Identification of bb-Jets Using a Deep-Sets-Based Flavour-Tagging Algorithm at the ATLAS Experiment

Master Thesis

UNIVERSITY OF FREIBURG

Department of Mathematics and Physics

Joschka Birk

Supervisor: Prof. Dr. Gregor Herten

Submission date: December 7, 2022

Acknowledgement

First, I would like to thank Prof. Dr. Gregor Herten for giving me the opportunity to complete both my Bachelor thesis and my Master thesis in his research group. Furthermore, I would like to thank PD Dr. Andrea Knue and Dr. Manuel Guth for their supervision, many insightful discussions and the support I received during this time. In addition, I would like to thank Alexander Froch for always helping me in case of any questions. I would also like to thank Ksenia Solovieva for her support throughout the course of my Master thesis. Finally, I would like to thank all the members of the AG Herten for the great time.

Abstract

Flavour tagging is a crucial tool in particle physics since it allows to identify the flavour of heavy hadrons inside hadronic jets. This is especially important for challenging analyses like the search for the $t\bar{t}H(\rightarrow b\bar{b})$ signal. This search suffers from a large irreducible $t\bar{t}+b\bar{b}$ background which contains the same particles in the final state as the signal. In the background process, a radiated gluon can split into a $b\bar{b}$ pair, which can lead to a single jet that contains two b-hadrons (bb-jet). A successful identification of bb-jets could allow for a better rejection of this background. However, this is not possible with common flavour-tagging algorithms. In order to enable bb-jet identification, an extended flavour-tagging is presented in this thesis. The design of this algorithm is built on the deep-learning-based flavour-tagging algorithm DL1d, which combines the output of multiple low-level algorithms and is used in the ATLAS experiment in Run III. One of those low-level algorithms is the DIPS algorithm, which is based on the machine-learning concept of Deep Sets. The DL1d tagger receives an additional output category for bb-jets, which is done in two stages. First, the DIPS algorithm itself is extended (bb-DIPS), which exploits the information of the tracks that are associated to the jet. Afterwards, an extended version of DL1d (bb-DL1d) is implemented, which combines the output of bb-DIPS with the output of two other low-level algorithms. Furthermore, optimisation studies regarding the definition of the tagger discriminant are performed. The final bb-DL1d tagger is able to successfully identify bb-jets while still performing well in single-b-tagging. At a 70 % single-b-jet identification efficiency, bb-DL1d achieves a bb-jet rejection of 8.7 ± 0.3. Moreover, the same algorithm outperforms the ATLAS DL1r tagger used in Run II, by reaching a 27 % larger c-jet rejection while performing equally well in light-flavour-jet rejection at a 70 % inclusive b-jet efficiency. Finally, first steps regarding a cut in a two-dimensional discriminant plane are presented, providing an outlook for future applications of bb-DL1d.

Kurzzusammenfassung

Flavour-Tagging ist ein wichtiges Tool in der Teilchenphysik, da es die Identifizierung des Flavours von schweren Hadronen in hadronischen Jets ermöglicht. Dies gilt insbesondere für anspruchsvolle Analysen wie die Suche nach dem $t\bar{t}H(\rightarrow b\bar{b})$ - Signal. Diese Messung leidet unter einem großen irreduziblen $t\bar{t}+b\bar{b}$ - Untergrund, der die gleichen Teilchen im Endzustand enthält wie der Signalprozess. Im Untergrundprozess kann sich ein abgestrahltes Gluon in ein $b\bar{b}$-Paar aufspalten, was zu einem einzigen Jet führen kann, der zwei b-Hadronen enthält (bb-Jet). Eine erfolgreiche Identifizierung von bb-Jets könnte eine bessere Unterdrückung dieses Untergrundes ermöglichen. Dies ist jedoch mit den üblichen Flavour-Tagging-Algorithmen nicht möglich. Um die Identifizierung von bb-Jets zu ermöglichen, wird in dieser Arbeit ein erweiterter Flavour-Tagging-Algorithmus präsentiert. Das Design dieses Algorithmus basiert auf dem Deep-Learning basierten Flavour-Tagging-Algorithmus DL1d, welcher die Ergebnisse mehrerer Low-Level Algorithmen kombiniert und im ATLAS-Experiment in Run III verwendet wird. Einer dieser Low-Level-Algorithmen ist der DIPS-Algorithmus, welcher auf dem Machine-Learning-Konzept der Deep Sets basiert. Der DL1d-Tagger erhält eine zusätzliche Klassifizierungskategorie für bb-Jets. Zunächst wird der DIPS-Algorithmus erweitert (bb-DIPS), welcher die Informationen der dem Jet zugeordneten Tracks ausnutzt. Danach wird eine erweiterte Version von DL1d (bb-DL1d) implementiert, welcher die Ergebnisse von bb-DIPS mit den Ergebnissen von zwei anderen Low-Level-Algorithmen kombiniert. Außerdem werden Optimierungsstudien zur Definition der Tagger-Diskriminante durchgeführt. Der endgültige bb-DL1d-Tagger ist in der Lage, bb-Jets erfolgreich zu identifizieren und gleichzeitig eine gute Leistung in Single-b-Tagging zu erzielen. Bei einer 70 % Erkennungseffizienz von Single-b-Jets erreicht bb-DL1d eine bb-Jet-Rückweisung von 8.7 ± 0.3. Darüber hinaus übertrifft derselbe Algorithmus den in ATLAS für Run II verwendeten DL1r-Tagger, indem er eine um 27 % höhere c-jet-Rückweisung erreicht, während er bei einer b-jet-Effizienz von 70 % die gleiche Leistung bei der Rückweisung von Light-Flavour-Jets zeigt. Abschließend werden erste Schritte bezüglich eines Schnitts in einer zweidimensionalen Diskriminantenebene vorgestellt, die einen Ausblick auf zukünftige Anwendungen von bb-DL1d geben.

Contents

1. Introduction

The Standard Model (SM) of particle physics is one of the most successful theories in physics. It describes the elementary particles that make up our universe and the interactions between these particles at very high precision. However, despite its success, many questions are still unsolved and thus both the SM itself and physics beyond the SM are an important area of current physics research.

One of the biggest achievements in experimental particle physics was the discovery of the Higgs boson in 2012, which was realised by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) [1, 2]. Finding a signal that is compatible with the Higgs boson as predicted by the SM allowed a much deeper understanding of our universe, since it is essential for our understanding of how massive particles acquire their mass. However, some properties of the Higgs boson and its interactions with other particles are still to be measured with high precision. One such measurement is the search for the $t\bar{t}H(\rightarrow b\bar{b})$ signal performed by the ATLAS experiment [3]. In the $t\bar{t}H(\rightarrow b\bar{b})$ process, a Higgs boson is produced in association with a top-antitop-quark pair and decays itself into a bottom-antibottom-quark pair. This production mode of the Higgs boson makes up only a small fraction of the overall produced Higgs bosons at the LHC. Hence, advanced analysis and reconstruction methods are required to observe this process.

In this thesis, a certain aspect of this analysis chain is studied in more detail, which is the method of jet flavour tagging. Quarks cannot be measured directly in the detector. Instead, they undergo the process of fragmentation and hadronisation, which leads to the formation of bound, colourless states, called hadrons. In a particle detector, this final state can be measured in the form of hadronic jets, which are clustered from the tracks of charged particles and energy deposits in the calorimeter. Flavour tagging allows to identify the flavour of the hadrons inside the jet and relies heavily on machine-learning techniques.

Flavour tagging is a crucial tool in the $t\bar{t}H(\rightarrow b\bar{b})$ measurement, due to the large number of quarks that are involved in this process. Moreover, the $t\bar{t}H(\rightarrow b\bar{b})$ measurement suffers from a large irreducible $t\bar{t}+b\bar{b}$ background. In this background process, a radiated gluon can split into a b-quark pair, potentially leading to the formation of a bb-jet, which is a jet that contains two b-hadrons. As motivated in Refs. [4, 5], being able to identify bb-jets could improve the identification of those background events and thus allow to define cleaner signal regions in the analysis.

The studies presented in this work investigate the design and training of a Deep-Sets-based flavour-tagging algorithm that is not just able to differentiate between the commonly used target classes used in flavour tagging, but is also able to identify if a jet was initiated by two bottom quarks.

An overview of the Standard Model is given in Chapter 2, followed by an introduction to physics at hadron colliders in Chapter 3. The ATLAS detector and the LHC at CERN are described in Chapter 4. An introduction to the core concepts of machine learning with deep neural networks is given in Chapter 5. The overall concept of flavour tagging as well as the corresponding algorithms that are used in ATLAS are introduced in Chapter 6. Afterwards, the training and performance studies of the extended flavour-tagging algorithm is presented in Chapter 7 and Chapter 8. Finally, the results of this thesis are summarised in Chapter 9.

2. The Standard Model of Particle Physics

The *Standard Model* (SM) is a theoretical framework that describes the fundamental interactions between elementary particles. It includes the electromagnetic, weak and strong force, while the gravitational force is not described by the SM.

The particle content of the SM is shown in Figure 2.1. It consists of fermions and their anti-particles, gauge bosons and the Higgs boson. Fermions are spin 1/2 particles and are further divided into leptons and quarks while gauge bosons carry a spin of 1 and represent the mediators of the fundamental forces. The Higgs boson is the only particle in the SM that has a spin of 0. Both leptons and quarks can be grouped in three generations, where the particles of each generation are heavier copies of the particles from the preceding generation. However, all visible matter that surrounds us contains almost exclusively first-generation fermions[1].

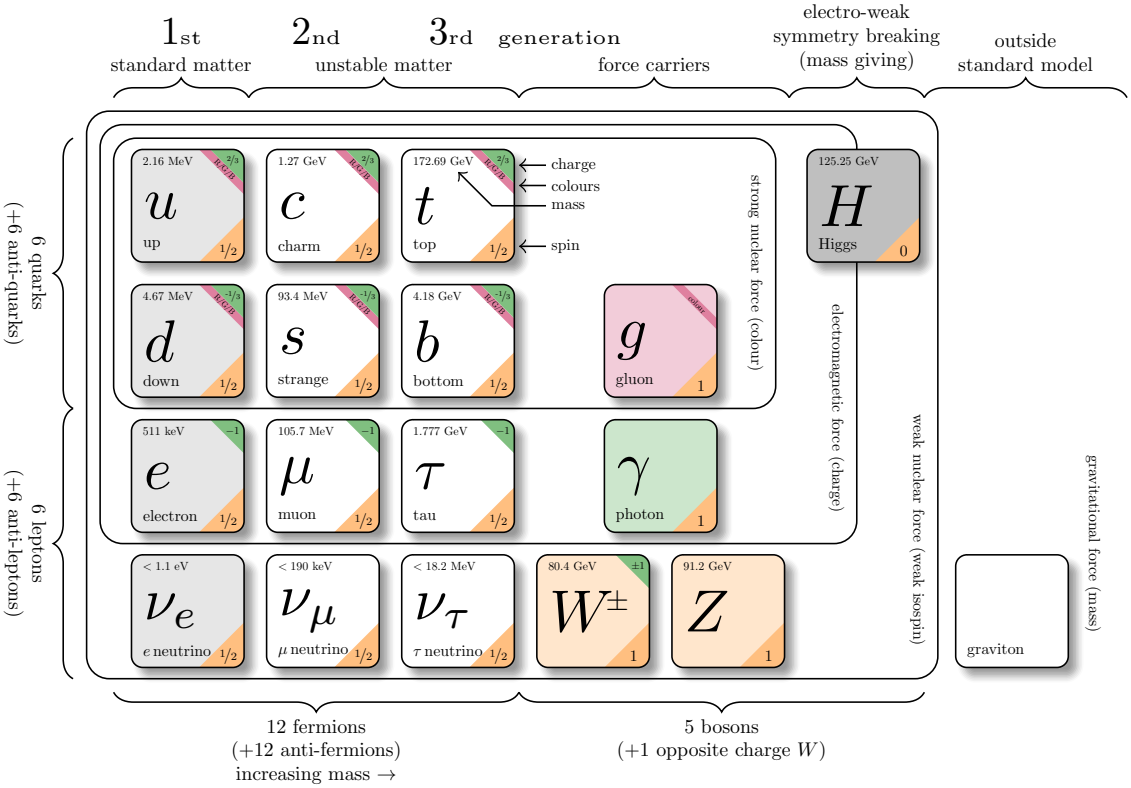


Figure 2.1.: Particle content of the Standard Model. Graphic adapted from [6, 7] and mass values taken from [8].

The electromagnetic force is mediated by the massless photon, which couples to all fermions that carry electric charge. The electric charge is usually expressed as multiples of the electric charge of the electron $e = 1.602\,176\,634 \times 10^{-19}\,\mathrm{C}$ [8]. While electrons, muons and taus carry an electric charge of $1\,e$, up-type quarks and down-type quarks carry an electric charge of $2/3\,e$ and $-1/3\,e$, respectively. The mediators of the weak force are three massive particles, namely the charged $W^{\pm}$ bosons and the electrically neutral Z boson. The $W^{\pm}$ bosons are the only particles that can change the flavour of quarks via the trans-

[1] Atoms consist of electrons and and nucleons, and the nucleons are made of up and down quarks.

ition $q \to Wq'$, where the flavour of the quark q is changed from up-type to down-type or vice versa. The corresponding transition probabilities $|V_{ij}|^2$ are described using the unitary Cabibbo–Kobayashi–Maskawa (CKM) matrix [9, 10], which is given by

$$V_{\mathrm{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} , \tag{2.1}$$

with the index of the entries indicating the change in flavour. The diagonal elements are measured to be very close to one, while the off-diagonal elements are rather small values. This means that flavour changing transitions via a $W^{\pm}$ boson usually happen within the same quark generation. This is most dominant in $|V_{tb}| = 0.999118$ [8] which shows that the top quark decays almost exclusively into a b-quark. Finally, the strong interaction is mediated by the gluons, which are massless particles that couple to colour charge. Only quarks and gluons carry colour charge, which is why these are the only particles taking part in this type of interaction.

The SM is a relativistic quantum field theory based on the idea of requiring local gauge invariance of the underlying Lagrange density $\mathcal{L}$, in the following referred to as *Lagrangian*. The gauge group of the SM is the non-abelian group

$$\mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y \, , \tag{2.2}$$

where C represents the colour charge, which is introduced in Quantum Chromodynamics, L stands for left-handed fields and Y for the weak hypercharge. The latter two form the foundation of the electroweak interaction, which is the unification of the electromagnetic and the weak interaction [11, 12, 13]. The dynamics of all known elementary particles are then described by the Lagrangian $\mathcal{L}_{\mathrm{SM}}$ that satisfies the gauge invariance mentioned above. This Lagrangian consists of the contributions from the electroweak theory, Quantum Chromodynamics, the Higgs sector and the Yukawa coupling. These different components of the SM are described in more detail in the following sections. If not further specified, the content of this chapter is based on references [14, 15, 16].

2.1. Quantum Electrodynamics

Quantum electrodynamics (QED) is the gauge theory of electromagnetic interactions. In relativistic quantum mechanics, a spin 1/2 particle of mass m is described by the Dirac equation, which is given by

$$(i\gamma^{\mu}\partial_{\mu} - m)\,\psi = 0 \, , \tag{2.3}$$

where γ^{μ} represents the gamma matrices and ψ is a four-component Dirac spinor. The corresponding Lagrangian is given by

$$\mathcal{L}_{\mathrm{Dirac}} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \, . \tag{2.4}$$

This Lagrangian is invariant under the global gauge transformation

$$\psi \to e^{i\theta}\psi \, , \tag{2.5}$$

where θ is an arbitrary real number representing a global change in the phase of the Dirac spinor. However, it is not invariant under a local gauge transformation [14]

$$\psi \to e^{i\theta(x)}\psi = e^{-i\,q\lambda(x)}\psi \, , \tag{2.6}$$

where the phase θ changes as a function of x. The definition $\theta(x) = -q\lambda(x)$ is a commonly chosen convention, such that the factor q represents the electric charge [14]. If the Lagrangian is required to be invariant under such a local gauge transformation, which corresponds to the abelian symmetry group $U(1)$, a gauge term has to be added to the Dirac Lagrangian, such that

$$\mathcal{L} = \underbrace{i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi}_{\mathcal{L}_{\text{Dirac}}} - \underbrace{q(\bar{\psi}\gamma^\mu\psi)A_\mu}_{\text{gauge term}} . \tag{2.7}$$

The gauge field A_μ, which represents the vector field of the massless[2] photon, transforms under the local gauge transformation from Equation 2.6 like

$$A_\mu \to A_\mu + \partial_\mu\lambda . \tag{2.8}$$

With this definition of A_μ, the gauge term then cancels the additional term that appears in $\mathcal{L}_{\text{Dirac}}$ when transforming ψ according to Equation 2.6. By also adding the free term for the vector field A_μ, the full Lagrangian for QED is given by

$$\mathcal{L}_{\text{QED}} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - q(\bar{\psi}\gamma^\mu\psi)A_\mu , \tag{2.9}$$

where $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. The QED Lagrangian therefore consists of the free terms of the Spinor field ψ, the free term of the photon field A_μ and an interaction term. However, the photon only couples to the field ψ with a strength proportional to the electric charge. As photons are not electrically charged, there is no photon-photon interaction. An alternative way of writing Equation 2.9 is to introduce a *covariant derivative* of the form [14]

$$D_\mu = \partial_\mu + iqA_\mu , \tag{2.10}$$

which results in

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} . \tag{2.11}$$

A global gauge transformation can therefore be extended to a local gauge transformation by introducing the covariant derivative D_μ from Equation 2.10, substituting ∂_μ with D_μ and adding the free term of the new field A_μ [14].

2.2. Quantum Chromodynamics

Quantum chromodynamics (QCD) is the theory that describes the strong interaction between particles that carry colour charge, which are quarks and gluons. The derivation of the Lagrangian follows the same approach as for QED. However, in this case the local gauge transformation under which the Lagrangian is required to be invariant, is a rotation of the quark colour field. The quark colour field can be written as

$$\psi^f = \left(\psi^f_R , \psi^f_G , \psi^f_B\right)^T , \tag{2.12}$$

where the index f represents the quark flavour and the quark colours are red (R), green (G) and blue (B). For each colour there is also an anti-colour, e.g. B (blue) and $\bar{B}$ (anti-

[2]The fact that the photon has to be massless is another requirement in order to achieve local gauge invariance. This is due to the fact that adding the Lagrangian of a spin 1 particle includes a mass term which has to vanish for the total Lagrangian to satisfy the local gauge invariance.

blue) cancel each other. In addition to that, the combination RGB is also colour neutral. The symmetry group of QCD is the non-abelian $SU(3)$ group and the resulting gauge transformation can then be written as

$$\psi^f \rightarrow e^{i\,\theta_a(x)\,\lambda_a}\,\psi^f \quad , \; a \in \{1, ..., 8\}\,, \tag{2.13}$$

with λ_a being the eight Gell-Mann matrices [17], which are the generators of the $SU(3)$ group. The local phase shift $\theta_a(x)$ is a real number for each of the eight components. The Gell-Mann matrices obey the commutation rule

$$[\lambda_\alpha, \lambda_\beta] = 2i f_{\alpha\beta\gamma}\lambda_\gamma\,, \tag{2.14}$$

where $f_{\alpha\beta\gamma}$ is the completely antisymmetric structure constant [17]. The local gauge invariance is then achieved by introducing the covariant derivative [14]

$$D_\mu = \partial_\mu + i g_s \lambda_a A^a_\mu\,, \tag{2.15}$$

where the index a runs again from zero to eight and the fields A^a_μ represent the vector fields of the eight gluons. The constant g_s quantifies the strength of the coupling and is related to the coupling constant α_s of the strong interaction via

$$g_s = \sqrt{4\pi\alpha_s}\,. \tag{2.16}$$

Using the field strength tensors $F^a_{\mu\nu}$, which are defined as

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - 2\,g_s f_{abc} A^b_\mu A^c_\nu\,, \tag{2.17}$$

the full Lagrangian for QCD can be evaluated to [14]

$$\mathcal{L}_{\text{QCD}} = i\bar{\psi}^f\gamma^\mu D_\mu\psi^f - m\bar{\psi}^f\psi^f - \frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu} \tag{2.18}$$

$$= i\bar{\psi}^f\gamma^\mu\partial_\mu\psi^f - m\bar{\psi}^f\psi^f - \frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu} - g_s(\bar{\psi}^f\gamma^\mu\lambda_a\psi^f)A^a_\mu\,. \tag{2.19}$$

Due to the structure of $F^a_{\mu\nu}$ from Equation 2.17, this Lagrangian contains interaction terms between gluons, which means that not only quarks but also gluons carry colour charge. This is due to the fact that $SU(3)$ is non-abelian.

A characteristic feature of the strong force is that its strength depends on the energy scale Q^2 of the corresponding interaction. The formula for $\alpha_s(Q^2)$ is given by [14]

$$\alpha_s(Q^2) = \frac{12\pi}{(33 - 2\,n_f)\ln\left(Q^2/\Lambda^2_{\text{QCD}}\right)}\,, \quad (Q^2 \gg \Lambda_{\text{QCD}})\,, \tag{2.20}$$

where n_f is the number of active quark flavours and Λ_{QCD} is a constant. A summary of measurements of $\alpha_s(Q^2)$ is shown in Figure 2.2.

This dependence of the strong coupling constant α_s on the energy scale, also referred to as *running coupling*, leads to a very distinct behaviour for particles which carry colour charge. At high energies, the coupling strength α_s becomes very small and hence quarks at high energies behave like free particles. This is called *asymptotic freedom*. On the other hand, with α_s becoming very large for small energies, colour-charged particles cannot exist as isolated particles, which is called *confinement*. The potential energy between two quarks increases when they are separated from each other. At a certain point, this potential energy is sufficiently large to create a quark-antiquark pair. This process is repeated until the energy is too low to produce new particles. The created particles form *hadrons*, which are colour neutral bound states. Given the previously mentioned structure of the colour

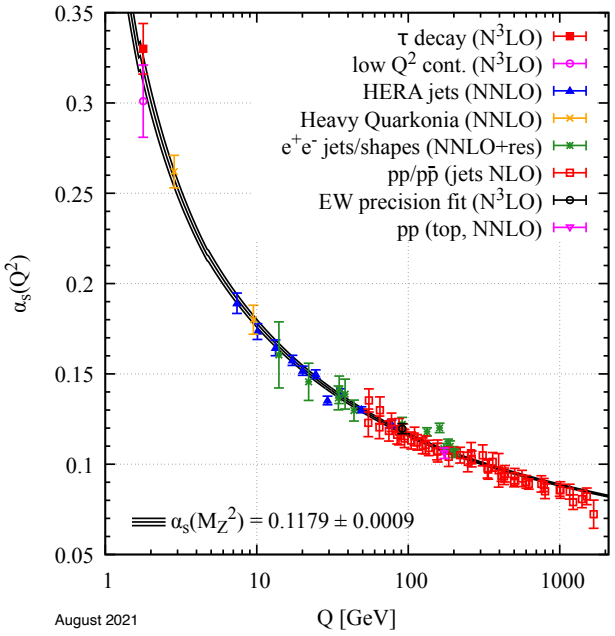


Figure 2.2.: Summary of measurements of α_s as a function of the energy scale Q. The respective degree of QCD perturbation theory used in the extraction of α_s is indicated in the brackets [8].

charge, these hadrons are either mesons, which consist of a quark and an anti-quarks with a colour-anti-colour combination or baryons, which consist of three quarks with different colour each [15].

2.3. Electroweak Unification

The electromagnetic force and the weak force were unified in the 1960s by Glashow, Salam and Weinberg [11, 12, 13]. This theory of electroweak (EW) interactions is based on the gauge group $SU(2)_L \otimes U(1)_Y$, where L stands for left-handed fields and Y represents the *hypercharge*. This definition of a more general charge relates the electric charge Q and the third component of the weak isospin I_3 via the Gell-Mann–Nishijima formula [18, 19]

$$Y = 2\,(Q - I_3)\ . \tag{2.21}$$

The fermions are grouped into weak isospin singlets and weak isospin doublets as summarised in Table 2.1. The distinction between left- and right-handed fields is important in EW interactions, since the coupling of the W boson has a *vector−axialvector* form. As a result of that, the W boson only couples to left-handed fermion states and right-handed anti-fermion states. The left-handed component ψ_L and the right-handed component ψ_R of a spinor ψ are defined as

$$\psi_L = \left(\frac{1-\gamma^5}{2}\right)\psi \quad \text{and} \quad \psi_R = \left(\frac{1+\gamma^5}{2}\right)\psi\,, \tag{2.22}$$

where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$.

The Glashow-Weinberg-Salam model introduces three massless gauge bosons W^1_μ, W^2_μ, W^3_μ with the coupling constant g and one gauge boson B_μ with the coupling constant $g'/2$. One can define two charged bosons $W^\pm$ as well as the neutral photon and Z boson as linear

Table 2.1.: Grouping of the fermions into isospin singlets and doublets. The quantum numbers are the electric charge Q, the third component of the weak isospin I_3 and the hypercharge Y. The subscripts L and R stand for left- and right-handed fields.

Fermion generations I	Fermion generations II	Fermion generations III	Q	I_3	Y
$\begin{pmatrix}\nu_e\\ e\end{pmatrix}_L$	$\begin{pmatrix}\nu_\mu\\ \mu\end{pmatrix}_L$	$\begin{pmatrix}\nu_\tau\\ \tau\end{pmatrix}_L$	0 -1	$+1/2$ $-1/2$	-1 -1
e_R	μ_R	τ_R	-1	0	-2
$\begin{pmatrix}u\\ d\end{pmatrix}_L$	$\begin{pmatrix}c\\ s\end{pmatrix}_L$	$\begin{pmatrix}t\\ b\end{pmatrix}_L$	$+2/3$ $-1/3$	$+1/2$ $-1/2$	$+1/3$ $+1/3$
u_R	c_R	t_R	$+2/3$	0	$+4/3$
d_R	s_R	b_R	$-1/3$	0	$-2/3$

combinations of these fields. The charged W bosons are given by

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp W_\mu^2\right) , \tag{2.23}$$

and the photon A and the Z boson by [15]

$$A_\mu = \sin\left(\theta_W\right) W_\mu^3 + \cos\left(\theta_W\right) B_\mu , \tag{2.24}$$

$$Z_\mu = \cos\left(\theta_W\right) W_\mu^3 - \sin\left(\theta_W\right) B_\mu , \tag{2.25}$$

where θ_W is the *Weinberg angle*, also called *weak mixing angle*. The Weinberg angle relates the two coupling constants of the electroweak interaction via the relation

$$\sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} . \tag{2.26}$$

The coupling between the fermions and the bosons of the EW theory is then introduced via a covariant derivative [15]. However, given that the $W^\pm$ bosons only couple to left-handed fermions, the covariant derivative looks slightly different for left- and right-handed fields. The resulting covariant derivatives D_μ^L and D_μ^R for left- and right-handed fields, respectively, are thus given by

$$D_\mu^L = \partial_\mu - \frac{1}{2} i\, g\, \sigma^a\, W_\mu^a - \frac{1}{2} i\, g'\, Y\, B_\mu , \tag{2.27}$$

$$D_\mu^R = \partial_\mu - \frac{1}{2} i\, g'\, Y\, B_\mu , \tag{2.28}$$

where σ^a are the three Pauli matrices. Using these definitions of the covariant derivatives, the following Lagrangian can be obtained:

$$\mathcal{L}_{\mathrm{EW}} = i\bar{\psi}_L \gamma^\mu D_\mu^L \psi_L + i\bar{\psi}_R \gamma^\mu D_\mu^R \psi_R - \frac{1}{4} W^{a\mu\nu} W_{\mu\nu}^a - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} . \tag{2.29}$$

In the above equation, the symbols $W_{\mu\nu}^a$ and $B_{\mu\nu}$ are the field strength tensors of W_μ^a and B_μ, respectively. Hence, the latter two terms in Equation 2.29 represent the free terms of the bosons and the first two terms in Equation 2.29 represent the coupling of the fermions

with the bosons of the EW sector. Adding mass terms for the gauge bosons to Equation 2.29 explicitly breaks the $SU(2) \otimes U(1)$ symmetry. Thus, in order to allow for massive gauge bosons, the masses of the gauge bosons are introduced via a formalism called the *Higgs mechanism*.

2.4. The Higgs Mechanism

An elegant mechanism that allows massive $W^{\pm}$ and Z bosons in the theory was found independently by P. Higgs [20, 21, 22], R. Brout and F. Englert [23] as well as G. S. Guralnik, C. R. Hagen and T. W. B. Kibble [24, 25]. The idea of this mechanism is to combine the local gauge symmetry with the *spontaneous symmetry breaking* of a continuous symmetry. For that, a doublet of complex scalar fields Φ with hypercharge $Y = 1$ is introduced

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} . \tag{2.30}$$

For a potential $\mathcal{V}$ of the form

$$\mathcal{V}(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda \left(\Phi^\dagger \Phi \right)^2 \tag{2.31}$$

with $\mu^2 < 0$ and $\lambda > 0$, the ground states of the scalar field Φ are located at

$$\Phi^\dagger \Phi = \frac{1}{2} \left(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2 \right) = -\frac{1}{2} \frac{\mu^2}{\lambda} =: \frac{v^2}{2} . \tag{2.32}$$

The ground states are therefore the infinite number of states that satisfy $\Phi^\dagger \Phi = v^2/2$, where v is called the vacuum expectation value. A graphical representation of this scenario is shown in Figure 2.3 for the case of two dimensions. Without loss of generality, one can choose the ground state with $\phi_1 = \phi_2 = \phi_4 = 0$, which is called the *unitary gauge*. The doublet then takes the form [15]

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} , \tag{2.33}$$

where $H(x)$ represents an oscillation of the Higgs field Φ in the direction of ϕ_3.

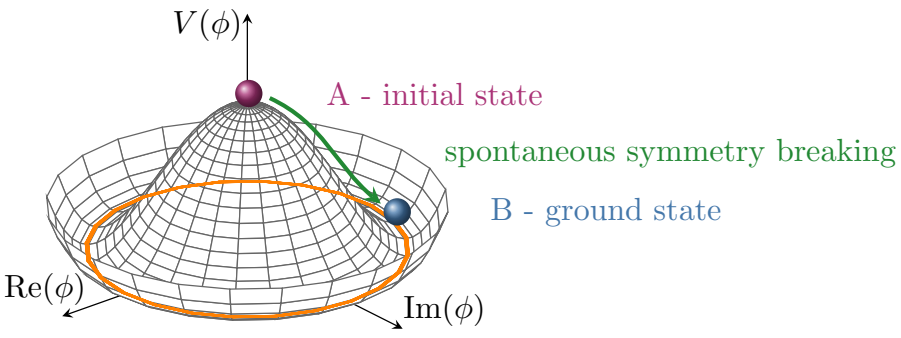


Figure 2.3.: Sketch of the process of spontaneous symmetry breaking in the case of a complex scalar field ϕ in a Mexican hat potential. The orange ring shows the infinite number of degenerate ground states. In the initial state A (magenta), the potential is symmetrical, but as soon as one of the ground states B (blue) is spontaneously chosen, the potential is not symmetrical any more. Adapted from [26].

Inserting this choice of ground state into Equation 2.31 results in

$$\mathcal{V}(\Phi_0) = -\lambda v^2 \left(\frac{v + H(x)}{\sqrt{2}}\right)^2 + \lambda \left(\frac{v + H(x)}{\sqrt{2}}\right)^4 . \tag{2.34}$$

With this ground state Φ_0, another term is added to the Lagrangian

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \Phi_0)^\dagger (D_\mu \Phi_0) - \mathcal{V}(\Phi_0) , \tag{2.35}$$

with the covariant derivative D_μ. Due to the spontaneous symmetry breaking, three mass terms emerge in the Lagrangian, which can be assigned to the massive $W^\pm$ and Z bosons. These masses are related to the vacuum expectation value of the Higgs field and the coupling constant g with the relations [15, 8]

$$m_W = \frac{1}{2} g\, v \quad \text{and} \quad m_Z = \frac{m_W}{\cos\theta_W} . \tag{2.36}$$

Furthermore, a new spin-0 particle with a mass of

$$m_H = \sqrt{-2\mu^2} \tag{2.37}$$

emerges, which is called the *Higgs boson*. Since μ is an unknown parameter in the theory, the mass of the Higgs boson could not be predicted.

Finally, the Higgs field interacts with the fermions via the *Yukawa coupling*, which relates the masses of the fermions m_f and the vacuum expectation value of the Higgs field v via a fermion-specific coupling constant λ_f [15]

$$m_f = \frac{\lambda_f}{\sqrt{2}} v . \tag{2.38}$$

The corresponding term in the SM Lagrangian is given for each fermion f by [15]

$$\mathcal{L}^f_{\text{Yukawa}} = -m_f \bar{\psi}\psi \left(1 - \frac{H}{v}\right) . \tag{2.39}$$

Since the top quark is the heaviest fermion in the SM, has the largest coupling with the Higgs boson. For a vacuum expectation value of $v = 246.22\,\text{GeV}$ and a top-quark mass of $m_{\text{top}} = (172.69 \pm 0.30)\,\text{GeV}$ [8] the Yukawa coupling of the top quark results in

$$\lambda_{\text{top}} = \sqrt{2}\,\frac{m_{\text{top}}}{v} \approx 0.992 . \tag{2.40}$$

3. Physics at Hadron Colliders

Particle colliders accelerate particles to velocities close to the speed of light and bring them to collision in order to study the fundamental interactions at the subatomic scale. Lepton colliders like the former Large Electron-Positron Collider (LEP) at CERN create very clean event structures which allow very precise measurements of the observed physics processes. The reason for this is that lepton colliders have a well-defined initial state and no hadronic interactions which leads to a small amount of background activity. However, circular lepton colliders are rather limited in the achievable collision energy. This is due to the fact that electrons loose a lot of energy as synchrotron radiation when moving on a circular trajectory[1]. Hadron colliders like the LHC at CERN or the former Tevatron collider at Fermilab have the advantage that the accelerated particles (usually protons) are much heavier than electrons, which reduces the amount of energy loss that is lost via synchrotron radiation drastically since this energy loss is proportional to $(E/m)^4$ with E being the energy and m the mass of the particle. While this allows to reach much higher collision energies, experiments at hadron colliders have to address the fact that hadrons are composite particles. This means that the individual constituents of these hadrons, called *partons*, interact in the collision. That gives rise to many possible interactions and rather complex event signatures in the detector with large backgrounds due to hadronic interactions.

3.1. Parton Distribution Functions

A hard-scattering process initiated by two hadrons with four-momenta P_1 and P_2 can be described with the factorisation theorem [27]. The cross-section σ of the process is then given by the integral over the parton-parton cross-section $\hat{\sigma}_{ij}(p_1, p_2, \hat{Q}^2)$ weighted with the so-called *parton distribution function* (PDF) $f_i(x, Q^2)$ [28]

$$\sigma\left(P_1, P_2\right) = \sum_{i,j} \int_0^1 \int_0^1 dx_1 dx_2 \, f_i\left(x_1, Q^2\right) f_j\left(x_2, Q^2\right) \hat{\sigma}_{ij}\left(p_1, p_2, \hat{Q}^2\right) . \tag{3.1}$$

In the above equation, $p_1 = x_1 P_1$ and $p_2 = x_2 P_2$ represent the momenta of the interacting partons which participate in the hard scattering and $\hat{Q}^2$ is the momentum transfer of the process. The PDF $f_i(x, Q^2)$ represents the probability that a parton of type i carries the fraction x of the momentum of the hadron, evaluated at the factorisation scale Q^2. In addition to the *valence quarks*, which are in the case of a proton two up-quarks and one down-quark, gluons and *sea quarks* also have to be taken into account in Equation 3.1. Sea quarks are virtual quark-antiquark pairs which only live for a very short time, but still carry part of the hadron momentum. With increasing momentum of the hadron momentum, the number of radiated quarks increases and thus also the number of sea quarks, which is why the PDFs depend on the factorisation scale Q^2 [8]. The value used for Q^2 is usually close to $\hat{Q}^2 = \hat{s} = x_1 x_2 s$ with s being the centre-of-mass energy of the collision. The PDFs of the partons inside a proton from the NNPDF3.0 global analysis [29] are shown in Figure 3.1 for two different factorisation scales. They are obtained from hard-scattering collisions combining recent results from the LHC with data from former particle accelerators like HERA [29].

[1] This is not a limitation for linear colliders as there is no angular acceleration in linear colliders.

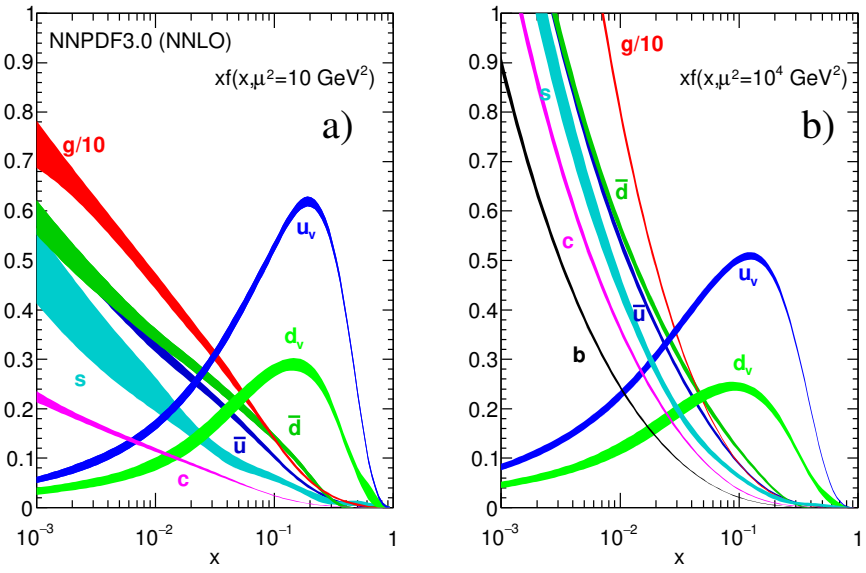


Figure 3.1.: Parton distribution functions $x \cdot f(x, Q^2)$ as a function of x, with x being the proton momentum fraction carried by the parton. The distributions u_v and d_v represent the PDFs of the valence quarks, while the other quark distributions correspond to the sea quarks. The plots show the results from the next-to-next-to-leading order (NNLO) NNPDF3.0 global analysis [29] at $Q^2 = 10\,\mathrm{GeV}^2$ and $Q^2 = 10^4\,\mathrm{GeV}^2$. Plots taken from [30].

3.2. Monte Carlo Generators

To test a proposed model, like the Standard Model of Particle Physics, theoretical predictions are compared to the observed data. To achieve this, theoretical predictions are approximated using Monte Carlo (MC) generators, which are programs that generate events based on the mathematical principle of MC simulation.

Starting from a hard collision, MC generators simulate the different stages which lead to the final-state particles that can be observed in an experiment. These steps are schematically shown in Figure 3.2. The first stage of the MC simulation process is the *hard subprocess*, which is a highly energetic collision of constituents of the colliding beams (centre red disc in Figure 3.2). The corresponding matrix element (ME) can be calculated perturbatively. For hadron-hadron collisions, the momenta of the colliding partons are obtained by sampling the parton distribution functions. The partons involved in the hard subprocess will emit further radiation in the form of photons and gluons, called *parton shower* (PS) (red, tree-like structure in Figure 3.2). This can happen both in the initial state, leading to *initial state radiation* (ISR), and in the final state, leading to *final state radiation* (FSR). The PS stops when the energy of a parton falls below a certain energy threshold. As mentioned in Section 2.2, coloured particles cannot exist in isolation, and instead form hadrons, which are colourless bound states (light-green discs in Figure 3.2). The process that describes the formation of hadrons is called *hadronisation*. However, the energy scale of hadronisation is small, which means that the coupling constant α_s of the strong force is very large. Thus, the modelling of the hadronisation process relies on non-perturbative calculations with models like the string model [32] and the cluster model [33]. These models have free parameters which are tuned (constrained) based on observed data.

The last step of the event generation is the decay of unstable hadrons that were formed in the hadronisation process (dark-green discs in Figure 3.2). In the area of collider event simulation, there are three main programs that offer all steps from the hard scattering simulation to the final state. These programs are Herwig7 [34, 35], Pythia8 [36] and

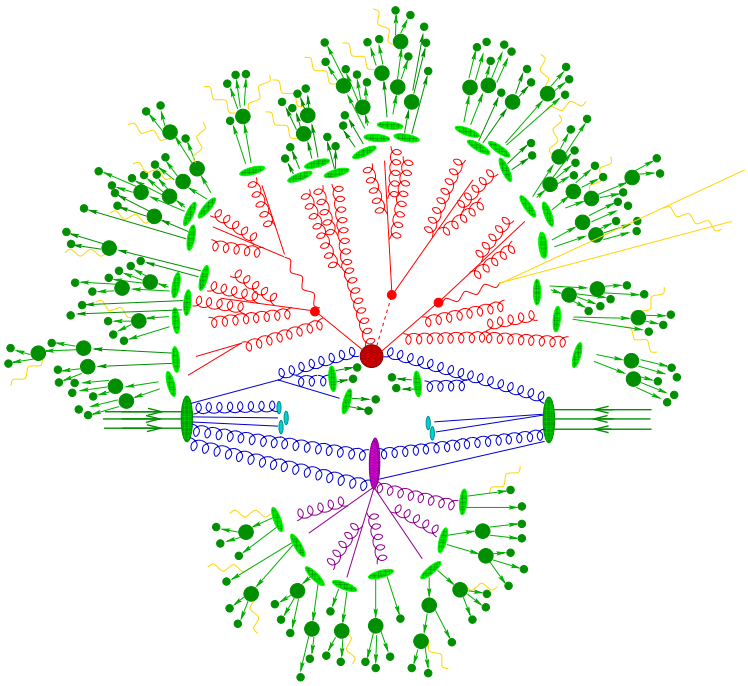


Figure 3.2.: Sketch of a hadron-hadron collision as simulated by a Monte Carlo event generator. The centre red point represents the primary hard collision. The parton showers are shown in the red tree-like structure around the primary hard collision. The purple disc indicates a secondary hard collision, the light green discs represent parton-to-hadron transitions and the dark green discs indicate hadron decays. Figure taken from Ref. [31]. Although not specified in Ref. [31], the particles in the sketch are consistent with the particles involved in the $t\bar{t}H(\to b\bar{b})$ process.

Sherpa [37]. In addition to that, there are other generators like PowhegBox [38, 39, 40, 41, 42] or MadGraph5_aML@NLO [43] which also provide ME calculations. The main programs currently used for top-quark analyses in ATLAS are PowhegBox, Sherpa and MadGraph5_aMC@NLO. Except for Sherpa, these are interfaced with Pythia8 or Herwig7 for PS and hadronisation.

3.3. Hadronic Jets

The hadronisation process leads to the formation of *hadronic jets*, in the following denoted as jets. Jets are cone-like structures in the detector which contain the tracks of many particles that are created during hadronisation. The jets used throughout all studies presented in this thesis are *Particle Flow* (PFlow) jets [44], which are jets reconstructed from both tracking and calorimeter information, using the anti-k_T algorithm [45] with a radius parameter $R = 0.4$. Furthermore, the jets used in the studies presented here have $p_{\mathrm{T}} \geq 20\,\mathrm{GeV}$. This lower cut on the jet p_{T} is applied since low-p_{T} jets are not in the valid calibration range [46]. However, while this cut is applied on the calibrated jet transverse momentum, the jet p_{T} shown throughout this thesis represent the partially calibrated transverse momentum of the jet. Furthermore, since jets with $p_{\mathrm{T}} > 250\,\mathrm{GeV}$ are very rare in the samples used in this thesis, only jets with $p_{\mathrm{T}} \leq 250\,\mathrm{GeV}$ are used. In addition to the p_{T} requirements, jets are required to have a pseudorapidity value of $|\eta| < 2.5$ corresponding

to the $|\eta|$-coverage of the inner detector of the ATLAS detector, which is described in more detail in Section 4.2.2.

