ATLAS PUB Note
ATL-PHYS-PUB-2025-045
3rd November 2025

Improving Tracking in Dense Environments with Transformers

The ATLAS Collaboration

This work presents a novel application of machine learning to the pattern-recognition stage of charged-particle reconstruction, enabling learned hit-to-track association within dense environments, such as the cores of high-p_{T} jets. Our Transformer-based architecture is based on a MaskFormer model that jointly optimises hit assignments and the estimation of the charged particles' properties. Trained and evaluated in dense environments the model delivers up to a 30% improvement in track-reconstruction efficiency over the standard ATLAS reconstruction when local particle density makes conventional reconstruction most challenging.

1 Introduction

Charged-particle reconstruction [1] is a core component of event reconstruction at the ATLAS experiment [2] and is closely linked to the experiment's overall physics reach. A characteristic feature of the energy-frontier programme is the presence of high-transverse-momentum (high-p_{T}) jets reaching up to several TeV. In dense environments, such as the cores of high-p_{T} jets, the close proximity of multiple charged particles leads to cluster merging and ambiguities in the assignment of clusters to tracks in the Inner Detector (ID) [3]. These effects reduce tracking efficiency, with losses increasing with jet energy. Although successive improvements to the seeding configuration [4], ambiguity solver, and dedicated pixel-classification networks [3, 5–7] have mitigated some of these challenges, performance recovery remains limited in dense environments.

In the current charged-particle reconstruction pipeline, a combinatorial Kalman filter (CKF) algorithm [8] builds candidate track hypotheses from any reasonable combination of clusters. These hypotheses are subsequently evaluated by an *ambiguity solver*, which removes redundant or inconsistent candidates to select the most likely set of tracks that correspond to the set of charged-particles in the event. Due to the sparsity (low occupancy) of the ID, the correct set of hypotheses can typically be identified by limiting how often each cluster is incorporated into a track. However, in dense environments, merged clusters may be correctly shared by several tracks, invalidating this assumption. The considerable set of conditional selection requirements used in the ambiguity solver are not sufficiently precise to resolve tracks for highly shared clusters. Therefore, within regions of interest (RoIs), defined as small angular regions around the momentum vector of hadronic calorimeter clusters [9], track hypotheses satisfying looser requirements can be accepted as reconstructed tracks but inefficiencies remain. One source derives from the limitation of the pixel-classification network which can determine if a shared cluster is compatible with being a merged cluster from one, two, or three or more particles. While that covers the majority of cases, in practice and especially at the highest energies reached by the Large Hardon Collider (LHC), up to 15 simulated particles can contribute to the same cluster. The reconstruction inefficiencies are exacerbated in the decay of high-energy b-hadrons as the decay vertex can occur at a non-negligible distance from the collision point leading to less separation between charged particles within the sensors.

While loosening ambiguity selections significantly improves efficiency, particularly within hadronic RoIs, there is a corresponding large increase in the fake rate and inefficiencies within the seeding stage remain [10]. This motivates the exploration of alternative methods, such as the machine-learning-based (ML-based) approach presented in this work that can reconstruct multiple tracks simultaneously.

ML-based reconstruction has been shown to provide substantial gains in similarly complex environments. For example, the CMS DeepJet architecture [11] employs deep learning to exploit high-dimensional correlations between tracks and clusters for jet flavour tagging, delivering significant improvements in dense jet cores. Adapting such approaches to ATLAS track reconstruction is non-trivial, as jets are not yet available during the tracking stage. Consequently, developing models that can operate on detector-level inputs directly, without reliance on reconstructed jets, offers a more practical path for ML-based tracking within the ATLAS reconstruction chain.

For the High Luminosity-LHC [12], the computational scaling of charged-particle reconstruction with the increase in the number of proton–proton (pp) collisions per bunch crossing motivates the exploration of new approaches. Graph neural networks (GNNs) have been studied in ATLAS [13, 14] and achieve high efficiency with low fake rates on large-scale datasets. However, these methods typically target scalability in high-multiplicity events and have not yet been optimised for local hit-level complexity in dense jet cores. They also rely on externally defined graph-construction procedures, such as module-map

or metric-learning approaches, and on separate graph-segmentation stages after inference to assemble tracks. These components are not optimised jointly with the network, limiting the degree of end-to-end reconstruction achievable.

