ML4EP
2025 Plan of Work

SFT POW Meeting - 22/1/2025

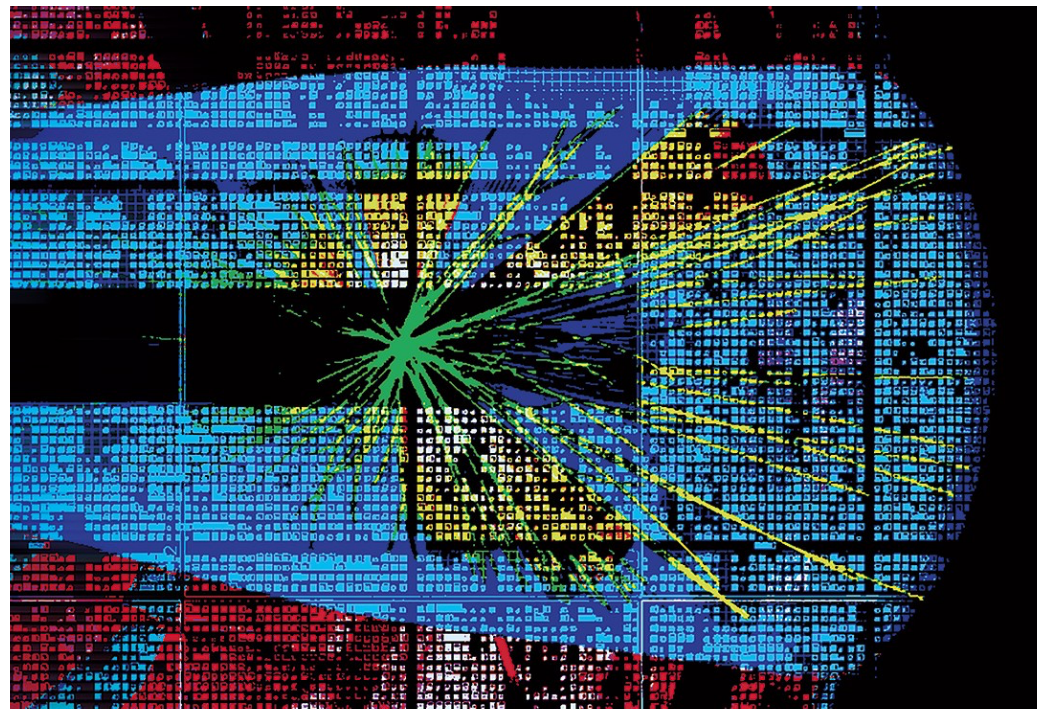
Lorenzo Moneta on behalf of the ML4EP team

The ML4EP Project

- Project started last year for hosting common AI/ML activities within SFT
- Initiated by building on the existing ML activities within the SFT projects:
  - ML for fast simulation
    - Developments of models for fast simulation of calorimeter showers
    - Collaboration with experiments in developing and deploying fast simulation ML models
  - ML software in ROOT
    - Maintain ROOT ML software (TMVA)
    - Interfaces for using external ML software with ROOT
    - Develop an efficient solution for C++ inference of ML models (SOFIE)

The NGT ML Activities

- Common (across multiple experiments) activities of the Next Generation Trigger project (WP1) and hosted within the SFT group
- Activities with ML focus that are part of ML4EP
  - **1.2**: Develop ML models (NN) for FPGA (**hls4ml**)
  - **1.3**: Implement algorithms for model compression and training
  - **1.7**: Develop interfaces for ML Inference on heterogeneous architectures



Project Goals

- Develop and maintain common ML software solutions required for the experiments
- Promote collaboration and give direction to the different SFT projects on AI/ML topics
  - sharing expertise and knowledge

- **Do not compete with existing industrial open-source tools, but be complementary and, when needed, facilitate their usage**

Project Organisation

- Topical weekly meetings for the different areas:
  - Fast simulation, ML software, NGT activities
- Bi-weekly/monthly meetings for all activities
  - for sharing knowledge and allowing synergies between activities
  - promote working across the different activities:
    - summer student was working on both diffusion models and benchmarking ML inference
- Reporting regularly to ROOT, Simulation and NGT meetings and workshops
- Organisation of common meetings/workshop with the community:
  - IML workshop or topical IML meetings
  - NGT workshops (e.g. hls4ml Community Forum)

Current Project Effort: Person Power

Persons available for ML4EP during 2024

- Fast Simulation:
  - 1 staff (*Anna Zabrowoska, on leave for ~50% of time in 2024)*
  - 2 fellows (*Piyush Raikwar, Peter McKeown*)
  - 3 summer/GSOC students
- ROOT ML:
  - 1 staff (*Lorenzo Moneta*)
  - 3 summer/GSOC students + one short-term student from Lituania
  - Contributions also from other ROOT members (*Vincenzo Padulano, Jonas Rembser*)
    - supervision of students and developments (BatchGenerator, RBDT, SBI)
- NGT (from September/October)
  - 1 staff (*Vladimir Loncar*)
  - 2 fellows (*Dimitrios Danopoulos, Roope Niemi*)
  - Contributions from FASTML community and experiments

Person Power in 2025

Expected changes in person power in 2025

- Fast Simulation:
  - +1 doctoral student (on EP/RD funding) (*to be hired*)
  - -1 fellow from mid-next year (*PR*)
- ROOT ML:
  - no change with respect to 2024
- NGT
  - +1 doctoral student for 1.7 (*Sanjiban Sengupta*)
  - +2 technical students for 1.2 and 1.3 (*Enrico Lupi, Anastasia Petrovych*)
- CERN Summer students and GSOC students will be requested for all activities

Achievements in 2024

2024 Achievements : Fast Simulation

- CaloDiT (diffusion model) for shower simulation
  - From 2024 POW:
    - Establish the best single-geometry diffusion model (**DONE**)
    - Work on inference optimisation (**DONE**)
    - Extend to different geometry and test adaptation capabilities, measure savings on training time (**DONE**)
  - *All items completed*
    - Demonstrated generalisation capabilities
    - Model inference optimisation using EDM[1] and consistency distillation[2] (single step diffusion)

