EXSCALATE 4COV

Exscalate4CoV: Towards an Exascale-Ready Docking Platform Targeting Urgent Computing against Pandemics

Keynote by
Cristina Silvano
Politecnico di Milano

POLITECNICO
MILANO 1863

MICRO 2020, 53rd IEEE/ACM International Symposium on Microarchitecture

Outline

- HPC-acceleration of Drug Discovery Pipelines
- Exscalate Platform
- Exscalate4COV European Project
- Ongoing and Future Activities

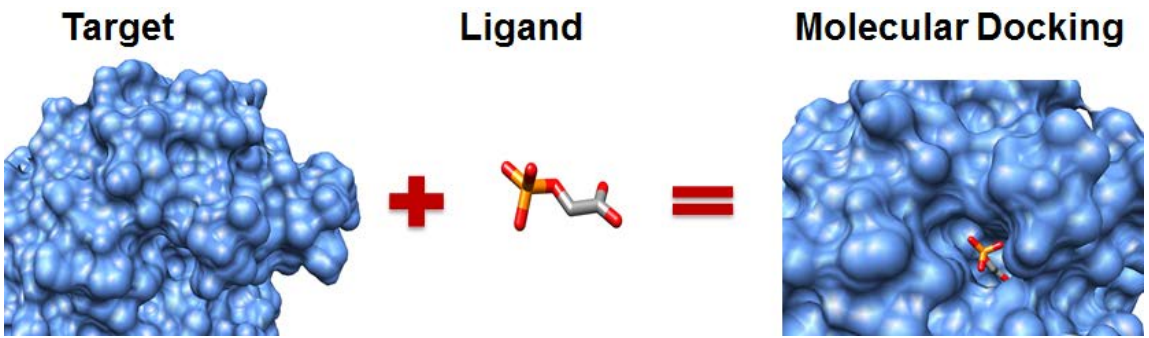
Urgent Computing against COVID-19 Pandemics

- The urgent search for therapies to fight the COVID-19 pandemic has imposed **supercomputers as technology enablers** to explore the wide chemical space of potential drug compounds faster than ever before.
- The COVID-19 pandemic has encouraged many researchers around the world to rethink the **in silico** computational phase of the **drug discovery process** (called ***virtual screening process***).

Need of HPC for Drug Discovery

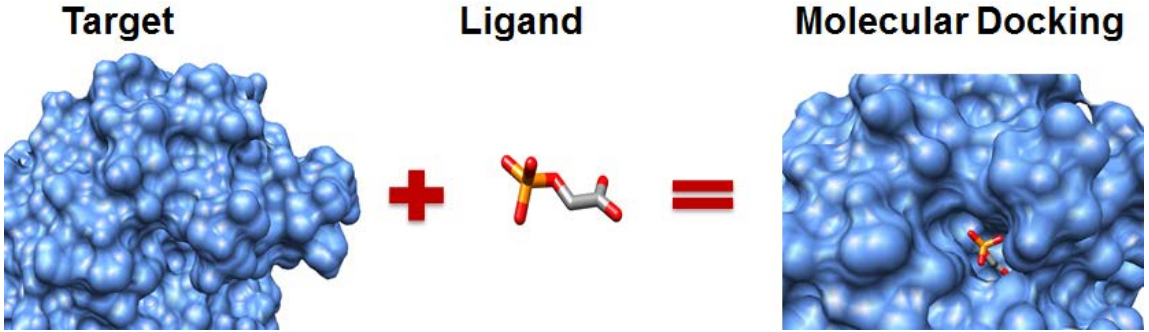
Molecular Docking for Virtual Screening

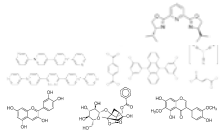
- The **molecular docking** problem consists of predicting **in silico** how well a small molecule compound (called **ligand**) interacts with a given 3D protein target such as a viral protein.



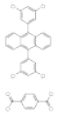
Molecular Docking for Virtual Screening

- The **molecular docking** problem consists of trying to predict **in silico** how well a small molecule compound (called **ligand**) interacts with the given 3D protein target such as a viral protein.



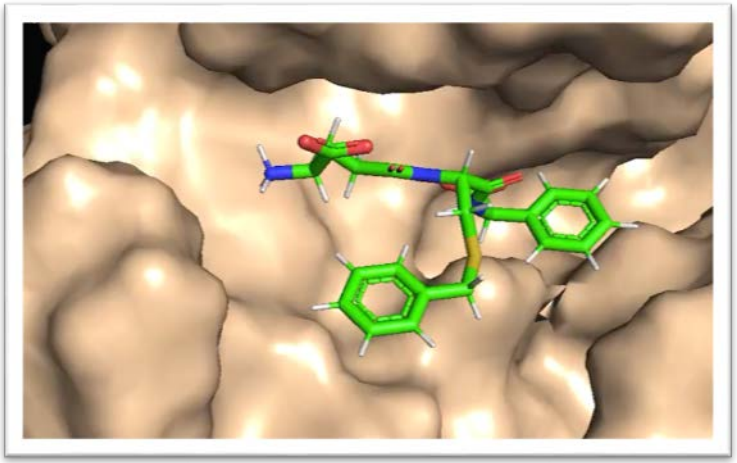


- Computational approaches are much faster and cheaper than **in vitro** approaches for **exploring** a large number of combinations of ligands and proteins and **filtering** the chemical space.

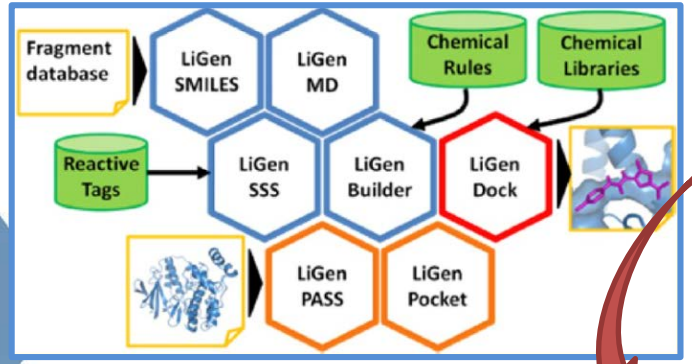


Molecular Docking for Virtual Screening

- **Molecular Docking** is the method to calculate the best 3D position of a small molecule (**ligand**) when bounded to the target viral protein.
  - Looking for shape complementarity
- The search is focused on some regions (**active binding sites** or **pockets)** that affect the chemical function of the protein
- **Scoring** function is used to evaluate the binding energy as a measure of the strength of the interactions and to rank the large set of ligands and poses.



Virtual Screening in ANTAREX HPC H2020 EU Project



LiGen software designed by Dompé Pharma & CINECA

Accelerate and tuning applications to scale out on energy and resource efficient HPC systems

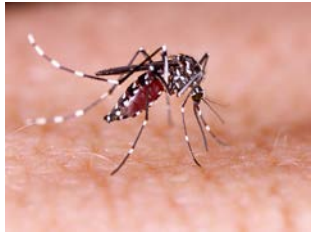
Chemical space

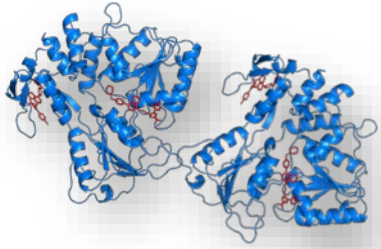
- Chemical Space: 500 Bio
- Peptide library: 120 Mio
- Commercial libraries: 6 Mio
- Safe-in-man library: 9 K

Exascale-ready Virtual Screening for Urgent Computing

ANTAREX 4ZIKA The Zika virus experiment

- Dompé generated a virtual chemical space of **1.2** Billion compounds and selected **26** binding sites of **5** Zika proteins
- LiGen application was deployed and scaled to occupy the full size Marconi A2 system at CINECA:
    - >3500 nodes Intel Xeon Phi KNL 7250 @1.40GHz
    - About 1 M Threads on **10 PetaFLOPS MARCONI**
- **January 2019:** *Run of one of the largest virtual screening experiments ever launched in terms of computational threads* ***(around 1 Million)*** *and size of the compound database screened* ***(1.2 Billion)***
- Run time to solution: 3.2 h for 1 out of 26 sites (run in Jan 2019)
    - ***Total Time to Solution: 3.5 days (84 h) for 26 sites***





periment

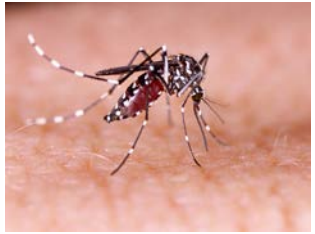
European Commission > Strategy > Shaping Europe's digital future > Projects >

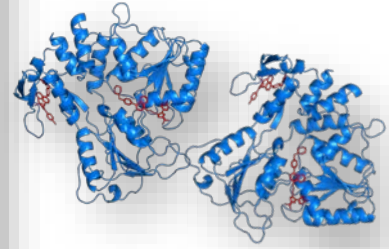
Shaping Europe's digital future

PROJECTS STORY | 23 April 2019

High performance computing used to develop new drugs against the Zika virus

ANTAREX, an EU-funded project, used the Marconi supercomputer to accelerate drug discovery.



