ML Fast Sim Developments

Peter McKeown, Piyush Raikwar, Anna Zaborowska

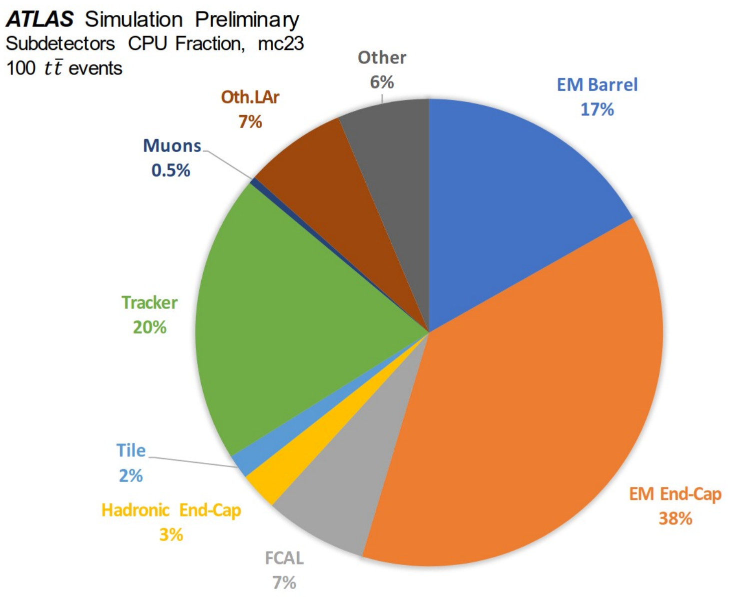
CERN, EP-SFT

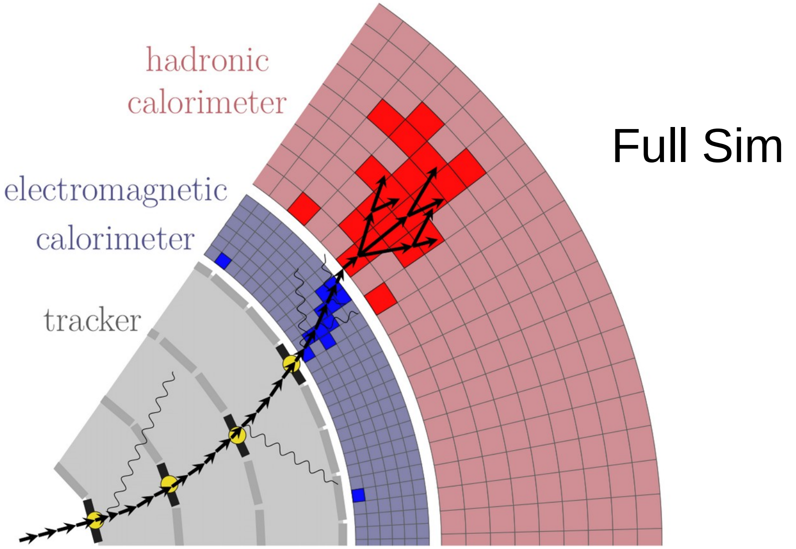
Geant4 Collaboration Meeting 2024

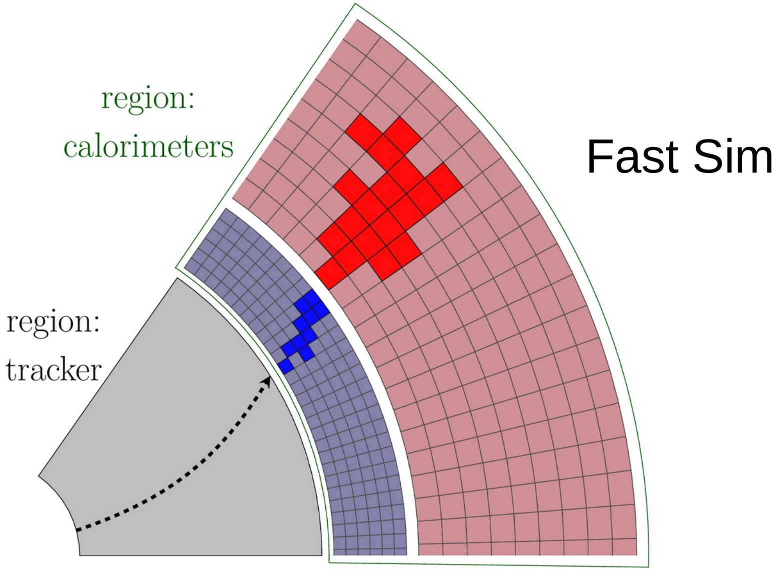
9.10.2024

Fast Simulation in HEP

- Current and future HEP experiments require ever larger quantities of simulated data
- Calorimeter shower simulation typically dominates compute time for full detector simulation
- Trade off some details from the full simulation for speed







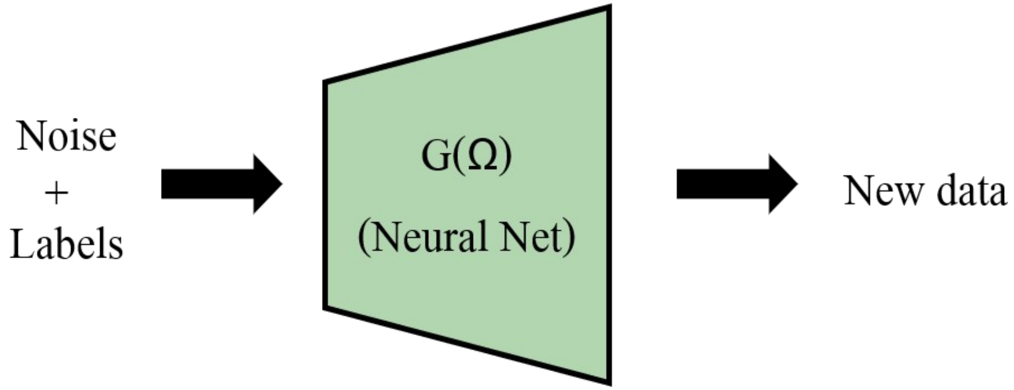
Optimizing the ATLAS Geant4 detector simulation for ACAT 2024, PLOT-SIMU-2024-03

ML Fast Simulation in HEP

- Generative ML models have seen significant attention for fast shower simulation
  - Used in production by ATLAS
  - Significant progress by LHCb and CMS

- Many developments have been **experiment specific**
  - Data representations
  - Models
  - Software ecosystems
- Difficult to propagate developments throughout the community

