PROCEEDINGS OF SCIENCE

Flavour-Tagging with Graph Neural Networks with the ATLAS Detector

Helena Santos, on behalf of the ATLAS Collaboration[a,*]

[a] *Laboratório de Instrumentação e Física Experimental de Partículas , Av. Prof. Gama Pinto 2, Lisboa, Portugal*

E-mail: helena@lip.pt

The identification of jets containing b-hadrons is key to many physics analyses at the Large Hadron Collider, including measurements involving Higgs bosons or top quarks, and searches for physics beyond the Standard Model. In this contribution, the most recent enhancements in the capability of ATLAS to separate b-jets from jets stemming from lighter quarks will be presented. The improved performance originates from the usage of state-of-the-art machine learning algorithms based on graph neural networks. A factor of more than two to reject light- and c-quark-initiated jets is observed compared to the current performance. Perspectives for the High-Luminosity LHC will be discussed.

ATL-PHYS-PROC-2025-123
05 December 2025

XXXII International Workshop on Deep Inelastic Scattering and Related Subjects (DIS2025)
24-28 March, 2025
Cape Town, South Africa

*Speaker

 ISSN 1824-8039 . Published by SISSA Medialab.

PoS(DIS2025)080

Heavy–flavour jets play a key role in the physics program of the LHC, particularly in analyses such as Higgs boson decays to $b\bar{b}$ and $c\bar{c}$ in the Standard Model and beyond, top-quark studies, and investigations of parton energy loss in the quark–gluon plasma. A distinctive advantage for identifying b-jets arises from the relatively long lifetime of b-hadrons, which produces characteristic signatures in the detector. These typically include large impact parameters of tracks, displaced secondary vertices and missing hits in the innermost tracking layers of the experiments. Such observables are reconstructed primarily from tracks associated with the jet and constitute the basis of flavour–tagging techniques.

ATLAS [1] flavour–tagging has evolved over time from deep neural-network classifiers to more advanced graph-based approaches. During LHC Run 2, the taggers have been based on deep neural network algorithms. In this paradigm, low-level algorithms first reconstruct track-related properties and displaced vertices, which are subsequently used as inputs to high-level classifiers, such as the DL1 series [2]. This two-step approach has demonstrated excellent capabilities, but it separates the tasks of low-level reconstruction and high-level jet classification. Currently, ATLAS is exploring Graph Neural Networks (GNN). In this approach, jets are represented as graphs, where the nodes correspond to tracks and the edges learn the relationships between them. The first ATLAS implementation, GN1 [3], is based on a Graph Attention Network architecture that aggregates information from neighbouring nodes. An improved model, GN2 [4], adopts a transformer-inspired attention mechanism, along with optimised training strategies such as a one-cycle learning-rate schedule, layer normalisation, and dropout, enhancing convergence speed and training stability. The GNNs developed for flavour–tagging are inherently multimodal, capable of processing heterogeneous inputs such as jets, tracks, and vertices. They are also multitask, simultaneously optimising objectives such as jet flavour classification, track origin prediction, and vertex finding. The network input consists of two jet-level features (p_{T} and η) and a variable-length array of track-level features [3]. During training, the GNN builds initial track representations, which are then iteratively refined by propagating information along graph edges. The total loss function guiding the training includes the jet flavour classification loss, complemented by auxiliary losses associated with vertex and track predictions. The updated track representations are used to predict the jet type, classify the origin of each track, and estimate vertex compatibility between track pairs. All algorithms discussed in these proceedings have been integrated into the ATLAS software [5].

The discriminant, $D_{\mathrm{b}} = \log[p_{\mathrm{b}}/(f_{\mathrm{c}}\, p_{\mathrm{c}} + (1 - f_{\mathrm{c}})\, p_{\mathrm{u}})]$, combines the algorithm output probabilities of a jet being a b-jet, a c-jet or a light-jet and the comparison between the legacy DL1r and the GN1 b-tagging discriminants is shown in Figure 1. The loosest 85% and tightest 60% working points are marked as solid lines for GN1 and dashed lines for DL1r. The GN1 model shifts the b-jet distribution to higher discriminant values, whereas the c-jet distribution is shifted to lower values of D_{b} when compared with DL1r, enhancing the separation.

