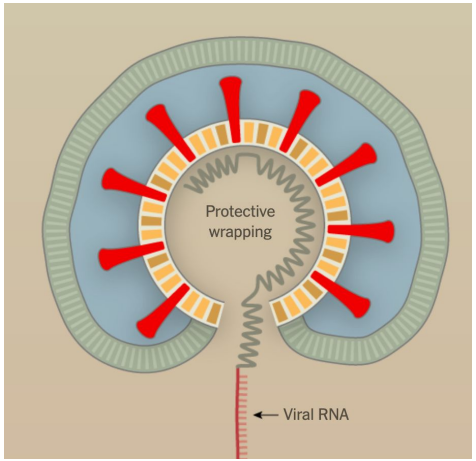
SARS-CoV-2 infection mechanism and the role of certain proteins

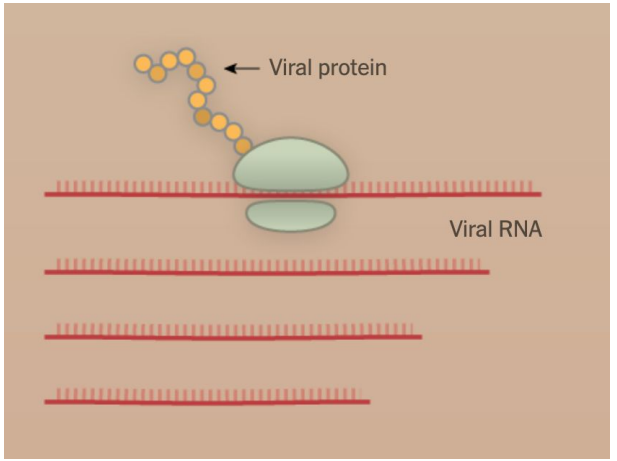


Viral infection (I)

- **Coronavirus Spike proteins** bind to **ACE2 enzyme** in human cells (e.g. lungs), which is involved in blood pressure control
  - Reduction in ACE2 enzyme function as a result of the infection may lead to higher blood pressure and tissue inflammation
  - Variability in the expression of this protein is believed to be at the origin of the severity variations with genetic components, age and sex differences
- Viral and human cell membranes are fused, then **RNA is injected into the host**
- Cells **protein production** mechanism (**ribosomes**) is **hijacked** by the viral RNA
  - SARS-CoV-2 RNA contains ~30k bases, compared to our 3 billion base-pairs

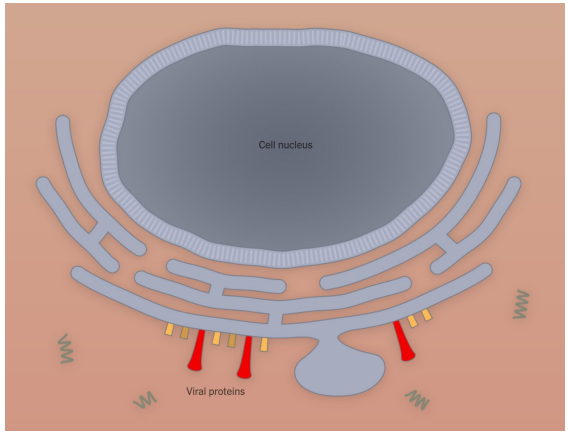


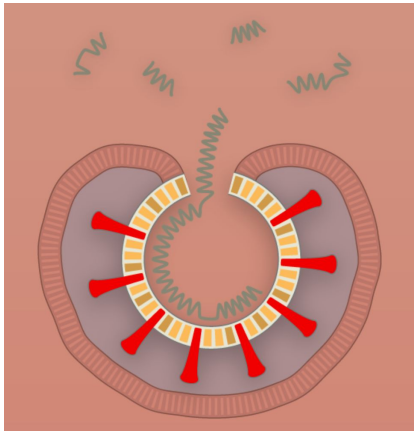


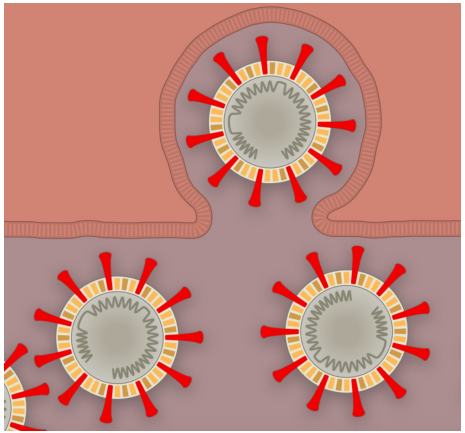


Viral infection (II)

- Infected cell starts producing **viral proteins** and replicas of the **genetic material**
  - Errors in RNA copy produce **viral mutations**
- New **copies of the virus are assembled** inside the infected cell and transported near the cell membrane
- **Viral copies are then released** to the organism
  - Each infected cell can produce up to **100,000 virus replicas in 24h**
  - New replicas can then **continue infecting** other cells or being **expelled from the body** (infecting droplets)

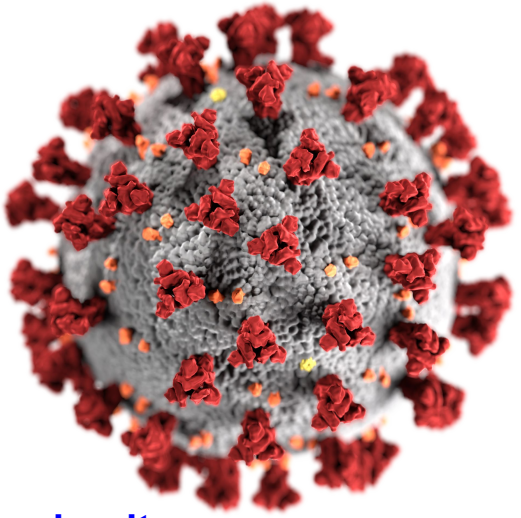






PIC
port d'informació
científica

SARS-CoV-2 proteins



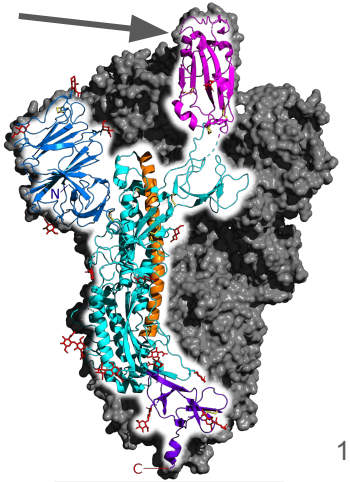
What are the **essential** proteins for SARS-CoV-2 functions?

