LIGATE - LIgand Generator and portable drug discovery platform AT Exascale

Invited Paper

G. Palermo
Politecnico di Milano - DEIB
Milan, Italy
gianluca.palermo@polimi.it

G. Accordi, D. Gadioli, Y. Zhang, E. Vitali, B. Guindani, D. Ardagna, C. Silvano
Politecnico di Milano - DEIB
Milan, Italy

A. R. Beccari, D. Bonanni, C. Talarico, F. Lughini, A. D. Biswas
Dompé Farmaceutici Spa
Napoli, Italy

J. Martinovic, M. Golasowski, P. Silva, A. Bohm, J. Beranek, J. Krenek, B. Jansik
IT4Innovations, VSB – TUO
Ostrava, Czech Republic

P. Thoman, P. Salzmann, T. Fahringer
University of Innsbruck
Innsbruck, Austria

T. Schwede, L. T. Alexander, G. Tauriello, J. Durairaj
Biozentrum, University of Basel
Basel, Switzerland

B. Cosenza, L. Crisci
University of Salerno
Fisciano (SA), Italy

A. Emerson, F. Ficarelli
CINECA
Bologna, Italy

E. Lindahl, S. Wingbermühle
KTH Royal Institute of Technology
Stockholm, Sweden

D. Gregori
E4 Computer Engineering Spa
Scandiano (RE), Italy

S. Coletti, E. Sana
Chelonia S.A.
Lugano, Switzerland

P. Gschwandtner
PH3 GmbH, University of Innsbruck
Innsbruck, Austria

ABSTRACT

The COVID-19 pandemic demonstrates that a top priority for society, now and in the future, is to be able to respond quickly to diseases with effective treatments. Among the new tools that pharmaceutical industries and researchers have in their hands nowadays, there are the extensive computer simulations capable of evaluating *in-silico* the interaction between possible drugs and the target proteins.

The central goal of the LIGATE project is to create and validate a leading application solution for drug discovery in High-Performance Computing (HPC) systems up to the exascale level. The overall project purpose is the automation of the drug design process, which is currently performed with substantial human effort throughout the different phases of the process: preparation of input parameters, management of data sets with billions of molecules, interaction with HPC queue management systems to handle jobs, and optimization of scoring function parameters and thresholds.

CCS CONCEPTS

• **Computing methodologies** → **Massively parallel and high-performance simulations**; • **Applied computing**;

KEYWORDS

HPC, Molecular Docking, Virtual Screening, Molecular Dynamics

1 INTRODUCTION

Pharmaceutical companies are investing in computer-aided drug discovery (CADD) to revolutionize the slow and expensive process of bringing new drugs to market. This technology uses computational tools to predict how molecules might interact with a target disease, analyze the vast amounts of data generated, and identify promising drug candidates. CADD has the potential to significantly reduce development times and costs, while increasing the success rate of new therapies.

Given its structure, drug discovery can get a major boost from the usage of High-Performance Computing (HPC). In fact, these powerful supercomputers are capable of handling heavy tasks such as virtual screening on an extreme scale, evaluating huge libraries of potential drug molecules. Moreover, HPC can also be used to speed up complex simulations, allowing researchers to model how these drug candidates might behave within the proteins of the target disease. The embarrassingly parallel nature of the virtual screening process and the exploitable parallelism within a single computation make this task suitable for exploiting large computing infrastructures composed of powerful compute nodes.

In this direction, the COVID-19 pandemic highlighted the critical role of HPC in accelerating drug discovery. During the pandemic, a significant effort emerged in the fight against the virus by exploiting supercomputing resources: the Exscalate4CoV project [10]. The Exscalate4CoV project was a European consortium aimed to identify potential treatments by virtual screening massive libraries of molecules against the SARS-CoV-2 virus. In particular, the two top European supercomputers at that moment, CINECA-Marconi100 and ENI-HPC5, were almost fully used to run this urgent computing campaign [3]. Within the same period, a similar experiment was run by exploiting the Summit supercomputer the Oak Ridge Leadership Computing Facility (OLCF) in the US [4].