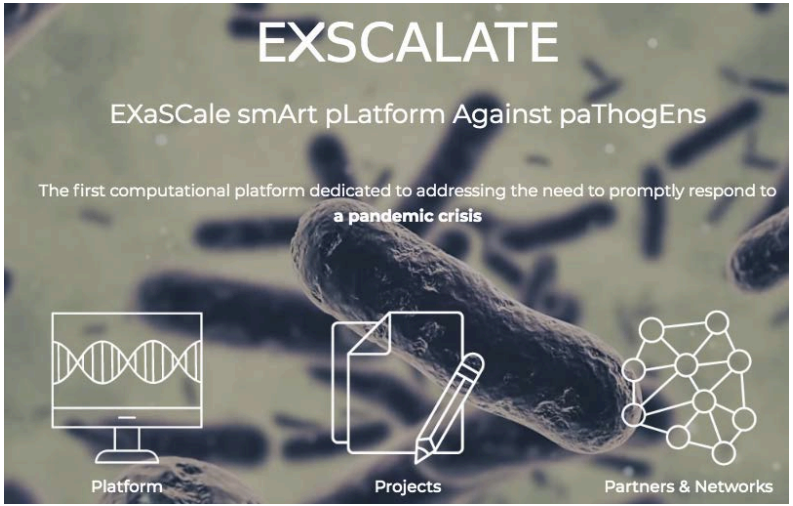


Experiment website:
https://www.exscalate.eu/en/projects.html#Antarex4Zika

EXaSCale smArt pLatform Against paThogEns

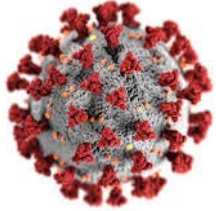
- ***April 2019:*** *Announced by Dompé as the first CADD platform for urgent computing to accelerate the response in case of pandemic crisis caused by viruses and multidrug-resistant bacteria.*

  - **In 2019,** further developments of LiGen ultra-fast virtual screening SW including the ANTAREX code base
  - Extension of the virtual chemical library up to **500 Billion ligands**



www.exscalate.eu

The beginning of COVID-19 pandemic outbreak



- **4 Jan 2020:** WHO tweeted on a cluster of pneumonia cases in Wuhan
- **11 Jan 2020**: Chinese media reported the first death from the novel coronavirus
- **23 Jan 2020:** Lockdown imposed in the city of Wuhan and Hubei region in China
- **30 Jan. 2020**: European Commission launched the First Emergency Call for research projects against the coronavirus.
- **12 Feb. 2020:** Exscalate4COV submitted to the Emergency Call and started the setup of first simulation activities.
- **21 Feb. 2020:** First patients infected in Italy followed by first red zones set up in the region around Milan in Italy.
- **4 March 2020:** Exscalate4COV project was granted with 3 Meur from the First EC Emergency Call (overall 18 projects were funded)
- **9 March 2020:** National lockdown imposed in Italy

EXSCALATE 4COV

The Exscalate4Cov (E4C) project aims to exploit the Exscalate drug discovery platform along with the most powerful supercomputing resources in Europe to empower smart in silico drug design.

Project Coordinator: Dr. Andrea Beccari, Dompé Pharmaceuticals SpA

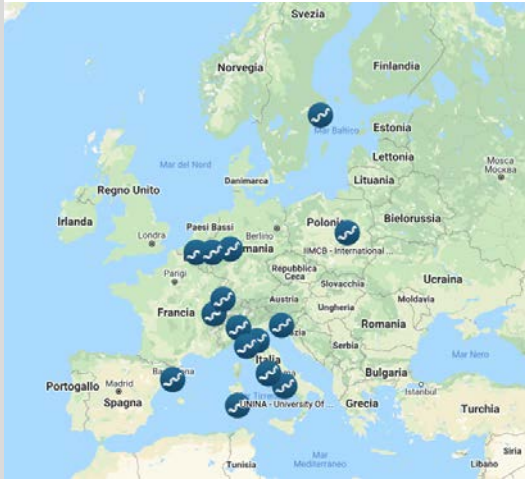
EXSCALATE 4COV

E4C is a fully integrated pipeline of in silico modeling and simulations followed by in vitro experimental validation to speed up the search for active compounds to be tested in clinical trials as new treatments for COVID19.

Project Coordinator: Dr. Andrea Beccari, Dompé Pharmaceuticals SpA

Exscalate4COV Consortium

- E4C project is led by **Dompé Pharma** & aggregates **18 EU partners** including 3 supercomputing centers: BSC, CINECA, FZ Juelich.
- **Main goal:** To develop an integrated pan European drug discovery infrastructure

