IRIS-HEP Fellowship Proposal
Maxym Naumchyk

Duration: July 1st to September 1st, 2022

Integrating ML algorithms for LHC data compression into the ESCAPE Virtual Research Environment

1. Context for this project

About 85% of the mass in the Universe consists of dark matter. However, we can't detect its existence by simple means, as it doesn't reflect light, and it can't be affected by magnetism or electricity. The only observational evidence of its existence can be seen through the gravitational pull it exerts on the normal mass that makes up the objects we can see, like stars and planets—and us. Dark matter can be produced in collisions of ordinary matter and observed in experiments at colliders and at particle accelerators, as well as in astrophysics experiments. For the most effective use of the information from all these different experiments, their data and workflows can be gathered within one platform. This is why a Dark Matter Test Science Project is being developed as a part of the ESCAPE project.

The goal of this Test Science Project is to gather dark matter analyses and results coming from the different fields, like experiments directly or indirectly detecting dark matter in the universe, particle colliders that can create dark matter and relevant theoretical models for the interpretation of results, within the same framework. This way, we will be able to produce dark matter summary plots with FAIR workflows and have a more complete picture of this subject. In doing this, we will contribute to the design of a prototype for the European Open Science Cloud using ESCAPE tools where we can store, distribute and provide data and analysis code to the scientific community all over the world.

2. Proposal

At Large Hadron Collider (LHC) detectors, there are approximately 1 billion particle collisions per second which are generating about one petabyte of data each second. Hence, storage is one of the main limiting factors to the recording of information from these collisions. To compress this data and improve the efficiency of the experiments, the innovative machine learning algorithms can be used.

As a part of my project, I will integrate one of the current state-of-the-art ML algorithms designed specifically for LHC data compression into the ESCAPE Virtual Research Environment as a part of the European Open Science Cloud.

It is worth mentioning that the Test Science Project is only a part of many projects within the European Open Science Cloud, and my work will open the possibility of extending the use of this algorithm to other experiments and fields.

3. Timeline

June 20th - June 24th: project proposal, reading "outreach" material and bachelor's thesis.
June 24th - July 1st: running ML algorithms on laptop and understanding what they do.
July 1st - July 13th: running ML algorithms on CERN resources.
July 13th - July 21st: learning and starting to work on the EOSC resources.
July 21st-22nd: CERN meeting on European Open Science Cloud (virtual), 5 minutes presentation on the plan and achievements.
July 25th - August 12th: running ML algorithms on EOSC resources.
August 12th - August 25th: wrap up (write report, documentation).
August 25th: end of project.

Also to be included:

- At the end of the project: IRIS-HEP fellow presentations
- September 2022: virtual EOSC meeting to present completed project

Used links

[1] "The Nature of Dark Matter," (2022) [https://www.lsst.org/science/dark-matter].
[2] Elena Cuoco, Caterina Doglioni, Kay Graf, Giovanni Lamanna and Samuel Meehan, "The ESCAPE Dark Matter Test Science Project," (2020) [https://pos.sissa.it/392/029/pdf].
[3] George Dialektakis, "Deep Autoencoders for ATLAS Data Compression," Google Summer of Code 2021 Project [https://zenodo.org/record/5482611#.YrCazi1Q1p-].