# ML_INFN project: status report and future perspectives

*Lucio* Anderlini[1], *Tommaso* Boccali[2], *Stefano* Dal Pra[3], *Doina Cristina* Duma[3], *Luca* Giommi[3,*], *Daniele* Spiga[4], and *Gioacchino* Vino[5]

[1]INFN Sezione di Firenze, Via Bruno Rossi 1, 50019 Sesto Fiorentino, Firenze (ITALY)
[2]INFN Sezione di Pisa, L.go Bruno Pontecorvo 3, 56127 Pisa (ITALY)
[3]INFN-CNAF, Viale Carlo Berti Pichat, 6/2, 40127 Bologna (ITALY)
[4]INFN Sezione di Perugia, Via Alessandro Pascoli 23c, 06123 Perugia (ITALY)
[5]INFN Sezione di Bari, Via Giovanni Amendola 173, 70126 Bari (ITALY)

**Abstract.** The ML_INFN initiative (*"Machine Learning at INFN"*) is an effort to foster Machine Learning (ML) activities at the Italian National Institute for Nuclear Physics (INFN). In recent years, artificial intelligence inspired activities have flourished bottom-up in many efforts in Physics, both at the experimental and theoretical level. Many researchers have procured desktop-level devices, with consumer-oriented GPUs, and have trained themselves in a variety of ways, from webinars, books, and tutorials. ML_INFN aims to help and systematize such effort, in multiple ways: by offering state-of-the-art hardware for ML, leveraging on the INFN Cloud provisioning solutions and thus sharing more efficiently GPUs and leveling the access to such resources to all INFN researchers, and by organizing and curating Knowledge Bases with production-grade examples from successful activities already in production. Moreover, training events have been organized for beginners, based on existing INFN ML research and focused on flattening the learning curve. In this contribution, we will update the status of the project reporting in particular on the development of tools to take advantage of High-Performance Computing resources provisioned by CNAF and ReCaS computing centers for interactive support to activities and on the organization of the first in-person advanced-level training event, with a GPU-equipped cloud-based environment provided to each participant.

## 1 Introduction

The development of Artificial Intelligence (AI) applications and in particular Machine Learning (ML) is rapidly changing the technological landscape of high-performance computing. Public and private research are investing in the development of processors, cloud solutions [1], software libraries, and statistical algorithms[2] reaching numerical models of unimaginable precision just a few months ago.

The Italian National Institute for Nuclear Physics (*Istituto Nazionale di Fisica Nucleare*, INFN) is the coordinating institution for nuclear, particle, theoretical, and astroparticle physics in Italy. It promotes, coordinates, and carries out scientific research as well as the technological development necessary for the activities in these sectors.

*e-mail: luca.giommi@cnaf.infn.it

In most of the research activities coordinated relevant to INFN, digital data are simulated, acquired, processed, or analysed. ML technologies are emerging as fundamental tools to deal with large amount of data and therefore are being investigated under many different perspectives in many different areas of the INFN research.

INFN has been revisiting its computing organization to support this vision. INFN Cloud is fully aligned with it and has already shown multiple results, with a roadmap for its evolution [3], covering ML and several other areas. The main drivers remain the needs of the scientific communities.

In this contribution, we will discuss ML_INFN, a technological initiative in its fourth year of activity aiming at fostering the adoption of modern data processing and ML techniques within the research fields relevant to INFN. In Section 2, we will discuss the computing infrastructure, built on top of INFN Cloud, providing interactive access to compute resources and hardware accelerators for developing, training, and sharing ML techniques. We will then discuss the organization of training events to support the adoption of modern techniques among students and young researchers (Section 3) and the effort to connect experts and experiences from the various areas in which INFN operates (Section 4). Conclusions and outlook are discussed in Sections 5 and 6.

## 2 Cloud infrastructure and hardware setup

One of the pillars of the ML_INFN initiative is the provisioning of a common, stable, and reliable ground for researchers involved in ML to develop, review and share their applications, crossing the borders between different communities, experiments and research domains.

