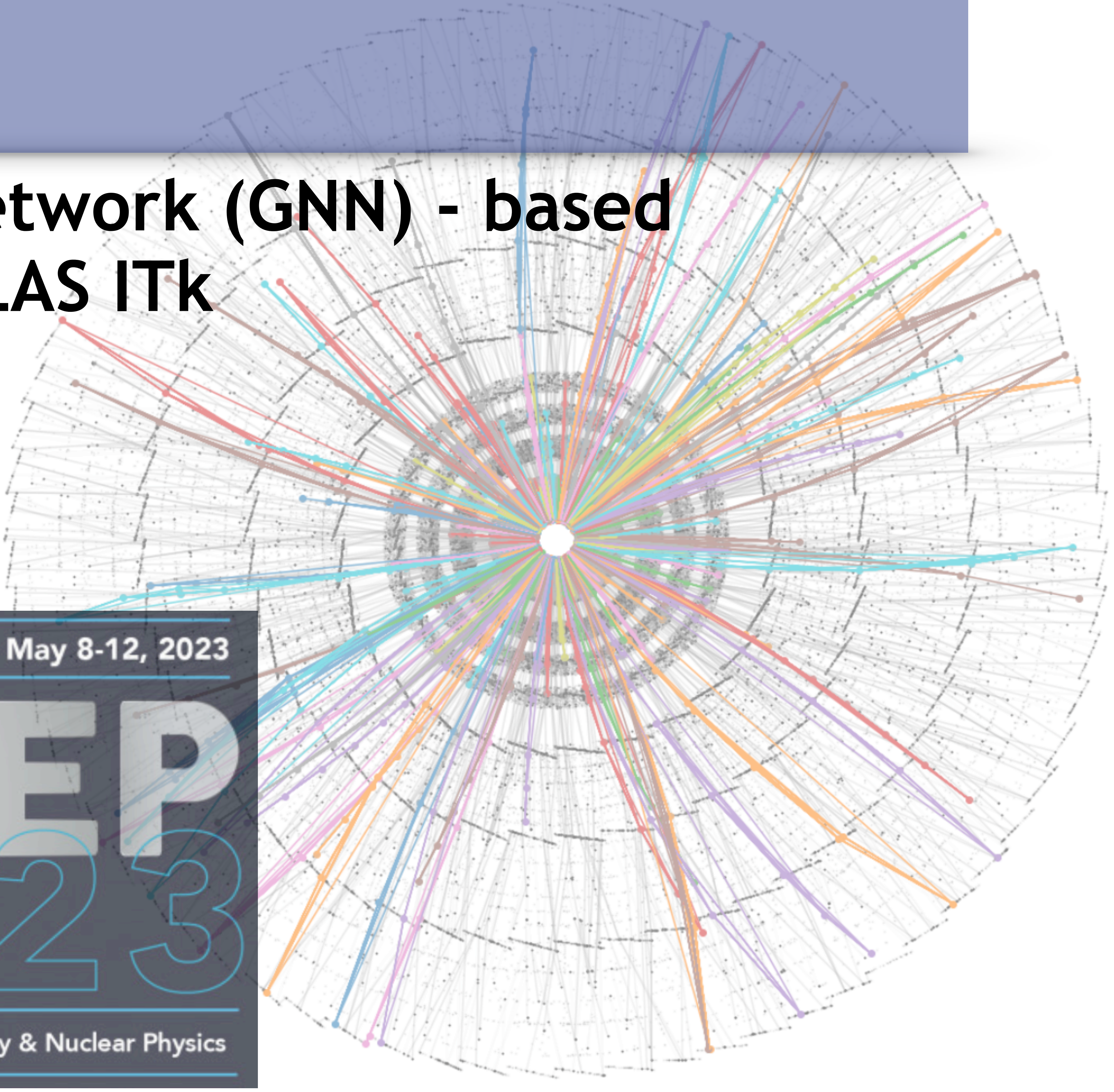
Novel fully-heterogeneous GNN designs for track reconstruction at the HL-LHC

Towards deployment of high performance Graph Neural Network (GNN) - based algorithms for charged particle track reconstruction in ATLAS ITk

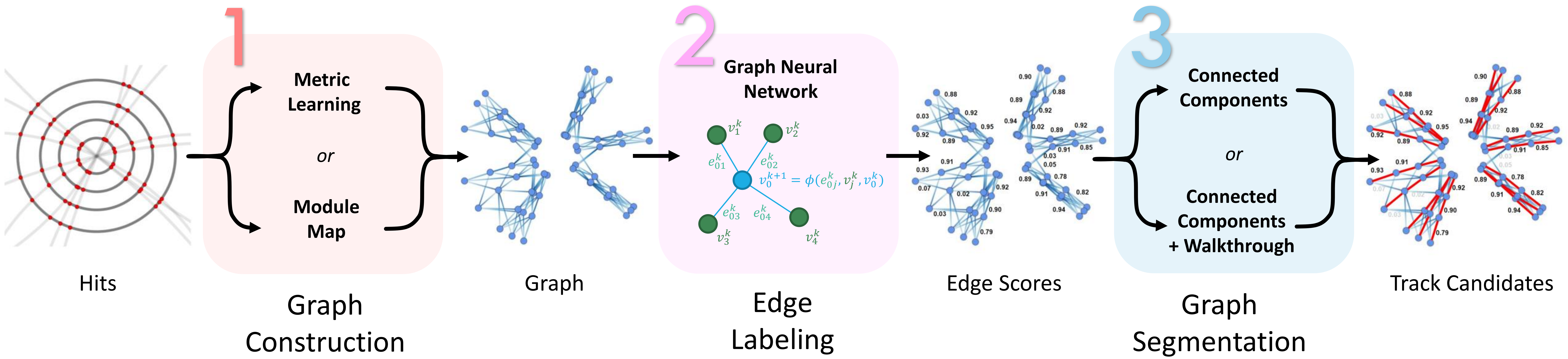


Sylvain Caillou

Laboratoire des 2 Infinis-Toulouse

GNN-based track reconstruction for HL-LHC

- Start of HL-LHC (2028) => Rate of collisions will be increased by an order of magnitude
- Current algorithms will not be able to cope with the complexity and rate of the data recorded
- Geometric Learning solution explored by Exa.TrkX Project and L2IT [P. Calafiura, CHEP2023] [X. Ju, CHEP2023]
- « ATLAS ITk Track Reconstruction with a GNN-based pipeline » C.Rougier (CTD2022)



1) Detector data represented as graph
- *Nodes* are hits in the detector
- *Edges* are *possible connections* between nodes
- *True edges* are connections between successive hits from *the same particle of interest*

2) GNN learns *deep geometric patterns* of the particle tracks and classifies edges between true and fake edges

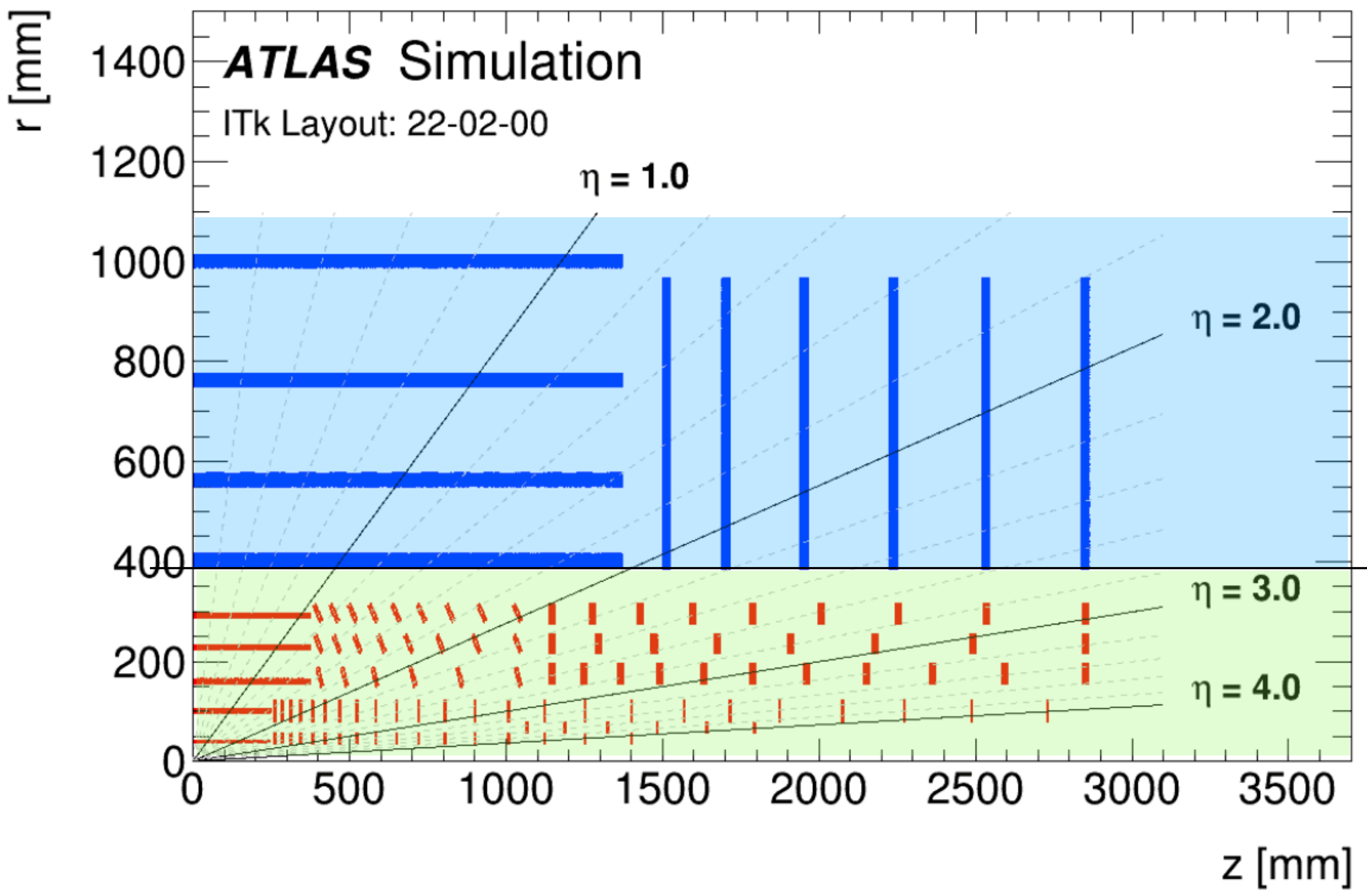
3) Final algorithm operates on scored graph to build *track candidates*

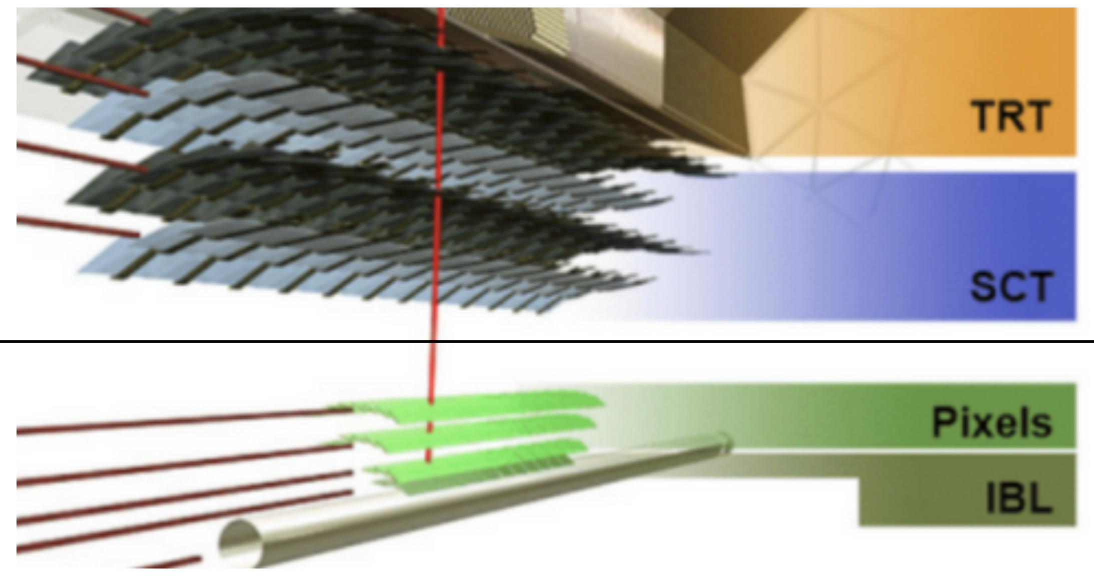
CTD2022 Data and graph representation

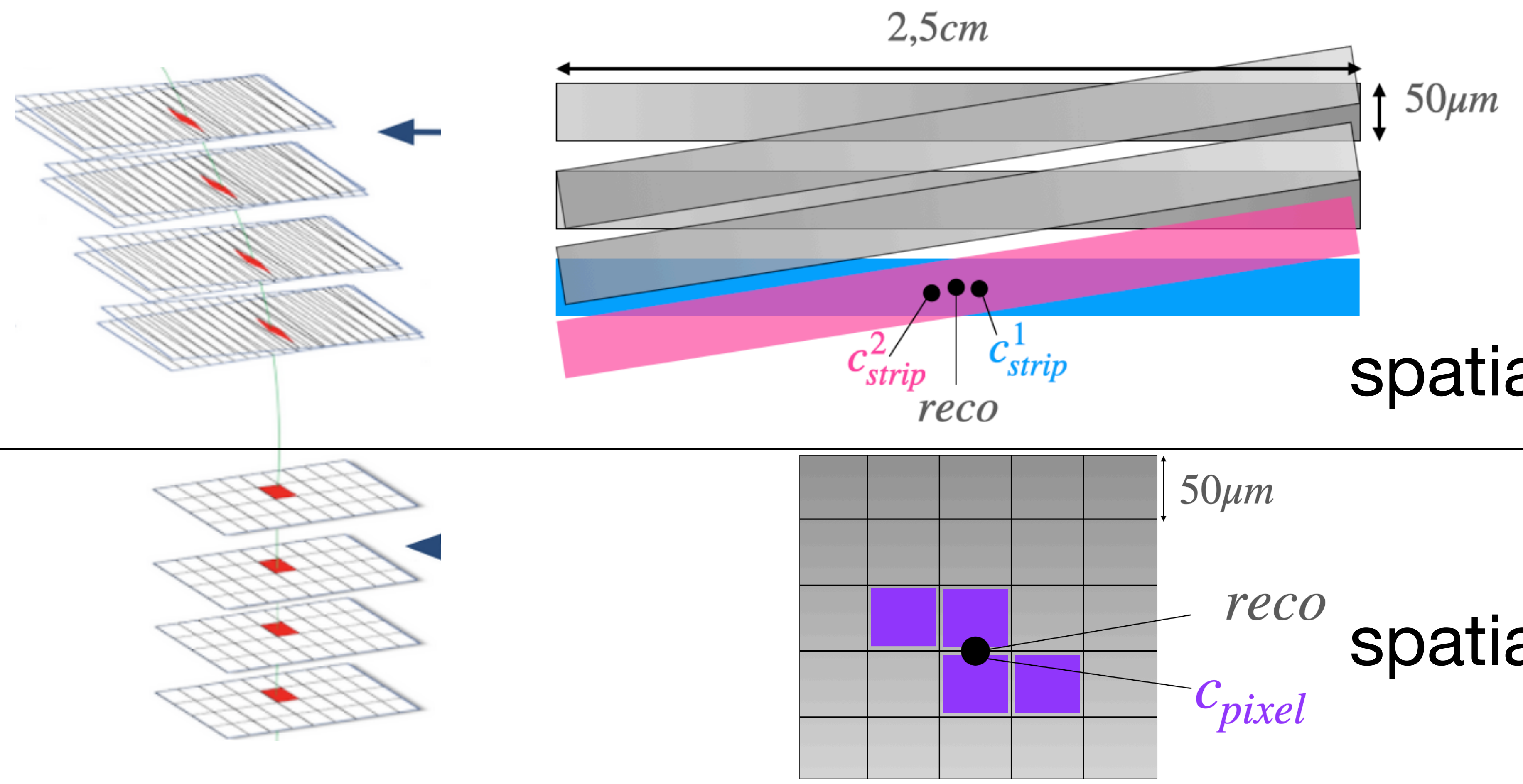
Results presented last year at CTD2022 (Princeton)
Nodes features are euclidian geometric coordinates of the Space Points
They are reconstructed from **2 different technologies** in the detector (i.e. heterogeneous data)

