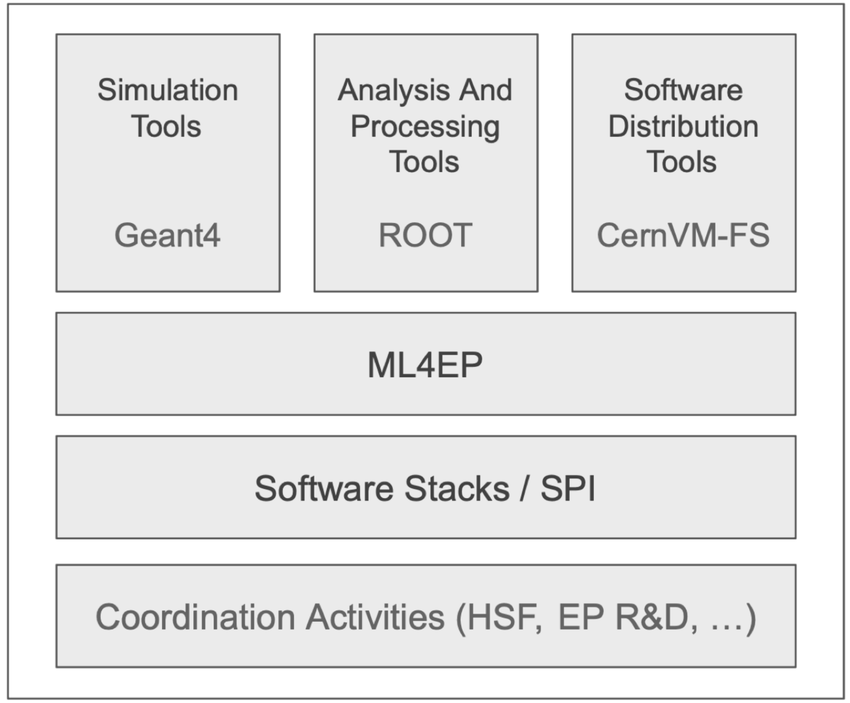
Introduction: ML4EP

- First Meeting of new project in SFT for common ML activities
  - **goal**: provide service and support to the experiment on common ML issues
- Initiated by building on existing ML activities:
  - **ML for fast simulation**
  - ML software in ROOT
    - **SOFIE** (DL inference)
    - **RBDT** (BDT inference)
    - **Batch generator**
    - **TMVA**



Stakeholders

ALICE
ATLAS
CMS
LHCb
EP R&D
IT projects
HSF
FCC
…

ML4EP Meeting

- Working meeting to monitor progress in current activities
- Plan for a bi-weekly meeting:
  - Thursday morning at 9:30 am?
    - or better in the afternoon for GSOC students?

- Today:
  - Introduction of different activities
  - GSOC students
  - Plan of work status

Slides for LHCC

ML4EP Project

- **Vision**
  - Building on existing activities in EP/SFT such as fast simulation and ROOT ML, develop and maintain common ML software solutions required for experiments and promote collaboration on AI/ML topics
- **Identified goals**
  - Development of ML models for fast simulation of calorimeter showers
    - for LHC experiments and future ones (e.g. FCC)
  - Integration of ML inference in experiment workflows
    - support heterogeneous architectures (CPU, GPU and FPGA)
  - Provide a common software pipeline for training ML models
  - Collaborate with the AI/ML community (IML and EuCAIF) on common efforts like:
    - development of foundation models for particle physics
    - maintain benchmark data and challenges for testing performance of algorithms
  - Host common ML activities of the Next Generation Trigger project

ML4EP Plans

- **Current activities and plans for near future**
  - Validation of diffusion model (based on transformer) for ATLAS and LHCb (hadronic) shower simulations.
  - Work on inference optimization of diffusion model
  - Extending inference support in ROOT SOFIE for complex ML models (GNN, transformers)
  - Benchmark inference in terms of CPU time and memory consumption of common ML models used by experiments (VAE, GNN, diffusion, and transformer models)
    - using different implementations: SOFIE, Tensorflow XLA, ONNXRuntime and PyTorch
    - abstract submitted to CHEP2024
- **Longer term plans**
  - Will include tasks from NGT using their new resources
  - Develop interfaces to ML inference for integration in reconstruction and high level trigger
  - Develop common software framework for training and hyper-parameter optimisation of ML models
    - including hardware-aware NN training
  - Work on fast inference on FPGA and GPU for complex ML models
  - Contribute to community efforts in fast simulation
    - organisation of CaloChallenge for algorithm benchmarks
    - integration of ML shower simulation models in FCCee detector simulation

Plan of Work presented in January

from A. Ribon presentation

Fast Simulation

The ML-related work items will be integrated into the new ML activity

- Develop transformer-based ML models
  - Establish the best single-geometry diffusion model
  - Work on inference optimisation
  - Extend to different geometries and test adaptation capabilities, measure savings on training time
- Experiment-specific work (in collaboration with members of the experiments)
  - LHCb
    - Find the best working model for hadronic showers (possibly a transformer-based model)
  - ATLAS
    - New Fellow (Peter Mckeown) will continue the work of D. Salamani on ML for ATLAS, implementing a data structure that allows to test VAE and transformer-based models
    - Co-supervise work of J. Beirer on FastCaloSimV2-based classical shower simulation
  - CMS
    - Implement data production sample with structure that allows to test transformer-based models on HGCal
- Others
  - Speed-up simulation of oriented crystals detector
  - Community efforts : CaloChallenge and Open Data Detector

Machine Learning

LM, JR

See Lorenzo's talk Vision for a new ML/AI activity !

Priority 1:

- Put RBatchGenerator in production
- Consolidate RBDT
- Support of integration of SOFIE in experiments Fast Simulation pipelines
- Add support in SOFIE for NVidia GPUs in CUDA
- Continue to add support for the ONNX operators requested by experiments

Priority 2:

- Make HLS4ML interoperable with SOFIE
- Streamline ROOT's inference interface, making it able to use models for Python ML frameworks (e.g. Keras/TF) directly

We want to support experiments inference (C++) for cases that are difficult to implement or require heavy dependencies.

We don't want to compete with existing industry tools for training.