OpenDataDetector

ColliderML

the High-Luminosity benchmark dataset

ColliderML Release 1

arXiv:2512.15230

Docs website

A. Zaborowska[1], D. Murnane[2], P. Gessinger[1], A. Salzburger[1], A. Stefl[1], D. Elitez[1], M. Raaholt[2], S. Skov[2], D. Amperiadou[1]

[1] CERN [2] University of Copenhagen

EP-SFT
Software Frameworks and Tools

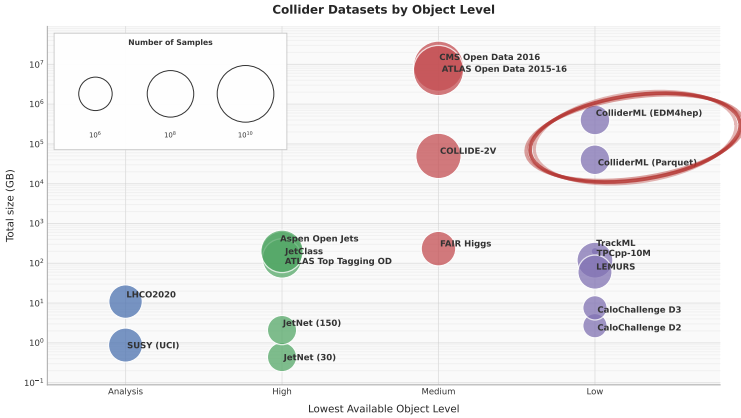
CHEP 2026, 28.05.2026

ColliderML Dataset

Docs website | arXiv:2512.15230 | ColliderML Release 1

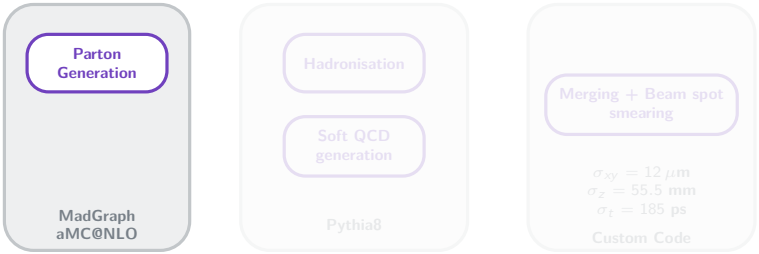
ColliderML is an **open**, experiment-agnostic benchmark for detector-aware ML at the HL-LHC.

- **HL-LHC conditions**: $\sqrt{s} = 14$ TeV with pile-up (PU) $\langle \mu \rangle = 200$ (and w/o PU).
- **Full chain**: generation, simulation, digitisation and reconstruction (track reco only in Release 1, more in future).
- **Multi-scale event record**: truth particles, detector hits, calorimeter cells and tracks.
- **Truth provenance**: low-level objects remain linked to the particles that produced them.
- **Release model**: low-level data first, higher-level reconstruction and challenge samples next.



Generation

implemented | partial / interface | not yet / planned



Standard Model

- ttbar*
- zee**, zmumu**
- diphoton
- multijet

Higgs

- ggf*
- dihiggs*
- ttH*

BSM / stress tests

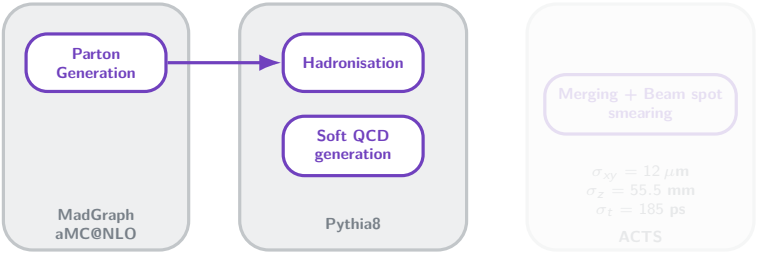
- higgs_portal***
- zprime***
- hiddenvalley***