- **In Geant4 we are perfectly placed to reach across experiments!**

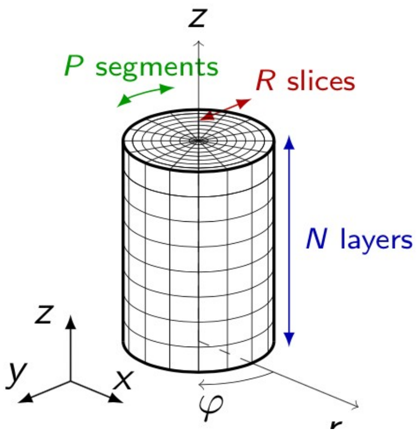


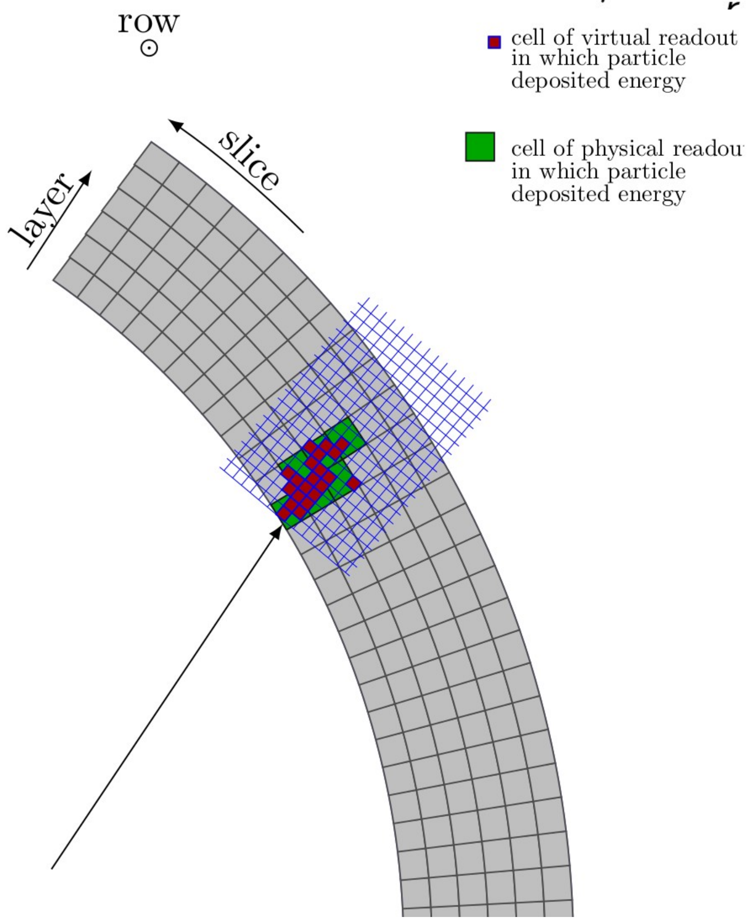
- … but models ultimately have to be evaluated in terms of **physics performance after reconstruction**
  - **Need to collaborate closely with experiments!**

ML Fast Simulation in Geant4

- Geant4 provides the extended Par04 example showing how to use **ML models in Geant4**
  - Virtual scoring mesh via parallel worlds
  - Inference libraries: ONNXruntime, libTorch, lwtnn
  - Can also be run on GPU (currently batch size 1)
- Datasets from Par04 provided the backbone for the CaloChallenge (2022)
  - Dataset 2: 6,480 voxels
  - Dataset 3: 40,500 voxels
- Provided a set of **common datasets and benchmarks** to enable the comparison of various ML models
- Total of **22** different models contributed
- Combined publication being finalised

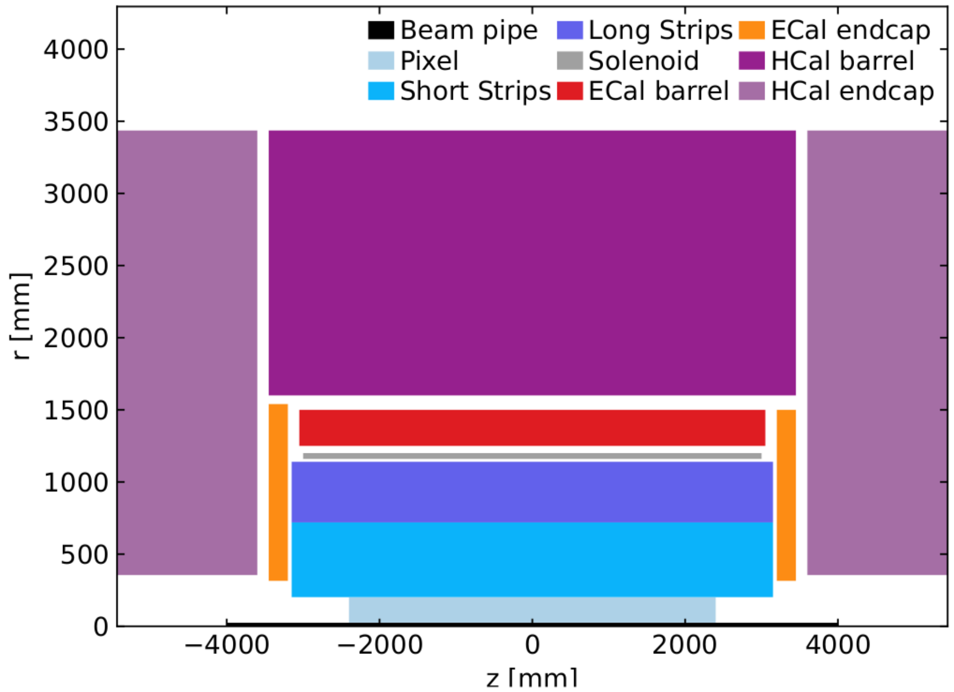
See Anna's talk at the previous Collaboration Meeting



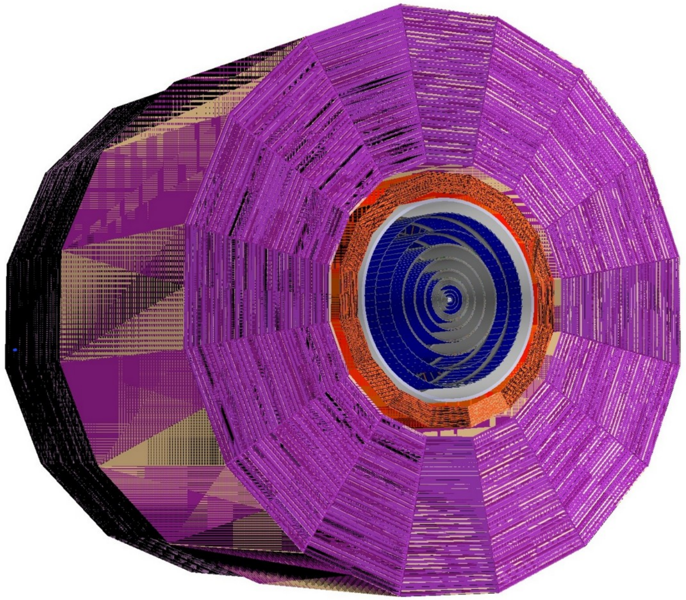


Open Data Detector (ODD)

