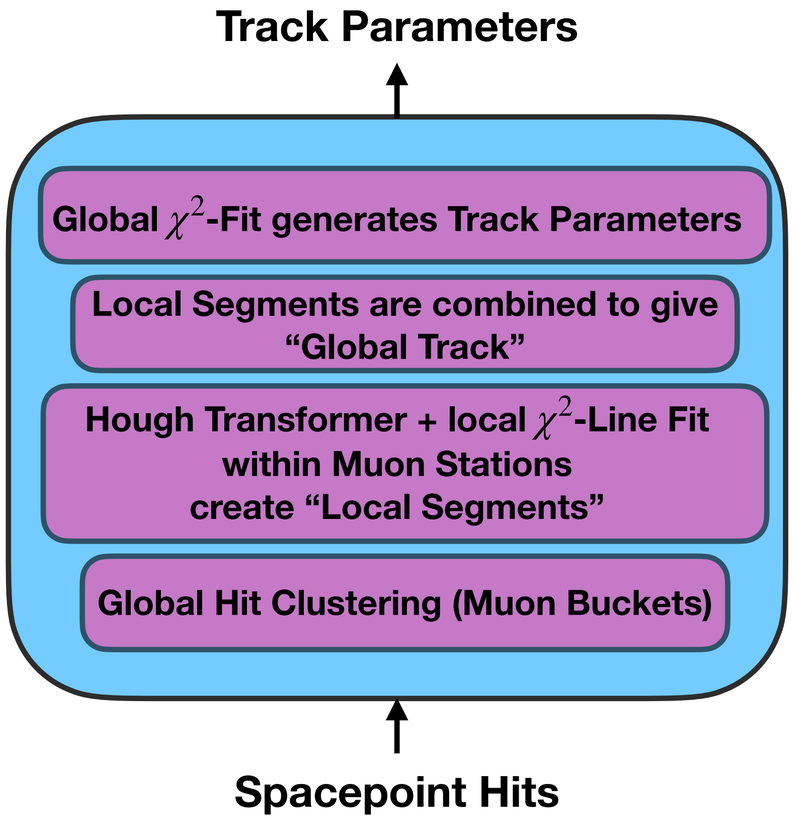
ML-Assisted Tracking in the ATLAS Muon Spectrometer

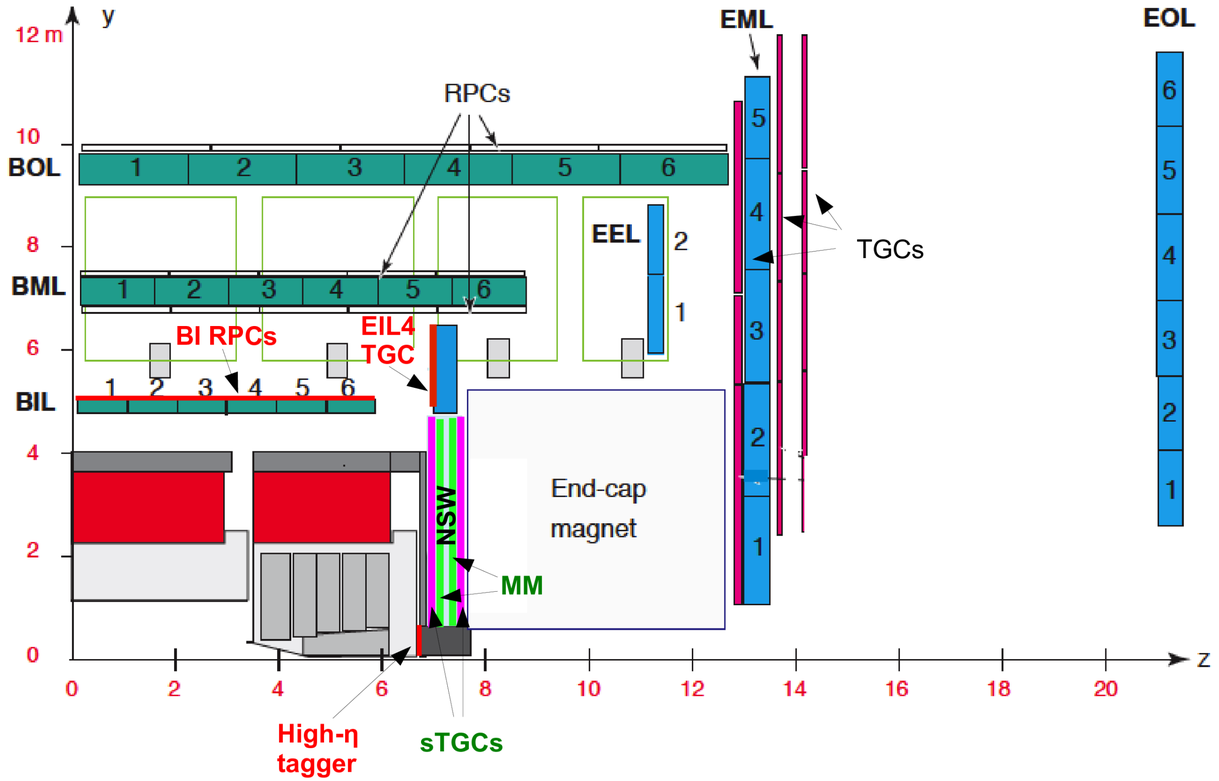
Jonathan Niklas Renusch on behalf of the ATLAS Collaboration