STRIP technology
Space Point reconstructed from 2 STRIP clusters

$$r^{reco}, \varphi^{reco}, z^{reco} = f_{strip}(c^1_{strip}, c^2_{strip})$$







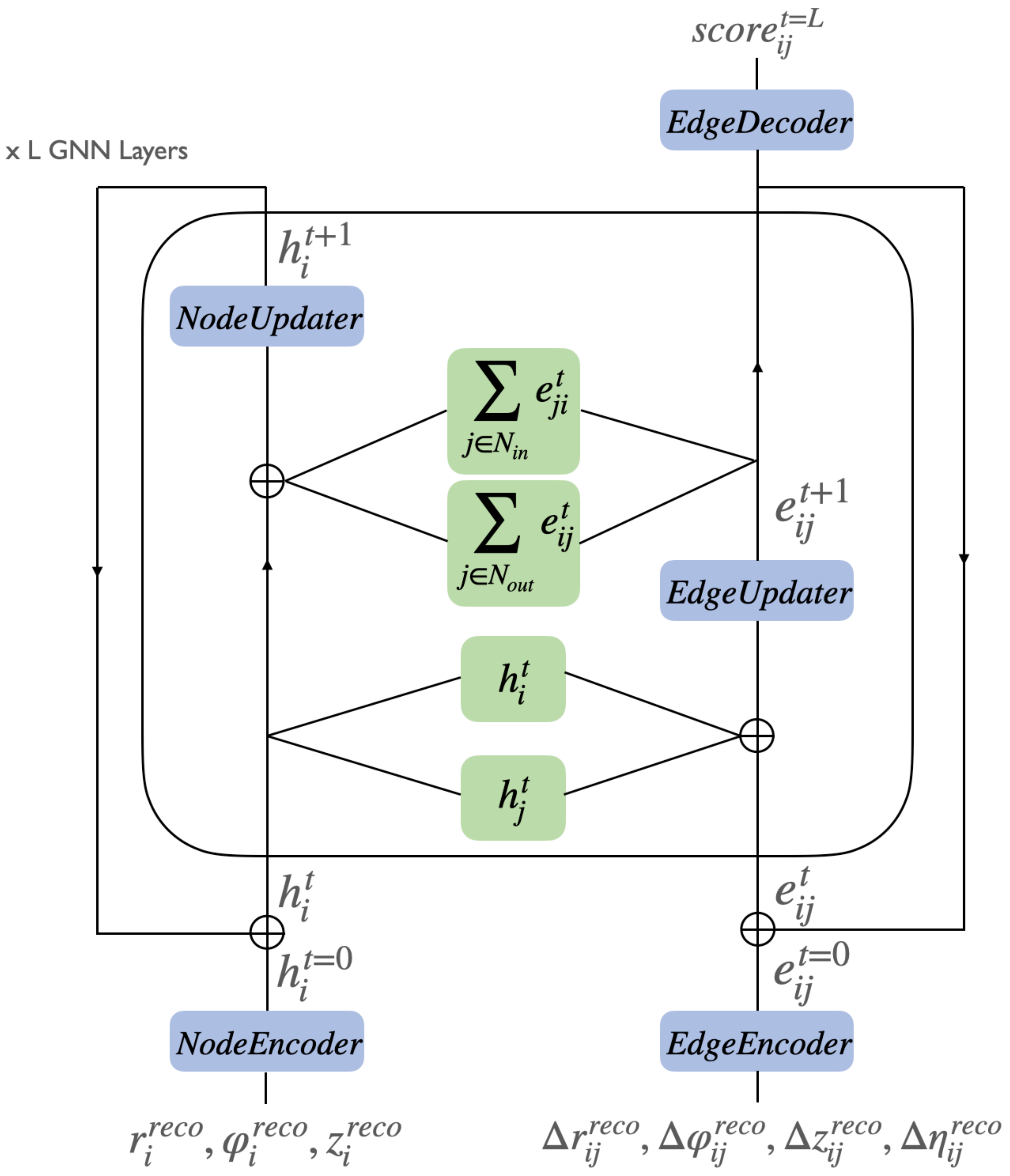
Low
spatial resolution

High
spatial resolution

PIXEL technology
Space Point reconstructed from 1 PIXEL cluster

$$r^{reco}, \varphi^{reco}, z^{reco} = f_{pixel}(c_{pixel})$$

CTD2022 Message Passing Neural Network



Decode (scoring) stage

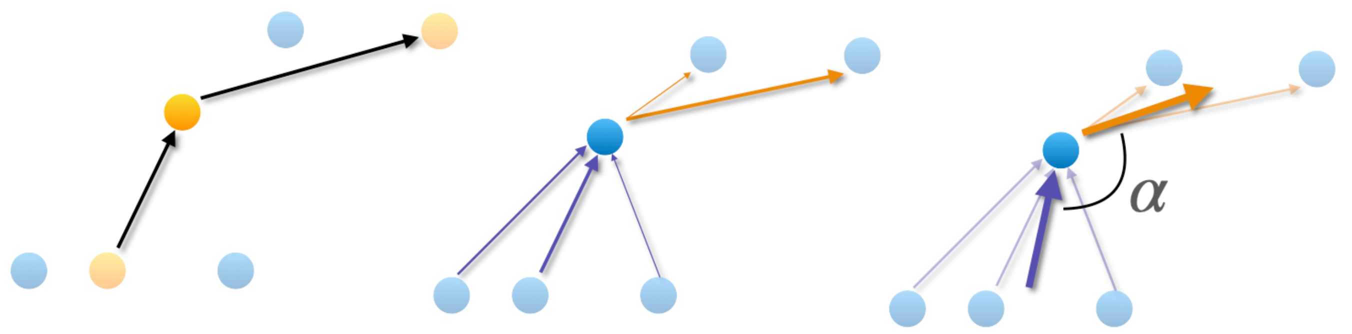
Edge latent is projected to a scalar value which is the score of the edge

$$\boxed{score_{ij}^{t=L} \leftarrow EdgeDecoder(e_{ij}^{t=L})}$$

GNN stage

- Edge latent is updated taking into account latent of source and destination nodes
- Node latent is updated from a ***separate* aggregation of incoming and outcoming edges**

$$e_{ij}^{t+1} \leftarrow \psi(e_{ij}^t \,|\, h_i^t \,|\, h_j^t)$$
$$h_i^{t+1} \leftarrow \phi(h_i^t \,|\, \sum_{j \in N_{in}} e_{ji}^t \,|\, \sum_{j \in N_{out}} e_{ij}^t)$$



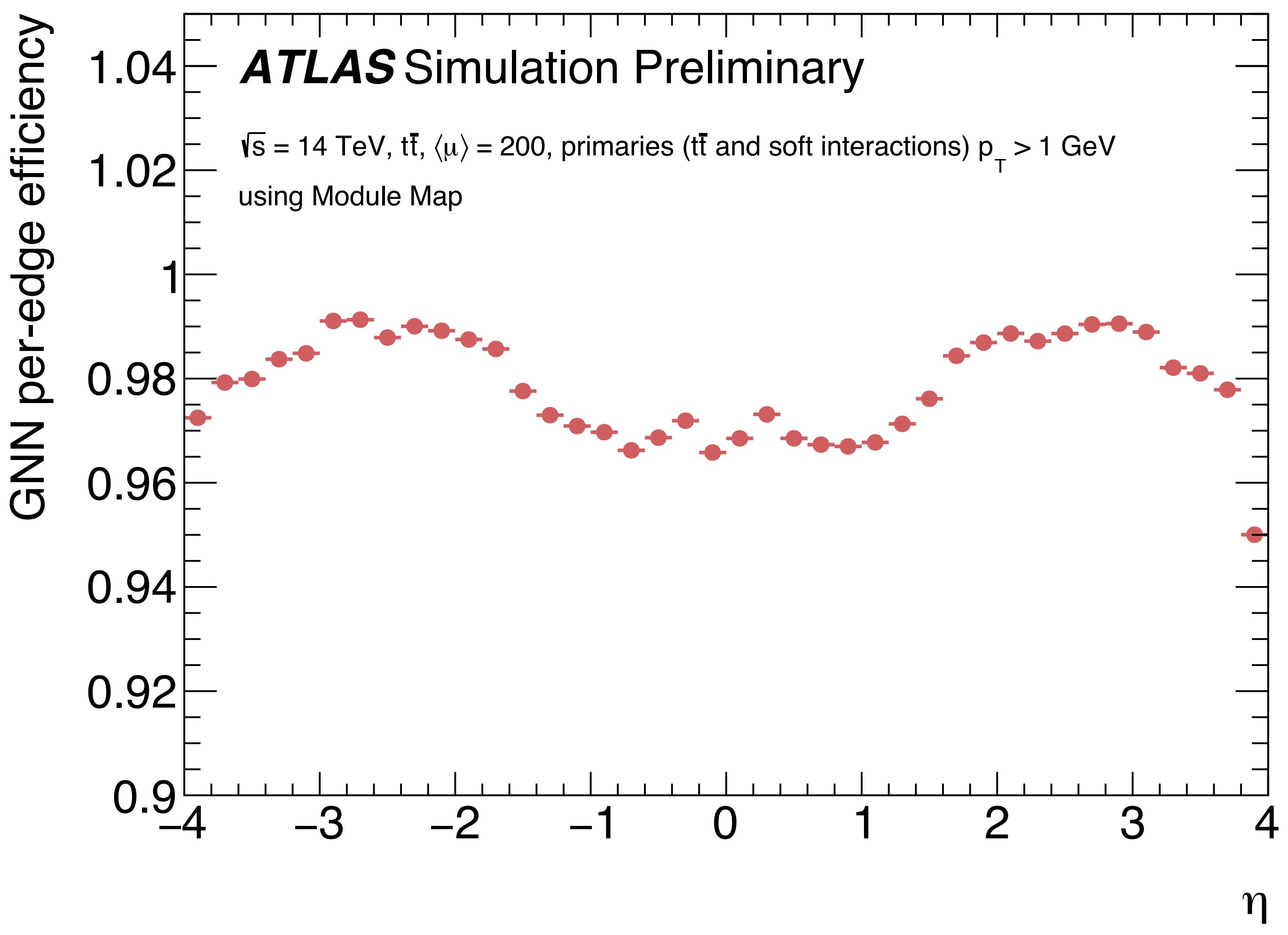
Encode stage

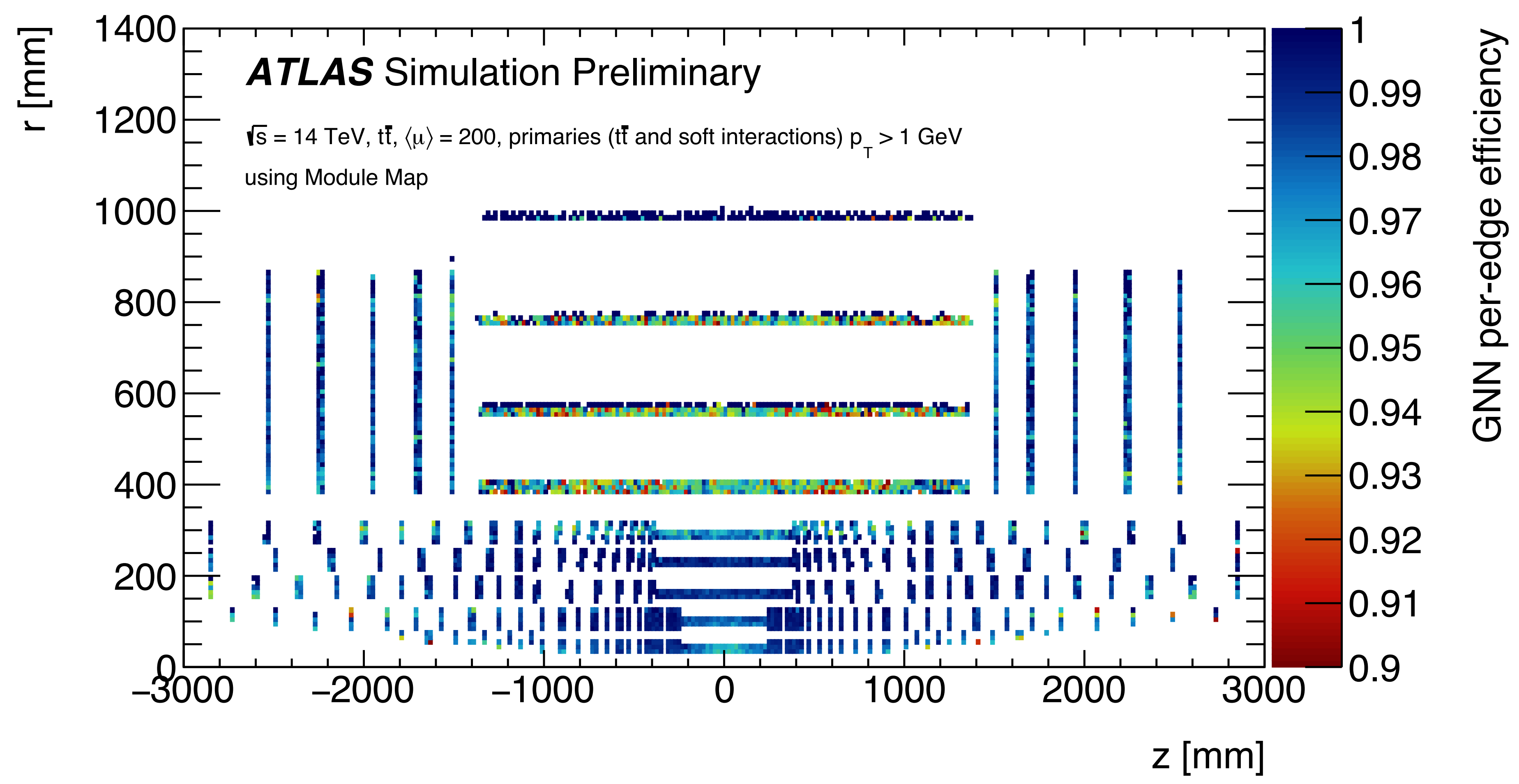
- Node euclidian features are projected into latent space
- Edge *preprocessed* features are projected into latent space

$$h_i^{t=0} \leftarrow NodeEncoder(r_i^{reco}, \varphi_i^{reco}, z_i^{reco})$$
$$e_{ij}^{t=0} \leftarrow EdgeEncoder(\Delta r_{ij}^{reco}, \Delta\varphi_{ij}^{reco}, \Delta z_{ij}^{reco}, \Delta\eta_{ij}^{reco})$$