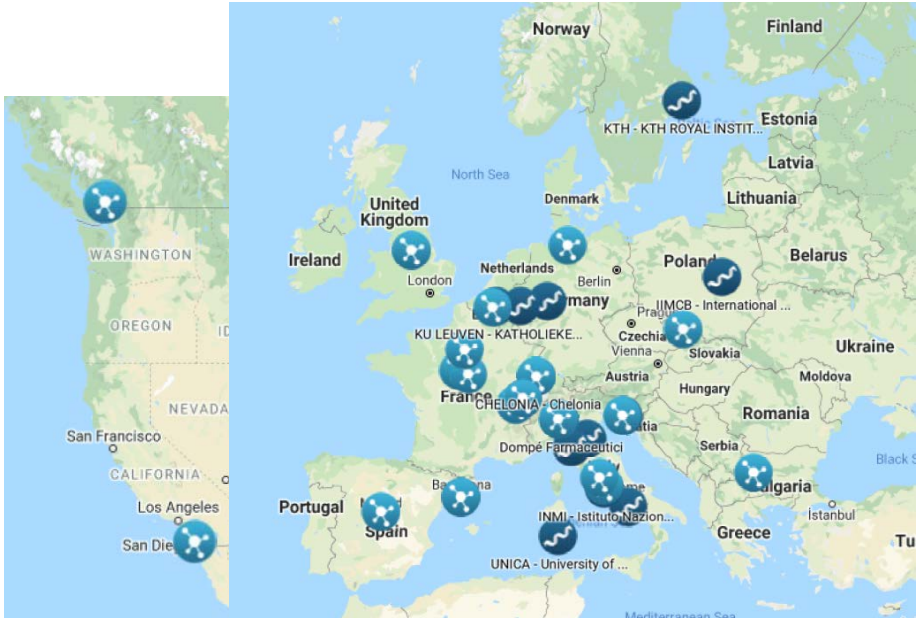


External Collaborations: the League & the Drugbox

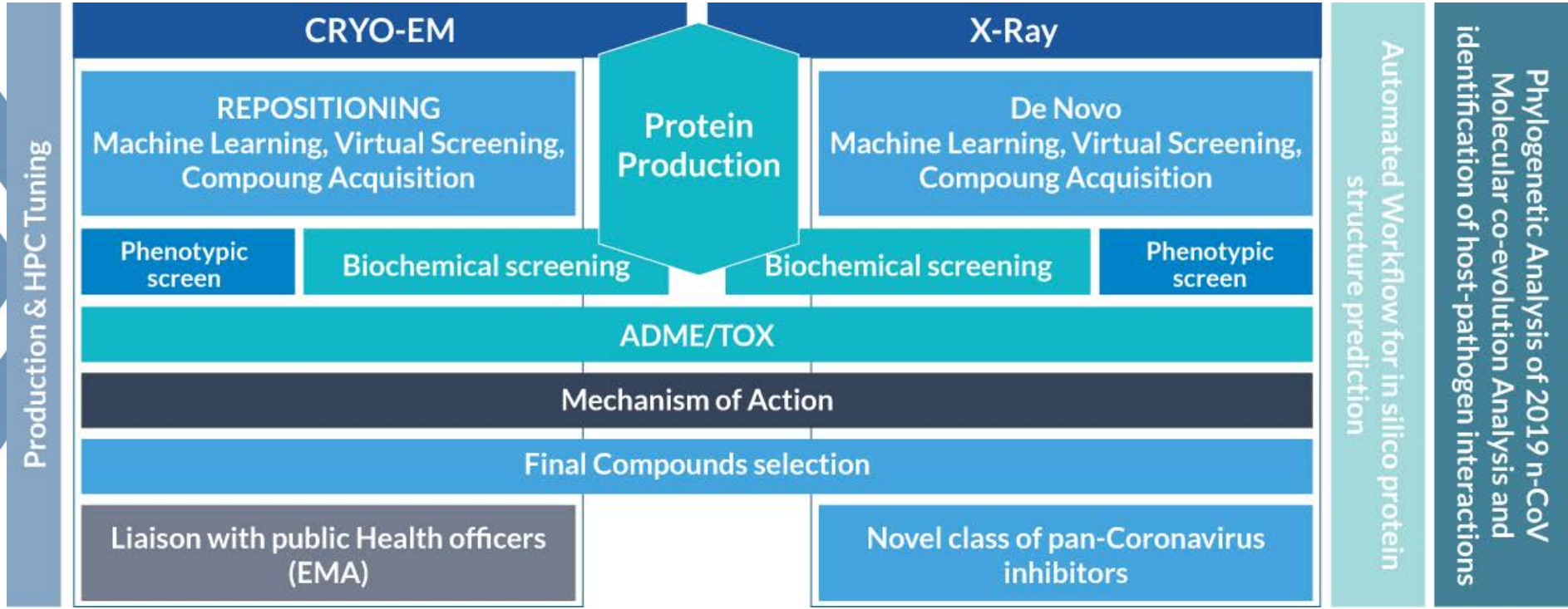
- **The League:** E4C includes more than 15 research centers and companies:
  - ENI, IT4Innovations, ESTECO, CECAM, Nanome, Dwave, Univ. of Basel, Dassault Systemes, Univ. Tuscia, GPCRmd, SAS, …

https://www.exscalate4cov.eu/league.html

- **The Drugbox** crowdsourcing initiative supported by EC and EFPIA (European Federation of Pharma Companies) was launched to collect molecules:
  - Pierre Fabre, Alfasigma, Almirall, Axxam, ESTEVE Pharma, Lilly, Sanofi, GreenPharma, Usheffield, CSIC Inst. of Food Science Research, Institut Cochin, Sofia Univ., ...



E4C Drug Discovery Workflow



E4C Task Force on High Performance Computing

Partners:

POLITECNICO MILANO 1863

CINECA

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

JÜLICH Forschungszentrum

INFN Istituto Nazionale di Fisica Nucleare

KTH VETENSKAP OCH KONST

From the "League":

Eni

IT4Innovations national supercomputing center

More than 60 PetaFlop/s aggregated!

Main goals:

- To setup the supercomputing infrastructure to run experiments
- Optimize porting and tuning application codes for Computer Aided Drug Discovery
- Accelerate the EXSCALATE virtual screening platform
- Accelerate the 3D protein structure resolution

EXSCALATE 4COV

June 22, 2020: Exscalate4Cov project can now count on the two most powerful supercomputers in Europe to empower smart in silico drug design.

Task Force 4: Coordinated by Politecnico di Milano

HPC-5 at ENI (Italy)

- **HPC-5 System Architecture:**
  No. 1 in Europe & No. 6 worldwide
- **Processors:** Xeon Gold 6252 24C 2.1 GHz
- **Cores:** 669 760
- **Accelerators:** NVIDIA Tesla V100
- **RAM:** 349 440 GN
- **Linpack Performance:** 35.5 PetaFlop/s
- **Peak Performance:** 51,7 PetaFlop/s
- **Internal Network:** Mellanox HDR Infiniband
- **Power:** 2,25 MW

MARCONI-100 at CINECA (Italy)

- **MARCONI-100 System Architecture:**
  No. 2 in Europe & No. 9 worldwide
- **Nodes:** 980
- **Processors:** 2x16 cores IBM POWER9 AC922 (Whiterspoon)/node at 2.6(3.1) GHz
- **Cores:** 347,776
- **Accelerators:** 4 x NVIDIA Volta V100 GPUs/node, Nvlink 2.0, 16GB
- **RAM:** 256 GB/node
- **Linpack Performance:** 21.6 PetaFLOPS
- **Peak Performance**: 29 Pflop/s Linpack
- **Internal Network:** Mellanox EDR Infiniband DragonFly++
- **Disk Space:** 8PB (raw) GPFS of local storage
- **Power:** 1,48 MW

Project Workflow: Phase 1

- **Drug Repurposing (Phase 1)**
  - To get a fast response to infected population by using safe-in-man drugs
  - Library of more than **400K molecules** (including 10K safe-in-man drugs and other natural products)
  - Each compound to be docked geometrically and scored
  - **Main outcome:** In June 2020, **Raloxifene** proposed to EMA for starting clinical trials after in-vitro tests showed it inhibits Sars-COV2

Project Workflow: Phase 1

- **Drug Repurposing (Phase 1)**
  - To get a fast response to infected population by using safe-in-man drugs
  - Library of more than **400K molecules** (including 10K safe-in-man dr
  - Each co
  - **Main o**
    for star
    Sars-CO

European Commission

EU Link

Home > Press corner > Coronavirus: Using European supercomputing, EU-funded resear

Available languages: English

Press release | 18 June 2020 | Brussels

Coronavirus: Using European supercomputing, EU-funded research project announces promising results for potential treatment*

EXSCALATE 4COV

Project Workflow: Phase 2

- Drug Repurposing (Phase 1)
  - Library of more than **400K molecules**
- **De-novo Drug Design (Phase 2)**
  - Huge de-novo chemical space up to **500 Billion molecules**
  - Each compound to be docked geometrically and scored
  - Only top-ranked compounds filtered to be passed to the next phases of the drug discovery pipeline

Challenge of the 2nd Phase of Virtual Screening

- To **screen** a **huge number of small molecules (ligands)** in a **reasonable time** on the **available resources**

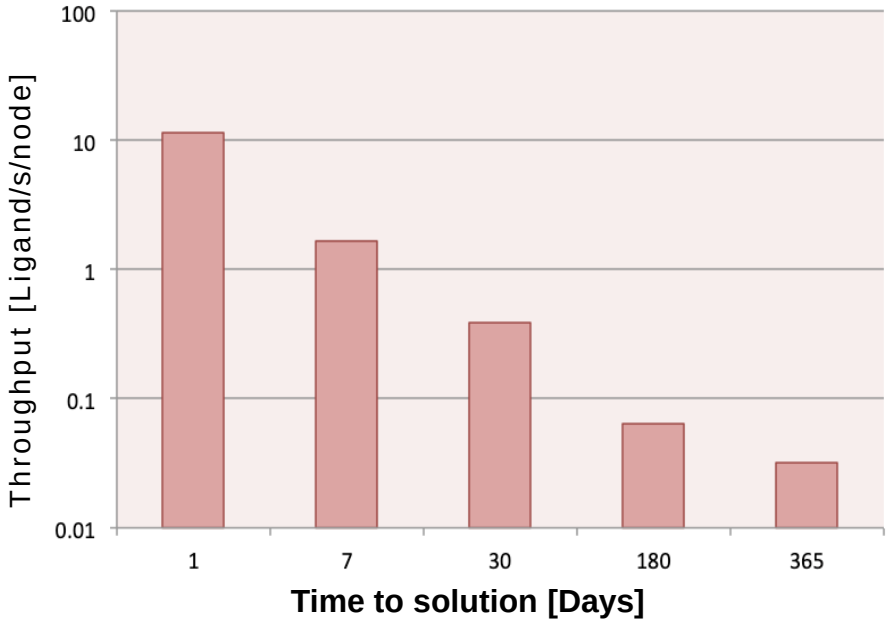
Time to Solution × **Available Resources** = **Processing Time per Ligand** × **Size of the Chemical Space**