Jonathan Niklas Renusch, 10.11.2025, jonathan.renusch@cern.ch

Context | ATLAS Muon Tracking

Legacy Event Filter Baseline | Excellent Technical Efficiency



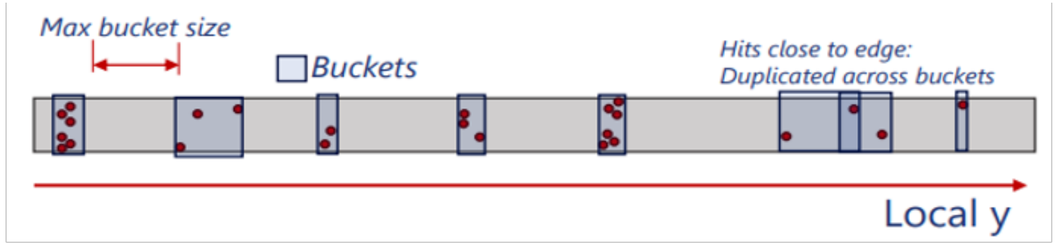


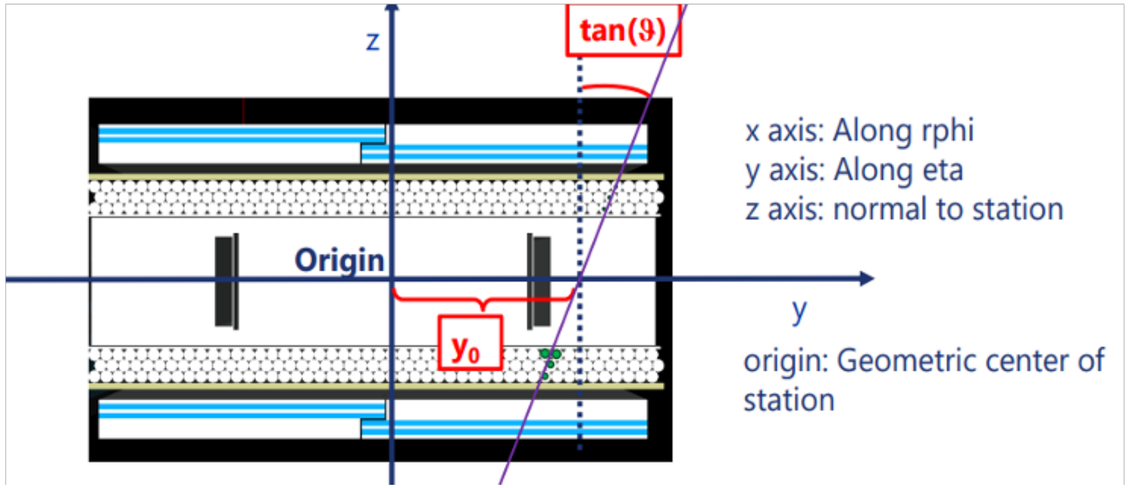
ATL-DAQ-PUB-2025-004

Bucketing

“Bucket-Based” Clustering as a Preprocessing Step

- Each “Bucket” consists of a cluster of Space Points based on geometric proximity and detector segmentation from the Muon Spectrometer Geometry
- Allowed length: maximum 2 m (longer regions get divided into multiple buckets)
- Reduces data access overhead and constrains the search region during reconstruction