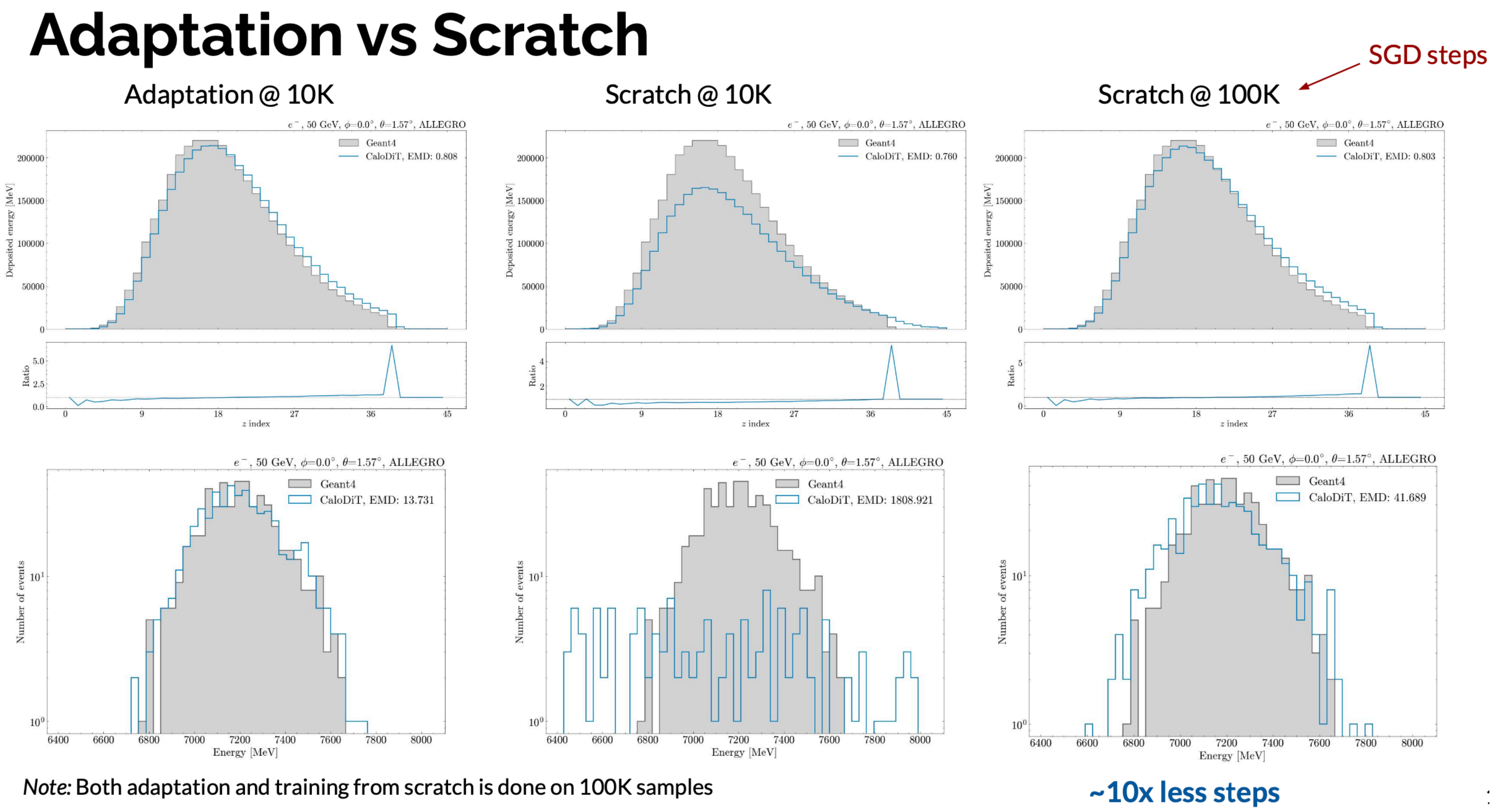
➡ Presentations at ACAT2024 and at ML4JETS

[1] EDM (Elucidating the Design Space of Diffusion-Based Generative Models): https://arxiv.org/abs/2206.00364

