ATLAS PUB Note
ATL-PHYS-PUB-2022-027
1st June 2022

Graph Neural Network Jet Flavour Tagging with the ATLAS Detector

The ATLAS Collaboration

Flavour tagging, the identification of jets originating from b- and c-quarks, is a critical component of the physics programme of the ATLAS experiment at the Large Hadron Collider. Current flavour tagging algorithms rely on the outputs of several low-level algorithms, which reconstruct various properties of jets using charged particle tracks, that are then combined using machine learning techniques. In this note a new machine learning algorithm based on graph neural networks, GN1, is introduced. GN1 uses information from a variable number of charged particle tracks within a jet, to predict the jet flavour without the need for intermediate low-level algorithms. Alongside the jet flavour prediction, the model predicts which physics processes produced the different tracks in the jet, and groups tracks in the jet into vertices. These auxiliary training objectives provide useful additional information on the contents of the jet and improve performance. GN1 compares favourably with the current ATLAS flavour tagging algorithms. For a b-jet efficiency of 70%, the light (c)-jet rejection is improved by a factor of ~1.8 (~2.1) for jets coming from $t\bar{t}$ decays with transverse momentum $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$. For jets coming from Z' decays with transverse momentum $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$, the light ($c$)-jet rejection improves by a factor ~6 (~2.8) for a comparative 30% b-jet efficiency.

1 Introduction

Flavour tagging, the identification of jets originating from b- and c-quarks, is a critical component of the physics programme of the ATLAS experiment [1] at the Large Hadron Collider (LHC) [2]. It is of particular importance for the study of the Standard Model (SM) Higgs boson and the top quark, which preferentially decay to b-quarks [3, 4], and additionally for several Beyond Standard Model (BSM) resonances that readily decay to heavy flavour quarks [5]. The significant lifetime of b-hadrons, approximately 1.5 ps [6], provides the unique signature of a secondary decay vertex which has a high mass and is significantly displaced from the primary vertex. Additional signatures of b-hadrons are the tertiary decay vertex, resulting from $b \rightarrow c$ decay chains, and the reconstructed trajectories of charged particles (henceforth simply referred to as tracks) with large impact parameters[1] (IPs). These signatures are primarily identified using tracks associated to jets. As such, efficient and accurate track reconstruction is essential for high performance flavour tagging.

This note introduces a novel algorithm, GN1, which uses Graph Neural Networks (GNNs) [7] with auxiliary training objectives, to aid the primary goal of classifying whether jets originate from b- or c-quarks (referred to as a flavour tagger). The concept is illustrated in Fig. 1. The use of GNNs offers a natural way to classify jets with variable numbers of unordered associated tracks, while allowing for the inclusion of auxiliary training objectives [8, 9].

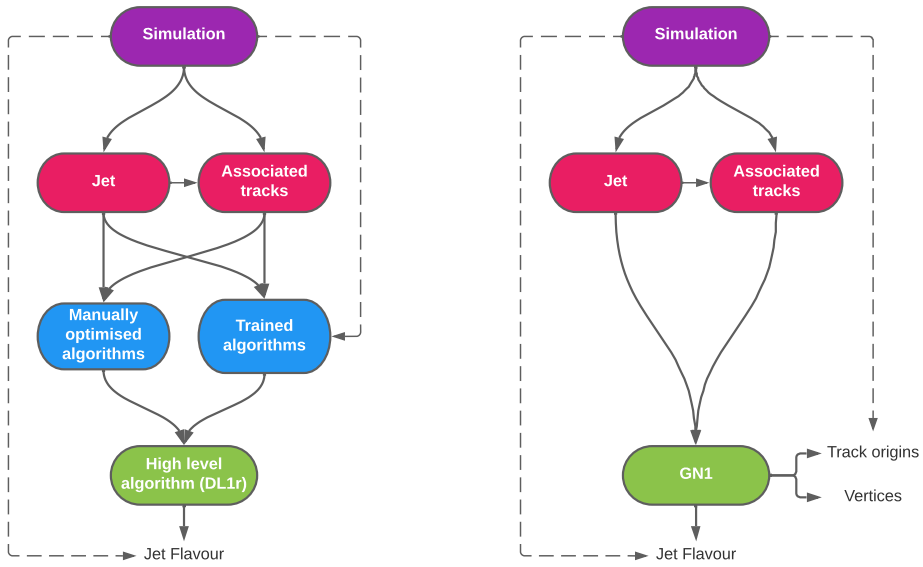

Figure 1: Comparison of the existing flavour tagging scheme (left) and GN1 (right). The existing approach utilises low-level algorithms (shown in blue), the outputs of which are fed into a high-level algorithm (DL1r). Instead of being used to guide the design of the manually optimised algorithms, additional truth information from the simulation is now being used as auxiliary training targets for GN1. The solid lines represent reconstructed information, whereas the dashed lines represent truth information.

The current ATLAS flavour tagger, DL1r [10], is a deep neural network which takes the outputs of a number of independently optimised "low-level" algorithms [11] as inputs. Each of these low-level algorithms

[1] The distance of closest approach from a track to the primary vertex.

makes use of tracks to reconstruct a particular aspect of the experimental signature of heavy flavour jets. The low-level algorithms can be manually optimised reconstruction algorithms, for example the SV1 and JetFitter algorithms that reconstruct displaced decay vertices, or trained taggers such as RNNIP and DIPS that use the IPs of a variable number of tracks to identify the flavour of the jet [11–14]. In contrast GN1 utilises a single neural network, which directly takes the tracks and some information about the jet as inputs. As such, it does not depend on any other flavour tagging algorithm, and a single training of the GN1 fully optimises all aspects of the algorithm.

GN1 is trained to understand the internal structure of the jet through the use of two auxiliary training objectives: the grouping of tracks originating from a common vertex, and the prediction of the underlying physics process from which each track originated. These auxiliary objectives are meant to guide the neural network towards a more complete understanding of the underlying physics, removing the need for the low-level algorithms, and therefore simplifying the process of optimising the tagger for new regions of phase space (e.g. c-tagging or high-p_{T} b-tagging), or when the detector or charged particle reconstruction algorithms are updated. The training targets for the primary and auxiliary objectives are extracted from "truth information", i.e. information only available in simulation, as opposed to reconstructed quantities available in both collision data and simulation.

In this note, the following benefits of this approach will be shown:

1. Improved performance with respect to the current ATLAS flavour tagging algorithms, with larger background rejection for a given signal efficiency.
2. The same network architecture can be easily optimised for a wider variety of use cases (e.g. c-jet tagging and high-p_{T} jet tagging), since there are no low-level algorithms to retune.
3. There are fewer flavour tagging algorithms to maintain.
4. Alongside the network's prediction of the jet flavour, the auxiliary vertex and track origin predictions provide more information on why a jet was (mis)tagged or not. This information can also have uses in other applications, for instance to explicitly reconstruct displaced decay vertices or to remove fake tracks.[2]

This note is organised as follows: a brief description of the ATLAS detector, object definitions and selections, and samples are provided in Section 2; details about the model architecture and training procedure are given in Section 3; and results are discussed in Section 4.

[2] A fake track is defined as a track with a truth-matching probability less than 0.5, where the truth-matching probability is defined in Ref. [15].

2 Experimental Setup

2.1 The ATLAS Detector

The ATLAS detector at the LHC covers nearly the entire solid angle around the collision point.[3] It consists of an inner tracking detector surrounded by a thin superconducting solenoid, electromagnetic and hadron calorimeters, and a muon spectrometer incorporating three large superconducting air-core toroidal magnets.

The inner-detector system (ID) is immersed in a 2 T axial magnetic field and provides charged-particle tracking in the range $|\eta| < 2.5$. The high-granularity silicon pixel detector covers the vertex region and typically provides four measurements per track, the first hit normally being in the insertable B-layer (IBL) installed before Run 2 [16, 17]. It is followed by the silicon microstrip tracker (SCT), which usually provides eight measurements per track. These silicon detectors are complemented by the transition radiation tracker (TRT), which enables radially extended track reconstruction up to $|\eta| = 2.0$. The TRT also provides electron identification information based on the fraction of hits (typically 30 in total) above a higher energy-deposit threshold corresponding to transition radiation. Reconstructed charged particles are assumed to have a charge of ± 1.

A complete overview of the ATLAS detector is provided in Ref. [1].

2.2 Object Definitions and Selection

The trajectories of charged particles are reconstructed as tracks from the energy depositions (hits) of the particles as they traverse the sensitive elements of the inner detector. Track selection follows the loose selection described in Ref. [14] and outlined in Table 1, which was found to improve the flavour tagging performance compared to previous tighter selections, whilst ensuring good resolution of tracks and a low fake rate [15]. The transverse IP d_0 and longitudinal IP z_0 are measured with respect to the hard scatter primary vertex, defined as the reconstructed primary vertex (PV) with the largest sum of the transverse momentum (p_{T}) of the associated tracks squared, $\sum p_{\mathrm{T}}^2$.

Jets are reconstructed from particle-flow objects [18] using the anti-k_T algorithm [19] with a radius parameter of 0.4. The jet energy scale is calibrated according to Ref. [20]. Jets are also required not to overlap with a generator-level electron or muon from W boson decays. All jets are required to have a pseudorapidity $|\eta| < 2.5$ and $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. Additionally, a standard selection using the Jet Vertex Tagger (JVT) algorithm at the tight working point is applied to jets with $p_{\mathrm{T}} < 60\,\mathrm{GeV}$ and $|\eta| < 2.4$ in order to suppress pileup contamination [21]. Tracks are associated to jets using a ΔR association cone, the width of which decreases as a function of jet p_{T}, with a maximum cone size of $\Delta R \approx 0.45$ for jets with $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ and minimum cone size of $\Delta R \approx 0.25$ for jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$. If a track is within the association cones of more than one jet, it is assigned to the jet which has a smaller $\Delta R(\mathrm{track}, \mathrm{jet})$.

