The rise of machine learning in large scientific facilities

Huilin Qu,[1,4] Feng Gao,[2,4] Manqi Ruan,[3] and Sen Qian[3,*]

[1]CERN, Geneva 1211, Switzerland

[2]Interuniversity Institute for High Energies, Université libre de Bruxelles, Bruxelles 1050, Belgium

[3]Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

[4]These authors contributed equally

*Correspondence: qians@ihep.ac.cn

Received: September 24, 2025; Accepted: November 5, 2025; Published Online: November 7, 2025; https://doi.org/10.59717/j.xinn-inform.2025.100006

Citation: Qu H., Gao F., Ruan M., et al. (2025). The rise of machine learning in large scientific facilities. *The Innovation Informatics* **1**:100006.

Large scientific facilities—such as the Large Hadron Collider (LHC) at CERN, the Five-hundred-meter Aperture Spherical Telescope (FAST) in China, the Laser Interferometer Gravitational-Wave Observatory (LIGO) in the US, advanced light sources worldwide, and proposed Higgs factories including Circular Electron Positron Collider (CEPC) and Future Circular Collider (FCC)—are engines of modern discoveries, probing the universe from the smallest particles to the grandest cosmic structures. However, these frontier facilities face two critical bottlenecks: the Data Deluge and the Complexity Barrier. Data volumes are soaring to the petabyte-exabyte scale, challenging rapid event reconstruction, signal identification, and long-term storage. Meanwhile, their operational parameter space of these facilities is immensely complex, making manual tuning costly and, thereby limiting scientific throughput.

Against this backdrop, machine learning (ML) has emerged as a key enabling technology, fundamentally transforming the operational strategies and data analysis paradigms of large scientific facilities. Nevertheless, this transformation is at an early stage. While ML brings great automation and insight, it also raises concerns about transparency, reproducibility, and interpretability—qualities essential for credible discovery. The key challenge is to integrate ML into large facilities without eroding these scientific principles.

MACHINE LEARNING IN ACTION

Real-time triggering and event building. In large scientific facilities, the real-time trigger is the first gate of data processing, selecting rare and meaningful signals under strict latency constraints. Traditional threshold-based systems are simple but rigid, whereas ML models capture complex, multi-dimensional features and reduce the chance of missing weak or unusual events.

At the LHC experiments (CMS and LHCb), hybrid architectures utilizing GPUs or FPGAs alongside CPUs have made ML models an integral part of trigger system.[1] At Belle II, lightweight neural networks have been deployed on FPGA hardware to achieve microsecond-level event discrimination. GPU architectures offer flexibility and rapid development, whereas FPGAs provide deterministic latency and power efficiency. In practice, the choice of platform depends on specific requirements—GPUs are preferred for high-throughput data processing, while FPGAs dominate low-latency front-end implementations. The main challenge is to maintain consistency and traceability between software-trained and hardware-deployed models over extended operation periods. Developing explainable and verifiable ML triggers will be essential to ensure performance and scientific reliability.

Closely related to triggering, event building—assembling detector signals into complete physics events—remains a major challenge for high-luminosity colliders. Because detector responses extend over finite timescales, multiple physics events can overlap in time, and detector noise may further complicate the event separation. ML models implemented on FPGAs or ASICs are expected to enhance both triggering and event building, sustaining performance under extreme data rates.

Event reconstruction and physics analysis. Event reconstruction Transforms raw detector signals into physically meaningful information, such as particle identity, energy, and trajectory. Deep learning models—particularly graph neural networks (GNNs) and transformers—are increasingly replacing classical methods by learning complex, non-linear dependencies directly from detector data.[1] GNNs are especially well-suited to the sparse and irregular geometries of tracking systems or large neutrino detectors, and they can effectively handle heterogeneous data types, providing a powerful framework for "particle-flow" reconstruction. In neutrino detectors, ML-based energy and angle reconstruction has been shown to improve physics performance at a level equivalent to a 10%–30% increase in exposure.

