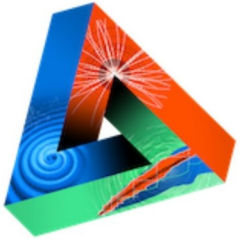
Working Group 4: Machine Learning and Artificial Intelligence Infrastructure

Andreas Ipp (TU Wien), Sascha Caron (Radboud University, Nikhef)

June 17, 2025



JENAA

Joint ECFA-NuPECC-APPEC Activities

EuCAIF

ECFA

European Committee for Future Accelerators

NuPECC

APPEC

Outline

- JENA and EuCAIF Working Groups
- Survey results
- White Paper with 12 recommendations

Strategic White Paper on AI Infrastructure for Particle, Nuclear, and Astroparticle Physics: Insights from JENA and EuCAIF

Sascha Caron,[*a,b] Andreas Ipp,[*c] Gert Aarts,[d] Gábor Bíró,[e,f] Daniele Bonacorsi,[g,h] Elena Cuoco,[g,h] Caterina Doglioni,[i] Tommaso Dorigo,[j,k] Julián García Pardiñas,[l] Stefano Giagu,[m] Tobias Golling,[n] Lukas Heinrich,[o] Ik Siong Heng,[p] Paula Gina Isar,[q] Karolos Potamianos,[r] Liliana Teodorescu,[s] John Veitch,[p] Pietro Vischia,[t] Christoph Weniger[u]

arXiv:2503.14192

JENA White Paper Executive Summary

Artificial Intelligence

We recommend establishing a dedicated organizational structure to coordinate strategic investments in Artificial Intelligence for fundamental physics. We also recommend conducting a feasibility study to compare a centralized Graphics Processing Unit (GPU) facility with federated and hybrid High Performance Computing infrastructures. Additionally, we advocate for supporting physics-specific large language models, foundation models, and benchmarking while prioritizing scalable computing, robust data infrastructure, and Machine Learning operations.

JENA White Paper on European Federated Computing
The JENA Computing Initiative, February 28, 2025

History of this working group

JENA Seminar in Madrid, May 2022

- Identification of joint computing requirements for the next decade.
- Discussions on the strategy and implementation of **European federated computing** at future large-scale research facilities.

JENA Computing Initiative workshop in Bologna, June 2023

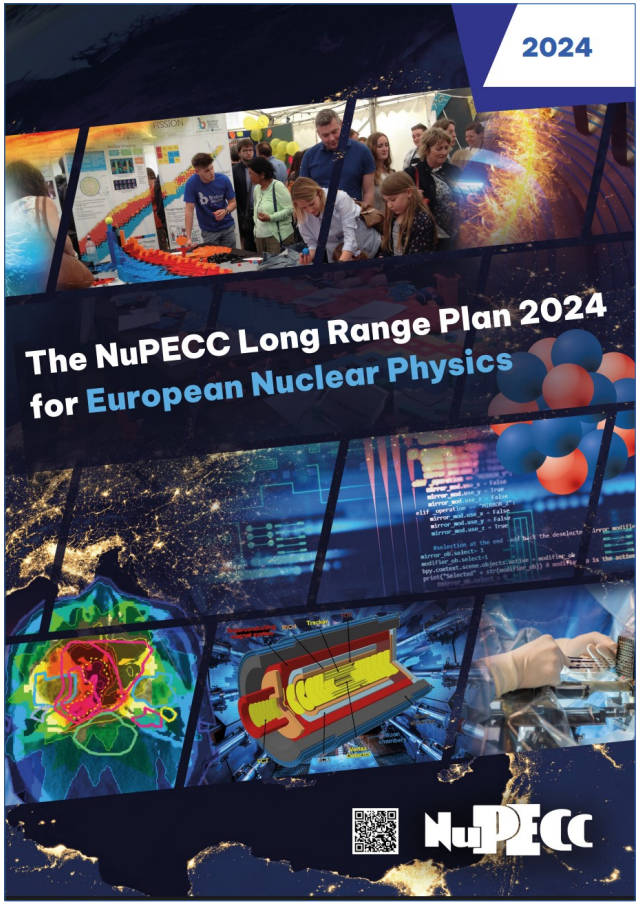
- Creation of five working groups to coordinate a white paper:
  - WP1: HTC, WLCG and HPC (HPC)
  - WP2: Software and Heterogeneous Architectures (Software)
  - WP3: Federate Data Management, Virtual Research Environments and FAIR/Open Data (Data)
  - WP4: Machine Learning (ML) and Artificial Intelligence (AI)
  - WP5: Training, Dissemination, Education (TDE)

EuCAIFCon in Amsterdam, May 2024

- Creation of five working groups
  - WG 1: Foundation models & discovery
  - WG 2: AI-assisted co-design of future ground- and space-based detectors
  - WG 3: FAIR-ness & Sustainability
  - WG 4: Machine Learning and Artificial Intelligence Infrastructure (JENA WP4)
  - WG 5: Building bridges - Community, connections and funding

