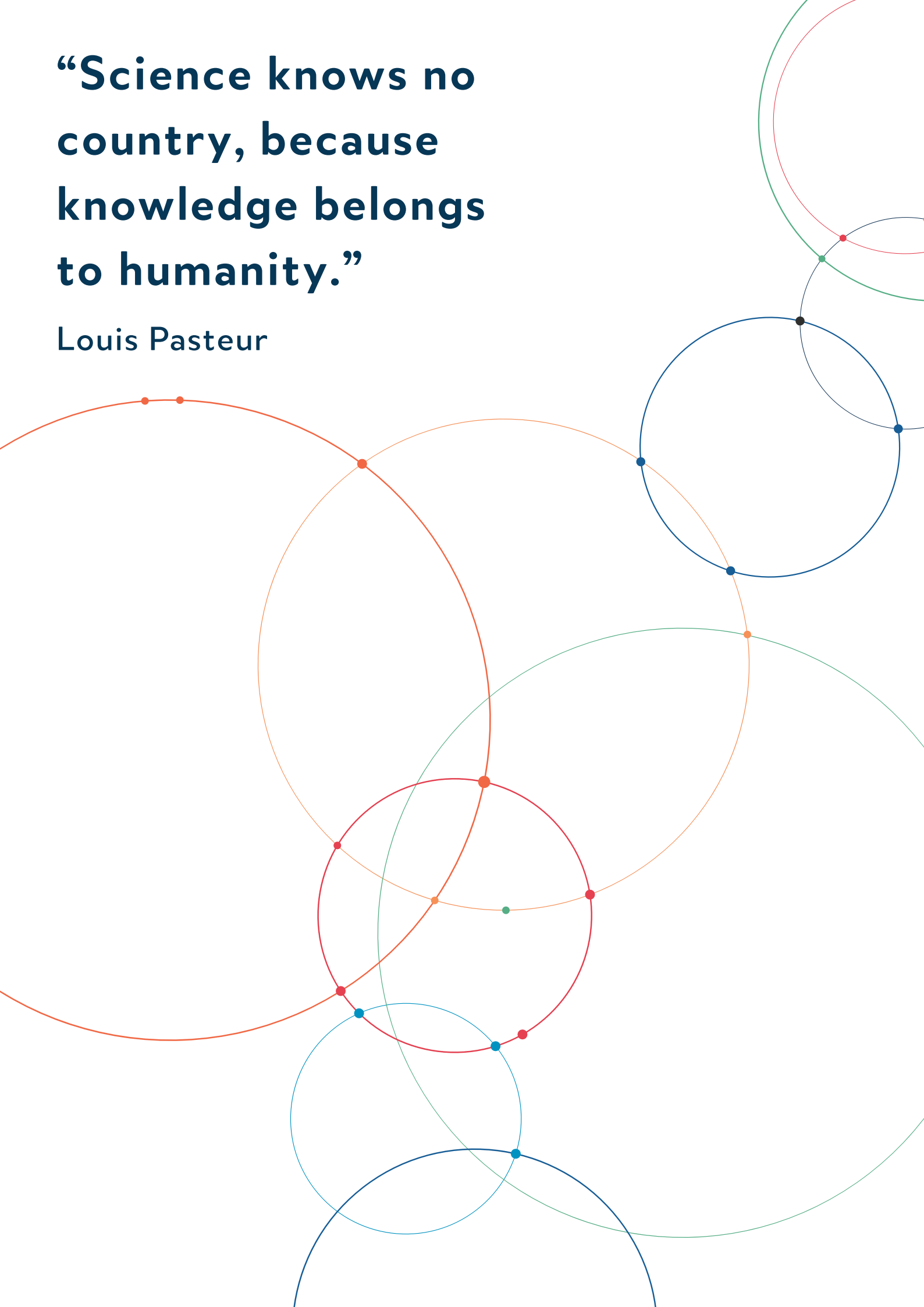
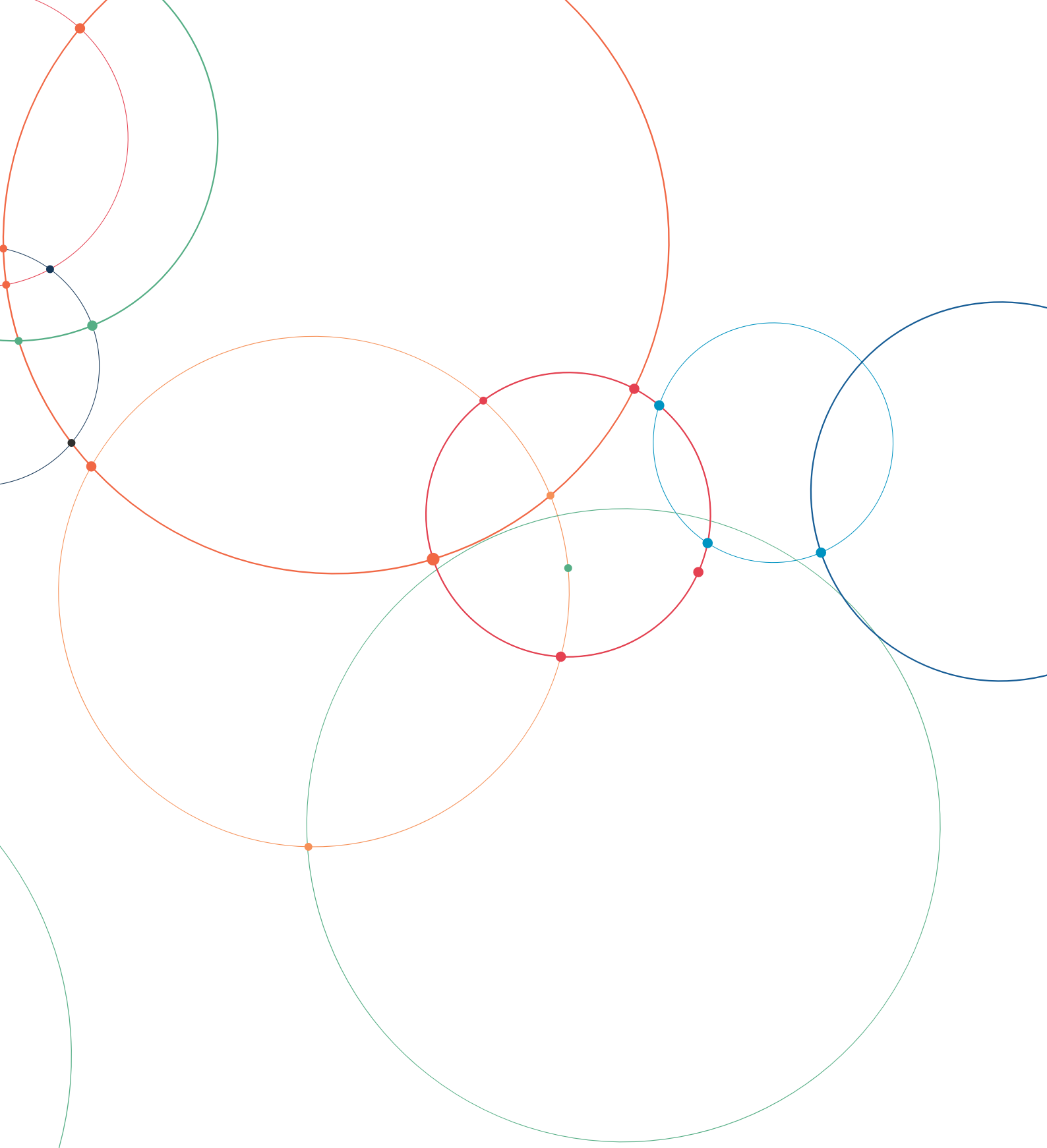
# 2023
# ANNUAL REPORT

CERN openlab

“Science knows no country, because knowledge belongs to humanity.”

Louis Pasteur

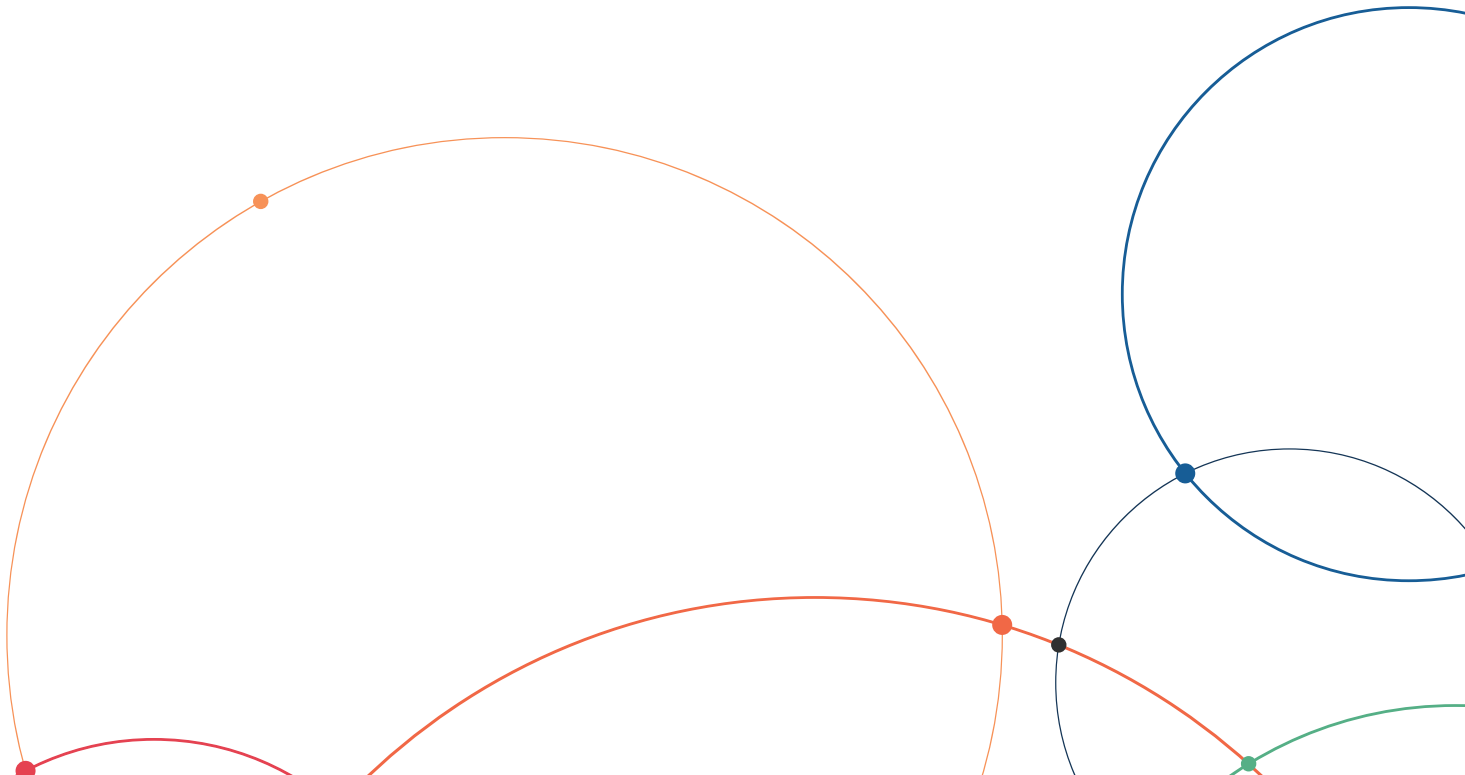




# CERN OPENLAB

## 2023 ANNUAL REPORT

# TABLE OF CONTENTS



# CERN OPENLAB

CERN openlab stands as a testament to over two decades of pioneering history and expertise. Established in 2001, it remains a ground-breaking **public-private partnership**, forging collaborations between leading ICT companies and research centres worldwide, uniting them with the forefront of scientific innovation at CERN. These partnerships fuel CERN researchers with invaluable opportunities and resources to push the boundaries of computing, essential for tackling the unprecedented challenges presented by monumental projects like the High-Luminosity Large Hadron Collider (HL-LHC) or the Square Kilometre Array (SKA).

For more than twenty years, CERN openlab has been the bedrock of partnership, providing a structured and collaborative framework for industry and research organisations to engage with CERN researchers. Its framework not only fosters collaboration but also facilitates industry participation and investment.

Collaboration is the core of CERN openlab, where industry and scientific researchers co-develop innovative solutions. Together, we embark on four primary missions: establishing strategic industry collaborations, fuelling technological innovation, exposing technology to researchers, and nurturing knowledge and growth in young STEM researchers.

Now embarking on its eighth phase, with 23 years of transformative history, CERN openlab embraces a new era of multidisciplinary collaboration, with fresh insights and activities. As the world evolves and new challenges emerge, our dedication to accelerate scientific computing, propelling technologies beyond HEP, and fostering positive societal impacts on a global scale remains unwavering.

**Maria Girone**
Head of CERN openlab

Since its inception, CERN openlab has fostered the development of big data scientific research through **four primary missions**:

## 1 Establishing strategic industry collaborations

CERN openlab projects provide an ideal **incubator for collaborations**, where they can be formed and longer-term partnerships can be built. They act as the first step in the establishment of strategic collaborations between CERN and other organisations interested in investing in the future of science and technology.

## 2 Fuelling technological innovation

CERN openlab serves as an **incubator for new technologies**. It forms a dynamic hub where CERN and its partners collaboratively push the boundaries of ICT technology. This cooperative synergy propels the co-development of new ideas and innovative solutions.

## 3 Exposing technology to researchers

CERN openlab provides access to new technologies available on the market to its members and the HEP community, **supporting critical tasks** of evaluation, adaptation, and benchmarking.

## 4 Nurturing knowledge and growth in young STEM researchers

CERN openlab plays a crucial role in **training the experts of tomorrow**. The CERN openlab summer student programme, supported through industry contributions, trains students through real-world, concrete, multi-disciplinary projects. Through various programmes and workshops, it equips the next generation of researchers with essential skills required to navigate the complex landscape of modern computing technologies.

**Maria Girone**
Head of CERN openlab

**Thomas Owen James**
CTO for AI and Edge Devices

**Antonio Nappi**
CTO for Platforms and Workflows

**Luca Mascetti**
CTO for Storage

**Luca Atzori**
CTO for Computing

**Killian Verder**
CTO Office

**Mariana Velho**
Chief Communications Officer

**Marina Banjac**
Junior Communications Officer

**Fariza Oulashova**
Junior Project Assistant

**Kristina Gunne**
Chief Administrative Officer

**Joelma Tolomeo**
Chief Financial Officer

# PHASE VII R&D TOPICS

## EXASCALE COMPUTING

Designing and operating distributed data infrastructures and computing centres poses challenges in areas such as networking, architecture, storage, databases, and cloud. These challenges are amplified and added to when operating at the extremely large scales required by major scientific endeavours. CERN is evaluating different models for increasing computing and data-storage capacity, in order to accommodate the growing needs of the LHC experiments over the next decade. All models present different technological challenges. In addition to increasing the on-premises capacity of the systems used for traditional types of data processing and storage, explorations are being carried out into a number of complementary distributed architectures and specialised capabilities offered by cloud and HPC infrastructures. These will add heterogeneity and flexibility to the data centres, and should enable advances in resource optimisation.

The next-generation of HPC technology offers great promise for supporting scientific research. Exascale supercomputers – machines capable of performing a quintillion, or a billion billion, calculations per second – are now becoming a reality. This change in the power of HPC technology, coupled with growing use of machine learning, will be vital in ensuring the success of future big science projects, such as the High-Luminosity Large Hadron Collider (HL-LHC).

