Naturwissenschaftlich-Technische Fakultät

Universität Siegen

UNIVERSITÄT SIEGEN

MASTERARBEIT

Material Interactions in Flavour Tagging of the ATLAS Experiment at the Large Hadron Collider

Nils Benedikt Krengel

Arbeit zur Erlangung des akademischen Grades
Master of Science

vorgelegt der

Experimentellen Teilchen- und Astroteilchenphysik
Department Physik

Gutachter
Prof. Dr. Markus Cristinziani
PD Dr. Carmen Diez Pardos

Version vom
1. August 2025

Dieser Forschungsbericht wurde als Masterarbeit von der Naturwissenschaftlich-Technischen Fakultät der Universität Siegen angenommen. Seit der Annahme sind höchstens geringfügige Änderungen vorgenommen worden. Der wissenschaftliche Inhalt der Arbeit ist unverändert.

Angenommen am:	29.07.2025
1. Gutachter:	Prof. Dr. Markus Cristinziani
2. Gutachterin:	PD Dr. Carmen Diez Pardos

Acknowledgements

This thesis is the culmination of a lot of work. Work I could not have done alone. So I would like to take a moment to thank all the people who played a part in bringing it together.

First and foremost I would like to thank Prof. Dr. Markus Cristinziani for giving me the opportunity to work in this exciting field of research, and for his guidance and advice throughout. I would also like to thank PD Dr. Carmen Diez Pardos for kindly agreeing to be the second reviewer of this thesis.

My special appreciation goes to the very knowledgeable experts who worked with me on this project. Many thanks go to Dr. Vadim Kostyukhin for being the supervisor of this project. I am especially grateful to Dr. Diptaparna Biswas, whose door was always open to me and who supported me with every problem I encountered along the way. Furthermore, I would like to thank Dr. Sam van Stroud and Dr. Nicole Michelle Hartmann for creating an extremely welcoming and friendly environment in the flavour tagging algorithms working group, and their support of my work in particular.

I am incredibly grateful to the many people I have the honour to call my friends. Be it outside or inside the institute, they helped me tremendously by providing not only advice and support relating to the work, but more importantly joy, warmth and lots of laughter.

All this would not have been possible without the love and support of my family, to whom I am deeply thankful for everything they have done and continue to do for me.

To everyone, thank you.

Contents

CHAPTER 1

Introduction

The best attempt to describe the fundamental building blocks of our universe is currently summarized in the Standard Model of Particle Physics [1]. This model encompasses elementary particles which make up matter and mediate the interactions of the electromagnetic, weak and strong forces. In order to observe and characterize these particles and to test the Standard Model and theories beyond, an environment of high energy is needed as well as sophisticated detector instrumentation.

The Large Hadron Collider (LHC) is the biggest and most powerful particle collider as of today [2]. The protons in the collider are currently accelerated to a centre-of-mass energy of 13.6 TeV and are brought to collision at different detectors around the collider. One of the general-purpose detectors at the LHC, and currently the largest volume particle detector ever constructed, is the ATLAS (A Toroidal LHC Apparatus) detector [3]. The ATLAS experiment was able to discover the Higgs boson in 2012 [4], a big missing piece in the Standard Model. Furthermore, the ATLAS experiment performs precision measurements of the Standard Model particles and their interactions while also searching for new phenomena, which might add missing pieces in our understanding of the universe.

The signals measured by the different systems of the ATLAS detector have to be reconstructed to physical particles. Comparing actual detector data to Monte-Carlo simulation events is essential to most analyses and these simulated events will go through the same reconstruction steps. One such step is the identification of particles. Quarks pose a unique challenge in this step as they cannot be as easily identified as electrons or muons for example, because they produce a whole spray of particles, which interact with and deposit their energy in the detector. This spray of particles is called a jet and can also be initiated by a gluon or the hadronic decay of a tau lepton. It is not obvious which kind of particle or which flavour initiated a jet, but for a lot of interesting questions in particle physics flavour is highly relevant.

Jet–flavour tagging aims to assign a flavour to a jet by exploiting the measured properties of the jet and the tracks associated to it. A clear distinction between all the different kinds of jets is not realized at the moment, but the tagger used by the ATLAS collaboration distinguishes between jets initiated by bottom, charm, lighter flavour quarks, and jets initiated by the hadronic decay of a tau lepton. Heavy-flavour quark jets, especially b-jets, have characteristic properties that make their identification relatively easy. That is why flavour tagging started out as b-tagging, as only a distinction between b-jets and all other jets was made. The characteristic properties of the heavier quark jets can also occur if a constituent of the jet interacts with the detector material, or a similar secondary phenomenon is part of the jet. This way jets originating from a lighter flavour quark might mimic the ones originating from heavier flavours and thus contribute to mistagging, when they contain secondary effects.

This thesis studies the influence of material interactions and other secondary effects on the ATLAS flavour tagging effort currently performed with a transformer encoder model [5]. A labelling scheme to classify the different secondary processes, including material interactions, is devised to perform these studies and to possibly mitigate the effects of secondaries on mistagging. To this purpose an additional classification objective is added to the flavour tagging model, which classifies the tracks inside a jet into the different categories of secondary origin.

CHAPTER 2

Physics and Technical Background

2.1 The Large Hadron Collider

The Large Hadron Collider, depicted in figure 2.1, covers a circumference of 27 km and lies up to 175 m deep below Geneva and its surrounding area. The main focus of the LHC are the proton–proton collisions, but its programme also includes heavy-ion collisions, mostly consisting of lead ions. It is the latest addition to the accelerator complex at CERN, where the particles go through separate linear and circular colliders before finally being stored in the LHC [2].

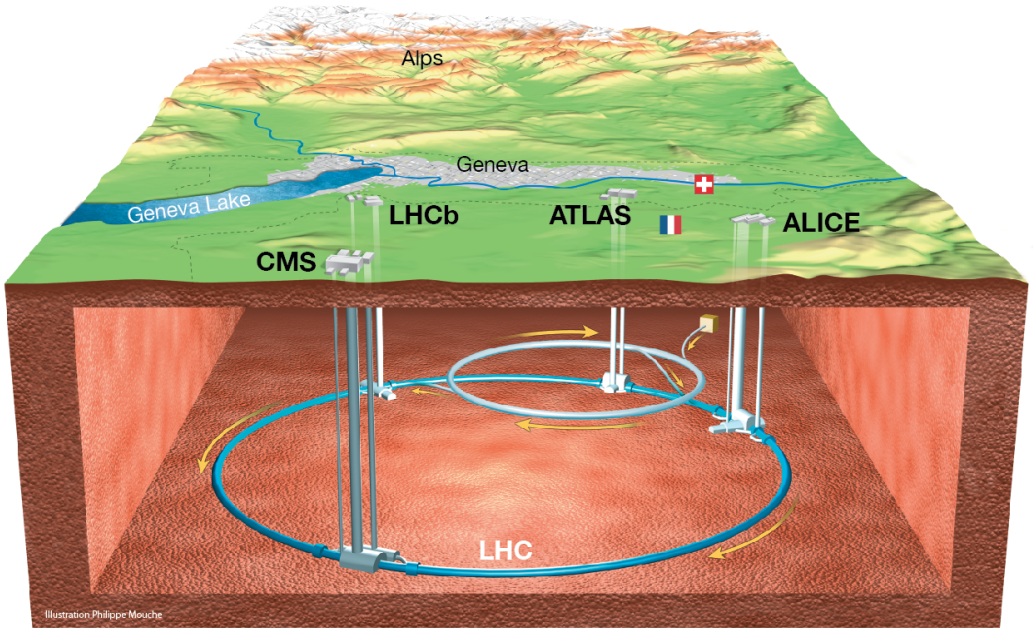


Figure 2.1: Overview of the Large Hadron Collider and its four largest experiments under the Geneva area. [2]

The particles are grouped together and stored in the ring as bunches. Two beams of multiple bunches circle the LHC in opposite directions and meet each other at the interaction points, where the detectors are located. The two beam pipes are held at ultra-high vacuum and the beams are kept on track by superconducting electromagnets, which bend the particles on the circular trajectory and focus the particle bunches. Once the quality of a beam is degraded by the collisions and other effects, it is diverted into a beam dump system as its energy of 350 MJ has an extreme destructive power [2].

Figure 2.2 provides an overview over the different runs of the LHC and of what is planned for the future decades. In Run 1, with centre-of-mass energies of 7 and 8 TeV, enough data could be gathered at these energies to reach one of the major goals of the project, the discovery of the Higgs boson. It was observed both in the ATLAS [4] and CMS [6] experiments. Aside from this important missing piece of the Standard Model, it was also deemed possible that the LHC might find particles beyond the Standard Model (e.g. SUSY particles), especially with the higher energies of 13 and 13.6 TeV in Run 2 and 3.

Although the SUSY particles were not found at these energies, a lot of searches for new particles are performed at the LHC experiments, including searches for dark matter candidates. Apart from these searches, the Standard Model, as it currently holds, is investigated by precise measurements of the properties of known particles, like the Higgs boson or the top quark, or by precise measurements of the interactions, especially rare processes. Finding a deviation from the Standard Model in these precision measurements is another gateway to the unanswered questions beyond our current understanding of the universe. To achieve even more precise measurements, the LHC will enter a High Luminosity phase, as shown in figure 2.2, further increasing the collision rate and energy to obtain more data.

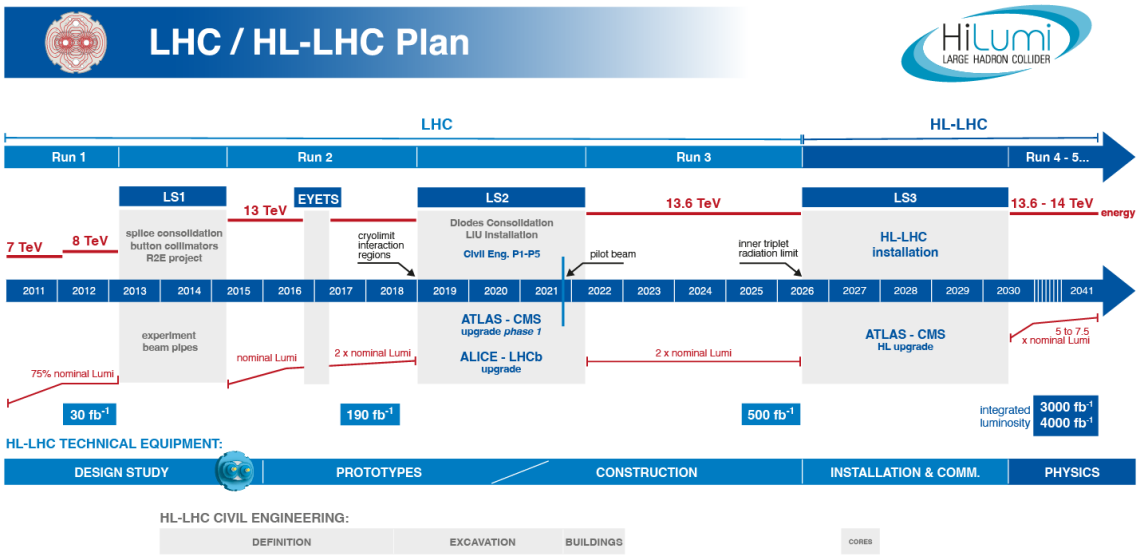


Figure 2.2: Overview of the LHC operation so far and of the currently planned upgrades and runs in the High Luminosity LHC (HL-LHC) phase. [7]

2.2 The ATLAS Detector

The ATLAS detector shown in figure 2.3 is designed in a cylindrical onion-like manner, so that particles are stopped or leave a unique signature at different layers according to their type. The different elements of the detector are either barrel-shaped, concentrically arranged around the beam pipe, or caps placed at the front or the back of the cylinder shape. It is symmetric in the forward and backward direction with respect to the interaction point, at which the proton bunches of the LHC are brought to collision, and it covers nearly the whole solid angle around the interaction point [8].

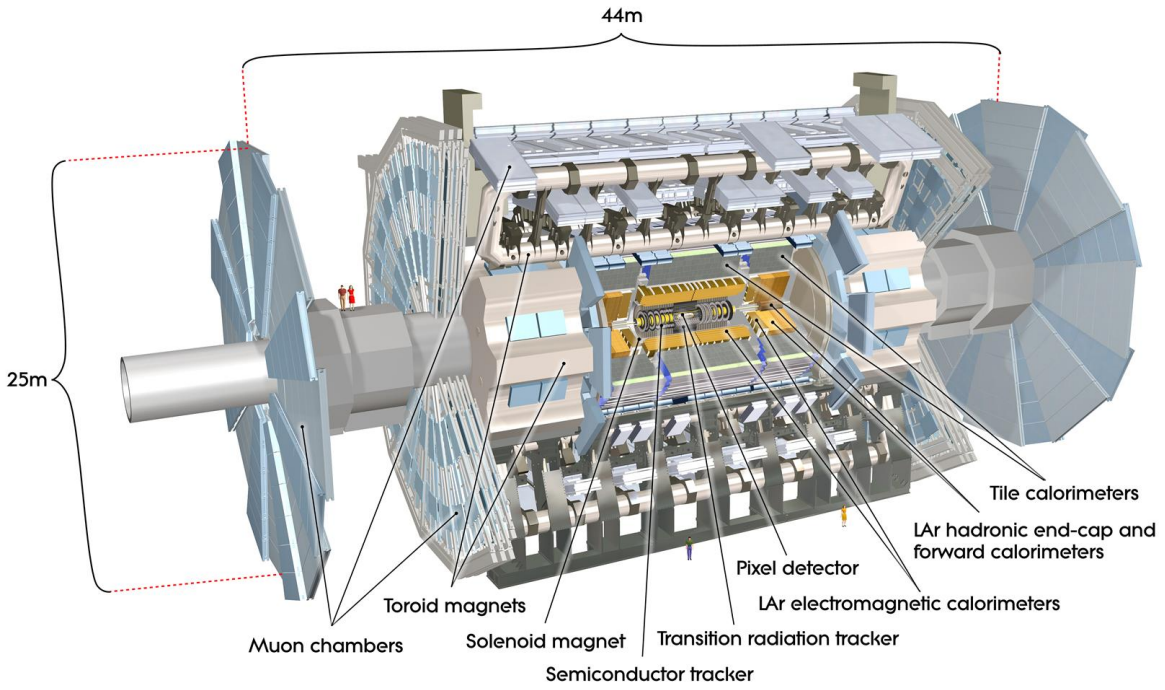


Figure 2.3: Cutaway diagram of the ATLAS detector showing the different subsystems. [9]

Surrounding the inner detector is a solenoid magnet providing the 2 T magnetic field B, which bends the trajectories of charged particles in the tracker with a bending radius r, so that the momentum p of these particles can be determined via $p[\mathrm{GeV}/c] = 0.3 \cdot B[\mathrm{T}] \cdot r[\mathrm{m}]$. The magnetic field for the outer muon system serves the same purpose and is provided by a superconducting toroidal magnet design, giving the ATLAS detector its name. Besides the magnets and the detectors, the Trigger and Data Acquisition (TDAQ) system and the Detector Control System (DCS) are also integral parts of the detector.

2.2.1 The Detector Layers

The inner detector consists of three subsystems, which together provide the tracking, momentum, and vertexing information. The innermost layer is populated by silicon pixel modules and has the highest accuracy of $10 \times 115\,\mu\mathrm{m}^2$ as it is closest to the interactions [8], with accuracies being noted in

R-ϕ and z. Around the pixel detector are the also semiconductor-based layers of silicon strips (SCT), which have a lower accuracy of $17 \times 580\,\mu m^2$ as they are further out from the beam pipe [8]. Both silicon detectors provide hit measurements of radius R, azimuthal angle ϕ, and displacement in beam direction z in the cylindrical coordinates. The outer part of the inner detector, the transition radiation tracker (TRT), consists of 4 mm straw tubes filled with an Ar and CO_2 mixture. As the straws are not subdivided lengthwise, the TRT only provides R and ϕ information, but the high number of straws leads to around 36 hits per track, whereas the pixel detector and the SCT contribute approximately 3–4 and 8–9 hits per track, respectively. Through transition-radiation photons the TRT additionally helps with electron identification [8]. In 2014 an additional layer of new pixel modules, the Insertable B-Layer (IBL), was installed between the beam pipe and the former first layer of the pixel detector [10].

The calorimeters are densely instrumented detectors, which measure the energies of particles by absorbing them fully as they manage to contain the showers these particles initiate. The electromagnetic calorimeter (ECAL) measures the energy of electrons and photons and the hadronic calorimeter (HCAL) measures the energy of hadrons. The ECAL is placed closer to the beam and is surrounded by the HCAL as it does not correspond to a full hadronic interaction length λ. The ECAL has a thickness of approximately 23 radiation lengths X_0 and the whole calorimeter has a thickness of approximately 10 hadronic interaction lengths λ [8]. In the barrel region, the ECAL consists of liquid argon (LAr) as active detector material and lead as absorber material. The tile calorimeter for the hadronic part of the barrel consists of steel as absorber material and scintillating tiles as active material. In the end-caps, both the ECAL and HCAL are based on liquid argon.

The muon spectrometer is the outermost layer of the detector as the muons traverse both the tracker and the calorimeter systems. It has dedicated chambers for two purposes. One purpose, the precise tracking of the muons, is realized with the monitored drift tube chambers and the cathode-strip chambers (multi-wire proportional chambers). The other purpose of the muon system is triggering. The resistive plate chambers (a gaseous parallel electrode-plate detector) and the thin gap chambers (also multi-wire proportional chambers) are utilized in the lowest level trigger, L1, so they have to provide the information on muon tracks very fast. With that, they also provide reliable bunch-crossing identification with a probability of $\geq 99\,\%$ [8].

2.2.2 The Material Distribution of the Detector

The material distribution or geometry is an important consideration for the design and operation of a detector. Interactions of the particles to be measured with inactive and even with active material of the detector complicate their reconstruction, so these effects and the material of the detector itself need to be well understood. Most of the secondary particles in the ATLAS detector are produced by nuclear interactions of primary particles (promptly produced in the proton–proton collisions) with the detector material [11]. Hadronic interactions happen between hadrons and nuclei of the material via the strong force. The other significant effect is the conversion of a photon to a pair of electron and positron, for which a Coulomb field of e.g. a nucleus needs to be present. This is also called gamma conversion.

Correctly modelling the conversion of photons plays an important role in calibrating the electromagnetic calorimeters, which measure the energies of electrons and photons [12]. The track reconstruction efficiency is also very sensitive to these secondary interactions [11]. Studies that quantify the material distribution of e.g. the tracker can utilize these secondary interactions by reconstructing their vertices [11]. An example of this is shown in figure 2.4, where the density of the vertices shows the structure

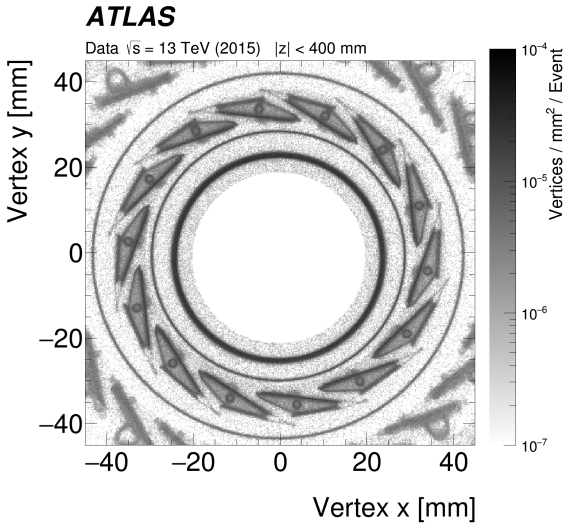


Figure 2.4: Distribution of hadronic interaction vertex candidates close to the beam pipe. [11]

of the tracker close to the beam line. An overview of the different detector and support layers of the pixel and strip detectors is provided in table 2.1.

These effects also play a role in the reconstruction of high-level objects based on the track and energy reconstruction. One example of this, the key part of this work, is the identification of jet flavour. Section 2.4.1 explains how hadronic interactions and gamma conversions can impact jet–flavour tagging by mimicking the characteristics often found in heavy-flavour jets. Another example are searches for new physics looking for the decay vertices of long-lived particles, where regions with a high material density are vetoed [13].

Table 2.1: Definition of the radial regions of the detector material. The corresponding z region is $|z| < 400\,\text{mm}$ for all the radial regions listed. [11]

Radial Region	Radial Range [mm]	Description
BP	22.5–26.5	beam pipe
IPT	28.5–30.0	inner positioning tube
IBL	30.0–40.0	IBL staves (for photon conversion: IPT+IBL+IST)
IST	41.5–45.0	inner support tube
PIX1	45.0–75.0	first pixel barrel layer
PIX2	83–110	second pixel barrel layer
PIX3	118–145	third pixel barrel layer
PSF	180–225	pixel support frame
PST	225–240	pixel support tube
SCT-ITE	245–265	SCT inner thermal enclosure
SCT1	276–320	first SCT barrel layer
SCT2	347–390	second SCT barrel layer
Gap1	73–83	material gap between PIX1 and PIX2
Gap2	155–185	material gap between PIX3 and PSF

2.3 Simulation and Reconstruction

2.3.1 General Overview

Monte-Carlo simulation plays a central role in the physics analyses performed by the ATLAS collaboration and in the whole of particle physics. Without it, the data taken in the detector could not offer nearly the amount of insight which is granted by the comparison. The simulation of signal and background processes enables the validation of models, estimation of backgrounds and the extraction of physical measurements. In the training of flavour-tagging models, the simulation data provides the training, testing and validation samples, as only in simulation the actual truth information is known. This enables the usage of supervised learning techniques. The Monte Carlo data is generated in multiple steps replicating what is happening in the detector [14]. An overview of this is provided in figure 2.5.

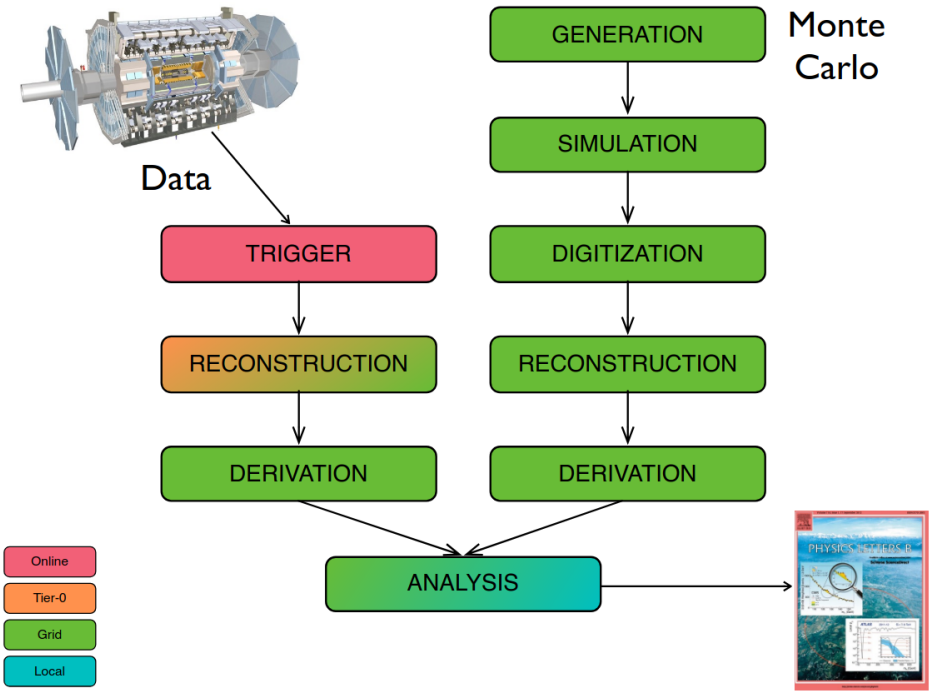


Figure 2.5: The flow of real detector data and Monte-Carlo simulation data up to the publication of a paper in the ATLAS collaboration. The coloured legend indicates at which computing level the different steps take place. [15]

The Monte-Carlo simulation starts with the event generation (referred to as generation). There are multiple generators in use describing the high energy collisions of the protons in the LHC. These describe the particles as four-momenta and handle their interactions and decays to model the physics processes that are intended to be studied. While creating an event, these generators also decide which particles are considered stable in the sense of them not decaying immediately and thereby travelling through the detector. After an event has been generated, the detector simulation (referred to as simulation) models how the stable particles pass through the detector and interact with its active and inactive material as described in section 2.3.2. For this purpose either Geant4 [16] is used or

AtlFast3 [17], which is less detailed, but faster than full simulations with Geant4. The hits in the detector systems provided by the simulation of the particle interactions with the detector can then be digitized. In the digitization the response of the detector systems to these interactions is emulated, noise is added, and the first level trigger is also simulated. After this step, the simulated data has the same format as real detector data and henceforth the same reconstruction can be applied to both. The reconstruction and derivation then reconstruct the real or simulated detector data into objects like tracks and jets and furthermore identify particles when possible, so that the actual physics process initiating the event can be studied. The data formats from both reconstructing actual data or simulation data after the digitization are ROOT files called Analysis Object Data (AOD), which become Derived AOD (DAOD) after derivation, but are still in the same format.

A visualization of the signatures left by different particles in the detector is shown in figure 2.6. The inner detector measures the tracks of charged particles. Electrons and photons initiate showers in the electromagnetic calorimeters, which are then absorbed to measure the total energy. Muons are able to traverse the whole detector, so the inner detector measures a track as well as the designated muon detectors. Neutrinos do not interact with the detector, because they only interact via the weak force and rarely. They can only be reconstructed by the missing energy in the transverse plane as due to conservation of momentum the total momentum in the transverse plane is approximately zero. In the case of more than one neutrino present in the event, an exact reconstruction is therefore not possible without further assumptions. Tau leptons can be identified via their hadronic decays, so they leave a track in the inner systems, and showers in the electromagnetic and hadronic calorimeters. The quarks produced in the elementary process do not traverse the detector, because quarks do not exist in an unbound state long enough. The quarks form bound states in a process called hadronization. One quark from the primary interaction produces multiple hadrons traversing the detector as a cone-shaped spray, a jet. These jets contain charged and neutral hadrons measured primarily in the HCAL, but also photons measured in the ECAL. While the identification and reconstruction of other particles exploit these differences in signatures, it is not trivial to differentiate which kind of particle initiated the jet, a gluon, the hadronic decay of a tau lepton, or the different flavours of quarks. To identify which kind of particle initiated a jet is exactly the task of flavour tagging, which will be discussed in section 2.4. How the jets are reconstructed is described in section 2.3.3 and how the tracks are reconstructed is described in section 2.3.4.

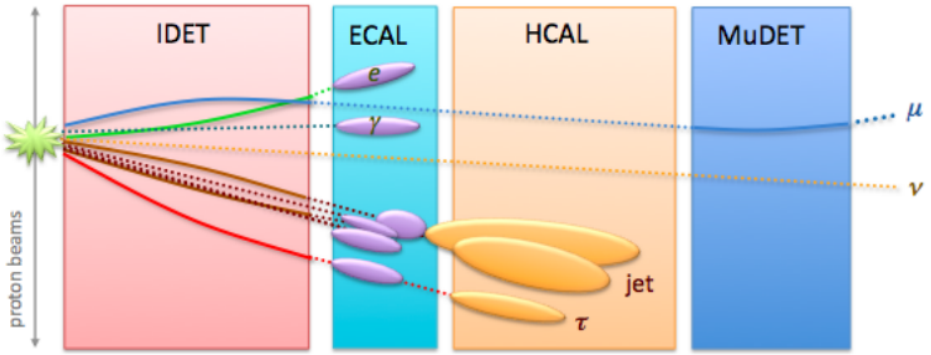


Figure 2.6: Diagram showing the signatures of different particles in the different detector systems of a particle detector. Dashed lines indicate that a particle does not interact with this part of the detector. [18]

2.3.2 Detector Simulation with Geant

The standard ATLAS simulation heavily builds on Geant4, which provides physics models and the infrastructure to transport particles through a geometry. The geometry of the ATLAS detector is constructed in the Geant4 format. But not only are the models and parameters of Geant4 chosen to optimally fit the ATLAS detector, there is also an extensive framework build around Geant4 to integrate the detector simulation into the rest of the ATLAS framework. Particle scoring, the detecting and storing of data from a particle passing through the detector, is done in Athena, which is the ATLAS software framework managing almost all ATLAS workflows. Each subsystem of the detector has its scoring tailored to its own performance. The simulated interactions of the particles manipulate or add to the Monte Carlo truth record, which is already defined during generation. The description of the simulation is based on Ref. [14].

Due to the sheer size of the ATLAS detector and the amount of particles per event, the abundance of secondary tracks produced in the detector simulation is too high. To select only interactions relevant to the studied physics there are strategies put into place, containing rules about which interactions to save. Most of these strategies are applied in the inner detector, which even with these rulings in place makes up the majority of the required file size. The inner detector strategies limit the storage needs of most processes like bremsstrahlung, photon conversion, ionization, hadronic interactions and decays by requiring the energy of the particle initiating that process to be above 500 MeV. There is one strategy for the calorimeter, only storing muon bremsstrahlung vertices if the primary muon has an energy above 1 GeV and the generated photon is above 500 MeV. When these criteria are satisfied, the incoming particle, outgoing particles, step information and the vertex are included in the truth record. Step information refers to how the particles are transported through the geometry model numerically via a stepping algorithm optimized in its parameters to particle type, energy and position in the detector. Besides the above mentioned strategies, other methods also limit the computational overhead. One example is the instant removal of neutrinos as they would require several thousand steps to leave the detector while practically never interacting. Another one are the applied range cuts, which check the expected range of a produced secondary particle beforehand and, if this range is smaller than the cut value, deposit the energy of the secondary particle at the end of the next step of the primary particle. There are a lot more considerations for an effective detector simulation in ATLAS concerning both computational resources and accurate modelling, detailed in Ref. [14].

The models describing the physics of particles interacting with material are often limited to a specific type of particle and energy range. Geant4 combines these models describing various scenarios into a few standard physics lists. The ATLAS collaboration only uses these physics lists provided by the Geant4 collaboration to ensure reproducibility of the results and the use of validated combinations of models, except for transition radiation, which is a crucial part of the tracking and therefore added to the physics lists in use [19]. ATLAS uses the QGSP_BERT, QGSP_EMV, and QGSP_BERT_HP lists, the first being used in the detector simulation production after 2008, the second before 2008 and the third being used for special neutron fluence studies. They contain models like the Quark-Gluon String Precompound model (QGSP) and the Bertini intranuclear cascade model (BERT) for hadronic physics, which give the lists their names. More detailed information on these lists can be found in the Geant4 documentation [20]. Different lists were studied for the ATLAS detector and the choice which to use was made based on how they agree with data. In this study, the following processes are of interest: Hadronic interactions between hadrons and the nuclei of the detector material, gamma conversions into an electron–positron pair in the presence of material, and the decay of long-lived hadrons like K^0_S and Λ.

The detector simulation outputs a hits file including some metadata describing the simulation, the requested truth information, and a collection of hits in each subdetector. Each subdetector selects, processes and records these hits consisting of energy deposited at a certain position and time. The next step in the ATLAS simulation infrastructure is the conversion of these hits into the detector response by the digitization software.

2.3.3 Jet Reconstruction

Quarks or gluons produced in the interaction of a proton–proton collision initiate a collimated shower of hadrons in the detector. The properties of the elementary particles taking part in the primary interaction are inferred from the properties of the jets, which are reconstructed out of the signatures these showers leave in the detector systems. Figure 2.7 shows how a single parton (quark or gluon) of the primary interaction results in a jet of particles, which are measured by the tracker (if charged) and are then absorbed by the calorimeter systems.