A computing farm composed of two servers is installed in the INFN CNAF Tier-1 computing center, equipped with eight NVIDIA Tesla T4 GPUs, five NVIDIA RTX 5000 GPUs, one NVIDIA Ampère A30 GPU, one NVIDIA Ampère A100 GPU. In the event of high pressure on GPU resources, CNAF may provide an additional server with three NVIDIA A100 GPUs. Such flexibility is granted by the INFN Cloud infrastructure, grouping the resources made available to ML_INFN in a dedicated *OpenStack tenancy* [4]. Ephemeral virtual machines (VMs) are then assigned to specific projects as agreed within a team of Cloud administrators with at least one representative per INFN unit. Cloud administrators take responsibility over the VMs assigned to their INFN unit, ensuring security updates and providing support to the developers. The configuration of the VMs to serve the typical ML workloads is obtained with Ansible [5] and TOSCA [6] templates, and a custom web dashboard developed and maintained in the framework of INFN Cloud.

The typical VM is equipped with 8 virtual cores, 64 GB of RAM, 512 GB of NVMe disk, and a GPU. The compute, memory, and disk resources can be expanded, and the model of GPU can be selected, through the INFN Cloud web dashboard when instantiating the VM. Splitting the A100 GPUs into partitions, as made possible by the Multi-Instance GPU (MIG) feature, can also be configured through the same dashboard.

While each VM is assigned to a single administrator (and therefore to an INFN unit) for security and maintenance, the user-basis is national, with authentication and authorization procedures managed with OIDC tokens provided by an instance of Indigo IAM [7] maintained by INFN Cloud. This was proven to foster interaction and collaboration between analysts and developers with different background and belonging to different communities.

Most developers access the resources through a customized JupyterHub [8] interface, spawning a Docker [9] container running a customized JupyterLab [10] image per connected user. Multiple containers can coexist on the same VM and run code simultaneously, and multiple Jupyter Notebooks can be executed in parallel in the same container. However, most ML frameworks are designed to take control of the entire GPU resource, making it impossible

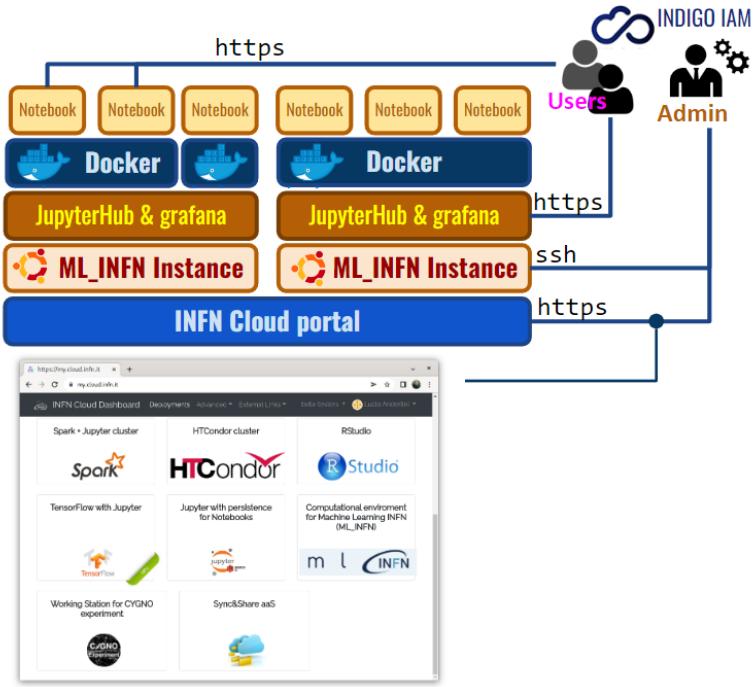


**Figure 1.** Conceptual software stack in the INFN Cloud middleware used to provision the compute resources of ML_INFN. A snapshot of the INFN dashboard as appearing in a web browser is also shown. See text for details.

to run multiple GPU-accelerated Jupyter Notebooks on the same VM. The arbitration on the allocation of the GPU resources through the users of the same VM is left to the self-organization of the developers collaborating on the project the VM is assigned to.

JupyterHub is configured to enable spawning custom Docker images downloading them from public or INFN image registries, such as DockerHub [11]. A default image is prepared as part of the ML_INFN project, providing a functional development environment configuring private and shared user areas, POSIX access to the CERN VM filesystem (`cvmfs`) [12] and a default Python environment with Pandas [13], TensorFlow [14], PyTorch [15], and other numerical computation libraries preinstalled. Developers are encouraged to customize the Docker image, for examples by adding libraries or selecting specific versions of the dependencies, to specialize the computing environment for their applications.

In those cases where long training campaigns are needed, Cloud administrators may enable a direct *Secure Shell* (`ssh`) connection to the VM to work around limitations due to authentication token lifespan and web connection instabilities. In four years of activity, only two projects requested and were granted direct ssh access to the VMs.