Recent developments in ML, particularly in computer vision, have introduced architectures capable of jointly identifying and segmenting multiple overlapping objects within an image. The MaskFormer architecture [15, 16] is one such model, combining a Transformer-based encoder–decoder network [17] with object-level queries (learnable vectors that each represent a potential object) that learn to represent distinct physical instances. Reinterpreting the image-segmentation problem in terms of particle-track reconstruction, hits can be treated as unordered inputs and tracks as individual objects to be predicted. This concept was first demonstrated using the open-source TrackML dataset [18], achieving a good efficiency of approximately 97% with sub-percent fake rates and sub-100 ms inference times on a single GPU [19]. The model eliminates the need for explicit graph construction and naturally supports shared-hit assignment through its attention mechanism, offering a unified and scalable approach to charged-particle reconstruction. Furthermore, the architecture has wide utility having been adopted for vertex reconstruction within flavour tagging and particle-flow reconstruction [20, 21].

Building upon this foundation, the present study adapts the Transformer-based reconstruction model to the dense environments characteristic of high-p_{T} jets in ATLAS. The model is trained and evaluated on simulated Run 2 conditions, using hits associated with all track hypotheses within a hadronic RoI. Using only these hits as inputs, it jointly performs hit-to-track assignment, regression of track parameters, and estimation of local track–hit properties such as incidence angles and cluster positions. This study focuses on the cluster-to-track assignment within track-hypothesis generation, as high-precision track parameters will ultimately be obtained from the existing ATLAS global χ^2 fit.

This proof-of-concept represents an initial step toward specialised ML-based tracking within the ATLAS reconstruction software framework. The approach complements the existing reconstruction by providing an alternative solution optimised for dense topologies, potentially reducing reliance on separate pixel-splitting or ambiguity-resolution stages. Future work will focus on further optimisation, validation with variable detector conditions, and application to collision data.

The model is described in Section 2 and the datasets are detailed in Section 3. Results are presented in Section 4.

2 Model Overview

The model presented in this study adapts a Transformer-based architecture originally developed for charged-particle reconstruction on the TrackML dataset [19]. That model demonstrated that modern attention mechanisms can perform track finding and fitting simultaneously, achieving high efficiency and low fake rates while avoiding the need for graph construction or combinatorial seeding. The central idea is to treat track reconstruction as an instance segmentation task. Each track is represented as an independent object predicted by the network, and each hit is assigned a probability of belonging to one or more such objects. This formulation naturally accommodates shared clusters and overlapping trajectories. In this adaptation, the model is trained to reconstruct all charged particles within a hadronic RoI, defined by the presence of a topological calorimeter cluster [9] with total transverse energy exceeding 150 GeV. A single

jet may contain multiple such RoIs. Only charged particles within $|\Delta\phi| < 0.05$ and $|\Delta\eta| < 0.05$ of an RoI are considered[1].

Two components of the model from Ref. [19] which targets reconstructing all $O(10^4)$ charged particles in an HL-LHC-like environment are not utilised in this work, where the model targets only charged particles in hadronic ROIs. Typical RoIs contain on average $O(10)$ particles and $O(100)$ clusters. Therefore, the dedicated hit-filtering stage meant to reduce the number of clusters considered for tracking is unnecessary here. In addition, while the regression output for track parameters is retained for completeness, these parameters are not required for the intended integration within the ATLAS reconstruction, where the global χ^2 fit will provide high-precision track parameters once the model is implemented in the ambiguity solver.

Moreover, the input representation is extended to include the pixel charge matrix (described below) in place of the cluster-level summary statistics, such as cluster length, used previously. Such detailed pixel-level information is available only in high-fidelity detector simulations. The TrackML dataset provides an approximate description of a high-energy physics detector and does not include the full charge-sharing structure of individual clusters.

2.1 Input Representation

For each RoI, all clusters from the high-granularity silicon pixel detector [22] and the silicon microstrip SemiConductor Tracker (SCT) [23] associated with reconstructed track hypotheses that enter the ambiguity solver are used as inputs to the model. These tracks are known as Silicon Space-point Seeded (*SiSp*) tracks. Each cluster is represented by a set of geometric and detector-level quantities:

- the global cylindrical coordinates (r, ϕ, z) and the local position on the module surface;
- the global cylindrical coordinates and orientation of the module;
- the angular separation in η and ϕ relative to the RoI axis, defined by a line from the median z position of the SiSp tracks within the RoI to the centre of the calorimeter topocluster;
- the detector layer, and whether it is in the barrel or endcap region;
- for pixel clusters, the charge collected on 7×7 matrix of pixels centered on the cluster barycentre and the corresponding pitch size in the longitudinal direction[2].

