BNL High Energy Physics (HEP) AI/ML Vision

July 2025 - DRAFT 1

Machine Learning (ML) – a suite of algorithmic techniques that extract patterns and approximate complex functions from data – has become a core component of modern HEP research, powering critical tasks across data taking, simulation, reconstruction, and analysis workflows. Artificial Intelligence (AI), broadly defined as systems exhibiting reasoning and adaptive capabilities, remains an emerging frontier for HEP, offering the potential to augment, optimize, and automate scientific reasoning at scale. This document articulates our vision for how AI and ML can transform the HEP landscape, enabling BNL's HEP program to deliver world-leading discovery science, steward essential U.S. research facilities, and train the next generation of scientific innovators through strategic investment in these technologies.

HEP problems often resemble non-deterministic polynomial time (NP-class) challenges, problems for which solutions are easy to verify but difficult to construct. ML, particularly Deep Learning, offers a pragmatic path forward by approximating solutions through training on representative data, effectively acting as a "cheat sheet" that highlights ML's potential to enhance performance and efficiency in addressing complex problems in HEP that are otherwise difficult to approach with traditional methods. However, in HEP, this "cheat sheet" is typically derived from simulations rather than real data. Because simulations are inherently imperfect, ML models trained on them can fail to generalize when applied to experimental observations. This simulation-to-data gap is a defining challenge in developing trustworthy, physics-informed ML, and calls for approaches that ensure reliable generalization across domains.

At BNL, our immediate focus is on developing ML tools that are robust, scalable, and ready for deployment in physics experiments, transforming promising R&D into reliable production systems. While early-stage research ("0 to 1") is essential for unlocking new methods, production deployment ("1 to 100") is where these methods must prove themselves under real-world constraints (e.g. simulation-to-data gap). Each phase presents distinct challenges: innovation versus integration, but both are indispensable. Our vision embraces this full arc: from conceptual breakthroughs such as the LS4GAN project at BNL, where unpaired generative models (e.g., UVCGAN) are used to refine simulations and quantify data/MC discrepancies, to robust production tools, such as ML models deployed on FPGAs for real-time decision-making with sub-10 μs latency in trigger systems.

In addition, we envision AI systems that integrate learning, planning, and reasoning that move beyond pattern recognition to support scientific thinking. While industry has led the development of large language models (LLMs), our goal is not to invent intelligence, but to discover and harness it for fundamental physics. Scientific workflows demand interpretability, uncertainty quantification, and domain consistency — features often underemphasized in industrial applications. By embedding physicists' reasoning, hypotheses, and judgment into these systems, we can guide AI toward solving challenges like the simulation-to-data gap. At BNL, the highly complex collision environments of RHIC, LHC, HL-LHC, and the upcoming EIC —

together with their massive, high-quality datasets curated across decades — offer an unmatched opportunity to advance this vision. These environments not only challenge traditional computational approaches, but also demand novel physics-guided AI methods to uncover subtle emergent phenomena and deepen our understanding of nature.

Current Status – Building a Foundation for Physics-Guided AI at BNL
BNL's HEP program spans the energy, intensity, cosmic, and theory frontiers—each presenting distinct challenges and unique opportunities for AI/ML innovation. Across these diverse areas, a unifying objective is emerging: to develop AI systems that not only automate, but also reason, adapt, and support scientific discovery with rigor and transparency.

A central pillar of this strategy is addressing the simulation-to-data gap, a long-standing barrier to robust scientific inference. BNL is leading efforts to improve simulation fidelity through domain adaptation techniques such as transfer learning and unpaired generative translation. For instance, the LS4GAN initiative applies models like UVCGAN to ProtoDUNE data, refining simulated outputs and identifying systematic mismatches between Monte Carlo and real detector responses. These techniques help align simulated distributions with experimental reality, mitigating bias and improving downstream analysis.

Complementing this, BNL is advancing hybrid ML + physics algorithms that embed physical principles—such as conservation laws or symmetries—directly into model architectures. Techniques like Graph Constraint Networks and equivariant neural networks ensure outputs are not only accurate but also interpretable and resilient to data/MC discrepancies. To further support uncertainty-aware decision-making, BNL is developing robust strategies based on ensemble learning, Bayesian inference, and auto-differential error propagation. These are being applied to key challenges across HEP, from cosmic shear estimation to SMEFT global fits.

BNL is also at the forefront of scientific-assistive AI, leveraging LLMs to assist with automated quality control in detector construction (e.g., 600k-channel DUNE electronics), intelligent assistance in detector operations (e.g., ATLAS), and symbolic reasoning tasks in theoretical cosmology. The long-term goal is to develop AI systems that integrate learning, reasoning, and planning—tailored for scientific research.

These efforts are further amplified through cross-program collaboration at BNL. For example, a new R&D initiative connects ATLAS and Electron-Ion Collider (EIC) physicists to explore AI-based data intelligence at the detector level. With LHC data rates approaching petabytes per second and storage costs dominating computing budgets, intelligent filtering at the DAQ stage is becoming essential. Drawing on its TDAQ expertise, BNL is applying deep learning to commodity FPGAs, achieving up to 100× reduction in data volume—yielding significant cost savings while preserving physics performance.

While the examples discussed here represent only a portion of BNL's growing efforts in AI, they highlight the emergence of a robust and expanding foundation for physics-guided AI—where data-driven methods are rigorously guided by scientific principles, domain-specific constraints,

and the need for interpretability and reproducibility. This fusion is essential to ensure AI systems contribute meaningfully and reliably to scientific discovery.