Based on the lessons learned from the Exscalate4CoV experience, the LIGATE project[1] [7] was conceived to design and develop an even more efficient European drug discovery platform, ready for a possible next pandemic and able to take advantage of recent HPC investments.

2 THE LIGATE CONSORTIUM

LIGATE is a three-year project co-funded by the European High-Performance Computing (EuroHPC) Joint Undertaking under the topic *Industrial software codes for extreme scale computing environments and applications*, and by national ministries of involved partners. It is a project led by Dompé Farmaceutici (Project Coordinator) and Politecnico di Milano (Scientific Coordinator), and aggregates overall 10 partners among Universities, Research Institutions, and Companies, including CINECA, VSB-IT4Innovations, E4 Computer Engineering, University of Salerno, University of Innsbruck, KTH - Royal Institute of Technology, University of Basel, PH3, and Chelonia Applied Science.

3 OBJECTIVES AND ACHIEVEMENTS

The overall objectives faced by the LIGATE project and related achievements reached can be listed as follows.

a) Design a CADD platform in the form of a highly scalable workflow capable of scaling up to a full Tier-0 supercomputer, in case of an urgent computing campaign.
The LIGATE CADD platform has been designed to consider a pipeline of the LiGen virtual screening application [3] and GROMACS for the Molecular Dynamic simulations and final ranking of the subset of drug candidates by their relative free energy of binding. During the project, an accurate design of the two phases relying on the embarrassingly parallel nature of the problem and an ad hoc meta-scheduler (named HyperQueue[2]) capable of managing a large number of heterogeneous tasks on HPC systems, were realized. In addition to that, we deeply analyzed and optimized the data management aspects within the workflow to avoid bottlenecks that could limit the process scalability.

b) Implement a code and performance portable virtual screening application, to make the proposed solution independent from the specific exascale HW, especially regarding GPU architectures and vendors.
Within LIGATE, we designed a *performance-portable* implementation of LiGen, originally written in C++ and CUDA to target only NVIDIA GPUs [3]. We first optimized the GPU implementation autotuning part of the code according to the characteristics of the input data [2]. Then, we ported the code to SYCL2020 to target also AMD GPUs and unlocking the possibility of running on the GPU partition of the CSC-LUMI supercomputer [1].

c) Develop an integrated automatic setup and execution of free energy calculations.
Starting from lists of thousands of docked compounds/poses, we made possible within GROMACS to automatically determine compound parameters, set up free energy calculations [6], execute the entire ensemble as a parallel run on HPC resources, and return free energies with a requested standard error. This permitted to automatize the free energy estimation step both for pure data collection and for large-scale screening.

d) Exploit state-of-the-art AI/ML techniques to further speed up the virtual screening phase and to precisely predict the scoring function used in the early stages of the CADD platform.
Within the project, we made use of the automatic setup for the free energy calculations to generate a large set of training data targeting the pose selection problem and the binding affinity prediction. The results obtained by the costly free energy computation derived by GROMACS [8] molecular dynamic simulations are used as ground truth for training and evaluation of advanced machine learning methods (i.e Convolutional Neural Networks, Graph Neural Networks, Mixture Density Networks).

e) Integrate the CADD platform within an industrial solution.
The final LIGATE solution including open source and non-open source IPs, has been adopted into an industrial context. The Dompé-owned EXSCALATE [3] platform inherits part of the work done in LIGATE and exploits the project results towards real-world targets.

4 PROJECT HIGHLIGHT: A PORTABLE DRUG DISCOVERY PLATFORM FOR URGENT COMPUTING

One of the project's main goals was to have a portable Computer Aided Drug Discovery platform capable of running on all the sets of EuroHPC supercomputers. This was thought for two main reasons. First, the need of not being limited by the available hardware. Second, the possibility of scaling an urgent computing campaign across multiple sites to use the maximum aggregate computing power possible.

To do so, we first developed a portable version of the LiGen software using SYCL [2]. The usage of SYCL permitted tackling the execution of the target code over non-NVIDIA GPUs. Indeed, the original LiGen code was fully optimized using CUDA since previous large-scale experiments were only on NVIDIA-based machines. Second, we integrated the solution using the LEXIS platform [5]. The LEXIS Platform enables easy access to High-Performance-Computing (HPC) and Cloud-Computing infrastructure for science, SMEs, and industry. It enables automatized workflows across Cloud and HPC systems, on-site or across different European supercomputing centers. It provides an intuitive web interface, REST API, and Python SDK for easy integration with other systems. For workflow orchestration, it uses Apache Airflow integrated with the HPC-as-a-Service HEAppE middleware [9], credential management, and the LEXIS distributed data interface (DDI) for distributed data and metadata management.

