Graph Neural Network Flavour Tagging and Boosted Higgs Measurements at the LHC

Samuel John Van Stroud
University College London

Submitted to University College London in fulfilment
of the requirements for the award of the
degree of **Doctor of Philosophy**

July 1, 2022

Declaration

I, Samuel John Van Stroud confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Samuel Van Stroud

Abstract

Here some useful packages are demonstrated. In particular, the `hepunit` package which adds additional units to `SIUnit`. A variety of jet measurements are made using data collected during the first year of 7 TeV proton-proton collisions from the general-purpose ATLAS experiment at the LHC. no more than 300 words

Impact Statement

impact statement 500 words link to ucl info

Acknowledgements

Here is an example of how to declare commands for use in a single file that will not be needed elsewhere. Additionally, it serves to illustrate the chapter referencing system.

Perhaps you might want to point out that Peter Higgs provided helpful advice for Chapter 1.

Preface

blah this is 300 TeV in text mode. this is 300 TeV in math mode.

Contents

1

Chapter 1

Theoretical Framework

- Introduce sm
- brief history
- current areas of study
- Reference relevenace to rest of thesis (studying hbb)

The Standard Model (SM) of particle physics is the theory describing all known elementary particles and their interactions via three of the four fundamental forces. Developed by merging the successful theories of classical quantum mechanics and relativity in the second half of the 20th century, the SM's position today at the centre of our understanding of the nature of the universe is firmly established by an unparalleled level of agreement between the predictions from the model and experimental results [1, 2].

The SM has predicted the discovery of the top and bottom quarks [3–5], the W and Z bosons [6], and the tau neutrino [7]. The last missing piece of the SM to be discovered was the Higgs boson, first posited in **X**. After its discovery in 2012 **citation**, much work has been ongoing on carrying out detailed measurements of its mass and interactions with other particles.

This thesis looks at understanding Higgs decays. . .

1.1 The Standard Model

- Introduce QFT
- Introduce SM Gauge symmetry
- List Contents of SM (different particles) masses and charges
- Write SM Lagrangian term break up LEW etc
- Walk through (or subsection) for each term

The SM is formulated in the language of Quantum Field Theory (QFT). In this framework, particles are localised excitations of corresponding quantum fields, which are operator-valued distribution across spacetime.

Central to QFT is the Lagrangian density which describes the kinematics and dynamics of the fields. Observations of conserved quantities are linked, via Noether's theorem, to symmetries which are expressed by the Lagrangian. Alongside Global Poincaré symmetry, the SM Lagrangian observes a local $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ gauge symmetry. The presence of gauge symmetries allows certain gauge transformations to be applied to fields, with the result that observable properties of the system are unchanged.

1.1.1 Quantum Electrodynamics

- Dirac equation, Lagrangian
- U(1) symmetry, transformation
- Follow through and interpretation of fields, conservation of electric charge

Consider a Dirac spinor field $\psi = \psi(x)$ and its adjoint $\overline{\psi} = \psi^\dagger\gamma^0$, where $\psi^\dagger$ is the Hermitian conjugate of ψ. The Dirac Lagrangian density is

$$\mathcal{L}_{\text{Dirac}} = \overline{\psi}(i\not{\partial} - m)\psi, \tag{1.1}$$

where $\partial\!\!\!/ = \gamma^\mu \partial_\mu$ denotes the contraction with the Dirac gamma matrices γ^μ, and summation over up-down pairs of indices is assumed. Application of the Euler-Lagrange equation on eq. (1.1) yields the Dirac equation

$$(i\partial\!\!\!/ - m)\psi = 0. \tag{1.2}$$

Suppose some fundamental symmetry that requires invariance under a $U(1)$ local gauge transformation

$$\psi \to \psi' = \psi e^{iq\alpha(x)}, \tag{1.3}$$

where α varies over every spacetime point x. Under this transformation, the Dirac equation transforms as

$$(i\partial\!\!\!/ - q\partial\!\!\!/\alpha(x) - m)\psi = 0. \tag{1.4}$$

For the Dirac equation to remain invariant under the transformation in eq. (1.3), a new field A_μ, which transforms as $A_\mu \to A'_\mu - q\partial_\mu\alpha$ must be added to the Dirac equation.

yielding the QED Lagrangain

$$\mathcal{L}_{\text{QED}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \overline{\psi}(iD\!\!\!\!/ - m)\psi. \tag{1.5}$$

1.1.2 Quantum Chromodynamics

1.1.3 The Electroweak Sector

The $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ is spontaneously broken to $SU(3)_C \otimes U(1)_\gamma$.

1.2 The Higgs Mechanism

- Motivation
- Walkthrough

1.2.1 Electroweak Symmetry Breaking

1.2.2 Fermionic Yukawa Coupling

Chapter 2

The Large Hadron Collider and the ATLAS Detector

2.1 Overview

The Large Hadron Collider (LHC) at CERN has extended the frontiers of particle physics through its unprecedented energy and luminosity. In 2010, the LHC collided proton bunches, each containing more than 10^{11} particles, 20 million times per second, providing 7 TeV proton-proton collisions at instantaneous luminosities of up to $2.1 \times 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

2.1.1 The ATLAS Detector

The ATLAS detector at the LHC covers nearly the entire solid angle around the collision point.[1] 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.

[1] 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 $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

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 [8, 9]. 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. [10].

2.2 Trigger system

An LHCb trigger table borrowed from `hepthesis` is shown in Table 2.1:

	L0	L1	HLT
Input rate	40 MHz	1 MHz	40 kHz
Output rate	1 MHz	40 kHz	2 kHz
Location	On detector	Counting room	Counting room

Table 2.1: Characteristics of the trigger levels and offline analysis.

2.3 Reconstructed Physics Objects

2.3.1 Tracks

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. [11] and outlined in table 2.2, 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 [12]. 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, $\sum {p_{\mathrm{T}}}^2$.

Table 2.2: 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 [12]. 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

2.3.2 Jets

Jets are reconstructed from particle-flow objects [13] using the anti-k_T algorithm [14] with a radius parameter of 0.4. The jet energy scale is calibrated according to Ref. [15]. 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 [16]. 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 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.

- Jet finding algorithms

2.3.3 Leptons

Chapter 3

Investigating Tracking Improvements

Todo:

- Check all info wrt to this PDG review

3.1 *b*-hadron Reconstruction

3.1.1 *b*-hadron Decay Topology

b-hadrons are quasi-stable bound states of quarks, where one of the quarks is a bottom quark (b quark). The proper lifetimes τ of the various b-hadrons are similar and relatively long, with $\tau{\sim}10^{-12}$s. This lifetime corresponds to a proper decay length $c\tau{\sim}300$ μm. In the rest frame of the detector, the typical b-hadron travels a distance $d = \beta\gamma c\tau$ before decaying, where at high energies $\gamma{\sim}E_B/m_B$. For a 1 TeV b-hadron, this gives $d{\sim}60$ mm - well beyond the radius of the first pixel layer (IBL) at 33 mm. At the LHC, b quarks are generated in the hard scattering of proton-proton (pp) collisions. They quickly hadronize into a b-hadron, which is often initially in an excited state due to the high energies of the pp collisions at the LHC ($\sqrt{s}$ = 13 TeV). The hadronisation process is hard - around 70-80% of the b quark's momentum goes into the b-hadron, with the rest being radiated as other particles. The excited b-hadron will quickly fragment (i.e. de-excite) by radiating particles, which are prompt (they are formed closed to the primary vertex). These fragmentation particles have an increasing multiplicity and collimation to the b-hadron axis as the p_{T} of the b-hadron increases. The de-excited b-hadron subsequently weakly decays to on average 4 or 5 particles (the multiplicity of the decay products of the weak decay of the b-hadron is unaffected by increases in the b-hadron p_{T}.).