## ARTIFICIAL INTELLIGENCE

The High-Luminosity LHC (HL-LHC), set to come online in 2029, will require roughly ten times the computing capacity we have today at CERN. Data-storage needs will also outstrip what it is possible to achieve with a constant investment budget by several factors. Even taking into account the expected evolution of technology, there will be a substantial shortage of IT resources. Thus, CERN openlab is exploring new and innovative solutions to help physicists bridge this resource gap, which may otherwise impact on the HL-LHC experimental programme.

Members of CERN's research community expend significant efforts to understand how they can get the most value out of the data produced by the LHC experiments. They seek to maximise the potential for discovery and employ new techniques to help ensure that nothing is missed. At the same time, it is important to optimise resource usage (tape, disk, and CPU), both in the online and offline environments.

Modern machine-learning technologies – in particular, deep-learning solutions – offer a promising research path to achieving these goals. Deep-learning techniques offer the LHC experiments the potential to improve performance in each of the following areas: particle detection, identification of interesting events, modelling detector response in simulations, monitoring experimental apparatus during data taking, and managing computing resources.

# QUANTUM COMPUTING

Following a pioneering workshop on quantum computing held at CERN in 2018, CERN openlab started a number of projects in quantum computing that are at different stages of realisation. These projects feed into the CERN Quantum Technology Initiative, launched in 2020.

While quantum computing should not be considered a panacea, this technology does hold significant potential:

It is poised to unlock unprecedented levels of computing power, improving our collective ability to simulate complex systems and understand the world around us.
It is likely to be much greener than classical forms of computing.
Quantum communication and encryption techniques are also exciting; these could, for example, play an important role in protecting citizens' privacy.
CERN is exceptionally well placed for quantum computing: as well as having experience with complex algorithms and with the physics that underpin this technology, the Organization has the required expertise in cryogenics, electronics, and materials. Nevertheless, we know that quantum computing technologies can only achieve their full potential for good if access is equitable and appropriate governance agreements and security systems are put in place. While the technology itself may be a decade or more away, the work to achieve these things needs to start now.

# BEYOND PARTICLE PHYSICS

By working with communities beyond high-energy physics, we are able to ensure maximum relevancy for CERN openlab's work, as well as learning and sharing both tools and best practices across scientific fields. Today, more and more research fields, such as medical research or space and Earth observation research, are driven by large quantities of data, and thus experience computing challenges comparable to those at CERN.

CERN openlab's mission rests on three pillars: technological investigation, education, and dissemination. Collaborating with research communities and laboratories outside the high-energy physics community brings together all these aspects. Challenges related to the life sciences, medicine, astrophysics, and urban/environmental planning are all covered in this section, as well as scientific platforms designed to foster open collaboration.

# PROJECTS OVERVIEW

CERN openlab projects for phase VII were integrated within the previous R&D topics. Each project compromises a project coordinator, technical team and collaboration liaisons, often collaborating with other groups at CERN, research institutes, or industry.

## EXASCALE COMPUTING

Real-time Data Processing for Level 1 Trigger: Scouting at CMS using CXL Memory-lake Architecture

Integration of Oracle cloud resources into CERN IT BC&DR project

Facilitate and Automatize Kubernetes Operations

Next Generation Archiver for WinCC OA

Heterogeneous Architectures Testbed

MADGRAPH5

Evaluation of Power CPU architecture for deep learning

Comtrade

Center of Excellence on AI and Simulation-based Engineering at Exascale (CoE RAISE)

InterTwin: An interdisciplinary Digital Twin Engine for Science

EMP2: Environmental Modelling and Prediction Platform

## ARTIFICIAL INTELLIGENCE

AI Models Registry in the Cloud

Data Analytics for Industrial Control Systems

Fast Detector Simulation

## QUANTUM COMPUTING

Quantum Databases for Dynamic Data Storage

## BEYOND PARTICLE PHYSICS

Digital Twin and Synthetic data in Healthcare

BIODYNAMO

# OVERVIEW OF INDUSTRY AND RESEARCH PARTNERS










# EXASCALE COMPUTING PROJECTS

EXASCALE COMPUTING

# REAL-TIME DATA PROCESSING FOR LEVEL 1 TRIGGER: SCOUTING AT CMS USING CXL MEMORY-LAKE ARCHITECTURE


**Emilio Meschi**
Project Coordinator

**Thomas Owen James**
**Giovanna Lazzari Miotto**
Technical Team

**Jason Adlard**
**Tony Brewer**
**Glen Edwards**
**Patrick Estep**
**Andrey Kudryavtsev**
Collaboration Liaisons from Micron




## PROJECT GOAL

The project described aims to use the Micron CXL-enabled memory devices as part of the ingestion and data processing chain for the L1 Scouting system at CMS, providing a coherent and seamless access to buffered data from multiple processors and compute accelerators, and a low-latency access/short term storage space for both raw and processed data at scale.

## BACKGROUND

The Compute Express Link (CXL) protocol is a new alternate protocol that can run over the standard PCIe physical layer, and dynamically multiplexes IO, cache and memory protocols. It is designed to empower a new generation of heterogeneous and disaggregated computing with efficient resource sharing, shared memory pools, enhanced movement of operands and results between accelerators and target devices, and significant latency reduction. CMS intends to profit from the capabilities of this new technology in the online processing solution for the L1 scouting data, and in doing so will pave the way for its utilisation in the wider community.

# PROGRESS

In 2023, our focus shifted from the prior Micron-CMS openlab project centred on deep learning inference acceleration, as outlined in the previous year's report, to a new initiative concentrating on CXL-enabled memory.

This transition prompted the design of a novel architecture for L1 Scouting online processing, wherein the RAM-disk is either substituted or augmented by a CXL-enabled "memory lake."

The initial stages called for a new demonstrator configuration leveraging Micron CXL memory modules. In September, two CXL 2.0, DRAM-based memory devices, each boasting 128 GB of memory, were installed in an AMD Genoa testbed server at CMS Point 5. Subsequent efforts focused on configuring and utilizing these modules, culminating in a comprehensive series of throughput and performance measurements. These evaluations employed standard and custom tools for various memory tiering management configurations, including NUMA balancing and transparent page placement.

In November 2023, key members of the CERN team engaged with Micron experts at the SuperComputing conference in Denver, USA. The teams used this opportunity to plan for the upcoming year, and exchange technical details and results.

# NEXT STEPS

The next line of investigation involves testing and understanding coherent memory sharing with dedicated accelerators, such as GPUs. Over the next 12 months, we will acquire a more extensive "memory lake" system, featuring expanded capacity and a CXL switch interconnect. This will propel us toward the ultimate goal of integrating CXL memory sharing into the CMS L1 Scouting system.

↓ The technical team (from the left to the right): Thomas Owen James, Giovanna Lazzari Miotto, Emilio Maschi.

EXASCALE COMPUTING

# INTEGRATION OF ORACLE CLOUD RESOURCES INTO CERN IT BC&DR PROJECT


**Exascale Computing**
R&D Topic

**Miroslav Potocky**
**Alexandros Stoumpis**
Technical Team

**Miroslav Potocky**
Project Coordinator

**Şengül Chardonnereau**
**Jérôme Designe**
**Sébastien Hurel**
**Stefan Jung**
**Cristobal Pedregal-Martin**
**Eva Dafonte Perez**
Oracle Collaboration Liaisons


## PROJECT GOAL

The aim is to establish a disaster recovery plan for CERN's crucial on-premises Oracle databases. This initiative focuses on enabling seamless switchover via Oracle Data Guard Database replicas operating asynchronously within Oracle Cloud Infrastructure. To secure replication transport, it employs private links across the GÉANT network, connecting to the Oracle Fast Connect endpoint situated in the Frankfurt datacenter.

## BACKGROUND

For a major incident impacting one of the CERN data centers leading to major part of infrastructure being unavailable for an extended period, it is necessary to have a strategy for where to build back the services deemed critical for CERN mission. Many could potentially be built back in the Cloud, considering data location and protection, as well as network bandwidth needs. In any case, a strategy and actions need to be developed and performed to allow for preservation of CERN critical data stored in on-premises Oracle databases and replicate them off-site (e.g. in Oracle Cloud Infrastructure), while still completely under full CERN control.

# PROGRESS

In 2023, the focus centered on enhancing Oracle Cloud Infrastructure (OCI) tenancy and streamlining automation tests for the Oracle Database replication process while assessing the performance of Disaster Recovery protocols.

A comprehensive overhaul of the Virtual Cloud Network configuration was executed, aligning it closely with CERN's internal network segmentation—dividing it into distinct subnets for general use, experiments, and technical purposes. Emphasis was placed on fortifying each subnet against unauthorized access, implementing robust logging, and devising detection infrastructure for the CERN computer security team.

Within this context, a Key Management System proof of concept was undertaken. Leveraging Oracle Key Vault software, it established a highly available encryption key and a 3rd party secret store, ensuring replication between on-premises infrastructure and OCI. Multiple replication tests were conducted, varying in data size from a few gigabytes to tens of terabytes. Measurements of latency and throughput were captured to effectively calibrate expectations for comprehensive disaster recovery reconstruction and subsequent failover tests.

Concurrently, the project implementation team pursued an expanded understanding of the OCI, engaging in various Oracle University trainings with the aim of obtaining formal certification credentials.

# NEXT STEPS

Production level database switch-over of all involved databases needs to be performed while capturing performance metrics to document required timeline for disaster recovery.
To conclude this project a complete tear-down and rebuild of off-site Oracle Cloud Infrastructure resources needs to be planned to validate configuration and documentation of all parts. In addition, security hardening and audit of data governance, access, and lifecycle - in cooperation with CERN Computer security team and CERN Data Protection Office - is necessary.

EXASCALE COMPUTING

# FACILITATE AND AUTOMATIZE KUBERNETES OPERATIONS

**Exascale Computing**

R&D Topic

**Adrian Karasinski**
**Antonio Nappi**

Technical Team

**Antonio Nappi**

Project Coordinator

**Eric Grancher** (CERN)
**Cristobal Pedregal-Martin** (Oracle)
**Garret Swart** (Oracle)
**Artur Wiecek** (CERN)

Collaboration Liaisons

ORACLE

## PROJECT GOAL

The main goal is to produce a tool that will help to validate, test and automate Kubernetes cluster upgrades. Currently the current process of validating a new Kubernetes version takes several weeks. We would like to simplify this and be able to run the same process in shorter time and in automated way.

## BACKGROUND

The main issue during upgrade to a newer version of Kubernetes cluster is that we cannot statically determine if our current Kubernetes workloads are going to break in the newer version because of API changes/deprecations. The only way to determine is to run it against a new Kubernetes cluster. When you have thousands of pods/resources this way to perform operations doesn't scale. The testing should not follow an empiric strategy but take advantage of a static analysis. This is the first step of a wider idea for statical analysis of service mesh dependencies.

# PROGRESS

Invent and develop kubernetes-diff application as working tool helping migration of workloads between different version of Kubernetes (K8s). As part of the project we have developed a tool that allows for the automated extraction of OpenAPI K8s schemas supported, which are then used by the application. The application allows detection of issues in Kubernetes workloads, against the chosen K8s cluster version. Because our tool can fully integrate with static system files or a running K8s cluster, it can be used in any scenario or context, for example, as a component of an automated pipeline, CI/CD system, script, or system terminal application. The result of the scan (differences/errors detection) is presented in popular formats such as JSON, YAML, and shows the results grouped in human-readable tables.

A big difficulty during the project proved to be the inconsistency of the kubernetes project with its own schemas, which means that a lot of K8s resources do not comply with the validation rules present in public OpenAPI, and refer to many different, hidden, private-based places in the code. As a result, time was spent on research on understanding and obtaining additional, hidden validation rules, which are essential if we are to make our tool overall usable.

Full description of the research was prepared in the course of the work, along with readme of scripts: runtime environment setup, setup metadata for debugging, support for DWARF, Delve, Go - enabling remote debugging of existing kube-apiserver instances (entry point for Kubernetes resources), which makes the experiment repeatable and possible to continue by anyone who repeats our setup and reverse engineering and debug engineering.

# NEXT STEPS

We will close the project at the beginning of 2024. We would like to make our investment repeatable for everyone and advertise it in order to increase awareness among the wider Kubernetes community to let them react.

It requires increased effort from several entities that must make several, intricate adjustments to the entire Kubernetes codebase - in fact, validation rules are actually found throughout the code. We have a situation where schemas don't match what's in the code and vice versa. Maintaining such a tool, forces us to manually review and diff the entire codebase every time new Kubernetes version is released.

EXASCALE COMPUTING

# NEXT GENERATION ARCHIVER FOR WINCC OA

**Exascale Computing**
R&D Topic

**Pedro Agostinho**
**Rafal Kulaga**
**Antonin Kveton**
**Ewald Sperrer**
Technical Team

**Rafal Kulaga**
Project Coordinator

**Pedro Agostinho**
**Ewald Sperrer**
**Christopher Stoegerer**
Siemens ETM Collaboration Liaisons

SIEMENS

## PROJECT GOAL

Our aim is to make control systems used for the LHC more efficient and smarter. We are working to enhance the functionality of WinCC OA (a SCADA tool used widely at CERN) and to apply data analytics techniques to the recorded monitoring data, in order to detect anomalies and systematic issues that may impact upon system operation and maintenance.

## BACKGROUND

The HL-LHC programme aims to increase the integrated luminosity – and hence the rate of particle collisions – by a factor of ten beyond the LHC's design value. Monitoring and control systems will therefore become increasingly complex, with unprecedented data throughputs.

Consequently, it is vital to further improve the performance of these systems, and to make use of data analytics algorithms to detect anomalies and anticipate future behaviour. Achieving this involves a number of related lines of work. This project focuses on the development of a modular and future-proof archiving system (NextGen Archiver – NGA) that supports different SQL and NOSQL technologies to enable data analytics. It is important that this can be scaled up to meet our requirements beyond 2023.

# PROGRESS

In 2023, the team focused on two main topics: preparation for the deployment of the NGA in all WinCC OA systems in ATLAS, LHCb and CMS and development of a backend for TimescaleDB.

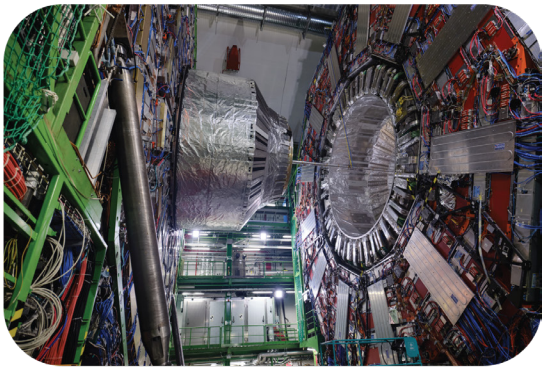
CERN-wide deployment of the NGA was initially planned for the Long Shutdown 3 (2026 – 2028). However, thanks to the very good experience with using the NGA in more than one hundred systems at the ALICE experiment, a decision to migrate already during Year-End Technical Stop 2023-24 has been taken by ATLAS, LHCb and CMS. The NGA team has supported this undertaking by providing the tools, support and performing database schema upgrades. All the steps of the process have been automated, allowing the interventions to be finished in unprecedentedly short time windows.

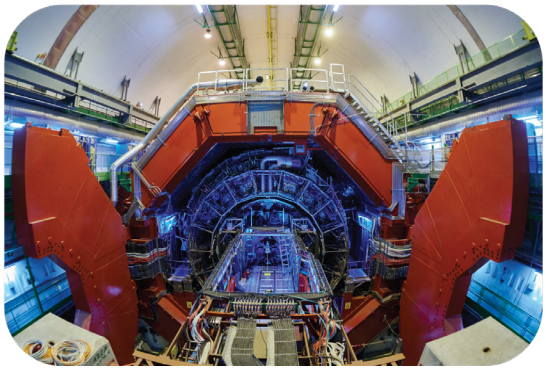
Based on the results of the performance and functional tests done in 2023, TimescaleDB is considered to be the best candidate to supersede InfluxDB as an alternative to Oracle. Significant progress has also been made on the TimescaleDB backend, paving the way for a preview release for pilot deployments in the first half of 2024.