One of LEXIS's significant features is the possibility of using multiple HPC clusters in one workflow execution. Within the LIGATE project, we exploited this feature by developing an urgent computing workflow across multiple sites. In this workflow, LiGen has been executed on the two EuroHPC clusters, IT4I-Karolina and CSC-LUMI. An image related to the LEXIS workflow is shown in Figure 1 where it is possible to see a clear fork in the workflow representing the execution on the two target machines. The input dataset, which is loaded through the LEXIS web interface or

[1]LIGATE website - https://www.ligateproject.eu/
[2]HyperQueue - https://github.com/It4innovations/hyperqueue
[3]EXSCALATE website - https://exscalate.com/

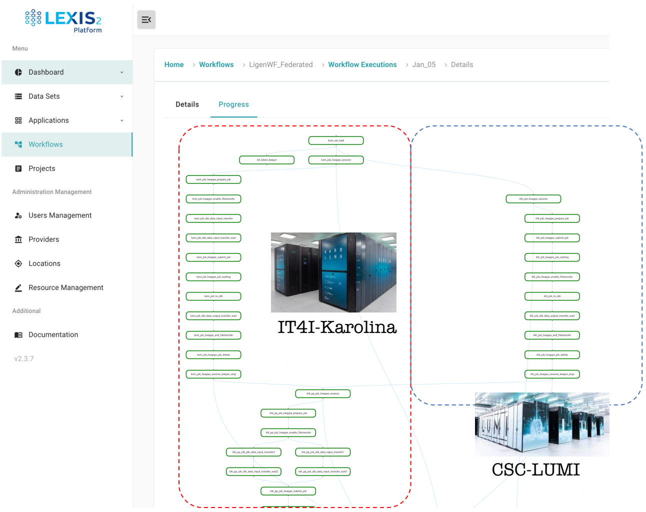

Figure 1: LEXIS multisite execution of the LiGen-based Drug-Discovery workflow on IT4I-Karolina and CSC-LUMI supercomputers

recovered from previous executions, is split between the two supercomputers and sent to the target locations. However, the workflow permits the inclusion of sensitive data that are directly stored on one of the clusters and that do not have to migrate somewhere else due to confidentiality issues. Finally, the LiGen output generated on both sites is staged to the IT4Innovations location, and a post-processing job takes the results, creates one combined output of the entire workflow, and stores it in the LEXIS Platform for further data analysis and use.

5 CONCLUSIONS

The LIGATE project aims to provide an automated solution for CADD on EuroHPC resources from the current pre-exascale architectures to the future exascale systems. The integration of best-in-class European software components will be made available to industrial as well as non-profit academic research and public institutions, as a Platform-as-a-Service solution for drug discovery.

In this document, we described the overall challenges and main achievements of the project. A more detailed highlight has been done on the final usage of the developed CADD workflow for an urgent computing scenario.

ACKNOWLEDGEMENTS

This project has received funding from EuroHPC-JU - the European High-Performance Computing Joint Undertaking - under grant agreement No 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, and Switzerland.

Moreover, we acknowledge EuroHPC Joint Undertaking for awarding us access to Karolina at IT4Innovations, Czech Republic (EHPC-DEV-2021D02-049), and to LUMI, hosted by CSC (Finland) and the LUMI consortium (EHPC-BEN-2022B12-001).

REFERENCES

[1] Gianmarco Accordi, Davide Gadioli, Luigi Crisci, Lorenzo Carpentieri, Biagio Cosenza, Andrea Beccari, and Gianluca Palermo. 2024. Unlocking performance portability on LUMI-G supercomputer: A virtual screening case study. *IWOCL - 12th International Workshop on OpenCL and SYCL* (2024).