Figure 1 presents a schematic representation of the virtualization and containerization infrastructure used to provision the GPU resources to the developers.

Finally, the monitoring dashboards of OpenStack, INFN Cloud and JupyterHub are complemented, and partially integrated, in a custom Grafana [16] dashboard and in a Streamlit [17] web application presenting the overall status and the recent history of the infrastructure.

## 3 Dedicated training opportunities

The infrastructure described in the previous section was used to provide computing resources for four training events, named *ML hackathons*. The events were designed in two flavours: *entry-level hackathons*, aiming at providing the conceptual and technical tools to start applying ML techniques to research data to the widest possible audience, and *advanced hackathons*, aiming at fostering the adoption of newer and more complicated methods to the research domains of relevance for INFN.

Entry-level hackathons are organized in virtual mode, using Zoom [18] as videoconferencing platform, in order to reduce the costs for the participants and reach an audience of students and young post-docs for whom travel fundings are more difficult to secure. Advanced hackathons, on the contrary, are designed to reach a more selected audience and to deepen particular aspects of ML, making *in-person* events the preferred option.

The first entry-level hackathon was organized in June 2021. Because of the extraordinary number of registrations received, exceeding the maximum number of participants already few hours after the registration was open, the event was then replicated in December 2021, instead of the planned *advanced hackathon*, whose organization would have been in any case affected by the spread of the pandemic.

The third edition was for an *advanced hackathon* organized in November 2022, at the Physics Department of the University of Bari.

One last entry-level hackathon was organized in June 2023, and a further advanced hackathon is being organized in Pisa in November-December 2023.

### 3.1 Entry-level hackathons

Entry-level hackathons are virtual-only events structured over three days. The first day is devoted to introductory lectures on the fundamentals of ML, Cloud computing and hands-on tutorials on how to access and profit from the computing infrastructure detailed in Section 2.

During the second day, applications of ML in INFN research are discussed, either by presenting simplified use cases in the form of Jupyter Notebooks, or in the form of full-fledged seminars. A large fraction of the afternoon is then reserved for the participants to focus on the material discussed in the previous sessions with the tutors available to answer questions and deepening particular aspects in one-to-one video calls, in small groups joining a parallel discussion room or with the full audience in the main videoconferencing room.

During the third day, participants are split into ten groups. A tutor, a VM, and an exercise are assigned to each group. Four different exercises were prepared for the entry-level hackathons, two presenting realistic applications in the analysis of CMS data, one focusing on the compression and anomaly detection of data collected by the gravitational wave observatories, in particular by Virgo, and one focusing on the automatic segmentation of CT scans using ML techniques.

Figure 2 shows the good distribution of the participants through the INFN units, research domain, and career level. Analysing the data, we observe that, while the number of participants remains more or less constant through the editions, saturating at the maximum capacity of sixty participants, the number of involved units decreases, with units historically more involved in ML developments significantly more represented than the others in the latest editions. This effect possibly reflects an evolution in the perception of ML: from a generic tool widening and expanding the education of a researcher, to a specialization domain in the fields of data analysis and software engineering.

At the end of each hackathon event, participants were encouraged to provide a feedback on their experience through a questionnaire. From the analysis of the responses and from

