GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Fakultät für Physik

Master's Thesis

Exploring Deep and Graph Neural Networks for Prompt Lepton Tagging at $\sqrt{s} = 13.6$ TeV with the ATLAS Experiment

prepared by

Tim Schlömer

from Cloppenburg

at the II. Physikalischen Institut

Thesis number:	II.Physik-UniGö-MSc-2024/08
Thesis period:	1st April 2024 until 30th September 2024
First referee:	Prof. Dr. Arnulf Quadt
Second referee:	PD Dr. Ralf Bernhard

Contents

1. Introduction

The Standard Model of particle physics is the state of the art theory describing the behaviour of fundamental particles and their interactions. However, despite its success, the Standard Model leaves many questions unanswered, meaning further research remains inevitable. In this thesis, alternate approaches to an advanced detection technique are studied. The Prompt Lepton Improved Veto (PLIV) is used for lepton tagging. This tool aims at enhancement of the discrimination between rare events and background processes by exploiting the properties of prompt leptons, electrons and muons originating directly from the decay of heavy particles. By refining the ability to distinguish prompt leptons from those arising from background sources, the sensitivity of experiments can be significantly enhanced and the understanding of the underlying physics can be improved. In this study, approaches for prompt lepton tagging with Deep and Graph neural networks are presented.
The thesis is structured into 6 chapters, starting with an introduction to the Standard Model and the theory of the relevant processes in Chapter 2. In Chapter 3, the experimental setup of the Large Hadron Collider and the ATLAS experiment is presented. Afterwards, in Chapter 4, prompt and non-prompt leptons are defined and their differences explained in detail. In the fifth chapter the fundamentals of the neural networks used in this thesis are presented. In Chapter 6 the results of the study on Deep and Graph neural networks are presented. In the end, a conclusion and an outlook are given.

2. The Standard Model of Particle Physics

The Standard Model of particle physics (SM) summarises the knowledge about particle physics. It contains all known elementary particles and the fundamental electromagnetic, weak and strong forces. The SM combines the U1, SU2, SU3 symmetries and predicts massive gauge bosons via the higgs mechanism. The associated Higgs boson was discovered in 2012 [1, 2] The SM is a very successful theory, predicting several particles before their discovery, most importantly the Higgs Boson. Despite its success, it is not complete, as gravity is not included in the SM, and neither dark matter nor dark energy can be described by it, despite them making up most of the universe.

2.1. Elementary particles and their interaction

The SM is the renormalisable gauge-invariant quantum field theory (QFT) with the symmetry $\mathrm{SU_C}(3) \times \mathrm{SU_L}(2) \times \mathrm{U_Y}(1)$. It is able to explain the existence of the massive gauge bosons via the higgs mechanism [3]. The strong interaction is described by quantum chromodynamics (QCD) with symmetry group $\mathrm{SU_C}(3)$, while $\mathrm{SU_L}(2)$ describes the weak interaction and $\mathrm{U_Y}(1)$ is the symmetry group for the electromagnetic interaction (quantum electrodynamics, QED) [4–7].
The SM differentiates between fermions with spin $\frac{1}{2}$ and bosons with integer spin. Fermions are further divided into leptons and quarks, with three generations being differentiated. Quarks carry a colour charge and interact strongly, while leptons do not. Per generation, there are two quarks and two leptons. In the first generation, there are the up and the down quark, the electron and its neutrino, as seen in Figure 2.1. These particles are all stable and make up the baryonic matter in our universe. In the second generation, there are the charm and the strange quark, the muon and its neutrino. The third generation consists of the top and the bottom quark, the tau and the tau-neutrino. In total, there are 12 fermions, as well as 12 antifermions with opposite electric charge.

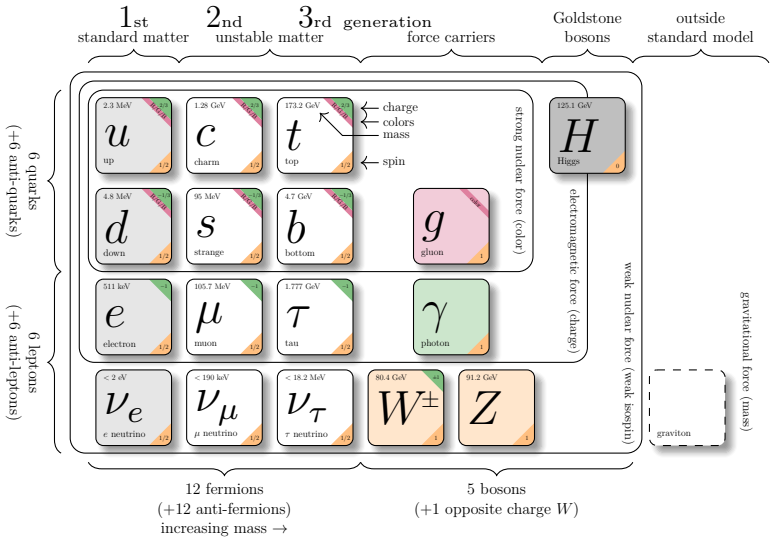

Figure 2.1.: Particle content of the SM of particle physics.

Two important concepts for the weak interaction are helicity, which is the projection of a particles spin onto its momentum, and chirality, which is an abstract quantum mechanical concept. Particles with positive chirality are called "right-handed", while particles with negative chirality are "left-handed". In the relativistic case, where the particles mass can be neglected, the helicity is equal to the chirality. The weak interaction couples to right- and left-handed particles differently. While the Z boson couples to both kinds, the W boson only couples to left-handed particles. This is described by weak isospin doublets or singlets. The right-handed fermions form singlets, while left-handed fermions are described by a doublet. The W boson only couples to left-handed particles, which means only left-handed particles can decay weakly.

Left-handed fermions are divided into up- and down-type fermions by the third component of their weak isospin (I_3). The up-type quarks with $I_3 = 1/2$ consist of up, charm and top quark. Meanwhile down, strange and bottom quark are the down-type ($I_3 = -1/2$) quarks. For the leptons, the neutrinos are the up-type leptons, and the electrically charged leptons (electron, muon, tau) are down-type.

It is important to note, that the particles in the isospin doublets are the weak eigenstates of the particles. These are related via the Cabibbo Kobayashi Maskawa (CKM) matrix to the mass eigenstates, which are usually used for particle description. The CKM matrix describes the mixing of the different flavours in the mass eigenstates, as seen in Equation (2.1). For example, the third generation doublet contains the top quark and the b' quark, and therefore describes the coupling of the top quark to the down quark (V_{td}), the strange

quark (V_{ts}) and the b quark (V_{tb}).

$$\begin{bmatrix} |d'\rangle \\ |s'\rangle \\ |b'\rangle \end{bmatrix} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \begin{bmatrix} |d\rangle \\ |s\rangle \\ |b\rangle \end{bmatrix} \tag{2.1}$$

This mechanism allows decays across generations, which is possible in weak interactions only.

The bosons are the force mediators in the SM. The first discovered boson was the photon, the mediator of the electromagnetic force, as one of the important discoveries that started the postulation of quantum mechanics [8, 9]. The next boson discovered was the gluon, the boson of the strong interaction, at DESY in 1978 [10]. The weak W and Z bosons were discovered in 1983 at CERN [11, 12], and were predicted by the electroweak theory formulated in 1968 [5–7]. The last addition was the Higgs boson in 2012, when it was discovered at CERN by ATLAS and CMS [1, 2].

The photon and the gluon have no mass and no electric charge. They mediate the electromagnetic force and strong force, respectively. The weak interaction is mediated by W and Z bosons. The W boson has either electric charge $+1$ or -1, a mass of 80.377 ± 0.012 GeV [13] and spin 1. The Z boson has electric charge 0 and a mass of 91.1876 ± 0.0021 GeV [14]. All of these bosons have spin 1, and are called vector bosons.

The last boson to be introduced into the SM is the Higgs boson, the boson of the Brout-Englert-Higgs mechanism [3]. The coupling between fermion fields and the Higgs field is responsible for the mass of the fermions. The mechanism was introduced to explain the masses of the W and Z boson, because without the spontaneous electroweak symmetry breaking of the Higgs mechanism, massive gauge bosons cannot exist, since they would spoil gauge invariance. The Higgs boson has no electric charge and spin 0, with a mass of 125.25 ± 0.17 GeV [15]. For fermions, the Yukawa-coupling descibes the coupling strength between the fermion field and the Higgs field, which leads to gauge-invariant masses for the fermions.

2.2. The top quark

The top quark was discovered in 1995 by DØ and CDF at the TEVATRON [16, 17] at FERMILAB in the US. Its measured electric charge of $\frac{2}{3}$ and spin $\frac{1}{2}$ are consistent with the SM prediction values [18]. The top quark is predicted to be an up-type quark, but so far there are no direct measurements of I_3.
The top quark is by far the heaviest elementary particle, with a mass of $172.08 \pm 0.39\,(\text{stat.}) \pm 0.82\,(\text{syst.})$ GeV [19], even higher than the massive bosons. This is of special interest, as it indicates a strong Yukawa-coupling in the order of unity, which has been measured as $Y_t = 1.16^{+0.07}_{-0.08}(\text{stat.})^{+0.23}_{-0.34}(\text{syst.})$[20]. This coupling is very sensitive to several theories beyond the standard model, e.g. two Higgs doublet models [21]. Therefore it is a very promising observable in the search for beyond standard model phenomena.
From the measured decay width of the top quark, a lifetime of $5 \cdot 10^{-25}\,\text{s}$ [22] has been calculated. This extremely short lifetime is less than the hadronisation $\Lambda_{had} \sim 10^{-23}\,\text{s}$ and the spin decorrelation ($\sim 10^{-21}\,\text{s}$) timescale. Therefore, the top quark will decay without forming hadrons, offering a possibility for studies of bare quarks. Also,the decay products of the top quark retain the spin information of the top quark. The CKM matrix shows that the top quark will almost always decay into a bottom quark as $V_{tb} \approx 1$ [23]. In these decays, a W boson is emitted, like in every flavour changing process.

2.2.1. Top Quark Pair Production

Top Quarks can to this day only be produced at hadron colliders, as there are no electron-positron colliders with a sufficient centre of mass energy of more than twice the top mass ($\approx$ 350 GeV) for $t\bar{t}$ production. Therefore, top quarks have only been produced at the TEVATRON, where they were discovered, and the LHC. The Feynman diagrams for $t\bar{t}$ production can be seen in Figure 2.2, with quark-antiquark annihilation being the dominant production mode at TEVATRON as it was a $p\bar{p}$ collider, and gluon-gluon fusion being dominant at the LHC.

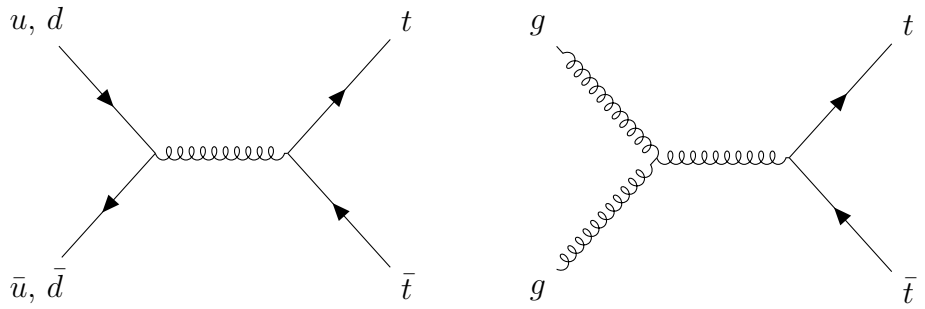

Figure 2.2.: Top quark pair production via quark-antiquark annihilation (left) and gluon-gluon fusion (right) in the s-channel. For gluon-gluon fusion, t- and u-channel diagrams contribute to the $t\bar{t}$ production as well.

2.2.2. Multi lepton processes

While top quarks can be produced in abundance at the LHC nowadays, some processes producing top quarks in association with other particles are rare. In Figure 2.3 the cross sections of several processes measured at the LHC are shown, which span five orders of magnitude. Observing these rare processes requires complex analyses that combine different decay channels and therefore different final states. These can then range from all-hadronic events, to final states requiring up to 4 leptons. For the number of events and the purity of the signal process in certain decay channels, depending on the number of leptons, simple estimates are possible by looking at the branching ratio of the weak bosons. Around 70 % of W/Z bosons decay into hadrons, with $\ell\,\nu$ making up the rest for the W, and $\ell\ell$ decays for Z bosons contributing 10 % [23]. This means that with a higher number of leptons, fewer events are available, as hadronic decays are more likely. At the same time, only very few processes produce prompt leptons. Therefore the number of background events drops as well, while the purity of the signal increases. This is used in many analyses, e.g. by requiring at least one lepton, removing a lot of background that is just producing jets.

For rarer processes, combinations of different channels are necessary to reach the required significance to observe them. This increases the importance of final states with several leptons, which are often ignored because they have low statistics and are difficult to reconstruct. A simple example for this would be the decay of a W-boson, where a neutrino is produced. While in a single lepton final state, the missing transverse energy can be used to estimate the neutrino, this is no longer possible for multiple neutrinos, as only their total sum of transverse energy would be the missing energy. This makes it significantly more difficult to properly handle the neutrinos.