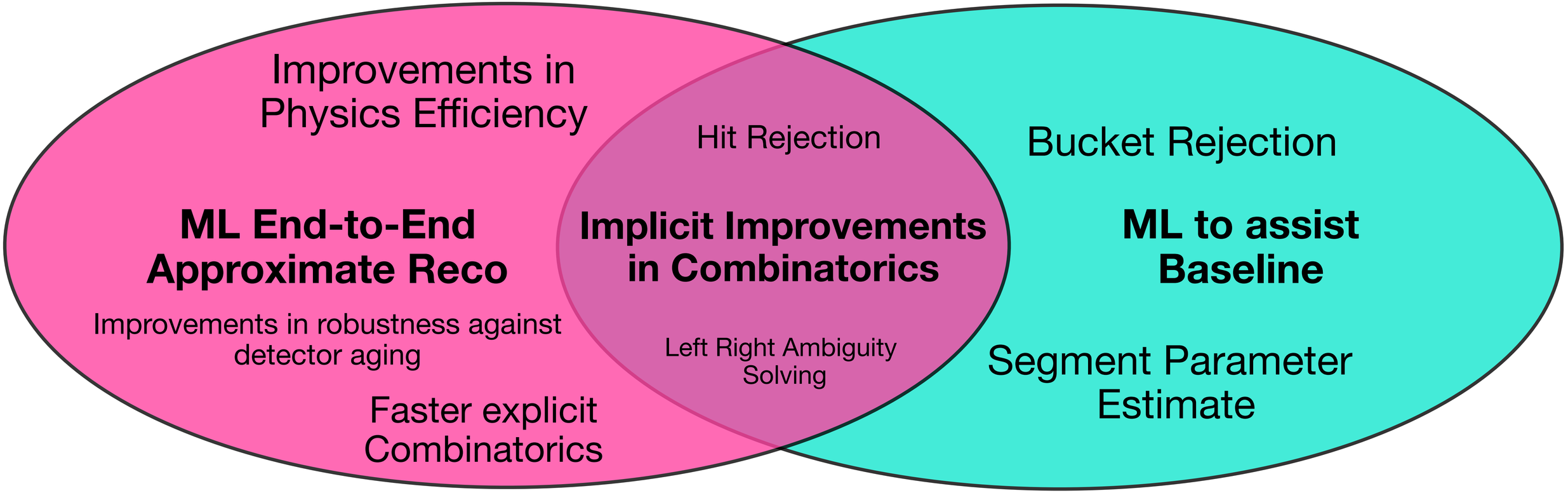




ATL-DAQ-PUB-2025-004

Context | ATLAS Muon Tracking

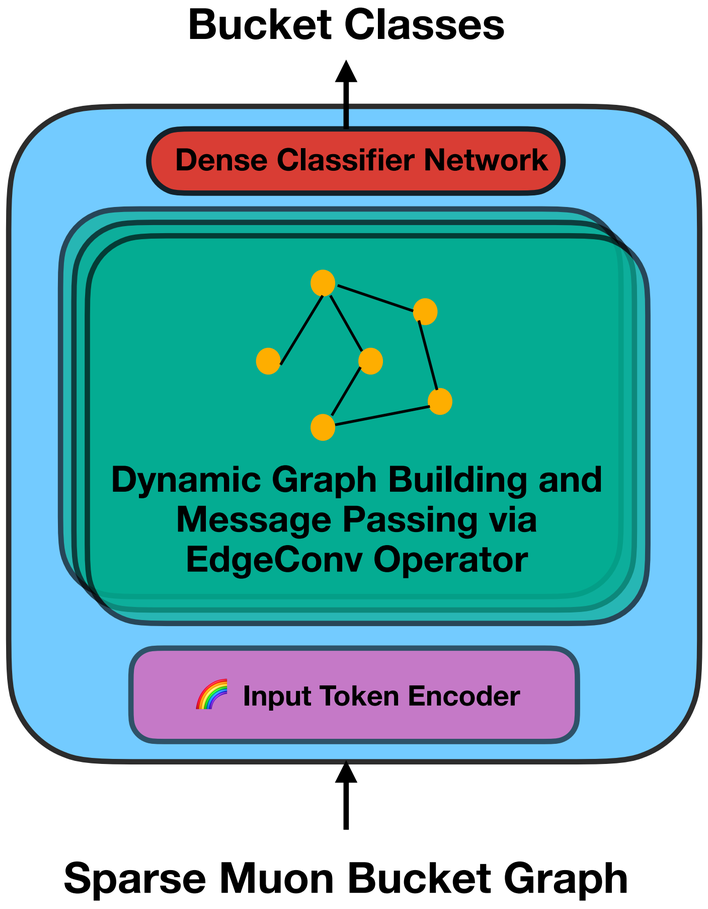
Where ML could help



GNN Filtering

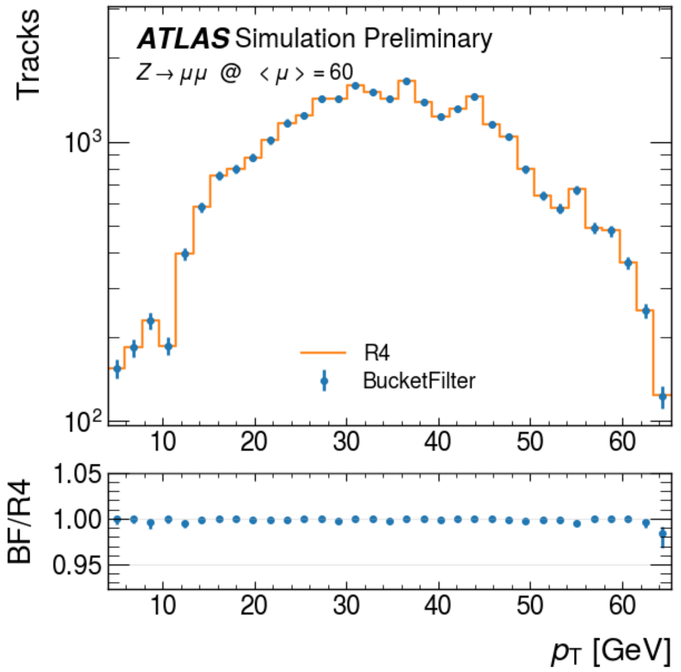
GNN based Filtering on Muon Buckets

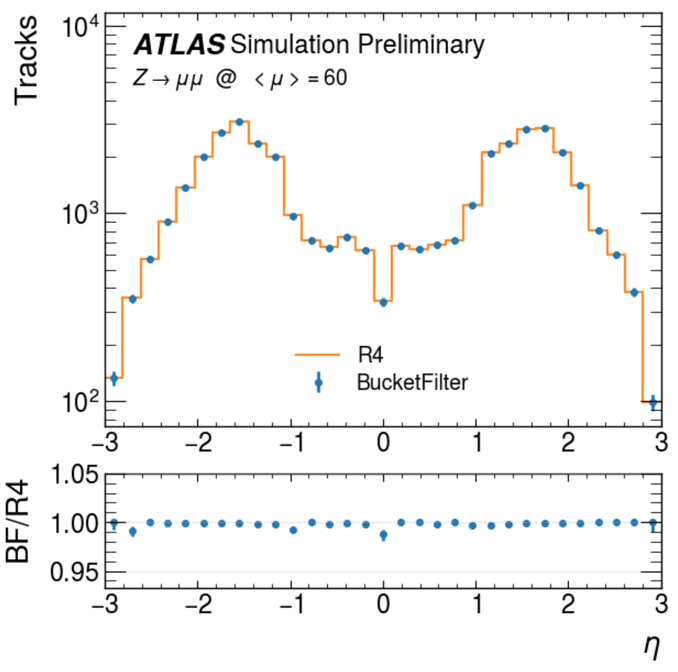
- The Bucket Filter GNN discriminates between buckets containing 0 “segments” (local line Fits of Muon Tracks), 1 segment and multiple segments
- Input Variables per Bucket: Centre of Bucket xyz
- The core operation relies on an adaptation of the EdgeConvolution operation as well as a Fourier Feature Expansion inspired by Computer Vision image preprocessing techniques (https://arxiv.org/pdf/1801.07829)
- It was trained on a mix of ***ATLAS*** Simulation $J/\psi, t\bar{t}$ and $Z \rightarrow \mu\mu$ events
- Network available in **ATHENA** 🥳