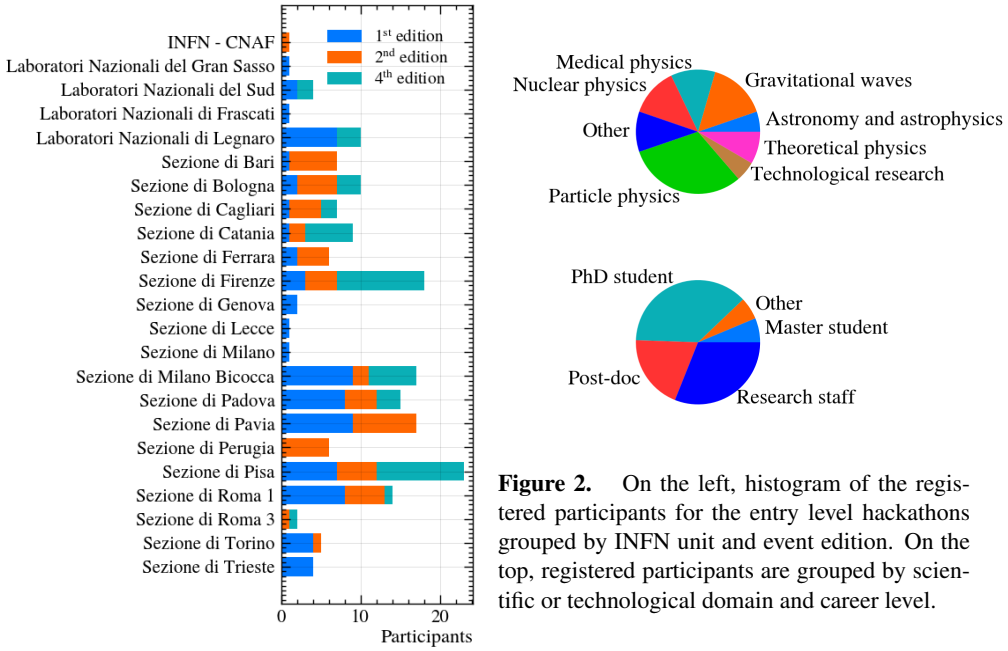


**Figure 2.** On the left, histogram of the registered participants for the entry level hackathons grouped by INFN unit and event edition. On the top, registered participants are grouped by scientific or technological domain and career level.

direct discussion with the participants, we believe that the demand for training opportunities on ML is indeed shifting towards more advanced or specialized topics and we recommend the organizers of future training events to tune the selection of the topics accordingly.

### 3.2 The third hackathon in Bari

In November 2022, a third edition of the hackathon, in its *advanced-level* format, was organized in Bari. The afternoon of the first day was devoted to a technological introduction to the ongoing evolution of the *High-Performance Computing* infrastructures available to the INFN researchers, in particular at CNAF (Bologna) and at ReCaS (Bari). The Cloud overlay connecting the two sites and the Cloud services designed to ease scientific computation tasks were also discussed, with live demonstrations and hands-on sessions. The second and the third days were entirely devoted to discuss advanced ML solutions applied to research tasks. In particular, Deep Learning architectures composed of multiple neural networks, such as autoencoders, adversarial models, and generative models were shown on the second day, presenting applications to the physics data analysis in CMS and in Medical Physics. Graph Neural Networks, Transformers, and concepts of Explainable Artificial Intelligence were discussed on the third day, practicing on Jet reconstruction data from CMS and on Genetic Sequences from a public data repository. Finally, the morning of the fourth day hosted seminars and demonstrations of active developments on ML of relevance for multiple of the application domains. In particular, Hyper-Parameter Optimization techniques using Bayesian Optimization and Genetic Algorithms, and technologies to deploy Deep Neural Networks on FPGAs were presented. For the hands-on sessions and the hackathon exercises, each participant was entitled to access a containerized JupyterLab instance with 8 dedicated virtual cores, 64 GB of memory and a 10-GB partition of an NVIDIA A100 GPU. The hardware resources were provided partially by CNAF and partially by ReCaS through a Cloud overlay making the differences in the underlying hardware and configuration transparent to the participants.

As for the entry-level event, a questionnaire was circulated through the participants and the responses analysed. Beyond the overall appreciation of the event (more than 92% of feedback was positive), the responses identify the presentation of advanced models, rather than their application to research problems, as the most relevant for the audience and the most appreciated. Given the valuable and abundant material available online, the feedback was partially surprising, and is being considered seriously for defining the agenda of the upcoming event in Pisa.

## 4 Community and Knowledge Base

To encourage the participation of researchers, students, and ML practitioners from the diverse and rich environment of INFN, and to create new opportunities for sharing their knowledge and works, ML_INFN organizes weekly virtual, national seminars where researchers present their work. The ML_INFN seminars are well received by the community and provide an opportunity to learn about the application of different ML algorithms to multiple research domains. Each seminar is dedicated to a different topic and lasts about one hour, including time reserved for questions and in-depth discussion. Moreover, summaries of the seminar topic, of the following discussion, and of the suggestions the speaker may receive from the audience are collected in a shared document.

Since the beginning of ML_INFN, 26 different seminars have been organized, covering the topics in Refs. [19–40].

The weekly seminars also represent an important opportunity for maintaining the community aligned on the availability and status of resources, on the plans for the upcoming future, and on the organization of the hackathons.