Another challenge in the reconstruction of these multi lepton final states are so called

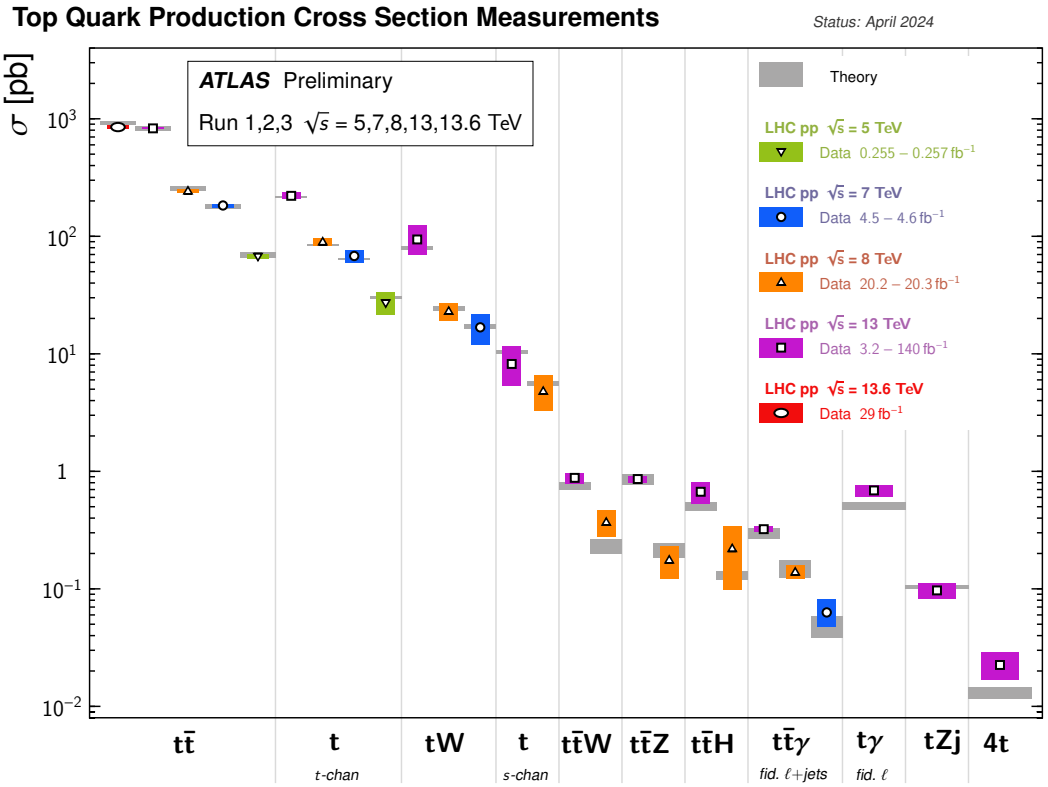

Figure 2.3.: Overview of measured cross sections of SM processes including top quarks at the LHC [24].

fake leptons, which mimic the signature of the prompt leptons from boson decays, but are produced by other processes like b-hadron decays, or are not even leptons at all. The properties and origins of prompt and non-prompt leptons are discussed extensively in Chapter 4. As an example, in Figure 2.4, data and prediction in a combined $t\bar{t}H$ and $t\bar{t}W$ analysis is shown. In both shown signal regions significant contributions from non-prompt electrons and muons, as well as other fake contributions as material conversion and charge misidentification can be seen. The contribution of the non-prompt leptons make up almost one third of the events in some bins. This shows how high precision prompt lepton tagging is key to perform such a multi lepton analysis succesfully.

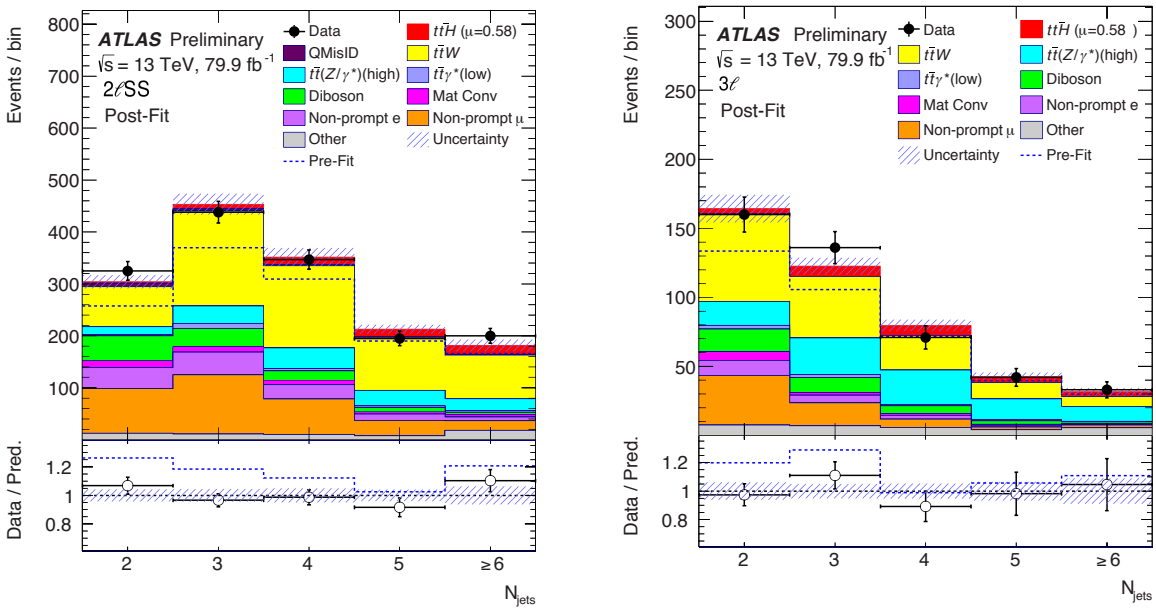

Figure 2.4.: Comparison between data and signal-plus-background prediction for the distribution of jet multiplicity in (a) the 2ℓ same sign channel and (b) the 3ℓ channel after event selection and before further event categorisation [25].

3. Experimental Setup

The experimental setup for this thesis is the ATLAS detector [26] at the Large Hadron Collider (LHC) [27], based at CERN in Geneva, Switzerland.

3.1. The Large Hadron Collider

The Large Hadron Collider is a circular hadron accelerator and collider with two beam pipes with four collision points. Most of the time, it operates as a proton-proton collider, but is also used to collide lead ions.
The LHC has several pre-accelerators, where the protons are accelerated step-by-step. Since 2020, the first accelerator is Linac4, where the proton beams are produced via acceleration of H^--Ions, which are stripped of their electrons on transition to the Proton Synchrotron Booster (PSB) where they reach an energy of 2 GeV. The next accelerators in line are the Proton Synchrotron (26 GeV) and the Super Proton Synchrotron (450 GeV). After these pre-accelerators, the proton beams with 450 GeV are injected to the LHC and accelerated to up to 6.8 TeV per beam. An overview of the LHC accelerator chain, the other accelerators and the experiments can be seen in Figure 3.1.
The LHC has two beam pipes with about 10,000 superconducting magnets along the beam line with a magnetic field strength up to 7.7 T. The two beam pipes are necessary, as the LHC is a proton-proton collider, and two beams are needed for the collisions. Beam pipes are kept at ultrahigh vacuum (less than 10 μPa) to minimise collisions with other particles, which would reduce the beam energy. At maximum energy, the protons reach more than 99.999 % of the speed of light around the accelerator ring, which has a circumference of 27 km. The proton beams consist of bunches of 10^{11} protons, separated by 25 ns [27].
Collisions then happen in one of the four interaction points, where the detectors recording the events are placed. These are ALICE [28], where lead ions are used to investigate the early universe, CMS [29], studying the SM and theories beyond it, LHCb [30], focusing on b quark physics, and ATLAS, also focusing on the SM and theories beyond, which will be explained in detail in the next section.

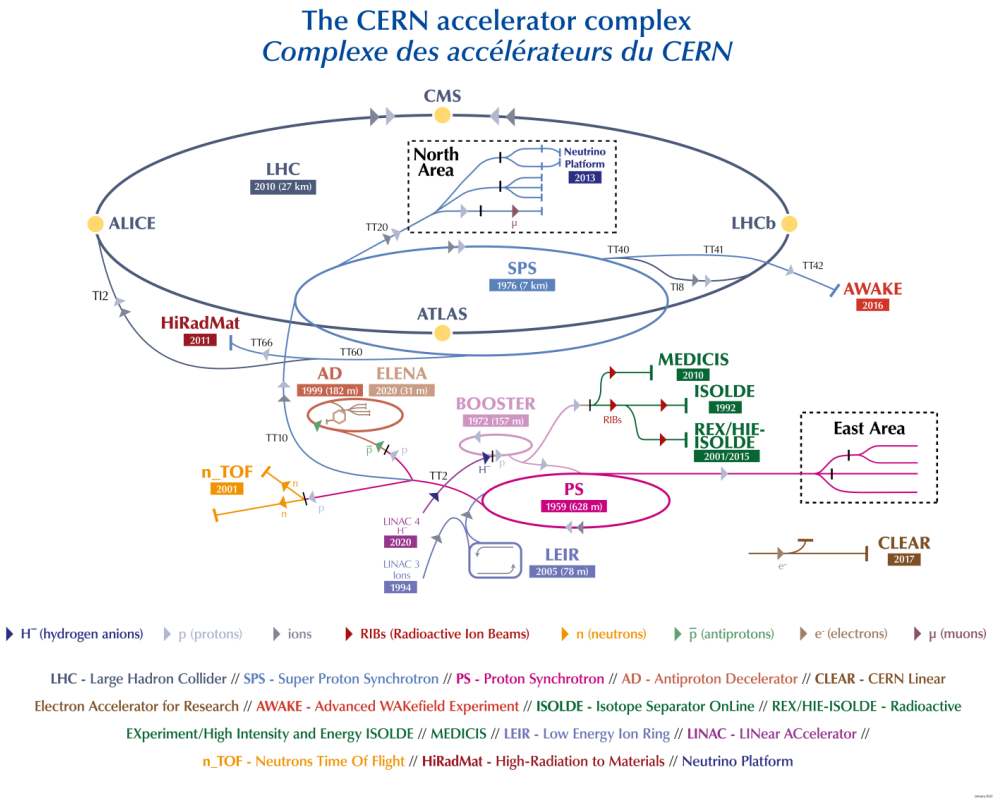

Figure 3.1.: Schematic depiction of the different accelerators and detectors around the LHC. © CERN

3.2. The ATLAS Detector

The ATLAS detector is the largest detector for particle physics at an accelerator. It is a general purpose detector, meaning it is built without a particular focus but instead can be used to investigate a broad range of processes. For this kind of detector, a coverage of the entire solid angle of 4π is the goal, which ATLAS achieves, except for the beam pipes themselves. It has a cylindrical form with a diameter of 25 m and a length of 44 m. To describe the detector, cylindrical coordinates are used, with the z-Axis along the beam line, the azimuthal angle φ and instead of the polar angle, the pseudorapidity $\eta = -\ln\left(\tan\frac{\theta}{2}\right)$. With the usage of the pseudorapidity, the masses of the particles are neglected, due the high energy achieved in the collisions.

At ATLAS, an instantaneous luminosity in the order of 10^{34} cm^{-2}s^{-1} can be achieved [26]. In Figure 3.2, a sketch of the ATLAS detector [26] can be seen. As a general purpose detector, it is able to record many different particle properties. For this purpose, it has multiple layers, as seen in Figure 3.3. The Inner Detector (ID) is embedded in a solenoid magnet with a field of 2 T. Around the magnet, the electromagnetic and

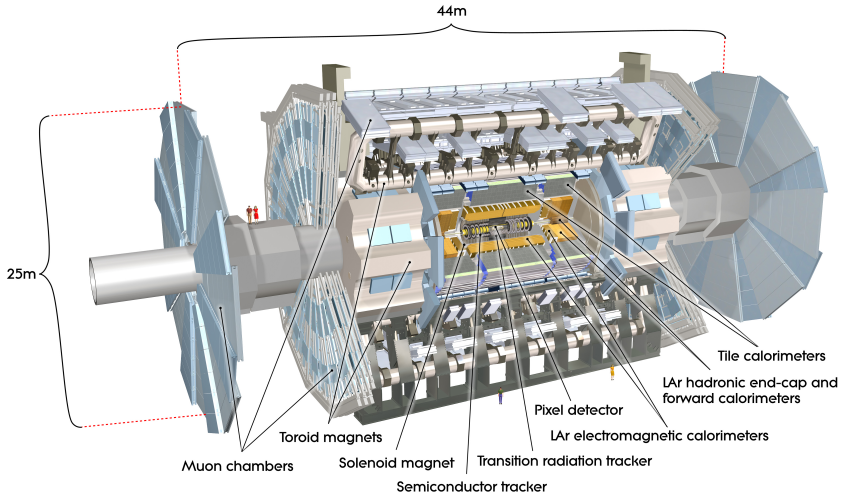

Figure 3.2.: Computer simulated image of the ATLAS detector. ©CERN

hadronic calorimeters are placed. The outermost layer is the muon spectrometer.

3.2.1. Inner Detector

The Inner Detector is the innermost layer of the detector, positioned as close to the beam line as possible, with a distance of only 3.3 cm. It measures the position of a particle in multiple layers, namely in the Pixel detector, the semi-conductor tracker and the transition radiation tracker [31]. In the Pixel detector, the position can be read out from hits in the detecting pixels, which are up to 50 by 400 μm small. The next layer is the semi-conductor tracker, which has silicon strips instead of pixels. Each strip measures 80 μm by 12 cm, spanning 61 m^2 in total. The transition radiation tracker consists of drift chambers (straws), each 4 mm in diameter and up to 144 cm long.
The ID is only able to record the tracks of charged particles, as it is based on electromagnetic interactions with the detector material and does not absorb particles. This can also be seen in Figure 3.3, where the muon, electron and proton leave tracks, while the electrically neutral particles do not. The magnetic field from the magnet around the ID bends the path of charged particles, therefore it is possible to reconstruct the charge (from the direction of the curvature) and momentum of the particles.
The tracking information is especially important to identify b-hadrons, because their relatively long life time leads to their decay products originating from a secondary vertex instead of the primary collision, making it possible to identify them, see chapter 3.3.

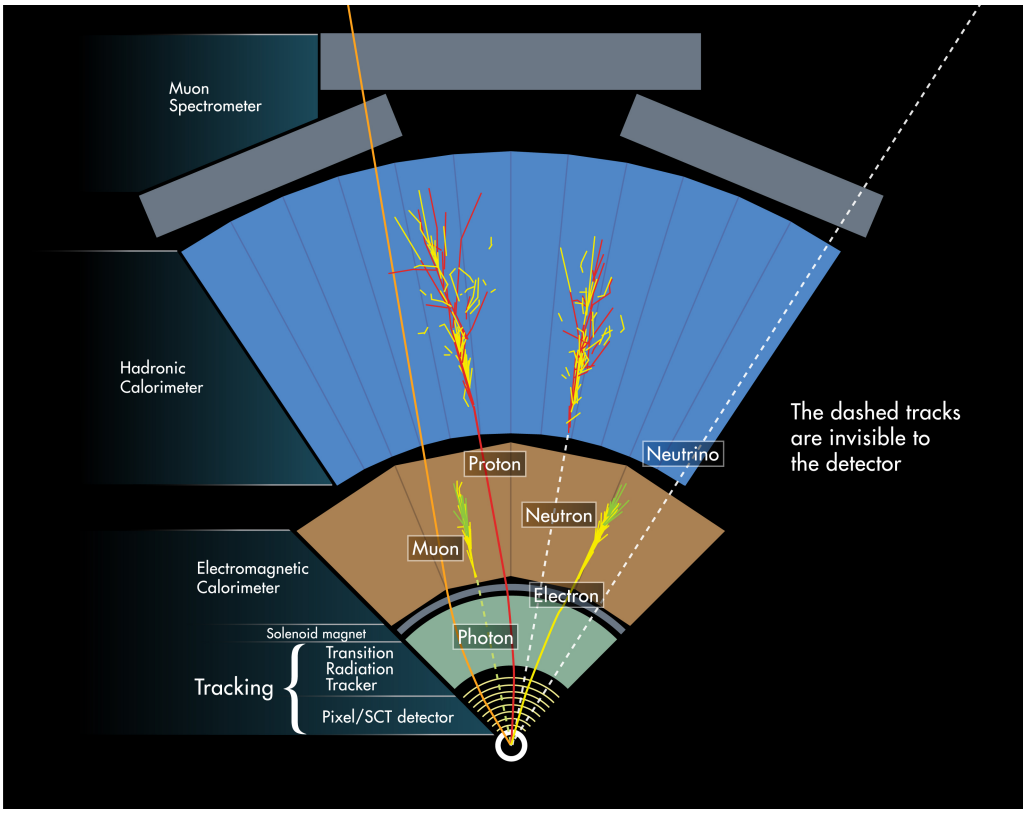

Figure 3.3.: Behaviour of particles in the detector © CERN

3.2.2. Calorimeter System

The purpose of the calorimeter is the measurement of the particle energy. In the calorimeters most particles deposit their energy, as seen in Figure 3.3. The exceptions are muons, as they are minimally ionising particles (MIPs), and neutrinos, which do not interact with the detector at all.

