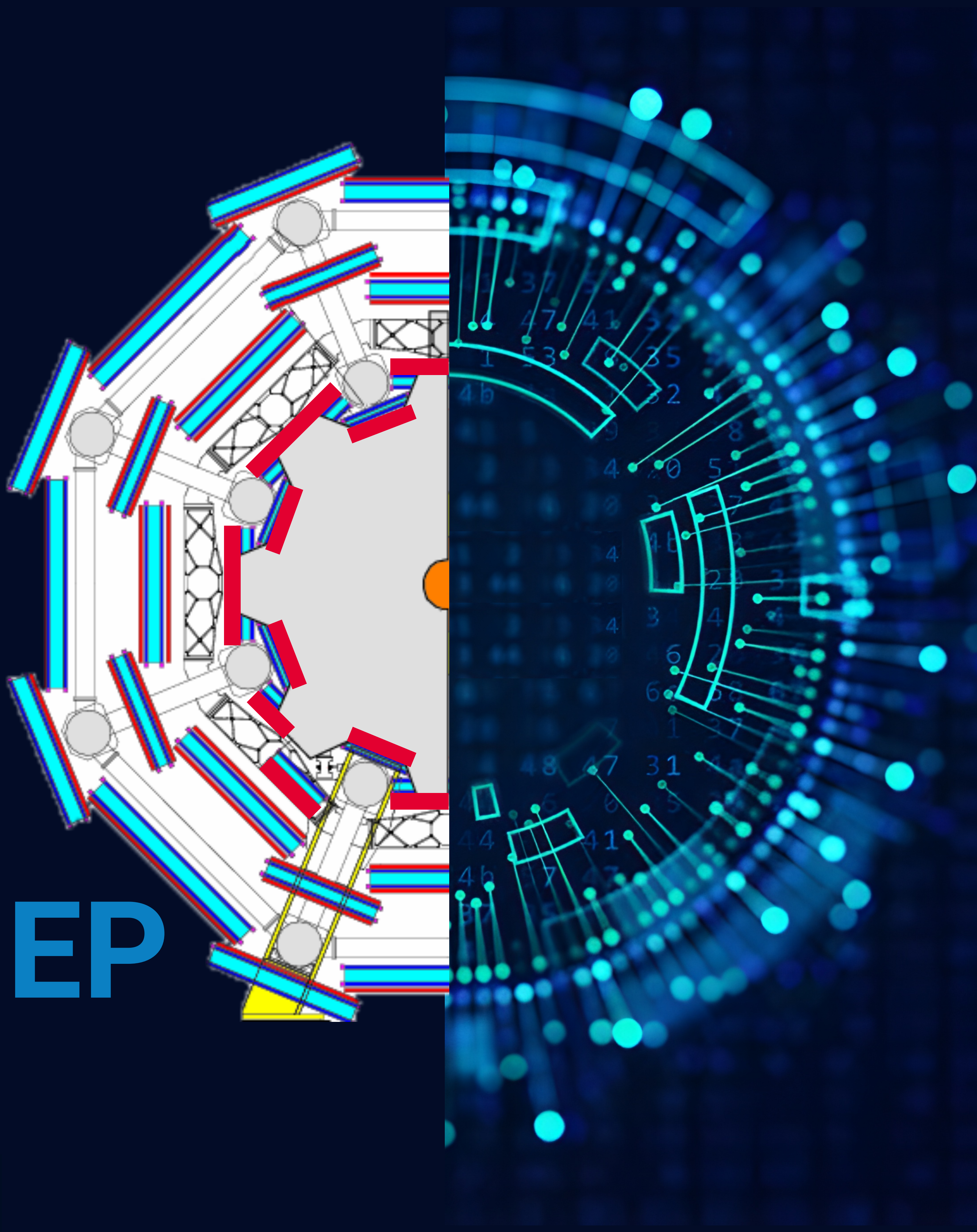
Model compression and simplification pipelines for fast and explainable deep neural network inference in FPGAs in HEP