Finally, a mailing list, *ml-infn-csn5@lists.infn.it*, and an Atlassian Confluence web repository [41] are used to share learning resources, materials, and for urgent communications. The Confluence page includes the pointers to the past and upcoming hackathon events, a list of the documented ML use cases of relevance to ML_INFN, the list of public datasets from a scientific background that can be used to prepare lectures and hands-on, but also to present and benchmark new models in publications. The lists of datasets and use cases are regularly expanded with new entries. In order to be inserted into the knowledge base a use case should have reproducible code and a public dataset available to serve as a starting point for newcomers to practice with the presented software solution and as a common reference to compare the performance of multiple algorithms.

## 5 Future developments and technological outlook

The provisioning model developed by ML_INFN, based on the static association of GPU-empowered VMs to research projects, has demonstrated the effectiveness of a Cloud approach to high-profile resource sharing by enabling a quick development of a relatively large number of applications by different, and sometimes competing communities. Nevertheless, the model should be improved to enable better scalability and increase the fraction of time the GPUs are actively used for interactive development or optimization campaigns. In particular, techniques to decouple the provision of the hardware accelerator from the filesystem where the algorithm is developed are critical. Indeed, the development cycle of ML applications, often led by Ph.D. students and young postdocs, is subject to long inactive periods. The ability to reallocate the GPU for, possibly opportunistic tasks, during the time slots of development inactivity would pave the way towards a more effective usage of the shared resources.

A first prototype of Kubernetes overlay designed to enable decoupled access to GPU and storage resources with acceptable degradation of the I/O performance is being developed.

The important experience of the CNAF team in managing job queues with complex priority policies with HTCondor will then be crucial to design a batch system opportunistically exploiting the GPUs not used for development activity. The investigation of other hardware accelerators, such as FPGAs and Quantum Processors provisioned in a Cloud environment, is also planned for the future evolutions of this research.

The outlined program has been submitted to the INFN as part of a new initiative, named `AI_INFN`, that, if approved, will be active for the next three years (2024-26).

## 6 Conclusion

Artificial Intelligence, and in particular ML, have transformed the technological landscape of digital data processing and analysis over the last ten years. ML_INFN is a technological initiative of INFN to foster the adoption of modern techniques in the context of the fundamental and applied research in which it is involved. The three pillars of ML_INFN are designed to ease the adoption of ML by INFN researchers by providing a shared computing infrastructure empowered with GPUs and provisioned through a Cloud infrastructure, by organizing training events in the form of *entry-level* and *advanced hackathons*, and by encouraging a network of competence across the geographically distributed units and laboratories of INFN and the multiple research communities, spanning from High Energy Physics, to Gravitational Wave analysis, Theoretical Physics, and Medical applications.

During the four years of activity, ML_INFN has set up a computing infrastructure that was demonstrated capable of supporting tens of concurrent Deep Learning Training processes accelerated through a Multi-Instance GPU partitioning of A100 graphic units, possibly made available by different computing centers across the country. Three *entry-level* hackathon events, organized as online events, reached more than 150 participants, well distributed through the INFN units and career levels. An *advanced* hackathon was organized in Bari and another is planned by the end of the year to take place in Pisa. In addition, 26 seminars on specific applications of ML to research activities of relevance for INFN have been organized. Part of them have been recorded and the following discussion made available to the group in the form of minutes. A website based on Atlassian Confluence was also set up to collect examples, realistic applications, and code snippets to enhance the opportunity of connections between researchers facing similar problems or adopting similar techniques in different domains of the INFN research.

The transformation of the computing landscape induced by the development of ML and artificial intelligence is still dawning. An ambitious program to evolve the Cloud infrastructure into a more proficient and effective computing facility for ML and GPU-accelerated digital data analysis applications is at the core of a new initiative, named AI_INFN, recently submitted to INFN. If approved, it will certainly keep contributing to the evolution of the scientific computing in Italy as ML_INFN did for the last four years.

## References


[1] T. Boccali, Rev. Phys. **4**, 100034 (2019)

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019), 1903.10563

[3] F. Fanzago et al., *INFN and the evolution of distributed scientific computing in Italy*, in *26th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2023)* (2023)

[4] *OpenStack*, https://www.openstack.org, accessed on 01/12/2023

[5] *Ansible*, https://www.ansible.com, accessed on 01/12/2023

[6] *TOSCA*, https://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.3/os/TOSCA-Simple-Profile-YAML-v1.3-os.html, accessed on 01/12/2023