NuPECC Long Range Plan 2024 for European Nuclear Physics

arXiv.org:2503.15575



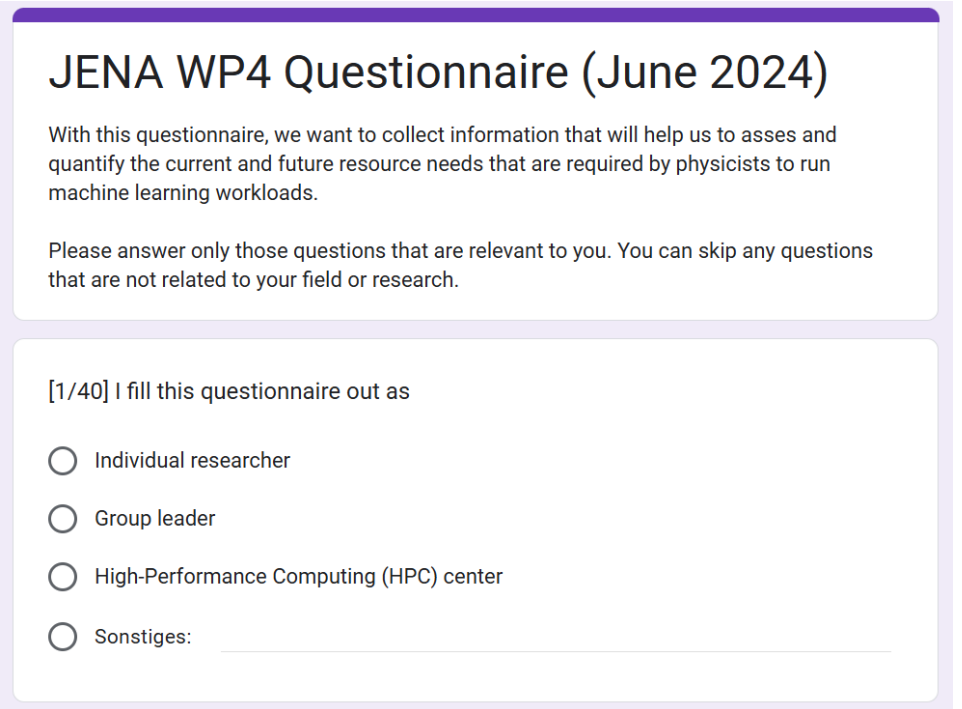
Recommendations Machine Learning and Artificial Intelligence

(Page 151 / 186)

- Transform ML prototypes into applications for production
- Foster data-sharing in nuclear physics
- Strengthen computational resources
- Train scientific foundation models

Working group progress

- March – June 2024:
  - About 35 participants in mailing list
  - Regular online meetings
  - Compile 40 questions
- June – September 2024:
  - Send out survey to EuCAIF, JENA and HPC communities
  - Collect 137 responses
- September – December 2024:
  - Discuss results and work out 12 recommendations in subgroups
  - Collect feedback from JENA and external experts
- December 2024 – February 2025:
  - Finalize WP4 White Paper
  - Finalize JENA White Paper
- April 2025:
  - Results presented at JENAS 2025 Joint Symposium, Didcot, Oxfordshire, UK
- May 2025:
  - White Paper submitted to IOP journal "Machine Learning: Science and Technology"



JENA WP4 Questionnaire (June 2024)

With this questionnaire, we want to collect information that will help us to asses and quantify the current and future resource needs that are required by physicists to run machine learning workloads.

Please answer only those questions that are relevant to you. You can skip any questions that are not related to your field or research.

[1/40] I fill this questionnaire out as

- ○ Individual researcher
- ○ Group leader
- ○ High-Performance Computing (HPC) center
- ○ Sonstiges:

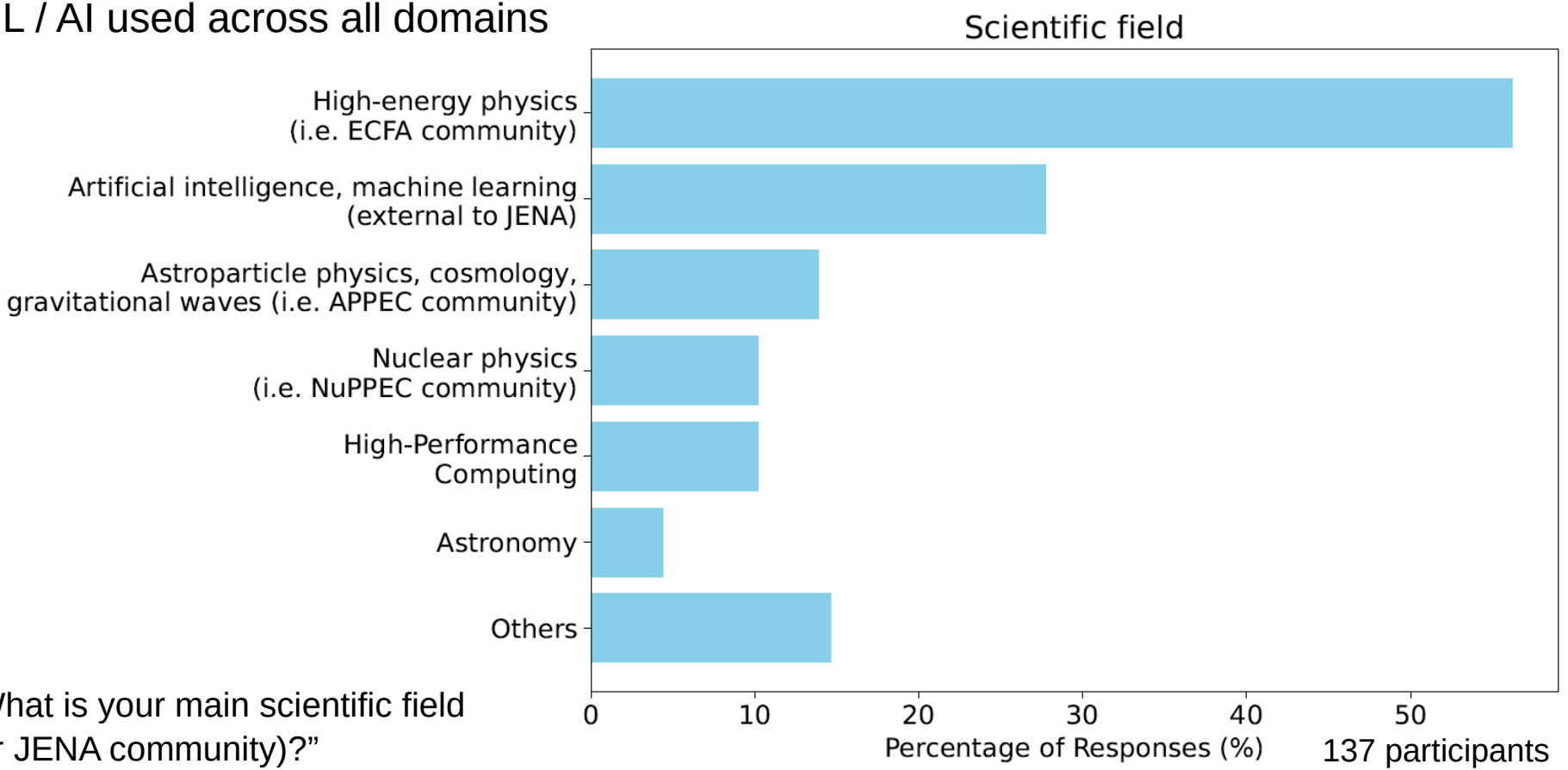
Related surveys by

SPECTRUM CoP

EuroHPC Joint Undertaking

Survey results

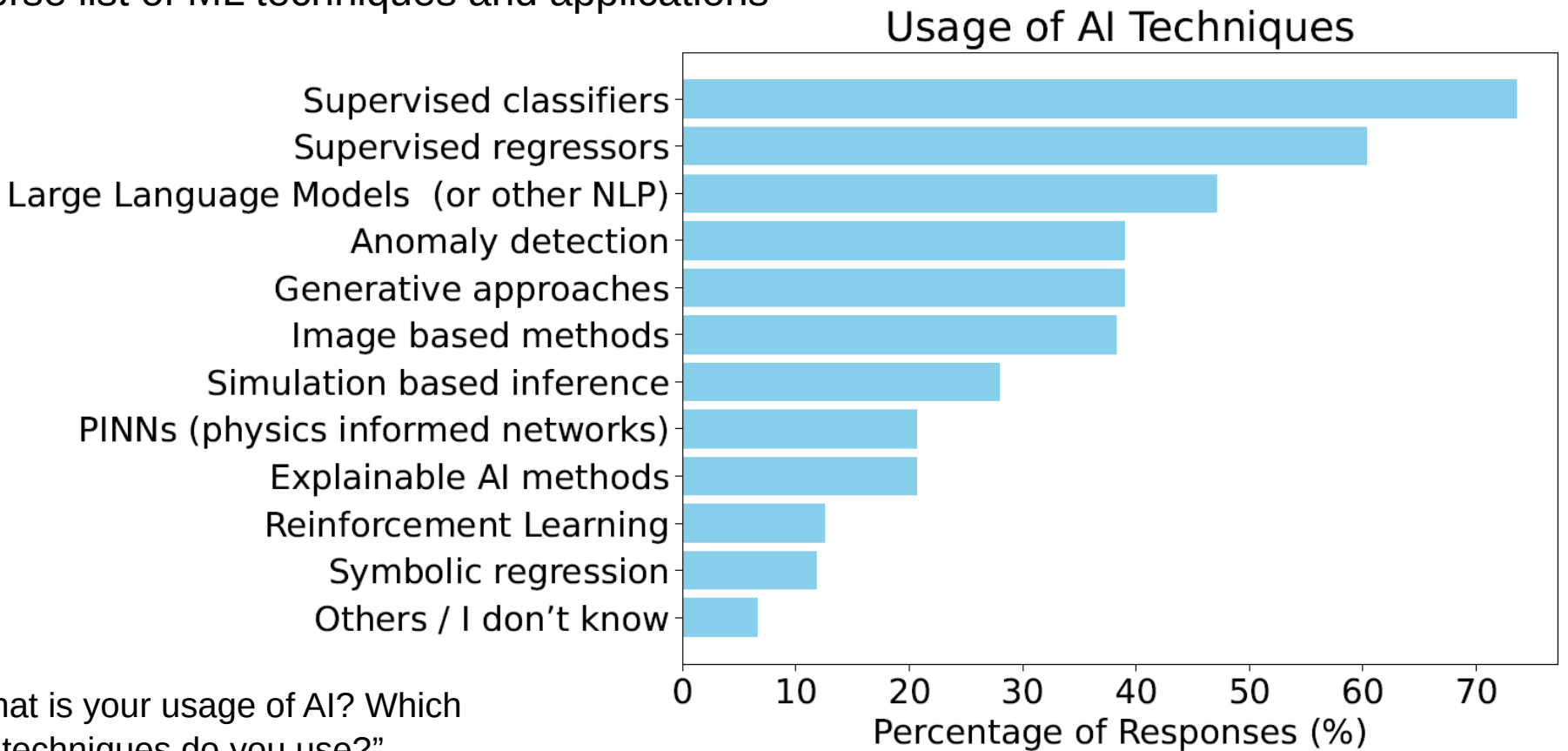
ML / AI used across all domains