Jet flavour labels are assigned according to the presence of a truth hadron within $\Delta R(\mathrm{hadron}, \mathrm{jet}) < 0.3$ of the jet axis. If a b-hadron is found the jet is labelled a b-jet. In the absence of a b-hadron, if a c-hadron is

[3] ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point in the centre of the detector and the z-axis along the beam pipe. The x-axis points from the interaction point 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 $\eta = -\ln\tan(\theta/2)$. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

Table 1: Quality selections applied to tracks, where d_0 is the transverse IP of the track, z_0 is the longitudinal IP with respect to the PV and θ is the track polar angle. Shared hits are hits used on multiple tracks which have not been classified as split by the cluster-splitting neural networks [15]. Shared hits on pixel layers are given a weight of 1, while shared hits in the SCT are given a weight of 0.5. A hole is a missing hit, where one is expected, on a layer between two other hits on a track.

Parameter	**Selection**
p_{T}	> 500 MeV
$\lvert d_0 \rvert$	< 3.5 mm
$\lvert z_0 \sin\theta \rvert$	< 5 mm
Silicon hits	≥ 8
Shared silicon hits	< 2
Silicon holes	< 3
Pixel holes	< 2

found the jet is called a c-jet. If no b- or c-hadrons are found, but a τ is found in the jet, it is labelled as a τ-jet, else it is labelled as a light-jet.

2.3 Datasets

To train and evaluate the model, simulated SM $t\bar{t}$ and BSM Z' events initiated by proton-proton collisions at a center of mass energy $\sqrt{s} = 13$ TeV are used. The Z' sample is constructed in such a manner that it has a relatively flat jet p_{T} spectrum up to 5 TeV and decays to an equal numbers of b-, c- and light-jets. The generation of the simulated event samples includes the effect of multiple pp interactions per bunch crossing with an average pileup of $\langle\mu\rangle = 40$, which includes the effect on the detector response due to interactions from bunch crossings before or after the one containing the hard interaction.

The $t\bar{t}$ events are generated using the PowhegBox [22–25] v2 generator at next-to-leading order with the NNPDF3.0NLO [26] set of parton distribution functions (PDFs) and the h_{damp} parameter[4] set to 1.5 times the mass of the top-quark (m_{top}) [27], with $m_{\mathrm{top}} = 172.5$ GeV. The events are interfaced to Pythia 8.230 [28] to model the parton shower, hadronisation, and underlying event, with parameters set according to the A14 tune [29] and using the NNPDF2.3LO set of PDFs [30]. Z' events are generated with Pythia 8.2.12 with the same tune and PDF set. The decays of b- and c-hadrons are performed by EvtGen v1.6.0 [31]. Particles are passed through the ATLAS detector simulation [32] based on GEANT4 [33].

For the $t\bar{t}$ events, at least one W boson from the top quark decay is required to decay leptonically. Truth labelled b-, c- and light-jets are kinematically re-sampled in p_{T} and η to ensure identical distributions in these variables. The resulting dataset contains 30 million jets, 60% of which are $t\bar{t}$ jets and 40% of which are Z' jets. While DL1r uses 70% $t\bar{t}$ jets and 30% Z' jets, the change in sample composition did not affect the final performance of GN1. To evaluate the performance of the model, 500k jets from both the $t\bar{t}$ and Z' samples, which are statistically independent from the training sample, are used. Track- and jet-level inputs are scaled to have a central value of zero and a variance of unity before training and evaluation.

[4] The h_{damp} parameter is a resummation damping factor and one of the parameters that controls the matching of Powheg matrix elements to the parton shower and thus effectively regulates the high-p_T radiation against which the $t\bar{t}$ system recoils.

3 Neural Network Architecture & Training

3.1 Model Inputs

GN1 is given two jet variables and 21 tracking related variables for each track fed into the network. The jet transverse momentum and signed pseudorapidity constitute the jet-level inputs, with the track-level inputs listed in Table 2. If a jet has more than 40 associated tracks, the first 40 tracks with the largest transverse IP significance[5] $s(d_0)$ are selected as inputs. Full track parameter information and associated uncertainties, along with detailed hit information, carry valuable information about the jet flavour. In the dense cores of high-p_{T} jets, tracks are highly collimated and separation between tracks can be of the same order as the active sensor dimensions, resulting in merged clusters and tracks which share hits [15]. Due to the relatively long lifetimes of b-hadrons and c-hadrons, which can traverse several layers of the ID before decaying and have highly collimated decay products, the presence of shared or missing hits is a critical signature of heavy flavour jets.

Dependence on the absolute value of the azimuthal jet angle ϕ is explicitly removed by providing only the azimuthal angle of tracks relative to the jet axis. The track pseudorapidity is also provided relative to the jet axis.

Since heavy flavour hadrons can decay semileptonically, the presence of a reconstructed lepton in the jet carries discriminating information about the jet flavour. In addition to the baseline GN1 model, the GN1 Lep variant includes an additional track-level input, leptonID, which indicates if the track was used in the reconstruction of an electron, a muon or neither. The muons are required to be combined [35], and the electrons are required to pass the *VeryLoose* likelihood-based identification working point [36].

3.2 Auxiliary Training Objectives

In addition to the jet flavour classification, two auxiliary training objectives are defined. Each auxiliary training objective comes with a training target which, similar to the jet flavour label, are truth labels derived from the simulation. The presence of the auxiliary training objectives improves the jet classification performance as demonstrated in Section 4.3.

The first auxiliary objective is the prediction of the origin of each track within the jet. Each track is labelled with one of the exclusive categories defined in Table 3 after analysing the particle interaction that led to its formation. Since the presence of different track origins is strongly related to the flavour of the jet, training GN1 to recognise the origin of the tracks may provide an additional handle on the classification of the jet flavour. This task may also aid the jet flavour prediction by acting as a form of supervised attention [37] - in detecting tracks from heavy flavour decays the model may learn to pay more attention to these tracks.

Displaced decays of b- and c-hadrons lead to secondary and tertiary vertices inside the jet. Displaced secondary vertices can also occur in light-jets as a result of material interactions and long-lived particle decays (e.g. K^0_{S} and Λ^0). The second auxiliary objective is the prediction of track-pair vertex compatibility. For each pair of tracks in the jet, GN1 predicts a binary label, which is given a value 1 if the two tracks in the pair originated from the same point in space, and 0 otherwise. To derive the corresponding truth

[5] Impact parameter significances are defined as the IP divided by its corresponding uncertainty, $s(d_0) = d_0/\sigma(d_0)$ and $s(z_0) = z_0/\sigma(z_0)$. Track IP significances are lifetime signed according to the track's direction with respect to the jet axis and the primary vertex [34].

Table 2: Input features to the GN1 model. Basic jet kinematics, along with information about the reconstructed track parameters and constituent hits are used. Shared hits, are hits used on multiple tracks which have not been classified as split by the cluster-splitting neural networks [15], while split hits are hits used on multiple tracks which have been identified as merged. A hole is a missing hit, where one is expected, on a layer between two other hits on a track. The track leptonID is an additional input to the GN1 Lep model.

Jet Input	**Description**
p_{T}	Jet transverse momentum
η	Signed jet pseudorapidity
Track Input	**Description**
q/p	Track charge divided by momentum (measure of curvature)
$\mathrm{d}\eta$	Pseudorapidity of the track, relative to the jet η
$\mathrm{d}\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the longitudinal plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the transverse plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$s(d_0)$	Lifetime signed transverse IP significance
$s(z_0)$	Lifetime signed longitudinal IP significance
nPixHits	Number of pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared pixel hits
nPixSplit	Number of split pixel hits
nSCTShared	Number of shared SCT hits
nPixHoles	Number of pixel holes
nSCTHoles	Number of SCT holes
leptonID	Indicates if track was used in the reconstruction of an electron or muon (only for GN1 Lep)

Table 3: Truth origins which are used to categorise the physics process that led to the production of a track. Tracks are matched to charged particles using the truth-matching probability [15]. A truth-matching probability of less than 0.5 indicates that reconstructed track parameters are likely to be mismeasured and may not correspond to the trajectory of a single charged particle. The "OtherSecondary" origin includes tracks from photon conversions, K^0_{S} and Λ^0 decays, and hadronic interactions.

Truth Origin	**Description**
Pileup	From a pp collision other than the primary interaction
Fake	Created from the hits of multiple particles
Primary	Does not originate from any secondary decay
fromB	From the decay of a b-hadron
fromBC	From a c-hadron decay, which itself is from the decay of a b-hadron
fromC	From the decay of a c-hadron
OtherSecondary	From other secondary interactions and decays

labels for training, truth production vertices within 0.1 mm are merged. Track-pairs where one or both of the tracks in the pair have an origin label of either Pileup or Fake are given a label of 0. Using the pairwise predictions from the model, collections of commonly compatible tracks can be grouped into vertices. The addition of this auxiliary training objective removes the need for inputs from a dedicated secondary vertexing algorithm.

Both auxiliary training objectives can be considered as "stepping stones" on the way to classifying the flavour of the jet. By requiring the model to predict the truth origin of each track and the vertex compatibility of each track-pair, the model is guided to learn representations of the jet which are connected to the underlying physics and therefore relevant for classifying the jet flavour.