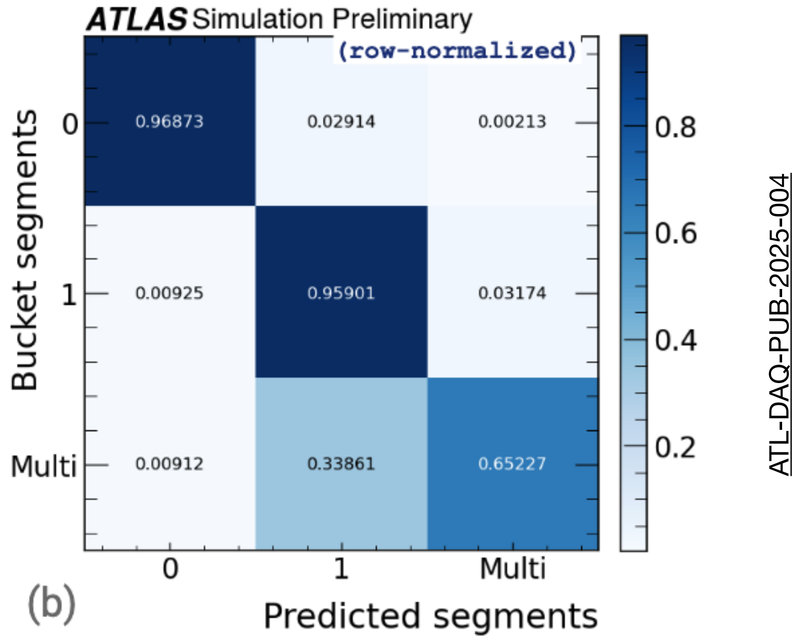


GNN Filtering

Performance







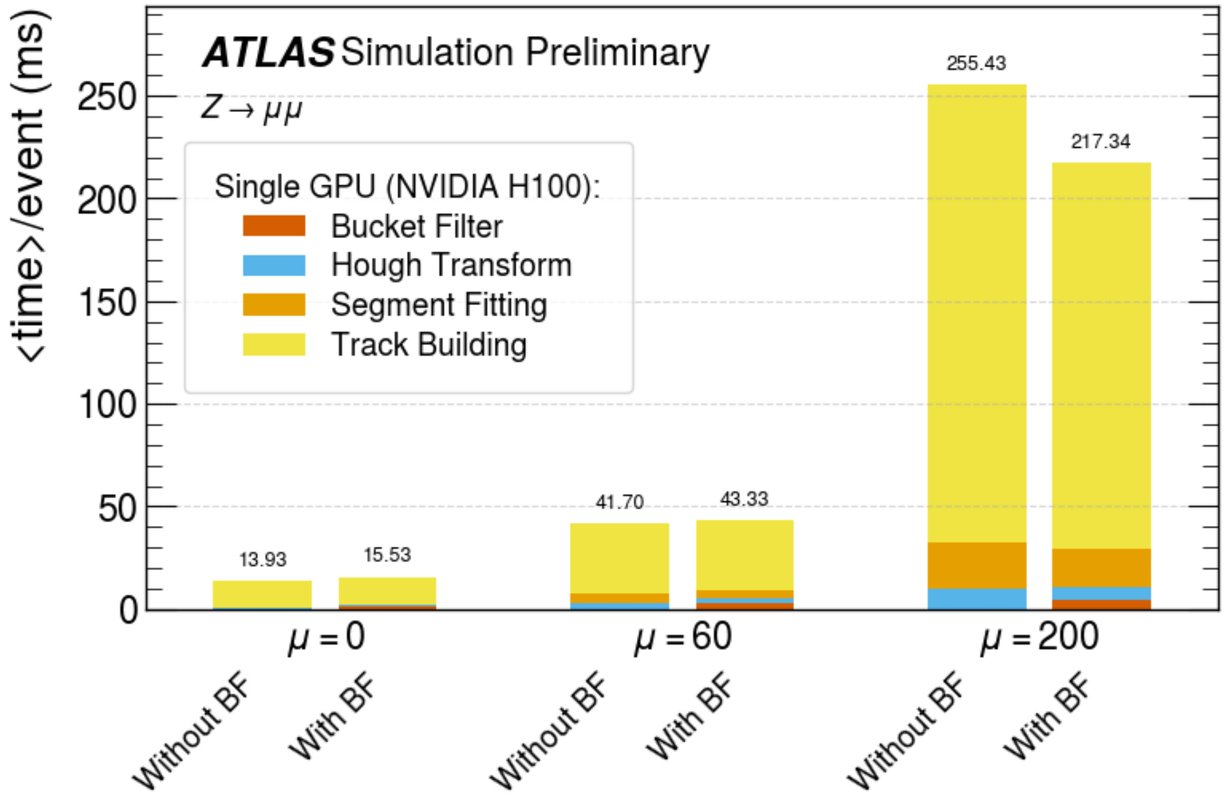
ATL-DAQ-PUB-2025-004

GNN Filtering introduces no bias into the reconstruction and can reject ~ 97 % of the buckets containing just noise

GNN Filtering

GNN Filtering integrated into Baseline

- Running the **baseline reconstruction** pipeline on top of the bucket filter allows for a combined **speed improvement** in the Hough Transform and Segment Fit of 42 ms
  - **Overall** speed **improvement** End-to-End: 15 %
- **Runtime** for the **Bucket Filter** including IO, graph building and inference measured on a NVIDIA H100 GPU: 4 - 6 ms

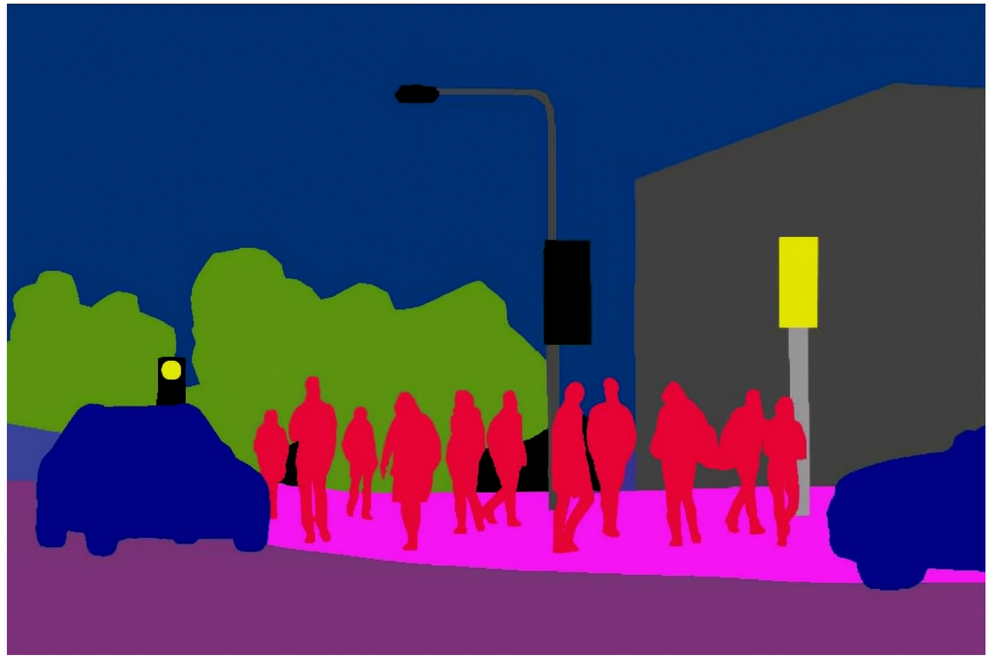


ATL-DAQ-PUB-2025-004

Idea | End-to-End ML for Muon Tracking

Can State-of-the-Art Image Segmentation find Muons?

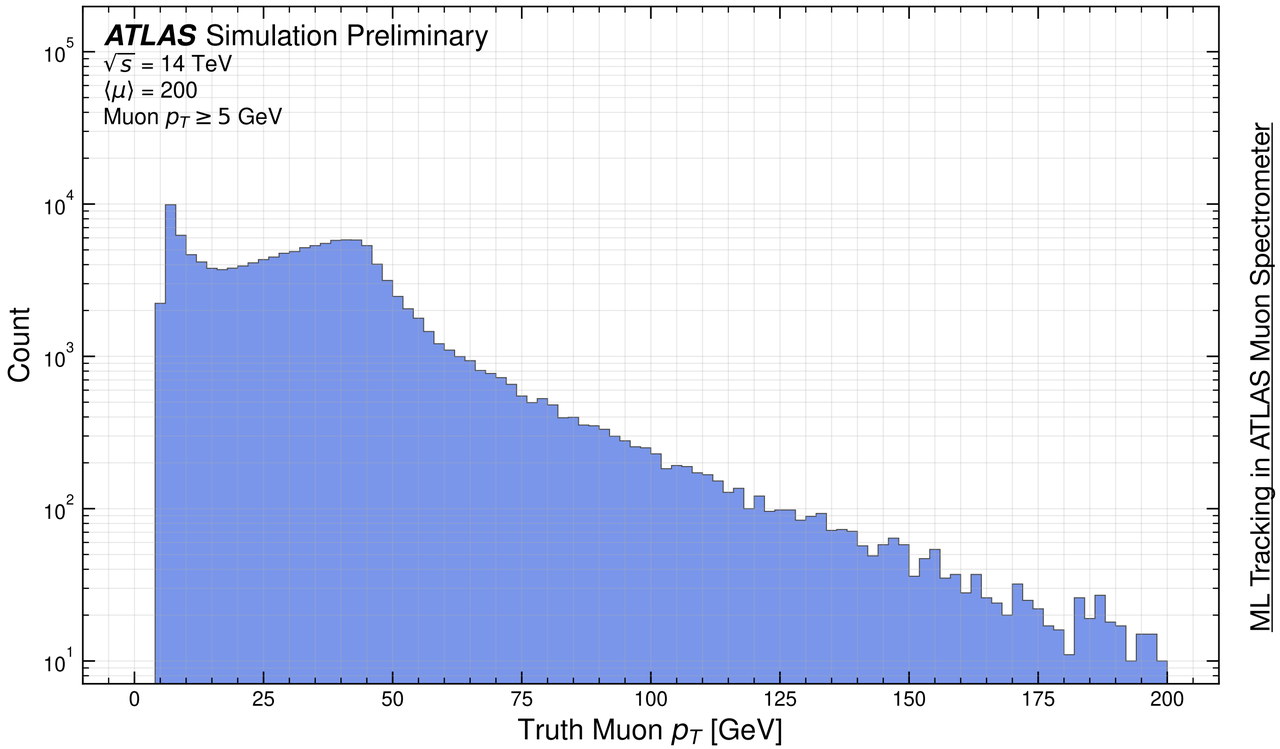
- **Approach:** Leverage MaskFormer open-source model by Meta AI, to treat charged particle tracking as a computer vision task (https://arxiv.org/pdf/2112.01527)
- **Collaboration:** Working closely with research group at UCL that published proof of concept on the TrackML dataset (https://arxiv.org/abs/2411.07149)