- **Spike** (**S**) proteins (red) allow the virus to attach to human cells.
- **Envelope** (**E**) proteins (yellow) help it get into human cells.
- **Membrane** (**M**) proteins (orange) give the virus its form
- **Nucleocapsid** (**N**) protein holds the RNA genome in the interior of the virus lipidic membrane (gray)

Some quotes(*) that illustrate for example the **importance of the S protein in fighting COVID19**:

"The virus binds to host cells through its **trimeric spike glycoprotein**"
"(...) authors show that **this protein binds at least 10 times more tightly** than the corresponding spike protein of severe acute respiratory syndrome (SARS)–CoV"
"The CoV **spike (S) glycoprotein is a key target** for vaccines, therapeutic antibodies, and diagnostics"
"characterization of the **prefusion S structure** would provide atomic-level information to guide vaccine design and development."

S1 subunit: ACE2 binding region

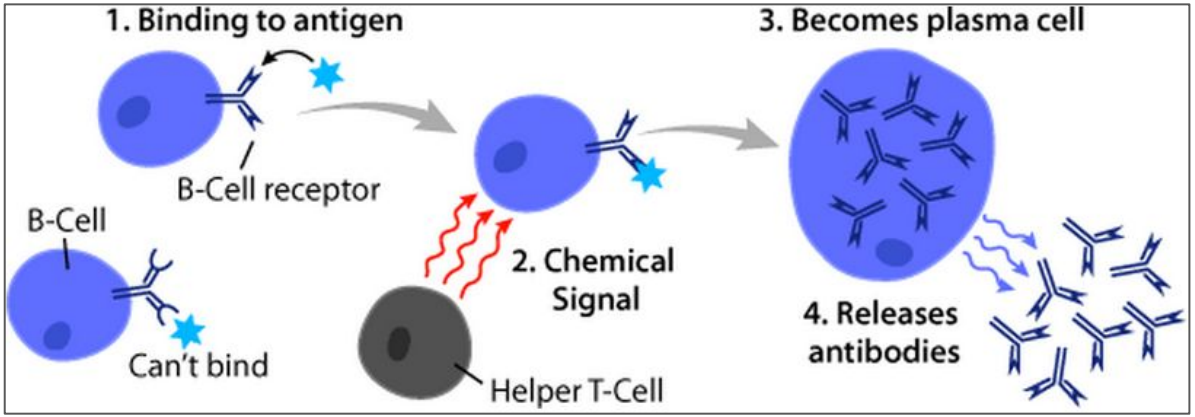


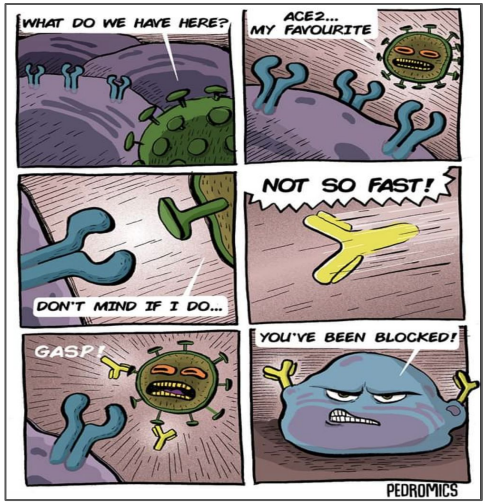
(*)https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164637/

Immune system reaction

The adaptive immune system response is an example of how protein biological function follows 3D structure (and can be inhibited!)

- Helped by T and NK cells, **B-lymphocytes** specialize to produce a single **antibody** as response to a single **antigen**
- Mature B-cells are released into the bloodstream, where they start producing specific **antibodies for S-proteins**.
- Antibodies **bind** to S-proteins in coronavirus membrane **preventing** it from binding to the human cells





PIC
port d'informació
científica

The Folding @ Home Project

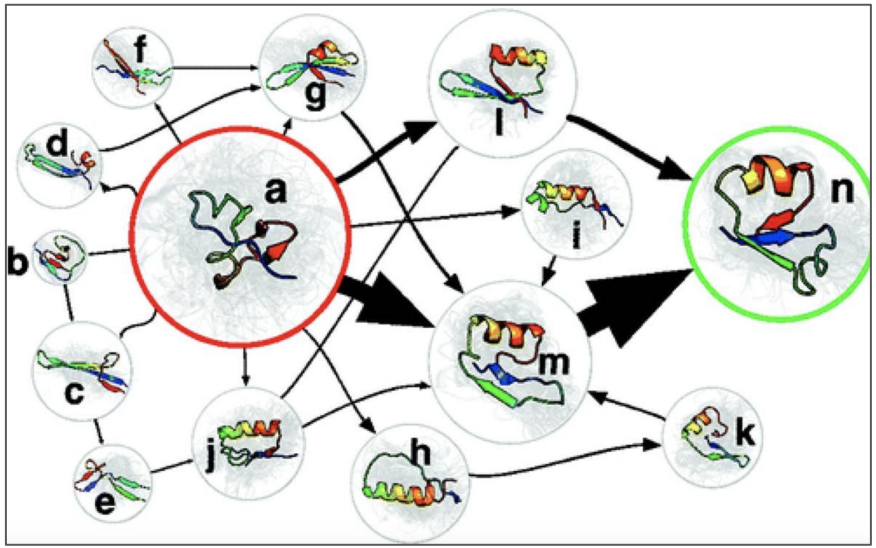
Folding @ Home

- **Folding@Home** (F@H) is a Citizen Science project (implemented as volunteer computing) dedicated to running simulations in order to predict protein folding into their 3D structure
- The project **started in 2000**, and by 2005 they were already involved in multiple fields such as antibiotics, Alzheimer's disease and cancer
- Folding@Home has been involved in multiple studies on **infectious** diseases (Ebola, Chagas), **neurodegenerative** diseases (Alzheimer's, Parkinson's), **diabetes**, and several forms of **cancer** (breast, kidneys)
    - Protein **misfolding** and aggregation often associated with **diseases**

- Some milestones include **computational power** from donors equivalent to 10 PFlops (2013), 100 PFlops (2016) and 1 EFlop by March 2020

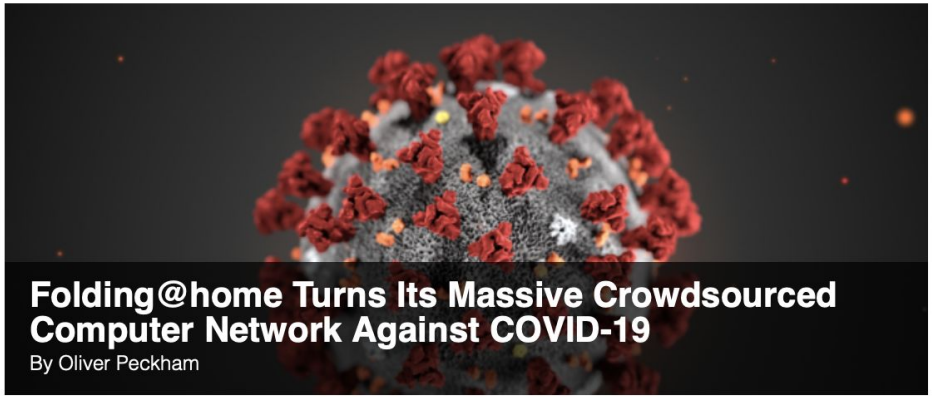
Folding @ Home

- **Folding@Home predicts protein structure** by running simulations on **random paths** of structure evolution, starting from an experimental structure of the protein, passing through intermediate **local configuration energy minima**
- **Markov state model**: determine **microstates** and their **transition prob.** at **discrete time intervals**
  - Initial dataset: seed the process with some static data of unfolded (or partially folded) proteins
  - Microstates: metastable states in the protein conformation, can be clustered into macrostates (compatible within a certain energy resolution)
  - Transition matrix determines rates and probable path of the protein folding