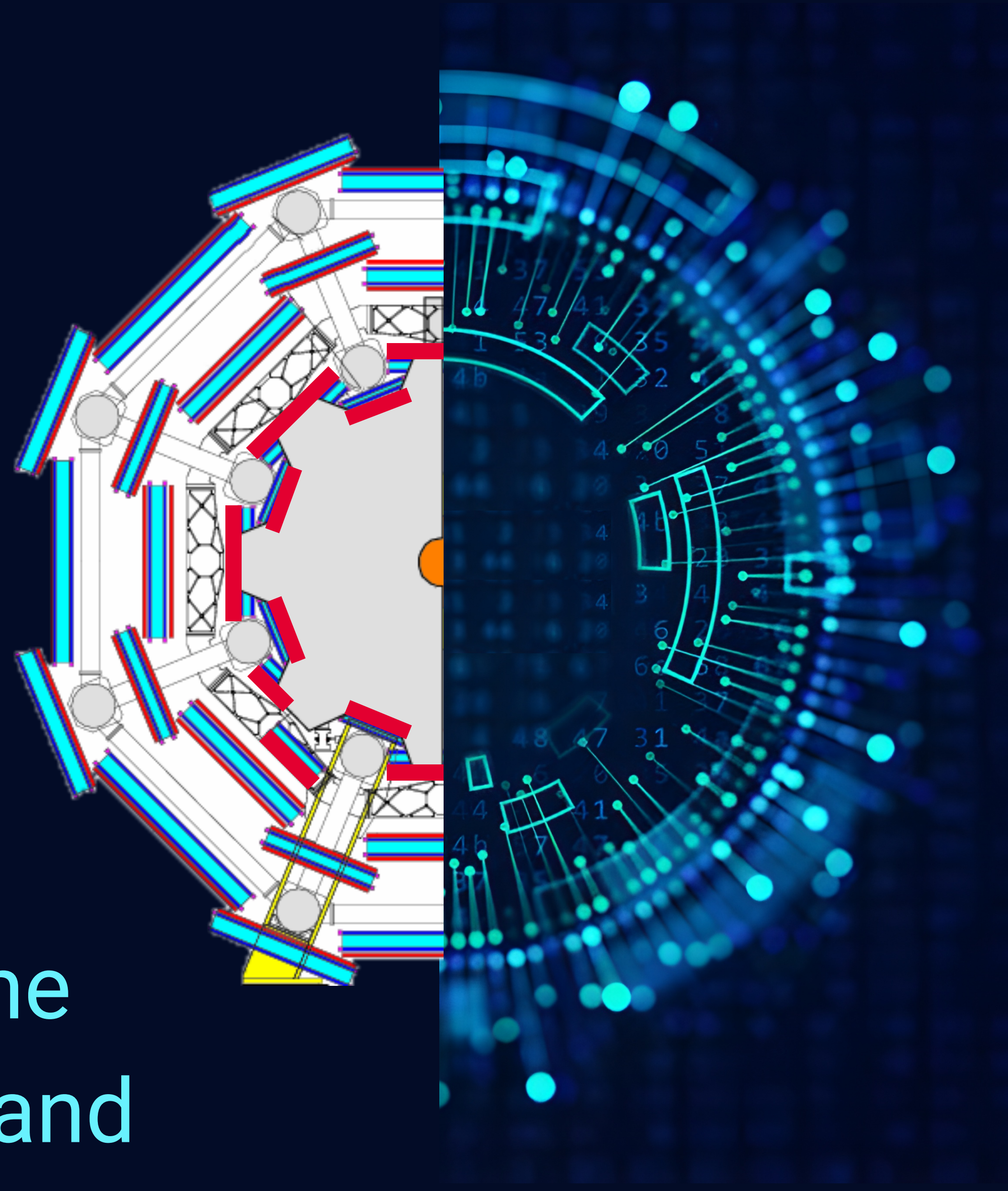
Stefano Giagu, **Graziella Russo**

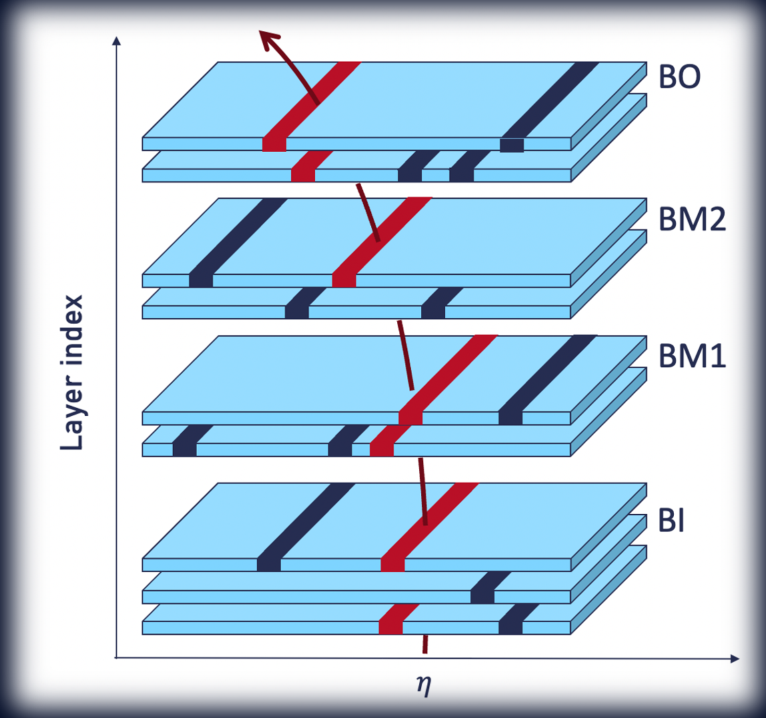
Sapienza Università di Roma

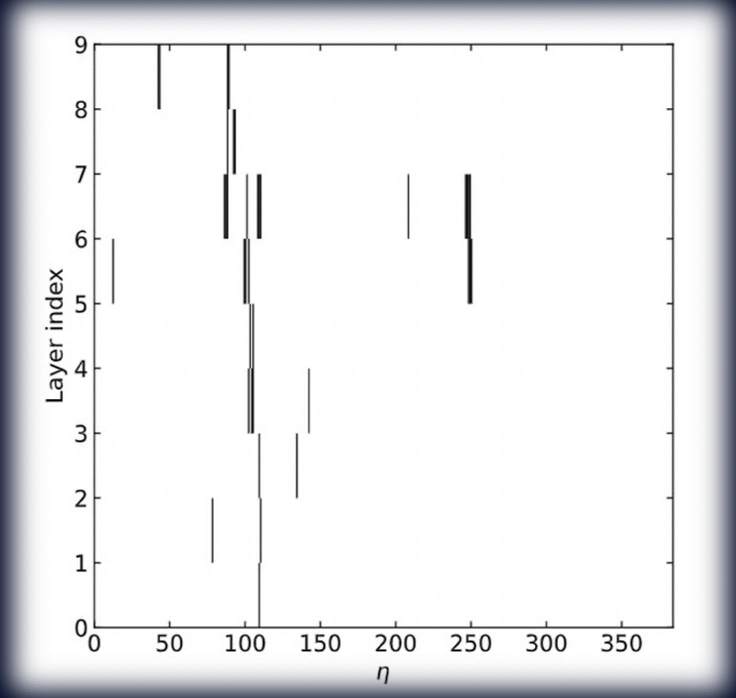
INFN

EuCAIFCon24 - 30/04/24

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

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

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

Studies on track finding algorithms based on machine learning with GPU and FPGA

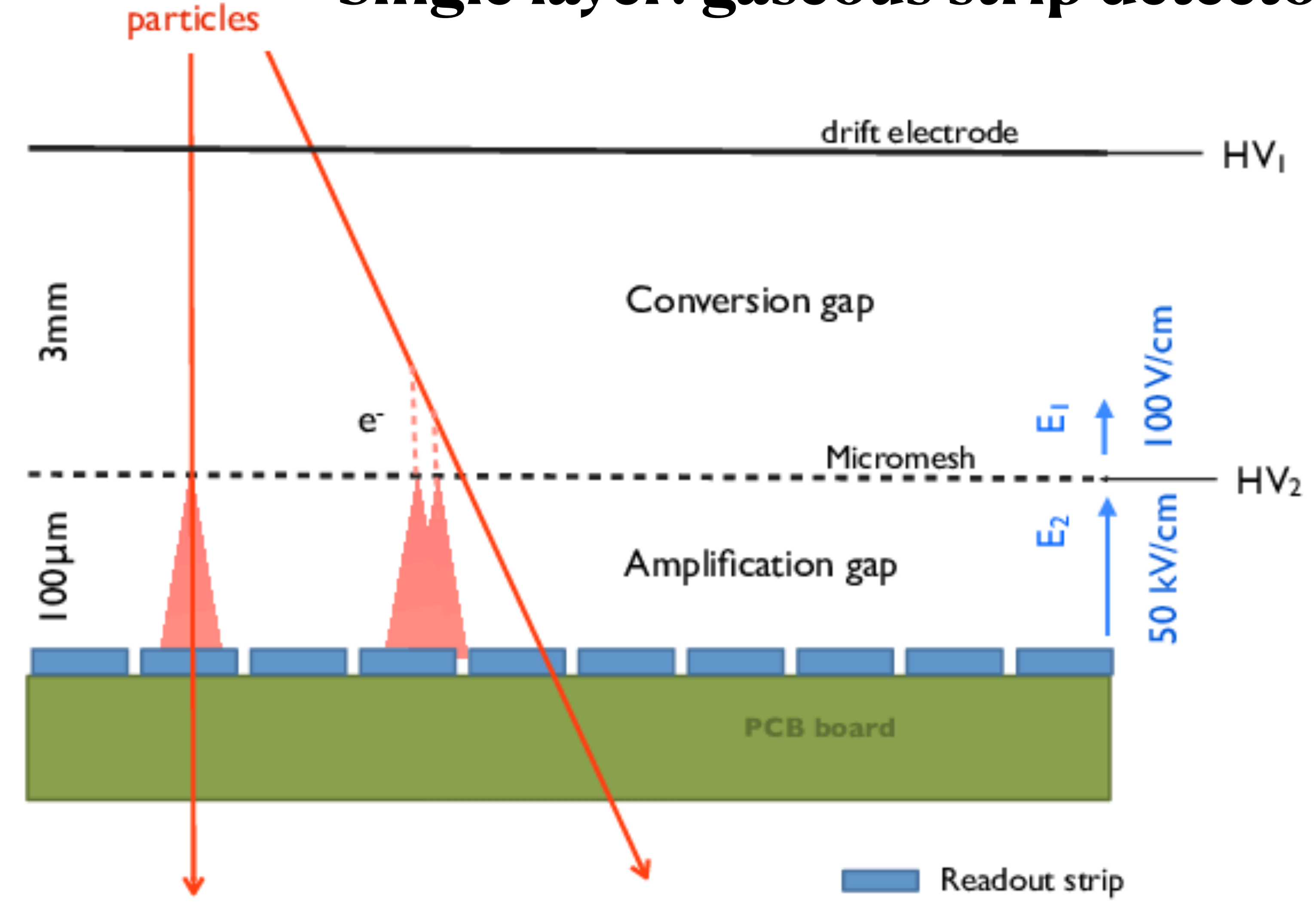
Maria Carnesale

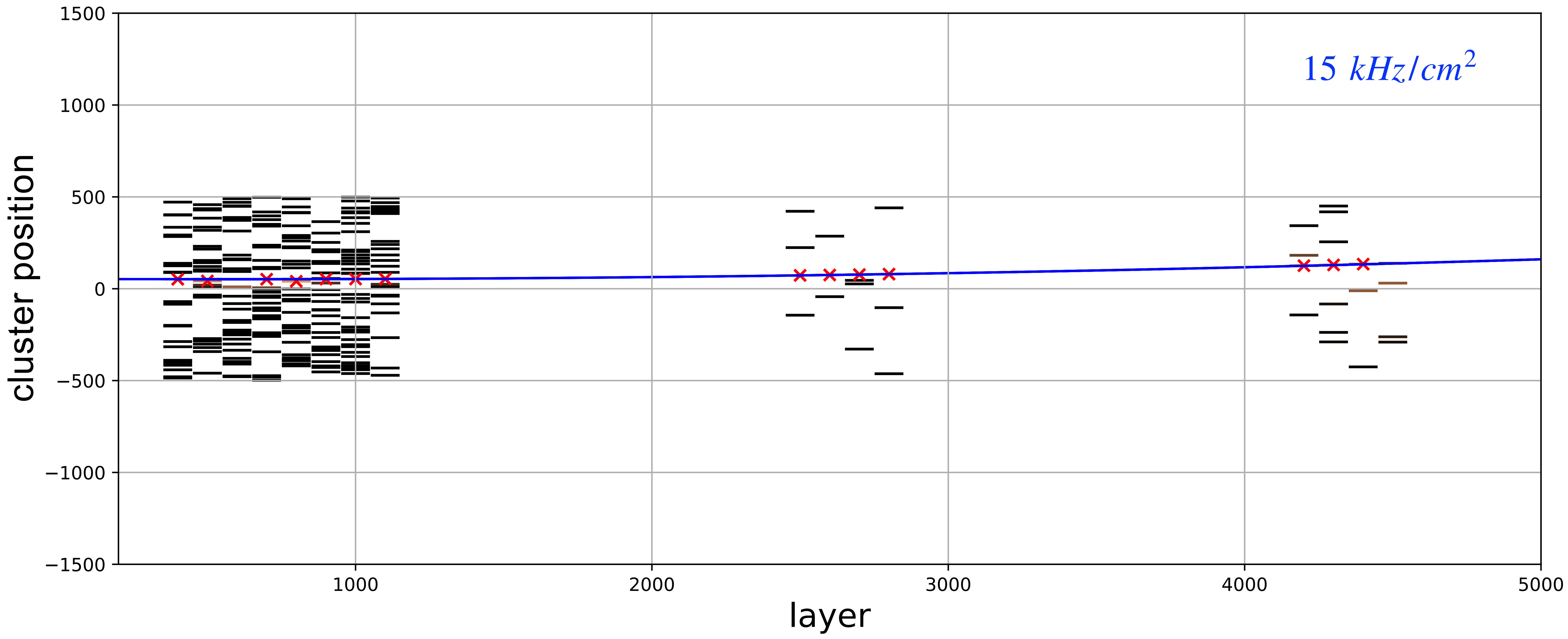
EuCAIFCon 2024 – Amsterdam – 30 Apr- 3 May 2024