Looking ahead, key frontiers include developing foundation models that generalize across detector types and physics regimes, as well as creating models that are more robust and interpretable. Meanwhile, ML-based reconstruction faces a long-standing calibration challenge—verifying that predicted quantities, such as reconstructed energy and direction, are consistent with

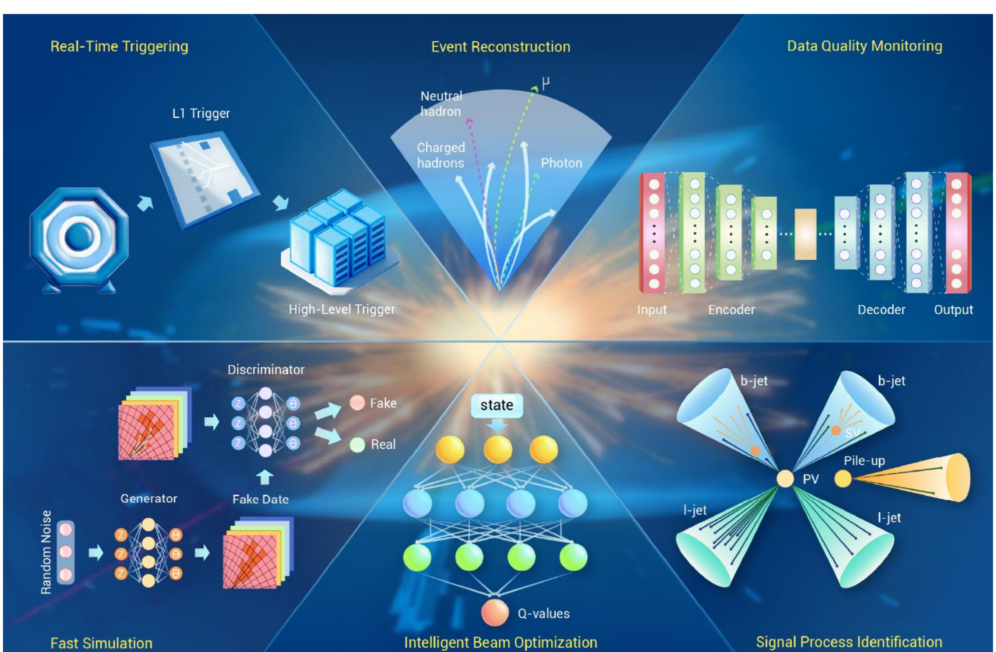

Figure 1. Schematic overview of machine learning applications in large scientific facilities.

the true physical values established by traditional methods or well-understood control samples. Establishing standardized calibration frameworks that align ML outputs to conventional reconstruction references will be crucial for these approaches to gain credibility in precision measurements.

Signal process identification. Particle physics experiments face the immense challenge of identifying rare signal events buried under overwhelming backgrounds, making signal-versus-background discrimination fundamental to discovery. A critical application at the LHC is jet tagging, where ML techniques have revolutionized the identification of jet origins —collimated sprays of particles produced in high-energy collisions. GNN- and transformer-based jet taggers, such as ParticleNet and Particle Transformer, have reduced background rates by over an order of magnitude by directly analyzing the tens to hundreds of particles per jet and exploiting intricate correlations among them. These approaches have become the new standard for jet tagging at the LHC, substantially enhancing the sensitivity of Higgs measurements. More recently, a new paradigm is emerging that shifts toward end-to-end signal-versus-background discrimination. This "holistic approach" treats all reconstructed particle information from an event as a single input to a GNN or transformer, enabling simultaneous disentangling of multiple physics processes.[2] By capturing a fuller picture of the event, this strategy promises significant improvements in measurement precision and represents a clear move toward comprehensive event-level analyses that maximize the discovery potential of current and future colliders.

Fast simulation. High-fidelity simulations play a vital role in large-scale experiments such as those at the LHC. However, traditional first-principles-based approaches (e.g., Geant4) are often too slow for high-luminosity operations. ML-based fast simulation has emerged as a promising solution, replacing the most computationally intensive steps with deep generative models. Architectures such as generative adversarial networks, variational autoencoders, normalizing flows, and, more recently, diffusion and flow-matching models—has been studied to emulate complex detector responses, achieving trade-offs between generation speed and sample fidelity.[3] Example applications include calorimeter shower generation in collider experiments, simulation of neutrino interactions, and modeling cosmic air showers.

Beyond accelerating simulations, a growing concern is the physical fidelity of ML-based generators. Because purely data-driven models may not explicitly enforce energy–momentum conservation or detector symmetries, they risk introducing subtle systematic biases. Recent studies therefore propose physics-informed generative models and uncertainty-aware simulators, which embed physical constraints into the model architecture and quantify prediction uncertainties. Meanwhile, the field is moving toward foundation models that can generalize across detector geometries and physics regimes, enabling transferable simulation capabilities for future large facilities.