ATLAS uses an electromagnetic and a hadronic calorimeter [26]. The energy of the particles is measured by measuring the energy of the induced shower in the calorimeter. The electromagnetic calorimeter uses liquid argon as the detector medium and is used for the detection of electrons and photons. The hadronic calorimeter consists of absorbing layers made from steel, and scintillating tiles as active material, and is used to measure the energy of hadrons, e.g. protons and neutrons. Protons and other charged hadrons leave a trace in the electromagnetic calorimeter and then produce showers in the hadronic calorimeter (Figure 3.3). As quarks cannot exist in an unbound state, they undergo parton showering and hadronisation forming a spray of particles which may be clustered to jets. These are reconstructed from the calorimeter measurements [26].

3.2.3. Muon Spectrometer

The muon spectrometer tracks the path of muons, similar to the ID. It measures the deflection of muons in a magnetic field generated by superconducting air-core toroids, with field integrals between 2.0 and 6.0 Tm. The muon spectrometer is composed of a set of precision chambers comprising three layers of monitored drift tubes. In the forward region, where background levels are highest, cathode-strip chambers complement the setup. For the muon trigger system, in the barrel region resistive-plate chambers are used, while the endcap regions use thin-gap chambers. With the combination of ID and muon spectrometer, muons are easily identified [26].

3.3. Object Reconstruction

Object reconstruction is an important part of the analysis of the data recorded with a detector. The measurements of the detector have to be attributed to particles. For this, definitions of objects like electrons, muons or jets are needed. These definitions are based on the expected behaviour of particles. With the object definitions, reconstruction of the particles from the data is possible. Observables used for object definitions include p_{T},η and tracks. Tracks are the trajectories of charged particles left in the ID and MS of the detector. They are constructed from the ID using the ATLAS New Tracking (NEWT) algorithm [32], using the SCT and pixel hits.

Electrons

Electrons leave a track in the ID, and produce a shower in the electromagnetic calorimeter. Therefore, to reconstruct an electron, energy clusters in the electromagnetic calorimeters which match with a reconstructed track are used [33]. The detector can detect electrons with $|\eta| < 2.47$, except for the $1.37 < |\eta| < 1.52$ region, since this is the transition between barrel and end-caps. This is also sometimes used to separate between barrel ($|\eta| < 1.37$) and end-cap ($|\eta| > 1.52$) electrons. They are required to have a transverse momentum larger than 10 GeV. The impact parameters d_0 and z_0 are also used in the definition, with d_0 divided by its uncertainty $\sigma(d_0)$ is required to be less than 5. Also, $|z_0 \sin(\theta)| < 0.5\,\mathrm{mm}$ is required. Additionally, the normalised transverse energy in a cone around the lepton candidate ($E_T\mathrm{TopoCone30}/p_T$) must be < 0.3. Additionally, a 'loose' likelihood-based

identification criterion is applied [33]. This is calculated via

$$L_{S(B)}(x) = \prod_{i=1}^{n} P_{S(B),i}(x_i)$$

with P being the probability density function [33]. Other working points included are 'very loose', 'medium' and 'tight'.

Muons

Muons are reconstructed by combining ID tracks with tracks in the muon spectrometer [34]. For muons, similar criteria to electrons are applied, with a few numerical changes, and the identification criterion now required to pass 'medium' quality. The muon has to have $|\eta| < 2.5$ and $d_0/\sigma(d_0) < 3$, with $|z_0 \sin(\theta)| < 0.5\,\text{mm}$ and $E_T\text{TopoCone30}/p_T < 0.3$ as for electrons.

Jets

As bare quarks cannot exist, they hadronize and produced multiple hadrons. These hadrons appear in collimated streams, which are then called jets. For top quark analyses, b quarks are of great importance. They can be identified using their considerable lifetime ($\sim 10^{-12}\,\text{s}$) [23], in which they travel a significant distance, differentiating them from other jets. They are required to have $p_\text{T} > 25$ GeV and $|\eta| < 2.5$.
Jets are reconstructed with the anti-k_t algorithm [35] with $R = \sqrt{\phi^2 + \eta^2} = 0.4$ as radius parameter. The b-tagging is done via the "DL1r" algorithm [36], which uses the reconstructed track and secondary as well as tertiary vertex information. The working point (WP) is at 70%, therefore the jet has to have a b-tagging discriminant value larger than for a 70 % efficient selection.

4. Prompt and Non-Prompt Leptons

4.1. Definition

When W and Z bosons decay into leptons, these leptons are called prompt, as they are coming from the primary vertex due to the short lifetime of these bosons. In contrast, "fake" leptons do not originate from these boson decays, and therefore need to be distinguished from the prompt leptons needed for the boson reconstruction. One large contribution to fake leptons are "non-prompt" leptons, which originate from secondary decays, e. g. b- or c-hadrons, also known as heavy flavour (HF) fakes. Other fake lepton sources are non-prompt leptons coming from light quarks (light flavour (LF) fakes), or electrons produced from photon conversion.
As top quarks almost exclusively decay to a W boson and a b quark, fake leptons from semileptonic b-hadron decays are an important background in any analysis using top quarks. This is especially relevant in analyses with multiple leptons in the final state, as discussed in Section 2.2.2.

4.2. Separation Variables

Prompt and non-prompt leptons can be distinguished based on multiple features, with the two main categories of observables being isolation and lifetime information. For isolation, a cone with a certain $\Delta R = \sqrt{(|\Delta\phi|^2 + |\Delta\eta|^2)}$ is used, and the energy or momentum inside the cone is summed up. Lifetime information is based on the position of the vertex where the lepton is produced.
An example of the signature of leptons can be seen in Figure 4.1. The left lepton originates from the primary vertex (short lifetime), and has no other tracks in the isolation cone around it (isolated), therefore it will be identified as prompt. The right lepton originates from a displaced secondary vertex (large lifetime), and has other tracks originating from this vertex in the isolation cone. This lepton is non-prompt.

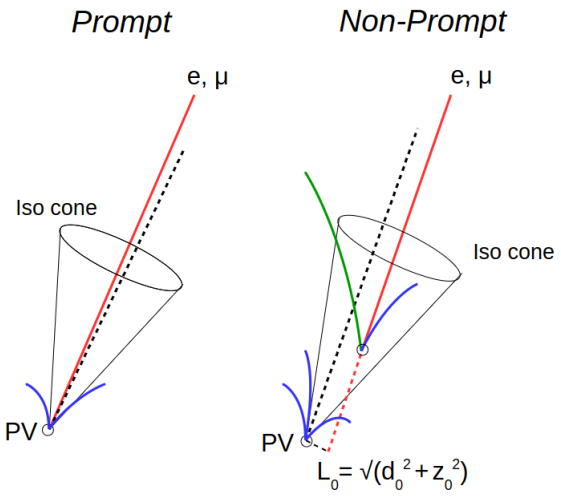

Figure 4.1.: Differences between prompt and non-prompt leptons.

4.3. PLIV in Run II

PLIV is the upgrade of the Prompt Lepton Veto (PLV) used in Run I. The goal of its development was an improvement of the performance for prompt lepton tagging by adding new observables. By doing studies on the properties of non-prompt leptons passing the PLV, new observables for lifetime and isolation were introduced.

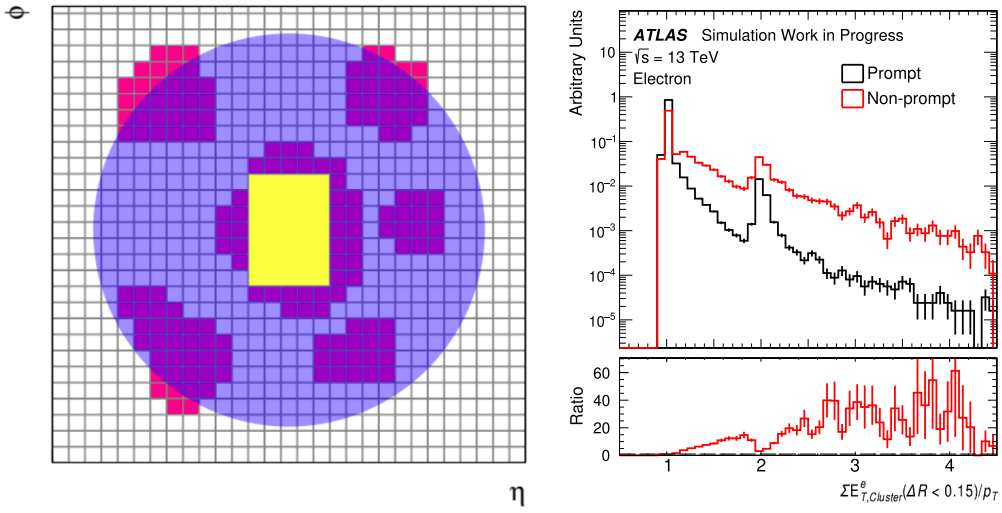

Figure 4.2.: Construction and distribution of the new observable $\Sigma E^{e}_{\mathrm{T,Cluster}}(\Delta R < 0.15)/p_{\mathrm{T}}$.

An example for the a new isolation variable for electrons is $\Sigma E_{\mathrm{T,Cluster}}(\Delta R < 0.15)/p_{\mathrm{T}}$. In Figure 4.2 on the left a sketch of the calorimeter clusters in the $\eta - \phi$ plane is shown.

For the usual calorimeter-based isolation variables the core energy (yellow) is taken as the lepton energy and excluded from the sum. For PLIV, the sum of the clusters in the centre of a smaller cone ($\Delta R < 0.15$) is calculated, and divided by the p_{T} of the electron. This keeps the information on the lepton p_{T} available. The difference in the distribution for prompt and non-prompt electrons is shown in Figure 4.2 on the right.
For lifetime information, fast vertex fitting [37] is used to reconstruct the secondary vertex of the B hadron decay. For further explanation, see Ref. [37]. One reconstructed vertex for the prompt lepton is expected, as there is a primary vertex. An additional secondary vertex is usually reconstructed for the non-prompt lepton. The $l^{longitudinal}_{\mathrm{SV\ to\ PV}}/\sigma$ observable refers to the longitudinal significance distance between the vertices.
An overview over all the included variables in PLIV for electrons can be seen in Section 6.1.

5. Neural Networks

Neural Networks (NN) are a powerful tool in the analysis of high energy physics events. With the increasing rarity of new processes, NNs are important to reach a high signal yield, while suppressing the background. The simplest network is the perceptron, invented in 1957 and inspired by the neurons in the human brain [38]. It takes an input $x = x_1, ..., x_n$, multiplies them with weights $w = w_1, ..., w_n$, sums them up, and applies a step function to determine the output, as shown in Figure 5.1. This way, the perceptron can perform a binary decision. One can now increase the complexity of the network by adding more neurons or "nodes".

5.1. Deep Neural Networks

Neural Networks, which consist of interconnected layers of nodes. These nodes are organised into three main types of layers: the input layer, which receives the raw input data, which is followed by hidden layers, which process the data through a series of transformations and the output layer in the end, which generates the final prediction or classification. If there are at least two hidden layers, the network is called "deep". If used for binary classification tasks in particle physics, one class is normally assigned a value of 0 (e.g. background), the other is assigned 1 (signal). One can then apply a threshold of 0.5, for example, to predict an event as one class or the other.
The single nodes in the network work similar to the perceptron explained above. Every node takes the outputs from the previous layer, applies weights and computes its own output. If every node in every layer receives the inputs of all nodes in the layer before, it is called fully connected. Additionally, a bias term can be added. The bias an offset that is added to the output of a layer. Onto the computed value of the node an activation function is applied. Activation functions introduce non-linearity into DNNs, allowing them to model complex, non-linear relationships in data. Without them, the network would behave like a linear model, regardless of the number of layers, severely limiting its

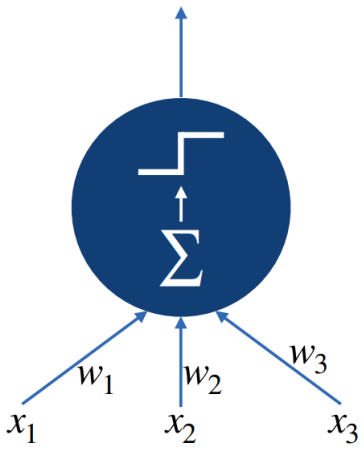

Figure 5.1.: Single perceptron which sums over the weighted inputs and applies a step function. The output is 1 if the threshold is passed, otherwise the output is 0.

capabilities. One important function is the Rectified Linear Unit function

$$ReLU(z) = \max(0, z),$$

which keeps the outputs from becoming negative. Other relevant activation functions are the Exponential Linear Unit (ELU)

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{else,} \end{cases}$$

which can range from -1 to positive infinity, and the tangens hyperbolicus

$$\tanh(x) = \frac{\sinh x}{\cosh x} = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}},$$

which can only take values between -1 and 1. Additionally, one can use the sigmoid function, which is especially used for the output layer

$$\sigma_{\mathrm{sig}}(z) = \frac{1}{1 + e^{-z}},$$

in binary classification. This makes sure that the output of the network always stays between 0 and 1, which is desired for classification tasks.
For the training of the network, a metric is needed to evaluate the prediction based on the true label. This metric is called the loss function, $J(\vec{\theta})$. For the binary classification task that will be performed, the binary cross-entropy loss will be used, which is defined

as

$$J(\vec{\theta})_{BCE} = \frac{1}{m}\sum_{i=0}^{m}\left[y_i \log\left(\hat{y}_i\right) + (1-y_i)\log\left(1-\hat{y}_i\right)\right],$$

with for every instance i, the true label being y_i, while the DNN prediction is $\hat{y}_i$, the size of the training sample is given by m and the parameters of the DNN are summarised as $\vec{\theta}$.

