Scalable Neural Network Models and Terascale Datasets for Particle Flow Reconstruction

One of the main approaches for event reconstruction at the Large Hadron Collider (LHC) currently relies on particle flow (PF), which combines hits across subdetectors, considering the full event to reconstruct all stable particles in the event. Given the planned High-Luminosity (HL) LHC program, as well as possible future experimental programs of e.g., the Future Circular Collider (FCC), computationally efficient and physically optimal evolutions of the PF-based event reconstruction need to be developed and tested.

Among various approaches, there has been considerable interest and development of Machine Learning (ML)-based reconstruction methods, including for full-event reconstruction. To support rapid progress of such approaches, it is beneficial to establish open datasets with sufficient realism and granularity for testing various types of approaches.

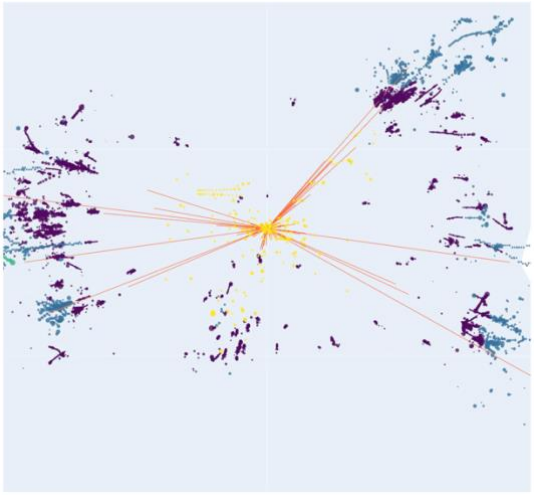

Figure 1: 3D visualization of the generator particles (targets) and the calorimeter hits in a single event.

In light of this, we describe, and make available, an extensive open dataset of physics events with full GEANT4 simulation, suitable for PF reconstruction, available in the EDM4HEP[1] format.

We generate dedicated events with Pythia8[2] and carry out a full detector simulation with GEANT4 using the Key4HEP framework[3]. In particular, we use the CLIC detector model[4], along with the Marlin reconstruction code[5], and the Pandora[6,7,8] package for a baseline particle flow implementation. Although the implementation is not specific to the detector model, the CLIC model is chosen since, to our knowledge, it is one of the most complete publicly available realistic detector models.

The datasets with all generator particles (training targets); reconstructed tracks, calorimeter hits and clusters (training inputs); as well as reconstructed particles from the baseline Pandora algorithm (for comparison) are saved in the EDM4HEP format. In addition, all associations between the aforementioned objects are saved in the standard format. Overall, the size of the dataset is approximately 2.5 TB.

This dataset is being used in studies of the Machine-Learned Particle-Flow (MLPF) algorithm[9,10,11] and new results are being prepared for publication in the near future. Any works using this dataset should cite the corresponding paper, once published.

---

[1] Frank Gaede et al. “EDM4hep and podio-The event data model of the Key4hep project and its implementation”. In: EPJ Web of Conferences. Vol. 251. EDP Sciences. 2021, p. 03026

[2] Christian Bierlich et al. “A comprehensive guide to the physics and usage of PYTHIA 8.3”. In: SciPost Physics Codebases (2022). DOI: 10.21468/SciPostPhysCodeb.8

[3] Gerardo Ganis, Clément Helsens, and Valentin Völkl. “Key4hep, a framework for future HEP experiments and its use in FCC”. In: The European Physical Journal Plus 137.1 (2022), p. 149.

[4] CLIC Collaboration. CLICdet: The post-CDR CLIC detector model. CLICdp note. 2017

[5] F. Gaede. “Marlin and LCCD: Software tools for the ILC”. In: Nucl. Instrum. Meth. A 559 (2006). Ed. by J. Blumlein et al., p. 177. DOI: 10.1016/j.nima.2005.11.138

[6] J. S. Marshall and M. A. Thomson. “The Pandora software development kit for particle flow calorimetry”. In: J. Phys. Conf. Ser. 396 (2012). Ed. by Michael Ernst et al., p. 022034. DOI: 10.1088/1742-6596/396/2/022034

[7] J. S. Marshall, A. Münnich, and M. A. Thomson. “Performance of particle flow calorimetry at CLIC”. In: Nucl. Instrum. Meth. A 700 (2013), p. 153. DOI: 10.1016/j.nima.2012.10.038

[8] J. S. Marshall and M. A. Thomson. “The Pandora software development kit for pattern recognition”. In: Eur. Phys. J. C 75 (2015), p. 439. DOI: 10.1140/epjc/s10052-015-3659-3

[9] Joosep Pata et al. “MLPF: Efficient machine-learned particle-flow reconstruction using graph neural networks”. In: Eur. Phys. J. C 81.5 (2021), p. 381. DOI: 10.1140/epjc/s10052-021-09158-w

[10] Joosep Pata et al. “Machine Learning for Particle Flow Reconstruction at CMS”. In: vol. 2438. 1. 2023, p. 012100. DOI: 10.1088/1742-6596/2438/1/012100

[11] E. Wulff, M. Girone, J. Pata, ” Hyperparameter optimization of data-driven AI models on HPC systems”. In: Journal of Physics: Conference Series 2438, 012092 (2023) DOI: 10.1088/1742-6596/2438/1/012092

The dataset consists of physical collision events as well as particle gun samples and is packaged in 43 tar archives with the naming convention <process_name>_<number>.tar for the physical samples and <process_name>.tar for the gun samples, where <process_name> refers to the name of the physics process and <number> is a running integer. Each tar archive contains ROOT[12] files where the physics events are saved in the EDM4HEP format. To process the data for ML tasks, the Python package uproot[13], which allows for convenient data loading of ROOT files into Python and NumPy objects, is recommended.

[12] Rene Brun and Fons Rademakers, ROOT - An Object Oriented Data Analysis Framework,
Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996,
Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86.
See also "ROOT" [software], Release v6.18/02, 16/06/2020

[13] Jim Pivarski, Pratyush Das, Chris Burr, Dmitri Smirnov, Matthew Feickert, Tamas Gal, Luke Kreczko, Nicholas Smith, Noah Biederbeck, Oksana Shadura, Mason Proffitt, benkrikler, Hans Dembinski, Henry Schreiner, Jonas Rembser, Marcel R., Chao Gu, Edoardo, Eduardo Rodrigues, … bfis. (2021). scikit-hep/uproot3: 3.14.4 (3.14.4). Zenodo. https://doi.org/10.5281/zenodo.4537826