Applicant: Arvind Tawker
Mentor: Dr Liv Helen Vage
Project duration: 12 Weeks
Proposed start date: 3rd July 2025

IRIS-HEP Fellowship Project Proposal

Development of Modular ML Pipeline Templates for High Energy Physics Applications

Overview and Motivation

High Energy Physics (HEP) research relies increasingly on machine learning (ML) to manage vast, complex datasets produced by modern experiments. Techniques such as deep neural networks, autoencoders, generative models, and graph-based networks have become essential tools for tasks ranging from signal/background classification and energy regression to anomaly detection and fast simulation. Yet despite many successes in HEP, ML pipelines are typically developed ad hoc, leading to duplicated effort and non-standardized approaches. With the High-Luminosity upgrades expected to yield larger and more detailed datasets, efficient, standardized, and scalable ML and data processing methods will be essential to ensure timely analysis, reproducibility, and broader accessibility across collaborations.

This project proposes developing a comprehensive GitHub repository of ML pipeline templates in which each branch hosts a self-contained pipeline tailored to a common HEP use-case. These pipelines will incorporate best practices (modular code, reproducibility, logging, CI/CD with pre-commit hooks) and tools to read HEP-specific data (ROOT files via Uproot, Awkward Array) and work with modern ML frameworks (PyTorch, PyTorch Lightning). Targeting a broad HEP audience and influenced by successful models such as the CMS ML Knowledge Base and the Exa.TrkX project, the repository will serve as a resource for building reusable, flexible pipelines for classification/regression, unsupervised anomaly detection, generative fast simulation, and graph-based modeling. This will also introduce industry-standard practices—such as configuration management, experiment tracking, and modular code design—to early-career and intermediate computational physicists, while dramatically reducing onboarding time for new members of ML-focused HEP teams by providing ready-to-use, well-documented, and domain-adapted pipeline templates grounded in community best practices.

Objectives

- **Establish a Multi-Branch Repository:**
  Create a central GitHub repository where each branch houses a unique, complete pipeline for a specific HEP use-case:
  - **Supervised Learning Pipeline:** For classification (e.g., signal vs. background) and regression tasks (e.g., momentum estimation).
  - **Unsupervised Learning Pipeline:** Utilizing autoencoders/VAEs for anomaly detection and exploratory data analysis.

    - **Generative Modeling Pipeline:** Employing GANs, VAEs, or normalizing flows for fast simulation of detector responses.
    - **Graph Neural Network Pipeline:** Using GNNs to process structured data (e.g., particle tracking, jet tagging).
    - *(Optionally) A Hyperparameter Optimization module will be integrated where possible.*
- **Standardize Best Practices:**
  By implementing configuration management (via Hydra/OmegaConf), comprehensive logging (W&B, TensorBoard), modular architecture, and coding standards (using pre-commit hooks), the templates will improve reproducibility and ease the transition from prototype to production.
- **Address HEP-Specific Challenges:**
  Each template will include data ingestion from ROOT using Uproot, appropriate preprocessing (e.g., normalization, error propagation), and sample evaluation techniques (ROC curves, confusion matrices, physics-specific metrics). This domain-specific design ensures the pipelines are directly applicable to HEP problems.

Proposed Timeline and Milestones (July–September)

July (weeks 1-3): Foundations and Repository Setup

- **Literature Review & Requirements Gathering:**
  Examine state-of-the-art ML applications in HEP through resources such as the HEP ML Living Review, CMS ML Knowledge Base and meetings/collaborations with HEP researchers.
- **Repository and Architecture Setup:**
  Establish a GitHub repository with a clear modular folder structure and core utilities (data handling, config management). This has already been done on a high-level.
- **Milestone:**
  Launch an initial branch with a supervised learning pipeline (classification/regression) using sample HEP data.

August (weeks 4-7): Pipeline Development

- **Implementation of Unsupervised & Generative Pipelines:**
  Develop an autoencoder-based anomaly detection pipeline and a generative modeling pipeline (using GANs/VAEs for fast simulation).
- **Integration of MLOps Tools:**
  Implement experiment tracking (W&B), CI/CD pipelines, and pre-commit hooks, ensuring reproducibility.
- **Milestone:**
  Three operational pipeline branches demonstrating diverse ML techniques applied to HEP data.

September (weeks 8-12): Graph-Based Pipeline and Project Finalization

- **Graph Neural Network Pipeline:**
  Create and test a graph-based pipeline for tasks such as jet tagging or track reconstruction using tools like PyTorch Geometric.
- **Validation, Documentation, & Outreach:**
  Validate pipelines with real or simulated HEP datasets; refine documentation, and prepare a tutorial and presentation for the fellowship committee.
- **Milestone:**
  Complete the repository with polished pipelines, comprehensive documentation, and demo results ready for community evaluation and further extension.

Key References and Resources

- **HEP ML Living Review:**
  https://iml-wg.github.io/HEPML-LivingReview/
- **CMS ML Knowledge Base:**
  https://cms-ml.github.io/documentation/index.html
- **MLOps Project Structure Guides:**
  https://mlops-guide.github.io/Structure/project_structure/
  https://mlops-guide.github.io/
- **PyTorch & PyTorch Lightning:**
  https://pytorch.org/
  https://lightning.ai/docs/pytorch/stable/
- **Weights & Biases (W&B):**
  https://wandb.ai/site/
- **Pre-commit Framework:**
  https://pre-commit.com/
- **Additional HEP Data Tools:**
  https://uproot.readthedocs.io/en/latest/