INFN
AI Strategy
2024

INFN
Istituto Nazionale di Fisica Nucleare

INFN
AI Strategy
2024

Istituto Nazionale di Fisica Nucleare

Executive Summary

Founded in 1951, the Istituto Nazionale di Fisica Nucleare (INFN) is the Italian research agency dedicated to studying the fundamental constituents of matter and the laws governing them. The INFN conducts theoretical and experimental research in subnuclear, nuclear, and astroparticle physics; it also performs studies in applied physics, by using the same techniques.

The mission of the institute includes advancing scientific knowledge, fostering innovation, and contributing to the global scientific community through its research and technological developments; to this purpose, it operates a large infrastructure of data centres.

In this context, the INFN funds, deploys, operates, and collaborates with facilities for large-scale AI model training and inference, leveraging national and EU funding opportunities. Interoperability of computing facilities is achieved through the standardisation of the authentication procedure and the federation of computing and storage resources.

The INFN funds initiatives that focus on foundational Artificial Intelligence (AI) and Quantum Machine Learning and actively promotes cross-fertilization with external entities through public and private collaborations (e.g. industrial partnerships and the refinement of public administration processes).

The INFN collaborates with national and international partners on AI-related projects, with particular attention to maintaining scientific reliability. This includes addressing issues like induced biases, ethics, and discrimination, to streamline and optimise productive and socially relevant processes.

Specialised training and support are offered to researchers and graduate students to promote a pervasive use of AI within INFN's research domains. Training opportunities are developed in collaboration with universities, such as through the Italian National PhD program on AI, and with other institutions.

The INFN believes that its expertise can have an immediate impact on the adoption of AI systems for the benefit of society as a whole.

To accelerate progress, the INFN welcomes national, European, and global initiatives that fund these activities (at the infrastructure, foundational and application layers – including opportunities for personnel training and advancement), also via partnerships with the industrial sector.

The INFN strives for the creation of a pan European initiative to foster and support the development of a prosperous AI ecosystem at the continental level, at the foundational and application levels, via:

- The realization of a state-of-the art distributed AI-tailored e-Infrastructure, under the coordination of the EuroHPC Joint Undertaking (JU) and designed to accommodate the use cases of research, society and productive system. The experience and the needs of the user communities, via current and additional EuroHPC Centers of Excellence (CoE), should be central in the design and operational choices for such systems. **In particular, the INFN aspires for a thematic Center of Excellence on Particle and Nuclear Physics, where the domain expertise on large scale computing and AI can be organized and shared with different domains within EuroHPC.**

- The funding of a **federated virtual institute** to aggregate national initiatives on AI: given the immaterial activities it would carry on once the infrastructure aspects are covered by the EuroHPC JU, the INFN thinks that this solution is to be preferred to a physical institute located in one or few locations. The institute would serve as the cornerstone of the European strategy for AI, organize the global plan of work and coordinate the high-level activities of the institutes and companies involved, which would be selected among the excellences with an inclusive domain representation. Among its activities and funding opportunities, large emphasis should be placed on researchers' training and mobility.

- Funding opportunities (via framework projects, ERC or directly by the above-mentioned institute) for calls intended to foster progress in AI, on the foundational and application levels. In particular, the INFN would welcome:
  - Opportunities for foundational studies on AI from the point of view of physics (complex systems and statistical mechanics)
  - Opportunities for application-level studies in the core domains of INFN research (AI on large scientific datasets, AI on applied physics, AI for theory)
  - Special attention to potentially disruptive technologies such as Quantum Machine Learning.

INTRODUCTION

The National Institute for Nuclear Physics (INFN) is the Italian research agency dedicated to the study of the fundamental constituents of matter and the laws that govern them. Operating under the supervision of the Ministry of Universities and Research, the INFN engages both in theoretical and experimental research in subnuclear, nuclear, and astroparticle physics. As a key player in international collaborations, INFN is a member of prominent institutions and laboratories such as the European Organization for Nuclear Research (CERN) in Geneva, the European Gravitational Observatory (EGO) in Pisa, the European Synchrotron Radiation Facility (ESRF) in Grenoble, the European Spallation Source (ESS) in Lund, among others, and is involved in numerous high energy, nuclear, and astroparticle physics collaborations and committees.
Over time, applied physics has also emerged as a vital research field for INFN, with contributions ranging from detector design to biomedicine (the INFN has established collaborations with several Scientific Institutes for Clinical Research and Treatment), and in the field of Cultural Heritage.