- Open Data Detector: an open-access detector for algorithmic development and benchmarking
- Detector is described with DD4hep - a detector description toolkit used in HEP which provides an interface to Geant4
- Originally developed for the Tracking machine learning challenge (2018)
- ECAL (Si-W) and HCAL (Fe-Sci) with detailed geometries now implemented
- Plan to release open datasets for the next community challenge
  - Would also provide possibility to benchmark after reconstruction via DD4hep



See Anna's talk at ML4Jets 2023



More Generic Models: Motivation and Datasets

- Aim to reduce the computational resources required for developing an ML fast sim model
- Explore a '**foundation model**' model approach:
  - Train the model once on a large dataset, consisting of numerous different detector geometries
  - Provide it to users for fast adaption to specific use case

- Need a **common shower representation**
  - Make use of the virtual scoring approach from Par04
  - Electromagnetic showers to begin with

- Currently explored geometries (1M showers each):
  - Par04 SiW
  - Par04 SciPb
  - ODD
  - FCCee: CLD
  - FCCee: Allegro

CaloDiT: Model Architecture

- **Diffusion Transformer** (CaloDiT) model developed in EP-SFT (P. Raikwar) in collaboration with CERN Openlab and IBM Research (inspired by arXiv:2212.09748)

- Diffusion model:
  - Learn to gradually remove noise from data to generate shower

- Attention
  - Mechanism for modeling long-range correlations (adopted from NLP applications)

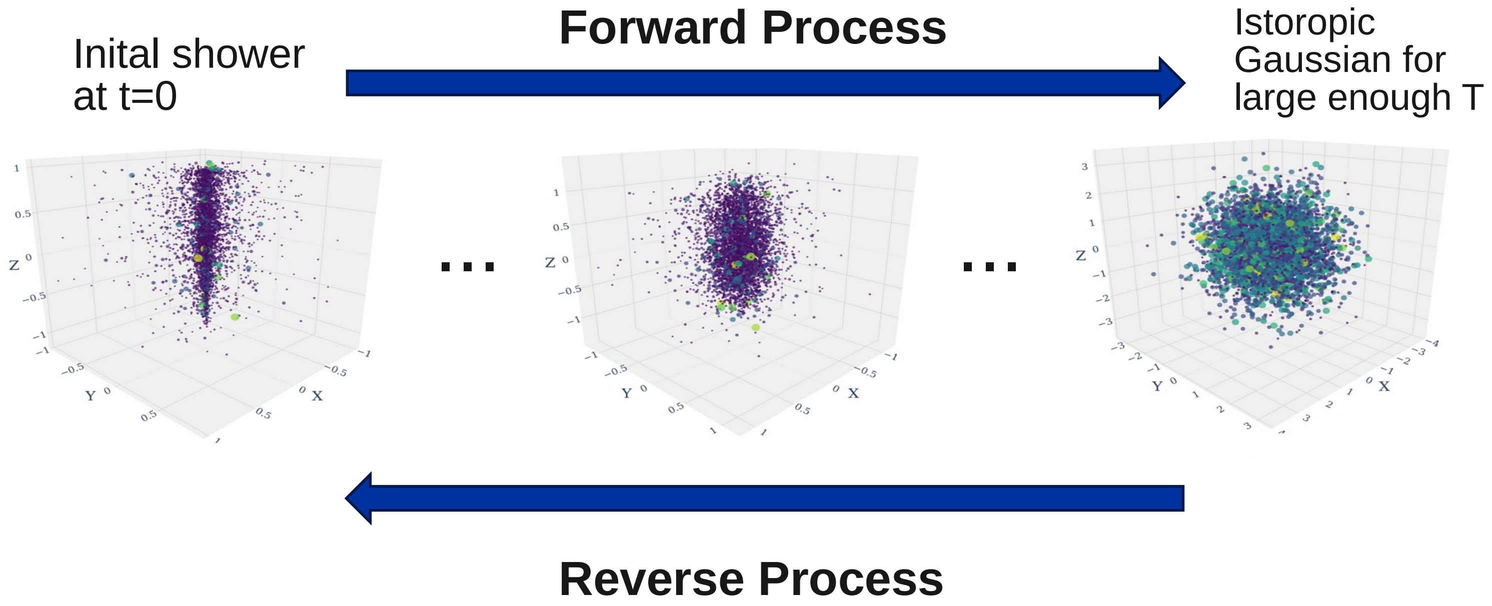


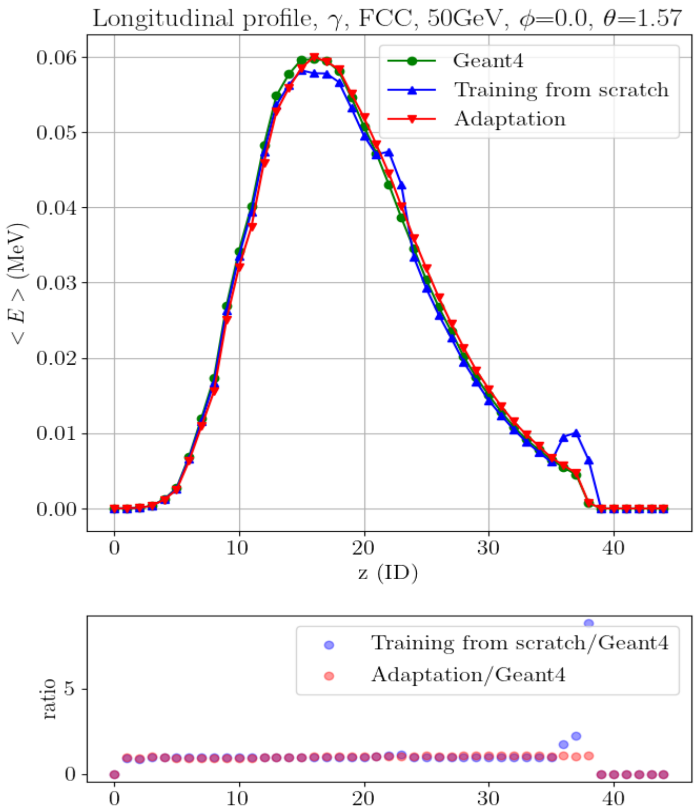
Figure adapted from
E. Buhmann, P.M. et al. JINST 18 (2023) 11