Example for semantic per-pixel image segmentation, generated using a diffusion model

Data

- Mix of PileUp 200 ***ATLAS*** Simulation $J/\psi, t\bar{t}$ and $Z \rightarrow \mu\mu$ events
- 40 true hits per event for an overall event size of 6900 hits (0.6 % purity)
- 97 % of the events contain ≤ 2 tracks originating from Muons (average 1.5 track per event)
- 1.4 Mio training events, 77 K validation events, 85 K testing events
- Criteria for selected tracks:
  - Minimum 3 true hits per track
  - $|\eta| \leq 2.7$
  - $p_T \geq 5$ GeV

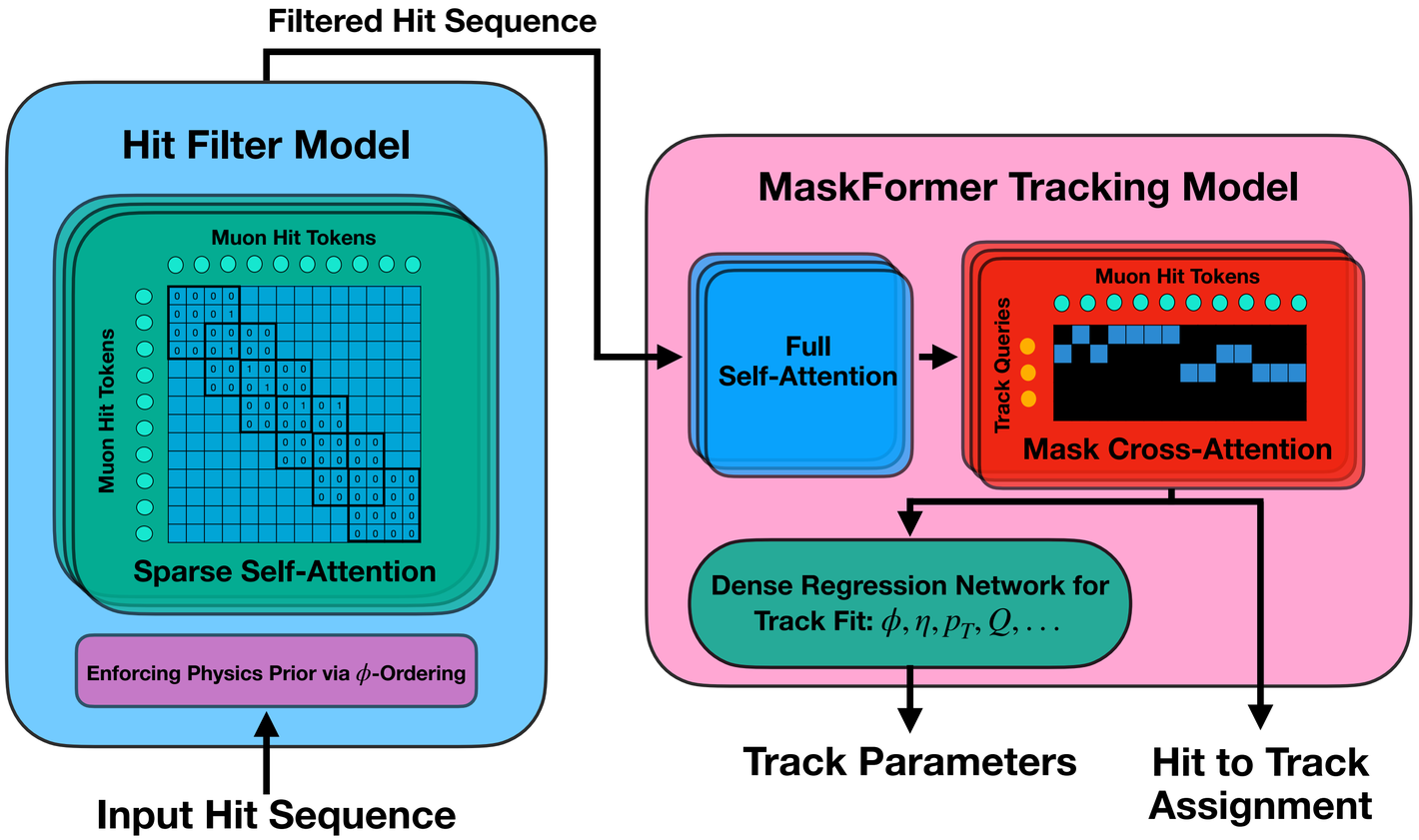