The jet labelling is performed by matching the b-hadrons, c-hadrons and τ leptons with a transverse momentum larger than 5 GeV to the jet within a $\Delta R = 0.3$ cone around the jet axis. If one or two b-hadrons are matched to the jet, the jet is labelled as single-b jet or bb-jet respectively. If this is not the case, but a c-hadron can be matched, the jet is labelled as c-jet. The same procedure follows for τ leptons, which yields τ-jets. However, τ-jets are not considered in the studies presented here. All remaining jet candidates are labelled as light-flavour jets. It should be noted that due to the hierarchy of the different jet categories in the jet-hadron matching, single-b jets and bb-jets can contain c-hadrons in addition to the matched b-hadron. Moreover, single-b jets, bb-jets and c-jets can contain tracks of τ-leptons.

Tracks are assigned to jets using the angular distance ΔR(track, jet) between the track and the jet axis. The maximum allowed $\Delta R(\text{track, jet})_{\text{max}}$ for a track to be associated with a jet decreases with the jet p_{T} and is approximately 0.45 for a jet with $p_{\text{T}} = 20$ GeV and 0.25 for $p_{\text{T}} = 200$ GeV.

3.4. Used MC Setup

The studies presented in this thesis make use of two MC samples, corresponding to two different physics processes. The ATLAS detector response simulation was performed for both samples with the full GEANT4 simulation [47].

The first sample includes $t\bar{t}$ events from proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. Only events from the lepton+jets channel and the dilepton channel are included. These events were generated using the POWHEG BOX v2 generator and the NNPDF3.0NLO PDF for the ME generation. The decays of b- and c-hadrons were simulated with the EVTGEN 1.6.0 program [48]. PYTHIA 8.230 [36] was used for the parton showering using the A14 set of tunable parameters [49] and the NNPDF2.3LO PDF [29]. The h_{damp} parameter[2] was set to $1.5\,m_t$, with m_t being the mass of the top quark.

The second sample contains Z+jets events from proton-proton collisions at $\sqrt{s} = 13$ TeV. In order to ensure that jets extracted from this sample are jets from a radiated gluon, only events with a leptonic decay of the Z boson are considered. Furthermore, to omit events with additional energy deposits in the calorimeters[3], only events where the Z boson decays into a pair of muons or neutrinos are used.

This sample was generated with the SHERPA 2.2.1 generator [37] and the NNPDF3.0NNLO PDF [29], providing next-to-leading order (NLO)-accurate ME calculations for up to two partons and ME accurate to leading order (LO) calculations for three and four partons. The PS was simulated with the SHERPA parton shower [50], using the parameters recommended by the SHERPA authors and the MEPS@NLO prescription [51, 52, 53, 54] .

[2]The resummation damping parameter h_{damp} is one of the parameters that controls the matching of POWHEG matrix elements to the parton shower and hence effectively regulates the high-p_{T} radiation against which the $t\bar{t}$ system recoils.

[3]The calorimeter system of the ATLAS detector will be discussed in Section 4.2.3.

4. The ATLAS Experiment at the Large Hadron Collider

To study the fundamental structure of the universe, physicists accelerate particles to speeds close to the speed of light and then bring them to collision. Such collider experiments can produce energy densities and temperatures close to those that prevailed shortly after the Big Bang. This allows not only to search for new types of particles, but also to study the interactions between elementary particles whose existence is already known. The studies presented in this thesis are carried out in the context of the LHC at the European Organization for Nuclear Research (CERN), which will be discussed in this chapter. In addition to that, the ATLAS experiment and its detector are explained in more detail.

4.1. The Large Hadron Collider

The LHC [55, 56, 57] is located in Geneva at CERN and is the highest-energy accelerator built so far. It has a circumference of 27 km and is able to produce proton-proton collisions with a centre-of-mass energy up to $\sqrt{s} = 13.6\,\text{TeV}$ ($\sqrt{s} = 13\,\text{TeV}$ during Run II). Additionally, the LHC can also accelerate and collide heavy ions. The acceleration of the hadrons is performed in several steps, using other accelerators located at the CERN complex. A sketch of the accelerator complex at CERN is shown in Figure 4.1.

Each accelerator increases the energy of the proton beam and then injects it to the next accelerator in the chain. Starting from a hydrogen bottle, electrons are separated

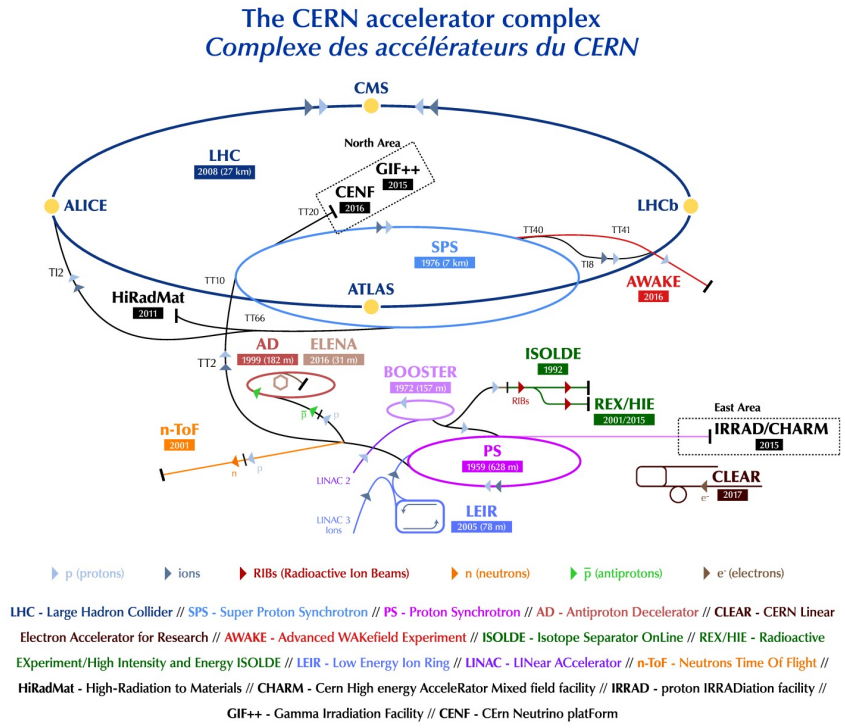


Figure 4.1.: Accelerator complex at CERN in August 2018 [58]. The accelerators used as pre-accelerators for the LHC are the Linac2, the Proton Synchrotron Booster (BOOSTER), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS).

from protons by passing through an electric field. The protons are then accelerated to an energy of 50 MeV in the Linac2 accelerator[1]. Afterwards, the Proton Synchrotron Booster increases the energy to 1.4 GeV. The beam is then injected to the Proton Synchrotron, where the protons are accelerated to an energy of 25 GeV. After this step, the Super Proton Synchrotron accelerates the protons to 450 GeV and then feeds the beam to the LHC. In the LHC, the protons are then accelerated to their final energy, which was 6.5 TeV in Run II, resulting in a centre-of-mass energy of 13 TeV. In the LHC, two beams circulate in opposite directions each consisting of 2808 proton bunches, where one bunch contains up to 1×10^{11} protons, which leads to around one billion pp collisions per second.

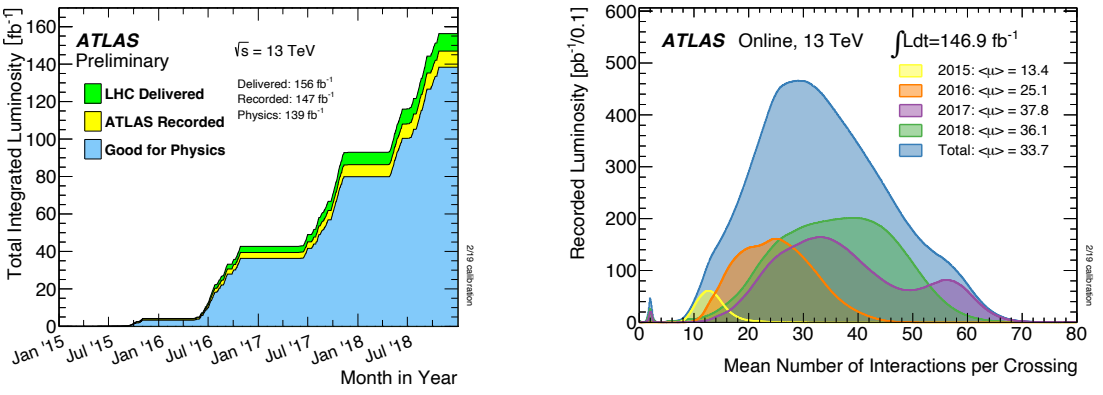
Along the LHC ring, there are four main collision points, each associated with a dedicated experiment. The LHCb experiment [59] focuses on measurements of CP violation and rare decays of B hadrons. In contrast, the ALICE experiment [60] is specialised on measurements of heavy-ion collisions to investigate the QCD sector of the Standard Model. The other two experiments, ATLAS [61] and CMS [62], are general-purpose detectors which are designed for the search for new particles and interactions as well as for precision measurements of the Standard Model. The amount of collisions that can be recorded and analysed by these experiments is described by the *instantaneous luminosity* L, which is defined as

$$L = \frac{1}{\sigma}\frac{dN}{dt}\,, \tag{4.1}$$

where σ is the cross-section and dN the number of events detected during the time interval dt. The total number N of recorded events is then given by the time integral of the instantaneous luminosity, also called *integrated luminosity* L_{int}, multiplied by the cross-section

$$N = \sigma \cdot \int dt\, L = \sigma \cdot L_{\text{int}}\,. \tag{4.2}$$

The integrated luminosity recorded with the ATLAS experiment during RUN II of the LHC operation is shown in Figure 4.2a. In order to obtain a large integrated luminosity



(a) **(b)**

Figure 4.2.: Luminosity measurements recorded with the ATLAS experiment during Run II of the LHC: **(a)** shows the integrated luminosity versus time and **(b)** the luminosity-weighted distribution of the average number of interactions per crossing split into the operation years of Run II [63].

[1]Starting with Run III, Linac2 is replaced with Linac4, which accelerates negatively charged H^- ions. The two electrons are stripped of during the injection from Linac4 to the Proton Synchrotron Booster.

and hence large datasets, a large instantaneous luminosity is desired. However, increasing the instantaneous luminosity is strongly associated with so-called *pile-up* events, where multiple proton-proton collisions overlap in the detected event. Pile-up events are further divided into in-time pile-up and out-of-time pile-up events. In-time pile-up means that multiple proton-proton collisions occur in one bunch crossing, while out-of-time pile-up reflects the case where particles from collisions from the bunch crossing before or after the collision of interest are recorded in the event. The amount of pile-up events can be quantified with the average number of pp interactions per bunch crossing, which is shown in Figure 4.2b for the years 2015 – 2018. These distributions, also called *pile-up profiles*, depend on the instantaneous luminosity, i.e. the slope of the integrated luminosity in Figure 4.2a. Therefore, since the instantaneous luminosity was not the same during the different operation periods of the LHC, the pile-up profiles is also different between the operation periods.

4.2. The ATLAS Detector

The ATLAS detector is a general-purpose detector installed 100 m below the surface at point 1 at CERN. It is used to record collision events that are used for a wide range of analyses performed by the ATLAS collaboration, which has over 3000 members from institutions around the world [64]. The research area of the ATLAS collaboration ranges from precision measurements of the Standard Model to searches for physics beyond the Standard Model like the existence of dark matter. The ATLAS detector is the largest volume particle detector ever built for a particle collider. It has a cylindrical shape, with a length of 46 m and a diameter of 25 m. A schematic representation of the full ATLAS detector is shown in Figure 4.3. It consists of several detector subsystems and strong magnetic fields, where each subsystem is dedicated to a component that is needed for the full reconstruction of the proton-proton collisions. In the following, the ATLAS coordinate system, the detector subsystems as well as the trigger and data acquisition system are explained in more detail.

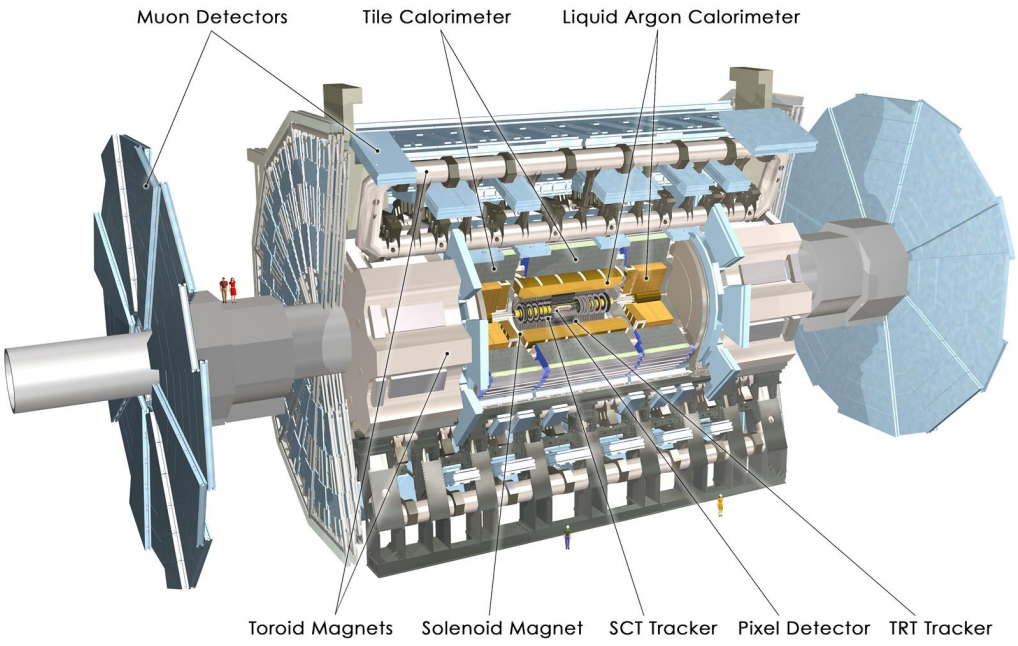
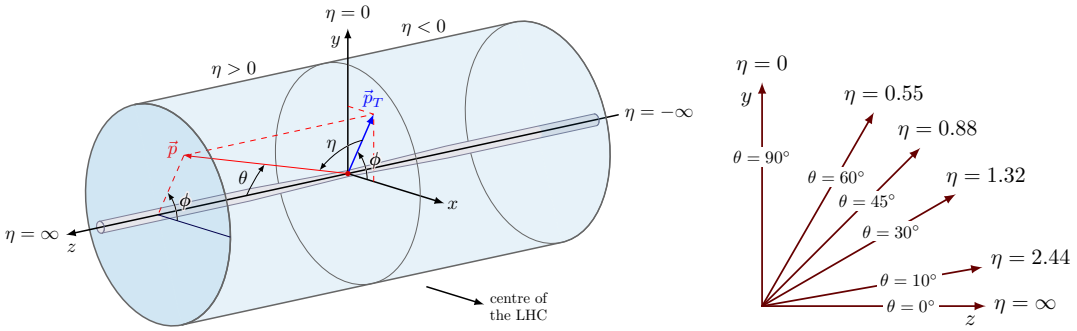


Figure 4.3.: Schematic representation of the ATLAS detector [65].



(a) **(b)**

Figure 4.4.: Illustration of the ATLAS coordinate system. Figure **(a)** shows a sketch of the coordinate system. The tube around the z-axis indicates the beam pipe and the red circle in the centre is the interaction point. Figure **(b)** shows the relation between the polar angle θ and the pseudorapidity η. Adapted from [66, 67].

4.2.1. Coordinate System

The ATLAS coordinate system is shown in Figure 4.4a. It is a right-handed coordinate system with the origin at the interaction point and the z-axis pointing along the beam pipe in the direction of the LHCb experiment. The x-axis points from the interaction point towards the centre of the LHC and the y-axis points upwards. Due to the detector's symmetry around the z-axis, coordinates are commonly described using the azimuthal angle ϕ and the *pseudorapidity* η, with the latter being a function of the polar angle θ, defined as

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) . \tag{4.3}$$

An illustration of this relation is shown in Figure 4.4b. Compared to the polar angle θ, the pseudorapidity has the advantage that differences $\Delta\eta$ are Lorentz invariant. The metric ΔR, which is used to describe the angular distance of vectors, is defined by combining the distance in the pseudorapidity and the distance in the transverse plane (xy-plane)

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} . \tag{4.4}$$

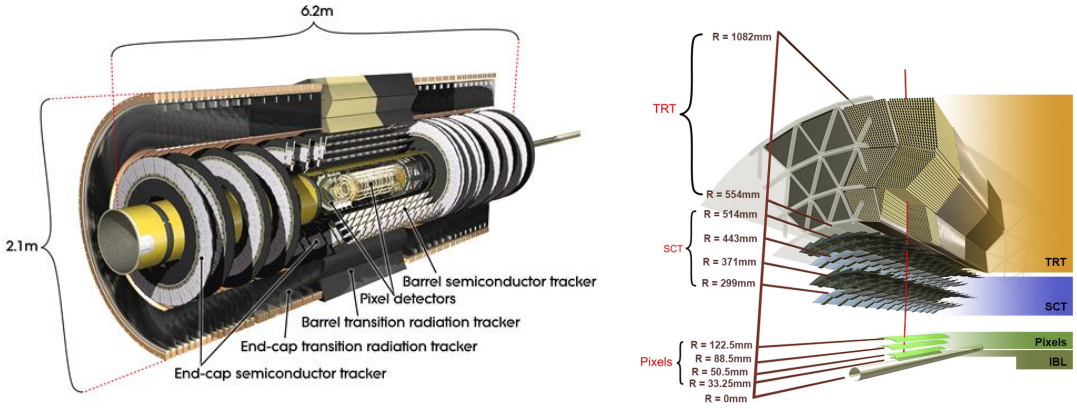
Since the z-axis of the coordinate system points along the beam line, the initial particles do not have a momentum component in the xy-plane, which is also called transverse plane. Thus, a commonly used variable to describe particles produced in the collision is the *transverse momentum* $\vec{p}_T$, which is the momentum component perpendicular to the z-axis

$$\vec{p}_T = \begin{pmatrix} p_x \\ p_y \end{pmatrix} . \tag{4.5}$$

This variable is also illustrated in Figure 4.4a.

4.2.2. Inner Detector

The inner detector (ID) [68, 69] is the innermost part of the ATLAS detector. It is very compact and designed for precise measurements of particle tracks, which means that it measures the direction, momentum and charge of electrically charged particles. By providing a precise measurement of particle tracks, the ID is the most important component for the reconstruction of primary and secondary decay vertices. This information, as well as fur-



(a) **(b)**

Figure 4.5.: The ATLAS inner detector: **(a)** the inner detector with the barrel and end cap components and **(b)** a cut-through illustration of the Pixel layers (including the insertable B-layer (IBL)), the TRT and the SCT [70].

ther track-related variables, are heavily exploited in flavour-tagging algorithms [2]. The inner detector consists of three main components namely the Pixel Detector, the semiconductor tracker (SCT) and the transition radiation tracker (TRT). A sketch of the inner detector can be seen in Figure 4.5a and a cut-through illustration of the three main components is shown in Figure 4.5b.

The innermost component, the Pixel Detector, is located only 3.3 cm from the LHC beam line. It consists of four layers in the barrel region and three disks in each of the end caps, featuring over 92 million silicon pixels in total [70]. The area of each pixel is 50 µm × 400 µm in the external layers and 50 µm × 250 µm for the innermost barrel layer, the insertable b-layer (IBL) [71]. The Pixel Detector covers a range up to $|\eta| < 2.5$.

The SCT fully surrounds the Pixel Detector and consists of six million micro-strip silicon sensors. The sensors are arranged in four layers in the barrel region and 18 planar end-cap discs, leading to a resolution of 16 µm in $R\phi$- and 580 µm in z-direction [68]. Like the Pixel detector, the SCT covers the $|\eta|$ range up to $|\eta| < 2.5$.

The TRT, which is the outermost component of the ID, is based on straw detectors. It consists of over 300 000 thin-walled drift tubes with a straw diameter of 4 mm, containing a gas mixture of Xe (70 %), CO_2 (27 %) and O_2 (3 %). Radiators are placed between the drift tubes, which are polypropylene/polyethylene fibres in the barrel region and polypropylene foils in the end-caps. Particles passing through the radiator emit transition radiation, generating photons that can be measured in the drift tubes. In contrast to the other two detector types of the inner detector, the TRT also provides information about the type of the detected particle. The reason for this is that the amount of the generated transition radiation depends both on the energy and the mass of the measured particle. Therefore, by combining the measurement from the TRT with the energy measurement from another detector component, the mass of the particle and therefore its type can be identified.

4.2.3. Calorimeter System

The ATLAS calorimeter system measures the energy of particles. All calorimeters used in ATLAS are sampling calorimeters, which consist of alternating layers of absorber material and active material. When particles interact with the absorber material, lower energy particles are created, leading to particle showers. The showering stops when the energy

[2] Flavour tagging is discussed in detail in chapter 6.

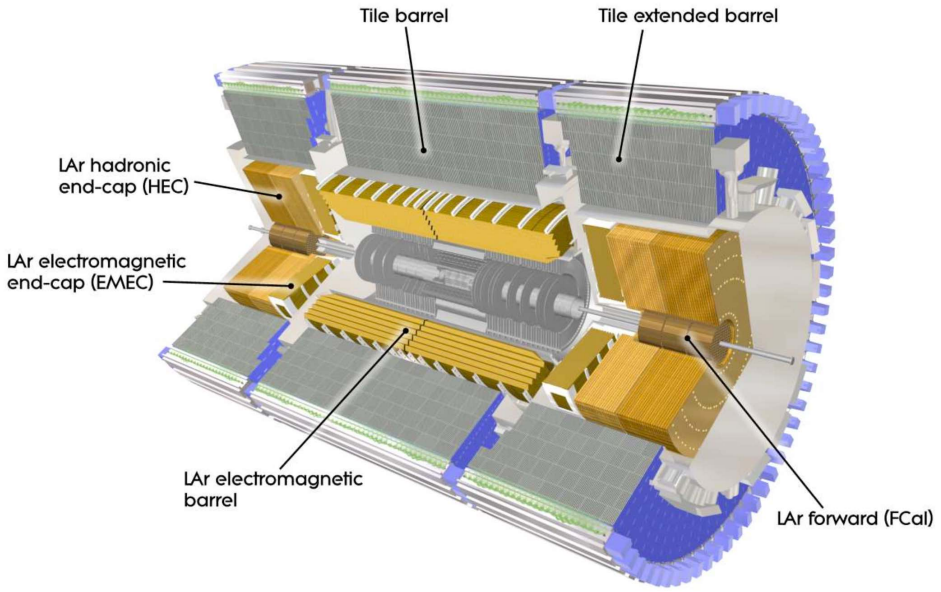


Figure 4.6.: Calorimeter system of the ATLAS detector [61].

of the particles becomes too small. Hence, calorimeters are required to have a sufficient thickness, such that the whole shower can be measured. The active material then measures the energy of these particle showers. The ATLAS calorimeter system is divided into the electromagnetic calorimeter and the hadronic calorimeter. While the electromagnetic calorimeter is designed to measure the energy of electrons and photons, the hadronic Calorimeters focuses on hadrons only. A sketch of the full ATLAS calorimeter system is shown in Figure 4.6.

Electromagnetic Calorimeters

The innermost part of the ATLAS calorimeter system is the electromagnetic calorimeter. It surrounds the ID and is made of lead plates as absorber material and liquid argon (LAr) as the active material. When the particles pass through the active layers, the LAr gets ionised and allows to measure the energy of the particles. The layers of the calorimeter are arranged in an accordion geometry in order to provide full coverage in the ϕ direction. The electromagnetic calorimeter consists of a barrel part and a part in each of the end-caps. The barrel part covers the region $|\eta| < 1.475$ and the electromagnetic end-cap calorimeter (EMEC) covers the region $1.375 < |\eta| < 3.2$.

Hadronic Calorimeters

There are three hadronic calorimeters, namely the LAr hadronic end-cap calorimeter (HEC), the LAr hadronic forward calorimeter (FCal) and the tile calorimeter. The HEC is a sampling calorimeter made of copper and liquid argon. It covers the region $1.5 < |\eta| < 3.2$ and is located behind the electromagnetic end-cap calorimeter. The hadronic forward calorimeter covers the region $3.1 < |\eta| < 4.9$, providing energy measurements of particles that are very close to the beam line [72]. It consists of three modules, where the module closest to the interaction point is optimised for electromagnetic measurement and uses copper as the absorber medium. The other two modules use tungsten as absorber medium and primarily measure the energy of hadronic interactions [61]. The tile calorimeter is divided into a central barrel part surrounding the LAr electromagnetic calorimeter and two so-called extended barrel parts that are located around the end-cap calorimeters, covering the region

$|\eta| < 1.7$. It consists of alternating layers of steel and scintillator plastic tiles, where steel is the absorber material and the scintillator the active material. The scintillators are read out via photomultiplier tubes.

4.2.4. Muon Spectrometer

Muons can penetrate both the inner detector and the calorimeters without loosing much energy. Therefore, the muon spectrometer is a detector system specifically designed for the measurement of the momentum and the charge of muons [73]. It is located around the calorimeters and covers a pseudorapidity range $|\eta| < 2.7$. The muon spectrometer uses four different types of detectors, which are monitored drift tubes (MDTs), cathode-strip chambers (CSCs), resistive plate chambers (RPCs) and thin gap chambers (TGCs). A schematic representation of the muon system and the placement of the different detectors is shown in Figure 4.7. Precise momentum measurements are achieved with the MDTs, which are located in both the barrel region and the end-caps, covering the range $|\eta| < 2.7$ and achieve an average tube resolution of 80 µm. CSCs are used for precision measurements in the inner layer of the end-cap region since they offer a higher rate capability. They cover the region $2 < |\eta| < 2.7$ and provide a resolution of 40 µm in the bending plane and 5 mm in the transverse plane [61, p.165]. The bending plane is the plane perpendicular to the magnetic field in which the particle moves. In addition to these precision muon tracking chambers, RPCs are installed in the barrel region and TGCs in the end-caps for triggering. These trigger chambers are used to decide if the measurements from the precision tracking chambers are read out and provide a second coordinate measurement. This combination allows using precise, but slower detector technology in the precision muon tracking chambers.

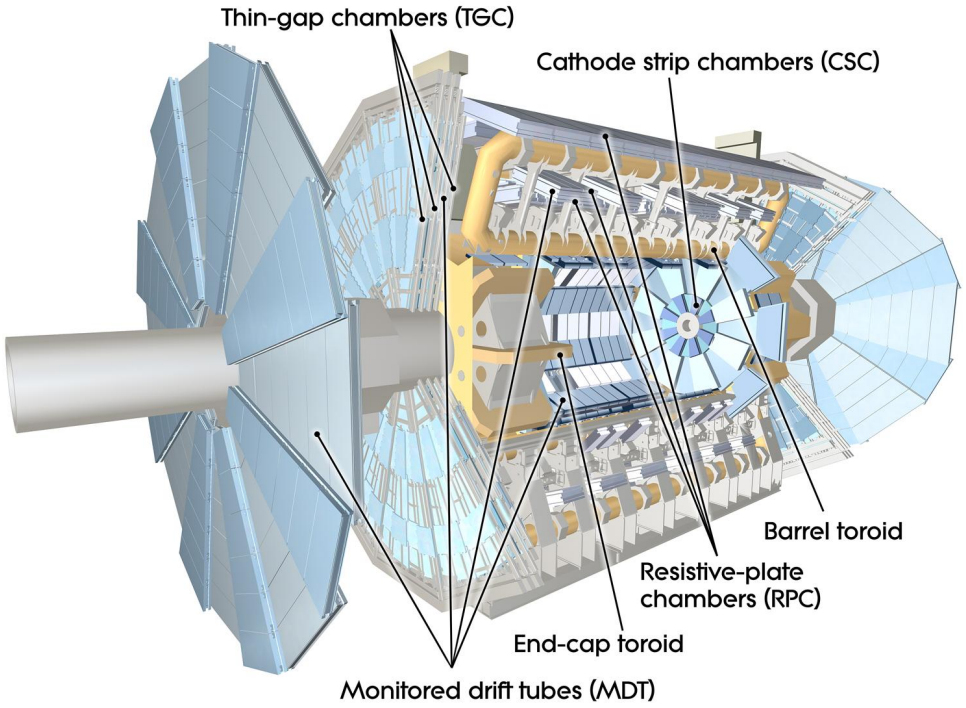


Figure 4.7.: Muon detector system of the ATLAS detector during Run I and Run II [74].

4.2.5. Magnet System

In order to obtain precise measurements of the momentum and the electric charge of charged particles, their trajectories are bent using strong magnets. The magnet system of the ATLAS detector consists of four superconducting magnets. A solenoidal magnet surrounds the inner detector and three toroidal magnets are installed in the muon spectrometer. The resulting field lines are illustrated schematically in Figure 4.8. All magnets are supercon-

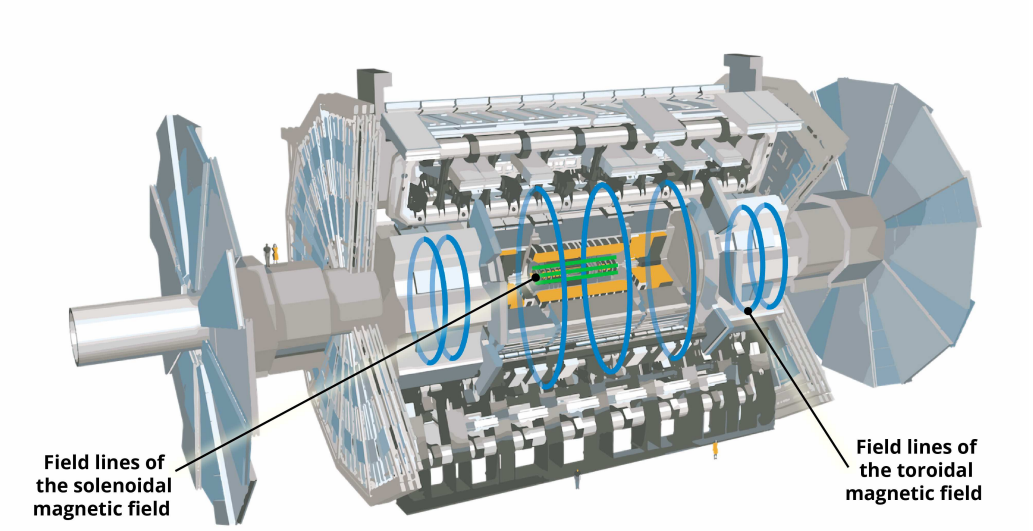


Figure 4.8.: Magnet system of the ATLAS detector [75].

ducting magnet systems which have to be cooled down to about 4.5 K. The central solenoid magnet is installed between the inner detector and the calorimeters and generates a 2 T strong homogeneous magnetic field parallel to the z-axis. Two of the toroid magnets are located in the end-caps, producing a magnetic field of approximately 1 T for the muon chambers. The third toroidal magnet is assembled around the calorimeter system and delivers a magnetic field which corresponds to a magnetic field integral between 2 Tm and 6 Tm across most of the detector [61].

4.2.6. Trigger System and Data Acquisition

With a collision rate of over 1 billion collisions per second, the recording of all events would lead to a data volume of over 60 terabytes per second. In order to reduce this data volume to a manageable amount, a sophisticated trigger system is used. The ATLAS trigger system used in Run II is shown in Figure 4.9 and consists of a Level-1 (L1) trigger system and a high-level trigger (HLT) [76, 77, 78].

The L1 trigger system is a hardware trigger which uses custom-made electronics that are installed at the detector. It uses information from the calorimeters and the muon trigger chambers to look for high transverse-momentum particles. Based on this information, it then makes a decision in less than 2.5 µs and reduces the event rate to about 75 kHz. It also provides *regions of interest* (RoI), which are regions in η and ϕ where the interesting features were found. The HLT then uses the full detector information to further analyse these RoIs. The HLT trigger is software based and reduces the event rate further to approximately 1 kHz, yielding the events that are saved for physics analyses.

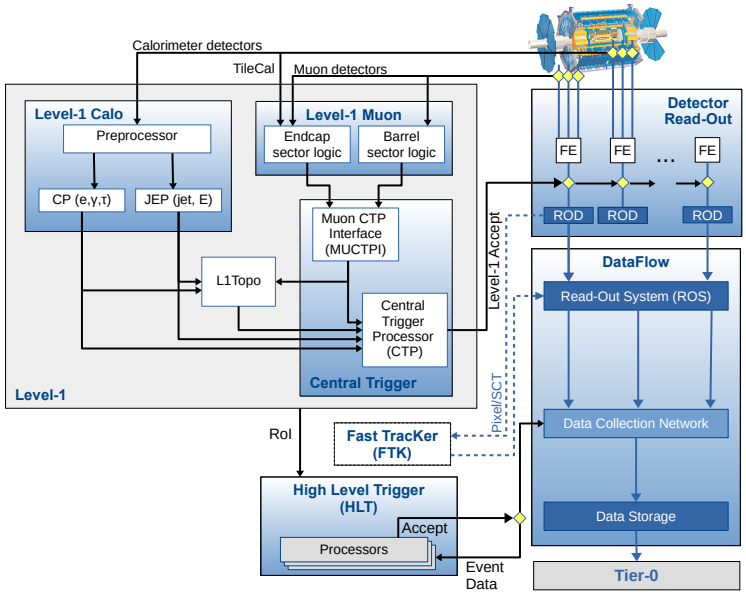


Figure 4.9.: Schematic representation of the ATLAS trigger system used in Run II [78].

5. Machine Learning with Neural Networks

The field of *machine learning* (ML) has evolved rapidly over the last decades. While most innovative methods are originally developed for natural language processing or image processing, many of those methods are adapted in high energy physics (HEP). This is due to the fact that particle physics analyses include many types of classification and regression problems, ranging from object and event reconstruction over object identification [79, 80, 81] and event selection [82, 83] to anomaly detection [84]. Detailed reviews about the impact of ML in particle physics can be found for instance in references [85, 86, 87] though there are many more. This chapter introduces the concept of machine learning, the mathematical foundation of artificial neural networks as well as the concept of Deep Sets, which represents a specific type of neural network architecture.

5.1. Core Concepts of Machine Learning

Machine learning is a subfield of artificial intelligence (AI). The goal of a ML algorithm is to fit a *model* to *training data*, such that the model can then be used to make predictions on data it has not seen before. A model is a mathematical function with many parameters that takes an input and yields an output corresponding to the given task. A famous example for such a task is the classification of handwritten digits from the MNIST dataset [88], where the model input is a 28x28 pixel image and the output indicates which digit is shown on the image. When dealing with a classification problem, the output of the algorithm is usually a vector with one entry for each of the target classes, where each entry indicates the probability that the element belongs to the corresponding class. For regression tasks, the output would be an estimate for the target function. The huge advantage of ML algorithms compared to traditional algorithms is that the algorithms itself finds a way to solve the given task whereas in non-ML algorithms the way of solving the task has to be explicitly defined in advance.

In *supervised learning*, the model is trained by showing it input features x and the correct output y. During the training process, the model then learns to make more and more accurate predictions $f_\theta(x) = \hat{y}$, where f_θ is a function with model parameters θ. This is shown schematically in Figure 5.1. The predictions of the model depend on its parameters, which are adjusted during the training. The training process itself is an iterative procedure, where the model parameters are modified in order to improve the accuracy of the model predictions. Unsupervised learning, semi-supervised learning, and reinforcement learning are additional disciplines that complement supervised learning. However, these other areas are not covered in the following since all the techniques utilised in this thesis belong to the category of supervised learning.

5.1.1. Training Process

The training of a model is accomplished by minimising the *loss function* $\mathcal{C}$ of a model, also sometimes referred to as *cost function*. This scalar function is a metric that indicates how accurate the model predictions are when compared to the truth labels of the training data. During training, the model is adjusted in such a way that the loss function is minimised. Depending on the task of the model, different definitions of the loss function have been shown to give the best training result. For regression tasks, the model is usually trained

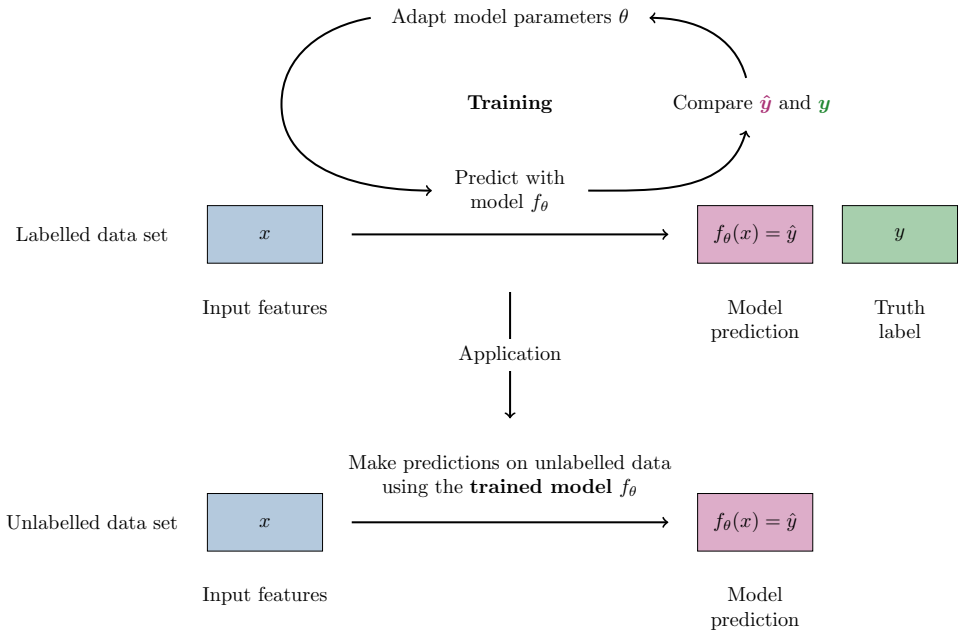


Figure 5.1.: Schematic representation of the main idea of supervised learning. By showing the model f_θ input features x and the corresponding correct output y of a labelled dataset, it learns to make predictions $\hat{y}$. The learning process is implemented by adjusting the model parameters θ such that the predictions becomes more accurate.

by minimising the mean squared error (MSE). The MSE is defined as

$$\mathcal{C}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \,, \tag{5.1}$$

where the sum over the index i represents the sum over all N items in the training dataset. For classification problems, the loss function of choice is usually the categorical cross entropy, which is defined as

$$\mathcal{C}_{\text{categorial-cross-entropy}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \cdot \log(\hat{y}_{ij}) \,. \tag{5.2}$$

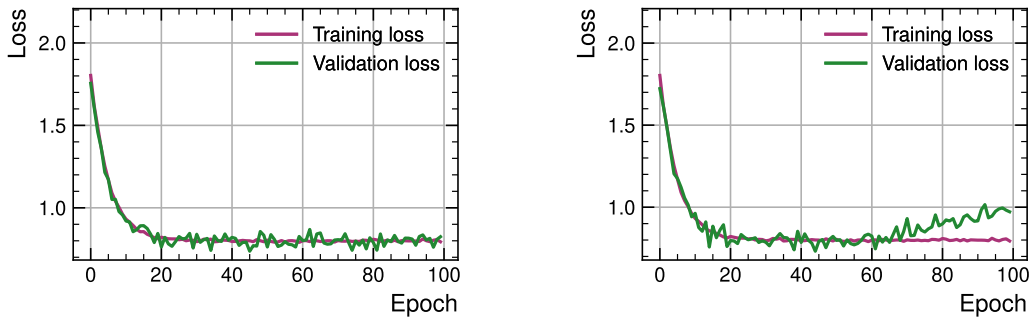
The index i runs over all N training items and the index j runs over all entries of the model output, which is a vector of length M.

Gradient Descent

If the model f_θ is a differentiable function, the loss is usually minimised using *gradient descent* (GD), which is an optimisation method that iteratively adjusts the parameters θ of the model. The idea of GD is to calculate the gradient of the loss function in the parameter space using backpropagation [89] and then update them by moving in the negative direction of the gradient. These steps are repeated for each iteration, which results in the update rule

$$\theta_{t+1} = \theta_t - \eta \, \nabla_\theta \, \mathcal{C} \,, \tag{5.3}$$

where t represents the number of the iteration and η the *learning rate* (LR). The goal is to adjust the parameters using this update rule until a suitable minimum of the loss is



(a) Example without overtraining

(b) Example with overtraining

Figure 5.2.: Illustration of overtraining. Both plots show the training (magenta) and validation (green) loss as a function of the training epoch. Figure **(a)** shows a model without overtraining, while figure **(b)** shows signs of overtraining after epoch 60, since the validation loss becomes significantly larger than the training loss, which means that the model performs much worse in the validation dataset compared to the training dataset.

reached. The size of the learning rate is a parameter that is chosen before starting the training process and should be chosen carefully. While a too large learning rate could have the effect that a suitable minimum of the loss function is leaped over, a too small learning rate reduces the training speed and the algorithm might get stuck in a shallow minimum of the loss. Furthermore, in order to allow for smaller gradient steps towards the end of the training, the learning rate can be reduced during the training process, which also allows for smoother convergence.

However, the calculation of the gradient with the whole training dataset becomes computationally expensive for large datasets. This is why the gradient is usually calculated for so-called *batches* of the training dataset. Each batch is required to be representative for the whole dataset, which can be achieved by shuffling the dataset before it is divided into the batches. A model is then trained over several *epochs*, where one epoch represents the gradient descent steps of all batches, such that the model has seen the whole training dataset. With the parameters being updated with the gradients calculated on these batches, much faster computing speed can be obtained. This version of gradient descent is called *stochastic gradient descent* (SGD).

Optimisation

While the gradient descent algorithm should not get stuck in a small local minimum, it should also not overshoot in terms of too large step sizes when reaching a suitable loss minimum in the parameter space. In addition to that, it is desired to reach the minimum in the smallest number of epochs possible. Depending on the optimisation algorithm, this is achieved by modifying the update rule from Equation 5.3 using the first or second moment of the gradients. One examples of such an algorithm is the ADAM optimiser [90].

Generalisation

In order to ensure that the model is not only performing well on the training dataset but also on data that was not used for the gradient calculation, the loss is also calculated for a labelled validation dataset. This ensures that the model generalises well, and shows comparable performance on items it was not trained on. If this is not the case, a model is *overtrained*, also sometimes referred to as *overfitted*. An example of such behaviour is illustrated in Figure 5.2.

Regularisation

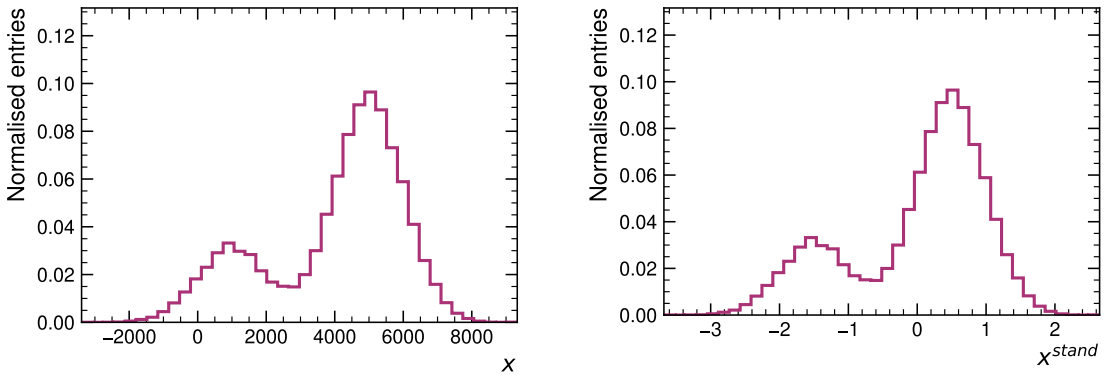
To avoid overtraining, different *regularisation* methods exist, most of which adjust the loss function by adding a penalty term which avoids individual parameters taking too large values. This reduces the complexity of the model and makes it more robust. For neural networks, there are also stochastic regularisation methods, which are discussed in section 5.2.