An additional feature that can be used in the training of NNs is Dropout for certain layers, which are then called dropout layers. This can be used to limit the impact of certain nodes and connections. In each training step, there is a probability that certain nodes will drop, i.e. not getting any inputs and therefore not producing any output. This can also happen for multiple steps back to back, as the dropout probability is independent of previous dropouts. This forces the network to use all nodes to keep the loss low even if important nodes drop out. This is only applied in the training, as in the evaluation and application, the network should be able to use its entire architecture.

For validation of the training, 2-fold cross validation is used. All events are split into two statistically independent subsets (folds) of equal size. Afterwards, one fold is used for the training. The remaining one is used for testing. From the training fold, 25% of the events are set aside as a validation subset. Both, the validation subset and the testing fold, are not used for the training of the DNN. The DNN is trained on the 75% of events remaining in training fold, with validation after each epoch on the validation set. After training, the other fold can be used as an independent measure of the performance of the Network, making sure the evaluation happens on events the network has not yet seen. Every fold is used as testing fold exactly once, so in the end, there are two models trained, and the final result is a combination of each of the predictions.

Another important part of training a DNN is the prevention of overtraining. This occurs if the DNN has learned the important features for the general predictions, but starts to pick up specific properties such as statistical fluctuations in the training dataset. This reduces the performance of the network on data other than the training data. Therefore, an early stopping mechanism is included, which stops the training before the specified maximum number of epochs, if there is no further improvement. This is done by setting a minimum improvement for the loss in the validation.

The Neural Networks used in this thesis are built using PYTORCH [39], and the optimiser used to update the connection weights is the NADAM optimiser [40] which is the ADAM optimiser [41] but with incorporated Nesterov momentum [42] using the principle of stochastic gradient decent (SGD). This optimiser is used to minimise the loss.

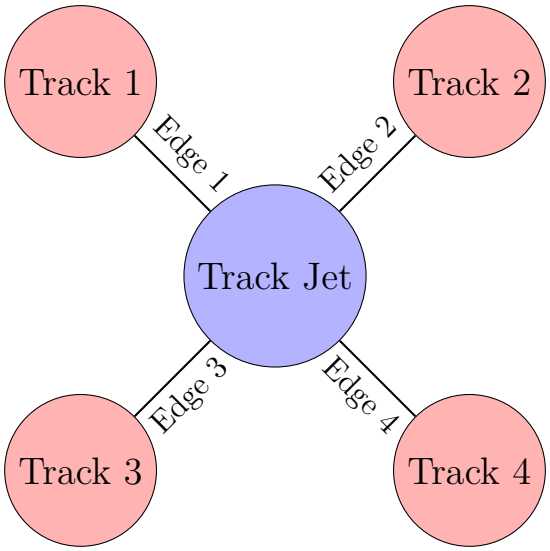

Figure 5.2.: Example of a Graph for an event with 4 tracks.

5.2. Graph Neural Networks

While DNNs operate well on constant structures such as grid like representations of data, for example in image classification, they struggle with more complex forms of data representation, e.g. graphs. Graph structures are used to model relationships between objects. A graph consists of two main components: nodes (also called vertices), which represent certain objects, and edges, which represent the connections or relationships between these objects. Graphs are highly versatile and can be used to model a wide range of systems across different domains. The power of graphs lies in their ability to represent and analyse complex systems by focusing on both the objects involved and the connections between them. An example for a graph can be seen in in Figure 5.2, which shows the graph for an event with one reconstructed track jet, and 4 reconstructed tracks. To each node and edge, multiple observables can be assigned. Ideally, they correspond to the object represented, as seen in Figure 5.3, which only shows two jets to keep it simple and clear. Graph Neural Networks [43], or "GNNs" are NNs that can handle graphs. They are able to use the graphs to understand the structure of the event, and habe the advantage of being able to handle inputs of variable length. This is especially useful while working with tracks, as the number of tracks is variable. A maximun number of tracks is set by defining this number of nodes, but if there are less tracks in the event, unnecessary nodes are pruned, i. e. removed from the input. That way the GNN can handle events with only two tracks, or as many as wanted, but reducing the complexity of the graph to the

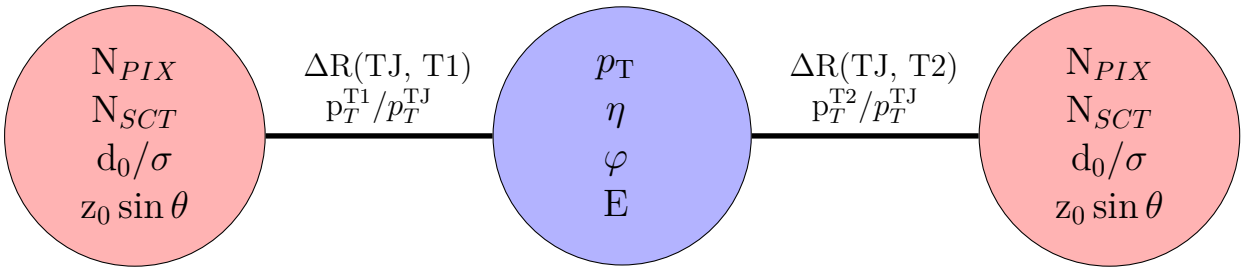

Figure 5.3.: Example of a filled Graph for an event with 2 tracks. An explanation of the observables is given in Table 6.1, with Track Jet being abbreviated as TJ, and Track 1/2 as T1/2. The Track Jet Variables are the four-vector of the Track Jet.

necessary level and without wasting computing power on padded values of non-existing tracks, as a DNN would have to.
There are different layouts of layers used in this analysis. The first one is GraphSAGE [44], which calculates the output

$$x_i' = W_1 x_i + W_2 \,\mathrm{mean}_{j \in \mathcal{N}(i)} x_j,$$

by combining the input, and the mean of the nodes around x_j, after multiplying a weight W with it. This is a relatively simple approach for a GNN layer. A more complex layer would be the Graph Convolutional Network (GCN) [45], given by

$$x_i' = \Theta^T \sum_{j \in \mathcal{N}(i) \cup i} \frac{e_{j,i}}{\sqrt{\hat{d}_j \hat{d}_i}} x_j,$$

with $\hat{d}_i = 1 + \sum_{j \in \mathcal{N}(i)} e_{j,i}$ where $e_{j,i}$ denotes the edge weight from source node j to target node i, for further explanation see Ref. [45]. Another layer used is the graph attentional operator (GAT), described by

$$x_i' = \sum_{j \in \mathcal{N}(i) \cup i} \alpha_{i,j} \Theta_t x_j,$$

as explained in detail in Ref. [46]. The last graph layer used is GraphConv [47], calculating the new values as

$$x_i' = W_1 x_i + W_2 \sum_{j \in \mathcal{N}(i)} e_{j,i} x_j,$$

with edge weights $e_{j,i}$ that connect the source node j with the target i.
For the final network, a hyperparameter optimisation is performed. This means that a number of different configurations of layers and network sizes are trained, and evaluated to choose the best possible architecture.

6. Prompt Lepton Tagging with Deep and Graph Neural Networks

PLIV was the primary tool for prompt lepton tagging in Run II of the LHC, relying on several dedicated observables related to lifetime and isolation of leptons. Since PLIV has been developed for Run II, an update for Run III has to be developed. For this update, a lot of potential approaches could be considered, starting from just using the established version and retrain it on new data, up to completely different architectures and types of neural networks. In this thesis, the approach of a DNN, and a GNN combined with a DNN, as such approaches are presented. The goal is an analysis of the potential of the presented approaches.
For the GNN approach, the GNN is used to analyse the tracks of the events, using the flexibility needed to incorporate the variable number of tracks in any given event. The DNN is trained on the other observables, which are mainly based around the lepton.

6.1. Sensitive Observables

The identification of scalar observables most sensitive to the electron origin is the starting point of any approach of lepton tagging. For this thesis, the observables used for the PLIV of Run II, as presented in Table 6.1 are reconsidered and analysed again. Depending on the type of the observable, they were used for either the GNN only or both NNs. The scalar observables are processed by DNNs in both cases, and therefore also called the global variables, while the vector-like observables are processed by the graph part of the GNN.

6.1.1. Global observables

The distributions of the observables that were most important in the training are shown in Figure 6.1, with the other observables being shown in Appendix A. For the most important observables the two dimensional separation is shown, to present the differences

Table 6.1.: Input variables for PLIV in Run II. jet_{track}^{lepton} refers to the closest track jet to the training lepton.

Input	Description	Type
ΔR(track, track jet)	The ΔR between ID track and jet_{track}^{lepton}	Vector
$p_T^{track}/p_T^{trackjet}$	p_{T} of ID track divided by p_{T} of jet_{track}^{lepton}	Vector
$z_0 \sin(\theta)$	longitudinal impact parameter scaled by $\sin(\theta)$	Vector
d_0/σ_{d_0}	transverse impact parameter divided by significance	Vector
N_{hit}^{PIX}	number of hits of the track in the pixel detector	Vector
N_{hit}^{SCT}	number of hits of the track in the SCT	Vector
$\max(l_{\mathrm{SV\ to\ PV}}^{longitudinal}/p_{\mathrm{T}})$	Max. SV longitudinal significance of the lepton	Scalar
$p_{\mathrm{T}}\mathrm{VarCone30}/p_{\mathrm{T}}$	Lep. isolation using ID tracks in cone $\Delta R < 0.3$	Scalar
$E_T\mathrm{TopoCone40}/p_{\mathrm{T}}$	Lep. isolation using topo. clusters in cone $\Delta R < 0.4$	Scalar
$\Sigma_{cluster}^{\Delta R<0.15} E_T/p_{\mathrm{T}}$	sum of cluster energy divided by lepton p_{T}	Scalar
N_{track} in track jet	Number of tracks clustered by the track jet	Scalar
p_T^{rel}	p_{T} along the track jet axis: p·sin(<lepton, track jet>)	Scalar
$p_T^{\mathrm{lepton\ track}}/p_T^{\mathrm{track\ jet}}$	lepton track p_{T} divided by track jet p_{T}	Scalar
ΔR(lepton, track jet)	ΔR between the lepton and the track jet axis	Scalar
p_T^{lepton} bin number	Index of the bin of lepton p_{T}	X
Track Jet ($p_{\mathrm{T}}, \eta, \phi$, E)	Four vector of the track Jet	Vector

in the distribution more clearly. For $E_T\mathrm{TopoCone40}/p_{\mathrm{T}}$ the distribution shows that charge flip and photon conversion electrons prefer values close to 0, and do not take values larger than 0.15. Meanwhile prompt and the light and heavy flavour decay electrons have a clear peak at 0, but have a broad distribution, especially seen in the relatively large last bin consisting of the overflow with values larger than 1. The separation plot supports this, showing a clear peak at 0 for the prompt and charge flip, and a more broad distribution for the light and heavy flavour decays. For the ΔR(lepton,track jet) it is similar, but for values larger than 0.2 almost only prompt leptons remain. But also here the separation shows a clear peak for prompt, photon conversion and charge flip electrons at 0, with the LF and HF decay electrons showing a more broad spectrum with a significantly lower peak at 0.

For the number of tracks in the track jets, most prompt leptons have no tracks in the track jet, because a track jet can only be constructed with at least 2 tracks. For prompt leptons, one track is expected, with multiple tracks being a possibility, for example if the track is distorted in any way. In the separation a clear preference for the LF and HF decays for more tracks can also be observed, while the other processes once again prefer the lower values. The impact parameters for prompt leptons are also expected to be small, as they are coming from the primary vertex, with the division of the significance, the shape is expected. This is also observable in the separation, with HF decay and charge

flip electrons having a very broad distribution, which only shows a slight peak at 0 for the heavy flavour decays, while charge flip electrons are the only process which has a low point of the distribution at 0.
These observables already show significant differences in the processes, which can then be combined and exploited by the NNs.

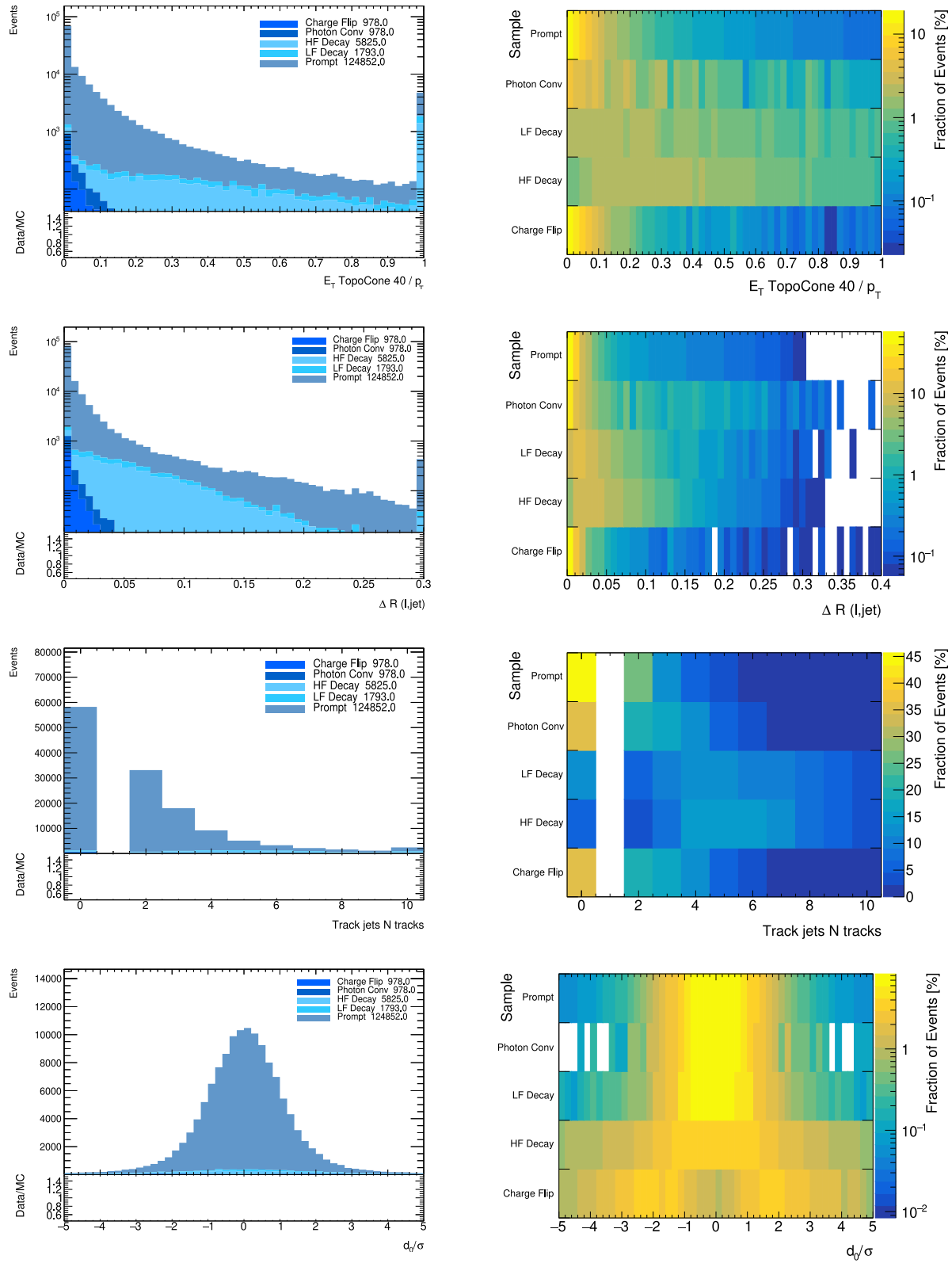

Figure 6.1.: Distributions and separation of the most important global observables.