3.3 Architecture

As discussed above, the GN1 model combines a graph neural network architecture [38] with auxiliary training objectives in order to determine the jet flavour. Coarse optimisation of the network architecture hyperparameters, for example number of layers and number of neurons per layer, has been carried out to maximise the tagging efficiency.

The model architecture is based on a previous implementation of a graph neural network jet tagger [9]. As compared to the previous approach, GN1 uses a only a single graph neural network and makes use of a more sophisticated graph neural network layer [39], described below. These changes yield improved tagging performance and a significant reduction in training time with respect to the previous approach.

The model takes jet- and track-level information as inputs, as detailed in Section 3.1. The jet inputs are concatenated with each track's inputs, as shown in Fig. 2. The combined jet-track vectors are then fed into a per-track initialisation network with three hidden layers, each containing 64 neurons, and an output layer with a size of 64, as shown in Fig. 3. The track initialisation network is similar to a Deep Sets model [40], but does not include a reduction operation (mean or summation) over the output track representations.

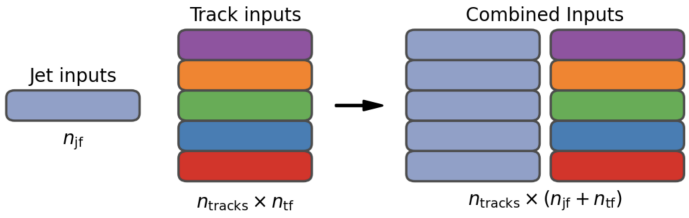

Figure 2: The inputs to GN1 are the two jet features ($n_{\text{jf}} = 2$), and an array of n_{tracks}, where each track is described by 21 track features ($n_{\text{tf}} = 21$). The jet features are copied for each of the tracks, and the combined jet-track vectors of length 23 form the inputs of GN1.

A fully connected graph is built from the outputs of the track initialisation network, such that each node in the graph neighbours every other node. Each node h_i in the graph corresponds to a single track in the jet, and is characterised by a feature vector, or representation. The per-track output representations from the initialisation networks are used to populate the initial feature vectors of each node in the graph. In each layer of the graph network, output node representations h'_i are computed by aggregating the features of h_i and neighbouring nodes $\mathcal{N}_i$ as described in Ref. [39]. First, the feature vectors of each node are fed into a fully connected layer $\mathbf{W}$, to produce an updated representation of each node $\mathbf{W}h_i$. These updated feature vectors are used to compute edge scores $e(h_i, h_j)$ for each node pair,

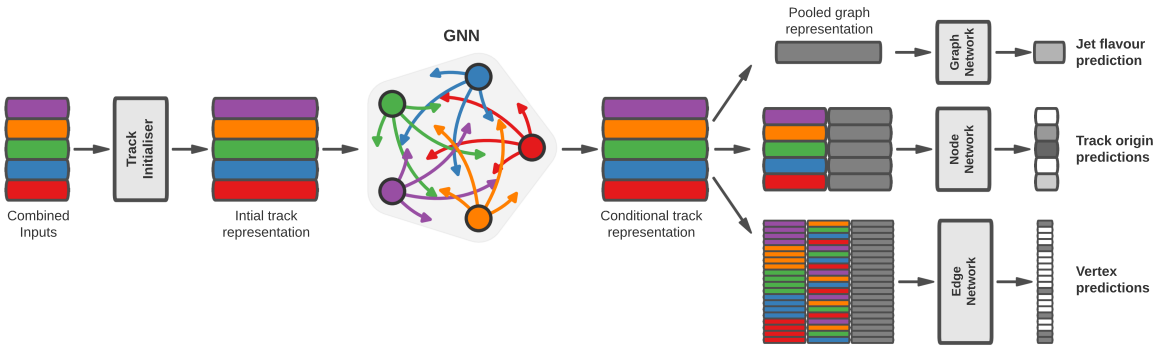

Figure 3: The network architecture of GN1. Inputs are fed into a per-track initialisation network, which outputs an initial latent representation of each track. These representations are then used to populate the node features of a fully connected graph network. After the graph network, the resulting node representations are used to predict the jet flavour, the track origins, and the track-pair vertex compatibility.

$$e(h_i, h_j) = \mathbf{a}^{\perp}\theta\left[\mathbf{W}h_i \oplus \mathbf{W}h_j\right], \tag{1}$$

where $\oplus$ denotes vector concatenation, θ is a non-linear activation function, and $\mathbf{a}$ is a second fully connected layer. These edge scores are then used to calculate attention weights a_{ij} for each pair of nodes using the softmax function over the edge scores

$$a_{ij} = \text{softmax}_j\left[e(h_i, h_j)\right]. \tag{2}$$

Finally, the updated node representation h_i' is computed by taking the weighted sum over each updated node representation $\mathbf{W}h_i$, with weights a_{ij}

$$h_i' = \sigma\left[\sum_{j\in\mathcal{N}_i} a_{ij}\cdot\mathbf{W}h_j\right]. \tag{3}$$

The above set of operations constitute a single graph network layer. Three such layers are stacked to construct the graph network, representing a balance between achieving optimal performance and preventing overtraining. The final output node feature vectors from the network are representations of each track that are conditional on the other tracks in the jet. The output representation for each track is combined using a weighted sum to construct a global representation of the jet, where the attention weights for the sum are learned during training. Three separate fully connected feedforward neural networks are then used to independently perform the different classification objectives of GN1. Each of the objectives makes use of the global representation of the jet. A summary of the different classification networks used for the various training objectives is shown in Table 4.

A node classification network, which takes as inputs the features from a single output node from the graph network and the global jet representation, predicts the track truth origin, as defined in Table 3. This network has three hidden layers containing 128, 64 and 32 neurons respectively, and an output size of seven, corresponding to the seven different truth origins.

Table 4: A summary of GN1's different classification networks used for the different training objectives. The hidden layers column contains a list specifying the number of neurons in each layer.

Network	Hidden layers	Output size
Node classification network	128, 64, 32	7
Edge classification network	128, 64, 32	1
Graph classification network	128, 64, 32, 16	3

An edge classification network, which takes as inputs the concatenated representations from each pair of tracks and the global jet representation, is used to predict whether the tracks in the track-pair belong to a common vertex. The edge network has three hidden layers containing 128, 64 and 32 neurons respectively, and a single output, which is used to perform binary classification of the track-pair compatability. These predictions are used for the auxiliary training objectives discussed in Section 3.2.

A graph classification network takes only the global jet representation as an input, and predicts the jet flavour. The graph classification network is comprised of four fully connected hidden layers with 128, 64, 32 and 16 neurons respectively, and has three outputs corresponding to the b-, c- and light-jet classes.

3.4 Training

The full GN1 training procedure minimises the total loss function L_{total}, defined in Eq. (4). This loss is composed of three terms: L_{jet}, the categorical cross entropy loss over the different jet flavours; L_{vertex}, the binary track-pair compatability cross entropy loss averaged over all track-pairs; and L_{track}, the categorical cross entropy loss for the track origin prediction. L_{vertex} is computed by averaging over all track-pairs in the batch, and L_{track} is computed by averaging over all tracks in the batch.

$$L_{\text{total}} = L_{\text{jet}} + \alpha L_{\text{vertex}} + \beta L_{\text{track}} \tag{4}$$

The different losses converge to different values during training, reflective of differences in the relative difficulty of the various objectives. As such, L_{vertex} and L_{track} are weighted by $\alpha = 1.5$ and $\beta = 0.5$ respectively to ensure they converge to similar values, giving them an equal weighting towards L_{total}. The values of α and β also ensure that L_{jet} converges to a larger value than L_{vertex} and L_{track}, reflecting the primary importance of the jet classification objective. In practice, the final performance of the model was not sensitive to modest variations in the loss weights α and β, or to pre-training using L_{total} and fine tuning on the jet classification task only. As there was a significant variation in the relative frequency of tracks of different origins, the contribution of each origin class to L_{track} was weighted by the inverse of the frequency of their occurrence. In L_{vertex}, the relative class weight in the loss for track-pairs where both tracks are from either a b- or c-hadron is increased by a factor of two as compared with other track-pairs.

The track classification and vertexing objectives are supplementary to the jet classification objective and trainings can be performed with either the node or edge networks, or both, removed, as discussed in Section 4.3. In these cases, the corresponding losses L_{vertex} and L_{track} are removed from the calculation of L_{total}. The resulting trainings demonstrate how useful the different auxiliary training objectives are for the primary jet classification objective.

GN1 trainings are run for 100 epochs on 4 NVIDIA V100 GPUs, taking around 25 mins to complete each epoch over the training sample of 30 million jets described in Section 2.3. The Adam optimiser [41] with an initial learning rate of 1e−3, and a batch size of 4000 jets (spread across the 4 GPUs) was used. Typically the validation loss, calculated on 500k jets, stabilised after around 60 epochs. The epoch that minimized the validation loss was used for evaluation. GN1 has been integrated into the ATLAS software [42] using ONNX [43], and jet flavour predictions for the test sample are computed using the ATLAS software stack.

4 Results

The performance of the GN1 tagger is evaluated for both b-tagging and c-tagging use cases, and for both jets with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ from the $t\bar{t}$ sample and jets with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$ from the Z' sample. Performance is compared to the DL1r tagger [10], which has been retrained on 75 million jets from the same samples as GN1. The input RNNIP tagger [13] to DL1r has not been retrained.