In the past two decades, the INFN has developed and operated (also profiting from EU funded initiatives) a large, distributed computing infrastructure to process the data generated by experimental and theoretical collaborations, in physics and in other scientific domains. This infrastructure is part of a broader global collaboration with the Worldwide LHC Computing Grid (WLCG) project and the European Grid Initiative (EGI), both aligned with the exascale computing vision fostered by the European High-Performance Computing Joint Undertaking (EuroHPC). A key strength of these federations has been the early development and adoption of a standardised set of protocols and tools for authorization, authentication, job management, and data access, facilitating interoperability between centres otherwise heterogeneous.
INFN also participates in the European Science Cloud Association (EOSC-A) to advance Open Science in the service of creating new knowledge, inspiring education, spurring innovation and promoting accessibility and transparency. EOSC aims to provide European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and reuse data, tools and services for research, innovation and educational purposes. INFN participates in several EOSC-related projects; in the Artificial Intelligence (AI) context, INFN is involved in AI4EOSC, which aims at further increasing EOSC tools and services by developing an array of AI, Machine Learning (ML) and deep learning models, bundled and customisable for ease of use to researchers.

AI, and in particular ML, has become a crucial technology and a research focus for INFN. While AI is highly relevant for social and industrial applications, the development of reliable models for scientific applications requires a dedicated effort. Compared to general-purpose AI applications, such as Natural Language Processing or Multimodal Generative Models, scientific ML has more stringent requirements for reproducibility, reliability, accuracy, and explainability.

At the national level, the INFN leads the *National Center for research on High Performance Computing, Big Data and Quantum Computing* (ICSC), a 320 million Euro initiative funded by the National Recovery and Resilience Plan (NRRP) under the Next Generation EU framework, aimed at modernising the Italian scientific computing infrastructure, also for the benefit of the productive system. The project involves 50 partners, including universities, research centres, and major Italian private companies. INFN is also one of the partners of the Italian national project funded by the NRRP on AI, the Future AI Research (FAIR), which connects the skills of the physics community with those of computer scientists. The research topics range from the simulation of the neuronal activity to the development of innovative AI-based solutions for the analysis of big data of physics experiments and applications in the medical field.

Fundamental experimental physics benefits from large volumes of collected and simulated data, along with a deep understanding of the effects of the experimental instruments and of the underlying physical models. These aspects make this field an ideal environment for advancing both the theoretical foundations and the applications of ML.
Unlike data sourced from, for instance, social media, High Energy Physics (HEP) data is highly uniform, accurately curated, and well understood by a large research community, thus enabling the design, development, and validation of AI models with a reduced bias. The INFN scientific community also has a strong interest in the mathematical aspects of the learning procedure, which can be investigated using tools and techniques from statistical mechanics and the physics of complex systems, potentially including realistic simulations of human brain-like models. This approach greatly enhances the potential for explainable and robust AI results, mitigating the so-called "black box effect" often encountered in large AI models.

The INFN also recognizes the growing importance and the opportunities in developing ML algorithms for Quantum Computing. This research area will enable the implementation of quantum algorithms via learnable quantum approximators, a field rooted in quantum physics on which INFN has an extensive expertise.

The INFN applied physics tradition continues to grow, with significant contributions to fields such as Medical Physics, Digital Healthcare, and Cultural Heritage preservation. These areas stand to benefit from INFN's computing infrastructure and global networks, which are poised to elevate their competitiveness on the international stage.

This position paper outlines the future development of what we call scientific AI within INFN, focusing on the infrastructure provisioning model integrated into the upcoming European HPC and EOSC landscape. Moreover, it discusses the realisation of software and algorithms for scientific applications, while addressing the external conditions necessary for success, including collaboration and access to adequate funding from Research Funding Organisations, from national to the EU levels.

