step2point: Optimising Point Cloud Representations for Fast Calorimeter Simulation

Peter McKeown[1,*], Anna Zaborowska, Martina Mozzanica, Siyu Chen, Tarun Tej Nandi, Piyush Raikwar, Saransh Chopra, Louis Vasseur, Matteo Casiraghi, Matteo Casadei, Salya Diallo

[1] CERN EP-SFT

[*] peter.mckeown@cern.ch

ML4Jets, 16.09.2026, Vienna

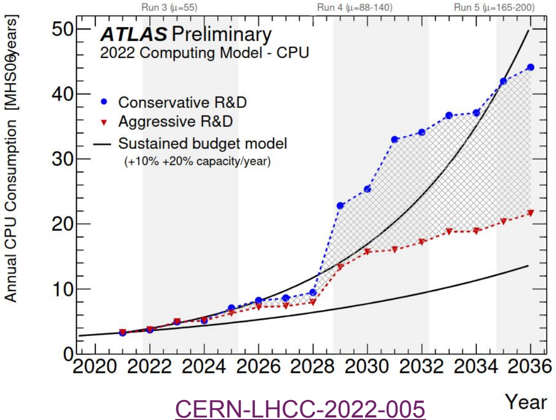
EP-SFT
Software Frameworks and Tools

GEANT4
A SIMULATION TOOLKIT

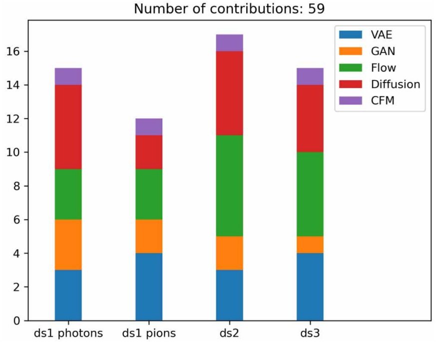
EP R&D

The Landscape of Fast Calorimeter Simulation

- Fast simulation is essential to meet the demands of future collider experiments
  - Generative models are the state-of-the-art modern tools
- The CaloChallenge compared a broad range of architectures
  - With a suite of datasets and validation metrics
- Implementations are already used in production or under development at the LHC experiments
  - See talks from Sitian and Florian



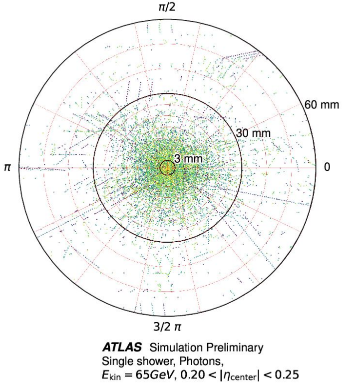
CERN-LHCC-2022-005



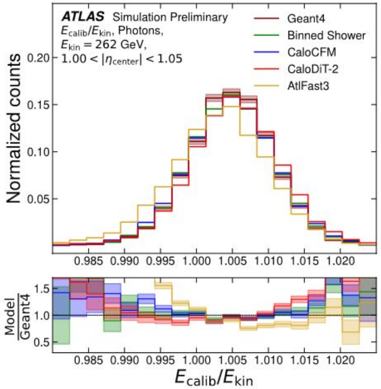
C. Krause *et al* 2025 *Rep. Prog. Phys.* **88** 116201

The Importance of Data Preparation

Example 1: Improving ATLAS Fast Sim



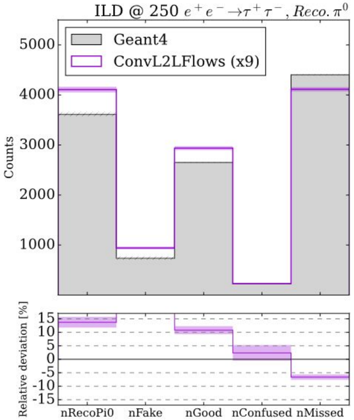
ATL-SOFT-PUB-2025-003



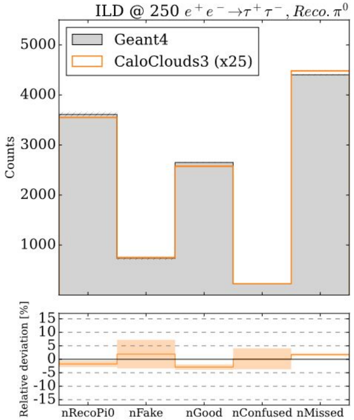
See talk from Florian!

- Improvements for ATLAS photon Fast Sim - new models and datasets
- Models can only be as good as the data they are trained on
  - **Optimising** shower **representations** is a **manually intensive**, **expert task**

Example 2: Highly Granular Calorimeter Showers



Regular Grid Model

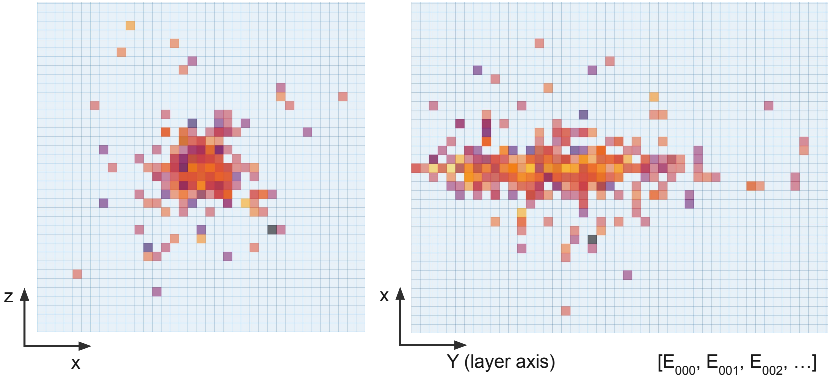


Point Cloud Model

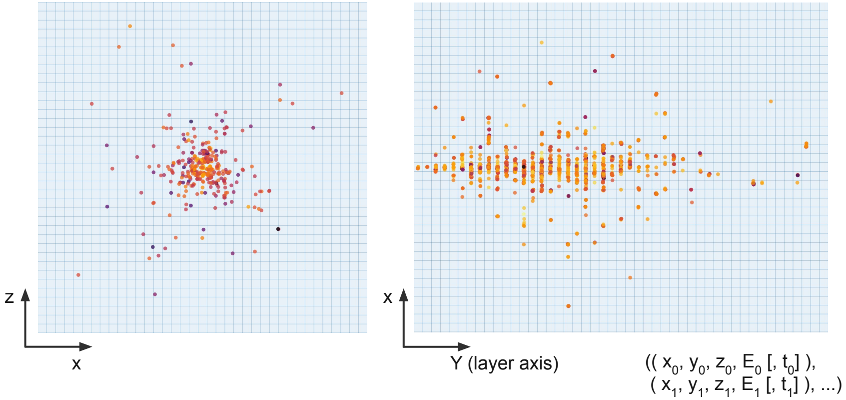
T. Buss et al., Accepted Phys. Rev. D

- Post-reco comparison of regular grid model to point cloud model
- Fewer representation constraints from **point clouds** can allow **better** post-reco **physics performance**