Due to their lifetimes, energetic b-hadrons can travel a significant distance from the primary pp interaction point before decaying to a spray of collimated stable particles. This signature is registered in the detector as a displaced jet. Due to the elements of the CKM matrix, b-hadrons decay with a high probability to D hadrons (which contain a c quark), which also have significant lifetimes - this can lead to reconstructed tertiary vertices in the jet core. The typical features of a b-jet, and in particular the large track impact parameter d_0 which can result from displaced decays, are shown in fig. 3.1. Many ATLAS analyses rely on a method of tagging jets instantiated by b quarks and rejecting jets created from other quarks (c and light flavours u, d, s). These "b-tagging" algorithms work by discriminating against the unique signatures of b-jets discussed above. b-tagging relies on the efficient and accurate reconstruction the tracks corresponding to the b-hadron decay products. These tracks are then used as inputs to vertex reconstruction algorithms and jet making algorithms.

3.1.2 *b*-hadron Decay Track Reconstruction

A necessary requirement for successful jet b-tagging is the efficient and accurate reconstruction of the charged particle trajectories in the jet. For high p_{T} jets (p_{T}

chapters/3.tracking/figs/b-jet-diagram.png

Figure 3.1: Diagram of a typical b-jet (blue) which has been produced along with two light jets (grey). The b-hadron has travelled a significant distance from the primary interaction point (pink dot) before its decay. The large transverse impact parameter d_0 is a characteristic property of the trajectories of b-hadron decay products.

chapters/3.tracking/figs/high-pt-b-tracks.png

Figure 3.2: As b-hadron p_{T} increases, the time of flight of the B increases, so tracks will have less room to diverge before reaching detector elements. To compound the problem, the collimation of the tracks increases. The detector may then be unable to resolve individual tracks.

$>$ 200 GeV) this task becomes difficult due to a combination of effects. As the jet energy increases, the track multiplicity of the jet increases due to the presence of additional fragmentation tracks. Tracks in the jet also become increasingly collimated as their inherited transverse momentum increases. Together, these two effects lead to a very high density of charged particles in the jet core, making reconstruction difficult. At high energies, the increased decay length of B (and D) hadrons means that decay products have less of an opportunity to diverge before reaching the first tracking layers of the detector. If the decay takes place very close to a detector layer, or if the decays are sufficiently collimated, hits left by nearby particles may not be resolved individually, leading to merged clusters (shown in fig. 3.2). Shared hits generally predict bad tracks. As such, shared hits are heavily penalised during reconstruction (and in particular as part of ambiguity solving). However, in the core of high p_{T} b-jets, where decay particles are displaced from the primary vertex and are highly collimated, the density of particles is high enough that the probability of clusters being merged increases dramatically. The presence of merged clusters requires that the corresponding tracks share hits (if they are to be reconstructed successfully), which may end up impairing the successfully reconstruction of the track. Furthermore, decays may also take place inside the tracking detectors themselves, which can lead to missing or wrong innermost cluster assignment. The combination of effects described above makes reconstructing tracks in the core of high p_{T} b-jets particularly challenging.

chapters/3.tracking/figs/overlay_po_nHitsOnIBL_Frac_B_pT.pdf	chapters/3.tracking/figs/overlay_po_nHitsOnPix_F
a	**b**

Figure 3.3: Hit multiplicities on the IBL (fig. 3.3a) and the all pixel layers (fig. 3.3b) as a function of the transverse momentum p_{T} of the reconstructed track. Tracks from the weak decay of the b-hadron are shown in red, while fragmentation tracks (which are prompt) are in blue. For each of these, standard tracks and pseudo-tracks are plotted. Hit multiplicities on the pseudo-tracks at high p_{T} due to the increased flight of the b-hadron. The baseline tracks have more hits than the pseudo-tracks, indicating that they are being incorrectly assigned additional hits.

Figure 3.4: Track reconstruction efficiency from b-hadron decay products for baseline ATLAS tracking (black), Bcut+Refit procedures applied (green), pseudo-tracking (blue), and for tracking where the ambiguity solver has been manually removed (orange).

Figure 3.5: The total number of pixel hits on tracks from b-hadron decays as a function of the production radius of the decay product. An excess of hits is assigned to the standard tracks in comparison to the ideal pseudo-tracks.

Concretely, then, the issues relating to high p_{T} b-hadron tracking can be factorised into two parts. The first part is a drop in track reconstruction efficiency. As mentioned, tracks originating from high energy b-hadron decay products can have a high rate of shared hits due to the number of particles present in a high p_{T} b-jet and their relative collimation. Additionally, tracks may be missing hits on the inner layers of the detector. This occurs primarily when the decay b-hadron decays inside the detector. These features of can make it difficult for B decay tracks to meet the ambiguity solver's stringent track quality requirements. As a result, many B decay tracks are rejected in the ambiguity solving stage, leading to a severe drop in tracking reconstruction efficiency. This is shown by the severe decrease in reconstruction efficiency visible when comparing baseline tracking with the ideal pseudo-tracks in fig. 3.4. This situation presents a problem: relaxing cuts on shared hits significantly degrades the ambiguity solver's power to reject bad tracks. However for b-hadron decay tracks it seems these same restrictions on shared hits are seriously impairing the reconstruction efficiency of good tracks. The second part of the problem is that, due to the high density of clusters available for assignment in the vicinity of the typical high energy b-hadron decay track, and also given the strong positive bias of the ambiguity solver towards those tracks with precise pixel measurements (especially the innermost IBL measurement), many b-hadron decay tracks are assigned incorrect inner layer hits. This is only a problem for those decay products which were produced inside the pixel detector as a result of a long-flying b-hadron, and so do not have a correct hit available for assignment (evidenced in fig. 3.8b). The incorrect hits may skew the parameters of the track, which can in turn mislead b-tagging algorithms. In particular, b-tagging algorithms rely heavily on the transverse impact parameter significance $d_0/\sigma(d_0)$ of the track. The quality of this measurement is expected to be adversely affected by wrong inner-layer hits on the track. This combination of reduced reconstruction efficiency and incorrectly assigned hits is thought to be the cause of the observed drop in b-tagging efficiency at high energies , although it is not clear which effect may dominate.

3.2 Pseudotracks and Ideal Tracks

Pseudotracking and ideal tracking are used as benchmarks of the best tracking possible given the ATLAS detector. Both pseudotracks and ideal tracks are constructed using truth information to group combinations of hits that have been left by the same truth particle. As a result, hit-to-track association and track reconstruction efficiency are both ideal (given the ATLAS detector). Ideal tracks represent a yet more idealised tracking scenario by correcting the cluster positions based on truth information, and smearing the cluster position based on the detector resolution.

When pseudotracking is run alongside standard tracking, those clusters which are shared on the reconstructed tracks run through the cluster splitting machinery. If a cluster is found to be compatible with being split, its definition is changed, and the pseudotracks use this definition too. As a result, pseudotracks can have split clusters.