6.1.2. Track Observables

For the GNN the several vector-like observables can be used, with most of them directly representing the additional track information. The other observable that is added is the maximum secondary vertex longitudinal significance of the electron, which describes the distance between the primary and a possible secondary vertex. The distribution of this observable is shown in Figure 6.2. It can be seen that the shape is asymmetric for all processes, with positive values being strongly preferred, and the peak of the distribution is at zero for most processes, even though there is one exception: For the heavy flavour fakes, the distribution is noticeably broader, and the peak is shifted from zero. This shows the expected behaviour of heavy flavour fakes, with them having a secondary vertex that is significantly displaced from the primary vertex, as the heavy quarks have a considerable lifetime. Therefore this observable will be very important to tagging prompt leptons in contrast to leptons originating from b- or c-jets.
One important decision to be made is the number of tracks to include into the GNN. In Figure 6.3 the length of the vectors of the pixel hit observable is shown. From this and the distributions for the other similar observables it was decided to move forward with including up to 6 tracks. This includes most of the events ($\approx 95\%$), while keeping the number of unnecessary nodes low. From the ratio it can be seen that there is a slight tendency for non-prompt processes to prefer a lower number of tracks. This would not be expected from LF- or HF-fakes, but for the charge flip and photon conversion fakes. The reason for this to not be compensated by the LF- and HF-fakes lies in the selection in Table 6.2, as most other tracks do not meet the requirements made.
Coming back to the track observables themselves, the normalised distributions of the number of pixel and SCT hits are shown in Figure 6.4, for each of the five processes. For tracks to be considered for the lepton tagging, they have to fulfil the selection criteria listed in Table 6.2. Looking at the distributions themselves, there are only slight fluctuations between the different sources. The ratios also stay very close to one, with the exception of some bins with very low bin content.
The shape of the distributions seem to be because of the applied selection, which requires a total of at least seven hits in the silicon detectors, all tracks with less than that are not considered. The peak of the Pixel hit distributions is at seven, with the peak of the SCT hit distributions at zero. All tracks with less than seven pixel hits have to have at least the difference as SCT hits. A track with six pixel hits, has at least one SCT hit, but could also have two. This explains the slight difference between the exact shapes of the distributions. For tracks with five pixel hits, two SCT hits are required. The absence of tracks with less than five hits is explained by the fact that there are no tracks with more

Table 6.2.: Requirements for the tracks to be considered for the lepton tagging

p_{T}[MeV]	$\lvert\eta\rvert$	$\lvert z_0 \sin\theta\rvert$ [mm]	ΔR(lepton, track)	N_{Si}	N_{Si}^{shared}	N_{Si}^{holes}	N_{pix}^{holes}
> 500	< 2.5	< 1	< 0.4 and $> 10^{-6}$	≥ 7	≤ 1	≤ 2	≤ 1

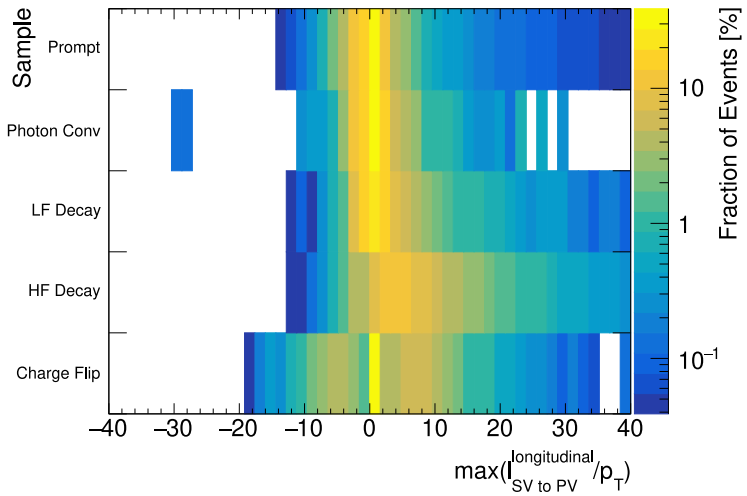

Figure 6.2.: Separation of the maximum secondary vertex longitudinal significance.

than two SCT hits.

For the impact parameters shown in Figure 6.5, the separation is also low. Both observables have a clear peak around zero for all processes. The ratios stay close to one near zero, with fluctuations only happening at the tails of the distributions, where small statistical fluctuations can impact the ratio. These should not be used to create a separation between the processes.

The last two observables are composed of track jet and track information, with the ΔR showing the angular distance, and the p_T^{rel} the relative transverse momentum of the track and track jet. Their distributions vary strongly for the different electron origins, as seen in Figure 6.6. For the ΔR distributions the prompt electrons have the highest peak, while light flavour and charge flip fakes show a less clear peak, and in the tails of their distribution there are larger. The heavy flavour and photon conversion fakes stay closer to the shape of the prompt distribution, but also show a slightly lower peak and a broader tail. Outside of the peak between two and four, the non-prompt electron density for every process is higher than the prompt one. For the relative p_{T} a similar effect can be observed. While in the first bin, the prompt leptons have the highest content, they drop to third most in the second bin, and from the third bin they have the lowest density. These two observables seem to offer the most separation power of the track observables.

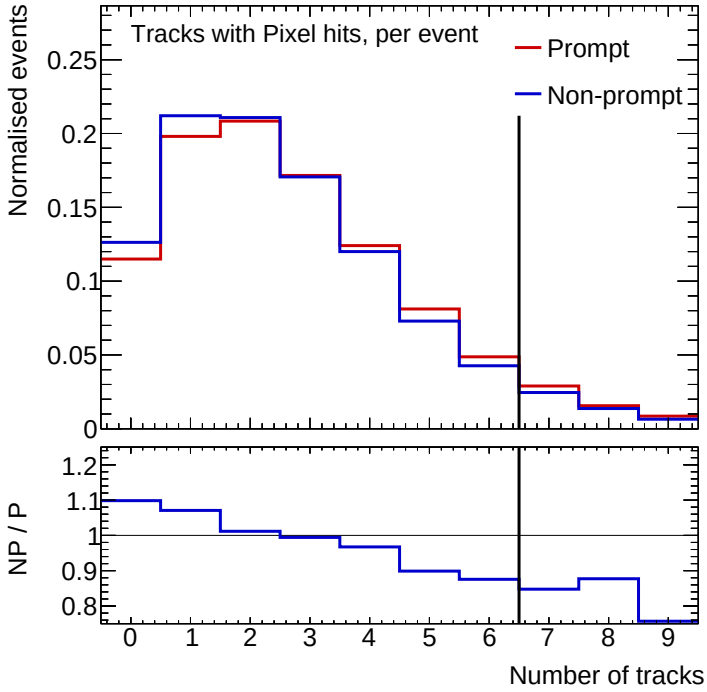

Figure 6.3.: Normalised distribution of the number of tracks with hits in the pixel detector per event for prompt and non-prompt electrons. The lower plot shows the ratio between non-prompt and prompt electrons per bin. The vertical line shows the limit of the six included tracks.

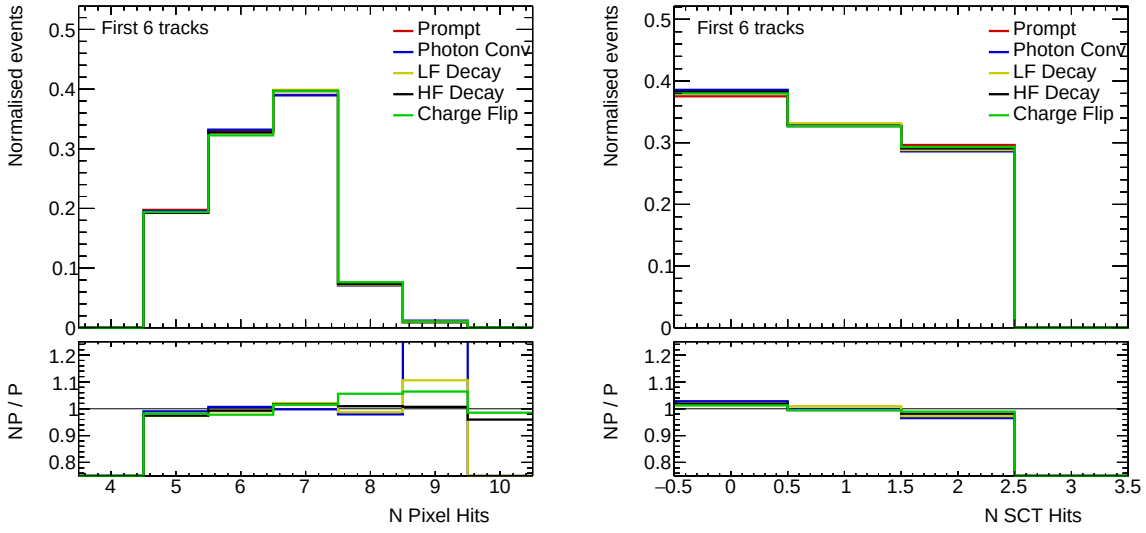

Figure 6.4.: Distributions of the number of pixel and SCT hits of the first 6 tracks.

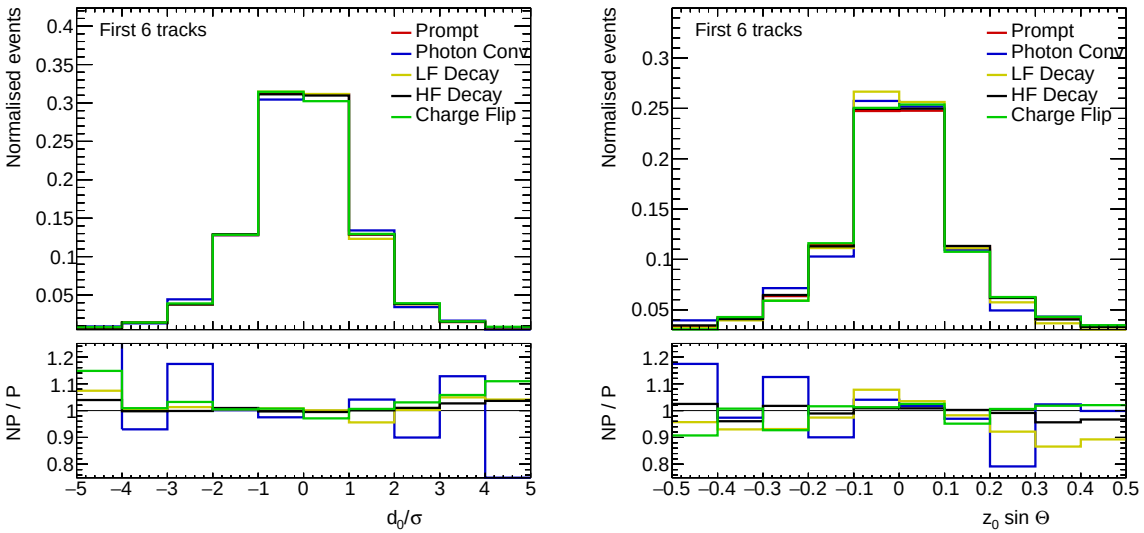

Figure 6.5.: Distributions of the impact parameters of the first 6 tracks.

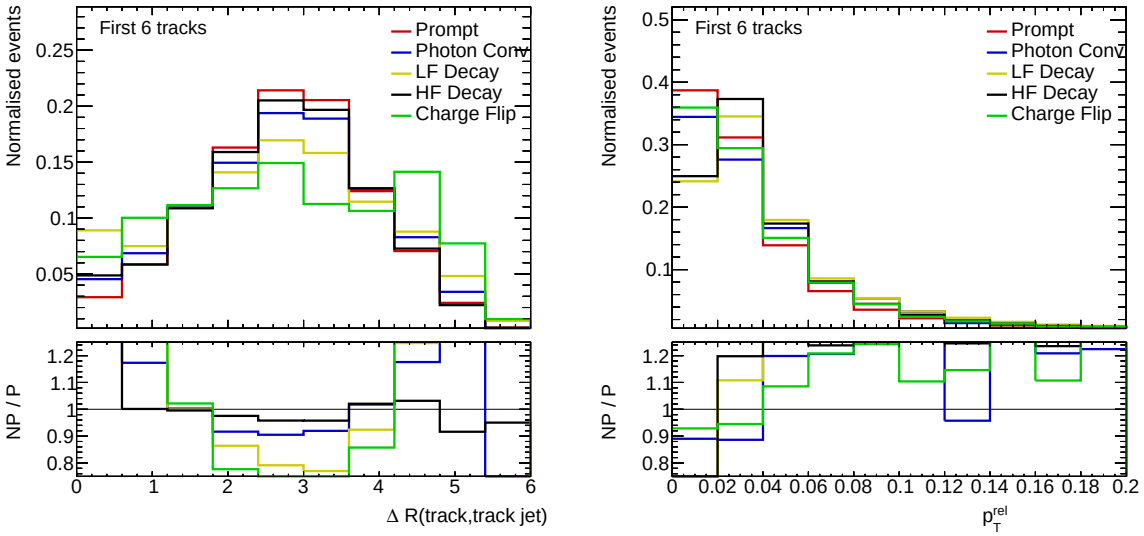

Figure 6.6.: Distributions of the ΔR and relative p_{T} between the track jet and the first 6 tracks.

6.2. Model Descriptions

For the training of the models the architecture and datasets have to be defined. For the DNN, the most important architecture part is the number of layers and nodes, as for all NNs. For the GNN, other important features come into play, as the layers can be very different in computational approach, and even which attributes are allowed for the nodes. Some graph layers allow for edge connections, some do not, and both are used in the model. Therefore the chosen architecture is important, especially for the GNN.