INFN SPECIFICITIES IN THE AI LANDSCAPE

Being primarily composed of physicists, the INFN community presents unique and complementary characteristics when compared to, for example, the Computer Science communities.
On the one hand, the theoretical foundations of AI, and ML in particular, intersect with the study of complex systems, a field where INFN has expertise, employing tools ranging from statistical methods to mathematical physics.
Quantum AI is also of considerable interest to INFN, through multiple directions. The development and deployment of quantum algorithms are already active research fields, with applications of Quantum Machine Learning (QML) in simulating quantum systems (e.g., for high energy collisions and particle physics detectors simulations), in utilising quantum processors for highly combinatorial algorithms (such as interpreting detector responses in terms of physics objects), and in several other application areas relevant to INFN research. Additionally, the INFN possesses a vast knowledge in the technological fields needed to develop quantum computers: vacuum, shielding, and cooling aspects are very similar to those needed for the design and the deployment of physics detectors and colliders.
On the other hand, the INFN collects, curates, and provides access to large datasets, derived from physics instruments and from detailed simulations of physical processes. Recently, the INFN has expanded its role to include the provisioning of sensitive data, such as large-scale medical databases, through the implementation of ISO-certified platforms.
One of the key distinctions between INFN's datasets and larger datasets, such as those generated from social data harvesting, lies in the level of understanding, in terms of the underlying model, and curation of the former. Physics analyses typically aim for percent-level precision or better, and achieving such accuracy requires an equally precise understanding of the data.
The INFN is therefore able to leverage these highly refined datasets (sometimes at the exabyte scale, such as in the HEP community) as training data for AI systems. This enables the testing, evaluation, and validation of explainable AI models to a degree that is not possible with noisier, less understood, or more biased datasets, even when those datasets are larger in size.

We believe that the development of Scientific AI and its application to INFN research domains is of fundamental importance, not only for the advancement of its research mission, but also for the broader advancement of AI as a field.

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE FOR SCIENTIFIC RESEARCH

While many research areas benefit from the rapid development of AI models for processing images and natural language, the INFN focuses on advancing foundational AI algorithms tailored for scientific research. These include quantum AI, explainable AI, and Physics-Informed ML, together with AI architectures for quantum computers, techniques for processing multidimensional data as graphs, and generative models for simulating synthetic data and detector responses.

Applying AI algorithms across different research domains requires a deep understanding of the specific context and challenges. Consequently, the selection and configuration of these algorithms are typically performed within highly specialised scientific communities. For example:

- AI is crucial for analysing data from present and upcoming instruments like the LHC and the High Luminosity LHC at CERN, and future projects like the Future Circular Collider and the Einstein Telescope for the detection of Gravitational Waves. Indeed, classical algorithms would limit the full potential of these multi-billion Euro research infrastructures;
- AI is revolutionising diagnostics in areas strongly related to medical physics, accelerating the acquisition and reconstruction of signals from imaging systems, and helping to improve the interpretation of large health-related datasets to support patient-specific diagnoses and therapeutic interventions;
- AI assists researchers in operating complex infrastructures like accelerator complexes and data centres, while also automating data acquisition and processing in the field of Digital Cultural Heritage, for tasks such as the inspection of artworks.

The INFN strongly supports the development of European networks and exchange programs for graduate and doctoral students, to encourage the identification of career paths focused on developing and deploying AI techniques for scientific research and its applications. These initiatives will help enhance the competitiveness of European research on the global stage.

DEVELOPMENT OF AN ARTIFICIAL INTELLIGENCE FRAMEWORK FOR MEDICAL RESEARCH

Medical data, unlike social science or HEP data, presents unique challenges, due to its heterogeneity, encoding the health status of an individual, and encompassing information from various sources and in various formats: imaging data, time series, medical records, textual notes, genetic profile, environmental factors, lifestyle details, and so on. Addressing these challenges requires technical innovation in collecting, curating, annotating, managing, and mining such data with AI, also following the prescription of GDPR tools.

High-quality annotated datasets are essential for developing a new generation of AI models with broader applicability, particularly in supporting clinical workflows. Substantial investments are required to build and maintain these datasets, as well as to train and validate AI models using them. Despite the availability of several AI-based products for diagnostic prediction or for automated segmentation of organs or lesions, many of which are already CE-marked, and the constant influx of new R&D solutions in the literature, a significant leap forward in this field is still required. The generalisation capabilities of these tools remain a topic of debate; moreover, current implementations often have a narrow scope, typically focusing on a single specific task, whereas clinicians seek AI systems that can predict more clinically meaningful endpoints, such as lesion malignancy, patient's response to therapy, and the likelihood of relapses. To achieve this goal, AI-based predictive models must be trained using a broad spectrum of health-related information, including clinical records, medical images, genetic profile, and other relevant data. Developing effective analysis strategies for handling heterogeneous and complementary data sources (multi-input data) requires dedicated effort. It is essential to build trust among patients and physicians, and this is closely tied to advancements in explainable-AI (XAI) methods for such complex algorithms. Despite significant progress in XAI research over the last few years, there remains a need for established quantitative strategies that are comprehensible to non-AI experts, such as physicians. Given the inherently multidisciplinary nature of this research, medical physicists can play a pivotal role as a bridge between different fields, standing at the intersection of established statistical methods and their clinical applications.