The taggers predict the probability that a jet belongs to the b-, c- and light-classes. To use the model for b-tagging, these probabilities are combined into a single score D_b, defined as

$$D_b = \log \frac{p_b}{(1 - f_c)p_l + f_c p_c}, \tag{5}$$

where f_c is a free parameter that determines the relative weight of p_c to p_l in the score D_b, controlling the trade-off between c- and light-jet rejection performance. This parameter is set to a value of $f_c = 0.018$ for the DL1r model, obtained through an optimisation procedure designed to maximise the c- and light-jet rejection of DL1r [10]. For the GN1 models a value of $f_c = 0.05$ is used, based on a similar optimisation procedure. The choice of f_c is arbitrary, with the different optimised values reflecting the relative c- versus light-jet rejection performance of the various taggers. A fixed-cut working point (WP) defines the corresponding selection applied to the tagging discriminant D_b in order to achieve a given inclusive efficiency on the $t\bar{t}$ sample.

The technical implementation of GN1 results in any jet with no associated tracks or exactly one associated track to be classified as a light-jet. The impact of this on the tagging performance of GN1 was found to be negligible, with 0.12% of b-jets in the $t\bar{t}$ sample and 0.02% of b-jets in the Z' sample affected. Of those, 89% of the b-jets in the $t\bar{t}$ sample and 98% of the b-jets in the Z' sample are classified as light-jets by DL1r at the 70% $t\bar{t}$ WP.

A comparison of the b-tagging discriminant D_b between DL1r and GN1 is given in Fig. 4. The shapes of the distributions are broadly similar for b-, c- and light-jets, however, the GN1 model shifts the b-jet distribution to higher values of D_b in the regions with the best discrimination. The GN1 c-jet distribution is also shifted to lower values of D_b when compared with DL1r, enhancing the separation and indicating that GN1 will improve c-jet rejection when compared with DL1r.

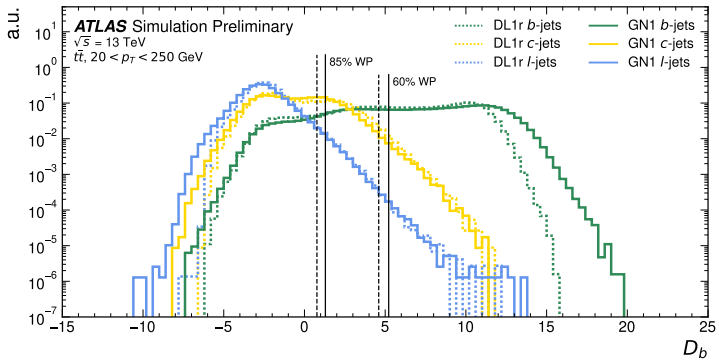

Figure 4: Comparison between the DL1r and GN1 b-tagging discriminant D_b for jets in the $t\bar{t}$ sample. The 85% working point (WP) and the 60% WP are marked by the solid (dashed) lines for GN1 (DL1r), representing respectively the loosest and tightest WPs used by analyses. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1. The distributions of the different jet flavours have been normalised to unity area.

4.1 b-tagging Performance

The performance of a b-tagging algorithm is quantified by its power to reject c- and light-jets for a given b-jet tagging efficiency, or WP. In order to compare the b-tagging performance of the different taggers for the b-jet tagging efficiencies in the range typically used by analyses, the corresponding c- and light-jet rejection rates are displayed in Figs. 5 and 6 for jets in the $t\bar{t}$ and Z' samples respectively. Four standard WPs with b-jet tagging efficiencies of 60%, 70%, 77% and 85% are used by physics analyses depending on their specific signal and background requirements. These WPs are defined using jets in the $t\bar{t}$ sample only. The b-jet tagging efficiencies for jets in the Z' sample are lower than the corresponding WPs calculated in the $t\bar{t}$ sample, due to the much higher jet p_{T} range in the Z' sample. For instance the WP defined to provide a 70% b-jet tagging efficiency on the $t\bar{t}$ sample results in a b-jet tagging efficiency of ∼30% on the Z' sample. To account for this, the range of b-jet tagging efficiencies displayed in Fig. 6 is chosen to span the lower values achieved in the Z' sample.

For jets in the $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$, GN1 demonstrates considerably better c- and light-jet rejection compared with DL1r across the full range of b-jet tagging efficiencies probed. The relative improvement depends on the b-jet tagging efficiency, with the largest improvements found at lower values. At a b-jet tagging efficiency of 70%, the c-jet rejection improves by a factor of ∼2.1 and the light-jet rejection improves by a factor of ∼1.8 with respect to DL1r. For high-p_{T} jets in the Z' sample with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$, GN1 also brings considerable performance improvements with respect to DL1r across the range of b-jet tagging efficiencies studied. Again, the largest relative improvement in performance comes at lower b-jet tagging efficiencies. At a b-jet tagging efficiency of 30%, GN1 improves the c-jet rejection by a factor of ∼2.8 and the light-jet rejection by a factor of ∼6. An increasing statistical uncertainty due to the high rejection of background affects the comparison at lower b-jet tagging efficiencies. It is estimated that for a b-jet tagging efficiency of 70% in the $t\bar{t}$ sample, ∼5% (∼30%) of the relative improvement in the c-jet (light-jet) rejection comes from loosening the track selection and for a b-jet tagging efficiency of 30% in the Z' the corresponding number is ∼10% for both c-jets and light-jets. Given the sophisticated exploitation of low-level information, further studies are needed to confirm if the performance gain is also observed in experimental data.

The GN1 Lep variant shows improved performance with respect to the baseline GN1 model, demonstrating

the additional jet flavour discrimination power provided by the leptonID track input. For jets in the $t\bar{t}$ sample, the relative c-jet rejection improvement with respect to DL1r at the 70% b-jet WP increases from a factor of ∼2.1 for GN1 to a factor of ∼2.8 for GN1 Lep. The improvement in light-jet rejection also increases from a factor of ∼1.8 to ∼2.5 at this WP. For jets in the Z' sample, the relative c-jet rejection (light-jet rejection) improvement with respect to DL1r increases from a factor of ∼2.8 to ∼3 (∼6 to ∼7.5) at a b-jet tagging efficiency of 30%. As shown in Fig. 7, the greatest improvement of GN1 Lep over GN1 is seen at low p_{T}.

The performance of the taggers is strongly dependent on the jet p_{T}. Charged particle reconstruction is particularly challenging within high-p_{T} jets [15]. The multiplicity of fragmentation particles increases as a function of p_{T}, while the number of particles from heavy flavour decays stays constant. Collimation of particles inside the jet increases and approaches the granularity of the tracking detectors, making it difficult to resolve the trajectories of different particles. Furthermore, at high p_{T}, heavy flavour hadrons will travel further into the detector before decaying. For hadrons which traverse one or more layers of the ID before decaying, the corresponding decay tracks may pick up incorrect hits, left by the hadron itself or fragmentation particles, in the inner layers of the detector, reducing the accuracy of the reconstructed track parameters. These factors contribute to a reduced reconstruction efficiency for heavy flavour tracks, and a general degradation in quality of tracks inside the core of a jet, which in turn reduces the jet classification performance.

In order to study how the b-jet tagging efficiency of the taggers varies as a function of jet p_{T}, the b-jet tagging efficiency as a function of p_{T} for a fixed light-jet rejection of 100 in each bin is shown in Fig. 7. For jets in the $t\bar{t}$ sample, at a fixed light-jet rejection of 100, GN1 improves the b-jet tagging efficiency by approximately 4% across all jet p_{T} bins. GN1 Lep shows improved performance with respect to GN1, in particular at lower p_{T}, with the relative increase in the b-jet tagging efficiency going from 4% to 8%. For jets in the Z' sample, GN1 has a higher b-jet tagging efficiency than DL1r across the p_{T} range, with the largest relative improvement in performance, approximately a factor of 2, found at jet $p_{\mathrm{T}} > 2\,\mathrm{TeV}$. GN1 outperforms DL1r across the entire jet p_{T} spectrum studied. The performance was also evaluated as a function of the average number of pileup interactions in an event, and was found to have no significant dependence on this quantity.