Folding @ Home

- COVID19: coronavirus protein folding boosted to major project soon after the start of the current crisis



Folding@home Turns Its Massive Crowdsourced Computer Network Against COVID-19

By Oliver Peckham

March 16, 2020

For gamers, fighting against a global crisis is usually pure fantasy – but now, it's looking more like a reality. As supercomputers around the world spin up to combat the coronavirus pandemic, the crowdsourced distributed computing platform Folding@home is setting its sights on coronavirus research, spurring a global movement to commit powerful home computers and gaming consoles to the cause.

PIC
port d'informació
científica

Folding @ Home: how to collaborate

CERN AGAINST COVID-19

- **F@H tasks**, known as Work Units (**WUs**), are downloaded from **workload servers** and processed by **F@H clients**, which usually run in the background of **donors** computers, using idle CPU cycles

- **Multiple donors** may join into **teams** (e.g. CMS as part of the CERN & LHC Computing team)

- Donors decide to contribute to a **cause**, each aggregating protein simulation workloads from multiple projects
  - Default cause = ANY
    - presently equivalent to COVID-19 due to its highest priority
    - COVID-19 can be explicitly selected

- The client is **highly configurable**:
  - Run with a number of levels of priority and % of CPU (e.g. on idle desktop CPUs)
  - Use CPUs, taking n_cores max
  - Use GPUs if available
  - Run N WUs then exit, or indefinitely
  - ...

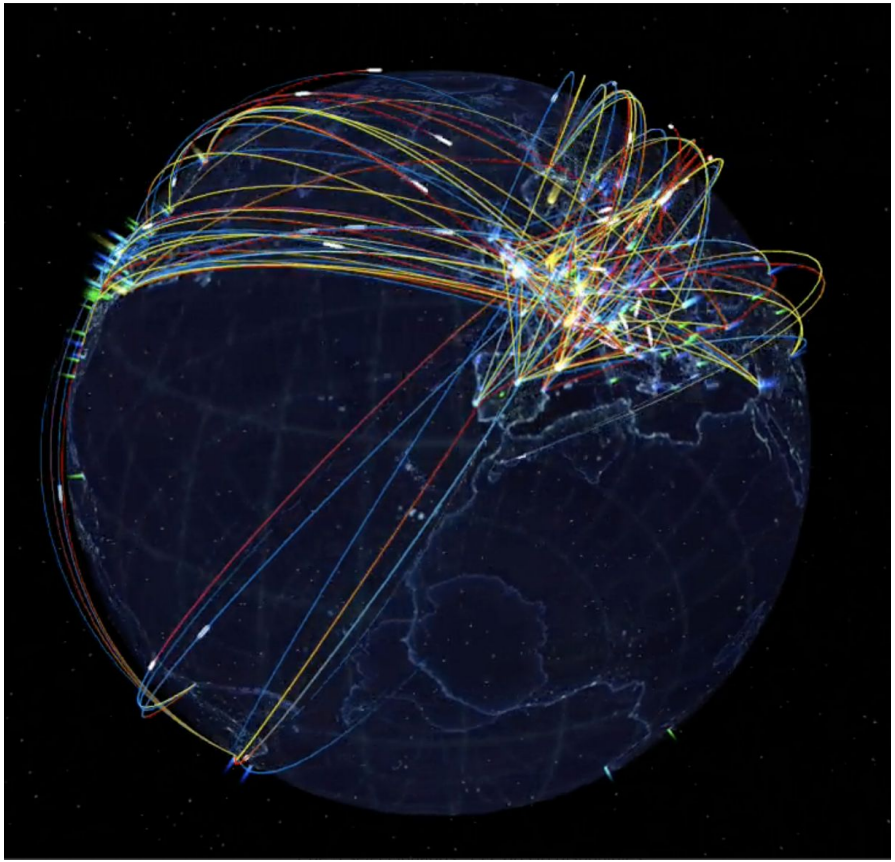
PIC
port d'informació
científica

Folding@Home workloads integration into WLCG

PIC
port d'informació
científica

WLCG
Worldwide LHC Computing Grid

CERN
AGAINST
COVID-19



The Worldwide LHC Computing Grid (WLCG) is a global computing infrastructure whose mission is to provide computing resources to store, distribute and analyse the data generated by the Large Hadron Collider (LHC), making the data equally available to all partners, regardless of their physical location.

Global collaboration

42 countries
170 computing centres
Over 2 million tasks daily
1 million computer cores
1 exabyte of storage

Spanish Tier-1 site for CMS, ATLAS & LHCb at
Port d'Informació Científica

PIC
port d'informació
científica

Folding @ Home at WLCG

- Given the current crisis, CERN created in March a CERN against Covid-19 task force, coordinating multiple aspects of how the CERN community at large could help in the crisis

- As part of this, on the computing front, the **use of CERN and WLCG computing power for Covid-19 related research was endorsed**

- Teams & resources related to CERN and WLCG Computing have been involved since in this effort in multiple ways:
  - Initiatives at national level (e.g. OSG in US, CNAF in IT, PIC in ES)
  - CERN
  - **CMS HLT farm and Grid resources**
  - Similarly ATLAS, ALICE and LHCb

F@H in CMS grid: how

- In order to run it in CMS Grid resources, we have packaged the **F@H client** together with a **job wrapper and config files**, in order to make it a suitable **payload** for the CMS Global Pool slots.