# NEXT STEPS

In 2024, the team will focus on implementing some missing features in the Oracle backend and continuing the development of the TimescaleDB backend. The performance of TimescaleDB at the scale of large CERN distributed WinCC OA systems (hundreds of nodes, millions of datapoints, retention policies spanning years) will be evaluated. Additionally, the flexibility of queries will be improved by allowing to better specify the source of the data.

↓ After the year-end technical stop 2023-2024, ATLAS, CMS, ALICE and LHCb are using the NextGen Archiver in all their product systems.





EXASCALE COMPUTING

# HETEROGENEOUS ARCHITECTURES TESTBED

**Exascale Computing**
R&D Topic

**Luca Atzori**
**Maria Girone**
**Krzysztof Michal Mastyna**
**Joaquim Santos**
**David Southwick**
**Eric Wulff**
Technical Team

**Luca Atzori**
**Maria Girone**
Project Coordinator

E4
COMPUTER ENGINEERING

## PROJECT GOAL

The project aims to provide a diverse hardware portfolio for comprehensive technology testing. Focused on assessing the efficacy of various architectures, this initiative aims to provide valuable insights into the practical utility of emerging technologies. By subjecting a spectrum of hardware configurations to real-world applications, the project seeks to establish benchmarks that guide the adoption of the most effective and efficient technologies.

## BACKGROUND

The project is pivotal for CERN and broader scientific communities. By creating a diverse portfolio of applications tailored for various hardware architectures, the project aims to enhance technology evaluation. This work is crucial for optimizing computational efficiency, fostering innovation, and ensuring that CERN and the wider scientific community stay at the forefront of technological advancements, ultimately advancing our capabilities in high-performance computing and scientific research.

# PROGRESS

The project has achieved significant milestones in evaluating cutting-edge technologies across diverse CPU platforms, including Intel® Xeon®Max, AMD EPYC™, Ampere® Altra®, and NVIDIA Grace™ as well as GPUs such as Intel®Flex and NVIDIA H100. The project extensively ran benchmarks with a focus on energy consumption measurements, providing a comprehensive assessment of efficiency.

Notable example applications and benchmarks such as HEPScore23 and MadGraph were thoroughly tested. Additionally, the testbed allowed extensive software testing for the CMS, ATLAS, and LHCb experiments. The support continuously given to the user community emerged as a crucial aspect, possibly the most important, fostering collaboration, offering valuable feedback, and ensuring seamless integration of new hardware.

This progress played a pivotal role in fostering collaboration with industry partners, offering valuable feedback on technology performance and shaping future developments. The project's outcomes are integral not only for advancing CERN experiments but also for guiding industry stakeholders in optimizing their technologies for real-world applications. The collaborative and open approach ensures seamless integration of new hardware, enhancing computational capabilities and fostering groundbreaking advancements in scientific research.

# NEXT STEPS

The project's next steps involve ongoing evaluation of emerging technologies, crucial for readiness in the High Luminosity LHC era. Continuous assessment of advancements in x86 and ARM CPUs, as well as GPU platforms, remains a priority. This forward-looking strategy ensures that the "Heterogeneous Architectures Testbed" stays at the forefront of technological innovation to meet the computational demands of the evolving scientific landscape.

↓From left to right: Eric Wulff, David Southwick, Joaquim Santos and Maria Girone.

EXASCALE COMPUTING

# MADGRAPH5


**Exascale Computing**
R&D Topic

**Stefan Roiser**
Project Coordinator

**Jorgen Teig**
Technical Team

**Igor Vorobtsov**
Intel Collaboration Liaisons


## PROJECT GOAL

GPUs have become ubiquitous in scientific data processing and provide e.g. being a vast part of the computing power of modern High Performance Computing Centers (HPCs). The goal of the Madgraph5 project with Intel was to leverage the computing power provided by hardware accelerators through porting the software onto GPUs and other devices using the Intel SYCL/ oneAPI portability tool and to compare its performance against native programming APIs such as Cuda or HIP.

## BACKGROUND

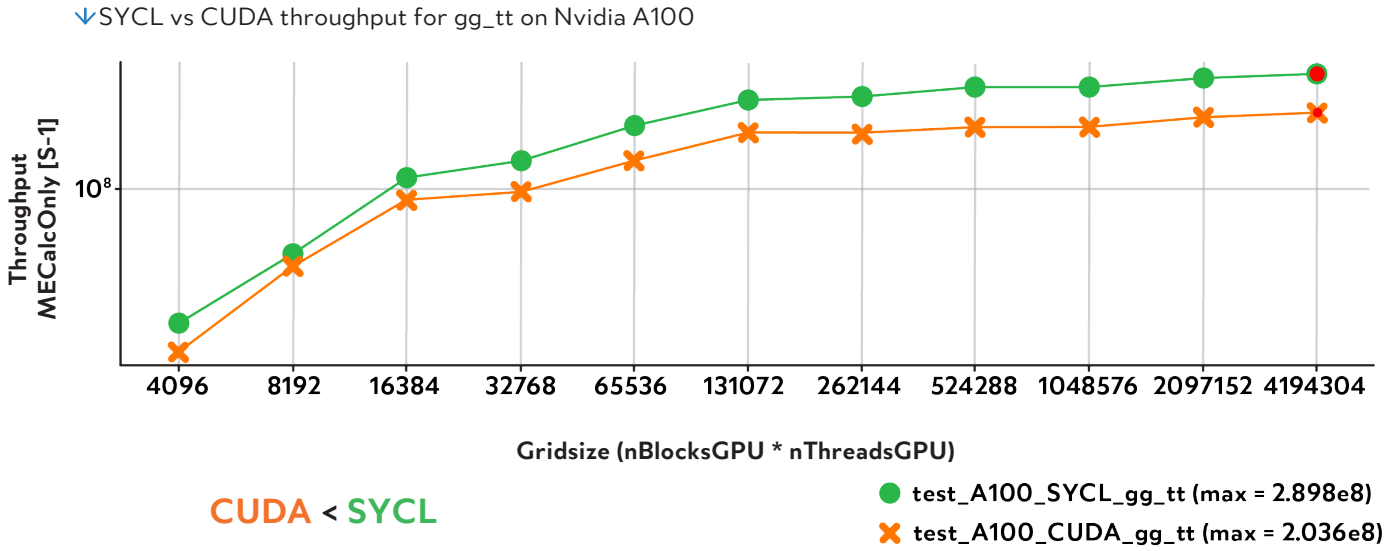
The Madgraph5_aMC@NLO event generator software package is used for the simulation of particle collisions, e.g. in the context of high energy physics experiments at the Large Hadron Collider (LHC) at CERN. With the forthcoming upgrade of the LHC (HL-LHC) the forecasted recorded data volume is expected to grow by one order of magnitude which also implies a major increase in the need of simulated data. To cope with these increased needs for simulated data the software packages need to be improved in terms of performance via the offloading of the computations to hardware accelerator devices such as GPUs.

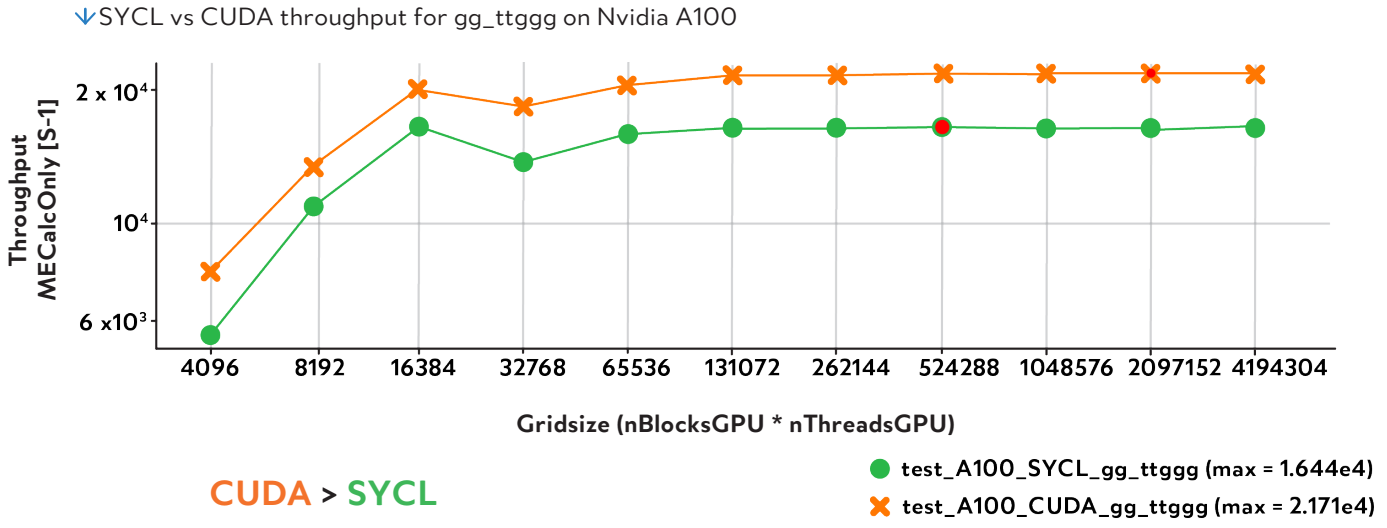
# PROGRESS

Madgraph5 is a versatile code generating tool which takes as input the physics processes to simulate and outputs source code (e.g. Fortran, C++, Python) to perform those simulations. Depending on the complexity of the underlying physics process the output may vary greatly in size and complexity. To cope with the forecasted needs for simulation at the HL-LHC, the compute intensive parts of the Madgraph5 software were re-engineered to output the Madgraph source code for parallel execution on GPUs via Intel SYCL/oneAPI as well as other native APIs. The porting of the code to SYCL started from the initial Cuda implementation in collaboration with University Catholique de Louvain and was first developed at CERN and later continued at Argonne National Labs. Together with the developments also a framework for automatic building and performance comparison of the software was put in place. The comparison was executed across over different physics processes with varying complexity and computing needs and across the different implementations of oneAPI/SYCL versus native GPU APIs and CPU implementations. The results of the performance comparison showed a better performance for oneAPI/SYCL over Cuda for simpler physics processes while Cuda outperforming oneAPI/SYCL for more complex processes.

# ACHIEVEMENTS

The project ended in January 2023 with the published comparison results. The SYCL code will be integrated in the future within the main project together with native APIs.



EXASCALE COMPUTING

# EVALUATION OF POWER CPU ARCHITECTURE FOR DEEP LEARNING

**Exascale Computing**
R&D Topic

**Marco Rossi**
Technical Team

**Maria Girone**
**Sofia Vallecorsa**
Project Coordinators

**Eric Aquaronne**
**Oliver Bethmann**
IBM Collaboration Liaisons

## PROJECT GOAL

We are investigating the performance of distributed training and inference of different deep-learning models on a cluster consisting of IBM Power8 CPUs (with NVIDIA V100 GPUs) installed at CERN. A series of deep neural networks is being developed to reproduce the initial steps in the data-processing chain of the DUNE experiment. In order to do so we have investigated how to adapt computer vision techniques to detector data analysis. More specifically, a combination of convolutional neural networks and graph neural networks are being designed for various tasks in the data processing chain of neutrino experiments: reducing noise, selecting specific portions of the data to focus on during the reconstruction step (region selector) and clustering the detector output into particle trajectories.

## BACKGROUND

Neutrinos are elusive particles: they have a very low probability of interacting with other matter. In order to maximise the likelihood of detection, neutrino detectors are built as large, sensitive volumes. Such detectors produce very large data sets. Although large in size, these data sets are usually very sparse, meaning dedicated techniques are needed to process them efficiently. Deep-learning methods are being investigated by the community with great success.

# PROGRESS & ACHIEVEMENTS

We have developed a series of deep neural network architectures based on a combination of two-dimensional convolutional layers and graphs, including a model inspired by the popular U-Net architecture. These networks can analyse both real and simulated data from protoDUNE and perform region selection, de-noising and clustering tasks, which are usually applied to the raw detector data before any other processing is run.

All of these methods improve on the classical approaches currently integrated in the experiment software stack and they are implemented using a platform agnostic format, ONNX Runtime, which allows the model to run on different hardware. In particular, in order to reduce training time and set up hyper-parameter scans, the training process for the networks is parallelised and has been benchmarked on the IBM Minsky cluster.

In accordance with the concept of data-parallel distributed learning, we trained our models on a total of twelve GPUs, distributed over the three nodes that comprise the test Power cluster. Each GPU ingests a unique part of the physics dataset for training the model.

This project has come to a conclusion in March 2023.

EXASCALE COMPUTING

# COMTRADE

**Exascale Computing**

R&D Topic

**Manuel Reis**
**Elvin Sindrilaru**

Technical Team

**Luca Mascetti**

Project Coordinator

**Ivan Arizanović**
**Branko Blagojević**
**Svetlana Milenković**
**Gregor Molan**
**Anđelina Petrović**

Comtrade Collaboration Liaisons


## PROJECT GOAL

This project is focused on the industry evolution of CERN's EOS large-scale storage system. The goal is to simplify the usage, installation, and maintenance of the system. The project also aims to add native support for windows systems, expand documentation, and implement new features/ integration with other software packages.

## BACKGROUND

Within the CERN IT department, the Storage and Data Management group (IT-SD) is responsible for the operations and development of the main data storage infrastructure. This infrastructure crossed in 2023 the 1 exabyte of raw disk storage and it is used to store the physics data generated by the experiments at CERN, as well as the files of all members of personnel.

EOS is a high-performance disk-based storage system developed at CERN. It is tailored to handle large data rates with low-latency metadata access from the experiments, while also running concurrent complex production workloads.

EOS is also the key storage component behind CERNBox, CERN's cloud-synchronisation service which allows to sync and share files on all major mobile and desktop platforms (Linux, Windows, macOS, Android, iOS), with the aim of providing offline availability to any data stored in the EOS infrastructure.

# PROGRESS

After successfully developing a native Windows client for EOS and a native drive interaction in the Windows environment, in 2023 Comtrade's team invested in optimizing and benchmarking multiple clients and distributed setup to further improve their EOS Windows Native Client (EOSwnc) performance and demonstrate the full capability of this new client.

A comprehensive dedicated full-stack environment, hosted at Comtrade, was deployed. Rigorous testing of various storage solutions was conducted using identical hardware to ensure the utmost accuracy in results. EOS, Ceph, Hadoop Lustre and IBM Spectrum Scale were tested and compared, showing excellent overall results. Measurements were made both with CentOS7 and ALMA8.8.

# NEXT STEPS

The next primary objective is to identify a strategic industry partner to demonstrate the EOS software, along with the additional dedicated resources, including comprehensive documentation and software, developed and produced by Comtrade. This future collaboration will aim to pioneer the creation of an initial prototype for a full-storage appliance solution, with plans to present it to other prospective customers in the near future.

↓The CERN openlab Comtrade team.

