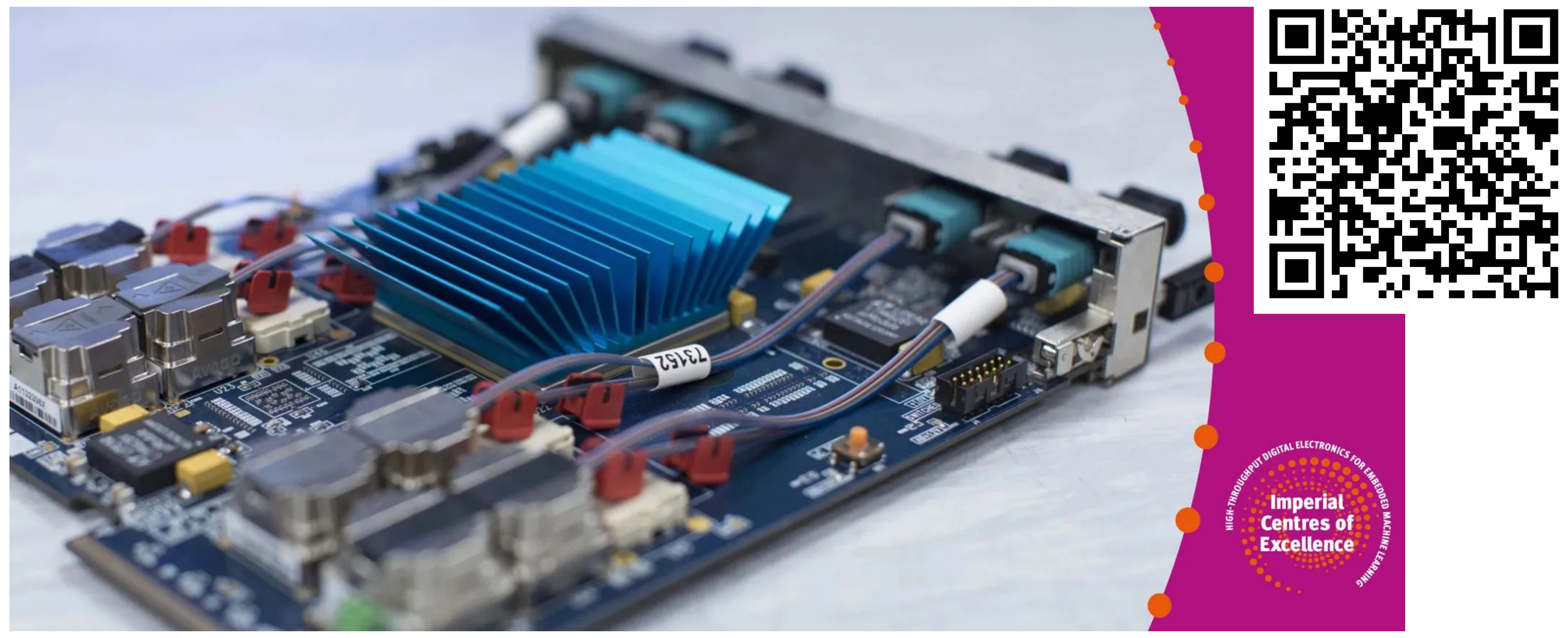
Centre for High-Throughput Digital Electronics and Embedded Machine Learning

An interdisciplinary collaboration focused on algorithms, programming and dataflow applications for high-throughput custom computing systems

The centre is focused on sharing and promoting expertise in algorithms, programming and dataflow applications for high-throughput custom computing systems. It hosts researchers studying the development and deployment of hardware, software and firmware, from a range of disciplines.

Our principal objectives are:

- Application of machine learning (ML) techniques on FPGA and other embedded platforms ("AI on chip") within both the scientific and commercial sectors.
- Development of postgraduate training in firmware and machine learning techniques


Imperial Centres of Excellence
High-Throughput Digital Electronics and Embedded Machine Learning

Case study: Embedded ML for discovery in particle physics

The LHC at CERN is already one of the most extreme data-processing environments on the planet, with latencies measured in μs and bandwidths measured in tens hundreds of Tbps. With each upgrade of the LHC at CERN, the physics reach becomes greater, and the data-processing challenges larger.