- F@H payloads are submitted as **simple HTCondor jobs** into a CMS schedd queue
  - no other CMS WM service required (although submission via CRAB also tested)

- The **client itself acts as a "pilot"**: once started, a workload server is contacted and a WU is fetched
  - Two layers of late binding: F@H client as payload to pilot, then WU to F@H client

- Configuration in use:
  - Each payload job running a single WU, then exit (1 CMS job = 1 F@H WU),
  - Cause=COVID19
  - Tested **8-core** and **4-core** payloads
  - Running as "backfill" to partially full Global Pool pilots (multicore, min 8-cores)
    - very **selective on sites whitelist**
    - **lower priority** compared to all other CMS workflows

F@H in CMS grid: how

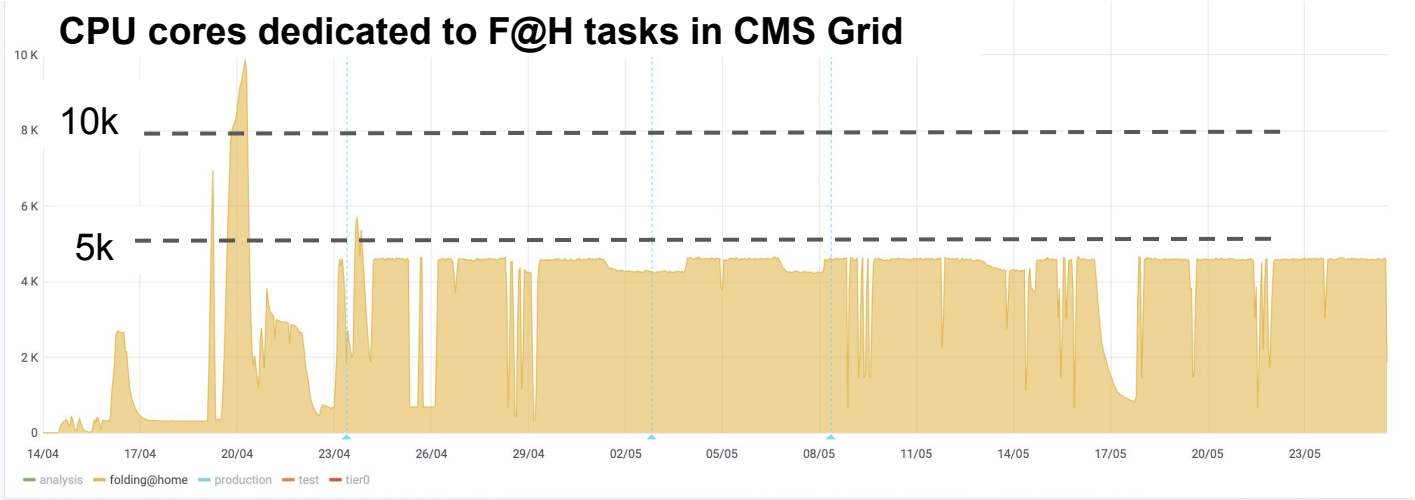
Some aspects that required **a bit of attention**:

- **Monitoring**: required as a prerequisite
  - solved by enabling the “folding@home” job type in the official monitoring
- **Max CPUs:** explicit limit to CMS grid donation
  - simply implemented as max running jobs at HTCondor schedd level
- **Removal of inefficient jobs:** kill jobs when no WU is assigned within initial 20 mins
- **Feeder:** a very simple WM mechanism to enable continuous running mode
  - Implemented in simple cron scripts in the schedd
- **Cleaner**: failing WUs produce log dumps of several GBs, clogging the schedd
  - Simple cleaner cron script

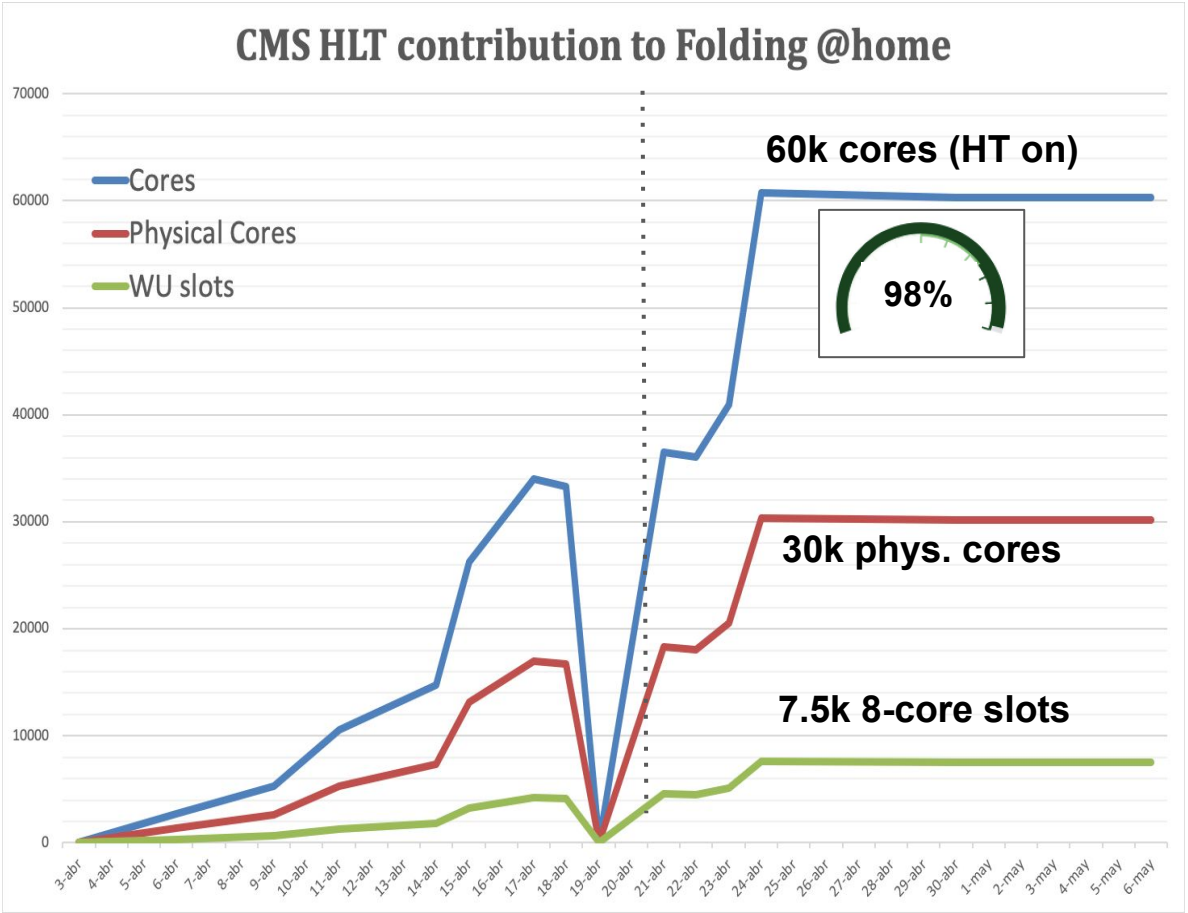
F@H in CMS grid: results

Integration of F@H into CMS grid **technically achieved** since early April, **running in sustained backfill mode with negligible impact on overall CMS activities**

- Initial integration tests of F@Home payloads
- Scaling tests peaking at 10k cores
- Backfill running mode, limited to tiny % of CMS Global Pool total resources