Challenge of the 2nd Phase of Virtual Screening

Time to Solution ✖ Available Resources = Processing Time per Ligand ✖ Size of the Chemical Space

ANTAREX 4ZIKA

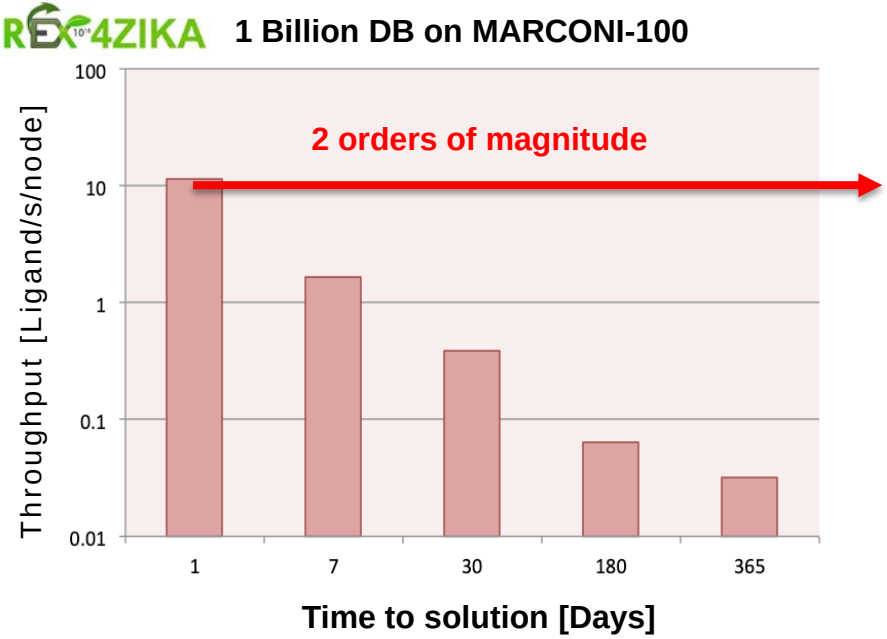
1 Billion DB on MARCONI-100

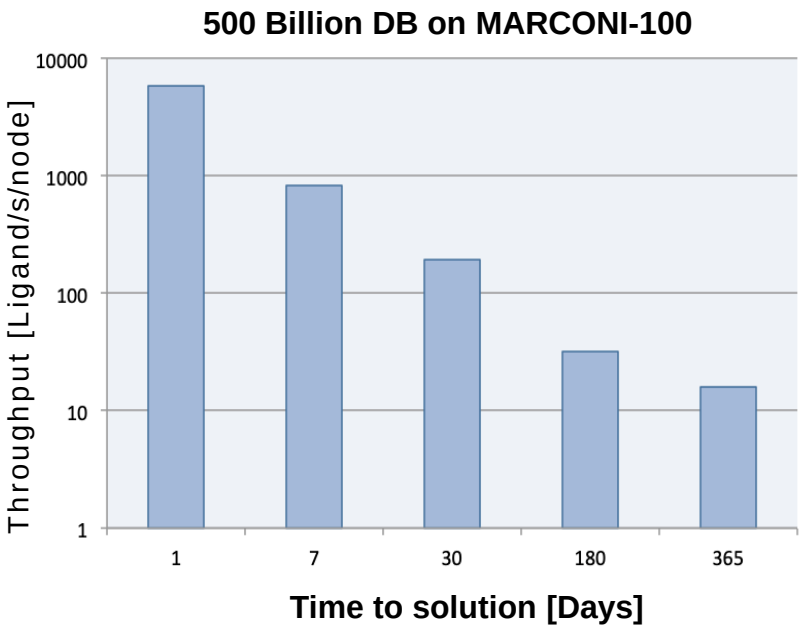


EXSCALATE 4COV

Challenge of the 2nd Phase of Virtual Screening

Time to Solution ✖ Available Resources = Processing Time per Ligand ✖ Size of the Chemical Space

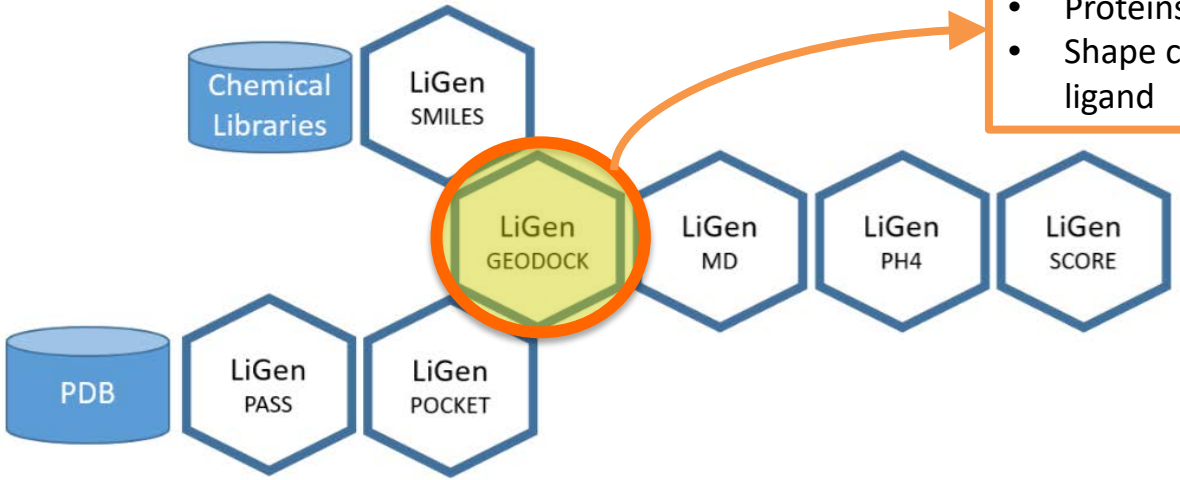




EXSCALATE Platform

High Performance Virtual Screening Platform:

- Tangible Chemical Library
- LIGEN: Exascale-ready modular CADD software

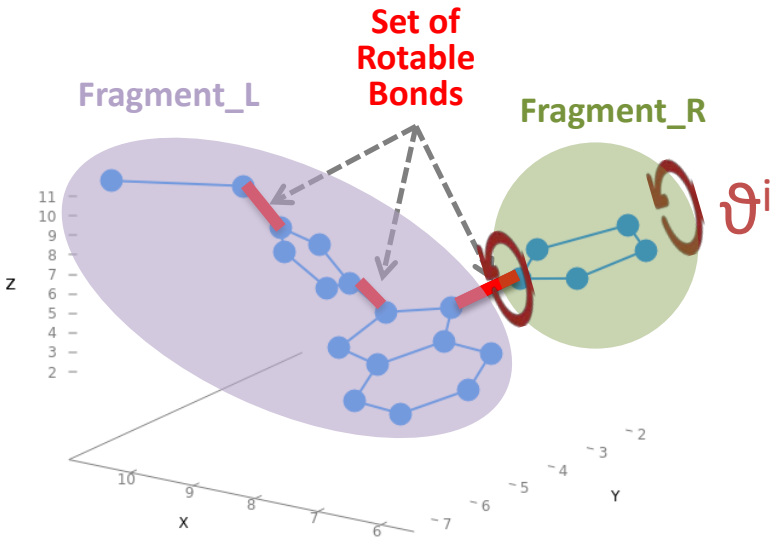


"Relaxed" Lock-and-Key Model

- Proteins as rigid bodies
- Shape complementarity with a flexible ligand

Molecular Docking: The Flexibility Challenge

- **Binding site (pocket) flexibility** managed as:
  - Multiple conformational states of the protein derived by MD simulations
  - Different conformations of the protein bring to **different pockets**
- **Ligand flexibility on rotatable bonds**