The charge matrix is the same format as used in the current pixel-classification network and density networks that predict up to three charged particle-sensor crossing points per cluster [3, 7]. Units for all input quantities are chosen such that their numerical values are of order one.

[1] ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the z-axis along the beam pipe. The x-axis points from the IP to the centre of the LHC ring, and the y-axis points upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z-axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$ and approximates the rapidity $y = \frac{1}{2}\ln\left(\frac{E+p_z c}{E-p_z c}\right)$ in the relativistic limit. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta y)^2 + (\Delta\phi)^2}$.

[2] In the some regions of the pixel modules, neighbouring pixels are connected through a shared readout channel, forming elongated "ganged" pixels. These pixels are typically 500 µm long in z (compared with 300 µm elsewhere), introducing a small ambiguity in the hit position and modifying the observed charge pattern [24].

2.2 Architecture

The model consists of two components: a *hit encoder* and a *track decoder*. The encoder processes a set of hits within a RoI, converting their spatial and detector-level features into learned feature embeddings. Positional features are encoded using Fourier features [25] to handle high-frequency spatial information. It employs a self-attention mechanism to allow each hit to aggregate information from all other hits, enabling the model to learn complex spatial correlations and relationships between hits.

The decoder follows the MaskFormer design [15, 16], consisting of a stack of Transformer layers operating on a fixed number of learnable object queries. Each query corresponds to a potential reconstructed track. The number of queries sets an upper limit on the number of tracks that can be reconstructed in a given RoI. We use 64 queries per RoI, which is sufficient for the vast majority of cases. Through cross-attention, the queries aggregate information from relevant hits in the encoded representation, while self-attention between queries allows global coordination between overlapping tracks. The decoder outputs three quantities per query:

1. a continuous mask over the input hits, representing per-hit probabilities for track membership;
2. a categorical score indicating whether the query corresponds to a valid track or to the NULL class; and
3. a regression vector containing the estimated track parameters.

During inference, the predicted masks over the input hits are thresholded at 0.5 to obtain binary hit assignments, and reconstructed tracks are accepted if their probability of belonging to the physical (non-NULL) class is greater than 0.75. The explicit representation of tracks, rather than pairwise hit relationships, naturally handles shared clusters. Unlike traditional algorithms, this procedure unifies track finding and fitting into a single optimisation step.

2.3 Training Objective

The target outputs for each RoI are sets of clusters corresponding to the same particle, identified using Monte Carlo truth information. Targets are restricted to charged truth particles originating from the pp collision with transverse momentum $p_{\mathrm{T}} > 1$ GeV and at least eight silicon clusters.

The model is trained in a supervised fashion using simulated truth-labelled hits within each hadronic RoI. A multi-task loss function combines three terms: binary cross-entropy (BCE) [26] losses for the mask prediction and track classification, and a smooth-L1 [27] loss for the regression of track parameters. The total loss is defined using an optimal bipartite matching [28] between predicted and target tracks to ensure permutation invariance in the ordering of object queries. In practice, this matching assigns each predicted track to the most compatible truth particle by minimising a measure of their difference based on mask similarity. This formulation allows the network to learn both the number and properties of tracks present in each RoI without any explicit seeding, edge construction, or post-processing.

2.4 Training Configuration

For this proof-of-concept study, the model is implemented in PyTorch and trained on simulated Run 2 ATLAS events as described in Section 3. Training and inference are performed on a single NVIDIA A100 GPU. The dataset consists of approximately 10 million ROIs, and is divided into 80% for training, 10% for validation, and 10% for testing. Although the model operates on individual RoIs, the dataset is split at the event level to prevent any information leakage between training and evaluation samples. Training is performed for 10 epochs using the AdamW [29] optimiser with a learning rate scheduler, using a maximum learning rate of 10^{-4} and a batch size of 100 RoIs. Loss terms for the mask, classification, and regression outputs are monitored separately to verify convergence. Model checkpoints are selected according to the minimum validation loss.

3 Dataset

The model is trained and evaluated on simulated pp collision events at $\sqrt{s} = 13$ TeV, corresponding to Run 2 detector and beam conditions. Events are produced using the full ATLAS detector simulation [2, 30] based on Geant4 [31].