Point Clouds vs Image-like Voxelisations



Voxelisation



Point Cloud

Voxelisation

- Showers treated as “images”
  - Regular grid of pixels/voxels
- Detector cells or finer

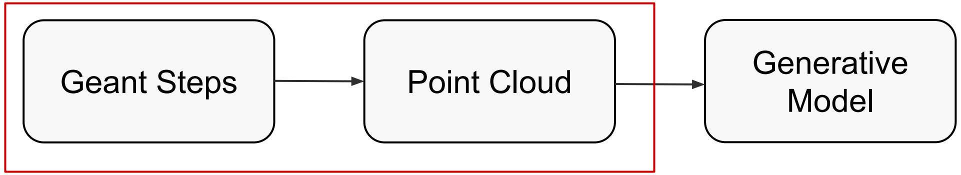
✓ Fixed number of voxels
✗ High sparsity- poor scaling to higher granularities

Point Cloud

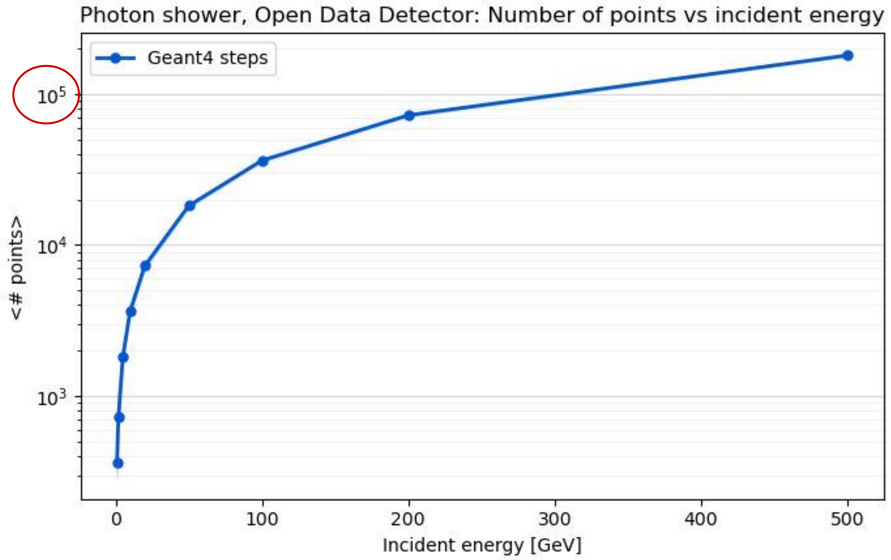
- Simulation provides much more detailed information (Geant4 steps)
- Can be detector-agnostic

✓ Efficient: only handle non-zero deposits
✓ Flexible: can preserve shower topology
✗ Variable number/position of deposits

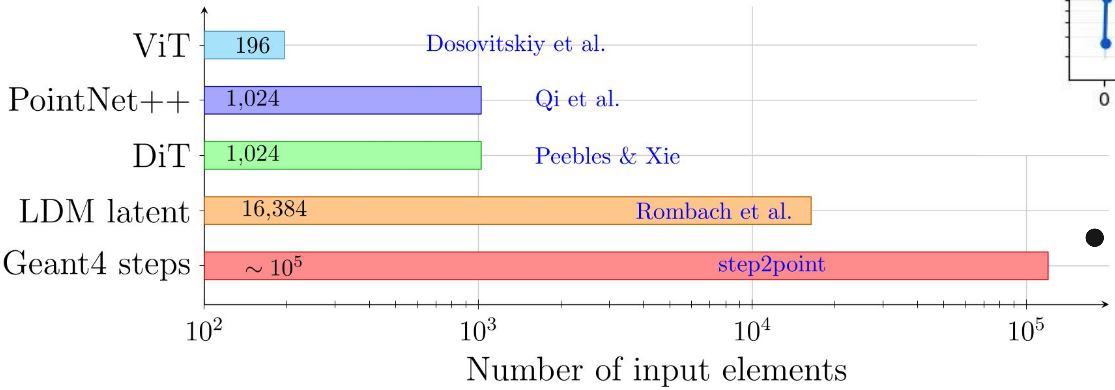
Why Not Just Use All the Steps?



Too large!



Generative models outside HEP



- Using raw Geant4 steps is unfeasible
  - Outside of designed input range
  - Can increase inference time