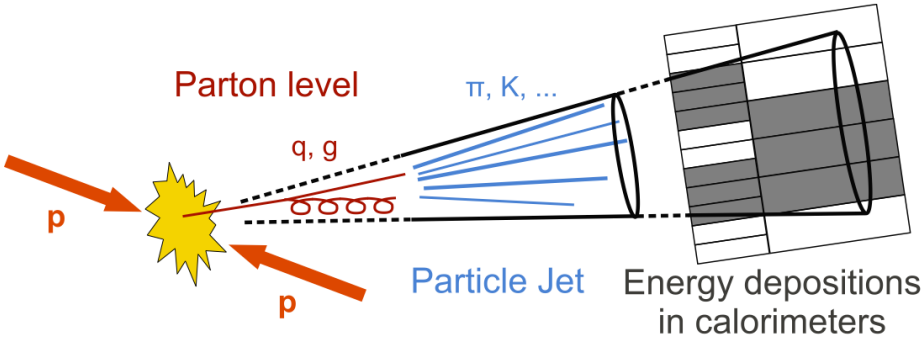


Figure 2.7: Schematic of the different levels of a jet starting as a parton (either quark or gluon), turning into a particle jet of mostly hadrons and manifesting in the detector by the energy deposits in the calorimeter. [21]

The reconstruction of jets begins with the formation of topological clusters from calorimeter signals of connected detector cells, described in Ref. [22]. These topo-clusters try to describe the energy deposition from particle showers in the calorimeter, but they usually do not contain the response to a single particle. They can contain either the full or partial response to a single particle or to multiple particles. The central observable for the clustering process is the cell signal significance defined as

$$\zeta_{\text{cell}}^{\text{EM}} = \frac{E_{\text{cell}}^{\text{EM}}}{\sigma_{\text{noise, cell}}^{\text{EM}}} \tag{2.1}$$

with $E_{\text{cell}}^{\text{EM}}$ being the cell signal and $\sigma_{\text{cell}}^{\text{EM}}$ being the expected noise in this cell, both measured at the electromagnetic (EM) energy scale. Cells with a signal significance larger than a parameter S (default $S = 4$) act as a seed for the proto-cluster. To these proto-clusters, their neighbours with a signal significance larger than parameter P (default $P = 0$) are added, and should they have a signal significance larger than N (default $N = 2$) their respective neighbours are also considered to be added. With this algorithm, the clusters are grown by including the neighbouring cells, which are defined as adjacent cells in the same layer or cells sharing a partial overlap in (η,ϕ) in adjacent layers. The proto-clusters obtained through this method can become too large to provide a good measurement

of the energy, because proto-clusters from different seeds can be merged in the formation process. Thus, after the merging, the clusters are split between signal peaks in a way that one cell can only be shared by two clusters. The resulting topo-clusters of both calorimeters can then be used for the reconstruction of electrons, photons, missing energy and jets.

The majority of ATLAS analyses in Run 1 of the LHC used jets built from topo-clusters [23]. But these jets do not exactly contain the energy of the particle produced in the primary interaction. One main reason for this is the non-compensating design of the ATLAS detector [8], which means that the energy of electrons and photons are measured accurately at the electromagnetic scale, but hadronic showers give a lower signal than the electromagnetic ones at the same energy. Another large contribution is pile-up. Due to the high luminosity at the LHC there are multiple interactions per bunch crossing, not only the hard-scatter interaction of interest. Additional proton–proton collisions in the same bunch crossing (in-time pile-up) or from other bunch crossings (out-of-time pile-up) leave remnant signal in the calorimeter cells. The jets have to be calibrated to parton level with a jet energy scale (JES) correction factor [24]. At the end of Run 1, it was found that taking the tracks associated to the jet into account improved the jet resolution [24].

Particle flow introduces a method to combine the measurements of the tracker and calorimeter, such that the jet is reconstructed from a collection of particle-flow objects and not only the topo-clusters. Utilizing the tracks has the advantages of an improved angular resolution for single charged particles, an improved momentum resolution for low-energy charged particles and an extended acceptance of softer particles, and the advantage of including low momentum particles, which are swept out of the jet cone before reaching the calorimeter [23]. Additionally, the signal of tracks coming from pile-up vertices instead of the primary vertex can be rejected. Thus, the tracker information on charged particles complements the ability of the calorimeters to reconstruct neutral and charged particles. To avoid the double counting of overlapping momentum measurements of the tracker and energy measurements of the calorimeter, the particle flow algorithm subtracts energy from the respective cells in the calorimeter.

The particle flow algorithm is outlined in figure 2.8. The algorithm uses topo-clusters and tracks, which must pass stringent quality criteria [23]. Starting from the largest transverse momentum p_{T}, tracks are matched to a given topo-cluster. The track momentum and the topo-cluster position are used to calculate the expected energy of the particle which created the track. Then the algorithm computes the probability that the particle deposited its energy in more than one topo-cluster and adds these clusters to the consideration. From these topo-clusters, the expected energy of the track particle is subtracted cell by cell and if the remaining energy in the track and topo-cluster system matches the expected energy, the remnants of the topo-cluster are removed. In the end, the algorithm provides a list of tracks and a list of both the modified and unmodified topo-clusters, which together are called particle-flow objects [23].

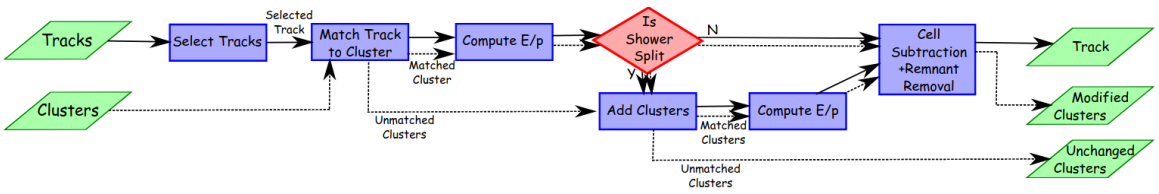


Figure 2.8: Flow diagram of the particle flow algorithm. [23]

Once the constituent objects are built, the jets have to be reconstructed. Important considerations for jet reconstruction are the infrared and collinear safety, which guarantee that either soft gluon emission (infrared) or the splitting of two particles moving in nearly the same direction (collinear) do not change the jet structure. To obtain jet definitions, which are stable and can be compared to theory calculations, the shape of the jets should not be influenced by soft radiation. One such algorithm, used with particle-flow objects as inputs in the jet reconstruction of ATLAS flavour tagging [25], is the anti-k_t jet clustering algorithm, which is described in detail in Ref. [26]. The basic quantities used by these algorithms are the distances d_{ij} between entities i and j and the distances d_{iB} between the entity i and the beam. These distances are defined as

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} , \tag{2.2}$$

$$d_{iB} = k_{ti}^{2p} , \tag{2.3}$$

where k_{ti} is the transverse momentum of particle i, Δ_{ij}^2 is defined as $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, and y_i and ϕ_i are the rapidity and azimuth of particle i, respectively. Aside from these observables, two parameters control this algorithm. The radius parameter R, also used by other jet reconstruction algorithms, can be thought of as the jet radius. The parameter p was added to govern the relative power of the energy versus geometrical (Δ_{ij}) scales with the anti-k_t algorithm using a value of $p = -1$. For $p = 1$ the inclusive k_t algorithm is recovered and for $p = 0$ it corresponds to the Cambridge/Aachen algorithm. The anti-k_t algorithm was an addition to sequential recombination jet algorithms as the two above, and improved on criteria including the infrared and collinear safety. In figure 2.9, a collection of jet reconstruction algorithms is compared on an event with few well separated hard particles and many soft particles at parton level. Coloured regions indicate where the uniformly distributed ghosts are clustered into a jet. In this context ghosts are artificial extremely soft particles, which do not affect the jet reconstruction, but are included to illustrate the regions assigned to a jet.

With the anti-k_t algorithm, hard particles accumulate soft ones long before the soft particles cluster together. Without hard neighbours within $2R$, a hard particle will result in a perfectly conical jet of radius R gathering all soft particles therein. Should two hard particles be within $R < \Delta_{12} < 2R$, the jets will not be perfectly conical. In the case of one hard particle having a significantly higher transverse momentum, the other jet will miss the overlapping part. For equal transverse momenta, the cones would be divided by a straight line and for approximately equal momenta both cones will be clipped. Two hard particles with $\Delta_{12} < R$ will cluster into the same jet. Should one transverse momentum be significantly larger, the jet will be conical, centred on the particle with the larger momentum, but should the momenta be roughly equal a more complex structure arises with a union of cones with a radius smaller than R around the hard particles plus a cone of radius R centred on the final jet. In figure 2.9 the key feature of the anti-k_t algorithm can be seen, being resilient with respect to soft radiation, but flexible with respect to hard radiation. In comparison, the k_t and Cambridge/Aachen algorithms show jagged borders sensitive to the soft particles, because they adapt more to the soft radiation. The SISCone algorithm yields regular single-particle jets, but the composite jets vary more in shape. Apart from these observations, there are also quantitative properties, which show the desirable behaviour of the anti-k_t algorithm related to the area, resummation, computing time and other more specific parameters. A more detailed description of these quantitative properties is found in Ref. [26].

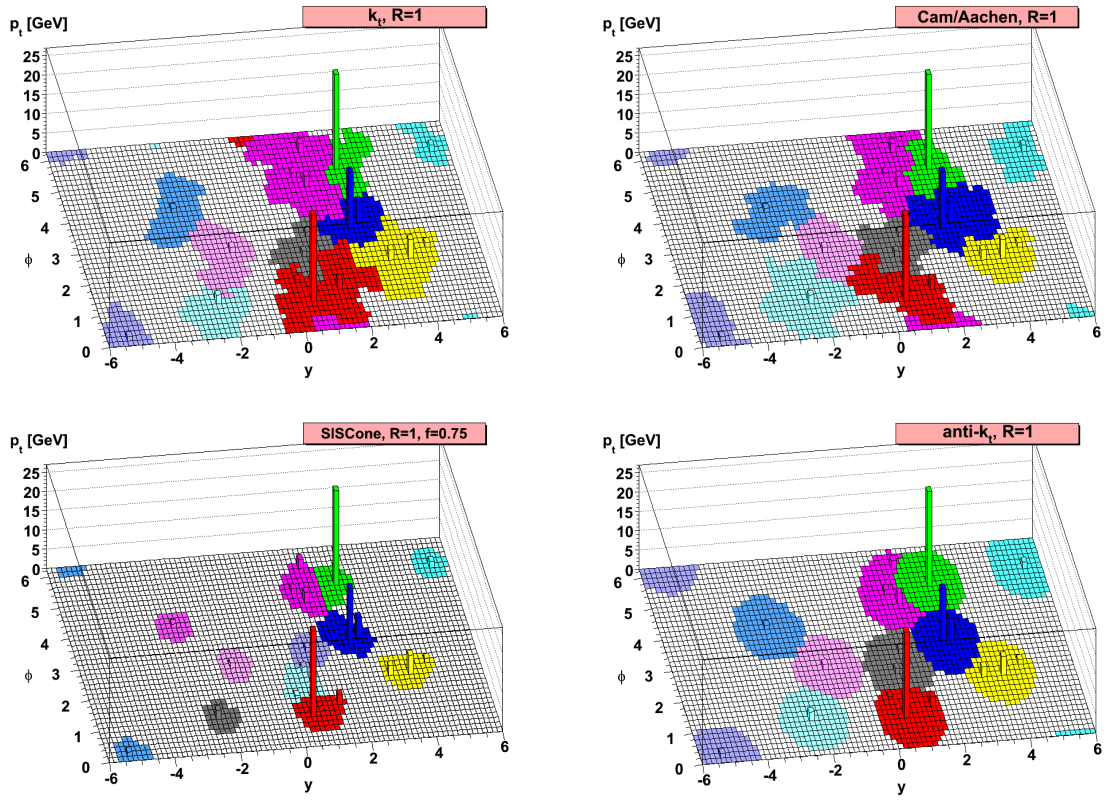


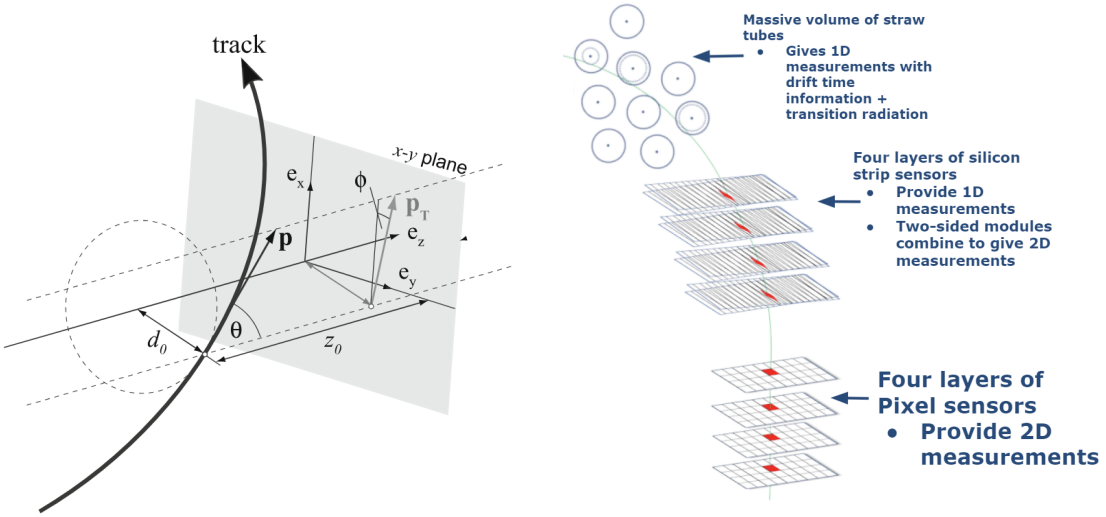
Figure 2.9: A comparison of the k_t, Cambridge/Aachen, SISCone and anti-k_t jet reconstruction algorithms at parton level example data containing few hard particles and many ghosts. [26]

The jets used in this study on flavour tagging are reconstructed from particle flow objects using the anti-k_t algorithm with a radius parameter of $R = 0.4$ and a jet energy scale calibration as described in Ref. [24]. Furthermore, all jets must have a pseudorapidity $|\eta| < 2.5$ and transverse momentum $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, while jets with $p_{\mathrm{T}} < 60\,\mathrm{GeV}$ and $|\eta| < 2.4$ must pass the tight working point of the Jet Vertex Tagger algorithm to suppress pileup [25]. The truth jet–flavour labels are assigned to a jet according to the truth hadrons present within $\Delta R(\text{hadron, jet}) < .3$ of the jet axis [25]. If a b-hadron is present, the jet is labelled as a b-jet. If a c-hadron is present and no b-hadron, the jet is labelled as a c-jet. In the absence of both a b- or c-hadron, but with a τ-lepton present, it is labelled as a τ-jet. The remaining jets are labelled as light-jets.

2.3.4 Track Reconstruction and Association

Tracks are the trajectories of charged particles detected by the inner systems of the ATLAS detector. Due to the magnetic field the charged particles travel along a helical path shown in figure 2.10(a), that can be parameterized by five quantities. The transverse and longitudinal impact parameters d_0 and z_0 are the distances of the point of the closest approach to the reference point, which is the averaged position of the proton–proton collisions. This point also serves for the determination of the azimuthal angle ϕ and the polar angle θ as well as the momentum p of the track. The momentum alone is

not a parameter, but rather the charge divided by the magnitude of momentum q/p. The figure also shows the coordinate system, in which the z-axis lays along the beam line, the x-axis points from the interaction point to the centre of the LHC ring, and the y-axis points upward [25].



(a) Diagram of the global track parameters d_0, z_0, ϕ, θ, and q/p in the ATLAS coordinate system. [27]

(b) Schematic of the hits left by a charged particle in the pixel, strip and TRT detectors. [28]

Figure 2.10: Parameterization of a track in the ATLAS coordinate system and depiction of the hits left by a track in the inner detector.

Tracks are reconstructed from the hits in the inner detector tracker systems, the pixel detector, the SCT, and the TRT, as shown in figure 2.10(b). Signals in the pixel and SCT detectors are grouped into clusters, which are then turned into space-points containing the three-dimensional measurement of a charged particle passing. The track reconstruction starts with the finding of seeds, triplets of space points in the pixel or SCT detector, which have to meet requirements for momentum and impact parameter. These candidates are then expanded to roads, along which compatible clusters are searched for in the other sensors of the silicon part of the tracker. Then the actual trajectory is constructed using a combinatorial Kalman filter [29]. With the track candidates found, overlaps between them and combinations of unrelated cluster (fake tracks), have to be resolved. In the ambiguity resolution, the tracks are assigned a quality. Tracks of lower quality sharing hits with higher quality ones are rejected. But some clusters can be assigned to multiple tracks as there can be topologies denser than the separation power of the sensors. To obtain a high-precision estimate of the track parameters a χ^2 fit is performed. After being fit with the silicon systems, the tracks can be extended into the TRT. They are fit again with a global χ^2 fit as this can provide additional measurements on the track, improving the momentum resolution and adding information for particle identification with the transition radiation effect. If the quality of the fit decreases with the TRT extension, e.g. by too many TRT outliers being present, the extension is rejected. This is referred to as the "inside-out" pass, which is optimized for particles produced in the primary proton–proton interactions.

The "outside-in" pass follows a similar scheme on the hits not included in the first, but unlike the "inside-out" it is performed starting with hits in the TRT, and it is only performed in regions of interest

defined by energy deposits in the electromagnetic calorimeter. Seeds of two space-points are then constructed near these TRT regions and the same road search, Kalman filter, ambiguity resolving and χ^2 fit is applied. This pass increases the acceptance on shorter tracks of particles produced further out from the beam line, e.g. electrons from gamma conversion [30].

Utilizing track information is crucial in the modern high-performing jet–flavour taggers [31]. There are selection criteria applied on possible track candidates, e.g. on their transverse momentum p_{T} or how many hits they left in the silicon part of the detector [25], before they are associated to a jet. There are multiple ways to associate tracks with a jet, one example being the ΔR association used in ATLAS flavour tagging. The angular distance ΔR is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}, \tag{2.4}$$

with the pseudorapidity $\eta = -\ln\tan(\theta/2)$. Tracks within this distance around the jet axis are associated with this jet. Should multiple associations be possible, the track is associated to the jet with the smallest ΔR. The width of ΔR varies with p_{T} of the jet, the maximum being $\Delta R \approx 0.45$ for jets with $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ and the minimum being $\Delta R \approx 0.25$ for jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$ [25].

A selected set of good quality tracks is also used in vertex reconstruction [32]. A vertex is the reconstructed position at which multiple particle tracks originate. There are multiple proton–proton interactions in a bunch crossing, but usually there is only one interesting interaction referred to as the hard-scatter, often having the largest transverse momentum. To isolate this interaction from the other interactions called pile-up and reconstruct it correctly, the primary vertex of this interaction is found on the beam line and reconstructed. Furthermore, the hard-scatter interaction can lead to secondary vertices either from the decay of particles, which are so long-lived that they decay significantly far away from the primary vertex, or from material interactions. As the flavour-tagging models studied in this work do not rely on reconstructed vertices explicitly, vertex reconstruction is not discussed in further detail.

2.4 Flavour Tagging

2.4.1 Introduction to ATLAS Flavour Tagging

Jet–flavour tagging is the effort to identify the type of particle from which a jet originated. Jets can be initiated by the hadronic decay of a τ-lepton, a gluon decay, or a decay of the different quark flavours, except for top quarks as they decay before they hadronize. Usually only the heaviest quark flavours (b and c) are identified, as a complete separation of the different incident particles is quite difficult. The identification of jets originating from heavy-flavour quarks is a very helpful asset to a lot of interesting physics analyses [8]. The top quark, the heaviest elementary particles, almost exclusively decays into a bottom quark and a W Boson, so the precision measurements around the top quark or other kind of physics utilizing the abundance of top quarks at the LHC benefit a lot from good b-tagging. As the coupling of the Higgs boson depends on the particle mass, it frequently decays into heavy-flavour quarks, so the identification of heavy-flavour quarks provides access to Higgs decays with a large branching ratio. As the LHC collides protons with protons, many jets are produced by Quantum ChromoDynamics (QCD) processes, which are not related to the interaction being studied. Good flavour tagging can also help in mitigating these backgrounds [31].

For a long time, b-tagging was the primary focus of flavour tagging, because of their significance in processes mentioned and because differentiating only b-jets from all other jets is easier than differentiating between b-jets, c-jets and all other jets, for example. The basic principles of flavour-tagging algorithms are described in Ref. [31]. The qualities of a b-jet, which can be exploited for the separation, are a result of the relatively long lifetime of the B mesons. Figure 2.11 shows how this long lifetime leads to the hadrons containing a b-quark decaying at a significant distance from the primary vertex. This can manifest itself in a secondary vertex and subsequently in larger impact parameters of the tracks, which originate from the decay of the hadron containing a b. The same holds true, though to a lesser extent, for the hadrons containing c-quarks. Since hadrons containing a b often decay into hadrons containing a c, there might even be an identifiable tertiary vertex inside the b-jet. Heavy-flavour jets also tend to have more tracks than lighter flavoured ones.

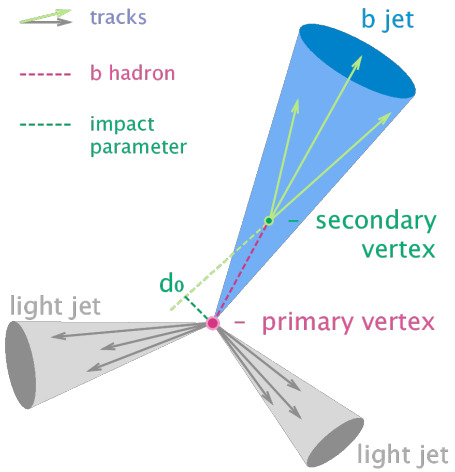


Figure 2.11: A diagram of the characteristics of a b-jet set apart from light jets by the presence of a secondary vertex and large impact parameters of the tracks. [33]

Several flavour-tagging algorithms are based on these features, either relying on the impact parameters of the associated tracks or trying to reconstruct secondary vertices. The IP2D tagger used by the ATLAS collaboration makes use of the transverse impact parameter significance d_0/σ_{d_0} and the IP3D additionally includes the longitudinal impact parameter significance $z_0 \sin\theta/\sigma_{z_0 \sin\theta}$. The log-likelihood ratio discriminants of the algorithms are used as inputs for high-level taggers. Similarly, the outputs of the secondary-vertex-tagging algorithm SV1, which reconstructs a single secondary vertex per jet, are used in high-level taggers. These outputs describe the secondary vertex, e.g. decay length and invariant mass, and are obtained by iterative χ^2 tests on the track-to-vertex matching. These low-level algorithms also contain vetoes to reject secondary particles from K^0_S or Λ decays, gamma conversions, and hadronic interactions with the detector material, because they share the characteristic high impact parameters and secondary vertices of b- and c-hadron decays as shown in figure 2.11. In SV1 this is done by rejecting two-track vertices compatible with K^0_S or Λ decays and only accepting vertices with an invariant mass less than 6 GeV. This rejection and other aspects were supplemented with the JetFitter algorithm, which tries to reconstruct the full b-hadron decay chain by exploiting the topological structure of b- and c-hadron decays inside the jet. The presence of soft leptons in the jet, which arise from semileptonic decays of b- and c-hadrons, is also an additional handle to discriminate heavy-flavour jets, but these jets only account for 20 % or 10 % of b- or c-hadron decays [34].

These algorithms, based on statistical interference, were expanded upon with trainable machine learning models. The RNNIP recurrent neural network tagger [31] could overcome the challenge of IP-based b-taggers, which had to make the assumption that the properties of the tracks in a jet are independent of the other tracks. The emergence of multiple tracks with large impact parameters out of a secondary or tertiary vertex in c- or b-hadron decays intrinsically correlates the properties of these tracks. The RNNIP algorithm, just by its recurrent nature, can learn the sequential dependencies of the tracks, of which it can take a variable amount. The model is fed similar quantities per track as used by the before mentioned algorithms, e.g. impact–parameter significances and distance between track and jet axis, and its outputs correspond to the probabilities of the jet being a b-jet, c-jet or light jet. Thus, with RNNIP the simultaneous tagging of b- and c-jets is already possible. With the output probabilities for the different flavours p_b, p_c, p_{light} and the c-jet fraction f_c a b-tagging discriminant function is defined as

$$D_{\text{RNNIP}} = \log\left(\frac{p_b}{f_c \cdot p_c + (1 - f_c)p_{\text{light}}}\right). \tag{2.5}$$

The c-jet fraction is not the exact relative amount of c-jets in a given sample, but rather a parameter, which governs the relative importance of c-jet and light jet rejection [31]. A discriminant for c-tagging is constructed analogously by switching p_b and p_c and introducing a fraction parameter for b-jets f_b.

The flavour-tagging performance was further enhanced with the use of deep-learning classifiers in the DL1 algorithm series [35], which are fully connected multi-layer feed-forward neural networks. They included the kinematic properties of the jet p_{T} and η and the outputs of the low-level algorithms described above. The samples are resampled such that jet p_{T} and η are uniformly distributed for each flavour class, to avoid that the classifier discriminates the different flavours by the differences in the kinematic distributions. For the training of these algorithms a hybrid sample of $t\bar{t}$ events and $Z' \rightarrow q\bar{q}$ events was used, with $t\bar{t}$ making up 70 % of the events. The outputs of RNNIP were added to the DL1 classifier as inputs to form the DL1r model. In the DL1d model, the RNNIP input was replaced by the DIPS (Deep Impact Parameter Sets) algorithm, which encoded the tracks in a jet in a permutation-invariant way as opposed to the sequential feeding of tracks in a fixed order as in RNNIP. This approach lead to a slight increase in performance, and more importantly, to significantly lower training and evaluation times by a factor of approximately 3 [36]. The discriminants of the DL1x models, which also provide effective flavour probabilities p_b, p_c, and p_{light}, are identical to the RNNIP discriminant in equation 2.5.

In these approaches, the properties of the reconstructed jets and tracks are used in low-level algorithms, either reconstructing physical properties of the jet system as in the vertex fitter or providing probabilities for the different flavours as in the impact–parameter algorithms. Through feeding the outputs of these algorithms into deep-learning models the performance of flavour tagging was enhanced [31]. In recent years a different approach was pursued by the ATLAS collaboration: Feeding the properties of the jets and associated tracks directly into the high-level models [25]. The physics context provided by e.g. the vertex fitters is not supplied to the model, but recovered in auxiliary tasks contributing to the training besides the main task of jet–flavour identification. This end-to-end approach has the practical advantage that only one algorithm has to be developed and maintained, instead of a handful of models. The GN1 model utilizes Graph Neural Networks (GNNs) [25] and the model studied in this work, GN2, follows the transformer architecture [5]. Before providing an overview of these models in section 2.4.3 and discussing their advantages over previous ones, a few of the machine learning concepts are introduced in section 2.4.2.

2.4.2 Machine Learning Concepts

Auxiliary learning is a useful machine learning strategy which aims to provide more problem-specific context to the model. This is done by having the model not only train with the main task, but also additional auxiliary tasks, which are related to the main task. If done right, the addition of these tasks can improve the performance of the models main purpose by acting as an inductive bias [37]. An example can be found in computer vision, where it was shown that depth estimation is a useful auxiliary task for semantic segmentation (labelling each pixel in an image with a semantic class) [38]. In the same work it was found that the other way around, adding semantic segmentation as an auxiliary task to depth estimation, did not help. Thus, challenging a model with additional tasks to improve its performance on the main task is not trivial. There is a lot of research concerned with what tasks should be trained together and how to optimize multitask learning. A typical framework consists of shared layers, to learn a unified representation for all tasks, followed by task-specific layers, while the loss function is a linear combination of each task [37]. The flavour-tagging models detailed in section 2.4.3 follow this typical framework.

Graph Neural Networks are models designed to work on data structures that can be modelled with a graph, which describes a set of objects (nodes) and their relationships (edges). In each GNN layer, the weights of the nodes and edges are updated via message passing between the nodes while not all nodes have to be connected necessarily. Due to the expressive power of graphs, GNNs can model complex relational data to embed or extract the features of the nodes, edges or the entire graph. An important difference to other powerful models which extract features from (locally) connected parts of the data, is the non-Euclidean character of the graph representation, whereas e.g. Convolutional Neural Networks (CNN) operate on Euclidean data like two-dimensional grids (images) or one-dimensional sequences (text) [39]. This has lead to successful applications of GNNs in a variety of fields.

Transformers are another class of deep-learning models, which work with sequential data and are set apart from other models by relying solely on attention mechanisms, while previous models of similar nature paired the attention mechanism with RNN or CNN elements. By getting rid of recurrence and convolution, the neural network becomes highly parallelizable, making it more resource efficient while the transformer also performs significantly better in e.g. translation tasks [40]. The heart of the transformer model is the Scaled Dot-Product Attention. Attention in a machine-learning context determines how important each part of the given sequence is relative to all the other parts. In Scaled Dot-Product Attention, this is realized by a set of queries packed in a matrix Q, a set of keys in matrix K and a set of values in matrix V, which result in the attention according to

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.6}$$

with the scaling factor d_k being the dimension of the queries and keys [40]. Usually, multiple attention heads are used, each projecting the queries, keys and values with weights corresponding to one head. The different heads lead to a multitude of representations, which are then concatenated. Typical transformer models used for example in language processing consist of an encoder and a decoder and contain layer normalization and feed-forward networks after the Multi-Head Attention with residual connections reaching around the attention and feed-forward layers [40].

In section 2.4.3 it will be shown how the application of these machine-learning methods managed to greatly improve the performance of the jet–flavour tagging in the ATLAS collaboration.

2.4.3 Current Flavour Tagging in ATLAS

In recent years, ATLAS flavour tagging moved beyond the hierarchical approach to flavour tagging, in which low-level algorithms extract physical information out of the data, which is then fed into a deep-learning neural network as described in section 2.4.1. One of the first all-in-one taggers is GN1 [25]. It is a Graph Attention Network, a GNN utilizing the attention mechanism. The architecture of GN1 is depicted in a schematic way in figure 2.12. The input to GN1 is done on a track-by-track basis, and in addition to the track properties the kinematic properties of the jet, p_{T} and $|\eta|$, are added to each track. The amount of tracks utilized per jet is limited to 40. If more tracks are associated to the jet than this limit, the first 40 tracks with the largest transverse impact parameter significance are chosen.

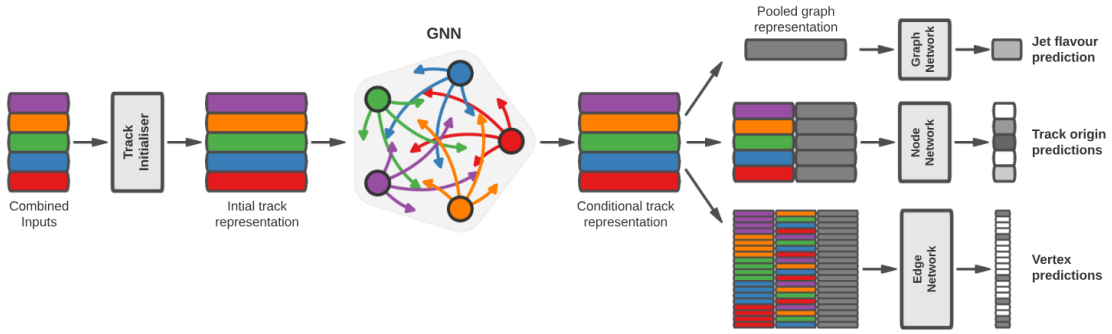


Figure 2.12: The network architecture of GN1. The colours represent individual inputs corresponding to one track, where each track is combined with the jet properties. These track inputs are fed into an initialization network, after which their representations make up the nodes of a fully connected graph. The conditional representations after the GNN are then fed into the networks for the jet–flavour prediction, track–origin prediction and vertex matching tasks. [25]

After an initial feed-forward neural network, the representations of the tracks act as the nodes of a fully connected GNN, so that all the relations between the tracks are considered by the model. With the graph model, a conditional representation of the tracks is acquired, which is then fed into different neural networks for each task. The physics context of the differences in jet structure depending on the flavour, although not provided by low-level algorithms, is recovered with the help of auxiliary tasks. Each task has a designated feed-forward neural network before the output layer, as is typical in the case of auxiliary learning. The main task, jet–flavour prediction, uses the representation of the whole graph to classify the flavour of the jet. The track–origin prediction auxiliary task tries to predict the type of process in which the charged particle matched with the track was produced, so it yields a prediction per track and uses the nodes of the GNN. The vertex prediction auxiliary task does not fit vertices, like JetFitter for example, but it finds vertices, so it outputs a probability per pair of tracks for them to share a vertex, and thus it relies on the edges between the nodes. The different classes of the truth origin are detailed in table 2.2. The truth origin category OtherSecondary is of particular interest for this work since it already partially contains the material interactions and decays of long-lived particles introduced in section 2.3.2. This label and its use for this work is discussed in more detail in section 3.1.

