An Extreme-Scale Virtual Screening Platform for Drug Discovery

D. Gadioli*, E. Vitali*, F. Ficarelli[x], C. Latini[x], C. Manelfi[+],
C. Talarico[+], C. Silvano*, A. R. Beccari[+], Gianluca Palermo*

*Politecnico di Milano, Dipartimento di Elettronica Informazione e Bioingegneria, Milan, Italy
[x] Cineca, Supercomputing Innovation and Application Department, Bologna, Italy
[+]Dompé Farmaceutici SpA, EXSCALATE, Napoli, Italy

ABSTRACT

Virtual screening is one of the early stages that aims to select a set of promising ligands from a vast chemical library. Molecular Docking is a crucial task in the process of drug discovery and it consists of the estimation of the position of a molecule inside the docking site. In the contest of urgent computing, we designed from scratch the EXSCALATE molecular docking platform to benefit from heterogeneous computation nodes and to avoid scaling issues. This poster presents the achievements and ongoing development of the EXSCALATE platform, together with an example of usage in the context of the COVID-19 pandemic.

CCS CONCEPTS

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

KEYWORDS

HPC, Molecular Docking, Virtual Screening

ACM Reference Format:
D. Gadioli*, E. Vitali*, F. Ficarelli[x], C. Latini[x], C. Manelfi[+],, C. Talarico[+], C. Silvano*, A. R. Beccari[+], Gianluca Palermo*. 2022. An Extreme-Scale Virtual Screening Platform for Drug Discovery. In *19th ACM International Conference on Computing Frontiers (CF'22), May 17–19, 2022, Torino, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3528416.3530872

1 INTRODUCTION

Drug discovery is a long process that usually involve *in-silico*, *in-vitro*, and *in-vivo* stages. The outcome of this process is a molecule, named *ligand*, that has the strongest interaction with at least one binding site of the protein, also known as *pocket*, that represents the target of the experiment. Domain experts expect this interaction to lead to a beneficial effect. Virtual screening is one of the early stages that aims to select a set of promising *ligands* from a vast chemical library [3]. The complexity of this operation is due to the ligand and pocket flexibility: both of them can change shape when they interact. Therefore, to estimate the interaction strength using a *scoring function*, we also need to predict the displacement of their atoms using a *docking* algorithm. This problem is computationally heavy, and it is well known in literature. Moreover, to increase the probability of finding promising candidates, we would like to increase the size of the chemical library as much as possible, exacerbating the complexity of the virtual screening. Since the evaluation is *in-silico*, we can design new molecules by simulating known chemical reactions. Therefore the chemical library size is limited only by the system's computational power.

In the context of urgent computing, where the time required to find a therapeutic cure should be as short as possible, we are re-designing LIGEN [1] the software for molecular docking that is part of the EXSCALATE platform with the goal of virtual screening as many ligands as possible in a given time budget.

2 THE EXSCALATE PLATFORM

The EXSCALATE platform [1] is composed of two main pillars: (i) a virtual library of target compounds, and (ii) a molecular docking software for high throughput virtual screening. The extremely large virtual chemical library is composed of hundreds of billions of compounds. We generate all the compounds starting from a database of millions of available commercial reagents that were combined using a set of robust synthetic reactions and truly achievable in one reaction step. The molecular docking software, named LIGEN, has been designed from scratch to target HPC machines. LIGEN is an integrated monolithic application to dock and re-score a chemical library, using MPI to scale out, *C++11* threads to scale up, and CUDA kernels to exploit the GPUs available in the target HPC cluster by accelerating the compute intensive sections.

Figure 2 shows an overview of LIGEN at different levels of abstractions: the computation model, the software stack, and its mapping on the underlying hardware. We have chosen to write an MPI application that implements the same asynchronous pipeline on each node. Therefore, we execute a single MPI process for each node available. Then, each process spawns a pipeline to carry out the elaboration using all the computation resources of its node. We use at least one native thread for each stage of the pipeline. The stages that perform I/O operations and the stage that track the computation progress are composed of a single thread. While the stage that parses a ligand from its description, the stage that clusters similar ligands in the same bucket, and the stage that performs the dock and score operations on the GPUs, are composed of multiple software threads. Since all the threads of the same stage interact with shared thread-safe queues, we have intra-node work-stealing.

Even if the problem is embarrassing parallel, the application throughput might generate high data transfer, posing challenges on the I/O that LIGEN needs to address [2]. In particular, to improve

[1] https://www.exscalate.eu/en/platform.html

CF'22, May 17–19, 2022, Torino, Italy

ACM ISBN 978-1-4503-9338-6/22/05.
https://doi.org/10.1145/3528416.3530872

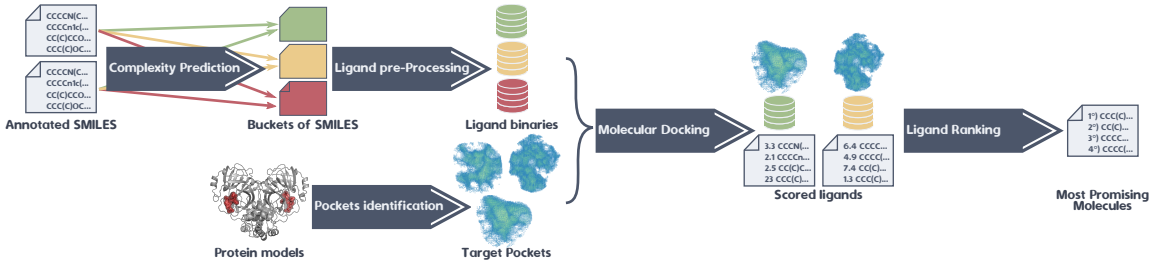

Figure 1: Overview of the designed workflow for extreme-scale simulations.