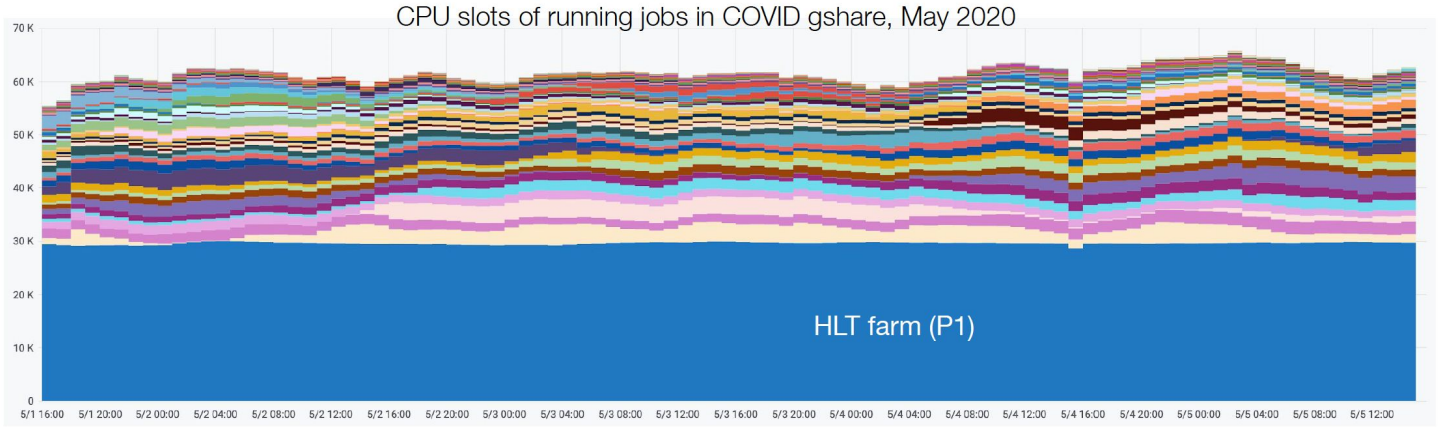
CMS donation to F@H with HLT



<u>Since April 21st</u>: CMS decision to **donate full HLT capacity to F@H with HT enabled**

ATLAS contribution similar to CMS

Stable running



- Since the start of this month, we have stable running with a total of 60k slots
  - Flat 30k from the unpledged HLT at P1
  - Another 30k distributed between the 55 contributing grid sites, representing 10% of their pledge

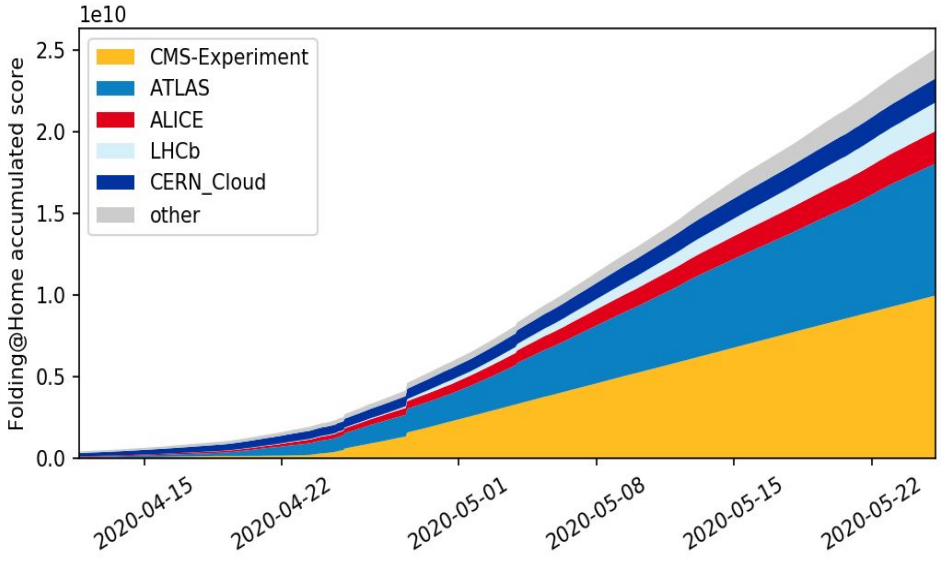
D. South et al., ATLAS and Covid-19, GDB, 6th May 2020

10

WLCG Computing Altogether

CERN & LHC Computing quickly rising to **top** contributors to F@H since this activity started (April)

CERN & LHC Computing team accumulated F@H score



Team: CERN & LHC Computing

Date of last work unit	2020-06-01 21:27:19
Active CPUs within 50 days	1,359,701
Team Id	38188
Grand Score	31,137,092,103
Work Unit Count	8,232,997
Team Ranking	25 of 253746
Homepage	http://public.web.cern.ch/public/
Fast Teampage URL	https://apps.foldingathome.org/teamstats/team38188.html

Team members

Rank	Name	Credit	WUs
29	CMS-Experiment	12,280,780,906	2,402,473
46	ATLAS_CPU	9,848,417,162	2,313,970
293	LHCbHLT	2,281,250,909	339,286
336	ALICE-FLP	2,014,228,852	177,221
426	CERN_Cloud	1,642,060,669	728,133
589	DESY-ZN_GPU	1,265,355,845	9,082
2,827	UC_ATLAS-ML	263,471,027	154,058
3,405	CMSDCS	211,903,168	23,351