ML algorithms for muon pattern recognition

- **Algorithms for cluster reconstruction and pattern recognition in gaseous strip detectors**
- Models tested are **Dense NN** (DNN) **Convolutional NN** (CNN)
  - DNN trained to identify clusters produced by muons in gaseous strip detectors
  - RNN/CNN trained to identify tracks in events with high occupancy

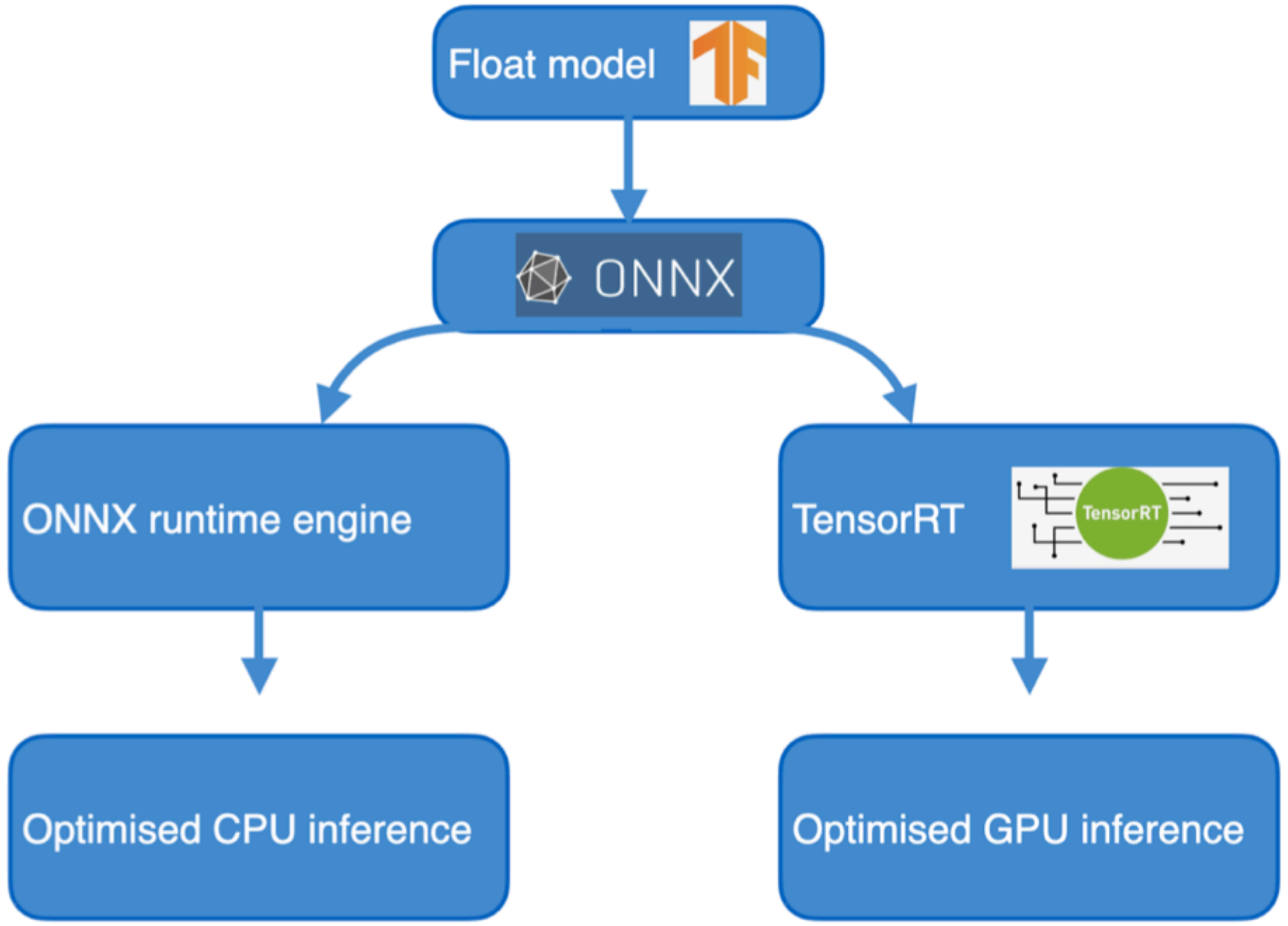
Single layer: gaseous strip detector



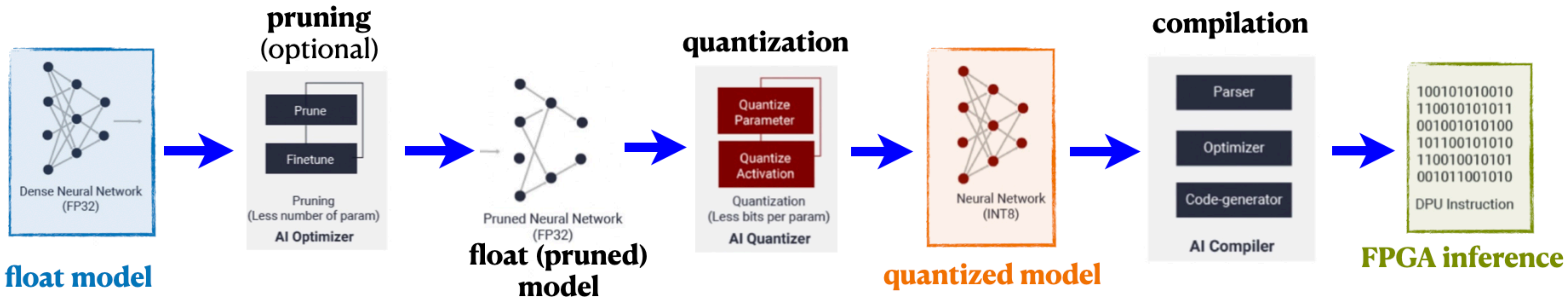


ML algorithms tested on CPU/GPU/FPGA

- **Study of inference time and performance on different architectures:**
  - **CPU**: using ONNX
    - Open Neural Network Exchange: open source framework that optimizes the usage of CPU resources
  - **GPU**: using tensor flow and tensorRT
    - Framework produced by NVIDIA to run optimized inference on GPU