Table 2.2: Truth origin categories serving as classes in the truth–origin classification task in GN1 and GN2. [25]

Truth Origin	**Description**
Pileup	From a proton 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 decay

The auxiliary tasks infer the physics context described in section 2.4.1, e.g. a b-jet having a displaced secondary vertex. The loss functions for the jet flavour and track–origin predictions is the categorical cross entropy and for the vertex prediction it is the cross-entropy loss. The total loss function of the model being minimized is a linear combination of these losses according to

$$L_{\text{total}} = L_{\text{jet}} + \alpha L_{\text{vertex}} + \beta L_{\text{origin}}, \tag{2.7}$$

with the choice $\alpha = 1.5$ and $\beta = 0.5$ to ensure the convergence of the individual losses to similar values, while L_{jet} is expected to be slightly larger than the other two as it is the loss of the main task [25]. An example prediction to illustrate how the model captures the physics context of the jet characteristics used for flavour prediction is provided in figure 2.13 with the prediction of a true b-jet containing not only a secondary, but also a tertiary decay vertex. The jet was correctly predicted as a b-jet and its substructure was also correctly recognized by the auxiliary tasks. The three true vertices were identified by the model, the correct origins were assigned to the tracks, and the model could recognize that the two pileup tracks do not share a vertex with any other track associated to the jet.

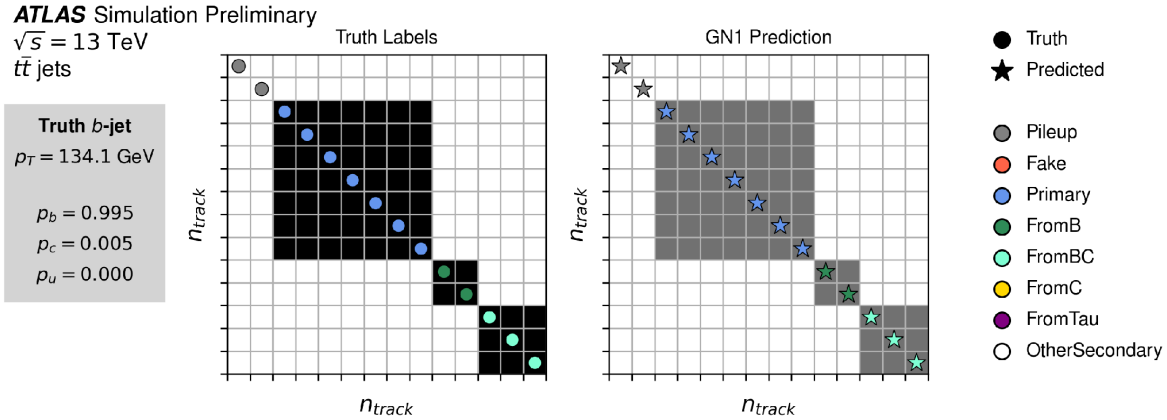


Figure 2.13: One prediction of the GN1 model for a true b-jet. The high p_b score shows that the model could correctly tag the b-jet. The black and grey squares indicate which tracks share a vertex and the colours of the tracks indicate the track origins according to the legend. The vertex and origin predictions were also completely correct in this example. [41]

The GN1 architecture was extended and changed to the closely related GN2 model. A few key changes lead to GN2 being more similar to a transformer encoder. The GATv2 attention was replaced by the Scaled Dot-Product Attention introduced in section 2.4.2, the number of attention heads increased from two to eight, layer normalization with dropout was added, and further changes were done leading to an increase of trainable parameters from 0.8M in GN1 to 1.5M in GN2 [5]. Initially, both GN1 and GN2 provided the jet–flavour probabilities p_b, p_c, and p_{light} and the b-tagging discriminant was constructed according to equation 2.5. However, while this work was performed, the GN2 model was expanded by the probability p_τ for jets from hadronic τ decays and so the discriminant follows

$$D_b = \log\left(\frac{p_b}{f_c \cdot p_c + f_\tau \cdot p_\tau + (1 - f_c - f_\tau) \cdot p_{\text{light}}}\right), \tag{2.8}$$

with f_τ describing the relative τ-jet importance. Overall, GN2 still follows the same layout as GN1 as shown in figure 2.12, and the tasks as well as the loss strategy remain the same. The input is also done in the same way with each track being complemented in its properties by the kinematics of the jet. An overview of the inputs to the model used for this work can be found in table 2.3.

Table 2.3: The inputs to the GN2 model used in this work. The jet inputs are attached to each track being fed into the network.

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)
dη	Pseudorapidity of the track, relative to the jet η
dϕ	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the transverse plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the longitudinal 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
leptonID	Indicates if track was used in the reconstruction of an electron or muon

In order to train this larger model, the amount of training jets was increased from 30M to 192M [5] in the GN2 training. Jets from a $t\bar{t}$ sample are used for the training in the $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ region and jets from a Z' sample for the $250 < p_{\mathrm{T}} < 5\,000\,\mathrm{GeV}$ region. The $t\bar{t}$ events are modelled using the Powheg Box event generator and adding parton shower, hadronisation, and underlying event via Pythia [5]. The Z' events used to populate the higher transverse momentum regime consider a hypothetical heavy Beyond Standard Model (BSM) particle called Z', which can decay into pairs of b-quarks, c-quarks, τ-leptons or lighter quarks [5]. These events are generated with Pythia directly. The decays of b- and c-hadrons are handled by EvtGen, as opposed to other long-lived hadrons, which are decayed in the detector simulation. This difference informs the labelling described in section 3.1.2. Both the $t\bar{t}$ and Z' events are produced at $\sqrt{s} = 13\,\mathrm{TeV}$ and at $\sqrt{s} = 13.6\,\mathrm{TeV}$. Different versions of the mentioned software packages are used for the different kinds of events and at the different energies [5].

In the usual reconstruction of these events for flavour tagging purposes Geant Thinning is applied. Skimming, slimming, and thinning are used to reduce the amount of data in the xAOD format. Skimming removes whole events from the data, thinning is the removal of individual objects and slimming is the removal of variables within a given object. Geant Thinning removes truth particles, which were written in the truth record by Geant, based on certain criteria. However, the tracks reconstructed out of these truth particles are kept. So the jets, which are being tagged, still contain the same tracks. However, for some tracks of detector simulation particles the information about the particle is missing. For this study and other flavour tagging and tracking work, a special data sample was produced, not including the Geant Thinning algorithms, allowing a more detailed investigation of material effects and their influence. The differences between the default reconstruction and the one without Geant Thinning are broadly discussed in chapters 3 and 4.

Individual jets will have different discriminant values (equation 2.8) and will follow a given distribution. Thus, to decide whether a jet is b-tagged or not, certain values have to be used as cuts. If the jet has a discriminant value above the cut, it is considered a b-jet, otherwise not. These cut values are used to define the working points. A handful of these working points are agreed upon and calibrated as all the supervised learning can only be done on simulation data with the truth information present. The usual quantities to evaluate a tagger within the ATLAS collaboration is the efficiency of b-tagging, i.e. how many true b-jets are actually tagged as b-jets, and the rejection of the other classes. The rejection is the inverse of the mistagging efficiency, i.e. the inverse of how many c-jets, τ-jets or light jets, respectively, are falsely tagged as b-jets. Figure 2.14 shows the large improvement in performance with the advent of GN2 over the hierarchical taggers of the DL1 series and over GN1.

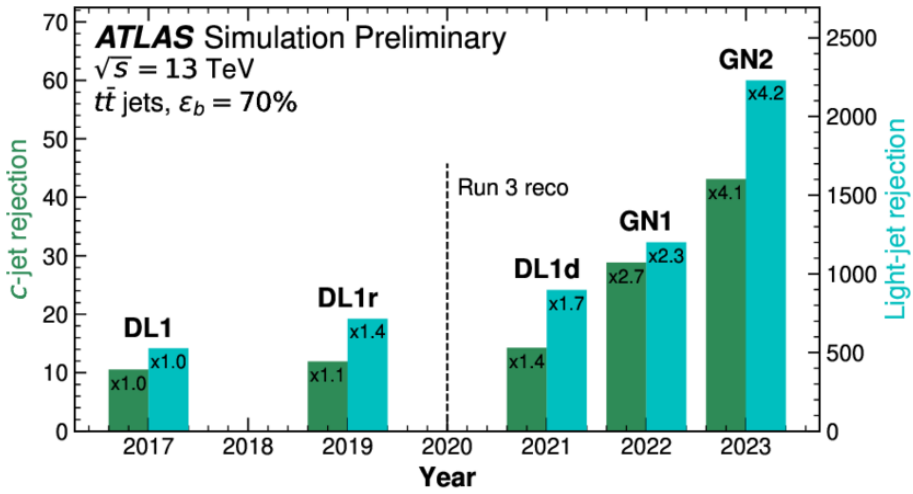


Figure 2.14: Comparison between the performance of the DL1 series taggers and the GN series taggers. All models were evaluated at a b-tagging efficiency of 70 % and the rejections for light and c-jets of DL1 were set as reference for the later models. [42]

CHAPTER 3

Material Interactions in Flavour Tagging

3.1 Secondary Origin Categorization

3.1.1 Motivation of the Categorization

The original motivation of this work was only identifying hadronic interaction tracks in jets, because light jets containing hadronic interaction vertices might be more prone to being mistagged as they share some characteristics central to the tagging of heavy-flavour jets as outlined in section 2.4.1. But this effort was expanded beyond only hadronic interactions to include all interactions, which are managed by the detector simulation. These interactions are commonly referred to as secondaries. This might be confusing as the decay vertices of heavy-flavour hadrons (containing a b or c) are also referred to as secondary vertices, but for the sake of brevity this work will use secondaries to refer to effects modelled by the detector simulation. Three distinct types of secondaries are depicted in figure 3.1, showing the similarity to the features of b-hadron decay shown in figure 2.11.

The core idea is for the flavour-tagging model to learn to consider the tracks of secondary origin in the prediction of jet flavour less. The low-level algorithms used for secondary-vertex reconstruction directly veto tracks of secondary origin as described in section 2.4.1. In this work, the model is expected to learn this itself, which is further detailed in section 4.1. In order for a model to be trained by supervised learning techniques, the corresponding truth information has to be present, so the track-by-track input to the model has to be expanded by a truth label. A labelling scheme has to be put in place, which should put tracks into categories which differentiate if the track is of a secondary origin or not. This was not realized as a binary classification, however, as a few common secondaries are distinguished, instead of only making up one category. Additionally, there is a category of tracks not from a secondary origin, and there is a category combining fake and pileup tracks. Motivating the distinction of hadronic interactions and gamma conversion is straightforward. These effects are only possible in the presence of detector material and can therefore be seen as nuisances mimicking the behaviour of heavy-flavour decays in light jets. The categories of long-lived particle (LLP) decays have a less physically sound motivation as the characteristic features of actual heavy-flavour jets are a result of the longevity of hadrons containing a b or a c. Bottom and charm hadron decays are not handled by the detector simulation, the LLP decays belonging to the secondaries are decays of strange hadrons. Additionally, not all strange hadrons are decayed by the simulation, most strange hadron decays are handled by the generator just as bottom and charm hadron decays. This depends on the

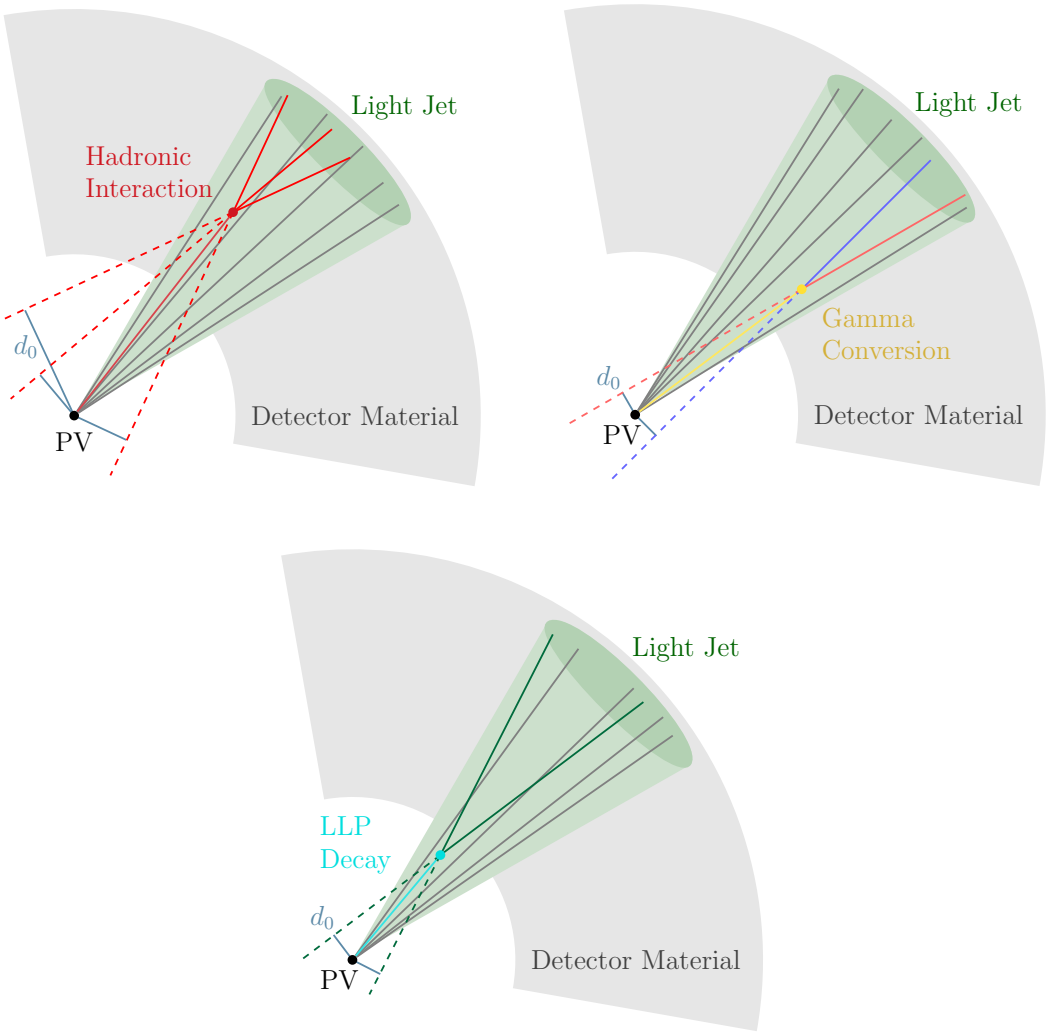


Figure 3.1: Three examples of secondaries, which might be found in a jet: Hadronic interaction, gamma conversion and the decay of a long-lived particle (LLP). They each have a secondary vertex displaced from the primary vertex (PV) leading to tracks with high impact parameters d_0.

lifetime of the strange hadrons. Two prominent strange hadrons with a long lifetime are K_S^0 mesons and Λ baryons. In the beginning of the categorization effort, both of these strange hadrons had their own category, while in later labelling schemes these two categories were broadened to include all strange meson or baryon decays performed by the simulation. Figure 3.2 shows how far away from the beam line certain hadrons are decayed either by the generator or by the simulation. In this and all following figures of the same style, statistical uncertainties are represented by shaded bands around the histogram lines.

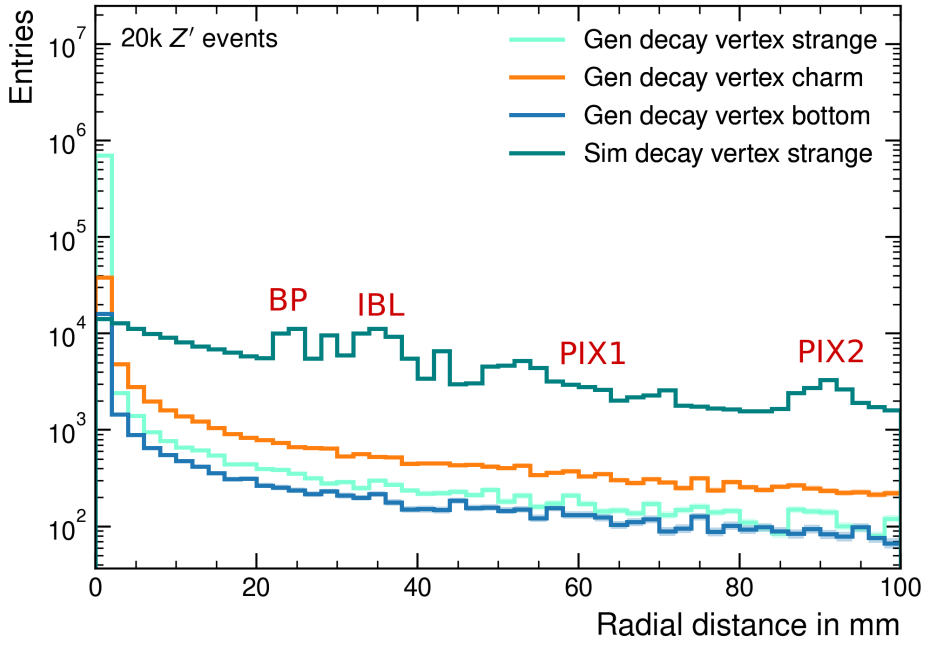


Figure 3.2: Radial distance of decay vertices to the beam line for different hadrons. Hadrons are split up depending on their quark content and whether they have been decayed by simulation or generation, where only strange hadrons are decayed by simulation.

Strange hadrons are decayed by the simulation in both material interactions (most probable hadronic interactions) and as decays-in-flight, which refers to the standard decay modes of strange hadrons. This becomes apparent in the figure as the decay spectrum of the strange hadrons decayed by the simulation is enhanced at positions of the detector material, which were introduced in section 2.2.2. The distinction between hadrons decayed by the generator versus hadrons decayed by the simulation is the foundation of this categorization, that is also why the decay-in-flight of strange hadrons done by the simulation has corresponding categories in the labelling scheme, although physically there is no clear distinction to the decay-in-flight of strange hadrons done by the generator. Nonetheless, because of the distinct decay modes of the strange mesons and baryons, it is also not completely unphysical to distinguish them. Another reason to include the simulation decay-in-flight processes is that the categorization is not perfect, as is further described in section 3.1.2.

3.1.2 Origin and Secondary Origin Labelling

A look at the current flavour-tagging models GN1 and GN2 described in section 2.4.3 shows that track origins are already utilized. The track–origin label, found in table 2.2, already provides truth information to the model, which is used for the training of the track–origin prediction auxiliary task. In the track–origin label, there already is a category OtherSecondary present, in which tracks of secondary origin should be placed. Thus, one possible implementation of a classification of secondary origins would be to make the OtherSecondary category more granular to differentiate between hadronic interactions, gamma conversion and decays-in-flight of long-lived particles. But a closer look at how this label is determined reveals, that working with the track–origin label will not suffice for the purposes of this work. Before describing the labelling scheme of the track origins and the secondary track origins, which are implemented separately, figure 3.3 already shows that an expansion of the origin label would not have sufficed. The figure shows how the ftagTruthOriginLabel is distributed in the ftagTruthSecondaryOriginLabel, a first attempt at a labelling of secondaries. The depicted secondary origins contain not only tracks which are of the OtherSecondary category in the track–origin label, there are significant contributions from the categories FromB, FromBC, FromC and minor ones from the other origin categories.

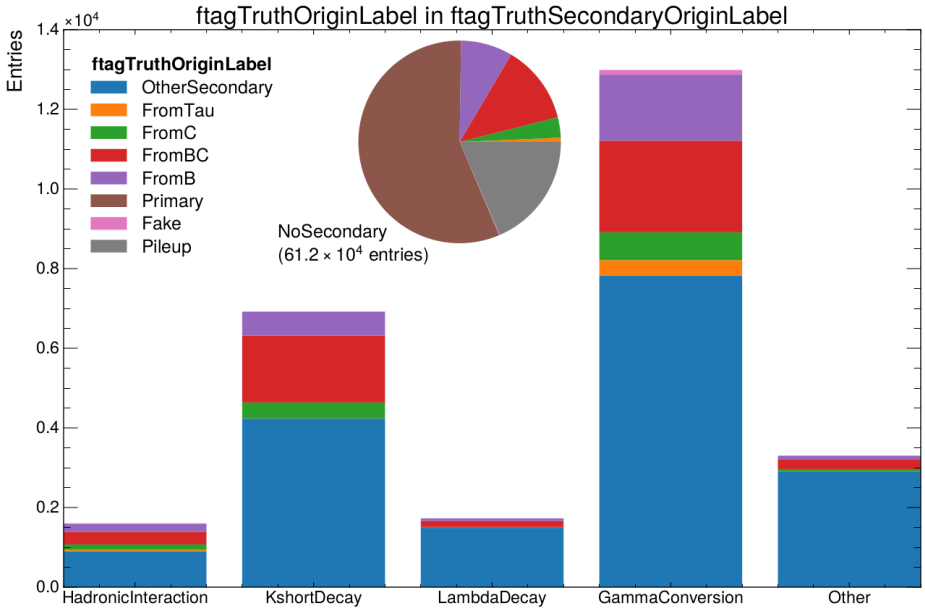


Figure 3.3: The distribution of track–origin label categories in the secondary track–origin categories. The NoSecondary category of the secondary origin label is displayed as a pie rather than another bar, because of the abundance of tracks in that category.

The secondary origin categories depicted in figure 3.3 were expanded by a NoTruth category, and the KshortDecay and LambdaDecay categories were extended to the more inclusive categories StrangeMesonDecay and StrangeBaryonDecay. The label for the secondary origin of tracks was renamed to ftagTruthSourceLabel (also referred to as track source) in the development of this categorization, because it is shorter and less prone to confusion with the origin label, but it is important to keep in mind that in principle it is also concerned with the origin of a track. To understand the source label itself and the differences to the origin label, an in-depth description of how both labels are acquired is necessary.

The main question in labelling the origin of a track, secondary or not, is which part of the decay chain determines the origin. The labels have to be exclusive as the existing origin prediction task in the GN2 is not a multi-class classification and the addition to the model was chosen to follow the same design. As an example, a bottom hadron can decay and one of its decay products can interact hadronically with the detector material. Which origin should be assigned to the product of this hadronic interaction depends on what the label is used for. In the case of the track–origin label the track should be labelled FromB in this example, in the case of the source label it should be HadronicInteraction. Both labelling schemes are realized in Athena, the main software repository used by the ATLAS collaboration. Two packages within Athena are utilized for this, the InDetTrackSystematicsTools package and the FlavorTagDiscriminants package. The former is more generally used, also by other working groups, while the latter is specifically used in the derivations of the data used for the flavour-tagging efforts. For the purposes of training and validating the models GN1 and GN2, the data in the DAOD format (Derived Analysis object data) are converted into HDF5 (Hierarchical Data Format version 5), abbreviated as h5. An overview of both labelling schemes is provided in figure 3.4.

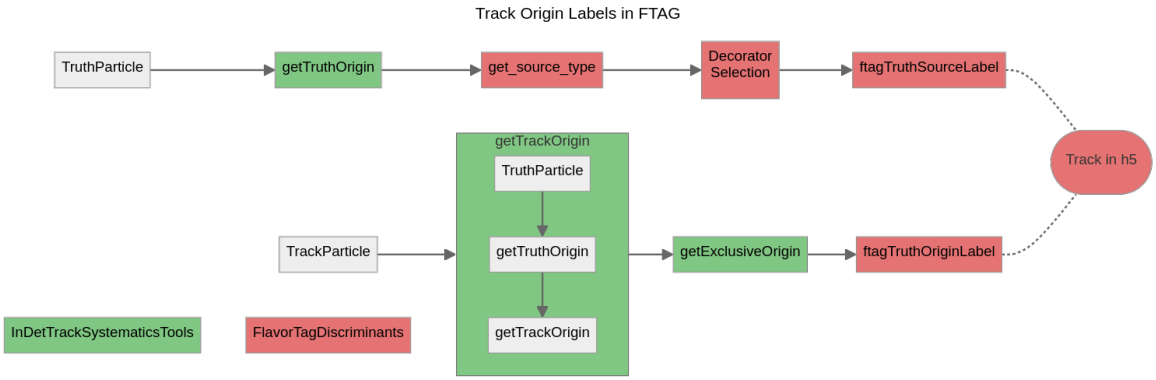


Figure 3.4: Overview of the track–origin and track-source labelling process utilizing the InDetTrackSystematics-Tools (green) and FlavorTagDiscriminants (red) packages of Athena.

Labelling starts with two fundamental object types, which are used in the ATLAS software framework, TruthParticle and TrackParticle, which contain the truth record information used. If the truth information of the particle is present, the TrackParticle contains a link to the corresponding TruthParticle it is associated with. The origin label relies on the TrackParticle and the getTrackOrigin method to determine the track origin. The getTrackOrigin method uses the TruthParticle object and getTruthOrigin method, which is shown in figure 3.5, but expands it by having additional functionality for pileup studies and by categorizing tracks as fake, if the matching probability is not sufficient. The source label directly uses the getTruthOrigin method as this suffices for this purpose. In this stage, the origin is not exclusive, but multiple categories are stored in a bitwise flag. Different origins are recorded by this flag, each corresponding to a digit in the bit, which is set to 1 if the respective conditions are met. The possible origins are: the secondaries KshortDecay, StrangeMesonDecay, LambdaDecay, StrangeBaryonDecay, GammaConversion, OtherDecay, HadronicInteraction, OtherSecondary; the decays BHadronDecay, DHadronDecay, and TauDecay; Fragmentation; OtherOrigin, which is also limited to particles from the detector simulation like the secondaries; Fake and Pileup, which are added in getTrackOrigin for TrackParticles only.

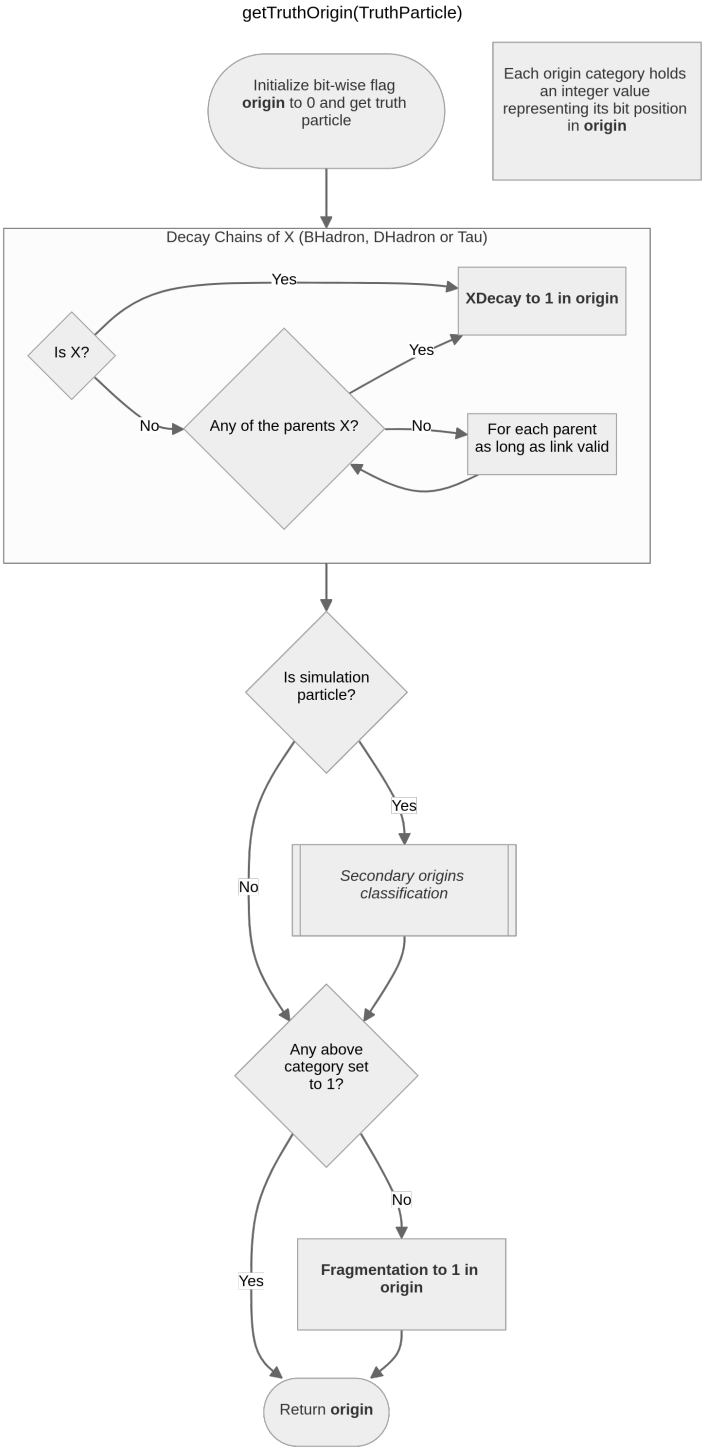


Figure 3.5: Flow diagram showing how the getTruthOrigin method constructs an origin for a provided TruthParticle. The secondary origins classification is found in figure 3.6.

The first categories to be checked in getTruthOrigin are BHadronDecay, DHadronDecay, and TauDecay by simply looking at the parent particle. These three categories are set apart from the others as only for them the whole decay chain upwards of the particle, which is currently being categorized, is scanned. Thus, the bit corresponding to one of these three decays is set to 1, respectively, even if the particle is not a direct daughter of such a decay, but further down the decay chain. Additionally, if a particle is a result of a charm hadron decay, which in turn was the daughter of a bottom hadron decay, both categories BHadronDecay and DHadronDecay are set to 1. After these three decays it is checked if the particle is a simulation particle or not. If yes, then the secondary origin classification differentiates the different secondary categories. If none of the origin categories conditions were met, the bit corresponding to Fragmentation is set to 1 in the origin.

The secondary origin classification shown in figure 3.6 is part of the getTruthOrigin method and is implemented such that the secondary origins are exclusive with respect to each other while the general origin is not, as already discussed. There are two exceptions to that, with StrangeMesonDecay and StrangeBaryonDecay being more general than KshortDecay and LambdaDecay, and so a particle belonging to the KshortDecay category is also in the StrangeMesonDecay category, for example. As the source label will only use the more general categories, the origin flag is still exclusive concerning the utilized categories of that label. If the parent of the particle is not in the truth record, no assumption about the origin of the particle can be made and therefore OtherOrigin is set to 1. GammaConversion is set to 1 in the case of the parent of the particle being a photon and the particle itself being an electron, which also includes positrons as they leave a similar signature as electrons. If a GammaConversion was not found to be true for the particle, the decay-in-flight categories are checked. Generally, the classification differentiates between decay-in-flight and hadronic interaction by using the number of children of the parent particle, two children corresponding to a decay-in-flight and more than two corresponding to hadronic interactions. If the parent of the particle has two children, it is checked whether it was a strange meson or baryon and, if either applies, StrangeMesonDecay or StrangeBaryonDecay is set to 1. Furthermore, if StrangeMesonDecay or StrangeBaryonDecay is set to 1, and if the particle itself is one of the most probable decay products of a K_S^0 or Λ, the KshortDecay or LambdaDecay categories are set to 1. In the case of the parent not being a strange hadron, but still having two children, the OtherDecay bit is set to 1 in the origin flag. Should none of the decay-in-flight categories apply, it is checked whether the parent has more than two children, then HadronicInteraction is set to 1, or not, then the particle is sorted into the OtherSecondary category.