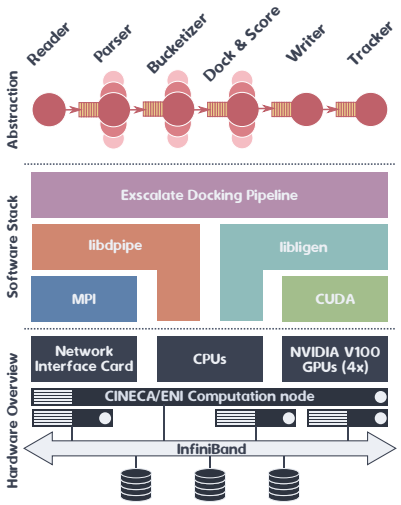

Figure 2: Overview of the EXSCALATE docking application.

the I/O efficiency we read and write the files sequentially. However, we statically split the input file among the MPI processes involved in the computation, negating intra-node work-stealing. Even if with this approach LIGEN is very I/O friendly, its throughput is equal to the one of the slowest process.

The proposed solution reaches a high throughput without relying on the node's local storage and it has been included in a more complex workflow to compensate for weak points that were limiting the extreme-scale scaling, e.g. the aforementioned work-stealing problem. Figure 1 depicts the designed workflow for extreme-scale simulations, which requires two different kinds of input from the domain knowledge and produces tin output the ranked list of most promising molecules. Regarding the input, on one hand, we require the binding sites of the target proteins. On the other hand, we require the chemical library of molecules that we want to evaluate (e.g. in SMILE format). To prepare the molecules to be screened in an optimal way for the docking phase, we pre-process and cluster them in buckets composed of elements with similar processing time. This phase is needed to balance the workload given the absence of the work-stealing across the nodes. The output of the docking phase includes the binding affinity estimation of each ligand with each protein/pocket. To finalize a screening experiment, these values require to be reordered by means of a ranking phase that decides which molecules are more interesting to be further analyzed.

3 THE 1 TRILLION DOCKING EXPERIMENT

In the context of the EXSCALATE4CoV European project[2], with the goal of finding new potential drugs against the COVID19 pandemic, we deployed the Exscalate platform in two HPC machines *CINECA-Marconi100* and *ENI-HPC5* with a combined throughput of 81 PFLOPS, to rank a chemical library of more than 70 billion ligands against 15 binding-sites of 12 viral proteins of Sars-Cov2. Overall, the experiment lasted 60 hours and it performed a trillion of docking operations, becoming the largest virtual screening campaign up to this moment. The knowledge generated by this experiment is publicly released through the MEDIATE website[3].

4 ONGOING WORK

Currently, the LIGEN code is portable only thanks to its C++ version, and thus only exploiting the CPU part of an HPC node Indeed, its GPU extension is deployable only on NVIDIA hardware given its CUDA code. To enable the deployment on any available type of architecture, we are currently moving the accelerated part using SYCL. Already in the past, we faced this problem using OpenACC [4]. However, we are finding the new SYCL standard a more suitable opportunity for our goals. In addition to that, we are also developing machine learning engines to be paired with the current pipeline to make more robust the pose selection and scoring phase within the code.

ACKNOWLEDGEMENTS

This work has received funding from EuroHPC-JU under the grant agreement No 956137 (LIGATE), and from H2020 Programme under the grant agreement No 101003551 (EXSCALATE4CoV).

REFERENCES

[1] Andrea R. Beccari, Carlo Cavazzoni, Claudia Beato, and Gabriele Costantino. 2013. LiGen: A High Performance Workflow for Chemistry Driven de Novo Design. *Journal of Chemical Information and Modeling* 53, 6 (2013), 1518–1527.

[2] Stefano Markidis, Davide Gadioli, Emanuele Vitali, and Gianluca Palermo. 2021. Understanding the I/O Impact on the Performance of High-Throughput Molecular Docking. In *2021 IEEE/ACM Sixth International Parallel Data Systems Workshop (PDSW)*. IEEE Computer Society, 9–14.

[3] Natarajan Arul Murugan, Artur Podobas, Davide Gadioli, Emanuele Vitali, Gianluca Palermo, and Stefano Markidis. 2022. A Review on Parallel Virtual Screening Softwares for High-Performance Computers. *Pharmaceuticals* 15, 1 (2022), 63.

[4] Emanuele Vitali, Davide Gadioli, Gianluca Palermo, Andrea Beccari, Carlo Cavazzoni, and Cristina Silvano. 2019. Exploiting OpenMP and OpenACC to accelerate a geometric approach to molecular docking in heterogeneous HPC nodes. *The Journal of Supercomputing* 75, 7 (2019), 3374–3396.

[2] https://www.exscalate4cov.eu/
[3] https://mediate.exscalate4cov.eu