PU only events also available

*: $\langle\mu\rangle = 0, 200$ **: $\langle\mu\rangle = 200$ ***: $\langle\mu\rangle = 0$

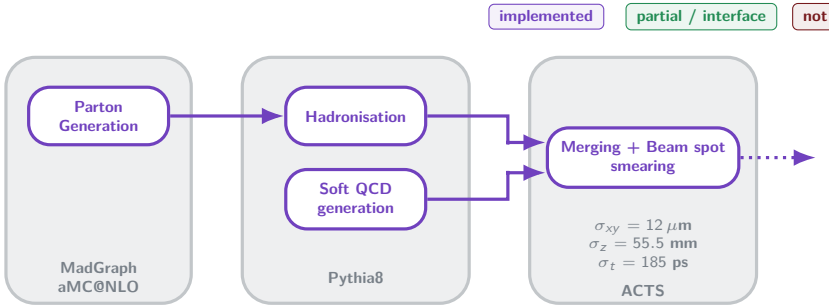
Generation

implemented | partial / interface | not yet / planned



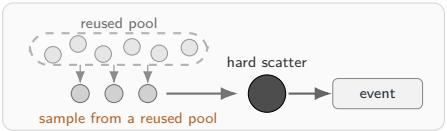
Generate every pile-up sub-event independently and combine without repetition before simulation

Generation

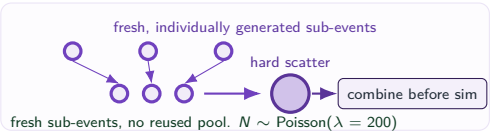


Finite beam spot: shift each sub-event randomly with Gaussian widths

Typical pile-up

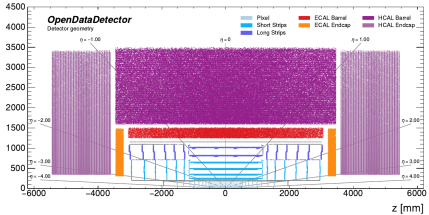


ColliderML

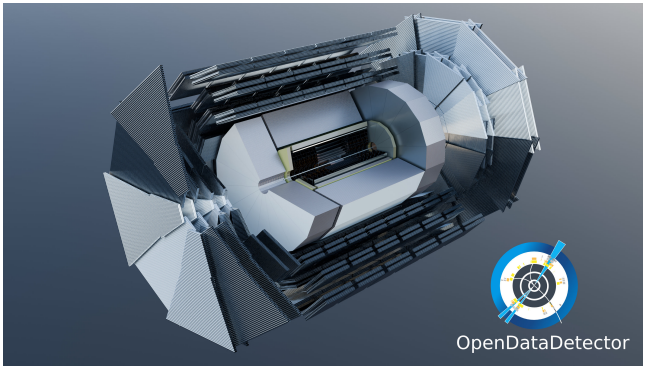


Open Data Detector

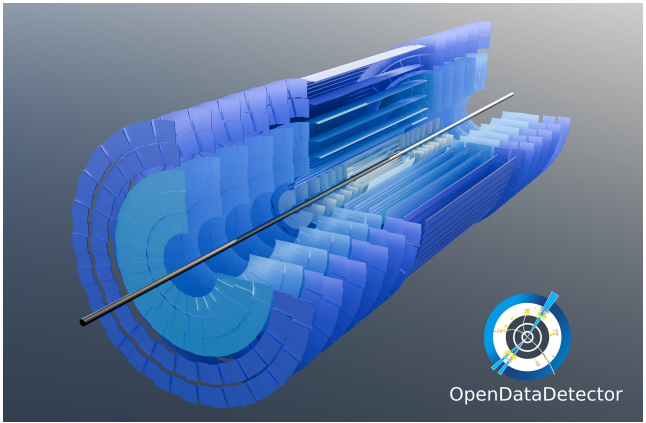
- **Experiment-agnostic** geometry for algorithm development.
- **HL-LHC-like** occupancy and material budget.
- **ITk-inspired tracker** with pixels and strips.
- **Granular ECAL/HCAL** for calorimeter and particle-flow studies.
- **3 T solenoidal field**; designed as a stable benchmark, not an experiment replica.