Looking ahead, it is absolutely critical that BNL researchers are empowered to fully integrate and leverage cutting-edge AI tools into their research. Researchers require comprehensive, flexible access to state-of-the-art AI models, including UI-based tools, programmatic APIs, and code-generation interfaces, along with institutional encouragement and training to use them effectively. These tools are already proving transformative to scientific productivity; researchers without access to them are increasingly at a competitive disadvantage. To stay at the forefront, BNL must establish agile and sustainable mechanisms to acquire, support, and evolve access to the rapidly shifting AI landscape.

Near-term focus (1–2 years) – Building capacity and quick wins
Our immediate priority is to broaden AI/ML adoption across every HEP program while removing practical barriers to entry. We will

- Deploy LLM-based assistants to streamline common scientific workflows, such as code generation, documentation, and data-quality checks, freeing researchers to focus on higher-level physics insight;
- Package proven ML algorithms (e.g., GNN vertexing, diffusion-based fast simulation) as containerized “recipes” that enable teams to fit their existing data into established network architectures, accelerating integration with current data taking, reconstruction, and analysis frameworks;
- Integrate LLMs into scientific tools and services using interfaces like the Model Context Protocol (MCP), enabling context-aware AI systems with real-time awareness of experimental conditions, improving usability, operational efficiency, and domain adaptability;
- Launch hands-on training sprints so staff and students can choose, tune, and validate networks on the data they know best.

These activities create an on-ramp from traditional workflows to ML-enhanced production, seeding the community and generating early physics pay-offs.

In parallel, we will begin developing and testing early reasoning-capable AI systems tailored to physics. This includes lightweight, domain-informed models—such as “toy” logic engines—for hypothesis generation, validation workflows, and structured interpretation of inputs and outputs. These prototypes will help both scientists and developers build intuition about what scientific reasoning in AI could look like. In parallel, domain-specific reasoning copilots will be deployed to assist with real-world detector operations—flagging anomalies, suggesting calibration strategies, and offering structured explanations of unexpected behavior. As these systems evolve, they will also contribute to a curated corpus of structured physics reasoning, laying both a practical and conceptual foundation for more intelligent, interpretable AI in the future.

Mid-term focus (3–5 years) – Integrating systems & addressing simulation-to-data gap
As foundational efforts take root, the next phase emphasizes growth—of people, systems, and scope. Practitioners who began by fitting data into known models will now design bespoke architectures around domain-specific data needs. Those who pioneered early-stage R&D (“0 to

1") will lead the transition of promising prototypes into robust, scalable tools ("1 to 100"). The goal is to build a self-sustaining ecosystem where model builders, physics analysts, and infrastructure experts co-develop solutions that are both scientifically rigorous and operationally viable. Strategic mentorship, tool standardization, and shared validation pipelines will help accelerate this evolution across HEP programs.

As prototypes mature, emphasis shifts to system integration and scientific robustness aiming to address the simulation-to-data gap challenge. We will scale up R&D lines funded through LDRD, SciDAC, and US-ATLAS computing into facility-grade services:

- **Simulation fidelity & domain adaptation**: LS4GAN-style unpaired translation and diffusion-based surrogates will be extended to DUNE, ATLAS, and RHIC to align Monte Carlo with real data, furnishing uncertainty budgets and improving analysis trust;
- **Hybrid ML + physics algorithms**: Architectures embedding physical priors—symmetries, conservation laws, or analytic constraints (e.g., Graph-Constraint Nets)—will supplant purely empirical models, boosting interpretability and resilience;
- **Reasoning AI:** Refine and expand reasoning-aware AI systems, evolving early copilots into robust scientific collaborators capable of supporting multi-step planning, adaptive diagnostics, and hypothesis-driven analysis across complex environments.

Building on this foundation, we will leverage the growing curated corpus of structured scientific reasoning to fine-tune LLMs that go beyond surface-level assistance—aligning closely with the logic, methodologies, and language of physicists. This shift from task automation to interpretable, domain-grounded AI will be enabled by continuous feedback loops between researchers and AI systems, fostering trust, accelerating discovery, and establishing a new model of co-evolving tools and scientific practice.

To support this vision with agility and scale, BNL must provide researchers with seamless access to shared AI-ready infrastructure—including GPUs, FPGAs, and emerging architectures such as neuromorphic chips. These resources will be tightly integrated with experimental and simulation workflows, enabling fast prototyping, iterative deployment, and real-time responsiveness. The goal is to ensure that innovations in AI—whether from within or outside the Lab—can be rapidly adopted, adapted, and deployed in pursuit of scientific discovery.

Long-term vision (5–10 years) – Toward scientific intelligence at scale
By the end of the decade we aim to establish BNL as the steward of globally important, AI-ready HEP data sets (ATLAS HL-LHC, EIC, DUNE, Belle II) and as a leader in physics-aware intelligence. Integrated systems will couple foundation models, bespoke ML toolkits, and classical algorithms into agentic workflows that can: propose detector settings, carry out fast what-if simulation, assess uncertainty, and even draft first-pass papers. Scientists will provide the high-value training signals—hypothesis trees, validation heuristics, failure cases—that elevate these systems from automated pattern recognizers to partners in discovery. Continuous co-training of models on fresh experimental data and curated reasoning traces will keep the intelligence relevant, adaptive, and trustworthy—enabling faster, more reliable interpretation of results across the entire HEP research portfolio.