[2] https://arxiv.org/abs/2303.01469

CaloDiT: Performances

- Adaption of the model vs training model from scratch



Adaptation vs Scratch

Adaptation @ 10K

Scratch @ 10K

Scratch @ 100K

SGD steps

Note: Both adaptation and training from scratch is done on 100K samples

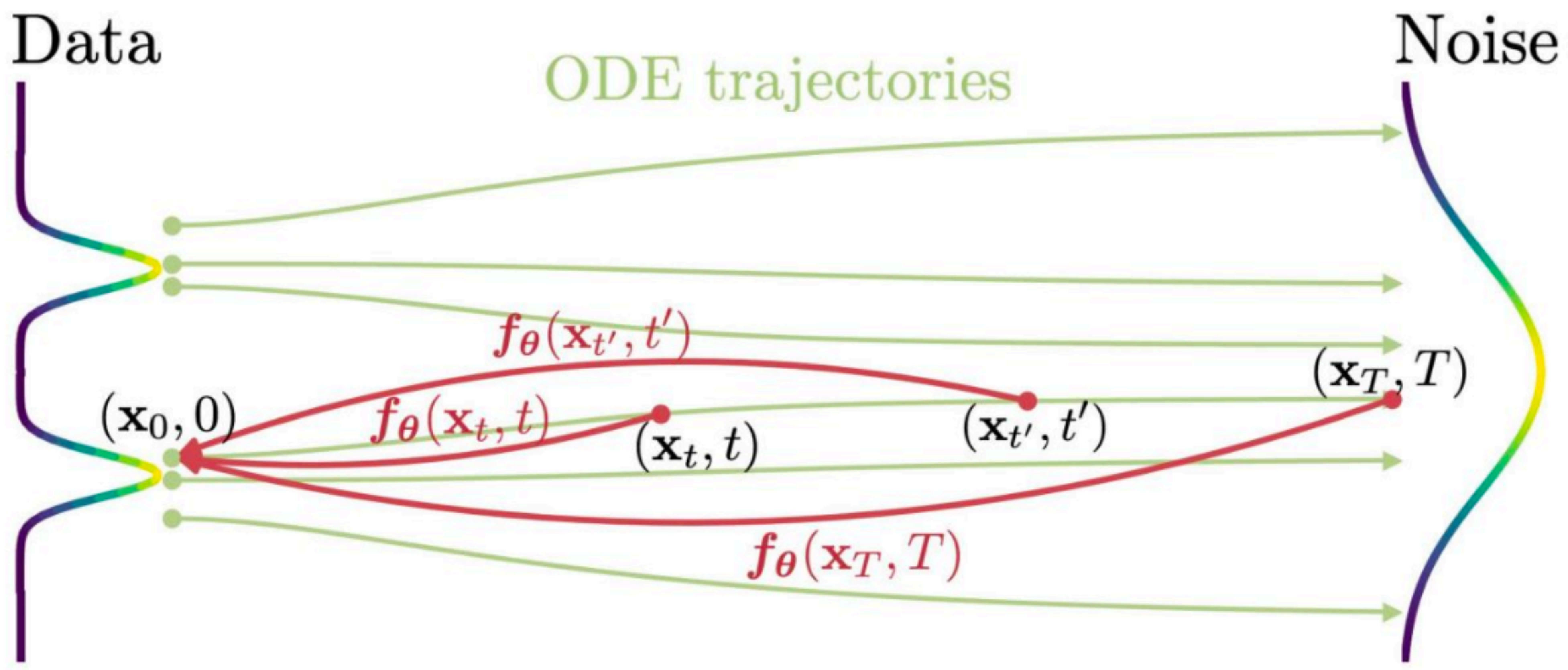
~10x less steps

Using in both cases 100k samples for training

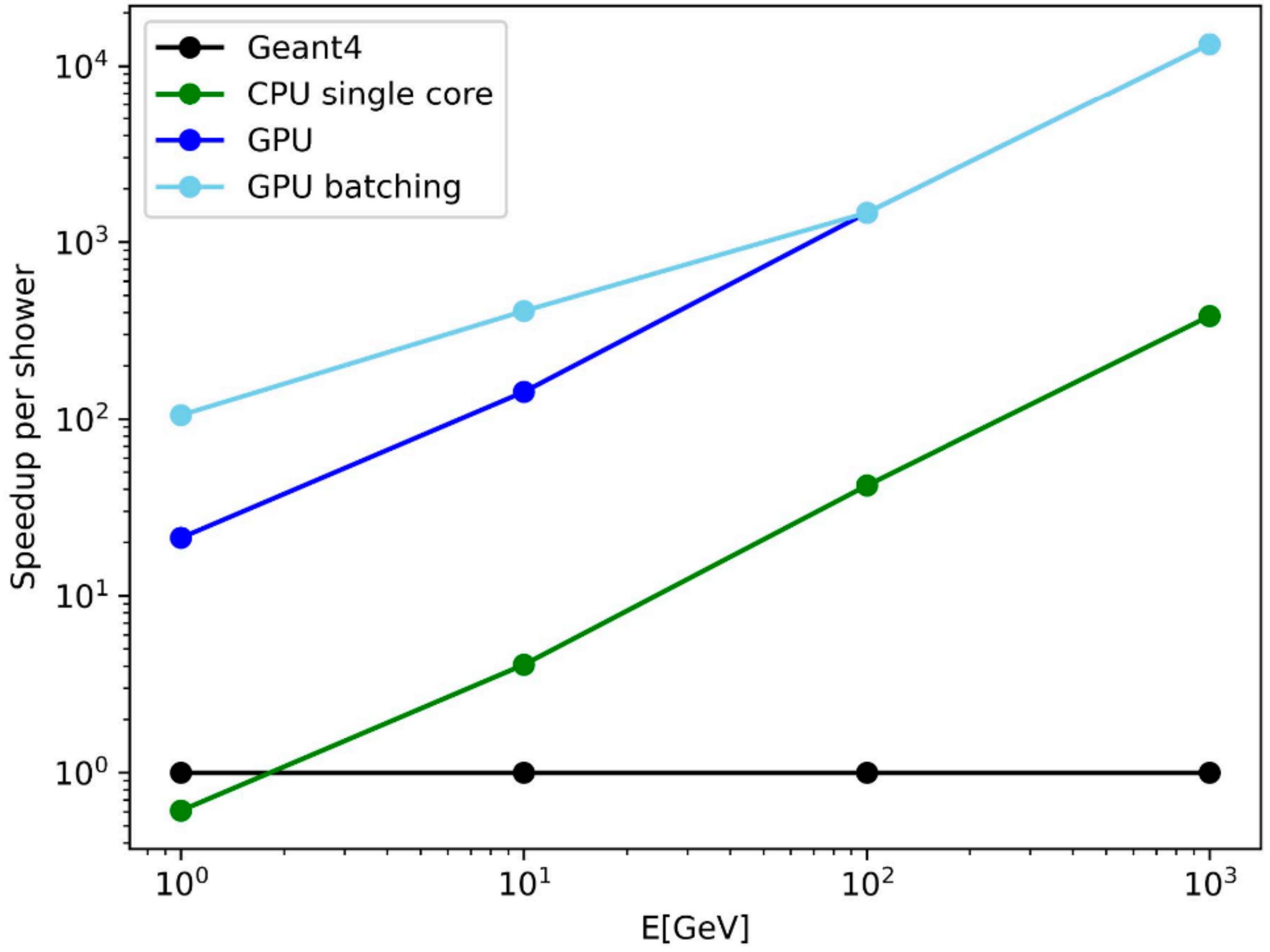
➡ Adaptation needs x10 less iterations in training

CaloDiT: Model Inference

- Diffusion models need multiple diffusion steps, slow sampling process.
- Consistency distillation:
  - a single diffusion step



https://arxiv.org/abs/2303.01469



Speedup relative to Geant4

Collaboration with the Experiments

- ATLAS
  - Support work on FastCaloSimV2 (classical fast simulation) (**DONE**)
  - Implement data structure allowing to test both VAE and transformer based model (CaloDiT) (**DONE**)
- LHCb
  - Find the best working model for hadronic shower (**NOT DONE**, deprioritised)
  - Validate EM fast simulation with a model based on Par04 (modified VAE) and planning at full-scale production (see presentation at CHEP24)
- CMS
  - Implement data production sample allowing to test models for HGCAL fast simulation (**NOT DONE**)

Other Experiment Collaborations

- Oriented Crystal Detectors
  - speed-up oriented crystal detector simulation (**PARTIALLY DONE**)
    - first tests completed moving to test the VAE model
- Future Collider Detectors
  - Produce single shower EM datasets (for FCCeeCLD and FCCeeALLEGRO) (**DONE**, new objective)
  - Developments in DDFastShowerML (**DONE**, new)
    - Implement Par04-like mesh placements of hits
    - Integration of CaloDiT for FCCeeCLD

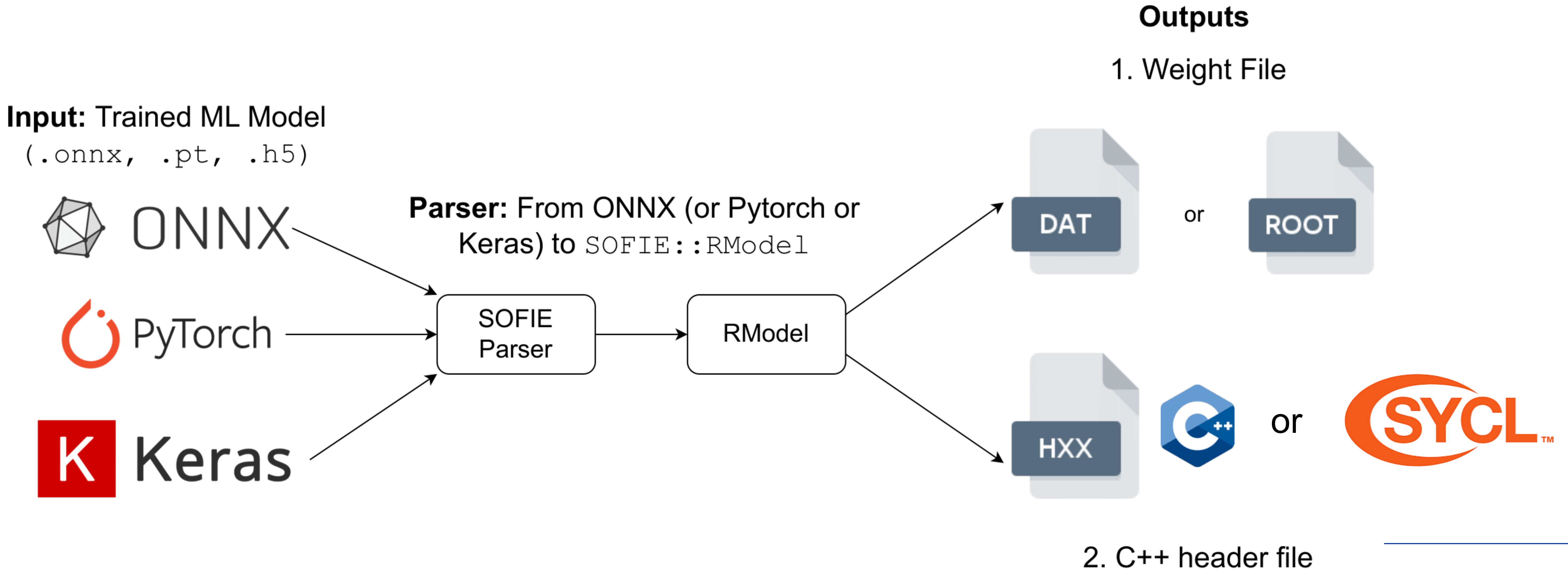
➡Presentations at ML4JETS (*PM*) and at FCC Workshop (*AZ*)

Community Efforts

- CaloChallenge:
  - Finalisation of CaloChallenge (with VAE and transformer models) (**DONE**)
  - Launch of a new challenge (**NOT DONE**, moved to 2025)
- Open data detector:
  - Demonstrator of ATLAS derived FastCaloSimV2 on ODD (**PARTIALLY DONE**)
  - Generation of combined tracker-calorimeter dataset (**NOT DONE**)
  - Generation of calorimeter single shower EM and hadronic (**DONE for EM**)