3.3 Investigating Improvements for High p_{T} B Tracking

An investigation into

3.3.1 Looser Track Cuts & Track Refit Procedure

A solution for the problem of wrong inner-layer hits on B tracks had previously previously been developed. This solution selects tracks which pass a b-jet Region of Interest (ROI) selection, and then removes the innermost hits on these tracks based on the result of a "refit" procedure. The refit procedure runs as follows. Each track is refitted without the innermost hit, and if there is a significant improvement in the fit quality (the χ^2 of the track fit divided by the number of degrees of freedom on the track n), the innermost hit is rejected and the new track is replaces the old. If the fit quality does not improve by a certain amount, the initial track is kept. This procedure is recursively applied. The b-jet ROI selection selects tracks that are matched within $dR < 0.14$ ($|\eta| < 0.1$, $|\phi| < 0.1$) of a CaloCluster with $E_T > 150$ GeV. The track itself must also pass a transverse momentum cut with $p_{\mathrm{T}} > 15$ GeV. The refit procedure was previously shown to lead to a reduction in the rate of wrongly assigned IBL hits on B decay tracks (see fig. 3.8b). However, this apparent improvement did not lead to an increase in b-tagging performance. It was found that the refit procedure also removed unacceptable numbers of good hits, degrading the quality of un-problematic tracks, shown in fig. 3.8a. This is likely the cause of the underwhelming b-tagging performance improvement.

The performance of both the ROI, and the hit removal using track fit information, is examined, and an attempt at improving the performance of the refit procedure is made. Results are discussed in the following two sections.

3.3.2 Region of Interest Optimisation

Selection cuts for the b-jet ROI were determined on a largely ad-hoc basis. An effort was made to systematically optimise the selection cuts. The decay tracks of B hadrons are tightly collimated with the B itself, with most decay products satisfying $dR(B, \text{track}) < 0.02$, as shown in fig. 3.6a. Meanwhile, calorimeter clusters relating to the B hadrons are generally found within $dR < 0.05$ of the B fig. 3.6b. In total, then, B decay tracks will usually be found within $dR < 0.07$ of the relevant calorimeter cluster, which suggests that the current $dR < 0.14$ is loose by a factor of two. Similar analysis of cluster and track energy distributions found that the related cuts were also loose, and so they were modified from $E_T > 150$ GeV to $E_T > 300$ GeV, and from $p_T > 15$ GeV to $p_T > 30$ GeV.

Additionally examined in the course of this work was the fake rate of the b-jet ROI. The distributions in fig. 3.7a demonstrate that most of clusters passing the $E_T > 150$ GeV selection were unable to be matched to a nearby B hadron using truth information. Clusters that pass the selection but do not correspond to energy depositions from B hadrons lead to fake ROIs. As a consequence of these distributions, tracks selected by the ROI are largely impure in the desired B hadron tracks.

The modified ROI was used to re-run the refit procedure. A comparison of of "standard" and "optimised" (using the optimised b-jet ROI) refit procedures is found in fig. 3.8. These results show that whilst tighter selection cuts did lead to a recovery of some good hits (fig. 3.8a), performance with respect to the baseline is still significantly degraded.

a **b**

Figure 3.6: Distributions of angular distance dR between B hadrons and their weak decays and other fragmentation tracks (fig. 3.6a), and the distribution of angular distance dR between B hadrons and the calorimeter clusters in the hadronic calorimeter (fig. 3.6b). In fig. 3.6a, the tracks from the weak decay of the B are significantly more collimated to the B than the other fragmentation tracks.

3.3.3 Fit Quality as a Discriminant for Wrong Hits

As mentioned, tracks selected by the ROI are refitted without their innermost hit, and, if an improvement in fit quality is observed, the hit is rejected. In order to test the effectiveness of this procedure, a dataset of two sets of tracks was produced. The first set contained unmodified baseline-reconstructed tracks. The second contained the same tracks as the first, but modifications made during reconstruction removed the innermost hit on each track. Then, using Monte Carlo (MC) truth information, a track-by-track fit quality comparison was made for tracks with good and wrong innermost hits.

It is clear from the distributions in fig. 3.7b that the fit quality improvement (measured by fractional change in χ^2/n of the track before and after the innermost hit is removed) is not a discriminating variable for wrong hits, and indeed attempted optimisations of the of the refit procedure based on these distributions were found to be ineffectual. While wrong hits are likely to degrade the track fit, it is also true that any additional measurement, good or wrong, constrains the track, and therefore removal of that measurement will be likely to lead to an increase in the χ^2/n of the track. Removing hits in this way is therefore problematic.

a **b**

Figure 3.7: The distribution of cluster transverse momentum, in fig. 3.7a for both clusters that were able (orange) and unable (blue) to be matched to a B hadron using MC truth information. The normalisation shows that the majority of clusters are not matched to B hadrons, resulting in fake ROIs. In fig. 3.7b, the fractional improvement in track fit quality (χ^2/n) is shown for all track (blue), tracks with good IBL hits (green), and tracks with wrong IBL hits (orange). The distributions are overlapping, suggesting that the χ^2/n improvement is not a good discriminator of good and wrong hits.

3.3.4 Conclusion

The work outlined in the two preceding sections has uncovered issues with both the b-jet ROI, and the methodology of identification and removal of wrong hits on tracks inside a given ROI. Attempts were made to optimise the selection cuts of the ROI, however the large background of energetic phenomena produced in collisions that are not B hadron related means that the ROI is largely unsuccessful in selecting a pure sample of likely B hadron candidates. An additional effort was made to improve the removal of wrong hits using other information in addition to the track fit improvement. Information such as the type and locations of its, and track d_0 were considered. While progress here was not insignificant, without substantial overhaul of the ROI to improve B purity, the results were not strong enough to demonstrate any viable solutions that would successful target and then improve B hadron decay tracks. Alongside the refit procedure, a "Bcut" cut scheme was suggested in order

a **b**

Figure 3.8: Distributions of good (fig. 3.8a) and wrong (fig:refit optimisation results sub2) hit assignment rates on the IBL for tracks using baseline tracking (black), the original unmodified refit procedure (green), and the refit procedure with an optimise set of ROI selection cuts (blue). The IBL lies at a radius of 33 mm from the beam pipe. Hence, particles produced with a production radius greater than this cannot leave good hits on the IBL.

to improve reconstruction performance. This consisted primarily of loosening the shared hit cuts in the ambiguity solver. While this did lead to a measurement increase in track reconstruction efficiency (see fig. 3.4), it was determined that the corresponding increase in fake tracks (i.e. those tracks for which the majority of hits do not come from a single truth particle) was too large to justify the implementation of the "Bcut" scheme. In conclusion, then, a different approach is required to address the problems discussed.

3.4 Global χ^2 Fitter Outlier Removal

This section documents ongoing progress into improving hit assignments using the Global χ^2 Fitter (GX2F) to prevent wrong hits from being assigned to tracks during the track fit. This is in contrast to the approach discussed in cref sec:refit, which attempts to identify and remove wrong hits after the reconstruction of the track (of which the track fit is a part). As part of the track fit, an outlier removal procedure is run, in which suspicious hits are indentified and removed. The GX2F code, as a relatively low-level component of track reconstruction, has not undergone significant modification for several years. During this time, a new tracking sub-detector, the

IBL, was installed, and subsequently precise detector alignments have been derived. The motivation for looking at the GX2F is that these changes may require re-optimisation of the GX2F code, and in particular the outlier removal procedures. Further motivation for this approach comes from the low rate of labelled outliers in baseline tracking. For example, while approximately 15% of B hadron decay tracks have a wrong IBL hit (a value which only increases with the p_{T} of the B), less than 1% of this tracks have had their IBL hit labelled and removed as an outlier.