The dataset is enriched with high-p_{T} jets from a Z' boson with a mass of 4 TeV, decaying with roughly equal probabilities into b-, c-, and light-quark jets generated using Pythia 8.243 [32] with the A14 [33] tune for the underlying event and the leading-order NNPDF2.3lo [34] parton distribution function set. A broad jet-p_{T} spectrum, approximately flat between 250 GeV and 1.5 TeV and extending to 3 TeV, is achieved by applying a weighting factor that broadens the natural Z' resonance width. The decays to $b\bar{b}$, $c\bar{c}$, and light-flavour quark pairs are set to equal branching fractions. Bottom- and charm-hadron decays are modelled using EvtGen 1.7.0 [35]. Additional pp interactions taking place simultaneously in the proton bunch crossing (pile-up) are not included in the simulated sample. Within dense environments, their contribution to merged clusters is minor [36].

Hadronic RoIs are defined by the presence of topological calorimeter clusters with a total transverse energy exceeding 150 GeV. On average, each event contains five RoIs. Typically, each jet contains a single RoI at its core if any, however, multiple RoIs can also originate from the same jet.

4 Results

4.1 Tracking Efficiency and Purity

The performance of the Transformer-based model is evaluated by comparing its reconstructed tracks, labelled as *MaskFormer* tracks, to those produced by the standard ATLAS reconstruction [1, 3], hereafter referred to as *Baseline* tracks. For reference, the set of track hypotheses that serve as input to the ambiguity solver, *SiSp* tracks, are also included. The reference for efficiency calculations are all primary truth particles that have contributed to at least eight ID clusters and satisfy $p_{\mathrm{T}} > 500$ MeV.

To provide a one-to-one correspondence between tracks and truth particles, a unique pairing is established within each RoI by maximising the intersection-over-union (IoU) of their associated clusters. IoU is defined as the ratio of the number of clusters common to both to the total number of unique clusters associated with

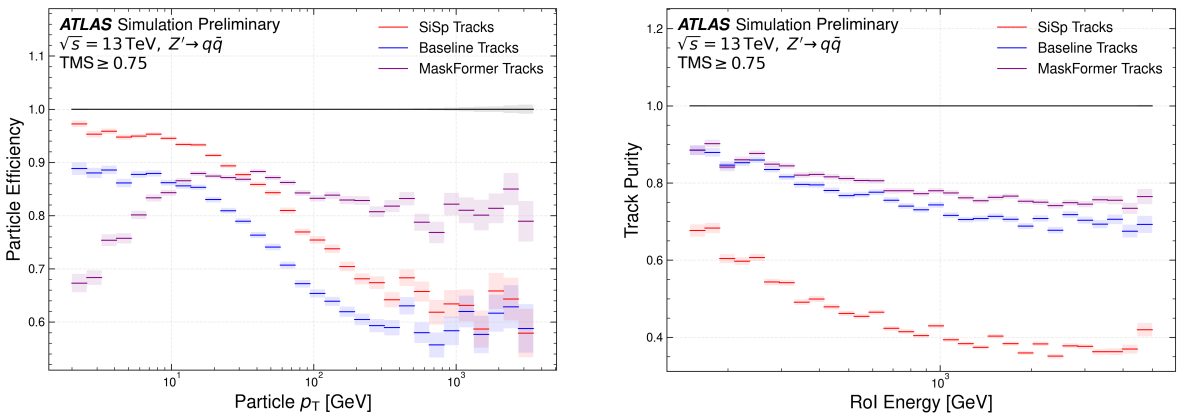
either. Track-to-particle reconstruction criteria follow the standard ATLAS procedure [37], which is based on the *truth-match score* (TMS). The TMS quantifies the weighted[3] fraction of clusters on a reconstructed track that originate from a given truth particle. A reconstructed track and truth particle are considered a *match* if they have a TMS ≥ 0.75. Based on this, a particle is considered as *efficient* if it is paired to a track that is also a match. A track is labelled as:

- *fake* if it is not matched to any truth particle,
- *duplicate* if it is matched to a truth particle but not selected in the one-to-one pairing, and
- *pure* if it is both matched and paired with the same truth particle.

Because the TMS threshold is intentionally strict, many tracks labelled as fake correspond to real charged particles that have been misreconstructed.

