ColliderML: Enabling Foundation Models in High Energy Physics Through Low-Level Detector Data

Daniel Murnane[1,3] and **Paul Gessinger**[2] and **Andreas Salzburger**[2] and **Anna Zaborowska**[2] and **Andreas Stefl**[2] and **Stine Kofoed Skov**[1] and **Marcus Raaholt**[1]
[1]**University of Copenhagen** [2]**CERN** [3]**Berkeley Lab**

Abstract

We introduce **ColliderML** , an open dataset of one million fully simulated proton-proton collisions at HL-LHC conditions, providing detector-level measurements across ten physics processes. Unlike existing fast-simulation datasets operating on high-level objects, ColliderML provides hits, energy deposits, and reconstructed tracks from realistic detector geometry under extreme pile-up ($\mu \approx 200$). We argue that foundation models trained on such low-level data represent the future of collider physics, and present ColliderML as the infrastructure to realize this vision.

1 The Foundation Model Opportunity

Machine learning in particle physics stands at an inflection point. The field has progressed from simple classifiers to sophisticated architectures Lönnblad et al. [1991], yet remains constrained by a fundamental limitation: **most public ML research operates on high-level physics objects rather than raw detector measurements**. This constrains the community to incremental improvements of existing reconstruction approaches rather than enabling transformative end-to-end approaches.

Foundation models—large-scale models trained on comprehensive low-level data—have revolutionized computer vision and natural language processing. In particle physics, such models could learn detector response, particle interactions, and physics signatures simultaneously, potentially discovering novel analysis strategies impossible with traditional pipelines. However, realizing this vision requires three prerequisites: (1) large-scale detector-level data, (2) realistic experimental conditions, and (3) diverse physics coverage. **No existing public dataset satisfies all three.**

Popular datasets like JetClass (100M events) and DarkMachines (1B events) use fast simulation, producing high-level jets or particles directly Qu et al. [2022], Aarrestad et al. [2022]. While valuable for specific tasks, they preclude learning detector response or low-level reconstruction. TrackML provided detector hits but covered only 10k events in the tracker alone—sufficient for a Kaggle competition but inadequate for foundation model training, with observed overfitting

AI in Science (AIS), 2025, Copenhagen, Denmark

Amrouche et al. [2019], Zhou et al. [2025]. Meanwhile, LHC collaborations generate billions of fully simulated events internally but do not release them publicly, creating a widening capability gap.

Recent work demonstrates the potential of end-to-end learning. The Maskformer architecture reconstructs jets directly from hits Gong et al. [2023], Higgsformer performs signal searches from raw detector data Zhou et al. [2025], and GNN-based tracking shows competitive performance with traditional algorithms Ju et al. [2021], ATLAS Collaboration [2022]. These proof-of-concept studies hint at a future where foundation models learn optimal representations from detector measurements, adapting to diverse downstream tasks without hand-crafted features. **ColliderML provides the infrastructure to systematically explore this this new direction.**

2 Dataset Design Philosophy

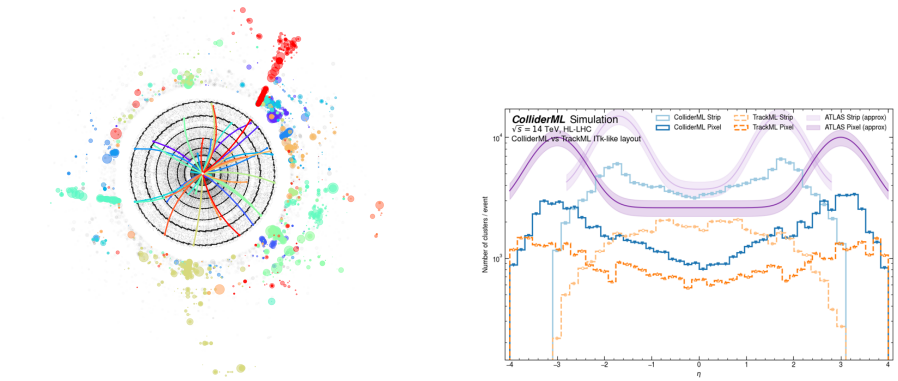
ColliderML is designed around three core principles: **realism**, **scale**, and **accessibility**.

Realism. All events use full Geant4 simulation Agostinelli et al. [2003] through the validated OpenDataDetector (ODD) geometry Gessinger-Befurt et al. [2023]—a realistic, experiment-agnostic detector combining design elements from ATLAS ITk, CMS HGCal, and future collider proposals. The inner tracker features pixel, strixel, and strip layers with accurate timing. Electromagnetic and hadronic calorimeters use realistic absorber materials and granular segmentation (5.1mm ECal cells, 30mm HCal cells). Events undergo proper digitization with geometric segmentation, thresholding, and detector response modeling.

Importantly, pile-up is generated uniquely for each event at mean $\mu = 200$, matching HL-LHC conditions. Unlike common practice, we avoid resampling pile-up events—a known source of train-test contamination ATLAS Collaboration [2022]. Sub-event IDs are preserved, enabling controllable pile-up scaling for curriculum learning and systematic studies of algorithm robustness.

Physics Coverage. One million events span ten processes: six Standard Model channels ($t\bar{t}$, $Z \to ee$, $Z \to \mu\mu$, $\gamma\gamma$, ggF Higgs, di-Higgs) and four BSM scenarios (SUSY gluinos, Z′, heavy neutral leptons, hidden valley). NLO matrix elements with modern shower matching ensure state-of-the-art parton modeling. Additionally, 1M single-particle gun events (e^-, μ^-, π^+, K^+, γ) enable detailed detector response studies.