Implementation

The outlier removal procedure for the pixel detector is described in this section. The states (also called measurements, or hits) on the track are looped over in order of increasing radial distance to the beam pipe. For each state, errors $\sigma(m_i)$ on the measurement of the transverse and longitudinal coordinates are calculated. These errors are dependent on the sub-detector which recorded the measurement (as some sub-detectors are more precise than others). Additionally, a residual displacement r_i between the predicted position of the track x_i (inclusive of the current measurement), and the position of the measurement itself, m_i, is calculated. The pull p_i on the track state due to the current measurement is calculated according to

$$p_i = \frac{m_i - x_i}{\sqrt{\sigma(m_i)^2 - \sigma(x_i)^2}} \ , \qquad r_i = m_i - x_i. \tag{3.1}$$

This pull is computed for the transverse and longitudinal coordinates of the measurement, and the maximum of the two is selected and checked to see if it exceeds a certain threshold. If it does, the hit will be removed, after some additional checks are made to confirm or deny the presence of the outlier. The threshold is set as a member variable `m_outlcut`. The results of varying this cut are described in section 3.4.1.

3.4.1 Cut Optimisation

A systematic variation of the cut point `m_outlcut` has been carried out. The results, demonstrating a reduction in wrong hit assignment whist keeping virtually all good hits assigned to tracks, are shown in fig. 3.9. The rate of wrong hits assigned to tracks decreases from 0.32 to 0.28 at the highest energies (12.5% reduction). Moreover, this result is obtained looking at all tracks inclusively, and the demonstrated improvement removes the need for a specific b-jet ROI (a requirement which led to problems outlined in section 3.3.2). These results hold when looking exclusively at B decay tracks. The fact that, as shown in fig. 3.8a, virtually all correctly assigned

a **b**

Figure 3.9: Profiles, as a function of parent B hadron p_{T}, of good (fig. 3.9a) and wrong (fig. 3.9b) hit assignment rates on the IBL for tracks using baseline tracking (black), and various looser values of the outlier cut.

hits are retained suggests that it may possible to relax this cut further. Tests are ongoing which will confirm this. The current GX2F treats all layers in the pixel detector in the same way - applying the same cut to each. While fig. 3.8a shows no adverse affects for hits on the IBL, when relaxing `m_outlcut` to a value of 1, some small reduction in good hit assignment efficiency was observed in other layers of the pixel detector, which are less precise. This difference in precision motivates the need to treat different layers in the pixel detector differently. To this end, layer-specific cutting capabilities for the GX2F are under development, which will allow each pixel layer to have their own cut point for outlier removal. Layer specific cuts will then be optimised to see if greater numbers of wrong hits can be successfully identified as outliers and removed, while maintaining high good hit assignment efficiency.

3.5 Tracking software validation

- tracking validation
- qspi validation

Chapter 4

Track Classification MVA

4.1 Machine Learning Background for Track Classification

4.2 Track Truth Origin Labelling

4.3 Fake Track Identification Tool

Probably talk about this model as a stepping stone to the general classifier

4.3.1 *b*-hadron Decay Track Identification Tool

Maybe don't need this section since it was talked about less

4.4 General Track Origin Classifier Tool

Culmination of this work in the general tool Martino has implemented

Applications:

- Frack to jet association

- Fake track studies (removal and for recommendations)

4.5 Conclusion

Improved with GNNs

Chapter 5

Graph Neural Network Flavour Tagger

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 $\sim$1.8 ($\sim$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 $\sim$6 ($\sim$2.8) for a comparative 30% b-jet efficiency.

5.1 Motivation

Flavour tagging, the identification of jets originating from b- and c-quarks, is a critical component of the physics programme of the ATLAS experiment [10] at the Large Hadron Collider (LHC) [17]. 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 [18, 19], and additionally for several Beyond Standard Model (BSM) resonances that readily decay to heavy flavour quarks [20]. The significant lifetime of b-hadrons, approximately 1.5 ps [21], 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 $b \to c$ 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) [22] 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. 5.1. 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 [23, 24].

The current ATLAS flavour tagger, DL1r [25], is a deep neural network which takes the outputs of a number of independently optimised "low-level" algorithms [26] as inputs. Each of these low-level algorithms 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, 26–28]. 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.

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

chapters/gnn_tagger/figs/GNN_compare_contrast.pdf

Figure 5.1: 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.

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 5.3; details about the model architecture and training procedure are given in section 5.4; and results are discussed in section 5.5.

[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. [12].

5.2 Graph Neural Network Theory

5.3 Experiemental Setup

5.3.1 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 $\sqrt{s} = 13\,\text{TeV}$ 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 [29–32] V2 generator at next-to-leading order with the NNPDF3.0NLO [33] set of parton distribution functions (PDFs). The h_{damp} parameter[3] is set to 1.5 times the mass of the top-quark (m_{top}) [34], with $m_{\text{top}} = 172.5\,\text{GeV}$. The events are interfaced to PYTHIA 8.230 [35] to model the parton shower, hadronisation, and underlying event, with parameters set according to the A14 tune [36] and using the NNPDF2.3LO set of PDFs [37]. 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 [38]. Particles are passed through the ATLAS detector simulation [39] based on GEANT4 [40].

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

[3]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.

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.

5.4 Model Architecture

5.4.1 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 5.1. If a jet has more than 40 associated tracks, the first 40 tracks with the largest transverse IP significance[4] $s(d_0)$ 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 [12]. 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 [42], and the electrons are required to pass the *VeryLoose* likelihood-based identification working point [43].

[4] 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 [41].

Table 5.1: 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 [12], 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 to reconstruct an electron or muon

5.4.2 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 5.5.3.

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 5.2 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 [44] - in detecting tracks from heavy flavour decays the model may learn to pay more attention to these tracks.

Table 5.2: 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 [12]. 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_S^0 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

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_S^0 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 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.

5.4.3 Architecture

As discussed above, the GN1 model combines a graph neural network architecture [45] 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 [24]. 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 [46], 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 5.4.1. The jet inputs are concatenated with each track's inputs, as shown in fig. 5.2. 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. 5.3. The track initialisation network is similar to a Deep Sets model [47], but does not include a reduction operation (mean or summation) over the output track representations.

chapters/gnn_tagger/figs/inputs_diagram.png

Figure 5.2: The inputs to GN1 are the two jet features ($n_{\mathrm{jf}} = 2$), and an array of n_{tracks}, where each track is described by 21 track features ($n_{\mathrm{tf}} = 21$). The jet features are copied for each of the tracks, and the combined jet-track vectors of length 23 form the inputs of GN1.

chapters/gnn_tagger/figs/full_arch.pdf

Figure 5.3: 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.

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. [46]. 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,

$$e(h_i, h_j) = \mathbf{a}^{\perp}\theta\left[\mathbf{W}h_i \oplus \mathbf{W}h_j\right], \tag{5.1}$$

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

$$a_{ij} = \mathrm{softmax}_j\left[e(h_i, h_j)\right]. \tag{5.2}$$

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}

$$h'_i = \sigma\left[\sum_{j\in\mathcal{N}_i} a_{ij}\cdot\mathbf{W}h_j\right]. \tag{5.3}$$

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 5.3.

Table 5.3: 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

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 5.2. 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.

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 5.4.2.

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.

5.4.4 Training

The full GN1 training procedure minimises the total loss function L_{total}, defined in eq. (5.4). 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.

$$L_{\text{total}} = L_{\text{jet}} + \alpha L_{\text{vertex}} + \beta L_{\text{track}} \tag{5.4}$$

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 5.5.3. 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 5.3.1. The Adam optimiser [48] 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 [49] using ONNX [50], and jet flavour predictions for the test sample are computed using the ATLAS software stack.