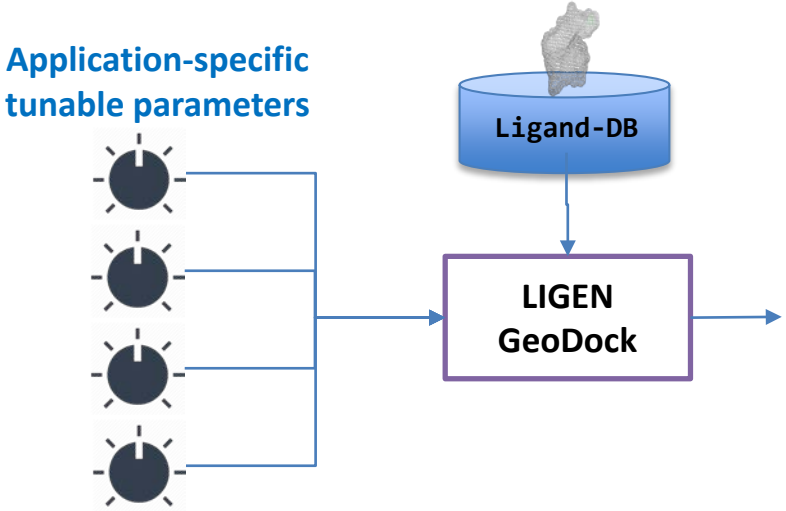
Ligan rigid roto-translation +

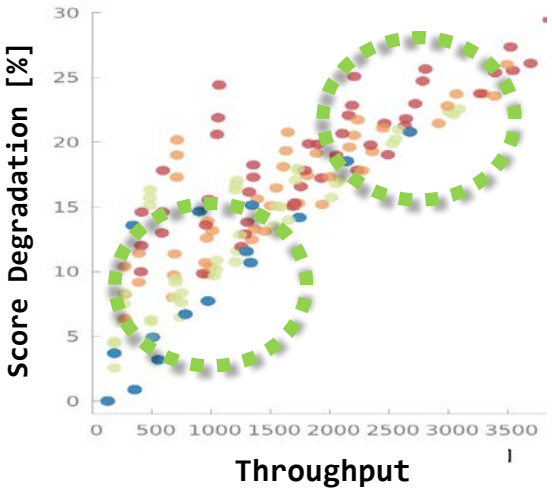
$$\vartheta=[\vartheta^1,\ldots,\vartheta^N]$$

EXSCALATE 4COV

Tuning Throughput vs. Accuracy

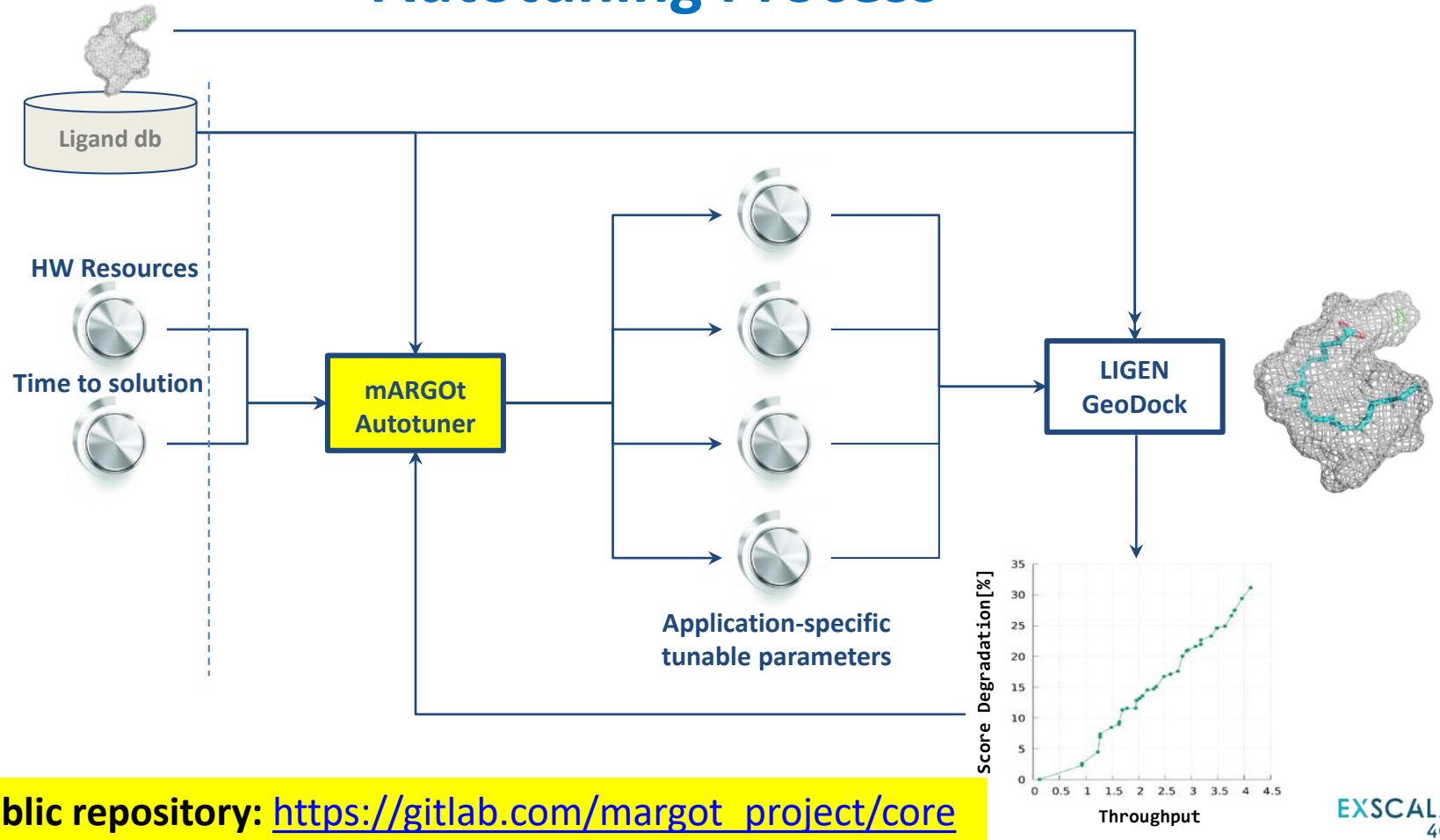
1. Screening huge library of compounds needs to be very fast but tolerate less accuracy
2. Screening a small set of compounds can be much slower but more accurate





Cristina Silvano et al., Autotuning and adaptivity in energy efficient HPC systems: the ANTAREX toolbox, CF 2018

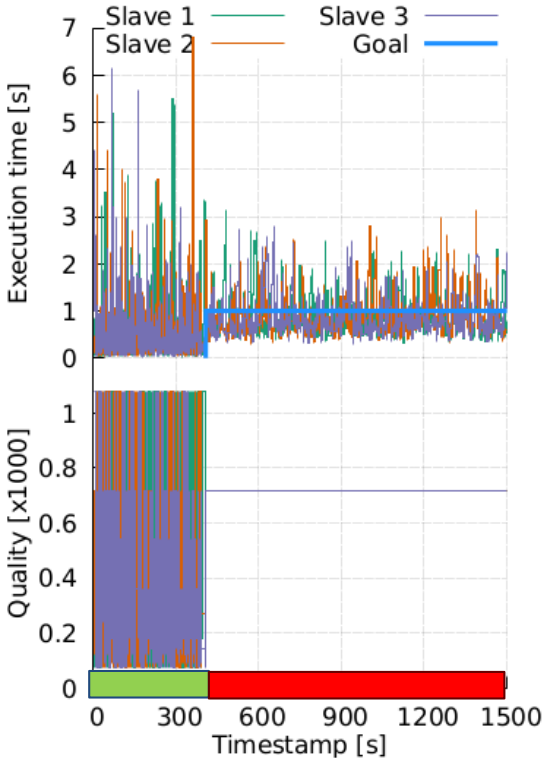
Autotuning Process



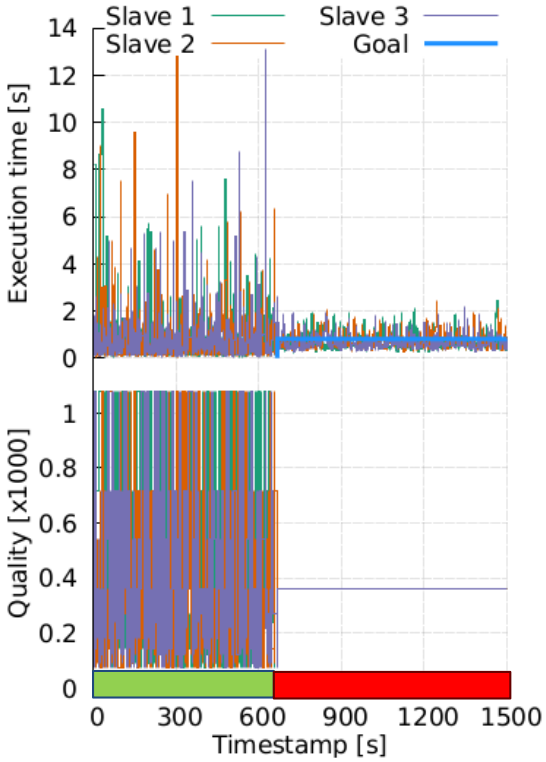
Public repository: https://gitlab.com/margot_project/core