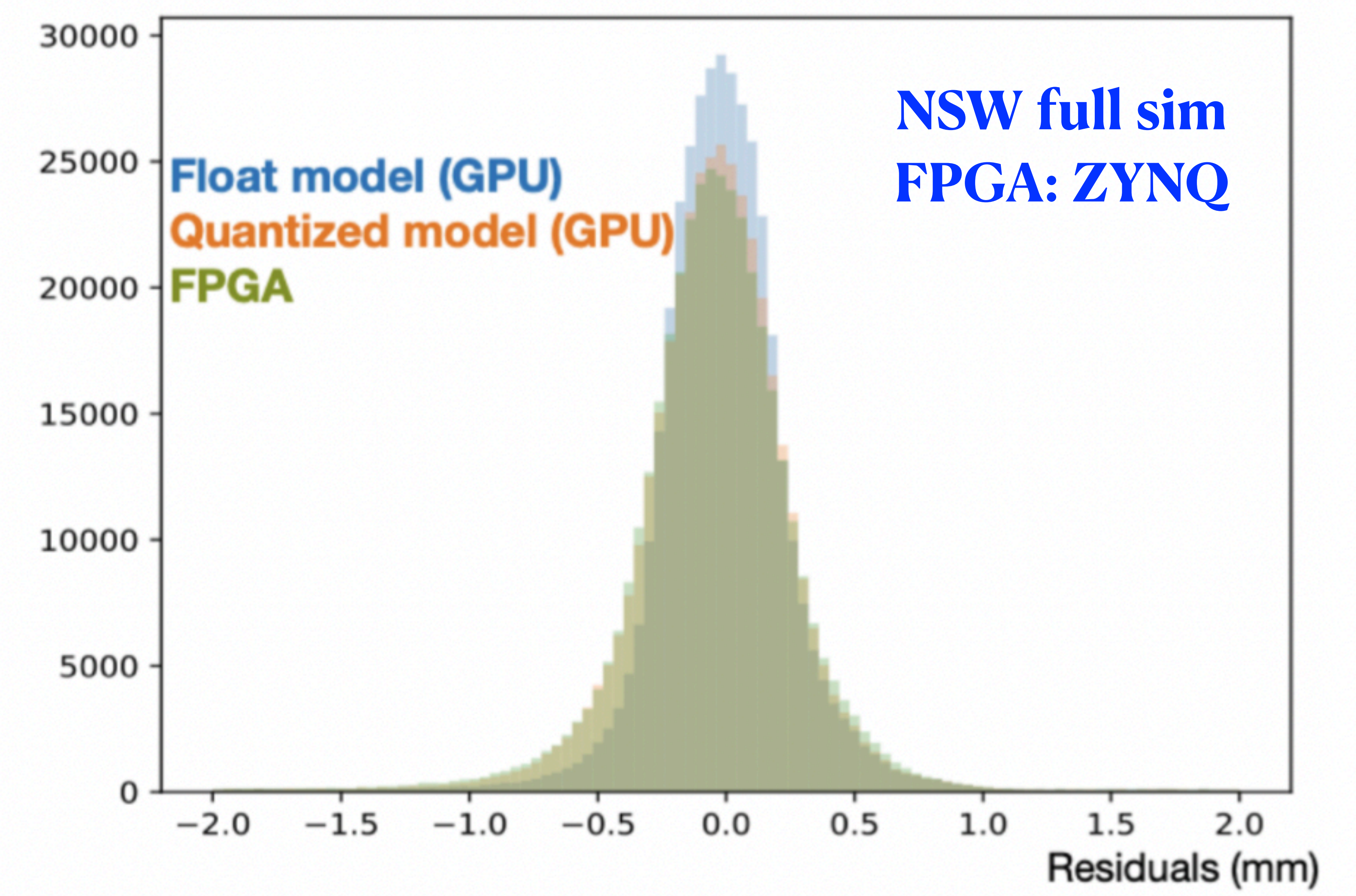


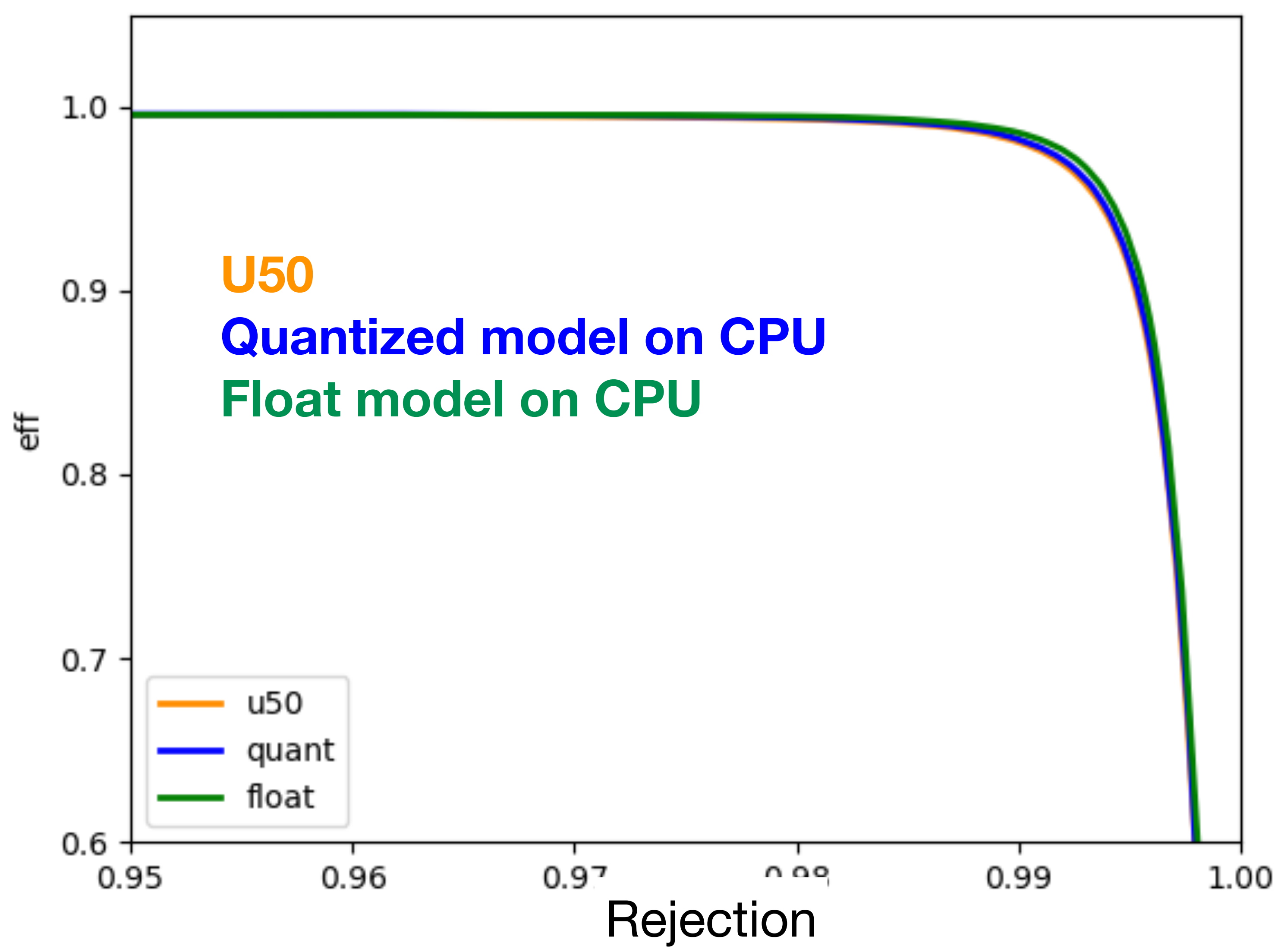
  - **FPGA**: using Vitis-AI workflow provided by Xilinx for inference acceleration or HLS4ML and vivado