Model | IO & Architecture

HEP Adaptation

23 Input Variables per Hit:

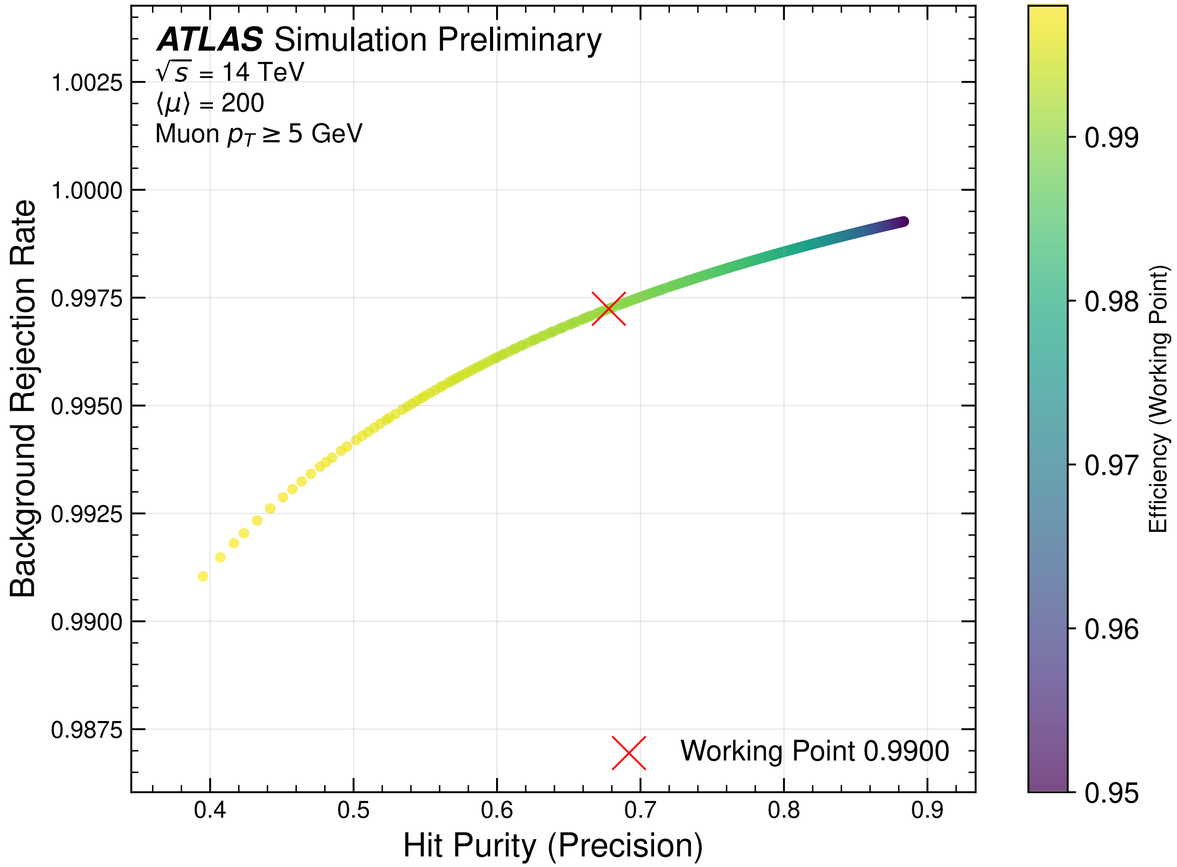
- xyz-coordinates of both ends of elongated detector technology
- Covariance: xx, xy, yx, yy
- Drift time
- Drift radius
- Detector geometry: station, channel, layer, technology, station phi/eta index
- Data Augmentation derived from xyz-coordinates: ϕ, theta, η, radial distance from beam axis, 3D distance from collision point



Filter Stage | Binary Classification

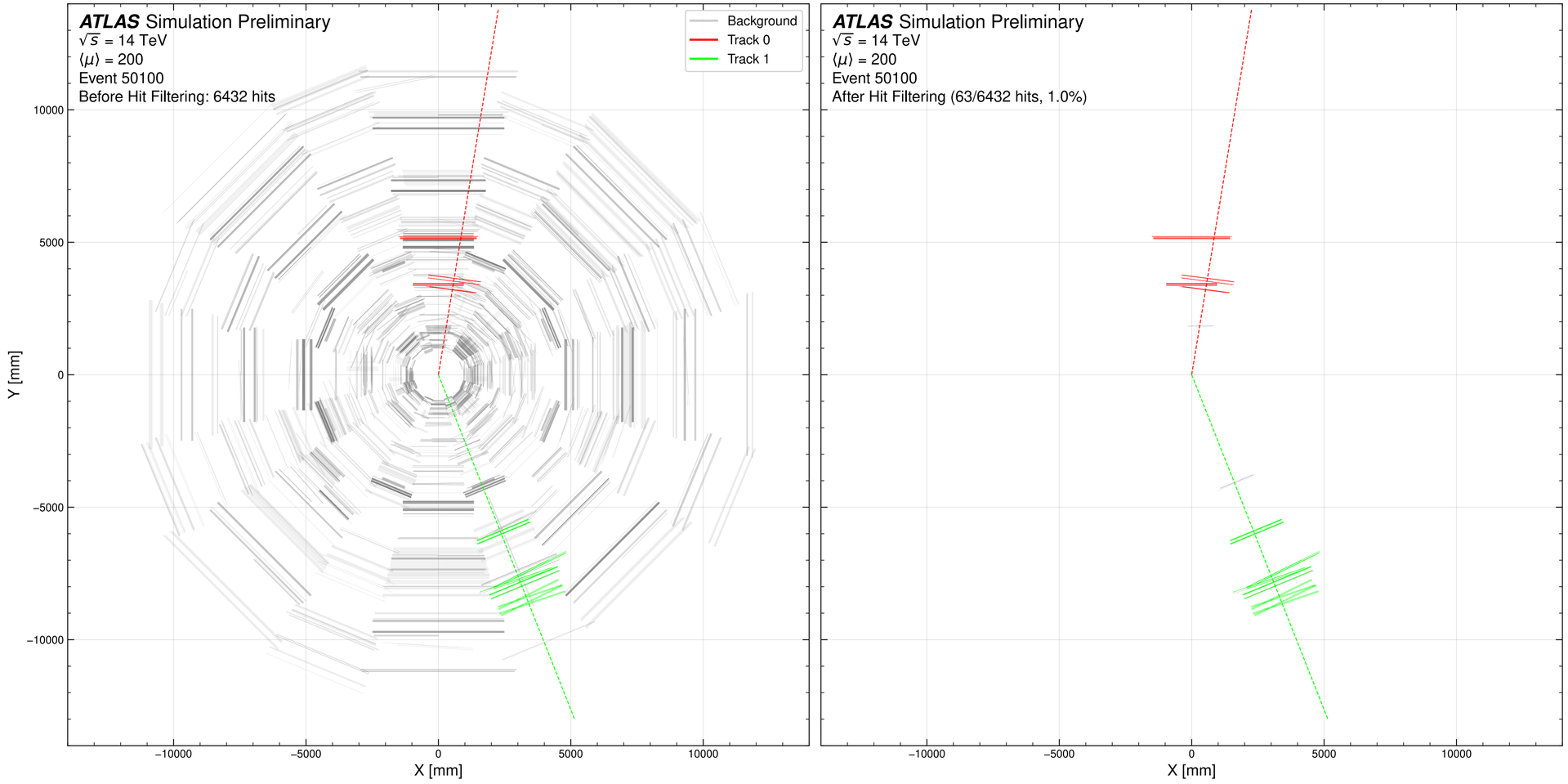
Attention is well suited to find Muon Trajectories