5.5 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 [25], which has been retrained on 75 million jets from the same samples as GN1. The input RNNIP tagger [28] 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

$$D_b = \log \frac{p_b}{(1-f_c)p_l + f_c p_c}, \tag{5.5}$$

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 [25]. 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. 5.4. 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.

chapters/gnn_tagger/figs/results/main/ttbar/ttbar_score_DL1r_GN12022

Figure 5.4: Comparison between the DL1r and GN1 b-tagging discriminant D_b for jets in the $t\bar{t}$ sample. The 85% 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.

5.5.1 *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. 5.5 and 5.6 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 $\sim$30% on the Z' sample. To account for this, the range of b-jet tagging efficiencies displayed in fig. 5.6 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-rejection improves by a factor of $\sim$2.1 and the light-jet rejection improves by a factor of $\sim$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-rejection by a factor of $\sim$2.8 and the light-jet rejection by a factor of $\sim$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, $\sim$5% ($\sim$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 $\sim$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-rejection improvement with respect to DL1r at the 70% b-jet WP increases from a factor of $\sim$2.1 for GN1 to a factor of $\sim$2.8 for GN1 Lep. The improvement in light-jet rejection also increases from a factor of $\sim$1.8 to $\sim$2.5 at this WP. For jets in the Z' sample, the relative c-rejection (light-jet rejection) improvement with respect to DL1r increases from a factor of $\sim$2.8 to $\sim$3 ($\sim$6 to $\sim$7.5) at a b-jet tagging efficiency of 30%. As shown in fig. 5.7, 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 [12]. 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. 5.7. 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.

5.5.2 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. (5.5) as chapters/gnn_tagger/figs/results/main/ttbar/ttbar_roc_btag.pdf

Figure 5.5: 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 $\sim$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.

chapters/gnn_tagger/figs/results/main/zprime/zprime_roc_btag.pdf

Figure 5.6: 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 $250 < p_{\mathrm{T}} < 5000$ 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.

chapters/gnn_tagger/figs/results/

chapters/gnn_tagger/figs/results/main/zprime/zp

Figure 5.7: 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.

$$D_c = \log \frac{p_c}{(1 - f_b)p_l + f_b p_b}. \tag{5.6}$$

A value of $f_b = 0.2$ is used for all models. Similar to section 5.5.1, 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 5.5.1, 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. 5.8 and 5.9. In fig. 5.8, 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 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-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. 5.9 shows the c-tagging performance on the jets in the Z' sample. Both GN1 and GN1 Lep perform similarly, improving the b-rejection by 60% and the light-jet rejection by a factor of 2 at the 25% c-jet WP.

5.5.3 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 5.4.2) 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. 5.10 and 5.11. 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 chapters/gnn_tagger/figs/results/main/ttbar/ttbar_roc_ctag.pdf

Figure 5.8: 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 $\sim$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.

chapters/gnn_tagger/figs/results/main/zprime/zprime_roc_ctag.pdf

Figure 5.9: 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.

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

5.5.4 Inclusion of Low-Level Vertexing Algorithms

GN1 does not include inputs from low-level tagging algorithms, including the vertexing tools SV1 and JetFitter [26]. Since these algorithms are known to improve the performance of DL1r, it was feasible that their inclusion in GN1 may further improve on the performance of the GN1 models. In a dedicated training of GN1 the SV1 and JetFitter tagger outputs were added to the GN1 jet classification network as an input, similar to their use in DL1r. These outputs include information on the reconstructed vertices, including the number of vertices, the vertex mass, displacement, and other properties. In addition, the index of the reconstructed SV1 or JetFitter vertices were included as two track-level inputs to GN1. The jet classification performance of this GN1 model was not significantly different to the baseline model, and in some cases the performance was slightly reduced. A dedicated look at the vertexing performance of GN1 with some comparisons to SV1 and JetFitter is found in section 5.5.5

5.5.5 Vertexing Performance

From the track-pair vertex prediction described in section 5.4.2, tracks can be partitioned into compatible groups representing vertices (see [24]). 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 chapters/gnn_tagger/figs/results/ablations/ttbar/ttbar_roc_btag.pdf

Figure 5.10: 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 b-jet tagging efficiencies usually probed in these regions.

chapters/gnn_tagger/figs/results/ablations/zprime/zprime_roc_btag.pdf

Figure 5.11: 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 b-jet tagging efficiencies usually probed in these regions.

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 5.4.2, 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.

More detail

In order to study the performance of the different vertexing tools inside b-jets, the truth vertex label of the tracks, discussed in section 5.4.2, is used. The reconstructed vertices from GN1, SV1 and JetFitter are compared to the target truth vertices in order to calculate the efficiencies of the different vertexing tools. 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.

There are several caveats to a comparison of the vertexing tools which are a result of the different approaches they take to vertexing. SV1 and JetFitter are designed to only find secondary vertices in the jet, whereas GN1 is also trained to determine which tracks in the jet belong to the primary vertex (the vertex of the hard scatter pp interaction). To account for this the GN1 vertex with the largest number of predicted primary tracks is excluded from the vertex finding efficiency calculation. While JetFitter and GN1 aim to resolve each displaced vertex inside the jet, such that secondary vertices from b-hadron decays are found separately to tertiary vertices from $b \to c$ decay chains, SV1 by design attempts to find a single inclusive vertex per jet. This inclusive vertex groups inclusive b-hadron decays. These are tracks from the b-hadron decay itself (FromB) and tracks from $b \to c$ decays (FromBC). In order to fairly compare the performance if the different tools, both the exclusive and inclusive vertex finding efficiency is studied. For the exclusive vertex finding case JetFitter and GN1 can be directly compared, while a comparison with SV1 is not possible due to aforementioned design constraints. The inclusive vertex finding performance of all three tools can be compared using the procedure outlined below.

The starting point for the secondary vertex finding efficiency in both the exclusive and inclusive cases is to select truth secondary vertices are those containing only inclusive b-hadron decays to be considered as initial targets. For exclusive vertex finding, these truth secondary vertices can be used directly as the demoninator for the efficiency calculation. Meanwhile for the inclusive efficiency all such truth secondary vertices in the jet are merged into a single inclusive target vertex. Correspondingly, for the inclusive vertex finding case, the vertices found by JetFitter are merged into a single vertex, and the vertices found by GN1 with at least one predicted inclusive b-hadron decay track are also merged similarly. SV1 does not require any vertex merging.

Next, in both cases for each truth secondary vertex, vertices in the jet found by the different vertexing tools are compared with the target truth vertex. The number of correctly and incorrectly assigned tracks is computed. In order to call a vertex efficient, it is required to contain at least 65% of the tracks in the corresponding truth vertex, and to have a purity of at least 50%. Single track vertices are required to have a purity of 100%. Additionally, for GN1 only, at least one track in the vertex is required to have a predicted heavy flavour origin.