4.2 c-tagging Performance

Since GN1 does not rely on any manually optimised low-level tagging algorithms, which may not have been optimised for c-tagging, tagging c-jets presents a compelling use case for GN1. To use the model for c-tagging, the output probabilities are combined into a single score D_c, defined similarly to Eq. (5) as

$$D_c = \log \frac{p_c}{(1 - f_b)p_l + f_b p_b}. \tag{6}$$

A value of $f_b = 0.2$ is used for all models. Similar to Section 4.1, performance of the different taggers is compared by scanning through a range of c-jet tagging efficiencies and plotting the corresponding b- and light-jet rejection rates. As in Section 4.1, WPs are defined using jets in the $t\bar{t}$ sample. Standard c-jet tagging efficiency WPs are significantly lower in comparison with the b-tagging WPs in order to maintain reasonable b- and light-jet rejection rates. This is reflected in the range of c-jet tagging efficiencies used in Figs. 8 and 9. In Fig. 8, which displays the c-tagging performance of the models on the jets in the $t\bar{t}$ sample, GN1 performs significantly better than DL1r. The b- and light-jet rejection improve most at lower c-jet

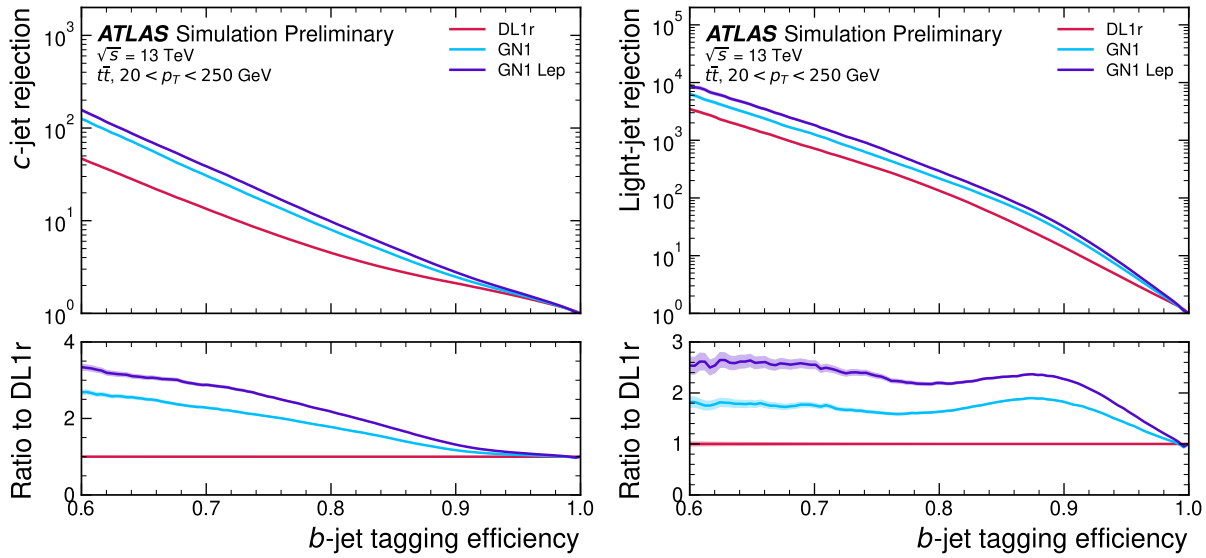

Figure 5: The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency for jets in the $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$. The ratio with respect to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. Binomial error bands are denoted by the shaded regions. At b-jet tagging efficiencies less than ~75%, the light-jet rejection becomes so large that the effect of the low number of jets is visible. The lower x-axis range is chosen to display the b-jet tagging efficiencies usually probed in these regions of phase space.

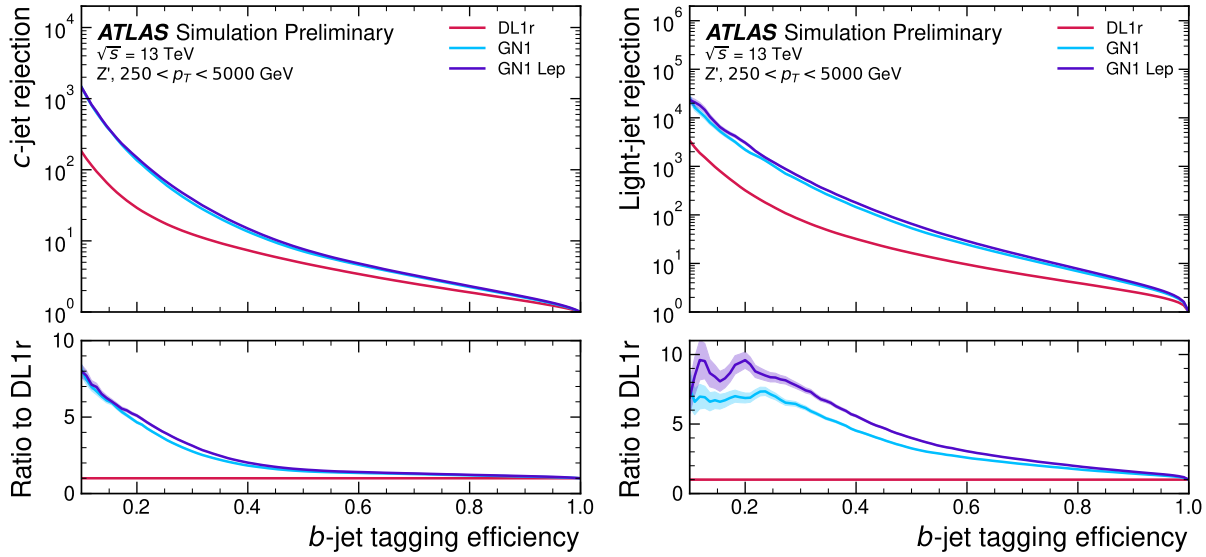

Figure 6: The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency for jets in the Z' sample with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$. The ratio with respect to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. Binomial error bands are denoted by the shaded regions. At b-jet tagging efficiencies less than ~20%, the light-jet rejection becomes so large that the effect of the low number of jets is visible. The lower x-axis range is chosen to display the b-jet tagging efficiencies usually probed in these regions of phase space.

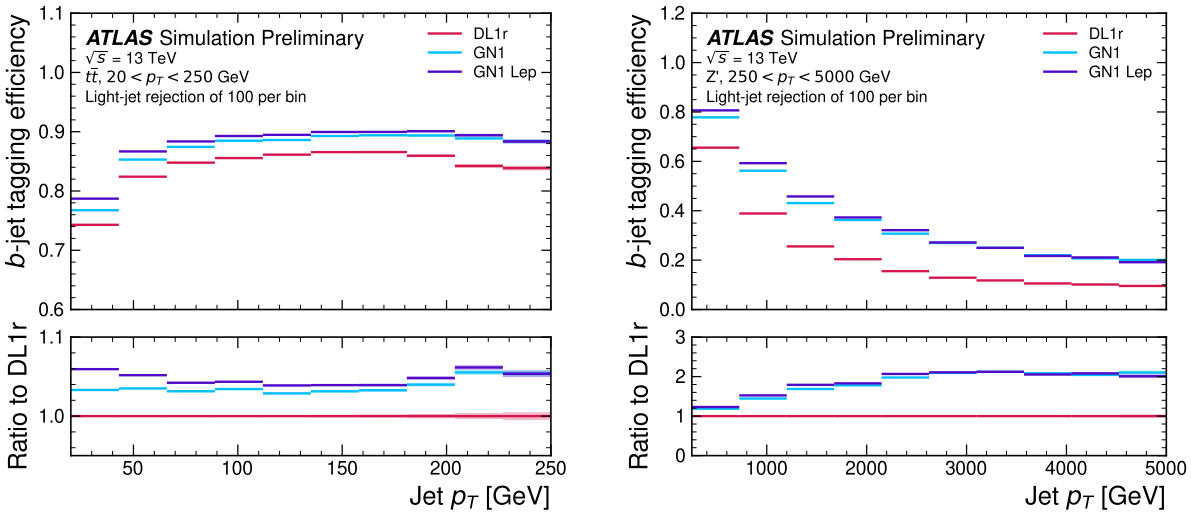

Figure 7: The b-jet tagging efficiency for jets in the $t\bar{t}$ sample (left) and jets in the Z' sample (right) as a function of jet p_{T} with a fixed light-jet rejection of 100 in each bin. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. Binomial error bands are denoted by the shaded regions.

tagging efficiencies, with both background rejections increasing by a factor of 2 with respect to DL1r at a c-jet tagging efficiency of 25%. GN1 Lep outperforms GN1, with the b-jet rejection (light-jet rejection) relative improvement increasing from a factor of 2 to 2.1 (2 to 2.3) at the 25% c-jet WP. Fig. 9 shows the c-tagging performance on the jets in the Z' sample. Both GN1 and GN1 Lep perform similarly, improving the b-jet rejection by 60% and the light-jet rejection by a factor of 2 at the 25% c-jet WP.

4.3 Ablations

Several ablations, the removal of components in the model to study their impact, are carried out to determine the importance of the auxiliary training objectives of GN1 to the overall performance. The "GN1 No Aux" variant retains the primary jet classification objective, but removes both track classification and vertexing auxiliary objectives (see Section 3.2) and as such only minimises the jet classification loss. The "GN1 TC" variant includes track classification but not vertexing, while "GN1 Vert" includes vertexing, but not track classification.

For jets in both the $t\bar{t}$ and Z' samples, the models without one or both of the auxiliary objectives display significantly reduced c- and light-jet rejection when compared with the baseline GN1 model, as shown in Figs. 10 and 11. For jets in the $t\bar{t}$ sample, the performance of GN1 No Aux is similar to DL1r, while GN1 TC and GN1 Vert perform similarly to each other. For jets in the Z' sample, the GN1 No Aux model shows a clear improvement in c- and light-jet rejection when compared with DL1r at lower b-jet tagging efficiencies. Similar to jets in the $t\bar{t}$ sample, GN1 TC and GN1 Vert perform similarly, and bring large gains in background rejection when compared with GN1 No Aux, but the combination of both auxiliary objectives yields the best performance.