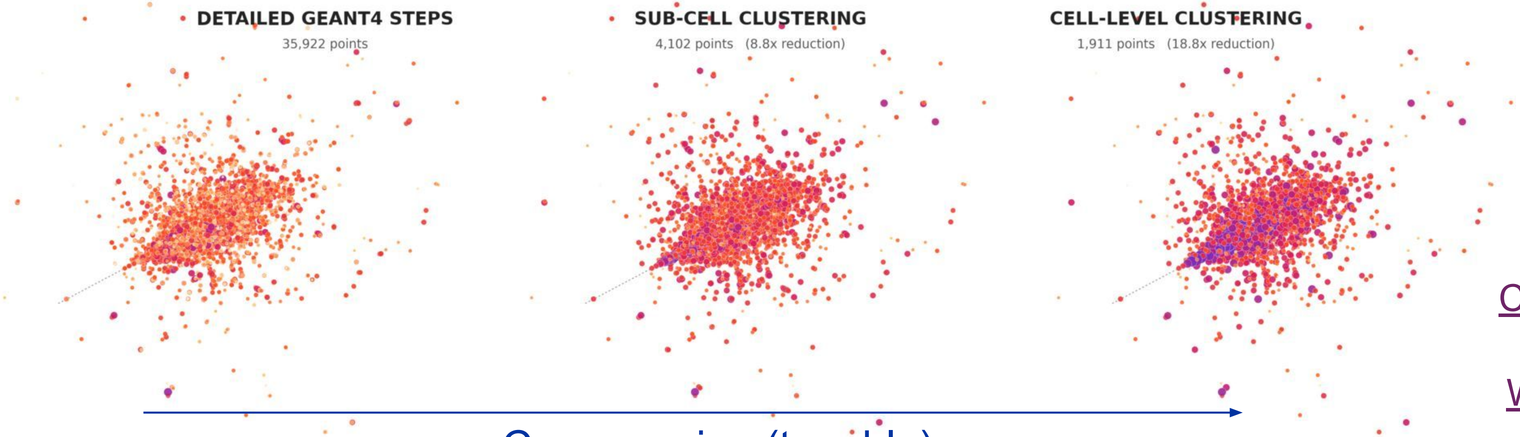
step2point

Geant Steps → step2point → Point Cloud → Generative Model

A library for turning detailed calorimeter shower deposits into **compact point-cloud representations** usable for ML-based fast simulation.

Goal 1: Find the **smallest set of points** that preserves the **physics of the shower.**
Goal 2: We aim to **provide infrastructure** that should/could work with **different detectors**.

Number of points



Compression (tunable)

[Code repo]

[Website]

Overview of the step2point Tool

01 Merge Within Cell

Sums all steps in the same detector cell into a single point.

02 Merge Within Regular Subgrid

Subdivides each cell into a regular N×M sub-grid, merging steps per sub-voxel.

03 Density-based Clustering in (x,y,z[,t]) Space

E.g.:

- HDBSCAN
- Agglomerative clustering

04 New Clustering Algorithm…

step2point can be extended by adding new clustering algorithms!

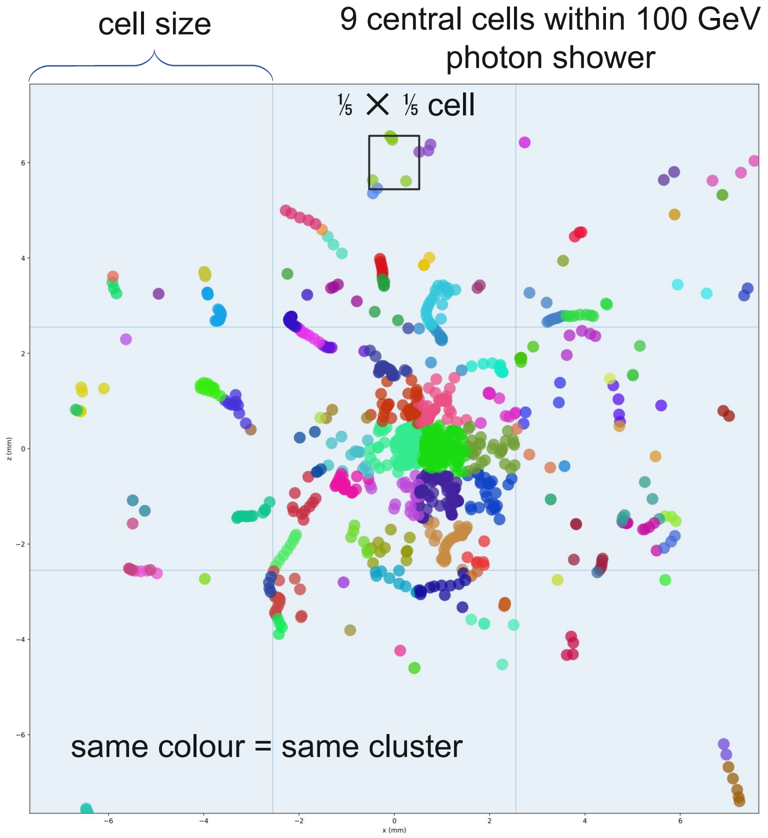
- Two different clustering modes are supported
  - Respect detector cells (default)
  - Allow cross-cell merging

- Aim to provide flexibility for users:
  - Tunable for target detector
  - Simple (simulation level) first validation included

Regular Sub Grid vs Density-based Clustering

- **02 - Regular subgrid**
  - Used to created data representations in existing ILD studies
  - Divide every cell into a fixed regular square subgrid, e.g. 5 x 5
  - Simple and deterministic
  - But may not be most efficient in general:
    - Pre-supposes cell geometry
    - Treats dense shower core and low-energy tails uniformly

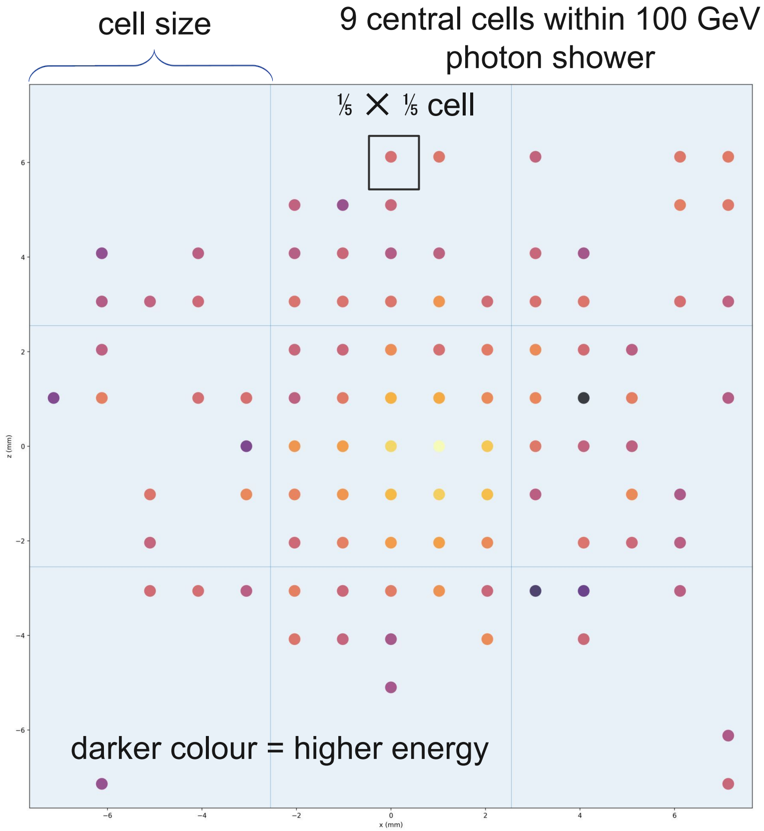
- Two modes provided:



Regular Sub Grid vs Density-based Clustering

- **02 - Regular subgrid**
  - Used to created data representations in existing ILD studies
  - Divide every cell into a fixed regular square subgrid, e.g. 5 x 5
  - Simple and deterministic
  - But may not be most efficient in general:
    - Pre-supposes cell geometry
    - Treats dense shower core and low-energy tails uniformly

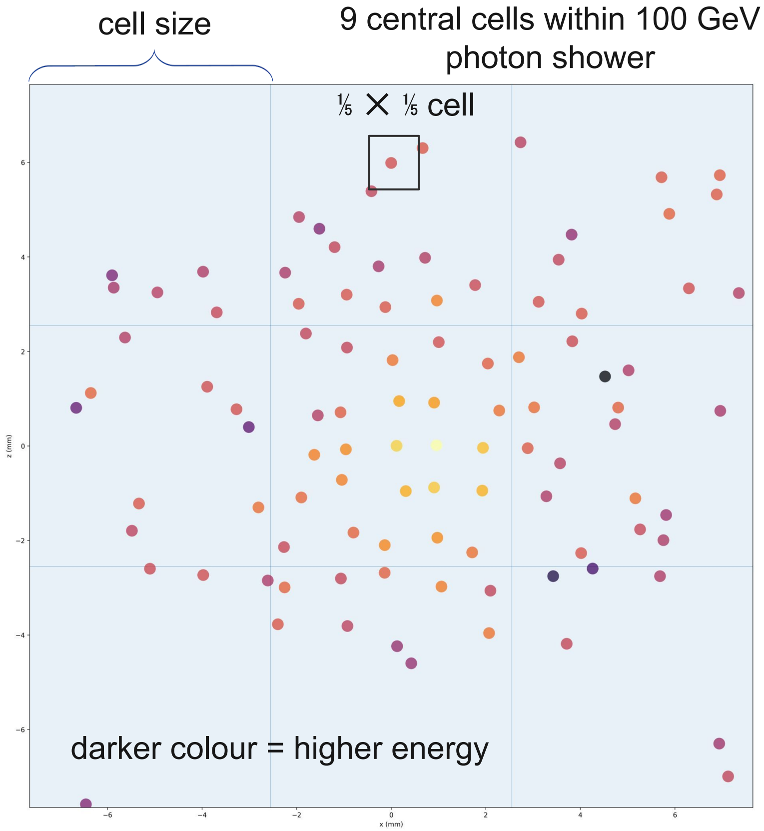
- Two modes provided:
  - Sub-grid **centre**



Regular Sub Grid vs Density-based Clustering

- **02 - Regular subgrid**
  - Used to created data representations in existing ILD studies
  - Divide every cell into a fixed regular square subgrid, e.g. 5 x 5
  - Simple and deterministic
  - But may not be most efficient in general:
    - Pre-supposes cell geometry
    - Treats dense shower core and low-energy tails uniformly

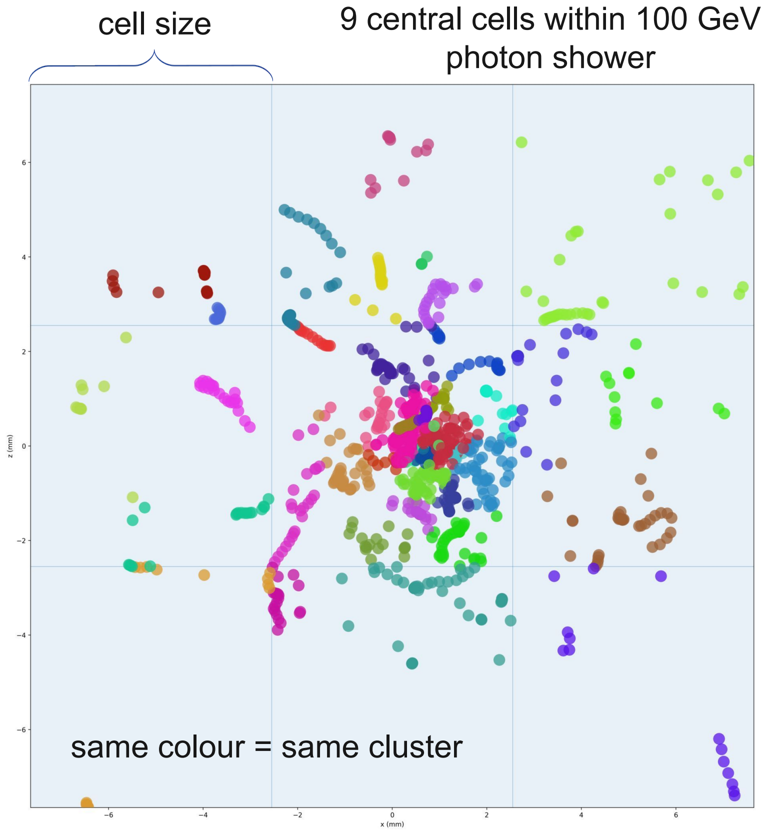
- Two modes provided:
  - Sub-grid centre
  - **Weighted barycentre**



Regular Sub Grid vs Density-based Clustering

- **03 - Density-based clustering (within cell)**
  - Cluster steps via local density, but only in the same detector cell
  - Dense regions represented by fewer weighted points
  - Isolated deposits remain separated
    - Limit distortions in shower tails
    - Retain hit multiplicities