CTD2022 GNN performance on ITk simulated data

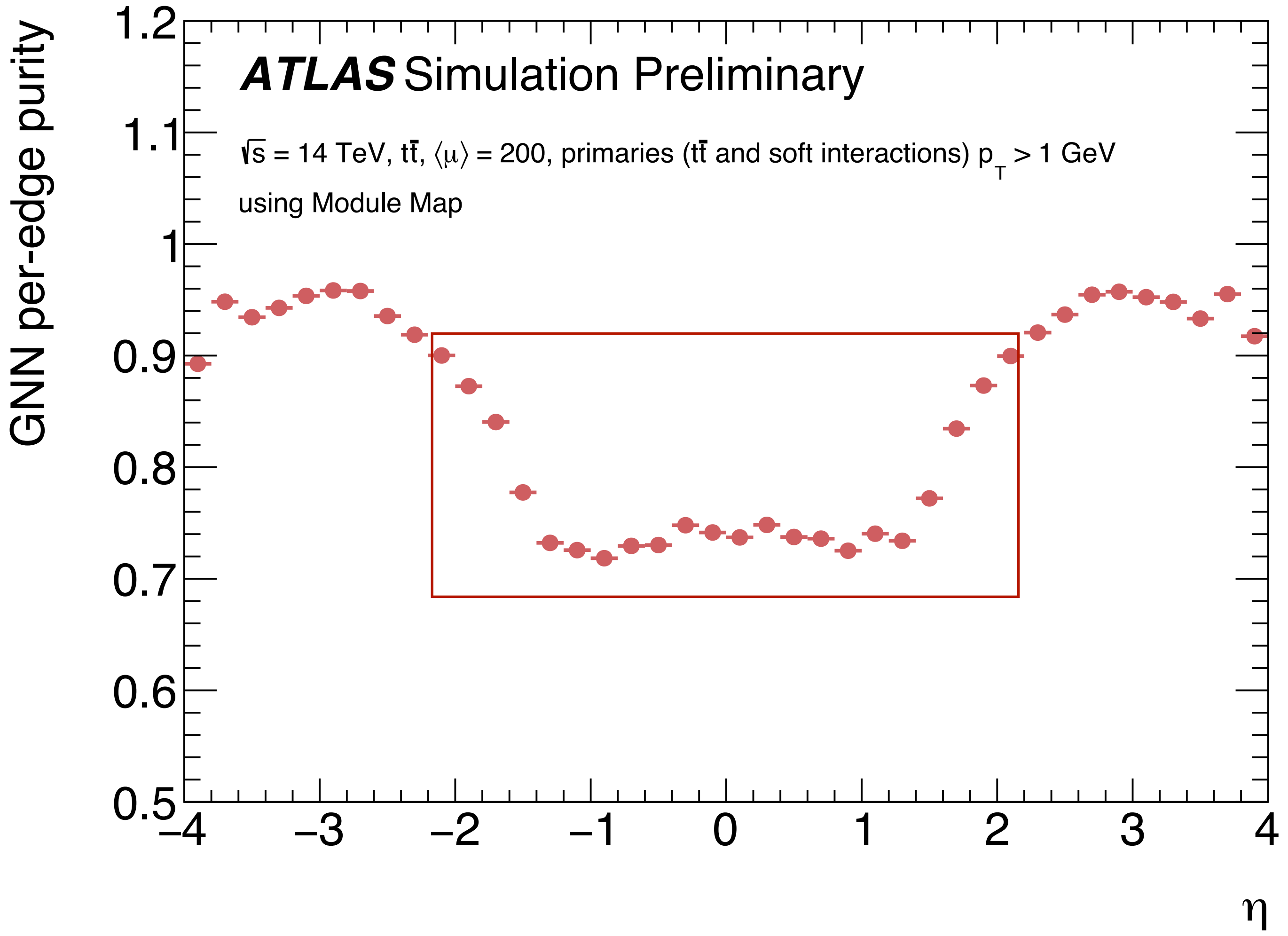
CTD2022

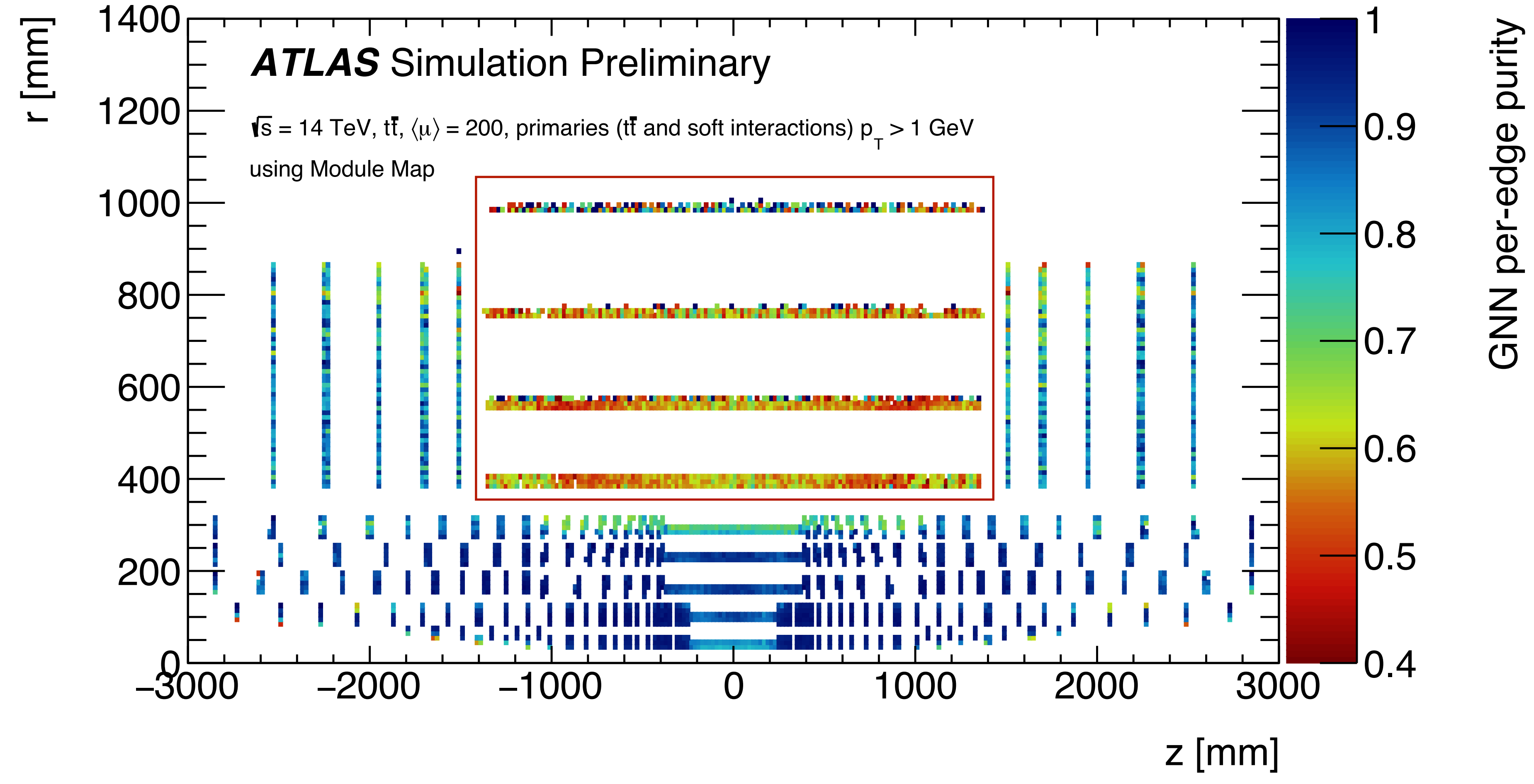




$$efficiency = \frac{n_{true>treshold}}{n_{true}}$$

- High efficiency: ~98%
- Slight drop of efficiency in the STRIP BARREL region

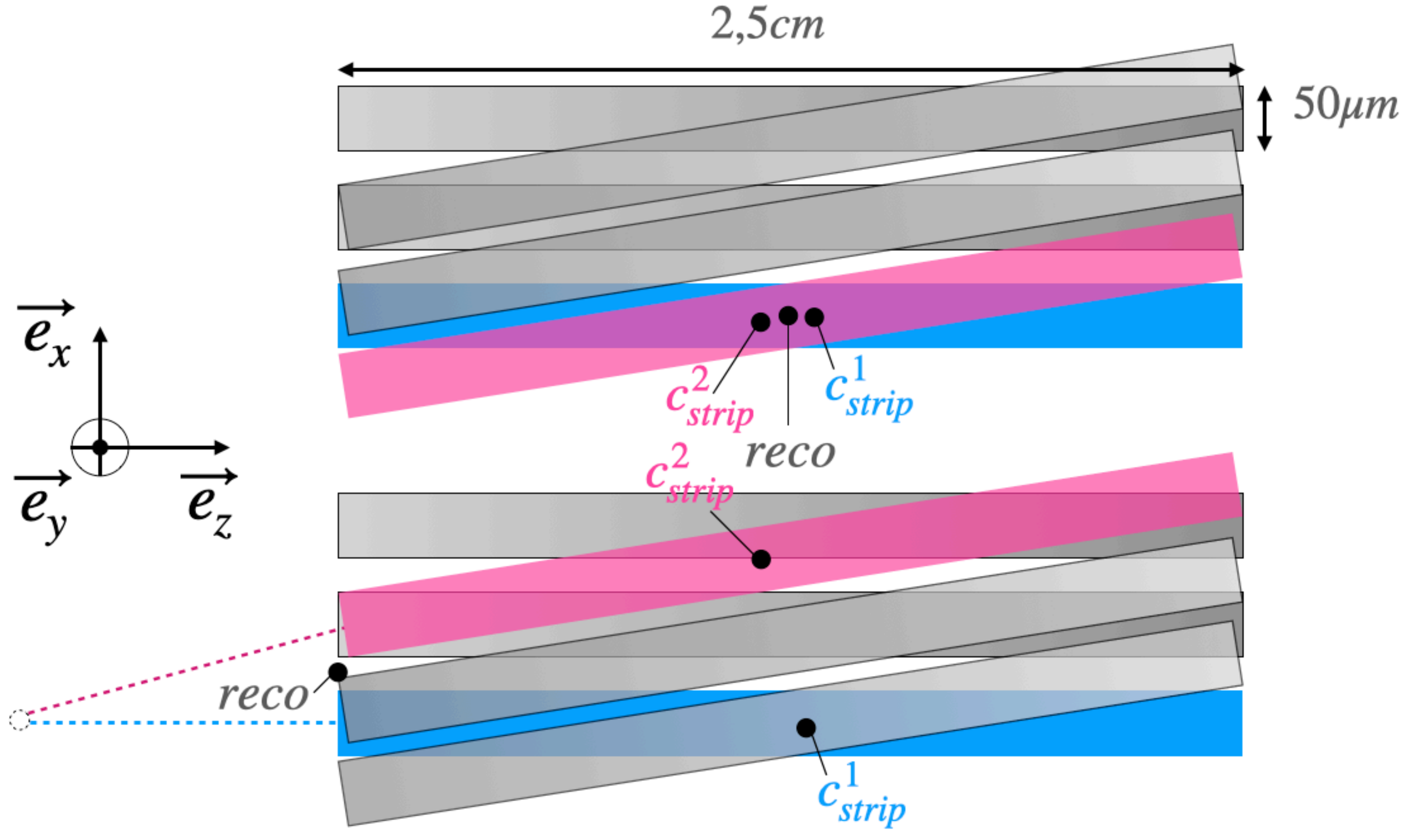


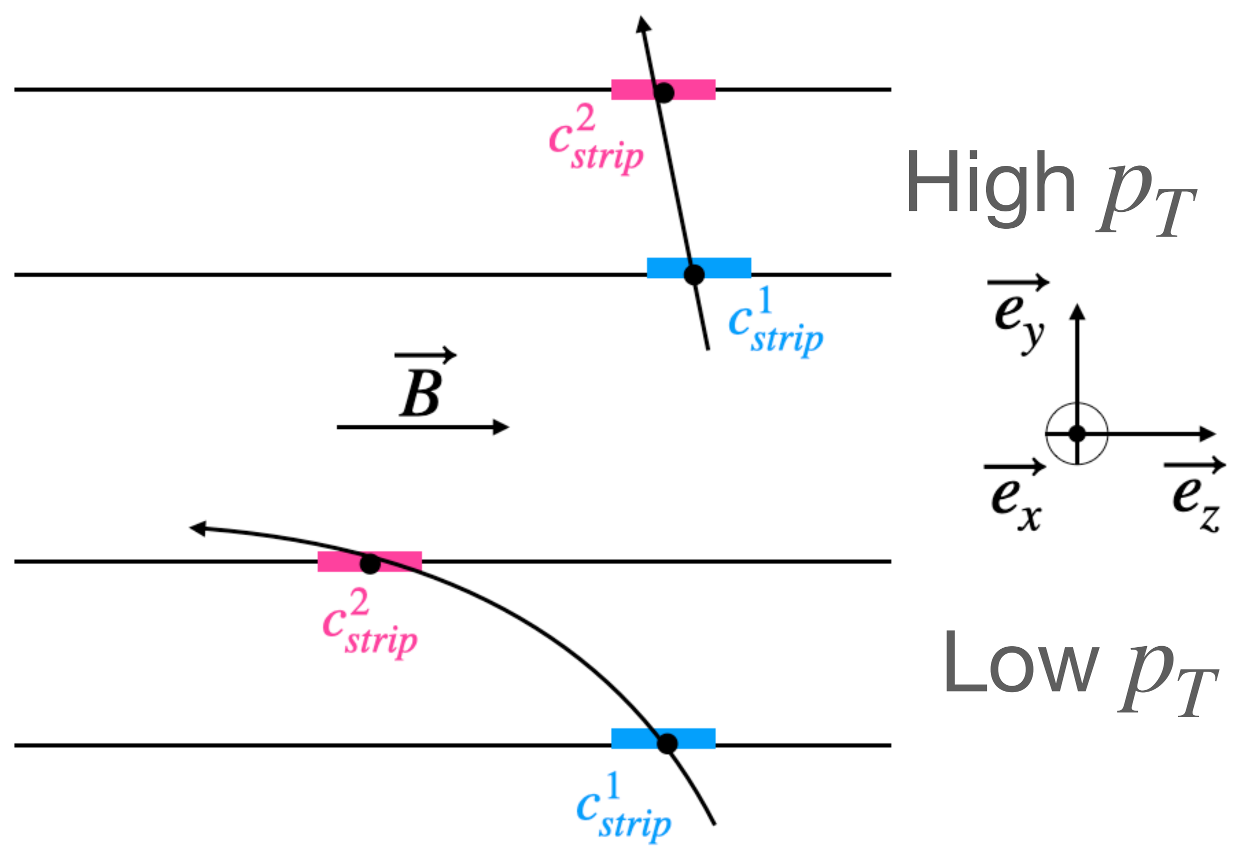


$$purity = \frac{n_{true>treshold}}{n_{true>treshold} + n_{fake>treshold}}$$

- Global purity ~85%
- Important drop of purity in the STRIP BARREL region

STRIP low spatial resolution and Heterogenous Data

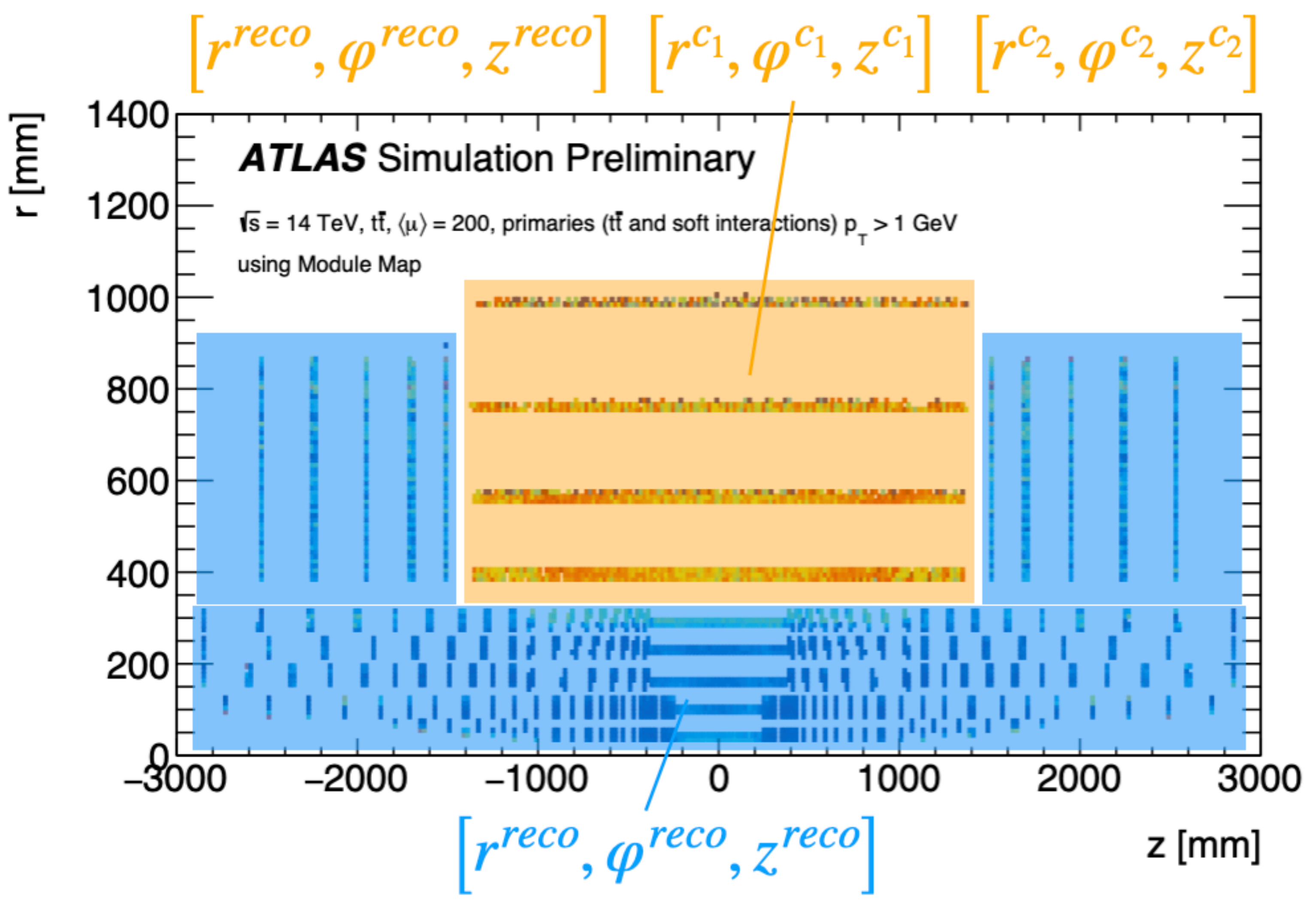




$$r^{reco}, \varphi^{reco}, z^{reco} = f_{strip}(c^1_{strip}, c^2_{strip})$$

The mechanism that led to the poor purity in CTD2022 results is understood: straight line approximation used in the ATLAS space point reconstruction leads to poor z resolution ($O(cm)$!) at low p_T

Impossible to exactly reconstruct space point position without knowing curvature of the track !