Vertex finding efficiencies for b-jets in the $t\bar{t}$ sample are displayed as a function of p_{T} separately for the inclusive and exclusive approaches in fig. 5.12. For b-jets in the $t\bar{t}$ sample with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$, the exclusive vertex finding efficiency of JetFitter and GN1 is relatively flat as a function of p_{T}. Of the truth secondary vertices in this p_{T} region, JetFitter efficiently finds approximately 40% and GN1 finds approimately 55%. When finding vertices inclusively the vertex finding efficiency is generally higher. An increased dependence on p_{T} is also visible for JetFitter and SV1. As the jet p_{T} increases from 20 GeV to 100 GeV, the efficiency of JetFitter increases from 55% to 65%. In the same range, the efficiency of SV1 increases from 55% to 75%. GN1 displays less dependence on p_{T} than JetFitter and SV1, efficiently

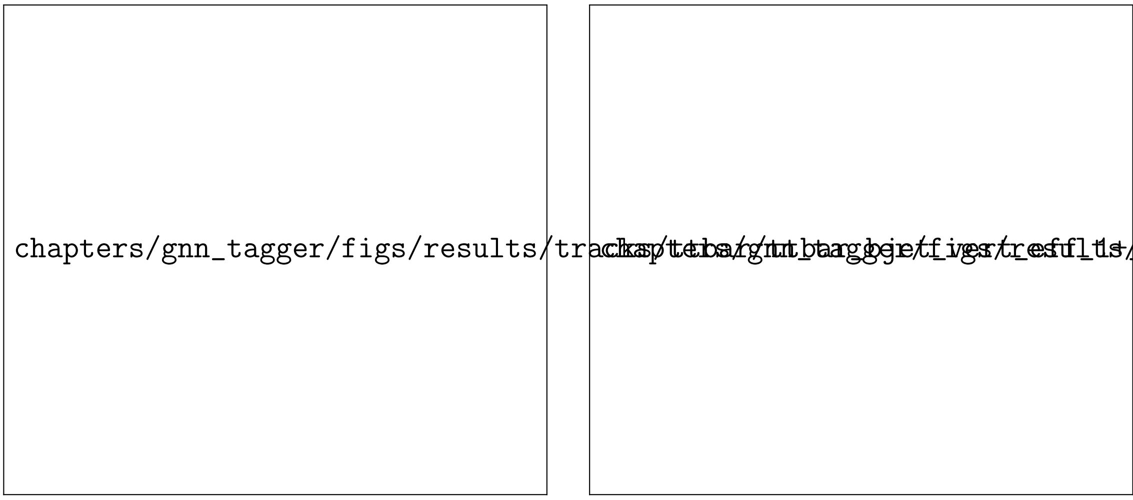

Figure 5.12: Vertex finding efficiency as a function of jet p_{T} for b-jets in the $t\bar{t}$ sample using the exclusive (left) and inclusive (right) vertex finding approaches. Efficient vertex finding requires the recall of at least 65% of the tracks in the truth vertex, and allows no more than 50% of the tacks to be included incorrectly. Binomial error bands are denoted by the shaded regions.

finding upwards of 80% of vertices in b-jets in this p_{T} region. For b-jets with $p_{\mathrm{T}} > 100\,\mathrm{GeV}$, JetFitter finds approximately 65% of vertices, SV1 finds appoimately 75% of vertices, and GN1 finds approximately 80% of vertices.

For b-jets in the Z' sample, the vertex finding efficiency drops steeply with increasing p_{T} up until $p_{\mathrm{T}} = 3\,\mathrm{TeV}$. GN1 outperforms SV1 and JetFitter across the p_{T} spectrum. In the first bin, the efficiency of GN1 is 75%, while the efficiencies of SV1 and JetFitter are around 60%. The efficiency of SV1 drops rapidly to almost zero above 3 TeV, while JetFitter and GN1 retain approximately 30% efficiency. fig. 5.13 compares the exclusive vertex finding efficiencies of JetFitter and GN1 for multi-track vertices. JetFitter finds 45-50% of vertices in b-jets in the $t\bar{t}$ sample, while GN1 finds 60-65%. For b-jets in the Z' sample, JetFitter finds 35% of vertices in the first bin, dropping to 20% of vertices above 2 TeV. GN1 finds 55% of vertices in the first bin, dropping to 30% above 2 TeV.

5.5.6 Track Classification Performance

As discussed in section 5.4.2, 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

Figure 5.13: Inclusive vertex finding efficiency for multitrack truth vertices in b-jets in the $t\bar{t}$ sample (left) and jets in the Z' sample (right) as a function of jet p_{T}. Efficient vertex finding requires the recall of at least 65% of the tracks in the truth vertex, and allows no more than 50% of the tacks to be included incorrectly.

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 5.4.1), 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 5.4, 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 5.4.3, 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.

Table 5.4: 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 5.4.3, 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**

fig. 5.14 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 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.

chapters/gnn_tagger/figs/results/tr[illegible]

chapters/gnn_tagger/figs/r[illegible]ks/zprime/

Figure 5.14: ROC curves for the different groups of truth origin labels defined in table 5.2 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.

Chapter 6

VHbb Analysis Preamble

6.1 Overview

The Higgs boson, discovered at the LHC in 2012, is predicted by the standard model to decay primarily to two b quarks, with a branching factor of 0.582 ± 0.007 [51]. Observation of this decay mode was recently reported by ATLAS [52]. Whilst the dominant Higgs production mode at the LHC is gluon-gluon fusion, this mode has an overwhelming QCD multijet background and so sensitivity to the Higgs is low. The $\mathrm{H} \rightarrow b\bar{b}$ observation therefore searched for Higgs bosons produced in association with a vector boson (W or Z). This production mechanism results in leptonic final states from the decay of the vector boson, allowing for leptonic triggering, whilst at the same time significantly reducing the multi-jet background.

A closely related analyses now searches for the $\mathrm{H} \rightarrow b\bar{b}$ decay of the Higgs boson, produced in association with a vector boson, when the vector boson and Higgs are highly boosted. The full Run-2 dataset is used for a total integrated luminosity of 139 fb^{-1}. The analysis is split into 0-, 1- and 2-lepton channels depending on the number of selected electrons and muons, to target the $\mathrm{ZH} \rightarrow \nu\nu bb$, $\mathrm{WH} \rightarrow \ell\nu bb$, $\mathrm{ZH} \rightarrow \ell\ell bb$ processes, respectively, where ℓ is an electron or muon. In all channels, events are required to have exactly two b-tagged jets, which form the Higgs boson candidate. At least one of the b-tagged jets is required to have p_{T} greater than 45 GeV. Events are further split into 2-jet or 3-jet categories depending on whether additional, untagged jets are present.

In the 0- and 1-lepton channels, the analysis is further split into signal and control regions. To leading order, there are no additional b-jets in the event other than the two coming from the reconstructed Higgs candidate. For this reason, there is a signal region veto (i.e. events are not accepted into the signal region) for events with additional b-tagged jets in the event. Events with additional b-tagged jets are included in the control region, which is highly pure in $t\bar{t}$ events. The control region is used to constrain the normalisation of the $t\bar{t}$ background.

Chapter 7

VHbb Boosted Analysis

7.1 Overview

7.2 Modelling Work

7.2.1 Background

Source of Uncertainty	Implementation
Renormalisation scale (μ_R)	Internal weights
Factorisation scale (μ_F)	Internal weights
PDF set	Internal weights
α_S value	Internal weights
Parton Shower (PS) models	Alternative samples
Underlying Event (UE) models	Alternative samples
Resummation scale (QSF)	Parameterisation
CKKW merging scale	Parameterisation

Table 7.1: Different sources of uncertainty (i.e. variations in the model) considered for V+jets background, and the corresponding implementation. For each uncertainty, acceptance and shape uncertainties are derived.

Alternative Samples

As mentioned, alternative samples of V+jets events was generated using MadGraph5_aMC@NLO+Pythia8, and the results are compared with the nominal Sherpa 2.2.1 samples. This allows for a comparison of different parton showering and underlying event models, and derivation of the systematic uncertainties on the nominal choice of models.

Internal Weight Variations

Nominal signal samples generated with Sherpa 2.2.1 include systematic variations of certain modelling parameters which are stored as alternative event weights. The samples contain event weight variations which correspond to variations of renormalisation scale μ_R, and factorisation scale μ_F, of 0.5 and 2 times the nominal value. Additionally stored is event weight variations corresponding to 30 different variations on the PDF and two variations of the strong coupling constant α_S. Variations of α_S were found to have negligible impact on the results of the analysis, and are not discussed further.