# CENTER OF EXCELLENCE ON AI AND SIMULATION-BASED ENGINEERING AT EXASCALE (CoE RAISE)

**Exascale Computing**

R&D Topic

**Maria Girone**
**Andreas Lintermann**

Project Coordinators

**Maria Girone**
**David Southwick**
**Eric Wulff**

Technical Team

**Marcel Aach** (Forschungszentrum Jülich)
**Naveed Akram** (The Cyprus Institute)
**Gabriele Cavallaro** (Forschungszentrum Jülich)
**Kurt de Grave** (Flanders Make)
**Andreas Lintermann** (Forschungszentrum Jülich)
**Arnis Lektauers** (Riga Technical University)
**Morris Riedel** (University of Iceland)
**Nikos Savva** (The Cyprus Institute)
**Eric Michael Sumner** (University of Iceland)
**Liang Tian** (University of Iceland)
**Eric Verschuur** (Delft University of Technology)

Collaboration Liaisons

## PROJECT GOAL

CERN leads Work Package 4 (WP4) which aims at the development and expansion of AI methods along representative use-cases from research and industry, which have a strong focus on data-driven technologies, i.e., analysing data-rich descriptions of physical phenomena. The outcomes are applicable to intelligent workflows including innovative AI methods and techniques, optimized on HPC-to-Exascale systems. The tasks contain the capabilities to evaluate prototype algorithms based on experimental and/or simulation data, code performance on Exascale HPC systems, and quality of data models.

## BACKGROUND

WP4 contains four tasks: Event reconstruction and classification at the CERN HL-LHC, led by CERN; Seismic imaging with remote sensing for energy applications, led by CYI; Defect-free additive manufacturing, led by FM; Sound Engineering, led by UOI.

The task led by CERN consists in developing a GPU native and AI-based algorithm for particle-flow reconstruction that can easily be accelerated by modern heterogeneous hardware. This algorithm, called Machine-Learned Particle-Flow (MLPF), is developed in collaboration with CMS and acts as a representative AI use case from HEP. Some of the most important contributions from this task include the implementation and execution of distributed training and large-scale hyperparameter optimization using HPC, significantly improving physics performance. Another area of work has been to optimize developed algorithms on various heterogenous architectures.

# PROGRESS

Significant progress has been made in terms of MLPF physics performance. One large contributor to this improvement has been the generation of new and larger datasets with a new ground truth definition. The use of large-scale distributed hyperparameter optimization has continued from previous years. Furthermore, the use of model performance prediction using Support Vector Regression (SVR) and Quantum-SVR has been implemented and applied successfully to the problem of tuning the hyperparameters of MLPF.

In 2023, the focus of the MLPF effort shifted from working on closed CMS data to an open dataset that we generated ourselves. The dataset consists of electron-positron collision events at a center of mass energy of 380GeV with full GEANT4 simulation, suitable for detector reconstruction and made publicly available in the EMD4HEP format. Using this dataset, a comparison between a graph neural network and a kernel-based transformer was carried out, demonstrating that both avoid quadratic memory allocation and computational cost while achieving realistic reconstruction. Furthermore, it was shown that hyperparameter tuning on HPC significantly enhanced the physics performance of the models, improving the jet transverse momentum resolution by up to 50% compared to the baseline hand-written algorithm. In addition, the resulting model is highly portable across a variety of hardware accelerators.

# NEXT STEPS

In T4.1, the MLPF studies on the open electron-positron collision dataset has been completed and focus will shift back to simulated CMS-based datasets with proton-proton collisions. A strategic decision has been made to migrate the optimization code of MLPF from TensorFlow to PyTorch. The reason for this is the superior support for cutting edge ML algorithms offered by PyTorch as well as its suitability for easy and fast development of new algorithms. The work has already started in late 2023 but will continue in 2024.

Next steps in the data challenge line of work is to finalize 200G connectivity testing with FZJ and then to carry out XrootD/Rucio dataset testing this spring. RTU tests with GÉANT will investigate UnicoreFTP transfer service in this same time period.

↓The CERN openlab CoE RAISE members at the All-Hands Meeting at CERN in January, 2023.

# **INTERTWIN:** AN INTERDISCIPLINARY DIGITAL TWIN ENGINE FOR SCIENCE

**Exascale Computing**

R&D Topic

**Matteo Bunino**
**Xavier Espinal**
**Enrique Garcia**
**Maria Girone**
**Kalliopi Tsolaki**
**Sofia Vallecorsa**
**Alexander Zoechbauer**

Technical Team

**EGI Foundation**

Project Coordination

**Isabel Campos** (CSIC)
**Charis Chatzikyriakou** (EODC)
**Levente Farkas** (EGI)
**Diana Gudu** (KIT)
**Andreas Lintermann** (FZJ)
**Paul Millar** (DESY)
**David Rousseau** (IJCLab, CNRS/IN2P3)
**Mario Rüttgers** (FZJ)
**Rakesh Sarma** (FZJ)
**Daniele Spiga** (INFN)

Collaboration Liaisons

interTwin

## PROJECT GOAL

The Interdisciplinary Digital Twin (interTwin) project is an ambitious initiative aimed at revolutionizing digital twin technology. At its core, interTwin seeks to co-design and implement a prototype of an open-source Digital Twin Engine (DTE), built upon open standards, facilitating seamless integration with application-specific Digital Twins (DTs). This innovative platform, rooted in a co-designed interoperability framework and the conceptual model of a DT for research, known as the DTE blueprint architecture, aims to simplify and accelerate the development of complex application-specific DTs. By extending the technical capabilities of the European Open Science Cloud with integrated modelling and simulation tools, interTwin not only fosters trust and reproducibility in science but also showcases the potential of data fusion with advanced modelling and prediction technologies. With a focus on ensuring quality, reliability, and verifiability of DT outputs, while simultaneously simplifying application development through AI workflow management and reinforcement of open science practices, interTwin stands as a pioneering endeavor at the forefront of interdisciplinary innovation.

## BACKGROUND

InterTwin develops and implements an open-source DTE that offers generic and customized software components for modeling and simulation, promoting interdisciplinary collaboration. The DTE blueprint architecture, guided by open standards, aims to create a common approach applicable across scientific disciplines. Use cases span high-energy physics, radio astronomy, climate research, and environmental monitoring. The project leverages expertise from European research infrastructures, fostering the validation of technology across facilities and enhancing accessibility. InterTwin aligns with initiatives like Destination Earth, EOSC, EuroGEO, and EU data spaces for continuous development and collaboration.

## PROGRESS

We've pinpointed the components of the CERN digital twin (DT) application. The first one uses the Monte Carlo (MC) based simulation framework called GEANT4. The second component is the deep learning component, known as the 3D Generative Adversarial Network (3DGAN). This component is designed to simulate particle interactions tailored to specific particle detector setups. We're initially concentrating on calorimeters use case, which are types of detectors that require the most computing power for simulations.

We've established the requirements for this use case, including thematic modules, and have explored more sophisticated generative models. We've also incorporated the latest 3DGAN component into our AI workflow tool. The capabilities of the DT workflow include generating training data, preprocessing this data before it's input into the machine learning model, storing both input and output data, enabling distributed training across multiple GPUs, conducting model inference, performing validation and quality checks, and facilitating continuous re-training to fine-tune the simulations.

Requirements of all use-cases about their specific AI/ML setup (model, data, infrastructure) have been collected. An analysis of those requirements led to the DTE blueprint architecture. A first version of a prototype was developed that includes basic machine learning functionalities like training, saving in a model registry using MLFlow, and inference. A toy preprocessing module has been created to achieve this. The workflow execution was tested in two ways: Using a python environment and using the Common Workflow Language. The prototype has been successfully tested locally, on CERN compute resources, and on the Julich HDF-ML cluster.

Finally, several use cases both from the Earth Observation and Physics domain have been successfully integrated: MNIST toy use case, which serves as an example other use cases can follow, CERN use case, and CMCC use case. The VIRGO use case's integration is currently on hold due to access policies. An MoU is currently being set up to solve this issue.

## NEXT STEPS

The upcoming steps involve exploring integration of our work with the MC based framework, and refining the data transformation processes. This includes integrating 3DGAN model into the MC framework, in case that this activity can be supported by the DTE, developing or incorporating tools for simultaneous training and optimizing hyperparameters (adjusting them as necessary for adversarial training), and selecting solutions that are best suited for the GAN use case, keeping in mind the specific characteristics of the computing hardware, such as accelerators and how they communicate between nodes. We plan to adopt a continuous training method that allows for the model to be updated as soon as new data become available.

Additionally, we're working on creating a customizable validation framework in partnership with experts from the High Energy Physics (HEP) community. This entails developing complex multivariate distributions that consider a wide variety of input conditions, and establishing validation techniques that can evaluate different performance metrics, such as accuracy and comparison to classical simulation methods (i.e., uncertainty estimation, coverage of the support space). We will be advancing the development of our DT's thematic modules, ensuring our software components work well with and meet the standards of DTE solutions created in other work packages. This involves integrating and rigorously testing our components for compatibility and compliance.

Our prototype was packaged in a Docker container, and an integration test with WP5 using their interLink prototype is currently underway. Furthermore, the prototype is getting extended to provide more advanced machine learning capabilities, such as distributed training.

After first experiments with hyperparameter optimization frameworks have been conducted, a production-ready implementation is planned within the next period. Moreover, additional use cases will be integrated in the next period.

Lastly, an easy-access user interface based on Jupyter notebooks using frameworks like, e.g., KubeFlow is planned.

EXASCALE COMPUTING

# **EMP2:** ENVIRONMENTAL MODELLING AND PREDICTION PLATFORM

**Exascale Computing**
R&D Topic

**Alberto Di Meglio**
**Ilaria Luise**
Technical Team

**Alberto Di Meglio**
Project Coordinator

**Christian Lessig** (ECMWF, Magdeburg University)
**Martin Schultz** (Juelich Supercomputing Center)
Collaboration Liaisons

## PROJECT GOAL

The proposed project aims to develop a proof-of-concept for a machine learning based digital twin of the atmosphere for environmental applications. To accomplish this, the project is subdivided into two main parts. The first segment will focus on the development of a machine learning based modelling core prototype, called AtmoRep, built on the concept of large scale representation learning applied to Earth System Science. In the second phase, the modelling core will be integrated into the digital twin architecture currently under development within the CERN IT department by the InterTwin project.

## BACKGROUND

The atmosphere and its dynamics have a significant impact on human well-being, from agricultural decision making, to policy making and the renewable energy sector. An accurate and equitable modeling of atmospheric dynamics is consequently of critical importance to allow for evidence-based decision making that improves human well being and minimizes adverse impacts for current and future generations. Very recently, AI-based models have shown tremendous potential in reducing the computational costs for numerical weather prediction. However, they lack the versatility of conventional models. The EMP2/AtmoRep project aims at developing an AI-based model of atmospheric dynamics for multi-purpose applications. The model will be implemented leveraging the concept of large-scale representation learning, so to encapsulate the information from the large amounts of available data. The implementation on the digital twin platform will make such information more accessible to the general public, allowing the users to easily develop their own applications in weather and climate.

# PROGRESS

In September 2023, the team publicly released a first prototype of the core model, which has been tested on multiple tasks in Earth System science, like weather forecasting, downscaling, spatio-temporal interpolation and precipitation rate corrections. The model, also referred to as AtmoRep, consists of a 3.5 billion parameter network, trained for several weeks in summer 2023 at the Juelich Supercomputing Center using 4M core hours in total. The model shows competitive skill in weather forecasting when compared to the newly released AI-based forecasting models and it out-performs the competitors for tasks such as downscaling or precipitation rate forecasting. One of the main innovations consists in a novel probabilistic loss, so the model outputs probabilistic ensembles for each downstream task.

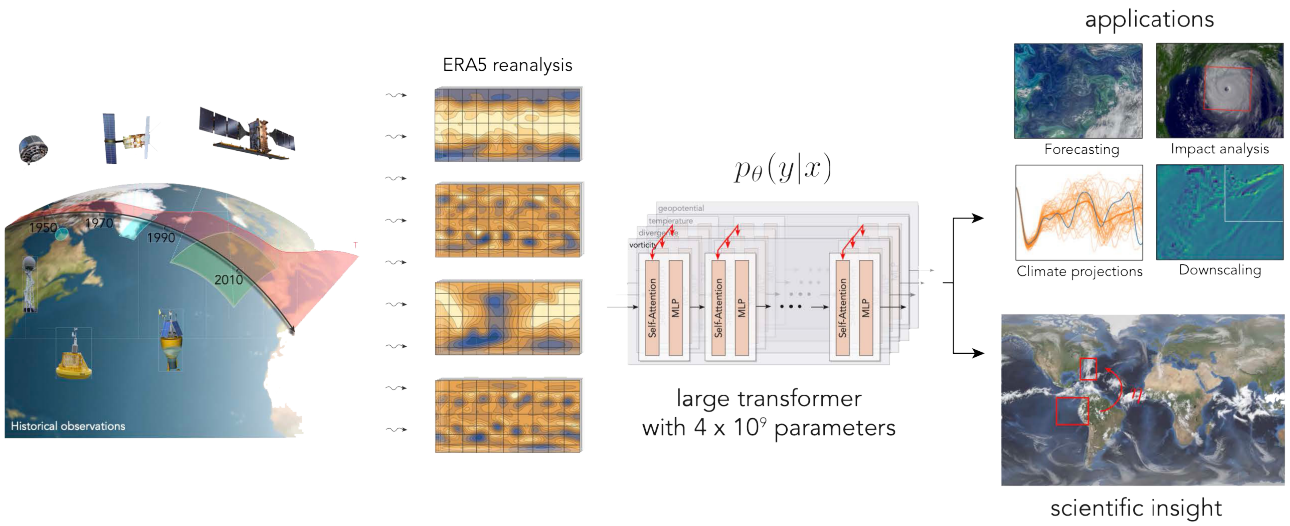
The CERN team was mainly responsible of implementing the analysis workflow of the downstream applications. This converged in the development of a dedicated analysis package together with collaborators at the Juelich Supercomputing Center. All the training code has been recently released on GitHub and made available to the public.

# NEXT STEPS

The next steps related to the modelling core development include the implementation of an autoregressive roll out mechanism to reach medium range weather forecasts (10-15 days) and the extension to multi-resolution dataset handling to go beyond the quarter degree spatial resolution forecasts available with the current setup.

The second part of the project will involve the implementation of the core model on the InterTwin digital twin architecture, in collaboration with InterTwin and the other members of the Digital Twin initiative at CERN.

Workflow of the AtmoRep core model, from the near to observation reanalysis data (ERA5), to the data-driven model based on large scale machine learning. The model is then used as backbone for multiple applications of interest in Earth System science.



# ARTIFICIAL INTELLIGENCE PROJECTS

# AI MODELS REGISTRY IN THE CLOUD

**Artificial Intelligence**
R&D Topic

**Renato Cardoso**
**Sofia Vallecorsa**
Technical Team

**Sofia Vallecorsa**
Project Coordinator

**Şengül Chardonnereau**
**Jérôme Designe**
**Allen Hosler**
**Sébastien Hurel**
**Cristobal Pedregal-Martin**
**Lyudmil Pelov**
**Bob Peulen**
**Garret Swart**
Oracle Collaboration Liaisons

## PROJECT GOAL

CERN is running a pilot project to demonstrate the possibility to optimize the ML training task for benchmark ML tasks of different sizes, complexity, and data. In this context, this project will focus on demonstrating the use of a ML model catalogue, evaluate the ML energy footprint of ML training and test the possibility of using large models (foundation models).

## BACKGROUND

With the increase of the utilization and complexity of ML algorithms at CERN it is necessary to investigate performance tracking and cost optimization, furthermore we need to ensure the models are generalizable and reusable. A ML catalogue aims to be a centralized place to store models, and able to be used for performance tracking, model sharing and reuse. On the other hand, from a data science perspective, the AI state-of-the-art trend is to use foundational model as a way for AI model generalization. The aim of these approaches will lead to a future 'sustainable' AI by improving the ML training and deployment efficiency.