“What is your main scientific field (or JENA community)?”

Survey results

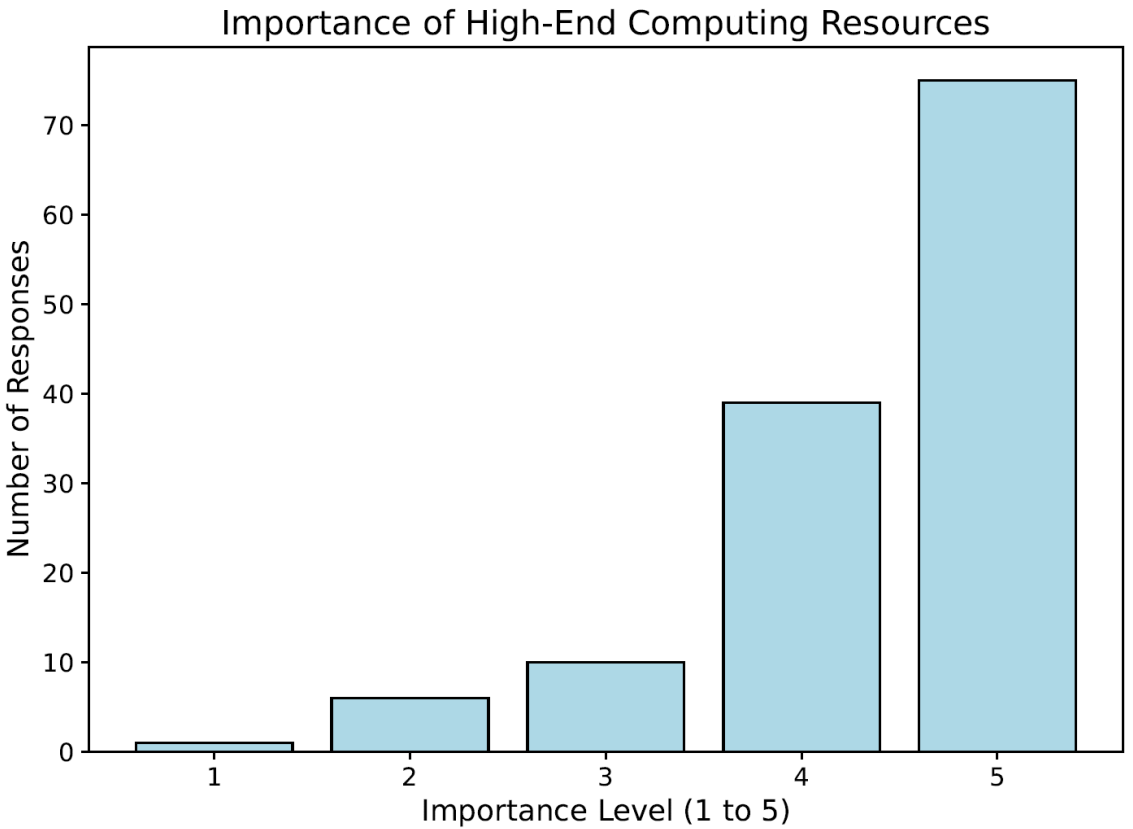
Diverse list of ML techniques and applications



