Semantic data handling for Monte-Carlo Event Generators

Jacan Chaplais

University of Southampton

12th September, 2023: PyHEP23

Motivations of a selfish PhD student

- Breaking into HEP pheno ⇒ overwhelming number of data formats
- Some semantic libraries exist, *eg.* `Particle` and `Vector` from `scikit-hep`
  - Don't unify whole event record, though
- Want to leverage knowledge of Numpy and Pythonic idioms
- Data driven understanding:
  - Query data over multiple particle attributes
  - Manipulate and easily visualise event record, including ancestry data
  - Save and read from disk in standardised high performance files
- Want code to be shared between all my scripts: publish as packages

Solution

- Packages should be standalone, but with high interoperability
  - Routines should take in and output objects with generic properties, *ie.* utilising `typing` and `collections.abc` from standard library

A package was created for each of the following tasks

1. Interfacing data generators with Python generator and NumPy (`showerpipe`)
2. Manipulating event record as heterogeneous data structure (graphicle)
3. Forming interactive visualisations of data structures (colliderscope)
4. Writing data to disk via HDF5 files (heparchy)

Features, caveats, and disclaimers

Features

- Efficient: mostly NumPy, SciPy, and Numba under the hood
- Pythonic: data structures and interfaces implement Python's data model

Caveat and disclaimer

- Event-by-event processing may be slightly less efficient
- Written almost exclusively by one student - may have rough edges!
  - Collaborators, GitHub issues, and pull requests are very welcome!