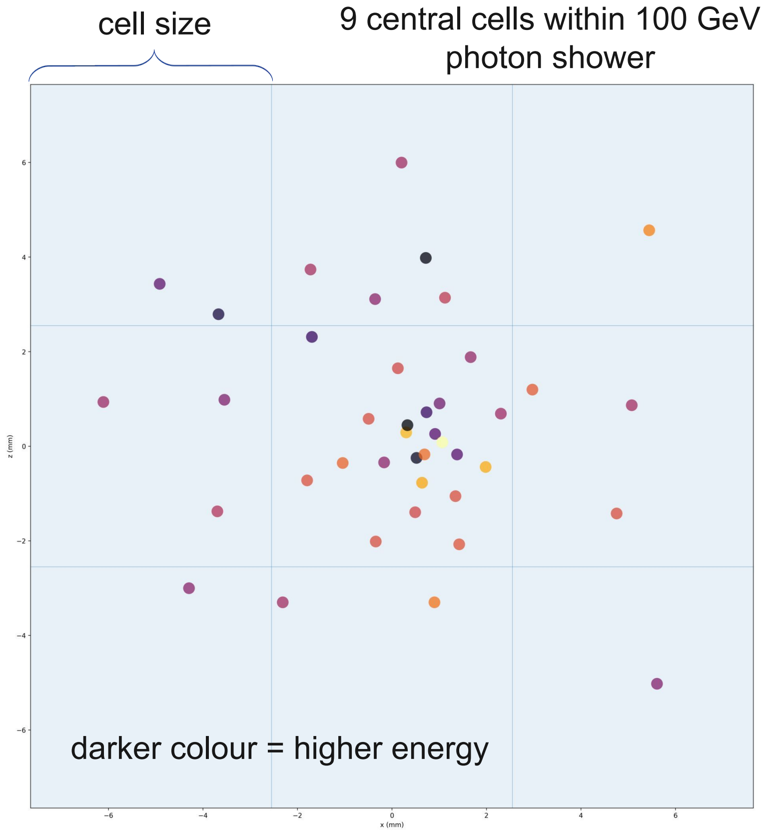
  - E.g. Hierarchical clustering with [HDBSCAN]



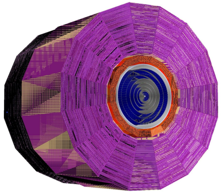
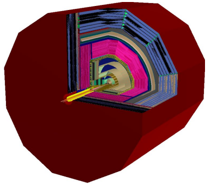
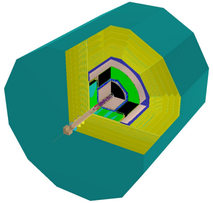
Regular Sub Grid vs Density-based Clustering

- **03 - Density-based clustering (within cell)**
  - Cluster steps via local density, but only in the same detector cell
  - Dense regions represented by fewer weighted points
  - Isolated deposits remain separated
    - Limit distortions in shower tails
    - Retain hit multiplicities

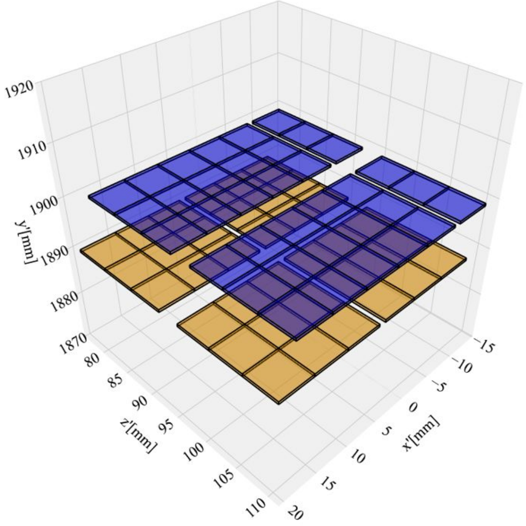
  - E.g. Hierarchical clustering with HDBSCAN



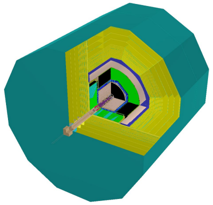
Currently Tested Electromagnetic Calorimeters

	Open Data Detector (ODD)	CLIC-like Detector (CLD)	International Large Detector (ILD)
			
Use case	Hadron collider	e+ e- collider	e+ e- collider
Materials	Si-W	Si-W	Si-W
Layers	48	40	30 (2 sampling fractions)
Cells	Square, 5.1 x 5.1 mm^2	Square, 5.1 x 5.1 mm^2	Square, 5.1 x 5.1 mm^2
Segmentation	Regular	Regular	Irregular

Currently Tested Electromagnetic Calorimeters



International Large Detector (ILD)



e+ e- collider

Si-W

30
(2 sampling fractions)

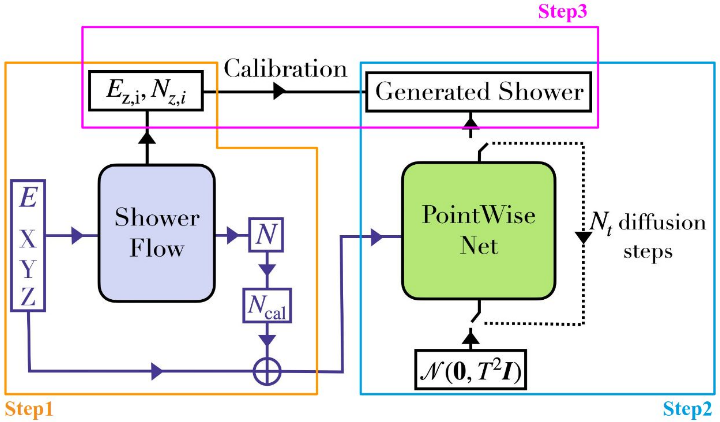
Square,
5.1 x 5.1 mm^2

Irregular

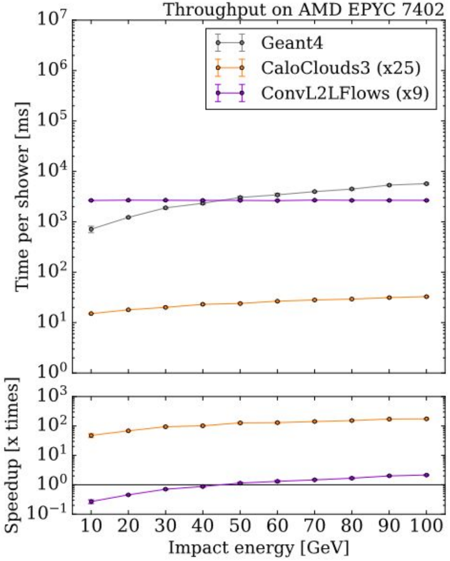
Case Study: CaloClouds3

- Compare different representations used to train CaloClouds3 model
- Architecture:
  - Step 1: Shower Flow- number of points and energy per layer
  - Step 2: Diffusion model generates points i.i.d
  - Step 3: Calibration of points according to Shower Flow predictions
- i.i.d sampling allows fast inference

- Training dataset from Benchmark paper:
  - ILD detector
  - Fixed incident position at calo face
  - Photons with energies 1-126 GeV
  - Incident angles varied in cone of up to 60°