## PROGRESS

For a thorough evaluation of the ML catalog provided by Oracle and its attached Accelerated Data Science (ADS) SDK, 2 use cases previously developed at openlab were used. Both consist of deep generative models, one for HEP and another for Earth Observation, with different levels of scale (number of network parameters and time to train). Apart from testing the platform and the usage of the model catalog, a study of energy consumption was conducted alongside to understand the environmental impact of training a machine learning model. More specifically, we tested the 2 generative models and logged the energy consumption during training, for multiple hardware options and multiple training and hardware optimizations. The second half of the year was dedicated to the exploration of a foundation model, using a diffusion model method for generation of calorimeter showers, alongside, and following state-of-the-art trends, we use transformers for better generalization. Currently we can generate, with high fidelity, showers like the ones produced by Geant4.

## NEXT STEPS

Our work in 2024 will be devoted to extending the generation approach, towards a foundation model, by being able to do more tasks than what it was trained to do.

↓Part of the AI Models Registry in the Cloud team.

ARTIFICIAL INTELLIGENCE

# DATA ANALYTICS FOR INDUSTRIAL CONTROL SYSTEMS

**Artificial Intelligence**

R&D Topic

**Jan Andrzej Bugajski**
**Abhit Patil**
**Fernando Varela Rodriguez**
**Jeronimo Ortola Vidal**

Technical Team

**Fernando Varela Rodriguez**

Project Coordinator

**Thomas Kaufmann**
**Christian Kern**
**Daniel Schall**
**Axel Sundermann**

Siemens AG Collaboration Liaisons

SIEMENS

## PROJECT GOAL

The project aims to enhance the efficiency and intelligence of the industrial control systems utilized by the CERN's accelerator complex. One of the main goals is to create a device monitoring platform, incorporating a web application prototype that uses edge computing technologies and real-time analytics to monitor control devices. Another goal is to test and deploy advanced control algorithms on an industrial edge device, aiming to boost the energy efficiency of control processes. As part of this initiative, Siemens' solutions will be assessed and fine-tuned to meet the particular requirements of the end-users at CERN.

## BACKGROUND

The planned HL-LHC upgrade is set to increase the particle collision data sample by tenfold relative to the existing LHC program. The associated control systems will grow in complexity with these enhancements. As a result, enhancing the current systems' functionality and sustainability becomes critical. This task encompasses several interconnected initiatives. One key subproject is devoted to developing a device monitoring platform that will oversee the hardware components of the industrial control system, employing edge computing technologies and real-time analytics. Additionally, another critical subproject aims to test and deploy sophisticated control algorithms on an industrial edge device to augment and optimize the control devices within the plant.

# PROGRESS

In 2023, several key milestones were achieved within the project. A working prototype of the web application was co-created with members from CERN and Siemens. The frontend component of the application enables end-users to configure and categorize control devices into a hierarchical tree-like structure and design rules for the individual categories or tree nodes. These rules are executed in real time, serving a crucial function: they hierarchically display the status of the entire control system. This real-time display facilitates easy navigation and efficient identification of errors within the system. A backend component of the application, based on finite state machines, was developed for rule execution. The application was designed to support third-party integrations, including compatibility with software for device monitoring developed by Siemens. Moreover, several Siemens monitoring software programs such as SINEC NMS, Machine Insight, and SIMATIC Automation Tool were evaluated for potential integrations using a test bench set up in the control systems laboratory at CERN.

In another subproject, we utilized industrial edge computing technology to operate an advanced control algorithm based on Model Predictive Control. This setup allows the algorithm to run directly on the edge computing device within a containerized environment, simplifying the algorithm development process and reducing the need for extensive control system components. This setup significantly lowers latency and decreases the load on the control system, consequently enhancing overall system efficiency.

# NEXT STEPS

The device monitoring project is set to continue in the upcoming years. Nonetheless, specific subprojects may transition to different focus areas based on the priorities established in collaboration with the company. These new areas include the application of large language models in the development of control software or the implementation of predictive maintenance for devices using historical data. A joint workshop between CERN and Siemens has been scheduled to outline specific use cases for the forthcoming year and the objectives for the eighth phase of CERN openlab.

↓Team members from CERN and Siemens captured at the CERN Control Centre during the CERN openlab Technical Workshop in 2023.

ARTIFICIAL INTELLIGENCE

# FAST DETECTOR SIMULATION

**Artificial Intelligence**

R&D Topic

**Emma Call** (Intel)
**Adel Chaibi** (Intel)
**Valeriu Codreanu** (SURF)
**Soumyadip Ghosh**(Intel)
**Kristina Jaruskova**
**Duncan Kampert** (SURF)
**Cai Maxwell** (SURF)
**Hans Pabst** (Intel)
**Damian Podarean** (SURF)
**Vikram A. Saletore** (Intel)
**Kalliopi Tsolaki**
**Sofia Vallecorsa**

Technical Team

**Sofia Vallecorsa**

Project Coordinator

**Vikram A. Saletore**

Intel Collaboration Liaisons

## PROJECT GOAL

The objective of this project is to optimize the training and inference of fast simulation models as well as to test the capabilities of the new hardware and software tools developed by Intel. This can help CERN better evaluate the capabilities and benefits of the generative deep learning models. The work is a joint effort of CERN openlab, Intel (CA, USA), and SURF (NL).

## BACKGROUND

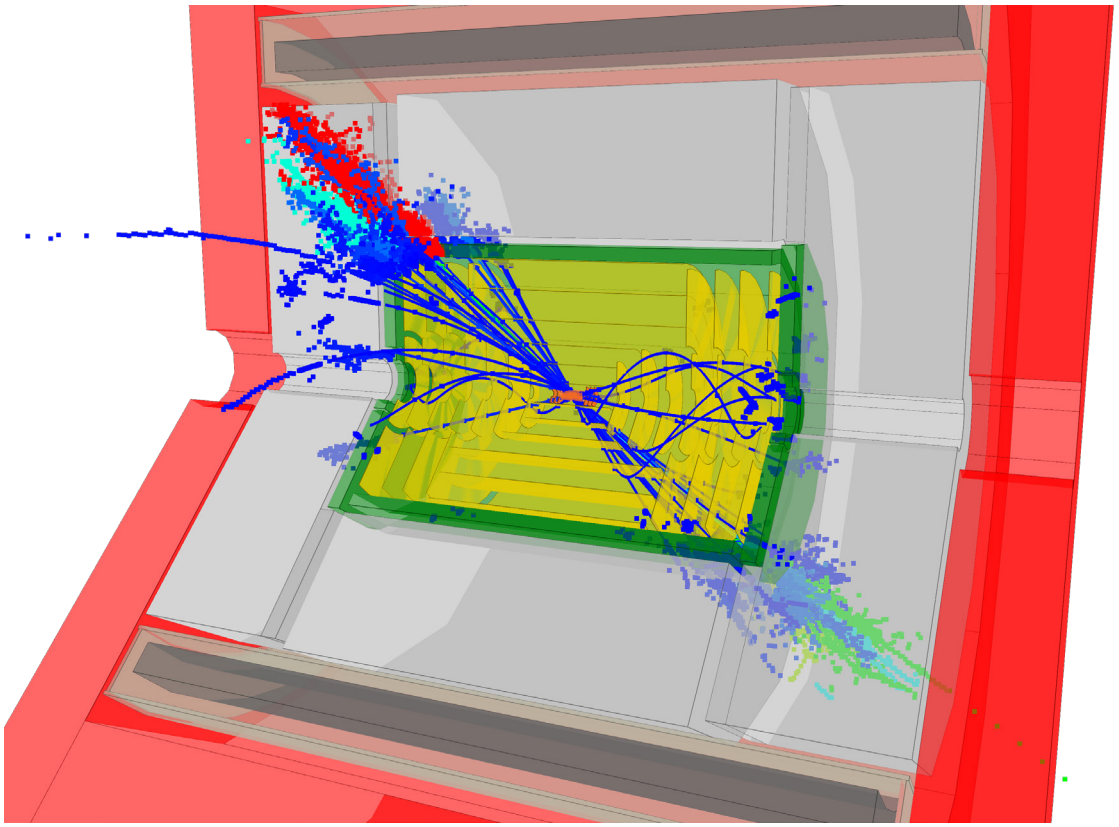
Detector simulations are an essential part of the HEP research. However, the current simulation techniques based on the Monte Carlo sampling are very time consuming. The number of simulations scales up with each upgrade of the LHC, unlike the available computing resources. The deep learning algorithms can provide detector simulations in a fraction of time and thus help overcome this bottleneck. Yet, the efficiency of the deep learning approach is significantly dependent on the hardware and software being used.

# PROGRESS & ACHIEVEMENTS

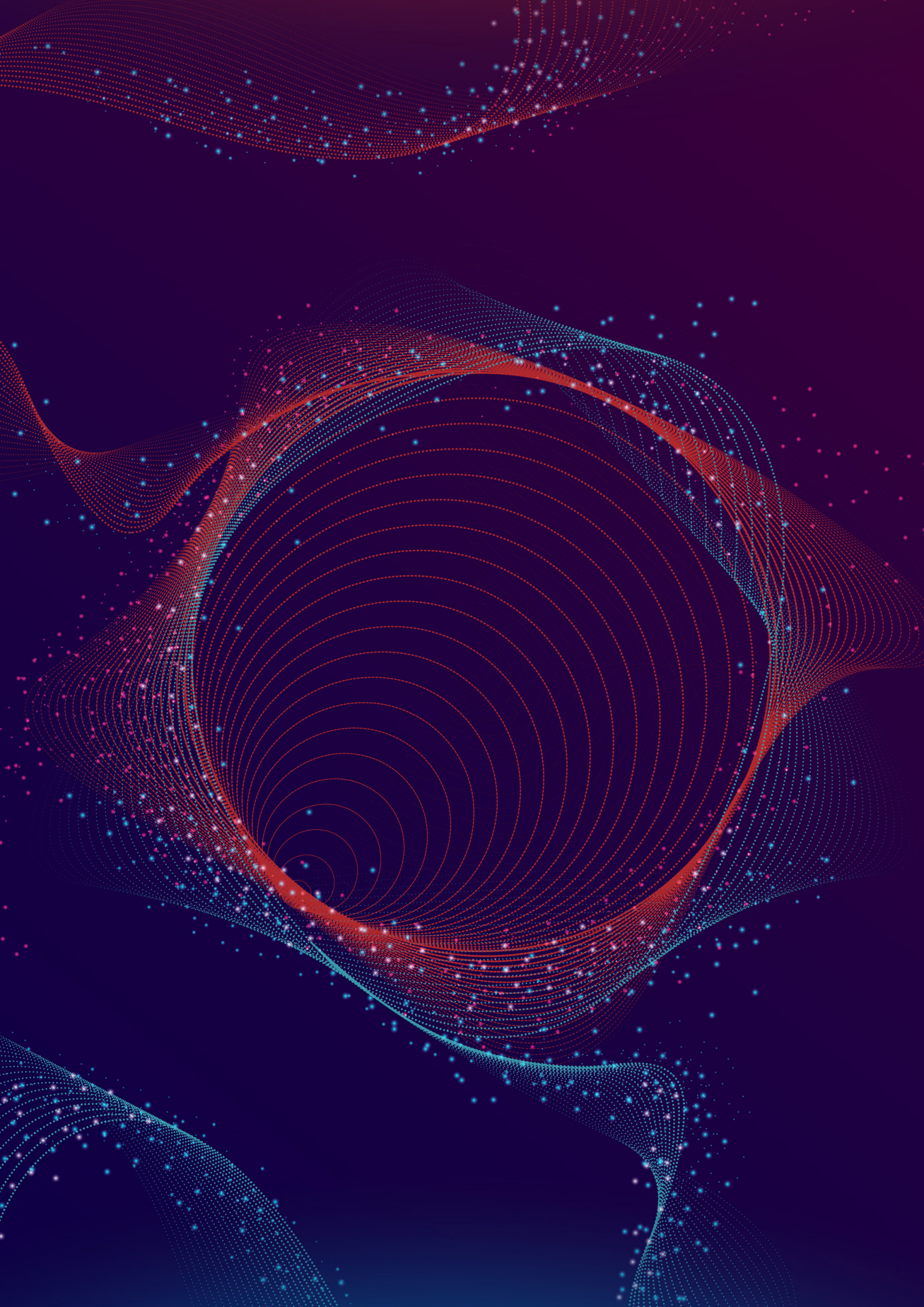
To continue the work started by the end of 2022, the focus in 2023 was mainly on optimizing the inference of the 3DGAN model for calorimeter simulations. We examined the quantization of the 3DGAN generator to INT8 precision while keeping the FP32 accuracy with the Intel® Neural Compressor tool. It enabled an automated partial quantization of the model (3 layers out of 7 converted to INT8) without any increase in the loss value. The inference performance of both version of the model (FP32 and partially quantized) was tested using the Intel® Xeon® Max Series CPUs. Using the Intel Neural Compressor, we obtained a 1.9x speedup on the inference. This results in a speedup of more than 8500x compared to the Geant4 simulations.

Later in 2023, the focus was on examining the possibility to integrate the 3DGAN model into a Geant4-based application simulating the entire detector. Conducting this study is relevant as it will provide the community with more detailed evaluation of gains from using the deep learning generators for detector simulation in a more realistic setup where we consider the exchange of information between the Geant4 simulator and the 3DGAN generator.

↓Fast simulation of the detector response to particles is an essential next step satisfy the increasing demands for simulations.



# QUANTUM COMPUTING PROJECTS



QUANTUM COMPUTING

# QUANTUM DATABASES FOR DYNAMIC DATA STORAGE

**Quantum Computing**
R&D Topic

**Michele Grossi**
**Carla Rieger**
**Sofia Vallecorsa**
**Martin Werner** (TUM)
Technical Team

**Sofia Vallecorsa**
Project Coordinator

**Gian Giacomo Guerreschi**
Intel Collaboration Liaisons

## PROJECT GOAL

This collaboration of CERN and Intel is exploring the usage of a quantum database for storage of classical and quantum data which may be initially of unknown length. Thus, we are aiming to create an algorithmic procedure for an efficient dynamic data storage.

## BACKGROUND

Quantum data itself is fragile, subject to decoherence and collapses when being measured. Due to the collapse of the quantum state upon quantum measurement, for future applications it will be needed to store the quantum state itself in form of a quantum memory for further data processing based on quantum algorithms. The idea of a quantum database is to store the individual quantum states containing the quantum or classical data itself in a superposition correlated with a label state which is used for indexing. In our project, we are mainly concerned with data obtained through experiments over an unknown temporal interval. Thus, such experimental data can be inherently of non-predefined length, e.g., the runtime of the experiment is not given initially. Hence, we are mainly concerned about its resource-efficient storage method and the specific data manipulation operations that are applied within a dynamical database.

## PROGRESS

In the year 2023, for this project we were focusing on the extension of the development of a theoretical model of the quantum database and algorithmic solutions for data manipulation with respect to different use-cases that go beyond high-energy physics. Within a quantum database the quantum state itself is stored instead of the classical data obtained though measurement and hence a quantum state collapse is avoided. Future quantum computing devices working with state-of-the-art quantum algorithms are supposed to make use of a quantum memory and would directly be able to process the stored quantum data. Hence, this technique of storage is also useful as a resource-efficient and compact quantum state preparation technique to obtain initial quantum states of, e.g., quantum machine learning models. Altogether, we are aiming to store as much information in our quantum state resulting from a temporally dynamical experimental setting and provide a set of manipulation techniques to operate within the database.

## NEXT STEPS

Current theoretical results are very promising. The goal is to provide a complete algorithmic toolkit of database manipulation operations of the dynamic quantum database which we have started to develop. Furthermore, we are going to explore specific use-cases and further possible applications of the quantum database model. For the future, we aim to add more operations and refine our solution.