- **Model Size**: 1.4 Mio Parameters
- **Purity:** ~ 0.6 % —> ~ 66.5 % (True hit efficiency 99 %):
  - **Background rejection rate:** 99.7 % (6900 —> 55 Hits)
  - 99.7 % of tracks remain reconstructable after filtering (retain more than 3 true hits)
- **Area under the ROC curve**: 0.9997
- **Speed** per event**:** measured on NVIDIA RTX 3090 excluding IO (cost ~ 1.5 K CHF):
  - For Batch size 200: 2.2 ms (GPU vRAM utilization < 80 %)
  - For Batch size 1: 6.3 ms (GPU vRAM utilization < 5 %)



Filter Stage | Binary Classification

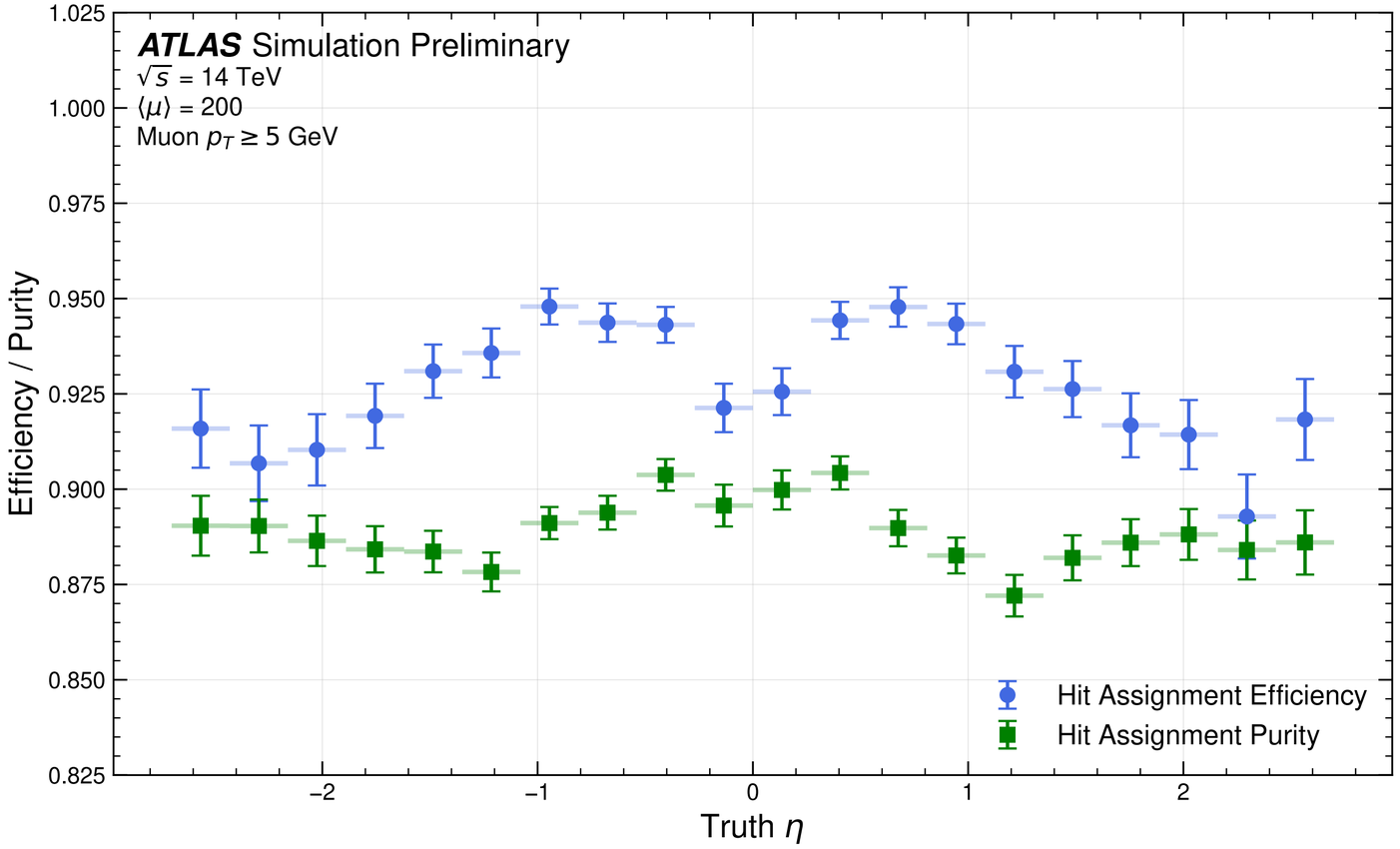
Performance as Event Displays Before & After Hit Filtering



Tracking Stage | Hit to Track Assignment

Performance as Purity and Efficiency per Track

- **Model size:** 97 K Parameter
- **Track Efficiency** for ML Regime: 98.0 % at fake rate of 5.1 %
- **Average** Hit Assignment **Efficiency**: 92.9%
- **Average** Hit Assignment **Purity**: 88.90%
- **Inference speed** per event on NVIDIA RTX 3090 excluding IO (cost 1.5 K CHF):
  - Batch size 200: ~ 0.07 ms (GPU vRAM utilization < 5 %)
  - Batch size 1: 13.6 ms (GPU vRAM utilization < 1 %)



Conclusions

- **End-to-End ML** shows potential to solve the combinatorial problem of track finding fast and converge on a Track Parameter Estimate within in 2.3 ms on cheap GPUs
- **The GNN bucket filter** manages to improve the baseline reconstruction from 255 ms to 217 ms (15 % improvement running on an H100 GPU), while introducing no bias into the baseline algorithm

What is next to come for End-to-End ML:

- Global Filtering should be tried as preprocessing to the baseline reco, due to purity improvement: 0.6 % —> 66.5 %
- Analysis Settings for LLP decays in the middle of the Muon Spectrometer
- Hit to Track assignment solving in Event Filter
- Assist for χ^2-Track Fit convergence parameters in Event Filter
- Improvements in robustness against detector aging
- Improvements in Physics Efficiency

Get in Touch if you would like to collaborate! Any help is welcome 🤝:
jonathan.renusch@cern.ch