CaloDiT: Results

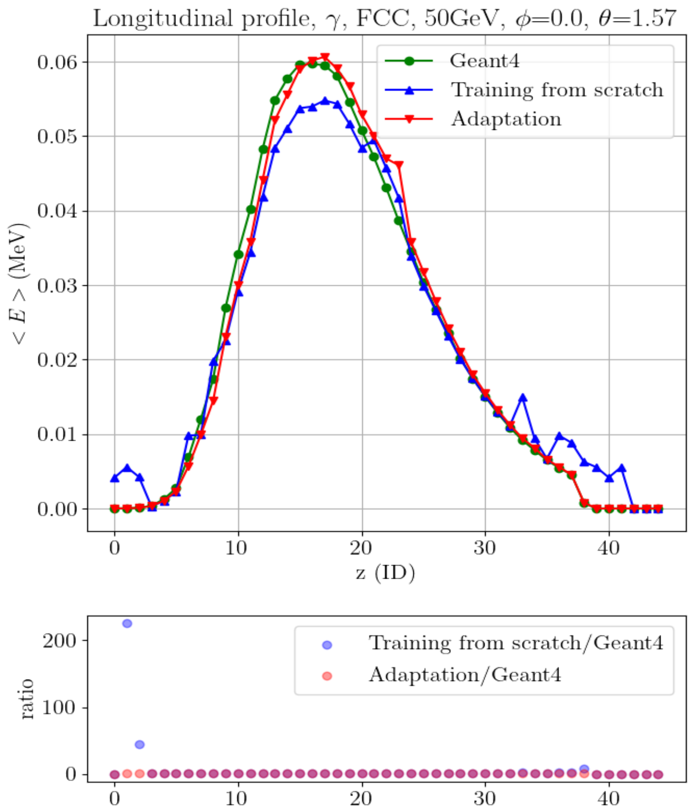
- Impressive performance in terms of physics observables
  - And adapting to new geometry is faster than training from scratch!
- However, iterative denoising (400 steps) makes inference slow...

At 200K samples
~25x less training time
<50% of the data

400K samples



200K samples

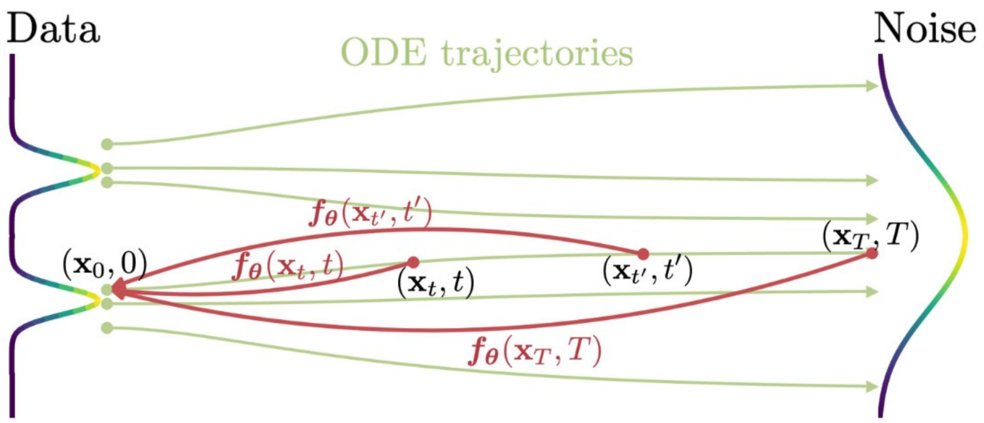


CaloDiT: Distillation

M. Piórczyński

- Number of diffusion steps dominates inference (i.e. shower generation) time- explored approaches to **distill** CaloDiT model
- With **consistency model**, maintain physics performance with **single diffusion step**
- **Significant speed-up** achieved with respect to full simulation

  - For **single photons** (standalone inference)*: - single core CPU **~1 order of magnitude faster**
    - **GPU usage** could (significantly) improve this yet further



Y. Song et al., *Consistency Models*, (2023) arXiv:2303.01469

* Details on timings in backup

DD4hep Integration: DDFastShowerML

- DD4hep toolkit widely used by future collider projects (FCC, CLIC, ILC, CEPC, IMCC, EIC…) via common **Key4hep** turnkey software stack
- Generic library DDFastShowerML recently included in Key4hep
  - Uses fast sim hooks in Geant4 via DDG4
  - Can be used with realistic, detailed detector models

- Aim for easy to use library which can accommodate all types of ML architectures

Trigger

- Fast Sim trigger
  - e.g. particle type, energy, geometry

Model

- Model-specific implementation of ML architecture
  - e.g. BIB-AE, Flow, Diffusion model

Inference

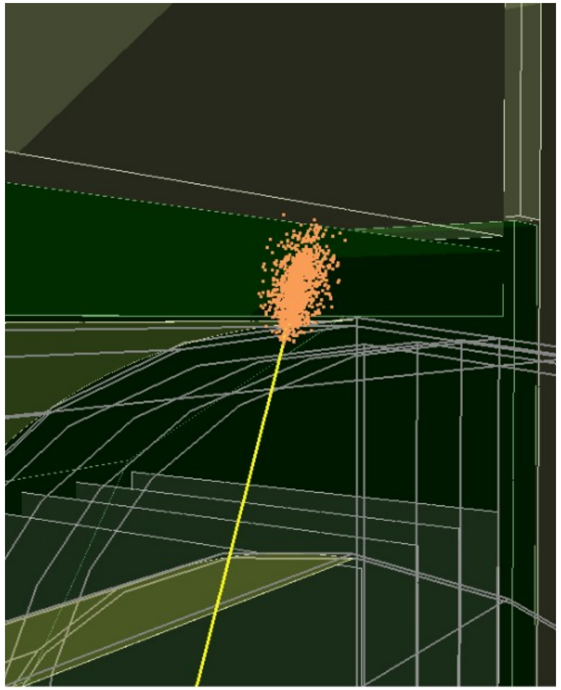
- Concrete inference in C++
  - ONNX, LibTorch etc…

