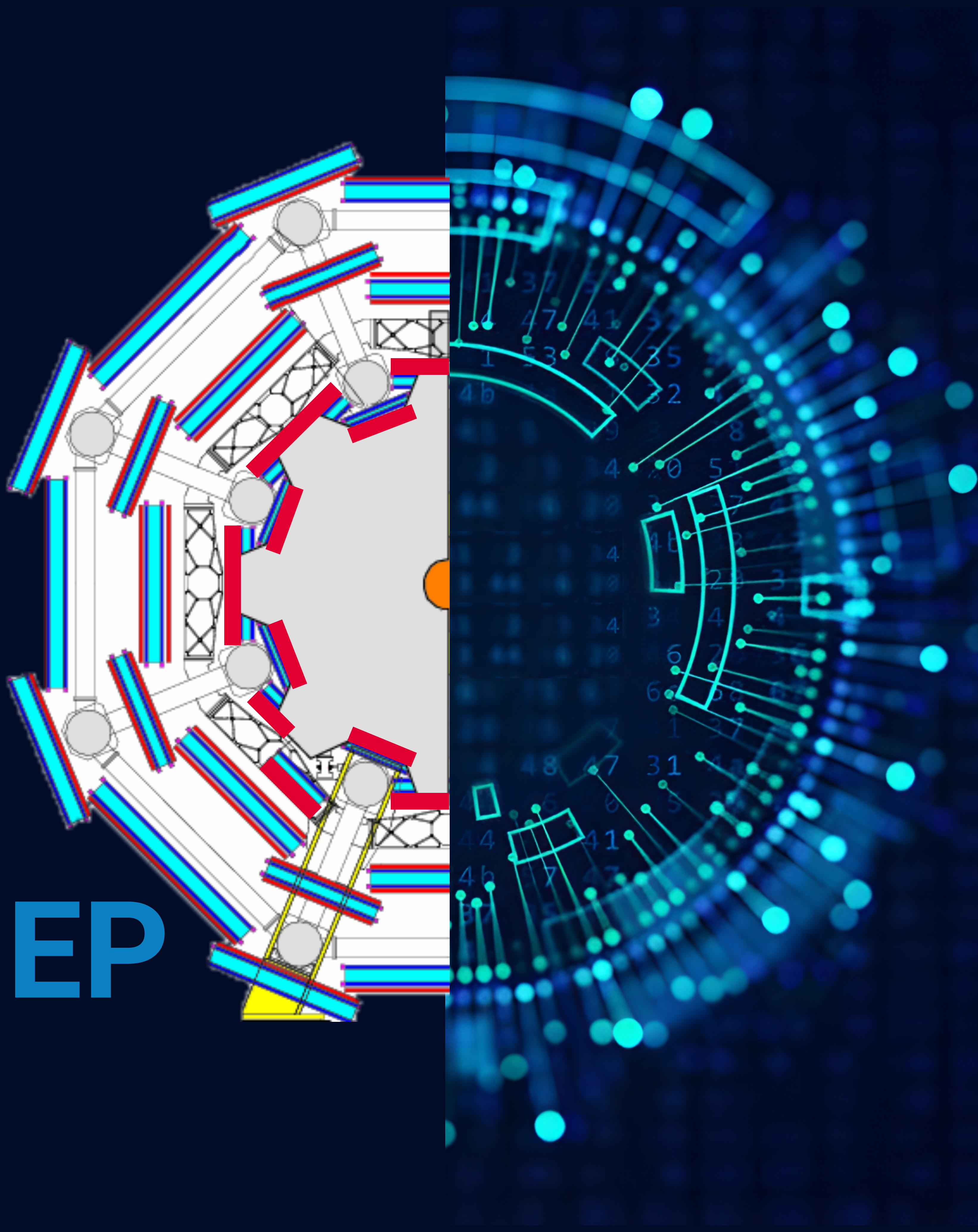
Model compression and simplification pipelines for fast and explainable deep neural network inference in FPGAs in HEP