Large Language Models (LLMs), although originally developed for text interpretation and generation, are increasingly proving to be valuable resources for image processing and analysis, particularly in the context of multimodal LLMs. Efforts to fine-tune LLMs for the medical domain, such as ClinicaBERT, BioBERT, and GeneGPT, are already underway; they can serve as foundational models for future developments, including in medical image analysis. However, LLM-based multi-modal solutions integrating medical image analyses have yet to be realised. The complexity and the heterogeneity of signals acquired by medical imaging devices, such as Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Electroencephalography (EEG), and so on, present significant challenges that require the knowledge and expertise of physicists, especially medical physicists.

To enable this development of a multi-input LLM capable of integrating accurate medical image analysis across different modalities, a robust data collection and management framework must be established at both the national and European levels, in accordance with the FAIR data principles. Training AI systems requires annotated data, which are difficult to obtain and use efficiently. Additionally, an adequate computing infrastructure is crucial for training and validating models on real clinical datasets.

COMPUTING INFRASTRUCTURE FOR ARTIFICIAL INTELLIGENCE

All the software stacks needed for AI require computing power that exceeds the capabilities of standard CPUs to ensure an efficient and timely computation, also from an energy efficiency standpoint. Graphics Processing Units (GPUs) are the solution most commonly employed for such tasks. Additionally, other specialised accelerators, such as Field-Programmable Gate Arrays (FPGAs) and Tensor Processing Units (TPUs), can offer further acceleration with significantly reduced latency and energy consumption.
Operating hardware accelerators within distributed computing infrastructures, which include AI-specialised hardware, standard hardware, and large data repositories, is a strategic research focus for INFN. This vision requires the development of middleware components that interconnect existing facilities to create a logically uniform infrastructure, which the INFN identifies as the "datalake". This infrastructure must also be capable of incorporating facilities dedicated to handling and processing sensitive data, certified for compliance with ISO standard.
Initiatives such as ICSC, alongside similar programs in other scientific fields, should receive robust support and be institutionalised as permanent and stable components of public research infrastructure.

It is also particularly important to secure funding opportunities for middleware development, addressing topics such as authentication and authorization mechanisms, the integration of Grid, Cloud, HPC, and Quantum systems, support for adapting and porting codes to heterogeneous computing environments, and the efficient access and transfer of large datasets in the datalake model.

Finally, INFN's strong tradition of collaboration with national and international partners from both academia and industry underscores the need to develop *trust networks* to facilitate globally authenticated access to specialised computing resources.

CONCLUSIONS AND VISION

INFN believes that AI is a fundamental resource for analyzing the enormous amounts of data produced by the experimental collaborations in which it participates, accelerating simulations and facilitating the process of discovery and interpretation of observed phenomena. In the application domain, the INFN fosters the adoption of AI in studies linked to life sciences, development of novel devices and for the design and operations of large equipment. At the same time, the INFN recognizes the increasing importance and impact of the AI, also beyond its research domains.
The INFN promotes the development of AI algorithms in collaboration with universities and other research institutions and industries, aiming for an AI that is transparent, ethical and adaptable suitable both for fundamental research and for multidisciplinary applications to the societal challenges.

The INFN supports the creation of a pan-European initiative, in the form of a federated virtual institute, aimed at promoting and supporting the development of a thriving AI ecosystem at continental level, both at fundamental and applicative levels, leveraging the existing HPC e-Infrastructure coordinated by the JU with user communities organised in the EuroHPC Centres of Excellence. In particular, the INFN aspires to create a Thematic Centre of Excellence on Particle and Nuclear Physics, where domain expertise on large-scale computing and AI can be organised and shared with different domains within EuroHPC.
Among its activities and funding opportunities, large emphasis should be placed on researchers' training and mobility.
The INFN finally considers that science-driven AI actions at the European level, should build on the success of large infrastructures (e.g. CERN, ESS, EGO, ESRF) and the vast communities behind them.