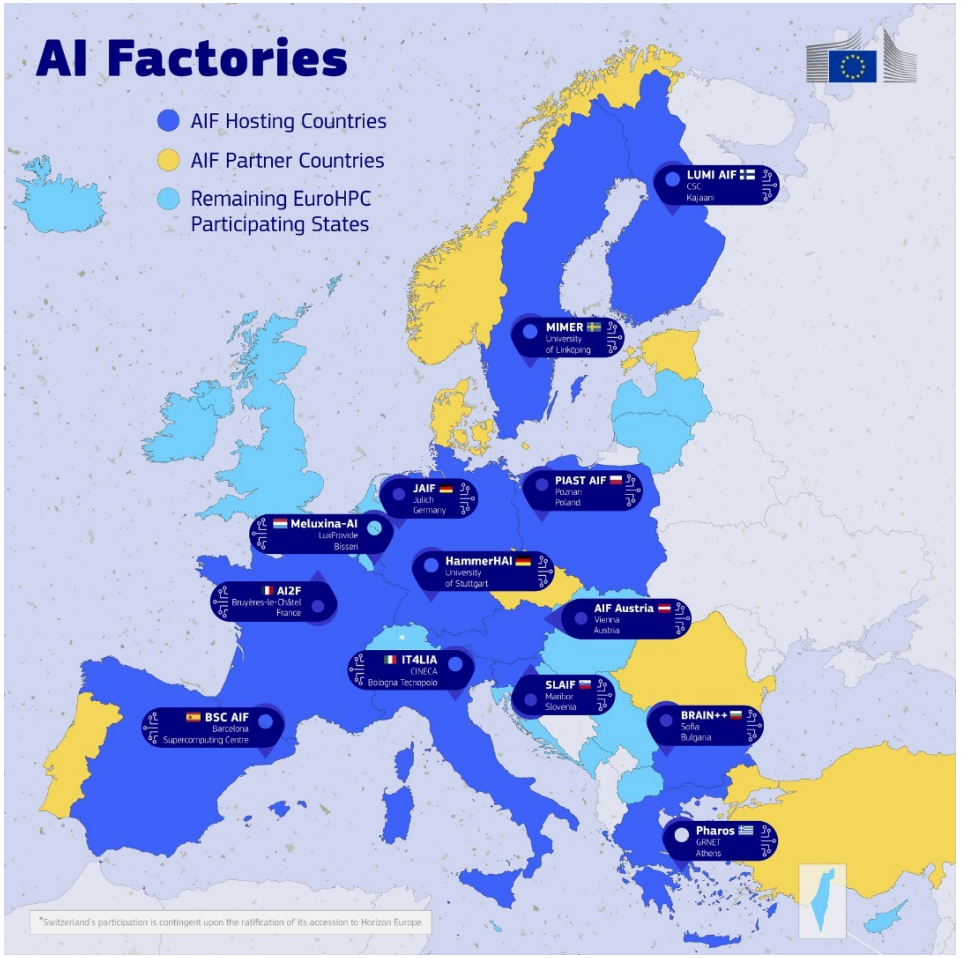
"What is your usage of AI? Which ML techniques do you use?"

Increasing need for compute

“How important would high-end/large GPUs (like the A100 etc) be for training large-scale deep learning models over the next five years?”



AI Factories and AI Gigafactories



The European Commission turns 13 scientific supercomputer centres into AI factories (max. €600m per AI factory, up to 25,000 AI processors)

3-5 gigafactories are planned (€20bn, €3-5bn each, >100,000 AI processors)
AI Continent Action Plan (April 9, 2025)

What are the needs of our scientific communities?

Are we well connected to or involved in the activities of the AI (giga-)factories?

Recommendations R1-2: How to scale?

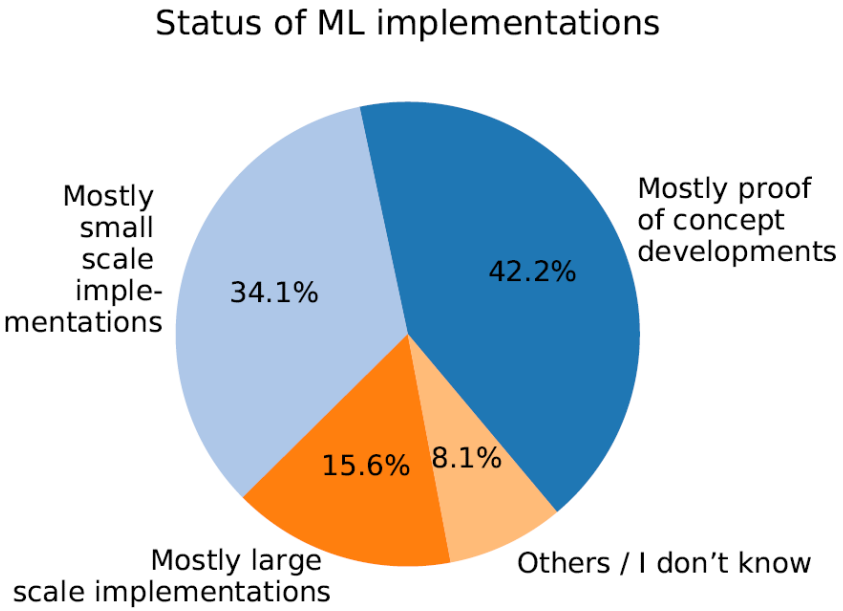
R1: Discussion on centralized large-scale GPU facility vs. federated HPC infrastructures