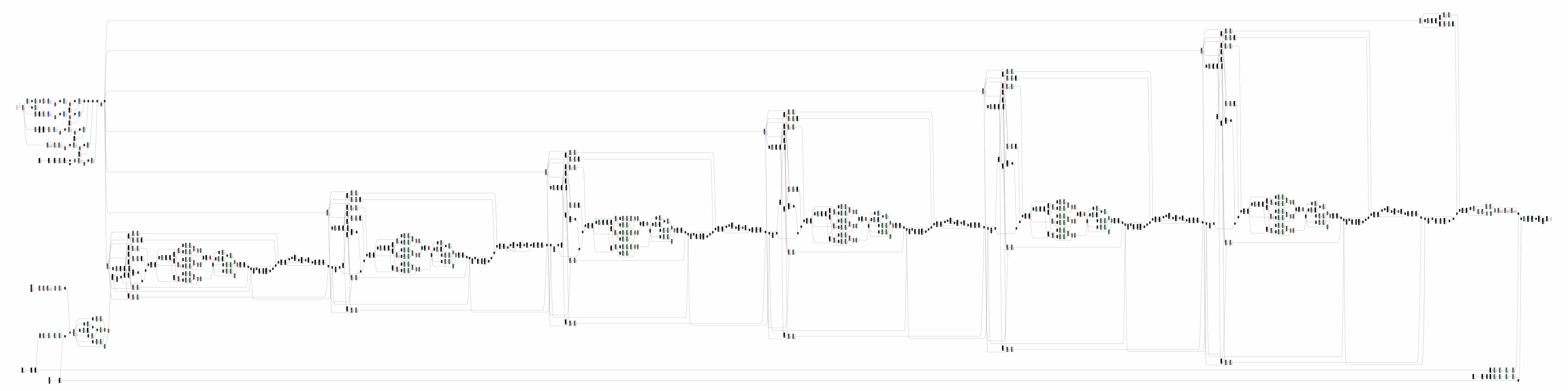
ROOT ML Software: SOFIE

- Add support in SOFIE for NVIDIA GPU's (**NOT DONE**)
  - already existing efficient solutions from NVIDIA (TensorRT)
  - will develop common interfaces for inference, as part of NGT 1.7
  - SOFIE has a SYCL prototype implementation
    - Presentation at ACAT2024
- Make SOFIE interoperable with hls4ml (**NOT DONE**)
  - low priority item, moved to 2025

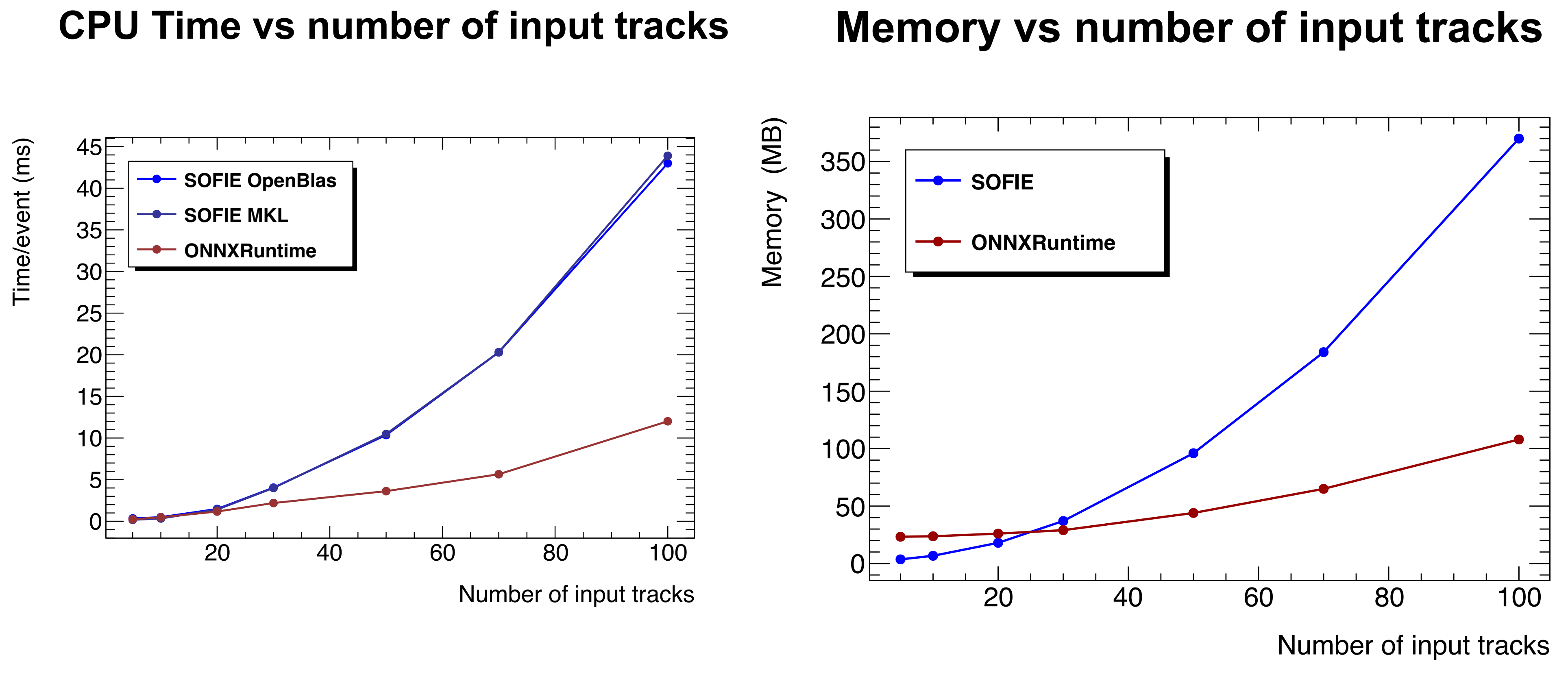


ROOT ML Software: SOFIE

- Extended operator support in SOFIE (**DONE**)
  - SOFIE can now parse ONNX based GNN's (ParticleNet (CMS) and GNN1 (ATLAS)) and CaloDiT model of fast simulation
    - can be used in experiment fast simulation applications
- Develop benchmarks with other ML inference solutions (ONNXRuntime and PyTorch) for both CPU and memory usage (**DONE, NEW**)
  - presentation at CHEP2024 (LM)