Geometry

- Concrete placement in detector geometry
  - Endcap, barrel etc…

DD4hep Integration: DDFastShowerML

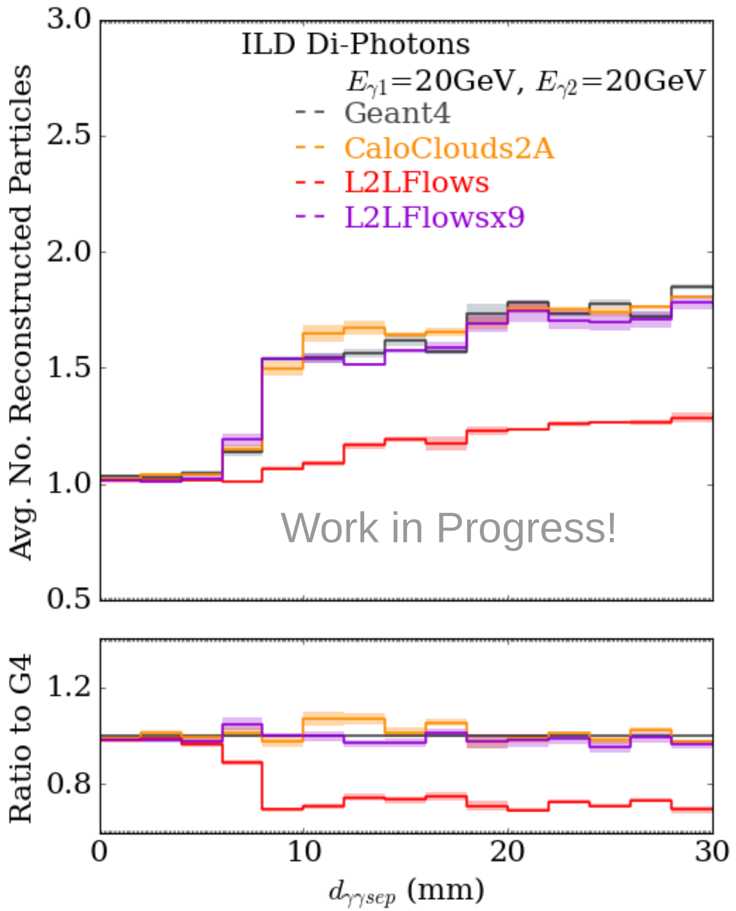
- DD4hep toolkit widely used by future collider projects (FCC, CLIC, ILC, CEPC, IMCC, EIC…) via common **Key4hep** turnkey software stack
- Generic library DDFastShowerML recently included in Key4hep
  - Uses fast sim hooks in Geant4 via DDG4
  - Can be used with realistic, detailed detector models
- Aim for easy to use library which can accommodate all types of ML architectures
- **Initial validation of CaloDiT** in scoring mesh (C. Zhu) integrated for FCCee CLD
  - Placement into detector readout (similar to G4 parallel worlds) is WIP

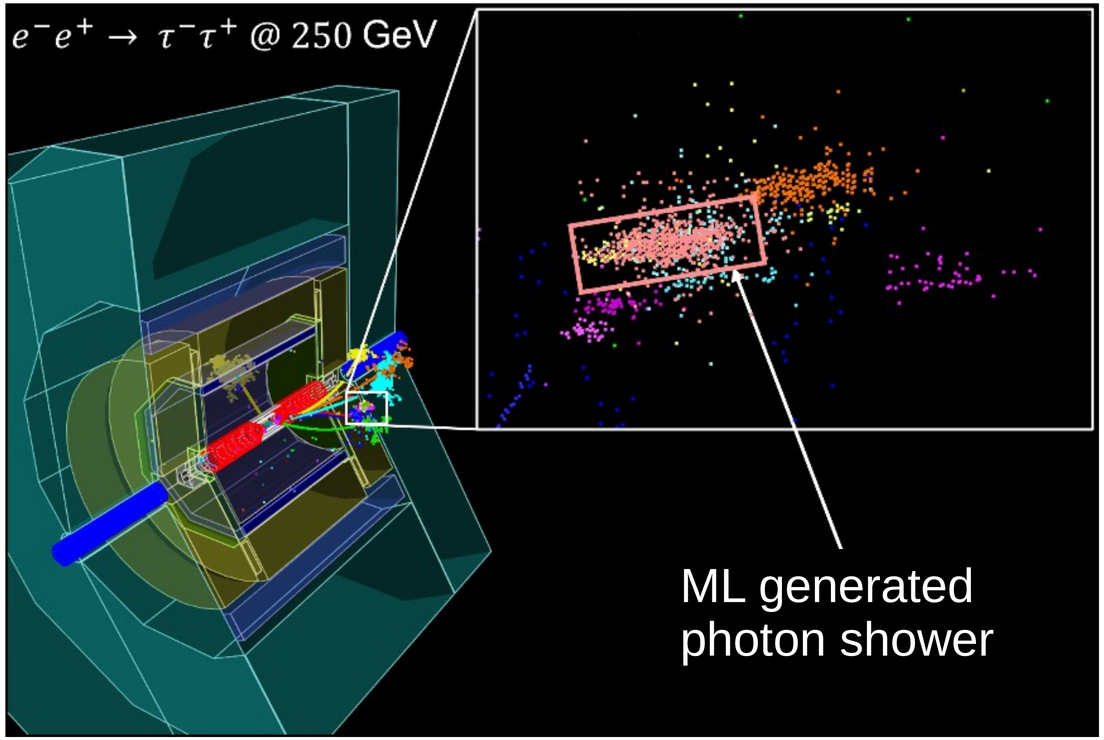


CaloDiT photon shower simulated in CLD with DDFastShowerML

Common Physics Benchmarks for Future Colliders

- With integration in DD4hep, now possible to define **common physics benchmarks**
- We are working with the community to start to define a common set





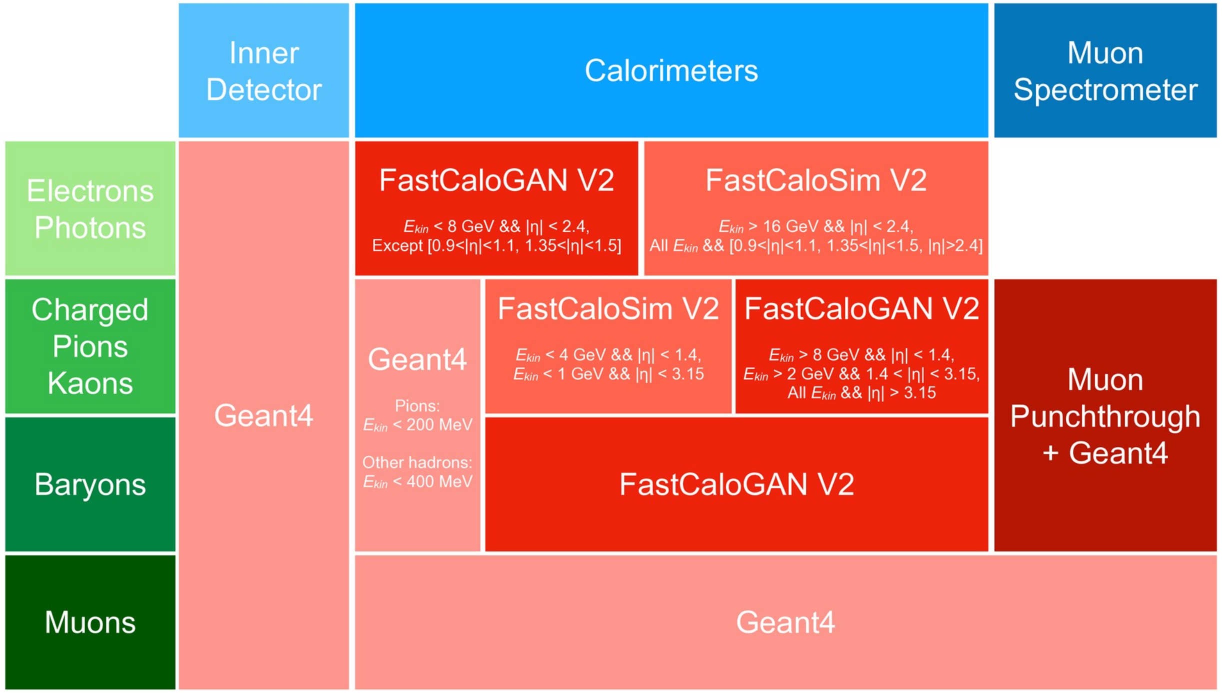
With input from T. Buss and A. Korol

Collaboration with Experiments: ATLAS

See more in Joshua's talk at ACAT 2024

Courtesy: J. F. Beirer

- ATLAS is already using generative models (FastCaloGAN) in production for Run 3 via **AtlFast3**
- Currently embedded in the Integrated Simulation Framework (**ISF**) to enable the use of multiple simulators in ATHENA