# BEYOND PARTICLE PHYSICS PROJECTS

BEYOND PARTICLE PHYSICS

# DIGITAL TWIN AND SYNTHETIC DATA IN HEALTHCARE

**Beyond Particle Physics**
R&D Topic

**Nicola Serra**
Project Coordinator

**Olivia Jullian Parra**
**Shrija Rajen Sheth**
**Sara Zoccheddu**
Technical Team

**Professor Milo Puham** (University of Zurich)
**Dr. Henock Yebyo** (University of Zurich)
Collaboration Liaisons

## PROJECT GOAL

The project's objective is to produce synthetic data mirroring specific characteristics of patients. This data will facilitate the calculation of outcome probabilities and counterfactual scenarios for different treatment options. Key anticipated outcomes include both potential side effects and prevention effects. The project places particular emphasis on patients with multimorbidity, aiming to enhance understanding and management of their complex health situations.

## BACKGROUND

Multimorbidity, characterized by the simultaneous occurrence of multiple chronic conditions in a single patient, poses a formidable challenge in modern healthcare, significantly impacting both treatment efficacy and healthcare costs. Traditional research methods, often reductionist in nature, inadequately address the complexities of treating such patients, as they typically focus on individual diseases. There is a critical need to adopt a more integrative approach, taking into account the interactions between various conditions and their treatments. We plan to utilize extensive observational data through advanced Machine Learning techniques, especially generative models and reinforcement learning. The same architecture we are studying is aimed at improving the data processing and analysis workflow at the LHCb experiment.

## PROGRESS

We have applied the Causal Inference framework to model the interactions between patients, treatments, and outcomes, which is essential for avoiding biases due to confounders. Our focus is on two types of data: synthetic, enabling precise mapping of relationships, and real data of varying quality. High-quality clinical trial data often have limited statistics, while observational data, though statistically rich, are noisier and less reliable.

Our strategy involves progressively incorporating noisier data. Generating epidemiologically inspired synthetic data, we've employed various architectures for analysis, with Conditional Variational Autoencoders being particularly noteworthy. This approach has been extended to real clinical trial data for HIV patients, assessing the benefits and risks of statin use, showing promising alignment with traditional methods.

Next, we plan to integrate reinforcement learning for causal graph modification and discovery. Additionally, we're adapting these methods for data quality monitoring within the LHCb collaboration. Automating these tasks can significantly enhance data collection efficiency and reduce the need for manual labor. This becomes challenging during new detector commissioning, where algorithms require continuous retraining. We propose applying Reinforcement Learning with human feedback for efficient and effective anomaly detection, balancing data-collection efficiency with human factors.

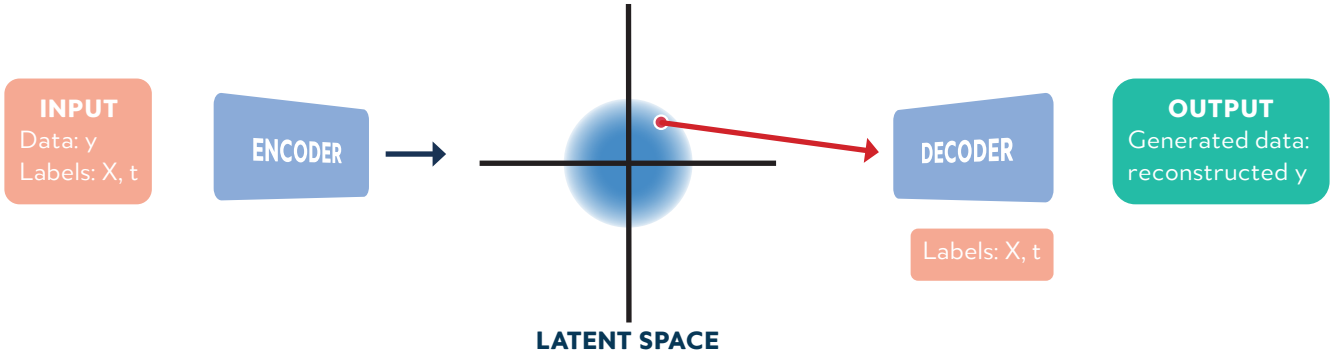
## NEXT STEPS

Moving forward, our goal is to expand the causal graph's complexity by incorporating a broader range of treatments and outcomes (both benefits and harms). We'll assess the efficacy of generative models for different causal graphs, adjusting their structures through reinforcement learning. This evaluation will initially employ synthetic data, with prospects of applying it to real data.

Simultaneously, we'll deepen our studies on the LHCb monitoring assistant, previously tested with synthetic data. The next step involves comparing its performance with data taken in Run1 and Run2.

↓Olivia Jullian Parra, member of the technical team.

↓Conditional Variational Autoencoder, one of the architectures studied to generate synthetic data of patients.



# BIODYNAMO

**Beyond Particle Physics**
R&D Topic

**Lukas Breitwieser**
Project Coordinator

**Lukas Breitwieser**
**Tobias Duswald**
Technical Team

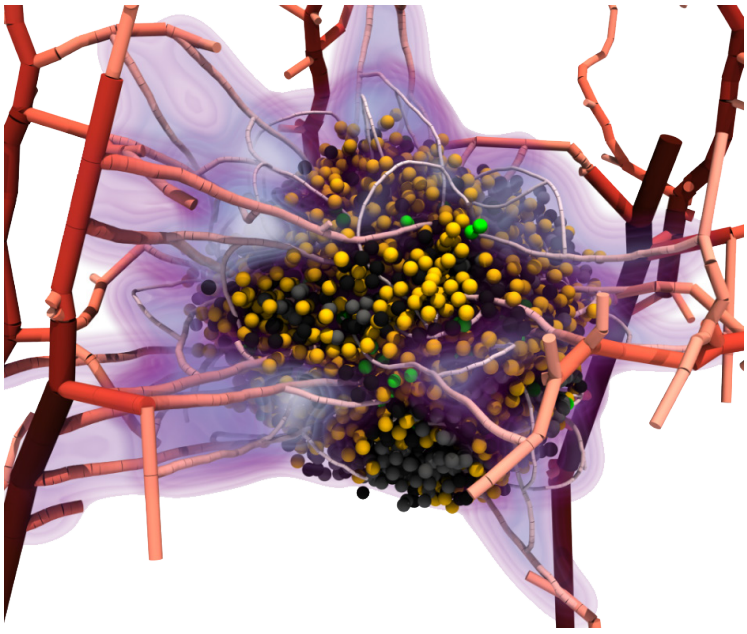
**Roman Bauer** (University of Surrey)
**Vasileios Vavourakis** (University of Cyprus)
Collaboration Liaisons

## PROJECT GOAL

The project's goal is to develop a platform that empowers scientists to effortlessly generate, execute, and visualize agent-based simulations. Utilizing cutting-edge computing technologies, the BioDynaMo platform will facilitate simulations of unprecedented scale and complexity. This capability opens avenues for addressing intricate scientific research inquiries with greater ease.

## BACKGROUND

In the life sciences community, computer simulation is gaining prominence for modeling intricate biological systems. While numerous specialized tools exist, creating a high-performance, versatile platform represents a significant advancement. CERN leverages its extensive expertise in large-scale computing, supported by funding through the Gentner stipend, ETH Zurich, and the CERN budget for knowledge transfer to medical applications, to collaborate on this unique platform development. The project aims to provide a comprehensive solution for simulating diverse biological scenarios.



→ Simulation of a vascular tumor growth under treatment. Credit: Tobias Duswald et al., https://doi.org/10.1016/j.cma.2023.116566. Used under CC BY 4.0 DEED

# PROGRESS

At the beginning of 2023, our work on performance optimizations was published at the PPoPP conference, the premier forum for leading work on all aspects of parallel programming. We show that with our optimizations BioDynaMo is more than 1000x faster than Cortx3D and NetLogo, while also achieving an impressive 9x greater efficiency than Biocellion (https://doi.org/10.1145/3572848.3577480).These enhancements empower our users to simulate larger models, explore model parameters, and accelerate model development. We are deeply honored to receive the Best Artifact Award, recognizing our work's exceptional quality and impact.

Further, our team published research in Computer Methods in Applied Mechanics and Engineering (CMAME) focusing on the complexities of cancer progression. We introduced a computational model simulating vascular tumor growth and responses to drug treatments in a 3D context. The project's goal is to develop a platform that empowers scientists to effortlessly generate, execute, and visualize agent-based simulations. The study investigated a therapy involving Doxorubicin and Trastuzumab. Published in CMAME, this work demonstrated BioDynaMo's capabilities by simulating vascular tumor growth with a volume of 400 mm3 using a total of 92.4 million agents.

# NEXT STEPS

Our plan is to make significant strides in enhancing the overall usability of our platform, with a dedicated focus on refining user experience and robustness. We will actively extend our documentation to ensure clarity and accessibility. Furthermore, we will continue our commitment to advancing various use cases in biology and beyond, continuously exploring innovative solutions and contributing to the broader scientific community.

↓Best Artifact Award ceremony at PPoPP'23 in Montreal, Canada.

# PUBLICATIONS

During phase VII, CERN openlab publications encompassed work on all the four Reasearch & Development topics. Many of these publications resulted from collaborative efforts among various groups within CERN, CERN openlab and external collaborators. The featured publications below provide an overview of CERN openlab extensive research spectrum.


Campos, I., Moltó, G., Jacob, A., Orviz, P., Caballer, M., Bunino, M., Zoechbauer, A., & Fiore, S. (2023). interTwin D6.2 First release of the DTE core modules (1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.10224213

Isabel Campos, Donatello Elia, Germán Moltó, Ignacio Blanquer, Alexander Zoechbauer, Eric Wulff, Matteo Bunino, Andreas Lintermann, Rakesh Sarma, Pablo Orviz, Alexander Jacob, Sandro Fiore, Miguel Caballer, Bjorn Backeberg, Mariapina Castelli, Levente Farkas, & Andrea Manzi. (2023). interTwin D6.1 Report on requirements and core modules definition (V1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.8036987

Lukas Breitwieser, Ahmad Hesam, Fons Rademakers, Juan Gómez Luna, and Onur Mutlu. 2023. High-Performance and Scalable Agent-Based Simulation with BioDynaMo. In Proceedings of the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP '23). Association for Computing Machinery, New York, NY, USA, 174–188. https://doi.org/10.1145/3572848.3577480

Tobias Duswald, Ernesto A.B.F. Lima, J. Tinsley Oden, Barbara Wohlmuth, Bridging scales: A hybrid model to simulate vascular tumor growth and treatment response, Computer Methods in Applied Mechanics and Engineering, Volume 418, Part B, 2024, 116566, ISSN 0045-7825, https://doi.org/10.1016/j.cma.2023.116566.

R. Cardosoet al., Deploying a machine learning model catalog at CERN, CHEP2023, Norfolk, USA (2023) https://indico.jlab.org/event/459/contributions/11656/

Aach, M. (Corresponding author), Wulff, E., Pasetto, E., Delilbasic, A., Sarma, R., Inanc, E., Girone, M., Riedel, M. & Lintermann, A., "A Hybrid Quantum-Classical Workflow for Hyperparameter Optimization of Neural Networks", ISC High Performance 2023, ISC2023.

Lessig et al., (2023). AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning. arXiv. https://arxiv.org/pdf/ 2308.13280.pdf

Kalliopi Tsolaki, Sofia Vallecorsa, David Rousseau, Isabel Campos, Yurii Pidopryhora, Sara Vallero, Alberto Gennai, & Massimiliano Razzano. (2023). interTwin D7.2 Report on requirements and thematic modules definition for the physics domain first version (V1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.8036997

Andrea Manzi, Levente Farkas, Kalliopi Tsolaki, Sofia Vallecorsa, Sara Vallero, Massimiliano Razzano, Javad Komijani, Yurii Pidopryhora, & Isabel Campos. (2023). interTwin D4.2 First Architecture design of the DTs capabilities for High Energy Physics, Radio astronomy and Gravitational-wave Astrophysics (1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.8321134

Farouk Mokhtar, Joosep Pata, Javier Duarte, Eric Wulff, Maurizio Pierini, Jean-Roch Vlimant "Progress towards an improved particle flow algorithm at CMS with machine learning", J. Phys.: Conf. Ser. (2024) https://arxiv.org/abs/2303.17657

Wulff E., Girone M., Southwick D., García Amboage J.P, Cuba E. "Hyperparameter optimization, quantum-assisted model performance prediction, and benchmarking of AI-based High Energy Physics workloads using HPC", J. Phys.: Conf. Ser. 21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2022) (2024) https://arxiv.org/abs/2303.15053

García Amboage J.P, Wulff E., Girone M., Pena T.F. "Model Performance Prediction for Hyperparameter Optimization of Deep Learning Models Using High Performance Computing and Quantum Annealing", EPJ Web Conf. 26th International Conference on Computing in High Energy & Nucelar Physics (CHEP2023) (2024) https://arxiv.org/abs/2311.17508

Aach, M. (Corresponding author), Wulff, E., Pasetto, E., Delilbasic, A., Sarma, R., Inanc, E., Girone, M., Riedel, M. & Lintermann, A., "A Hybrid Quantum-Classical Workflow for Hyperparameter Optimization of Neural Networks", ISC High Performance 2023, ISC2023.

Lessig et al., (2023). AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning. arXiv. https://arxiv.org/pdf/ 2308.13280.pdf

Kalliopi Tsolaki, Sofia Vallecorsa, David Rousseau, Isabel Campos, Yurii Pidopryhora, Sara Vallero, Alberto Gennai, & Massimiliano Razzano. (2023). interTwin D7.2 Report on requirements and thematic modules definition for the physics domain first version (V1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.8036997

Andrea Manzi, Levente Farkas, Kalliopi Tsolaki, Sofia Vallecorsa, Sara Vallero, Massimiliano Razzano, Javad Komijani, Yurii Pidopryhora, & Isabel Campos. (2023). interTwin D4.2 First Architecture design of the DTs capabilities for High Energy Physics, Radio astronomy and Gravitational-wave Astrophysics (1 Under EC review). Zenodo. https://doi.org/10.5281/zenodo.8321134

A. Patil, D. Daniel, F. Ghawash, K. Kaufmann, C. Kern, B. Schofield, A.S. Sündermann and F. Varela, Leveraging Local Intelligence to Industrial Control Systems through Edge Technologies, Proc. 19th Int. Conf. on Accelerator and Large Experimental Physics Control System (ICALEPCS'23), Cape Town, South Africa, p. 793–798. https://icalepcs2023.vrws.de/posters/tupdp102_poster.pdf

L. Morelli, C. Kern and A. Patil, Monitoring of CERN Industrial Control System using Hierarchical Finite State Machine based approach (16 March). Presented at CERN openlab Technical Workshop, Geneva, 2023.

Valassi et al., (2023). Speeding up Madgraph5 aMC@NLO through CPU vectorization and GPU offloading: towards a first alpha release arXiv. https://arxiv.org/abs/2303.18244

Hageboeck et al., (2023). Madgraph5_aMC@NLO on GPUs and vector CPUs Experience with the first alpha release. arXiv. https://arxiv.org/abs/2312.02898

# PRESENTATIONS

Olivia Jullian Parra, Machine Learning for multimorbidity causal inference (17 March). Presented at the CERN openlab Technical Workshop, Geneva 2023.

Sara Zoccheddu, Conditional Variational Autoencoders in Healthcare (31 August). Presented at the LHCb Summer Student Presentation, Geneva, 2023.

J. Santos, K. Mastyna, Heterogeneous architectures testbed - (16 March). Presented at CERN openlab Technical Workshop, Geneva, 2023.