5.1.2. Preprocessing

Before training a model, the input data is usually preprocessed. A commonly used step in preprocessing is the *standardisation* of the input variables. The corresponding transformation is defined by

$$x_i \quad \longrightarrow \quad x_i^{\text{stand}} = \frac{x_i - \mu(x_i)}{\sigma(x_i)}\,, \tag{5.4}$$

where $\mu(x_i)$ is the mean and $\sigma(x_i)$ the standard deviation of the non-preprocessed variable x_i. The resulting distribution of the transformed variable x_i^{stand} has a mean of 0 and a standard deviation of 1. While this modifies the range of the variable, the shape of the distribution remains unchanged by this transformation, which is shown in Figure 5.3. This is done in order to avoid individual input variables dominating over others in terms of their numerical value, which would mean that different variables can dominate in the backpropagation process. While the model parameters can generally account for different ranges of different input variables, standardisation makes training with GD more stable. This is due to the fact that the range of the input features also affects the gradient step size.



(a) Before standardisation **(b)** After standardisation

Figure 5.3.: Illustration of the standardisation of a variable x. Figure **(a)** shows the initial distribution of a variable x and figure **(b)** shows the distribution after applying the standardisation transformation. The standardised distribution has mean 0 and standard deviation 1.

When a model is trained for a classification task, another important component of preprocessing is the handling of imbalanced classes. Since the loss is calculated as a sum or average over the whole batch, having unequal amounts of representative training items for different classes results in a training bias towards the dominating classes.

To overcome this issue, the training dataset can be composed by picking the same number of randomly selected items of each class from the original data set. If certain classes are

heavily underrepresented, this class can be *resampled*, adding items of this class to the training dataset multiple times. Another way of dealing with imbalanced classes is to assign larger weights to items of the underrepresented classes. However, resampling leads in general to a more stable training [91].

5.2. Neural Networks

Neural networks (NNs) are the most widely used type of machine learning algorithm, providing extraordinary performance in a large variety of applications. The fundamental idea of neural networks is to have several layers of nodes, which can be divided into the input layer, the hidden layers and the output layer as shown in Figure 5.4. Each node carries a real numerical value, called the *activation* of a node. The activation values in the input layer are the values of the input variables. The information from the input layer is then processed layer by layer, where each node is connected to all nodes of the preceding and the subsequent layer. The connection between two nodes is called *weight* and indicates how strongly a node influences the activation of the corresponding node in the next layer.

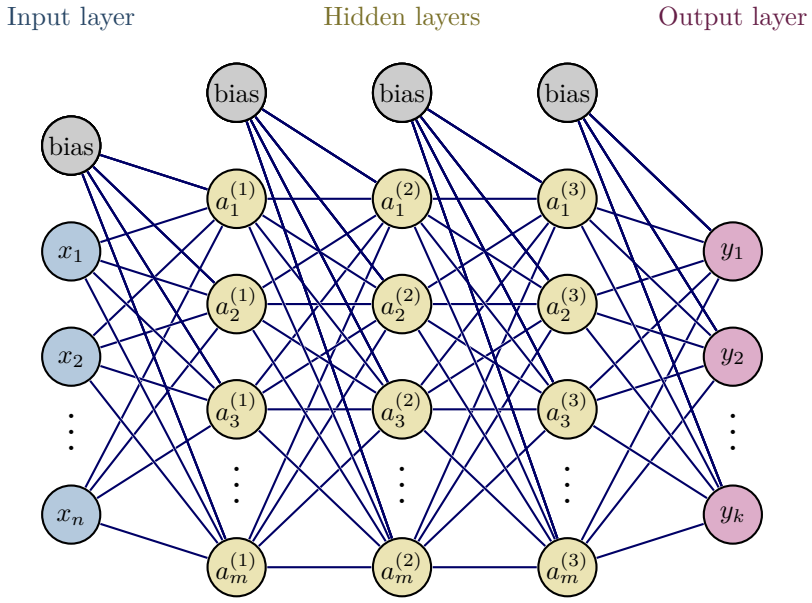


Figure 5.4.: Schematic representation of a feed-forward neural network with n input features, three hidden layers of size m and a k-dimensional output layer. The blue discs represent the input layer, the yellow discs the hidden layers and the red discs the output layer. Adapted from [92].

The activation $a_j^{(\ell)}$ of the j^{th} node in layer ℓ is then calculated by applying an *activation function* σ to the weighted sum of the activations from the previous layer. In addition to the weighted sum, each node has its assigned bias $b_j^{(\ell)}$, which is also added before the activation function is applied. With these definitions, the activation of the node j in layer ℓ is given by

$$a_j^{(\ell)} = \sigma\left(w_{j,1}^{(\ell)}\, a_1^{(\ell-1)} + w_{j,2}^{(\ell)}\, a_2^{(\ell-1)} + \ldots + w_{j,n}^{(\ell)}\, a_n^{(\ell-1)} + b_j^{(\ell)}\right) \tag{5.5}$$

$$= \sigma\left(\sum_{i=1}^{n} w_{j,i}^{(\ell)}\, a_i^{(\ell-1)} + b_j^{(\ell)}\right). \tag{5.6}$$

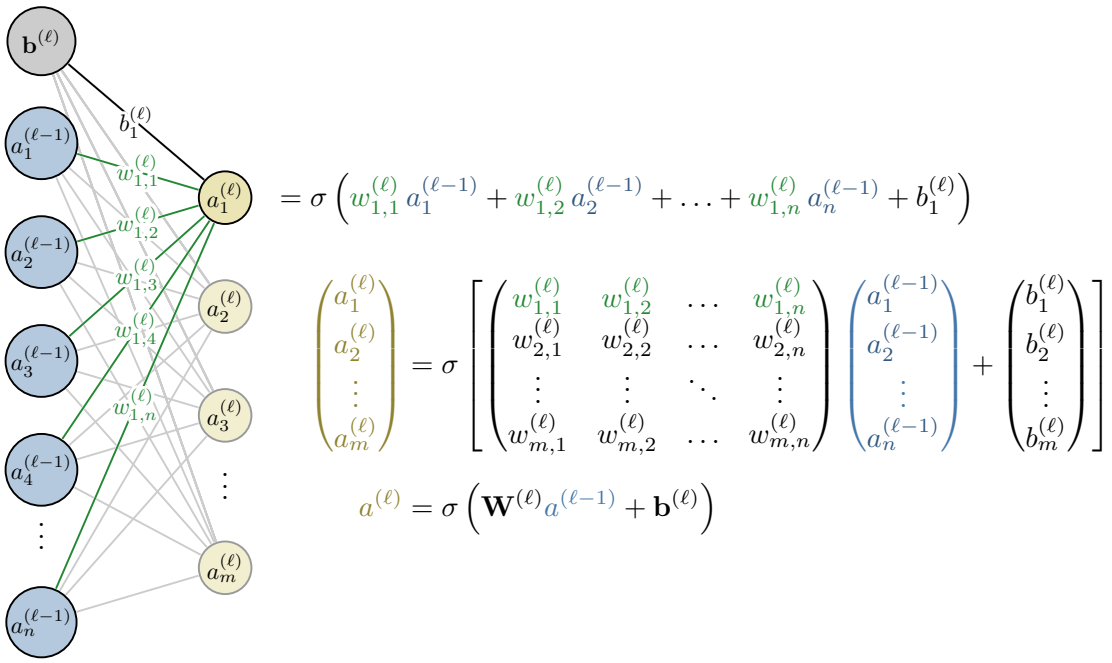


Figure 5.5.: Illustration of how the activations of a NN layer are calculated using matrix multiplication. The value of $a_j^{(\ell)}$ is the activation of the j^{th} node in layer number ℓ. The bias of this node is indicated by $b_j^{(\ell)}$. The weights between layer number $\ell - 1$ and layer ℓ are $w_{i,j}^{(\ell)}$, where the index i is the index of the node in layer $\ell - 1$ and j is the index of the node in layer ℓ. For better readability, the upper index of the weights and the bias is not shown in the graph. Adapted from [92].

By writing the weights $w_{j,i}^{(\ell)}$ between layer ℓ and layer $\ell - 1$ as a matrix $\mathbf{W}^{(\ell)}$, the activations in layer ℓ can simply be calculated using matrix multiplication. This calculation is illustrated in Figure 5.5 and can be written as

$$a^{(\ell)} = \begin{pmatrix} a_1^{(\ell)} \\ a_2^{(\ell)} \\ \vdots \\ a_m^{(\ell)} \end{pmatrix} = \sigma\left(\underbrace{\mathbf{W}^{(\ell)} a^{(\ell-1)} + \mathbf{b}^{(\ell)}}_{:=z^{(l)}}\right). \tag{5.7}$$

The specific design of the network architecture, i.e. the number and size of the layers, defines the number of trainable parameters, which are the weights and the biases of the NN. Neural networks with multiple large hidden layers are called *deep neural networks*, also referred to as *deep learning*.

Non-trainable parameters, which define the training process itself and components of the model which are not modified during the training, are called *hyperparameters*. Such hyperparameters are for example the activation functions used for the different layers, parameters of the optimiser and the used regularisation method, the number of training epochs, the size of the batches used in SGD or the learning rate. A good tuning of these parameters is essential since the hyperparameters heavily influence how successfully a model is trained. The activation function used for most layers is the rectified linear unit (ReLU)

function [93]. The definition of this function and the corresponding derivative are given by

$$\mathrm{ReLU}(z) = \begin{cases} 0 \text{ for } z < 0 \\ z \text{ for } z \geq 0 \end{cases} \tag{5.8}$$

$$\mathrm{ReLU}'(z) = \begin{cases} 0 \text{ for } z < 0 \\ 1 \text{ for } z \geq 0 \end{cases} \tag{5.9}$$

The simplicity of the ReLU function and its derivative lead to a drastic improvement in computing speed during gradient descent. However, while the ReLU activation function is usually the choice for hidden layers, other activation functions are used for the output layer. In classification tasks, the number of nodes in the output layer is usually equal to the number of target classes. The activation of these nodes is then used as an indication for how likely an item belongs to a certain class.

Two common activation functions for this use-case are the *sigmoid* function and the *softmax* function. The sigmoid function, which is defined as

$$\mathrm{Sigmoid}(z) = \frac{1}{1 + e^{-z}}\,, \tag{5.10}$$

returns values between 0 and 1 and is applied to each node separately. The softmax function on the other hand takes as input the vector z of the whole layer, where z_j is the value carried by the j^{th} node before applying the activation function as defined in Equation 5.7. The output value of the j^{th} component after applying the softmax function is then given by

$$\mathrm{Softmax}(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}\,. \tag{5.11}$$

Compared to the sigmoid function, the softmax function has the advantage that by construction the output of the softmax function sums up to 1. As a result of that, the output values can be interpreted as probabilities. Additionally, the softmax function is more sensitive to differences in the values of the vector z of a layer. This can be seen in the following example, where both the softmax and the sigmoid function are applied to a vector $z = [6, 3, -5]^T$, resulting in[1]

$$\mathrm{Softmax}\left(\begin{bmatrix} 6 \\ 3 \\ -5 \end{bmatrix}\right) = \begin{bmatrix} 0.95 \\ 0.05 \\ 0.00 \end{bmatrix} \quad \text{and} \quad \mathrm{Sigmoid}\left(\begin{bmatrix} 6 \\ 3 \\ -5 \end{bmatrix}\right) = \begin{bmatrix} 1.00 \\ 0.95 \\ 0.01 \end{bmatrix}\,. \tag{5.12}$$

In addition to the regularisation methods discussed in subsection 5.1.1, neural networks can make use of further methods like *dropout* [94] or *batch normalisation* [95].

Dropout is a regularisation method where a certain fraction of the weights of the neural network are set to zero. This means that for example in a NN layer with a dropout rate of 10 %, the dropout method randomly selects 10 % of the weights between the current and the next layer and sets them to zero. The resulting effect is that the model is forced to not rely too much on certain connections and as a result of that becomes more robust.

Batch normalisation on the other hand applies the standardisation transformation introduced in Equation 5.4 after each layer of the neural network. This means that the activation values of each batch are shifted to mean zero and re-scaled to standard deviation 1 after the activation function was applied. Batch normalisation allows using larger learning rates while also acting as a regulariser [95].

[1] The resulting values are rounded to two decimals for better readability.

5.3. Deep Sets

Deep Sets [96] are a special type of model architecture that allows to perform machine learning on tasks that are defined on sets. A set $\{x_i\}$ is defined as a collection of elements x_i where the ordering of the elements does not have any meaning. This invariance regarding the order of the elements is also referred to as *permutation invariance*. Therefore, when writing a set $\{x_i\}$ with n elements as a list, changing the order of the elements in that list does not change the overall set, which means

$$\{x_1, x_2, \ldots, x_n\} \quad \overset{\text{Swap } x_1 \text{ and } x_2}{=} \quad \{x_2, x_1, \ldots, x_n\} \quad .$$

This property is promising for applications in HEP, since measurements in particle physics often yield sets of objects, but there is no generally reasonable ordering of these objects. A schematic representation of the concept of Deep Sets is shown in Figure 5.6. It is based on the decomposition of a set function $f(\{x_i\})$ into

$$f(\{x_i\}) = \rho\left(\sum_{i=1}^{n} \phi(x_i)\right) , \tag{5.13}$$

where ϕ is a function that operates on instances x_i of the set $\{x_i\}$ and ρ operates on objects of the shape of the output of ϕ. Although the functions ϕ and ρ might be any functions that have this feature, using deep neural networks for both ϕ and ρ is a popular choice. The ϕ network is the same for all elements of the set, which means that it is one network that it is applied to all elements in parallel as illustrated in Figure 5.6, but the weights are the same.

Since the summation in Equation 5.13 is invariant under permutations of the elements of the set $\{x_i\}$, Deeps Sets offer a permutation invariant algorithm architecture. It should be noted though, that the summation operation is just one possible implementation. In general, other aggregation functions like the average or the max could be used as well.

The idea of this architecture is therefore to first transform the representation of the set instances x_i into an intermediate representation that is suitable for aggregation using the network ϕ. The aggregated output $\sum_i \phi(x_i)$ can then be used to obtain the desired output, i.e. the classification of the whole set.

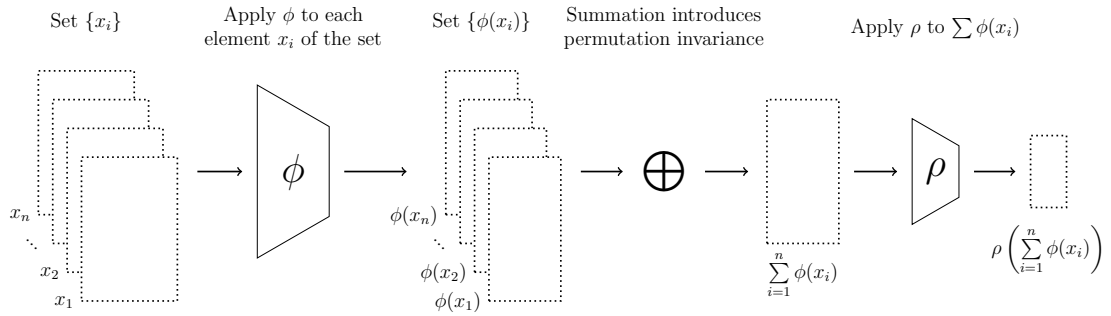


Figure 5.6.: Schematic representation of the Deep Sets architecture.

6. Flavour Tagging

Due to colour confinement, quarks cannot appear as free particles and therefore cannot be measured directly in collider experiments. Instead, the quarks hadronise in the detector and appear in the form of hadronic jets. To reconstruct the physical process that took place in the hard interaction, the type of hadrons inside the jets is identified. This technique is called *flavour tagging*, as it tries to tag the jets according to the flavour of the quarks inside these hadrons. The concept of flavour tagging and the corresponding algorithms that are studied in this thesis are introduced in this chapter.

6.1. Overview

Jets containing c-hadrons or b-hadrons have a characteristic structure that differs from the structure of light-flavour jets, which represents the comprehensive class of jets that are neither matched to c- or b-hadrons nor to τ-leptons. b-hadrons have a lifetime of around 1.5 ps [8], which is caused by the fact that the entry V_{tb} in the CKM matrix is close to 1. As a result of that, flavour oscillations of b-quarks to quarks of the first two generations are suppressed and since a transition of a b-quark to a t-quark is kinematically not possible, b-quarks are forced to decay via the suppressed off-diagonal channels of the CKM matrix. Having such a long lifetime, the b-hadron travels some distance away from the *primary vertex* (PV) before it decays itself. For a b-hadron with $p_{\mathrm{T}} = 50\,\mathrm{GeV}$, this *decay length* is approximately 5 mm. Thus, the point where the b-hadron decays represents a *secondary vertex* (SV) that is displaced by a few millimetres from the PV. Furthermore, if the b-hadron decays into a c-hadron, then the point where the c-hadron decays represents a *tertiary vertex*. An illustration of two light-flavour jets and one b-jet is shown in Figure 6.1.

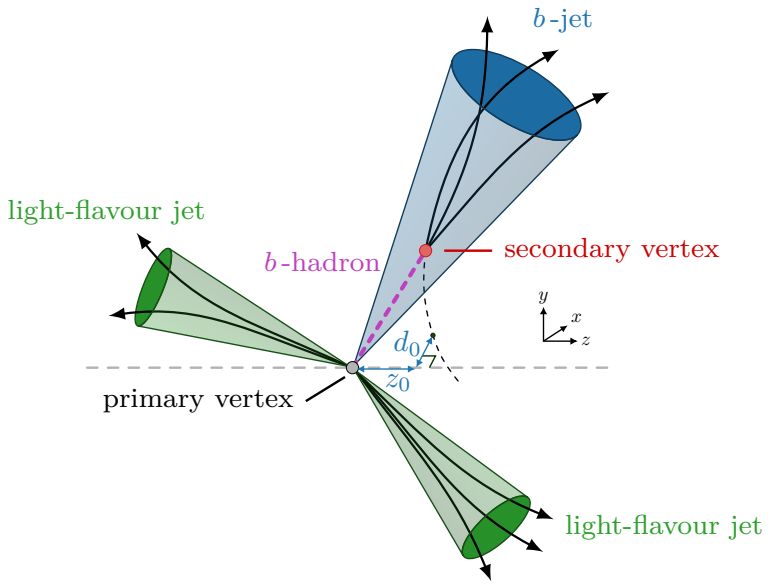
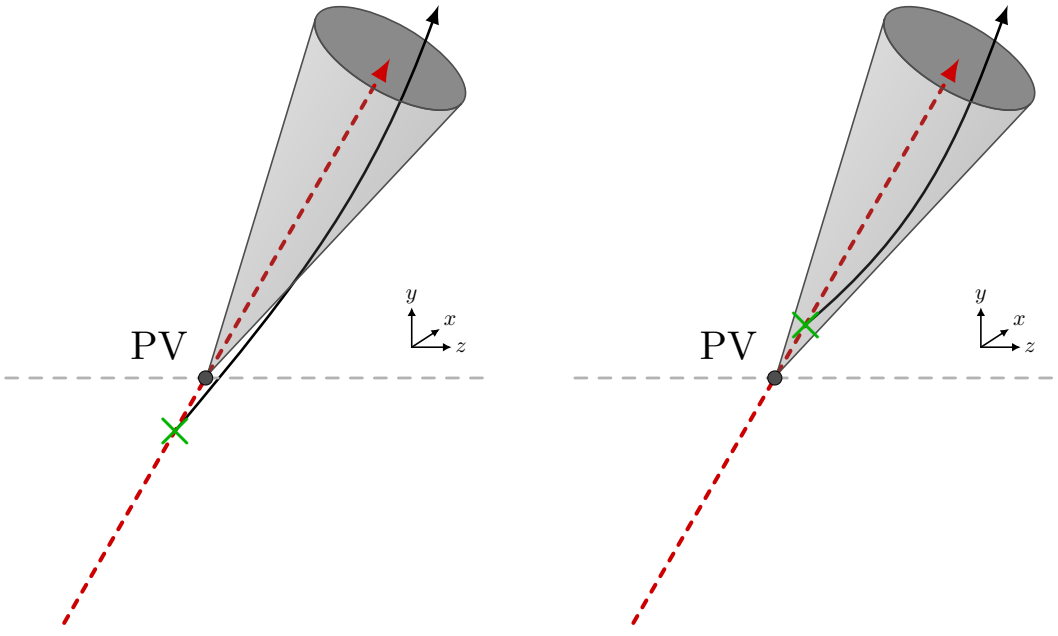


Figure 6.1.: Schematic illustration of two light-flavour jets and a b-jet. The distance travelled by the b-hadron is shown in purple and tracks are shown as solid lines with arrow heads. The transverse (longitudinal) impact parameter d_0 (z_0) is shown for one of the tracks of the b-jet. Adapted from [97].

c-jets, which are jets that contain c-hardons but no b-hadrons, have similar topologies. However, c-hadrons have shorter lifetimes than b-hadrons since c-quarks can transition into s-quarks via the large diagonal entry V_{cs} of the CKM matrix. As a result of that, the corresponding SV in a c-jet is closer to the PV than in the case of a b-jet. c-hadrons have a smaller mass than b-hadrons, which is also reflected in its decay products. Finally, since top quarks decay before hadronisation processes take place, there are no top-quark jets. Instead, top-quark decays result in a b-jet which that is accompanied by either two light-flavour jets or a charged lepton and a neutrino.

The most commonly used flavour tagging algorithms are optimised for the identification of b-jets, called *b-tagging* algorithms. These algorithms exploit variables that reflect the characteristics of b-jets mentioned above to make a prediction if a jet is likely to originate from a b-quark or not. The input variables of b-tagging algorithms are usually either variables that describe the jet as a whole, like the p_{T} of the jet and information about the SV, or information about the tracks that are found within the jet. This track information can for example be the *impact parameter* (IP), which is a measure for the shortest distance between the PV and the track of interest. The *transverse impact parameter* d_0 is defined as the shortest distance between the extended track and the z-axis, as shown in Figure 6.1. The *longitudinal impact parameter* z_0 is the z-coordinate of the track at the point where d_0 is evaluated. In this thesis, a lifetime signing convention is used for the significance of both the transverse and the longitudinal IP. If the point where the track crosses the jet axis is located upstream to the PV along the jet axis, the sign is positive, negative otherwise [98]. An illustration of the two cases is shown in Figure 6.2. Furthermore, the longitudinal IP is commonly multiplied by $\sin\theta$, where in this context θ is the difference $\Delta\phi$ in the r-ϕ plane of the ATLAS coordinate system [99].



(a) Negative IP significance sign **(b)** Positive IP significance sign

Figure 6.2.: Illustration of the lifetime signing convention that is used for the impact parameter significance. The dashed red arrow represents the jet axis and the black arrow the track of interest. The green cross is the point where the track crosses the jet axis.

6.2. ATLAS b-Tagging Algorithms

The b-tagging algorithms used in ATLAS are divided into low-level taggers and high-level taggers. As illustrated in Figure 6.3, this can be thought of as two stages of tagging, where the output of the low-level taggers is used as input for the high-level tagger, which then gives the final tagger output. The low-level algorithms are tuned to specific tasks, and are grouped in two categories.

The first category is dedicated to the reconstruction of displaced vertices, with the SV1 [100] algorithm focusing on the reconstruction of an inclusive secondary vertex and the JetFitter [101] algorithm being designed for the reconstruction of the full b- or c-hadron decay chain. The second category includes algorithms that exploit IP information. These are the IP2D and IP3D algorithms [98], as well as the RNNIP [102] and the Deep Impact Parameter Sets (DIPS) [103] algorithms. The IP2D algorithm uses only the information of the transverse IP significance d_0/σ_{d_0} while the IP3D algorithm additionally uses the longitudinal IP significance $z_0 \sin\theta/\sigma_{z_0 \sin\theta}$. The discriminant of both IP2D and IP3D is a log-likelihood ratio of the probabilities that a track corresponds to correspond to a certain class. The probabilities are calculated using template histograms that are obtained from MC simulation and the final discriminant on jet level is the sum over the log-likelihood values of all tracks associated with this jet [98]. The RNNIP algorithm uses a Recurrent Neural Network (RNN) architecture and treats jets as a sequence of tracks. The DIPS algorithm, which is the successor algorithm to RNNIP, uses the concept of Deep Sets and therefore treats jets as a set of tracks.

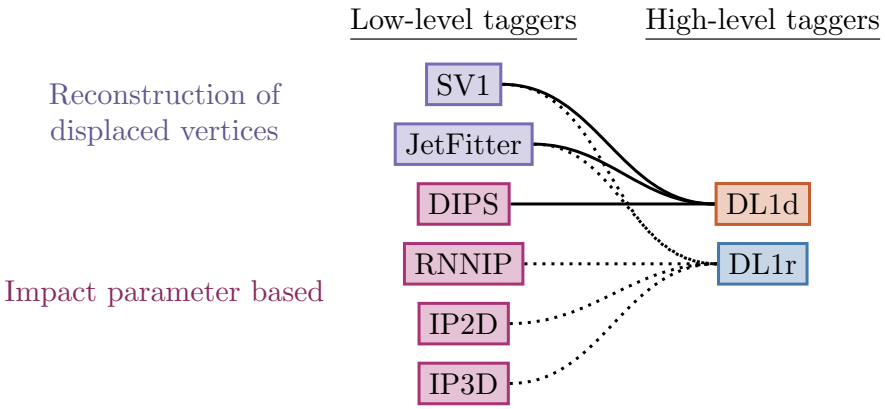


Figure 6.3.: Two-level structure of the ATLAS b-tagging algorithm. The output from the low-level taggers is used as the input for the high-level taggers. The two different high-level algorithms DL1d and DL1r both use the information of SV1 and JetFitter, but use different IP-based algorithms. While DL1r uses the information from IP2D, IP3D and RNNIP, the DL1d algorithm only exploits the information from the DIPS algorithm.

The output of the low-level algorithms and information about the jet kinematics are then fed to the high-level taggers of the DL1 family, which use a deep-neural-network architecture. Depending on the use of either the RNNIP or the DIPS algorithm, the overall tagger is called DL1r (with RNNIP) or DL1d (with DIPS). Furthermore, the DL1d algorithm does not use the information from the IP2D and IP3D algorithms, which are used in DL1r. The DL1r algorithm is used within ATLAS for analyses using data that was recorded during Run II of the LHC. The DL1d algorithm, on the other hand, is designed for analyses working with Run III data, but will also be part of a new Run II software release that aims at harmonising the measurements performed at Run II and Run III.

The algorithms are trained on jets that are extracted from MC simulated collision events. Hadrons with a transverse momentum larger than 5 GeV are matched to jets within an

Table 6.1.: Definition of the jet classes used for the algorithm training. The jet labelling is performed by matching b- and c-hadrons to the jet within a cone of size $\Delta R = 0.3$. The colours shown in the table are used throughout this thesis.

Class name	Colour	Description
b-jet		Jet that is matched to at least one b-hadron.
single-b jet		Jet that is matched to exactly one b-hadron.
bb-jet		Jet that is matched to at least two b-hadrons.
c-jet		Jet that is matched to at least one c-hadron (and not matched to a b-hadron).
light-flavour jet		Jet that is neither matched to a c- or b-hadron nor to a τ lepton.

angular distance of $\Delta R = 0.3$ around the jet axis as explained in section 3.3. An overview of the jet class definitions used in this thesis is given in Table 6.1.

The neural-network-based algorithms, i.e. RNNIP, DIPS, DL1r and DL1d, are usually trained with three jet classes namely light-flavour jets, c-jets and b-jets. Corresponding to the number of target classes, the algorithms have a three-dimensional output layer using the softmax activation function. Each output node can then be interpreted as the probability that a jet corresponds to a certain class, resulting in

$$\text{tagger output} = \begin{bmatrix} p_{\text{light}} \\ p_c \\ p_b \end{bmatrix} = \begin{bmatrix} \text{probability that this is a light-flavour jet} \\ \dots \qquad\qquad\qquad\qquad c\text{-jet} \\ \dots \qquad\qquad\qquad\qquad b\text{-jet} \end{bmatrix} . \tag{6.1}$$

The three-dimensional output is then translated into a one-dimensional discriminant which separates light-flavour jets and c-jets from b-jets. This *b-tagging discriminant* is defined as

$$D_b = \log\left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light}}}\right) , \tag{6.2}$$

where f_c is a parameter that can be tuned after the training for optimal performance, especially regarding a trade-off between light-flavour-jet rejection and c-jet rejection. With this discriminant definition, b-jets are expected to have positive D_b values while light-flavour jets and c-jets are expected to peak at negative values, as illustrated in Figure 6.4. The final result, i.e. the decision if a jet is b-tagged or not, is then defined by a cut D_b^{cut} on D_b, resulting in

$$\begin{aligned} \text{jets with } D_b > D_b^{\text{cut}} \quad &\rightarrow \quad b\text{-tagged} \\ \text{jets with } D_b \leq D_b^{\text{cut}} \quad &\rightarrow \quad \text{not } b\text{-tagged} \end{aligned}$$

An illustration of such a cut at a b-jet efficiency of 70 %, also called the 70 % *working point*, is shown in Figure 6.4 as a vertical purple line. The b-jet efficiency ϵ_b represents the fraction of b-jets whose discriminant value D_b is larger than a certain cut value

$$\epsilon_b = \frac{N_b(D_b > D_b^{\text{cut}})}{N_b(\text{total})} , \tag{6.3}$$

where $N_b(D_b > D_b^{\text{cut}})$ is the number of b-jets with a D_b value larger than the cut value and $N_b(\text{total})$ is the number of all b-jets. Furthermore, the rejection is defined as the inverse of

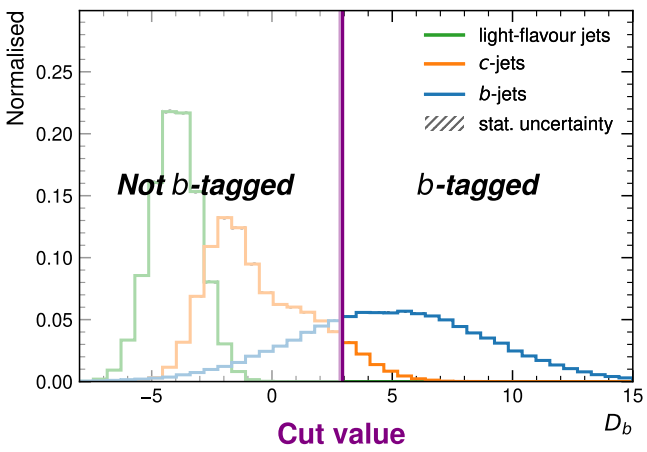


Figure 6.4.: Illustration of the cut on the b-tagging discriminant. The purple vertical line represents the cut value that corresponds to a b-jet efficiency of 70 %, also called the 70 % working point. The distributions do not represent the results of an existing b-tagging algorithm but are instead created from random numbers for illustration purposes.

the efficiency, thus for example the c-jet rejection is given by

$$r_c = \frac{1}{\epsilon_c} = \frac{N_c(\text{total})}{N_c(D_b > D_b^{\text{cut}})}\,, \tag{6.4}$$

with $N_c(D_b > D_b^{\text{cut}})$ being the number of c-jets with a D_b value larger than the cut value and $N_c(\text{total})$ is total number of c-jets. The rejection is commonly plotted as a function of the b-jet efficiency to visualise the performance of the tagger for different working points. One could also use the value of $1 - \epsilon_i$ as a measure for the amount of rejected background. However, the differences in $1 - \epsilon_i$ between two taggers are hard to see when the efficiency ϵ_i is close to 0, while the definition in Equation 6.4 is very sensitive to such comparisons.

6.2.1. DIPS

The DIPS tagger [103] is based on the concept of Deep Sets, which was introduced in section 5.3, and exploits the information of up to 40 tracks associated to the jet within a certain ΔR cone around the jet axis. The DIPS algorithm builds up on earlier applications of the Deep Sets formalism in particle physics, known as energy flow networks [104]. The input variables of DIPS are the transverse and longitudinal IP, the corresponding IP significances, the jet p_{T} fraction carried by the track, the opening angle between the jet and the track as well as several variables that represent the number of hits in certain components of the inner detector. In the hit variables, *split hits* refer to hits which are identified to be created by multiple charged particles during the track reconstruction at the pattern recognition level [105, 106]. *Shared hits* refer to hits that are shared by multiple tracks and are not already marked as split hits [107]. A full list of the variables that are used by the DIPS algorithm is shown in Table 6.2.
A schematic representation of the DIPS algorithm architecture is shown in Figure 6.5. The input consists of a set $\{x_i\}$ of up to 40 vectors, each vector containing the 15 track variables from Table 6.2. These vectors are then processed by a first neural network ϕ, where the same network is applied to each track (i.e. they are processed in parallel), resulting in a set $\{\phi(x_i)\}$, where each item $\phi(x_i)$ is a vector with 128 entries. These vectors can be thought of as a new, more abstract representation of the track information. To obtain permutation invariance, a summation is performed over all items of the set $\{\phi(x_i)\}$. This

Table 6.2.: Input variables of the DIPS algorithm.

Variable name	Description
d_0	Transverse impact parameter
$z_0 \sin\theta$	Longitudinal impact parameter multiplied by $\sin\theta$
s_{d_0}	d_0/σ_{d_0}: Transverse impact parameter significance
$s_{z_0 \sin\theta}$	$z_0 \sin\theta/\sigma_{z_0 \sin\theta}$: Longitudinal impact parameter significance
$\log p_{\mathrm{T}}^{\mathrm{frac}}$	$\log(p_{\mathrm{T}}^{\mathrm{track}}/p_{\mathrm{T}}^{\mathrm{jet}})$: Logarithm of the fraction of the jet p_{T} carried by the track
$\log \Delta R(\mathrm{track}, \mathrm{jet})$	Logarithm of the opening angle ΔR between the track and the jet axis
IBL hits	Number of hits in the IBL
PIX1 hits	Number of hits in the next-to-innermost pixel layer
Shared IBL hits	Number of shared hits in the IBL
Split IBL hits	Number of split hits in the IBL
nPixHits	Combined number of hits in the pixel layers (including IBL hits)
Shared pixel hits	Number of shared hits in the pixel layers (including shared IBL hits)
Split pixel hits	Number of split hits in the pixel layers (including split IBL hits)
nSCTHits	Combined number of hits in the SCT layers
Shared SCT hits	Number of shared hits in the SCT layers

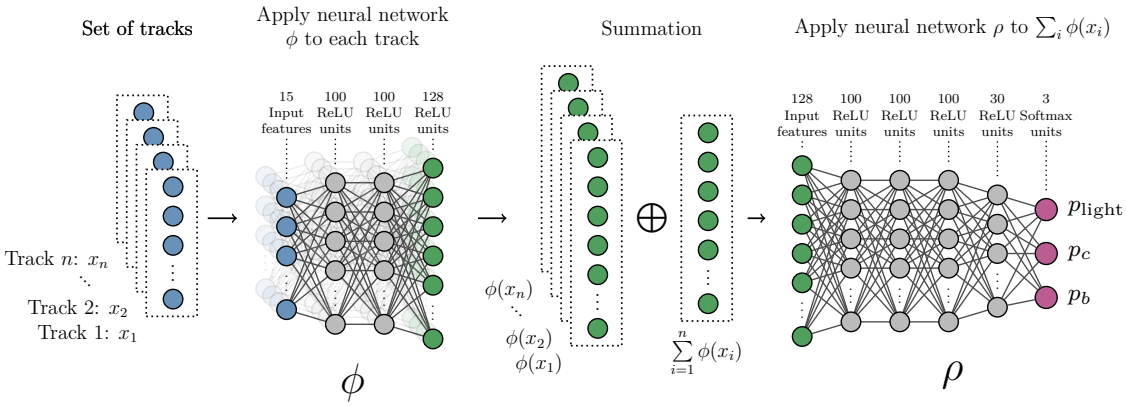


Figure 6.5.: Schematic representation of the DIPS algorithm. The number of nodes in each layer is reduced for better readability. The dotted boxes including individual layers of the neural networks are used to illustrate the transition from the set structure to the permutation-invariant representation. The input is a set of n vectors x_i, where each vector represents the information of one track. The maximum accepted number of tracks is 40. If a jet contains less than 40 tracks, the remaining entries are masked, which means that $\phi(x_i)$ of such tracks is a vector of zeros.

aggregated result from the ϕ network is then fed into the second neural network ρ, which consists of several layers. The output layer of the ρ network has the structure introduced in Equation 6.1. If a jet contains less than 40 tracks, the remaining entries are masked, which means that $\phi(x_i)$ of such tracks is a vector of zeros. Furthermore, these masked tracks do not contribute to the loss of the model.

First studies of the DIPS algorithm already showed that it outperforms the RNNIP algorithm [103]. Additionally, the use of the Deep Sets architecture is well motivated for the structure of the input, since this allows to treat the tracks as a set. In the RNNIP algorithm, on the other hand, the RNN architecture enforces the selection of an order of the input tracks. This ordering is done by sorting the tracks by their $|s_{d_0}|$ values, which

might not be the optimal ordering.

6.2.2. DL1d

The DL1d algorithm is a deep neural network, which combines the information of the low level taggers. The architecture, which is shown in Figure 6.6, consists of an input layer, eight hidden layers where the number of nodes decreases with each layer, and an output layer with one node for each target class. The input for this high-level tagger is

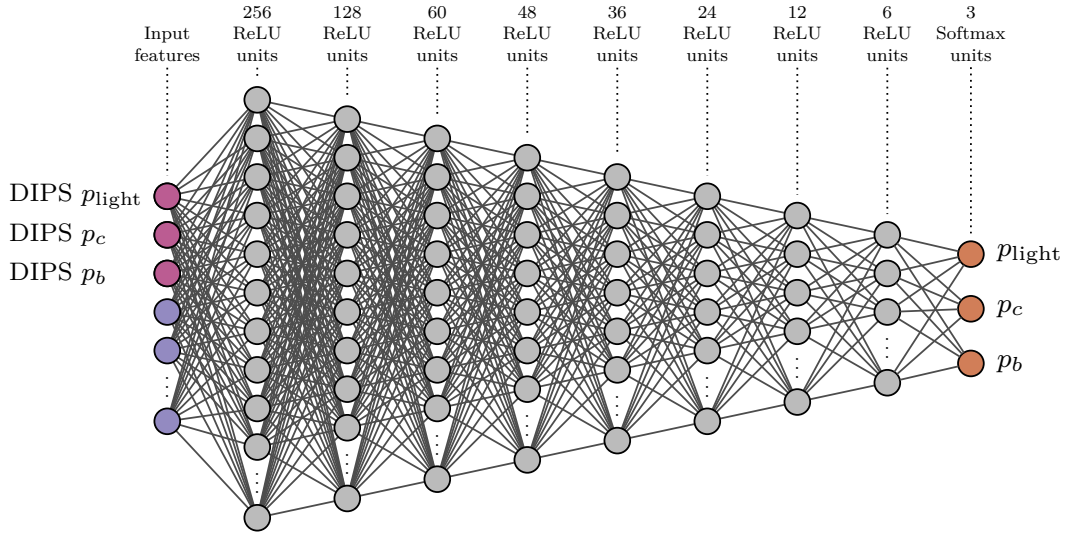


Figure 6.6.: Schematic representation of the DL1d algorithm. The number of nodes in the hidden layers decreases with each layer.

the output from the DIPS, JetFitter and SV1 algorithms, as well as the p_{T} and the pseudorapidity $|\eta|$ of the jet. A full list of the DL1d input variables, grouped by their origin, is given in Table 6.3. If the low-level taggers JetFitter or SV1 fail to reconstruct a displaced vertex, the corresponding variables in Table 6.3 are filled with default values. To take this into account in the high-level tagger, the two binary variables $\text{JF}_{\text{isDefault}}$ and $\text{SV1}_{\text{isDefault}}$ are added and indicate if the jet variables are filled with default values for the corresponding low-level tagger. The output layer of DL1d has the same shape as the DIPS output layer. This makes it possible to directly compare the predictions of the two algorithms, even though DIPS is a low-level algorithm. Given that the DL1d algorithm takes the output from DIPS as input, it is expected to further improve the prediction from the DIPS algorithm by combining it with the jet information listed in Table 6.3.

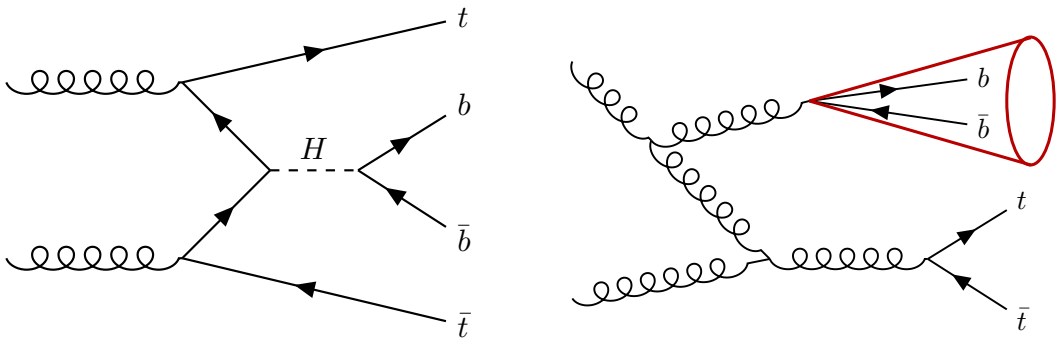
6.3. Extended Algorithms for bb-Jets

The algorithms introduced so far are usually trained with an inclusive b-jets class, which means that a jet is considered to be a b-jet if at least one b-hadron can be matched to a jet within an angular distance of $\Delta R = 0.3$. This means that no distinction is made between jets that contain one b-hadron and jets that contain two or more b-hadrons. While this inclusive b-jet class is sufficient for most use-cases, there are special cases where the number of b-hadrons inside the jet is of great interest.

One example where the number of b-hadrons inside a jet is rather interesting, is the search for the $t\bar{t}H(\rightarrow b\bar{b})$ signal, which suffers from a large irreducible $t\bar{t}+b\bar{b}$ background [3]. The Feynman diagrams of both the signal process and the $t\bar{t}+b\bar{b}$ background process are shown in Figure 6.7. The signal process is the production of a Higgs boson in association with

a top quark and an anti-top quark. The Higgs boson then decays into a $b\bar{b}$ pair, which means that the final state consists of a top quark, an anti-top quark, a bottom quark and an anti-bottom quark. In the background process, a $t\bar{t}$ pair is produced via gluon-gluon fusion with an additional gluon being radiated which splits into a $b\bar{b}$-pair. This means that the $t\bar{t} + b\bar{b}$ background process has the same final state particles as the $t\bar{t}H(\rightarrow b\bar{b})$ signal process.

Since gluons are massless particles, the $b\bar{b}$ pair which is created from gluon splitting is rather collimated, leading to a bb-jet as illustrated in Figure 6.7b. The $b\bar{b}$ pair from the decay of the Higgs boson in the signal process is expected to be less collimated, since the Higgs boson is a massive particle. Figure 6.8 shows the predicted fractions of background events with at least one single-b jet or bb-jet. Being able to identify bb-jets and to differentiate them from single-b jet, could help with the identification of background events that contain bb-jets. These events are denoted in Figure 6.8 as $t\bar{t} + B$ and make up approximately 18 % of all background events with at least one single-b jet or bb-jet.



(a) $t\bar{t}H(\rightarrow b\bar{b})$ signal **(b)** $t\bar{t} + b\bar{b}$ background

Figure 6.7.: Feynman diagrams of the $t\bar{t}H(\rightarrow b\bar{b})$ signal and the dominating background process.