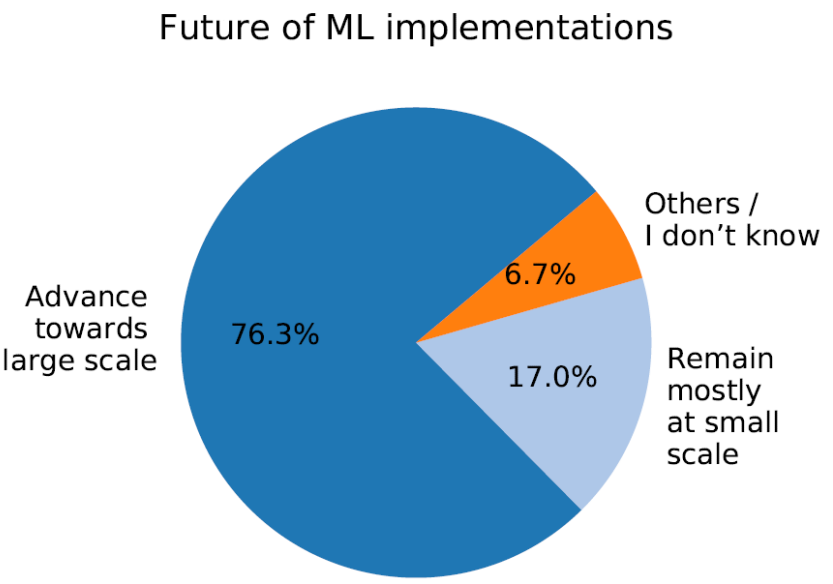
(R1) Convene dedicated discussions with national research groups and funding bodies to assess and compare the feasibility of a **centralized large-scale GPU facility versus federated and hybrid high-performance computing (HPC) infrastructures**, supported by working groups developing detailed implementation plans for both options, with the aim of accelerating the deployment of a scalable AI infrastructure.

R2: Scalable data infrastructure initiative, develop platforms for distributed workloads

(R2) Establish a **scalable data infrastructure** initiative by creating shared repositories and tools, and **developing platforms for distributed workloads.** These efforts need targeted funding programs and a concrete community-driven structure to ensure widespread adoption and collaboration in AI research.

From proof of concept to large scale models





“In your current work, what is the predominant status of your machine learning implementations?”

“How do you foresee the status of ML implementations evolving in your field over the next five years?”

R3-4: From R&D to production

R3: Encourage transition from R&D into production

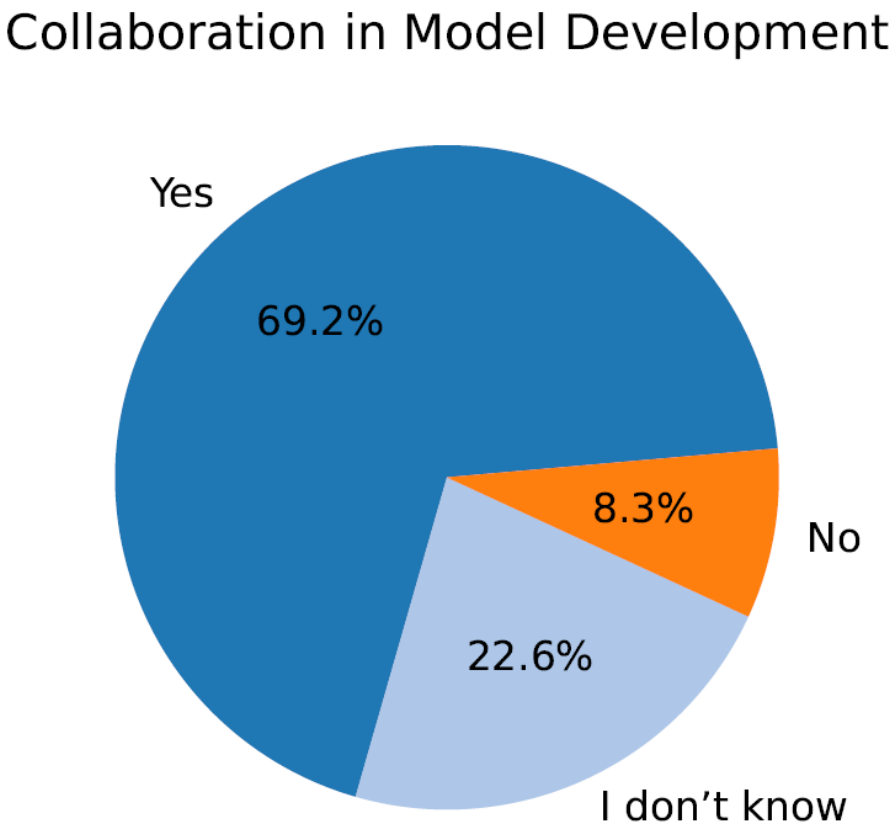
(R3) Encourage funding to **transition AI-driven R&D activities into production-ready applications** within established experimental workflows, focusing on adopting best practices to achieve practical, scalable improvements without requiring a complete system overhaul.