M. Potocky, High level plan for integration of Oracle cloud resources into CERN IT BC&DR project (16 March), Presented at 2023 CERN Openlab Technical Workshop, Geneva, 2023

M. Potocky, Protecting CERN data from ransomware with Oracle Cloud (November 29), Presented at Oracle Global Leaders Program Customer Meeting – EMEA, Porto, 2023

T. James, Fast ML inference in FPGAs for the Level-1 Scouting system at CMS (25 Sep). Presented at Fastml2023: Fast Machine Learning for Science, London, 2023. URL: https://indico.cern.ch/event/1283970/contributions/5554347/attachments/2720900/4731423/fastml-tjames-2023-v2.pdf

K. Jaruskova, K. Tsolaki, S. Vallecorsa, D. Kampert, V. Saletore, Generative Models for Simulation (17 March). Presented at CERN openlab Technical Workshop, Geneva, 2023.

M. Bunino, A. Di Meglio, M. Girone, K. Tsolaki, S. Vallecorsa, A. Zoechbauer, The interTwin project - Prototyping an interdisciplinary Digital Twin Engine (16 March). Presented at CERN openlab Technical Workshop, Geneva, 2023.

M. Bunino, A. Zoechbauer, itwinAI: Empowering Scientific AI Development and Deployment (27 September). Presented at IBERGRID 2023, Madrid, 2023.

I. Luise, EMP2 - Environmental Modelling and Prediction Platform (December). Presented at KT applied AI internal workshop, Geneva, 2023.

I. Luise, Environmental Applications at CERN: Focus on the AtmoRep project (November). Presented at ETH AI+Environment Summit, Zurich, 2023.

C. Lessig, AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning (September). Presented at the Large-scale deep learning for the Earth System workshop, Bonn 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (June). Presented at PASC 2023, Davos, 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (May). Presented online at Google Air Quality Journal Club, 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (May). Invited talk presented online at NVIDIA research, 2023.

C. Lessig , AtmoRep: Large scale representation learning of atmospheric dynamics (May). Presented as invited talk at ECMWF, Bonn 2023.

E. Wulff, F. Mokhtar, D. Southwick, M. Zhang, M. Girone, J. Duarte, J. Pata, Scalable Neural Network Models and Terascale Datasets for Particle-Flow Reconstruction (6 November 2023). Presented at the ML4Jets Conference, Hamburg, 2023.

M. Girone, E. Wulff, WP4: Data-Driven Use-Cases at Exascale (18 January 2023). Presented at the CoE RAISE All-Hands meeting, Geneva, 2023. E. Wulff, D. Southwick, M. Girone, A. Lektauers, J.P. García Amboage, E. Cuba, Event Reconstruction and Classification at the HL-HLC (18 January 2023). Presented at the CoE RAISE All-Hands meeting, Geneva, 2023.

E. Wulff, M. Girone, J.P. García Amboage, J. Pata, Distributed Hyperparameter Optimization using HPC systems (16 March 2023). Presented at the CERN openlab Technical Workshop, Geneva, 2023.

D. Southwick, HPC Benchmarking for Exascale (16 March 2023). Presented at the CERN openlab Technical Workshop, Geneva, 2023.

J.P. García Amboage, E. Wulff, M. Girone, T.F. Pena, Model Performance Prediction for Hyperparameter Optimization of Deep Learning Models Using High Performance Computing and Quantum Annealing (8 May 2023). Presented at the 26th International Conference on Computing in High Energy & Nucelar Physics (CHEP2023)

E. Wulff, Hyperparameter Optimization for Deep Learning using High Performance Computing (18 July 2023). Presented at the CERN openlab Summer Student Lecture Programme, Geneva, 2023.

M. Girone, E. Wulff, WP4: Data-Driven Use-Cases at Exascale (28 August 2023). Presented at the Second CoE RAISE All-Hands meeting, Hveragerði, Iceland, 2023.

E. Wulff, D. Southwick, M. Girone, A. Lektauers, J.P. García Amboage, E. Cuba, Event Reconstruction and Classification at the HL-HLC (28 August 2023). Presented at the Second CoE RAISE All-Hands meeting, Hveragerði, Iceland, 2023.

D. Southwick, Data Challenges (29 August 2023). Presented at the Second CoE RAISE All-Hands meeting, Hveragerði, Iceland, 2023.

D. Southwick, HPC Integration in Data Intensive Scienc (27 August 2023). Presented at the 7th edition of the cross-disciplinary International Summer School INFIERI, São Paulo, 2023.

J.P. García Amboage, E. Wulff, M. Girone, Accelerating hyperparameter optimization using performance prediction on a heterogeneous HPC system (14 July 2023). Presented at the CERN Computing Seminar in the openlab Series, Geneva, 2023.

S. Hageboeck et al, Madgraph5_aMC@NLO on GPUs and vector CPUs: experience with the first alpha release, (8 May) Presented at 26th International Conference on Computing in High Energy & Nuclear Physics (CHEP 2023), Norfolk, 2023 https://indico.jlab.org/event/459/contributions/11829/

A. Valassi et al, Speeding up Madgraph5_aMC@NLO through CPU vectorization and GPU offloading: towards a first alpha release, (24 October), Presented at 21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT), Bari, 2022 https://indico.cern.ch/event/1106990/contributions/4997226/

J. Teig, Automated performance profiling of event generation on heterogeneous architectures comparing CUDA and oneAPI/SYCL, (16/17 March 2023), Poster presented at Openlab Technical Workshop, Geneva, 2023. https://indico.cern.ch/event/1225408/

Aach, M. (Corresponding author), Wulff, E., Pasetto, E., Delilbasic, A., Sarma, R., Inanc, E., Girone, M., Riedel, M. & Lintermann, A., "A Hybrid Quantum-Classical Workflow for Hyperparameter Optimization of Neural Networks" (22 May 2023). Presented at ISC High Performance 2023, (ISC2023).

D. Southwick, M. Girone, E. Wulff, M. Bunino, A. Zoechbauer, HPC and CERN: Integration and Challenges (23 October 2023). Presented at the CERN Computing Seminar in the openlab Series, Geneva, 2023.

J.P. García Amboage, E. Wulff, M. Girone, M. Aach, A. Delilbasic, E. Pasetto, R. Sarma, M. Riedel, A. Lintermann, Distributed Hybrid Quantum-Classical Performance Prediction for Hyperparameter Optimization (23 November 2023). Presented at the 7th Quantum Techniques in Machine Learning Conference (QTML), Geneva, 2023.

A. Patil, Data Analytics for Industrial Control Systems (16 March). Presented at CERN openlab Technical Workshop, Geneva, 2023. https://cern.ch/6pv68

A. Karasinski, Migration between Kubernetes versions doesn't have to be error-prone (16 March), Presented at 2023 CERN Openlab Technical Workshop, Geneva, 2023

R. Cardoso, Unlock Enterprise-Grade Machine Learning on Oracle Cloud Featuring CERN (19 September), Presented at Oracle Cloud World, Las Vegas, 2023

T. James, Real-time deep learning inference and FPGA based processing for level 1 trigger scouting at CMS (16 March). Presented at CERN Openlab Workshop, Geneva, 2023. URL: https://indico.cern.ch/event/1225408/contributions/5243978/attachments/2613031/4515255/tjames_openlab_160323.pdf

M. Bunino, A. Di Meglio, M. Girone, K. Tsolaki, S. Vallecorsa, A. Zoechbauer, interTwin (16 March). Presented at CERN openlab Technical Workshop, Geneva, 2023.

A. Manzi, interTwin: Co-design an prototyping an interdisciplinary Digital Twin Engine for Science (21 June). Presented at EGI 2023, Poznań, 2023.

I. Luise, K.Tsolaki, A. Zoechbauer, Digital Twins: introduction and use cases (07 August). Presented at CERN openlab Summer Student Lecture Programme, Geneva, 2023.

I. Luise, AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning (16 December). Presented at NeurIPS, San Diego, 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (April). Presented at EGU23, Vienna, 2023.

I. Luise, EMP2 - Environmental Modelling and Prediction Platform (March). Presented at CERN OpenLab technical workshop, Geneva, 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (March). Presented at the IntelliAQ workshop, Bonn, 2023.

C. Lessig, AtmoRep: Large scale representation learning of atmospheric dynamics (February). Presented at the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, online, 2023.

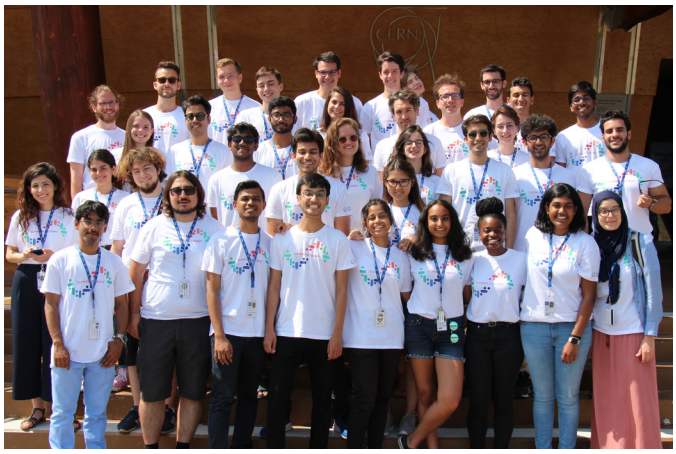
# TRAINING & EDUCATION

As a part of the education and training programme, CERN openlab **runs various initiatives that support participation of young scientists and other research organisations**. For example, the CERN openlab summer student programme provides undergraduate and master's level students with an opportunity to work on one of the R&D projects for nine weeks under experts' supervision.

Apart from that, the public has open access to CERN openlab **lectures that cover a wide range of computing topics, from AI to exascale computing and quantum technologies**. For community development, CERN openlab offers regular specialised technical training to members of the scientific community. Recent examples include NVIDIA GPU programming with CUDA hackathons, Intel software tools, hands-on training for unified programming and AI/ML models, and containerisation and virtualisation training including Kubernetes.

When preparing the future of scientific computing, it is vital to **ensure that the computing specialists of the future have the right skills to enable them to fully capitalise on new, innovative technologies**. Through projects, lectures, and workshops, computer scientists are being equipped with indispensable knowledge that inspires scientific advancement and fuels innovation.

↓ Students from past CERN openlab summer student programmes



# TECHNICAL WORKSHOP

CERN openlab holds an **yearly Technical Workshop where members of CERN openlab engage with industry members and the ICT community** to showcase the work being done, review of the R&D projects carried out during the past year and discuss future plans. This event **features technical talks, poster sessions and technology tracks dedicated to our industrial partners with invited speakers**. It is a good opportunity for industry partners to meet with the students and fellows working on common projects.

In 2023, the Annual CERN openlab Technical Workshop saw the end of Phase VII and kicked-off Phase VIII with a passing of testimony from Alberto Di Meglio (Head of CERN openlab during Phase VII) to Maria Girone (current Head of CERN openlab).

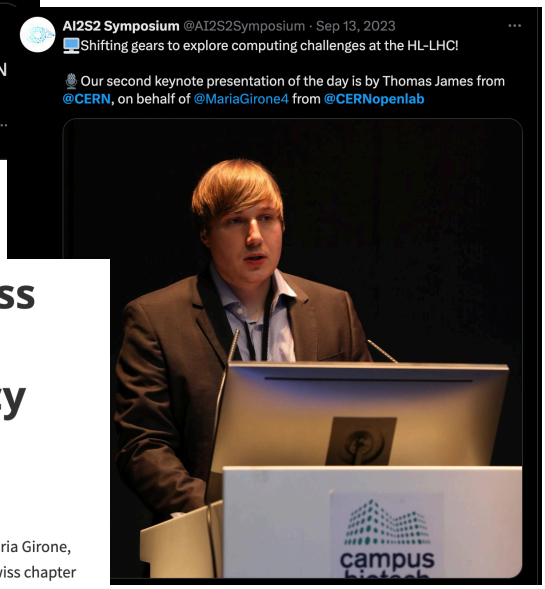
↓ CERN openlab Technical Workshop 2023


# COMMUNICATION & OUTREACH

CERN openlab has a **dedicated communication team that works extensively to promote and communicate its partnerships and the work emerging from its R&D projects**. CERN openlab is present in various social media channels (Facebook, Twitter, LinkedIn Group), it is occasionally highlighted on CERN social media accounts, as well as in partners and other important industry channels.

Disseminating its research to the public is an important mission for CERN openlab, demonstrating the importance of the R&D work being developed not only to the ICT community but to society in general.

↓ Some examples of CERN openlab coverage in 2023

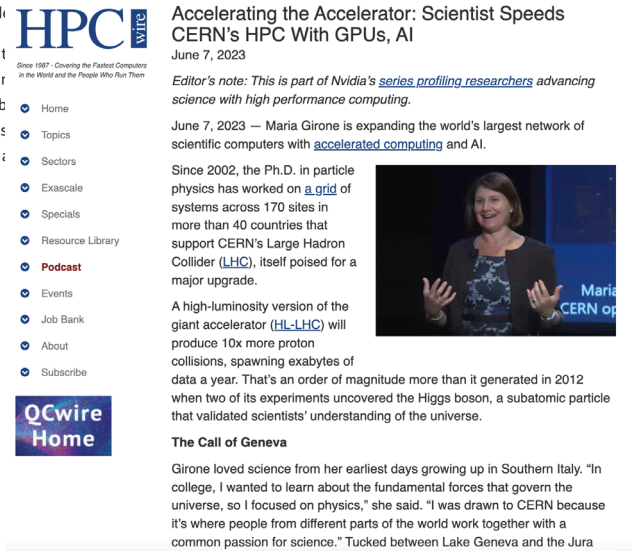

## CERN openlab CTO co-founds Swiss chapter of Women in High-Performance Computing advocacy group

17 FEBRUARY, 2023 | By Andrew Purcell

The four founders of the new chapter of Women in HPC, photographed during a special networking session at last year's Platform for Advanced Scientific Computing (PASC22) Conference in Basel, Switzerland. (Image: CERN)

CERN openlab's Chief Technology Officer, Maria Girone, is one of four founding members of a new Swiss chapter of the Women in HPC (WHPC) advocacy group. The announcement comes on the International Day of Women and Girls in Science, which is dedicated to reducing gender dispa
levels of scientific end

Women in HPC works t
performance computir
the University of Edinb
Women in HPC organis
and provides support
working in this field.



# CERN OPENLAB
# PHASE VIII (2024-2026)

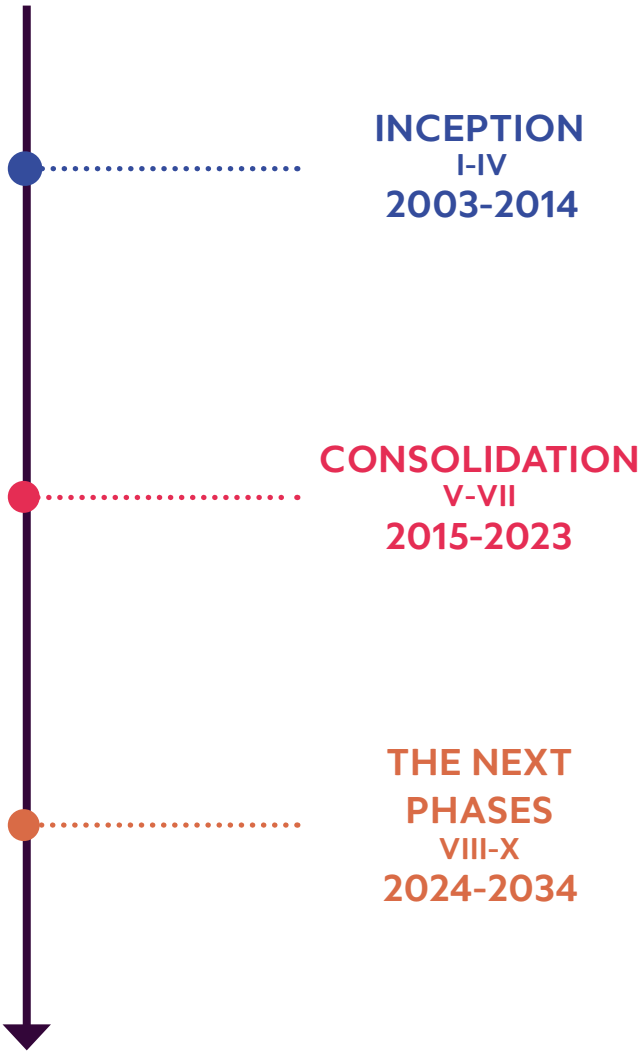
CERN openlab operates within structured **three-year phase cycles** designed to systematically assess technological evolution, anticipate future needs, and delineate overarching thematic priorities. This approach ensures the maintenance of a relevant and current research programme, fostering effective collaborations and innovative advancements.

