DIANA Fellowship Proposal: Finalizing H5HEP, a ROOT-like file format based on HDF5, targeted toward intermediate stages of big-data analysis and outreach

H5HEP (Hdf5 with Heterogeneous Entries in Parallel)[1] is a file format built on HDF5 for use with and inspired by High Energy Physics datasets and the ROOT analysis toolkit. The package is entirely pythonic, though the final data files are pure HDF5, meaning another language could easily interface with the data. H5HEP uses `h5py` and defines a file structure that allows the user to either a) extract "columns" of data or b) loop over "events", similar to how a HEP analyst would use the ROOT `TTree` object. Reading and writing of these files is supported. It is currently being used for a multi-experiment outreach effort[2], as well as for intermediate stages of a current CMS analysis. This file format has the potential to offer another alternative to ROOT files, while still retaining most of the functionality of those files.

Development on H5HEP begin in earnest in Summer 2017, but there are a number of lingering features that should be added before a wide-spread release. A proof-of-principle interface to *awkward arrays* has already been implemented. A more complete version will be tested and deployed as part of this project. Additional feature requests will also be implemented, including functions to embed text documentation in to the files themselves. The suite of unit tests and examples will be extended and formalized performance tests will be developed. The end goal of this project is to submit a paper to *JOSS: The Journal of Open Source Software*[3].

The undergraduate student will work under the supervision of a physicist, the primary developer of H5HEP. Candidates should have a working knowledge of python and git/Github. Experience with Github workflows, continuous integration (CI), and software documentation is a plus, but not required. Familiarity with large datasets, HEP or otherwise, is also not required but considered favorably. The anticipated duration of the project is the three month period May - July, 2021, although there is some flexibility related to the exact start and finish dates.

Matt Bellis (CMS, Siena College) will supervise the student. A timeline with deliverables is provided on the next page.

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

Timeline

- **Weeks 1-2**. Develop a familiarity with the current code base, including unit tests and state of the documentation.
- **Weeks 3-4**. Add functionality for meta information at the file level and individual parameter level. Develop unit tests and add to documentation.
- **Weeks 5-6**. Improve functionality to interface with *awkward arrays*, both reading and writing. Develop unit tests and add to documentation.
- **Weeks 7-9**. Expand test cases and examples, including astronomical and the bioinformatics FASTA format. Develop unit tests and add to documentation.
- **Weeks 10-11**. Develop and run profiling tests to compare to other file formats. Add to documentation.
- **Weeks 12-13**. Finalize JOSS paper and submit.

At the end of the project, the student will present their work at an IRIS-HEP topical meeting and depending on the resources of their home institution, have the opportunity to present at relevant national or international meetings and workshops.