EXSCALATE 4COV

mARGOt GeoDock: Autotuning Results



Small pocket (1b9v)



Large pocket (1cvu)

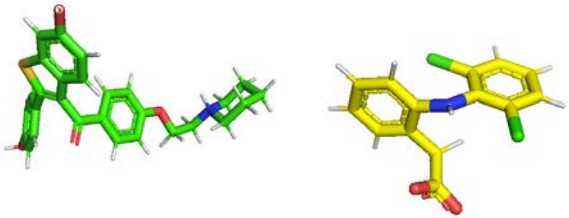
Learning Time

Inference Time

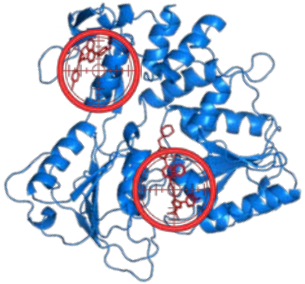
D. Gadioli, E. Vitali, G. Palermo, C. Silvano, "mARGOt: a Dynamic Autotuning Framework for Self-aware Approximate Computing", **IEEE Transactions on Computers,** May 2019.

Embarrassing parallel processing

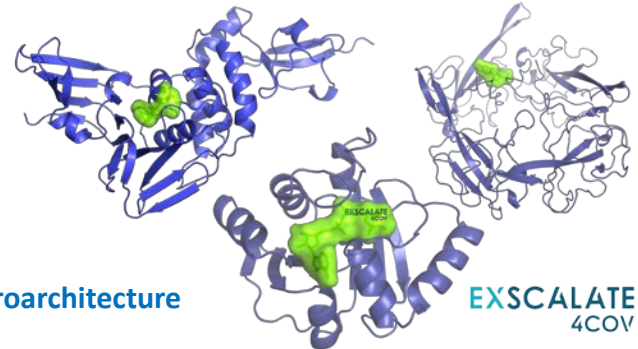
- We can process in parallel **different ligands** in a pocket



- We can process in parallel **different pockets** within a protein

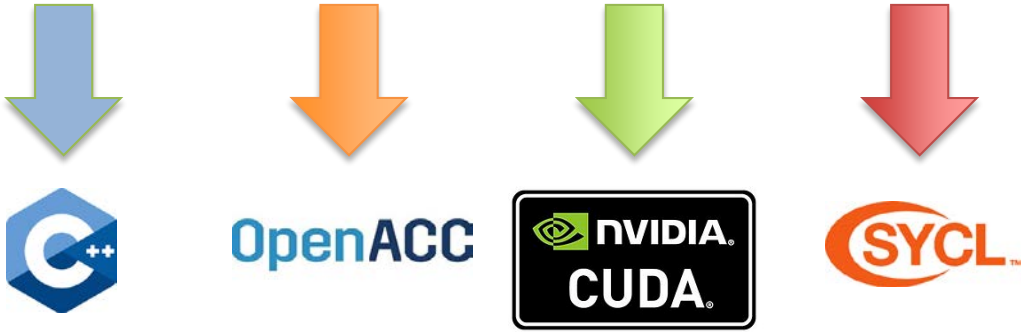


- We can process in parallel **different proteins** and different states of the proteins

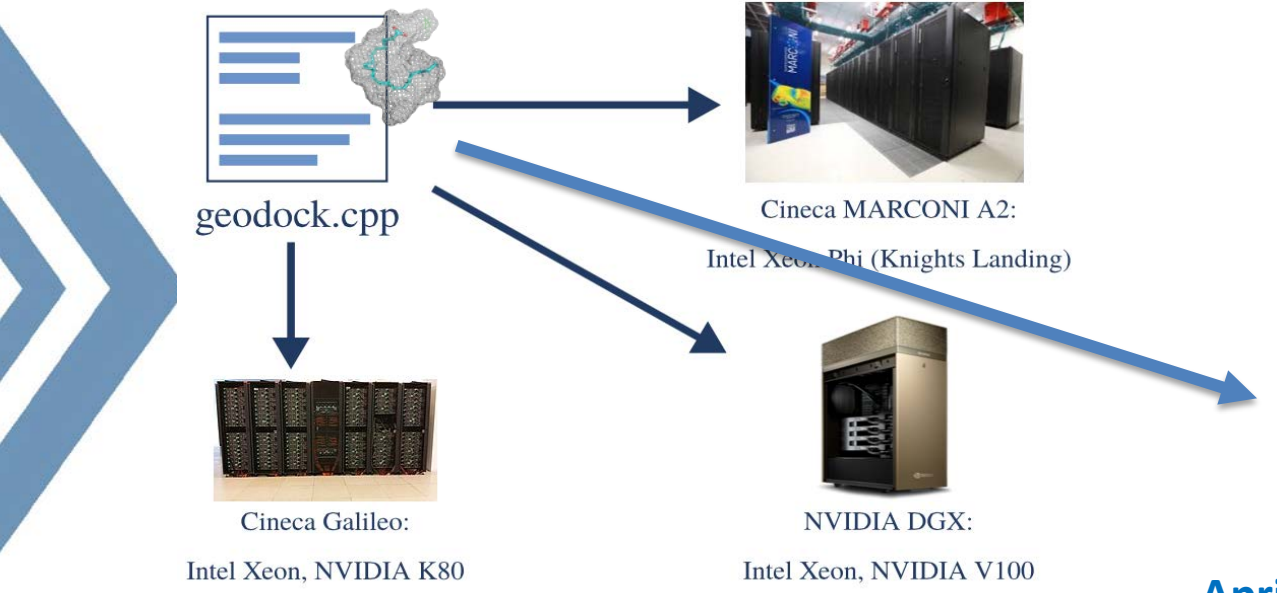


Portable Urgent Computing Platform

- Original code written in C/C++ CPU code using MPI backbone for CINECA MARCONI (Intel Xeon Phi / Knights Landing)
- Goal in 2020: To get a portable virtual screening platform for an urgent computing scenario on heterogeneous (CPU/GPU) HPC systems



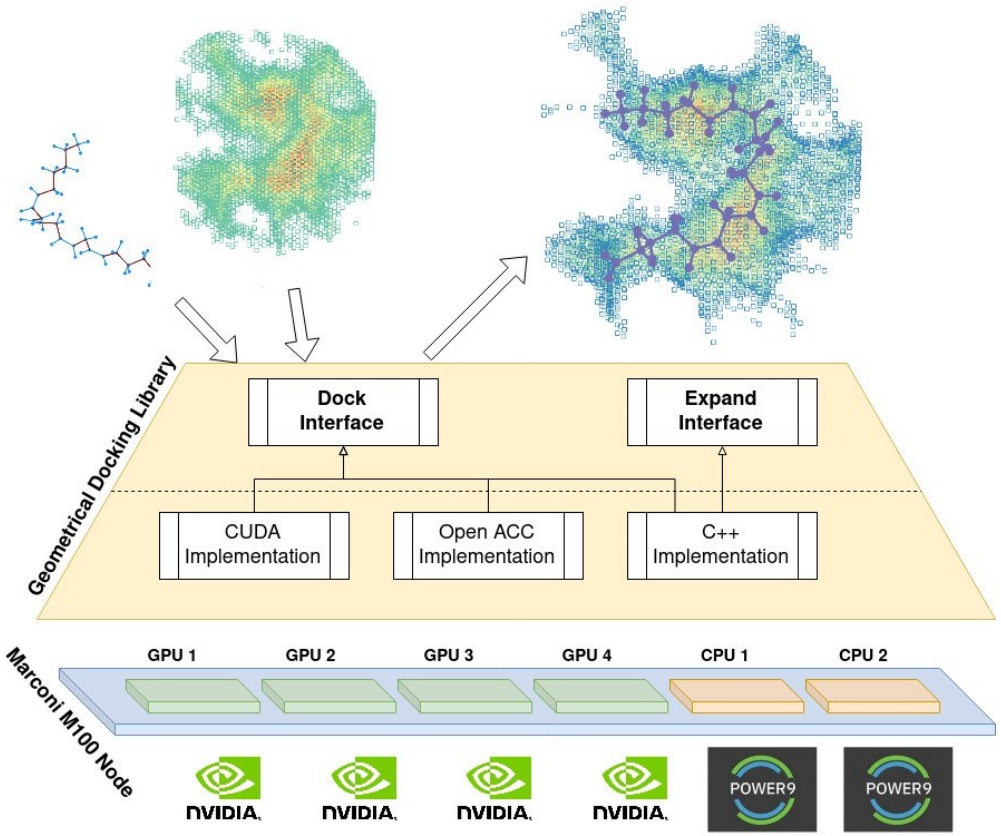
Porting EXSCALATE on Heterogeneous Nodes