6.2.1. DNN

The base approach for this thesis is a DNN, as they work quite stable and reliable if their training dataset is large enough. Therefore, it makes sense to use them as a baseline for other possible approaches. Therefore, a DNN with just 2 layers with 30 nodes each is trained, the parameters can be seen in Table 6.3. Additionally, models with 2 layers of 50, 3 layers of 30, and 3 layers of 100, 6 layers of 30 and 2 layers of 20 nodes were trained. The observables used in the training can be seen in Table 6.4. Additionally, some similar variables such as E_TTopoCone20 and E_TTopoCone40 as well as $\Sigma_{cluster}^{\Delta R<0.15} E_T$ were provided, which are the base versions of some PLIV observables, to see whether they offer additional separation power without the normalisation to the $p_{\mathrm{T}}$of the electron.
For the initial training a MC set of $\approx 135,000$ electrons is used. Of those electrons, 92.88 % are classified as prompt. The second largest contribution are heay flavour decays with 4.33 %, and then light flavour deacys with 1.33%. The rest is made up of photon conversion and charge flip fakes. The underlying samples are $t\bar{t}$ non-all hadronic samples, and therefore all contain at least one lepton. The events were generated using the PowHegBox-2 [48] at NLO with the NNPDF3.0NLO PDF set [49] and the damping factor h_{damp} set to 1.5 times the top mass. The events are interfaced to Pythia 8.308 [50] using the A14 [51] tune and a NNPDF2.3LO [49] PDF set.
The reason for some inputs missing in the DNN is mainly that these observables were either not available in the derivation at that time, such as the secondary vertex information necessary for $\max(l_{\mathrm{SV\ to\ PV}}^{longitudinal}/p_{\mathrm{T}})$, which only became available during the later part of this analysis, or that they are not easily includable in the DNN, such as the number of hits in the pixel and semiconductor tracker, as they are of variable length, which is not possible for a DNN input ad hoc.
An approach including the track observables is presented additionally to the DNN, where the 6 variables per track were just added into the DNN, applying "padding". Padding means filling up empty variables with a default value, in this case -99. This makes it pos-

Table 6.3.: Overview of the main DNN features.

Feature	Value
Nodes	30,30
Loss	Binary_Crossentropy
Optimiser	Nadam
Epochs	10000
Learning rate	10^{-3}
Dropout	30%

Table 6.4.: Input variables reused for the Neural Networks. jet_{track}^{lepton} refers to the closest track jet to the training lepton. The GNN edges are drawn between the track jet node and the track nodes.

Input	Included in
ΔR(track, track jet)	GNN-Edge
$p_T^{track}/p_T^{trackjet}$	GNN-Edge
$z_0 \sin(\theta)$	GNN-Node
d_0/σ_{d_0}	GNN-Node
N_{hit}^{PIX}	GNN-Node
N_{hit}^{SCT}	GNN-Node
$\max(l_{\text{SV to PV}}^{longitudinal}/p_{\mathrm{T}})$	GNN-Global
p_{T}VarCone30/p_{T}	DNN & GNN-Global
E_TTopoCone40/p_{T}	DNN & GNN-Global
$\Sigma_{cluster}^{\Delta R<0.15} E_T/p_{\mathrm{T}}$	DNN & GNN-Global
N_{track} in track jet	DNN & GNN-Global
p_T^{rel}	DNN & GNN-Globa
$p_T^{\text{lepton track}}/p_T^{\text{track jet}}$	DNN & GNN-Global
ΔR(lepton, track jet)	DNN & GNN-Global
Track Jet ($p_{\mathrm{T}}, \eta, \phi$, E)	GNN-Node

sible for the DNN to process inputs that are smaller than expected, as the cost of adding inputs that do not have any information. Therefore the addition of the track observables could be beneficial or disadvantageous.

6.2.2. GNN

As the baseline DNN approach is limited by the inherent features and limitations of DNNs, a second approach in the form of a GNN is presented. The GNN offers the flexibility of variable length for the variables. This opens up a more promising way to include the track observables. There is still a hard limit on the maximum number of tracks included, with a cutoff defined at a maximum of 6 tracks, as explained earlier. If there are less tracks in the event, pruning is applied. This is done to keep the computations as simple as possible by pruning away all nodes that are not filled. This could be either due to them not existing at all, or just incomplete reconstruction. The presented approach also combines the already shown DNN approach with the GNN, as the final model is a combination of both DNN and GNN features.
For the GNN, the full list of observables listed in Table 6.4 is included, with the track variables either being assigned to the nodes (Impact Parameters, ID hits) or to the edges between the track jet as the central node (ΔR, relative p_{T}). All other variables are assigned as global variables, being processed by a DNN.
The training set has been increased for the GNN, with a total of 523,602 electrons. The largest fraction by far is the prompt electrons, making up $\approx 90\%$, with the second largest fraction heavy flavour decays only making up $\approx 7.5\%$. Light flavour decays only make up $\approx 1.4\%$, while photon conversion and charge flip contribute less than one percent each. The underlying events are still non-all hadronic $t\bar{t}$ decays, generated as described before.
The architecture of the GNN was determined via hyperparameter optimisation with 250 different models. The chosen model is shown in Table 6.5, with the relevant functions having been explained in Section 5.2. The Graph part consists of six graph layers, with four different kinds of layers being used, and a different number of Graph channels for each. Different activation functions are used after each Graph layer.
The DNN that is handling the global inputs has two layers with 95 and 70 nodes, respectively. The layers have different activation functions in ReLU and sigmoid. The outputs of the Graph Network and the Global Deep Network are combined by two final layers with 20 and 30 nodes, activated by ELU and ReLU. The final output activation is then once again the sigmoid function, making sure the output is between 0 and 1.

Table 6.5.: Architecture of the GNN

Feature	Value
Graph Layers	GCN, SAGE, GAT, Graph, GCN, GCN
Channels	20,90,60,25,55,60
Global Nodes	95,70
Final Nodes	20,30
Activation Graph	ReLU, tanh, ReLU,tanh, ELU, ReLU
Activation Global	ReLU, sigmoid
Activation Final	ELU,ReLU
Activation Output	sigmoid
Loss	Binary Cross-entropy
Optimiser	Nadam
Epochs	10000
Learning rate	0.009
Dropout	20 %

6.3. Training

The training and the validation of the models are important steps for every analysis using neural networks. The networks have to be checked for over- and underfitting, and the relevance of the observables has to be discussed as well.

6.3.1. DNN

The DNN was trained in two folds, which show no significant differences between them, as seen in the loss curves in Figure 6.7. Both converge relatively fast, and do not show signs of over- or underfitting. Therefore it can be assumed that the model is working as intended.
To asses the relevance of the observables, the importance and the correlation between the observables are considered. The permutation importance in Figure 6.8 shows that the most important observable is the number of tracks per track jet. This is probably based on the expectation that prompt leptons will leave less tracks than non-prompt leptons inside jets, that also produce tracks. The next most important observable is the impact parameter, which measures the distance of the object from the primary vertex, which is expected to be large for b-quarks, the main source of heavy flavour fakes. The third most important is the E_TTopoCone30/p_{T} variable, which is an isolation variable divided by the lepton p_{T}. Almost equally important was the ΔR between lepton and track jet, so the angular distance between lepton and jet, another isolation variable.
On the left of Figure 6.8 the correlation between the variables is shown. This matrix is reduced to the observables with considerable correlations, as non correlated variables do not need close monitoring. From the matrix, it can be seen that the p_{T}VarCone30/p_{T} and p_{T}VarCone20 have a correlation of 1, meaning that one of them can be removed, as they do not hold additional information. Other highly correlated observables are E_TTopoCone30/p_{T} and E_TTopoCone30. Both of these pairs are expected to have high correlation, as one of them is just the other divided by p_{T}. Observables that have non-obvious correlation are the ΔR between lepton and track jet, with the number of tracks in the track jet and the electron track p_{T} divided by the track jet p_{T}.

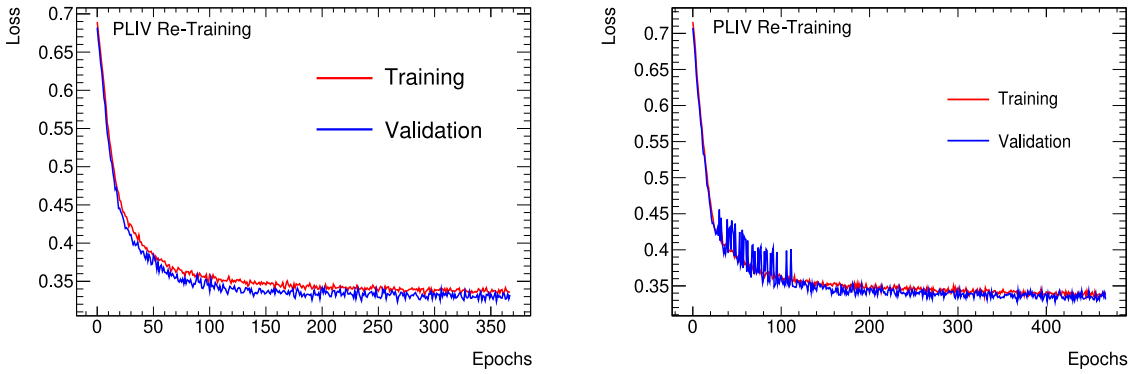

Figure 6.7.: Loss for the two folds of the training, with Fold 0 on the left, and Fold 1 on the right.

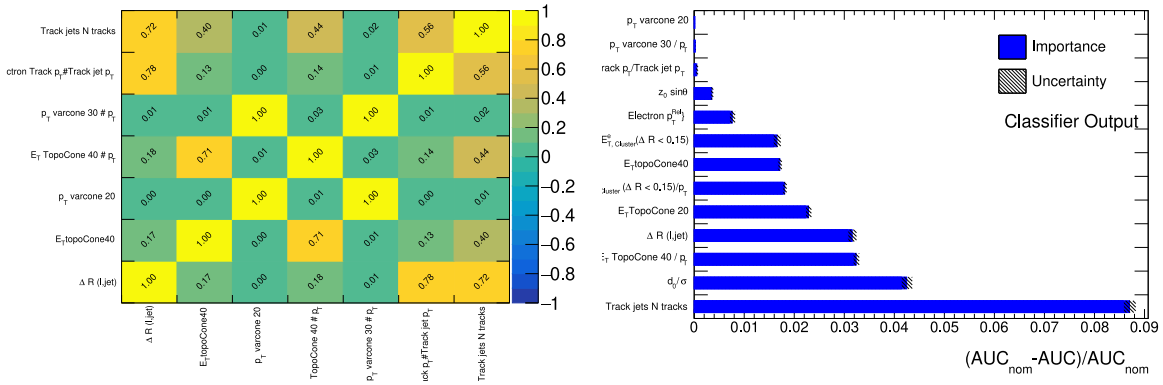

Figure 6.8.: On the left, the correlation matrix of the observables with considerable correlations, leaving out observables with low correlation. On the right, the relative importance of the input variables.

6.3.2. GNN

To asses the validity of the GNN models, the loss curves of the models are shown in Figure 6.9. The loss for both folds shows that the models converge, and a plateau in the loss of the validation is reached. The loss of the training set is still decreasing, but since the validation loss stays stable, there is no overfitting. Overall, the loss is behaving as desired, and no signs of unwanted behaviour can be seen. It can therefore be concluded that the model is valid and can reasonably be used for lepton tagging.
For the validation of the sensitive observables, the permutation importance of the inputs is considered again, which can be seen in Figure 6.10. The p_TVarCone30/ p_{T} observable has the highest importance with the p_T^{rel} being second most important. For the track observables the impact parameter d_0/σ of several tracks is most important. The energy and p_{T} of the track jet also show their importance. From the permutation importance it can be seen that no observable is completely dominating the decision making, and that the global network and the graph network observables contribute to the output score significantly.

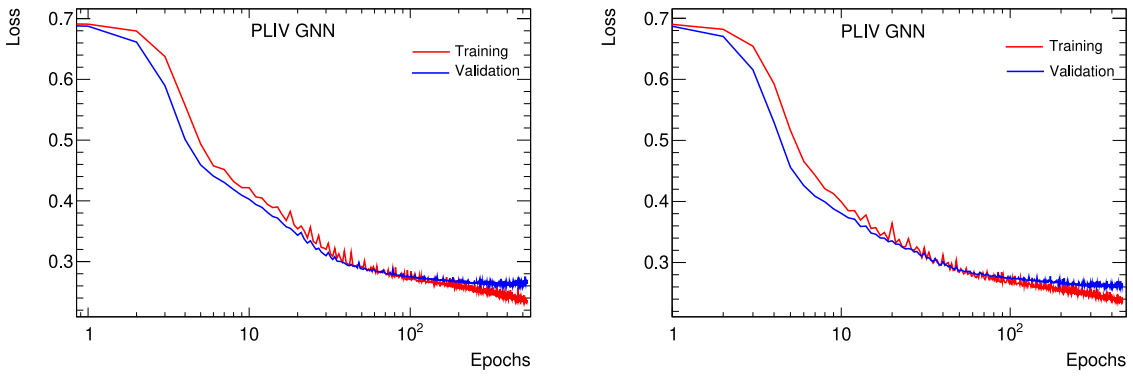

Figure 6.9.: Loss for the two folds of the training, with Fold 0 on the left, and Fold 1 on the right.

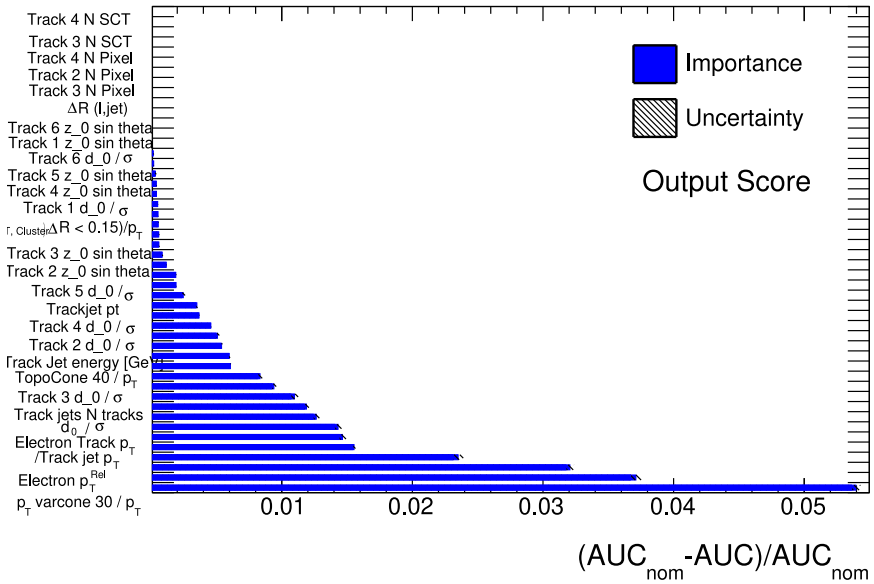

Figure 6.10.: Permutation importance of the inputs. The p_TVarCone30/p_{T} observable has the highest importance. The second most import was the relative p_{T} of the electron.