At the High-Luminosity LHC (c. 2030), it will be necessary to deploy ML techniques in FPGAs to find signatures of interesting physics hidden under ever more background.

For example: We have <10μs to form ~250 tracks from ~15,000 data points, distinguish genuine tracks from "fakes", associate tracks with an origin, associate tracks with data from other detectors, and make a decision about whether to keep or discard the data for this event.

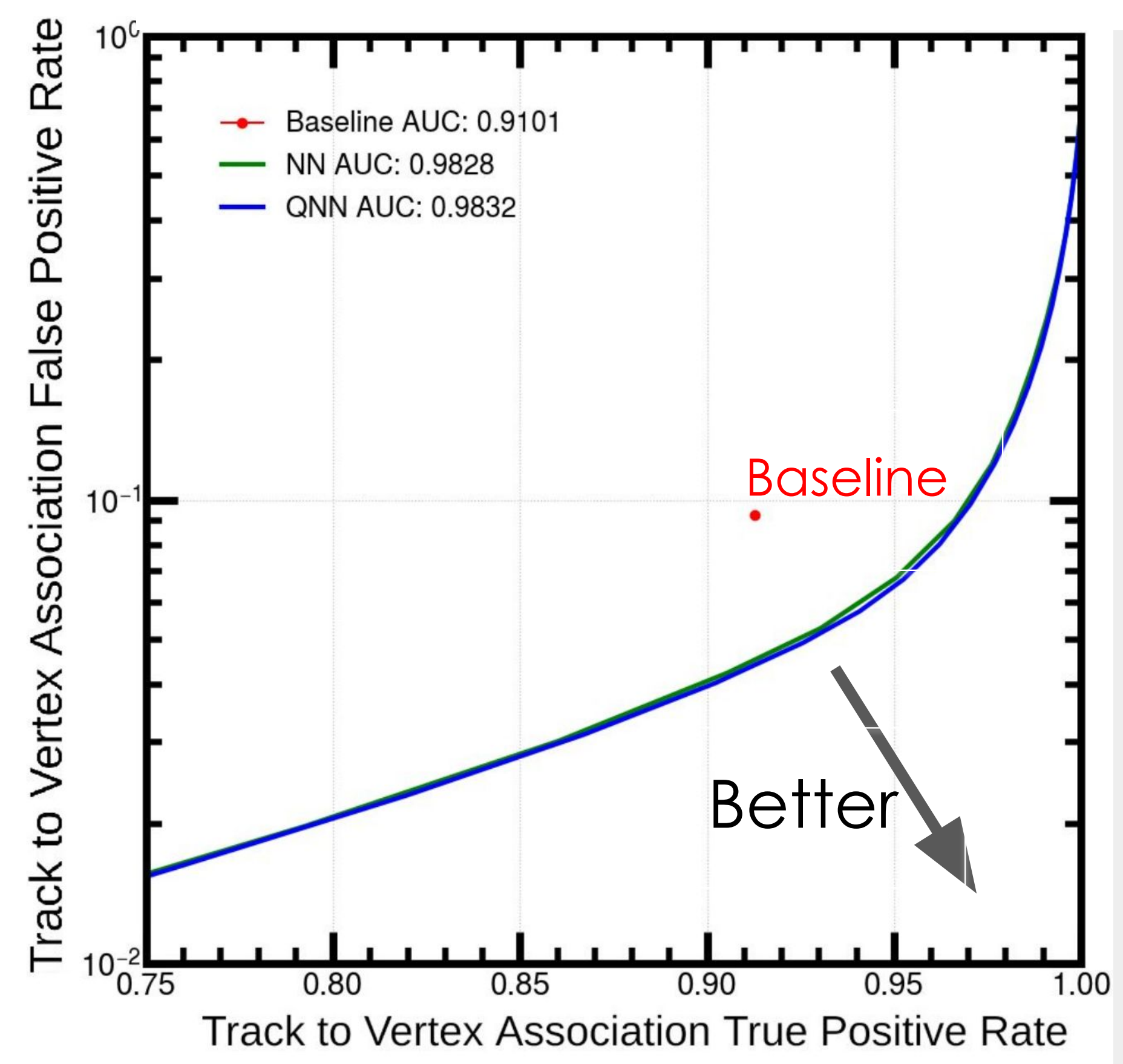
Mixture of classical algorithms, such as Kalman Filtering, and ML approaches: BDTs, CNNs, DNNs. The Centre is currently also exploring GNNs.