R4: Fund Machine Learning Operations (MLOps) personnel

(R4) Allocate **dedicated funding to establish and support specialized Machine Learning Operations (MLOps) personnel** to streamline the integration and ensure the sustainable maintenance of AI models within production workflows. This effort should encompass the development of community-wide standards, tools, and platforms to effectively manage the entire lifecycle of machine learning models.

Large ML models

“Should we collaborate more in the development of large-scale ML models (e.g. foundation models) for physics?”



R5-6: Collaborate on Large Physics Models

R5: Create Science Large Language Models

(R5) Invest in the **creation of "science Large Language models (LLMs)"** tailored to the unique challenges of fundamental physics and science, balancing the use of commercial tools for general tasks with specialized models for domain-specific needs. This requires dedicated funding, access to large-scale GPU infrastructure, and collaborative frameworks to enable transparent, efficient, and impactful AI solutions.

EuCAIF WG 1: Foundation models & discovery

MIAPbP
Munich Institute for Astro-, Particle and BioPhysics

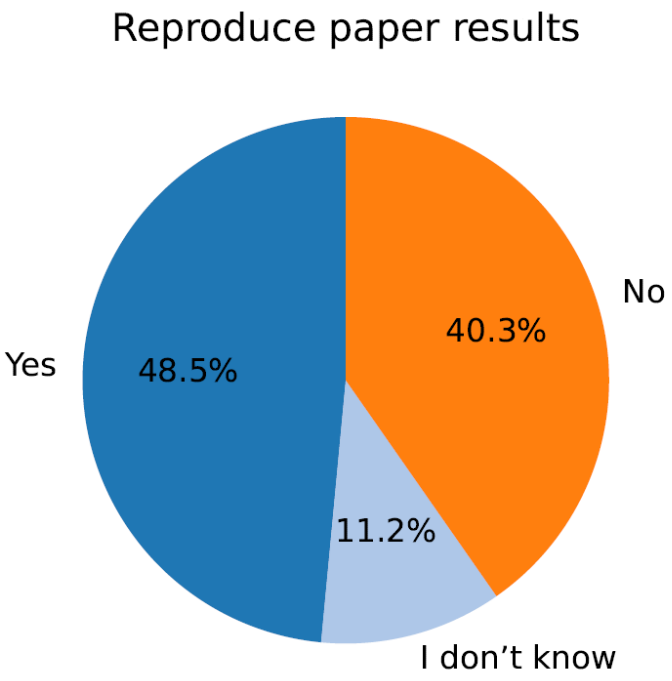
Munich Sept 2025 Workshop:
Build Big or Build Smart: Examing Scale and Domain Knowledge in Machine Learning for Fundamental Physics

Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models
arXiv:2501.05382

R6: Develop foundation models

(R6) Establish dedicated funding schemes and a collaborative structure to develop community-driven **foundation models trained on domain-specific data to learn meaningful representations serving a large variety of downstream tasks**. This effort should identify representative benchmarks, extendible in complexity and realism by integrating both synthetic and real-world data to address domainshift issues, leverage physics-informed augmentations, ensure models are rooted in scientifically relevant tasks, and foster automation, explainability and interpretability to accelerate AI advancements in the field, and to develop a well-defined AI demonstrator for the wider AI community.

Reproducabilty and benchmarks



R7: Create AI benchmarks in fundamental physics

(R7) Establish a dedicated effort to **develop and maintain extensible benchmarks for various AI tasks in fundamental physics**, such as event classification, parameter inference, tracking and anomaly detection. Support efforts to encourage researchers to share well-documented surrogate models to promote reusability and collaboration to drive innovation and standardisation in this area.

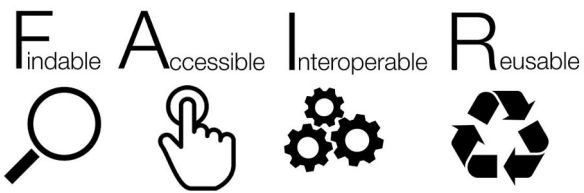
“Did you ever manage to reproduce someone’s paper results / retrain their model?”