Figure 1 shows the model achieves comparable overall efficiency to the standard reconstruction and a large improvement, up to 30%, for high-p_{T} tracks in dense jet cores. The model outperforms the SiSp collection at high p_{T}, overcoming limitations in the track-seeding stage. At low p_{T}, performance is reduced due to the limited training statistics in that regime. The model is primarily designed to recover tracks with a large number of shared clusters, while the standard reconstruction remains optimal for isolated or low-p_{T} tracks. The reduced performance at low p_{T} is therefore expected and reflects the complementary role of the two reconstruction approaches.

Since the p_{T} of a fake track is not well-defined, the purity is shown as a function of the energy of the RoI, which is strongly correlated with the particle multiplicity and density. The purity of the learned reconstruction is slightly better than the standard reconstruction across the full energy range, indicating that the combined rate of fake and duplicate tracks is lower.

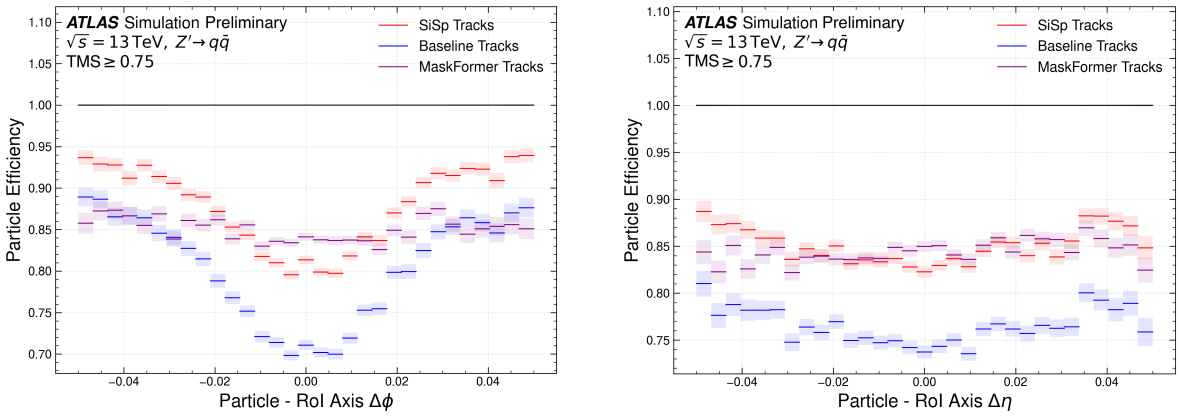


(a) (b)

Figure 1: (a) Reconstruction efficiency of charged particles in a hadronic RoI as a function of the transverse momentum of the corresponding truth particle. (b) Reconstructed track purity as a function of the hadronic RoI energy. The Transformer-based model (purple) is compared with the standard ATLAS reconstruction (blue) and the set of tracks entering the ambiguity solver (red). The reference for efficiency calculation is all primary truth particles that have contributed to at least eight ID clusters and satisfy $p_{\mathrm{T}} > 500$ MeV.

[3] Pixel to SCT cluster weights are 2 to 1.

Figure 2 shows efficiency versus the azimuthal and polar angular separation to the RoI axis. In the ID barrel, pixels have a smaller pitch in the ϕ direction (50 µm) compared to the η direction (250 µm or 400 µm depending on the layer). Along $\Delta\phi$, a clear decrease can be seen in the the Baseline and SiSp track efficiency where the merging rate increases sharply. In contrast, the model's efficiency is approximately flat. The model's performance is below the Baseline at larger angular separations from the RoI, where the average track p_{T} drops below 10 GeV. Along $\Delta\eta$, the larger pitch size and the narrow RoI definition ($|\Delta\phi| < 0.05$ and $|\Delta\eta| < 0.05$) result in less of a change in merging rate over the span of the ROI. Again, the model's efficiency is roughly flat and achieves efficiencies comparable to the SiSp track collection.