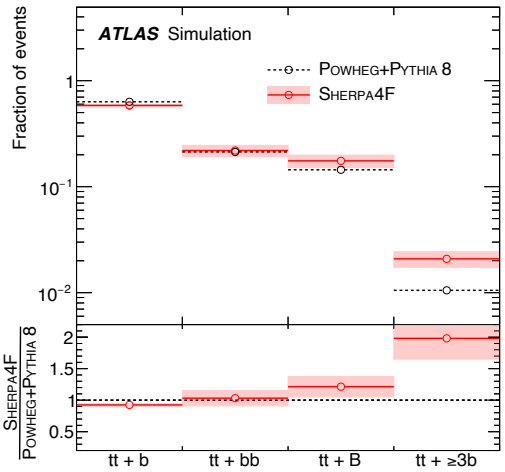


Figure 6.8.: Predicted fractions of the different classes of background events in the ATLAS $t\bar{t}H(\rightarrow b\bar{b})$ search [108]. The events are split into different categories, corresponding to the number of single-b jets and bb-jets in the event (in addition to the jets from the $t\bar{t}$ decay). $t\bar{t} + b$ refers to events with one single-b jet, $t\bar{t} + b\bar{b}$ refers to events with two single-b jets, $t\bar{t} + B$ are events with one bb-jet and $t\bar{t} + \geq 3b$ refers to all other events with at least one single-b jets or bb-jet.

In order to achieve bb-jet identification, both the DIPS and the DL1d algorithm can be extended by an additional output node, dedicated to bb-jets. The output layer of the algorithms is then given by

$$\text{tagger output} = \begin{bmatrix} p_{\text{light}} \\ p_c \\ p_{\text{single-}b} \\ p_{bb} \end{bmatrix} = \begin{bmatrix} \text{probability that this is a} & \text{light-flavour jet} \\ \cdots & c\text{-jet} \\ \cdots & \text{single-}b \text{ jet} \\ \cdots & bb\text{-jet} \end{bmatrix}, \tag{6.5}$$

which means that the output node that initially corresponded to the inclusive b-jet class is split into a single-b jet node and a bb-jet node. In the following, the extended DIPS and DL1d algorithms will be referred to as bb-DIPS and bb-DL1d, respectively.

The ideal outcome of this extension is an extended b-tagging algorithm, which can still be used for b-tagging with the typical definition of b-jets, but at the same time is able to identify bb-jets. The discriminant for b-tagging is then

$$D_b = \log\left(\frac{p_{\text{single-}b}}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light}}}\right), \tag{6.6}$$

where instead of the probability p_b which corresponds to the inclusive b-jet class the single-b-jet probability $p_{\text{single-}b}$ is used. Studies regarding the inclusion of p_{bb} in the D_b definition from Equation 6.6 are presented in section 8.3.

The discriminant that is used to separate bb-jets from single-b jets can be defined as

$$D_{\text{single-}b} = \log\left(\frac{p_{\text{single-}b}}{p_{bb}}\right). \tag{6.7}$$

Both p_{light} and p_c are not included in the definition of the single-b-jet discriminant $D_{\text{single-}b}$, since this discriminant is dedicated to separating single-b jets from bb-jets only.

This approach of training an algorithm that is capable of both b-tagging and the additional identification of bb-jets differs from other efforts like the DeXTer tagger [109], which exploits jets that are re-clustered with a large radius parameter and omits the c-jet category.

Table 6.3.: Input variables of the DL1d algorithm.

Origin	Variable name	Description
Jet kinematics	p_{T}	Jet p_{T}
	$\lvert\eta\rvert$	Jet $\lvert\eta\rvert$
JetFitter	$m(\mathrm{JF})$	Invariant mass of tracks from displaced vertices
	$f_E(\mathrm{JF})$	Jet energy fraction of the tracks associated with the displaced vertices
	$S_{xyz}(\mathrm{JF})$	Significance of the average distance between PV and displaced vertices
	$N_{\geq 2\text{-trk vertices}}(\mathrm{JF})$	Number of vertices with more than one track
	$N_{1\text{-trk vertices}}(\mathrm{JF})$	Number of single-track vertices
	$N_{\mathrm{TrkAtVtx}}(\mathrm{JF})$	Number of tracks from multi-prong displaced vertices
	$N_{2\mathrm{TrkVtx}}(\mathrm{JF})$	Number of two-track vertex candidates (prior to decay chain fit)
	$\Delta R(\vec{p}_{\mathrm{jet}}, \vec{p}_{\mathrm{vtx}})(\mathrm{JF})$	ΔR between the jet axis and the vectorial sum of momenta of all tracks attached to displaced vertices
	$\eta_{\mathrm{trk}}^{\mathrm{min,max,avg}}(\mathrm{JF})$	Minimum, maximum and average pseudorapidity of all tracks in the jet
	$N_{\mathrm{TrkAtVtx}}(2^{\mathrm{nd}})(\mathrm{JF})$	Number of tracks associated to 2^{nd} vertex
	$m_{\mathrm{Trk}}(2^{\mathrm{nd}})(\mathrm{JF})$	Invariant mass of tracks associated to 2^{nd} vertex
	$E_{\mathrm{Trk}}(2^{\mathrm{nd}})(\mathrm{JF})$	Energy of charged tracks associated to 2^{nd} vertex
	$f_E(2^{\mathrm{nd}})(\mathrm{JF})$	Jet energy fraction of the tracks associated with the 2^{nd} vertex
	$L_{xyz}(2^{\mathrm{nd}})(\mathrm{JF})$	Distance of 2^{nd} vertex from PV
	$L_{xy}(2^{\mathrm{nd}})(\mathrm{JF})$	Transverse distance of 2^{nd} vertex from PV
	$\eta_{\mathrm{trk}}^{\mathrm{min,max,avg}}(2^{\mathrm{nd}})(\mathrm{JF})$	Minimum, maximum and average pseudorapidity of tracks at the 2^{nd} vertex
SV1	$N_{\mathrm{TrkAtVtx}}(\mathrm{SV})$	Number of tracks used in the secondary vertex
	$m(\mathrm{SV})$	Invariant mass of tracks at the secondary vertex
	$N_{2\mathrm{TrkVtx}}(\mathrm{SV})$	Number of two-track vertex candidates
	$f_E(\mathrm{SV})$	Energy fraction of the tracks associated with the secondary vertex
	$\Delta R(\vec{p}_{\mathrm{jet}}, \vec{p}_{\mathrm{vtx}})(\mathrm{SV})$	Angular distance ΔR between the jet axis and the direction of the 2^{nd} vertex from the PV
	$L_{xy}(\mathrm{SV})$	Transverse distance between PV and SV
	$L_{xyz}(\mathrm{SV})$	Distance between PV and SV
	$S_{xyz}(\mathrm{SV})$	$L_{xyz}(\mathrm{SV})$ divided by its uncertainty
DIPS	p_{light}	Light-flavour jet probability output from DIPS
	p_c	c-jet probability output from DIPS
	p_b	b-jet probability output from DIPS

7. The Extended bb-DIPS Tagger

This chapter presents the studies performed with the bb-DIPS tagger that was introduced in the previous chapter. A schematic representation of the bb-DIPS algorithm architecture is shown in Figure 7.1, where the only difference compared to the default DIPS architecture is the number of output nodes. The input variables are presented in Section 7.1, followed by an in-depth explanation of the training sample preparation in Section 7.2. Afterwards, the results of the algorithm training are shown in Section 7.3 and compared to a training of the default DIPS algorithm. Finally, optimisation studies related to the underlying track selection are shown in Section 7.4.

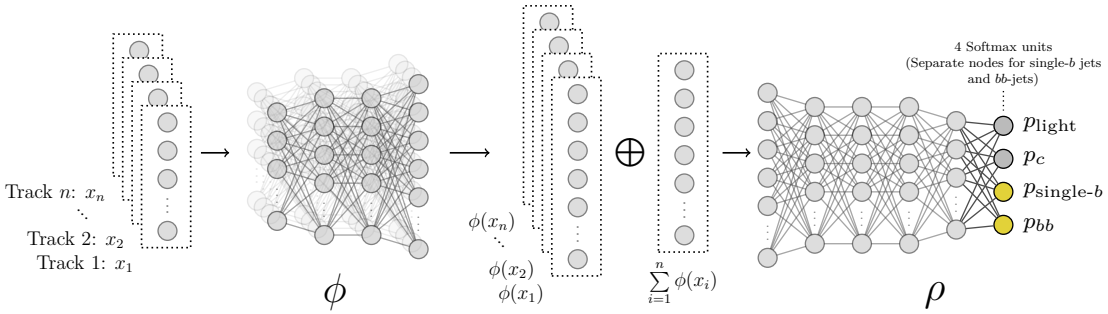


Figure 7.1.: Schematic representation of the bb-DIPS algorithm. The algorithm has the same architecture as the DIPS algorithm except for the output layer, which is four-dimensional in the bb-DIPS algorithm.

7.1. Input Variables

The training and evaluation of the bb-DIPS tagger is performed with jets that are produced in $t\bar{t}$ and Z+jets events. The samples were produced using the MC generator setup discussed in Section 3.4. The Z+jets samples are used to increase the number of bb-jets available for the training of the algorithm. An example Feynman diagram for a $Z+bb$-jet event is shown in Figure 7.2, where a gluon is radiated from an incoming quark and then splits into a $b\bar{b}$-pair. If the two b-hadrons are strongly collimated, they can be reconstructed as one single bb-jet. While light-flavour jets, c-jets and single-b jets are taken from the $t\bar{t}$ sample only,

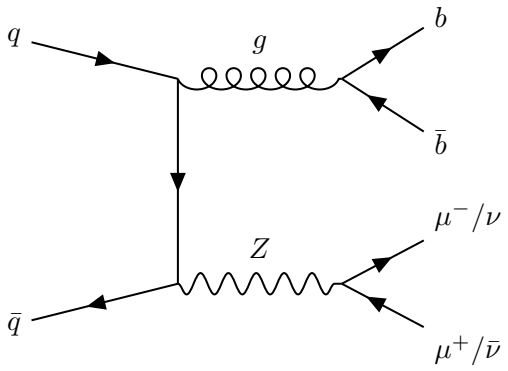


Figure 7.2.: Example Feynman diagram of a $Z + bb$-jet event where the Z boson decays into a $\mu^-\mu^+$ or $\nu\bar{\nu}$ pair.

bb-jets are taken from both samples in order to increase the size of the training sample.

Tracks are required to pass certain selection criteria. Each track is required to have a minimum p_{T} of 1 GeV and a pseudorapidity $|\eta|$ smaller than 2.5. The absolute value of the transverse and longitudinal impact parameter of the track has to be smaller than 1 mm and 1.5 mm respectively. Tracks are required to have at least 7 hits in the silicon layers (Pixel Detector and SCT combined). No more than one space point shared by multiple tracks is allowed in the silicon layers. In the pixel detector, a space point corresponds to a pixel hit. In the SCT, on the other hand, a space point is defined by two overlapping SCT strips to make it more comparable to a pixel space point. Furthermore, no more than two hits are allowed to be missing in the silicon layers combined and no more than one missing hit is allowed in the pixel layers. The selection criteria are summarised in Table 7.1 and are in the following referred to as *standard track selection*.

Table 7.1.: Definition of the standard track selection.

Variable	Criterion	Description
$p_{\mathrm{T}}^{\mathrm{track}}$	$> 1\,\mathrm{GeV}$	Minimum allowed track p_{T}
$\lvert d_0 \rvert$	$< 1\,\mathrm{mm}$	Transverse IP has to be below threshold
$\lvert z_0 \sin(\theta) \rvert$	$< 1.5\,\mathrm{mm}$	Longitudinal IP has to be below threshold
$\lvert \eta \rvert$	< 2.5	Maximum allowed track $\lvert \eta \rvert$
$N_{\text{Silicon hits}}$	≥ 7	At least 7 hits in the silicon layers (pixel and SCT)
$N_{\text{Pixel holes}}$	< 2	No more than 1 missing hit in the pixel layers
$N_{\text{Silicon shared}}$	< 2	No more than 1 space point shared by multiple tracks in silicon layers
$N_{\text{Silicon holes}}$	< 3	No more than 2 missing hits in silicon layers

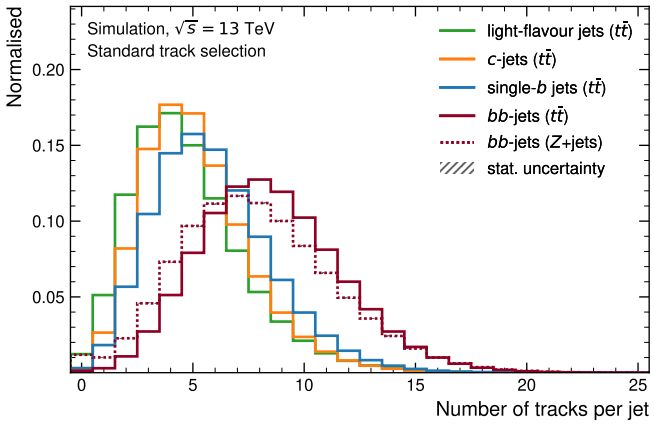


Figure 7.3.: Normalised distribution of the number of tracks per jet. Only tracks which pass the standard track selection are included in the plot.

The distribution of the number of selected tracks per jet is shown in Figure 7.3 for the different jet classes. Light-flavour jets contain on average 4.8 tracks while c-jets and single-b jets have more tracks with an average of 5.2 and 5.9 tracks per jet, respectively. By far the most tracks per jet are seen in bb-jets, with an average of 8.5 for bb-jets from $t\bar{t}$ events and 7.9 for bb-jets from Z+jets events. The p_{T} distribution of the individual jet classes is shown in Figure 7.4. In order to compare the number of tracks per jet as a function of the jet p_{T},

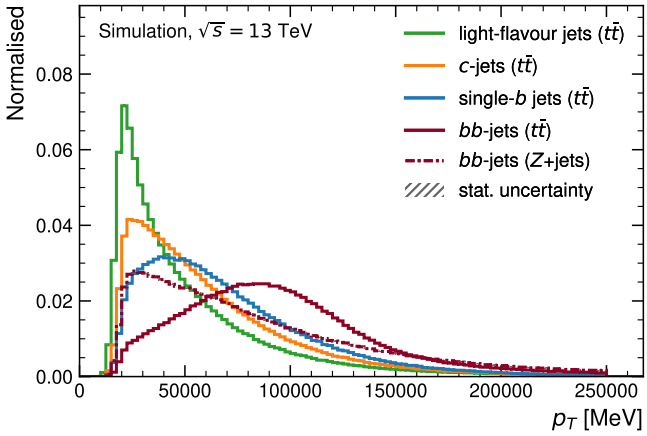


Figure 7.4.: Normalised distribution of the jet p_{T} for the different jet classes used in the training.

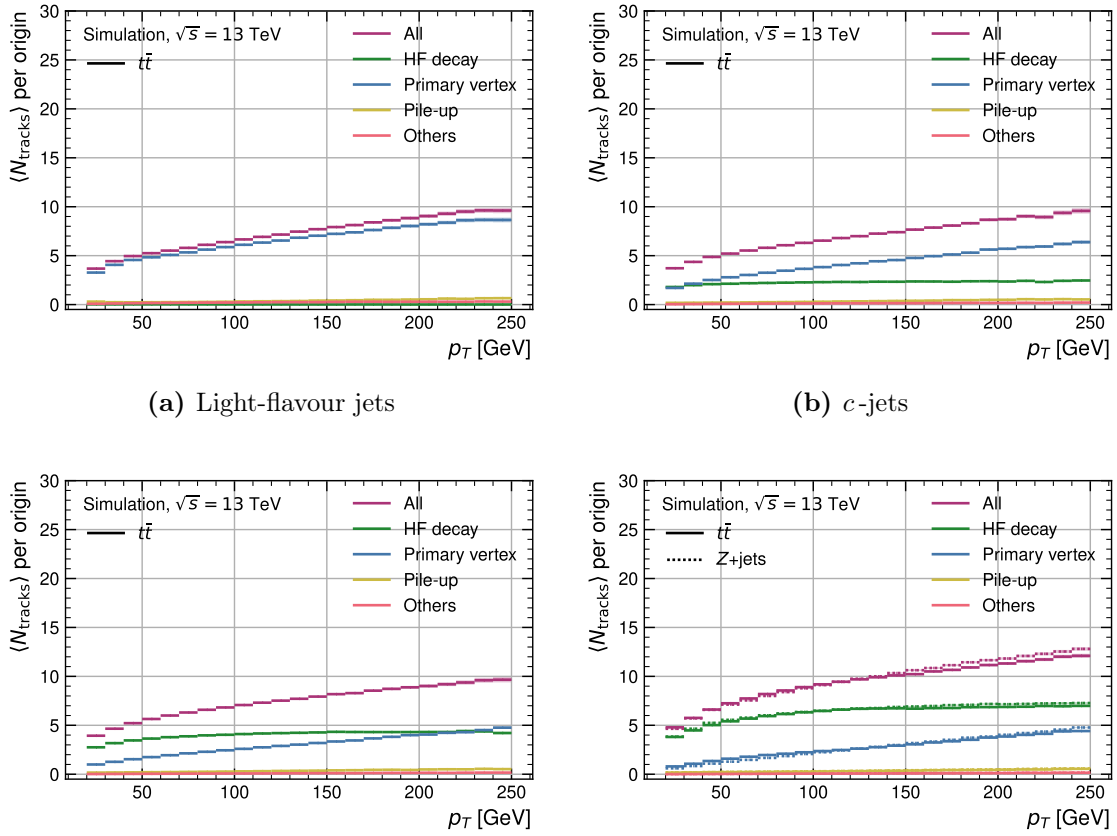
the average number of tracks is calculated in bins of p_{T}. Furthermore, this average value is also calculated for each type of origin of the tracks individually, to investigate where the tracks originate from. The corresponding plot is shown for each jet flavour separately in Figure 7.5.

The average number of tracks that originate from the primary vertex is denoted as $\langle N^{\mathrm{PV}}_{\mathrm{tracks}} \rangle$. Tracks from heavy flavour (HF) decays, denoted as $\langle N^{\mathrm{HF}}_{\mathrm{tracks}} \rangle$, are tracks that originate from the decay of a b- or c-hadron and tracks from pile-up are denoted as $\langle N^{\text{pile-up}}_{\mathrm{tracks}} \rangle$. The remaining tracks, listed as $\langle N^{\mathrm{other}}_{\mathrm{tracks}} \rangle$ are tracks from other secondary interactions, from the decay of τ-leptons and tracks that are created from the hits of multiple particles. The average number of tracks for each jet category and each track origin, calculated over the full range of the jet p_{T}, is listed in Table 7.2.

As shown in Figure 7.5, the number of tracks increases with the jet p_{T} for tracks of all categories except from a small decrease in the number of pile-up tracks in light-flavour jets from $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ to $p_{\mathrm{T}} = 50\,\mathrm{GeV}$.

Table 7.2.: Average number of tracks per jet corresponding to the different jet-flavour categories. The numbers are obtained with the standard track selection. The statistical uncertainty of these values are negligible, and thus not shown in the table. The columns represent the average number of tracks from the decay of a heavy-flavour hadron $\langle N^{\mathrm{HF}}_{\mathrm{tracks}} \rangle$, from the primary vertex $\langle N^{\mathrm{PV}}_{\mathrm{tracks}} \rangle$, from pile-up $\langle N^{\text{pile-up}}_{\mathrm{tracks}} \rangle$, and the average number of tracks that originate from other secondary interactions, the decay of a τ-lepton or which are created from the hits of multiple particles $\langle N^{\mathrm{other}}_{\mathrm{tracks}} \rangle$.

Jet flavour	$\langle N_{\mathrm{tracks}} \rangle$	$\langle N^{\mathrm{HF}}_{\mathrm{tracks}} \rangle$	$\langle N^{\mathrm{PV}}_{\mathrm{tracks}} \rangle$	$\langle N^{\text{pile-up}}_{\mathrm{tracks}} \rangle$	$\langle N^{\mathrm{other}}_{\mathrm{tracks}} \rangle$
light-flavour jets ($t\bar{t}$)	4.8	0.02	4.4	0.3	0.1
c-jets ($t\bar{t}$)	5.2	2.1	2.8	0.2	0.1
single-b jets ($t\bar{t}$)	5.9	3.6	2.0	0.2	0.05
bb-jets ($t\bar{t}$)	8.5	6.0	2.2	0.3	0.06
bb-jets (Z+jets)	7.9	5.7	1.8	0.3	0.06



(a) Light-flavour jets

(b) c-jets

(c) Single-b jets

(d) bb-jets

Figure 7.5.: Average number of tracks as a function of the jet p_{T} for different track origins individually. The statistical uncertainty of each bin is indicated as a coloured band, but is too small to be seen in most bins. The inclusive class (all) represents the total average number of tracks per jet for each p_{T} bin.

The overall number of tracks per jet increases almost linearly with the jet p_{T}. The number of tracks from pile-up is very small for all jet categories, with the average over the full p_{T} range being either 0.2 or 0.3. Moreover, with $\langle N_{\text{tracks}}^{\text{other}} \rangle = 0.1$ for light-flavour jets and c-jets and $\langle N_{\text{tracks}}^{\text{other}} \rangle = 0.05$ for single-b jets and $\langle N_{\text{tracks}}^{\text{other}} \rangle = 0.06$ for bb-jets, the number of tracks from other secondary interactions, from the decay of τ-leptons, or tracks which are created from the hits of multiple particles, is very small as well.

As expected, light-flavour jets contain almost no tracks from HF decays for the full p_{T} range. The average number of tracks from HF decays in light-flavour jets, calculated over the full p_{T} range, is $\langle N_{\text{tracks}}^{\text{HF}} \rangle = 0.02$. The very few tracks from HF decays that are assigned to light-flavour jets can originate from HF hadrons that have a p_{T} smaller than 5 GeV, which is the minimum hadron p_{T} requirement in the jet-hadron matching. Consequently, the tracks in light-flavour jets are predominantly from the PV.

Like in light-flavour jets, the dominating track category in c-jets is also the category of tracks from the PV. However, with $\langle N_{\text{tracks}}^{\text{PV}} \rangle = 2.8$, the average number of tracks from the PV is smaller in c-jets compared to light-flavour jets, where the average over the full p_{T} range is $\langle N_{\text{tracks}}^{\text{PV}} \rangle = 4.4$. The average number of tracks from HF decays in c-jets is almost constant over the full p_{T} range with a value slightly above 2.

Single-b jets and bb-jets are dominated by tracks from HF decays. Also, the average

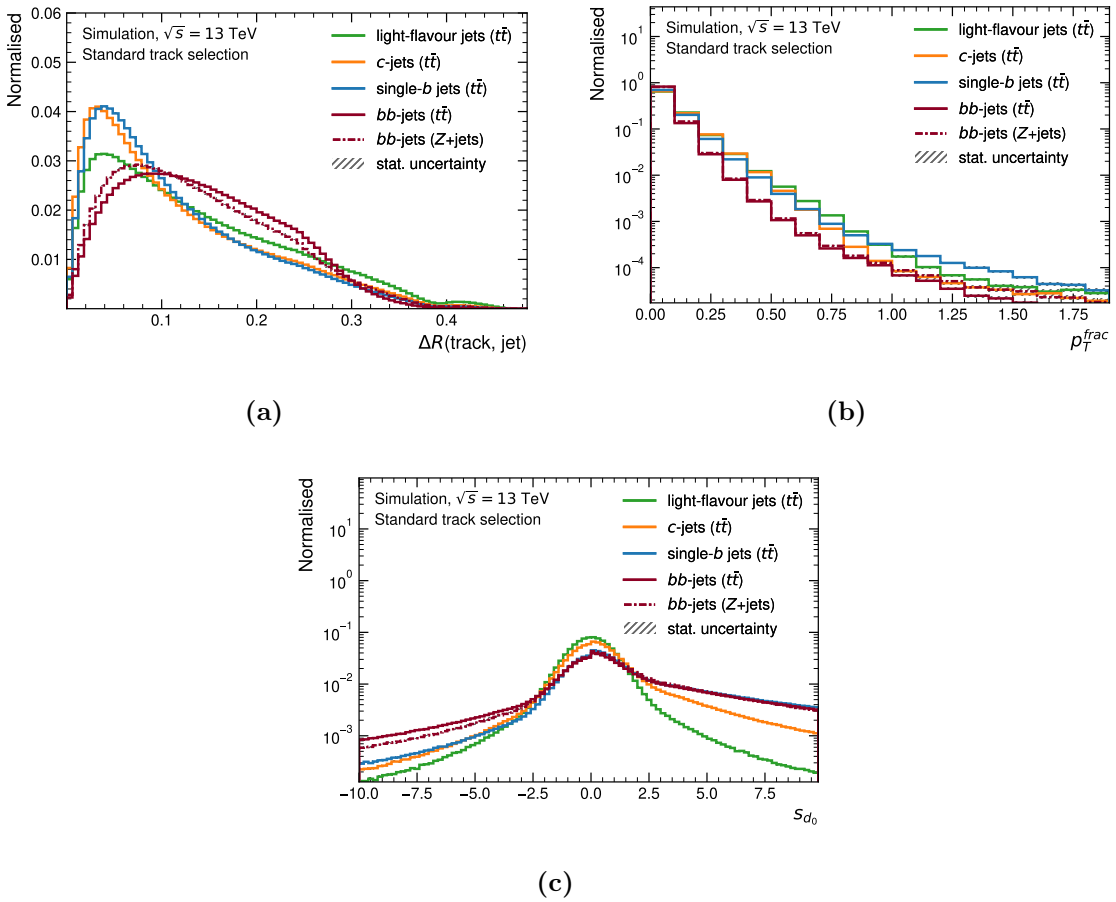


Figure 7.6.: Normalised distributions of three track variables that are used by the DIPS tagger: **(a)** the angular distance between the jet axis and the track axis ,ΔR(track, jet), **(b)** the fraction of the jet p_{T} carried by the track ,$p_{\mathrm{T}}^{\mathrm{frac}}$, and **(c)** the lifetime-signed significance of the transverse IP, s_{d_0}. The DIPS algorithm uses the logarithm of ΔR(track, jet) and $p_{\mathrm{T}}^{\mathrm{frac}}$, but to show the values of the underlying physical quantity, the distributions are shown here before the logarithmic function is applied to the values.

number of tracks from this origin is nearly constant for a jet p_{T} above 100 GeV. bb-jets from $t\bar{t}$ and Z+jets events contain on average 6 and 5.7 tracks from HF decays respectively, while single-b jets contain on average only around 3.6 tracks from HF decays. This large difference is expected since bb-jets contain an additional b-hadron. The fact that the number of tracks from HF decays in bb-jets is not exactly twice the size of the corresponding number in single-b jets indicates that not all tracks of the two b-hadrons inside the bb-jets are assigned to the jet. The number of tracks from the PV is nearly identical for both single-b jets and bb-jets over the full p_{T} range. Due to the overall p_{T} dependence of the number of tracks per jet, the bb-jets from the Z+jets sample have on average a smaller number of tracks, since these bb-jets have on average a smaller p_{T} than the bb-jets from the $t\bar{t}$ sample, as can be seen in Figure 7.4.

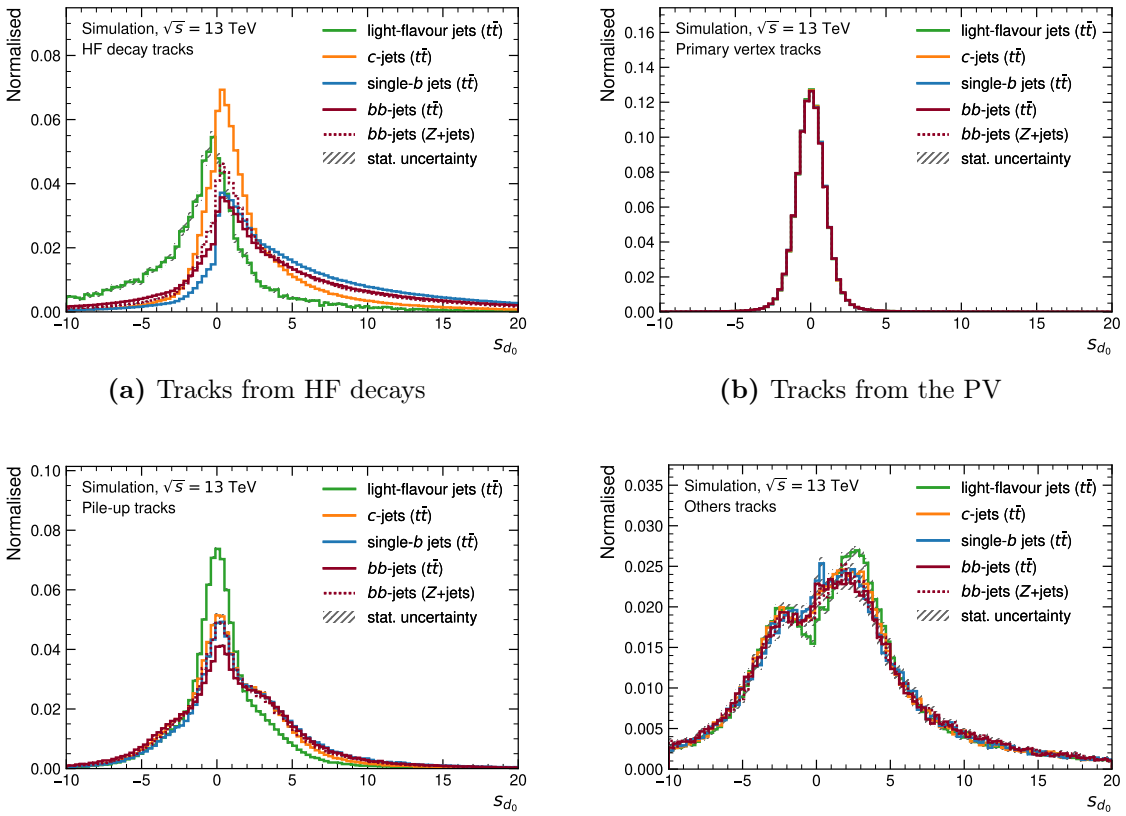
Three of the track variables, which are used as input features for the DIPS algorithm, are shown in Figure 7.6. The plots show the angular distance ΔR(track, jet) between the track and the jet axis, the variable $p_{\mathrm{T}}^{\mathrm{frac}}$, which is the fraction of the jet p_{T} carried by the track and the lifetime signed significance of the transverse impact parameter s_{d_0}. Furthermore, the distributions of all variables used by DIPS algorithm are shown in Figure 7.8.

The angular distance between the jet and the track, which is shown in Figure 7.6a, has a much larger mean value for bb-jets compared to the other jet classes. The reason for this are

the tracks from HF decays, of which the bb-jets have more than jets of the other categories. Additionally, since bb-jets contain at least two b-hadrons, the jet axis is usually between the direction of flight of the two b-hadrons. As a result of that, the angular distance between the tracks from the decay of these hadrons tends to be larger than the angular distance between tracks from HF decays and the jet axis in c-jets and single-b jets.

The $p_{\mathrm{T}}^{\mathrm{frac}}$ distribution in Figure 7.6b shows that tracks in bb-jets carry on average a smaller fraction of the jet p_{T} compared to the other classes. This might be due to the fact that the jet p_{T} is distributed to more particles. Furthermore, it should be noted that this variable only compares the scalar values of the jet p_{T} and the track p_{T}. Therefore, values larger than 1 are possible for $p_{\mathrm{T}}^{\mathrm{frac}}$, since only the vector sum of the transverse momentum of the jet constituents has to add up to the overall p_{T} of the jet. For both ΔR(track, jet) and $p_{\mathrm{T}}^{\mathrm{frac}}$, the natural logarithm is applied to the values before using them as input for the algorithm. This is done because these variables have long tail towards larger values and transforming the distribution with the logarithmic function improves the convergence time of the training [103].

In order to explain the different shapes of the s_{d_0} distributions in Figure 7.6c, the s_{d_0} distributions are shown in Figure 7.7 separately for tracks from the different categories introduced before. As can be seen in Figure 7.7b, tracks from the PV have the same s_{d_0} distribution for all jet flavours, with a narrow peak centred at $s_{d_0} = 0$. This is expected, since by definition tracks from the PV are expected to cross the jet axis close to the PV and thus have a small IP. Since the tracks in light-flavour jets are almost exclusively tracks from the PV, the overall distribution of s_{d_0} in Figure 7.6c for light-flavour jets is almost the same as the distribution drawn for tracks from the PV only. In contrast, tracks from HF



(a) Tracks from HF decays

(b) Tracks from the PV

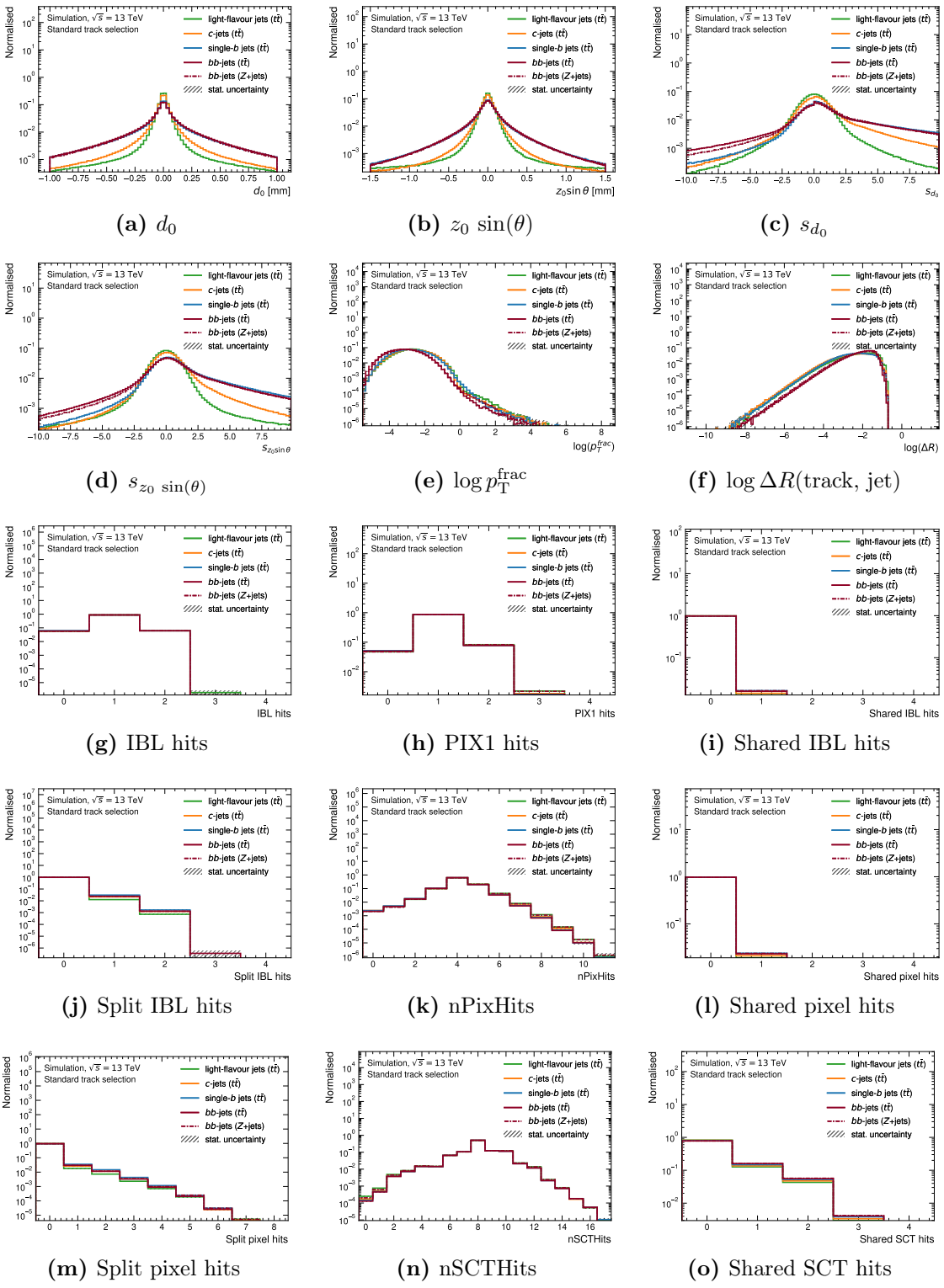
(c) Tracks from pile-up

(d) Other tracks

Figure 7.7.: The distributions of the lifetime signed significance of the transverse IP for tracks of the different categories: **(a)** tracks from HF decays, **(b)** tracks from the PV, **(c)** tracks from pile-up and **(d)** the remaining tracks.

decays predominantly cross the jet axis in front of the PV, corresponding to a positive sign of s_{d_0}. The corresponding distributions in Figure 7.7a for c-jets, single-b jets and bb-jets therefore shows a clear step at $s_{d_0} = 0$ and long tails towards larger values. As listed in Table 7.2, bb-jets have on average nearly twice as many tracks from HF decays as single-b jets. However, the additional tracks seen in bb-jets have a smaller tendency towards positive lifetime signs. The normalised distribution of s_{d_0} for bb-jets in Figure 7.6c therefore shows a larger tail towards negative values. Since the majority of the tracks in single-b jets and bb-jets are tracks from HF decays, the shape of the s_{d_0} distribution in Figure 7.7a dominates the shape of the overall distribution for these jet classes in Figure 7.6c. As the number of tracks from pile-up and the remaining tracks from the "others" category only make up a very small fraction of the tracks in all jets, the corresponding s_{d_0} distributions of these tracks, shown in Figure 7.7c and Figure 7.7d, do not have a large impact on the inclusive distributions in Figure 7.6c.

The differences in the distributions of the input variables between the different jet classes and further hidden feature correlations that are not directly visible in the distributions can be used by the DIPS algorithm to identify the flavour of the jets.



(a) d_0 **(b)** $z_0 \sin(\theta)$ **(c)** s_{d_0}

(d) $s_{z_0 \sin(\theta)}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R(\text{track, jet})$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure 7.8.: Distributions of the DIPS input variables, all of which are track variables. These distributions are obtained with the standard track selection. The variable definitions can be found in Table 6.2.

7.2. Dataset Preprocessing

Before starting the training of either the DIPS or the bb-DIPS algorithm, the sample is split into training, validation and test samples. The split is performed using the event number N_{event} of the event each jet was extracted from. The event number is a unique integer number on event-level within each sample and is not correlated to any physical observable. The assignment is then given by:

$$\begin{array}{llll} \text{jets from events with} & N_{\text{event}} \bmod 6 \leq 3 & \rightarrow \text{ used for} & \text{training} \\ \text{...} & N_{\text{event}} \bmod 6 = 4 & \rightarrow \text{ ...} & \text{validation} \\ \text{...} & N_{\text{event}} \bmod 6 = 5 & \rightarrow \text{ ...} & \text{testing} \end{array}$$

This splitting ensures that the different samples used for the training, validation and testing are orthogonal to each other, which is crucial to test the model for generalisation.

7.2.1. Training Sample

The training sample that is obtained with the aforementioned selection based on the event number associated to the jet consists of highly imbalanced classes, with bb-jets making up only 0.6 % of all jets in the sample. Training the bb-DIPS algorithm with such highly imbalanced classes would result in a large bias towards the dominating classes. Most notably, the algorithm would presumably completely ignore the minority class of the bb-jets, since the loss from Equation 5.2 is calculated via the average over the training sample and thus the average performance could be good even if the algorithm would perform poorly on bb-jets. Furthermore, the initial kinematic distributions (p_{T} and $|\eta|$) are different between the jet classes, which could introduce a bias towards certain kinematic regions.

Resampling

To solve these problems, the training sample is further processed by applying a resampling technique. The first goal of this method is to end up with equal kinematic distributions for jets of all classes. The second goal is to obtain the same number of jets of each class. An illustration of the resampling procedure in the case of 4 jet classes is shown in Figure 7.9. The procedure starts by calculating the target distribution in the p_{T}-$|\eta|$ plane. For trainings with 3 classes, this target distribution is the p_{T}-$|\eta|$ distribution of the b-jets while for trainings with 4 classes the p_{T}-$|\eta|$ distribution of the single-b jets is chosen as the target.

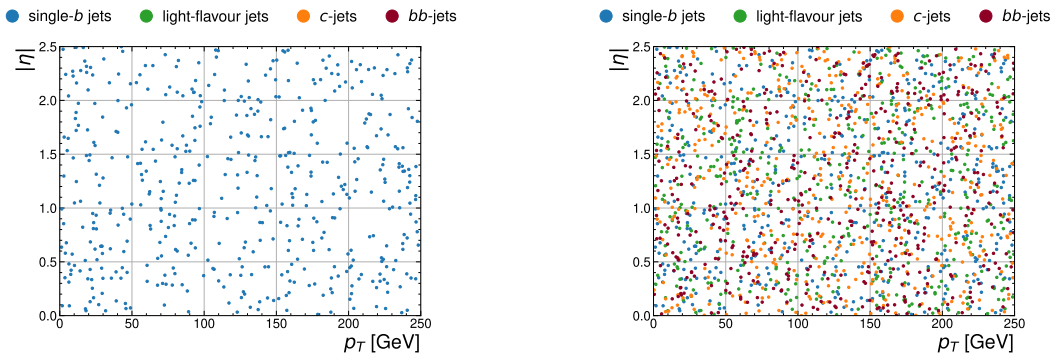
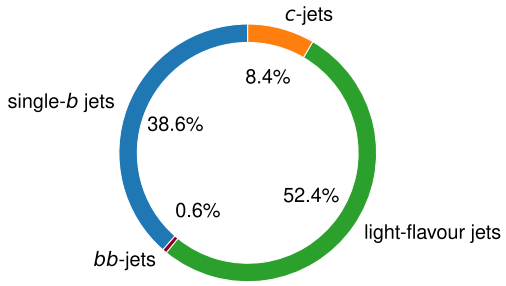


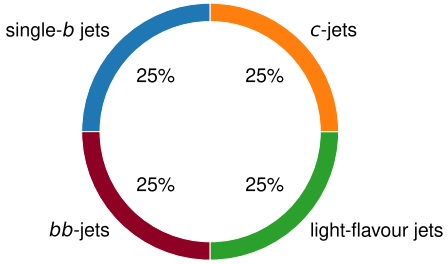
Figure 7.9.: Schematic illustration of the two-dimensional resampling in the case of four jet classes. The target distribution is calculated using the single-b jets **(a)** and then the other classes are sampled to match the p_{T}-$|\eta|$ distribution of the single-b jets.

Afterwards, jets are sampled from each class with an acceptance rate according to the target distribution, and the process is repeated until a certain number of jets N_{target} is obtained. Jets are therefore allowed to be selected multiple times. The number N_{target} is set to 4 million, which results in 12 million jets for trainings with 3 classes (DIPS) and 16 million jets for trainings with 4 classes (bb-DIPS). With around 2.3 million bb-jets, these jets are on average oversampled by a factor of around 1.7. Furthermore, 212 million light-flavour jets, 34 million c-jets and 156 million single-b jets are available for training, which means that jets from these classes are rarely oversampled.