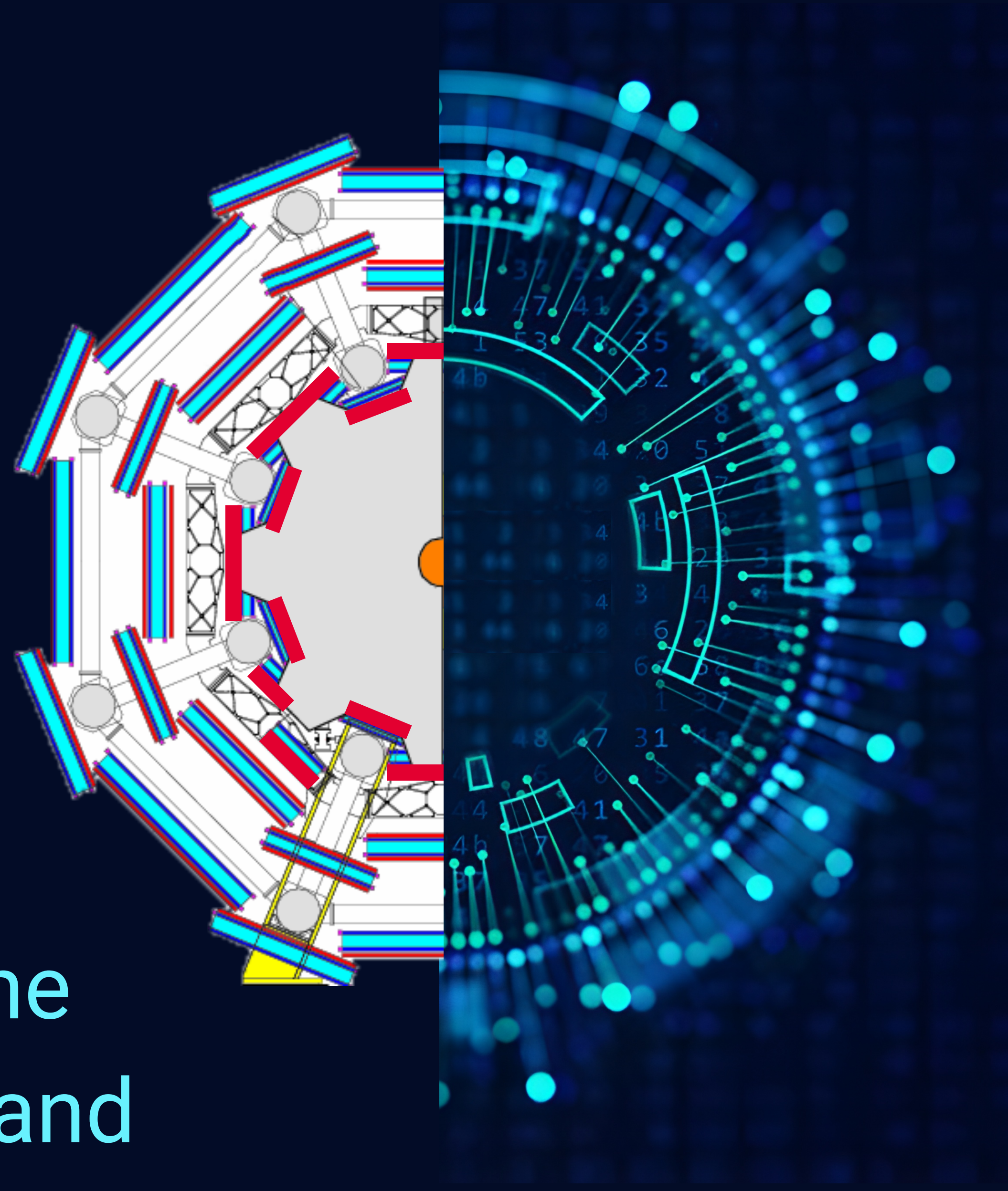
Stefano Giagu, **Graziella Russo**

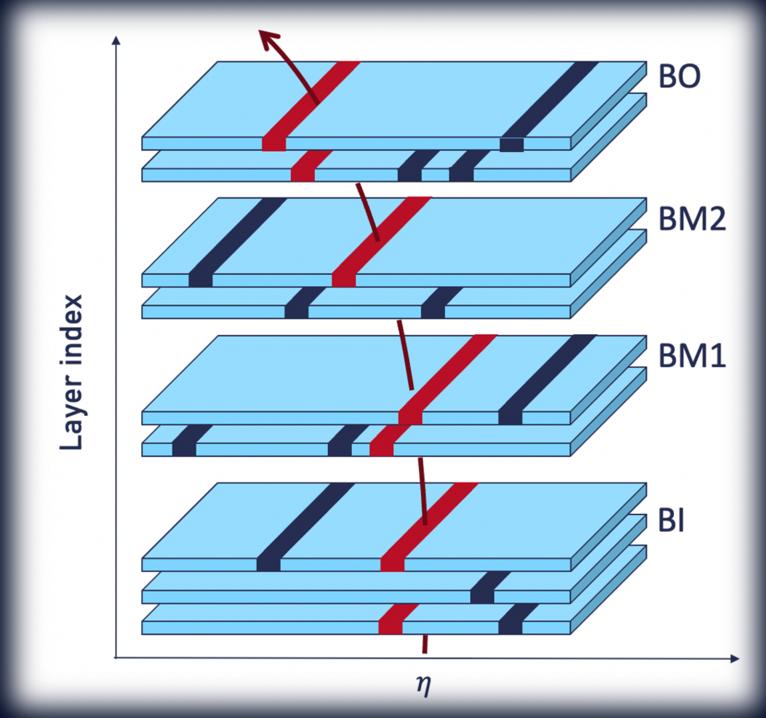
SAPIENZA UNIVERSITÀ DI ROMA

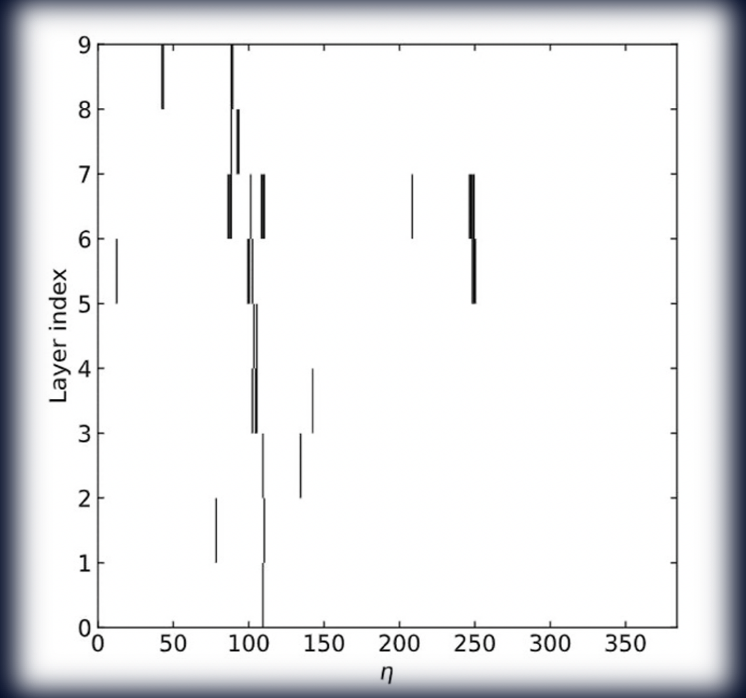
EuCAIFCon24 - 30/04/24

INFN

HiLumi upgrade (2026-29) → x5 LHC instantaneous luminosity → upgrade also in the ATLAS Muon Spectrometer







ML for trigger pattern recognition

Muon tracks as black-and-white 9x384 or 4x384 images, input for **CNN with around 1k parameters** that predicts the transverse momentum p_T, pseudo rapidity η, the charge and the number of muons (up to 3)

Challenges

- Fit within the XCV13P FPGA resources
- Maximum latency $\sim 400\ ns$
- Fake efficiency (= trigger efficiency on noisy events) < 2 ‰

Compression Techniques

First muon p_T efficiency

- **Quantization aware training** (QAT) with QKeras
- **Knowledge Distillation** (KD)

Results, ***Explainability studies*** and FPGA synthesis… on the poster board 51

Graziella Russo

EuCAIFCon24 - 30/04/24