Accessibility. Data is provided in two formats: EDM4hep ROOT files (100TB) for full truth preservation, and HDF5 files (1TB) for ML workflows. A lightweight Python library handles downloads and data manipulation. All simulation and reconstruction code is open-source, with comprehensive documentation for reproduction. Data is mirrored across US (NERSC) and European (EOS) facilities.



(a) An example event showing inner tracker hits, reconstructed tracks, and calorimeter deposits under $\mu \approx 200$ pile-up.

(b) Hit multiplicities closely match predicted ATLAS ITk occupancy Aad [2025], unlike simplified datasets.

Figure 1: ColliderML provides realistic detector complexity for foundation model development.

3 Foundation Model Research Directions

ColliderML enables several transformative research directions previously inaccessible to the public ML community.

End-to-End Reconstruction. Traditional pipelines decompose reconstruction into stages: tracking, clustering, particle flow, jet finding, and analysis. Each stage introduces assumptions, inefficiencies, and information loss. Foundation models could learn optimal representations directly from hits and energy deposits, jointly optimizing for downstream physics tasks. Early work shows promise—we demonstrate GNN-based tracking approaching ACTS performance (**??**), and the inclusion of timing information further improves efficiency, validating detector R&D priorities.

Multi-Task Learning. A single foundation model pre-trained on ColliderML could simultaneously perform tracking, clustering, jet reconstruction, and signal classification. This amortizes the computational cost of learning detector response while enabling transfer learning across tasks. The diverse physics coverage (SM and BSM processes) provides rich supervision for learning general representations.

Controllable Complexity via Pile-up Scaling. Unlike datasets with fixed complexity, ColliderML events can be downsampled from the hard-scatter vertex plus n pile-up vertices. This enables systematic studies of algorithmic scaling behavior and curriculum learning strategies—training first on low pile-up,

then progressively increasing difficulty. We propose this as a standard protocol replacing arbitrary energy cuts or simplified geometries.

Robustness and Domain Adaptation. Real detectors exhibit calibration drifts, dead channels, and alignment uncertainties. Foundation models trained on comprehensive low-level data could learn robust features less sensitive to such variations. Future ColliderML releases could include systematic variations, enabling research on domain adaptation for deployment in real experiments.

4 Broader Impact and Future Directions

Beyond technical capabilities, ColliderML democratizes access to realistic collider simulation. Historically, only collaboration members with institutional computing resources could perform such studies. By providing this data publicly with minimal barriers to entry, we enable:

- **Algorithm innovation** from the broader ML community, unconstrained by collaboration computing policies
- **Education and training** for students entering the field, who can experiment with realistic data before joining experiments
- **Reproducible research** with standardized benchmarks, facilitating fair comparisons across methods
- **Cross-pollination** between HEP and other domains (astronomy, medical imaging) facing similar reconstruction challenges

Planned Extensions. Release 2 adds calorimeter topoclusters, Pandora particle flow objects, and reconstructed jets with flavor tagging and energy regression baselines—enabling research on higher-level reconstruction while maintaining access to low-level data. Future releases will incorporate muon detection, expand physics coverage, and potentially include systematic variations for robustness studies.

Community Engagement. We envision ColliderML as living infrastructure. The dataset is version-controlled, and improvements to simulation or reconstruction can be propagated to future releases. We encourage community contributions of baselines, tutorials, and novel architectures.

Data Access. Full documentation, download instructions, and reproduction recipes available at `https://www.colliderml.com`.

References

G. e. a. Aad. Expected tracking performance of the atlas inner tracker at the high-luminosity lhc. *Journal of Instrumentation*, 20(02):P02018, Feb. 2025. ISSN 1748-0221. doi: 10.1088/1748-0221/20/02/p02018. URL `http://dx.doi.org/10.1088/1748-0221/20/02/P02018`.

T. Aarrestad et al. The dark machines anomaly score challenge: Benchmark data and model independent event classification for the large hadron collider. *SciPost Phys.*, 12:043, 2022. doi: 10.21468/SciPostPhys.12.1.043.

S. Agostinelli et al. Geant4—a simulation toolkit. *Nucl. Instrum. Meth. A*, 506: 250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8.

S. Amrouche et al. The tracking machine learning challenge: Accuracy phase. *arXiv preprint*, 2019.

ATLAS Collaboration. Atlas itk track reconstruction with a gnn-based pipeline. ATL-ITK-PROC-2022-006, 2022. URL `https://cds.cern.ch/record/2816063`.

P. Gessinger-Befurt et al. Opendatadetector tracking system: Design and prototyping. In *IEEE NSS/MIC 2023*, 2023.

J. Gong et al. Maskformers for particle physics: End-to-end secondary vertex reconstruction. *Phys. Rev. D*, 108:076025, 2023.

X. Ju et al. Graph neural networks for particle tracking and reconstruction. In *ICLR 2021 Workshop on AI for Science*, 2021.

L. Lönnblad, C. Peterson, and T. Rögnvaldsson. Using neural networks to identify jets. *Nucl. Phys. B*, 349:675–702, 1991. doi: 10.1016/0550-3213(91)90190-V.

H. Qu, C. Li, and S. Qian. Particle transformer for jet tagging. *arXiv preprint*, 2022. Presents the 100M-jet *JetClass* dataset.

F. Zhou et al. Higgsformer: Transformer-based end-to-end search for higgs pair production from raw detector hits. *arXiv*, 2025. Preprint, 2025-08-26.