It is notable that the GN1 No Aux model matches or exceeds the performance of DL1r without the need for inputs from the low-level algorithms. This indicates that the performance improvements enabled by GN1 appear to be able to compensate for the removal of the low-level algorithm inputs. The GN1 TC and

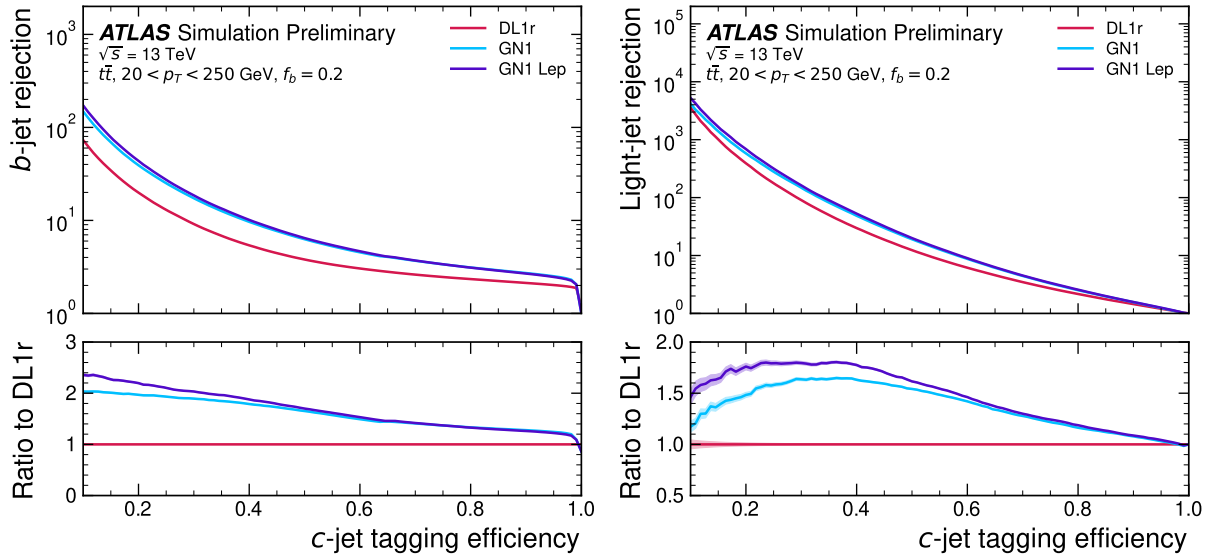

Figure 8: The b-jet (left) and light-jet (right) rejections as a function of the c-jet tagging efficiency for $t\bar{t}$ jets with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$. The ratio to the performance of the DL1r algorithm is shown in the bottom panels. Binomial error bands are denoted by the shaded regions. At c-jet tagging efficiencies than ~25%, the light-jet rejection becomes so large that the effect of the low number of jets is visible. The lower x-axis range is chosen to display the c-jet tagging efficiencies usually probed in these regions of phase space.

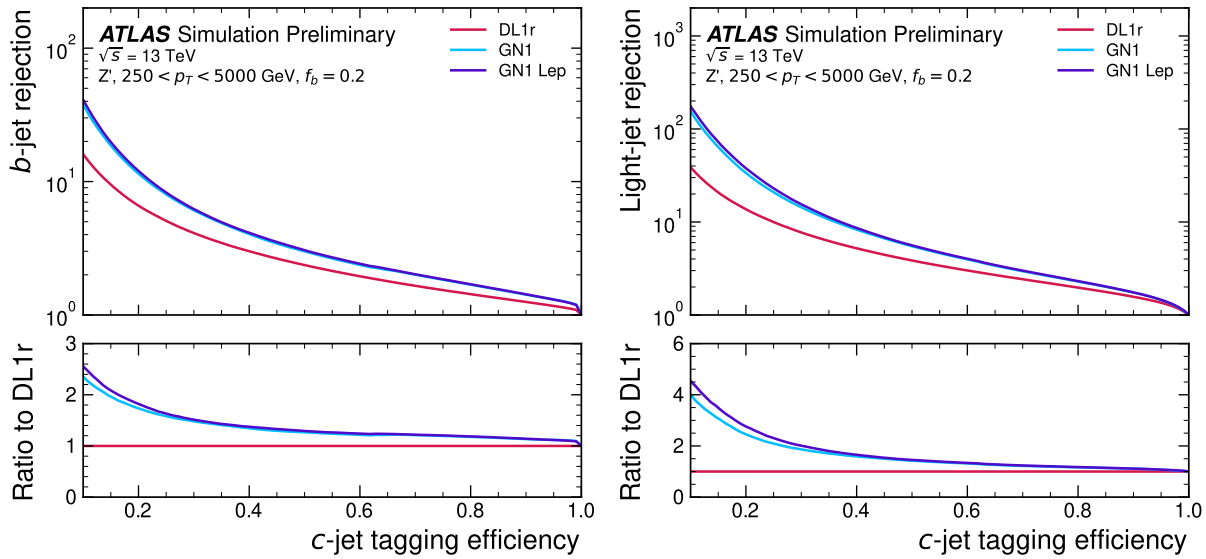

Figure 9: The b-jet (left) and light-jet (right) rejections as a function of the c-jet tagging efficiency for Z' jets with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$. The ratio to the performance of the DL1r algorithm is shown in the bottom panels. Binomial error bands are denoted by the shaded regions. The lower x-axis range is chosen to display the c-jet tagging efficiencies usually probed in these regions of phase space.

GN1 Vert variants each similarly outperform DL1r, demonstrating that both contribute to the overall high performance of the baseline model.

4.4 Vertexing Performance

From the track-pair vertex prediction described in Section 3.2, tracks can be partitioned into compatible groups representing vertices (see [9]). As such, GN1 is able to be used to perform vertex "finding", but not vertex "fitting", i.e. the reconstruction of a vertex's properties, which currently still requires the use of a dedicated vertex fitter. In order to study the performance of the different vertexing tools inside b-jets, the truth vertex label of the tracks, discussed in Section 3.2, are used. To estimate the efficiency with which GN1 manages to find vertices inclusively, vertices from GN1 containing tracks identified as coming from a b-hadron are merged together and compared to the inclusive truth decay vertices that result from a b-hadron decay (where if there are multiple distinct truth vertices from a b-hadron decay they are also merged together). Vertices are compared with the target truth vertex and the number of correctly and incorrectly assigned tracks is computed. Since secondary vertex information is only recovered for reconstructed tracks, an efficiency of 100% here denotes that all possible secondary vertices are recovered given the limited track reconstruction efficiency. A vertex is considered matched if it contains at least 65% of the tracks in the corresponding truth vertex, and has a purity of at least 50%. GN1 manages to achieve an inclusive reconstruction efficiency in b-jets of ∼80%, demonstrating that it effectively manages to identify the displaced vertices from b-hadron decays.

4.5 Track Classification Performance

As discussed in Section 3.2, one of the auxiliary training objectives for GN1 is to predict the truth origin of each track in the jet. Since the equivalent information is not provided by any of the existing flavour tagging tools, as a benchmark a multi-class classification multilayer perceptron (MLP) is trained on the same tracks used for the baseline GN1 training. The model uses the same concatenated track-and-jet inputs as GN1 (see Section 3.1), but processes only a single track at a time. The model is comprised of five densely connected layers with 200 neurons per layer, though the performance was not found to be strongly sensitive to changes in the network structure. To measure the track classification performance, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve is computed for each origin class using a one versus all classification approach. The AUCs for the different truth origin classes are averaged using both an unweighted and a weighted approach. The unweighted mean treats the performance of each class equally, while the weighted mean uses the fraction of tracks from each origin as a weight. As seen in Table 5, GN1 outperforms the MLP, both at $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ for jets in the $t\bar{t}$ sample, and at $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$ for jets in the Z' sample. For tracks in jets in the $t\bar{t}$ sample, GN1 can reject 65% of fake tracks while retaining more than 99% of good tracks. The GN1 model has two advantages over the MLP which can explain the performance improvement. Firstly, the mixing of information between tracks, enabled by the fully connected graph network architecture as discussed in Section 3.3, is likely to be beneficial since the origins of different tracks within a jet are to some extent correlated. Secondly, the jet classification and vertexing objectives can be considered auxiliary to the track classification task, and may bring improved track classification performance with respect to the standalone MLP.

Fig. 12 shows the track origin classification ROC curves for the different track origins for jets in both the $t\bar{t}$ and Z' samples. In order to improve legibility of the figure, the heavy flavour truth origins have been combined weighted by their relative abundance, as have the Primary and OtherSecondary labels. In

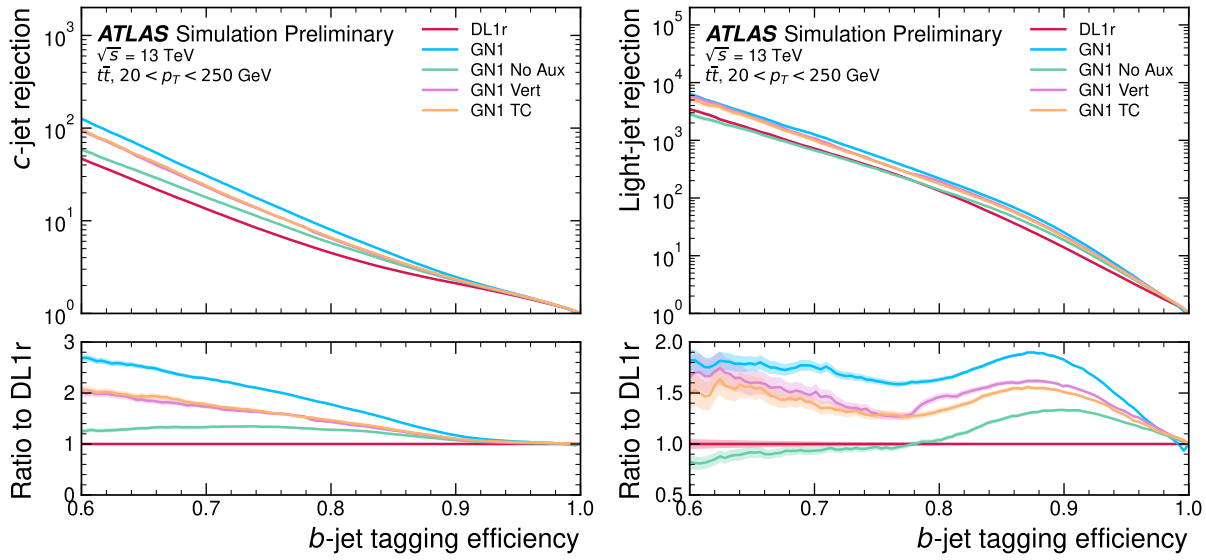

Figure 10: The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency for $t\bar{t}$ jets with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$, for the nominal GN1, in addition to configurations where no (GN1 No Aux), only the track classification (GN1 TC) or only the vertexing (GN1 Vert) auxiliary objectives are deployed. The ratio to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1. Binomial error bands are denoted by the shaded regions. At b-jet tagging efficiencies less than ~65%, the light-jet rejection become so large that the effect of the low number of jets are visible. The lower x-axis range is chosen to display the efficiencies usually probed in these regions of phase space.