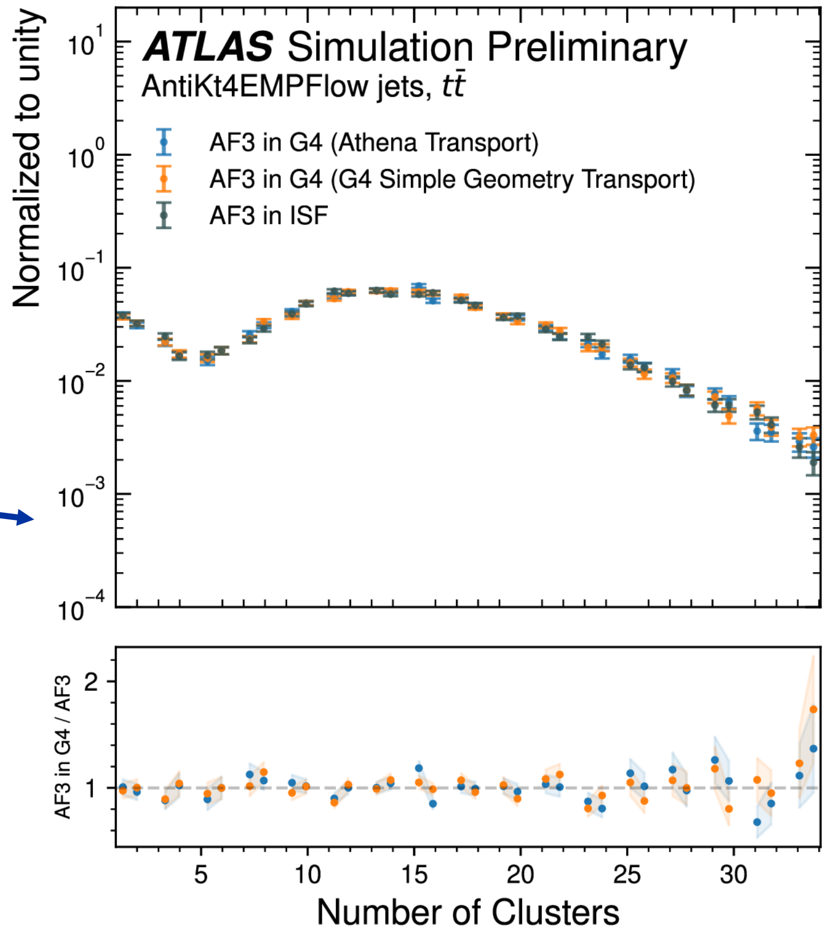
PLOT-SIMU-2024-04

Collaboration with Experiments: ATLAS

See more in Joshua's talk at ACAT 2024

Courtesy: J. F. Beirer

- ATLAS is already using generative models (FastCaloGAN) in production for Run 3 via **AtlFast3**
- Currently embedded in the Integrated Simulation Framework (**ISF**) to enable the use of multiple simulators in ATHENA
- Significant progress made on migrating to **Geant4 fast sim hooks**!
- Recently strengthened collaboration between EP-SFT and ATLAS Simulation Group to prepare the next generation of ATLAS FastCaloSim
  - Informed by CaloChallenge, **compare set of different models** to current FastCaloGAN (including CaloDiT)



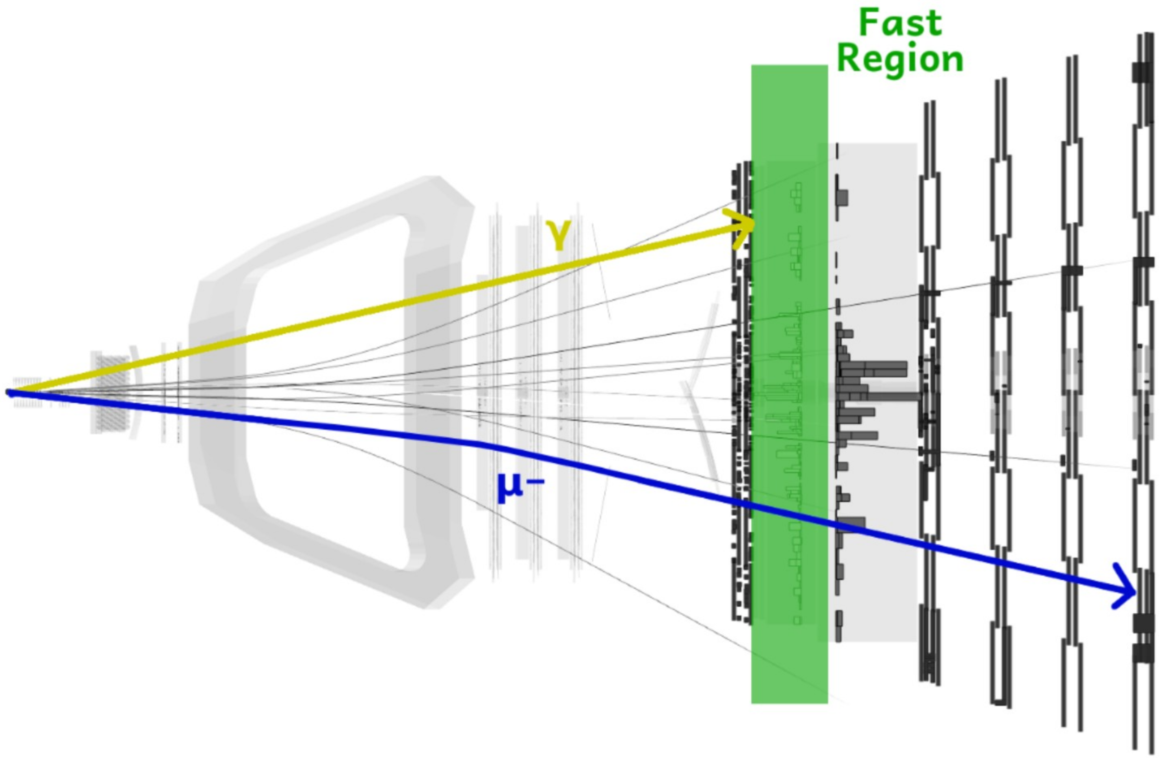
PLOT-SIMU-2024-04

Collaboration with Experiments: LHCb

Courtesy: M. Mazurek

- Significant progress on integrating **CaloChallenge-like** geometries into **Gaussino**
- Have so far explored a custom VAE as a 'pilot' model for e+/- and gamma for p=0.1-1000 GeV
  - In future can **explore other CaloChallenge models**