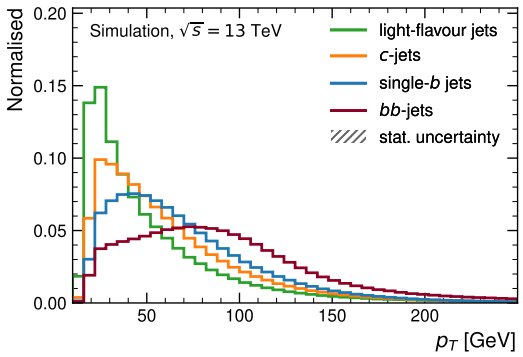
It should be noted that this means that the trainings performed with these samples do not exploit the full statistics of the non-bb-jet classes, since only a small fraction of the



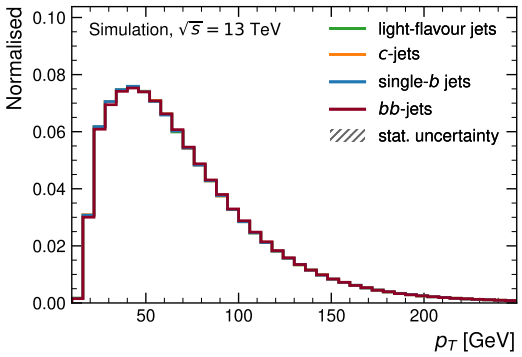
(a) Class composition before resampling



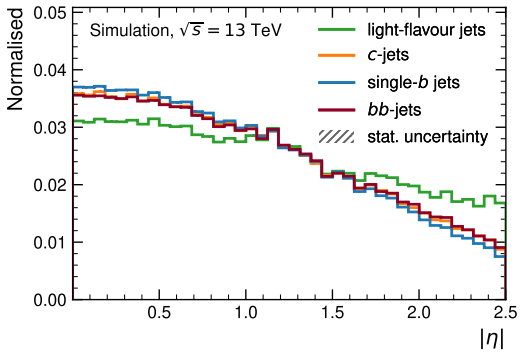
(b) Class composition after resampling



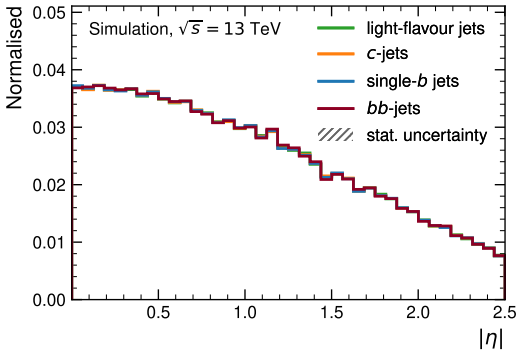
(c) Jet p_{T} before resampling



(d) Jet p_{T} after resampling



(e) Jet $|\eta|$ before resampling



(f) Jet $|\eta|$ after resampling

Figure 7.10.: Illustration of the resampling performed for the training sample. The figures on the left side (figures **(a)**, **(c)** and **(e)**) show the distributions before the resampling, and the figures of the right side (figures **(b)**, **(d)** and **(f)**) show the distributions after the resampling.

available jets are used in those cases. The sample composition and the p_{T}-$|\eta|$ distributions, both before and after resampling, are shown in Figure 7.10 for the case of 4 jet classes.

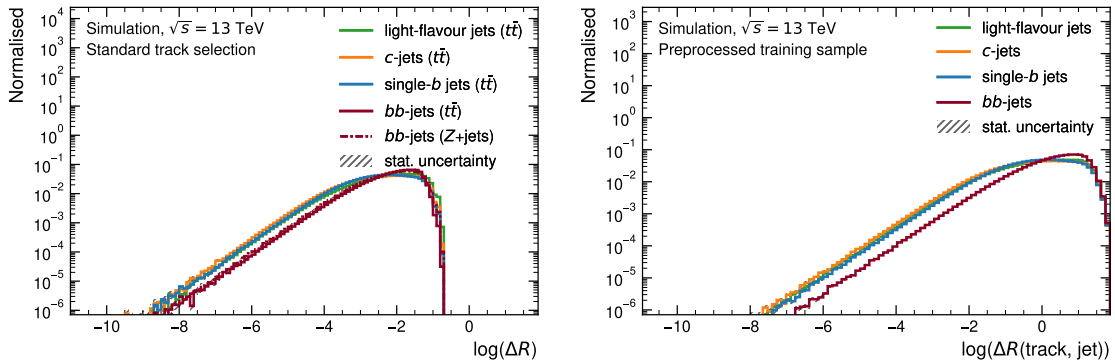
Variable Standardisation

After the resampling, some variables are shifted and scaled according to the standardisation transformation from Equation 5.4 that was introduced in Section 5.1.2. Variables which already have a mean near zero are not transformed. A short summary table of which variables are standardised is given in Table 7.3. The distributions of the preprocessed

Table 7.3.: Summary of which DIPS input variables are standardised in the preprocessing.

Standardised variables	Not standardised variables
$\log p_{\mathrm{T}}^{\mathrm{frac}}$	s_{d_0}
$\log \Delta R$	$s_{z_0 \sin\theta}$
nPixHits	IBL hits
nSCTHits	PIX1 hits
d_0	Shared IBL hits
$z_0 \sin\theta$	Split IBL hits
	Shared pixel hits
	Split pixel hits
	Shared SCT hits

variables are shown in the appendix in Figure A.1. The corresponding distributions of the training sample for the DIPS training with the inclusive b-jet class is shown in Figure A.2. The distribution of the $\log \Delta R$(track, jet) variable is shown in Figure 7.11 both before and after the standardisation. While the distribution before the preprocessing is located at negative values only, the distribution after the preprocessing is located around zero. Since the standard deviation of $\log \Delta R$(track, jet) is already close to 1 before the standardisation, there is not much change in the width of the distributions.



(a) Before preprocessing **(b)** After preprocessing

Figure 7.11.: The $\log \Delta R$(track, jet) distribution for each jet class before **(a)** and after **(b)** the preprocessing. The distribution corresponding to bb-jets after the preprocessing shows the joint distribution of bb-jets from $t\bar{t}$ and Z+jets.

In addition to the shifting and scaling, small changes in the shape of the distributions can be seen for some variables, one example being the variable $\log \Delta p_{\mathrm{T}}^{\mathrm{frac}}$. This is caused by the resampling, since the distributions before the preprocessing are obtained with jets from different kinematic regions, while the jets after the resampling all have the same distribution in the p_{T}-$|\eta|$ plane. If a variable is correlated with the jet p_{T}, the distribution of this variable changes when the resampling procedure is applied.

7.2.2. Validation and Test Samples

During the training, the algorithms are evaluated on two different validation samples. The first validation sample is also processed with the same resampling procedure as the training sample. This ensures that the fractions of the different jet classes are the same in the training sample and the validation sample. This validation sample is in the following referred to simply as *validation sample*. The second sample that is used for validation during training is a *non-resampled validation sample*, which only contains jets from $t\bar{t}$ events in the fractions they appear in those events. It should be noted that the jets in the non-resampled validation sample are obtained from the same pool of jets as the jets for the other validation sample. Therefore, these two validation samples are not orthogonal to each other. While the validation sample randomly selects jets from the pool of jets with $N_{\mathrm{event}} \mod 6 = 4$ during the resampling procedure, the non-resampled validation sample contains the first 300k jets that are written out from the sample of $t\bar{t}$ events. The overlap between the two validation samples is therefore very small, but not zero. After training, the algorithms are evaluated on a *test sample*. Like the non-resampled validation sample, the test sample contains jets from $t\bar{t}$ events without any resampling. A summary of the different samples is shown in Table 7.4.

Furthermore, the shifting and scaling factors calculated for the standardisation of the training sample are applied to the validation and test samples before being processed by the algorithm.

Table 7.4.: Overview of the different samples used for training, validation and testing.

Sample name	N_{jets}	$N_{\mathrm{event}} \mod 6$	Resampled
Training sample	4M per jet class	≤ 3	✔
Validation sample	300k in total	$= 4$	✔
Non-resampled validation sample	300k in total		✖
Test sample	4M in total	$= 5$	✖

7.3. Algorithm Training and Evaluation

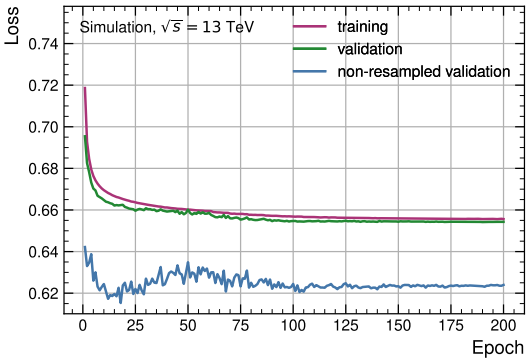
In order to investigate the effect of adding a bb-jet class to the DIPS algorithm, two trainings are presented in this section. The first training is a baseline training of DIPS which uses the standard b-tagging classes, i.e. light-flavour jets, c-jets and the inclusive b-jets category. While the objective of the studies in this thesis is to use the extended jet classes, this baseline training is performed to have a reference of a DIPS model trained with the usual jet classes. All trainings presented in this thesis are implemented with TensorFlow 2.9.1 [110] using the Keras 2.9.0 [111] API. The training sample is created using the same procedure as explained in Section 7.2 but without the bb-jets from the Z+jets sample. The second training is the result obtained with bb-DIPS, i.e. when using light-flavour jets, c-jets, single-b jets and bb-jets as the target classes of the algorithm. Both algorithms use the network architecture presented in Section 6.2.1, with the only difference being the number of output nodes corresponding to the target classes of the model. The full list of hyperparameters of both setups is shown in Table 7.5. Both models use a dropout rate of 0.05 in all hidden layers of the ρ network, while only using dropout with a rate of 0.05 in the last layer of the ϕ network. Several combinations of dropout rates were studied, and the combination presented here was found to lead to the best agreement between the accuracy of the model predictions in the training and validation sample.

Table 7.5.: Hyperparameters of the DIPS and the bb-DIPS model. The two models use the same hyperparameters, except for the definition of the target classes and thereby the number of the output nodes.

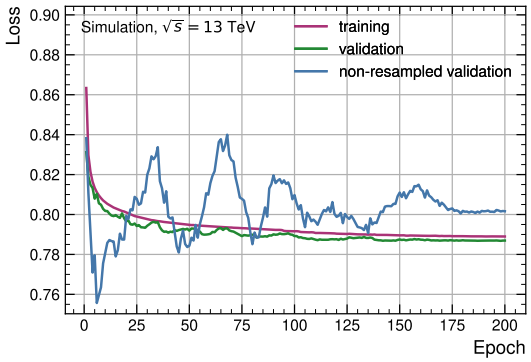
Parameter	DIPS	bb-DIPS
Target classes	light-flavour jets c-jets b-jets	light-flavour jets c-jets single-b jets bb-jets
Size of training sample	12M (4M per class)	16M (4M per class)
Output nodes	$[p_{\text{light}}, p_c, p_b]$	$[p_{\text{light}}, p_c, p_{\text{single-}b}, p_{bb}]$
$N_{\text{nodes}}^{\text{hidden}}$ ϕ	[100, 100, 128]	
$N_{\text{nodes}}^{\text{hidden}}$ ρ	[100, 100, 100, 30]	
Dropout rates in ϕ	[0, 0, 0.05]	
Dropout rates in ρ	[0.05, 0.05, 0.05, 0.05]	
Learning rate	0.001	
Batch size	15000	
Batch normalisation	✔	
Learning rate reducer	✔	
Training epochs	200	
Optimiser	Adam	
Loss function	Categorical cross entropy	

7.3.1. Training Results

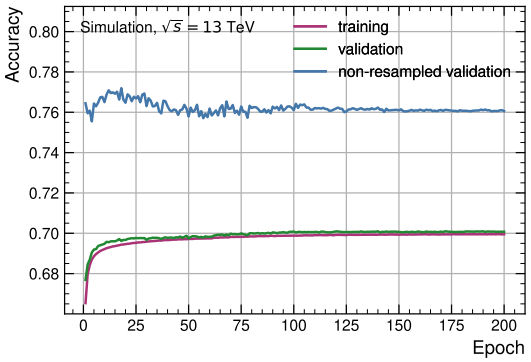
The loss and accuracy curves of both trainings are shown in Figure 7.12. In both trainings the loss of the validation sample (green line) agrees well with the loss of the training sample (magenta line). The validation loss is even slightly smaller than the training loss. This can be explained by the usage of dropout, which is only applied during the training and not in the validation step. Hence, the predictions made on the training sample are calculated with some connections in the neural network turned off. The predictions on the validation samples on the other hand are calculated with all connections, resulting in slightly better performance compared to the training.



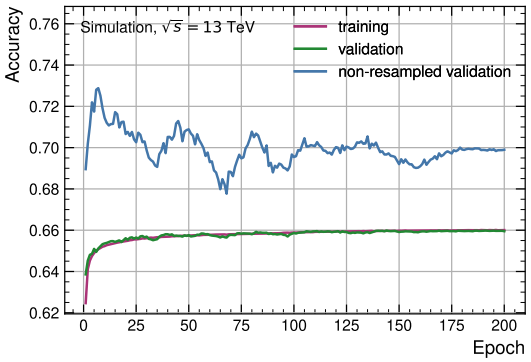
(a) DIPS loss



(b) bb-DIPS loss



(c) DIPS accuracy

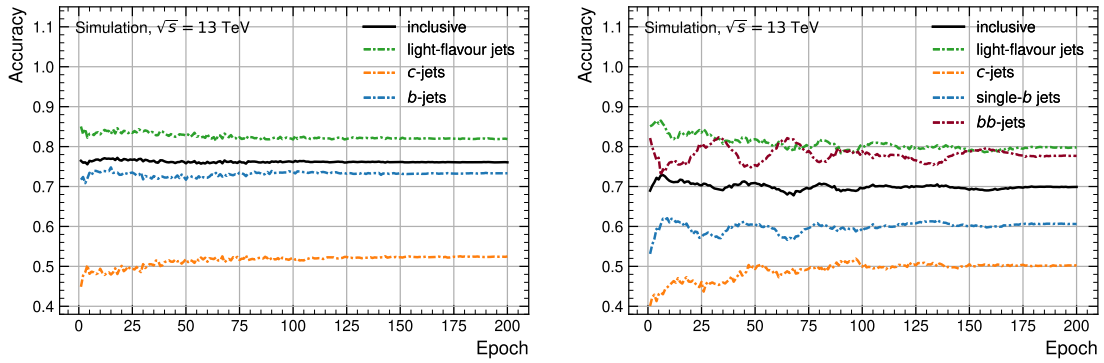


(d) bb-DIPS accuracy

Figure 7.12.: Loss and accuracy curves as a function of the training epoch: figure **(a)** and figure **(b)** show the loss of DIPS and bb-DIPS while figure **(c)** and figure **(d)** show the accuracy of the two models. All figures show the corresponding metric for the training sample (magenta) and the two different validation samples (green and blue).

The accuracy also shows very good agreement between the performance on the training sample and on the validation sample. The training is not performed over more than 200 epochs, since both the loss and the accuracy already show convergence with this number of training epochs. The converged value of the accuracy is around 70 % for DIPS while the accuracy of the bb-DIPS model converges towards 66 %. The DIPS training loss converges to approximately 0.655 while the bb-DIPS training loss converges to a value of 0.790. This shows that the model performance is worse with the additional bb-jet class when all target classes are to be distinguished from each other.

Both the loss and the accuracy of the non-resampled validation sample (blue) show an offset to the corresponding metric calculated on the training sample. This offset is caused



(a) DIPS **(b)** bb-DIPS

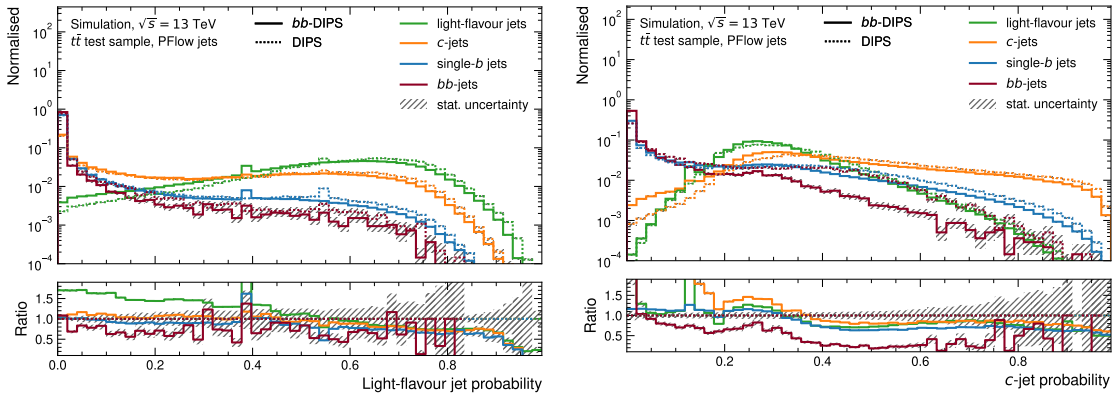
Figure 7.13.: Validation accuracy as a function of the training epoch for each jet class individually in the non-resampled validation sample. Figure **(a)** shows the corresponding decomposition for the DIPS training and figure **(b)** for the bb-DIPS training.

by the composition of the non-resampled validation sample and the fact that the model performance is different for the different jet classes. Since this sample has imbalanced classes, the different jet classes do not contribute with equal fractions to the average loss as defined in Equation 5.2. In contrast, for both the training sample and the validation sample, each jet class contributes with the same fraction to the overall loss of the model.

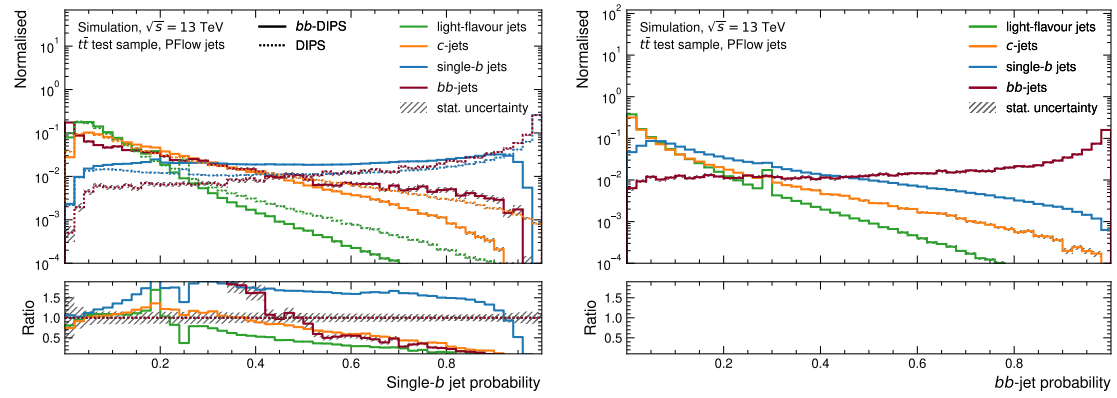
Furthermore, the loss of the non-resampled validation sample in Figure 7.12b shows large oscillations over the training process. This is also caused by the imbalanced classes in this sample.

Figure 7.13 shows the accuracy for each jet class from the non-resampled validation sample individually. The accuracy is calculated by comparing the model prediction with the true label of each jet. The model prediction is the class corresponding to the output node with the maximum value for a given jet. If the output node with the largest activation value corresponds to the truth label, a jet is counted as correctly predicted. As can be seen in Figure 7.13b, the accuracy of each individual class shows oscillations which align with the oscillations seen in the loss curve in Figure 7.12b. These oscillations do not show up in the training of DIPS, with the class-specific accuracies shown in Figure 7.13a. The reason for this could be that the classification task becomes more complex with the additional bb-jet class in bb-DIPS. When the bb-DIPS model improves in the classification of bb-jets, the accuracy for single-b jet and light-flavour jet classification decreases and vice versa. Since the overall accuracy over the whole non-resampled validation sample is dominated by the accuracy on single-b jets and light-flavour jets, these oscillations are translated into the overall accuracy. This is less pronounced in the validation sample, since the overall loss and accuracy is calculated with equal numbers of jets from the different classes. Furthermore, the fact that the classification task becomes harder with four target classes is also reflected in the values of the converged loss and accuracy curves.

The last training epoch is chosen for the evaluation, as there is no other epoch with significantly smaller loss values. The raw output distributions of the corresponding two models are shown in Figure 7.14. The distribution is shown for each jet class and each output node. Since the p_b node of the DIPS model is expected to be comparable to the $p_{\text{single-}b}$ node of the bb-DIPS model, the corresponding distributions are drawn in the same plot. The solid lines show the output of the bb-DIPS model and the dotted lines show the DIPS output. While the p_{light} and p_c distributions shift slightly towards smaller values in bb-DIPS, the $p_{\text{single-}b}$ distribution from bb-DIPS in Figure 7.14c shows significant changes compared to the p_b distribution from DIPS, which is drawn in the same plot. The p_b



(a) p_{light} **(b)** p_c



(c) p_b (DIPS) and $p_{\text{single-}b}$ (bb-DIPS) **(d)** p_{bb}

Figure 7.14.: Comparison of the raw output of the DIPS tagger with 3 output classes (dotted line) and the bb-DIPS tagger with 4 output classes (solid line): **(a)** the light-flavour jet probability p_{light}, **(b)** the c-jet probability p_c, **(c)** the single-b jet probability $p_{\text{single-}b}$ (bb-DIPS) and b-jet probability p_b (DIPS) and **(d)** the bb-jet probability p_{bb}. The bb-jet probability in figure **(d)** is only shown for the bb-DIPS tagger, since this information is not available with the default DIPS output. The ratio is calculated with respect to the DIPS prediction for each jet flavour individually.

distribution from DIPS is rather similar for single-b jets and bb-jets, which is expected as they both correspond to the b-jets target class in DIPS. The $p_{\text{single-}b}$ distribution of bb-DIPS on the other hand shows the intended separation between single-b jets and bb-jets, since this model has a dedicated target class for both single-b jets and bb-jets. Finally, the p_{bb} distribution in Figure 7.14d is only shown for bb-DIPS, since this information does not exist in the DIPS model. The spikes in the p_{light} distribution (Figure 7.14a) and the p_{bb} distribution (Figure 7.14c) originate from jets with no tracks. The corresponding spikes in p_c (Figure 7.14b) and $p_{\text{single-}b}$ (Figure 7.14c) are not clearly visible in the plot, since they are hidden in the bulk of the distribution. The DIPS predictions for jets with no tracks are $p_{\text{light}} \approx 0.56$, $p_c \approx 0.19$ and $p_b \approx 0.25$. The bb-DIPS predictions for jets with no tracks are $p_{\text{light}} \approx 0.40$, $p_c \approx 0.13$, $p_{\text{single-}b} \approx 0.18$ and $p_{bb} \approx 0.29$.

7.3.2. Performance in Inclusive b-Tagging

The b-tagging discriminant D_b for both DIPS and bb-DIPS is shown in Figure 7.15. By introducing the additional target class with bb-DIPS, the separation between the different

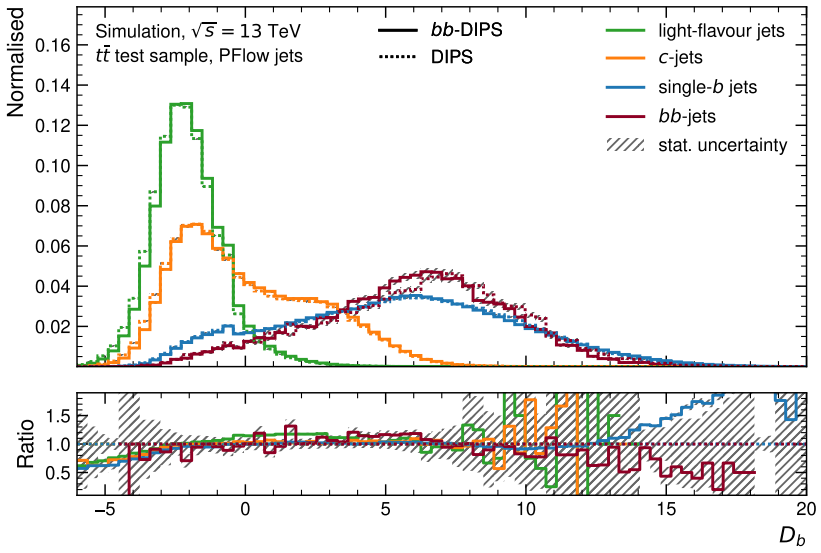


Figure 7.15.: Comparison of the b-tagging discriminant D_b obtained from the DIPS tagger with 3 output classes (dotted lines) and the bb-DIPS tagger with 4 output classes (solid lines). The lower panel shows the ratio between the bb-DIPS histograms and the DIPS histograms for each jet class separately. The f_c value, which defines the weight of p_c and p_{light} in the discriminant, is set to 0.005. Jets with no tracks are located at $D_b \approx -0.80$ for DIPS and at $D_b \approx -0.78$ for bb-DIPS.

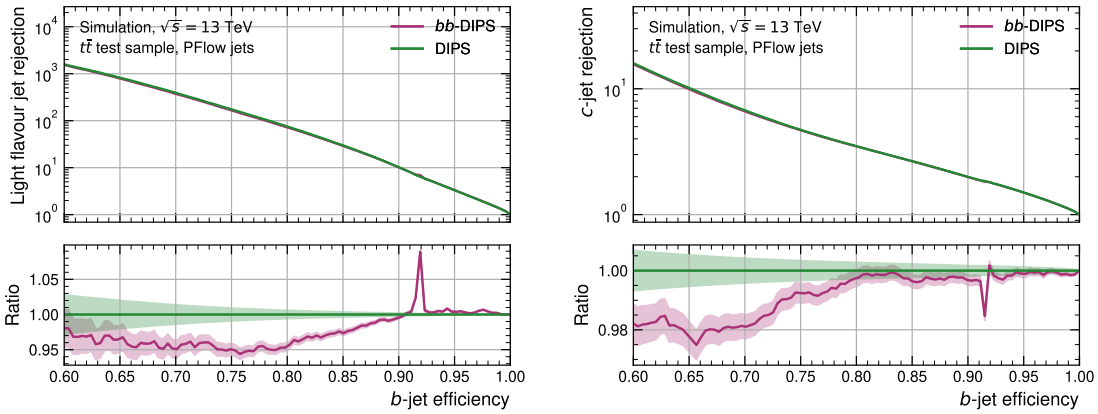
histograms is only slightly affected. The separation $\text{Sep}(h_1, h_2)$ between two equally binned, normalised histograms h_1 and h_2 is calculated using the formula

$$\text{Sep}(h_1, h_2) = \frac{1}{2} \cdot \sum_{i=1}^{N} \frac{(h_{1,i} - h_{2,i})^2}{h_{1,i} + h_{2,i}}, \tag{7.1}$$

with $h_{j,i}$ being the height of bin i of histogram h_j and N begin the total number of bins. With this definition, two non-overlapping histograms have a separation of 1 while two exactly overlapping histograms have a separation of 0. The separation values corresponding to the histograms in Figure 7.15 are listed in Table 7.6. The largest difference is seen in the separation between bb-jets and c-jets, which decreases from $(51.9 \pm 0.4)\,\%$ to $(50.6 \pm 0.4)\,\%$. The difference in separation is within two standard deviations. The separation values corresponding to other combinations of signal (single-b jets and bb-jets) and background (light-flavour jets and c-jets) have a much smaller difference and agree within one standard deviation. Hence, the extension of the tagger by an additional bb-jet node only leads to a small decrease in separation power between c-jets and bb-jets.

Table 7.6.: Separation between the background classes (light-flavour jets and c-jets) and the signal classes (single-b jets and bb-jets) in the D_b distribution. Since D_b is the discriminant for b-tagging with the inclusive b-jets class, both single-b jets and bb-jets are part of the signal.

	Sep(single-b, light)	Sep(single-b, c)	Sep(bb, light)	Sep(bb, c)
DIPS	$(72.1 \pm 0.1)\,\%$	$(39.7 \pm 0.1)\,\%$	$(82.6 \pm 0.5)\,\%$	$(51.9 \pm 0.4)\,\%$
bb-DIPS	$(72.0 \pm 0.1)\,\%$	$(39.5 \pm 0.1)\,\%$	$(82.5 \pm 0.5)\,\%$	$(50.6 \pm 0.4)\,\%$

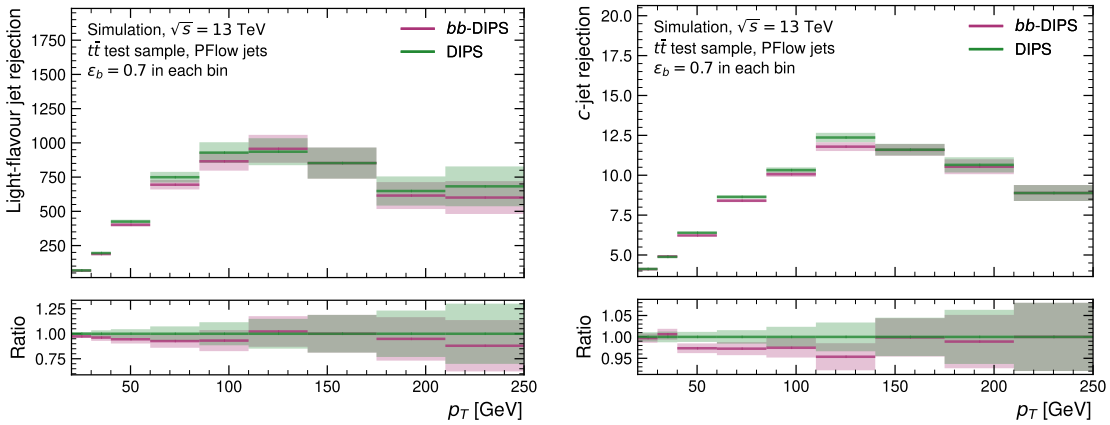


(a) **(b)**

Figure 7.16.: Light-flavour-jet rejection as a function of the b-jet efficiency **(a)** and c-jet rejection as a function of the b-jet efficiency **(b)** for DIPS (green) and bb-DIPS (magenta). The b-jet efficiency refers to the efficiency of the inclusive b-jet class which includes both single-b jets and bb-jets. The lower panels show the ratio with respect to the DIPS line (green). The f_c value in the discriminant D_b is set to 0.005. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

The light-flavour-jet rejection and the c-jet rejection curves corresponding to the discriminant distributions in Figure 7.15 are shown in Figure 7.16 as a function of the b-jet efficiency, where the b-jet efficiency corresponds to the inclusive b-jet class. The light-flavour-jet rejection of bb-DIPS is smaller than the one of DIPS for b-jet efficiencies smaller than 90 %, which can be seen in the lower panel in Figure 7.16a. However, the largest difference is only around 5 % compared to DIPS. The difference is even smaller for c-jet rejection, which can be seen in Figure 7.16b. For b-jet efficiencies larger than 80 % the c-jet rejection of the two taggers is nearly identical. For smaller b-jet efficiencies, the c-jet rejection is better for DIPS, but the largest relative difference is only around 2 %. The spikes in Figure 7.16a and Figure 7.16b at a b-jet efficiency slightly above 0.91 are an artefact caused by the location of jets with no tracks, which are located at $D_b \approx -0.80$ for DIPS and at $D_b \approx -0.78$ for bb-DIPS. It should be noted that the rejection is a very sensitive metric in the way that very small changes in the underlying discriminant distribution lead to rather large changes in the corresponding rejection curves.

The p_{T} dependence of both the light-flavour-jet rejection and the c-jet rejection is shown in Figure 7.17, where the corresponding rejection values are calculated for a b-jet efficiency of 70 % in each bin of p_{T}. The performance of both models increases at smaller jet p_{T} until it drops for $p_{\mathrm{T}} > 140\,\mathrm{GeV}$. This overall trend of the p_{T}-dependence of both the light-flavour jet rejection and the c-jet rejection is consistent with previous studies of the DIPS algorithm [103] as well as the DL1r algorithm [46].



(a) **(b)**

Figure 7.17.: Light-flavour-jet rejection **(a)** and c-jet rejection **(b)** calculated in bins of the jet p_{T} for DIPS (green) and bb-DIPS (magenta). Each bin shows the corresponding rejection for a b-jet efficiency of 70 % in that bin. The b-jet efficiency refers to the inclusive b-jet class which includes both single-b jets and bb-jets. The lower panels show the ratio of the rejection with respect to DIPS (green). The f_c value in the discriminant D_b is set to 0.005. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

7.3.3. Performance in bb-Jet Identification

Using the p_{bb} information from the bb-DIPS model, single-b jets and bb-jets can be separated using the discriminant $D_{\text{single-}b}$ as defined in Equation 6.7. The corresponding distributions for single-b jets and bb-jets are shown in Figure 7.18. The separation between the two distributions is (46.4 ± 0.4) %.

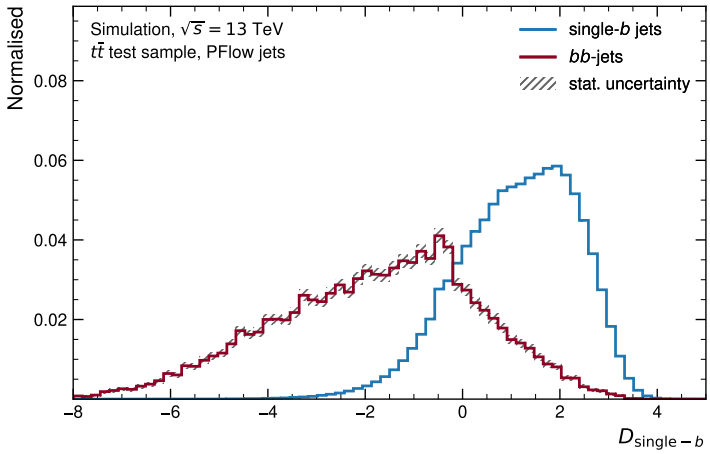


Figure 7.18.: $D_{\text{single-}b}$ distribution of the bb-DIPS tagger. This information is not available with the default DIPS output, since that tagger does not have the p_{bb} output node.

Moreover, applying a cut on $D_{\text{single-}b}$ at a single-b jet efficiency of 70 % results in a bb-jet rejection of 7.62 ± 0.17. Although the $D_{\text{single-}b}$ discriminant is only designed to separate between single-b jets and bb-jets, it also separates bb-jets from the other two classes, namely light-flavour jets and c-jets. The corresponding separation values are listed in Table 7.7 and are approximately 50 %. The $D_{\text{single-}b}$ distributions of all jet classes, including the

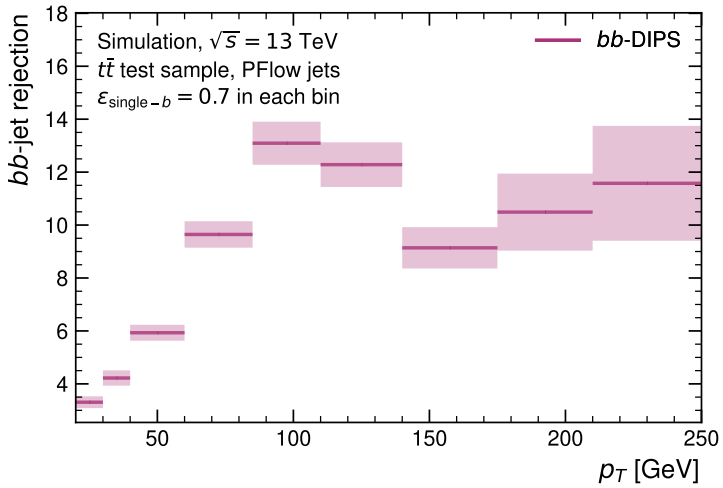


Figure 7.19.: bb-jet rejection of the bb-DIPS algorithm at a 70 % single-b jet efficiency for different bins in p_{T}. The cut value in $D_{\text{single-}b}$ is chosen such that a single-b jet efficiency of 70 % is obtained in each bin separately. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

distributions of light-flavour jets and c-jets, are shown in the appendix in Figure A.3.

The bb-jet rejection is shown for bins of the jet p_{T} in Figure 7.19, which shows a similar p_{T} dependence as the light-flavour-jet rejection and the c-jet rejection shown in Figure 7.17. To illustrate this more clearly, the discriminant $D_{\text{single-}b}$ is shown in Figure 7.20 for single-b jets and bb-jets from different p_{T} ranges. The corresponding separation values for the different

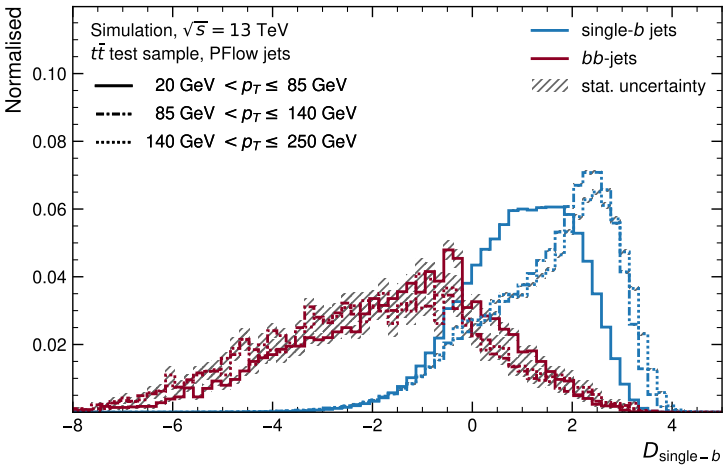


Figure 7.20.: $D_{\text{single-}b}$ distribution of the bb-DIPS tagger for jets of three different regions of the jet p_{T}. Each linestyle corresponds to a certain p_{T} region, with the corresponding values of p_{T} indicated in the upper left of the figure.

p_{T} regions are listed in Table 7.7. The $D_{\text{single-}b}$ distribution of bb-jets shifts to smaller values for the two higher p_{T} regions (dash-dotted and dotted lines). However, when splitting the test sample into slices of p_{T}, the statistical uncertainties of the distributions become rather large. Furthermore, the $D_{\text{single-}b}$ distribution of the single-b jets peaks at noticeably larger values for single-b jets with $p_{\text{T}} > 85\,\text{GeV}$. Therefore, applying a cut on the distribution corresponding to the jets with higher p_{T} results in a cut value that is located at larger values, cutting away a larger fraction of the bb-jets, which then yields the larger bb-jet rejection.

This can also be seen in the separation values. The separation between single-b jets and bb-jets is $(41.0 \pm 0.6)\,\%$ for jets in the low-p_{T} region ($20\,\mathrm{GeV} < p_{\mathrm{T}} \leq 85\,\mathrm{GeV}$) while the separation is much larger in the two p_{T} regions with $85\,\mathrm{GeV} < p_{\mathrm{T}} \leq 140\,\mathrm{GeV}$ and $140\,\mathrm{GeV} < p_{\mathrm{T}} \leq 250\,\mathrm{GeV}$, with a separation of $(54.9 \pm 0.7)\,\%$ and $(51 \pm 1)\,\%$, respectively.

Table 7.7.: Separation of the $D_{\text{single-}b}$ distributions of the different jet classes for different ranges of the jet p_{T}. The values are calculated using the bb-DIPS model.

p_{T} range	Sep(bb, single-b)	Sep(bb, light)	Sep(bb, c)
$20\,\mathrm{GeV} < p_{\mathrm{T}} \leq 250\,\mathrm{GeV}$	$(46.4 \pm 0.4)\,\%$	$(51.1 \pm 0.5)\,\%$	$(51.5 \pm 0.5)\,\%$
$20\,\mathrm{GeV} < p_{\mathrm{T}} \leq 85\,\mathrm{GeV}$	$(41.0 \pm 0.6)\,\%$	$(45.3 \pm 0.6)\,\%$	$(46.2 \pm 0.6)\,\%$
$85\,\mathrm{GeV} < p_{\mathrm{T}} \leq 140\,\mathrm{GeV}$	$(54.9 \pm 0.7)\,\%$	$(65.2 \pm 0.7)\,\%$	$(61.7 \pm 0.7)\,\%$
$140\,\mathrm{GeV} < p_{\mathrm{T}} \leq 250\,\mathrm{GeV}$	$(51 \pm 1)\,\%$	$(62 \pm 1)\,\%$	$(56 \pm 1)\,\%$

7.4. Track Selection Optimisation Studies

Since the bb-DIPS algorithm uses the information of the selected tracks of each jet, the corresponding track selection defines what information will be made available to the algorithm. Loosening the criteria of the track selection results in more selected tracks per jet, and hence provides more information for the algorithm to work with. However, the selection has to filter out enough tracks such that only tracks that are believed to contain meaningful information about the overall jet are selected. In this section, studies regarding a loosening of the previous track selection are presented. This looser track selection is used for the DIPS tagger in ATLAS trained for Run 3 of the LHC and is also studied in Ref. [103].

7.4.1. Loose track selection

In order to obtain more tracks per jet, several criteria of the standard track selection are loosened. The affected criteria are the minimal p_{T} requirement which is decreased from 1 GeV to 0.5 GeV as well as the maximum allowed values for the transverse and longitudinal impact parameters, which are increased from 1 mm to 3.5 mm and from 1.5 mm to 5 mm respectively. This modified track selection is referred to in the following as *loose track selection*. An overview of the criteria for the two different track selections is shown in Table 7.8.

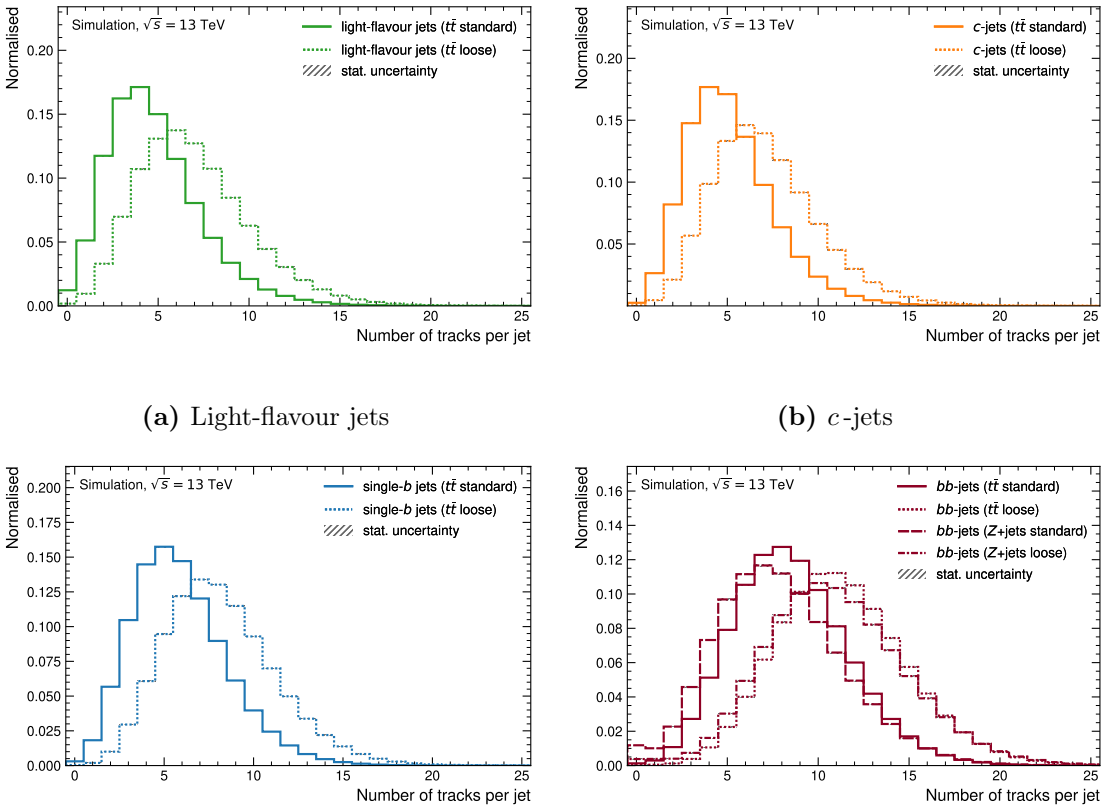
Table 7.8.: Overview of the criteria used in the two different track selections.

Variable	Standard selection	Loose selection
$p_{\mathrm{T}}^{\mathrm{track}}$	$> 1\,\mathrm{GeV}$	$> 0.5\,\mathrm{GeV}$
$\lvert d_0 \rvert$	$< 1\,\mathrm{mm}$	$< 3.5\,\mathrm{mm}$
$\lvert z_0 \sin(\theta) \rvert$	$< 1.5\,\mathrm{mm}$	$< 5\,\mathrm{mm}$
$\lvert \eta \rvert$	< 2.5	
$N_{\text{Silicon hits}}$	≥ 7	
$N_{\text{Pixel holes}}$	< 2	
$N_{\text{Silicon shared}}$	< 2	
$N_{\text{Silicon holes}}$	< 3	

The resulting distribution of the number of selected tracks per jet is shown in Figure 7.21. For better visibility, each jet flavour is shown in a separate plot. For all jet flavours, the distributions shift towards larger values as expected. Furthermore, the shape of the distributions changes slightly, with the distributions corresponding to the loose track selection having a larger variance.

Light-flavour jets, c-jets and single-b jets have on average 2 more tracks with the loose track selection compared to the standard track selection. The increase is even larger for bb-jets, which have on average around 3 more tracks in the loose track selection. The average number of tracks from the track categories discussed before are shown in Table 7.9 in comparison to the numbers obtained with the standard track selection. The p_{T} dependence of the average number of tracks from each of the track origins is shown in the appendix in Figure A.4.