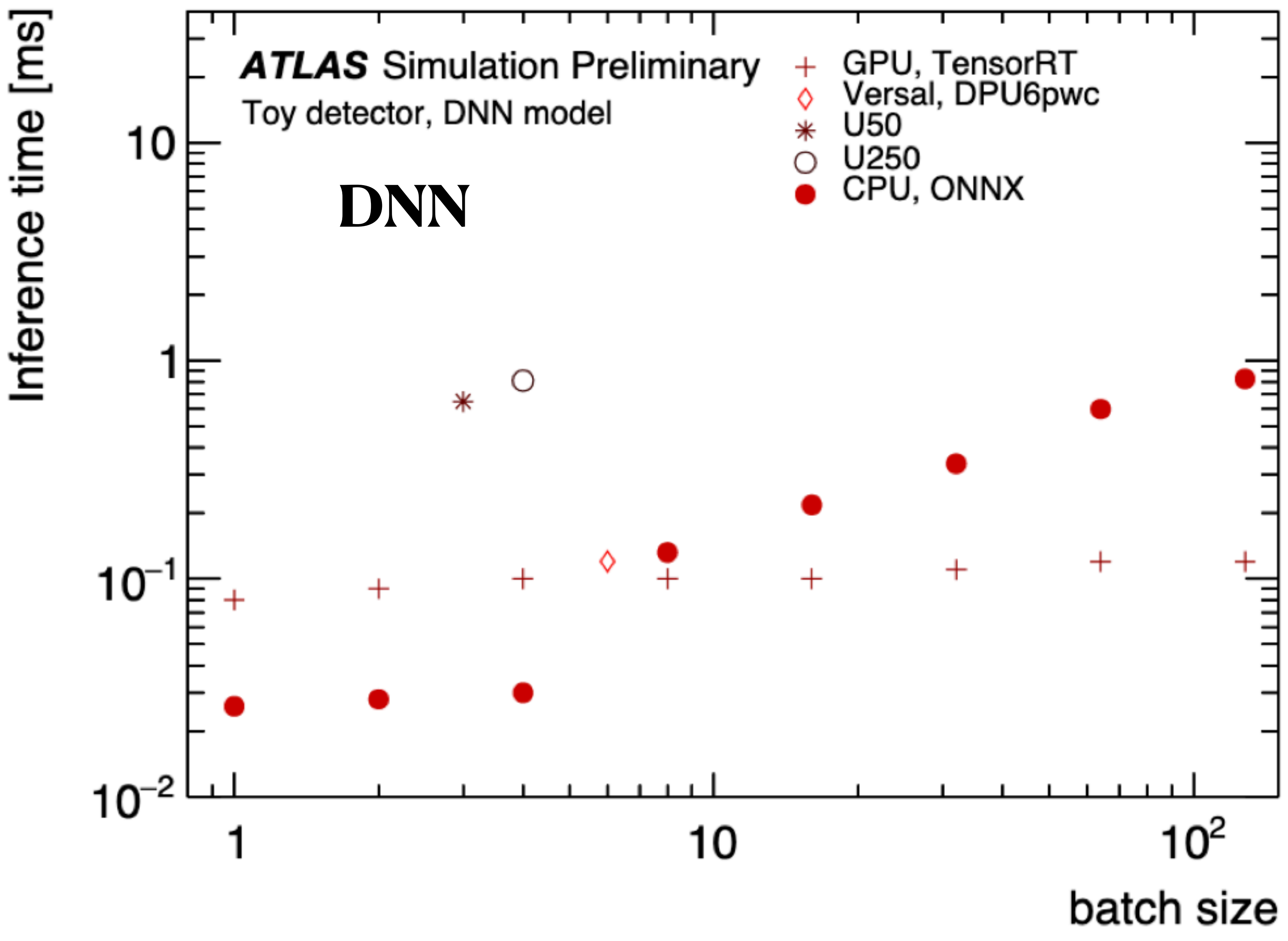


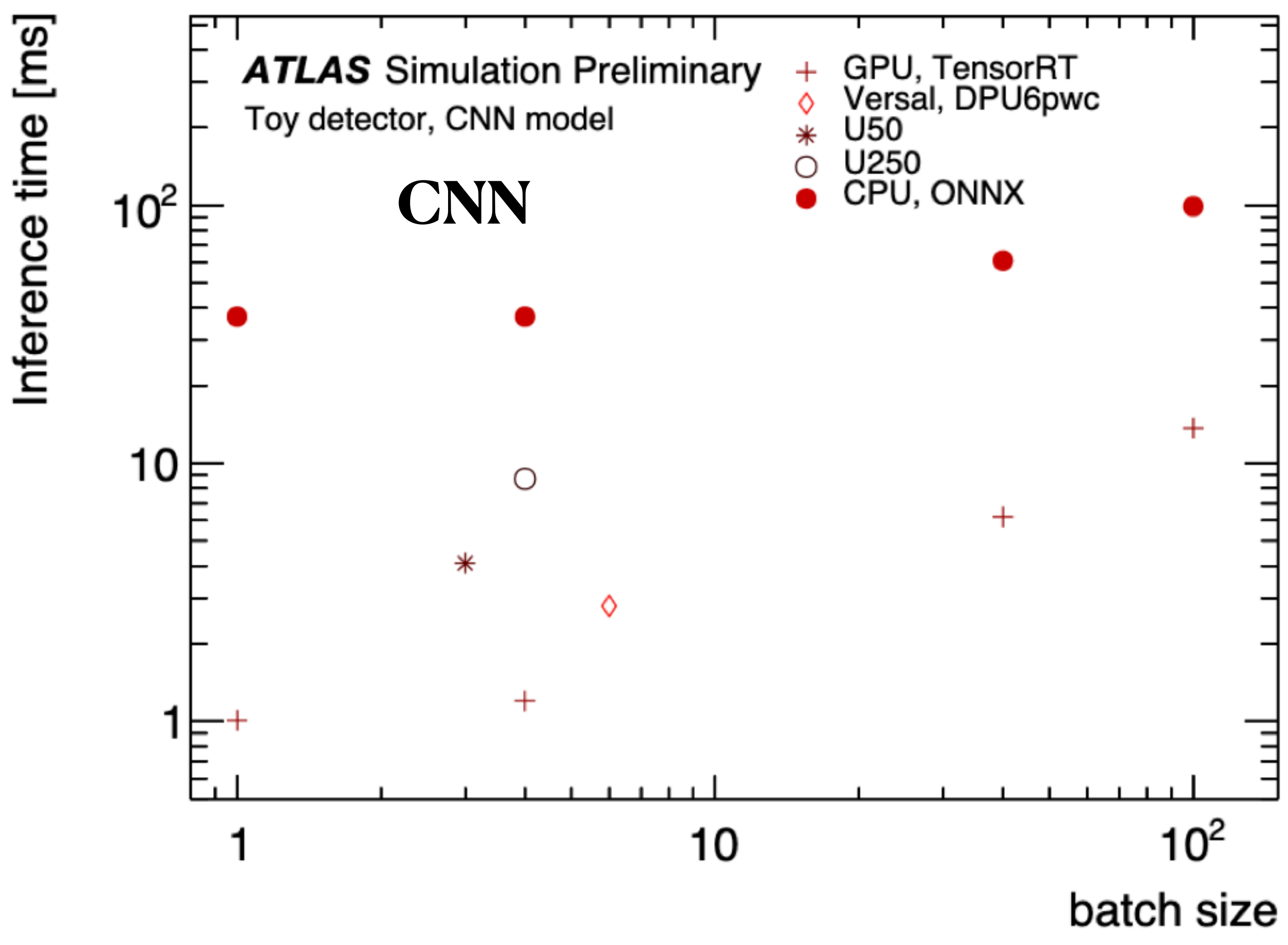
Timing and performance on CPU/GPU/FPGA

- Comparing CPU (ONNX) / GPU (TensorRT) / FPGA (Vitis AI) for DNN and CNN inference
- Small resolution degradation after quantisation
- Same performance of quantised model on CPU/GPU and FPGA









- Batch size (number of events processed in parallel) is fixed in the case of FPGA, free for CPU/GPU
- FPGA times are not a simulation

Adaptive Machine Learning on FPGAs: Bridging Simulated and Real-World Data in High-Energy Physics