Figure 2 compares the b-tagging performance of the DL1d, GN1, and GN2 algorithms in a $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250$ GeV and in a Z' sample with $250 < p_{\mathrm{T}} < 6000$ GeV [4]. DL1d is a modification of DL1r [2] in which RNNIP [6], a recurrent neural network to learn track impact-parameter correlations, is replaced by DIPS [7], a Deep Sets architecture that models the jet as a set of tracks. In turn, GN2 upgrades GN1. This model already improves significantly over DL1d, but GN2 transcends it with an improved training strategy and a more efficient transformer architecture. Across the full b-jet efficiency range, GN2 provides the highest c-jet and light-jet rejection. In the

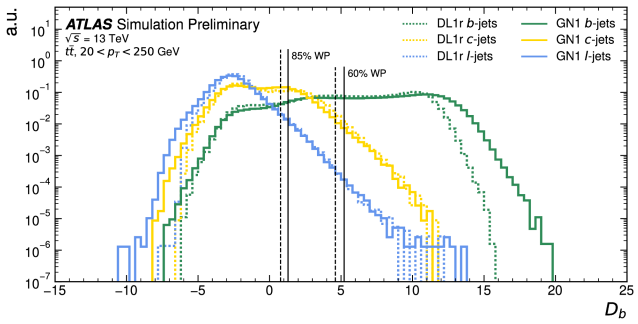

Figure 1: Comparison of DL1r and GN1 b-tagging discriminants D_b for $t\bar{t}$ jets, showing 85% (loosest) and 60% (tightest) working points as solid (GN1) and dashed (DL1r) lines. Discriminants use $f_\mathrm{c} = 0.018$ (DL1r) and 0.05 (GN1), with all jet flavour distributions normalised to unit area. More details can be found in Ref. [3].

$t\bar{t}$ sample at the 70% working point, GN2 improves light-jet rejection by a factor of ∼2 and c-jet rejection by a factor of ∼3 relative to DL1d. In the Z' sample at the 30% working point, the gains are even larger, with light-jet rejection improving by a factor of ∼4 and c-jet rejection by a factor of ∼3.5.

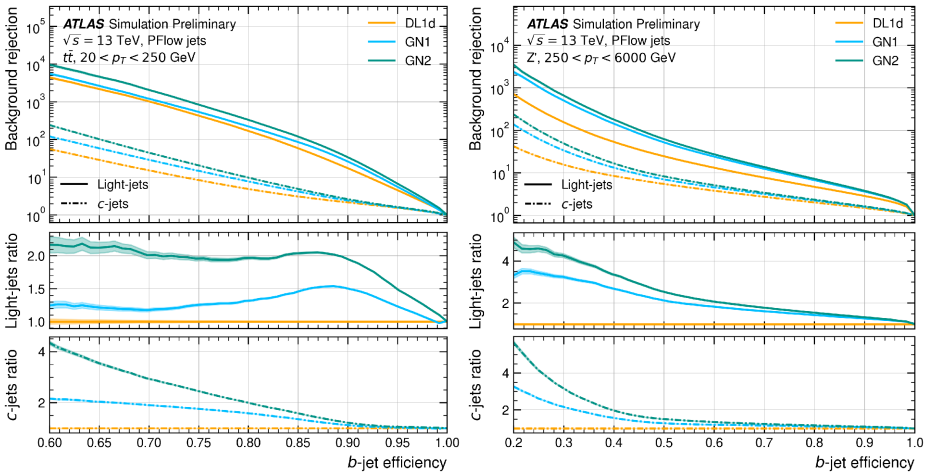

Figure 2: The c-jet and light-jet rejections as a function of the b-jet tagging efficiency for: (left) jets in a $t\bar{t}$ sample with $20 < p_\mathrm{T} < 250$ GeV and (right) jets for in a Z' sample with $250 < p_\mathrm{T} < 6000$ GeV. Bottom panels display the ratio to DL1d performance. Tagging discriminants use $f_\mathrm{c} = 0.018$ (DL1d), 0.05 (GN1), and 0.1 (GN2). Shaded regions indicate binomial errors. More details can be found in Ref. [4].