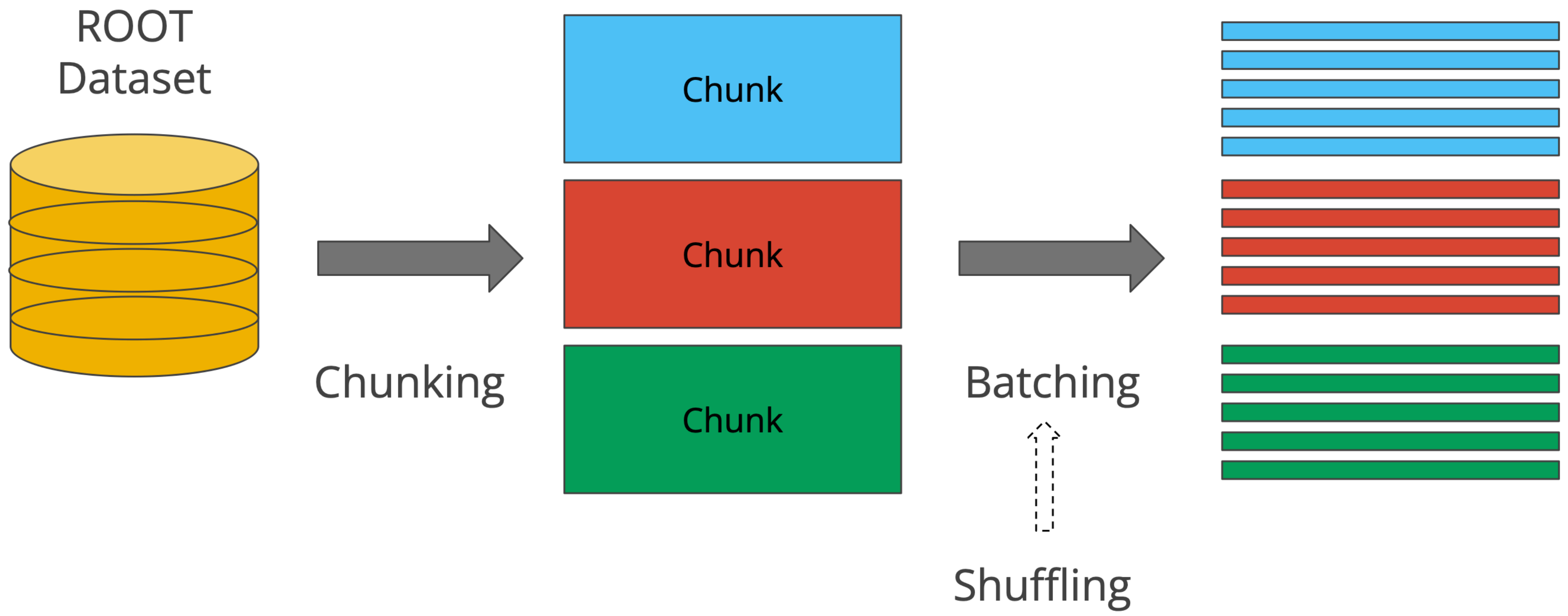
CaloDiT model



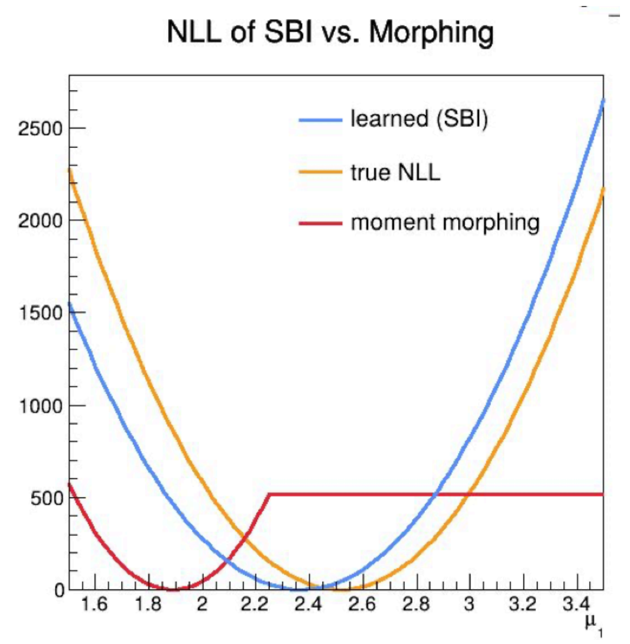
Benchmark results using ParticleNet (GNN tagger of CMS)

Other ROOT ML Software

- Completed first version of RBatchGenerator (**DONE**)
  - efficiency way of creating batches of data for ML training directly from ROOT files
  - integrated with RDataFrame
  - presented at CHEP2024

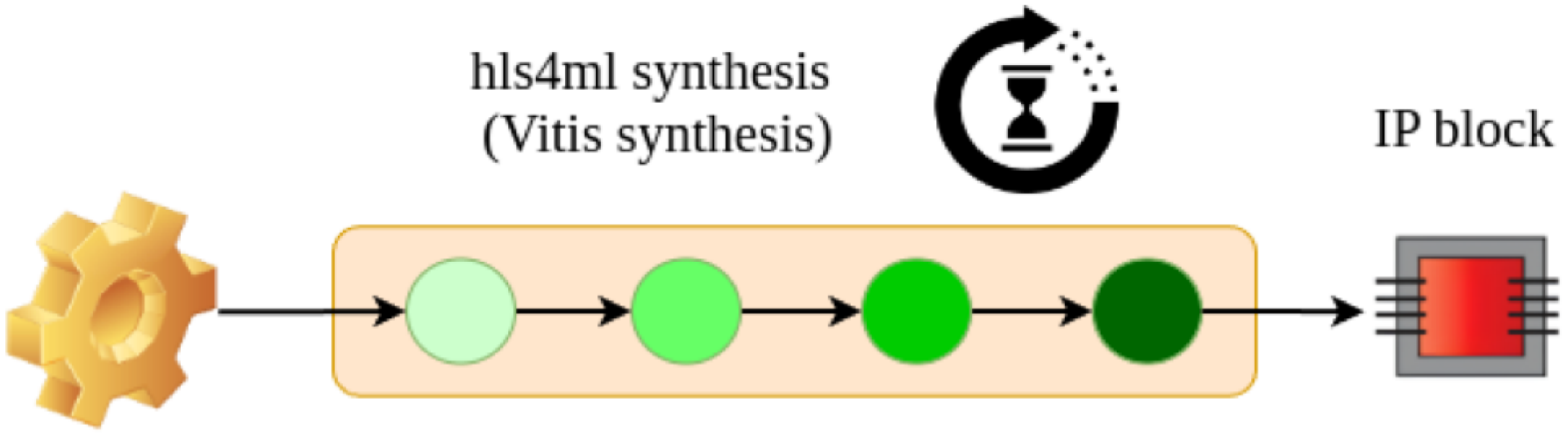


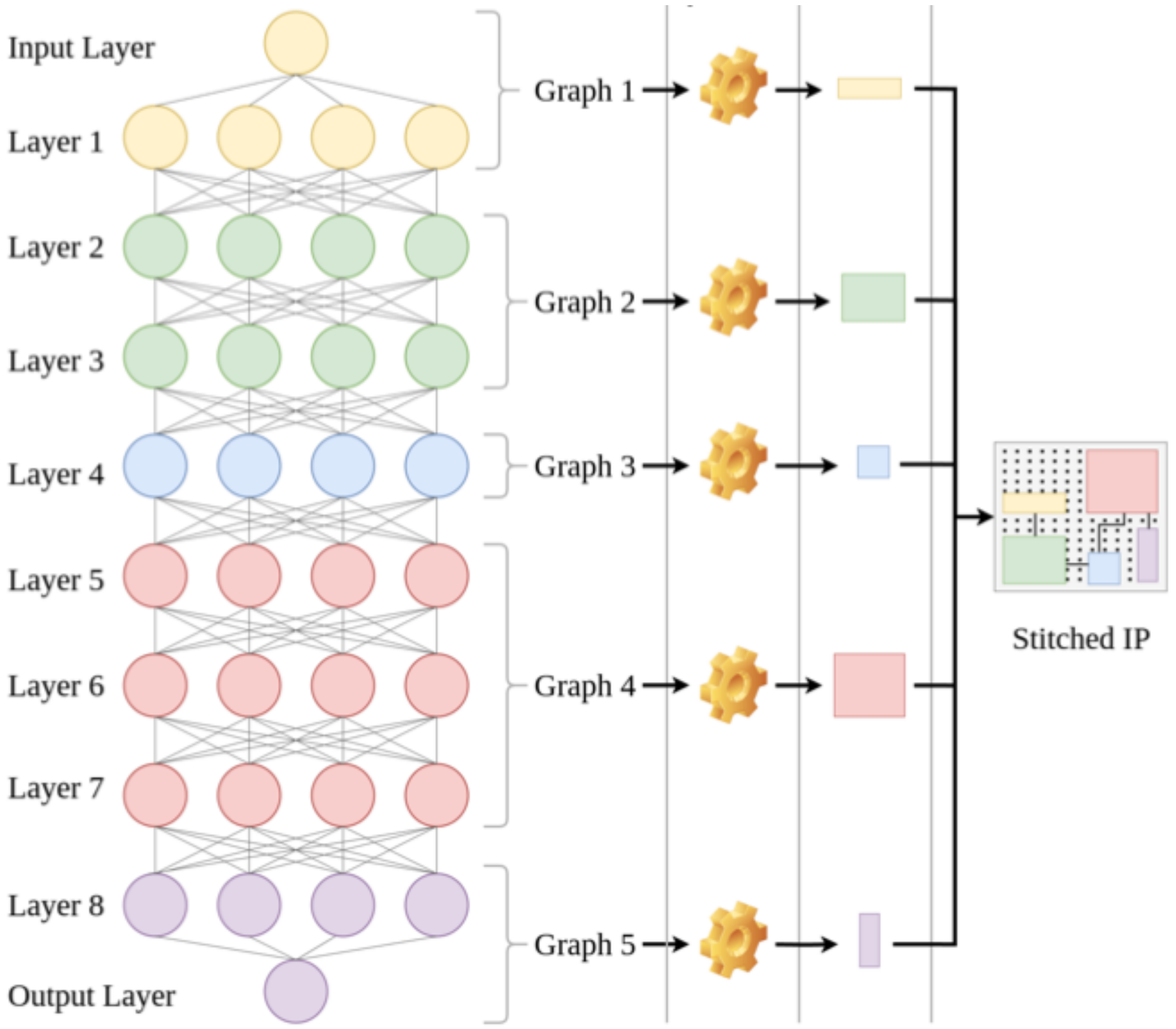
- Released a new implementation of RBDT (efficient inference for BDT)(**DONE**)
- First prototype of SBI with RooFit (**DONE, NEW**)
  - Support of Python functions in RooFit to integrate ML tools
  - presentation at CHEP2024