April 2020: CINECA MARCONI-100
2x 16 IBM Power9 cores/node
4 NVIDIA Volta V100 GPUs/node

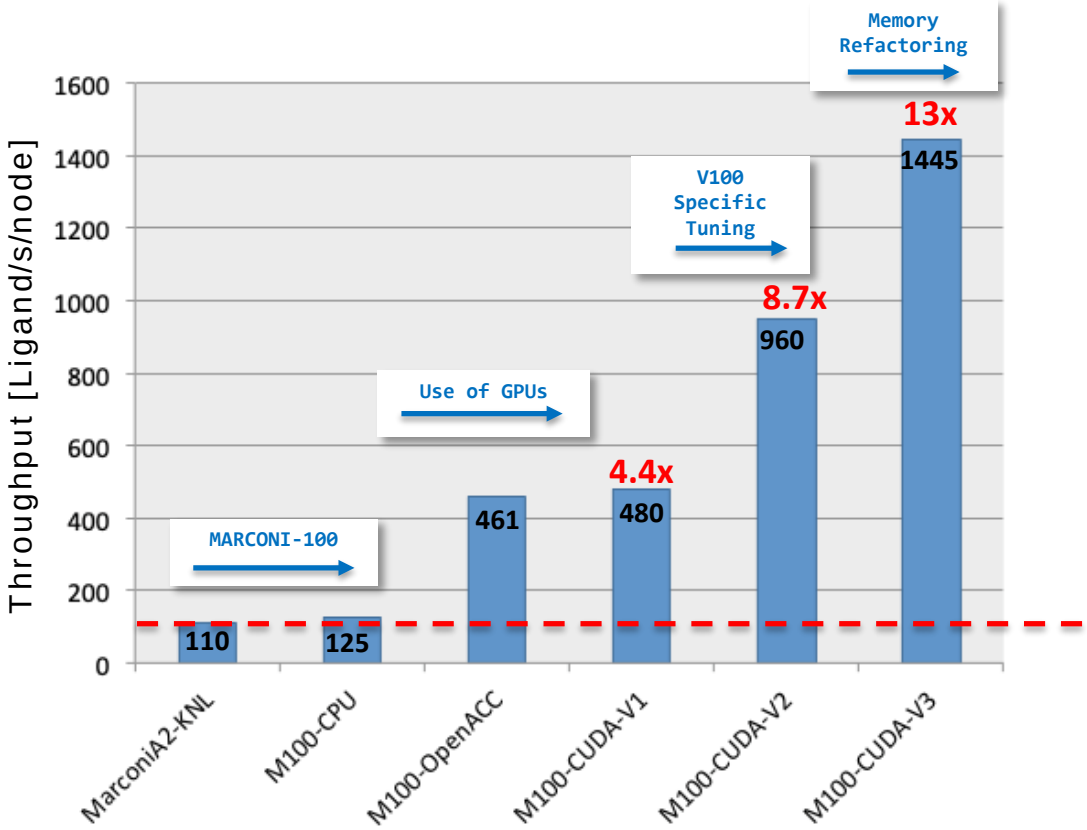
E. Vitali, D. Gadioli, G. Palermo, A. Beccari, C. Cavazzoni, C. Silvano «Exploiting OpenMP and OpenACC to accelerate a geometric approach to molecular docking in heterogeneous HPC nodes». J on Supercomput. 75(7): 3374-3396 (2019)