The way the classification differentiates decay-in-flight and hadronic interaction by the number of children of the parent is a weakness of the secondary origin classification. There are decay channels of hadrons with more than two children and with the way hadronic interactions are handled in the simulation, there can be only one or two children emerging in the truth record, although, in reality, hadronic interactions tend to involve more particles as the hadron interacts with a nucleus [11]. Over the course of this work, it was tried to fix this weak point and replace this condition by a more physically accurate solution. An improvement on this condition was planned to be along the lines of how decay-in-flight and hadronic interaction were separated in the material study of the ATLAS detector [11], from which figure 2.6 was taken. In this material study, the mass of the secondary vertex m_{SV} is used to veto K_S^0 or Λ decays for hadronic interactions, for example. In a decay-in-flight, the mass of the secondary vertex m_{SV} should be equal to the masses of these strange hadrons $m_{K_S^0}$ or m_Λ within a certain mass error margin, when calculating m_{SV} under the assumption that the two daughter particle are pions in the case of K_S^0 or a proton and a pion in the case of Λ. No assumptions on the type of daughter particles would have to be made in the context of the truth labelling, as they are known within the truth record. Implementing a similar condition in the secondary origin classification

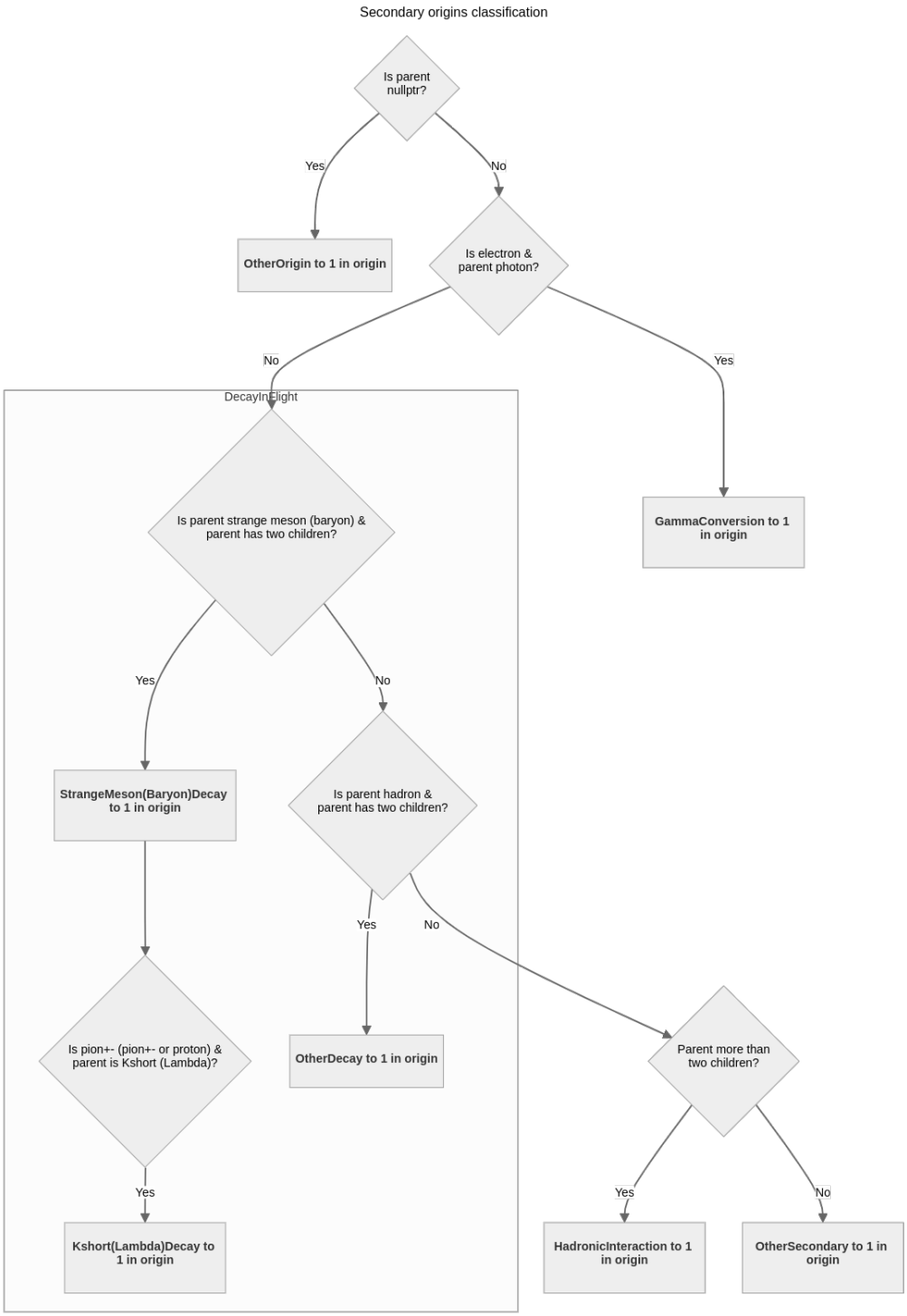


Figure 3.6: Flow diagram showing the classification of secondary origins inside the getTruthOrigin method.

was not possible, at least that late in the reconstruction and derivation flow. The problem with the implementation was that not all daughter particles, or at least their truth information, were present for the decays handled by the detector simulation. One part of this is the Geant Thinning described in section 2.4.3. To realize a more accurate condition to separate decay-in-flight and hadronic interaction processes, a categorization has to occur earlier in the code, but this was beyond the scope of this study.

Once the origin flag is built, either by getTruthOrigin directly in the case of the source label, or by getTrackOrigin in the case of the origin label, the bitwise origin flag has to be converted into an exclusive label. For the origin, the method getExclusiveOrigin serves this purpose. This method prioritizes the categories in the order of how they are listed in table 2.2, e.g. a particle with an origin including both BHadronDecay and HadronicInteraction interaction is assigned the FromB category, as FromB is listed before OtherSecondary in the table. There is one special case, the FromBC category, which is for particles coming from decay of a charm hadron, which itself was a daughter of a bottom hadron decay. The FromC category only includes daughters of charm hadrons, which were not daughters of bottom hadrons. This separation is important as the tertiary decay vertex of a charm hadron inside a b-jet can be a helpful feature for the identification of b-jets as described in section 2.4. Thus, a particle for which the bit in the origin flag corresponding to BHadronDecay is 1, but the bit for DHadronDecay is 0, ends up in FromB, while a particle for which both are 1 ends up in FromBC, and a particle only ends up in FromC, in the case of BHadronDecay being 0 and DHadronDecay being 1. This prioritization leads to particles with secondary origins only being placed in the OtherSecondary category when none of the origins before applies and results in what is shown in figure 3.3. Hence, the origin label does not capture all secondaries in its OtherSecondary category and is not useful for the study of the influence of secondaries of flavour tagging. Especially since secondary origins are already very rare, as is seen in section 3.2, loosing a third of the tracks of secondary origins to more abundant categories in the origin label would be counterproductive for the studies performed in this work.

For the source label, the origin flag is made exclusive by the get_source_type method added to FlavorTagDiscriminants as part of this work. This method is based on the mutually exclusive secondary categories (the more inclusive StrangeMesonDecay and StrangeBaryonDecay are used), so no prioritization between them has to be implemented. Particles without truth information are assigned to the NoTruth category. Particles, which are not from detector simulation processes, are assigned to the NotSecondary category. Then the HadronicInteraction, StrangeMesonDecay, StrangeBaryonDecay and GammaConversion all correspond to the same category in the origin flag and the categories OtherOrigin, OtherDecay, and OtherSecondary are put together in the Other category in the source label. Unfortunately, a bug was introduced during this work, which lead to particles with the OtherSecondary bit set to 1 in the origin flag not being considered simulation particles. The bug was only found and fixed after the studies were done. It leads to the very low number of tracks in the Other category of the source label as seen in section 3.2.

The origin labelling is applied to the TrackParticle, while the source labelling is applied to the corresponding truth particle. Thus, a decorator is needed, which decorates the track with properties of the TruthParticle, including the source label. The decorator includes an additional condition, which sorts tracks into NoTruth, independent of the origin flag of the TruthParticle, if they are not stable or have a p_{T} below 500 MeV and are not a charm hadron.

The source label categories and their classification are summarized in table 3.1. The truth information contained in this label is used to study the effects that tracks of secondary origin have on jet–flavour tagging in section 3.2 and are utilized in an effort to improve the performance of the ATLAS jet–flavour taggers, as detailed in chapter 4.

Table 3.1: Overview over the track source label classifying tracks into NoTruth, NotSecondary, and different secondary origins, which are highlighted by being written in a bold font.

Source	Condition
NoTruth	no association with truth particle possible (e.g. PU)
NotSecondary	associated with truth particle, but not a simulation particle
HadronicInteraction	parent has a number of children > 2
StrangeMesonDecay	parent is strange meson, parent has 2 children
StrangeBaryonDecay	parent is strange baryon, parent has 2 children
GammaConversion	parent is photon, particle is electron
Other	a simulation particle, but not in the above categories

3.2 Characteristics of Jets Containing Secondary Interactions

With the truth information regarding secondary origins of tracks put in place by the implementation of the source label, tracks of secondary origin and jets containing these tracks can be studied. The motivation for this work was that light jets containing these secondary processes share the characteristics of heavy-flavour jets used in their identification: the secondary vertex, the subsequently larger impact parameters of the tracks, and a larger number of tracks. Thus, the secondary processes classified by the source label might contribute significantly to the misidentification of light jets as heavy-flavour jets. To confirm if that is indeed true, some characteristic quantities are investigated.

The track impact–parameter is one such quantity. In figure 3.7, the distributions of track impact parameters of tracks associated with jets are shown. On one hand the distinction between these distributions between tracks in light jets and tracks in b-jets is shown, on the other hand the distinction is drawn between tracks of generator particles and tracks of simulation particles. The difference in the d_0 distribution is even more pronounced for the latter distinction. Thus, the tracks being of secondary origin or not has more discriminative power than the tracks being part of a heavy-flavour jet or not with respect to the impact parameter. This effect might seem exaggerated in figure 3.7, where each histogram is normalized to unit area. There are significantly fewer tracks of simulation particles than of generator particles.

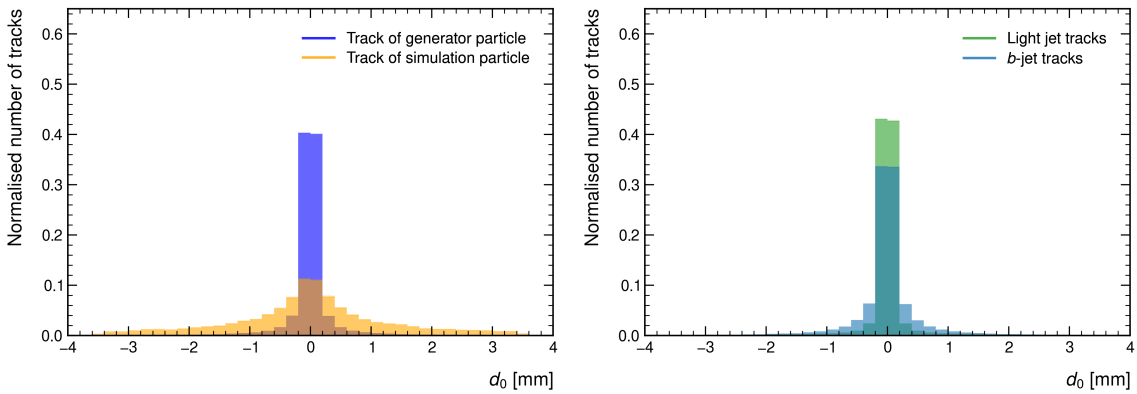


Figure 3.7: Track impact–parameter distributions, the left plot distinguishing tracks of a generator or simulation particle and the right plot distinguishing tracks in a light or heavy-flavour jet.

Another characteristic feature of heavy-flavour jets, the larger number of tracks, is depicted in figure 3.8. Similar distinctions are made as for the track impact–parameters, only that this feature is of jets themselves. Thus, to quantify the effect of phenomena modelled by the detector simulation, the jets are distinguished by having no track of secondary origin at all (only NoTruth and NotSecondary) and having at least one track of secondary origin. Comparing this distinction to the one between light and heavy-flavour jets shows a very similar separation for both.

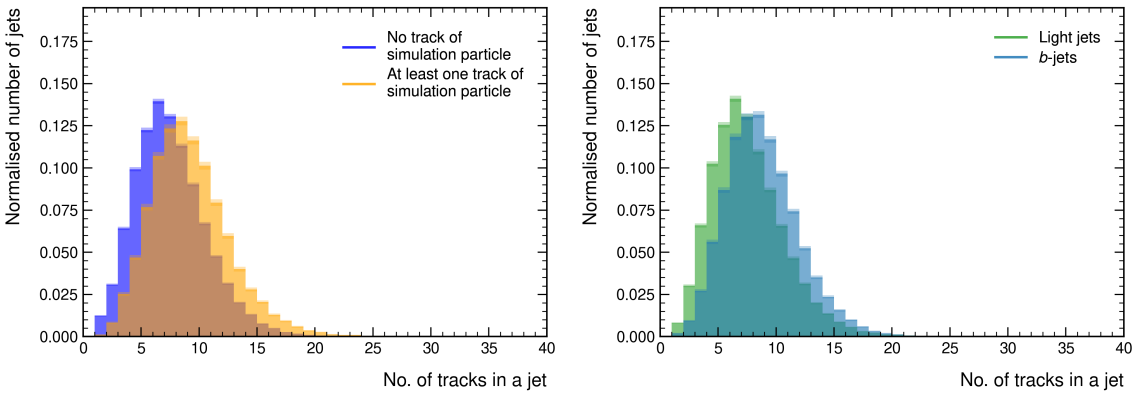


Figure 3.8: Distribution of the number of tracks in a jet, the left plot distinguishing jets with or without a track of a simulation particle and the right plot distinguishing light or heavy-flavour jets.

At high p_{T}, the efficiency of b-tagging is decreased, mainly due to a deterioration in track reconstruction, especially in the jet core [43]. An additional complication is that at higher transverse momentum there are also more tracks of secondary origin. This can be seen in figure 3.9, which shows the relative amount of secondary tracks in light jets for the $t\bar{t}$ and the Z' samples introduced in section 2.4.3. The $t\bar{t}$ sample is dedicated for a lower p_{T} range, while and the Z' sample covers a higher p_{T} range. In the case of $t\bar{t}$, the special sample without Geant Thinning is also available. The expected effect of jets containing more secondary tracks when they have a larger transverse momentum is very noticeable for the $t\bar{t}$ No Geant Thinning sample (NGT). For this sample, the fraction of secondary tracks in the jet keeps increasing with p_{T}, while for the default sample the fraction remains constant. This difference is to be expected as the secondary tracks, for which the truth information about them being secondary has been thinned, are not counted as secondaries, but are still counted for the total amount of tracks in the jet in the default sample. Because no NGT sample was produced for a comparison to the default Z' sample, it is unclear if for very large p_{T} the relative amount of secondary tracks really decreases slightly, as indicated by the figure. Additionally, only reconstructed tracks, which also pass the selection criteria and are associated to a jet, are taken into account. For higher p_{T} jets the reconstruction efficiency of secondary tracks is also decreased and the increased amount of secondary phenomena contributes to the denser environment of tracks in the jet.

The expectation that light jets containing tracks of secondary origin are similar to heavy-flavour jets in their characteristic features, is fulfilled. Thus, the question remains, how the currently deployed model deals with jets containing these secondaries. Figure 3.10 shows the GN2 b-tagging discriminant introduced in section 2.4.3, but without the τ classification, as this was added to the model later. The five plots show how the light-, c-, and b-jets are distributed in the discriminant with the distinction between jets containing no secondary track and jets containing at least one secondary track of each category. The categories contain KshortDecay and LambdaDecay as this was investigated when the

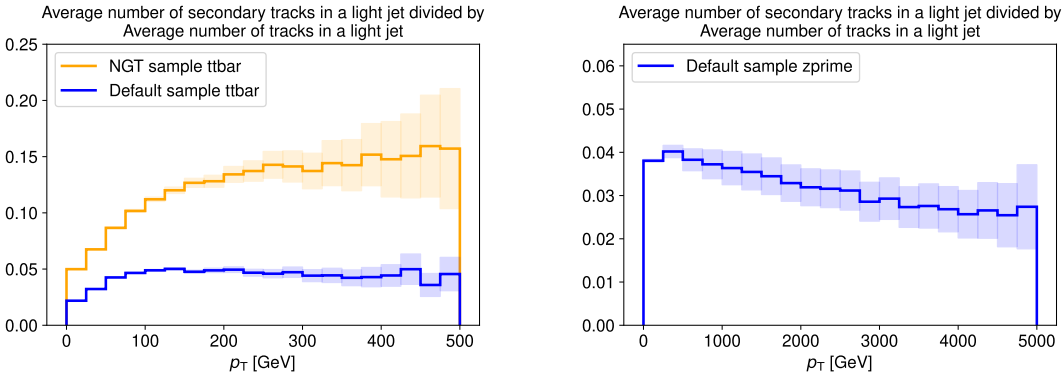


Figure 3.9: Relative amount of secondary tracks within jets as a function of p_T of the $t\bar{t}$ (left) and Z' (right) samples. The default $t\bar{t}$ sample is compared to the one without Geant Thinning (NGT).

label was in a preliminary form. Clearly the model can separate the different jet flavours better when there are no secondary tracks present in the jet, independent of which kind of secondaries. The smallest deteriorating effect on the separation of the flavours comes from the GammaConversion source and the largest from the HadronicInteraction source as seen by how far apart the two discriminant distributions are for jet with and without secondaries. A possible reason for these differences are the signatures of these effects and how similar they are to a decay of a heavy-flavour hadron. Gamma conversions have a very clear signature, while hadronic interactions can differ quite a lot between each other. It is also important to keep in mind that GN2 is a large transformer model and thus capable of picking up these relations without making them explicit. Thus, the model might already have learned to distinguish between the secondary vertex of a b-hadron decay and a gamma conversion. The idea of this work was to adapt the model such that it picks this context up more directly as detailed in chapter 4. The original expectation that light jets containing secondaries appear as b-jets to the model is quite clear in the plots: The distribution of light jets containing secondaries is shifted to the right compared to light jets not containing secondaries. In addition to that, the figure shows that b-jets with tracks of secondary origin are also harder to identify as b-jets, probably because their signature is not as clear as for b-jets without secondary origin tracks.

Figure 3.11 shows the abundance of tracks in the different source categories. There are very few tracks in the secondary categories as most are NotSecondary or do not have the truth information relevant for the categorization available and thus end up in NoTruth. The low number of secondary tracks will present a challenge in chapter 4. The NGT sample holds up to expectation of containing more relevant truth information as it contains fewer tracks in the NoTruth category and significantly more tracks in the HadronicInteraction and GammaConversion categories, approximately three times the amount. Here, it is important to note, that the depicted data was acquired with the faulty derivation containing a bug discussed in section 3.1, which leads to tracks with the OtherSecondary bit set to 1 in the origin not being considered secondary, although they should be labelled as Other in the source label. Thus, there should be more tracks in the Other category. However, figure 3.11 is representative of the data used in the studies detailed in chapter 4. In an earlier stage of this work, when the source label was called the secondary origin label, the amount of tracks in the Other category was approximately the same as in the HadronicInteraction category, but since the secondary origin label used KshortDecay and LambdaDecay over the more general StrangeMesonDecay and StrangeBaryonDecay, more tracks

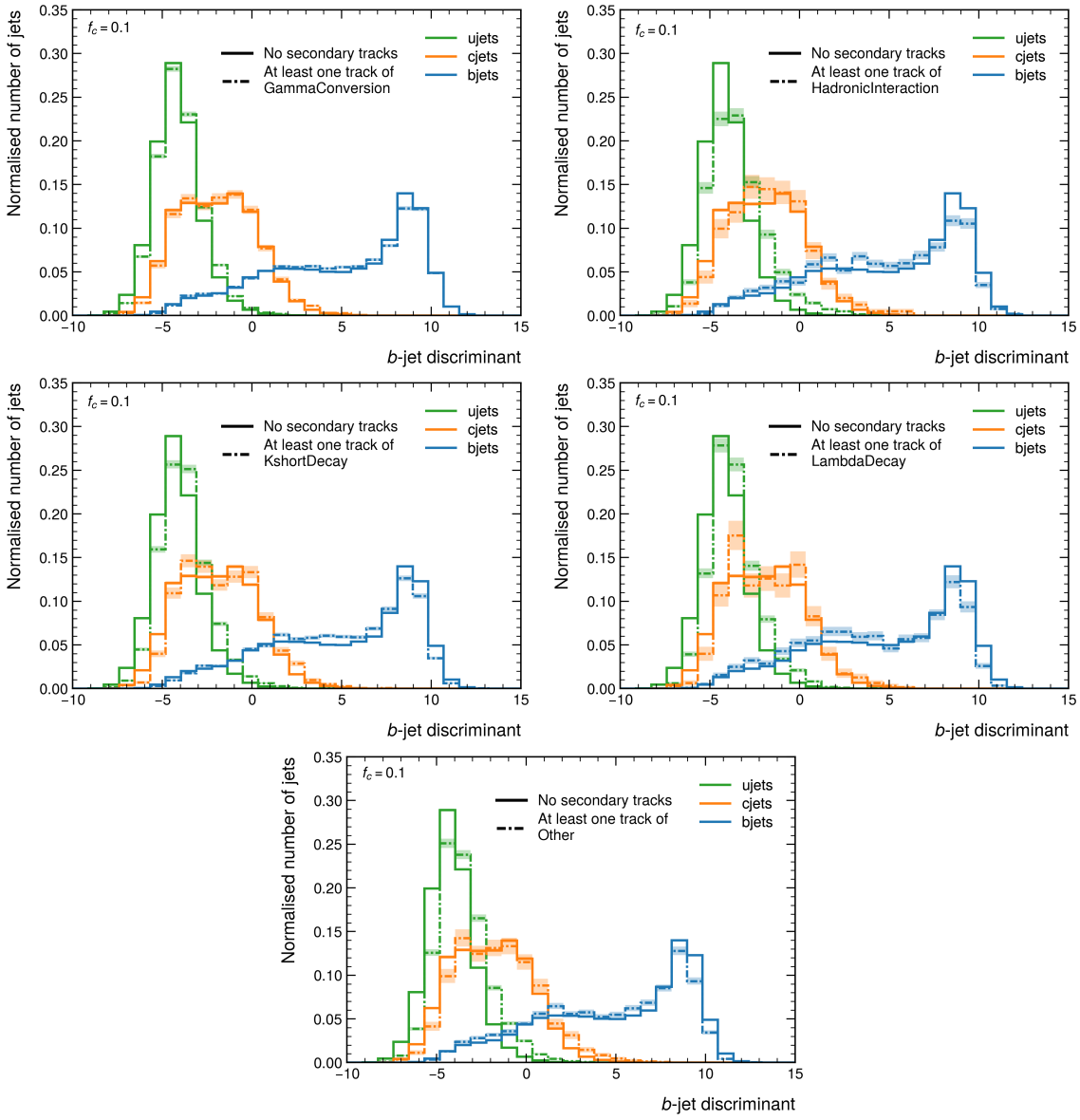


Figure 3.10: The b-tagging discriminant of the GN2 model for light jets (ujets), c-jets (cjets), and b-jets (bjets). The five distributions show jets containing no secondary and jets containing at least one secondary of the categories GammaConversion, HadronicInteraction, KshortDecay, LambdaDecay, and Other.

are classified as Other as they would for the source label. Taking all secondaries together, a bit over 20 % of all jets in the default $t\bar{t}$ sample include at least one secondary track. For the NGT $t\bar{t}$ sample it is up to 40 % of the jets, as can be seen in figure A.1 in the appendix.

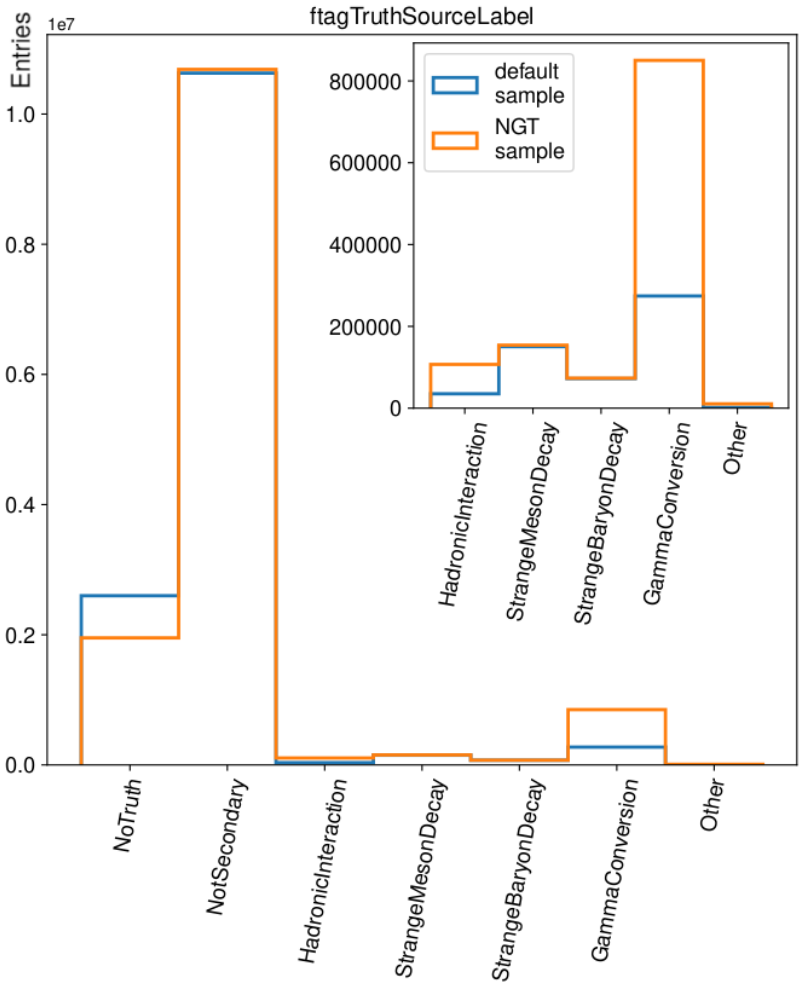


Figure 3.11: Total number of tracks in each of the source label categories with approximately 13.8 million tracks of the default and NGT reconstruction of the $t\bar{t}$ sample respectively. The inset plot shows the secondary categories in more detail as they contain very few tracks compared to the NoTruth and NotSecondary categories.

Also concerning the abundance of secondary tracks, an interesting feature of the jets would be a difference in how many tracks of a certain secondary category they contain depending on the flavour of the jet. In the appendix, figure A.2 and A.3 show this for the $t\bar{t}$ and Z' sample respectively. Light, c-, and b-jets only show minor differences in this respect and with the overall small amount of secondaries it is not probable that these differences play any role in the distinction between jet flavour. The τ-jets are very different from the other flavours in this regard, because they also differ from the other kind of jets in general, being more narrow and having low track and energy multiplicities [44]. These two figures also show that significantly more jets contain secondary tracks in the Z' sample, especially of the material interaction categories HadronicInteraction and GammaConversion, due to the higher transverse momentum.

To make sure that the No Geant Thinning reconstruction does not differ from the default reconstruction used for flavour-tagging purposes, except for containing more truth information and thus improving the secondary origin track labelling, the input variables of jets and tracks for GN2 were compared. Figure 3.12 shows this comparison for the jet input variables, transverse momentum p_{T} and pseudorapidity η, and additionally compares the number of tracks per jet and the number of secondary tracks per jet. As expected, the two input variables and the total number of tracks per jet show no significant difference, while the number of secondary tracks does. The twenty track input variables are compared in a similar fashion in the figures A.4, A.5, A.6, A.7, and A.8 in the appendix. They also do not show significant differences in the physical observables.

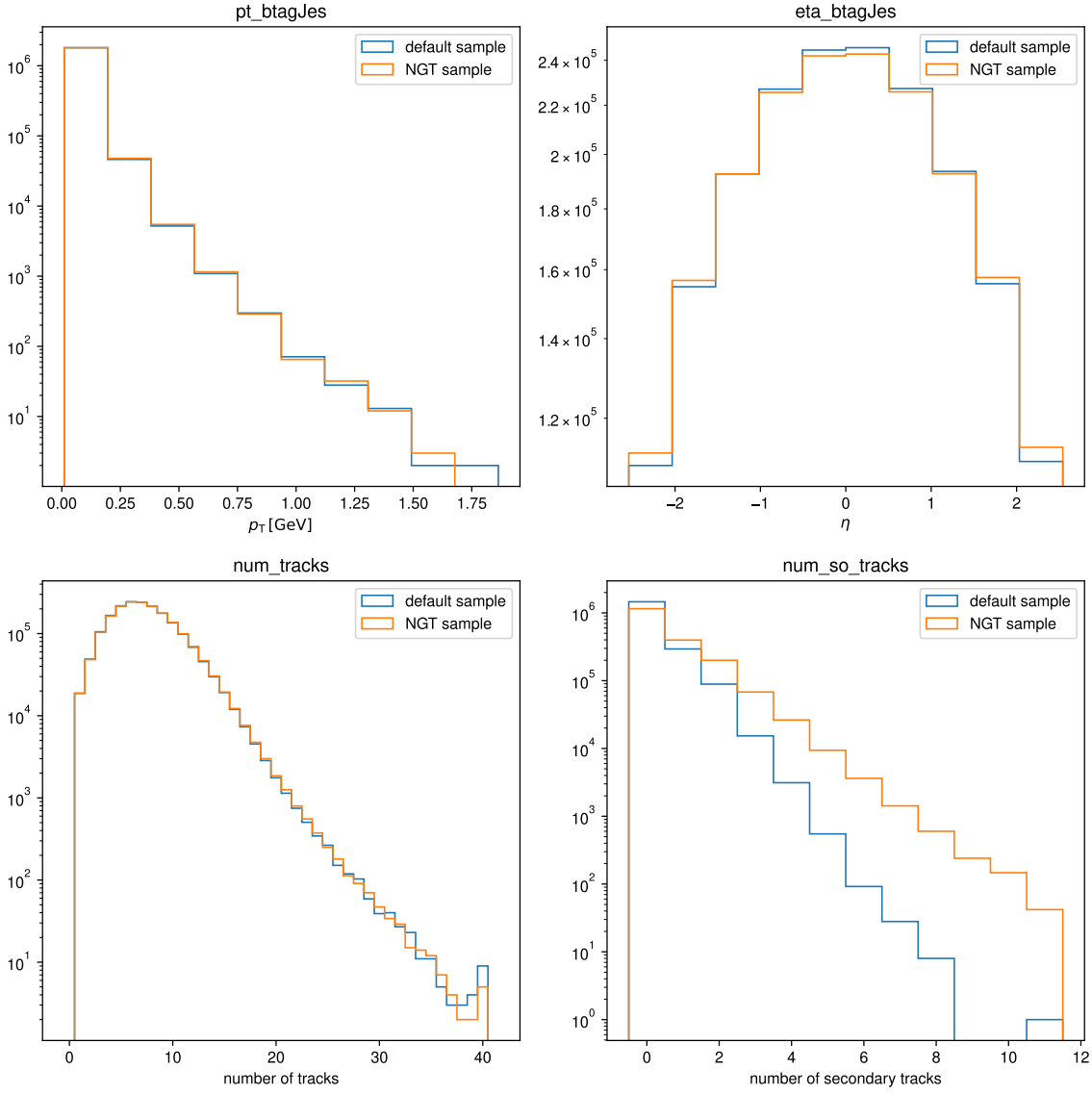


Figure 3.12: Histograms of the two jet input variables p_{T} and η for GN2, the number of tracks per jet, and the number of secondary tracks per jet, comparing the default and NGT $t\bar{t}$ sample.

Overall, jets containing tracks of secondary origin share the same characteristics that set the heavy-flavour jets apart and make them easy to tag. The deteriorating impact of secondary origin tracks in jets on the ability of the model to differentiate between the flavours is observed in the b-tagging discriminant. Additionally, it is shown how rare tracks of secondary origin are, which can be improved by using data without the Geant thinning applied. With the additional truth information available in NGT data, approximately three times more tracks coming out of material interactions, i.e. hadronic interactions and photon conversions, can be correctly identified as such. It is also shown that, apart from the additional truth information, the NGT data is identical to the one going through the default processing of the data.

CHAPTER 4

Adding Secondary Origin Classification to Flavour Tagging

4.1 Expanding the Model Architecture and Retraining

With the labelling scheme for secondary tracks in place and the expected behaviour of jets containing them confirmed, an approach to mitigate the impact of these secondary effects can be worked out. The approach studied extensively in this work is the addition of a further auxiliary task to the GN2 model as depicted in figure 4.1. Section 2.4.3 already discusses how auxiliary tasks can be used to inject physical context into a learning model directly. The two already implemented auxiliary tasks supporting the jet–flavour prediction are the vertex prediction task, trying to match tracks pairwise, and the track–origin prediction. The new track–source prediction auxiliary task shares an identical architecture to the track–origin prediction task due to their close relation, which was discussed in detail in section 3.1.2. The added task classifies tracks into the source categories that can be found in table 3.1: NoTruth, NotSecondary, HadronicInteraction, StrangeMesonDecay, StrangeBaryonDecay, GammaConversion, Other.