6.4. Performance Comparison

The base model trained for this study was the DNN with 2 layers of 30 nodes each. The distribution of the DNN output score is shown in Figure 6.11. It already shows that the network is generally able to separate the prompt and non-prompt leptons, as seen in the clear peaks of the non-prompt leptons at a low score and the prompt leptons with a high score. Notably, in the interval between a score of 0.6 and 0.9, there are basically no fake electrons, but they reappear at scores slightly higher than 0.9, together with the prompt lepton peak.
To asses where the DNN is working well, the separation is considered. The separation in one dimension, so just differentiating between prompt and non-prompt leptons, can be seen on the left side of Figure 6.12, while the right side shows the two dimensional separation between the different kinds of fake leptons. The one-dimensional separation shows that the fraction of prompt events is low for low scores, and becomes the larger fraction at a score of 0.55, with a clear peak at 0.9, with the non-prompt leptons showing the opposite behaviour, similar to Figure 6.11. The total separation is 61.04%. For the two dimensional separation clear peaks are visible for the prompt leptons at around 0.9, while the LF- and HF-fakes peak near zero. For the photon conversion and charge flip fakes the distribution is not as clear, but relatively broad and almost uniform. This shows the DNN works best for the prompt leptons and LF/HF fakes, as it is tagging them correctly, but worse for photon conversion and charge flip fakes who are not clearly identified as fakes.
The confusion matrix in Figure 6.13 shows the percentage of prompt and non-prompt leptons being assigned correctly. Notably signal events are correctly classified as signal about 90 % of the time, while non-prompt leptons are only classified as such in 84 % of cases. This shows there is especially room for improvement for the correct tagging of non-prompt leptons.
To compare the different sized DNNs, their ROC curves are compared in Figure 6.14. It can be seen that there are only very small differences between all of the models. Therefore it has to be concluded that increasing the size of the DNN is not able to yield better results, even a slightly smaller model reaches the same result. This could mean that the available observables do not offer more information than already used.
To evaluate the inclusion of the track observables in the DNN, the ROC Curves of the DNN without, and with 3, 6 or 9 tracks are shown in Figure 6.15. It can be seen that the network performs worse with added variables, which is obviously unwanted behaviour, as the network should at least be performing similar, if the added information offers no additional separation power. Instead, the added information affects the prediction

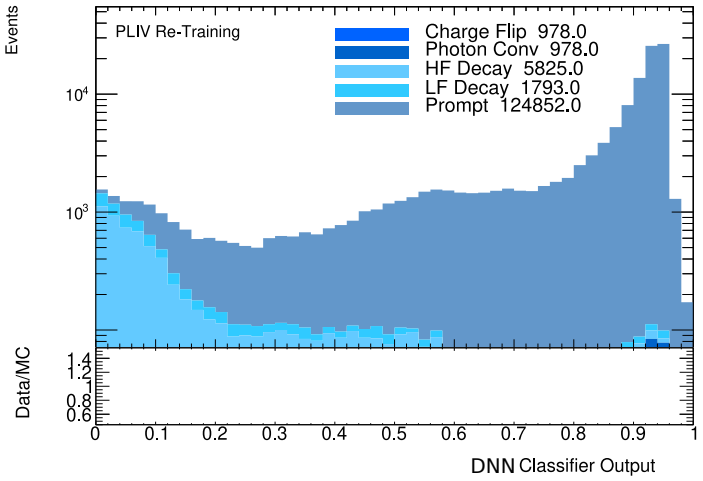

Figure 6.11.: Distribution of the DNN output score.

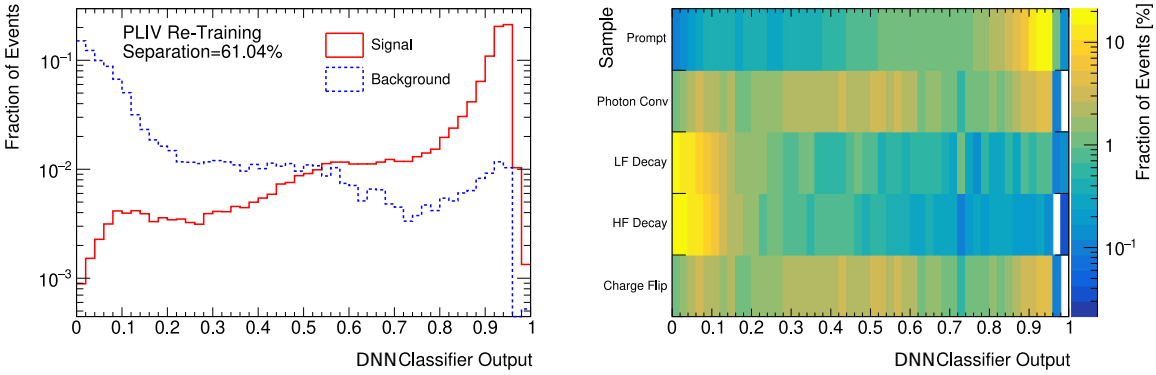

Figure 6.12.: Separation of the prompt and non-promt leptons in 1D and 2D.

power negatively. This could be either due to the padding leading to unwanted results, as the network gets confused by the unfilled tracks, or due to the massive increase in size and complexity in the network, as 18, 36 or 54 input variables are added. Due to this behaviour, the track variable inclusion is presented separately from the initial DNN results. It can only be concluded that the DNN is not suitable for the inclusion of the track observables.

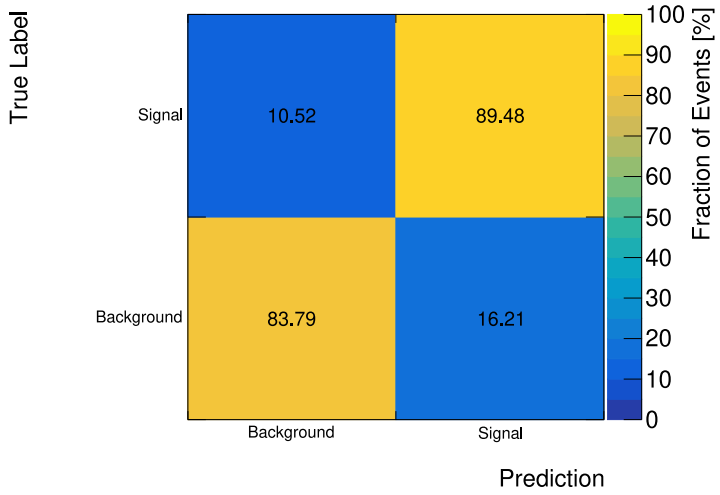

Figure 6.13.: The confusion matrix of the DNN, showing how the DNN classifies certain events, with signal referring to prompt, and background to non-prompt electrons.

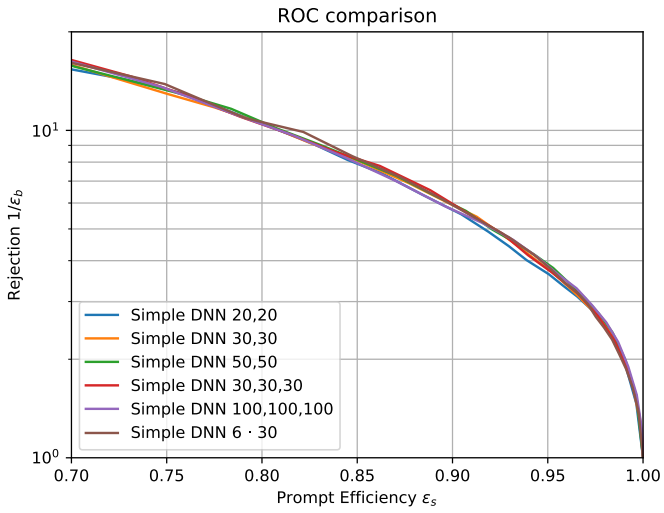

Figure 6.14.: ROC curve for the different DNN models.

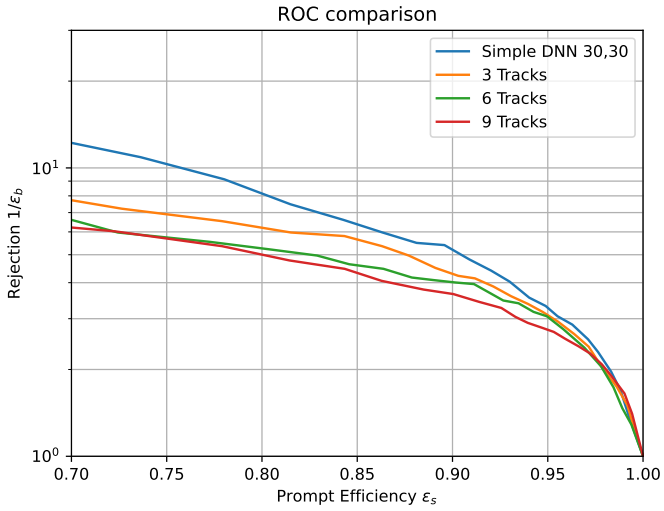

Figure 6.15.: ROC Curve for the DNN with added Track Variables for 3, 6 and 9 Tracks.

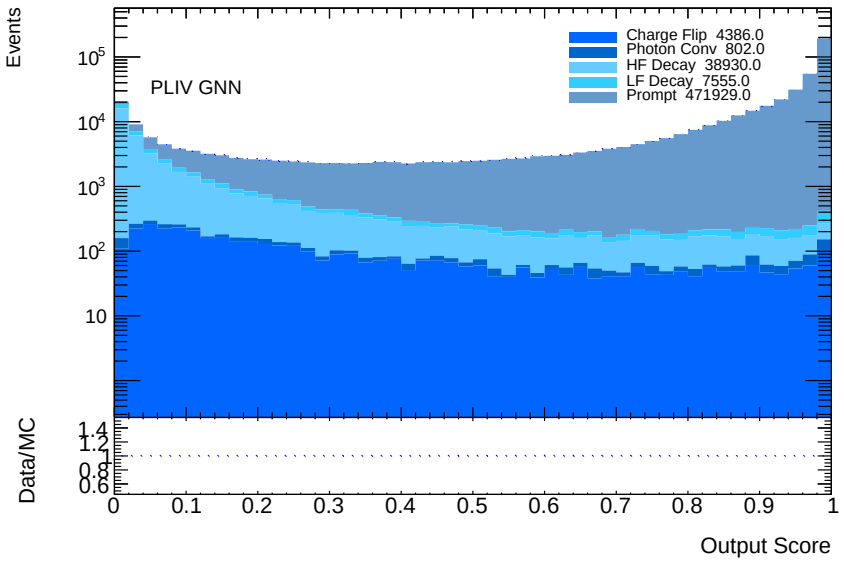

Figure 6.16.: Distribution of the GNN Score.

For the evaluation of the GNN, the first thing to look at is the output of the GNN, which is shown in Figure 6.16. From there, several things can already be seen. The distributions of charge flip and photon conversion electrons are pretty uniform, meaning the GNN is not good at classifying these electrons. For the light and heavy flavour fakes and the prompt leptons the distributions show clear preferences for the desired values. The fake distribution here has a clear peak at zero, and gets smaller towards one. On the other hand, the score of the prompt leptons has a peak at one, and at 0 there are almost no prompt electrons. This is supported by looking at the two-dimensional separation seen in Figure 6.17, which makes the distributions even more clear. This shows the GNN is good at separating prompt electrons from light and heavy flavour fakes. The peaks at the desired values of those three processes are very clear in this representation. For photon conversion and charge flip fakes the distributions do not have one clear peak each, but instead pretty uniform distributions. For photon conversion, there seem to be two slight peaks at both ends of the spectrum.

The confusion matrix in Figure 6.18 shows an accuracy close to 90 % for assigning the correct label to a given electron. For the prompt electrons, there is only a slight improvement in contrast to the DNN, but for non-prompt electrons the correct tagging percentage rises to close to 90 % as well. This already shows that the GNN is performing better than the DNN.

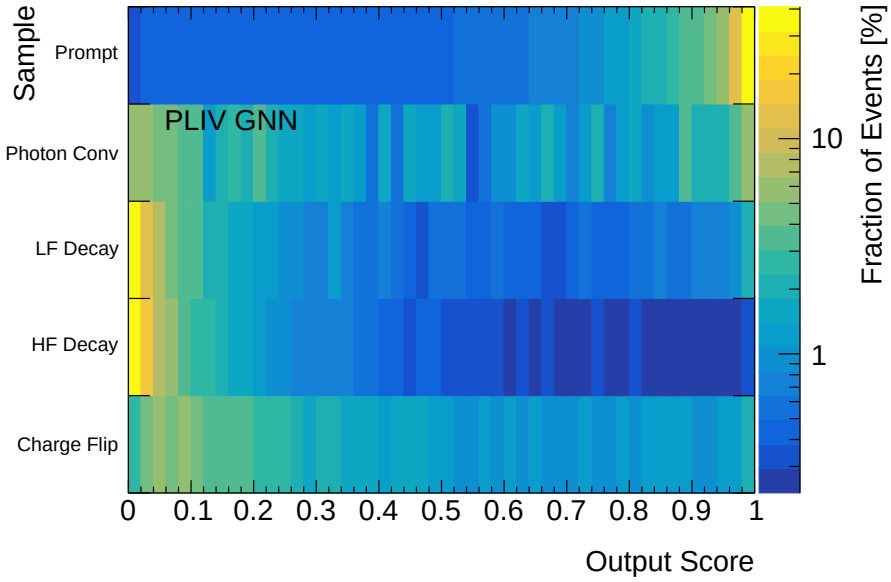

Figure 6.17.: 2D separation of the output score.

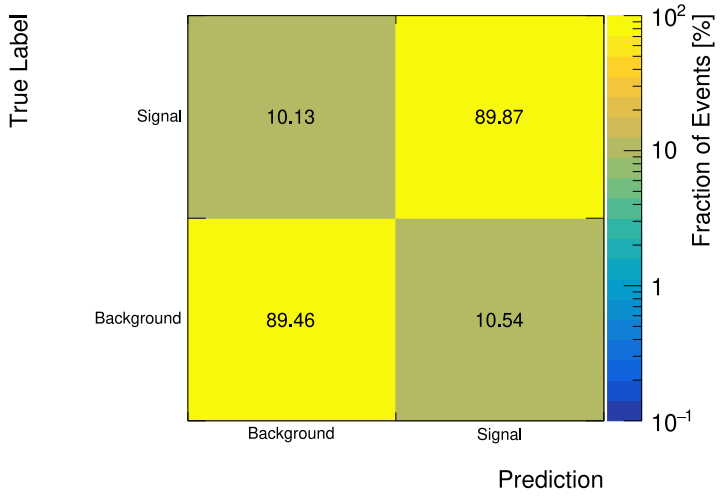

Figure 6.18.: Confusion matrix for the GNN, with signal corresponding to prompt, and background refering to non-prompt electrons.