- Detailed **physics validation** for 4 different channels with significant electromagnetic component
- High level of agreement between Geant4 and ML
  - More details in Michał's upcoming CHEP talk



Poster-2021-1058

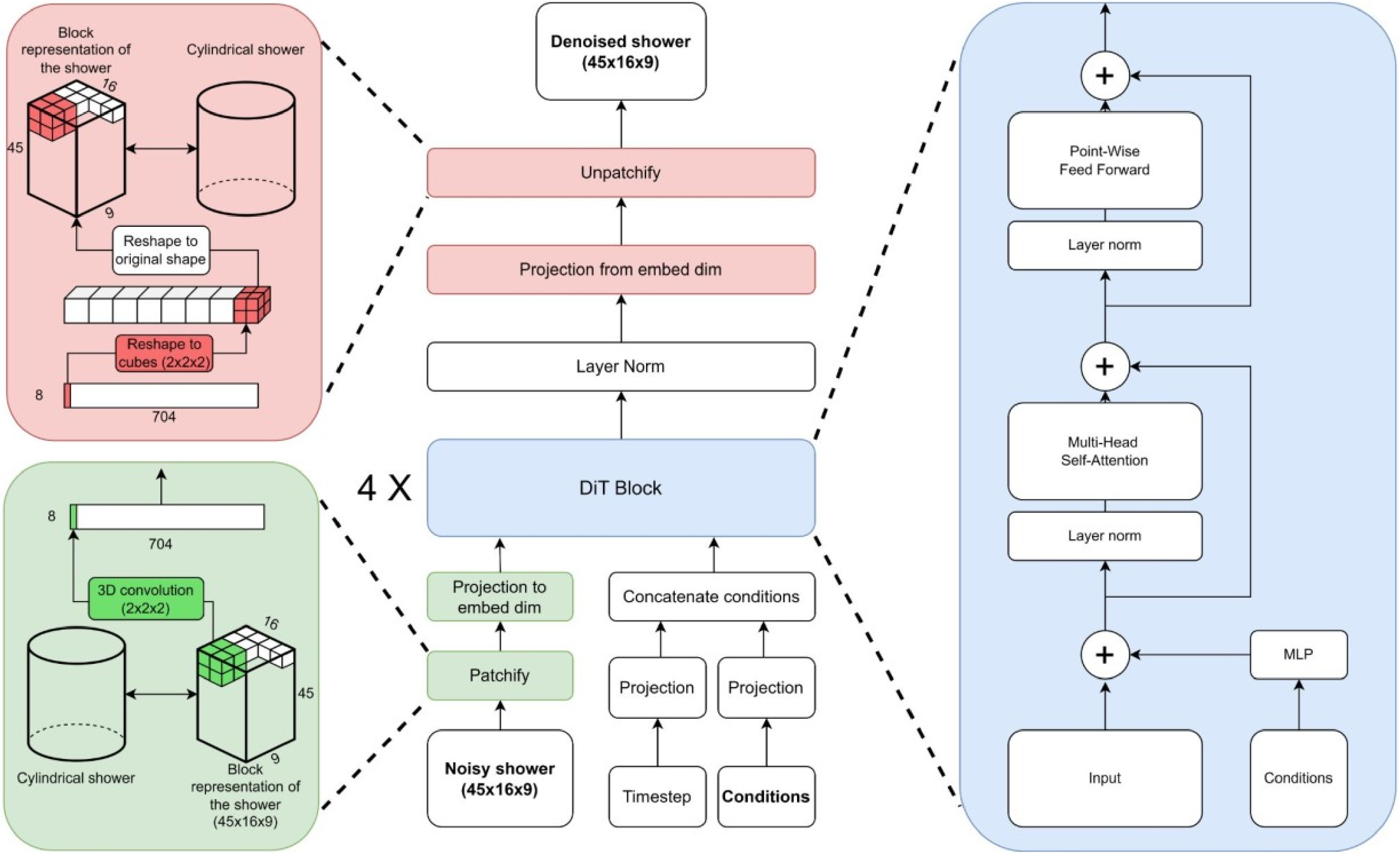
Summary

- Significant progress on ML fast sim
- Supporting community efforts:
  - CaloChallenge (heavy use of Par04) in final write-up stage
  - Significant contributions to the Open Data Detector
- Model R&D efforts
  - Exploring the potential of a more general approach to fast sim (CaloDiT)
  - Distillation of CaloDiT to achieve speed-up

- Directly collaborating with experiments
  - ATLAS: develop next generation of fast calorimeter simulation
  - LHCb: Detailed physics validation of CaloChallenge-like fast sim
  - Future Colliders: contributions to common library and physics benchmarks
  - CMS: Actively engaging to support fast sim efforts