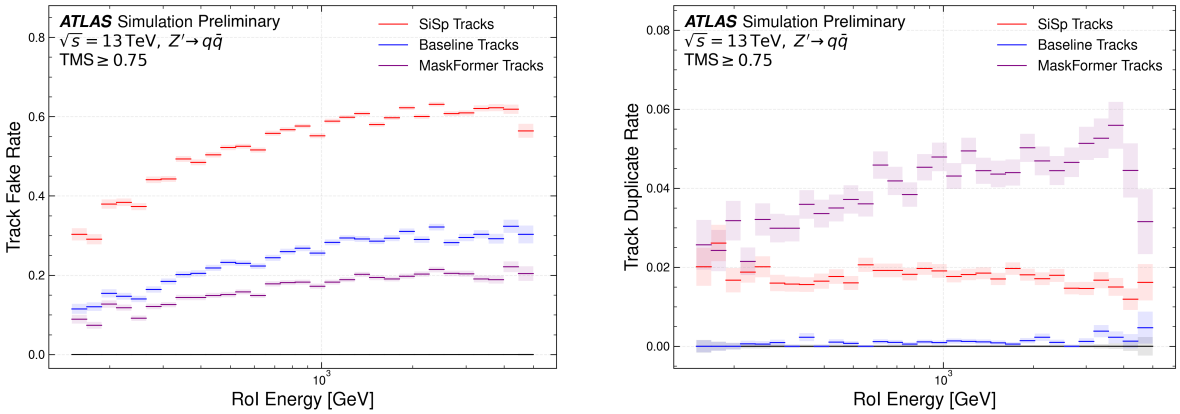
(a) (b)

Figure 2: Reconstruction efficiency of charged particles as a function of (a) $\Delta\phi$ and (b) $\Delta\eta$ between the corresponding truth particle and the RoI axis. The Transformer-based model (purple) is compared with the standard ATLAS reconstruction (blue) and the set of tracks entering the ambiguity solver (red). The reference for efficiency calculation is all primary truth particles that have contributed to at least eight ID clusters and satisfy $p_{\mathrm{T}} > 500$ MeV.

4.2 Duplicate and Fake Tracks

The rate of fake tracks and duplicate tracks is shown in Figure 3. The fake rate remains below 20%, an improvement relative to the standard reconstruction. Note, the Baseline fake rate rises to about 5% when using a looser TMS > 0.5 criterion with the purity of the learned reconstruction being slightly better. The tighter threshold ensures that only tracks with a high fraction of correctly assigned clusters are counted as genuine, providing a more stringent assessment of reconstruction quality in dense environments where cluster sharing is frequent.

The Transformer-based model exhibits a duplicate rate of 3–5% in dense regions arising from multiple reconstructed tracks being assigned to the same truth particle. Ongoing work focuses on refining the training loss and implementing query-suppression mechanisms to reduce the duplicate rate without degrading efficiency.



(a) (b)

Figure 3: (a) Fake-track rate and (b) duplicate-track rate as a function of hadronic RoI energy. The Transformer-based model (purple) is compared with the standard ATLAS reconstruction (blue) and the set of tracks entering the ambiguity solver (red).

4.3 Cluster Sharing

A key advantage of the approach presented in this note is its ability to robustly handle shared clusters amongst multiple track hypotheses, thereby removing the need for the current networks that identify merged clusters. Figure 4 shows the model successfully assigns shared clusters to multiple tracks, preserving efficiency where the standard reconstruction would discard them. The confusion matrix for the Transformer-based model shows a narrower spread around the diagonal compared to the SiSP candidates, indicating a higher fraction of correctly assigned clusters. In the current reconstruction, the pixel-classification network can label a cluster as compatible with up to three or more charged particles. Such clusters, corresponding to the highest multiplicity category, are permitted to be used by no more than four reconstructed tracks. Although a few special cases allow limited exceptions, this hard limit within the ambiguity solver is a primary constraint in dense environments. In contrast, the learned model integrates this contextual information across layers, enabling it to perform well in such highly populated regions, overcoming the limitations of the baseline reconstruction.