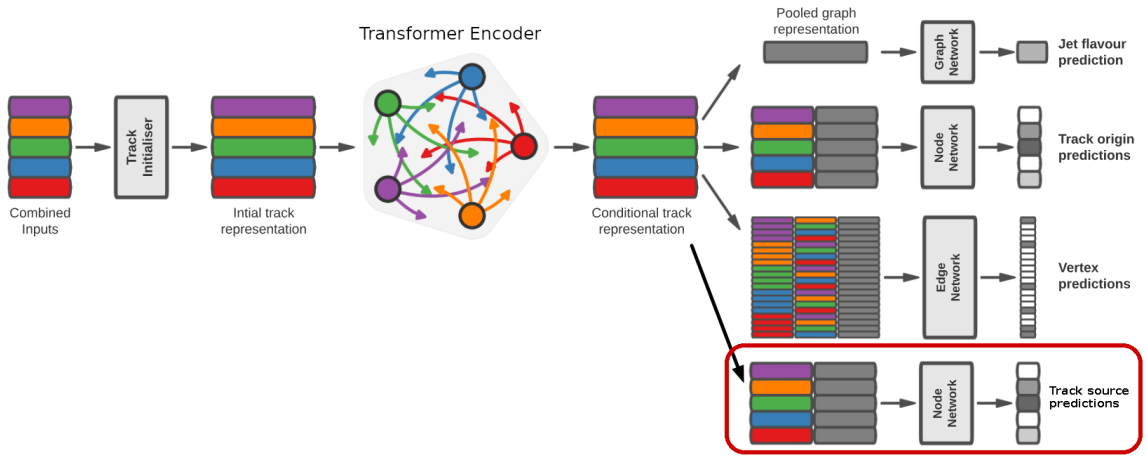


Figure 4.1: The network architecture of GN2 expanded by the source auxiliary task indicated in the red box. The colours represent individual inputs corresponding to one track (combined with jet properties).

The input of the model consists of the jet variables, which are concatenated with the variables of the up to 40 leading tracks associated with the jet. Both, the jet and track variables, are listed in table 2.3. The tracks are ranked according to absolute track impact–parameter significance. These combined inputs are then fed into a per-track initialization network of a single hidden layer and an output layer of 256 nodes. Next is a transformer encoder, which consists of eight layers with eight attention heads and an embedding size of 256. The output of the transformer is projected down to 128 dimensions and provides a conditional per-track representation after the initial track representations could incorporate information of the other tracks within the jet. A global representation of the jet is achieved by attention pooling and this global jet representation as well as the conditional track representation are then provided as inputs to the networks of the specific tasks. Each of the task networks consist of three hidden layers with sizes 128, 64, and 32 respectively. The main task of jet classification only uses the global representation of the jets and has four outputs p_b, p_c, p_u, and p_τ for the tagging discriminants. The auxiliary tasks also use the track embeddings. The origin–prediction task has eight output categories for the origins in table 2.2, the vertexing task has a binary output layer, and the added source–prediction task has seven output categories. The SiLU activation function is used across the model and the tasks use cross entropy loss. This implementation of the GN2 model is not identical to the final version of GN2 [5] as this work was done while GN2 was being finalized and work on GN3 [45] already started. For example, the final GN2 version still uses the ReLU activation, while the SiLU activation used here was a change introduced in the development of GN3. But, very significant changes to what kind of inputs are used and which tasks the models contained have been introduced later in the GN3 development, so that the model investigated here is still similar to GN2.

The total loss function of the model being minimized is a linear combination of the task losses according to

$$L_{\text{total}} = L_{\text{jet}} + \alpha L_{\text{vertex}} + \beta L_{\text{origin}} + \gamma L_{\text{source}} \tag{4.1}$$

with the choice $\alpha = 1.5$ and $\beta = 0.5$ as described in section 2.4.3. The loss of the source task is added and brings with it the additional parameter γ. Different values for this parameter were studied in this work and it will be referred to as task weight (TW) in the following. For the source task weight γ, choices of 0.5, 1, and 2 were tested. The parameters α, β, and γ should be chosen such that the individual losses are approximately equal and that the loss of the jet–flavour prediction task being slightly larger than the rest, since it is the main task of the model.

In addition to the weight of the source–prediction task, the weighting of the different categories inside the source task loss is also important. Contrary to the origin prediction, where no weights are applied to the different origin categories, different weighting schemes are applied to the source classes. Either the source classes are not weighed, they are weighed according to their relative abundance, or they are weighed with the square root of the factor of their relative abundance. These schemes will be referred to as CWnone, CWfull, and CWsqrt. Section 3.2 showed how imbalanced the track source classes are and that the interesting secondary categories contain few tracks. Thus, if no weighting strategy is put into place, it is a good strategy for the model to simply classify all the tracks as NotSecondary or NoTruth. The different strategies will yield models with significantly different performance in the source–prediction task.

To adequately judge the impact of the source auxiliary task on the model, a model within identical architecture, but without the source task, has to be trained alongside the model being tested under the same conditions and with identical data. The data flow in the Monte-Carlo simulation process was presented in section 2.3.1. The flavour tagging specific part begins with special derivation

formats taking the AODs into the DAOD format with Athena software packages, part of which were introduced in section 3.1.2 for the truth labels assigned within the derivation. Once the DAODs are available, the algorithm development in flavour tagging is decoupled from the ATLAS analysis software frameworks and further processing of the data is done with software developed and maintained by the flavour-tagging group, which is tailored to the needs of the development process. Figure 4.2 shows the workflow within the flavour-tagging environment.

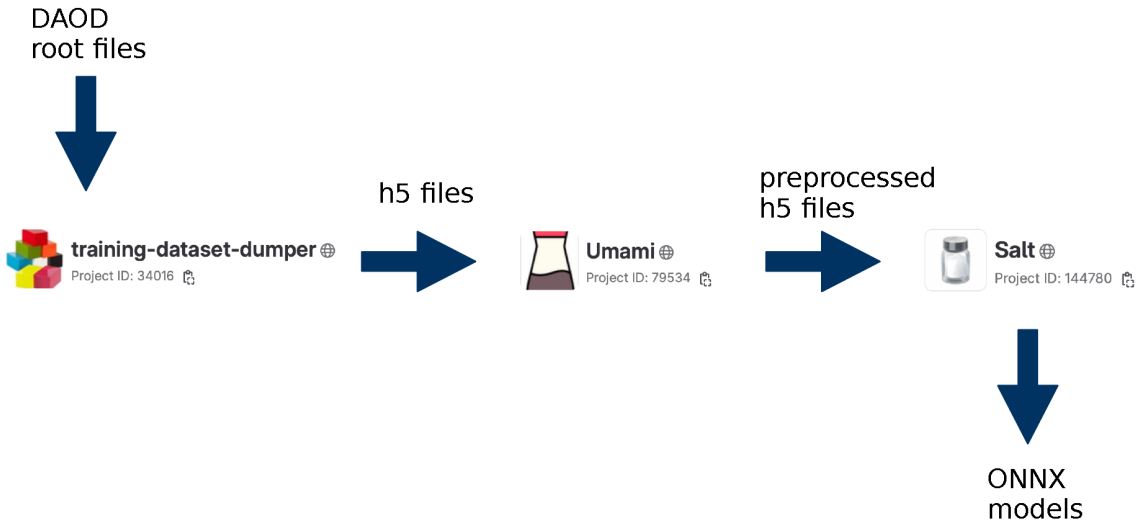


Figure 4.2: The workflow of the ATLAS flavour-tagging development including the training dataset dumper, Umami PreProcessing, and Salt. [46]

The training dataset dumper (TDD) is the first part of this flavour tagging specific chain [47]. It takes data in the DAOD format and outputs files in the HDF5 format. In the flavour tagging case, the jets are stored with the associated tracks and the TDD takes a configuration to choose which variables in the DAOD file are saved. Besides the input variables and truth labels of both the jets and associated tracks needed for the training of a model, additional information might be stored for studies and comparison, e.g. the p_b, p_c, p_u, and p_τ scores of older versions of the model. Through the selection of variables in the TDD configuration and the h5 format, the output is considerably smaller than the input DAODs. The flavour-tagging ecosystem is built around the h5 format, because it is designed to store and structure large datasets and because it is well-supported across software like Python, which is widely used in the FTAG pipeline. This makes the flavour-tagging software easily accessible and contributes to the fast turnaround time in the development of the taggers.

The produced h5 files are then fed into the Umami PreProcessing (UPP) software [48]. Umami is a framework which was used to train many of the machine learning based taggers in ATLAS, e.g. DL1x and DIPS. Although the training for the current flavour-tagging work is not done in Umami, the preprocessing developed for it is still in use. In addition to classic preprocessing steps, like the scaling and shifting of variables and the removal of outliers, UPP is geared towards the physical context of jet–flavour tagging. Inside the configuration file provided to UPP, the exact amount of jets per flavour is chosen, and thus the bias introduced by different amounts of jets from the respective flavour is avoided. This also enables the easy usage of jets from different samples, e.g. the $t\bar{t}$ and Z' samples commonly used for the different p_T regions. UPP also enables resampling of the jets in the

kinematic distributions of p_{T} and η, which has to be applied to avoid the kinematic bias introduced by the differences in the distributions between jet flavours. After preprocessing, the different flavours are approximately identically distributed across the kinematic range. UPP also provides the relative amounts for the different categories in the truth labels, which can be further used as class weights in the training. The most abundant class is assigned to have a value of one and the others are assigned a factor as the ratio to the most abundant class. The output of this preprocessing is also split into train, validation, and test datasets, for which the splitting ratio is determined in the UPP configuration.

The training of the jet–flavour taggers is done in Salt [49]. Salt is built on Pytorch Lightning and enables the direct building of machine learning models via configuration files. It was designed for the development of GN1 and GN2, but it was also made available for older taggers like DL1x. Salt allows the building and modification of taggers on a very high level through the configuration files, contributing to the accessibility and fast turnaround time, just like TDD and UPP. It supports multi-modal and multitask models as the building process in the configuration files is done by specifying the building blocks of the model one after the other and detailing their architecture, e.g. number of nodes in a layer or number of attention heads in the transformer part. Training is then executed by providing the configuration and the datasets acquired from the earlier steps. It is possible to monitor the training process via comet [50], tracking quantities like the total or individual losses during the training. Salt saves the model at defined points during the training as checkpoints and saves metadata about the training itself, containing also the configurations. The models built and trained in this way can also be exported to the Open Neural Network Exchange (ONNX) format, which is used in C++ environments like Athena. Thus, although the flavour-tagging development takes place outside the usual ATLAS software frameworks, the models can be integrated seamlessly. However, the data can also be directly evaluated with a model saved in the checkpoints of the training.

The data used to train GN2 was introduced in section 2.4.3. For this study, a smaller amount of data was taken through the workflow. The various model setups investigated in this work were trained on three different datasets. For the comparison with other ongoing work in the ATLAS flavour tagging effort at the time, a first dataset of 18.6 million $t\bar{t}$ and 7.4 million Z' jets was used, which will be referred to as default. As the special production without the Geant Thinning in the AOD reconstruction did not contain Z' events, only 18.6 million $t\bar{t}$ jets were used to form the dataset, which will be referred to as NGT. Models trained on this NGT sample could not be compared fairly to models trained on the default dataset as they contain different total amount of jets and cover a different p_{T} range, so another 18.6 million $t\bar{t}$ jets dataset was constructed with the default reconstruction. This will be referred to as OTTB (only $t\bar{t}$). Each of these datasets were split in such a way that 80 % of the jets are used for training and 10 % for testing and validation respectively. All $t\bar{t}$ sets are composed of 4.5 million b- and c-jets respectively, 9 million light jets and 625 thousand τ-jets. The Z' part of the default dataset consists of 1.8 million b- and c-jets respectively, 3.6 million light jets and 250 thousand τ-jets. For the source–prediction task, table 4.1 shows the class weights in the loss as they are calculated by UPP. The table again shows the increased abundance of secondaries in the NGT data as shown in section 3.2, but it also compares a pure $t\bar{t}$ sample (OTTB) with the merged $t\bar{t}$ and Z' sample (default).

The preprocessing of the datasets and the training of the various model setups was performed on the OMNI-cluster of the University of Siegen [51], providing ample storage and GPU access, which was utilized during the training. The TDD and UPP configurations were not noticeably different compared to the commonly used one, but the configuration of Salt includes the expansion of the model architecture. The model configurations are available in the repository in Ref. [52], which is a fork of the Salt repository. For all three datasets, default, NGT, and OTTB, the model was trained only

Table 4.1: Weights for the source classification task classes when applied in full for the default, NGT, and OTTB datasets.

Source	**Default**	**NGT**	**OTTB**
NoTruth	4.72	5.86	4.33
NotSecondary	1.00	1.00	1.00
HadronicInteraction	147.31	100.90	311.12
StrangeMesonDecay	89.15	69.00	70.22
StrangeBaryonDecay	178.41	145.04	144.96
GammaConversion	46.57	12.65	38.81
Other	5 509.32	982.89	5 981.48

once without the additional track source–prediction task. An overview of the different setups, which include the additional source task, trained in this study, is given in table 4.2. In the training process, the different choices of the task weight parameter was informed by the resulting tagging performance of the model and by the losses themselves, as the source task loss should be approximately equal to the losses of the other auxiliary tasks. The choices of class weighting schemes were also informed by the tagging performance of the models, and by the performance of the source classification task, as these weights have a large influence on the specific task itself. No formal hyperparameter optimization was performed for the determination of the task weight parameter, the choice of the class weight scheme, and neither for the parameters already included in the architecture, which was expanded for these trainings.

Table 4.2: Trained model configurations including the track source–prediction task with different task weights (TW) and class weights (CW) across the different datasets. If the setup was trained and evaluated it is noted with a tick (✓), otherwise it is left blank.

Dataset	**TW**	**CW**		
		none	sqrt	full
Default	0.5	✓		✓
	1		✓	
	2	✓	✓	✓
NGT	0.5			✓
	1		✓	
	2	✓	✓	✓
OTTB	0.5			
	1		✓	
	2		✓	

4.2 Results of Retraining

The impact of the source auxiliary task on the performance of the model is evaluated with the Receiver Operator Curves (ROC curves) shown in figures 4.3 and 4.4, of which the former shows the evaluation with the $t\bar{t}$ jets and the latter with the Z' jets in the testing split of the default dataset. Both figures show the curves for the model trained on the default dataset without the additional source auxiliary task and for all the trained model setups with the additional task. Since this work focuses on the b-tagging with GN2, the model outputs p_b, p_c, p_τ, and p_{light} are evaluated in the b-tagging discriminant given in equation 2.8. The f_c and f_τ values chosen for the calculation of the discriminant score are $f_c = 0.2$ and $f_\tau = 0.01$, the recommended values for GN2 [5]. The ROC curves are acquired by placing cuts on the discriminant and calculating the b-tagging efficiency, how many true b-jets are correctly predicted as b-jet, and the rejections of the other flavours, which is the inverse of these flavours being falsely tagged as b-jets. With the supervised learning methods utilized in the current approach to flavour tagging, the models can only be trained on Monte-Carlo simulation, because the truth information has to be available. But as the physics modelling, e.g. the distributions of and the correlations between the input variables in the model, is not perfect, the model might perform differently on collision data. This problem is attended to by dedicated calibration analyses, which measure the tagging efficiency directly for a set of pre-defined operating points (OPs) [5]. For the GN2 model inclusive b-jet tagging efficiencies of 65 %, 70 %, 77 %, 85 %, and 90 % were defined and the model calibrated at these values [5]. The OPs are not determined by the b-jet tagging efficiencies, but rather as fixed cuts in the discriminant, resulting in the listed inclusive efficiencies for a reference $t\bar{t}$ sample. Thus, for different samples the efficiencies resulting from these discriminant cuts might differ. In the Z' sample, the same cuts in the discriminant yield efficiencies of approximately 22 %, 29 %, 40 %, 63 %, and 81 %, depending on the exact sample. The efficiencies at the OPs in the $t\bar{t}$ sample are nearly identical to the inclusive b-jet tagging efficiencies listed. The model is calibrated at multiple OPs, because different physics analyses profit from different trade-offs in tagging efficiency and background rejection, depending on their physics cases. These points inform at which b-jet efficiencies the ROC curves with the different source task setups should be compared.

In the $t\bar{t}$ sample (figure 4.3), two setups of the model with the source task clearly outperform the standard model without the source task, TW2-CWsqrt and TW2-CWnone. They are consistently better at rejecting light jets with an improvement of up to 25 % and show a very similar rejection of c-jets to the model without the source task. The other combinations of task weight parameter and class weighting schemes yield mixed results in the light jet rejection, performing significantly worse for low b-jet efficiencies, but better at high b-jet efficiencies, including the highest two of the five operation points. In the c-jet rejection, the other models including the source task are either also similar to the model without the source task, or perform significantly worse.

Similar observations are made for the Z' sample (figure 4.4). The TW2-CWsqrt and TW2-CWnone setups of the source task outperform the standard model very clearly in the light jet rejection, reaching close to an improvement of factor two at the lower OPs. The rejection of c-jets is still very similar for most models, but, interestingly, the otherwise well-performing TW2-CWsqrt leads to a decrease of 5 % to 10 % in the c-jet rejection of the Z' sample.

Every source task setup benefits the rejection of τ-jets in $t\bar{t}$ and Z' jets. As the τ-jets represent the smallest fraction in the dataset, a model closer to the physics context of flavour tagging might benefit their modelling the most. Additionally, figure A.2 and A.3 in the appendix show that τ-jets are quite different from the other flavours, because they contain fewer tracks of secondary origin.

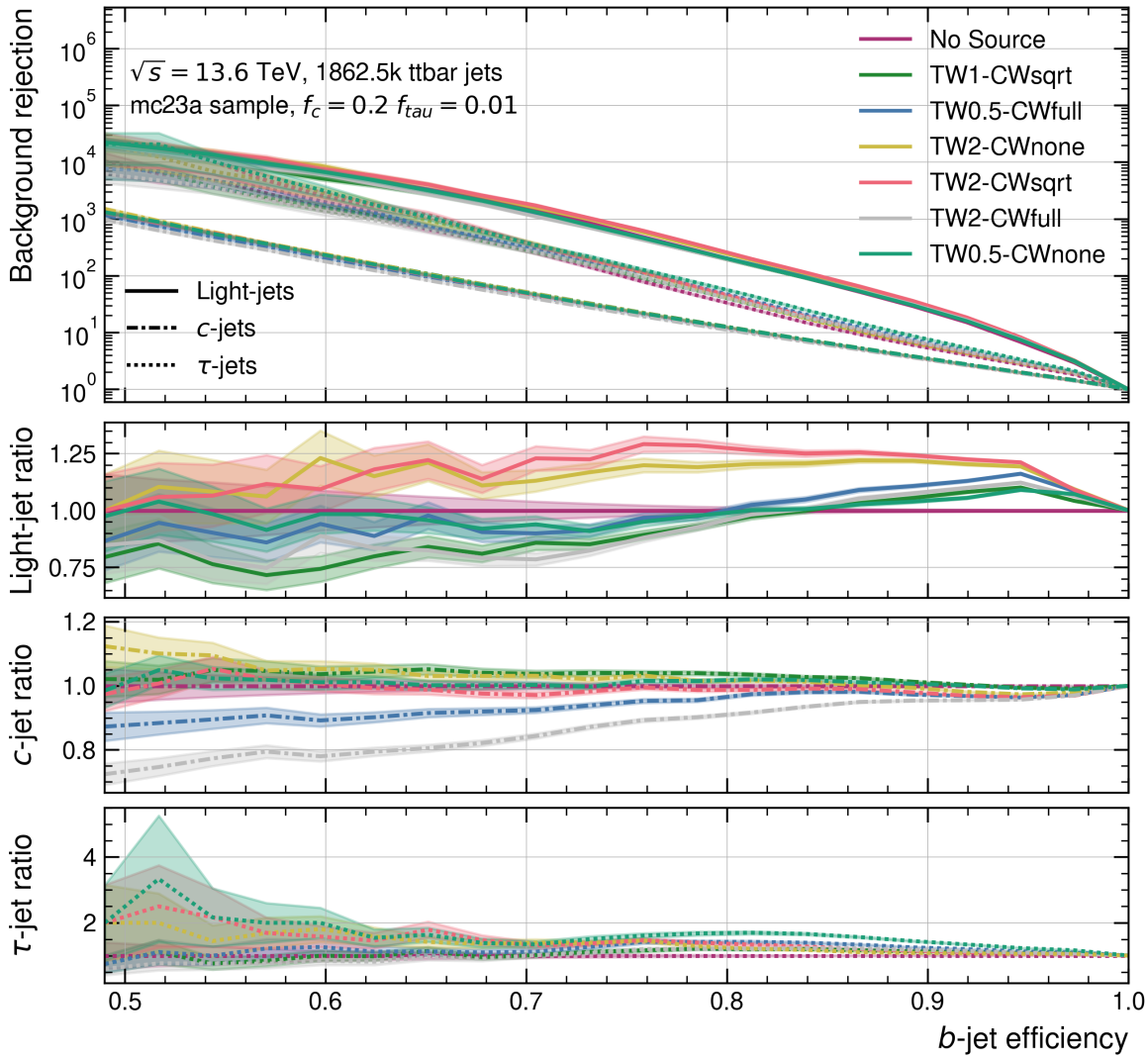


Figure 4.3: ROC curves of the b-jet tagging efficiency and the background rejections of the model trained without the additional source task and all setups with the additional source task trained on the default dataset. All models are evaluated on the $t\bar{t}$ part of the test split.

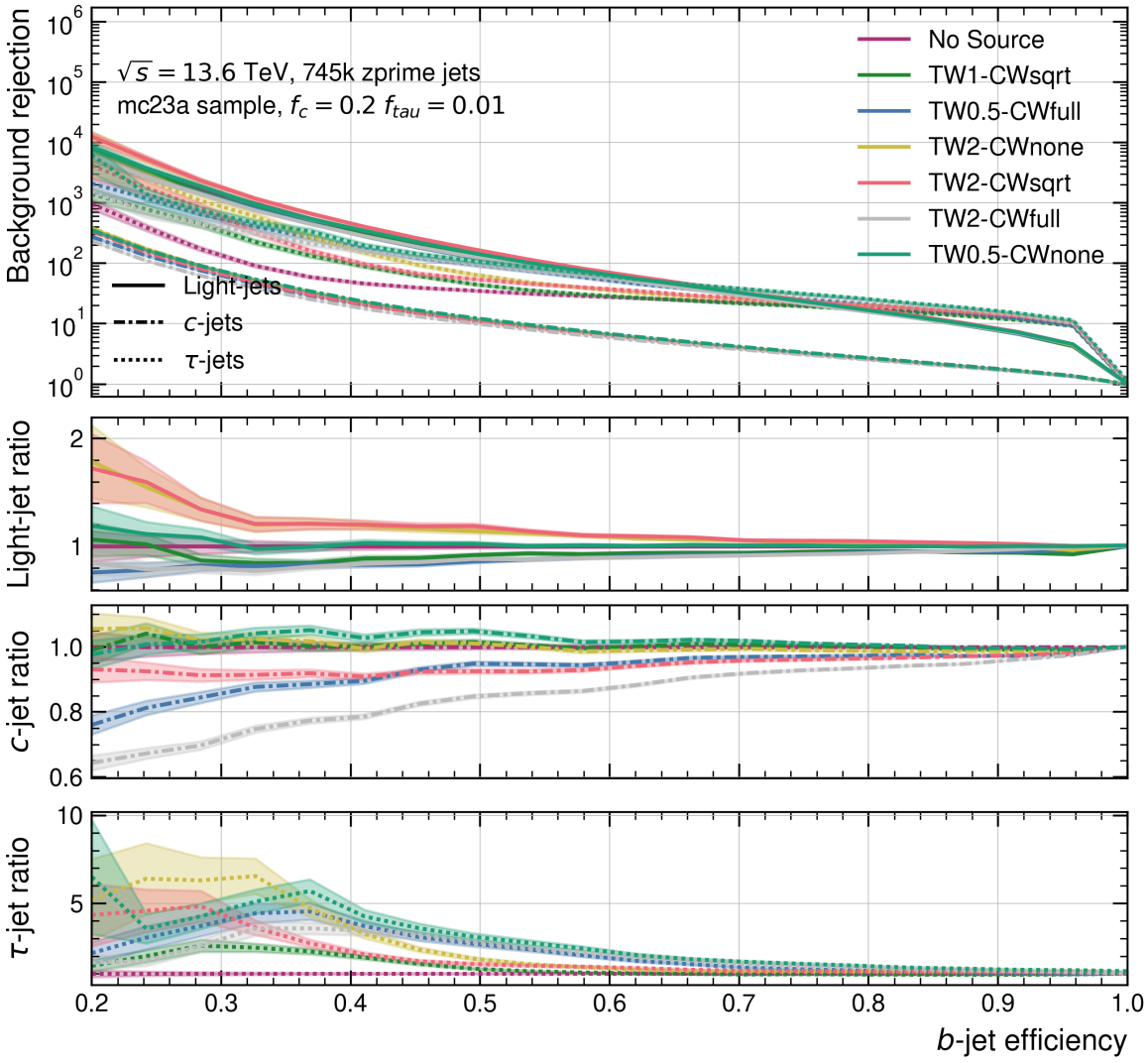


Figure 4.4: ROC curves of the b-jet tagging efficiency and the background rejections of the model trained without the additional source task and all setups with the additional source task trained on the default dataset. All models are evaluated on the Z' part of the test split.

The ROC curves show the performance of the different models in the main task of the model, tagging b-jets as such, while rejecting as much of the other flavours as possible. To assess the performance of the source classification task itself in the different models, confusion matrices are used. Figure 4.5 shows the confusion matrix for the TW2-CWsqrt model and figure 4.6 the confusion matrix for the TW2-CWnone model, both evaluated on $t\bar{t}$ data. These matrices are normalized to 1 per row, which means that for each true category of the source label it is shown how the tracks in that category are distributed in the predictions over all the categories. The confusion matrices for these models evaluated on default Z' data show very similar results and can be found in the appendix in figure A.9 for the TW2-CWsqrt model and in figure A.10 for the TW2-CWnone model.

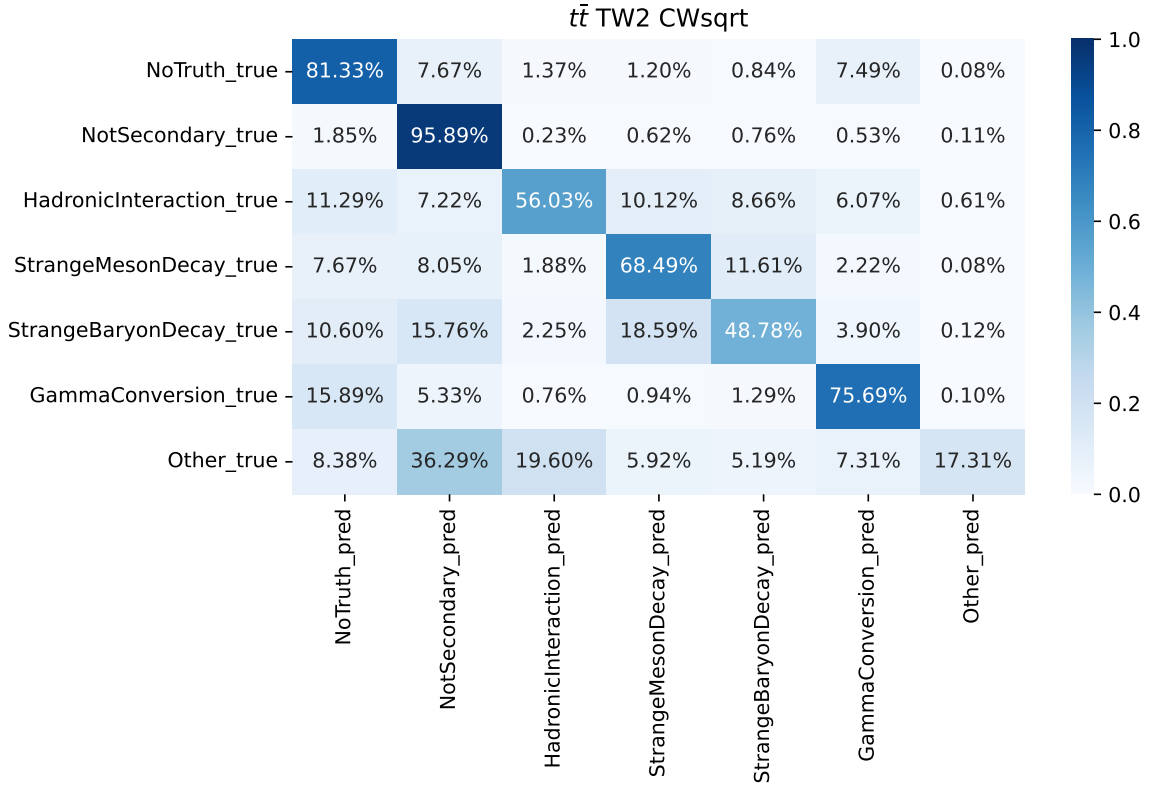


Figure 4.5: Confusion matrix of the source prediction in the TW2-CWsqrt model evaluated on default $t\bar{t}$ data.

As the two models shown here improved the performance of jet–flavour prediction significantly, good performance in the source–prediction task would confirm the assumption motivating the addition of the source–prediction task. The assumption is that the classification of secondary tracks helps the model to distinguish between secondary vertices from heavy-flavour hadron decays and from material interactions and the decays-in-flight of other long-lived hadrons. For figure 4.5, showing model TW2-CWsqrt, this is the case. The confusion matrix shows a well-performing source prediction, especially considering the large class imbalance. This could explain the performance boost in the main task. However, the model TW2-CWnone in figure 4.6, does not confirm this assumption. This model performs poorly in the prediction of secondary origins of tracks, but still improved the main task performance. Thus, it appears that the improvement in performance of jet–flavour prediction is

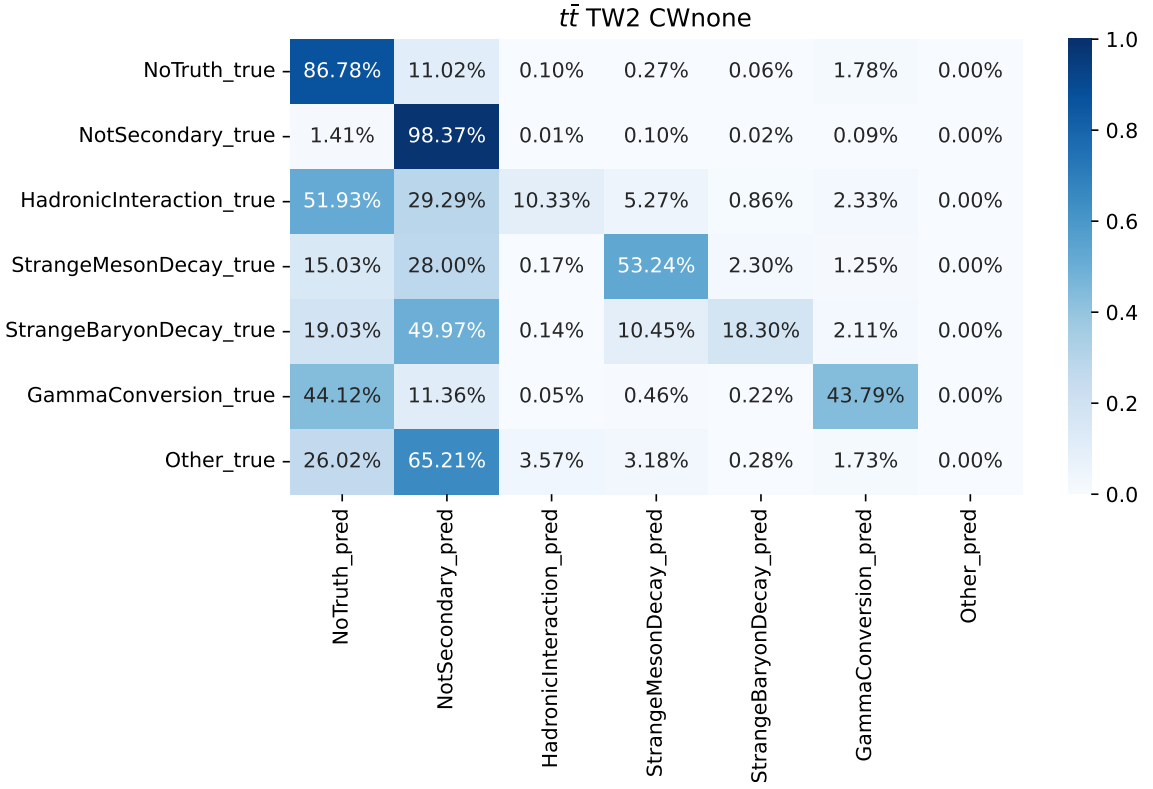


Figure 4.6: Confusion matrix of the source prediction in the TW2-CWnone model evaluated on default $t\bar{t}$ data.

independent of a well-performing source prediction and that the assumption of this venture does not hold. Possibly, the differences per model in jet–flavour prediction shown with the ROC curves are a result of statistical fluctuations. Yet, it is also possible that the NoTruth and NotSecondary divide is sufficient to benefit the performance of the model. The TW2-CWnone model in figure 4.6 does not classify the secondaries well, but it still manages to distinguish NoTruth and NotSecondary well. This could be helpful as NoTruth contains pileup tracks for example, so the source task would support the origin–prediction task in its classification described in section 2.4.3.