NGT Activities

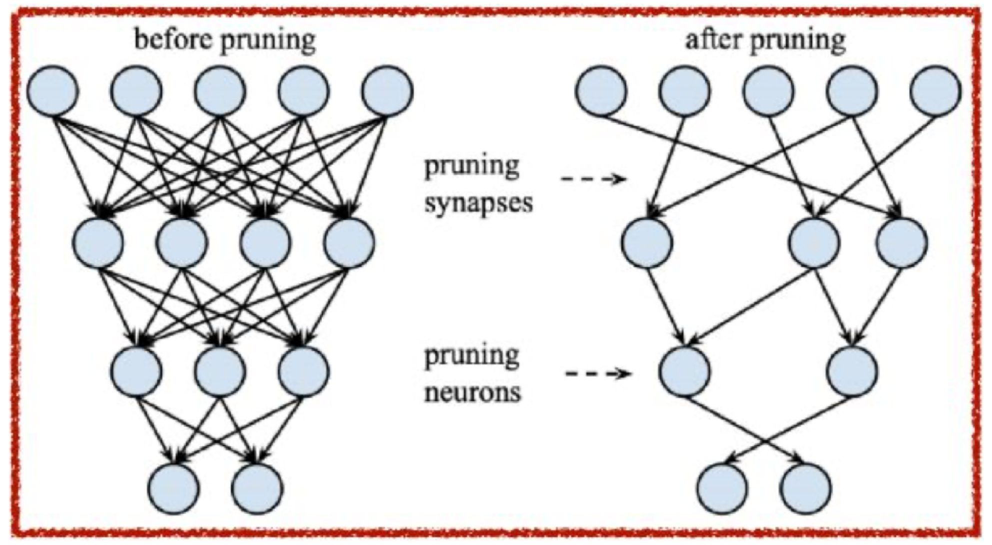
- Newcomers founded by NGT started only in September/October
- **NGT 1.2**
  - Goal: develop hardware-efficient neural networks and tools for FPGA and promote collaborative effort
    - hls4ml : automatic translation of trained ML models to optimised HLS code
  - Organised community workshop in September (milestone of NGT project)
    - received key feedback from the community
  - First achievements:
    - reducing hardware synthesis time (synthesis of complex models can take hours)
      - partition model in hls4ml into smaller independent subgraphs, which can be synthesised in parallel

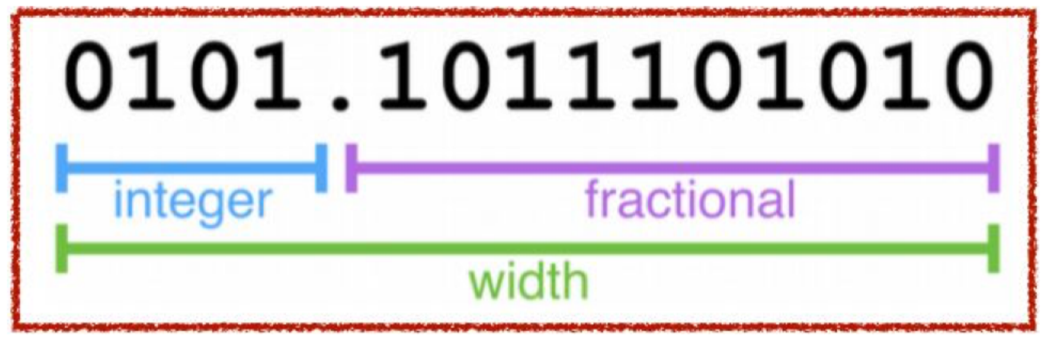


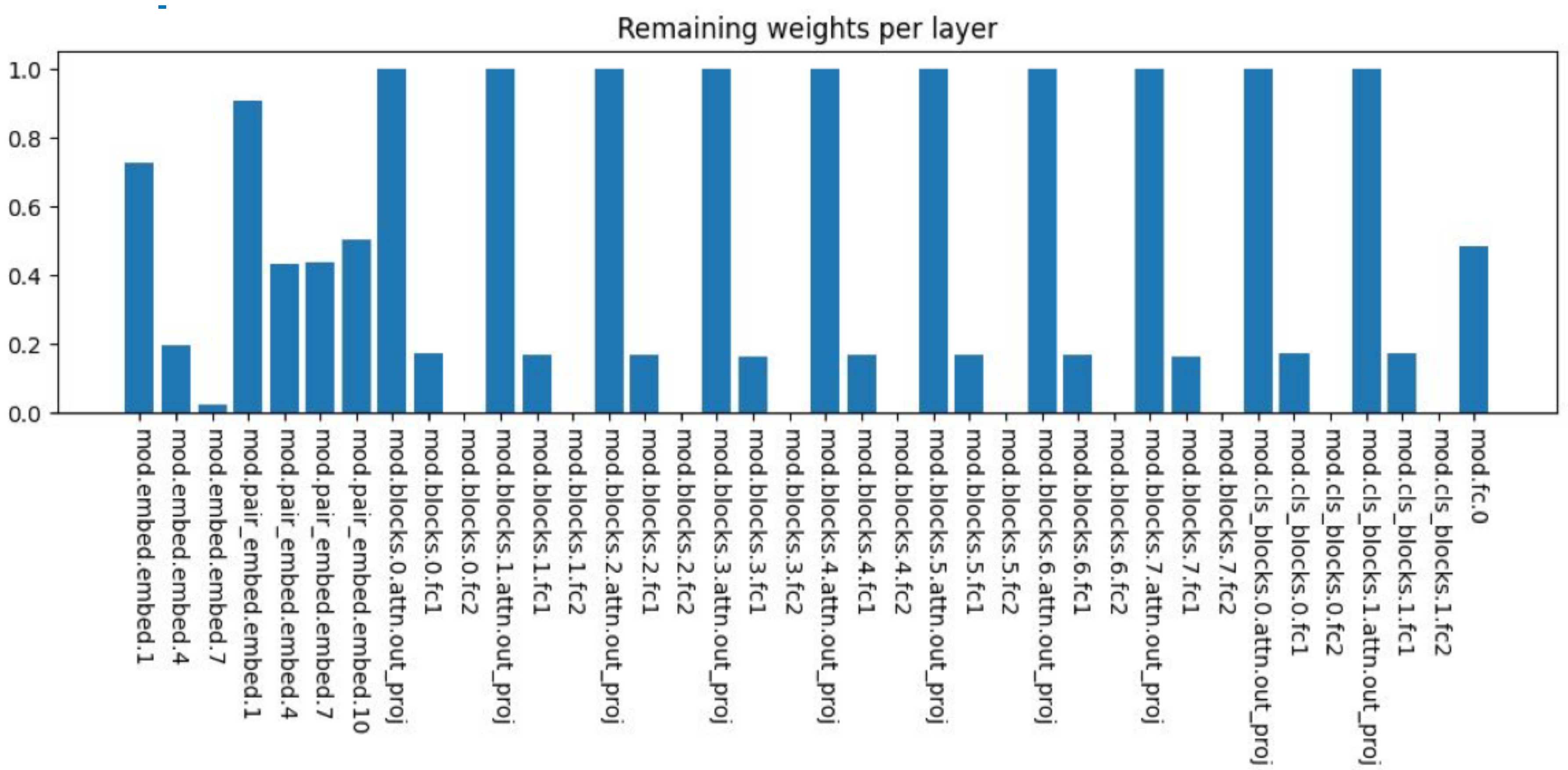


NGT Activities (1.3)