By loosening the cuts in the track selection, the jets of all categories contain on average one more track from pile-up. The average number of tracks from other secondary interac-



(a) Light-flavour jets **(b)** c-jets

(c) Single-b jets **(d)** bb-jets

Figure 7.21.: Comparison of the resulting number of selected tracks per jet when using the standard track selection or the loose track selection. Each figure shows the distribution of the number of tracks per jet for both selections but only for one jet flavour for better visibility.

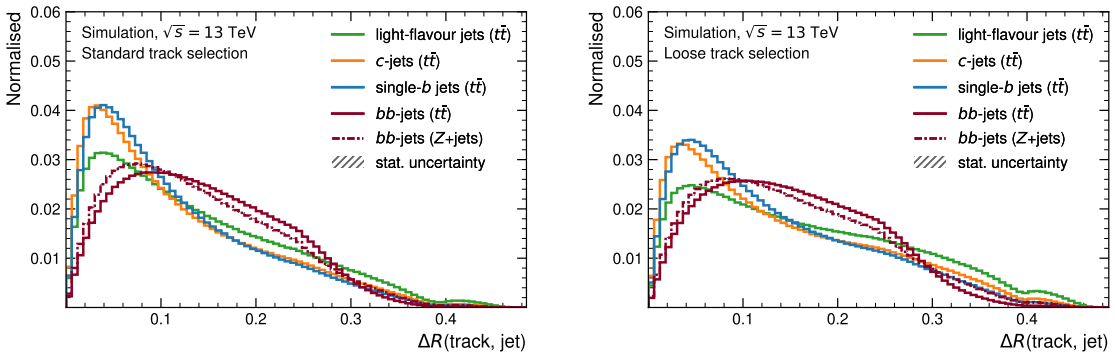
tions or the decay of τ-leptons increases only slightly, with the maximum increase of 0.2 seen for light-flavour jets. The increase in the average number of tracks from the PV ranges from 0.5 for single-b jets to 0.8 for light-flavour jets. Most notably, bb-jets from Z+jets events have on average 1.3 more tracks from the PV in the loose track selection. bb-jets from $t\bar{t}$ events have on average 1.1 more tracks from HF decays compared to the average number obtained with the standard track selection. The increase in tracks from HF decays is smaller in single-b jets and c-jets with an increase of 0.7 and 0.2 tracks.

Given that the increase in the average number of tracks from pile-up is approximately the same for all jet categories, bb-jets benefit the most from this loosening of the cut criteria in terms of the increase of track information, since the number of other tracks from other origins increases the most for these jets.

The distribution of ΔR(track, jet) is shown in Figure 7.22 for both track selections. For all jet flavours the distribution shifts to larger ΔR(track, jet) values. This is expected, since tracks with a large IP usually also have a rather large angular distance from the jet axis. Hence, loosening the cut on the impact parameter yields mostly additional tracks with large ΔR(track, jet). The distributions of all non-preprocessed input variables as obtained with the loose track selection are shown in Figure 7.23.

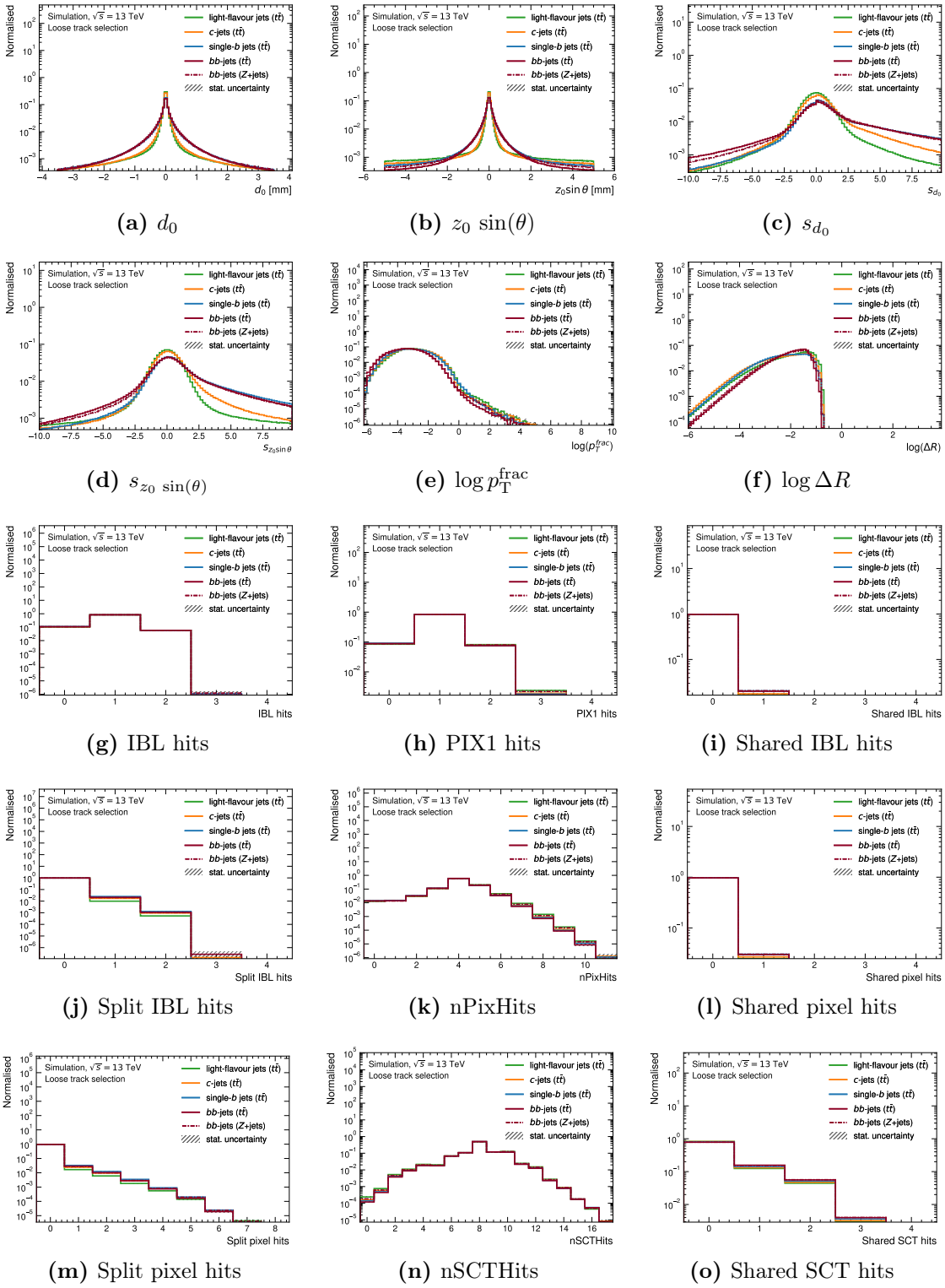
Table 7.9.: Average number of tracks per jet corresponding to the different jet-flavour categories. The statistical uncertainty of these values are negligible, and thus not shown in the table. The columns represent the average number of tracks from the decay of a heavy-flavour hadron $\langle N_{\text{tracks}}^{\text{HF}} \rangle$, from the primary vertex $\langle N_{\text{tracks}}^{\text{PV}} \rangle$, from pile-up $\langle N_{\text{tracks}}^{\text{pile-up}} \rangle$, and the average number of tracks that originate from other secondary interactions, the decay of a τ-lepton or which are created from the hits of multiple particles $\langle N_{\text{tracks}}^{\text{other}} \rangle$.

Jet flavour	Track selection	$\langle N_{\text{tracks}} \rangle$	$\langle N_{\text{tracks}}^{\text{HF}} \rangle$	$\langle N_{\text{tracks}}^{\text{PV}} \rangle$	$\langle N_{\text{tracks}}^{\text{pile-up}} \rangle$	$\langle N_{\text{tracks}}^{\text{other}} \rangle$
Light-flavour jets ($t\bar{t}$)	standard	4.8	0.02	4.4	0.3	0.1
	loose	7.0	0.04	5.2	1.5	0.3
c-jets ($t\bar{t}$)	standard	5.2	2.1	2.8	0.2	0.1
	loose	7.1	2.3	3.4	1.2	0.2
Single-b jets ($t\bar{t}$)	standard	5.9	3.6	2.0	0.2	0.05
	loose	8.2	4.3	2.5	1.2	0.1
bb-jets ($t\bar{t}$)	standard	8.5	6.0	2.2	0.3	0.06
	loose	11.3	7.1	2.7	1.3	0.1
bb-jets (Z+jets)	standard	7.9	5.7	1.8	0.3	0.06
	loose	10.9	7.0	2.4	1.4	0.1



(a) **(b)**

Figure 7.22.: Comparison of the angular distance between the track and the jet axis for the two different track selections: **(a)** with the standard track selection and **(b)** with the loose track selection.



(a) d_0 **(b)** $z_0 \sin(\theta)$ **(c)** s_{d_0}

(d) $s_{z_0 \sin(\theta)}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

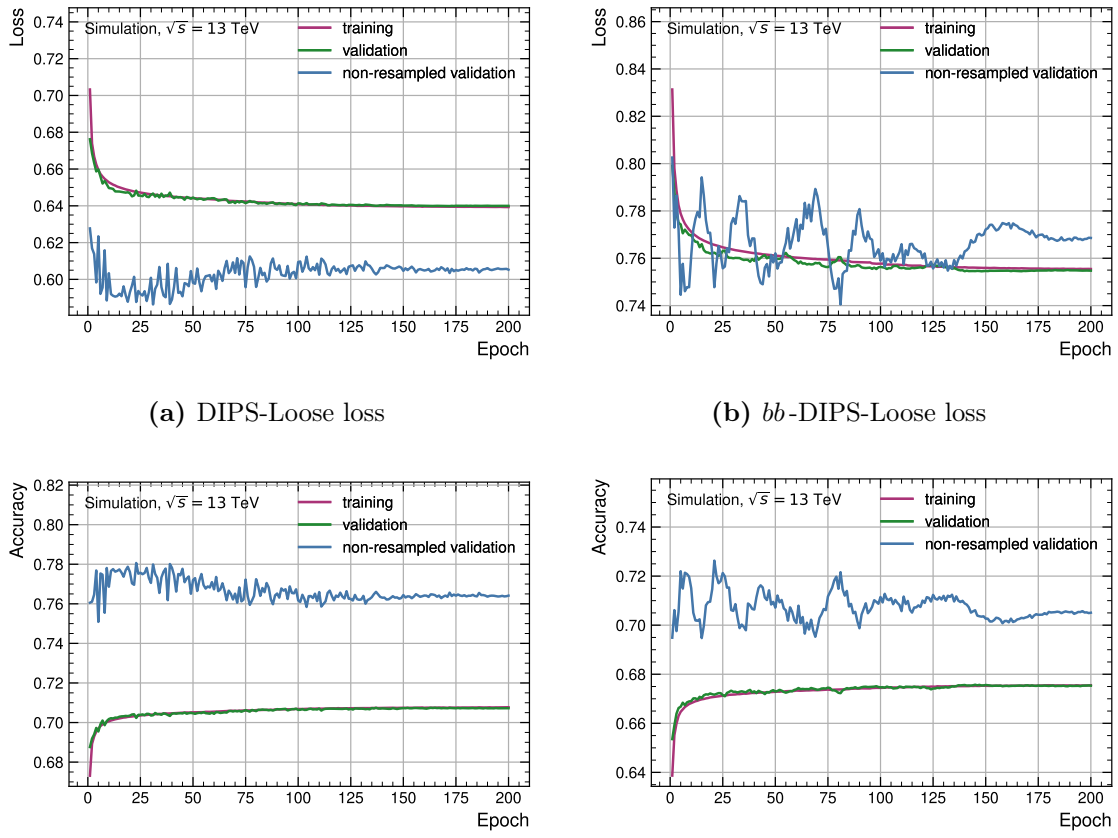
(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure 7.23.: Distributions of the DIPS input variables. These distributions are obtained with the loose track selection.

7.4.2. Training Results

An additional training of the bb-DIPS algorithm is carried out using the track information corresponding to the loose track selection. In the following, this configuration is referred to as bb-DIPS-Loose. As before, a reference training is performed with the inclusive b-jet target class, referred to as DIPS-Loose. Both the training sample and the validation sample are created using the same procedure as explained before, but with the loose track selection. The distributions of the input variables after the preprocessing procedure are shown in the appendix in Figure A.5 for the case of 4 target classes and in Figure A.6 for the case of 3 target classes. The hyperparameters of the bb-DIPS-Loose and DIPS-Loose algorithms are the same as for their standard-track-selection equivalent (bb-DIPS and DIPS). The evolution of the loss and the accuracy during the training process is shown in Figure 7.24 for both DIPS-Loose and bb-DIPS-Loose.



(a) DIPS-Loose loss

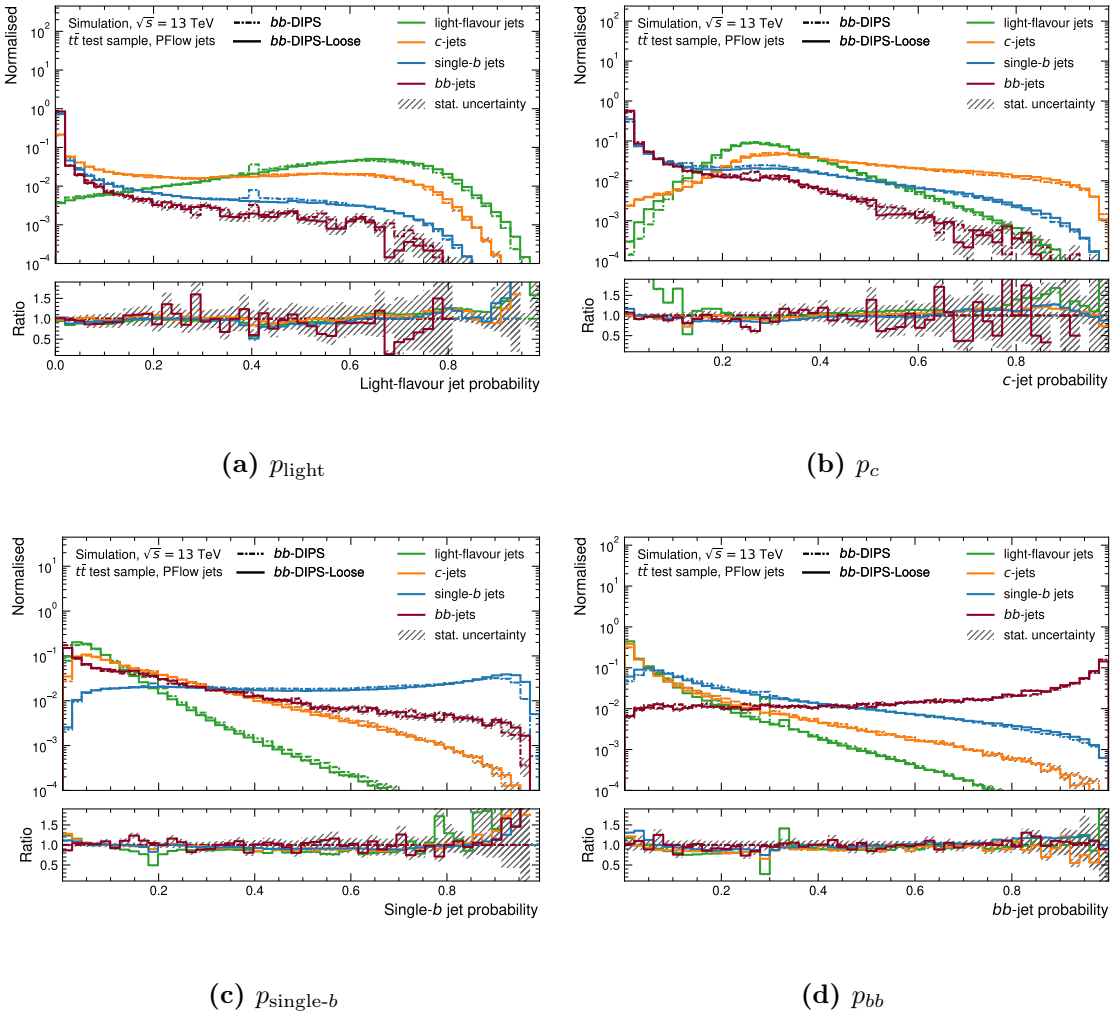
(b) bb-DIPS-Loose loss

(c) DIPS-Loose accuracy

(d) bb-DIPS-Loose accuracy

Figure 7.24.: Loss and accuracy as a function of the training epoch of the DIPS-Loose and bb-DIPS-Loose training. The figures show the corresponding metric for the training sample (magenta) and for the two different validation samples (green and blue).

Very good agreement of both the loss and the accuracy is seen between the training sample and the validation sample for both taggers. The converged values of both the loss and the accuracy of DIPS-Loose and bb-DIPS-Loose show a small improvement compared to the corresponding values obtained in the DIPS and bb-DIPS trainings in Section 7.3.1. The behaviour of the loss and accuracy of the non-resampled validation sample is caused by the same reasons as discussed in Section 7.3.1. In both cases the model from the last epoch is chosen for evaluation. The raw tagger output is shown in Figure 7.25. Compared to the bb-DIPS model, the bb-DIPS-Loose model especially gains in separation power in the



(a) p_{light} **(b)** p_c

(c) $p_{\text{single-}b}$ **(d)** p_{bb}

Figure 7.25.: Comparison of the raw tagger output of the bb-DIPS tagger (dash-dotted line) and the bb-DIPS-Loose tagger (solid line): **(a)** the light-flavour jet probability p_{light}, **(b)** the c-jet probability p_c, **(c)** the single-b jet probability $p_{\text{single-}b}$ and **(d)** the bb-jet probability p_{bb}. The ratio is calculated with respect to the bb-DIPS prediction for each jet flavour individually.

$p_{\text{single-}b}$ node (Figure 7.25c), where single-b jets obtain larger values with bb-DIPS-Loose. Furthermore, single-b jets and bb-jets shift to larger values in the p_{light} node and the p_c node, which is an improvement as well.

7.4.3. Performance in Inclusive b-Tagging

The distributions of the D_b discriminant are shown in Figure 7.26 for both bb-DIPS-Loose and bb-DIPS. The loose track selection leads to a better separation, with single-b jets and bb-jets being shifted to larger values and light-flavour jets and c-jets being shifted to smaller values, compared to the distributions of the bb-DIPS model. The separation values are listed in Table 7.10. All separation values achieved with bb-DIPS-Loose increase compared to bb-DIPS, with the largest improvement being the separation between bb-jets and c-jets which increases from $(50.6 \pm 0.4)\,\%$ to $(54.9 \pm 0.4)\,\%$. This improvement in the separation in the discriminant D_b translates into a significant improvement in both the light-flavour-jet rejection and the c-jet rejection. The corresponding curves of both rejections as a function of the b-jet efficiency are shown in Figure 7.27. For a b-jet efficiency of 70 %,

Table 7.10.: Separation between the background classes (light-flavour jets and c-jets) and the signal classes (single-b jets and bb-jets) in the D_b distribution. Since D_b is the discriminant for b-tagging with the inclusive b-jets class, both single-b jets and bb-jets are part of the signal.

	Sep(single-b, light)	Sep(single-b, c)	Sep(bb, light)	Sep(bb, c)
bb-DIPS	$(72.0 \pm 0.1)\,\%$	$(39.5 \pm 0.1)\,\%$	$(82.5 \pm 0.5)\,\%$	$(50.6 \pm 0.4)\,\%$
bb-DIPS-Loose	$(74.4 \pm 0.1)\,\%$	$(43.0 \pm 0.1)\,\%$	$(84.2 \pm 0.5)\,\%$	$(54.9 \pm 0.4)\,\%$

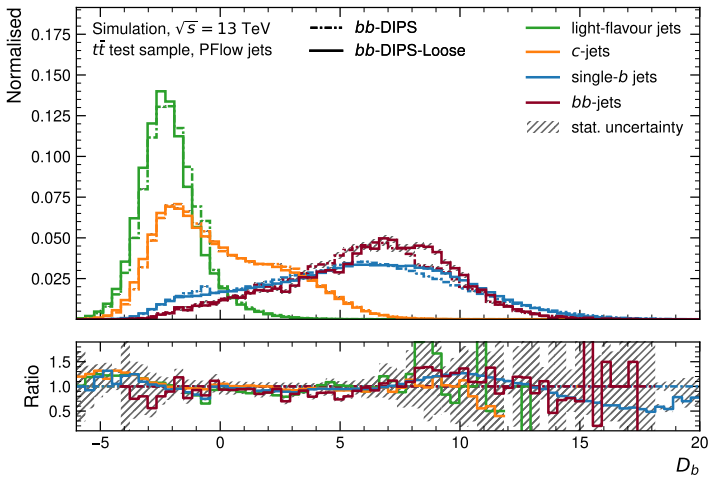
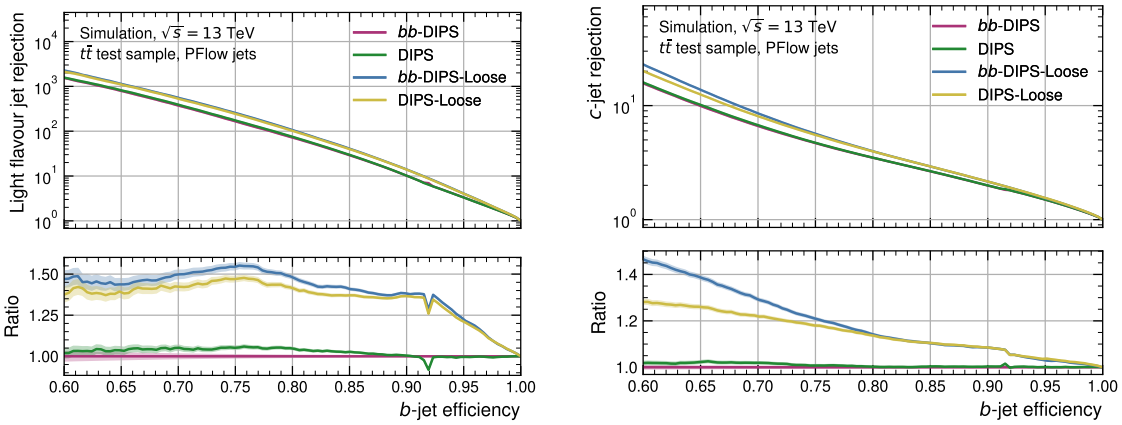


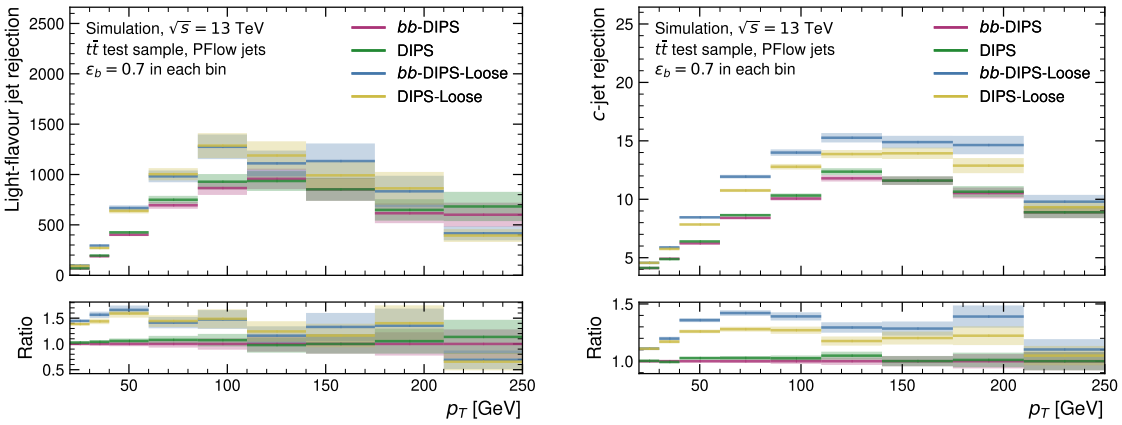
Figure 7.26.: Comparison of the b-tagging discriminant distribution D_b of the bb-DIPS-Loose tagger (solid lines) and the bb-DIPS tagger (dotted lines). The ratio is calculated with respect to bb-DIPS and the f_c value of both models is set to $f_c = 0.005$.

the light-flavour-jet rejection is 50 % larger in bb-DIPS-Loose and the c-jet rejection at this b-tagging working point is 30 % better than with the bb-DIPS model. Furthermore, while introducing the additional bb-jet class to the model decreased the performance when using the standard track selection, the bb-DIPS-Loose model performs even better than DIPS-Loose in both light-flavour-jet rejection and c-jet rejection. It should however be noted, that the presented training of DIPS-Loose is performed with 12M jets (4M per target class) to obtain a fair comparison in terms of the size of the training sample. When using the inclusive b-jet class (DIPS and DIPS-Loose), the training sample size could be increased drastically since it is the bb-jet class that limits the size of the training sample in the case of four target classes. While this is not investigated here, this could lead to improved performance of DIPS and DIPS-Loose. The p_T dependence of both the light-flavour-jet rejection and the c-jet rejection at a 70 % b-jet efficiency is shown in Figure 7.28. The models which use the loose track selection achieve a c-jet rejection which is better over the full p_T range, whereas the light-flavour-jet rejection only improves for $p_\mathrm{T} < 110\,\mathrm{GeV}$.



(a) **(b)**

Figure 7.27.: Light-flavour-jet rejection **(a)** and c-jet rejection **(b)** as a function of the b-jet efficiency for bb-DIPS (magenta), DIPS (green), bb-DIPS-Loose (blue) and DIPS-Loose (yellow). The b-jet efficiency refers to the efficiency of the inclusive b-jet class which includes both single-b jets and bb-jets. The lower panels show the ratio for each algorithm with respect to the bb-DIPS line (magenta). The f_c value in the discriminant D_b is set to $f_c = 0.005$ for all models. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.



(a) **(b)**

Figure 7.28.: Light-flavour-jet rejection **(a)** and c-jet rejection **(b)** calculated in bins of the jet p_{T} for bb-DIPS (magenta), DIPS (green), bb-DIPS-Loose (blue) and DIPS-Loose (yellow). Each bin shows the corresponding rejection for a b-jet efficiency of 70 % in that bin. The b-jet efficiency refers to the inclusive b-jet class which includes both single-b jets and bb-jets. The lower panels show the ratio for each algorithm with respect to the bb-DIPS line (magenta). The f_c value in the discriminant D_b is set to $f_c = 0.005$. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

7.4.4. Performance in bb-Jet Identification

The distribution of the single-b jet discriminant $D_{\text{single-}b}$ is shown in Figure 7.29. The separation of single-b jets and bb-jets in $D_{\text{single-}b}$ increases from $(46.4 \pm 0.4)\,\%$ to $(47.4 \pm 0.4)\,\%$ when using the loose track selection.

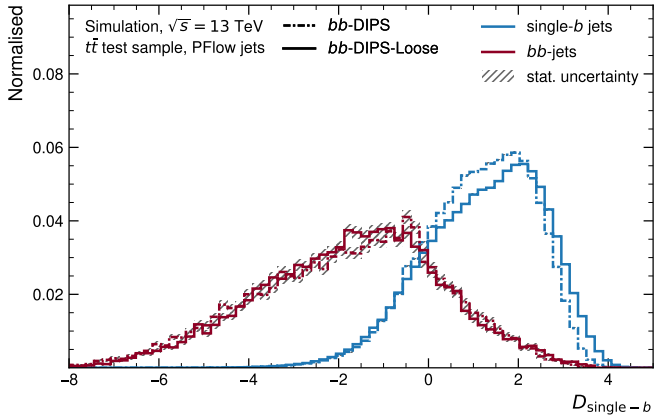
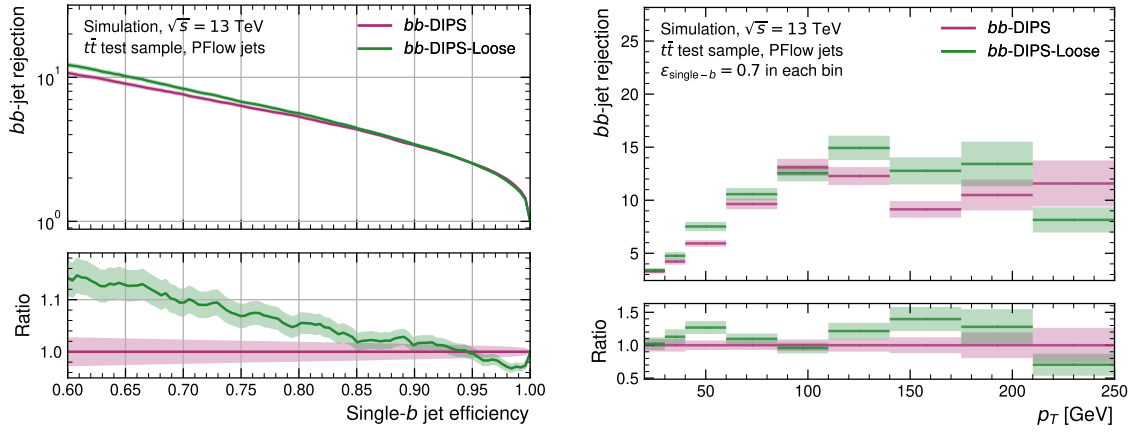


Figure 7.29.: Comparison of the discriminant distribution $D_{\text{single-}b}$ of the bb-DIPS tagger (dash-dotted line) and the bb-DIPS-Loose tagger (solid line).

This leads to a small increase in the bb-jet rejection for single-b jet efficiencies smaller than 85 %, which is shown in Figure 7.30a. At a single-b jet efficiency of 70 %, a bb-jet rejection of 8.4 ± 0.2 is achieved with bb-DIPS-Loose, which is a 10 % improvement compared to bb-DIPS. The p_{T}-dependence of the bb-jet rejection is shown in Figure 7.30b for both bb-DIPS and bb-DIPS-Loose. No clear trend can be seen in this comparison. However, the loose track selection leads to either similar or better performance over the full p_{T} range except for the last bin.



(a) **(b)**

Figure 7.30.: bb-jet rejection of bb-DIPS-Loose and bb-DIPS: **(a)** the bb-jet rejection as a function of the single-b jet efficiency and **(b)** the bb-jet rejection at a 70 % single-b jet efficiency for different bins in p_{T}. The cut value in $D_{\text{single-}b}$ is chosen such that a single-b jet efficiency of 70 % is obtained in each bin separately. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

8. The Extended bb-DL1d Tagger

In order to obtain the extended bb-DL1d tagger, the architecture of the DL1d algorithm is adjusted both in the input layer and in the output layer. As illustrated in Figure 8.1, instead of using the three-dimensional output of the DIPS tagger, the output from the bb-DIPS-Loose tagger is fed into the bb-DL1d algorithm joined by the other variables introduced in subsection 6.2.2. This is shown in Figure 8.1 with the yellow nodes. Furthermore, the output layer is modified the same way as for bb-DIPS-Loose, which means that instead of having three output nodes, the bb-DL1d tagger has four, which is indicated with the pink nodes in Figure 8.1.

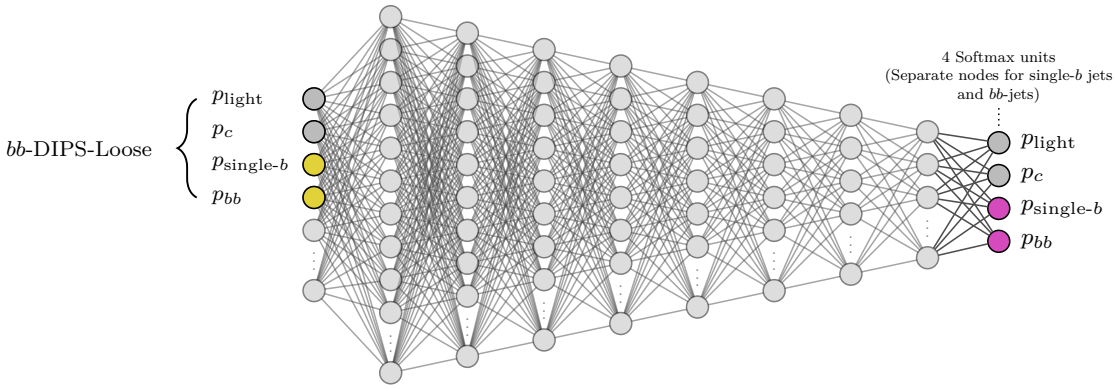
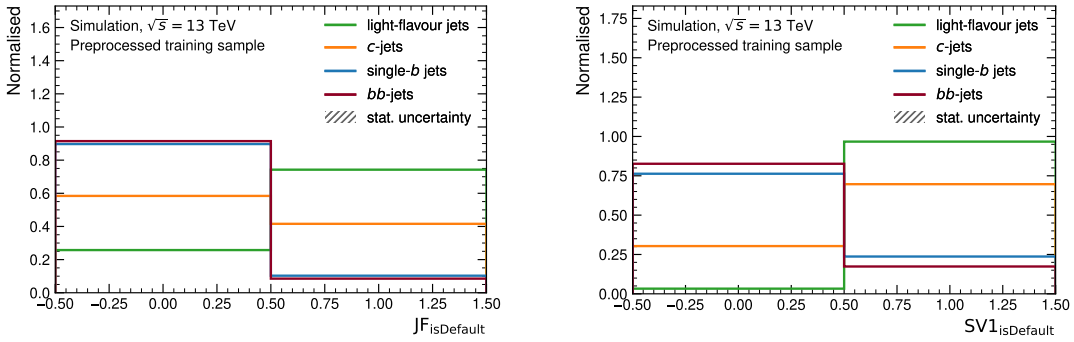


Figure 8.1.: Illustration of the architecture of the bb-DL1d algorithm. In comparison to the default DL1d architecture, bb-DL1d takes the output of bb-DIPS-Loose as input, which means that there are the two inputs $p_{\text{single-}b}$ and p_{bb} from bb-DIPS-Loose, instead of only p_b from DIPS (indicated in yellow). The output of bb-DL1d also features a dedicated $p_{\text{single-}b}$ and p_{bb} node (indicated in pink) instead of the p_b node in the default DL1d architecture.

8.1. bb-DL1d Input Variables

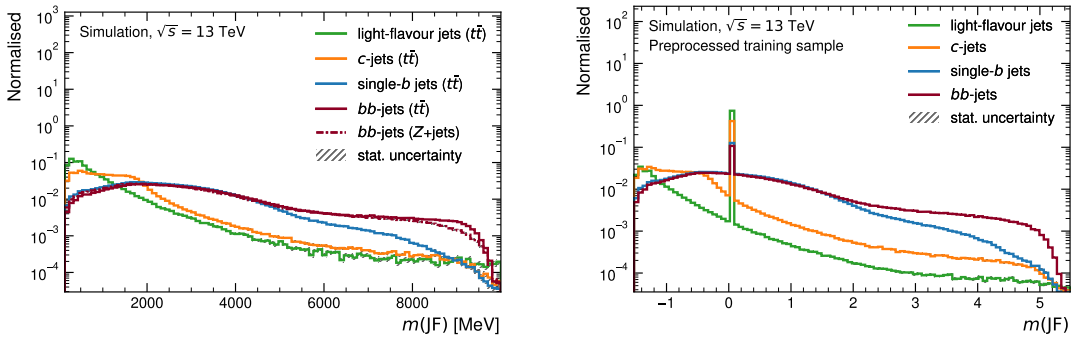
The non-preprocessed input variables of the bb-DL1d algorithm are shown in the appendix in figures A.7 and A.8 and the definition of all variables can be found in Table 6.3. As mentioned in subsection 6.2.2, the JetFitter and SV1 algorithms do not always succeed in the reconstruction of displaced vertices. If no displaced vertices are reconstructed, the variables that are obtained from these algorithms are filled with default values. The information whether a jet is filled with default values for either JetFitter or SV1 is stored in the binary variables $\text{JF}_{\text{isDefault}}$ and $\text{SV1}_{\text{isDefault}}$, which are also handed as an input variable to the algorithm. The distributions of the two variables are shown in Figure 8.2 for the training sample. Since both JetFitter and SV1 are algorithms which reconstruct displaced vertices, mostly light-flavour jets are filled with default values. In c-jets, both JetFitter and SV1 succeed more often, resulting in a smaller fraction of jets filled with default values compared to light-flavour jets. Both single-b jets and bb-jets have comparably small fractions of jets where the algorithms fail to reconstruct displaced vertices which is why $\text{JF}_{\text{isDefault}}$ and $\text{SV1}_{\text{isDefault}}$ is zero for most of these jets. For most variables, the default value is the mean of the inclusive distribution. However, for some variables a more physically motivated default value is chosen. The energy fraction variables $f_E(\text{SV})$, $f_E(2^{\text{nd}})(\text{JF})$ and $f_E(\text{JF})$ are set to zero, since the energy fraction carried



(a) **(b)**

Figure 8.2.: Distributions of the binary variables that indicate if the JetFitter or SV1 algorithm returns default values for the given jet. Figure **(a)** shows the distribution of JetFitter and figure **(b)** shows the corresponding distribution of SV1. The distributions correspond to the training sample.

by displaced vertices should be zero if no displaced vertex is found. Furthermore, the default value of the variables $S_{xyz}(\text{JF})$ and $N_{\text{TrkAtVtx}}(2^{\text{nd}})(\text{JF})$ is set to zero, and the default value of $N_{\geq 2\text{-trk vertices}}(\text{JF})$, $N_{\text{TrkAtVtx}}(\text{JF})$, $N_{1\text{-trk vertices}}(\text{JF})$, $N_{2\text{TrkVtx}}(\text{JF})$, $N_{2\text{TrkVtx}}(\text{SV})$ and $N_{\text{TrkAtVtx}}(\text{SV})$ is set to -1. The training sample consists of the same jets that are used for the training of *bb*-DIPS-Loose, including the resampling procedure. Furthermore, all input variables are scaled and shifted according to the standardisation procedure, except for the variables JF$_{\text{isDefault}}$ and SV1$_{\text{isDefault}}$. The resulting distributions of the preprocessed input variables are shown in the appendix in figures A.9 to A.11 for *bb*-DL1d and in figures A.12 to A.14 for DL1d. The distribution of the invariant mass of tracks from displaced vertices $m(\text{JF})$ is shown in Figure 8.3 both for the individual jet classes before the preprocessing (Figure 8.3a) and for the individual jet classes after the preprocessing (Figure 8.3b). The default values are filled during the preprocessing, which is why the spike at $m(\text{JF})$ is



(a) **(b)**

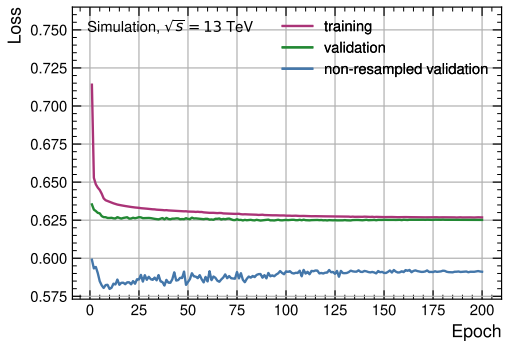
Figure 8.3.: Distribution of the invariant mass of tracks from displaced vertices $m(\text{JF})$ before **(a)** and after **(b)** the preprocessing. The default values are filled during the preprocessing, which is why the spike at $m(\text{JF})$ is not included in figure **(a)**.

not included in Figure 8.3a where only jets for which JetFitter does not fail are shown. While the distribution of $m(\text{JF})$ in Figure 8.3b peaks close to 0 MeV for light-flavour jets, the distribution shifts towards larger values for the other jet classes. The largest values of $m(\text{JF})$ are seen in *bb*-jets, which is consistent with the fact that these jets contain tracks from the displaced vertices of two *b*-hadrons.

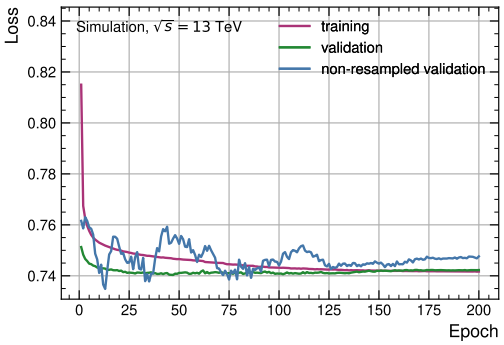
8.2. Training Results

The hyperparameters which are used for the bb-DL1d training and the reference DL1d training are shown in Table 8.1. Except for the number of target classes, the only hyperparameter that differs between DL1d and bb-DL1d is the dropout rate. The dropout rate is 4 % in all hidden layers except the first two layers in bb-DL1d. In DL1d, larger dropout rates were needed to achieve good agreement between training and validation loss, which is why the dropout rate is 10 % in the last 6 hidden layers in the DL1d model.

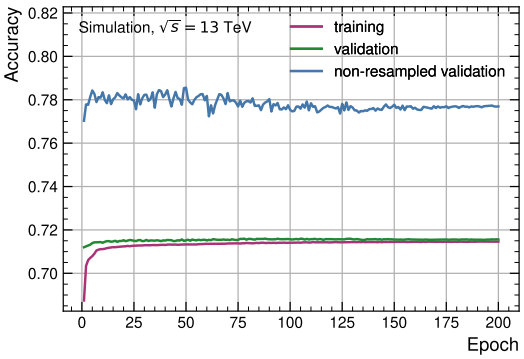
No dropout is applied in the first two hidden layers of either DL1d or bb-DL1d. This choice is made in order to allow the network to use the information of $\mathrm{JF}_{\mathrm{isDefault}}$ and $\mathrm{SV1}_{\mathrm{isDefault}}$, since this information is needed to correctly process the remaining jet variables related to JetFitter and SV1. Combinations with other dropout rates were tested as well, but were found to lead to worse agreement between the training and validation sample or worse performance in general. The accuracy and loss of the bb-DL1d model are shown in Figure 8.4 as a function of the training epoch, showing good agreement between the training and the hybrid validation sample. The model corresponding to the last training epoch is chosen for evaluation.



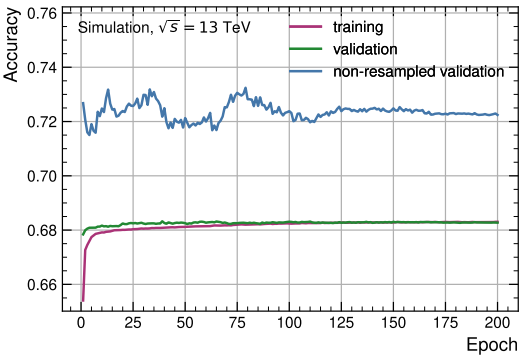
(a) DL1d loss



(b) bb-DL1d loss



(c) DL1d accuracy



(d) bb-DL1d accuracy

Figure 8.4.: Training metrics of the DL1d and bb-DL1d training: **(a)** the loss of the DL1d training and **(b)** the loss of the bb-DL1d training as a function of the training epoch. Both figures show the loss for the training sample (magenta) and for the two different validation samples (green and blue).

Table 8.1.: Hyperparameters of the DL1d and bb-DL1d model.

Parameter	DL1d	bb-DL1d
Target classes	light-flavour jets c-jets b-jets	light-flavour jets c-jets single-b jets bb-jets
Output nodes	$[p_{\text{light}}, p_c, p_b]$	$[p_{\text{light}}, p_c, p_{\text{single-}b}, p_{bb}]$
Size of training sample	12M (4M per class)	16M (4M per class)
$N_{\text{nodes}}^{\text{hidden}}$	[256, 128, 60, 48, 36, 24, 12, 6]	
Dropout rates	0.1 in the last 6 hid. layers	0.04 in the last 6 hid. layers
Learning rate	0.0005	
Batch size	15000	
Batch normalisation	✖	
Learning rate reducer	✔	
Training epochs	200	
Optimiser	Adam	
Loss function	Categorical cross entropy	

8.3. Discriminant Optimisation Studies

The raw output of the bb-DL1d tagger is shown in Figure 8.5, showing good separation in each of the output nodes.