Parameterisation Methods

While the inclusion of internal weight variation in MC event generators has decreased simulation times and increased available statistics, there are in Sherpa 2.2.1 currently some sources of systematic uncertainty that are unable to be stored as internal weight variations due to technical limitations. Two such systematics relate to the choice of CKKW matrix element merging scale, and resummation scale (QSF). The generation of high statistics alternative samples is a time consuming process, as is typically not done for all samples for every new generator release. A method to parameterise the systematic variation using one sample, and to then apply this parameterisation to another sample, has been developed by the ATLAS SUSY group [53]. This method was used to derive CKKW and QSF uncertainties for the nominal Sherpa 2.2.1 sample, using a previous (lower statistic) Sherpa 2.1 alternative sample. The resulting uncertainties were studied and found to be negligible in comparison with systemics from other sources.

Shape Uncertainties

In order to derive shape uncertainties (which as the name suggests affect shapes but not overall normalisations of distributions), the following procedure is carried out. Normalised distributions of the reconstructed Higgs candidate mass m_J are compared for the nominal sample and variations. For each variation, the ratio of the variation to nominal is calculated, and an analytic function is fit to those sources of variation which have a ratio deviating from unity. If different analysis regions or channels show the same pattern of variation, a common uncertainty is assigned. An example of a significant source of uncertainty, arising from choice of factorisation scale μ_R is shown in fig. 7.1. An exponential function has been fitted to the ratio of the normalised distributions. Two different analysis regions (medium and high $p_T{}^V$ bins) are shown. The difference of the shape of the variation means that two separate uncertainties have to be added in the fit, and applied individually in each $p_T{}^V$ region.

chapters/6.vhbb_boosted/figs/1L_Wh

a **b**

Figure 7.1: Normalised distributions of leading fat jet mass m_J for medium (fig. 7.1a) and high (fig. 7.1b) $p_T{}^V$ analysis regions for W+heavy-flavour-jets (merged in heavy flavours, high and low purity signal regions) in the 0 lepton channel. The renormalisation scale μ_R has been varied by a factor of 2 ("1up") and 0.5 ("1down"). An exponential function has been fit to the ratio.

Acceptance Uncertainties

Several different types of acceptance uncertainties have been calculated. These are implemented as nuisance parameters in the fit and for the most part account for the migration of events between different analysis regions. The list acceptance uncertainties relevant to the V+jets processes are given summarised below.

- **Overall normalisation:** only relevant where normalisation cannot be left floating (i.e. determined in the fit).
- **SR-to-CR relative acceptance:** the uncertainty on the normalisation of the signal region due to events migrating between the signal and control regions.
- **HP-to-LP relative acceptance:** the uncertainty on the normalisation of the high-purity (HP) signal region due to events migrating between the high- and low-purity signal regions.
- **Medium-to-high p_{T^V} relative acceptance:** describes any 'shape' effect in p_{T^V} distribution, given that the analysis only uses two p_{T^V} bins (medium and high).
- **Flavour relative acceptance:** for each flavour Vxx, where $xx \in$ {bc,bl,cc} the ratio of Vxx/Vbb events is calculated. This corresponds to the uncertainty of Vbb events due to the miss-tagging of other flavours Vxx.

The uncertainties on different systematics are summed in quadrature to give a total uncertainty on each region. A summary of the different acceptance uncertainties that were derived in this way and subsequently applied in the fit are given in table 7.2. An effort has been made, wherever possible, to harmonise similar uncertainties across different analysis regions and channels.

7.2.2 Vector Boson + Jets Modelling

The background processes involving W or Z boson decays into leptons (including those in which the W boson arises from a top-quark decay) are collectively referred to as electroweak (EW), or V+jets, backgrounds. W+jets events are most relevant to the 1-lepton channel via the leptonic decay of W$\rightarrow \ell\nu$. In the event of W$\rightarrow \tau\nu$, and subsequent decay of the τ, or the lack of the successful reconstruction of the

e or μ, W+jets can also contribute to the 0-lepton channel. Meanwhile, Z+jets contributes primarily to the 0- and 2-lepton channels via the processes Z$\rightarrow \nu\nu$ and Z$\rightarrow \ell\ell$ respectively.

Modelling is used to predict the outcomes of the analysis and to assess the impact of sources of different systematic uncertainty. Signal and background modelling has has primarily consisted of using Monte Carlo (MC) generators to produce simulated events. The uncertainties on the simulated output must be well understood to perform a successful analysis. To achieve this, a set of "nominal" samples are first defined as a reference to which different variations can be compared. The nominal samples are chosen as the best possible representation of the underlying physical process. "Alternative" samples are used to understand the systematic uncertainties on the nominal samples. To generate an alternative sample, some aspect of the model is varied, and the simulation is re-run. A comparison back to the nominal sample gives a handle on the systematic uncertainty associated with the model parameter which was changed. Detailed information can be found in [54]. In order to access uncertainties associated with the use of MC generators, variations of the data are produced using alternative generators or variation of nominal generator parameters. The variation of nominal generator parameters can in certain cases be implemented using internal weight variations stored alongside the nominal events, and in other cases a new independent sample must be generated. The nominal generator used for V+jets events is SHERPA 2.2.1, while MADGRAPH5_aMC@NLO+PYTHIA8 (which uses different parton showering models) is used as an alternative generator. As production of large MC samples is computationally expensive, a feature of state of the art simulation packages is to store some sources of variation as internal event weights, which can be generated alongside the nominal samples, saving computation time. Several sources of uncertainty, summarised in table 7.1, have been assessed.

V+jets Acceptance Uncertainties				
Boson	**W**		**Z**	
Channel	0L	1L	0L	2L
Vbb Norm.	30%	-	-	-
SR/CR	90%†	40%†	40%	-
HP/LP	18%		18%	-
High/Medium p_T^V	30%	10%*	10%	
Channel Extrap.	20%	-	16%	-
Vbc/Vbb	30%			
Vbl/Vbb	30%			
Vcc/Vbb	20%			
Vcl Norm.	30%			
Vl Norm.	30%			

Table 7.2: V+jets acceptance uncertainties. W+jets SR/CR uncertainties marked by † are correlated. The 1L W+jets H/M uncertainty marked by $*$ is applied as independent and uncorrelated NPs in both HP and LP signal regions. The 0L W+jets Wbb Norm uncertainty is only applied when a floating normalisation for Wbb cannot be obtained from the 1L channel. A 30% uncertainty for Zbb norm is applied in the 1L channel when a floating normalisation for Zbb cannot be obtained from the 0L or 2L channels.

7.2.3 Diboson Modelling

7.3 Fit Studies

7.3.1 Fit Model

A global profile likelihood fit is used to extract the signal strength μ and its significance from the data. This statistical setup treats each bin as a Poisson counting experiment. The combined likelihood over N bins, without considering sources of systematic uncertainty, is given by

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} \exp\left[-(\mu s_i + b_i)\right], \tag{7.1}$$

where s_i (b_i) is the expected number of signal (background) events in bin i, and n_i is the number of events observed in data in bin i. The presence of systematic uncertainties which can affect the expected numbers of signal and background events necessitates the addition of nuisance parameters (NPs), θ, to the likelihood. Each source of systematic uncertainty for V+jets samples discussed in the previous section was implemented as a NP θ_j in the fit. The presence of NPs modifies the likelihood as

$$\mathcal{L}(\mu) \to \mathcal{L}(\mu, \theta) = \mathcal{L}(\mu) \times \mathcal{L}(\theta)\ , \quad s_i \to s_i(\theta)\ , \quad b_i \to b_i(\theta), \tag{7.2}$$

where

$$\mathcal{L}(\theta) = \prod_{\theta_j \in \theta} \frac{\exp\left[-\theta_j^2/2\right]}{\sqrt{2\pi}}. \tag{7.3}$$

Post-fit m_J distributions in the high-purity medium p_{T^V} regions for the 0- and 2-lepton channels are shown in fig. 7.2. The plots show large falling backgrounds, predominantly made up of W+jets and Z+jets events, and a signal distribution corresponding to the Standard Model Higgs boson peaking around $m_H = 125$ GeV.

a

b

Figure 7.2: Post-fit distributions for the 0-lepton (fig. 7.2a) and 2-lepton (fig. 7.2b) channels in the high purity medium p_{T^V} region, obtained in the combined conditional $\mu = 1$ fit to data. The last bin of each plot is an overflow bin.