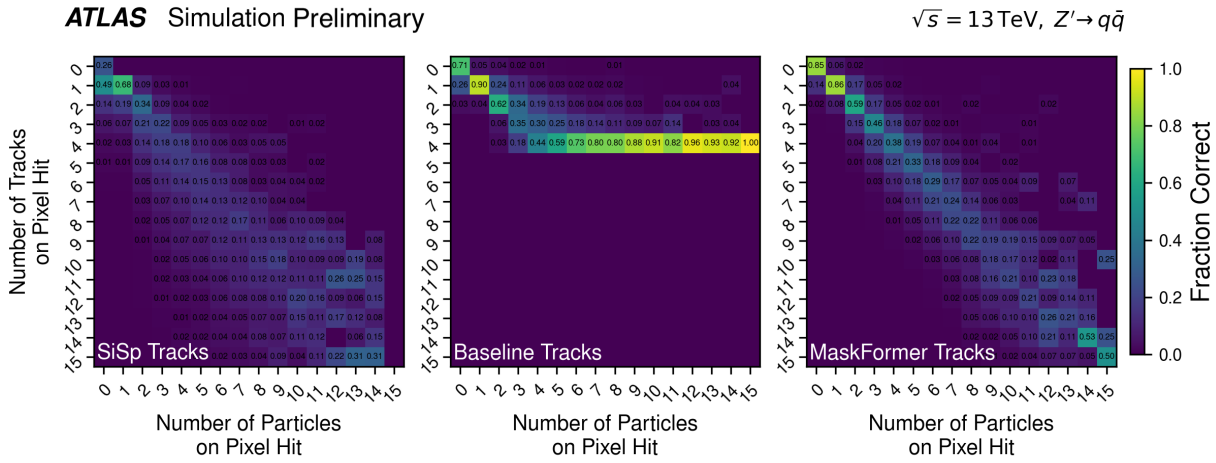


Figure 4: The cluster confusion matrices comparing the number of particles contributing to a pixel hit against the number of predicted tracks using the pixel hit for the set of tracks (from left to right) entering the ambiguity solver, from the standard ATLAS reconstruction, and the Transformer-based model. The values are normalised by column to the number of true particles. The learned reconstruction is able to have a good rate of correct matches for clusters created by a large number of particles.

5 Conclusion

The Transformer-based model demonstrates that an end-to-end, attention-based reconstruction can perform effective track finding in dense environments typical of high-p_{T} jets. It complements the standard reconstruction, by significantly improving track reconstruction efficiency up to 30% in dense environments where hit sharing is frequent. Although duplicate rates remain higher than the baseline reconstruction, simple reduction techniques such as removing tracks that have a majority of clusters in common with one other track and tuning of the probability cuts applied to the model have yet to be explored. Therefore, the Transformer-based method provides a robust proof of concept for specialised tracking in dense environments. Future studies will focus on reducing duplicate tracks and integrating the model as a specialised module within the ATLAS ambiguity-solver framework. This will enable large-scale validation, including studies of track-parameter resolution, using both simulated and collision data.

References

[1] ATLAS Collaboration, *Software Performance of the ATLAS Track Reconstruction for LHC Run 3*, Comput. Softw. Big Sci. **8** (2024) 9, arXiv: 2308.09471 [hep-ex] (cit. on pp. 2, 6).

[2] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003 (cit. on pp. 2, 6).

[3] ATLAS Collaboration, *Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2*, Eur. Phys. J. C **77** (2017) 673, arXiv: 1704.07983 [hep-ex] (cit. on pp. 2, 4, 6).

[4] ATLAS Collaboration, *ATLAS Run 3 charged particle track seed finding performance*, ATL-PHYS-PUB-2023-034, 2023, URL: https://cds.cern.ch/record/2882156 (cit. on p. 2).

[5] ATLAS Collaboration, *A neural network clustering algorithm for the ATLAS silicon pixel detector*, JINST **9** (2014) P09009, arXiv: 1406.7690 [hep-ex] (cit. on p. 2).

[6] ATLAS Collaboration, *Training and validation of the ATLAS pixel clustering neural networks*, ATL-PHYS-PUB-2018-002, 2018, URL: https://cds.cern.ch/record/2309474 (cit. on p. 2).

[7] E. E. Khoda, *ATLAS pixel cluster splitting using Mixture Density Networks*, tech. rep., CERN, 2019, URL: https://cds.cern.ch/record/2687968 (cit. on pp. 2, 4).

[8] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, Nucl. Instrum. Meth. A **262** (1987) 444 (cit. on p. 2).

[9] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Eur. Phys. J. C **77** (2017) 490, arXiv: 1603.02934 [hep-ex] (cit. on pp. 2, 3).

[10] ATLAS Collaboration, *Tracking efficiency studies in dense environments*, tech. rep., 2022, URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/IDTR-2022-03/ (cit. on p. 2).

[11] CMS Collaboration, *CMS DeepJet Tagger: Improved Jet Flavour Tagging with Deep Learning*, tech. rep., CMS Collaboration, 2023, URL: https://cds.cern.ch/record/2854609 (cit. on p. 2).

[12] I. Zurbano Fernandez et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, **10/2020** (2020), ed. by I. Béjar Alonso et al. (cit. on p. 2).

[13] C. Biscarat, S. Caillou, C. Rougier, J. Stark and J. Zahreddine, *Towards a realistic track reconstruction algorithm based on graph neural networks for the HL-LHC*, EPJ Web Conf. **251** (2021) 03047, arXiv: 2103.00916 [physics.ins-det] (cit. on p. 2).