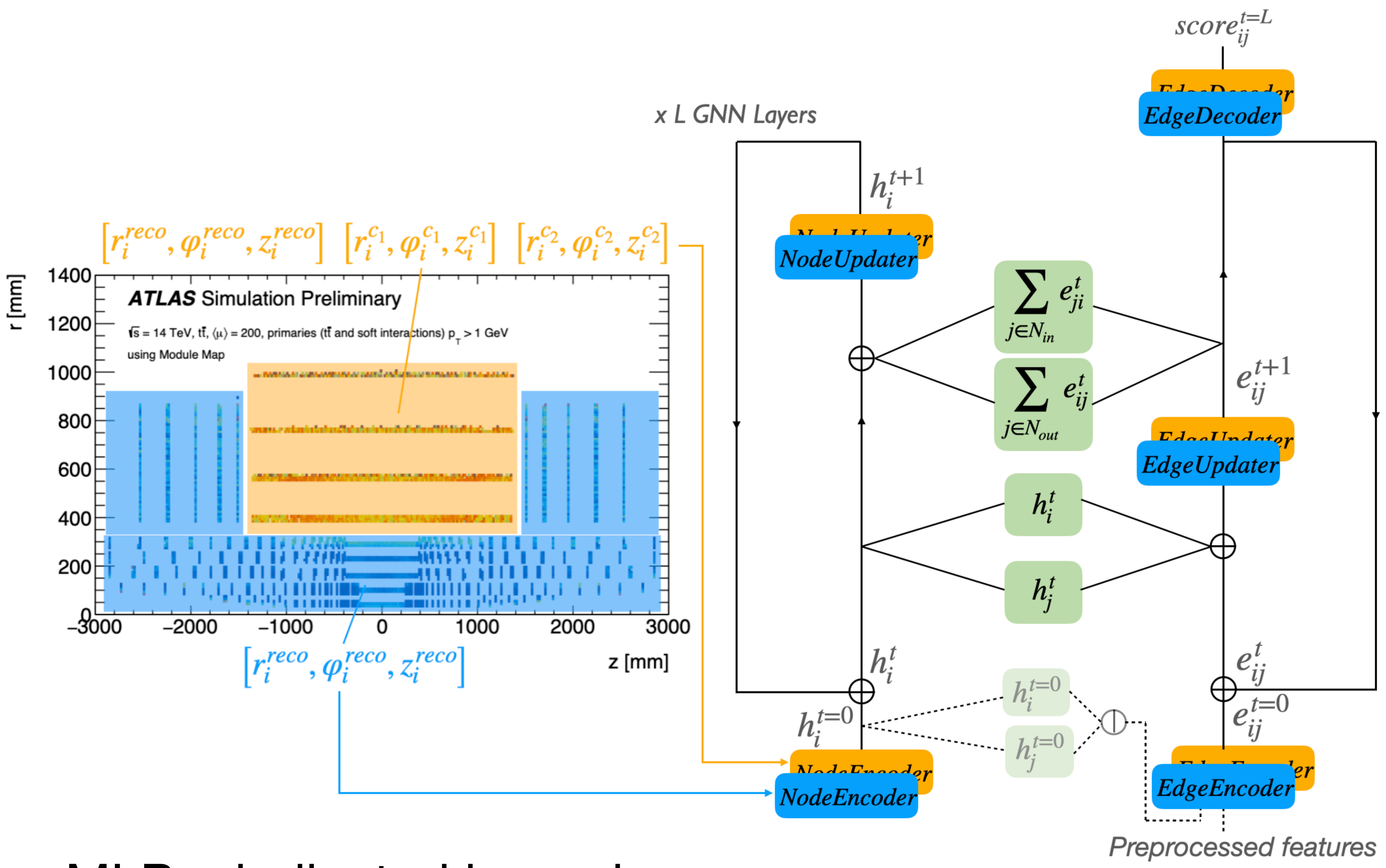
Key Idea: Give as node features in STRIP BARREL the **STRIP clusters data**
Hopefully the GNN will be able to learn a **better representation of the Space Point**

GNN combines space points into tracks, i.e. has access to curvature !

Heterogeneous data and GNN designs

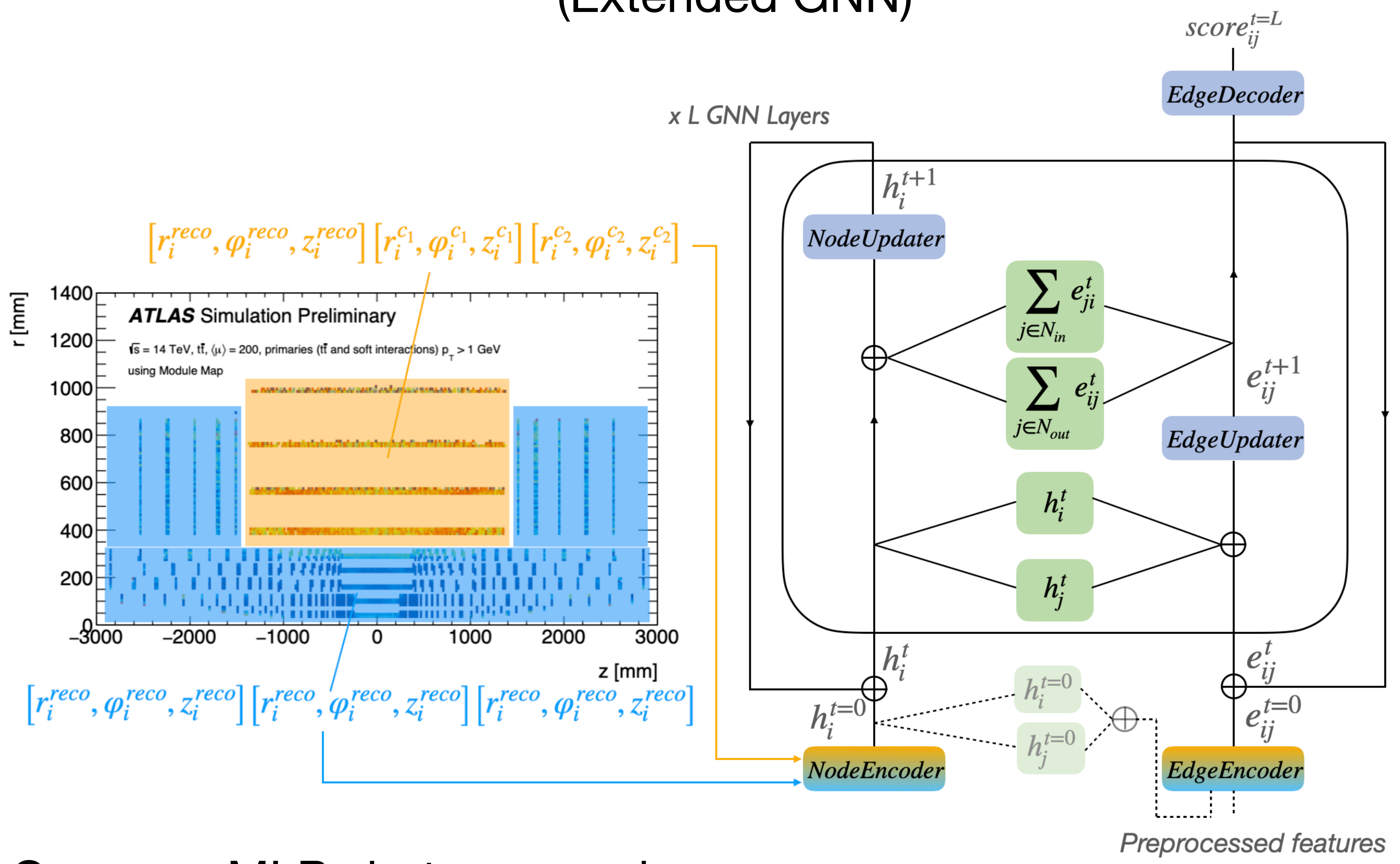
Two possible designs to handle Heterogenous Data

Heterogeneous Data + Heterogeneous GNN



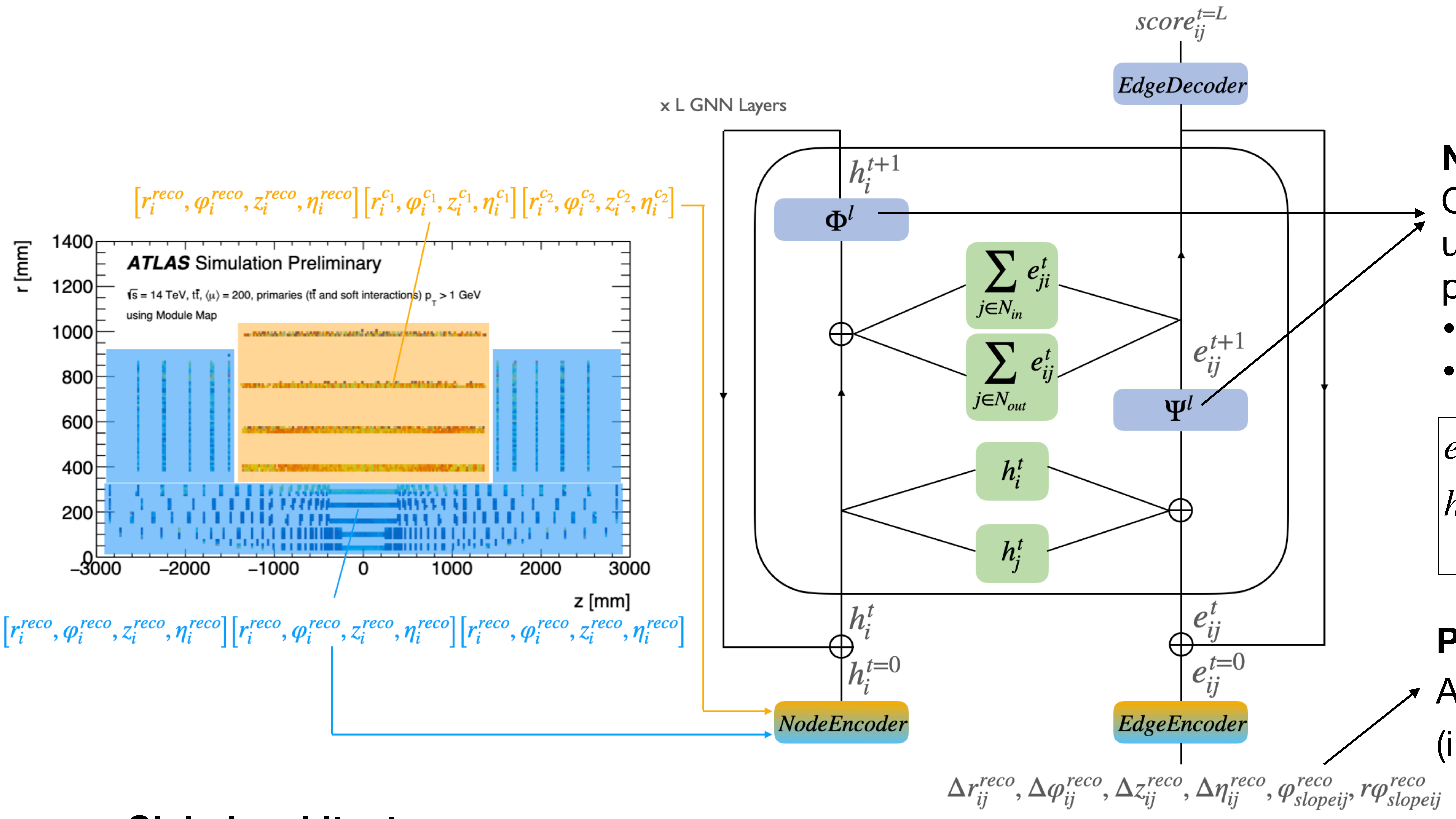
MLPs dedicated by region
+Specialized MLP per heterogenous region
-Hard to train / synchronise the learning for efficient Message Passing between region

Heterogeneous Data + Homogeneous GNN (Extended GNN)



Common MLPs between regions
+Common learning for efficient Message Passing between region
-No specialized MLP per heterogenous region

Additional GNN improvements



Non recurrent GNN layers

One instance of edge updater and node updater per GNN layer (No shared parameters)

- Better function estimation
- Avoid vanishing gradients

$$e_{ij}^{t+1} \leftarrow \psi^l(e_{ij}^t \,|\, h_i^t \,|\, h_j^t)$$
$$h_i^{t+1} \leftarrow \phi^l(h_i^t \,|\, \sum_{j \in N_{in}} e_{ji}^t \,|\, \sum_{j \in N_{out}} e_{ij}^t)$$

Preprocessed edge features

Add φ_{slope} which is closely related to p_T (informative for the curvature of the tracks)