7.4 Conclusion

Work has been carried out as part of the boosted VHbb analysis group to understand, and implement in the global profile likelihood fit, systematic uncertainties on V+jets samples. This background modelling work is an essential part of the success of the analysis. So far the fit has proved stable with the inclusion of the V+jets uncertainties, and detailed studies are now underway to determine the causes behind any observed pulls of the added NPs. Additional work is ongoing to help with the derivation of uncertainties on diboson samples, another important background. The analysis is already advanced, and is now progressing into its final stages. Publication is expected in the new year.

Chapter 8

VHbb Legacy Analysis

8.1 Overview

Chapter 9

Conclusion

Appendix A

Combining Multiple Triggers

Colophon

This thesis was made in $\LaTeX\,2_\varepsilon$ using the "hepthesis" class [55].

Bibliography

[1] L. Morel, Z. Yao, P. Cladé, and S. Guellati-Khélifa, Nature 588, 61 (2020).

[2] T. Sailer *et al.*, Nature 606, 479 (2022).

[3] CDF, F. Abe *et al.*, Phys. Rev. Lett. 74, 2626 (1995), hep-ex/9503002.

[4] D0, S. Abachi *et al.*, Phys. Rev. Lett. 74, 2632 (1995), hep-ex/9503003.

[5] S. W. Herb *et al.*, Phys. Rev. Lett. 39, 252 (1977).

[6] UA1, G. Arnison *et al.*, Phys. Lett. B 122, 103 (1983).

[7] DONUT, K. Kodama *et al.*, Phys. Lett. B 504, 218 (2001), hep-ex/0012035.

[8] ATLAS Collaboration, (2010).

[9] B. Abbott *et al.*, JINST 13, T05008 (2018), 1803.00844.

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

[11] ATLAS Collaboration, Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS, ATL-PHYS-PUB-2020-014, 2020.

[12] ATLAS Collaboration, Eur. Phys. J. C 77, 673 (2017), 1704.07983.

[13] ATLAS Collaboration, Eur. Phys. J. C 77, 466 (2017), 1703.10485.

[14] M. Cacciari, G. P. Salam, and G. Soyez, JHEP 04, 063 (2008), 0802.1189.

[15] ATLAS Collaboration, Phys. Rev. D 96, 072002 (2017), 1703.09665.

[16] ATLAS Collaboration, Tagging and suppression of pileup jets with the ATLAS detector, ATLAS-CONF-2014-018, 2014.

[17] L. Evans and P. Bryant, JINST 3, S08001 (2008).

[18] ATLAS Collaboration, Phys. Lett. B 786, 59 (2018), 1808.08238.

[19] ATLAS Collaboration, Phys. Lett. B 784, 173 (2018), 1806.00425.

[20] ATLAS Collaboration, JHEP 03, 145 (2020), 1910.08447.

[21] Particle Data Group, M. Tanabashi *et al.*, Phys. Rev. D 98, 030001 (2018).

[22] P. W. Battaglia *et al.*, arXiv preprint arXiv:1806.01261 (2018).

[23] J. Shlomi *et al.*, The European Physical Journal C 81 (2021).

[24] H. Serviansky *et al.*, Set2graph: Learning graphs from sets, 2020, 2002.08772.

[25] ATLAS Collaboration, Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run, ATL-PHYS-PUB-2017-013, 2017.

[26] ATLAS Collaboration, Eur. Phys. J. C 79, 970 (2019), 1907.05120.

[27] ATLAS Collaboration, Secondary vertex finding for jet flavour identification with the ATLAS detector, ATL-PHYS-PUB-2017-011, 2017.

[28] ATLAS Collaboration, Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment, ATL-PHYS-PUB-2017-003, 2017.

[29] P. Nason, Journal of High Energy Physics 2004, 040–040 (2004).

[30] S. Frixione, G. Ridolfi, and P. Nason, Journal of High Energy Physics 2007, 126–126 (2007).

[31] S. Frixione, P. Nason, and C. Oleari, Journal of High Energy Physics 2007, 070–070 (2007).

[32] S. Alioli, P. Nason, C. Oleari, and E. Re, Journal of High Energy Physics 2010 (2010).

[33] NNPDF, R. D. Ball *et al.*, JHEP 04, 040 (2015), 1410.8849.

[34] ATLAS Collaboration, Studies on top-quark Monte Carlo modelling for Top2016, ATL-PHYS-PUB-2016-020, 2016.

[35] T. Sjöstrand *et al.*, Comput. Phys. Commun. 191, 159 (2015), 1410.3012.

[36] ATLAS Collaboration, ATLAS Pythia 8 tunes to 7 TeV data, ATL-PHYS-PUB-2014-021, 2014.

[37] R. D. Ball *et al.*, Nucl. Phys. B 867, 244 (2013), 1207.1303.

[38] D. J. Lange, Nucl. Instrum. Meth. A 462, 152 (2001).

[39] ATLAS Collaboration, Eur. Phys. J. C 70, 823 (2010), 1005.4568.

[40] GEANT4 Collaboration *et al.*, Nucl. Instrum. Meth. A 506, 250 (2003).

[41] ATLAS Collaboration, JINST 11, P04008 (2016), 1512.01094.

[42] ATLAS Collaboration, Muon reconstruction performance in early $\sqrt{s} = 13$ TeV data, ATL-PHYS-PUB-2015-037, 2015.

[43] ATLAS Collaboration, Eur. Phys. J. C 79, 639 (2019), 1902.04655.

[44] D. Hwang *et al.*, Self-supervised auxiliary learning with meta-paths for heterogeneous graphs, 2020.

[45] J. Shlomi, P. Battaglia, and J.-R. Vlimant, Machine Learning: Science and Technology 2, 021001 (2021).

[46] S. Brody, U. Alon, and E. Yahav, arXiv e-prints , arXiv:2105.14491 (2021), 2105.14491.

[47] M. Zaheer *et al.*, Deep sets, 2018, 1703.06114.

[48] D. P. Kingma and J. Ba, (2014).

[49] ATLAS Collaboration, (2021).

[50] J. Bai *et al.*, Onnx: Open neural network exchange, `https://github.com/onnx/onnx`, 2019.

[51] D. de Florian *et al.*, arXiv e-prints , arXiv:1610.07922 (2016), 1610.07922.

[52] M. Aaboud *et al.*, Physics Letters B 786, 59–86 (2018).

[53] J. K. Anders and M. D’Onofrio, CERN Report No. ATL-COM-PHYS-2016-044, 2016 (unpublished).

[54] A. S. Bell and F. Lo Sterzo, CERN Report No. ATL-COM-PHYS-2018-505,

2018 (unpublished).

[55] A. Buckley, *A class for typesetting academic theses*, 2010.