- **NGT 1.3**
  - goal: develop model compression methods and common interface to users to make it easy to adopt model compression techniques
    - develop software library and tools for hardware-aware training of neural networks making use of compression methods
    - work on pruning and quantisation
  - First achievements:
    - Investigation of pruning technique
      - 4 different algorithms impleme (with YAML configuration)
    - Testing on commonly used mode
      - ResNet and ParT







Plans for 2025

Fast Simulation: CaloDiT

- Further developments of CaloDiT
  - Priority 1:
    - Optimise the architecture and tune hyperparameters, trade-off between accuracy and inference speed
    - Integrate in Geant4 11.4 release and publish work in a journal paper
  - Priority 2:
    - Explore speeding up inference beyond using the consistency model
    - Test hadronic shower simulations with voxelisation and test on full simulation and then test point cloud model
      - plan to request a GSOC student working on point cloud model

Fast Simulation: Experiment Collaborations

Collaboration with the experiments

- ATLAS:
  - priority 1:
    - Provide help and support with FastCaloSim (classical parametrisation)
    - Test CaloDiT on ATLAS on both electromagnetic (multiple eta slices) and hadronic showers
      - need to explore best data representation to use for hadronic showers (e.g. point cloud)
    - Improve energy modelling using a normalising flow model on top of CaloDiT
  - priority 2:
    - Explore irregular voxalisation
    - Generalize inference code (working now only for GAN)
    - Perform comparison of different generative models at the reconstruction level (summer student project)
- CMS
  - Test CaloDiT for HGCal simulation (priority 2, if manpower is available)
- LHCb
  - Support usage of CaloDiT (priority 1)

Fast Simulation: Other Items

- Future Detectors
  - Priority 1:
    - Ensure Par04 models are working with FCCeeCLD and FCCeeALLEGRO
    - Continue developments of DDFastShowerML and combine fastsim with reconstruction for physics validation
  - Priority 2:
    - Provide infrastructure for hadronic shower
- Open Data Detector and Community Efforts
  - Priority 1:
    - Support the work on FastCaloSim demonstrator on ODD
    - Generation of combined tracker-calorimeter dataset
    - Launch next edition of CaloChallenge with new data, new detector (ODD), different data representation of the same showers (both EM and hadronic) and including reco validation

ML Software: SOFIE

- **SOFIE**
  - already we have a large number of operators, implement new ones according to needs from experiments
    - interest in using SOFIE also from non-CERN experiments (Belle-II, ePIC)
- Priority 1
  - Integrate SOFIE inference in Geant4 Par04 example
  - Perform Memory and CPU optimisations
    - optimise memory usage of memory analysing the generated computational graph
    - optimise processing time by profiling inference code and provide new more efficient operator implementations when needed
  - Prototype GPU porting using ALPAKA (part of NGT 1.7)
- Priority 2
  - Interoperability with HLS4ML, support converting a HLS4ML model to SOFIE

ML Software: Other Items

- Maintain benchmark of different ML inference solutions
- Promote the batch generator as a convenient interface for training
  - integrate into the currently developed training framework (priority 2)
    - b-hive from CMS and Salt/FTAG from ATLAS
    - integration with ml.cern.ch (based on kubeflow)
- Support ML workflows for Simulation-Based Inference (priority 2)
  - Integration of ML with statistical tools (RooFit)

NGT 1.7: Interfaces for ML Inference

- ML activity of 1.7: Main objective is to develop interfaces for ML inference for heterogenous architectures
  - initial discussions with LHC experiments to understand their needs
- Investigate various ML inference solutions for models used in experiments high level trigger (event filter)
- Plan to work on these initial items (priority 1)
  - Prototype GPU porting with ALPAKA in SOFIE
  - Develop interfaces to TensorRT and ROCm for NVidia and AMD GPUs
  - Explore also other possibilities (e.g. AITemplate, XLA) and SONIC
  - Benchmark the different solutions using the HEP HLT experiment models

NGT 1.2: ML for FPGA

- Priority 1
  - Finalize initial version of IP splitting and merge into the main codebase
  - Study use of Versal platform to offload parts of the individual IPs to the AI engines
  - Investigate sparse tensor representations on FPGA to foster future synergy with T1.3 on hardware-aware pruning techniques
- Priority 2
  - Explore advanced corner cases of IP splitting (residual connections, multiple inputs/outputs, per-IP reloadable weights, automating optimal split selection…)
  - Deploy a synthesis service platform to address user needs and automate synthesis tests
  - Finalize the hls4ml internal model format for potential integration with SOFIE

NGT 1.3: Model Compression

- Priority 1
  - Expand the current unstructured pruning methods evaluations to more HEP-based models, preferably in collaboration with experiments and SFT and write a publication on the findings
  - Move on to structured pruning method evaluation, selecting methods suitable for FPGA and perhaps GPU implementations
  - Investigate combination of quantization methodologies with pruning by integration of S-QUARK quantization framework (developed by Caltech) or implementation of "FitCompress" algorithm (developed within KT CEVA project)
  - Restructure the code into a library that can be reused, or integrated into CMS/ATLAS training frameworks (b-hive and Salt) and provide tutorials
- Priority 2
  - Integrate metric visualisation with Kubeflow (ml.cern.ch) for interactive usage

Summary

- Successful year in 2024 with many objectives achieved
  - Very good start of the NGT activities, a lot done in just a few months
- Having a very ambitious program for 2025
  - hoping for a fruitful year despite some reduction foreseen in manpower (e.g fast simulation)
- Continue collaboration with experiments and AI/ML community
  - ML plays an increased role in experiment computing software
- In the longer term we could contribute (if manpower available) to additional items:
  - develop and maintain common software framework for ML training
  - be involved in challenges and maintaining benchmark datasets
  - development and evaluation of foundation models for HEP
    (see recent overview paper on Large Physics Models)

Backup Slides

POW Items in 2024

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.