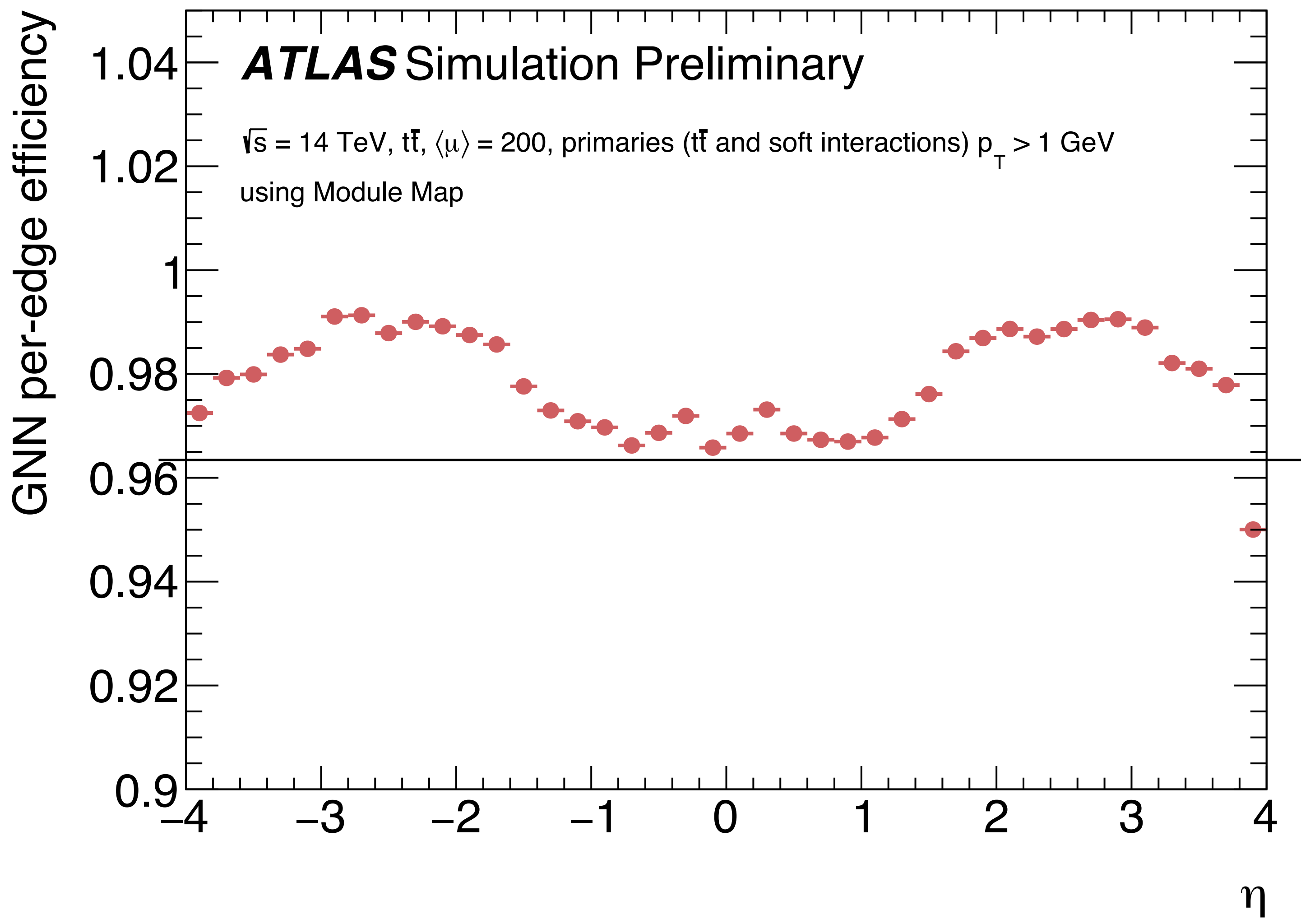
Global architecture

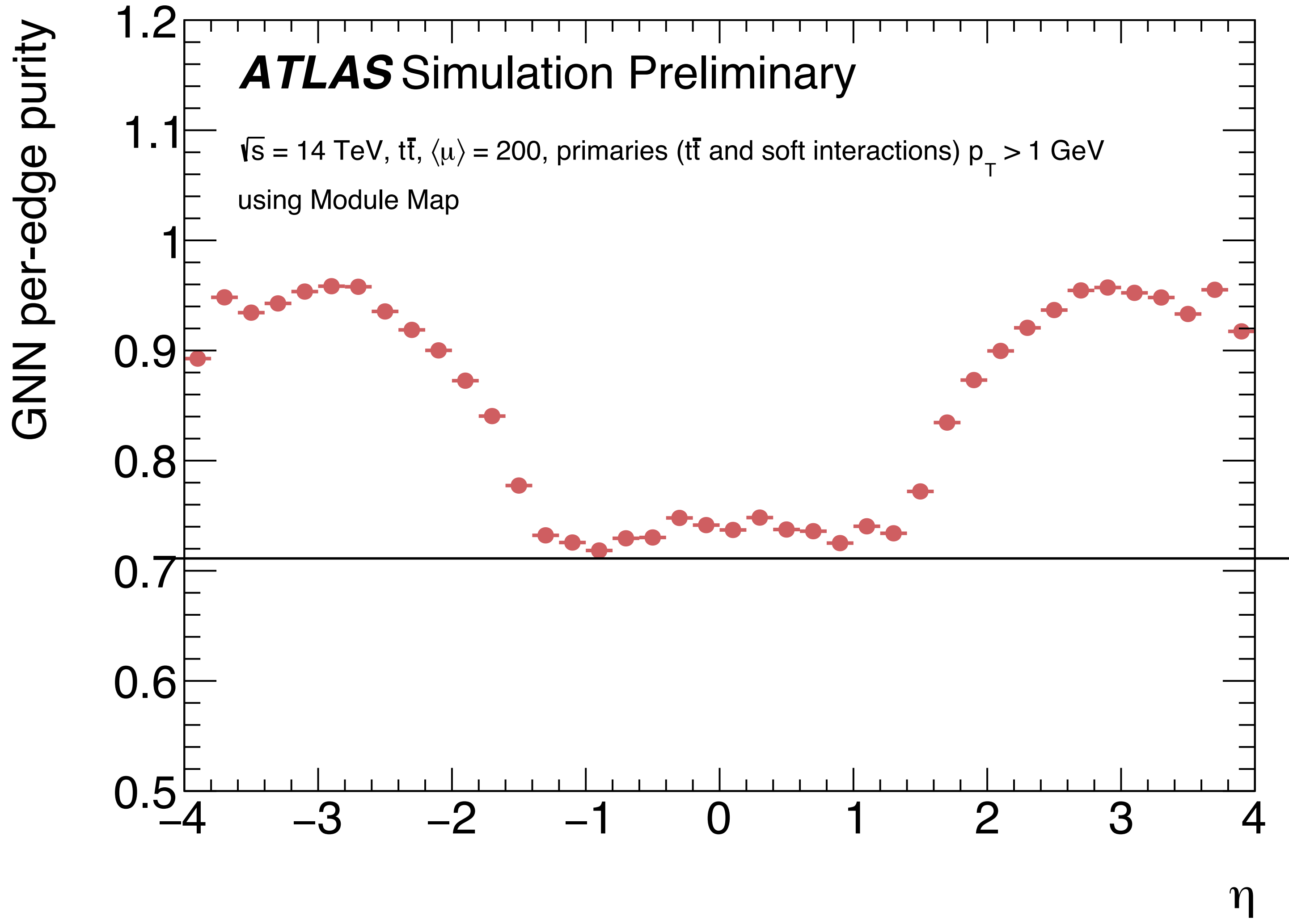
Move to 3 layers per MLP (vs 2 layers per MLP in the CTD2022 GNN model)

Add BatchNorm layers to make the training faster and more stable

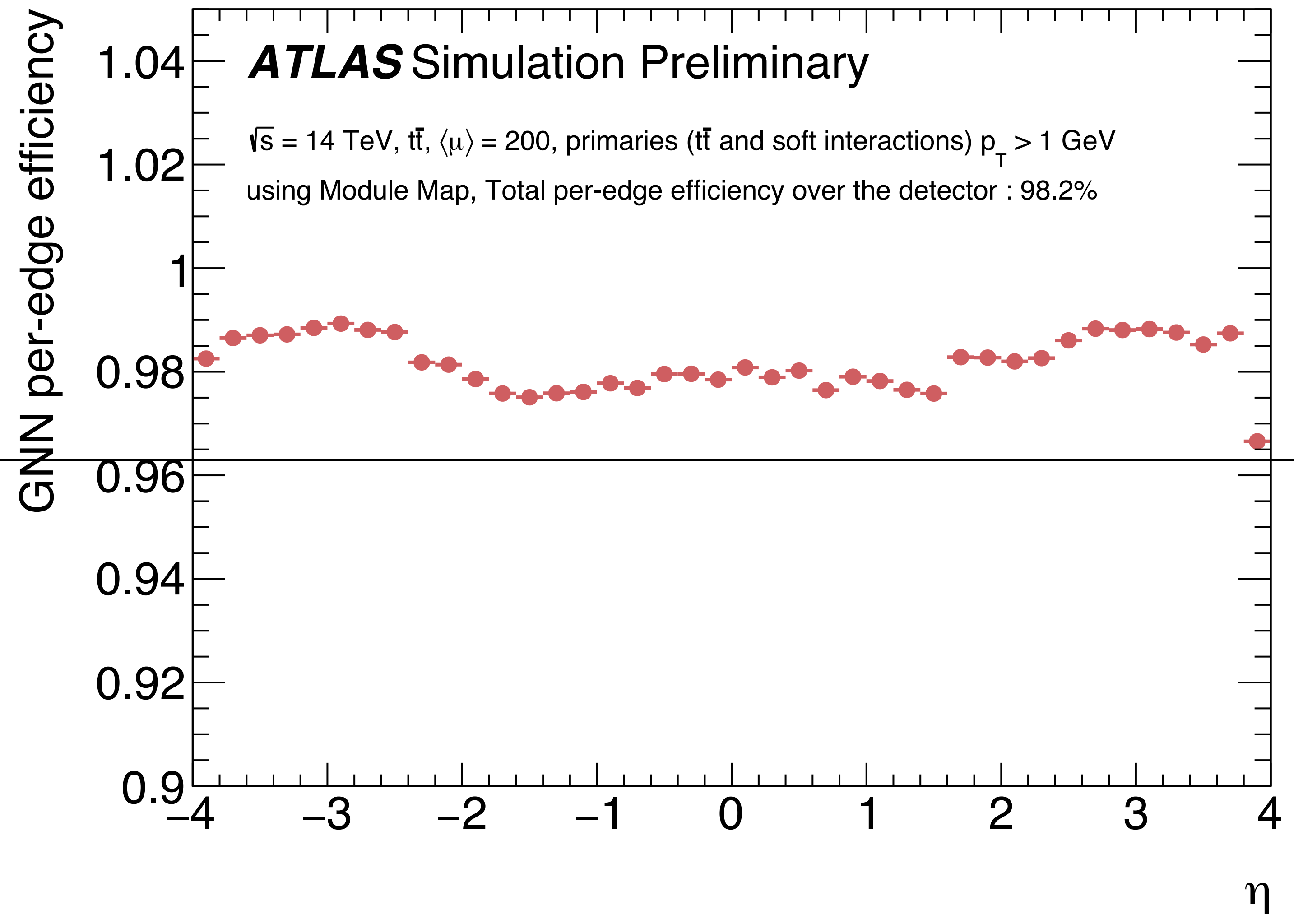
Efficiency and purity vs η

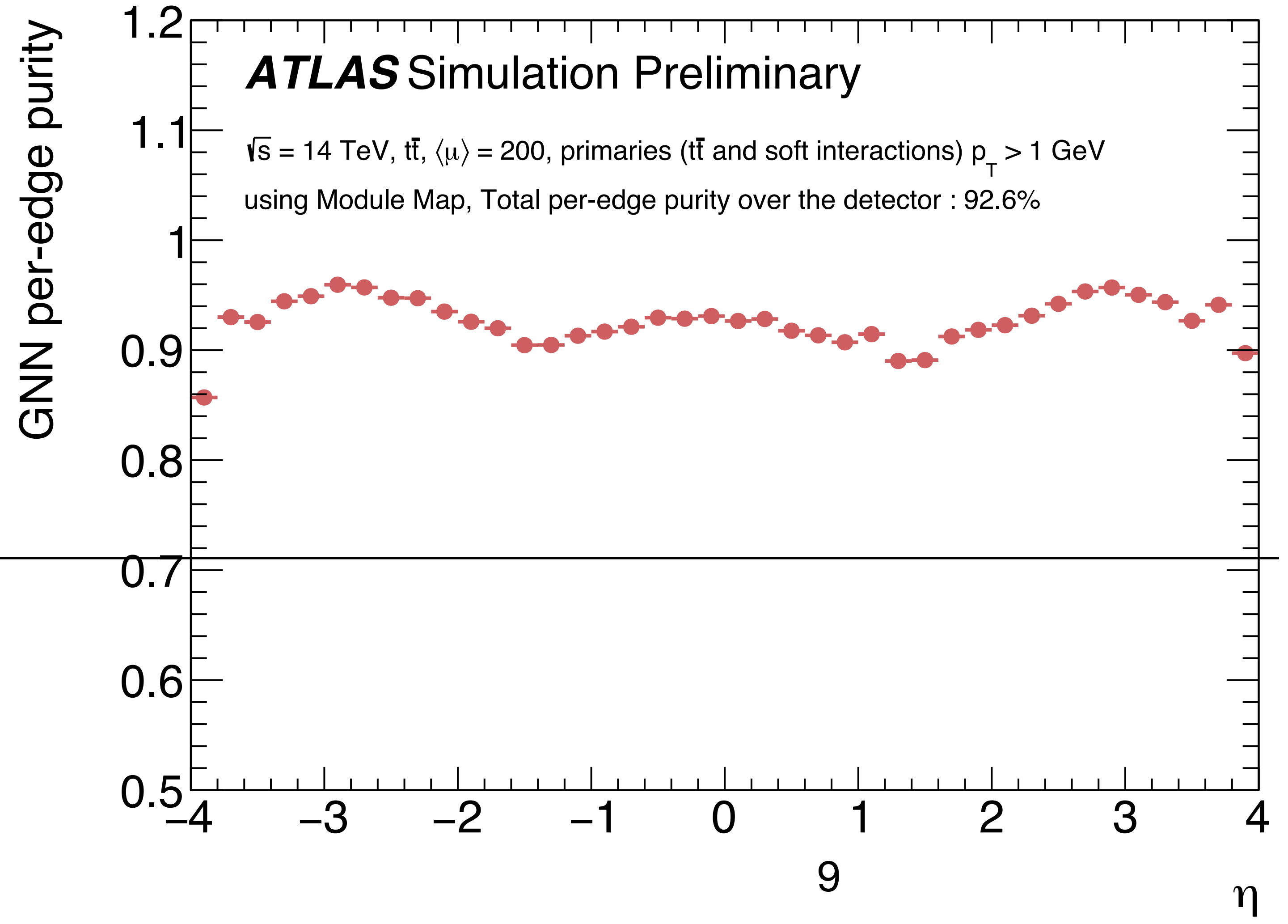
CTD2022





Heterogeneous Data + Extended GNN





$$efficiency = \frac{n_{true>treshold}}{n_{true}}$$

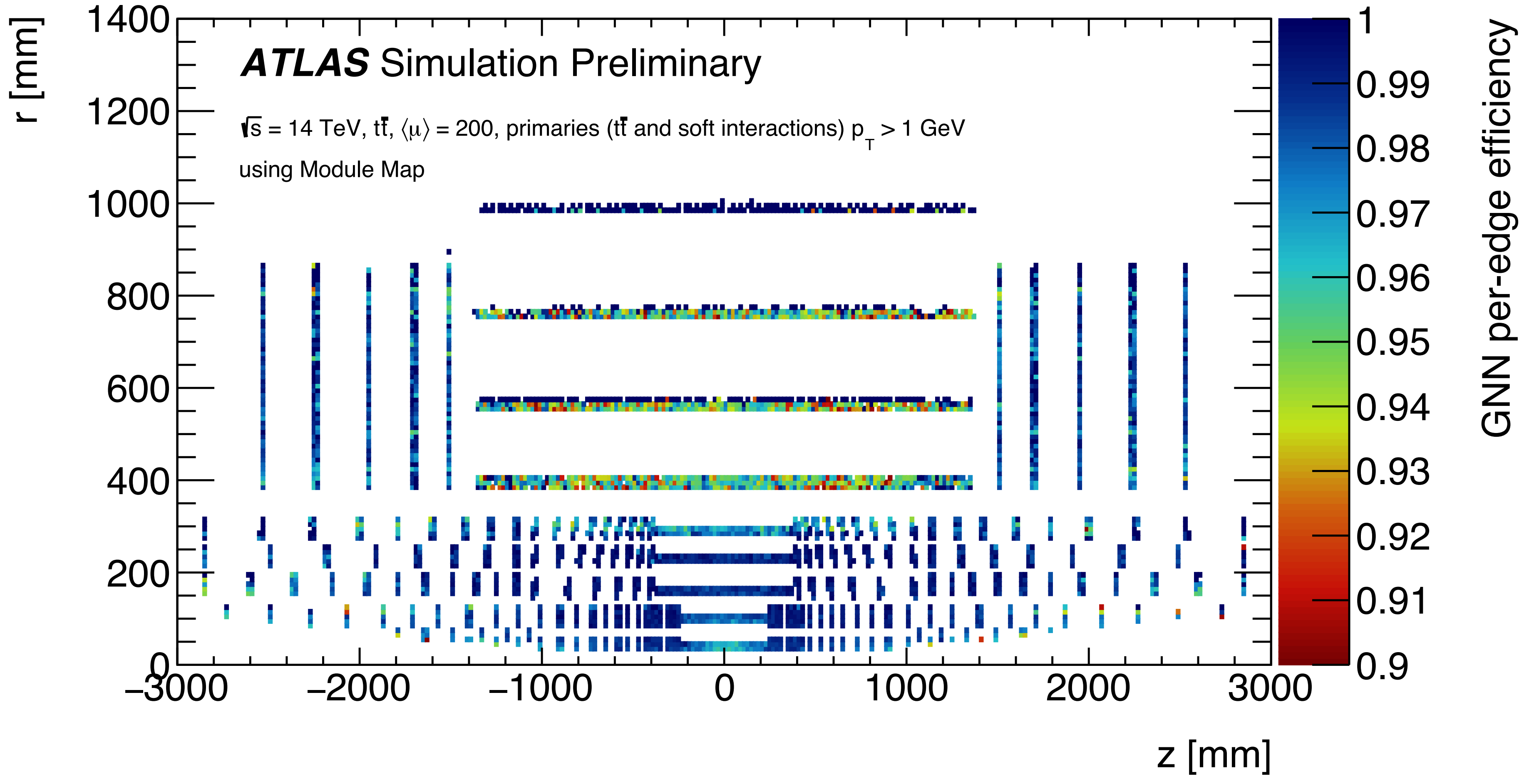
Similar efficiency for the same threshold (0.5)