[14] ATLAS Collaboration, *Computational Performance of the ATLAS ITk GNN Track Reconstruction Pipeline*, ATL-PHYS-PUB-2024-018, 2024, URL: https://cds.cern.ch/record/2914282 (cit. on p. 2).

[15] B. Cheng, A. G. Schwing and A. Kirillov, *Per-Pixel Classification is Not All You Need for Semantic Segmentation*, (2021), arXiv: 2107.06278 [cs.CV] (cit. on pp. 3, 5).

[16] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov and R. Girdhar, *Masked-attention Mask Transformer for Universal Image Segmentation*, (2022), arXiv: 2112.01527 [cs.CV] (cit. on pp. 3, 5).

[17] A. Vaswani et al., *Attention Is All You Need*, (2017), arXiv: 1706.03762 [cs.CL] (cit. on p. 3).

[18] P. Calafiura et al., 'TrackML: A High Energy Physics Particle Tracking Challenge', *2018 IEEE 14th International Conference on e-Science (e-Science)*, 2018 344 (cit. on p. 3).

[19] S. Van Stroud et al., *Transformers for Charged Particle Track Reconstruction in High Energy Physics*, (2024), arXiv: 2411.07149 [hep-ex] (cit. on pp. 3, 4).

[20] S. Van Stroud et al., *Secondary vertex reconstruction with MaskFormers*, Eur. Phys. J. C **84** (2024) 1020, arXiv: 2312.12272 [hep-ex] (cit. on p. 3).

[21] D. Kobylianskii et al., *GLOW: A Unified Particle Flow Transformer*, (2025), arXiv: 2508.20092 [hep-ex] (cit. on p. 3).

[22] ATLAS Collaboration, *ATLAS Inner Tracker Pixel Detector: Technical Design Report*, ATLAS-TDR-030; CERN-LHCC-2017-021, 2017, URL: https://cds.cern.ch/record/2285585 (cit. on p. 4).

[23] ATLAS Collaboration, *Operation and performance of the ATLAS semiconductor tracker in LHC Run 2*, JINST **17** (2022) P01013, arXiv: 2109.02591 [physics.ins-det] (cit. on p. 4).

[24] ATLAS Collaboration, *ATLAS Pixel Detector: Technical Design Report*, ATLAS-TDR-11; CERN-LHCC-98-013, 1998, URL: https://cds.cern.ch/record/381263 (cit. on p. 4).

[25] M. Tancik et al., 'Fourier features let networks learn high frequency functions in low dimensional domains', *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546 (cit. on p. 5).

[26] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016, URL: http://www.deeplearningbook.org (cit. on p. 5).

[27] R. Girshick, *Fast R-CNN*, 2015, arXiv: 1504.08083 [cs.CV], URL: https://arxiv.org/abs/1504.08083 (cit. on p. 5).

[28] S. Guthe and D. Thuerck, *Algorithm 1015: A Fast Scalable Solver for the Dense Linear (Sum) Assignment Problem*, ACM Trans. Math. Softw. **47** (2021), ISSN: 0098-3500 (cit. on p. 5).

[29] I. Loshchilov and F. Hutter, 'Decoupled Weight Decay Regularization', *International Conference on Learning Representations*, 2017, URL: https://api.semanticscholar.org/CorpusID:53592270 (cit. on p. 6).

[30] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv: 1005.4568 [physics.ins-det] (cit. on p. 6).

[31] S. Agostinelli et al., *GEANT4 – a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250 (cit. on p. 6).

[32] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph] (cit. on p. 6).

[33] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on p. 6).

[34] NNPDF Collaboration, R. D. Ball et al., *Parton distributions with LHC data*, Nucl. Phys. B **867** (2013) 244, arXiv: 1207.1303 [hep-ph] (cit. on p. 6).

[35] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. A **462** (2001) 152 (cit. on p. 6).

[36] ATLAS Collaboration, *Clustering and Tracking in Dense Environments with the ATLAS Inner Tracker for the High-Luminosity LHC*, ATL-PHYS-PUB-2023-022, 2023, URL: https://cds.cern.ch/record/2867615 (cit. on p. 6).

[37] ATLAS Collaboration, *Early Inner Detector Tracking Performance in the 2015 Data at* $\sqrt{s} = 13$ *TeV*, ATL-PHYS-PUB-2015-051, 2015, URL: https://cds.cern.ch/record/2110140 (cit. on p. 7).