[7] *Indigo IAM*, https://github.com/indigo-iam/iam, accessed on 01/12/2023

[8] *JupyterHub*, https://jupyter.org/hub, accessed on 01/12/2023

[9] *Docker*, https://www.docker.com, accessed on 01/12/2023

[10] *JupyterLab*, https://jupyter.org, accessed on 01/12/2023

[11] *DockerHub*, https://hub.docker.com, accessed on 01/12/2023

[12] *CernVM-FS*, https://cernvm.cern.ch/fs/, accessed on 01/12/2023

[13] *Pandas*, https://pandas.pydata.org, accessed on 01/12/2023

[14] *TensorFlow*, https://www.tensorflow.org, accessed on 01/12/2023

[15] *PyTorch*, https://pytorch.org, accessed on 01/12/2023

[16] *Grafana*, https://grafana.com, accessed on 01/12/2023

[17] *Streamlit*, https://streamlit.io, accessed on 01/12/2023

[18] *Zoom*, https://zoom.us, accessed on 01/12/2023

[19] L. Layer, T. Dorigo, G. Strong (2023), 2301.10358

[20] L. Anderlini, M. Barbetti, PoS **CompTools2021**, 034 (2022)

[21] G.C. Strong, Mach. Learn. Sci. Tech. **1**, 045006 (2020), 2002.01427

[22] L. Giannini, Ph.D. thesis, Scuola normale superiore di Pisa, Pisa, Scuola Normale Superiore (2020)

[23] M. Migliorini, R. Castellotti, L. Canali, M. Zanetti, Comput. Softw. Big Sci. **4**, 8 (2020), 1909.10389

[24] C. Ariza-Porras, V. Kuznetsov, F. Legger, Comput. Softw. Big Sci. **5**, 5 (2021), 2007.03630

[25] A. Bombini, F. Bofías, C. Ruberto, F. Taccetti, Rendiconti Lincei. Scienze Fisiche e Naturali **34**, 867 (2023)

[26] M. Canaparo, E. Ronchieri, EPJ Web Conf. **214**, 05007 (2019)

[27] C. Scapicchio, A. Retico, F. Lizzi, M. Fantacci, Physica Medica **104**, S42 (2022)

[28] D. Spiga, D. Ciangottini, M. Tracolli, T. Tedeschi, D. Cesini, T. Boccali, V. Poggioni, M. Baioletti, V.Y. Kuznetsov, EPJ Web Conf. **245**, 04024 (2020)

[29] L. Anzalone, T. Diotalevi, D. Bonacorsi (2022), 2202.00424

[30] L. Clissa, Ph.D. thesis, Bologna U. (2022)

[31] S. Mariani, L. Anderlini, P. Di Nezza, E. Franzoso, G. Graziani, L.L. Pappalardo, J. Phys. Conf. Ser. **2438**, 012107 (2023)

[32] G. Grosso, Ph.D. thesis, Padua U. (2023)

[33] V. Kuznetsov, L. Giommi, D. Bonacorsi, Comput. Softw. Big Sci. **5**, 17 (2021), 2007.14781

[34] G. Carloni, A. Berti, C. Iacconi, M. Pascali, S. Colantonio, *On the Applicability of Prototypical Part Learning in Medical Images: Breast Masses Classification Using ProtoPNet* (2023), pp. 539–557, ISBN 978-3-031-37659-7

[35] G. Grosso, N. Lai, M. Letizia, J. Pazzini, M. Rando, L. Rosasco, A. Wulzer, M. Zanetti, Machine Learning: Science and Technology **4**, 035029 (2023)

[36] M. Barbetti, *Lamarr: LHCb ultra-fast simulation based on machine learning models deployed within Gauss*, in *21th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI meets Reality* (2023), 2303.11428

[37] F. Vaselli, A. Rizzi, F. Cattafesta, G. Cicconofri (CMS), Tech. rep., CERN, Geneva (2023), https://cds.cern.ch/record/2858890

[38] S. Giagu, L. Torresi, M. Di Filippo, Front. in Phys. **10**, 909205 (2022)

[39] M. Lorusso, D. Bonacorsi, D. Salomoni, R. Travaglini, D. Michelotto, D.C. Duma, P. Veronesi, PoS **ICHEP2022**, 243 (2022)

[40] F. Viola, B. Martelli, D. Michelotto, E. Fattibene, A. Falabella, S. Dal Pra, L. Morganti, L. Dell'Agnello, D. Bonacorsi, S.R. Tisbeni, EPJ Web Conf. **245**, 07008 (2020)

[41] *ML-INFN dashboard*, `https://confluence.infn.it/display/MLINFN/ML-INFN+Dashboard`, accessed on 01/12/2023