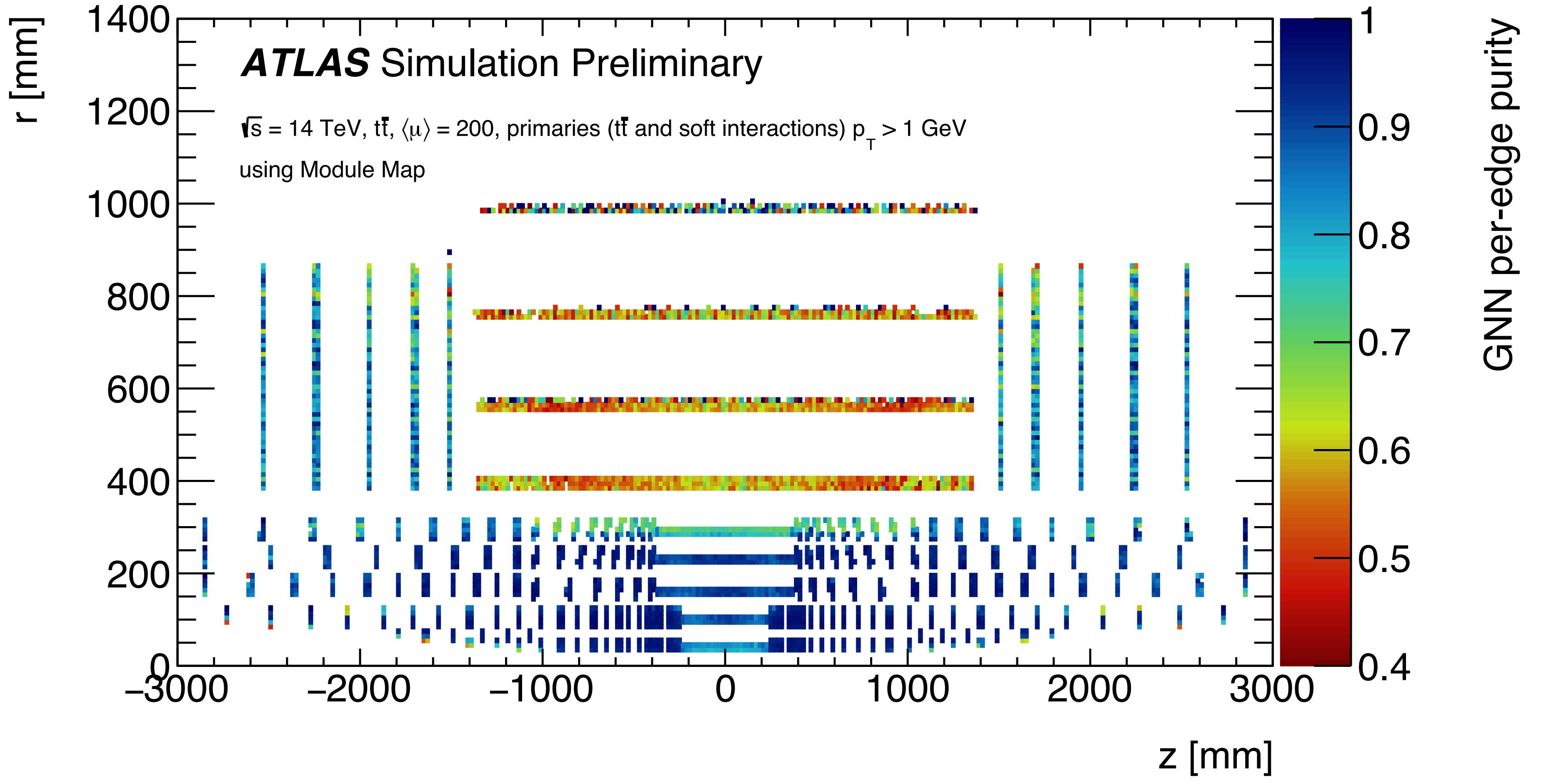
$$purity = \frac{n_{true>treshold}}{n_{true>treshold} + n_{fake>treshold}}$$

Significant improvement of purity in the CENTRAL region

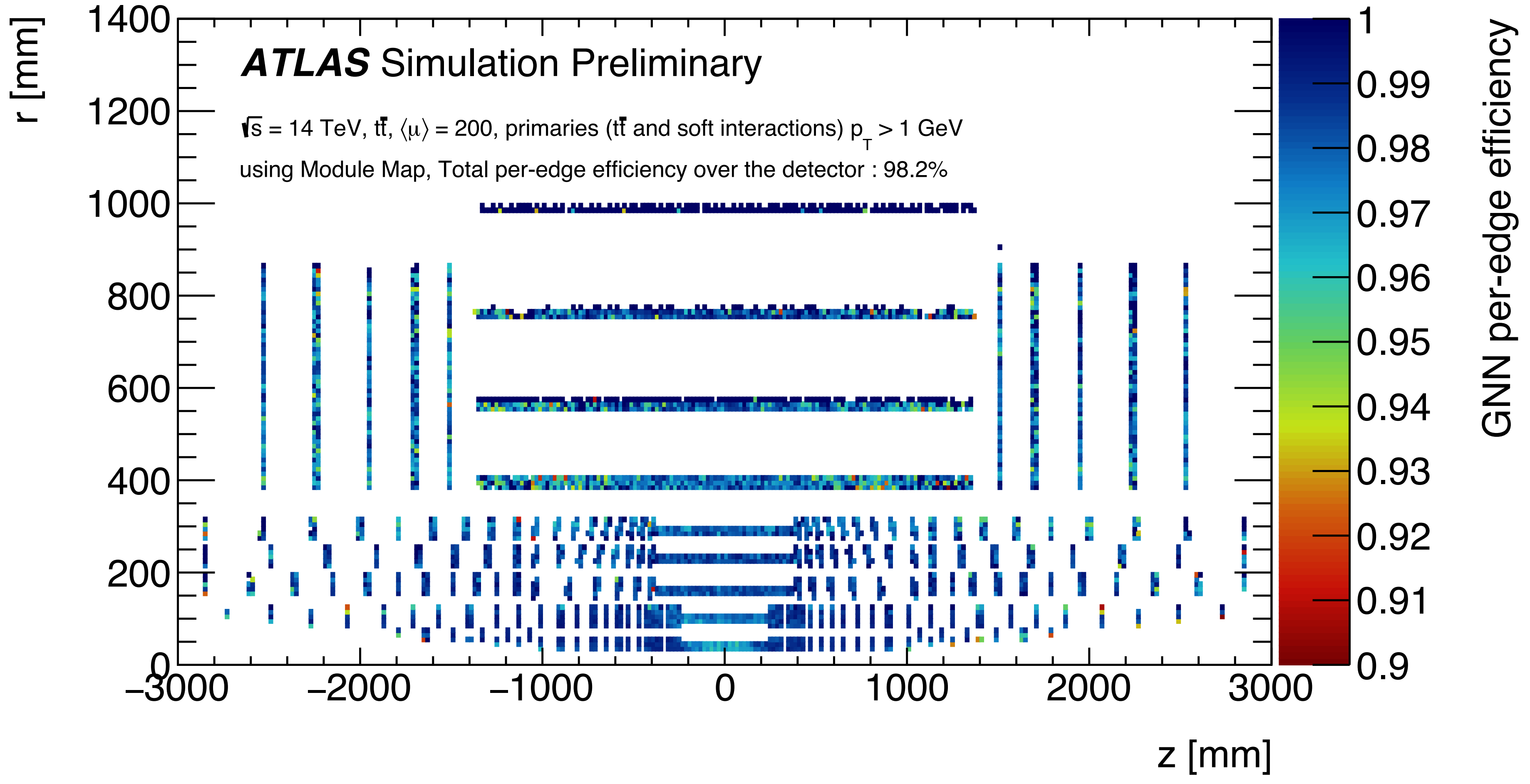
Efficiency and purity vs (r, z)

CTD2022



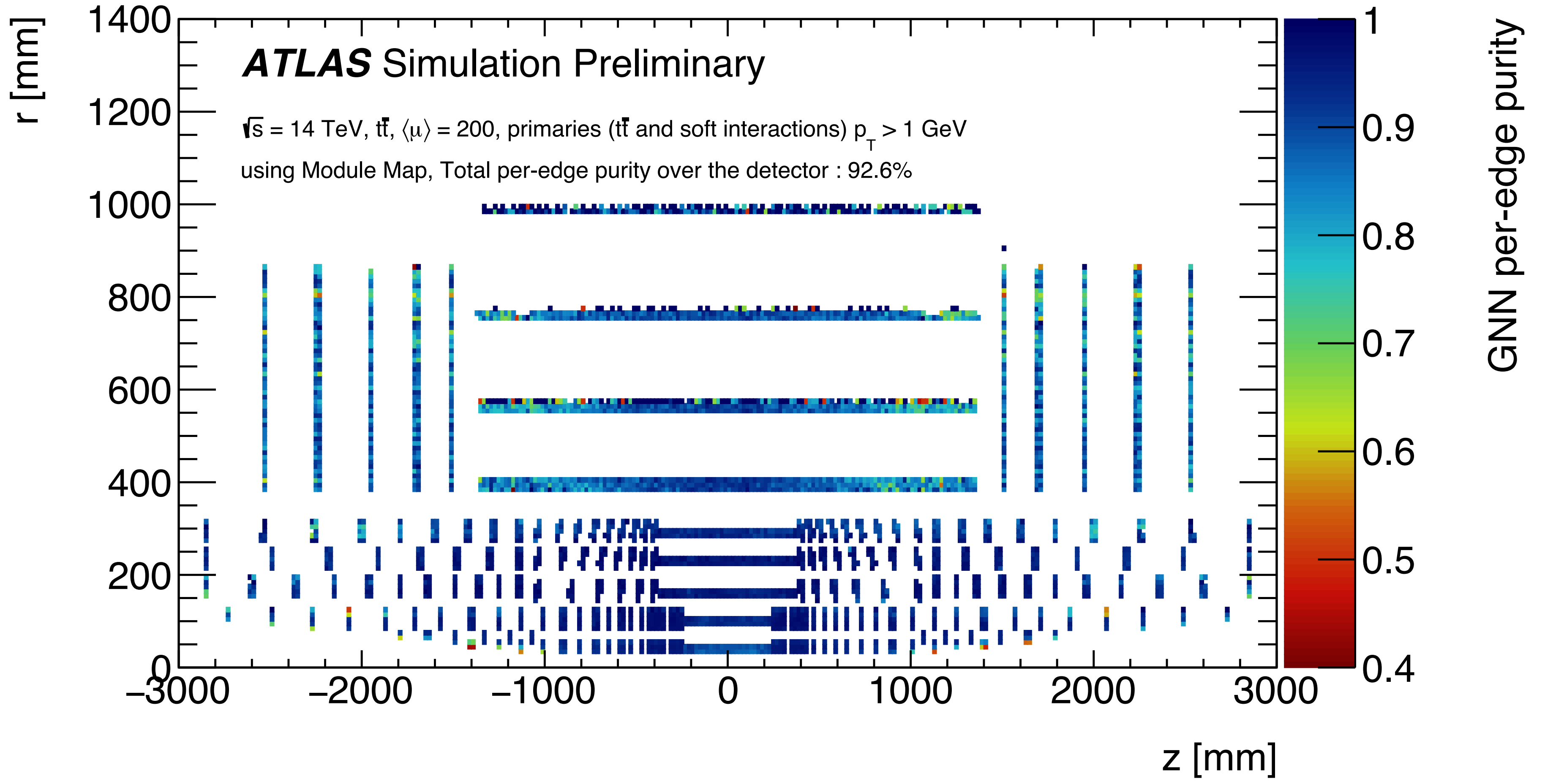


Heterogeneous Data + Extended GNN



$$efficiency = \frac{n_{true>treshold}}{n_{true}}$$

- Global efficiency of ~98%
- Efficiency more uniform in BARREL region



$$purity = \frac{n_{true>treshold}}{n_{true>treshold} + n_{fake>treshold}}$$

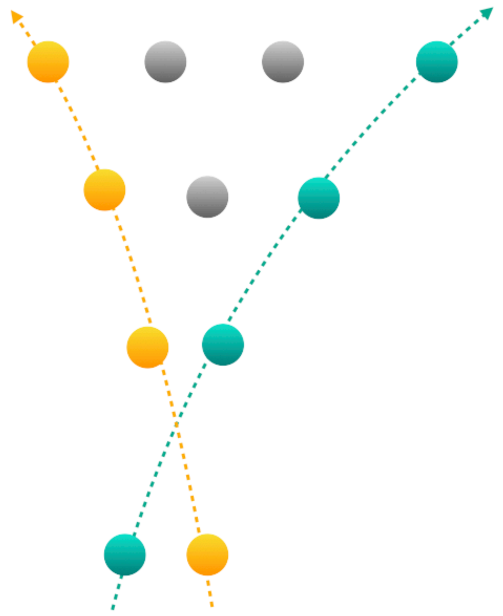
- **Significant improvement of purity in the STRIP BARREL region from ~40% to ~80%**
- Global purity of ~95%

Towards high performance track reconstruction algorithm

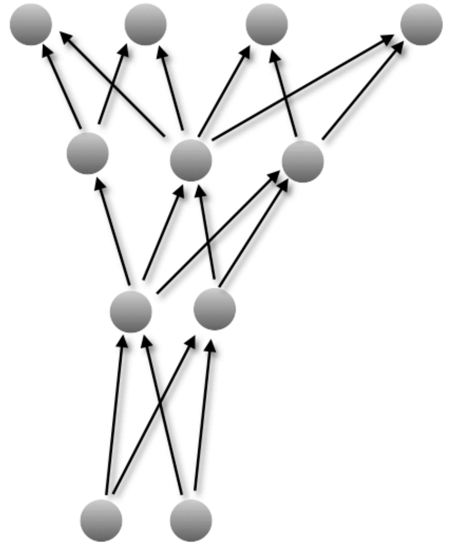
What is the link between GNN performance and track reconstruction ?