from A. Ribon presentation

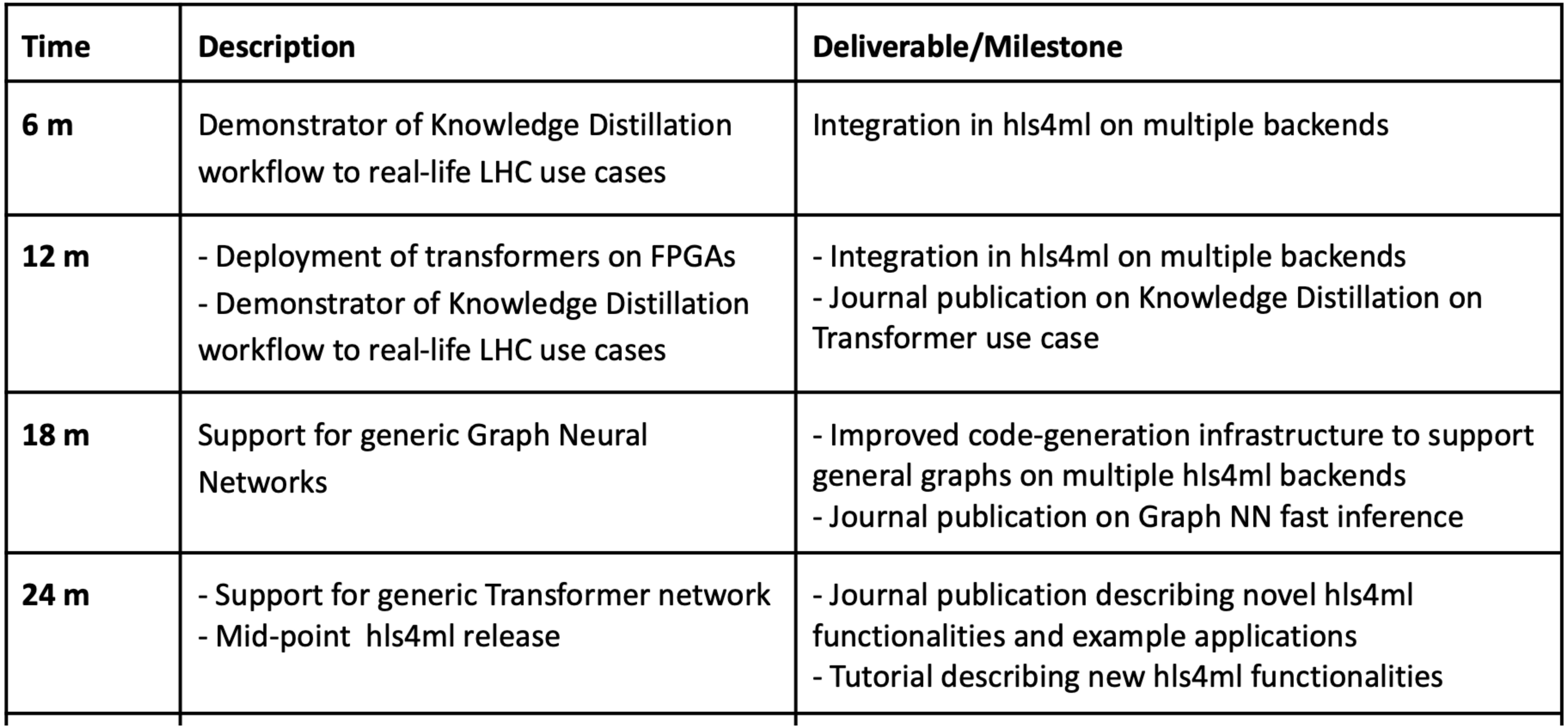
Fast Simulation

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

NGT Plans - 1.2

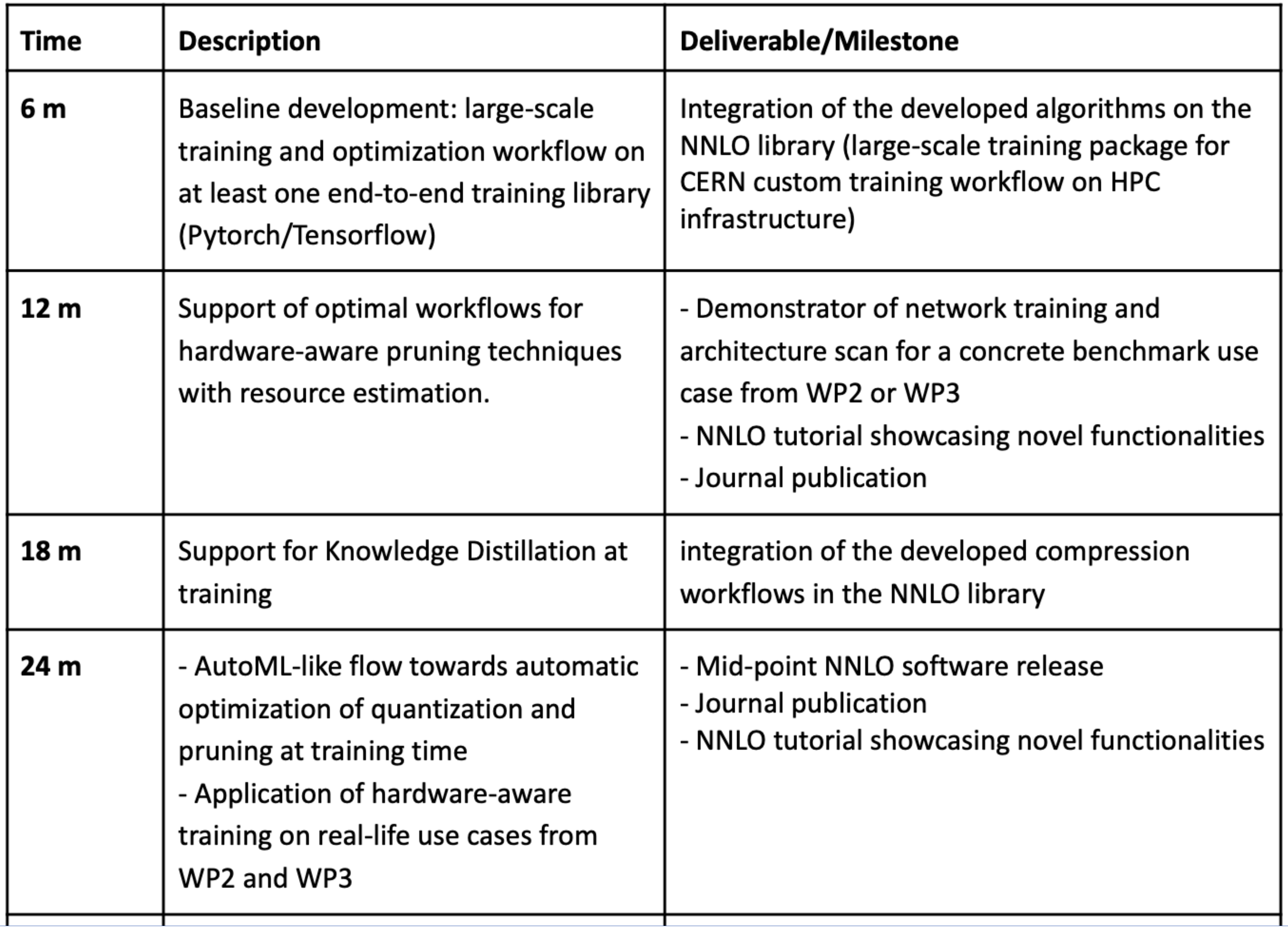
- Milestones from NGT proposal



Time	Description	Deliverable/Milestone
6 m	Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	Integration in hls4ml on multiple backends
12 m	- Deployment of transformers on FPGAs - Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	- Integration in hls4ml on multiple backends - Journal publication on Knowledge Distillation on Transformer use case
18 m	Support for generic Graph Neural Networks	- Improved code-generation infrastructure to support general graphs on multiple hls4ml backends - Journal publication on Graph NN fast inference
24 m	- Support for generic Transformer network - Mid-point hls4ml release	- Journal publication describing novel hls4ml functionalities and example applications - Tutorial describing new hls4ml functionalities

NGT Plans - 1.3

- Milestones from NGT proposal



Time	Description	Deliverable/Milestone
6 m	Baseline development: large-scale training and optimization workflow on at least one end-to-end training library (Pytorch/Tensorflow)	Integration of the developed algorithms on the NNLO library (large-scale training package for CERN custom training workflow on HPC infrastructure)
12 m	Support of optimal workflows for hardware-aware pruning techniques with resource estimation.	- Demonstrator of network training and architecture scan for a concrete benchmark use case from WP2 or WP3 - NNLO tutorial showcasing novel functionalities - Journal publication
18 m	Support for Knowledge Distillation at training	integration of the developed compression workflows in the NNLO library
24 m	- AutoML-like flow towards automatic optimization of quantization and pruning at training time - Application of hardware-aware training on real-life use cases from WP2 and WP3	- Mid-point NNLO software release - Journal publication - NNLO tutorial showcasing novel functionalities