ODD = public detector benchmark

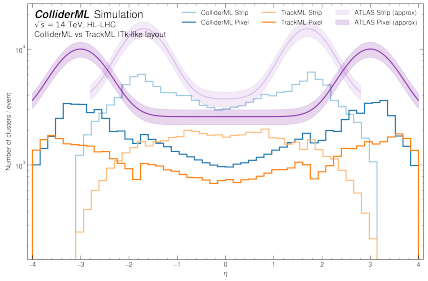


Open Data Detector: Tracker



Tracking detector

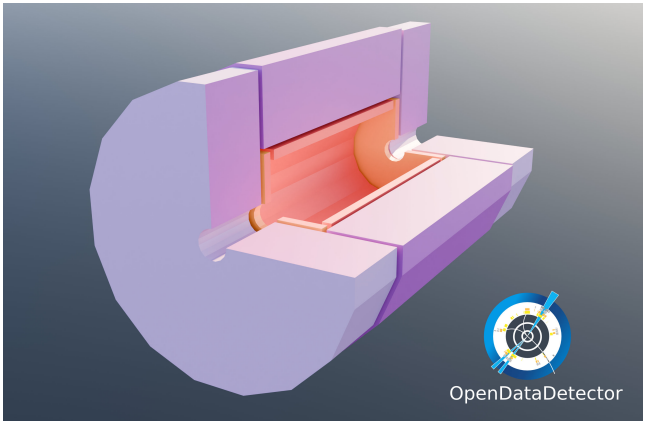
- Barrel and endcaps covering tracks up to $|\eta| \leq 3.5$.
  - at least 12 sensitive layers for $|\eta| < 3$,
  - and 8 layers for $|\eta| < 3.5$.
- **Pixel layers** in the inner region.
- **Short and long strips** in the outer tracker.
- Introduced realistic material to stress-test tracking.



Open Data Detector: Calorimetry

Highly-granular calorimetry inspired by future detectors.

- **ECAL**: silicon sensors with tungsten absorber.
  - 48 silicon layers,
  - $5.1 \times 5.1\,\mathrm{mm}^2$ cells;
- **HCAL**: scintillator cells with steel absorber.
  - 36 scintillator layers,
  - $30 \times 30\,\mathrm{mm}^2$ cells.
- Release 1 exposes **thresholded cell energies** and **truth contributions**.
- Full calibration and PF-driven hadronic calibration remain key next steps.



Simulation & digitisation

implemented | partial / interface | not yet / planned

Simulation

DD4hep + Geant4

Tracker Digitisation

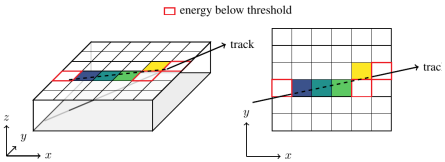
ACTS

Track Reconstruction

ACTS

Calorimeter Digitisation

k4ODD



Per-cell thresholds:
50 keV (ECAL), 250 keV (HCAL)
No full energy calibration yet