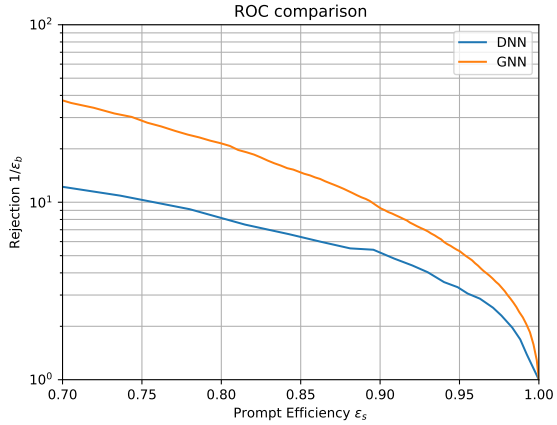

Figure 6.19.: ROC curve comparison between the DNN and the GNN.

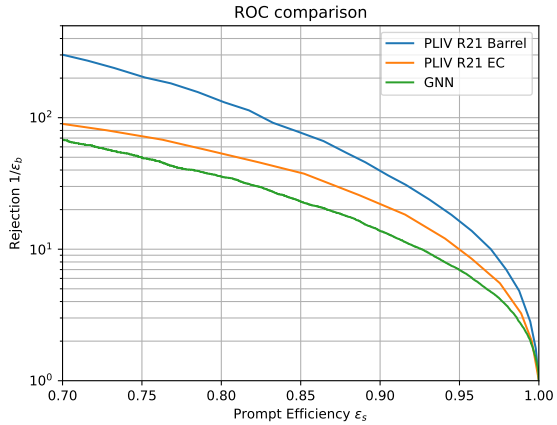

Figure 6.20.: ROC curve comparing the GNN and PLIV. This is only evaluated on prompt leptons and heavy flavour fakes to create an even comparison with PLIV.

To complete the comparison of the DNN and the GNN, their ROC curves can be seen in Figure 6.19. It can be seen that the GNN is outperforming the DNN in background rejection for any prompt efficiency. At a prompt efficiency of 70 %, the GNNs rejection is higher by a factor of 2.5 to 3. As the global observables are available to both approaches, and both process them via a DNN, the difference has to lie in the track observables.
Finally, since the base of the entire study is PLIV, it should be compared to it. This is done in Figure 6.20, comparing the GNN to PLIV. To make a fair comparison, the evaluation here is done on only prompt and heavy flavour fake electrons, as this was the basis for the evaluation on PLIV. It can be seen that the GNN is able to get close to the rejection the end-cap version of PLIV could reach. The barrel version of PLIV is still performing better by a factor of 4 at 70% prompt efficiency.

7. Conclusion and Outlook

The approaches to prompt lepton tagging presented in this thesis show that alternate methods of lepton tagging can work as well. The DNN approach shows that a lot can already be gained from the global observables alone, but also shows how the DNN struggles with the variable number of tracks. The comparison between DNN and GNN proves that it is important to include the available track information, and that neural networks that can work with variable input lengths have an advantage over less flexible architectures. The addition of the track observables and secondary vertex information for the GNN has increased the performance significantly. If evaluated on the same basis as PLIV, the performance already is in the right order of magnitude. On the other hand this approach is still significantly worse than the PLIV, despite the inclusion of the same observables.
As for reasons the GNN is not able to reach the same accuracy as PLIV, the first thing to be mentioned should be that the training sample for the GNN was significantly smaller than the samples that PLIV was trained on. An important thing to note is that both networks presented were trained only on $t\bar{t}$ samples. The inclusion of other samples, such as Z+jets, should definitely be investigated. Additional processes might give additional insights to which observables are important to lepton tagging. Additionally, another possible reason for the difference could be the non-existing separation of barrel and end-cap electrons. For PLIV, two completely separate models are used for the tagging of those, while the presented approaches both do not make this differentiation. It would make sense to investigate whether there are significant differences between barrel and end-cap electrons, to see if it makes sense to keep this split.
With new approaches to lepton tagging in Run III already on the way, an approach similar to the one presented is already being used. The tagging networks GN1 and GN2 [52] are GNNs trained to be used for tagging of several objects, especially jets (b- and c-tagging). Currently GN1 is used for jet flavour tagging, beating the standard b tagging algorithms such as DL1r [53] by a factor of at least 2 in the rejection of light- and c-jets, it is discussed to use this tool for prompt lepton tagging. If these networks shows similar performance in lepton tagging as in jet tagging, they will certainly be considered as a valid and powerful tool for lepton tagging.

A. Observable Distributions

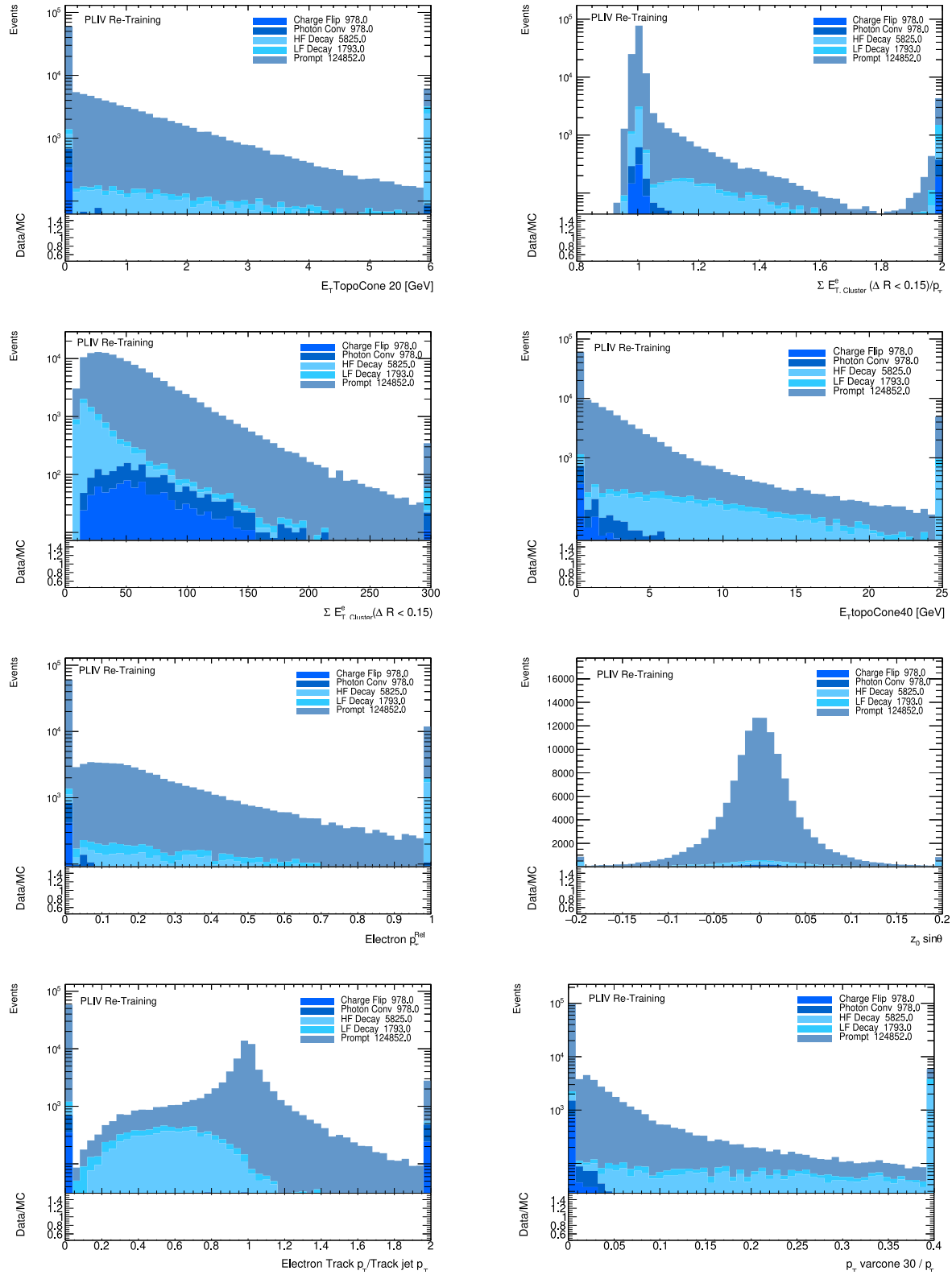

Figure A.1.: Distributions of the other observables used in training.

Bibliography

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716**, 1 (2012)

[2] CMS Collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, Phys. Lett. B **716**, 30 (2012)

[3] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13**, 508 (1964)

[4] W. J. Marciano, H. Pagels, *Quantum Chromodynamics: A Review*, Phys. Rept. **36**, 137 (1978)

[5] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[6] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[7] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519**, 367 (1968)

[8] M. Planck, *On the Law of Distribution of Energy in the Normal Spectrum*, Annalen Phys. **4**, 553 (1901)

[9] A. Einstein, *Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt*, Annalen Phys. **322(6)**, 132 (1905)

[10] D. P. Barber, et al., *Discovery of Three Jet Events and a Test of Quantum Chromodynamics at PETRA Energies*, Phys. Rev. Lett. **43**, 830 (1979)

[11] UA1 Collaboration, *Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at* $\sqrt{s} = 540$ *GeV*, Phys. Lett. B **122**, 103 (1983)

[12] UA1 Collaboration, *Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider*, Phys. Lett. B **126**, 398 (1983)

[13] ATLAS Collaboration, *Measurement of the W-boson mass in pp collisions at* $\sqrt{s}$ = 7 TeV *with the ATLAS detector*, Eur. Phys. J. C **78(2)** (2018)

[14] Arnaudon et al., *Measurement of the mass of the Z boson and the energy calibration of LEP*, Phys. Lett. B **307** (1993)

[15] ATLAS Collaboration, *Measurement of the Higgs boson mass in the H→ZZ→ 4 ℓ and H→ γγ channels with* $\sqrt{s}$ = 13 TeV *pp collisions using the ATLAS detector*, Phys. Lett. B **784**, 345 (2018)

[16] DØ Collaboration, *Observation of the Top Quark*, Phys. Rev. Lett. **74**, 2632 (1995)

[17] CDF Collaboration, *Observation of top quark production in* $\bar{p}p$ *collisions*, Phys. Rev. Lett. **74**, 2626 (1995)

[18] ATLAS Collaboration, *Measurement of the top quark charge in pp collisions at* $\sqrt{s}$ = *7 TeV with the ATLAS detector*, JHEP **11**, 031 (2013)

[19] A. Castro (ATLAS, CMS), *Top Quark Mass Measurements in ATLAS and CMS*, in *12th International Workshop on Top Quark Physics* (2019)

[20] M. Negrini (ATLAS, CMS), *Recent measurements of the top-quark mass and Yukawa coupling using the ATLAS and CMS detectors at the LHC*, PoS **EPS-HEP2021**, 479 (2022)

[21] G. C. Branco, et al., *Theory and phenomenology of two-Higgs-doublet models*, Phys. Rept. **516**, 1 (2012)

[22] Bigi et al., *Production and decay properties of ultra-heavy quarks*, Phys. Lett. B **181**, 157 (1986)

[23] Particle Data Group, *Review of Particle Physics*, PTEP **2022**, 083C01 (2022)

[24] *Top cross section summary plots - April 2024*, Technical report, CERN, Geneva (2024)

[25] ATLAS Collaboration, *Analysis of* $t\bar{t}H$ *and* $t\bar{t}W$ *production in multilepton final states with the ATLAS detector* (2019)

[26] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3**, S08003 (2008)

[27] L. Evans, P. Bryant, *LHC Machine*, JINST **3(08)**, S08001 (2008)

[28] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, JINST **3**, S08002 (2008)

[29] CMS Collaboration, *The CMS Experiment at the CERN LHC*, JINST **3**, S08004 (2008)

[30] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3**, S08005 (2008)

[31] G. Gorfine (ATLAS), *Alignment of the ATLAS Inner Detector Tracking System*, in *Meeting of the Division of Particles and Fields of the American Physical Society (DPF 2009)* (2009)

[32] T. Cornelissen, et al., *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)* (2007)

[33] ATLAS Collaboration , *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data*, JINST **14(12)**, P12006 (2019)

[34] ATLAS Collaboration , *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at* $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **81(7)**, 578 (2021)

[35] M. Cacciari, G. P. Salam, G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04**, 063 (2008)

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

[37] P. Billoir, S. Qian, *Fast vertex fitting with a local parametrization of tracks*, Nucl. Instrum. Meth. A **311**, 139 (1992)

[38] F. Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, Report: Cornell Aeronautical Laboratory, Cornell Aeronautical Laboratory (1957)

[39] A. Paszke, et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (2019)

[40] T. Dozat, *Incorporating Nesterov Momentum into Adam*, in *Proceedings of the 4th International Conference on Learning Representations*

[41] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization* (2017)

[42] Y. Nesterov, *A method of solving a convex programming problem with convergence rate 0(1/k2)*, Dokl. Akad. Nauk SSSR 269 (1983)

[43] F. Scarselli, et al., *The Graph Neural Network Model*, IEEE Transactions on Neural Networks **20(1)** (2009)

[44] W. L. Hamilton, R. Ying, J. Leskovec, *Inductive Representation Learning on Large Graphs* (2018)

[45] T. N. Kipf, M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks* (2017)

[46] P. Veličković, *Graph Attention Networks* (2018)

[47] C. Morris, et al., *Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks* (2021)

[48] S. Frixione, G. Ridolfi, P. Nason, *A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, JHEP **2007(09)**, 126 (2007)

[49] R. Ball, et al., *Parton distributions for the LHC run II*, JHEP **2015(4)** (2015)

[50] T. Sjöstrand, et al., *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191**, 159 (2015)

[51] ATLAS Collaboration , *ATLAS Pythia 8 tunes to 7* TeV *data* (2014)

[52] A. Duperrin (ATLAS), *Flavour tagging with graph neural networks with the ATLAS detector*, in *30th International Workshop on Deep-Inelastic Scattering and Related Subjects* (2023)

[53] *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, Technical report, CERN, Geneva (2017)

Erklärung nach §17(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen: Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.

Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den 14. Juli 2025

(Tim Schlömer)