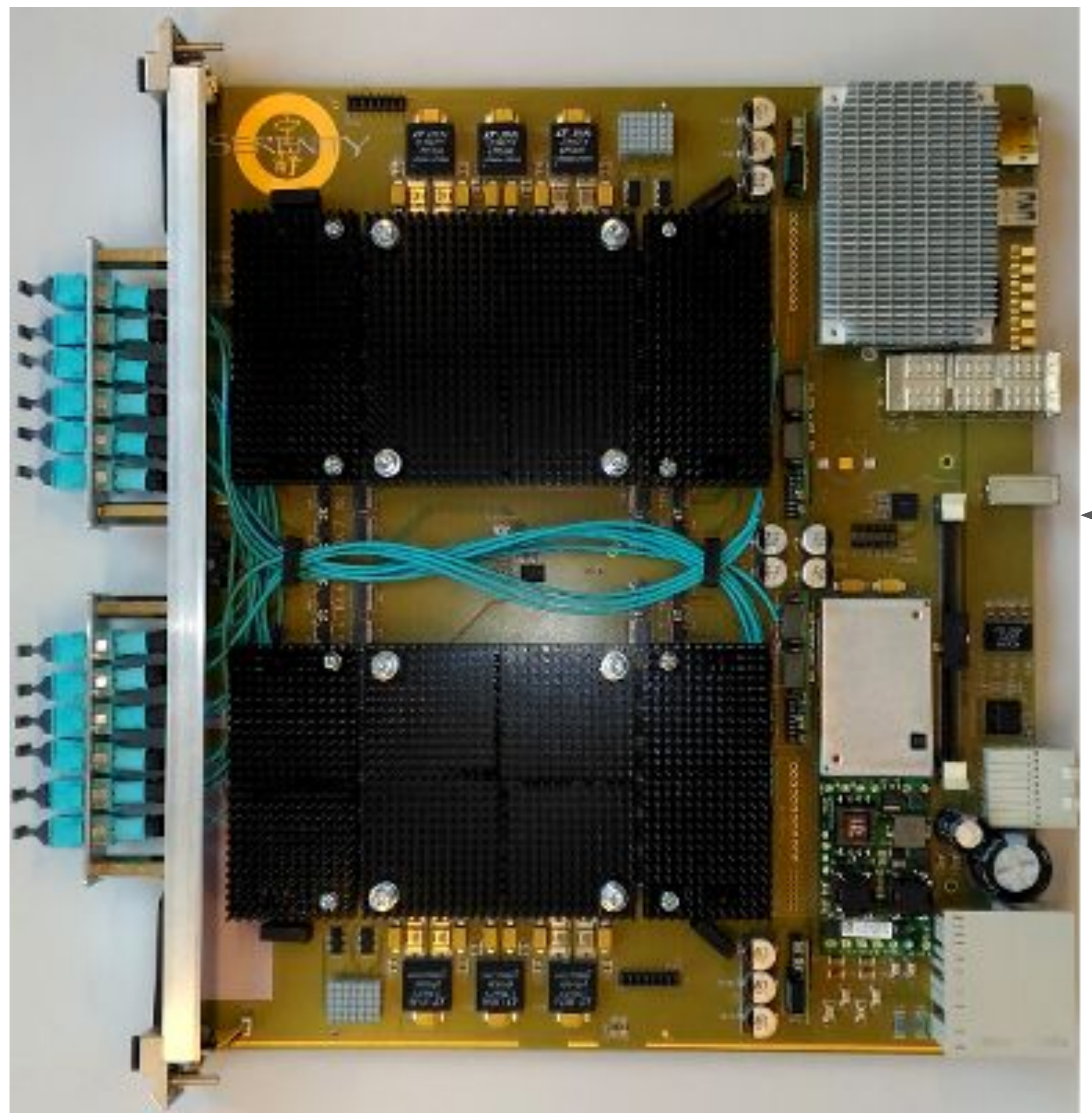
HL-LHC: At-a-glance

- **1 Higgs event in 10^{10}**
- **Up to 250 background events obscuring any interesting event**
- **New data every 25ns**
- **Processing time: <10μs**