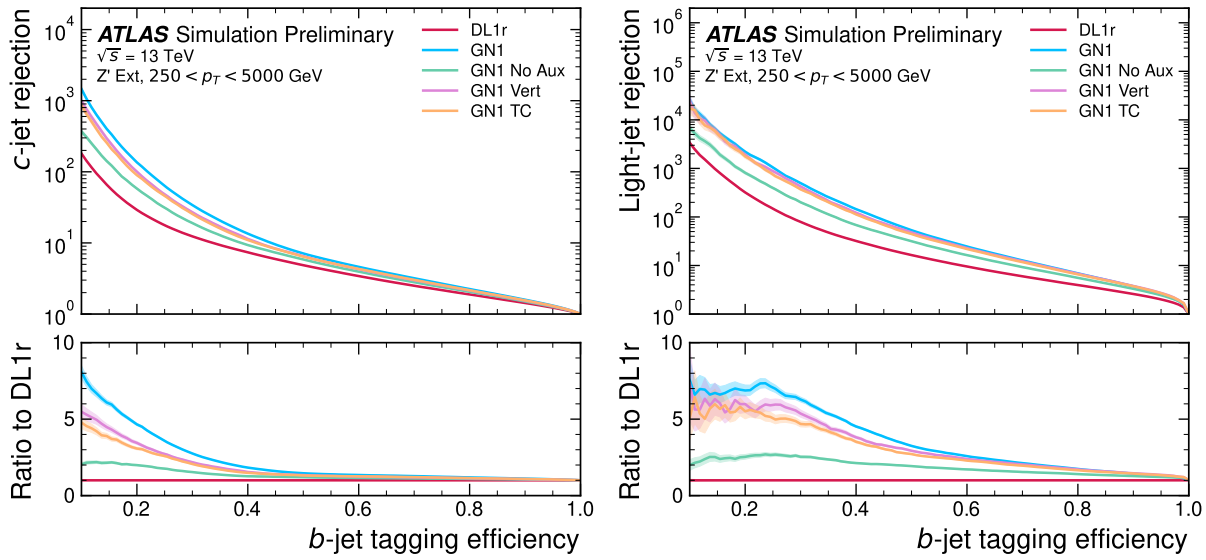

Figure 11: The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency for Z' jets with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$, for the nominal GN1, in addition to configurations where no (GN1 No Aux), only the track classification (GN1 TC) or only the vertexing (GN1 Vert) auxiliary objectives are deployed. The ratio to the performance of the DL1r algorithm is shown in the bottom panels. A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1. Binomial error bands are denoted by the shaded regions. At b-jet tagging efficiencies less than ~25%, the light-jet rejection become so large that the effect of the low number of jets are visible. The lower x-axis range is chosen to display the efficiencies usually probed in these regions of phase space.

Table 5: The area under the ROC curves (AUC) for the track classification from GN1, compared to a standard multilayer perceptron (MLP) trained on a per-track basis. The unweighted mean AUC over the origin classes and weighted mean AUC (using as a weight the fraction of tracks from the given origin) is provided. GN1, which uses an architecture that allows track origins to be classified in a conditional manner as discussed in Section 3.3, outperforms the MLP model for both $t\bar{t}$ and Z' jets.

		AUC	
		Mean	Weighted
$t\bar{t}$	MLP	0.87	0.89
	GN1	**0.92**	**0.95**
Z'	MLP	0.90	0.94
	GN1	**0.94**	**0.96**

jets in both the $t\bar{t}$ and Z' samples, the AUC of the different (grouped) origins is above 0.9, representing good classification performance. Fake tracks, followed by pileup tracks, are the easiest to classify in both samples.

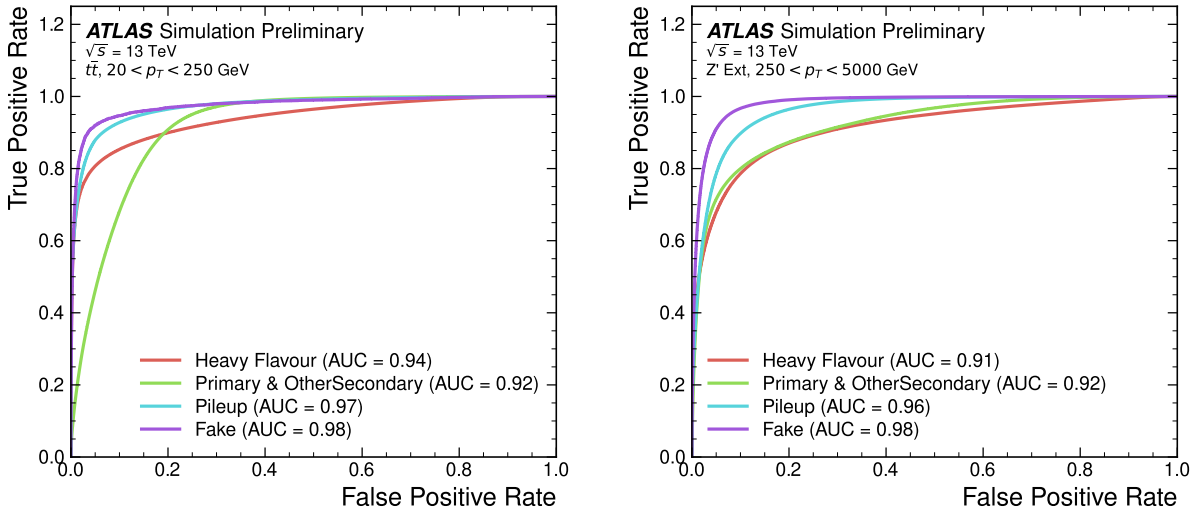

Figure 12: ROC curves for the different groups of truth origin labels defined in Table 3 for jets in the $t\bar{t}$ sample (left) and jets in the Z' sample (right). The FromB, FromBC and FromC labels have been combined, weighted by their relative abundance, into the Heavy Flavour category, and the Primary and OtherSecondary labels have similarly been combined into a single category. The mean weighted area under the ROC curves (AUC) is similar for both samples.

5 Conclusions

A novel jet tagger, GN1, with a graph neural network architecture and trained with auxiliary training targets, is presented and now fully implemented in the ATLAS software. GN1 is shown to improve flavour tagging performance with respect to DL1r, the current default ATLAS flavour tagging algorithm, when compared in simulated collisions. GN1 improves c- and light-jet rejection for jets in the $t\bar{t}$ sample with

$20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ by factors of ~2.1 and ~1.8 respectively at a b-jet tagging efficiency of 70% when compared with DL1r. For jets in the Z' sample with $250 < p_{\mathrm{T}} < 5000\,\mathrm{GeV}$, GN1 improves the c-jet rejection by a factor of ~2.8 and light-jet rejection by a factor of ~6 for a comparative b-jet efficiency of 30%. Previous multivariate flavour tagging algorithms relied on inputs from low-level tagging algorithms, whereas GN1 needs no such inputs, making it more flexible. It can be easily fully optimised via a retraining for specific flavour tagging use cases, as demonstrated with c-tagging and high-p_{T} b-tagging, without the need for time-consuming retuning of the low-level tagging algorithms. The model is also simpler to maintain and study due to the reduction of constituent components. GN1 demonstrates improved track classification performance when compared with a simple per-track MLP and an efficiency of ~80% for inclusive vertex finding in b-jets. The auxiliary track classification and vertex finding objectives are shown to significantly contribute to the performance in the jet classification objective, and are directly responsible for the improvement over DL1r. Further studies need to be undertaken to verify the performance of GN1 on collision data.

References

[1] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003 (cit. on pp. 2, 4).

[2] L. Evans and P. Bryant, *LHC Machine*, JINST **3** (2008) S08001 (cit. on p. 2).

[3] ATLAS Collaboration, *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*, Phys. Lett. B **786** (2018) 59, arXiv: 1808.08238 [hep-ex] (cit. on p. 2).

[4] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, Phys. Lett. B **784** (2018) 173, arXiv: 1806.00425 [hep-ex] (cit. on p. 2).

[5] ATLAS Collaboration, *Search for new resonances in mass distributions of jet pairs using* $139\,fb^{-1}$ *of pp collisions at* $\sqrt{s} = 13\,TeV$ *with the ATLAS detector*, JHEP **03** (2020) 145, arXiv: 1910.08447 [hep-ex] (cit. on p. 2).

[6] M. Tanabashi et al., *Review of Particle Physics*, Phys. Rev. D **98** (3 2018) 030001, URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001 (cit. on p. 2).