R8: Sustainability

R8: Raise awareness to environmental impact

(R8) Investigate and adopt benchmarks that are suitable for fundamental sciences to raise **awareness of the environmental impact** of large AI models. Consider collective mitigation strategies such as optimising widely used frameworks and models and their interfaces to existing software frameworks, as well as individual strategies that lead to minimal/acceptable performance loss. Cooperate with infrastructure and computing sites to minimise carbon costs of compute-intensive AI tasks.

R9: FAIR compliance in publication criteria and practises



(R9) Develop activities aiming to **integrate FAIR compliance into publication criteria and practices**, recognise and incentivise the FAIR compliant work in policy and funding measures as well as career progression, build community awareness through training and collaboration, and support the development of technical tools and standards to facilitate the adoption of the FAIR principles.

R10: Training

R10: Training courses and summer schools

(R10) Fund the development and **organization of practical training courses and summer schools** to equip researchers with the skills to implement open research and reproducibility requirements, incorporating examples and industry perspectives. Facilitate partnerships with industry to sponsor training events and provide placements for early-career and senior researchers, enhancing their AI and data science expertise while fostering connections between fundamental science and commercial applications.

LaVA – Lattice Virtual Academy

Data & Software Carpentries

ErUM Data Hub

R11-12: Interdisciplinarity and coordination

R11: Interdisciplinary research and knowledge transfer

(R11) **Establish interdisciplinary research initiatives** that bring together physicists, AI specialists, software engineers, HPC experts, and potentially experts from other related fields, to tackle large-scale projects. Provide dedicated funding to support **cross-domain knowledge transfer** through workshops, training programmes and open source collaboration. Invest in shared repositories and computing platforms to enable data sharing, modelling development and collaboration between different disciplines.

EuCAIF

EuCAIFCon 2025:
16.-20. June 2025,
Cagliari, Sardinia, Italy

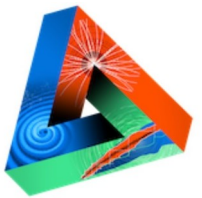
Propose dedicated joint meeting on AI Strategy in Fundamental Physics to bring together JENA bodies & European labs (CERN, DESY, etc.)

R12: Organisational structure to coordinate strategic investments in AI

(R12) **Establish and support a dedicated organisational structure** to coordinate strategic investments in AI for fundamental physics to accelerate the development and deployment of innovative AI technologies tailored to the specific challenges of the field. Existing initiatives like the European Coalition of AI for Fundamental Physics (EuCAIF) can serve as a model for such efforts.

Conclusions

- Insights from JENA and EuCAIF Working Groups
- White Paper on AI infrastructure with 12 recommendations based on survey results, discussions and feedback
- What do we need **organization-wise**?
  - R1: Discussion on centralized vs. federated HPCs
  - R5-7: Collaboration on large physics models
  - R11-12: Organizational structure for investments in AI
- What do we need **funding-wise**?
  - R2: Scalable infrastructure
  - R3-4: Fund R&D to production, need ML Ops personnel
  - R8-10: Resources for sustainability, FAIR, training

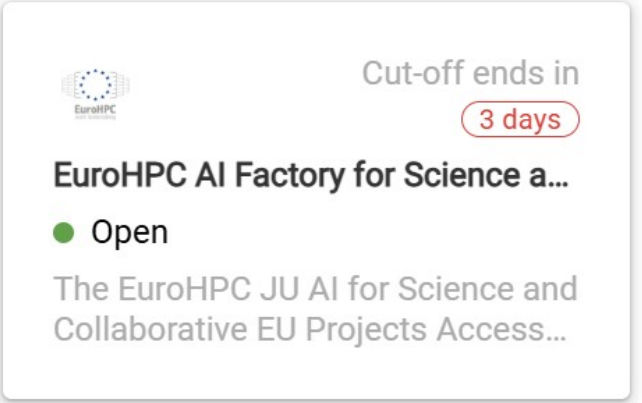


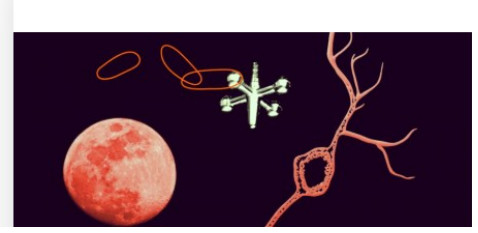
JENAA
Joint ECFA-NuPECC-APPEC Activities

EuCAIF

How shall we proceed with this working group?

- How can we enable the recommendations?
- Monitor future development?
- Common applications to EuroHPC, AI factories and AI gigafactories?
  https://access.eurohpc-ju.europa.eu/
- Collaborations and doctoral networks on large-scale ML?
  https://marie-sklodowska-curie-actions.ec.europa.eu/actions/doctoral-networks
- Should we continue? Transform into different working group? Other ideas?





Doctoral Networks call schedule

- The call is **open**.
- This call will close on **25 November 2025**.