LiGen GeoDock Library deployed on a MARCONI-100 node



EXSCALATE 4COV

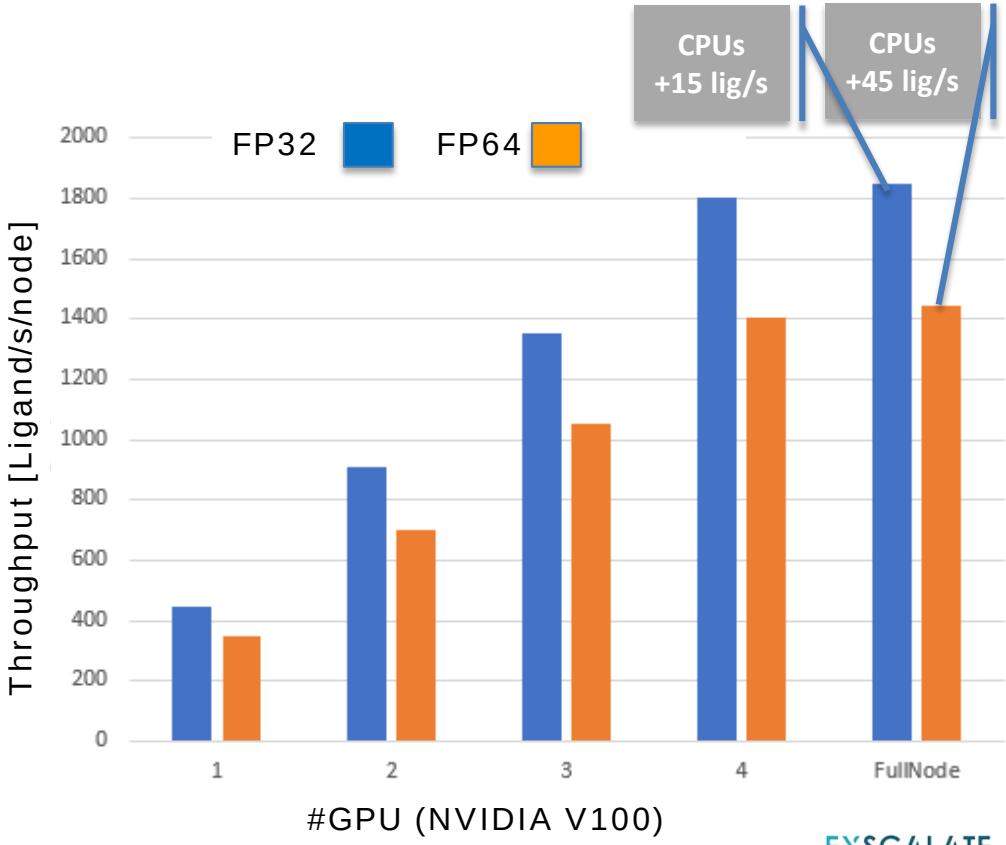
GPUs Acceleration on Marconi-100 node



EXSCALATE 4COV

Precision Tuning on MARCONI-100 node

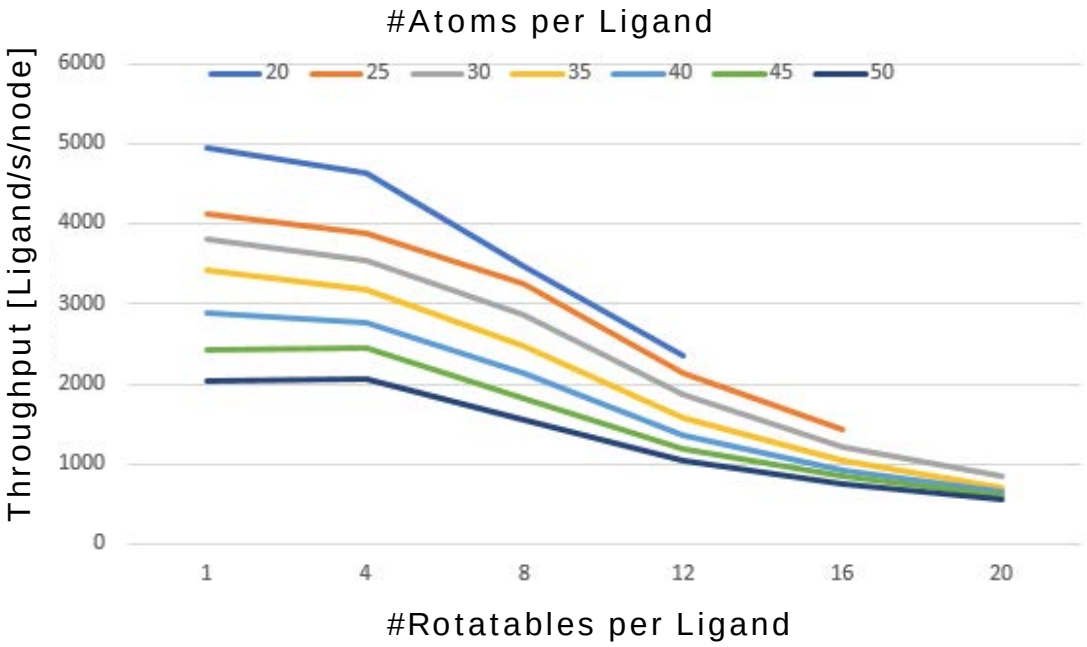
- LiGen GeoDock Library was originally designed with FP 64-bit precision
- Tuning **precision vs accuracy** at kernel level on GPUs:
  - 25% speedup
- Scaling GPUs & tuning precision from FP64 to FP32:
  - +100 Ligand/s per GPU
  - negligible CPUs contribution



EXSCALATE 4COV

Ligand-dependent Performance Analysis

Throughput by increasing the number of rotatable bonds per ligand given different number of atoms per ligand



EXSCALATE 4COV

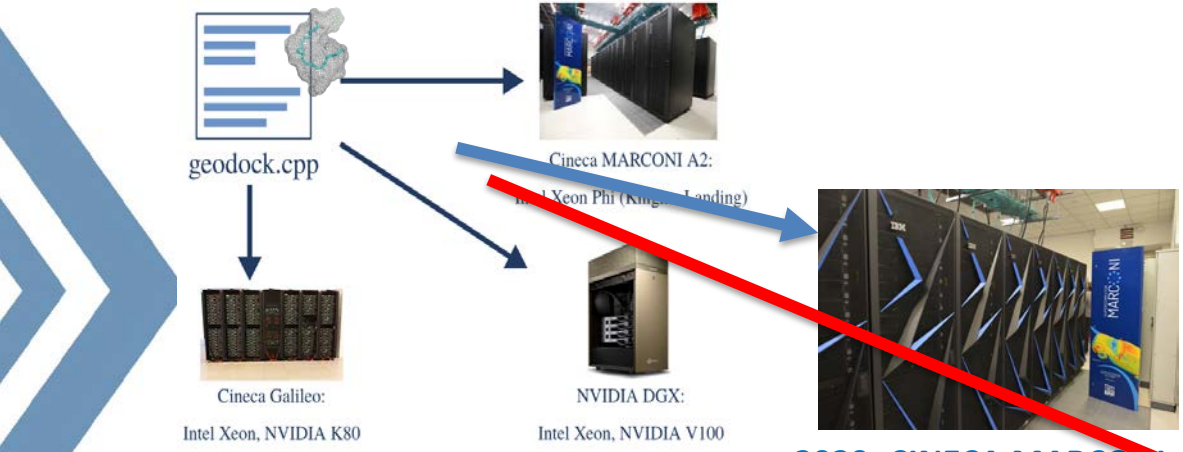
GPU-based acceleration results on MARCONI-100

- **Challenge:** Speeup Exscalate platform on **MARCONI-100** by exploiting the heterogenous nodes of MARCONI-100
- **Result: We reached a throughput of 2k ligands per second per node making feasible a large experiment to screen 500 B ligands in approx. 3 days (70h) on full MARCONI-100 at 2M ligands per second.**

April 2020: CINECA MARCONI-100
2x 16 IBM Power9 cores /node
4 NVIDIA Volta V100 GPUs / node

2 orders of magnitude faster than ANTAREX 4 Zika experiment and SotA techniques

Porting on Heterogeneous Nodes: The Next Challenges



2020: CINECA MARCONI 100
2x 16 IBM Power9 cores /node
4 NVIDIA Volta V100 GPUs / node

EuroHPC Joint Undertaking

2021: CINECA LEONARDO

Atos Bull Sequana
NVIDIA GPUs

The Path towards the European Exascale: CINECA Leonardo Supercomputer with Atos and NVIDIA

- ✓ **15 October 2020:** Announced the new **"Leonardo"** supercomputer (240 Meur) funded by EuroHPC & Italian Ministry of University & Research
- ✓ **2021:** Leonardo will be installed at Tecnopolo di Bologna (Italy).
- ✓ Leonardo architecture: Atos BullSequana XH2000 nodes, each with 4 NVIDIA Tensor Core GPUs and a single Intel CPU, NVIDIA Mellanox HDR 200 Gb/s Infiniband network and 100 PetaByte storage.
- ✓ Leonardo planned to reach **250 PetaFlop/s** Linpack performance (approx. **10 times MARCONI-100**, currently No. 9 in TOP500)
- ✓ Leonardo will features **14 000 NVIDIA Ampere GPUs** to deliver **10 ExaFlop/s** of FP16 AI performance

https://www.hpc.cineca.it/news/announcement-award-procurement-leonardo-supercomputer

LIGATE EuroHPC Project

LIgand Generator and portable drug discovery platform AT Exascale

- 3-year project coordinated by Dompé (11 partners) to be started in Jan. 2021.
- **Main goals:**
  - To accelerate a portable and tunable drug discovery platform ready for exascale HPC systems to respond promptly to worldwide pandemic crisis.
  - **Code** and **Performance Portability** on HPC systems based on heterogeneous accelerators
  - Develop **Artificial Intelligence** techniques to further speedup the virtual screening phase on HPC systems
  - Apply **Data Analytics** technologies to the huge ligands and proteins databases

LIGATE – EuroHPC-03-2019 Industrial software codes for extreme scale computing environments and applications

HPC and AI in Drug Discovery Pipelines

- Acceleration of the computational drug discovery pipeline can be reached by GPU-based acceleration, massive parallelism, **synergy between HPC and AI techniques.**
- **AI algorithms** can use the huge amounts of data generated by the HPC drug discovery pipelines to train Deep Learning models to predict:
  - Simulation results of the docking between the drug compounds against the target proteins under analysis.
  - Binding energy between the ligands and the target proteins
  - Next series of molecular Dynamics simulations
  - New drug candidates in AI-augmented molecular design

Ongoing and Future Activities

Continue to accelerate the search for molecules effective against the COVID-19

Continue to develop an effective tool for countering future pandemics to be consolidated over time

EXSCALATE platform is a computing platform to address the need of responding promptly to a pandemic crisis not only today, but also in the future due to its scalability, configurability and portability.

Cristina Silvano

cristina.silvano@polimi.it

POLITECNICO MILANO 1863

Gianluca Palermo

gianluca.palermo@polimi.it

EXSCALATE Team


Andrea Beccari

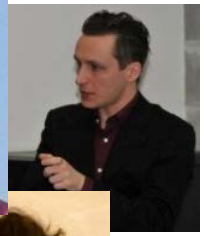

Carmine Talarico

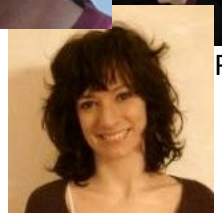
Dompé

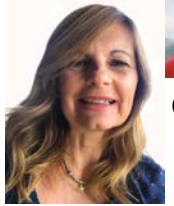

Candida Manelfi

CINECA


Carlo Cavazzoni


Federico Ficarelli


Chiara Latini


Cristina Silvano


Gianluca Palermo

POLITECNICO MILANO 1863


Emanuele Vitali


Davide Gadioli

EXSCALATE 4COV | exscalate4cov.eu

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003551.