Simulation & digitisation

implemented | partial / interface | not yet / planned

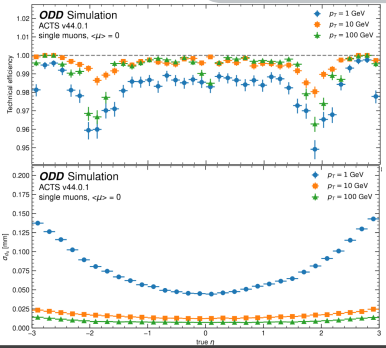
Simulation
DD4hep + Geant4

Tracker Digitisation
ACTS

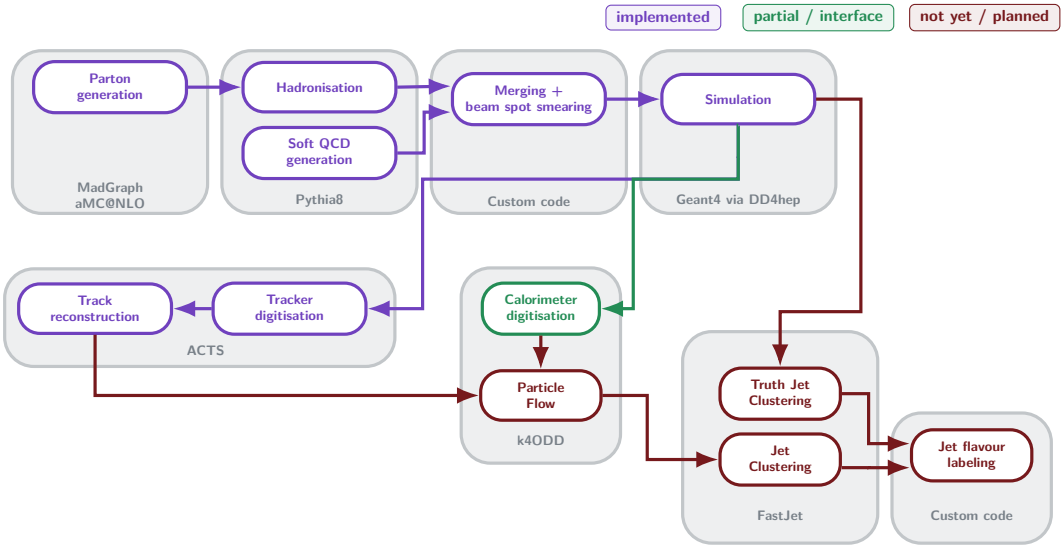
Track Reconstruction
ACTS

Tracks added
to EDM4hep files

Calorimeter
Digitisation
k4ODD



Full workflow



ColliderML Release 1

ColliderML Dataset Release 1 @ Hugging Face (Parquet)

@ NERSC (ROOT)

Object	Content
Particles	4-momentum, charge, PDG ID, production vertex, ancestry and truth-perigee parameters.
Tracker Hits	Measured/true positions, timing, detector hierarchy IDs, generating particle link.
Calo Hits	Cell energy and position, truth-particle contributions with energy fractions.
Tracks	Helix parameters, associated hit IDs, majority-particle truth matching.

Direct Hugging Face access

```
from datasets import load_dataset
ds = load_dataset(
    "CERN/ColliderML-Release-1",
    "dihiggs_pu0_particles", )
```

Convenience library

```
pip install colliderml
```

Quick access

```
pip install colliderml

colliderml download \
  --channels ttbar \
  --pileup pu0 \
  --objects particles,tracker_hits,calo_hits,
     tracks \
  --max-events 200
```

Includes tools for visualisation, particle ancestry, production regions and pile-up downsampling.

ColliderML Release 1

ColliderML Dataset Release 1 @ Hugging Face (Parquet)

@ NERSC (ROOT)

Object	Content
Particles	4-momentum, charge, PDG ID, production vertex, ancestry and truth-perigee parameters.
Tracker Hits	Measured/true positions, timing, detector hierarchy IDs, generating particle link.
Calo Hits	Cell energy and position, truth-particle contributions with energy fractions.
Tracks	Helix parameters, associated hit IDs, majority-particle truth matching.

