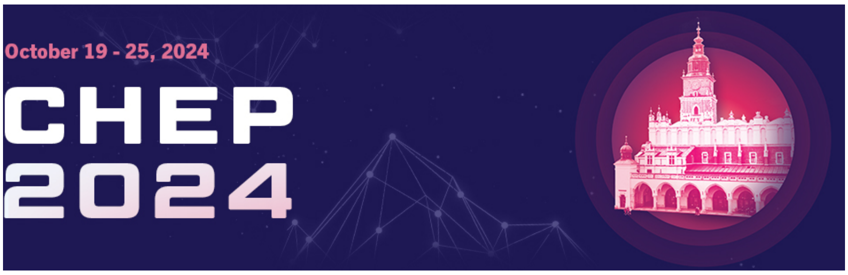

Contribution ID: **356** Type: **Talk**

Benchmark Studies of ML Inference with TMVA SOFIE

Wednesday 23 October 2024 13:30 (18 minutes)

Within the ROOT/TMVA project, we have developed a tool called SOFIE, that takes externally trained deep learning models in ONNX format or Keras and PyTorch native formats and generates C++ code that can be easily included and invoked for fast inference of the model. The code has a minimal dependency and can be easily integrated into the data processing and analysis workflows of the HEP experiments.
This study presents a comprehensive benchmark analysis of SOFIE and prominent machine learning frameworks for model evaluation such as PyTorch, TensorFlow XLA and ONNXRunTime. Our research focuses on evaluating the performance of these tools in the context of HEP, with an emphasis on their application with typical models used, such as Graph Neural Netwarks for jet tagging and Variation auro-encoder and GAN for fast simulation. We assess the tools based on several key parameters, including computational speed, memory usage, scalability, and ease of integration with existing HEP software ecosystems. Through this comparative study, we aim to provide insights that can guide the HEP community in selecting the most suitable framework for their specific needs.

Primary authors: PANAGOU, Ioanna Maria; MONETA, Lorenzo (CERN); SHAH, Neel (Indian Institute of Technology Madras (IN)); WOLLENHAUPT, Paul (Georg August Universitaet Goettingen (DE)); SENGUPTA, SANJIBAN

Presenter: MONETA, Lorenzo (CERN)

Session Classification: Parallel (Track 5)

Track Classification: Track 5 - Simulation and analysis tools