[7] P. W. Battaglia et al., *Relational inductive biases, deep learning, and graph networks*, arXiv preprint arXiv:1806.01261 (2018) (cit. on p. 2).

[8] J. Shlomi et al., *Secondary vertex finding in jets with neural networks*, The European Physical Journal C **81** (2021), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-021-09342-y (cit. on p. 2).

[9] H. Serviansky et al., *Set2Graph: Learning Graphs From Sets*, 2020, arXiv: 2002.08772 [cs.LG] (cit. on pp. 2, 8, 17).

[10] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013, 2017, URL: https://cds.cern.ch/record/2273281 (cit. on pp. 2, 11).

[11] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s}$ = 13 TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: 1907.05120 [hep-ex] (cit. on pp. 2, 3).

[12] ATLAS Collaboration, *Secondary vertex finding for jet flavour identification with the ATLAS detector*, ATL-PHYS-PUB-2017-011, 2017, URL: https://cds.cern.ch/record/2270366 (cit. on p. 3).

[13] ATLAS Collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, ATL-PHYS-PUB-2017-003, 2017, URL: https://cds.cern.ch/record/2255226 (cit. on pp. 3, 11).

[14] *Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS*, tech. rep., CERN, 2020, URL: https://cds.cern.ch/record/2718948 (cit. on pp. 3, 4).

[15] 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. 3–7, 13).

[16] ATLAS Collaboration, *ATLAS Insertable B-Layer: Technical Design Report*, ATLAS-TDR-19; CERN-LHCC-2010-013, 2010, URL: https://cds.cern.ch/record/1291633 (cit. on p. 4), Addendum: ATLAS-TDR-19-ADD-1; CERN-LHCC-2012-009, 2012, URL: https://cds.cern.ch/record/1451888.

[17] B. Abbott et al., *Production and integration of the ATLAS Insertable B-Layer*, JINST **13** (2018) T05008, arXiv: 1803.00844 [physics.ins-det] (cit. on p. 4).

[18] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: 1703.10485 [hep-ex] (cit. on p. 4).

[19] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: 0802.1189 [hep-ph] (cit. on p. 4).

[20] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 072002, arXiv: 1703.09665 [hep-ex] (cit. on p. 4).

[21] ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, ATLAS-CONF-2014-018, 2014, URL: https://cds.cern.ch/record/1700870 (cit. on p. 4).

[22] P. Nason, *A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms*, Journal of High Energy Physics **2004** (2004) 040, ISSN: 1029-8479, URL: http://dx.doi.org/10.1088/1126-6708/2004/11/040 (cit. on p. 5).

[23] S. Frixione, G. Ridolfi and P. Nason, *A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, Journal of High Energy Physics **2007** (2007) 126, ISSN: 1029-8479, URL: http://dx.doi.org/10.1088/1126-6708/2007/09/126 (cit. on p. 5).

[24] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, Journal of High Energy Physics **2007** (2007) 070, ISSN: 1029-8479, URL: http://dx.doi.org/10.1088/1126-6708/2007/11/070 (cit. on p. 5).

[25] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, Journal of High Energy Physics **2010** (2010), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP06(2010)043 (cit. on p. 5).

[26] R. D. Ball et al., *Parton distributions for the LHC run II*, JHEP **04** (2015) 040, arXiv: 1410.8849 [hep-ph] (cit. on p. 5).

[27] ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling for Top2016*, ATL-PHYS-PUB-2016-020, 2016, URL: https://cds.cern.ch/record/2216168 (cit. on p. 5).

[28] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph] (cit. on p. 5).

[29] 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. 5).

[30] R. D. Ball et al., *Parton distributions with LHC data*, Nucl. Phys. B **867** (2013) 244, arXiv: 1207.1303 [hep-ph] (cit. on p. 5).

[31] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. A **462** (2001) 152 (cit. on p. 5).

[32] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv: 1005.4568 [physics.ins-det] (cit. on p. 5).

[33] GEANT4 Collaboration, S. Agostinelli et al., *GEANT4 – a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250 (cit. on p. 5).

[34] ATLAS Collaboration, *Performance of b-jet identification in the ATLAS experiment*, JINST **11** (2016) P04008, arXiv: 1512.01094 [hep-ex] (cit. on p. 6).

[35] ATLAS Collaboration, *Muon reconstruction performance in early $\sqrt{s}$ = 13 TeV data*, ATL-PHYS-PUB-2015-037, 2015, URL: https://cds.cern.ch/record/2047831 (cit. on p. 6).

[36] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s}$ = 13 TeV*, Eur. Phys. J. C **79** (2019) 639, arXiv: 1902.04655 [hep-ex] (cit. on p. 6).

[37] D. Hwang et al., *Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs*, 2020, URL: https://arxiv.org/abs/2007.08294 (cit. on p. 6).

[38] J. Shlomi, P. Battaglia and J.-R. Vlimant, *Graph neural networks in particle physics*, Machine Learning: Science and Technology **2** (2021) 021001, ISSN: 2632-2153, URL: http://dx.doi.org/10.1088/2632-2153/abbf9a (cit. on p. 8).

[39] S. Brody, U. Alon and E. Yahav, *How Attentive are Graph Attention Networks?*, arXiv e-prints, arXiv:2105.14491 (2021) arXiv:2105.14491, arXiv: 2105.14491 [cs.LG] (cit. on p. 8).

[40] M. Zaheer et al., *Deep Sets*, 2018, arXiv: 1703.06114 [cs.LG] (cit. on p. 8).

[41] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, (2014), URL: https://arxiv.org/abs/1412.6980 (cit. on p. 11).

[42] *The ATLAS Collaboration Software and Firmware*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-SOFT-PUB-2021-001: CERN, 2021, URL: https://cds.cern.ch/record/2767187 (cit. on p. 11).

[43] J. Bai, F. Lu, K. Zhang et al., *ONNX: Open Neural Network Exchange*, https://github.com/onnx/onnx, 2019 (cit. on p. 11).

Appendix

A b-tagging Peformance

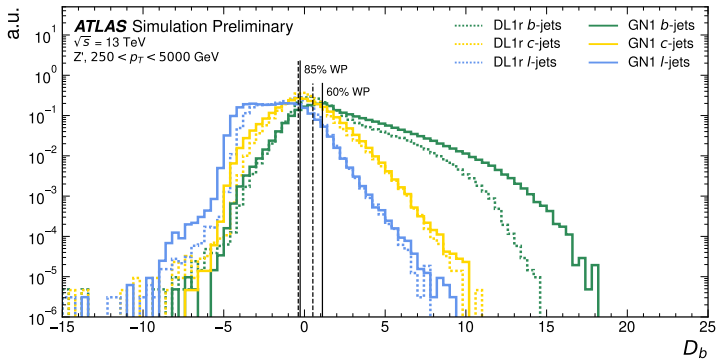

Figure 13: Comparison between the DL1r and GN1 b-tagging discriminant D_b for jets in the Z' sample. The 85% working point (WP) and the 60% WP are marked by the solid (dashed) lines for GN1 (DL1r). A value of $f_c = 0.018$ is used in the calculation of D_b for DL1r and $f_c = 0.05$ is used for GN1 and GN1 Lep. The distributions of the different jet flavours have been normalised to unity area.

B c-tagging Peformance

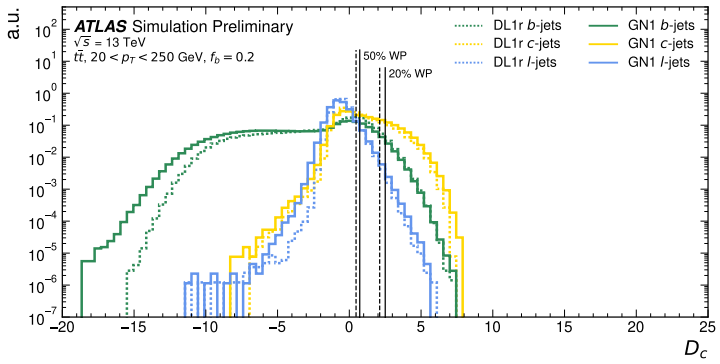

Figure 14: Comparison between the DL1r and GN1 c-tagging discriminant D_c for jets in the $t\bar{t}$ sample. The 50% working point (WP) and the 20% WP are marked by the solid (dashed) lines for GN1 (DL1r). The distributions of the different jet flavours have been normalised to unity area.

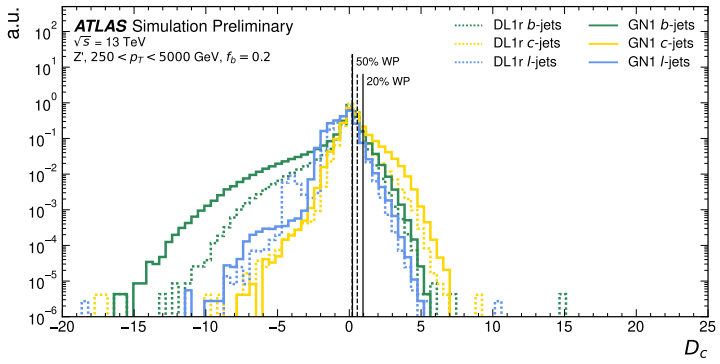

Figure 15: Comparison between the DL1r and GN1 c-tagging discriminant D_c for jets in the Z' sample. The 50% working point (WP) and the 20% WP are marked by the solid (dashed) lines for GN1 (DL1r). The distributions of the different jet flavours have been normalised to unity area.