Graph construction

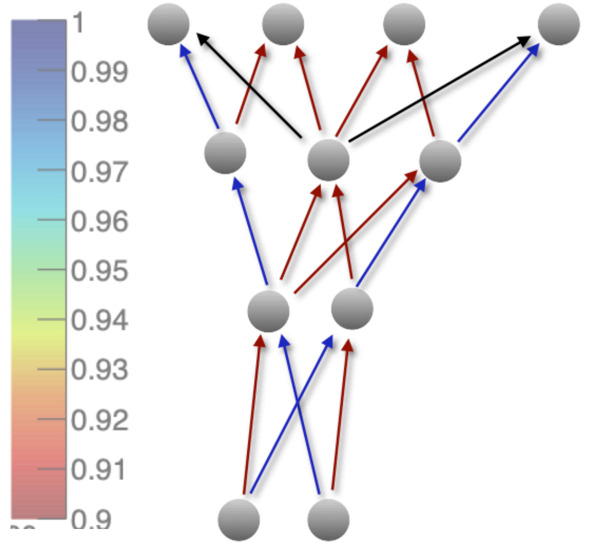
Raw Data



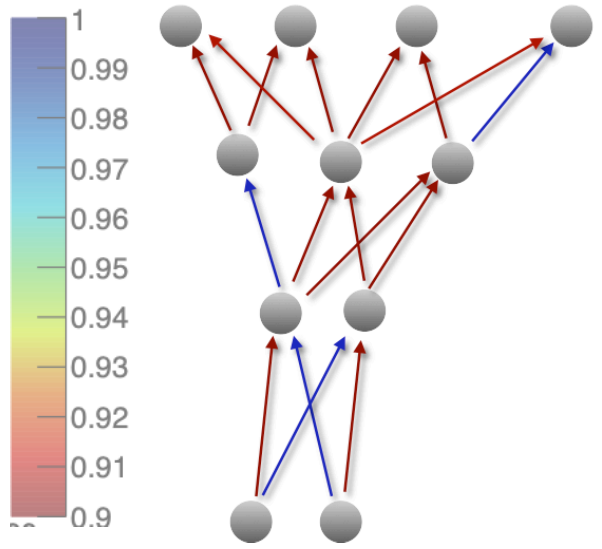
Graph data



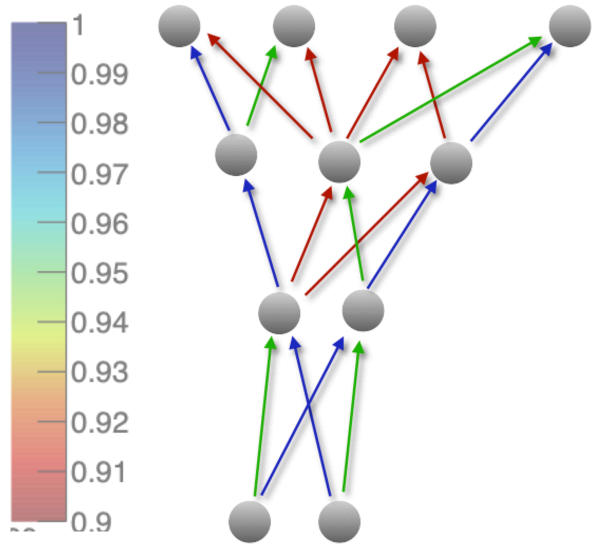
GNN edge classification



High GNN efficiency
High GNN purity

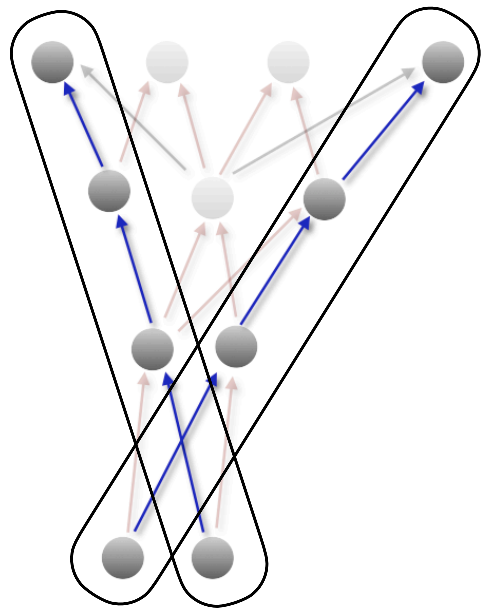


Low GNN efficiency
High GNN purity

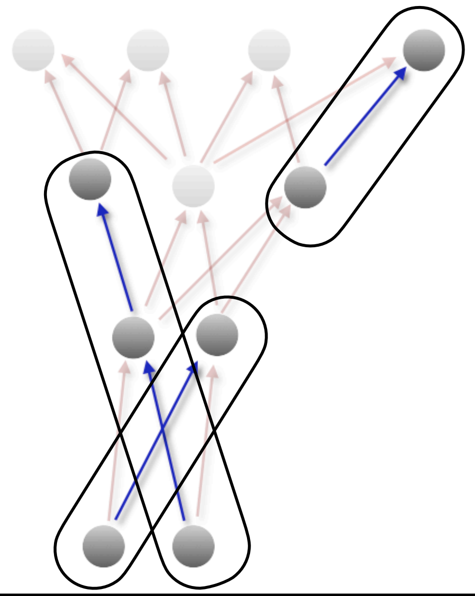


High GNN efficiency
Low GNN purity

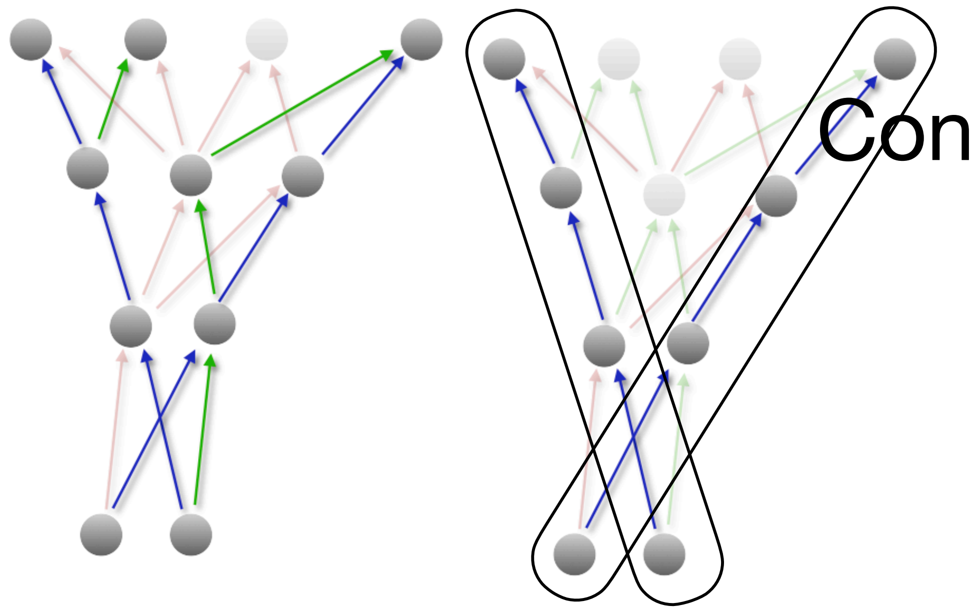
Track reconstruction



Connected Components (**CPU** & **GPU**)
Easy reconstruction



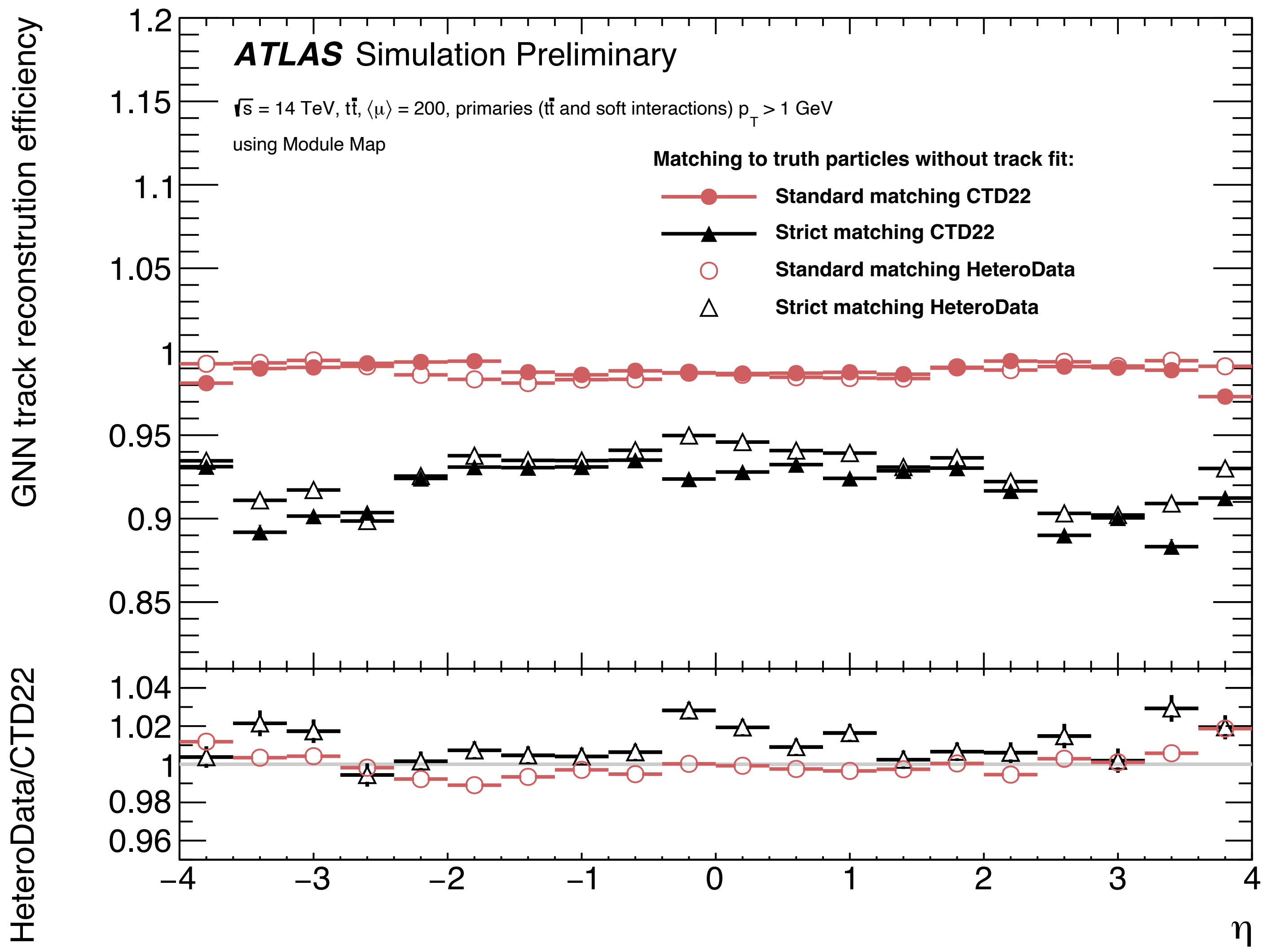
Connected Components (**CPU** & **GPU**)
Holes in tracks

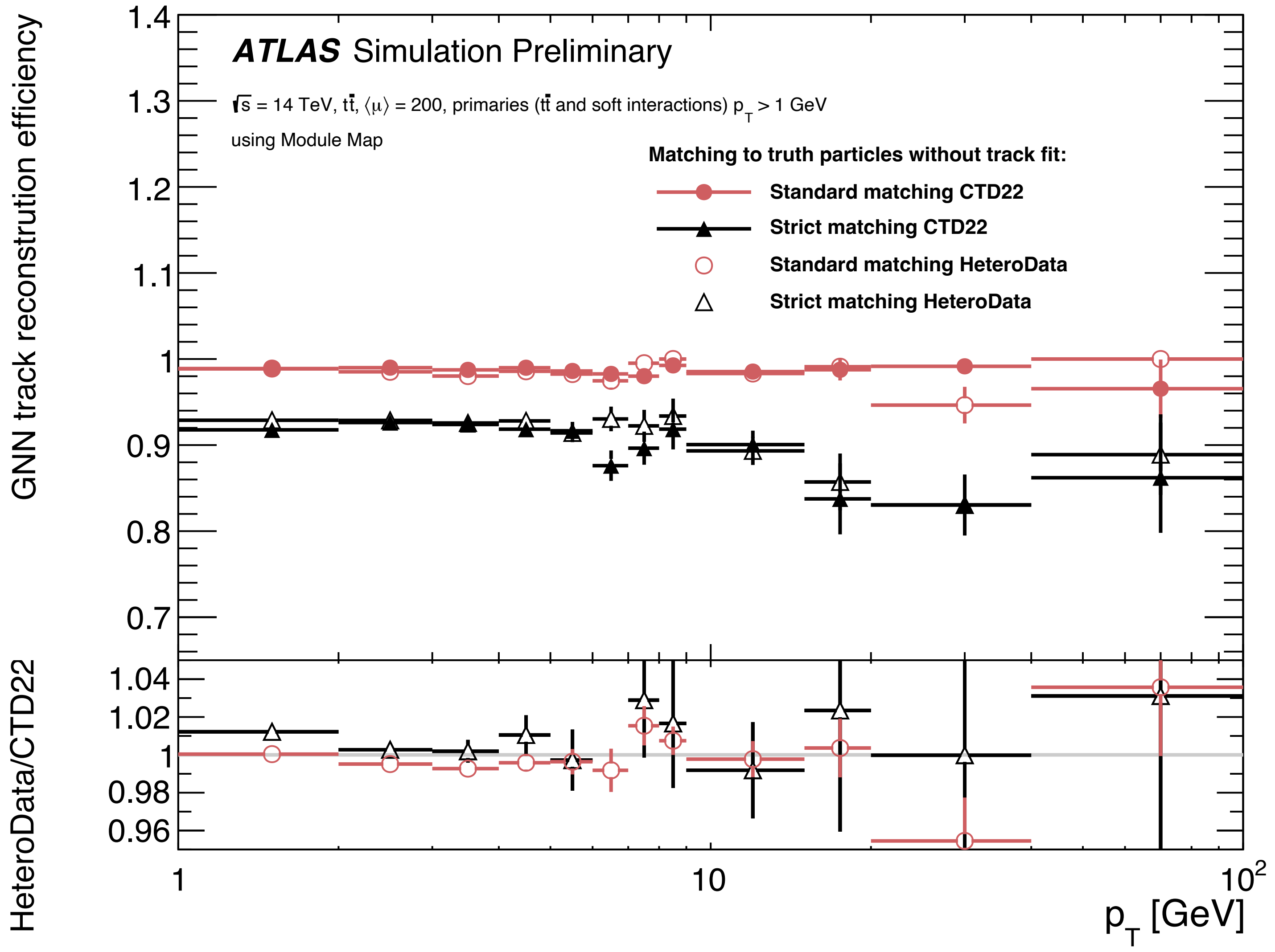


Connected Components (**CPU** & **GPU**)
+
Walkthrough algorithm (**CPU**)
Hard reconstruction

The higher GNN efficiency and purity are, the easier and the faster track reconstruction is !

Towards high performance track reconstruction algorithm

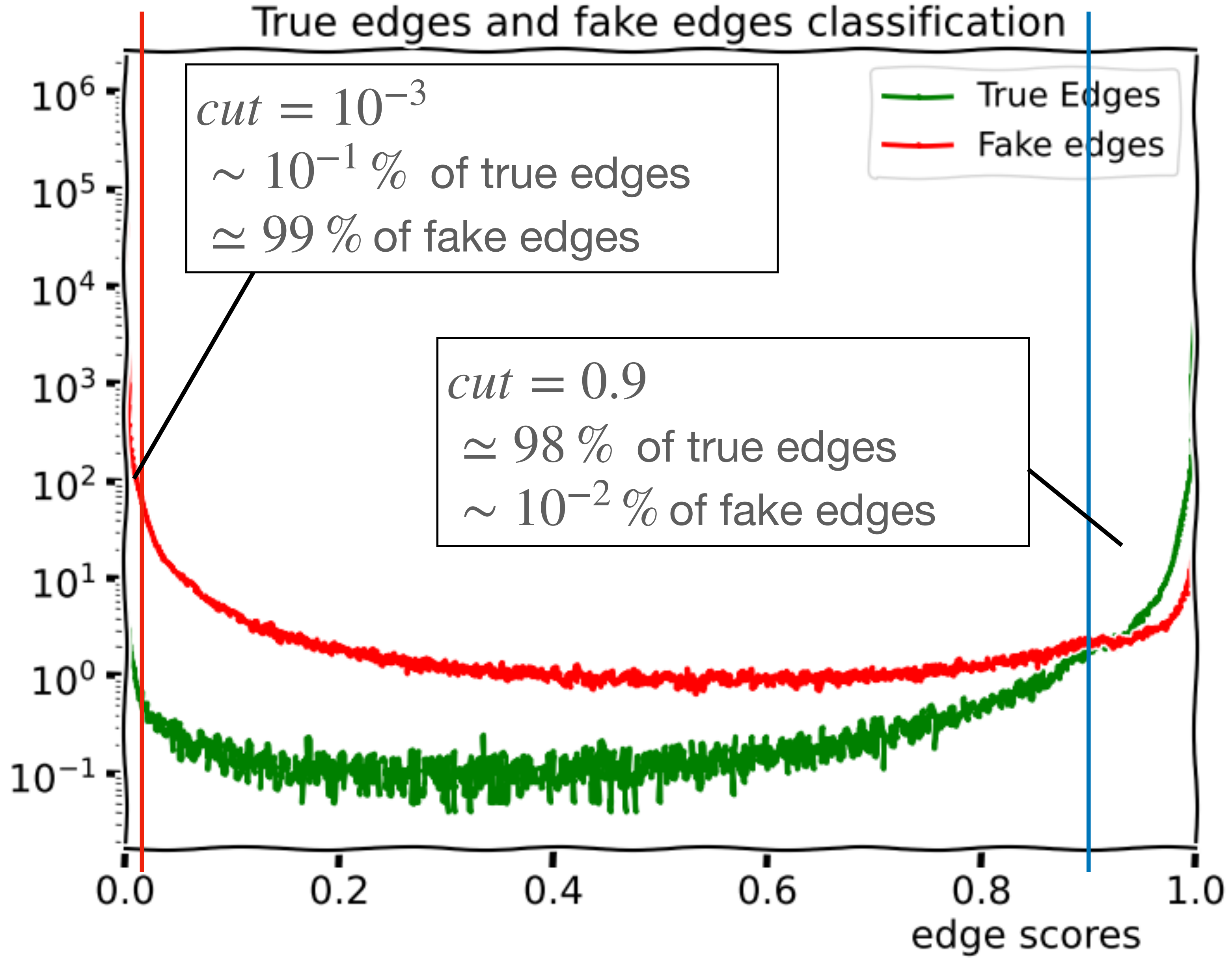




For CTD 2022 even with the poor GNN purity in the STRIP BARREL it was possible to get **excellent track reconstruction performance** at the cost of the computation time of the walkthrough algorithm

With the new high GNN efficiency and purity it is now possible to get 80% of perfect tracks and 95% of standard matching tracks with a simple Connected Components (very important as Connected Components algorithm **can be easily accelerated on GPU**). Walkthrough used only for small subset of tracks.