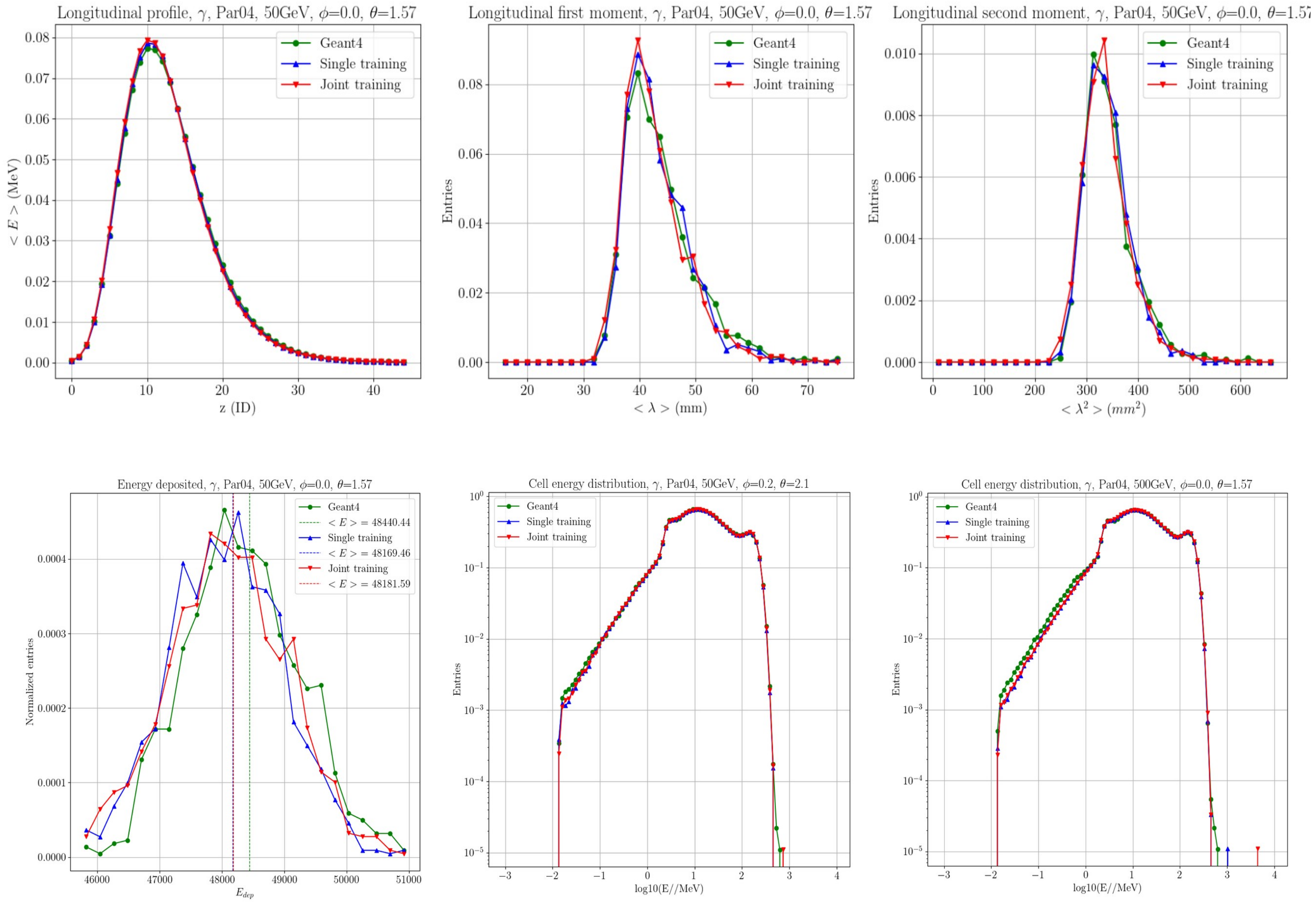
Backup

CaloDiT: Model Architecture



CaloDiT: Results

- Additional observables (Par04): training on multiple different geometries vs one



CaloDiT: Distillation

M. Piórczyński

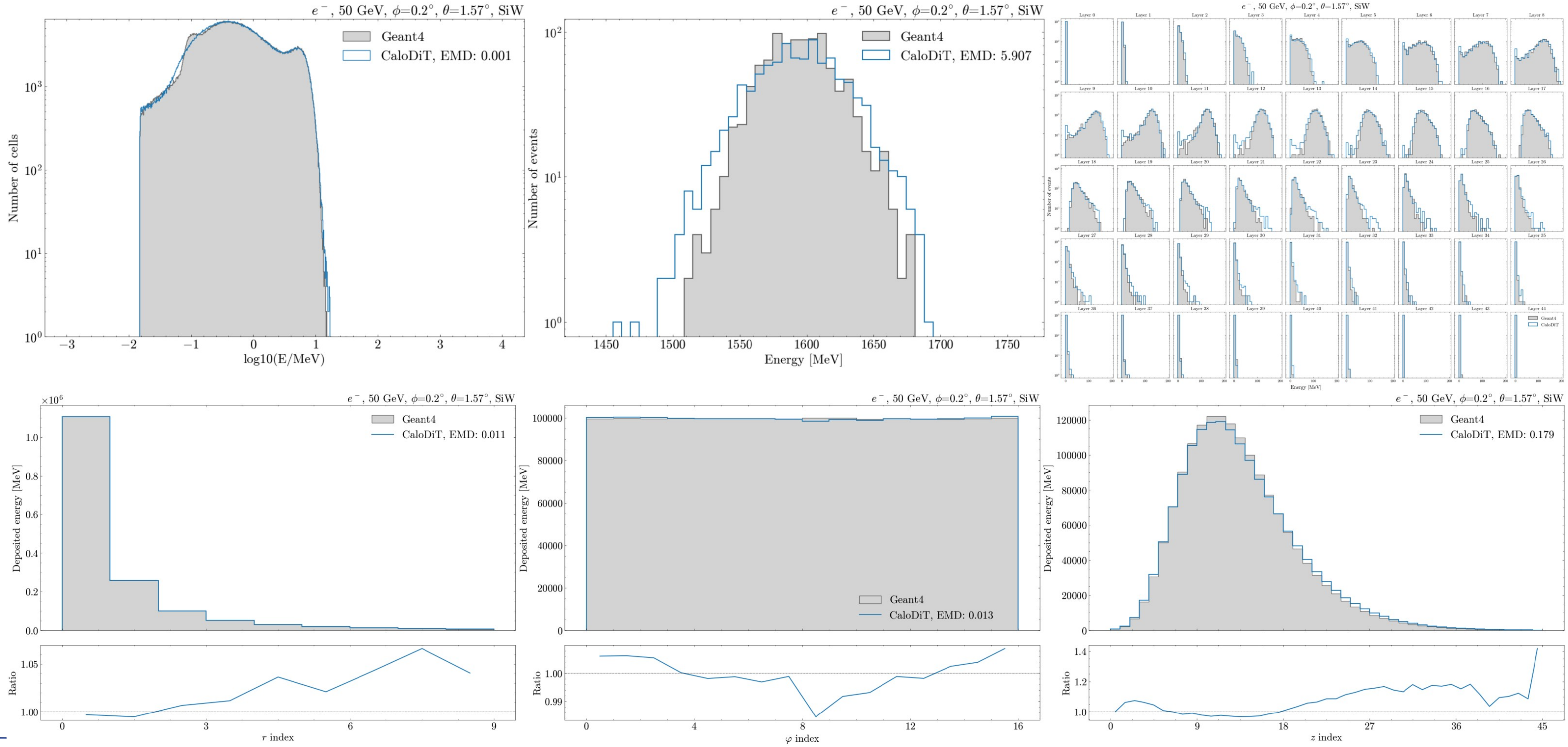
- Number of diffusion steps dominates inference (i.e. shower generation) time- explored approaches to **distill** CaloDiT model
- With **consistency model**, maintain physics performance with **single diffusion step**

- **First look at single photon** timings
- CPU: AMD EPYC 7282 16-core processor
- NVIDIA Quadro RTX 8000 with 48 GB of memory
- Caveat: **Model inference timings are standalone!**

Method	Device	Batch size	Time/Shower [ms]	Speed-up	Energy Range
Geant4 (Par04/ODD Geos)	CPU (single core)	N/A	1800-2300	x1	1-100 GeV (flat) (`)
			18300-22000	x1	1-1000 GeV (flat) (``)
CaloDiT (1 step consistency)	CPU (single core)	1	158.7±0.9	x11-14	`
				x115-139	``
	CPU (multi-core)	1	25.4±0.3	x71-91	`
				x720-866	``
	GPU	64	1.31±0.01	x1374-1756	`
				x13969-16794	``

CaloDiT: Distillation

M. Piórczyński



DD4hep Integration: DDFastShowerML