PIC
port d'informació
científica

Folding@Home against COVID19 highlights

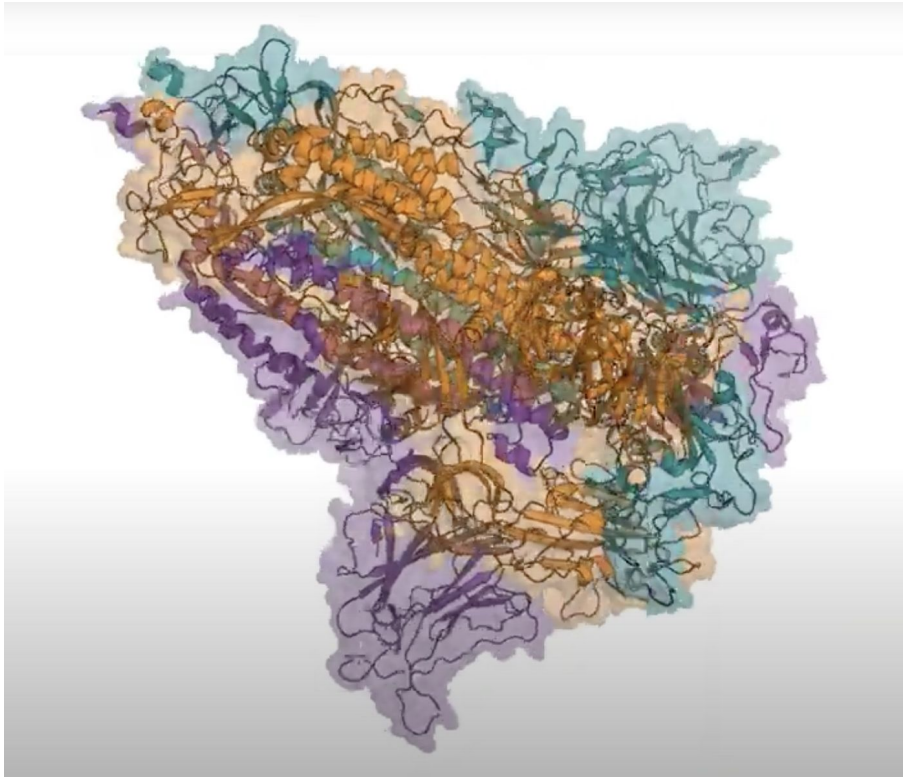
Folding @ Home against COVID19

CAPTURING THE COVID-19 DEMOGORGON (AKA SPIKE) IN ACTION

April 3, 2020
by Greg Bowman

The spike of the SARS-CoV-2 virus (shown below) is one particularly appealing target for designing therapeutics to combat the COVID-19 disease. It is actually comprised of three identical proteins arranged in a circle. Many copies of the spike protrude from the surface of the virus, where they wait to encounter a protein on the surface of many human cells, called ACE2. Binding of a spike to ACE2 initiates a series of events that ultimately allow the virus to enter the human cell. Therefore, therapeutics that bind the spike in a manner that blocks its interaction with ACE2 could provide a valuable means to prevent infection.

Prevent coronavirus **infection** to human cells



Folding @ Home against COVID19

Inhibit effective **viral protein** replication in infected cells

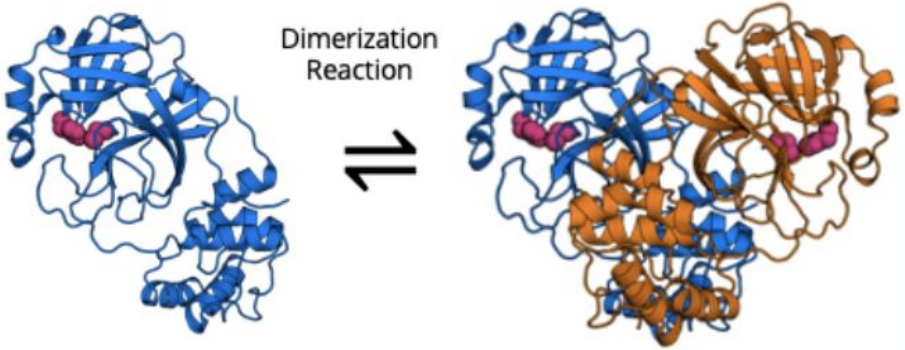
3CL-Protease (3CL-PRO)

What does 3CL-Pro do?

3CL-Pro cuts the immature SARS-CoV2 protein from one piece into many mature viral proteins. It is more active after it assembles into a two-component unit (dimer).

How is F@h helpful here?

We are looking for pockets in the single unit and the dimer (blue and orange) that open, due to protein motions, where a drug could bind. We are also calculating how strongly certain molecules bind to the active site. This would prevent the virus from efficiently replicating. 3CL-Pro is one of the major drug targets across many ongoing studies.



With what project IDs is 3CL-Pro associated?

Compound binding simulations p14350-14399 and p14600-14699
Pocket opening simulations p14582, p14584, p14542, 14592 and p14543

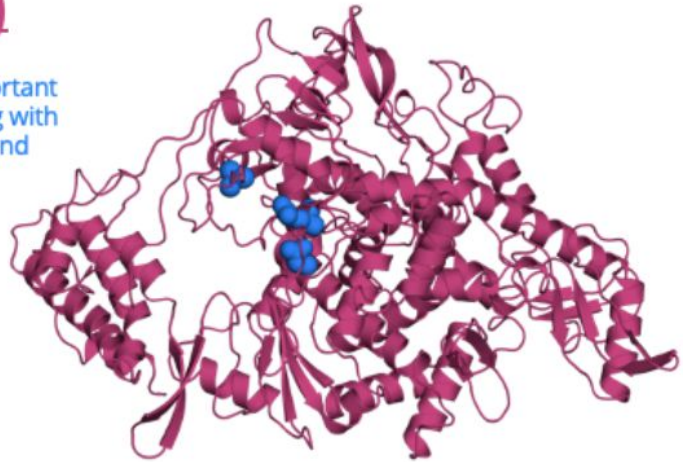
6Y2E

Zhang et al., *Science* 2020
DOI: 10.1126/science.abb3405

Folding @ Home against COVID19

RNA Dependent RNA Polymerase (Nsp12)

A region important for interacting with the genome and Remdesivir



With what project IDs is Nsp12 associated?

CPU: p14412, p16424, p16432, p16500, p16501, p16402
GPU: p14436, p14437

What does Nsp12 do?

Nsp12 copies the SARS-CoV-2 genome to make new viral particles.

How is F@h helpful here?

Nsp12's essential nature make it a great drug target. Drug binding could inhibit the production of new viral genomes and thus viral spread in the body. We are looking for pockets that open, due to protein motions, where a drug could bind. One such example, Remdesivir is in clinal trials. Mutations to Nsp12 are known to make Remdesivir ineffective. F@h can help understand how Nsp12 interacts with this drug and how mutations disrupt that interaction.

Inhibit effective **viral RNA** replication in infected cells

PIC
port d'informació
científica

SUMMARY

Summary & Conclusions

- Protein structures essential to understanding their biological functions
- SARS-CoV-2 infection and replication mechanisms depend on a number of proteins,
  - some of them similar, but more effective, than previous diseases (e.g. SARS-CoV)
- Folding@Home is a volunteer computing project devoted to the study of protein structures in relation to multiple diseases
  - Rapidly growing in resources and almost fully dedicated to COVD19 since the start of the crisis
- CERN community at large involved in many aspects of the fight against COVID19
  - CERN & WLCG Computing contribution to coronavirus studies with Folding@Home
- Folding@Home against COVID19 highlights include studies on how to inhibit cell infection, along with preventing effective coronavirus replication