The poor performance of the TW2-CWnone model is due to the missing class weights. As described in section 4.1, the huge class imbalance in the source label leads to the model not being punished for predicting the very rare secondary origins as the abundant NoTruth or NotSecondary categories. All trained CWnone models show the poor performance of the source prediction in the confusion matrices, while CWsqrt models show a good classification. Utilizing the class weight in full in the CWfull models leads to an even better classification of the secondaries, but the performance of the NotSecondary prediction is reduced by approximately 5 %, which leads to an overall worse performing task, because of the large amount of tracks in the NotSecondary category. Thus, for a well-performing source prediction, the weighting scheme with the square root of the class weights should be preferred.

The classification of the Other category is performing poorly through all the models, because few tracks are in this category. Interestingly, in the CWnone models, which show a poor classification of the secondary origins, the material interaction categories HadronicInteraction and GammaConversion

tend to be misclassified as NoTruth rather than NotSecondary, although the NotSecondary category is the most abundant. This hints at a similarity of material interaction secondaries and pileup tracks, while the tracks from long-lived strange-hadron decays are more similar to the NotSecondary tracks. Even in CWnone models the GammaConversion and StrangeBaryonDecay categories retain a very good predictive capability in the source task. The prediction of GammaConversion and StrangeBaryonDecay sources perform the best out of the secondaries throughout the different setups. This might be due to the decay of strange baryons and the conversion of photons having distinct characteristics, which make them more easily identifiable.

The impact of the class weighting choice becomes apparent in the performance of the source task. The other variation of the trained models, the task weight, manifests itself in the loss of the source task during the training. Figure 4.7 shows the losses of the models discussed in detail so far, TW2-CWsqrt and TW2-CWnone, and the loss of the TW1-CWsqrt model. In the comparison between TW2-CWsqrt and TW1-CWsqrt, the direct impact of the task weight parameter is observed, as the loss of the source task in the TW2-CWsqrt model is approximately twice as high as in the TW1-CWsqrt model. But the class weights also have a large impact on the loss of this task. The difference between the losses of TW2-CWsqrt and TW2-CWnone is even slightly larger than a factor of two.

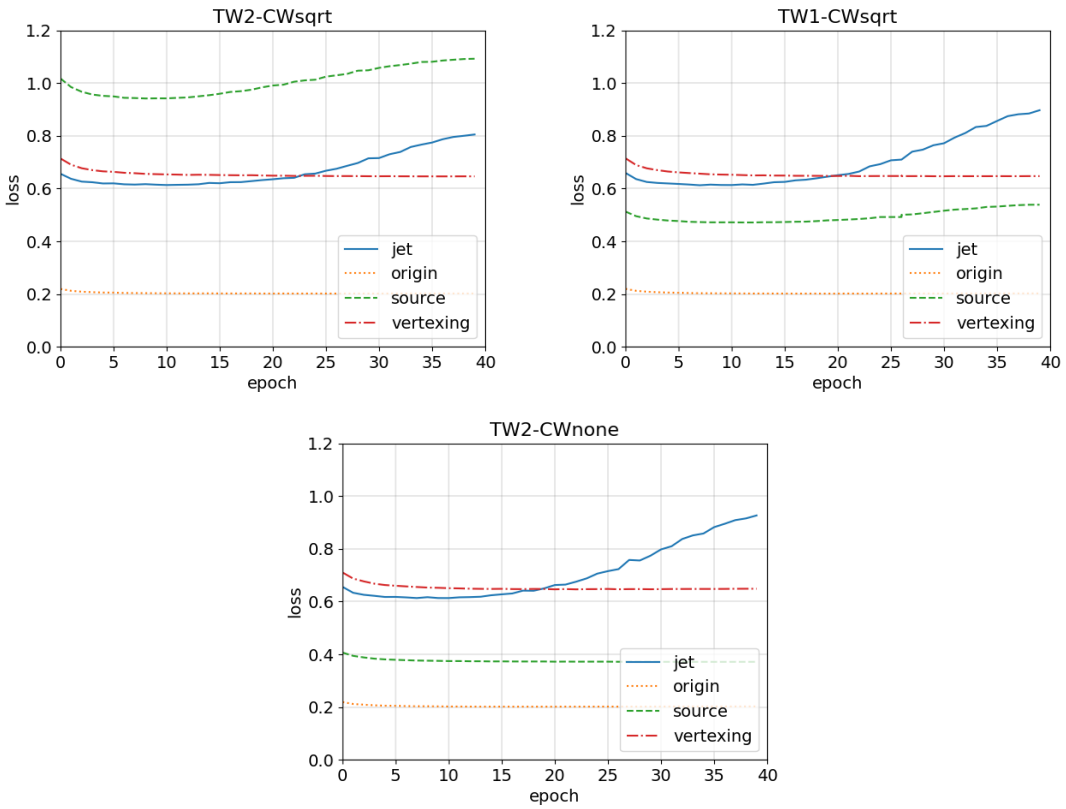


Figure 4.7: Losses as a function of training epoch for the jet flavour, track origin, vertex, and track source prediction tasks for three of the trained models: TW2-CWsqrt, TW2-CWnone, and TW1-CWsqrt.

In general, the goal discussed in section 2.4.3 of the jet–flavour prediction being the dominant loss and the auxiliary tasks being approximately equal, but below the main task, is not met. In all three models shown in figure 4.7, the vertexing task loss is larger than the jet–flavour prediction loss for the first fifteen to twenty epochs of training. The origin task loss is consistently small compared to the other tasks. This observation might motivate a different choice for the parameters α and β, the weights of the origin and vertexing task in the linear combination of the losses, for the GN2 model, as they were chosen to be equal to the ones in use for GN1 [5]. But since GN3 does not use the linear combination approach to the loss anymore, such studies might not be warranted in the current flavour-tagging development.

In section 3.2, the impact of secondaries on the unaltered GN2 model was studied by considering the discriminant distributions of the different jet flavours in figure 3.10. The discriminants of the different flavours are not as well separated for jets containing secondaries as compared to jets not containing secondaries. The idea motivating this work was to mitigate this effect by adding a task categorizing the secondaries. Figures 4.8 and A.11 in the appendix show the comparison between the discriminant distributions for the jet flavours of the model trained without and with the additional source task, specifically the TW2-CWsqrt setup. Figure 4.8 shows this comparison on the $t\bar{t}$ data and figure A.11 on the Z' data. A direct comparison of these discriminant distributions to figure 3.10 is not appropriate, because the discriminant scores in this figure were acquired from a training of an early version of the GN2 algorithm without the τ-jet class and on the full dataset.

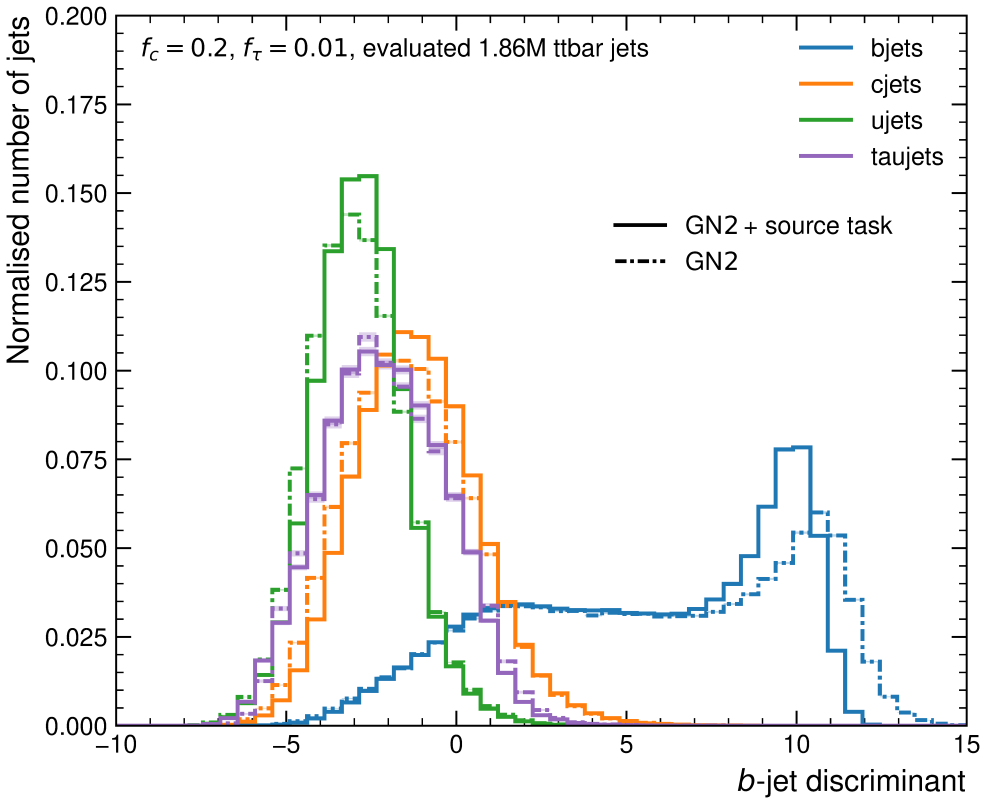


Figure 4.8: The b-tagging discriminant of the TW2-CWsqrt model and the model without source task evaluated on the $t\bar{t}$ data for light jets (ujets), c-jets (cjets), b-jets (bjets), and τ-jets (taujets).

The anticipated impact of the source task on the discriminant is not observed, the distributions are not better separated. At a first glance in the $t\bar{t}$ data, the opposite seems to apply. For the b-jets the distribution does not reach as high in the discriminant score with the addition of the source task, and the light jet distribution is less populated in the very low discriminant scores. However, the distributions are not merely shifted, while retaining their form, they are narrower. The observed performance improvement of the ROC curves is found in the tails of the distributions around 0 in the b-tagging discriminant. The upper tails of the light jet and τ-jet distributions are shorter as well as the lower tail of the b-jet discriminant. For the Z' data, a similar picture arises, but the fact, that the TW2-CWsqrt model performed worse in c-jet rejection than the model without source task, can be observed in the upper tail of the discriminant distribution of the c-jets.

The desired improvement could have manifested in light jets containing secondaries forming a discriminant distribution closer to light jets without secondaries, such that the gap observed in figure 3.10 gets narrower, and more light jets have lower b-tagging discriminant values. But figure 4.9 shows that this is not the case. It splits the discriminant distributions for the light jets in the $t\bar{t}$ evaluation data for both, the TW2-CWsqrt model and the model without source task, into containing secondaries or not. The desired effect is not observed. Thus, the model still considers light jets containing secondaries more b-jet like than light jets without, even with the addition of the source auxiliary task.

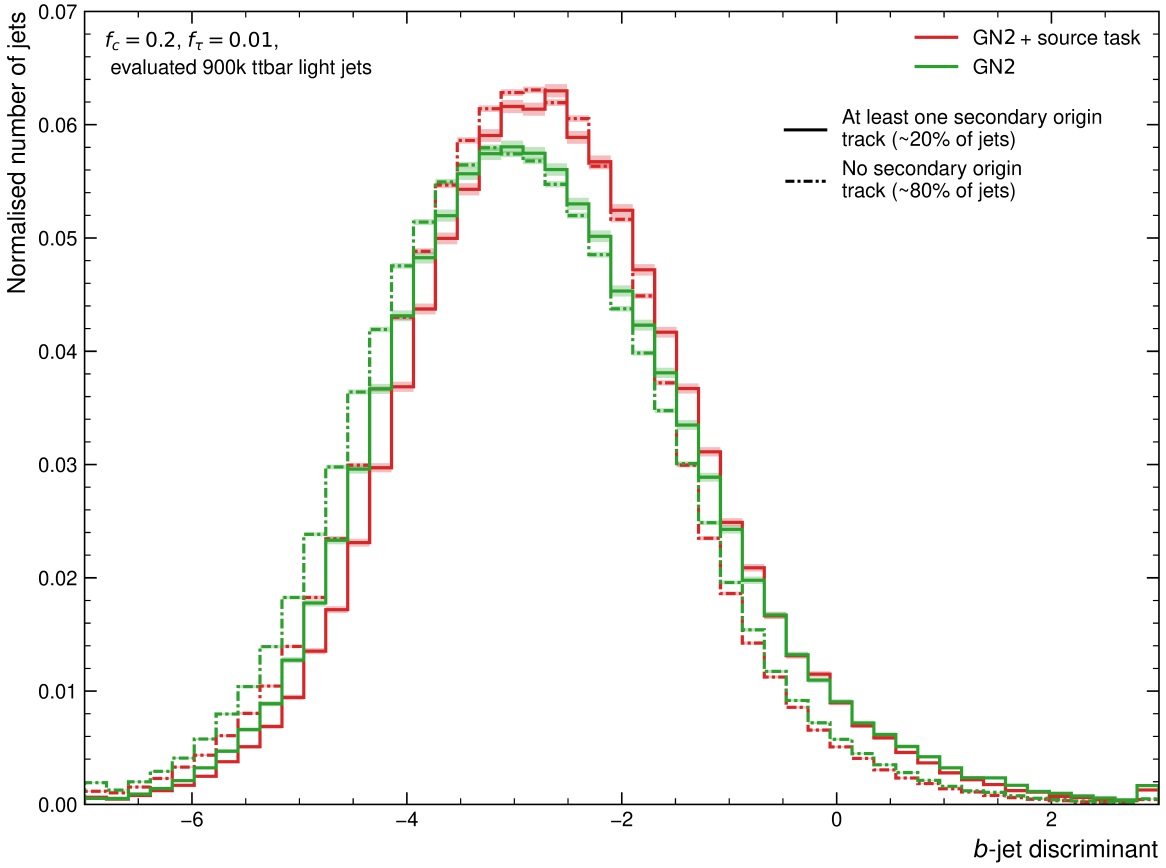


Figure 4.9: The b-tagging discriminant of the TW2-CWsqrt model and the model without source task for light jets, split for each model into jets containing at least one track of secondary origin and jets containing none.

Figure 4.10 also illustrates this by showing the average relative amount of secondaries in light jets over the b-tagging discriminant for both the TW2-CWsqrt model and the model without the source task. This figure shows the $t\bar{t}$ data with the corresponding Z' figure A.12 in the appendix. A higher relative amount of secondaries clearly correlates with higher discriminant scores, as is expected. But the addition of the source task has no significant impact on this correlation. On the lower end of the discriminant, a small surplus of secondaries is observed, but besides that, the histograms match very closely, especially in the bins without large errors.

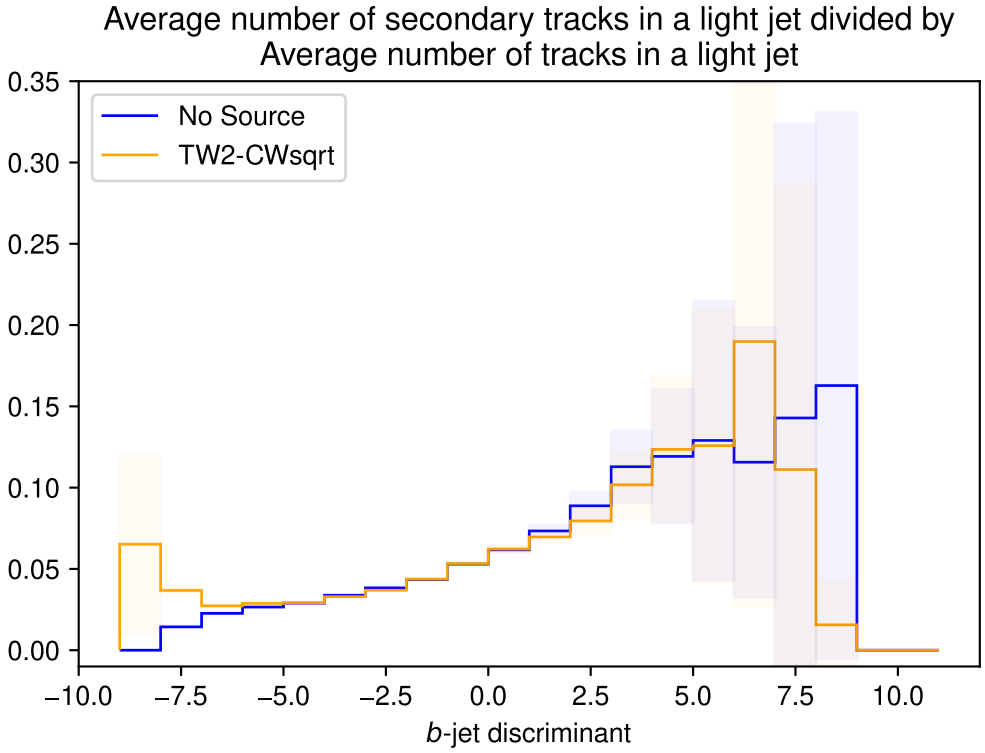


Figure 4.10: Relative amount of secondary tracks in a jet over the b-tagging discriminant for the TW2-CWsqrt model and the model without source task, evaluated on $t\bar{t}$ data.

To summarize, all the results show that the model is definitely capable to identify and categorize tracks of secondary origin, as it shows relatively good accuracies in the source prediction, especially when considering the class imbalance in the training data and the fact that the source prediction is only meant to be an auxiliary task. The performance of the models trained in this work regarding the source prediction is comparable to the capabilities of GN3 in the origin prediction [45]. Furthermore, a potential improvement of the jet–flavour prediction through the addition of the source–prediction auxiliary task is observed. Although the impact on the performance of the main task is quite volatile with respect to the chosen task weight and class weighting scheme, two setups provided a consistent 25 % improvement in light jet rejection. Most model setups yielded a worse performing flavour-tagging model, but a potential performance boost of this magnitude warrants further study, especially since it was realized without altering other parts of the architecture.

4.3 Results of Retraining on No Geant Thinning Data

The models trained on NGT data are expected to perform better in the source–prediction task, and thus possibly also perform better overall, because the increased availability of truth information leads to less class imbalance by labelling fewer tracks as NoTruth and more tracks as of secondary origin, especially HadronicInteraction and GammaConversion, as seen in table 4.1 or figure 3.11. The model performance in jet–flavour tagging, shown in figure 4.11, confirms this expectation in part, as more variations of the model including the source task show a consistent improvement over the model without. Except for the TW2-CWfull model, the models with the source task show an improved light jet rejection either at or even below the lowest operating point, which is at 65 % b-jet efficiency. They also show either similar or improved performance in c-jet rejection. The τ-jet rejection also shows similar performance between the models with and without source task. Thus, when comparing this performance to the trainings on default data in figure 4.3, the expectation for the NGT training is fulfilled, as nearly all variations of the model including the source task show a consistent improvement in the rejections of the other flavoured jets. Interestingly, a significant improvement in τ-jet rejection like in the default dataset is not observed for the models trained on the NGT data. All in all, the expected improvement is observed within the comparison between NGT models.

A direct comparison between the models trained on the NGT dataset and models trained on the default dataset is not insightful due to the Z' sample not being available without Geant Thinning. As described in section 4.1, the OTTB dataset was introduced exactly for direct comparisons between data including and excluding the Geant Thinning. Comparing the models trained on the OTTB data amongst themselves shows that the addition of the source task leads to a severe deterioration of the rejection of light and τ-jets, and a similar rejection of c-jets, as seen in the ROC curves in figure A.13 in the appendix. Since the two models with the source task trained on OTTB data have the same setups as the best ones trained on default and NGT, respectively, the clearly worse performance is surprising.

One problem might be the composition of the loss, which is discussed below. Another more general problem, applying to all trainings, is the low number of jets used for training. After this work was already past the stage in which the models were trained, a number of jets was recommended for studies on developing the flavour-tagging algorithms, which was 30 million. The default dataset comes close to this with approximately 26 million jets, but the NGT and OTTB dataset only consist of 18.6 million jets, slightly more than half of the recommendation. The benefit of the extra physics context provided by the additional task might only take effect for more training data, because additional tasks also increase the complexity within the model and training process. Additionally, the models trained on NGT and on OTTB data share the exact same architecture as the ones trained on the default dataset, and this architecture was optimized for $t\bar{t}$ and Z' jets.

The comparison between NGT and OTTB models is depicted in figure 4.12. It shows a worse performance for the NGT models, regardless of whether the source task is included or not, compared to the OTTB model without the source task. The c- and τ-jet rejections are clearly worse, while the light jet rejection is ambiguous, but worse for most of the OPs. Why the model without source task trained on the OTTB dataset performs the best out of the NGT and OTTB models is not clear. Especially since the only difference between the NGT and OTTB dataset is that there is more truth information available in the NGT dataset, but still the NGT NoSource model performs worse than the OTTB NoSource model. There is one caveat to this comparison, because the ROC curves are created by evaluating the models on the type of dataset on which they were trained, so the OTTB and NGT models are neither trained nor evaluated on the exact same data. In section 3.2, it is demonstrated that

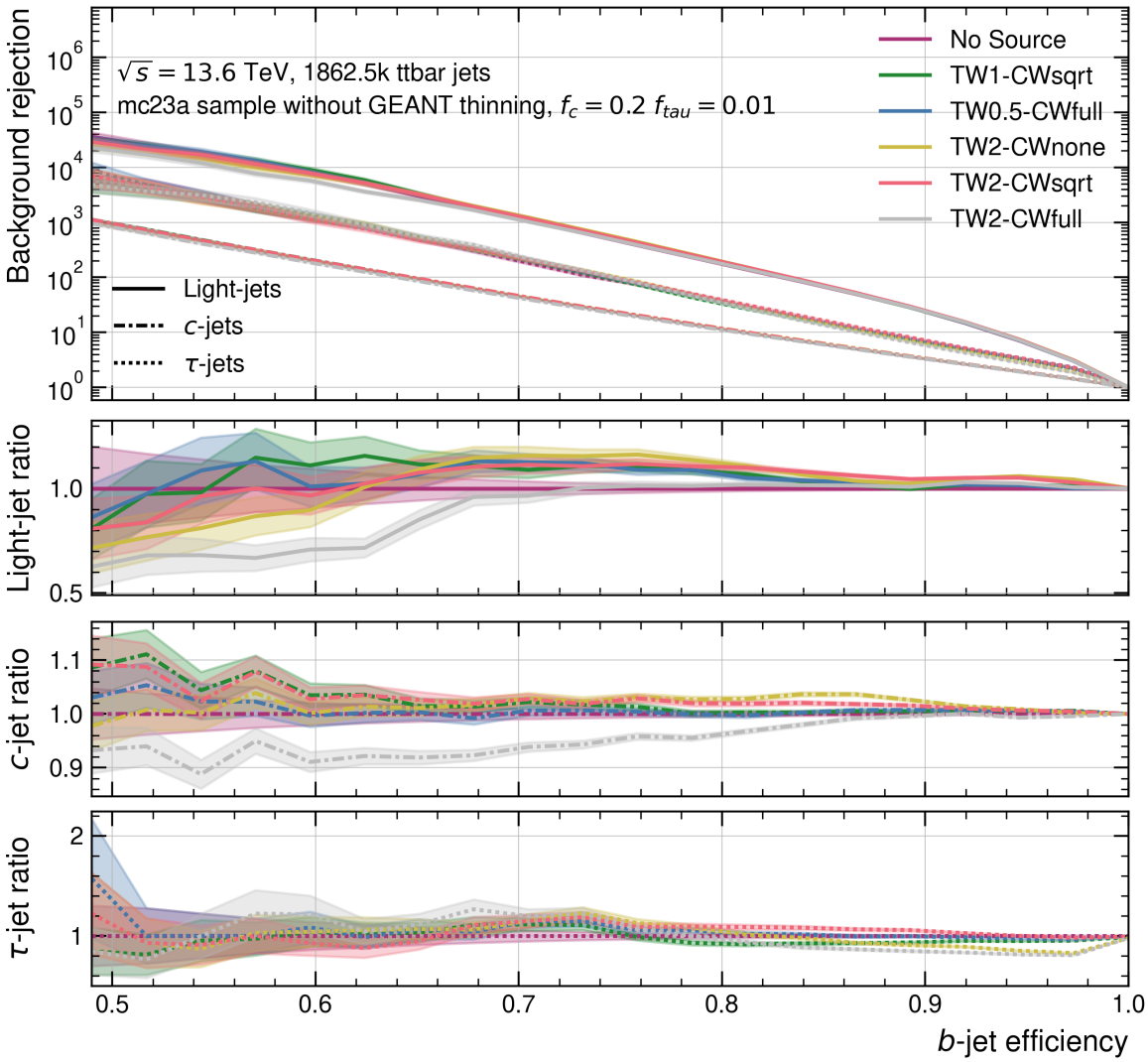


Figure 4.11: ROC curves of the b-jet tagging efficiency and the background rejections of the model trained without the additional source task and all setups with the additional source task trained on the NGT $t\bar{t}$ dataset.

there is no difference in the variables used by the model, whether Geant Thinning is applied or not, so the data used for the evaluation should only differ due to statistical effects.

Independently of the performance in the main jet–flavour prediction task, the models trained on NGT data should perform better in the source prediction, because of the reduced class imbalance. For a direct comparison, the confusion matrix for both datasets are shown for the TW2-CWsqrt model, in figure 4.13 with the NGT dataset, and in figure 4.14 with the OTTB dataset. Both show a relatively good performing source task, considering it is an auxiliary task, just like some models trained on the default dataset discussed in section 4.2. One noticeable difference between the NGT and OTTB source task performance is the significant improvement of accuracy in the GammaConversion category by 17 %. The GammaConversion prediction is even more accurate than the NoTruth prediction, despite also having less tracks in the NGT dataset, which points to the photon conversion having a distinct signature, which the model can identify well. Although the number of tracks in the HadronicInteraction category tripled similar to the number of tracks in the GammaConversion category, the prediction for HadronicInteraction tracks does not improve significantly, only by 3 %. StrangeMesonDecay and StrangeBaryonDecay predictions perform very similarly between NGT and OTTB, which fits the expectation as the number of tracks in these categories are also very similar with or without Geant Thinning applied.

The concerns regarding the loss of the trainings on the default dataset get worse for the OTTB and NGT trainings. Figure 4.15 shows the individual losses of the tasks over the epochs of the trainings for the TW2-CWsqrt in the NGT and in the OTTB datasets. The loss of the vertexing task, which was slightly larger than the loss of the jet–flavour prediction task for the training on the default data, is significantly larger for the NGT and OTTB trainings. Additionally, the source–prediction task loss is also significantly larger. The fact that these two components of the overall loss are larger than the loss of the main task might explain the unexpected degradation of performance in the main task. A close look at the OTTB loss shows that the minimum of the source–prediction task is few epochs earlier than the minimum of the jet–prediction task. This might lead to a model being selected as the best one by having the overall smallest loss, which is actually premature with respect to the jet–flavour prediction. Thus, the good performance of the source prediction and the negative impact of the addition of the source task on the jet–flavour prediction might be explained.

The expectations placed in the models trained with the NGT dataset are fulfilled at least partially. Reducing the class imbalance by approximately tripling the amount of tracks in the GammaConversion and HadronicInteraction categories of the source label led to a sizeable improvement in the GammaConversion prediction. Surprisingly, a similar improvement for the HadronicInteraction category is not observed, which can probably be attributed to hadronic interactions being more complicated processes. The increase in performance of the jet–flavour prediction thanks to the additional source task is way more robust when comparing different models trained on the NGT data. A comparison of the models trained on the NGT data with the ones trained on the OTTB data yields unexpected results. The model without the source task trained on OTTB data, containing less truth information, outperforms all models trained on NGT data. This might be due to the low number of jets in the training or a result of the absence of Z' jets in the OTTB and NGT datasets.

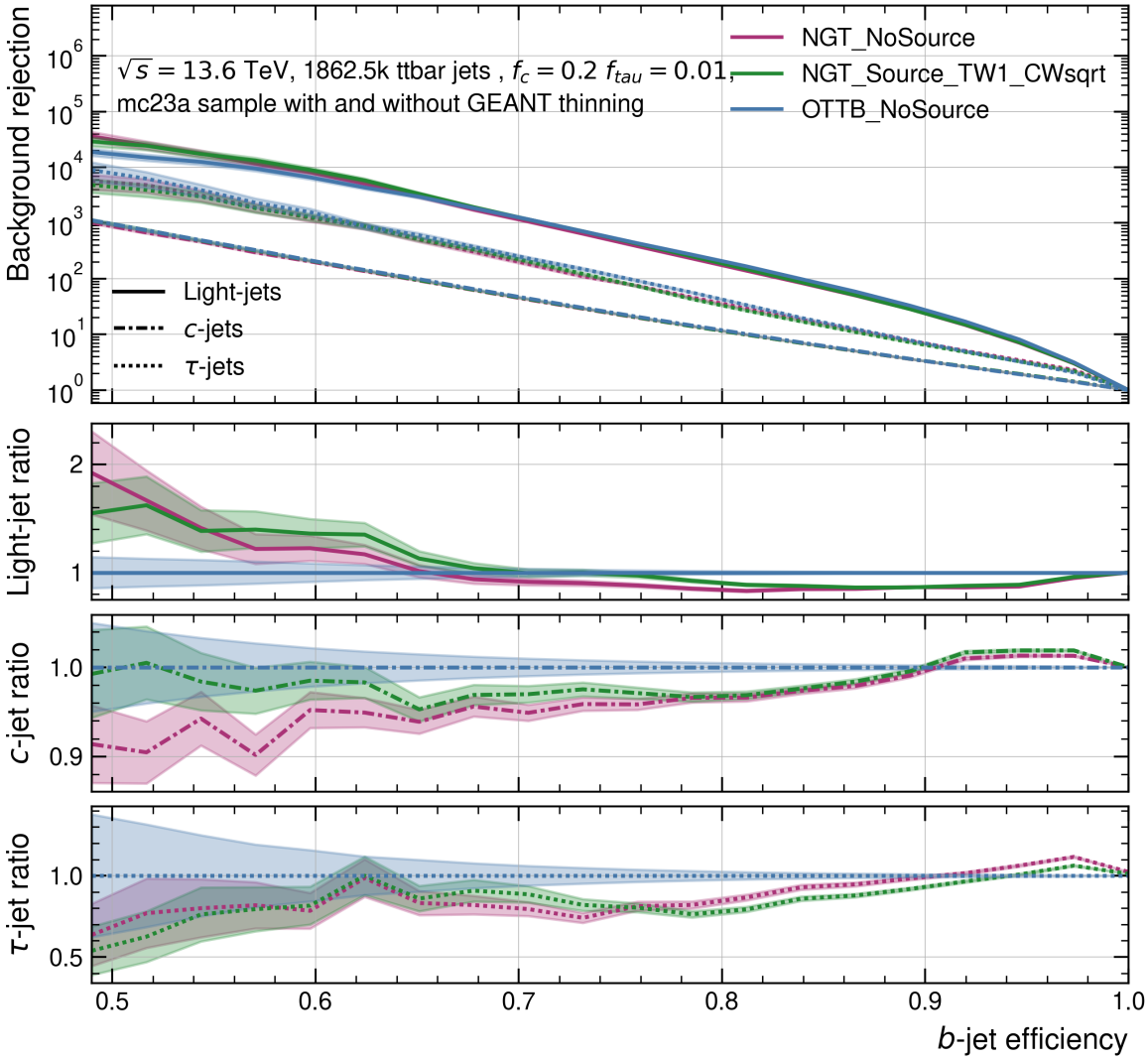


Figure 4.12: ROC curves of the b-jet tagging efficiency and the background rejections of the models without source task trained on the NGT and OTTB data and the TW1-CWsqrt model trained on NGT data.