- **Bandwidth: 300Tbps**

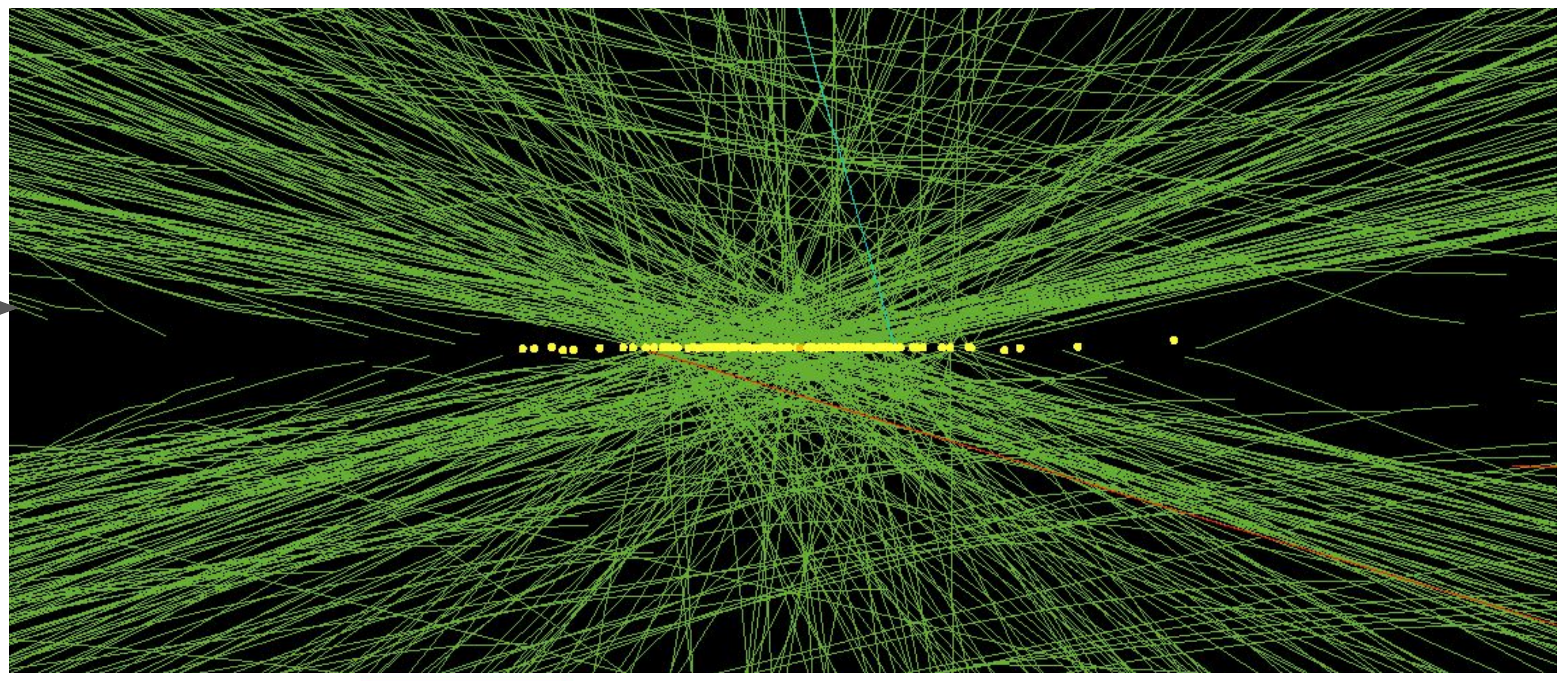
A study into adapting a classical algorithm for locating the primary vertex to use machine-learning approaches, and then optimizing it for FPGA implementation:

- **1000 parameter network**
- **End-to-end latency: 108ns**
- **55% less misidentification**
- **Improved energy and position resolution**
- **Track & vertices now have quantifiable likelihoods**





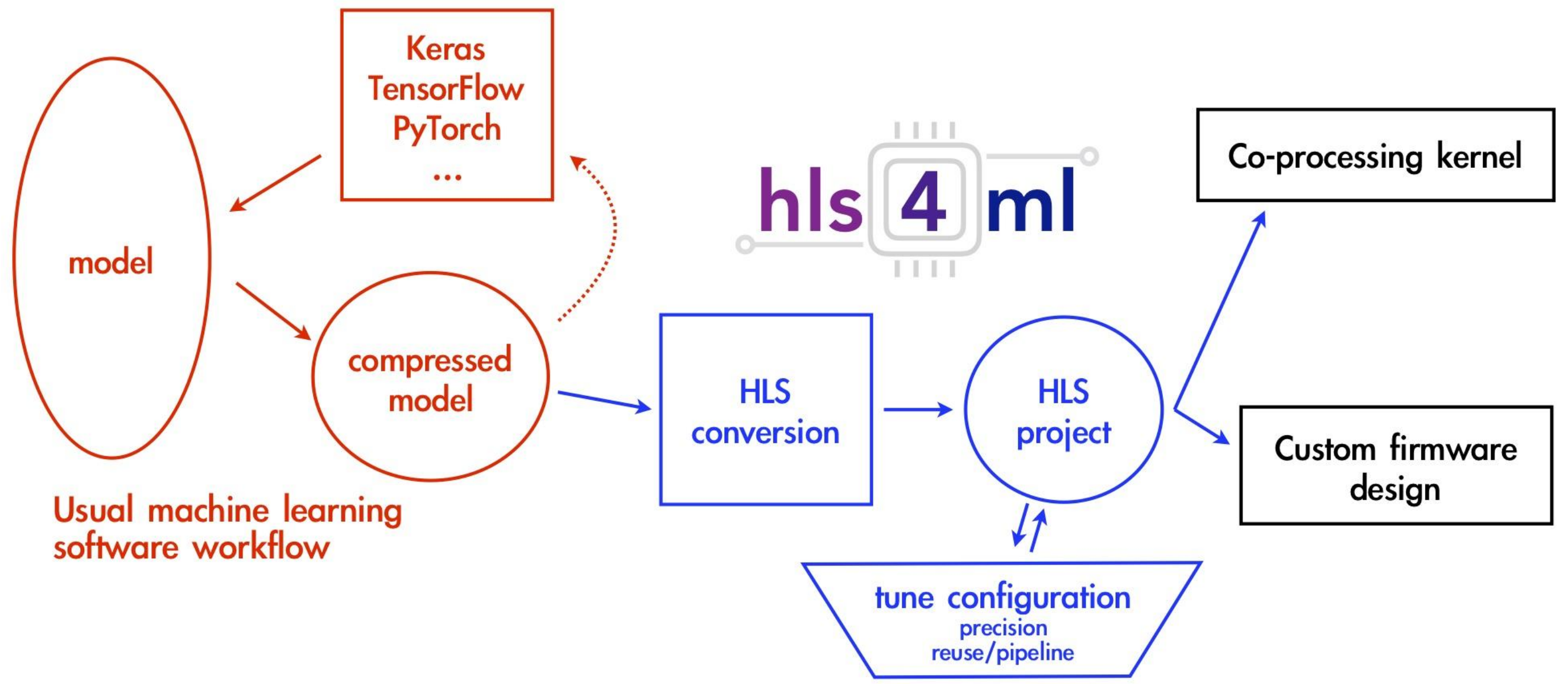
A simulated HL-LHC collision with 130 Vertices (yellow) - Can you tell if any of them contains "interesting" physics?