A. Korol, PhD Thesis, U. Hamburg



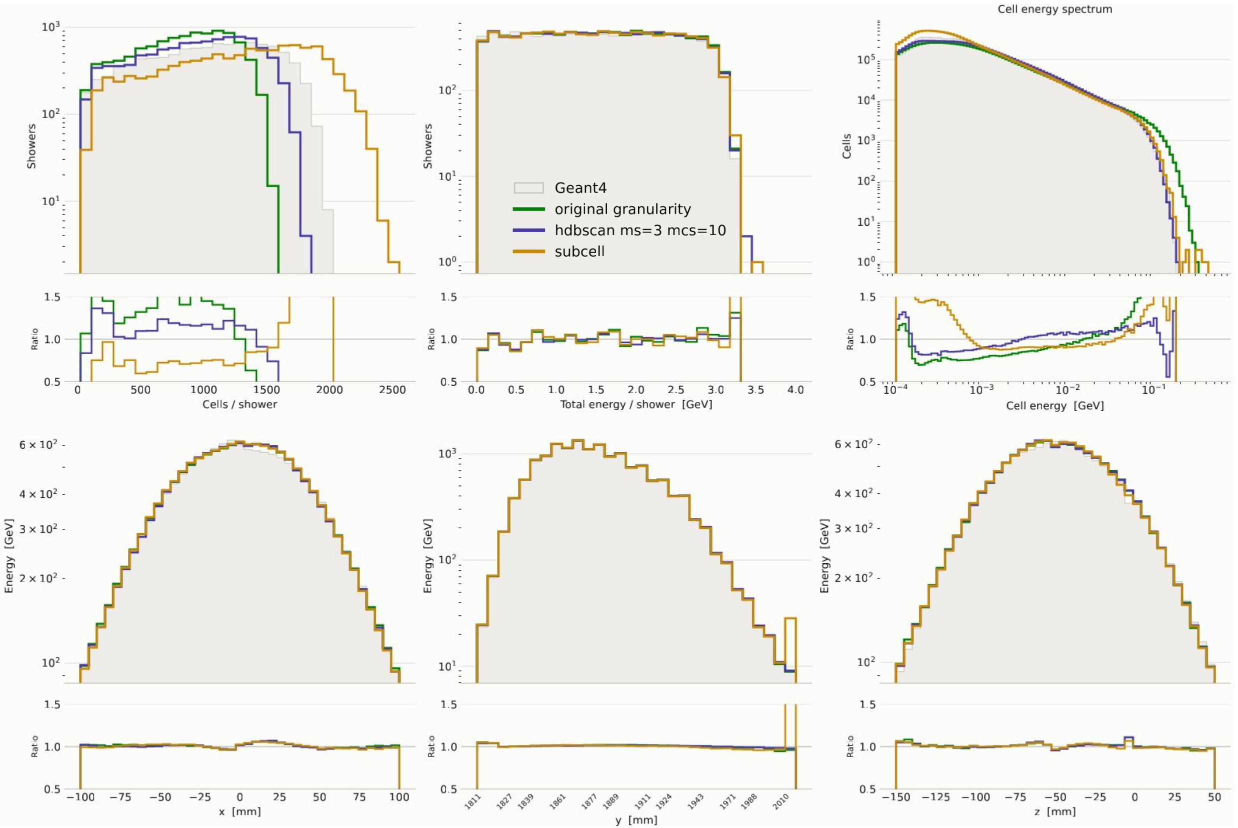
T. Buss et al., Accepted Phys. Rev. D

Case Study: CaloClouds3 Training Results

- Training performed with 350k showers

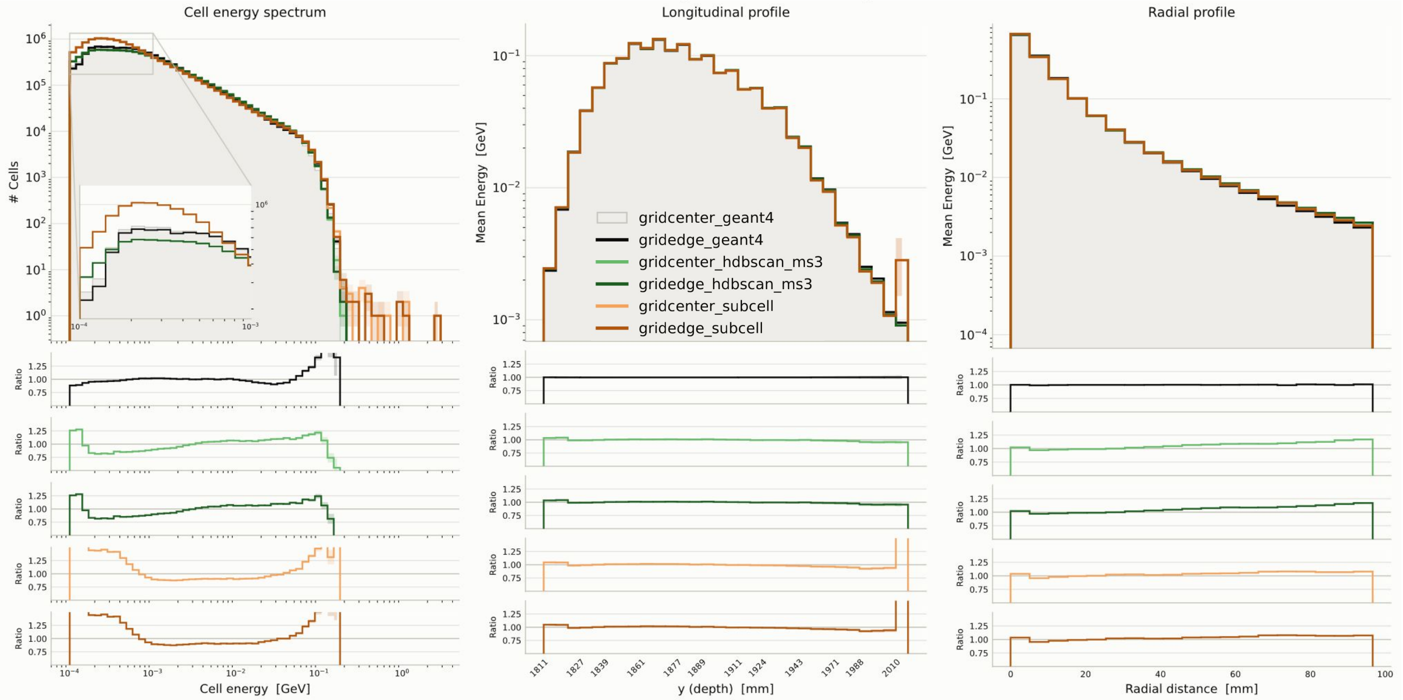
Point Cloud level	Mean number of points	Max number of points
Cell	1209	2302
Subcell	3758	7985
Hdbscan	1799	3555

