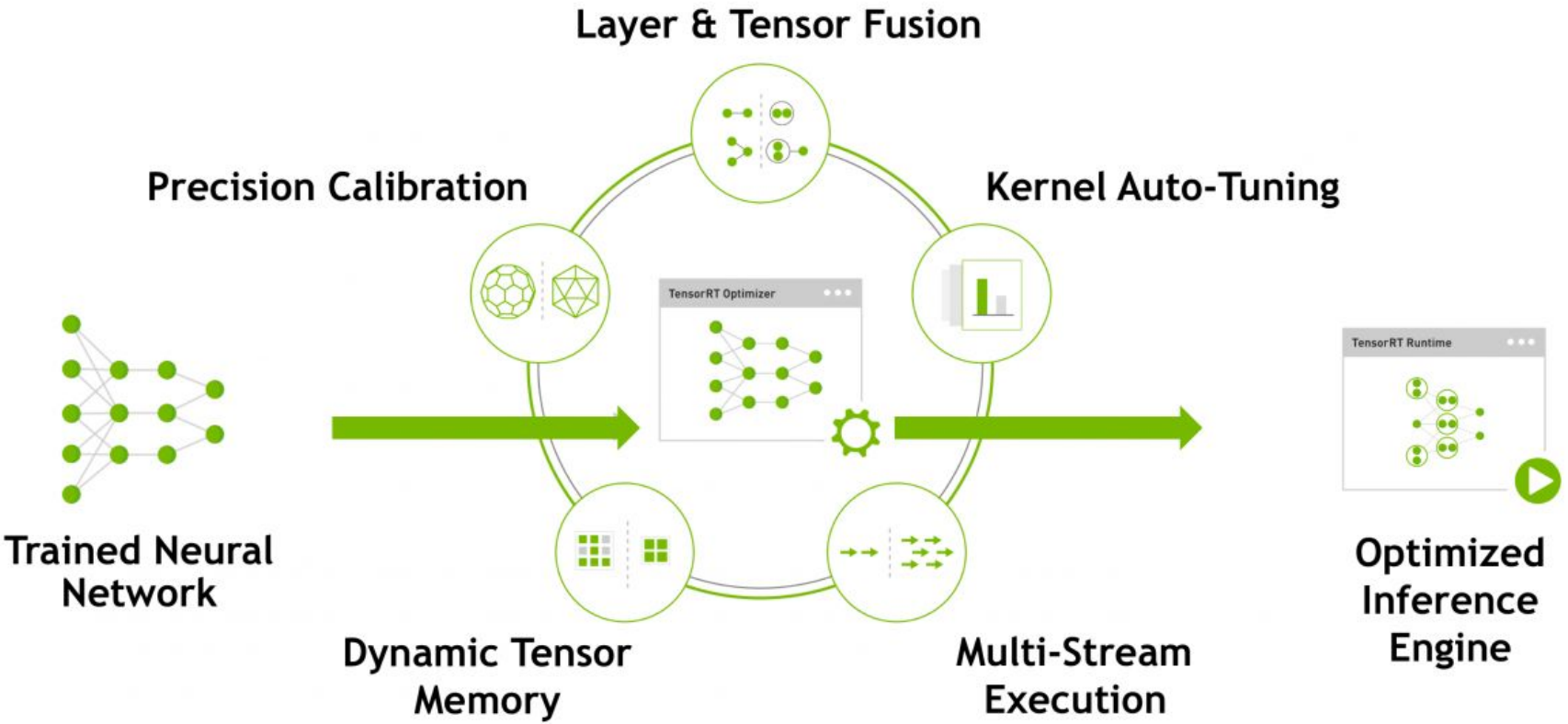
Fast inference in CMSSW with NVIDIA TensorRT

A. Di Pilato, A. Di Florio

Goal

- Integrate fast inference in CMSSW on GPU with **NVIDIA TensorRT**: https://developer.nvidia.com/tensorrt
- NVIDIA TensorRT is a platform for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime that delivers low latency and high-throughput for deep learning inference applications.
- Test the performance on the model developed for Particle ID and Energy Regression in HGCAL within TICL framework.


Layer & Tensor Fusion
Precision Calibration
Kernel Auto-Tuning
TensorRT Optimizer
TensorRT Runtime
Trained Neural Network
Dynamic Tensor Memory
Multi-Stream Execution
Optimized Inference Engine