Mattia Cerrato & **Marius Köppel**

Institute of Computer Science Johannes Gutenberg-Universität Mainz

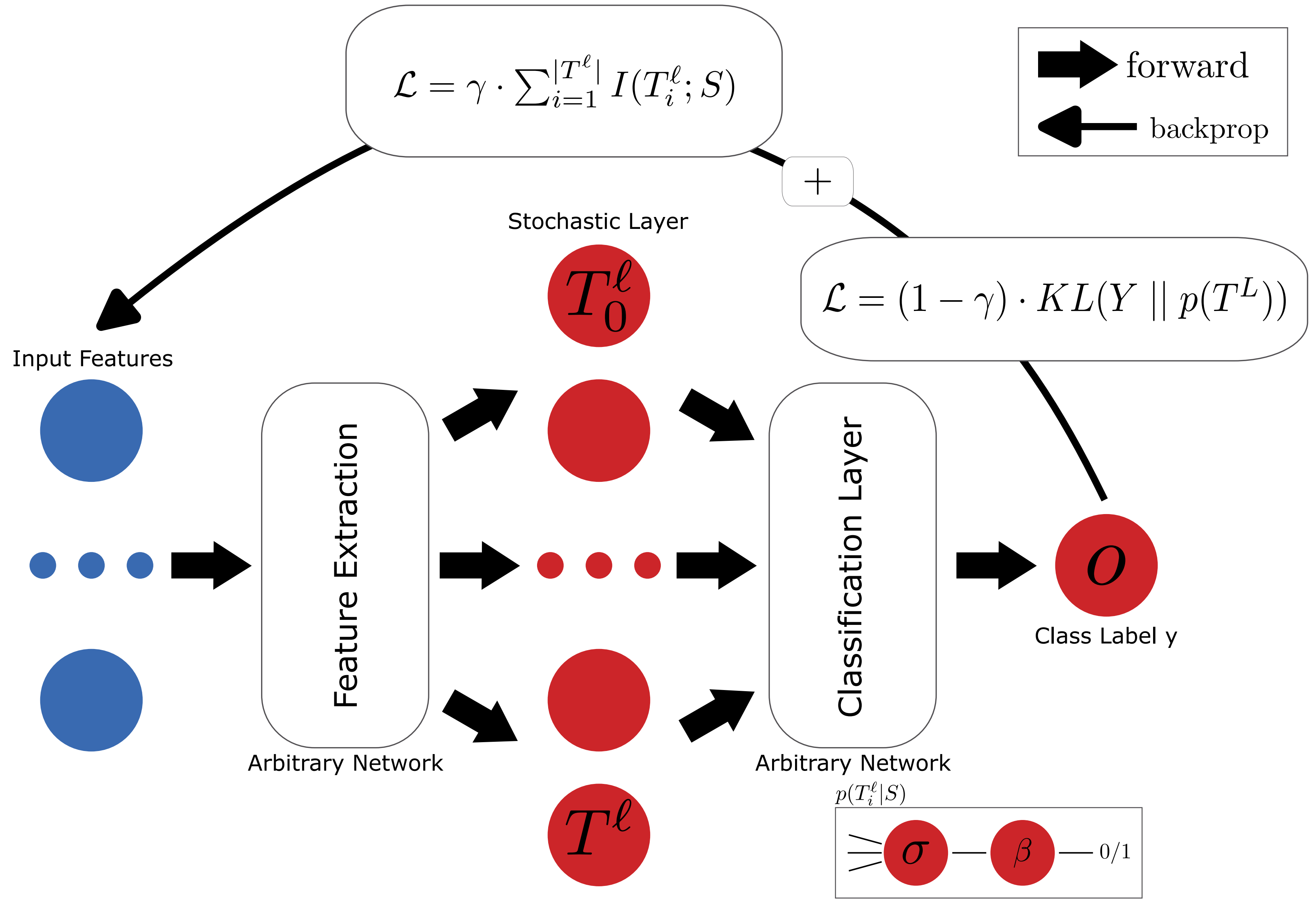
Institute for Particle Physics and Astrophysics ETH Zürich

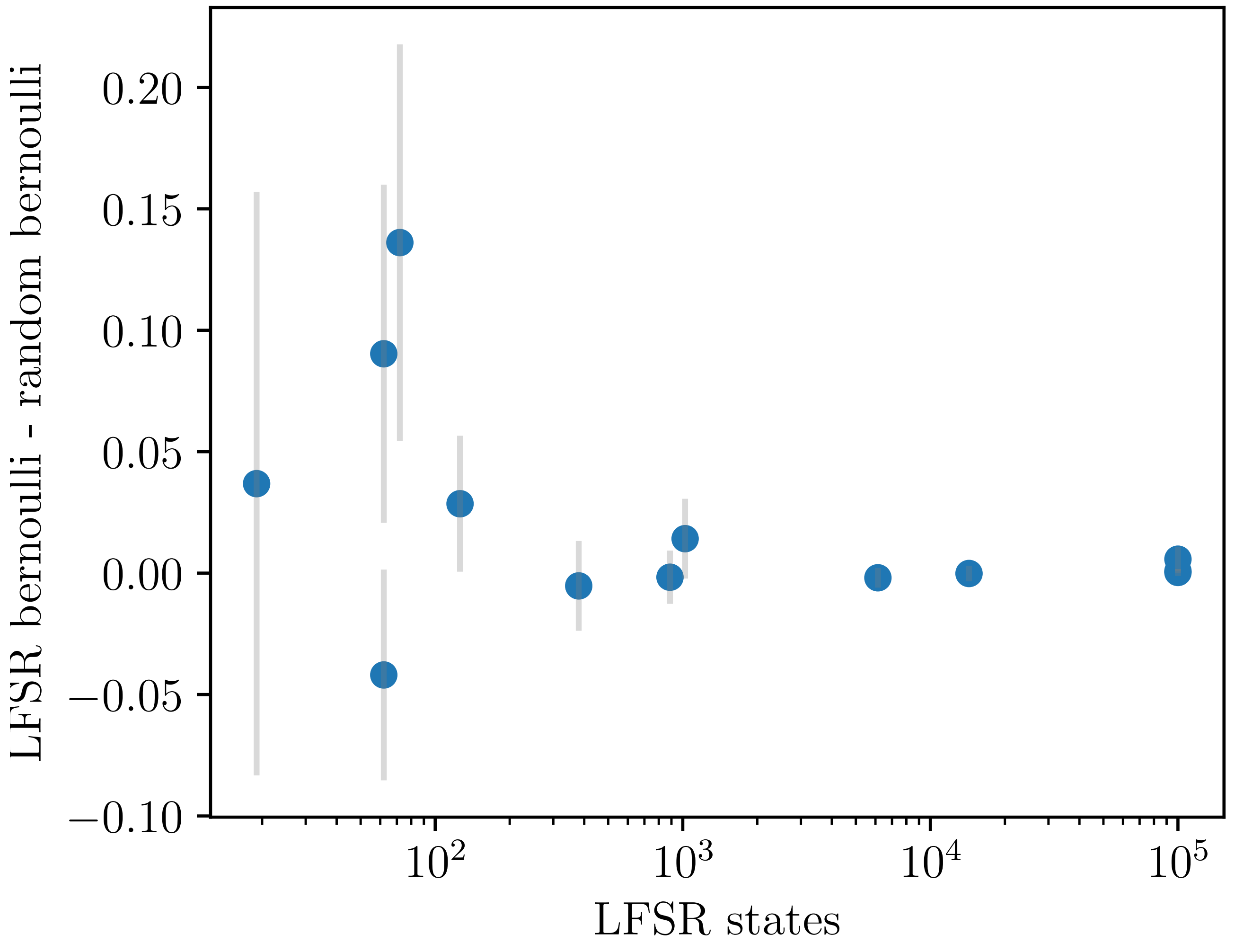
Marius Köppel 30.04.2024

ETH zürich JG|U

Motivation

- How to train **invariant representations** to bridge simulation and real-data?
- Use **binary stochastically** activations to treat as random variables
- Adversarial classifier need around **2x neurons** domain transformation task
- **Stochastic quantization** neurons by Bernoulli
- On FPGAs use **linear-feedback-shift-register** for Bernoulli distribution





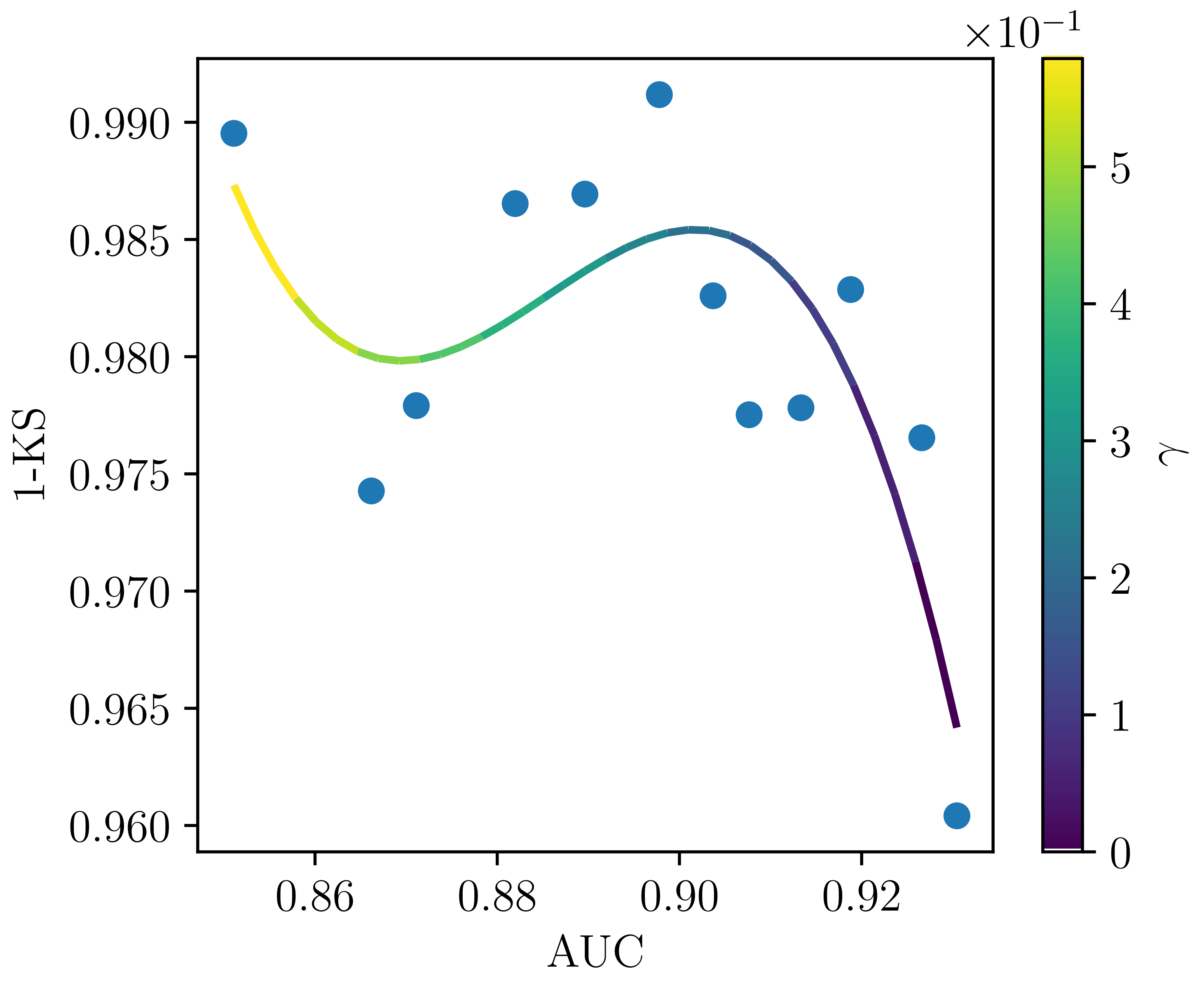
$$H(T_i^l) = -(1-\theta_i^l)\cdot \log_2(1-\theta_i^l) - \theta_i^l \cdot \log_2(\theta_i^l)$$