Figure 3 compares DL1r to multiple GN1 model variants, including the baseline model and ablations with auxiliary objectives removed. The models without one or both of the auxiliary objectives show significantly reduced c- and light-jet rejection when compared with the baseline GN1 model. GN1 No Aux (the baseline GN1 model without auxiliary losses) performs similarly to DL1r, while GN1 TC (with a track-classification auxiliary task) and GN1 Vert (with a track-pair vertex compatibility auxiliary task) exhibit comparable results. The combined auxiliary objectives yield optimal performance. Remarkably, GN1 No Aux achieves performance comparable or superior to

DL1r despite excluding low-level algorithm inputs, suggesting GN1's architectural improvements effectively compensate for this removal. Both GN1 TC and GN1 Vert variants individually outperform DL1r, confirming that each auxiliary objective contributes significantly to the baseline's higher performance.

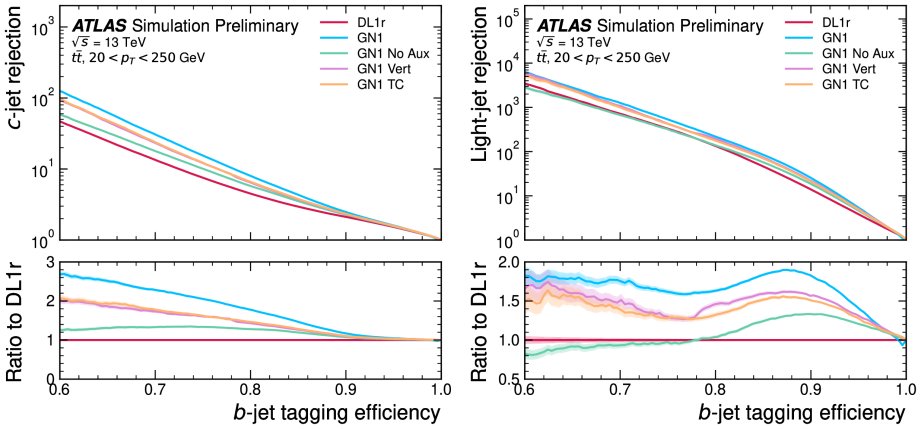

Figure 3: The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency for jets in a $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250$ GeV, comparing DL1r and GN1 configurations with different auxiliary objectives: nominal, no auxiliaries (GN1 No Aux), track classification only (GN1 TC), and vertexing only (GN1 Vert). Bottom panels show the ratio to DL1r performance. Discriminants use $f_{\mathrm{c}} = 0.018$ (DL1r) and 0.05 (GN1), with shaded regions indicating binomial errors. The restricted b-jet tagging efficiency focuses on typical b-jet efficiencies. More details can be found in Ref. [3].

The High-Luminosity LHC (HL-LHC), scheduled to begin operation in 2030, will reach an instantaneous luminosity of up to $7.5\text{x}10^{34}$ cm^{-2}s^{-1} with $\langle\mu\rangle$ values up to 200, significantly exceeding Run 3's peak luminosity of $2.5\text{x}10^{34}$ cm^{-2}s^{-1} and $\langle\mu\rangle$ of 55. These extreme conditions will challenge the ATLAS tracking detector with unprecedented radiation levels and occupancy. To maintain at least the current tracking performance, the existing inner detector will be replaced by an all-silicon inner tracker (ITk) [9]. GN1 has been trained on simulated samples with HL-LHC conditions and the ITk geometry. The left panel of Figure 4 shows the b-tagging efficiency as a function of jet p_{T} at fixed light-jet rejection of 100 in each bin for jets in a Z' sample. While GN1 achieves nearly 90% b-efficiency at $p_{\mathrm{T}} = 250$ GeV, the performance declines at high-p_{T} but maintains approximately 50% efficiency at $p_{\mathrm{T}} = 5$ TeV. The right panel displays b-tagging efficiency as a function of $|\eta|$ for jets in a $t\bar{t}$ sample at the same rejection. GN1 maintains stable performance (>75% efficiency) up to $|\eta| < 3.5$. Increased pile-up track contamination in hard-scatter jets shows minimal impact on flavour–tagging performance. These results demonstrate the model's ability to generalize to a completely new detector geometry.