[2] Gianmarco Accordi, Davide Gadioli, Emanele Vitali, Luigi Crisci, Biagio Cosenza, Andrea Beccari, and Gianluca Palermo. 2024. Out of kernel tuning and optimizations for portable large-scale docking experiments on GPUs. *The Journal of Supercomputing* (2024). https://doi.org/10.1007/s11227-023-05884-y

[3] Davide Gadioli, Emanuele Vitali, Federico Ficarelli, Chiara Latini, Candida Manelfi, Carmine Talarico, Cristina Silvano, Carlo Cavazzoni, Gianluca Palermo, and Andrea Rosario Beccari. 2023. EXSCALATE: An Extreme-Scale Virtual Screening Platform for Drug Discovery Targeting Polypharmacology to Fight SARS-CoV-2. *IEEE Transactions on Emerging Topics in Computing* 11, 1 (2023), 170–181.

[4] Jens Glaser, Josh V Vermaas, David M Rogers, Jeff Larkin, Scott LeGrand, Swen Boehm, Matthew B Baker, Aaron Scheinberg, Andreas F Tillack, Mathialakan Thavappiragasam, Ada Sedova, and Oscar Hernandez. 2021. High-throughput virtual laboratory for drug discovery using massive datasets. *The International Journal of High Performance Computing Applications* 35, 5 (2021), 452–468.

[5] Martin Golasowski, Jan Martinovič, Jan Křenek, Kateřina Slaninová, Marc Levrier, Piyush Harsh, Marc Derquennes, Frédéric Donnat, and Olivier Terzo. 2022. *The LEXIS Platform for Distributed Workflow Execution and Data Management.* 17 – 35 pages. https://doi.org/10.1201/9781003176664-2

[6] Magnus Lundborg, Jack Lidmar, and Berk Hess. 2021. The accelerated weight histogram method for alchemical free energy calculations. *Journal of Chemical Physics* 154 (2021), 204103. https://doi.org/10.1063/5.0044352

[7] G. Palermo, G. Accordi, D. Gadioli, E. Vitali, C. Silvano, B. Guindani, D. Ardagna, A. R. Beccari, D. Bonanni, C. Talarico, F. Lughini, J. Martinovic, P. Silva, A. Bohm, J. Beranek, J. Krenek, B. Jansik, B. Cosenza, L. Crisci, P. Thoman, P. Salzmann, T. Fahringer, L. T. Alexander, G. Tauriello, T. Schwede, J. Durairaj, A. Emerson, F. Ficarelli, S. Wingbermühle, E. Lindhal, D. Gregori, E. Sana, S. Coletti, and P. Gschwandtner. 2023. Tunable and Portable Extreme-Scale Drug Discovery Platform at Exascale: the LIGATE Approach. In *Proceedings of the 20th ACM International Conference on Computing Frontiers (CF '23)*. Association for Computing Machinery, New York, NY, USA, 272–278. https://doi.org/10.1145/3587135.3592172

[8] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R. Shirts, Jeremy C. Smith, Peter M. Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinform.* 29, 7 (2013), 845–854.

[9] Vaclav Svaton, Jan Martinovic, Jan Krenek, Thomas Esch, and Pavel Tomancak. 2020. HPC-as-a-Service via HEAppE Platform. *Advances in Intelligent Systems and Computing* 993 (2020), 280 – 293. https://doi.org/10.1007/978-3-030-22354-0_26

[10] Giulio Vistoli, Candida Manelfi, Carmine Talarico, Anna Fava, Arieh Warshel, Igor V. Tetko, Rossen Apostolov, Yang Ye, Chiara Latini, Federico Ficarelli, Gianluca Palermo, Davide Gadioli, Emanuele Vitali, Gaetano Varriale, Vincenzo Pisapia, Marco Scaturro, Silvano Coletti, Daniele Gregori, Daniel Gruffat, Edgardo Leija, Sam Hessenauer, Alberto Delbianco, Marcello Allegretti, and Andrea R. Beccari. 2023. MEDIATE - Molecular DockIng at homE: Turning collaborative simulations into therapeutic solutions. *Expert Opinion on Drug Discovery* 18, 8 (2023), 821–833. https://doi.org/10.1080/17460441.2023.2221025 PMID: 37424369.