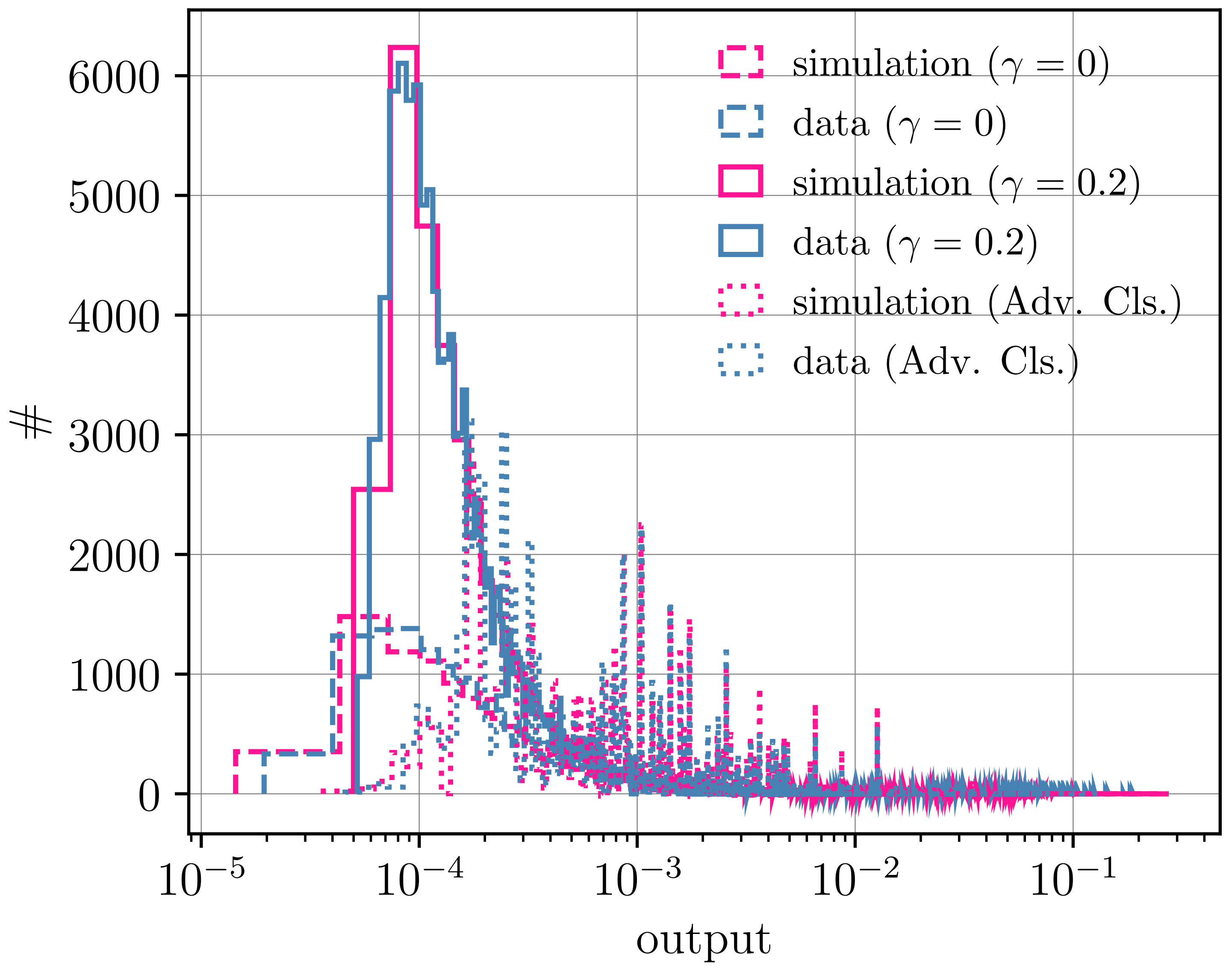
$$\sum_{i=1}^{|T_l|} I(T_i^l; S) \geq I(T_l; S)$$

[1] https://arxiv.org/abs/2208.02656

Experiments

- We analyze the method on LHCb data [2]
- Method can outperform **adversarial classifier** & **normal NN**
- Method obtains stable **invariant/accuracy tradeoffs**
- Use method to be **pile-up** invariant
- Revisit **information bottleneck theory**

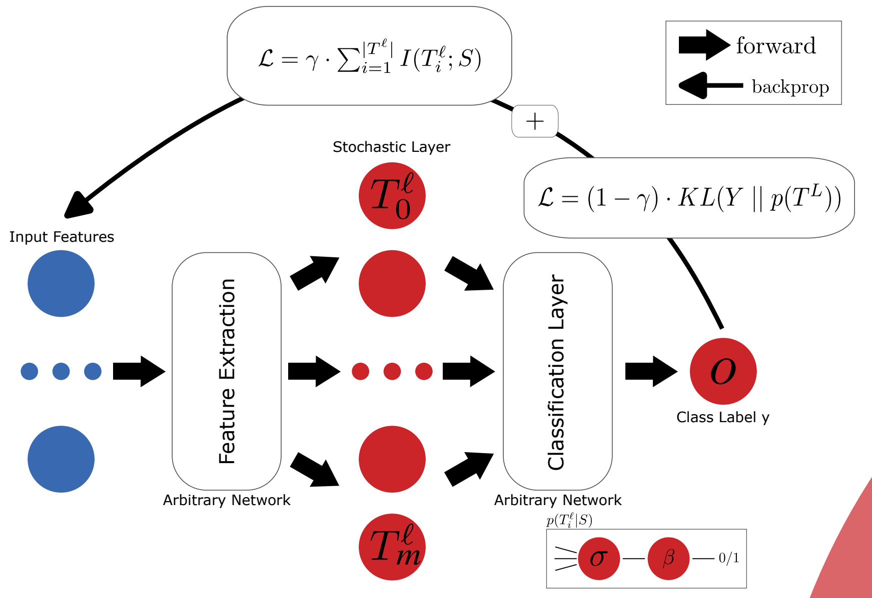




Adaptive Machine Learning on FPGAs: Bridging Simulated and Real-World Data in High-Energy Physics

Mattia Cerrato[1] & Marius Köppel[2]
[1]Institute of Computer Science Johannes Gutenberg-Universität Mainz
[2]Institute for Particle Physics and Astrophysics ETH Zürich

Abstract
- Use **binary stochastically activations** to treat neurons as random variables [1]