In summary, graph neural networks significantly outperform deep neural networks in jet flavour–tagging. In the GN1 algorithm, both auxiliary objectives - track and vertex classifications - are shown to be crucial, as their removal degrades performance. Notably, GN1 with no auxiliary objectives already surpasses DL1r, despite the absense of low-level inputs, demonstrating the architecture's effectiveness. The upgraded GN2 model demonstrates even better performance, rejecting light–jets up to factor of ∼4 compared to DL1d. GNN's ability to learn optimal representations

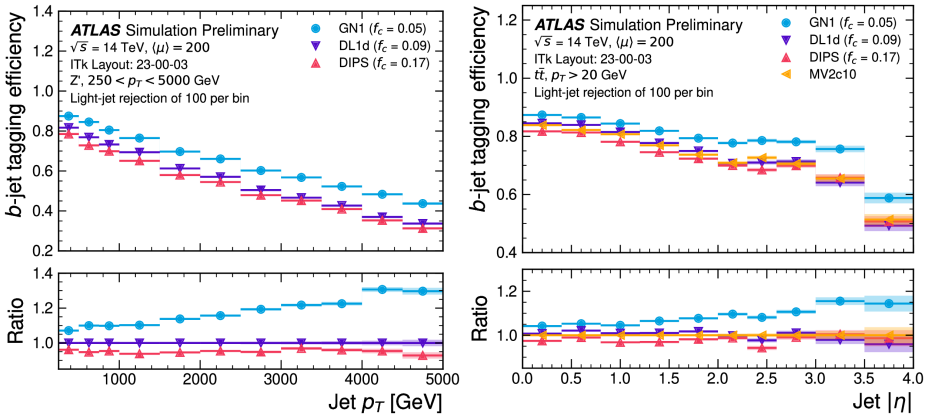

Figure 4: Left: b-jet tagging efficiency as a function of jet p_{T} for jets in Z' sample at fixed light-jet rejection of 100, with ratio to DL1d performance shown in the bottom panel. Right: The b-jet tagging efficiency for jets in the $t\bar{t}$ sample as a function of jet $|\eta|$ for $t\bar{t}$ jets at the same rejection, with ratio to previous benchmark MV2c10. Shaded regions indicate binomial errors. More details can be found in Ref. [8].

from track features allows robust generalization to new detector geometries, establishing it as a powerful solution for HL-LHC conditions. In these challenging pileup regimes, the b-jet identification is enhanced by 30% at high-p_{T} and 15% in forward regions ($\eta > 2.5$), while maintaining stable performance up to $|\eta| \sim 3.5$.

References

[1] ATLAS Collaboration, *JINST* **3**, S08003 (2008).

[2] ATLAS Collaboration, *Eur.Phys.J.* C83 (2023) **681**, arXiv:2211.16345[physics.data-an].

[3] ATLAS Collaboration, *ATL-PHYS-PUB-2022-027* (2022), https://cds.cern.ch/record/2811135/.

[4] ATLAS Collaboration, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2023-01/.

[5] ATLAS Collaboration, *ATL-SOFT-PUB-2021-001* (2021), https://cds.cern.ch/record/2767187.

[6] ATLAS Collaboration, *ATL-PHYS-PUB-2017-003* (2017), https://cds.cern.ch/record/2255226.

[7] ATLAS Collaboration, *ATL-PHYS-PUB-2020-014* (2020), https://cds.cern.ch/record/2718948.

[8] ATLAS Collaboration, *ATL-PHYS-PUB-2022-047* (2022), https://cds.cern.ch/record/2839913.

[9] ATLAS Collaboration, *ATLAS-TDR-030; CERN-LHCC-2017-021* (2017), https://cds.cern.ch/record/2285585.