- **Low cut** to remove the majority of fake edges
- **High cut** to get **very high purity** and get « easy » tracks with Connected Components

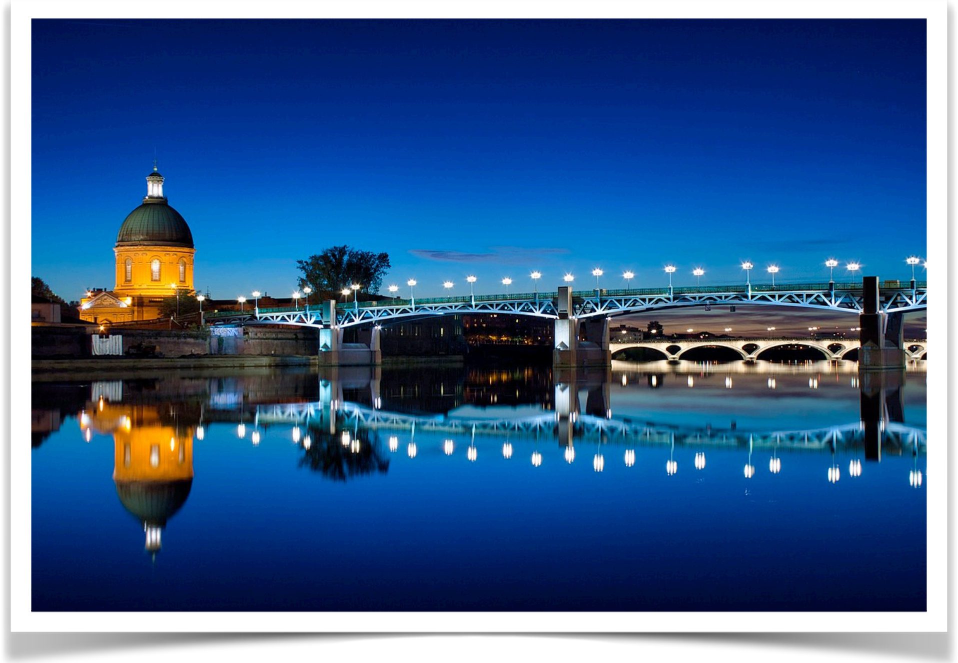


Summary and further steps

- Exa.TrkX and L2IT R&D collaboration to understand and handle heterogenous data has led to high performance GNN models in ATLAS ITk simulated data
- The high level of GNN performance leads to a very high performance track reconstruction full algorithm in terms of **track efficiency and purity** and **computation time**
- Further steps:
  - **GNN R&D** (Heterogeneous GNN model, GNN filter, ambiguity resolution with GNN transformer)
  - Pursue GNN pipeline **software integration** in ACTS & Athena
  - **Optimization** and **acceleration** of graph construction and track reconstruction on CPU and GPU

We developed a CommonFramework for GNN tracking R&D:

https://github.com/GNN4ITkTeam/CommonFramework



See you @CTD 2023 (Oct 10–13, 2023) in Toulouse, France !

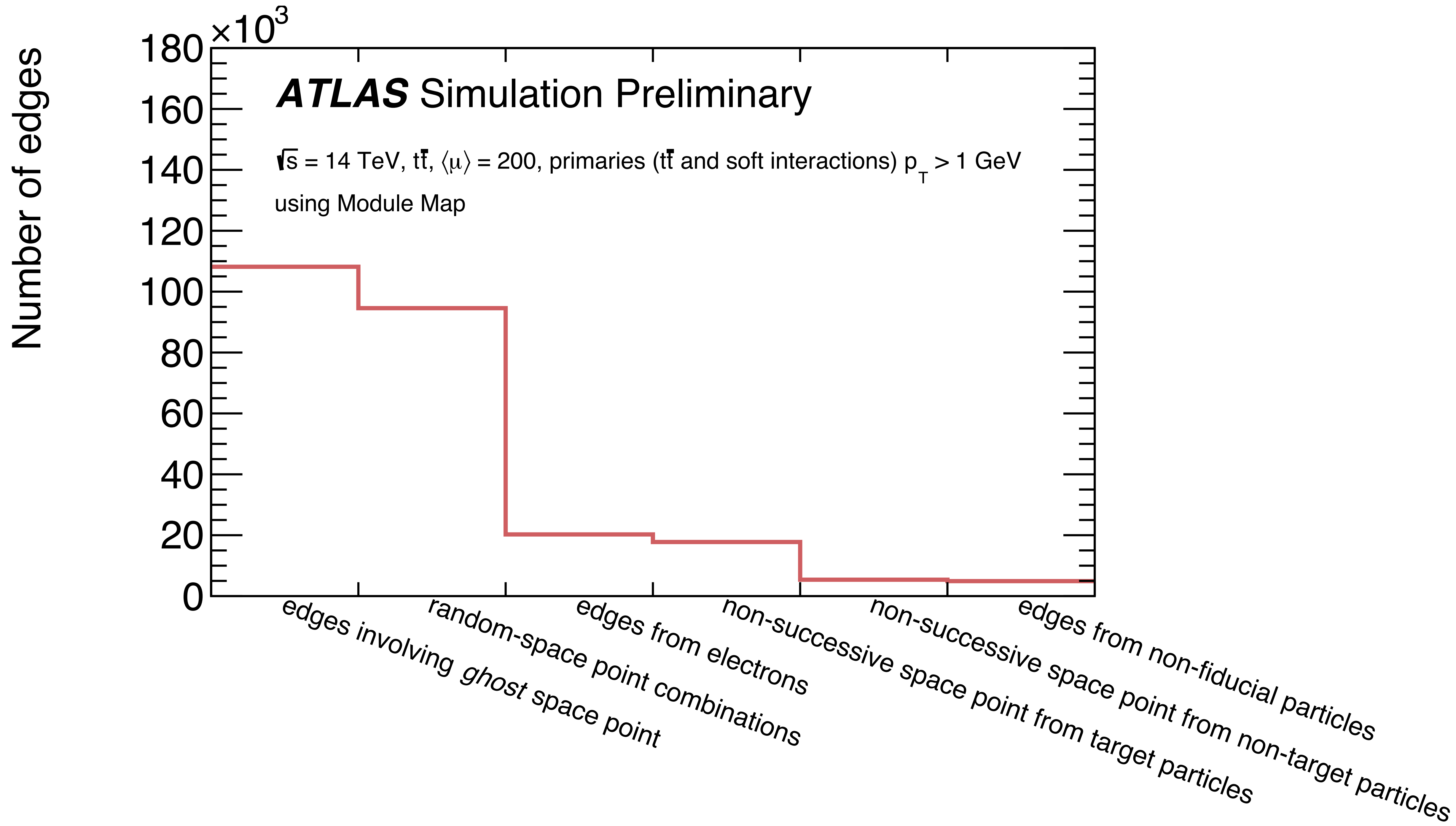
Thank you for your attention

Any question ?

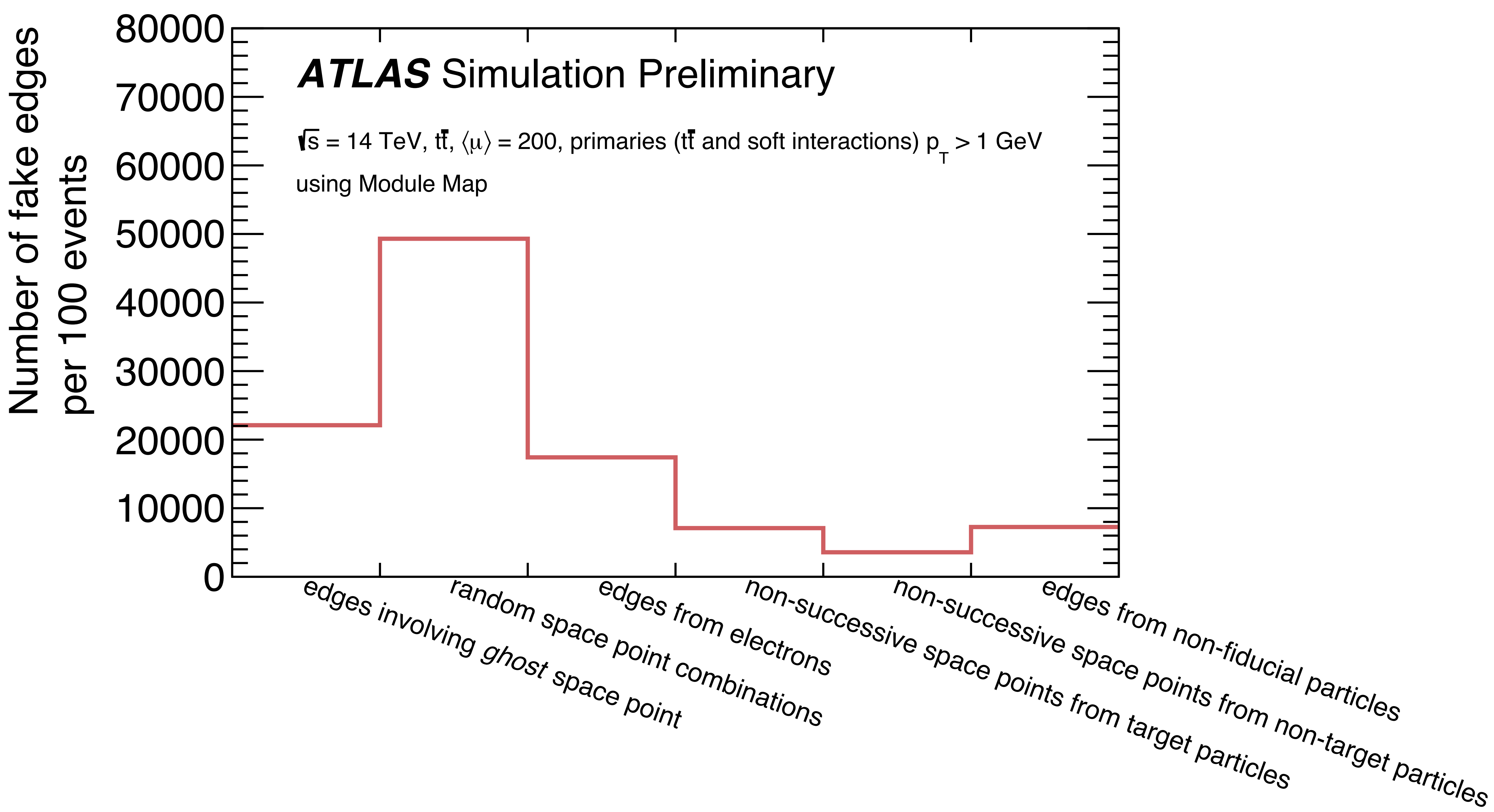
Backup

Origin of fake edges CTD 2022 vs new results

CTD2022

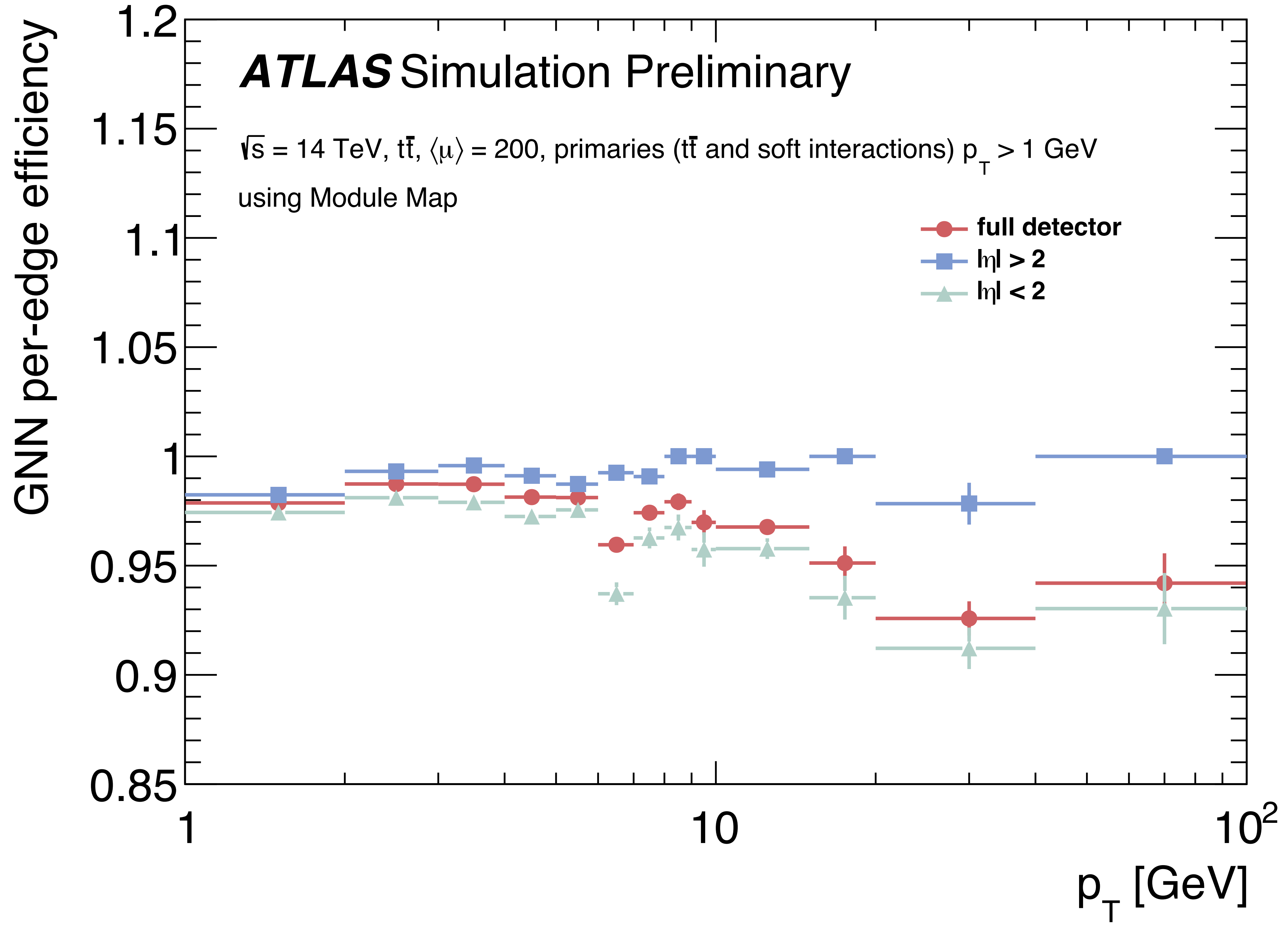


Heterogeneous Data + Extended GNN



GNN efficiency vs pT

CTD2022



Heterogeneous Data + Extended GNN