CERN, along with the High Energy Physics (HEP) community, is gearing up for the HL-LHC program and the future of particle physics. Building upon the achievements of the initial decade of LHC operations, marked by swift advancements in computing architectures, infrastructures, algorithms, and disruptive technologies, CERN openlab is poised to harness these changes. With its extensive network spanning industry and academia, CERN openlab is uniquely positioned to validate and integrate emerging technology capabilities while facilitating access to novel infrastructures.

To address the impending challenges, **CERN openlab is reviewing its operational framework**, notably restructuring industry membership into two tiers: Associate and Partner (refer to membership levels and governance sections).

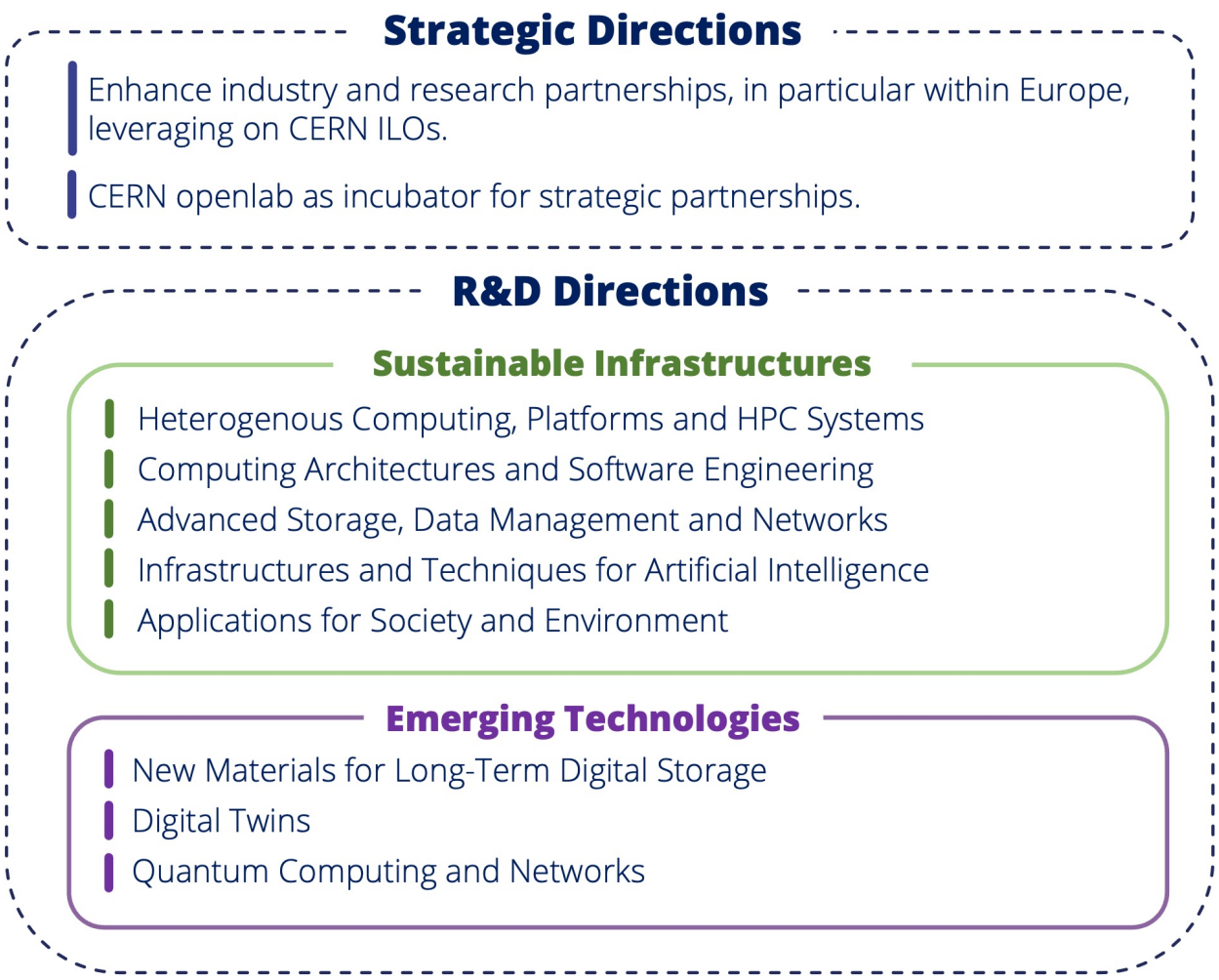
CERN openlab will build on its consolidated agile mechanism that facilitates the definition of practical projects, to remain an efficient vehicle for innovation. Projects typically span between 1 to 3 years, an optimal duration to yield actionable outcomes aligned with the latest technological advancements. This framework represents an **ideal mechanism to nurture initial collaborations that can evolve into long-term programmes, but also to promote new partnerships and cultivate diverse networking avenues**.

In the coming years, CERN openlab will introduce additional tools to facilitate sustained, long-term collaborations, serving as a platform not only for partnership preparation but also as a "training ground" for IT professionals seeking to expand their skill sets into project management and broader coordination roles.



Throughout its history, CERN has been at the forefront of big data scientific research, with CERN openlab playing a pivotal role in tackling the associated computing challenges. By fostering collaborations with industry and research organisations, **CERN openlab empowers the HEP community in its research endeavours**. In response to the evolving landscape of scientific research, including the advent of exascale computing, CERN openlab spearheads efforts to enhance and scale up IT infrastructure to tackle the upcoming data challenges.

CERN openlab **objectives** are to pioneer sustainable and innovative computing solutions, harness AI and heterogeneous computing for environmental benefits, and foster collaboration and technology transfer between industry and the scientific community. Through collaboration with diverse stakeholders, including the HEP community, other scientific disciplines, and technology providers, CERN openlab fosters co-development of solutions and co-design of infrastructure. This collaborative approach drives innovation and advancement for all parties involved. Moreover, it enables partners to leverage solutions from HEP to address challenges in other fields, ensuring maximal relevance and **impact**. Projects within the CERN openlab framework are dedicated to accelerating computing for science, particularly under the **R&D directions of "Sustainable Infrastructures" and "Emerging Technologies"**.



## WITH THANKS TO

all partners who have collaborated with CERN openlab activities and everyone who has contributed to the content and production of this document.

## ALL IMAGES PROVIDED BY THE CERN AUDIO-VISUAL PRODUCTION SERVICE, CERN OPENLAB COMMUNICATION TEAM OR CERN OPENLAB MEMBERS, EXCEPT:

page 38, 46, 50 - Freepik

## GET IN TOUCH

openlab.cern

facebook.com/cernopenlab

x.com/cernopenlab

openlab-communications@cern.ch

https://www.linkedin.com/showcase/cerncomputing

WEBSITE FACEBOOK TWITTER LINKEDIN

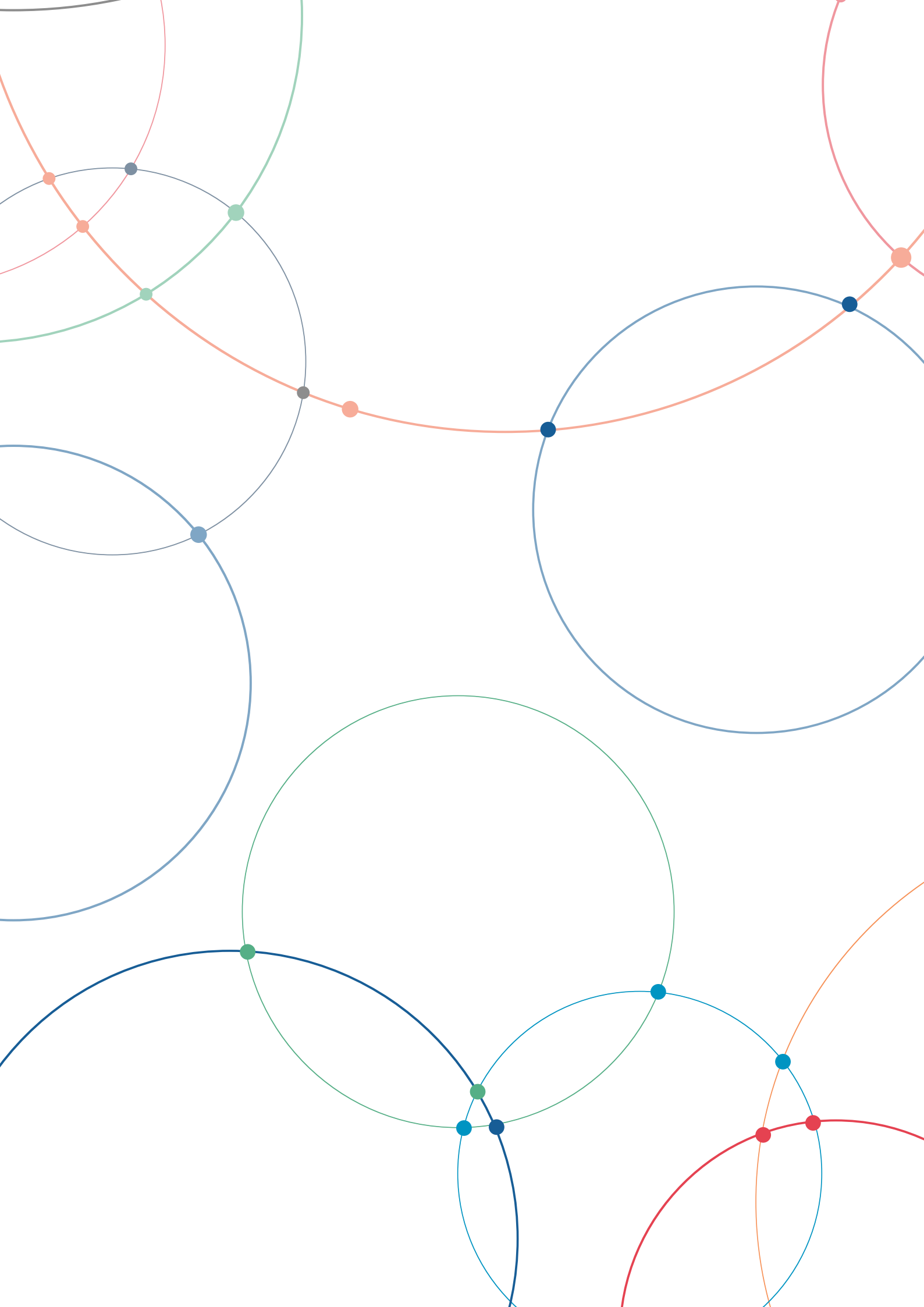
**Editors**
**Maria Girone**
(Head of CERN openlab)

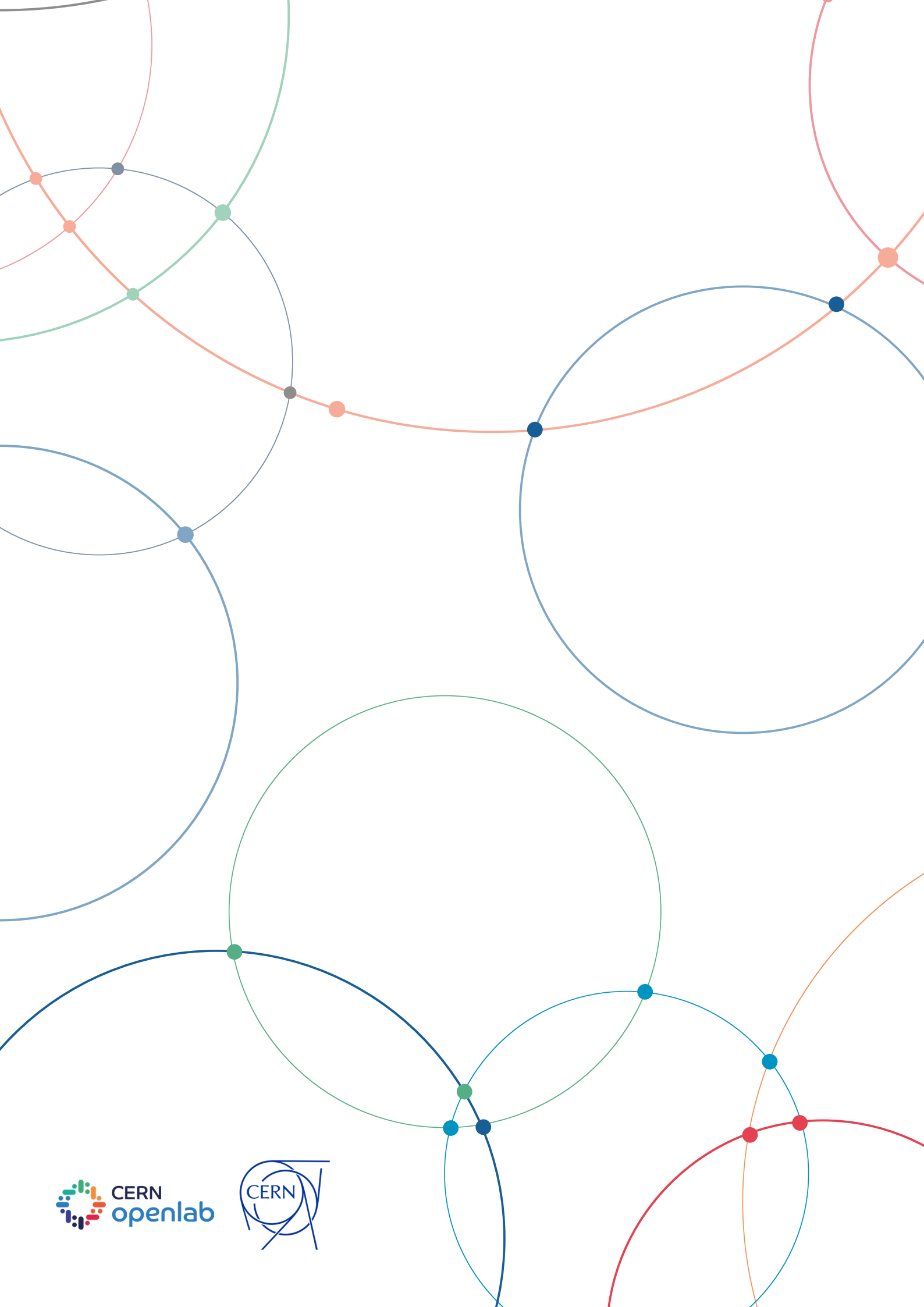
**Mariana Velho**
(CERN IT Communications & CERN openlab Chief Communications Officer)

**Graphic Design & Layout**
**Mariana Velho**
(CERN IT Communications & CERN openlab Chief Communications Officer)

**ISBN**
978-92-9083-649-0 (Digital)
978-92-9083-650-6 (Printed)

**Published by CERN**




CERN
openlab
CERN