For this final tagger, optimisation studies are performed regarding the choice of the discriminant. As introduced in section 6.3, the default b-tagging discriminant is given by

$$D_b = \log\left(\frac{p_{\text{single-}b}}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light}}}\right) . \tag{8.1}$$

This definition of D_b includes the free parameter f_c, which can be tuned to achieve the desired trade-off between light-flavour-jet rejection and c-jet rejection. To investigate the performance of the bb-DL1d tagger for different values of f_c, a scan is performed over different values for this parameter. For each f_c value, the light-flavour-jet rejection and the c-jet rejection are calculated at the 70 % b-tagging working point.

Additionally, the use of the bb-jet probability p_{bb} in the discriminant is investigated as well. Since bb-jets are part of the inclusive b-jet category, including p_{bb} in the numerator in Equation 8.1 could further improve the b-tagging performance.

8.3.1. Variants of the Discriminant Definition

Two different variants of the discriminant D_b are studied. The first variant introduces a new parameter f_{bb} in the same manner as the f_c parameter, with the difference that f_{bb} gives a weight to $p_{\text{single-}b}$ and p_{bb}, instead of p_{light} and p_c. The resulting definition of this

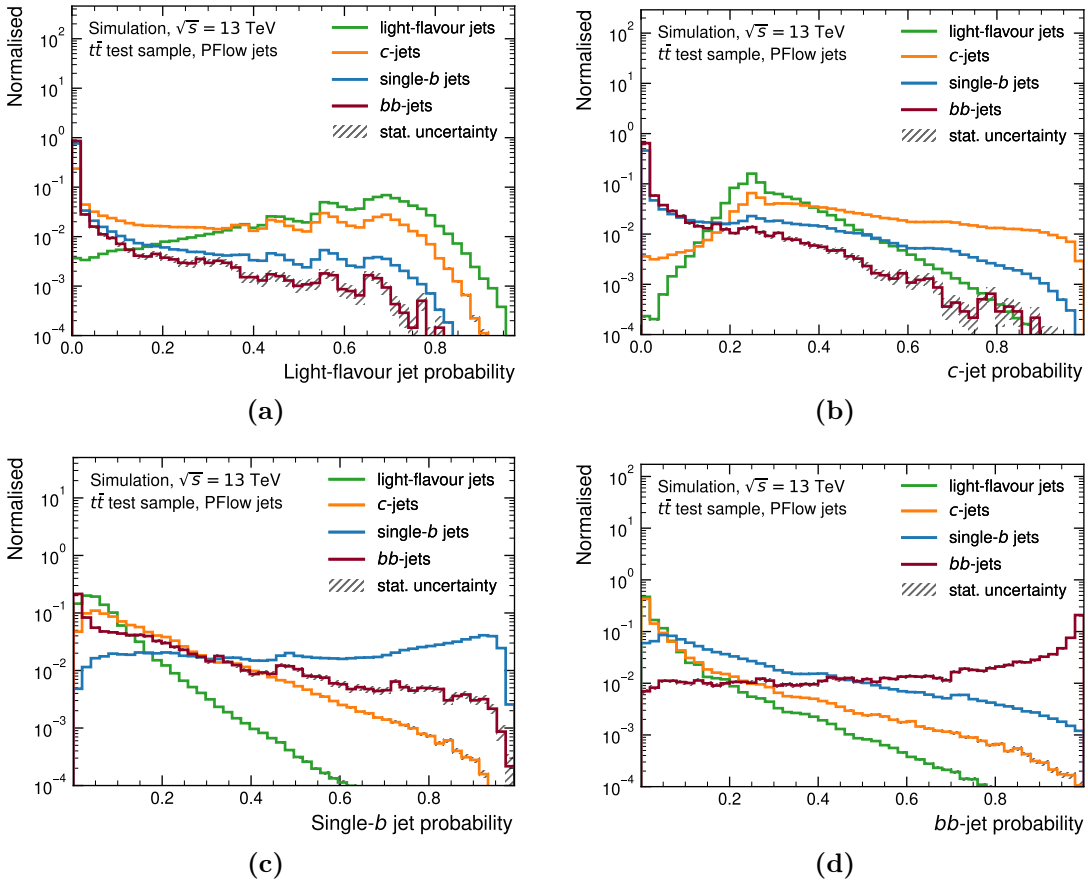


Figure 8.5.: Output distributions of the bb-DL1d tagger: **(a)** the light-flavour jet probability p_{light}, **(b)** the c-jet probability p_c, **(c)** the single-b jet probability $p_{\text{single-}b}$ and **(d)** the bb-jet probability p_{bb}. The distributions are obtained with the $t\bar{t}$ test sample, which consists of 4M jets.

discriminant is given by

$$D_b^{\text{variant-1}} = \log\left(\frac{(1-f_{bb})\cdot p_{\text{single-}b} + f_{bb}\cdot p_{bb}}{f_c\cdot p_c + (1-f_c)\cdot p_{\text{light}}}\right). \tag{8.2}$$

The f_c scan curves for four different values of f_{bb} are shown in Figure 8.6, with different line styles corresponding to different f_{bb} values. More possible values of f_{bb} were tested, but only a few representative steps are illustrated in the plot for better visibility. The f_c scans show that the best performance is achieved with $f_{bb} = 0$, since both the light-flavour-jet rejection and the c-jet rejection values of the corresponding curve are the largest for all f_c values.

The second variant of D_b is to add the p_{bb} prediction to the numerator weighted with the faction f_{bb}, but keeping the weight of p_b fixed to one. The corresponding definition is given by

$$D_b^{\text{variant-2}} = \log\left(\frac{p_{\text{single-}b} + f_{bb}\cdot p_{bb}}{f_c\cdot p_c + (1-f_c)\cdot p_{\text{light}}}\right). \tag{8.3}$$

The same f_c scans are performed as for the first variant of D_b, which are shown in Figure 8.7. Like for the first variant, no improvement is seen when adding the p_{bb} prediction to the b-tagging discriminant. Therefore, the initial definition of D_b is used in the following.

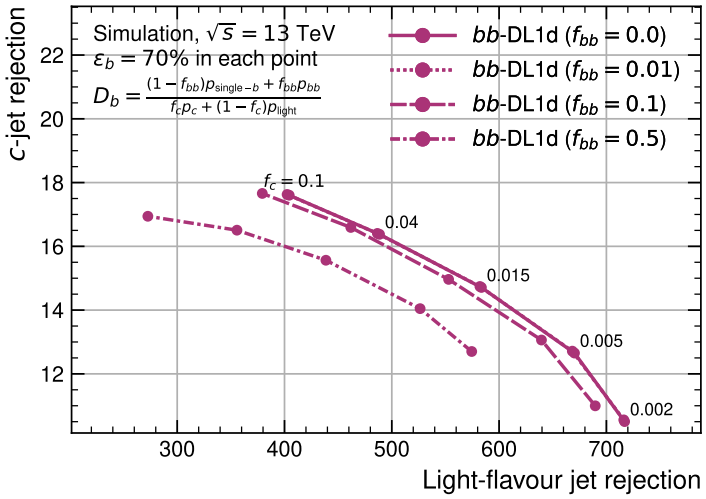


Figure 8.6.: Discriminant optimisation scan for bb-DL1d when including the p_{bb} information ($f_b = (1 - f_{bb})$). The discriminant definition is shown in the upper left of the plot. Each line corresponds to a scan over the f_c value at a fixed f_{bb} value. Each point in the plot represents the c-jet rejection and light-flavour-jet rejection obtained at a 70 % b-jet efficiency with a certain f_c (and f_{bb}) value. The f_c values are drawn in the plot next to the calculated points for the case with $f_{bb} = 0$. The lines represent the direct connection between the calculated points. The dotted line ($f_{bb} = 0.01$) is hidden under the solid line ($f_{bb} = 1$).

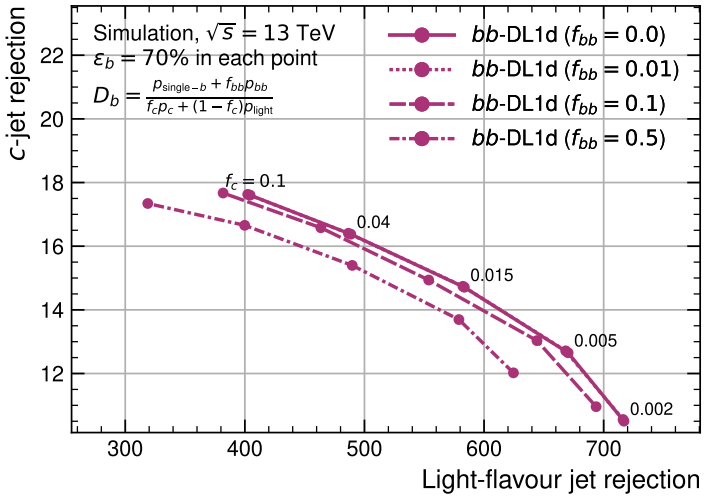


Figure 8.7.: Discriminant optimisation scan for bb-DL1d when including the p_{bb} information but keeping the p_b contribution constant ($f_b = 1$). The discriminant definition is shown in the upper left of the plot. Each line corresponds to a scan over the f_c value at a fixed f_{bb} value. Each point in the plot represents the c-jet rejection and light-flavour-jet rejection obtained at a 70 % b-jet efficiency with a certain f_c (and f_{bb}) value. The f_c values are drawn in the plot next to the calculated points for the case with $f_{bb} = 0$. The lines represent the direct connection between the calculated points. The dotted line ($f_{bb} = 0.01$) is hidden under the solid line ($f_{bb} = 1$).

8.3.2. Optimisation of the f_c Parameter

Using the initial definition of D_b, an f_c scan is performed to compare the bb-DL1d performance to both DL1r and DL1d. The DL1d model used here is the reference training of DL1d with 3 classes and the DL1r model used for comparison is taken from Ref. [46]. The corresponding curves are shown in Figure 8.8. As before, the scan is performed by calculating the light-flavour-jet rejection and c-jetrejection values corresponding to the 70 % b-jet efficiency for the different f_c values.

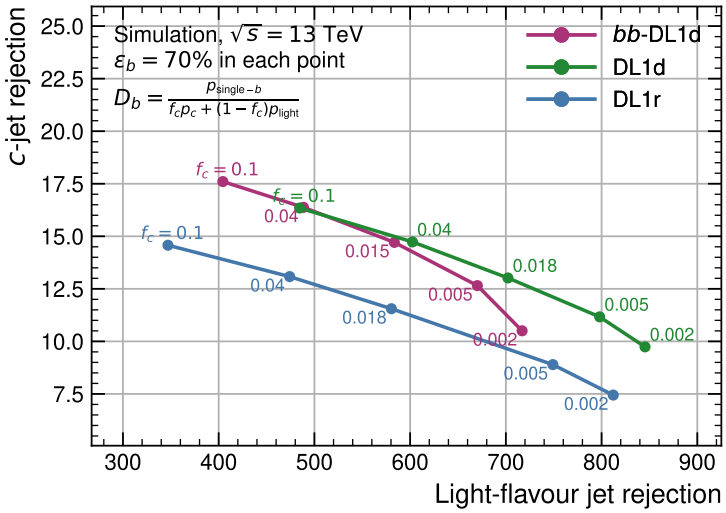
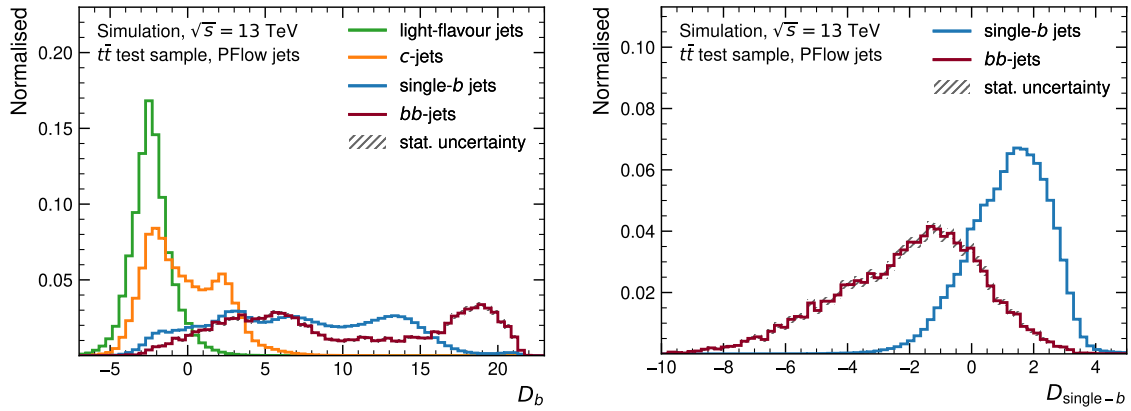


Figure 8.8.: Discriminant optimisation scan for bb-DL1d compared to DL1r and DL1d The discriminant definition is shown in the upper left of the plot. Each line corresponds to a scan over the f_c value. Each point in the plot represents the c-jet rejection and light-flavour-jet rejection obtained at a 70 % b-jet efficiency with a certain f_c. The f_c values are drawn in the plot next to the calculated points. The lines represent the direct connection between the calculated points. For DL1r and DL1d the b-jet probability p_b is used in D_b instead of $p_{\text{single-}b}$.

With f_c set to 0.015, bb-DL1d achieves the same light-flavour-jet rejection as DL1r at $f_c = 0.018$, while being 27 % better than DL1r in c-jet rejection. Furthermore, comparing bb-DL1d at $f_c = 0.015$ with DL1d at $f_c = 0.04$ shows that bb-DL1d can achieve the same c-jet rejection as DL1d while performing only 3 % worse in light-flavour-jet rejection. However, as already mentioned in subsection 7.4.3, while the size of the training sample for bb-DL1d is limited by the comparable small number of available bb-jets, the DL1d model could be trained with a larger training sample. Training the algorithm with more jets in the training sample could lead to an improvement of the DL1d performance, leading to a much better performance than with bb-DL1d. The D_b distribution of bb-DL1d with the optimised $f_c = 0.015$ and the $D_{\text{single-}b}$ distribution are shown in Figure 8.9. Compared to the D_b distribution of bb-DIPS-Loose, bb-DL1d results in are a discriminant distribution where single-b jets and bb-jets are located over a larger range.



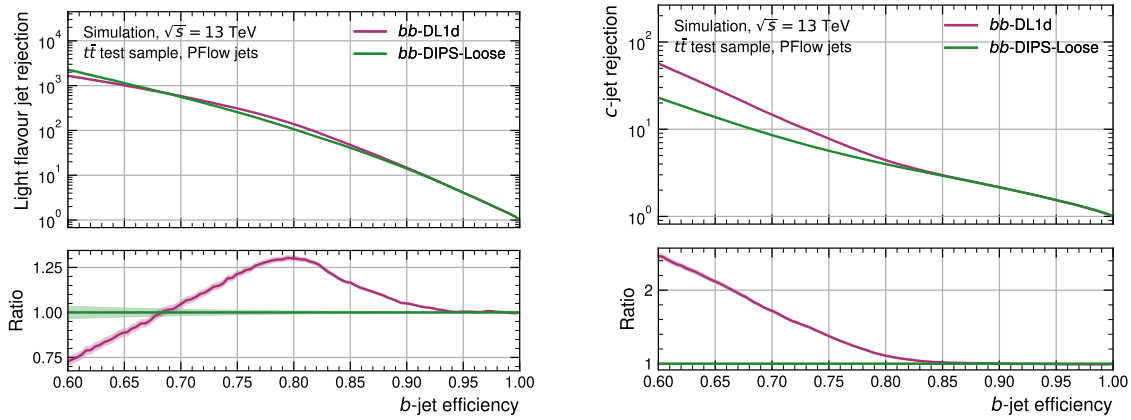
(a) **(b)**

Figure 8.9.: Optimised D_b and $D_{\text{single-}b}$ discriminant distributions obtained with bb-DL1d: **(a)** the b-tagging discriminant D_b with $f_c = 0.015$ and **(b)** the discriminant used for bb-jet identification.

8.4. Performance of the Optimised bb-DL1d Tagger

8.4.1. Comparison to bb-DIPS-Loose

The comparison of the performance of the bb-DL1d tagger to its underlying low-level tagger bb-DIPS-Loose is shown in Figure 8.10. The bb-DL1d algorithm shows significant improvement in c-jet rejection and also in light-flavour-jet rejection for b-jet efficiencies larger than 70 %. At the 70 % b-jet efficiency working point, the c-jet rejection of bb-DL1d is approximately 70 % larger than the c-jet rejection of bb-DIPS-Loose. However, the bb-DL1d performs worse in light-flavour-jet rejection for smaller b-jet efficiencies.



(a) **(b)**

Figure 8.10.: b-tagging performance improvement of bb-DL1d compared to the low-level tagger bb-DIPS-Loose. $f_c = 0.015$ is used for bb-DL1d and $f_c = 0.005$ for bb-DIPS-Loose. The ratio is calculated with respect to bb-DIPS-Loose. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

Furthermore, small improvements are seen in the performance of bb-jet identification. At a single-b jet efficiency of 70 %, bb-DL1d achieves a bb-jet rejection of 8.7 ± 0.3, which is a 4 % improvement compared to bb-DIPS-Loose. The bb-jet rejection is shown as a function of the single-b jet efficiency in Figure 8.11.

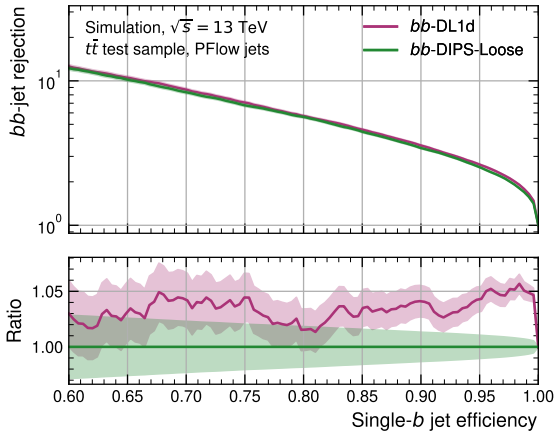
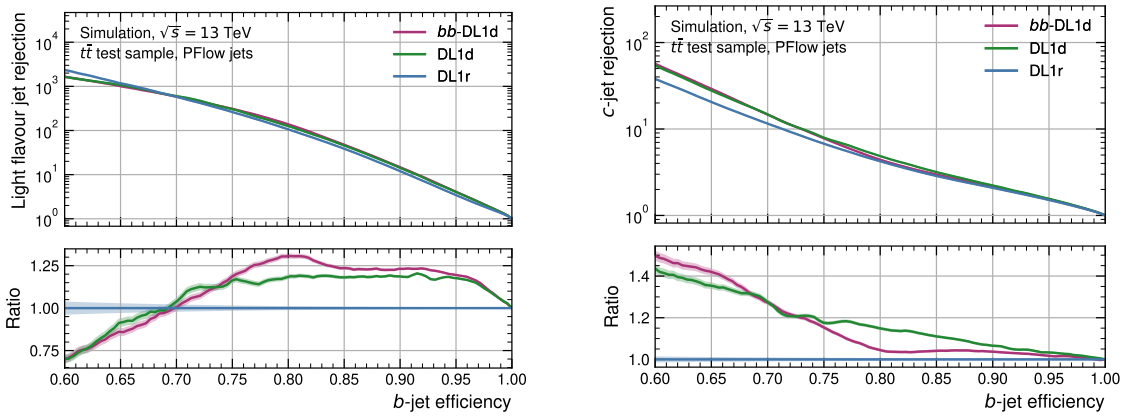


Figure 8.11.: bb-jet identification performance improvement of bb-DL1d compared to the low-level tagger bb-DIPS-Loose. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

8.4.2. Comparison to DL1r and DL1d

The light-flavour-jet rejection and the c-jet rejection of bb-DL1d, DL1d and DL1r is shown in Figure 8.12. The values used for f_c are $f_c = 0.015$ for bb-DL1d, $f_c = 0.018$ for DL1r and 0.04 for DL1d. With this choice, the light-flavour-jet rejection is the same for bb-DL1d and DL1r at a 70 % b-jet efficiency and the c-jet rejection is the same as the one of DL1d at this working point. For b-jet efficiencies larger than 70 %, the light-flavour-jet rejection is slightly better in bb-DL1d compared to the light-flavour-jet rejection of DL1d. Both the bb-DL1d and the DL1d tagger perform approximately 25 % and 20 % better than DL1r at a b-jet efficiency of 85 %. The decrease of the light-flavour-jet rejection of both models from the DL1d family towards smaller b-jet efficiencies shows the same trend as in the comparison between bb-DL1d and bb-DIPS-Loose in Figure 8.10. Both bb-DL1d and DL1d are better in c-jet rejection for the full range of the b-jet efficiency. For b-jet efficiencies above 70 %, however, the DL1d model performs better in c-jet rejection.



(a) **(b)**

Figure 8.12.: b-tagging performance comparison of bb-DL1d, DL1r and DL1d. The ratio is calculated with respect to DL1r. The f_c value is set to 0.015 for bb-DL1d, to 0.018 for DL1r and to 0.04 for DL1d. The coloured bands indicate the statistical uncertainties on the rejection, which are calculated using binomial uncertainties.

8.5. Outlook for Potential Applications of bb-DL1d

Using the two-dimensional plane spanned by the two discriminants D_b and $D_{\text{single-}b}$, the two tasks of b-tagging and bb-jet identification could be combined. By first applying a cut on D_b, jets are given a label if they are b-tagged at this working point. The jets that are b-tagged can then be further separated by applying another cut, this time on $D_{\text{single-}b}$. The ideal case of such a two-dimensional cut is shown in Figure 8.13. The cut on D_b (vertical cut in Figure 8.13) separates light-flavour jets and c-jets from single-b jets and bb-jets. The second cut on $D_{\text{single-}b}$ (the horizontal cut in Figure 8.13) further splits the discriminant plane, resulting in the four regions A, B, C and D. In the ideal case, the regions A and C contain mostly light-flavour jets and c-jets, single-b jets are expected to be located in region B and bb-jets in region D.

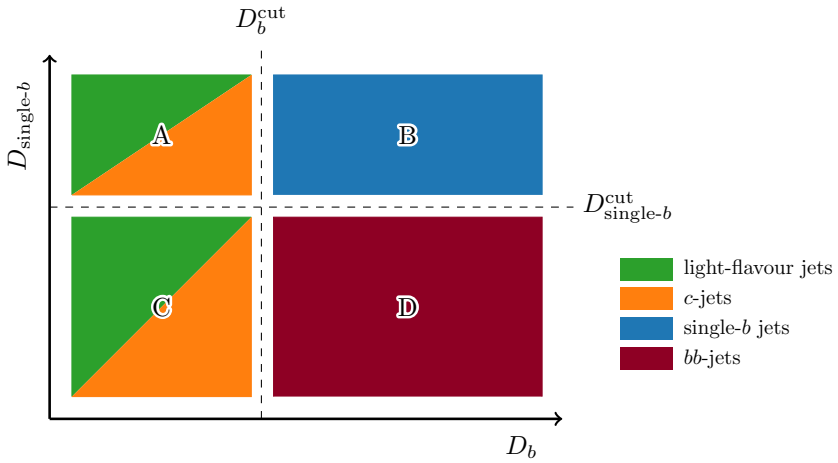


Figure 8.13.: Illustration of the ideal case of a 2D-discriminant, spanned by D_b and $D_{\text{single-}b}$. The colour of the different regions indicates the jet flavours that are desired to populate this region.

This two-dimensional plane is shown with a contour plot in Figure 8.14. Equipotential lines are shown for each jet category individually in the 2D plane, and the projections on the x- and y-axis are the 1D distributions of D_b and $D_{\text{single-}b}$, respectively. The cuts on D_b and $D_{\text{single-}b}$ are shown for the 70 % b-jet efficiency and single-b jet efficiency, respectively. For each jet category the fraction of jets populating a certain region are calculated and listed in Table 8.2. More than 99 % of all light-flavour jets and slightly more than 93 % of the c-jets populate the regions A and C. By definition of the chosen cut value, 70 % of the jets from the inclusive b-jets class are located in the regions B and D. While 51.7 %

Table 8.2.: Fraction (in percent) of jets in the different regions of the two-dimensional discriminant.

Jet category	A	B	C	D
light-flavour jets	59.9	0.1	39.9	0.1
c-jets	69.5	5.0	23.7	1.8
single-b jets	18.3	51.7	11.8	18.2
bb-jets	3.4	8.2	17.9	70.5

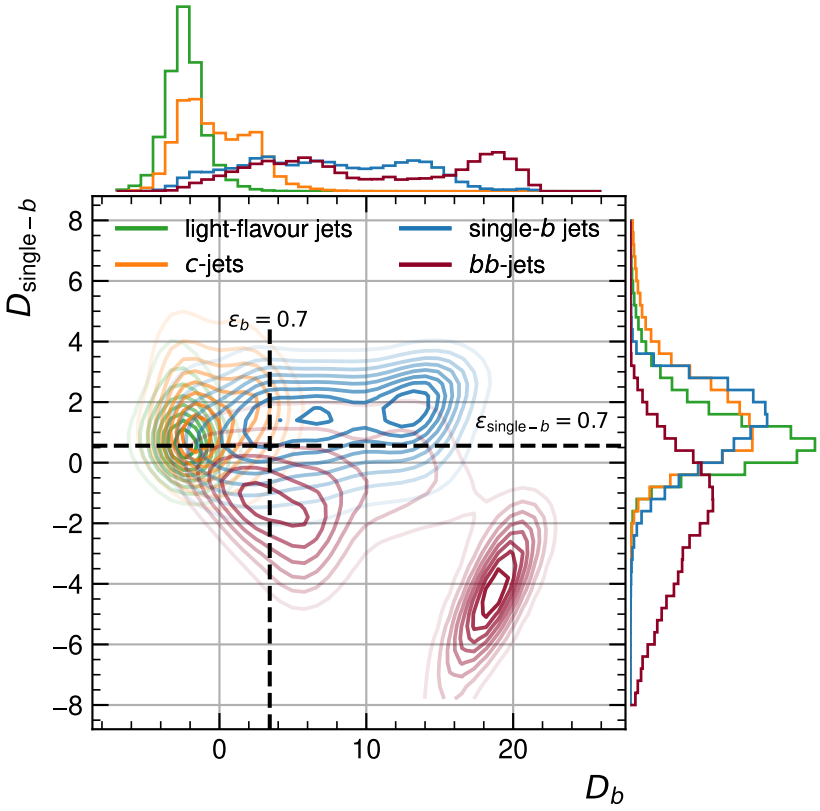


Figure 8.14.: bb-DL1d 2D discriminant distribution spanned by D_b and $D_{\text{single-}b}$. Equipotential lines are shown for each jet category individually in the 2D plane, and the projections on the x- and y-axis are the 1D distributions of D_b and $D_{\text{single-}b}$, respectively. The intensity of the equipotential lines indicate the height of the bins in the 2D-plane. The scale is adjusted for each flavour, such that full intensity corresponds to the maximum value in this histogram. The intensity of the lines increases linearly with the population density. The distributions are obtained with the $t\bar{t}$ test sample, which consists of 4M jets. The optimised f_c value of $f_c = 0.015$ is used.

of the single-b jets are located in region B, only 18.2 % of them are located in region D. Furthermore, the majority of the bb-jets populate region D with 70.5 % of these jets being located in this region.

While there is leakage into the different regions, the majority of the jets populate the dedicated regions in the 2D plane. This shows that a very clear separation between the different jet classes can be obtained with the two-dimensional discriminant. Furthermore, this allows to keep the overall cut procedure regarding b-tagging by separating single-b jets and bb-jets after the cut on D_b is applied.

9. Conclusion

Flavour tagging is an essential tool in particle physics. Usually, flavour-tagging algorithms are trained to distinguish between three target classes, namely light-flavour jets, c-jets and b-jets. In the b-jet category, no distinction is made between jets from one b-hadron (single-b jets) or two hadrons (bb-jets). However, some analyses could benefit from an extension of the three main categories. The $t\bar{t}H(\rightarrow b\bar{b})$ measurement is one such analysis. This measurement suffers from a large irreducible $t\bar{t}$+$b\bar{b}$ background, where a radiated gluon can split into a $b\bar{b}$-pair, potentially leading to a jet that contains two b-hadrons. Therefore, in the context of the search for the $t\bar{t}H(\rightarrow b\bar{b})$ signal, a flavour-tagging algorithm with a dedicated target class for bb-jets is of special interest, since an identification of these jets could improve the background rejection in the analysis.

With this goal in mind, two extended flavour-tagging algorithms were studied in this thesis. The extended algorithms are based on the DL1d algorithm, which is the ATLAS flavour-tagging algorithm for RUN III. DL1d follows a two-level approach, where the output of multiple low-level taggers is combined into a high-level tagger. The low-level algorithm which exploits impact-parameter information of the tracks that are associated with the jet is the Deep-Sets-based DIPS tagger [103]. The output of the DIPS algorithm, as well as the output of two further low-level algorithms, is then processed by the deep-learning-based high-level tagger DL1d. This structure is similar to the DL1r tagger [46] used by ATLAS in Run II, with the main difference being that DL1r uses the RNN-based low-level tagger RNNIP [102] instead of DIPS.

Extended versions of both DIPS and DL1d, denoted as bb-DIPS and bb-DL1d, were successfully trained, and their performance was studied in comparison to the usual approach of having three target classes. In this context, several extensions and improvements have been made to the ATLAS flavour-tagging framework, paving the way for a continuation of the studies presented here. Additionally, as one of its main developers, several contributions were made to the open-source Python module PUMA [112], which enables consistent evaluation of flavour-tagging algorithms in ATLAS.

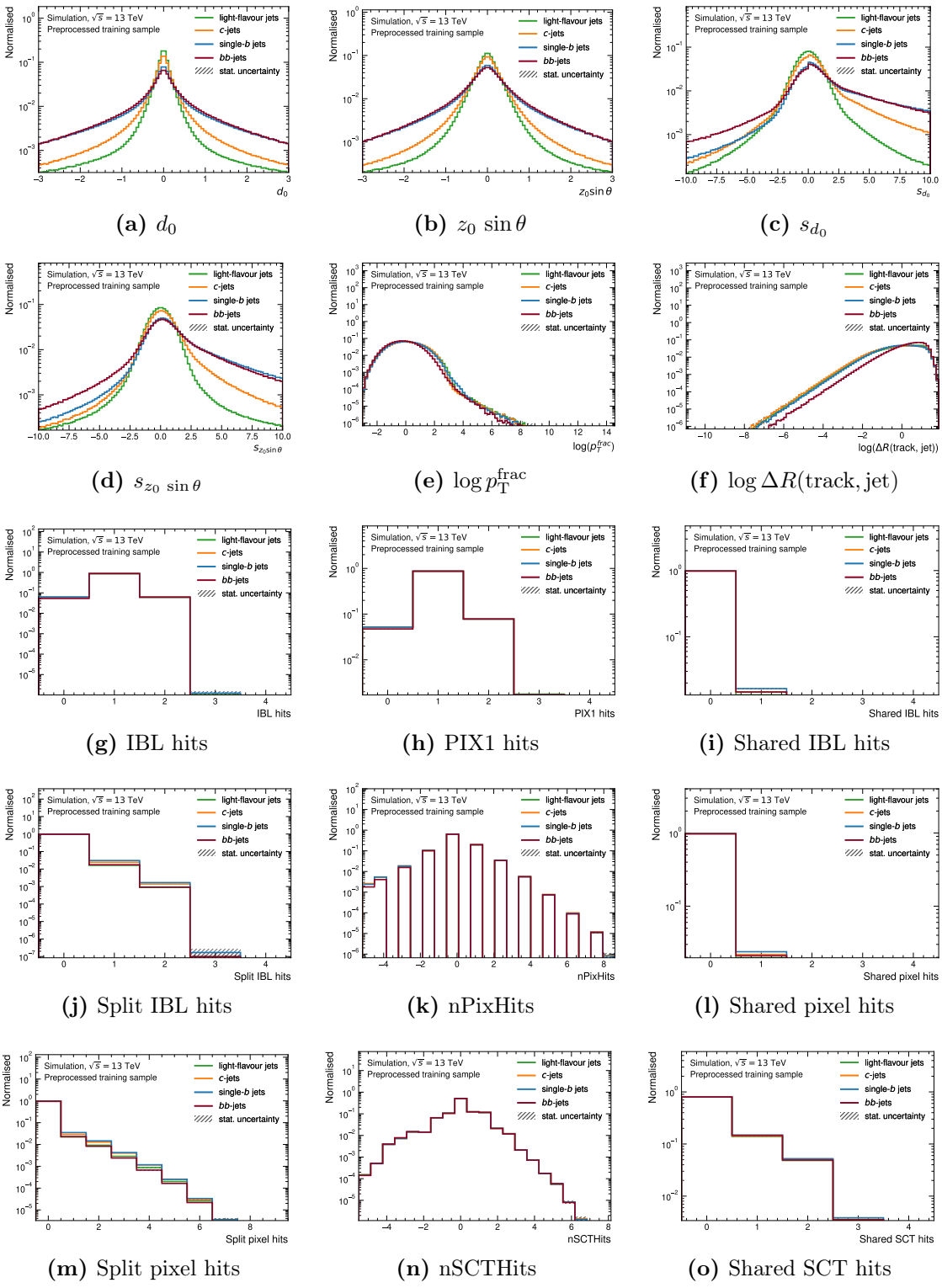
The bb-DIPS algorithm, which uses the information of up to 40 tracks that are associated with the jet, was studied for two different track selections, similar to the studies presented in Ref. [103]. The bb-DIPS and bb-DL1d algorithms are evaluated on jets from MC simulated $t\bar{t}$ events. In order to increase the number of available bb-jets for the training of the algorithm, the training sample is enriched with bb-jets from Z+jets events. As for the DIPS algorithm, the performance of the tagger that uses the looser track selection, bb-DIPS-Loose, clearly outperforms the bb-DIPS model with the tighter track selection. This can be explained by the increase in information that is obtained when loosening the selection criteria on the tracks. The bb-DIPS-Loose tagger achieves a bb-jet rejection of 8.4 ± 0.2 at a 70 % single-b-jet efficiency. Moreover, bb-DIPS-Loose even slightly outperforms the DIPS-Loose baseline model which does not have the dedicated output node for bb-jets, in both light-flavour-jet rejection and c-jet rejection. However, it should be noted that the DIPS-Loose model presented here is trained with the same number of jets per class as the bb-DIPS-Loose model, which is a rather small training sample for the training of an algorithm with the traditional target classes.

With the final bb-DL1d tagger, which uses the output of bb-DIPS-Loose and two further low-level taggers as input, a bb-jet rejection of 8.7 ± 0.3 is achieved at a 70 % single-b-jet efficiency. Besides, comparisons regarding the b-tagging performance have shown that bb-DL1d clearly outperforms the DL1r tagger [46] by achieving a 27 % larger c-jet rejection while still performing equally well as DL1r in light-flavour-jet rejection at a 70 % b-jet efficiency. Comparisons of bb-DL1d to a baseline training of DL1d have shown that introducing a dedicated bb-jet class to this model architecture leads to a small decrease in performance compared to the standard architecture with only three target classes. The bb-DL1d tagger is a substantial advancement of previous studies [4, 5] in this direction, and has been developed from the beginning of this thesis within the ATLAS flavour-tagging framework.

The performance of all models could presumably be improved by using larger training samples, with the limiting factor being the number of bb-jets. This, however, requires the simulation of further events. Additionally, no dedicated hyperparameter optimisation was performed for the models presented this thesis. Given that this could lead to an improvement of the performance of this tagger, this option should be considered for future studies. Furthermore, new developments of flavour-tagging algorithms based on graph neural networks [113] and transformers could be considered in future studies for bb-jet identification as well.

Nonetheless, first studies regarding a two-dimensional cut in the plane spanned by the two discriminants D_b and $D_{\text{single-}b}$ are promising for a successful application of the bb-DL1d tagger. Next steps in this direction could be an optimisation of the 2D cut definition and finally the calibration of the tagger for a dedicated analysis like the ATLAS $t\bar{t}H(\rightarrow b\bar{b})$ analysis.

A. Appendix



(a) d_0 **(b)** $z_0 \sin\theta$ **(c)** s_{d_0}

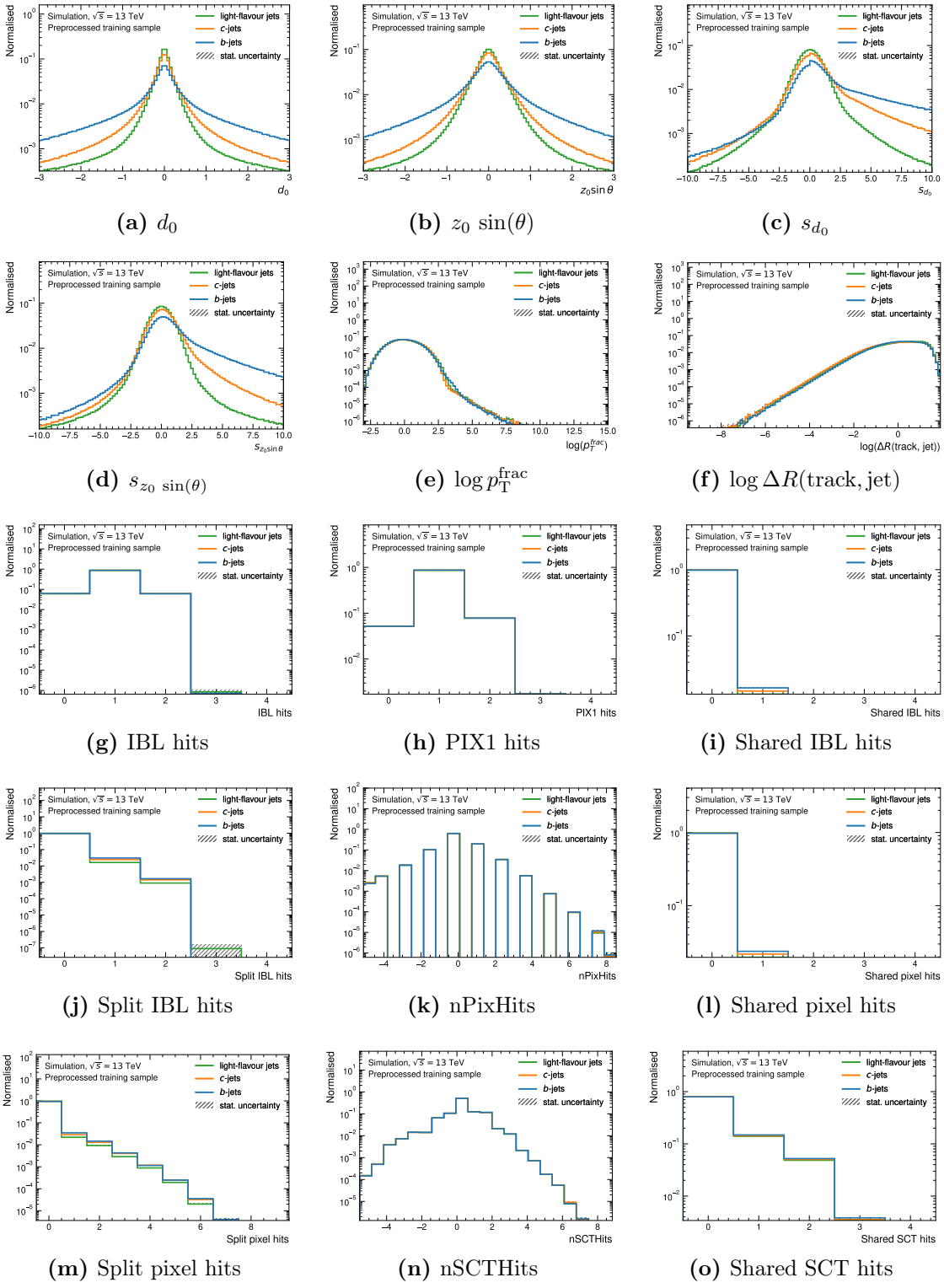
(d) $s_{z_0 \sin\theta}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R(\mathrm{track}, \mathrm{jet})$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure A.1.: Distributions of the preprocessed *bb*-DIPS input variables. These distributions are obtained with the standard track selection. The variable definitions can be found in Table 6.2.



(a) d_0 **(b)** $z_0 \sin(\theta)$ **(c)** s_{d_0}

(d) $s_{z_0 \sin(\theta)}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R(\mathrm{track}, \mathrm{jet})$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure A.2.: Distributions of the preprocessed DIPS input variables. These distributions are obtained with the standard track selection. The variable definitions can be found in Table 6.2.

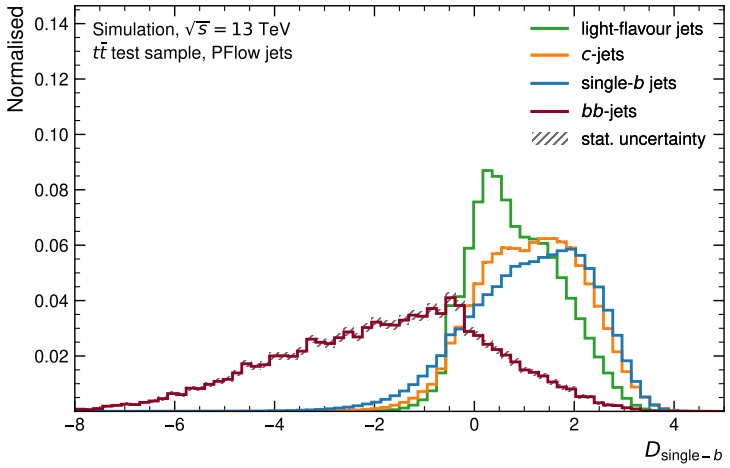
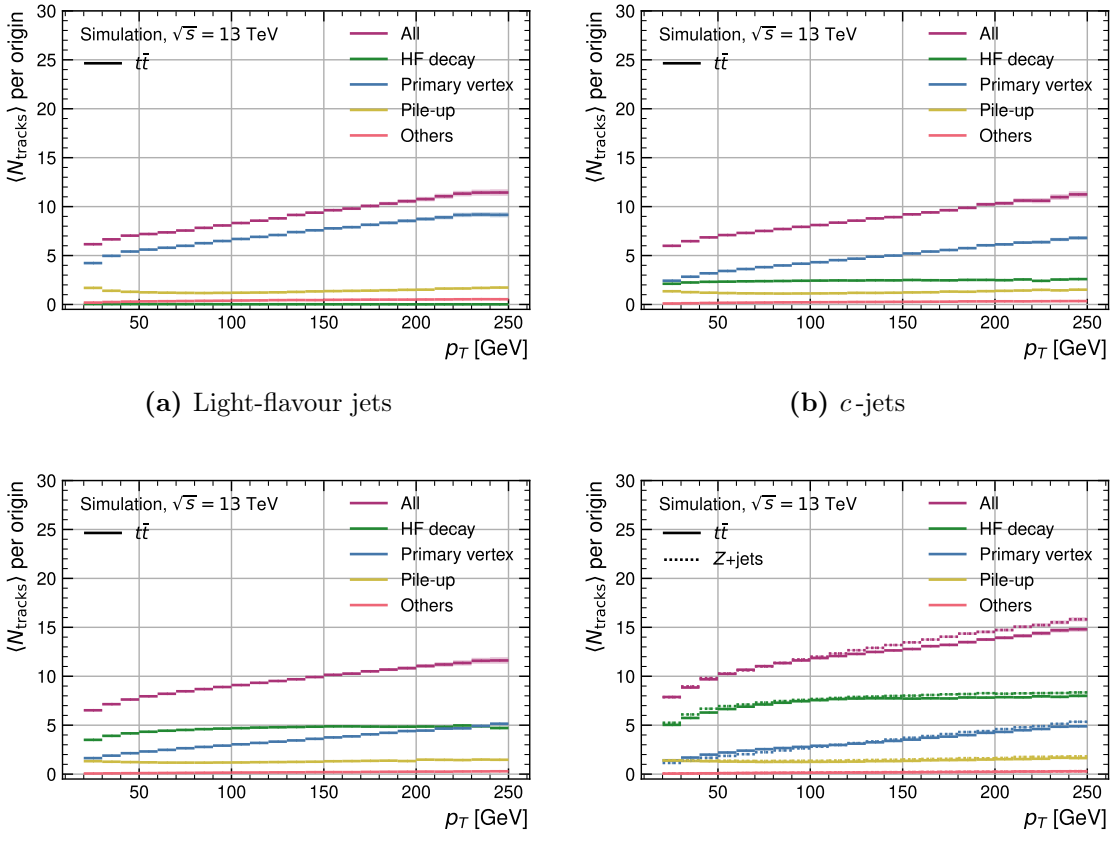


Figure A.3.: $D_{\text{single-}b}$ distribution of the bb-DIPS tagger.