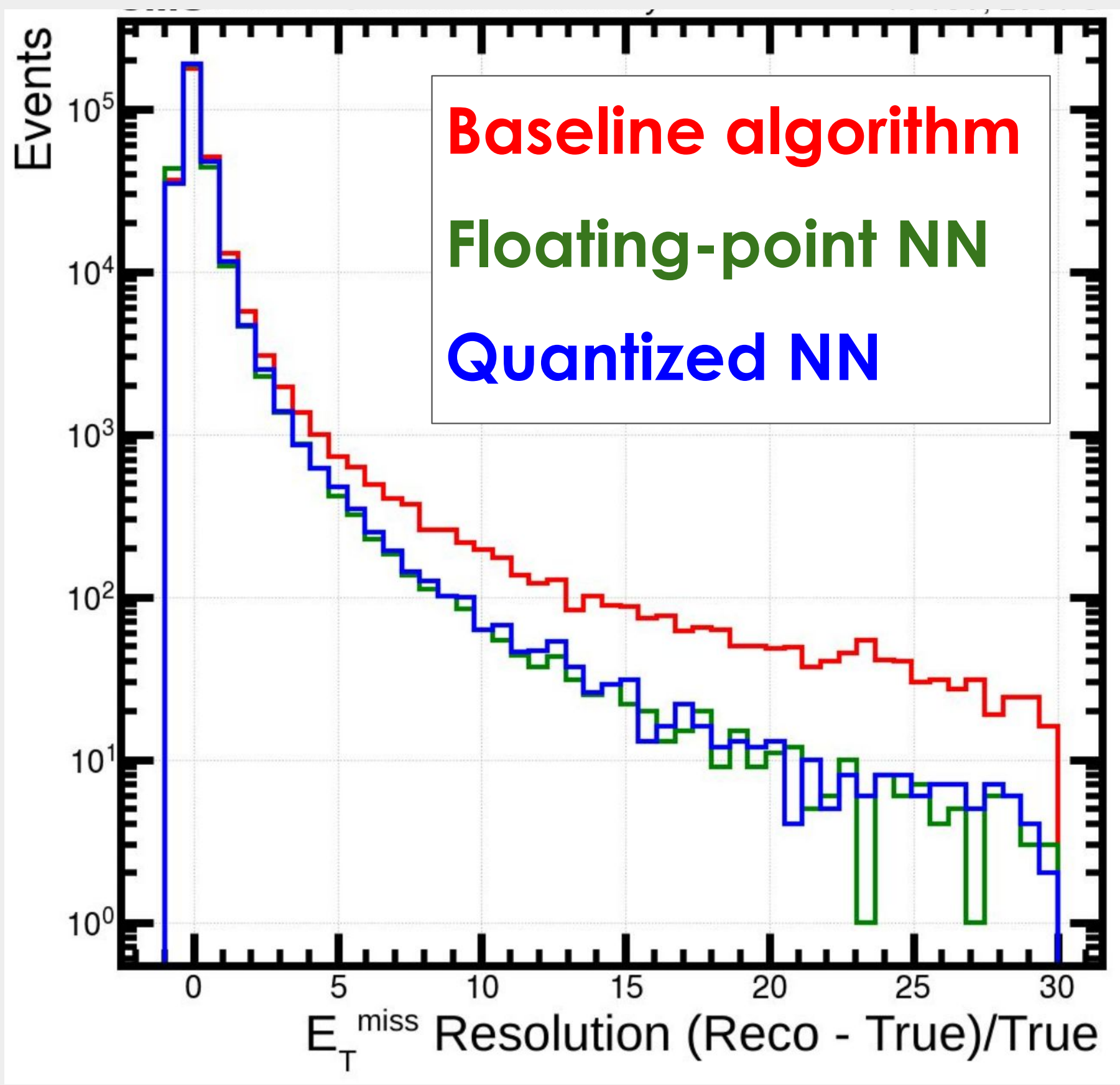


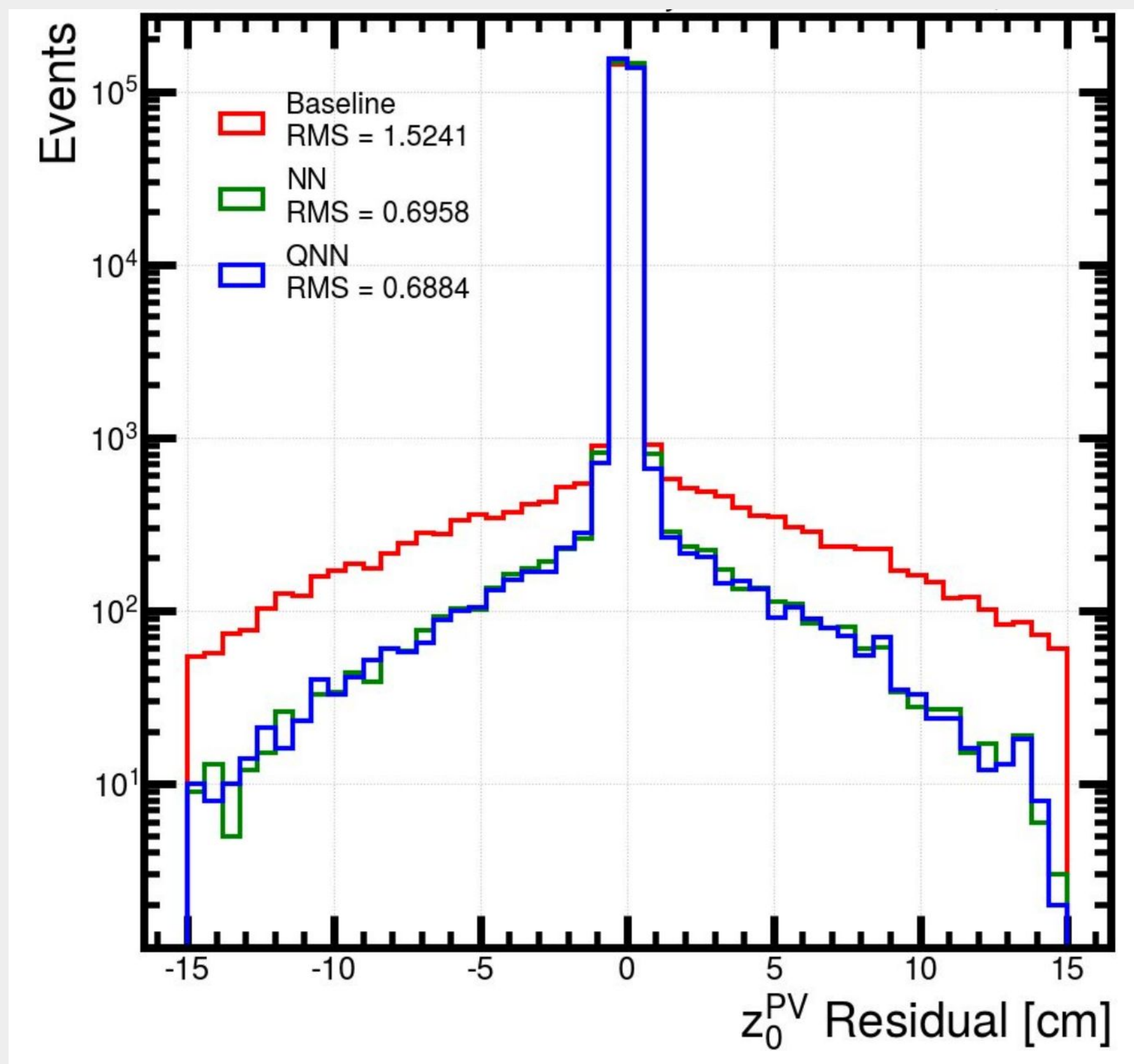
Serenity: A 7Tbps datastream processor developed at Imperial College London

The HLS4ML toolkit adapts CPU-trained models for use in FPGAs, including pruning and the use of quantized weights to optimize networks for the logic resources available in the chip.

https://fastmachinelearning.org/hls4ml/







Andrew W. Rose | awr01@imperial.ac.uk | 05/2022