Intelligent beam optimization

Intelligent beam optimization. In facilities such as accelerators and colliders, beam stability and precise control of parameters are essential for experimental efficiency. Traditional methods, relying on manual tuning or simple models, are often time-consuming and struggle to adapt to dynamic changes. In recent years, ML has become an essential tool for efficient tuning: at BEPCII, a control method based on deep reinforcement learning has been integrated into daily operations to optimize the collision point.[4] At LCLS, researchers have employed neural network surrogate models as priors for Bayesian optimization, significantly reducing the time required to find optimal solutions in high-dimensional parameter spaces. These applications demonstrate that ML not only reduces the burden on operators but is also driving the evolution of beam control from "experience-driven" to "intelligent and adaptive," making it a core component of future operational models.

These developments mark a shift from experience-driven operation toward intelligent, data-driven control. Yet challenges remain: ensuring the stability and safety of ML agents during live beam operation, maintaining interpretability of the learned policies, and validating performance across different accelerator conditions. Achieving trustworthy autonomy will require hybrid frameworks that combine physics-based models with adaptive learning—laying the groundwork for self-optimizing and self-calibrating scientific facilities.

Data quality monitoring. ML has greatly improved data quality monitoring (DQM) in large scientific facilities. By learning the baseline behavior of detectors across thousands of channels, it can recognize subtle anomalies in real time, enhancing both sensitivity and automation.

At the CMS experiment, a semi-supervised autoencoder has been deployed for real-time monitoring of the electromagnetic calorimeter, achieving high sensitivity with few false alarms.[5] Similar methods are now being tested in neutrino and synchrotron facilities. While ML has made DQM more efficient, ensuring that automated alerts remain interpretable and physically meaningful is still essential. Future systems will likely combine predictive analytics with expert validation, balancing intelligence with reliability.

CHALLENGES AND FRONTIERS

Despite the broad potential of ML techniques, its full integration into large scientific facilities still faces challenges. The foremost challenge lies in computational scalability: as data volumes approach the exascale, algorithms must be both resource-efficient and hardware-aware. Co-designing ML models with accelerator platforms such as FPGAs is increasingly essential to achieve real-time performance within energy and latency constraints.

Ensuring scientific reliability is equally important. ML systems must be calibrated and validated against physical truth, so data-driven correlations are not misinterpreted as new phenomena. Quantifying uncertainties, enforcing physical constraints, and maintaining interpretability under complex detector conditions is central to building trust in ML-based analyses.

Looking ahead, the next paradigm shift is toward greater autonomy. This involves integrating ML-driven controls, fast simulations, and real-time analysis into closed-loop systems, creating "self-driving" facilities that can intelligently optimize their operation. The development of foundation models that generalize across detector geometries and physics regimes will unlock the full potential of large scientific facilities and accelerate scientific discoveries. This evolution will establish ML not merely as a tool, but as a foundational partner in navigating the complexities of modern science.

REFERENCES

1. Boggia L., Cocha C., Giasemis F., et al. (2025). Review of machine learning for real-time analysis at the Large Hadron Collider experiments ALICE, ATLAS, CMS and LHCb. *arXiv* preprint. DOI:10.48550/arXiv.2506.14578
2. Zhu Y., Liang H., Wang Y., et al. (2025). Holistic approach and advanced color singlet identification for physics measurements at the high-energy frontier. *arXiv* preprint. DOI:11783.10.48550/arXiv.2506.11783
3. Krause C.G., Faucci Giannelli M., Kasieczka G., et al. (2025). CaloChallenge 2022: A community challenge for fast calorimeter simulation. *arXiv* preprint. DOI:10.48550/arXiv.2410.21611
4. Fan J. and Wang J. (2024). Online reinforcement learning control for maintaining an optimum beam collision on an electron–positron collider. *Phys. Rev. Accel. Beams* **27**:122801. DOI:10.1103/PhysRevAccelBeams.27.122801
5. Harilal A., Park K. and Paulini M. (2025). Anomaly detection based on machine learning for the CMS electromagnetic calorimeter online data-quality monitoring. *EPJ Web Conf.* **320**:00048. DOI:10.1051/epjconf/202532000048

AUTHOR CONTRIBUTIONS

All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.