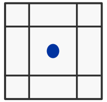
- Similar performance to Subcell with fewer points
- Reducing number of points could improve:
  - Model size (memory footprint)
  - Inference time

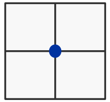


Additional Checks: Varying Incident Position

- In a realistic simulation, particles hit the calorimeter at varying incident positions and angles
- Modelling arbitrary incident position within cell
  - Amount of training data can be reduced

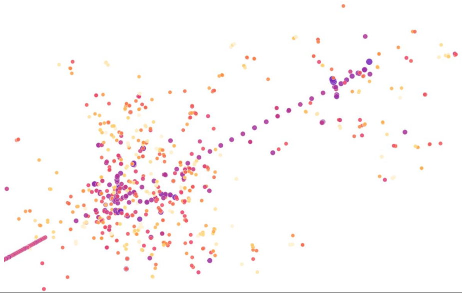






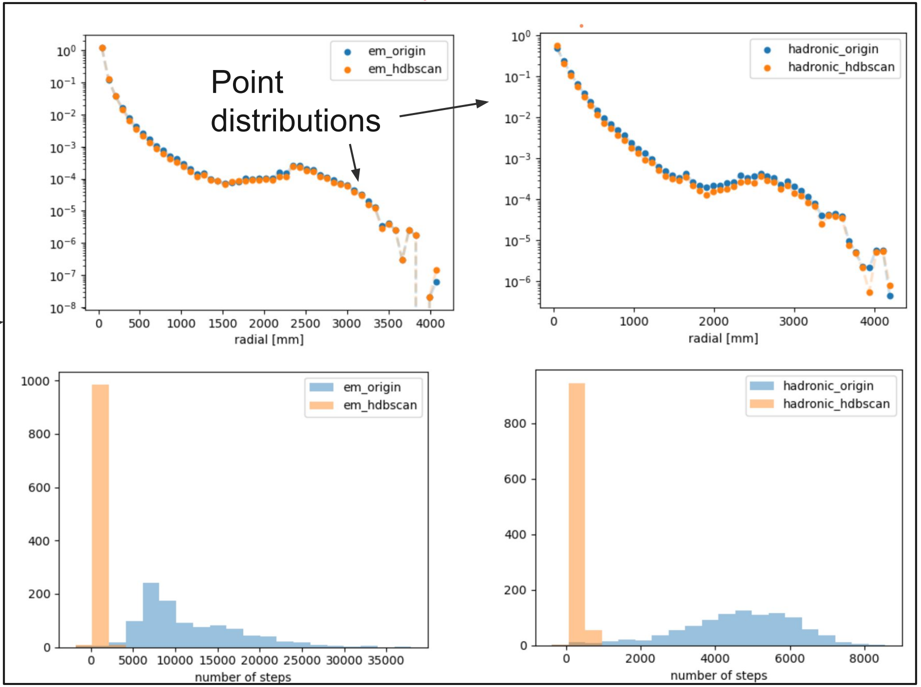
Outlook- What About Hadrons?

10 GeV pi^-



- Significantly more challenging than EM showers:
  - Mixture of EM and hadronic interactions
  - Large event-to-event fluctuations
  - Complex topology and time evolution

- Explored so far:
  - Naive clustering of the whole shower
  - Clustering EM and hadronic contributions independently



- Next steps:
  - Placement into detector cells
  - Test training of a point cloud model

Summary

- Step2point: general library for optimising point cloud representations for fast calorimeter simulation

- Benefits of point clouds:
    - More geometry independence
    - Transferable models (see e.g.: arXiv:2608.18233)
    - More efficient representation

- Density based clustering improve CC3 results - reduction in points for comparable quality:
    - Reduced model size
    - Reduced inference time

- Next steps:
    - Explore approaches for clustering hadronic showers and test generative model training
    - Apply a highly-granular point cloud representation to LHC experiments

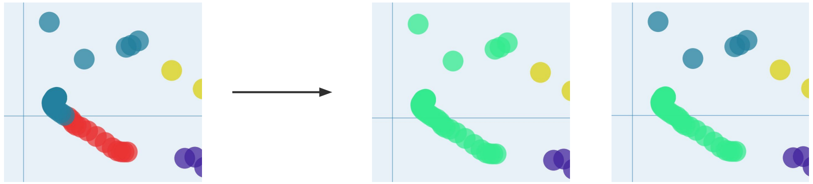
Backup

Clustering Modes

- **Respect detector cells** (default):
  - Only merge steps falling within one cell
  - Intrinsically preserve detector-level performance*
  - All algorithms have options to respect detector-cell boundaries
    - Regular grids (requires detector info)
    - Density-based clustering
    - Greedy agglomerative clustering

- **Allow cross-cell merging**:
  - Allows for higher reduction of points
  - Can be implemented:
    - Irrespective of detector structure - risky (shifting points between cells can easily distort physics)
    - As post processing to cluster points at the edges of cells (work in progress)

* assuming a fixed incident position within a cell (i.e. no later translation to exploit detector symmetries)



HDBSCAN clustering

Algorithm
Hierarchical Density-Based Spatial Clustering of Applications with Noise applied to 3D (or 4D with time) shower coordinates.

Clustering space
(x, y, z) positions — optionally extended with time t via --use-time.

Merge scope
Controls which steps compete to form clusters:

- system — whole detector
- system_layer — within each layer independently
- cell_id — within each detector cell

Outlier Policy
Outlier steps not assigned to any cluster are summed to the points of the nearest cluster.