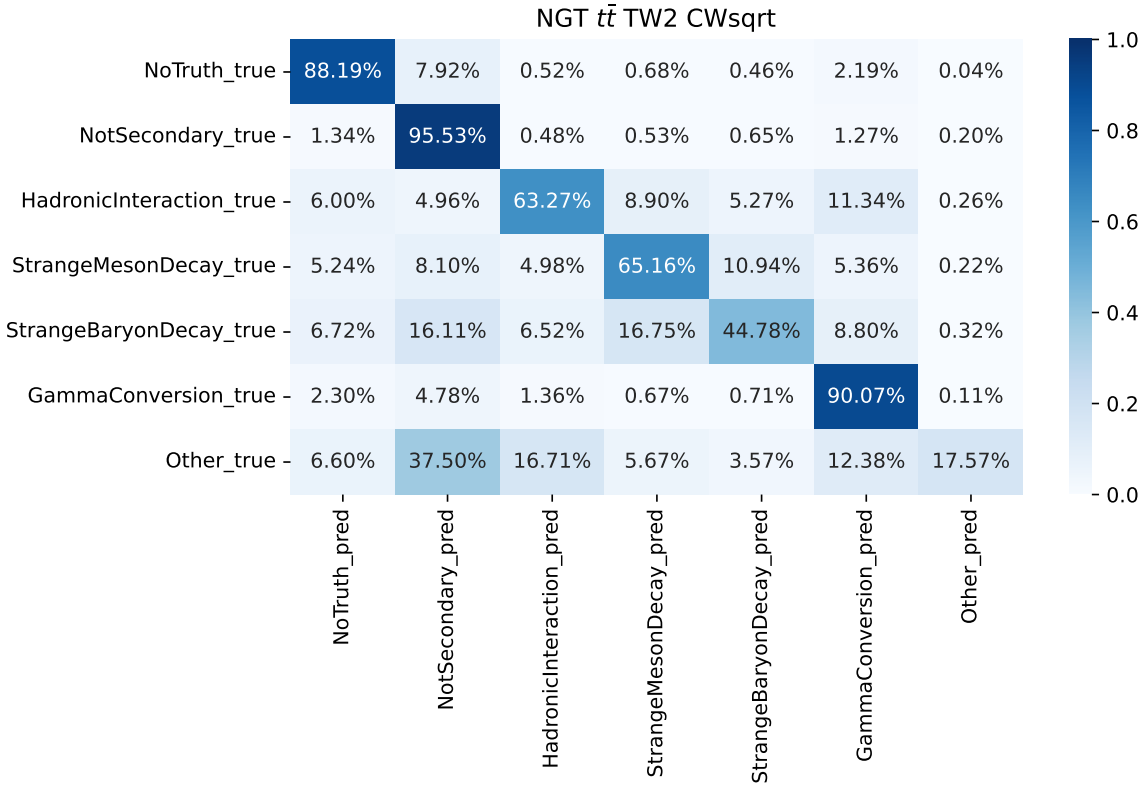


Figure 4.13: Confusion matrix of the source prediction in the TW2-CWsqrt model trained on the NGT data.

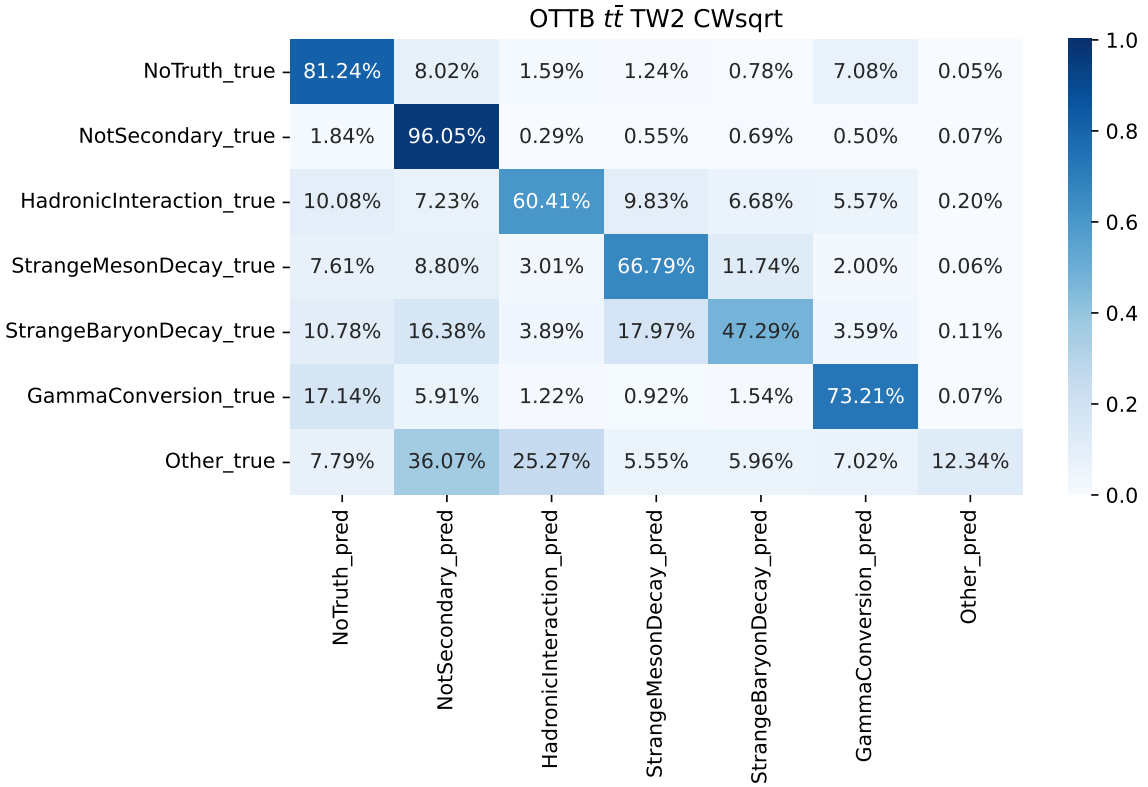


Figure 4.14: Confusion matrix of the source prediction in the TW2-CWnone model trained on OTTB data.

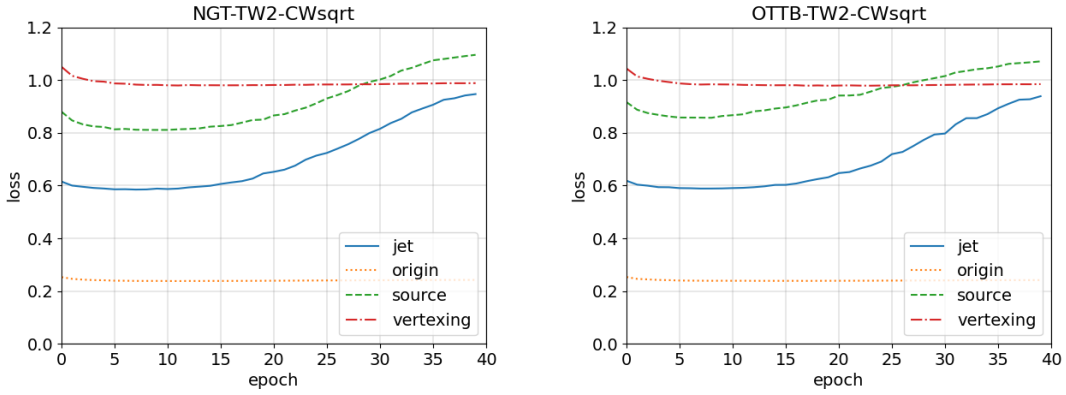


Figure 4.15: The losses for the jet flavour, track origin, vertex, and track–source prediction tasks for the TW2-CWsqrt models trained on the NGT (left) and OTTB (right) data.

CHAPTER 5

Conclusions and Outlook

Flavour tagging within ATLAS has seen a huge boost in performance by deploying a Graph Neural Network architecture within the GN1 model and expanding this architecture to a Transformer Encoder in the GN2 tagger. In contrast to previous taggers, the GNx models are designed in an "end-to-end" approach, directly being provided with the variables of the jets and tracks, whereas the input to previous taggers were not only observed variables, but also the output of more low-level algorithms. These low-level algorithms, e.g. IP2D and SV1, concern themselves with characteristic features of the jets like the impact parameter of the contained tracks or the reconstruction of a secondary vertex inside the jet, when possible. These characteristics are crucial in the identification of jet flavour and the information, which was gained through the low-level algorithms in previous taggers, is still utilized in the newer neural network based models by the usage of auxiliary tasks. In GN1 and GN2 the track–origin prediction and the vertexing auxiliary tasks are used to recover the information about the structure of the jets and thus provide more physical context to the model, supporting the main task of jet–flavour prediction.

The defining qualities of heavy-flavour jets, e.g. a secondary or even a tertiary vertex in the jet and tracks with large impact parameter, also emerge in jets containing tracks of secondary origin. The decay-in-flight of long-lived strange hadrons, the hadronic interactions of hadrons with the detector material, or the conversion of photons, are all secondary effects modelled within the detector simulation. The motivation of this work was the expectation that light jets containing these secondaries are more likely to be misidentified as b-jets by the flavour-tagging models. In order to investigate this a labelling scheme was put into place, which categorizes tracks into NoTruth, NotSecondary, HadronicInteraction, StrangeMesonDecay, StrangeBaryonDecay, GammaConversion, and Other. This label, called track source, was used to confirm that the presence of secondaries in jets makes it harder for the model to identify the flavour of a jet. Additionally, the label can be used to employ different strategies to mitigate this effect. However, the labelling is not ideal, and would benefit from a more sophisticated way to differentiate between decay-in-flight and material interaction processes. Working this out was beyond the scope of a master's thesis project and would necessitate a larger effort in the collaboration, including experts in reconstruction and detector simulation.

A first approach to mitigate the deteriorating effects of secondaries on flavour tagging, using the implemented labelling, was studied and consisted of adding a further auxiliary task to the flavour-tagging model. This source task classifies the tracks of a jet into the categories of the source track label. One major challenge of this approach is the huge class imbalance resulting from the low number

of secondary tracks. Multiple models were trained, varying the weight of the additional auxiliary task in the linear combination of the losses and varying the weighting scheme of the source classes. While the source task itself performed well in most of these models, the impact of the additional task on the jet–flavour tagging performance was not unambiguously positive. There are setups improving the performance consistently, especially the rejection of light jets, as expected, but also setups with worse performing flavour tagging. Within the No Geant Thinning sample, offering more truth information on particles from material interactions, the source task has a consistent positive effect on the flavour predictions. However, the comparison of the NGT to a comparable default dataset did not confirm this inclination. But the potential improvement in flavour tagging through the source task was observed and warrants further investigation.

With the truth information put into place, performing further studies within the flavour-tagging workflow will be straightforward. The amount of jets used for further investigations should be increased, as it was a concern in the studies performed in this work. Should studies with NGT data be performed, they should also contain high p_{T} events like in the Z' samples. While this work was ongoing, the flavour tagging model was developed even further. Studying the effect of the source–prediction task on the GN3 model would be interesting as it uses different inputs and a new strategy of combining the losses of different tasks. Furthermore, modifications of the rest of the architecture alongside the addition of an auxiliary task could be investigated.

A different approach to handle secondaries was also proposed at the beginning of this work, for which the time of the project did not suffice. Instead of incorporating the identification of secondaries into the flavour-tagging model, a standalone tool could be developed for this purpose, possibly using an autoencoder architecture. The output of this could also be used by flavour-tagging models. One big advantage of this would be a more complete view of the secondary processes, because the tracks included in the flavour-tagging data have to pass stringent conditions to even be considered and have to be associated to a jet. The standalone tool would work on a much broader selection of tracks and the identification of tracks coming out of secondary processes would be the main task of this tool.

This thesis demonstrated how the continuous improvements of jet–flavour taggers within ATLAS lead to a situation, in which the consideration of seemingly minor or relatively rare effects, such as material interactions and other secondaries, can have a beneficial effect on the performance. Taking secondaries or effects of similar magnitude into account might even become necessary in future efforts to improve flavour tagging and similar challenges, enabling more precise measurements of known fundamental physics and a better chance at reaching for what is yet unknown.

APPENDIX A

Additional Figures

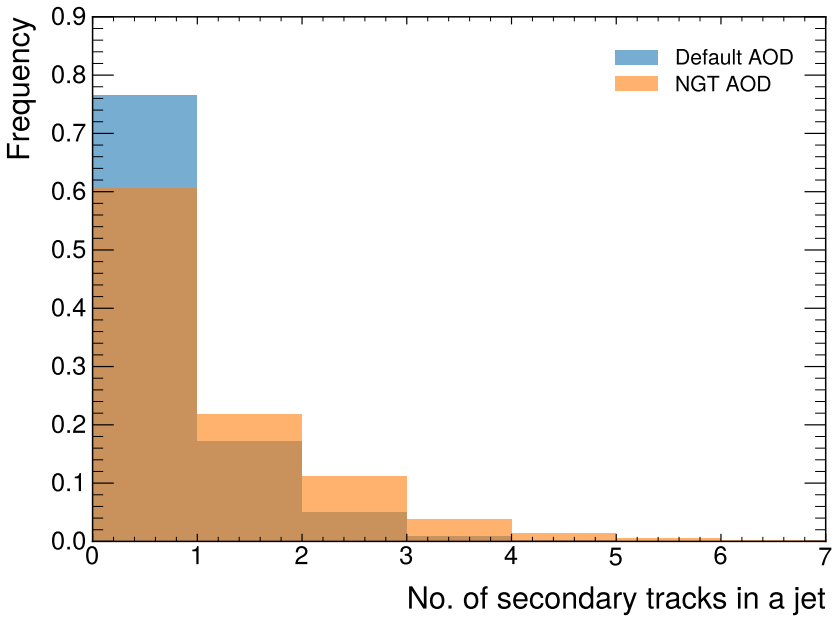


Figure A.1: Number of secondary tracks per jet for the default and the NGT $t\bar{t}$ sample.

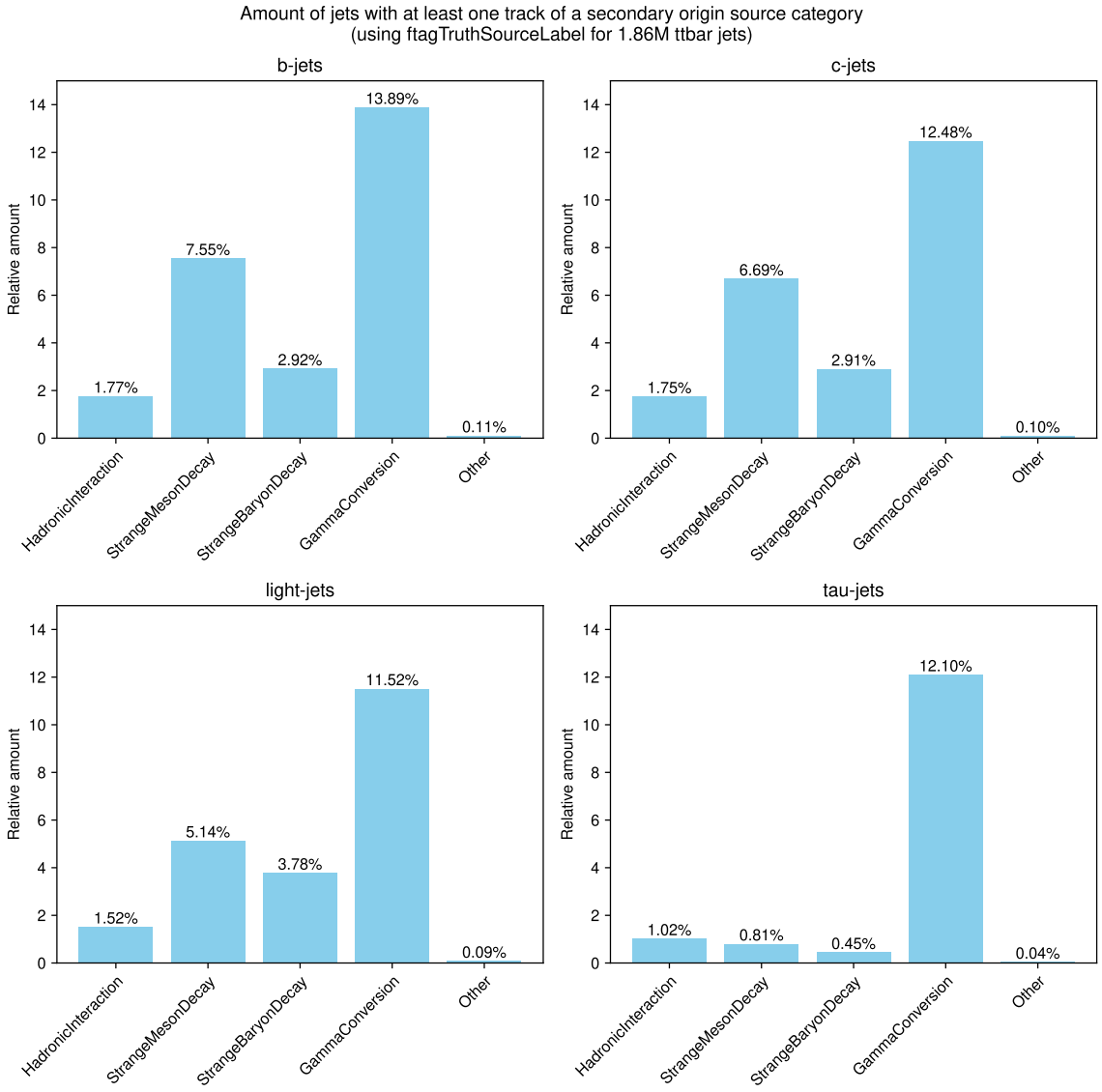


Figure A.2: Relative amount of jets containing at least one track of the secondary categories for the different flavours of jets in the $t\bar{t}$ sample.

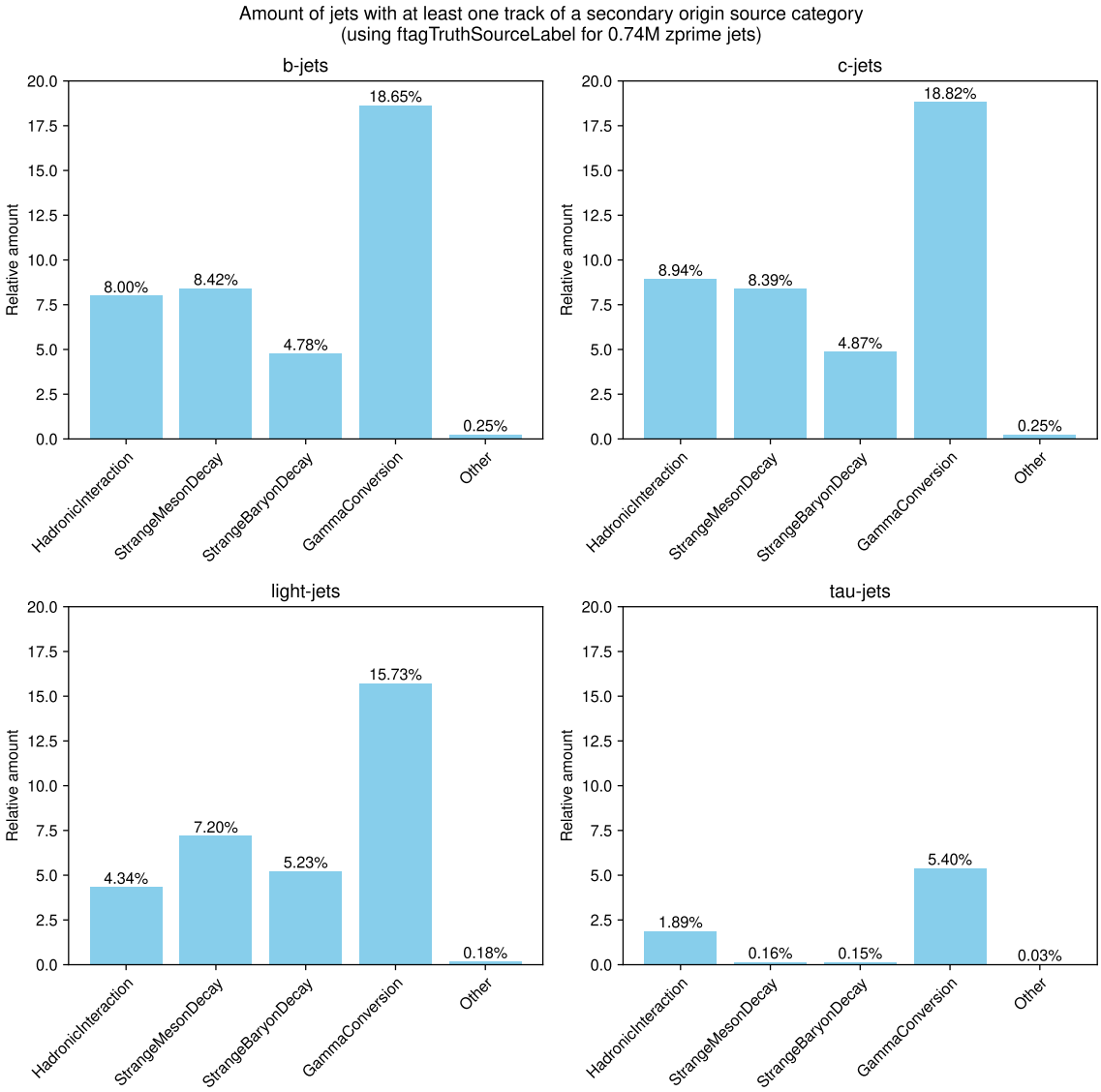


Figure A.3: Relative amount of jets containing at least one track of the secondary categories for the different flavours of jets in the Z' sample.

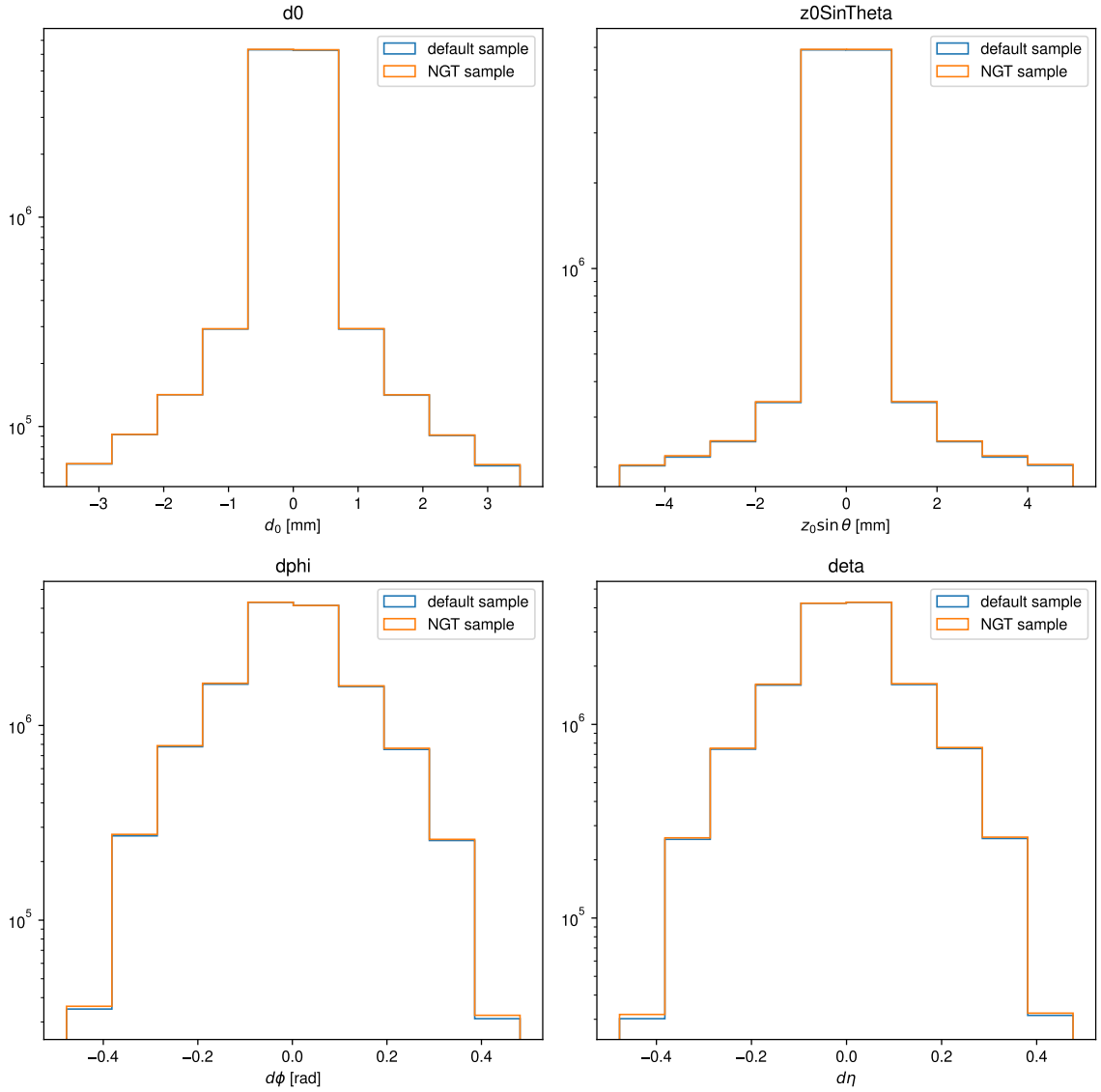


Figure A.4: Histograms of the four track input variables d_0, $z_0 \sin\theta$, $d\phi$, and $d\eta$ for GN2 comparing the default derivation and NGT derivation of the $t\bar{t}$ sample.

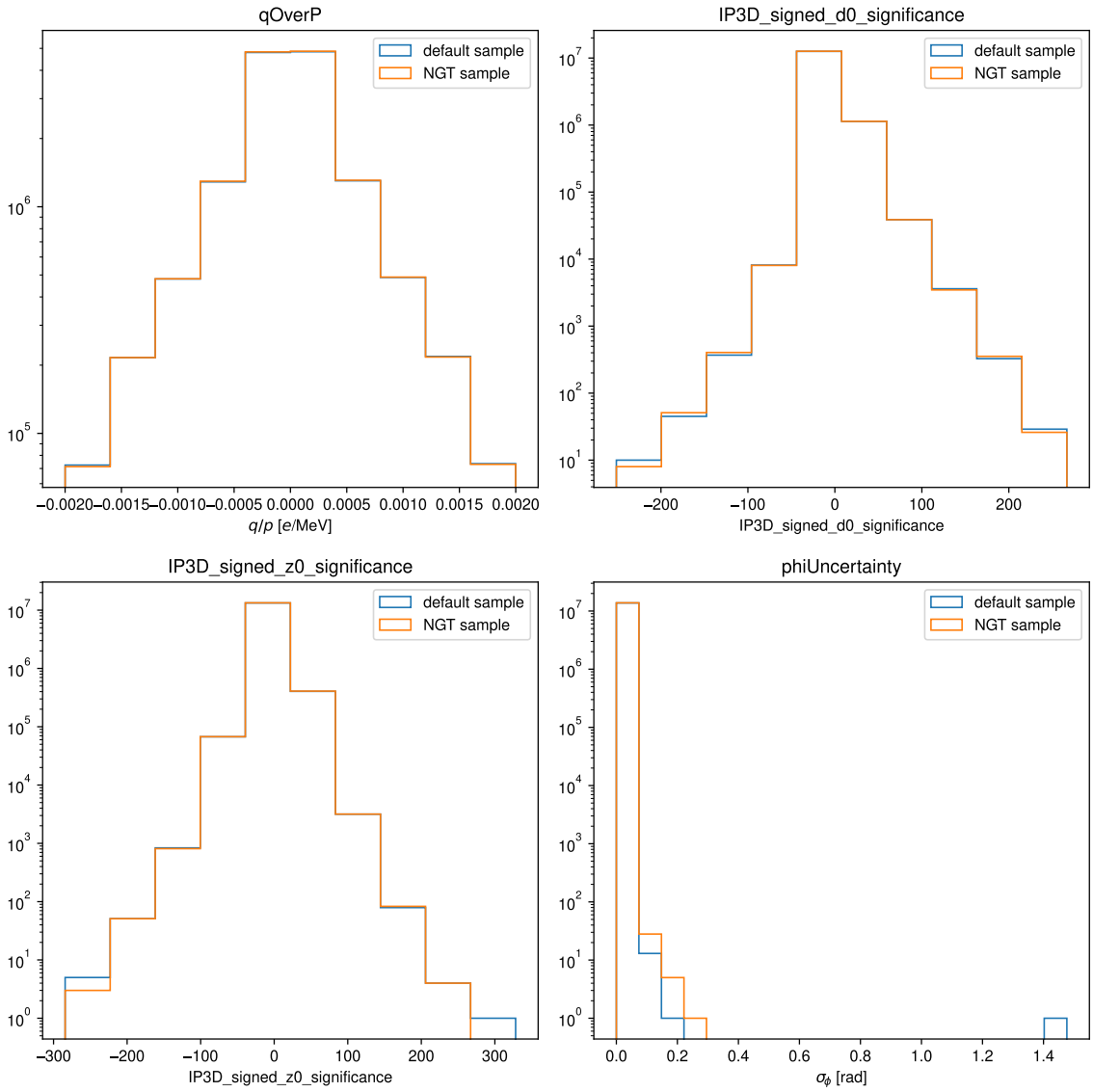


Figure A.5: Histograms of the four track input variables q/p, IP3D_signed_d0_significance, IP3D_signed_z0_significance, and σ_ϕ for GN2 comparing the default derivation and NGT derivation of the $t\bar{t}$ sample.

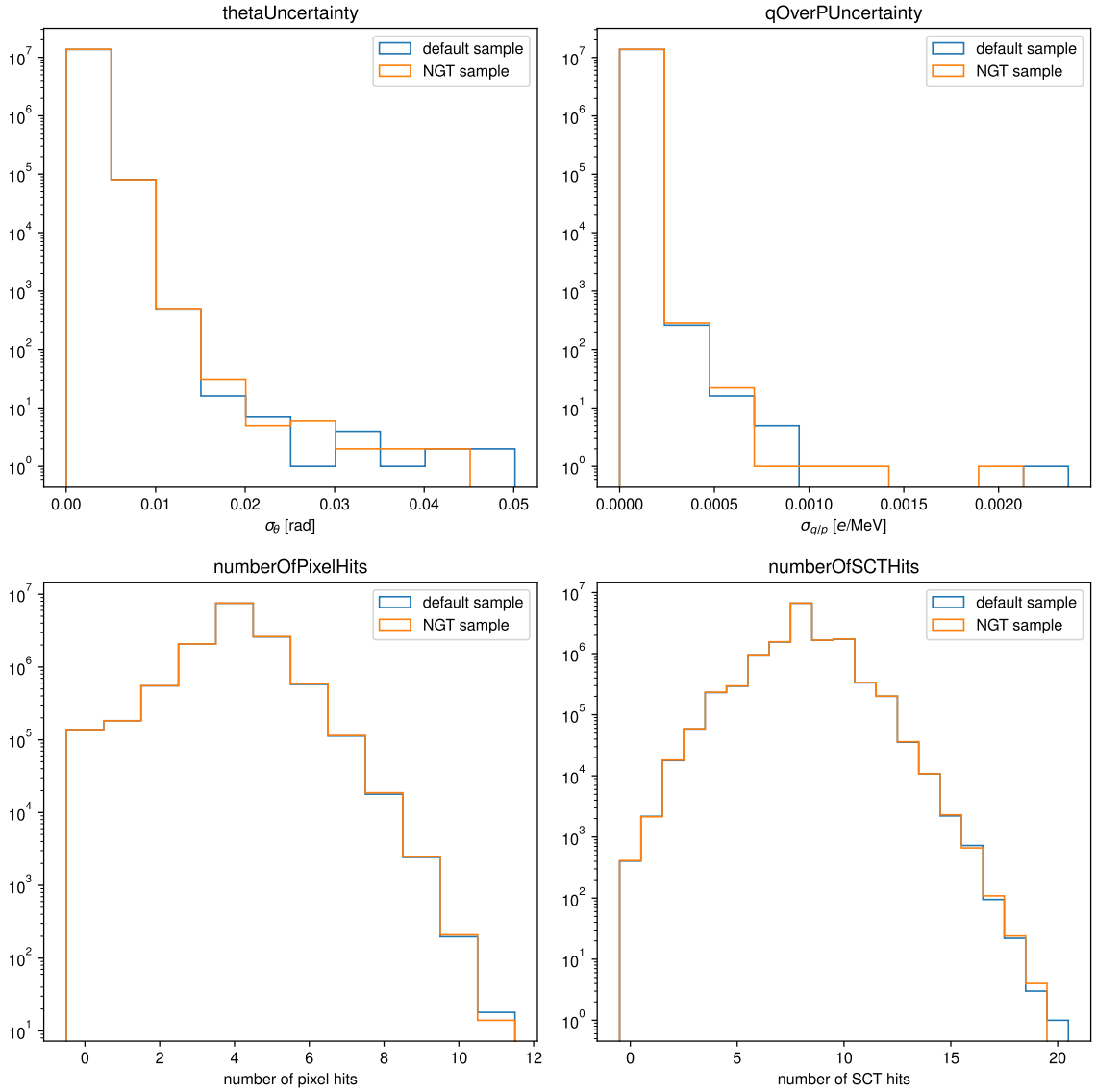


Figure A.6: Histograms of the four track input variables σ_θ, $\sigma_{q/p}$, number of pixel hits, and number of SCT hits for GN2 comparing the default derivation and NGT derivation of the $t\bar{t}$ sample.

