An extensible, experiment-agnostic file format to facilitate educational access to HEP datasets

Matt Bellis[1]

[1]Siena College, Loudonville, NY

September 1, 2020

There is few things we love more than engaging young people in our research. This can take the form of public talks, or high school visits, or for those lucky enough to be near a national lab, tours of our research facilities. To take high school students or undergrads or the general public on a deeper dive into work however, usually involves a multi-day workshop or a few weeks set aside in a class. In these situations students will either interact with event displays from the experiments or real data in the form of ROOT files and pre-written scripts or highly curated datasets that reduce the data to something that can work in a spreadsheet. These cases all have their place, however it cuts out the students who have a basic understanding of programming, but maybe not the details of the experiment's particular ROOT data structure.

With the growth of python as an introductory programming language at both the high school and undergraduate level, there are more and more students who can perform simple analysis, once they can actually access the data. For the past six (6) years, I've maintained a website, the Particle Physics Playground[1] that provides simplified data and starter scripts that allow the users to play with small datasets from multiple experiments. The data was originally stored in a very structured text file that was quite fragile and did not allow for much extension. Two (2) years ago, my students and I moved the data to a homegrown format that is is built on top of HDF5, `h5hep`[2]. This format and the library to access it can return individual events with any structure as standard python dictionaries, greatly reducing the intellectual overhead required of students. We have used this data format to store data from CMS, CLEO, BaBar, and most recently IceCube.

We propose adding some additional functionality to `h5hep`, improving the documentation, and putting forth as an option for the HEP community for eductional and outreach efforts. We have demonstrated its use with Colab notebooks, which lowers the threshold for middle and high school students with Chromebooks. There are more formal speed/efficiency tests to be performed as

[1] https://particle-physics-playground.github.io/
[2] https://github.com/mattbellis/h5hep

well as simply testing it out with other experiments' datasets. If the community can converge around an open, supported, non-ROOT file format, it would make sharing educational resources that much easier.