PIC
port d'informació
científica

SPARE SLIDES

PIC
port d'informació
científica

Abstract

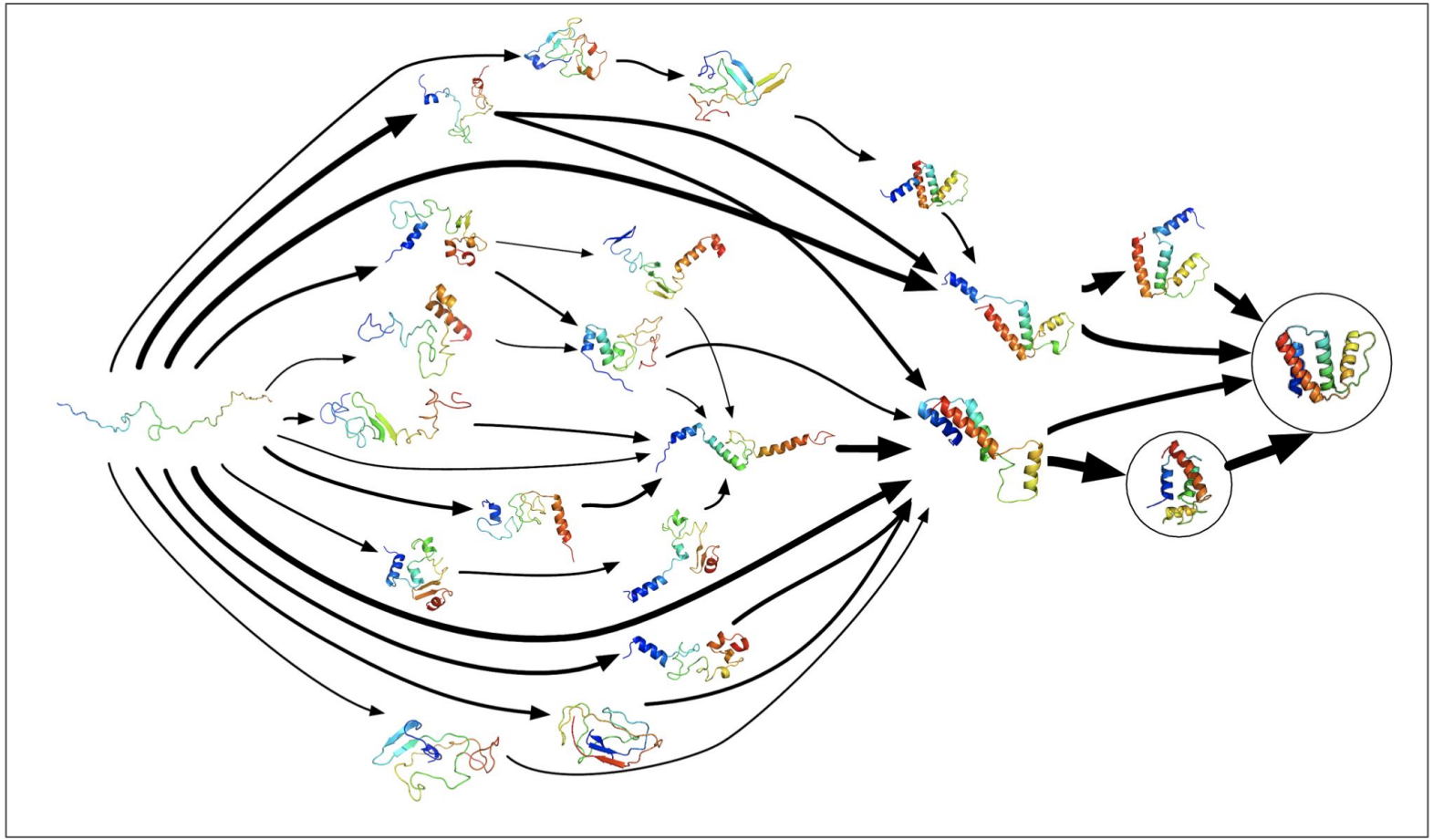
The Folding@Home project is devoted to computational simulations of tridimensional protein structures in relation to multiple human diseases, from breast cancer to ebola. Folding@Home is at the same time an example of a Citizen Science project, in which individuals and institutions get involved in volunteer computing, donating their work and available computing resources in an altruistic way. In the current COVD19 crisis, all efforts have shifted to simulating SARS-CoV-2 proteins in order to better understand infection mechanisms and develop new drugs to prevent or treat it.

Eager to contribute, multiple initiatives have appeared from diverse Science domains, including Particle Physics. Experts in Distributed Computing from CERN and the Worldwide LHC Computing Grid community have adapted their resources, their slot allocation and workload management infrastructures so that Folding@Home tasks can be executed on resources nominally dedicated to the reconstruction, simulation and analysis of LHC collisions. Starting with their contribution to Folding@Home in early April, and quickly growing their contribution in terms of resources and reliability of the infrastructure, the CERN & WLCG Computing team has since become a key donor in the computational front of the fight against COVID19 pandemic.

PIC
port d'informació
científica

Folding @ Home

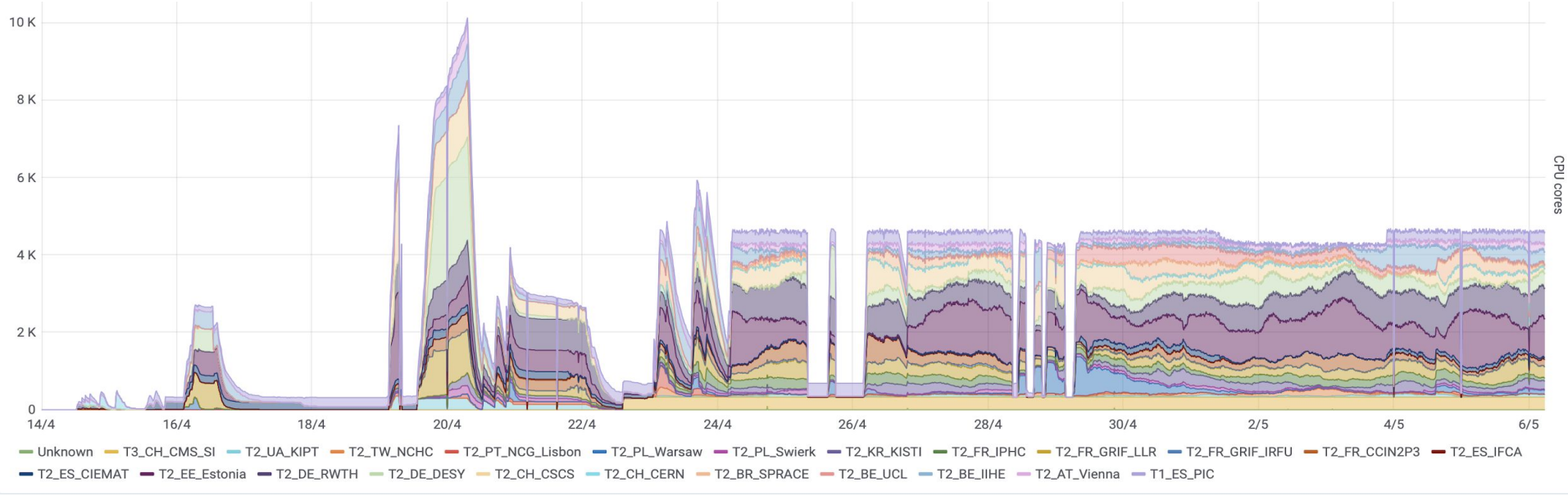
CERN
AGAINST
COVID-19



F@H in CMS grid: results

Distribution of F@H tasks in the CMS Global Pool:

- Running in backfill mode on **CMS sites not involved in other initiatives** (e.g. no US or IT sites included in whitelist)
- Enabling F@H in the CMS Global Pool also a **means for sites willing to donate** a percentage of their CPU with **no initial local efforts** (e.g. **Spanish sites** PIC+CIEMAT+IFCA)
- **Idle CPU in CMS WM+SI hosts** also reconfigured for this purpose



F@H in CMS grid: results

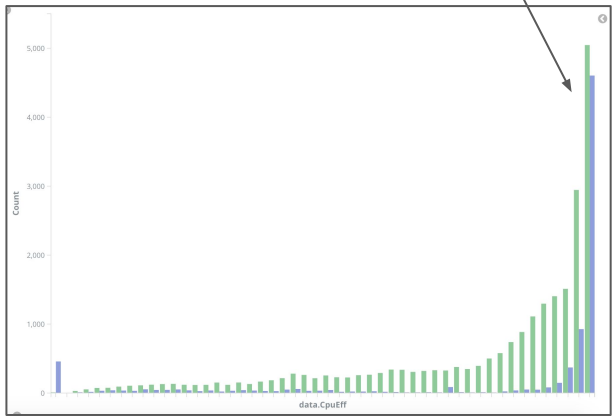
Performance analysis for F@H WUs as payloads:
high CPU efficiency, low memory, short jobs: **ideal for backfill**

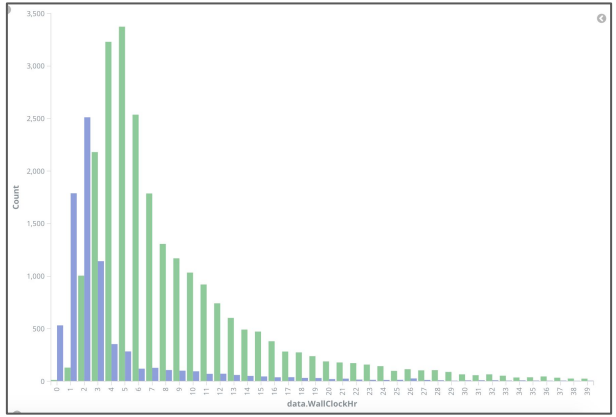
4-core **8-core**

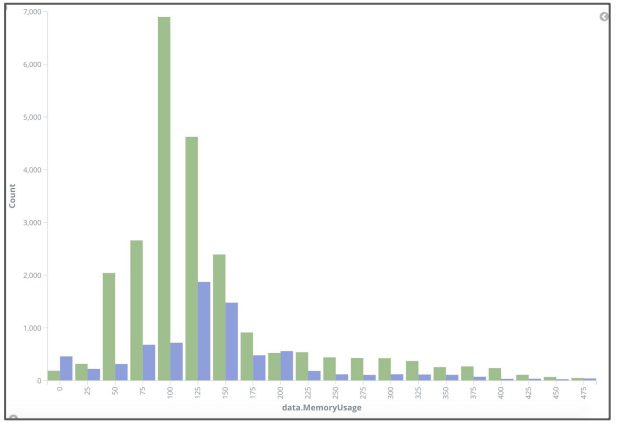
99% for 8-core

2 - 3h for 8-core WUs
5 - 6h for 4-core (but longer tails)

Very low memory usage compared to our workflows: 100 - 125 MB







CPU eff (%)

Walltime (h)

Peak memory (MB)

References (I):

- https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome_coronavirus_2
- https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/
- SARS-CoV-2 complete genome: https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3
- The Architecture of SARS-CoV-2 Transcriptome https://www.sciencedirect.com/science/article/pii/S0092867420304062
- Coronavirus: https://www.nytimes.com/2020/04/01/health/coronavirus-illustration-cdc.html
- Infection: https://www.nytimes.com/interactive/2020/03/11/science/how-coronavirus-hijacks-your-cells.html
- Desarrollo y tipos de vacunas (ElPais): https://elpais.com/elpais/2020/05/30/ciencia/1590828979_735960.html
- Respuesta inmune (ElPais): https://elpais.com/elpais/2020/04/20/ciencia/1587379836_984471.html
- B lymphocytes: https://en.wikipedia.org/wiki/B_cell
- INFORME DEL GRUPO DE ANALISIS CIENTÍFICO DE CORONAVIRUS DEL ISCIII: https://www.conprueba.es/sites/default/files/informes/2020-06/FACTORES%20DE%20RIESGO%20EN%20LA%20ENFERMEDAD%20POR%20SARS-CoV-2%20%28COVID-19%29_2.pdf
- ACE2 enzyme: https://en.wikipedia.org/wiki/Angiotensin-converting_enzyme_2
- Male-Female variations: https://elpais.com/ciencia/2020-05-17/puede-una-sola-enzima-explicar-por-que-el-coronavirus-mata-mas-a-hombres-que-a-mujeres.html
- ¿Por qué la Covid-19 puede llegar a ser mortal en unas personas y en otras no causar síntomas? https://www.vozpopuli.com/altavoz/next/COVID19-llegar-mortal-personas-sintomas_0_1348366661.html

References (II)

- Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164637/
- Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102550/
- Protein structure: https://en.wikipedia.org/wiki/Protein_structure_prediction
- Protein structure: https://www.nature.com/scitable/topicpage/protein-structure-14122136/
- Protein structure prediction software: https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software
- Protein data bank: https://www.rcsb.org
- COVID-19/SARS-CoV-2 Resources: https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52
- Wash your hands! https://pdb101.rcsb.org/learn/videos/fighting-coronavirus-with-soap

References (Folding@Home)

- Folding@Home timeline: https://foldingathome.org/project-timeline/
- Folding@Home and COVID19: **https://foldingathome.org/covid19/**
- **https://www.hpcwire.com/2020/03/16/foldinghome-turns-its-massive-crowdsourced-computer-network-against-covid-19/**
- https://foldingathome.org/2020/04/03/capturing-the-covid-19-demogorgon-aka-spike-in-action/
- https://foldingathome.org/2020/03/30/covid-19-free-energy-calculations/
- https://foldingathome.org/2020/05/25/going-after-the-mysterious-sars-cov-2-envelope-protein/
- https://mujeresconciencia.com/2020/05/31/una-animacion-de-las-maravillas-del-mundo-molecular/?fbclid=IwAR1eJT-smCQN7p1J7VjCf8j7MOWMvYblVXQmGu57Pd97BOgeU46QdFN67mM