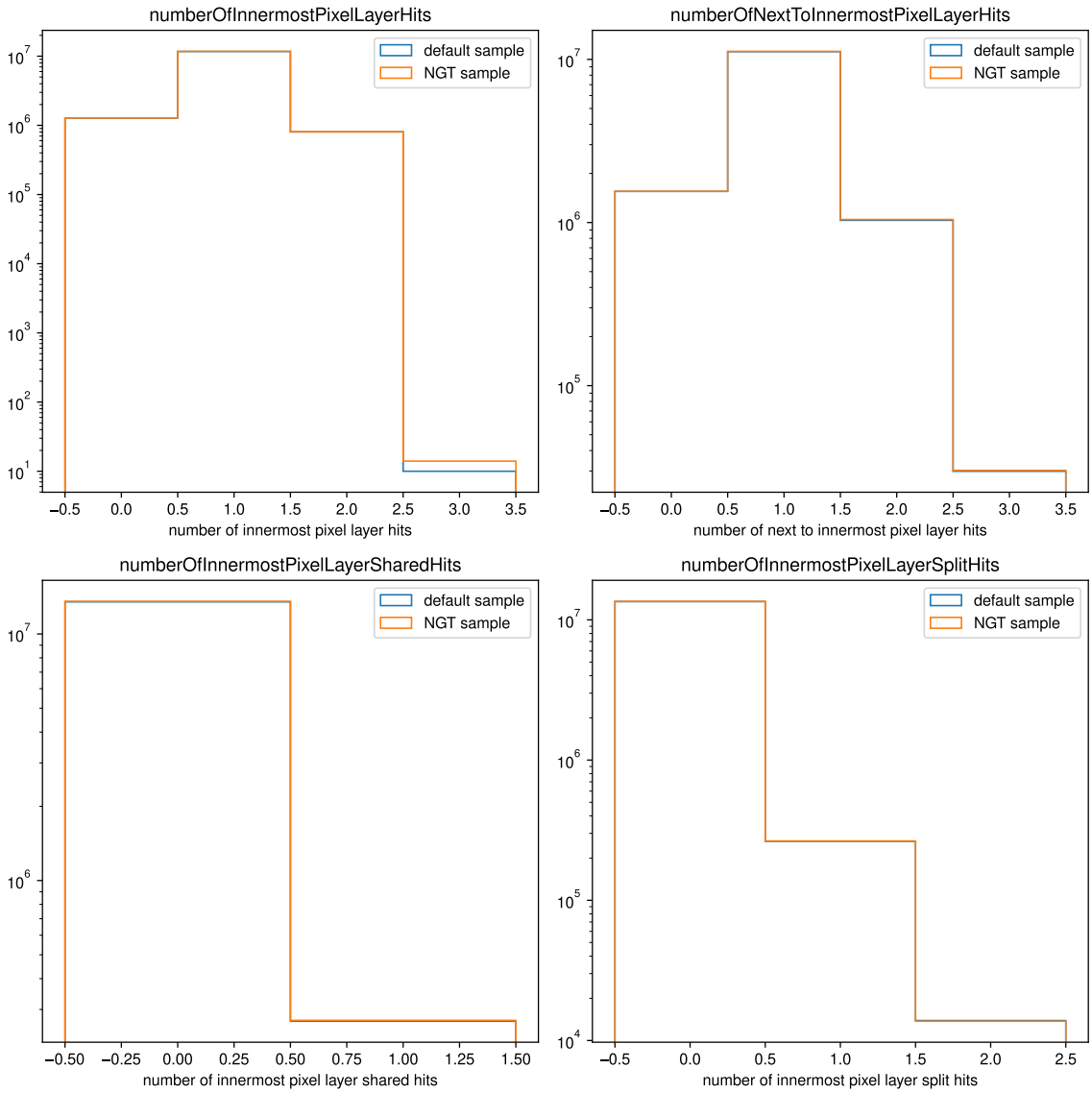


Figure A.7: Histograms of the four track input variables number of innermost pixel layer hits, number of next to innermost pixel layer hits, number of innermost pixel layer shared hits, and number of innermost pixel layer split hits for GN2 comparing the default derivation and NGT derivation of the $t\bar{t}$ sample.

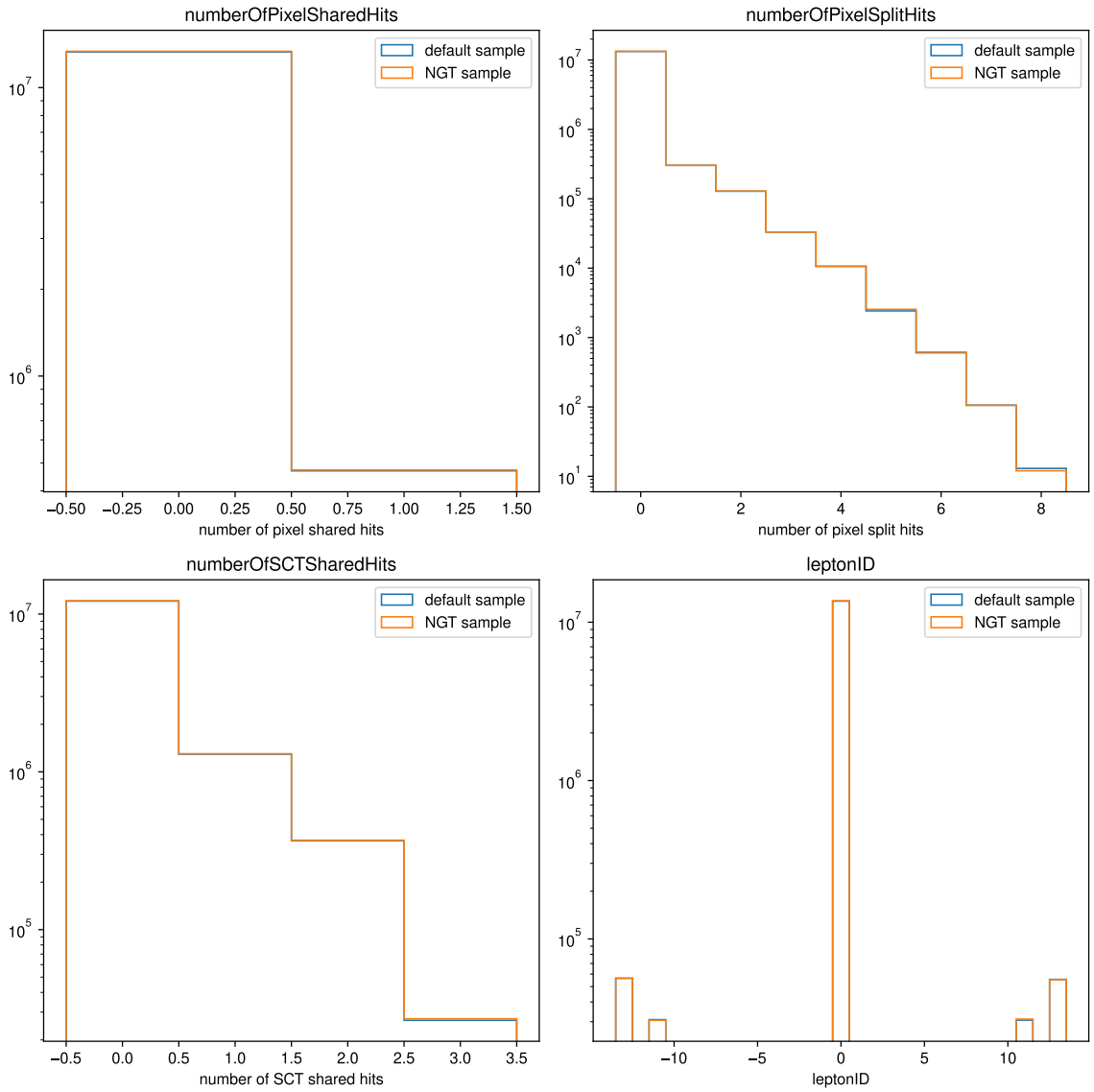


Figure A.8: Histograms of the four track input variables number of pixel shared hits, number of pixel split hits, number of SCT shared hits, and leptonID for GN2 comparing the default derivation and NGT derivation of the $t\bar{t}$ sample.

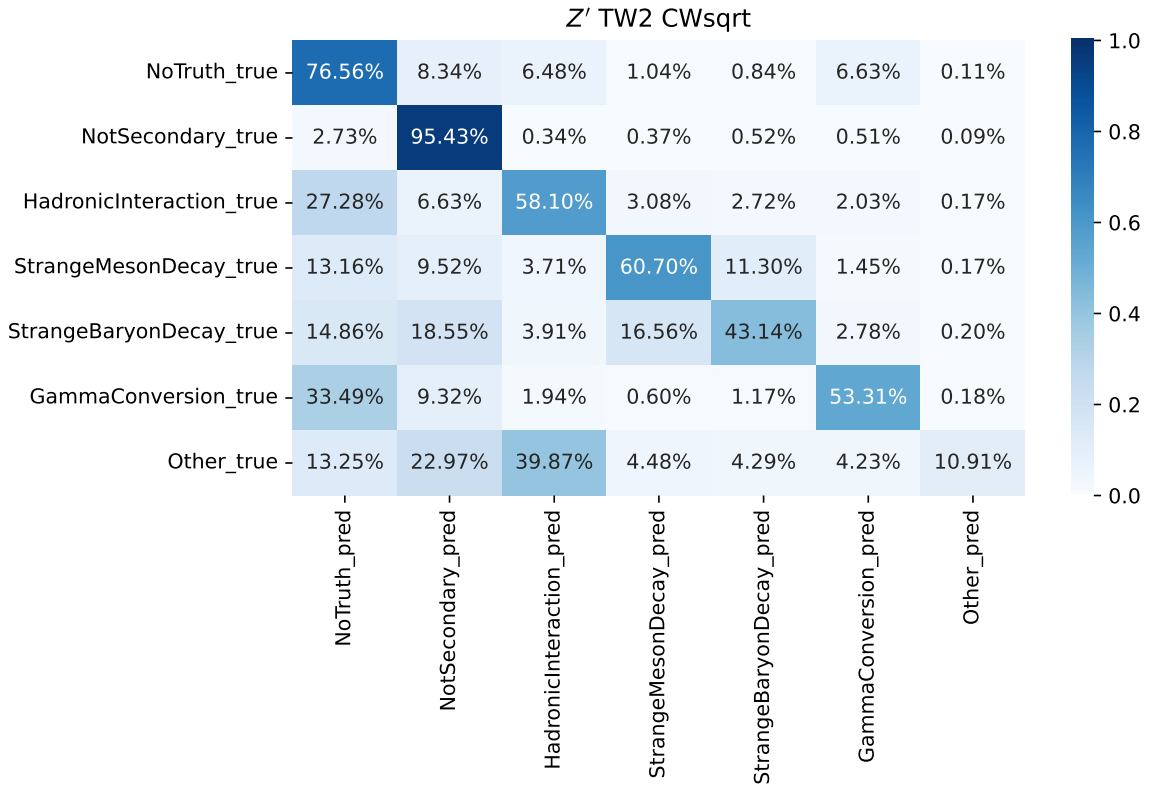


Figure A.9: Confusion matrix of the source prediction in the TW2-CWsqrt model evaluated on default Z' data.

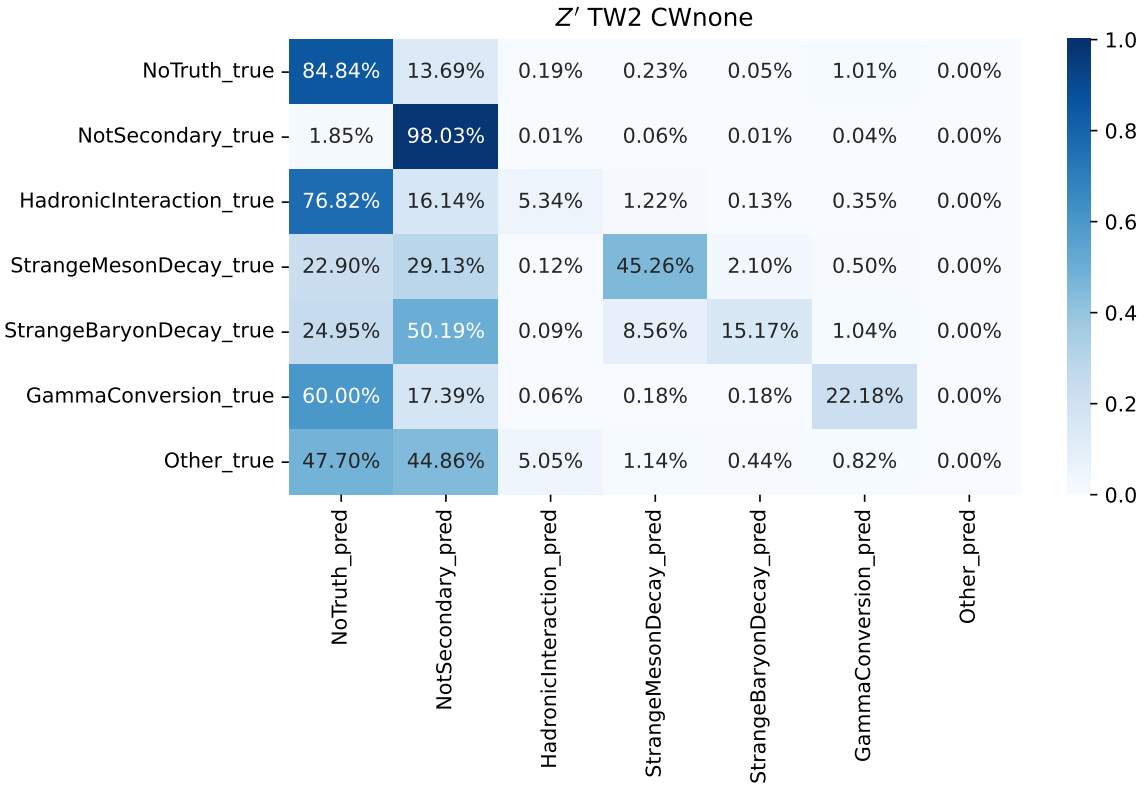


Figure A.10: Confusion matrix of the source prediction in the TW2-CWnone model evaluated on default Z' data.

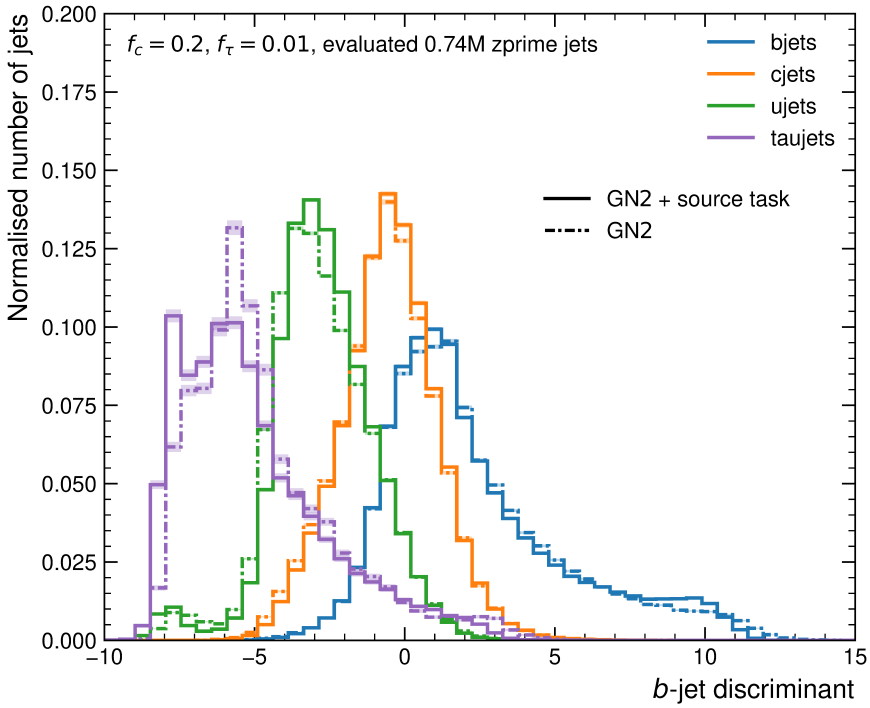


Figure A.11: The b-tagging discriminant of the TW2-CWsqrt model and the model without source task evaluated on the Z' data for light jets (ujets), c-jets (cjets), b-jets (bjets), and τ-jets (taujets).

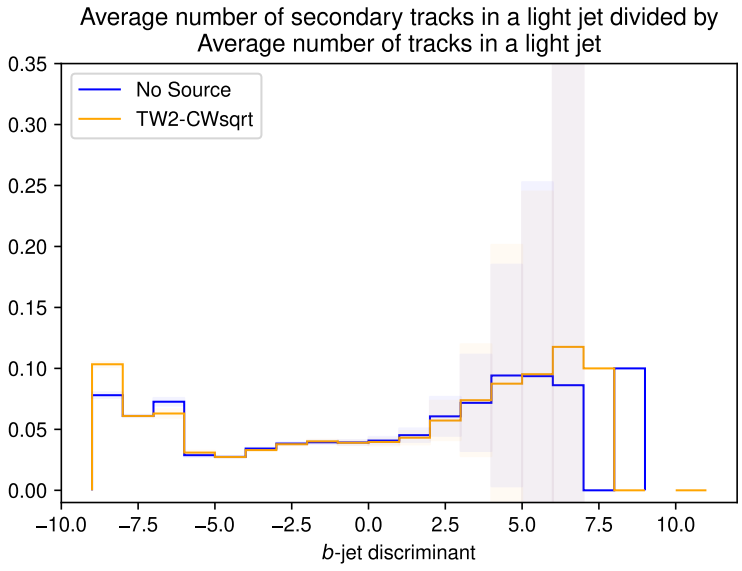


Figure A.12: Relative amount of secondary tracks in a jet over the b-tagging discriminant for the TW2-CWsqrt model and the model without source task, evaluated on Z' data.

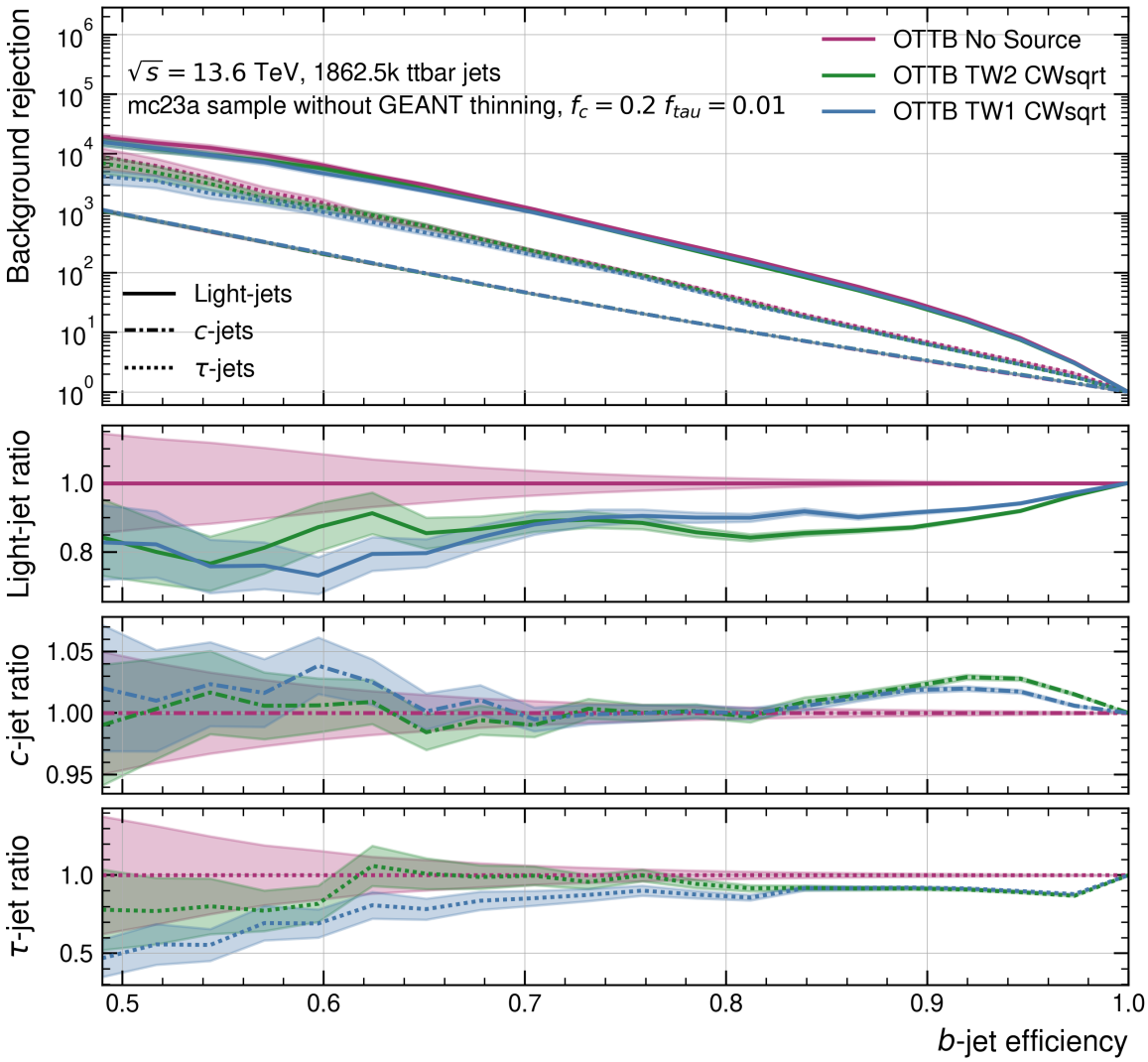


Figure A.13: ROC curve of the b-jet tagging efficiency and the background rejections of the model trained without the additional source task and all setups with the additional source task trained on the OTTB $t\bar{t}$ dataset.

Bibliography

[1] S. Navas et al. (Particle Data Group), *Review of Particle Physics*, Phys. Rev. D **110** (2024) 030001, and 2025 update, URL: https://link.aps.org/doi/10.1103/PhysRevD.110.030001.

[2] O. Brüning, H. Burkhardt and S. Myers, *The Large Hadron Collider*, Progress in Particle and Nuclear Physics **67** (2012) 705, URL: https://www.sciencedirect.com/science/article/pii/S0146641012000695.

[3] *CERN website*, URL: https://home.cern/science/experiments/atlas (visited on 02/02/2025).

[4] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012) 1, URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[5] ATLAS Collaboration, *Transforming jet flavour tagging at ATLAS*, 2025, arXiv: 2505.19689 [hep-ex], URL: https://arxiv.org/abs/2505.19689.

[6] ATLAS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B **716** (2012) 30, URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[7] *HL-LHC*, URL: https://project-hl-lhc-industry.web.cern.ch/content/project-schedule (visited on 22/04/2025).

[8] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation **3** (2008) S08003, URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[9] *ATLAS open data*, URL: https://opendata.atlas.cern/docs/documentation/introduction/introduction_ATLAS (visited on 07/02/2025).

[10] Y. Takubo, *ATLAS IBL operational experience*, PoS **Vertex2016** (2017) 004.

[11] ATLAS Collaboration, *Study of the material of the ATLAS inner detector for Run 2 of the LHC*, Journal of Instrumentation **12** (2017) P12009, URL: http://dx.doi.org/10.1088/1748-0221/12/12/P12009.

[12] ATLAS Collaboration,
Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data,
The European Physical Journal C **74** (2014) 3071,
URL: http://dx.doi.org/10.1140/epjc/s10052-014-3071-4.

[13] ATLAS Collaboration, *Search for massive, long-lived particles using multitrack displaced vertices or displaced lepton pairs in pp collisions at $\sqrt{8}$ TeV with the ATLAS detector*,
Physical Review D **92** (2015) 072004,
URL: https://doi.org/10.1103/PhysRevD.92.072004.

[14] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*,
The European Physical Journal C **70** (2010) 823,
URL: http://dx.doi.org/10.1140/epjc/s10052-010-1429-9.

[15] J. Catmore, *The ATLAS data processing chain: from collisions to papers*, 2016,
URL: https://indico.cern.ch/event/472469/contributions/1982677/attachments/1220934/1785823/intro_slides.pdf (visited on 05/05/2025).

[16] Geant4 Collaboration, *Geant4*, 2025,
URL: https://geant4.web.cern.ch/ (visited on 06/05/2025).

[17] ATLAS Collaboration, *AtlFast3: The Next Generation of Fast Simulation in ATLAS*,
Computing and Software for Big Science **6** (2022) 7,
URL: http://dx.doi.org/10.1007/s41781-021-00079-7.

[18] J. Boyd, *LHC Run-2 and Future Prospects*, 2019, URL:
https://indico.cern.ch/event/798971/contributions/3414162/attachments/1903821/3144264/StPetersburg-talk_jboyd.pdf (visited on 12/05/2025).

[19] A. Rimoldi et al., *First Report of the Simulation Optimization Group*, tech. rep.,
All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-SOFT-PUB-2008-002: CERN, 2008, URL: https://cds.cern.ch/record/1097789.

[20] Geant4 Collaboration, *QGSP-BERT physics list*, 2025, URL:
https://geant4.web.cern.ch/documentation/dev/plg_html/PhysicsListGuide/reference_PL/QGSP_BERT.html#hadronic-component (visited on 16/05/2025).

[21] CMS Collaboration, *Jets at CMS and the determination of their energy scale*, URL:
https://cms.cern/news/jets-cms-and-determination-their-energy-scale (visited on 13/05/2025).

[22] ATLAS Collaboration,
Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1,
The European Physical Journal C **77** (2017) 490,
URL: http://dx.doi.org/10.1140/epjc/s10052-017-5004-5.

[23] ATLAS Collaboration,
Jet reconstruction and performance using particle flow with the ATLAS Detector,
The European Physical Journal C **77** (2017) 466,
URL: http://dx.doi.org/10.1140/epjc/s10052-017-5031-2.

[24] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Physical Review D **96** (2017) 072002, URL: http://dx.doi.org/10.1103/PhysRevD.96.072002.

[25] ATLAS Collaboration, *Graph Neural Network Jet Flavour Tagging with the ATLAS Detector*, tech. rep., CERN, 2022, URL: https://cds.cern.ch/record/2811135.

[26] M. Cacciari, G. P. Salam and G. Soyez, *The anti-kt jet clustering algorithm*, Journal of High Energy Physics **2008** (2008) 063, URL: http://dx.doi.org/10.1088/1126-6708/2008/04/063.

[27] T. G. Cornelissen et al., *Updates of the ATLAS Tracking Event Data Model (Release 13)*, tech. rep., CERN, 2007, URL: https://cds.cern.ch/record/1038095.

[28] ATLAS Collaboration, *ATLAS Software Documentation*, 2024, URL: https://atlassoftwaredocs.web.cern.ch/internal-links/tracking-tutorial/idoverview/ (visited on 08/05/2025).

[29] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, Nuclear Instruments and Methods A **262** (1987) 444, URL: https://www.sciencedirect.com/science/article/pii/0168900287908874.

[30] G. Gaycken for the ATLAS collaboration, "Track and vertex reconstruction performance of the ATLAS detector", *Proceedings of the 42nd International Conference on High Energy Physics (ICHEP 2024)*, Geneva, 2024, URL: https://cds.cern.ch/record/2914607.

[31] ATLAS Collaboration, *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, The European Physical Journal C **83** (2023) 681, URL: http://dx.doi.org/10.1140/epjc/s10052-023-11699-1.

[32] ATLAS Collaboration, "Development of ATLAS Primary Vertex Reconstruction for LHC Run 3", *Connecting the Dots and Workshop on Intelligent Trackers*, 2019, arXiv: 1910.08405 [hep-ex].

[33] N. Bartosik, *B-tagging diagram*, 2025, URL: https://en.wikipedia.org/wiki/B-tagging#/media/File:B-tagging_diagram.png (visited on 19/05/2025).

[34] S. Mondal and L. Mastrolorenzo, *Machine learning in high energy physics: a review of heavy-flavor jet tagging at the LHC*, The European Physical Journal Special Topics **233** (2024) 2657, URL: http://dx.doi.org/10.1140/epjs/s11734-024-01234-y.

[35] M. Lanfermann for the ATLAS collaboration, *Deep Learning in Flavour Tagging at the ATLAS experiment*, tech. rep., 2018 764.

[36] ATLAS Collaboration, "Deep Sets for Flavor Tagging on the ATLAS Experiment", *Proceedings of Connecting The Dots 2020 (CTD 2020)*, 2020, URL: https://cds.cern.ch/record/2721094.

[37] Y. Liu, B. Zhuang, C. Shen, H. Chen and W. Yin,
Auxiliary Learning for Deep Multi-task Learning, Preprint, not peer-reviewed, 2019,
arXiv: 1909.02214 [cs.CV], URL: https://arxiv.org/abs/1909.02214.

[38] T. Standley et al., "Which Tasks Should Be Learned Together in Multi-task Learning?",
Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, vol. 119, Proceedings of Machine Learning Research, PMLR, 2020 9120, URL: http://proceedings.mlr.press/v119/standley20a.html.

[39] J. Zhou et al., *Graph neural networks: A review of methods and applications*,
AI Open **1** (2020) 57, ISSN: 2666-6510, URL:
https://www.sciencedirect.com/science/article/pii/S2666651021000012.

[40] A. Vaswani et al., *Attention Is All You Need*,
Advances in neural information processing systems **30** (2017) 5998.

[41] A. Duperrin, *Flavour Tagging with Graph Neural Network with the ATLAS Detector*,
Conference slides, Presented at the 30th International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS 2023), East Lansing, USA, 2023,
URL: https://cds.cern.ch/record/2855275.

[42] M. Draguet, "Flavour Tagging with Graph Neural Network at ATLAS",
Proceedings of 42nd International Conference on High Energy Physics — PoS(ICHEP2024), vol. 476, 2025 1002, URL: https://cds.cern.ch/record/2912358/.

[43] ATLAS Collaboration, *Simulation-based extrapolation of b-tagging calibrations towards high transverse momenta in the ATLAS experiment*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-003: CERN, 2021, URL: https://cds.cern.ch/record/2753444.

[44] ATLAS Collaboration, *Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at* $\sqrt{s} = 8$ *TeV*,
The European Physical Journal C **75** (2015) 303,
URL: http://dx.doi.org/10.1140/epjc/s10052-015-3500-z.

[45] ATLAS Collaboration, *GN3: Multi-task, Multi-modal Transformers for Jet Flavour Tagging with the ATLAS Detector: The GN3 Flavour Tagging Algorithm*,
ATLAS Internal Note ATL-COM-PHYS-2025-470, CERN, 2025,
URL: https://cds.cern.ch/record/2933750.

[46] J. Wagner, "Salt Tutorial", Flavour Tagging Workshop, 2023,
URL: https://indico.cern.ch/event/1311519/overview (visited on 25/07/2025).

[47] ATLAS FTAG, *TDD Documentation*, 2025,
URL: https://training-dataset-dumper.docs.cern.ch/ (visited on 17/06/2025).

[48] ATLAS FTAG, *UPP Documentation*, 2025,
URL: https://umami-hep.github.io/umami-preprocessing/ (visited on 17/06/2025).

[49] ATLAS FTAG, *Salt Documentation*, 2025,
URL: https://ftag-salt.docs.cern.ch/ (visited on 17/06/2025).

[50] Comet ML, Inc., *comet*, 2025,
URL: https://www.comet.com/site/ (visited on 18/06/2025).

[51] ZIMT, Universität Siegen, *OMNI-Cluster*, 2025,
URL: https://cluster.uni-siegen.de/ (visited on 18/06/2025).

[52] N. B. Krengel, *Salt Fork with Source Task*, 2024,
URL: https://gitlab.cern.ch/nkrengel/salt/-/tree/source-aux-task/salt/configs?ref_type=heads (visited on 18/06/2025).