Backup

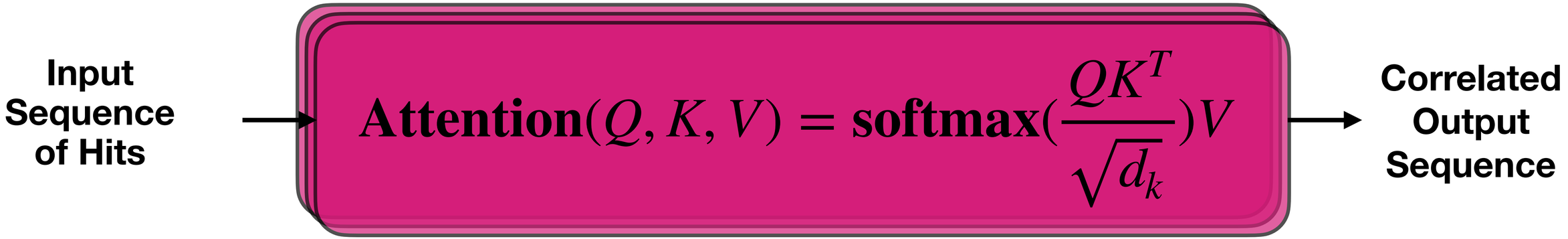
Next Steps | End-to-End ML

Just getting started

- Manpower for now limited to Master's student's 4 month effort + Markus Elsing —> find more people
- Make the networks run in ATHENA
- Bayesian Hyperaparameter tuning
- Improvements in Regression
- Left right ambiguity solving for track fit
- Speed studies: Torch Compile, TensorRT, Quantization, Pruning, Model Distillation

Get in Touch if you would like to collaborate! Any help is welcome 🤝:
jonathan.renusch@cern.ch

Model | Machine Learning Background

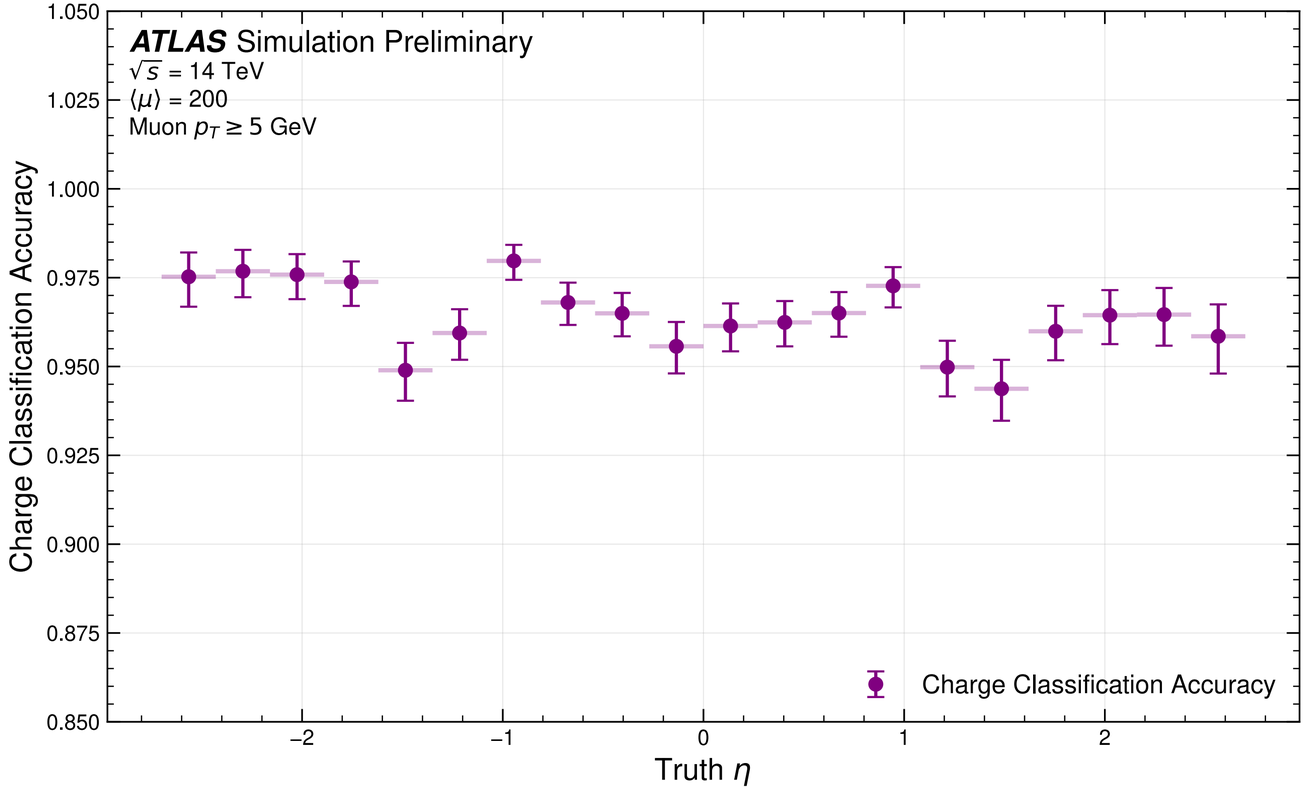


- Attention Mechanism calculates context-agnostic attention scores within n^2 “attention-matrix”
- Paper “Attention is all you need” enabled modern Machine Learning revolution, by proposing the “Transformer” architecture build around the Attention Mechanism 2017 (https://arxiv.org/pdf/1706.03762)
- Attention-based Transformer Models are leading for major Deep Learning benchmarks:
  - Natural Language Processing: GPT, Gemini, Claude, etc.
  - Computer Vision: MaskFormer, SAM, DINO, etc.
- Longevity: Attention Mechanism is a “sustainable technology” choice, as it is likely to become faster “on its own”
  - Significant resources are funneled towards improving Hardware and Algorithm (Google, Meta, Anthropic, OpenAI, Academic labs, to name a few)

Tracking Stage | Track Parameter Estimate

Work in Progress on Track Parameter Estimate for η, ϕ, q and p_T

- Charge classification shows potential
- Track parameter estimate for η, ϕ and p_T is convergent but not yet competitive with the baseline χ^2-Fit
- Charge Classification Accuracy for balanced (50:50) sample: 96.35%



Tracking Stage | Hit to Track Assignment

Performance as Double Matching Efficiency

- **Double** Matching **Efficiency** (50% Working Point): 94.59%
- **Inference speed** per event on NVIDIA RTX 3090 excluding IO:
  - Batch size 200: ~ 0.07 ms (GPU vRAM utilization < 5 %)
  - Batch size 1: 13.62 ms (GPU vRAM utilization < 1 %)