Direct Hugging Face access

```
from datasets import load_dataset
ds = load_dataset(
    "CERN/ColliderML-Release-1",
    "dihiggs_pu0_particles", )
```

Convenience library

```
pip install colliderml
```

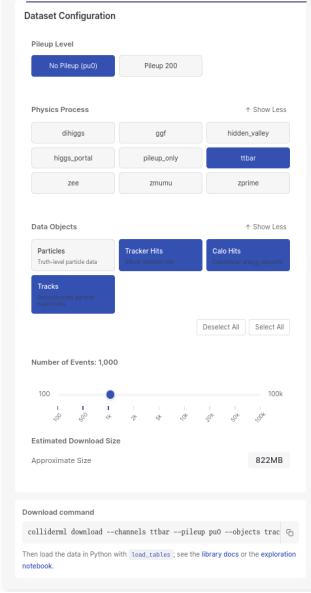
Quick access

```
pip install colliderml

colliderml download \
  --channels ttbar \
  --pileup pu0 \
  --objects particles,t
     tracks \
  --max-events 200
```

Includes tools for visualisat
production regions and pile

Configurator @ website



DIY dataset extensions

Facilitating self-production:

- Reproduce samples locally for other physics channels;
- Pile-up can be sampled from the central production;
- Build controlled ML stress tests with modified conditions;

Locally:

```
pip install "colliderml[sim]"
```

```
colliderml.simulate(preset="ttbar-quick")
print(result.run_dir)
print([p.name for p in result.list_files()])

# Explicit parameters: bundled (channel, events
    , pileup) triple.
colliderml.simulate(channel="higgs_portal",
    events=10, pileup=10)
```

- Larger samples may require larger resources, e.g. NERSC Perlmutter;
- Results hosted directly at HuggingFace;
- Simple credit schema implemented
  1 credit $\approx$ 1 node-hour $\approx$ 100 events at $\langle\mu\rangle = 0$
  New users get 10 credits on first login
  Benchmark wins will grant credits

Remotely:

```
pip install "colliderml[remote]"
```

```
huggingface-cli login     # interactive
# or
export HF_TOKEN=hf_xxxxxxxx
```

```
import colliderml

result = colliderml.simulate(
    preset="higgs-portal-quick",
    remote=True)
print(result.remote_request_id)
```

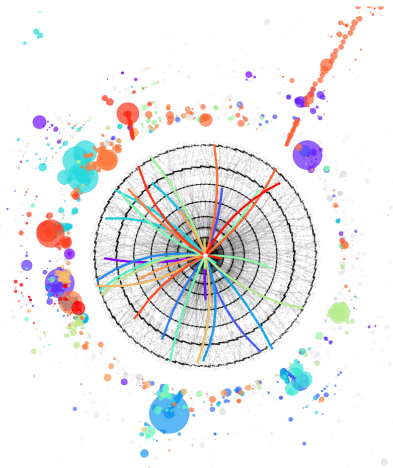
Next steps

Planned **dataset** extensions (still within Release 1):

- **Miscalibrations/missing processes** from the planned list.
- **Single particle** samples: finalising kinematics to accommodate different use cases (eg. fast calorimeter simulation)

Work in progress on items extending beyond Release 1:

- **Calorimeter calibration**: available recipe and constants are not yet the final detector calibration.
- **Higher-level objects**: PF candidates, jets and flavour labels are planned extensions.
- Addition of a **muon system**: signal from muon chambers.



Summary

1. **ColliderML** is an open, experiment-agnostic benchmark for **detector-aware ML** at the HL-LHC.
2. Release 1 provides **fully simulated**, **digitised** events with particles, tracker hits, calorimeter hits and tracks.
3. Its distinctive ingredients are

   truth provenance,
   non-reused pile-up sub-events
   and a **realistic** ODD detector geometry.
4. The pipeline is already useful for **tracking**, **calorimetry** clustering and low-level ML; particle flow, jets, flavour labels and challenge samples are next.
5. ColliderML is designed as a community benchmark.

Open detector-level data for realistic ML at the HL-LHC.