Key Parameters

`--min-cluster-size`

Minimum steps to form a cluster (e.g. 5)

`--min-samples`

Core-point density threshold (e.g. 3)

`--merge-scope`

cell_id

`--outlier_policy`

nearest_cluster

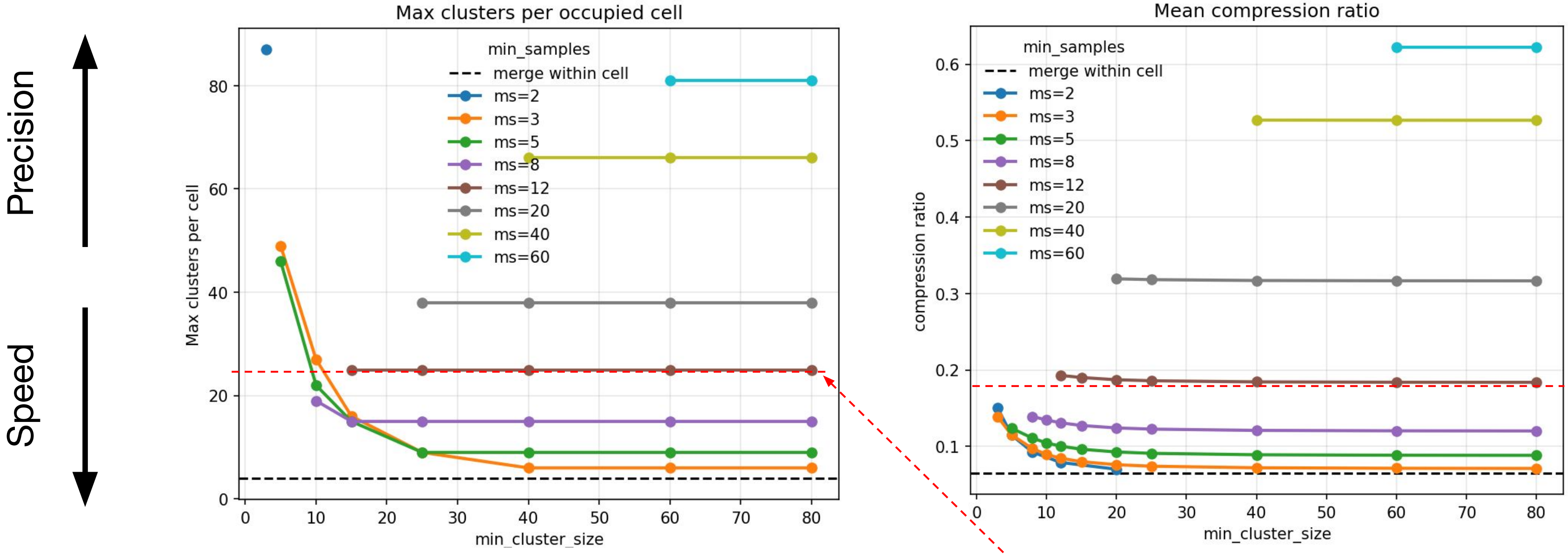
`--compact-xml`

Detector geometry for cell ID decoding

`--collection-name`

EDM4hep collection to read

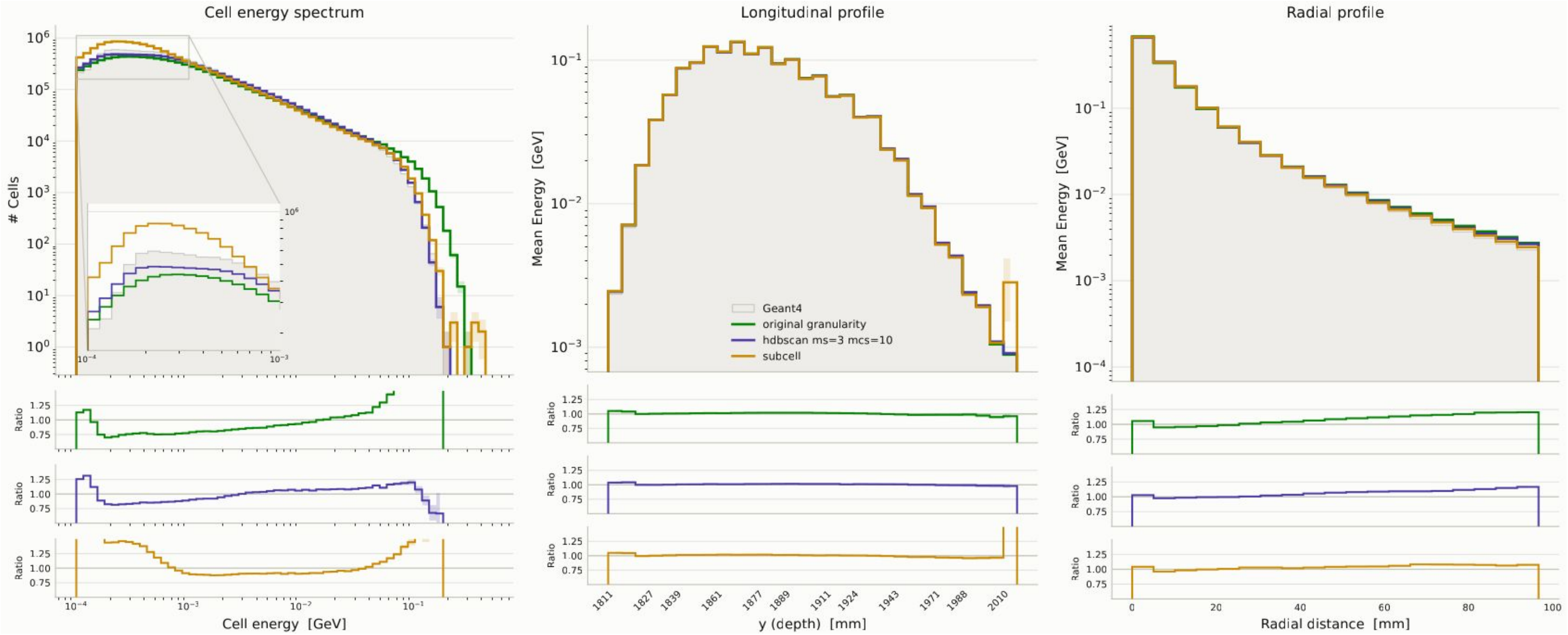
Case Study: HDBScan Parameters



⅕ ✕ ⅕ cell regular subgrid

- Scan across HDBScan parameters
  - Minimum cluster size and minimum number of samples
  - Select ms = 3, mcs = 10

CaloClouds3 Training Additional Results



CaloClouds3 Training Additional Results