Applicant: Deimantė Juknevičiūtė
Mentor: David Lange
Project Duration: 8 weeks
Proposed Start Date: 6 July 2026

IRIS-HEP Fellows Program Project Proposal

Creating a fast analysis format for/from EDM4hep data

Project Description

The Future Circular Collider (FCC) is a proposed large-scale project at CERN that aims to conduct detailed studies of the Higgs boson and its interactions, search for new particles, new forces and possible dark matter candidates. With substantial gains in precision and sensitivity, it would enable measurements beyond the reach of current colliders, including the Large Hadron Collider (LHC) and its high-luminosity upgrade [1].

The FCC community, including studies of the FCC-ee design, uses EDM4hep and associated software to describe the data model for ongoing simulation studies. EDM4hep defines a series of C++ classes representing detector and physics objects corresponding to each aspect of the simulation and event reconstruction pipeline [2]. While important for detailed studies, these data structures are inefficient for data analysis users who instead would benefit from having fast to analyze columnar data. As FCC-ee aims to collect more than a trillion Z-boson decays, speed of analysis codes is particularly important.

The scale of the FCC-ee physics program presents a significant computing challenge. Given the projected collection of more than 10^{12} Z-boson decays, analysis workflows must process very large numbers of events efficiently. Compact columnar formats will improve analysis throughput and reduce the time required to obtain physics results.

Goals

This project aims to build a prototype and assess the performance of a columnar format derived from the EDM4hep format that is appropriate for FCC-ee users. Methods to convert data stored in EDM4hep into compact structures will be investigated. The CMS NanoAOD [3] format will be used as an example for comparison and design. ROOT RNTuple [4] will potentially be used as a storage backend. The Key4hep software ecosystem [5] and FCC-ee Monte Carlo simulation samples will be used to develop and evaluate the proposed format. The resulting format will be evaluated through example Python analysis scripts, validation studies and throughput measurements to assess its suitability for future FCC analyses.

Project Timeline

Weeks 1-2

- Become familiar with the CMS NanoAOD data format using CMS Open Data.
- Study example FCC EDM4hep datasets and the EDM4hep data model.

Weeks 3-4

- Design a Python script for creating an RNTuple-based columnar format from EDM4hep data.
- Test the conversion workflow on representative FCC datasets.

Weeks 5-6

- Create Python analysis examples for the newly created columnar format.
- Perform validation studies and throughput measurements using the developed scripts.

Weeks 7-8

- Document the results and summarize the performance of the prototype.
- Determine possible improvements for future development.

About Me

I am an undergraduate student at Vilnius University, Faculty of Physics. Through my studies, I have developed a strong interest in scientific research, particularly data analysis and modern physics. This project provides me with an opportunity to contribute to the development of scientific software and gain experience working with large datasets.

References

[1] CERN, "Future Circular Collider" [https://home.cern/science/accelerators/future-circular-collider].

[2] F. Gaede et al., "EDM4hep – a common event data model for HEP experiments," PoS ICHEP2022 (2022) 1237 [https://edm4hep.web.cern.ch/].

[3] CMS Collaboration, "CMS NanoAOD Format" [https://opendata.cern.ch/docs/cms-getting-started-nanoaod].

[4] ROOT Collaboration, "ROOT Data Analysis Framework" [https://root.cern.ch].

[5] Key4hep Collaboration, "Key4hep Software Ecosystem" [https://key4hep.github.io].