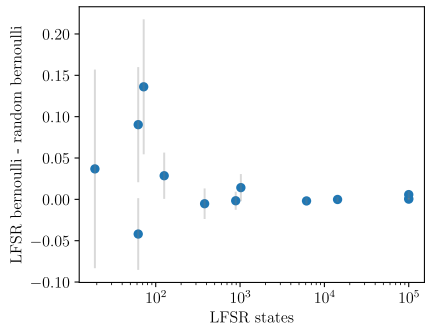


Introduction
- Be invariant to **differences** in **simulation** and **data**
- **Adversarial classifier** need **2x** neurons to perform domain transformation task
- Computing of mutual information in **deterministic** neural networks is **hard** or even **impossible**

Method
- **Stochastic quantization** neurons by **Bernoulli**

$$H(T_i^l) = -(1-\theta_i^l) \cdot \log_2(1-\theta_i^l) - \theta_i^l \cdot \log_2(\theta_i^l)$$

- On **FPGAs** use linear-feedback-shift-register for Bernoulli distribution

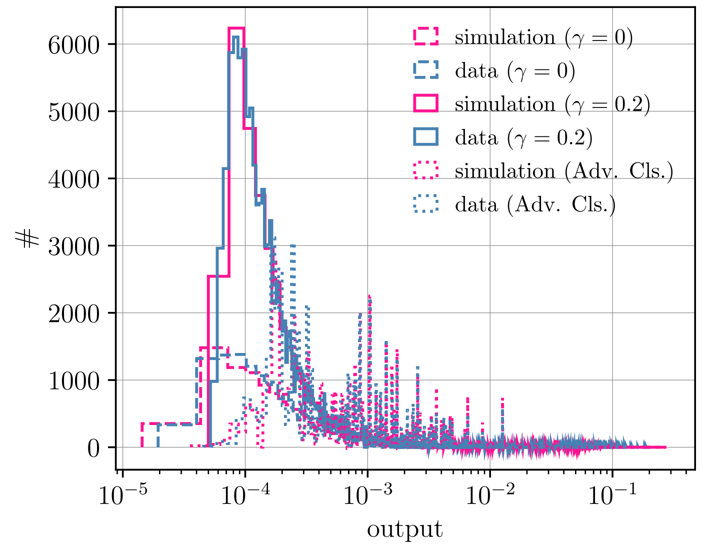


$$H(T_i^l) \qquad H(S)$$

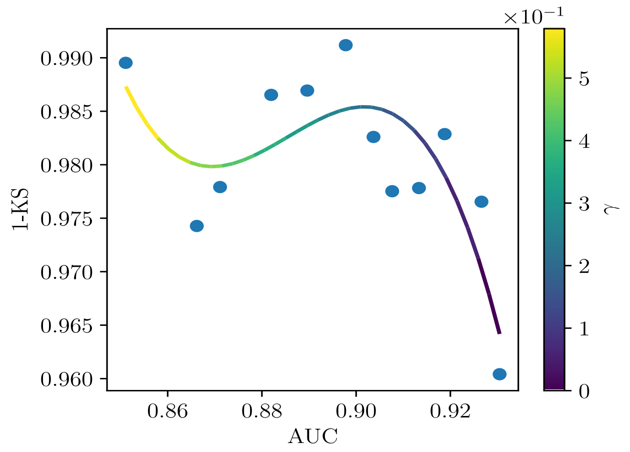
Direct minimization of mutual information in a full precision network using binary stochastically-activated layers for obtaining invariant representations

$$\sum_{i=1}^{|T_l|} I(T_i^l; S) \geq I(T_l; S)$$

Experiments
- We analyze the method on **LHCb data** [2]



Results
- Method can **outperform adversarial classifier** & **normal NN**
- Method obtains **stable** invariant/accuracy **tradeoffs**



Outlook
- Use method to be **pile-up** invariant
- Revisit **information bottleneck** theory

References
[1] https://arxiv.org/abs/2208.02656
[2] https://www.kaggle.com/competitions/flavours-of-physics

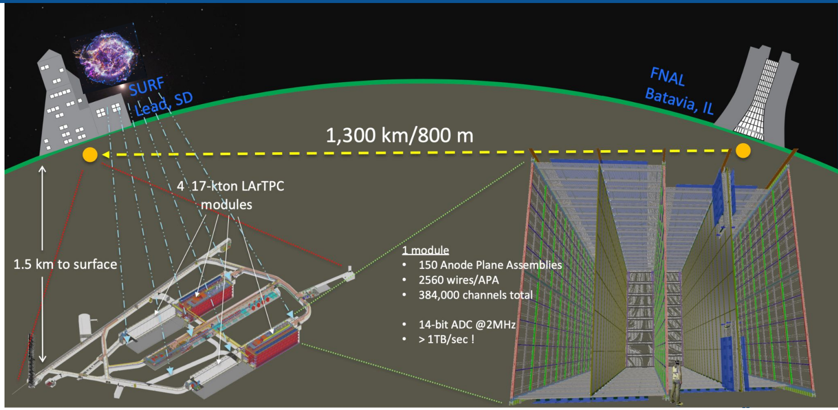
ETH zürich JG|U

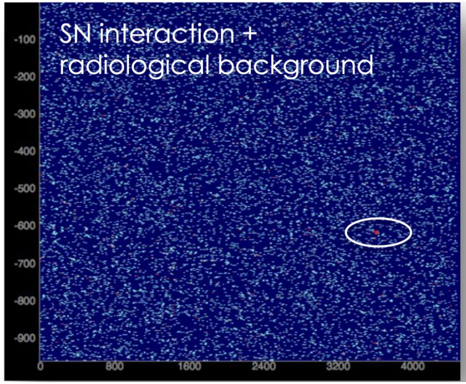
Real-Time Detection of Low-Energy Events for the DUNE Data Selection System using ML

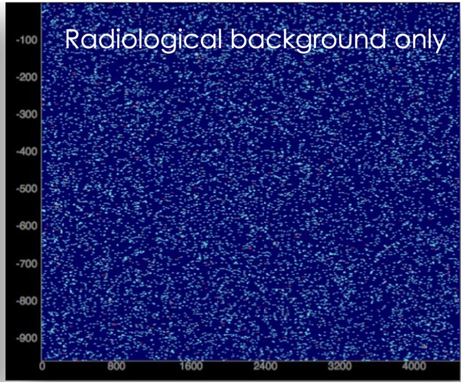
EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE EuCAIFCon 2024

The Deep Underground Neutrino Experiment (DUNE) is a next-generation experiment for neutrino science at the Fermi National Accelerator Laboratory in Batavia, Illinois.

- DUNE high-resolution “video” stream: up to 4x200 cell volumes, 11.5 MP frames per 2.25ms, 12-bit resolution, total of ~40 terabits/s.
- Designed for 95% trigger efficiency on a supernova burst.
- Early trigger & SN pointing from LE ν.
  - Hard to distinguish, Multiplicity and Clustering not efficient.
  - Differentiate between ν-LE types.
  - Delay in SN light - a few mins to days.
  - Very rare (~1/100 yr) - accuracy is important.
- Improve signal efficiency for solar ν.
  - Low sensitivity due to high Background noise and high threshold.
- Data reduction $O(10^4)$ is a necessity.
- Power consumption, heat, space an issue.
- 2DCNN on FPGA a potential solution.

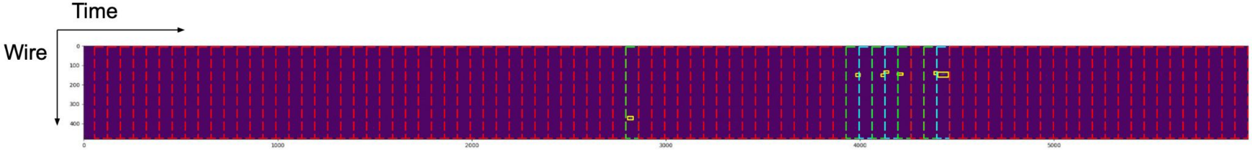






Real-Time Detection of Low-Energy Events for the DUNE Data Selection System using ML

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE EuCAIFCon 2024



- ML algorithm for real-time data processing and trigger from a stream of LArTPCs data.
- Continuous read-out, arranged into “frames” and selected data is sent for further processing.
- Denoise + Downsize + 2DCNN
- Classify ν-LE events in real time with ≥ **90%** efficiency, reject noise background (NB) images with $\gg$**99.99%** efficiency.
- Each incoming 480 x 64 image must be processed within **32 μs** to avoid queuing.
- **HLS code injection** : to reach the meet latency and resource requirements.
- A detailed study of various implementations.
- A viable solution for DUNE readout.