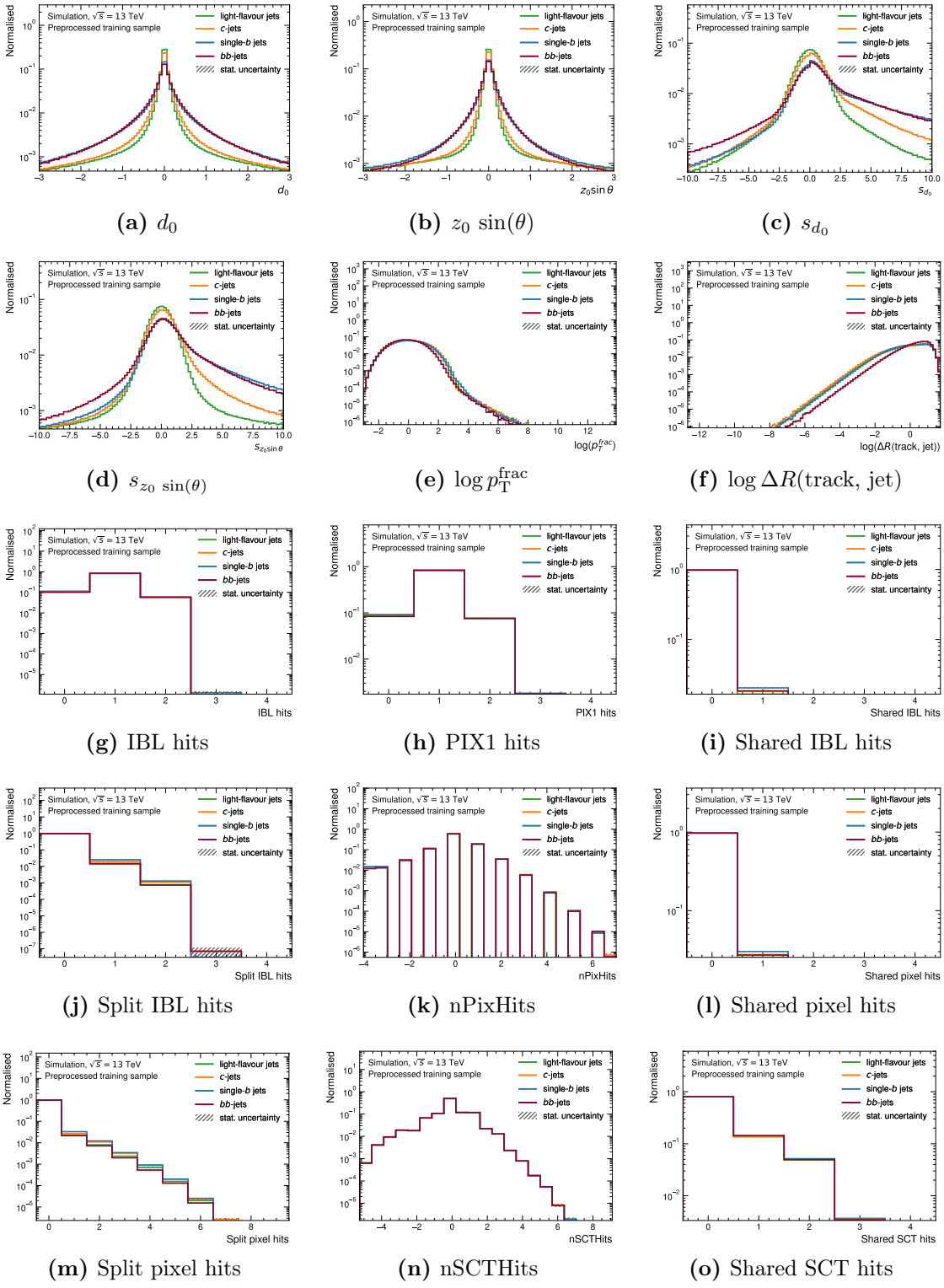
(a) Light-flavour jets

(b) c-jets

(c) Single-b jets

(d) bb-jets

Figure A.4.: Average number of tracks as a function of the jet p_{T} for different track origins individually using the loose track selection. The statistical uncertainty of each bin is indicated as a coloured band, but is too small to be seen in most bins. The inclusive class (all) represents the total average number of tracks per jet for each p_{T} bin.



(a) d_0 **(b)** $z_0 \sin(\theta)$ **(c)** s_{d_0}

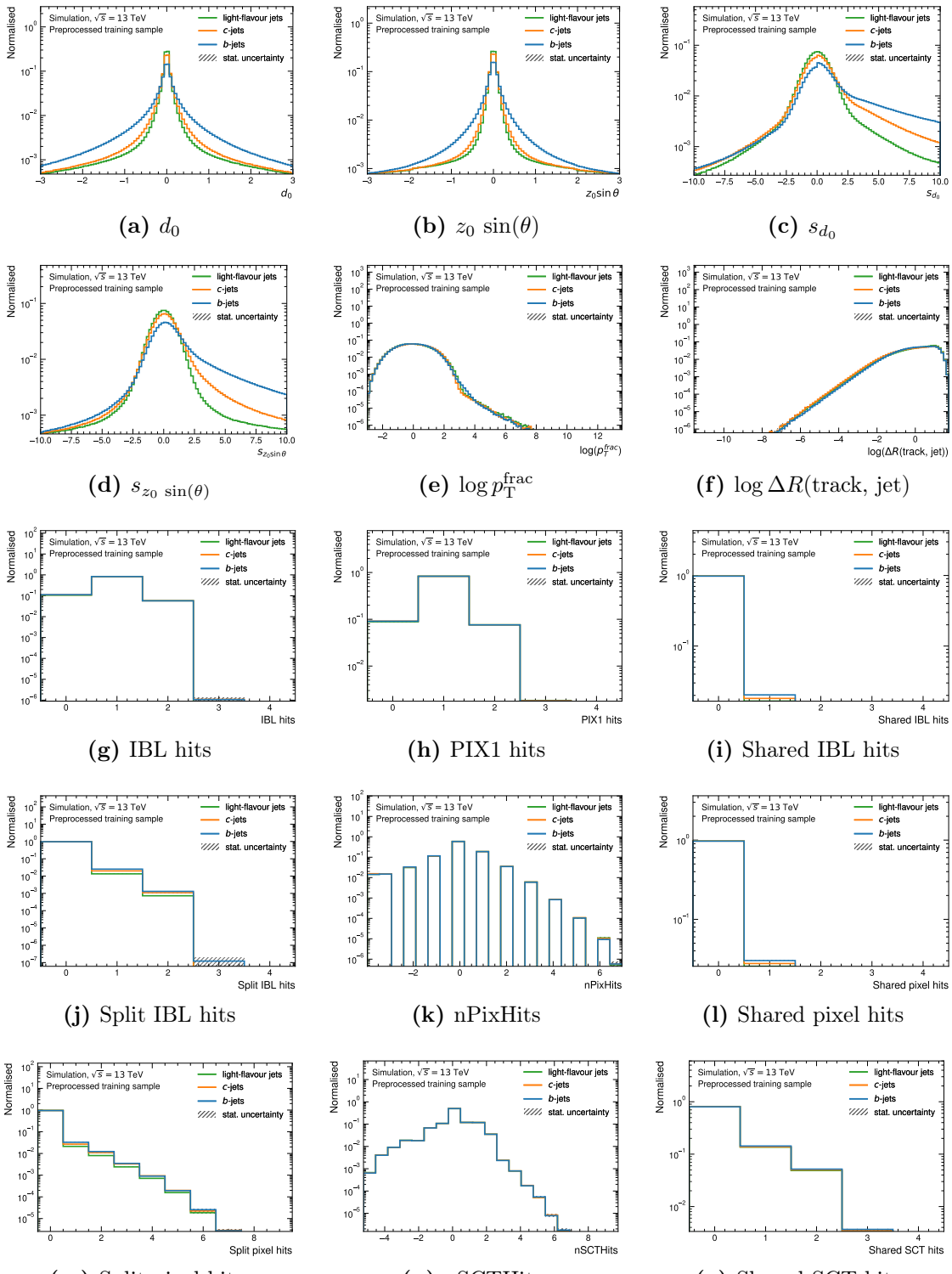
(d) $s_{z_0 \sin(\theta)}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R(\text{track, jet})$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure A.5.: Distributions of the preprocessed *bb*-DIPS-Loose input variables. These distributions are obtained with the loose track selection. The variable definitions can be found in Table 6.2.



(a) d_0 **(b)** $z_0 \sin(\theta)$ **(c)** s_{d_0}

(d) $s_{z_0 \sin(\theta)}$ **(e)** $\log p_{\mathrm{T}}^{\mathrm{frac}}$ **(f)** $\log \Delta R(\text{track, jet})$

(g) IBL hits **(h)** PIX1 hits **(i)** Shared IBL hits

(j) Split IBL hits **(k)** nPixHits **(l)** Shared pixel hits

(m) Split pixel hits **(n)** nSCTHits **(o)** Shared SCT hits

Figure A.6.: Distributions of the preprocessed DIPS-Loose input variables. These distributions are obtained with the loose track selection. The variable definitions can be found in Table 6.2.

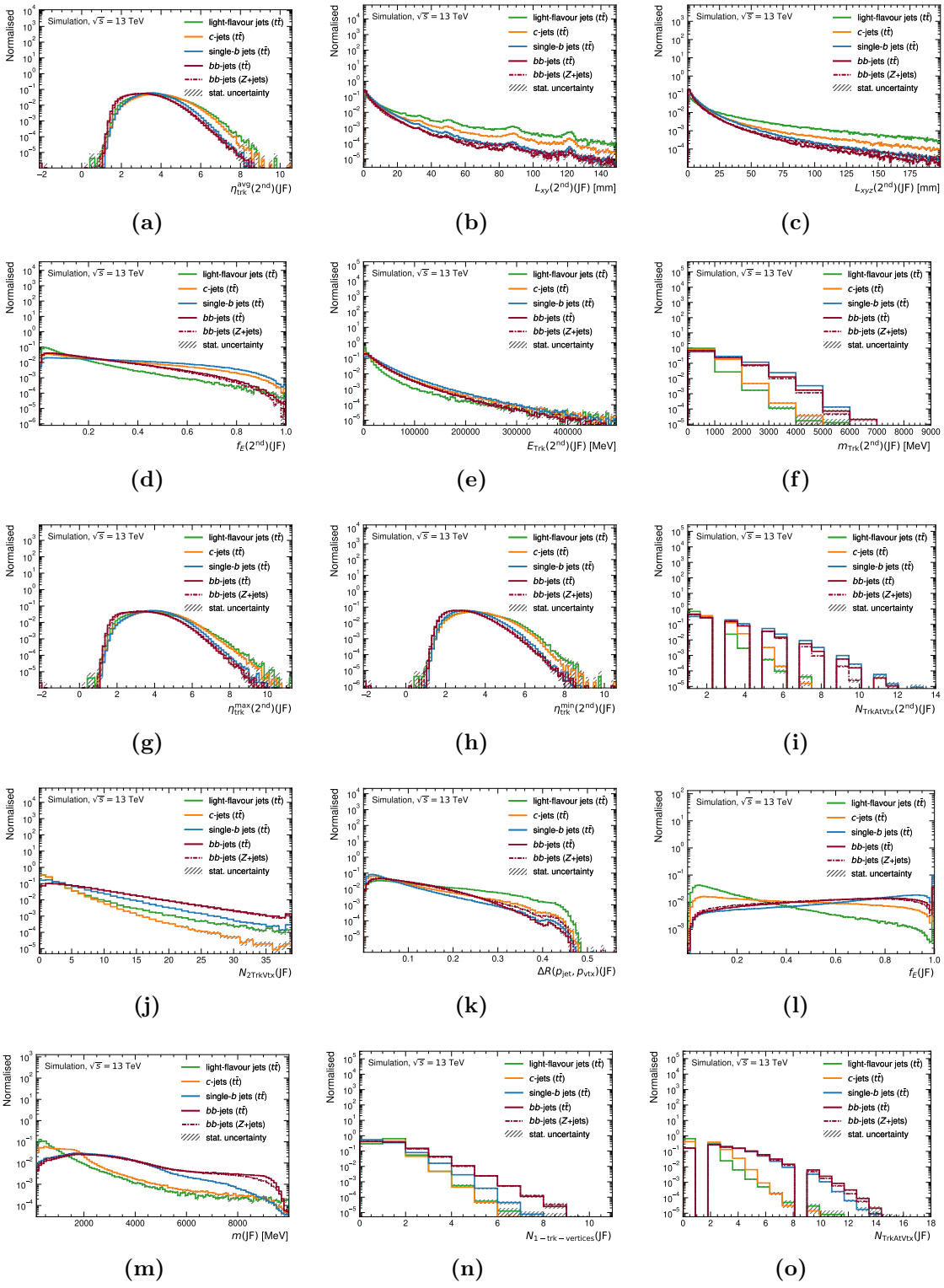


Figure A.7.: Distributions of the non-preprocessed *bb*-DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.

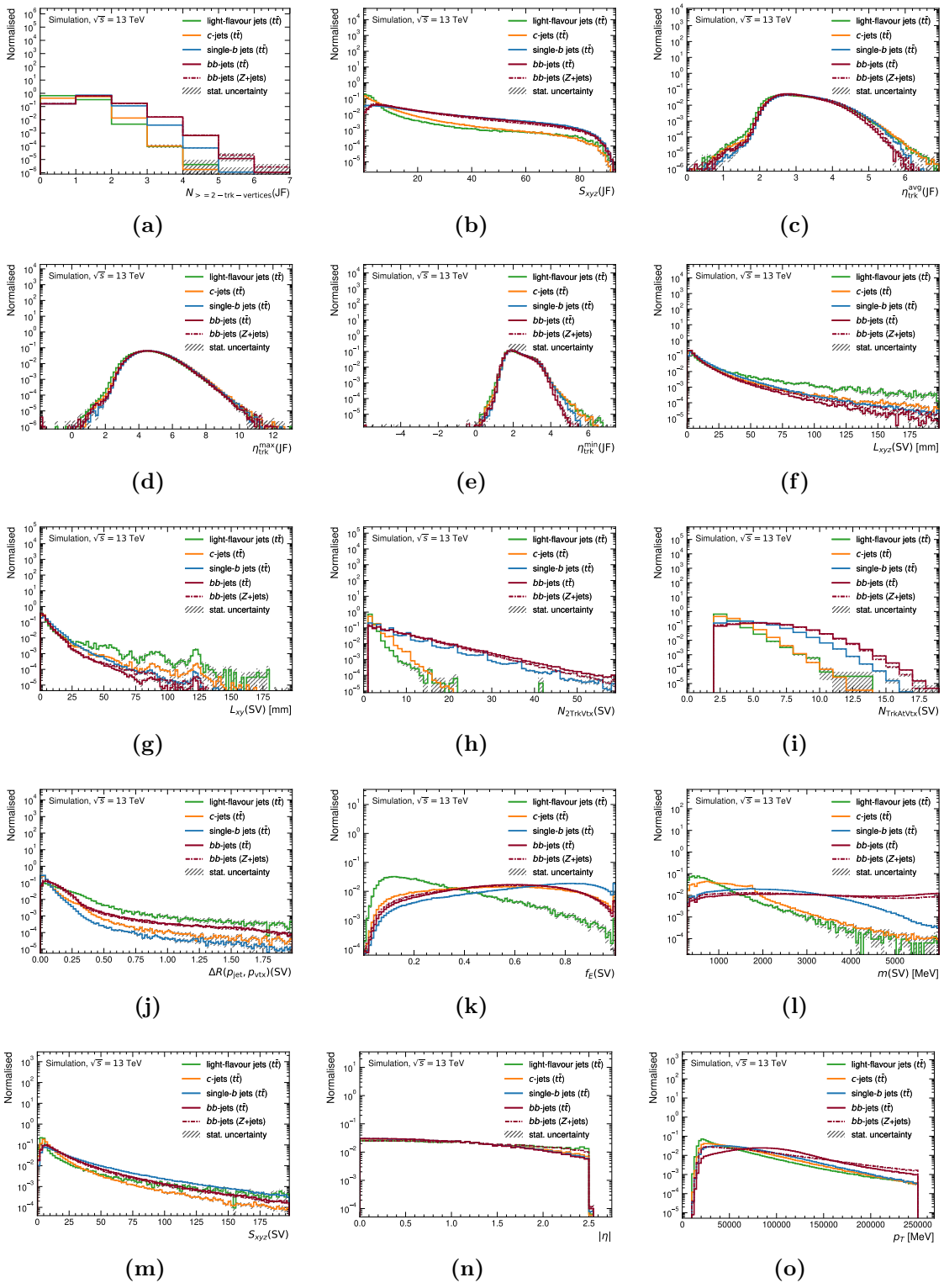
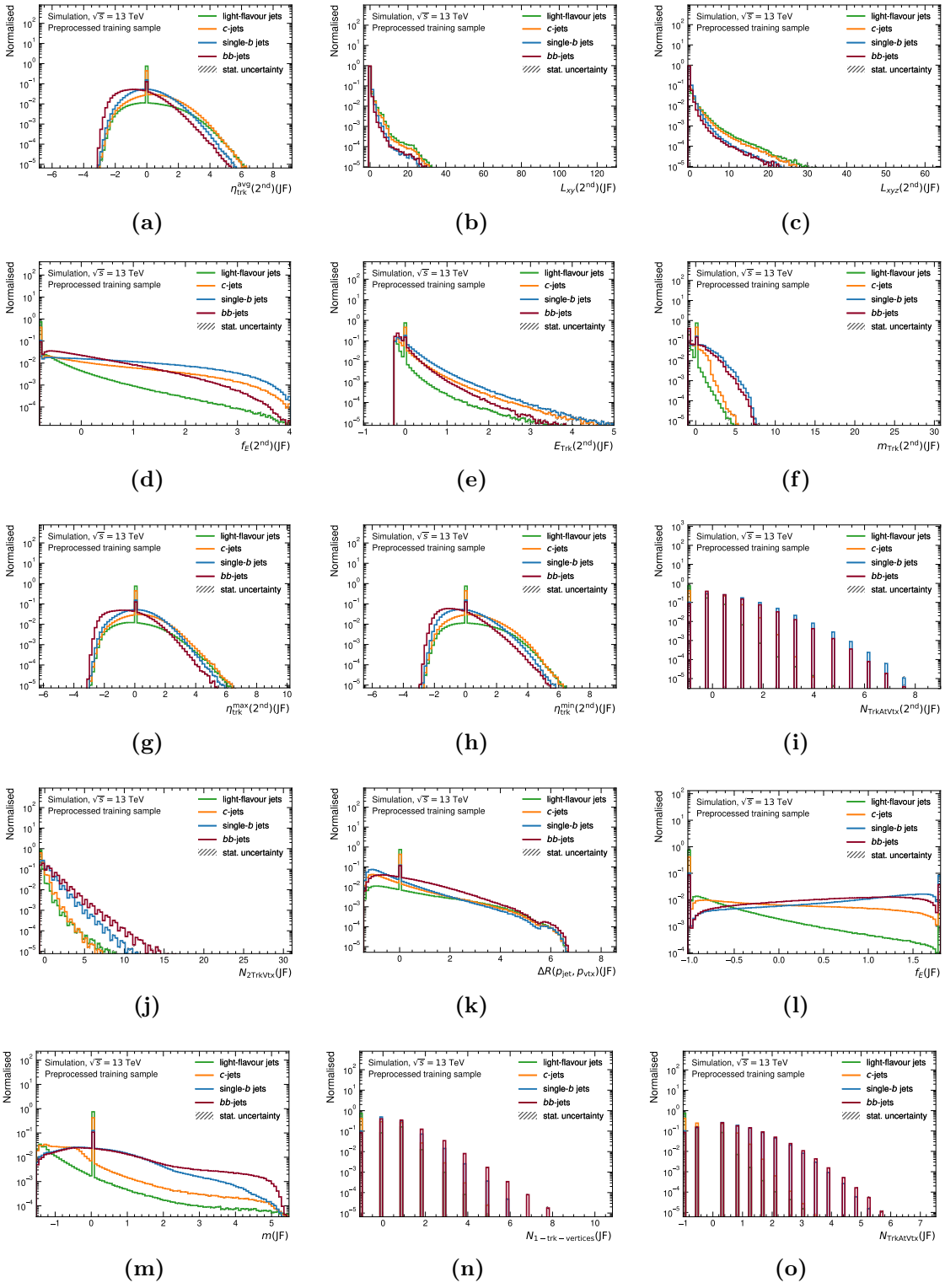


Figure A.8.: Distributions of the non-preprocessed bb-DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.



(a) **(b)** **(c)**

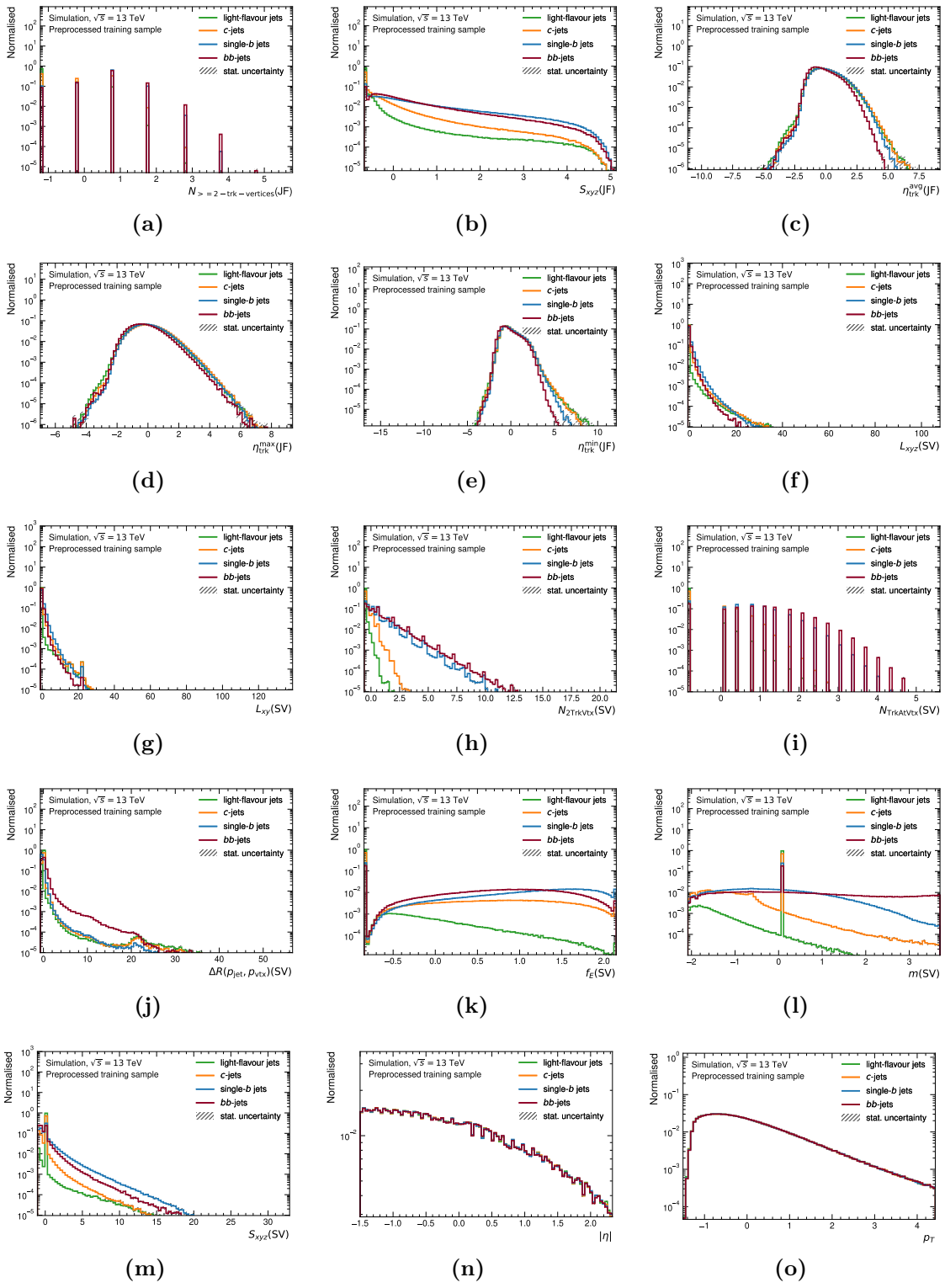
(d) **(e)** **(f)**

(g) **(h)** **(i)**

(j) **(k)** **(l)**

(m) **(n)** **(o)**

Figure A.9.: Distributions of the preprocessed *bb*-DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.



(a) (b) (c)

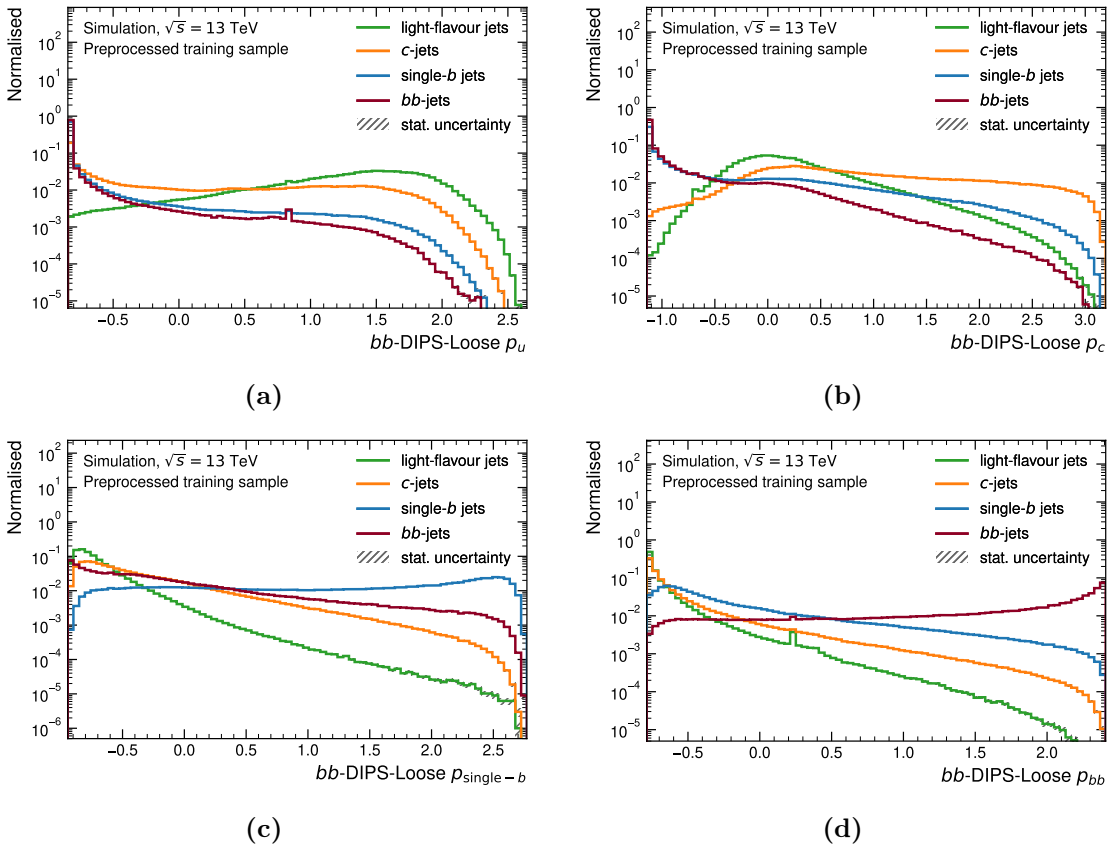
(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure A.10.: Distributions of the preprocessed *bb*-DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.



(a) **(b)**

(c) **(d)**

Figure A.11.: Distributions of the bb-DIPS-Loose output after being preprocessed for the bb-DL1d training.

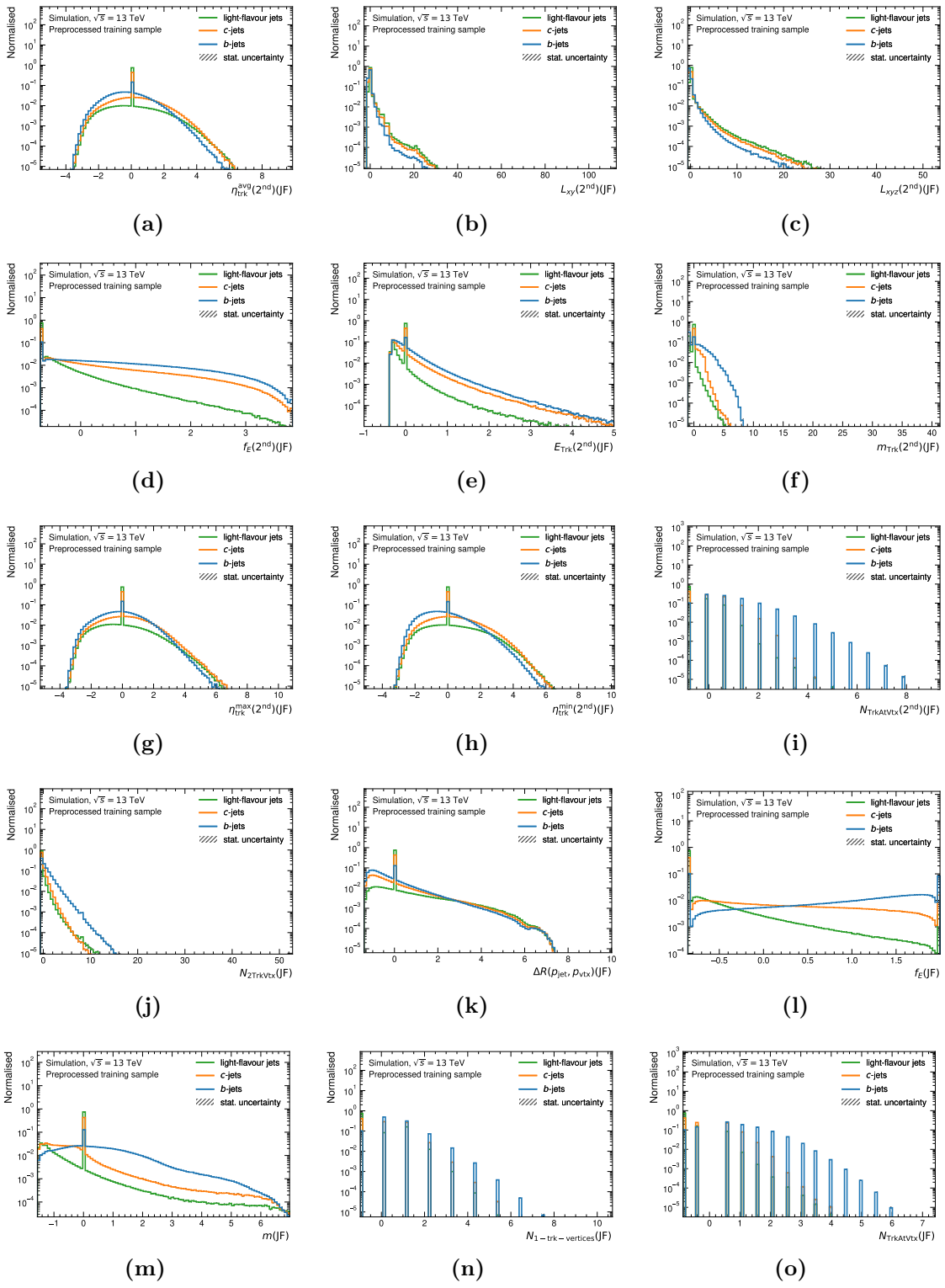


Figure A.12.: Distributions of the preprocessed DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.

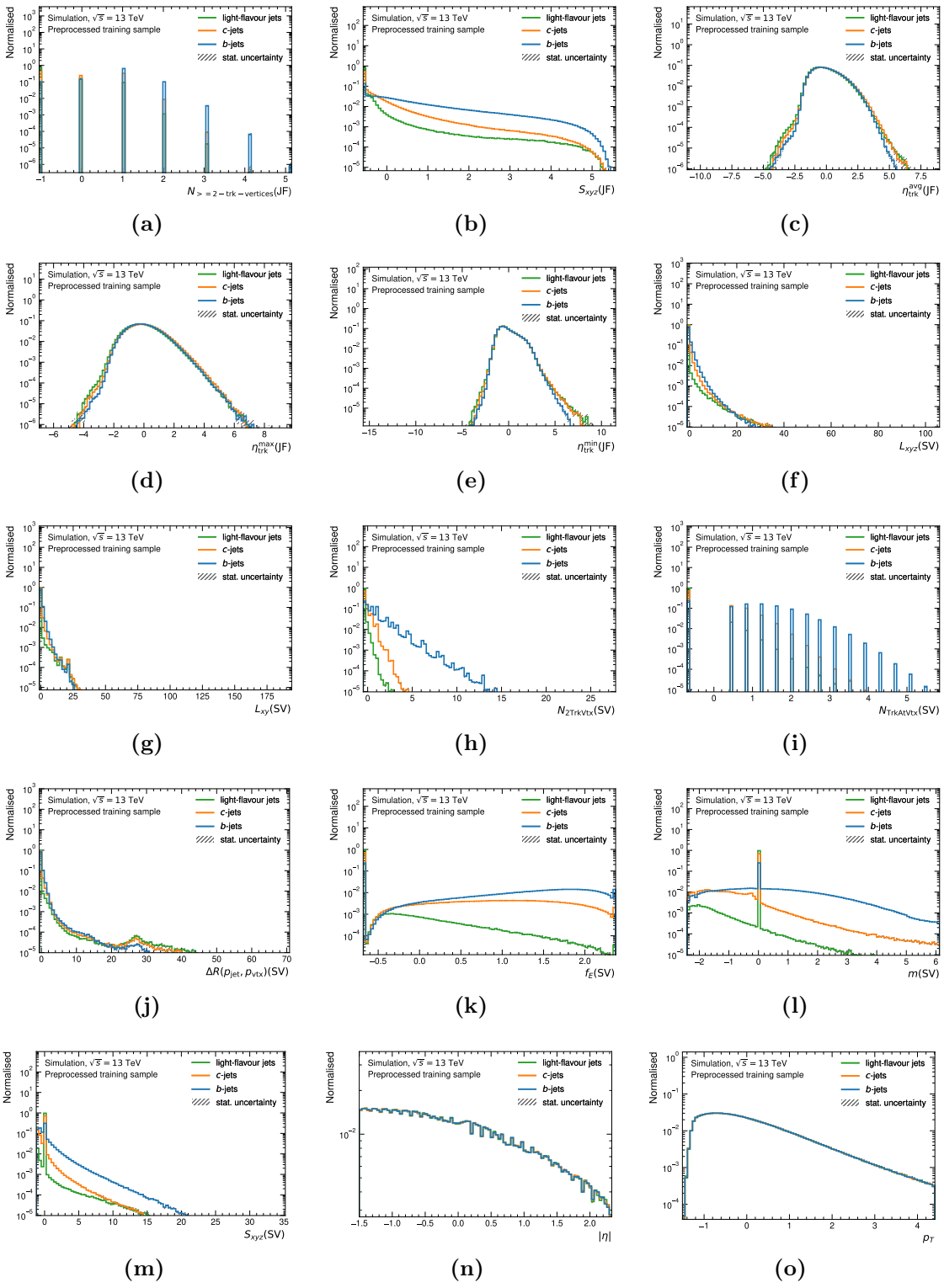
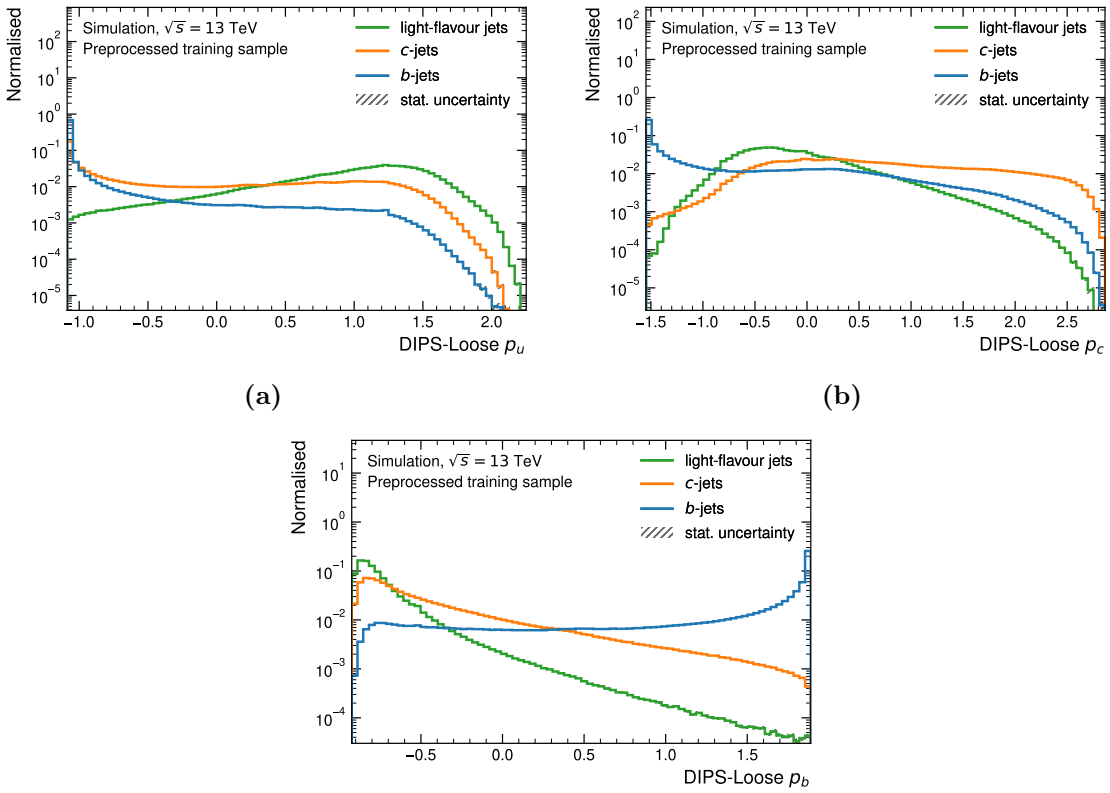


Figure A.13.: Distributions of the preprocessed DL1d input variables. Only showing 15 variables. The variable definitions can be found in Table 6.3.



(a) **(b)**

(c)

Figure A.14.: Distributions of the DIPS-Loose output after being preprocessed for the DL1d training.

B. References

[1] The ATLAS Collaboration. 'Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC'. In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020.

[2] The CMS Collaboration. 'Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC'. In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021.

[3] The ATLAS collaboration. 'Measurement of Higgs boson decay into b-quarks in associated production with a top-quark pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector'. In: *Journal of High Energy Physics* 2022.6 (June 2022). DOI: 10.1007/jhep06(2022)097.

[4] Manuel Guth. 'Search for $t\bar{t}H$ ($H \to b\bar{b}$) Production in the Lepton + Jets Channel and Quark Flavour Tagging with Deep Learning at the ATLAS Experiment.' Presented 25 Mar 2021. 2021. URL: https://cds.cern.ch/record/2765038.

[5] Thea Engler. 'Training of an extended b-tagging algorithm with neural networks'. 2020.

[6] David Galbraith. *UX: Standard Model of the Standard Model*. Accessed: 2022-11-06. URL: http://davidgalbraith.org/portfolio/ux-standard-model-of-the-standard-model/.

[7] Carsten Burgard. *Standard model of physics*. Accessed: 2022-11-06. URL: https://texample.net/tikz/examples/model-physics/.

[8] R.L. Workman et al. (Particle Data Group). 'Review of Particle Physics'. In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097.

[9] Nicola Cabibbo. 'Unitary Symmetry and Leptonic Decays'. In: *Phys. Rev. Lett.* 10 (12 June 1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531.

[10] Makoto Kobayashi and Toshihide Maskawa. 'CP-Violation in the Renormalizable Theory of Weak Interaction'. In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652.

[11] Sheldon L. Glashow. 'Partial-symmetries of weak interactions'. In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2.

[12] Abdus Salam. 'Gauge unification of fundamental forces'. In: *Rev. Mod. Phys.* 52 (3 July 1980), pp. 525–538. DOI: 10.1103/RevModPhys.52.525.

[13] Steven Weinberg. 'A Model of Leptons'. In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[14] David J Griffiths. *Introduction to elementary particles; 2nd rev. version*. Physics textbook. New York, NY: Wiley, 2008. URL: https://cds.cern.ch/record/111880.

[15] Michael E. Peskin and Daniel V. Schroeder. *An Introduction To Quantum Field Theory*. CRC Press, May 2018. DOI: 10.1201/9780429503559.

[16] Cliff Burgess and Guy Moore. *The Standard Model - A Primer*. Cambridge: Cambridge University Press, 2007. ISBN: 978-0-521-86036-9.

[17] Murray Gell-Mann. 'Symmetries of Baryons and Mesons'. In: *Phys. Rev.* 125 (3 Feb. 1962), pp. 1067–1084. DOI: 10.1103/PhysRev.125.1067.

[18] M. Gell-Mann. 'The interpretation of the new particles as displaced charge multiplets'. In: *Il Nuovo Cimento* 4.S2 (Apr. 1956), pp. 848–866. DOI: 10.1007/bf02748000.

[19] Tadao Nakano and Kazuhiko Nishijima. 'Charge Independence for V-particles*'. In: *Progress of Theoretical Physics* 10.5 (Nov. 1953), pp. 581–582. ISSN: 0033-068X. DOI: 10.1143/PTP.10.581.

[20] P.W. Higgs. 'Broken symmetries, massless particles and gauge fields'. In: *Physics Letters* 12.2 (1964), pp. 132–133. ISSN: 0031-9163. DOI: https://doi.org/10.1016/0031-9163(64)91136-9.

[21] Peter W. Higgs. 'Broken Symmetries and the Masses of Gauge Bosons'. In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[22] Peter W. Higgs. 'Spontaneous Symmetry Breakdown without Massless Bosons'. In: *Phys. Rev.* 145 (4 May 1966), pp. 1156–1163. DOI: 10.1103/PhysRev.145.1156.

[23] F. Englert and R. Brout. 'Broken Symmetry and the Mass of Gauge Vector Mesons'. In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[24] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble. 'Global Conservation Laws and Massless Particles'. In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[25] T. W. B. Kibble. 'Symmetry Breaking in Non-Abelian Gauge Theories'. In: *Phys. Rev.* 155 (5 Mar. 1967), pp. 1554–1561. DOI: 10.1103/PhysRev.155.1554.

[26] Janosh Riebesell. *Higgs Potential*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. 2021. URL: https://tikz.net/higgs-potential/.

[27] J C Collins and D E Soper. 'The Theorems of Perturbative QCD'. In: *Annual Review of Nuclear and Particle Science* 37.1 (1987), pp. 383–409. DOI: 10.1146/annurev.ns.37.120187.002123.

[28] R.K. Ellis, W.J. Stirling and B.R. Webber. *QCD and Collider Physics*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1996. ISBN: 9780521581899. URL: https://books.google.de/books?id=IIJlQgAACAAJ.

[29] Richard D. Ball et al. 'Parton distributions for the LHC run II'. In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). DOI: 10.1007/jhep04(2015)040.

[30] Particle Data Group. 'Review of Particle Physics'. In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104.

[31] Stefan Höche. *Introduction to parton-shower event generators*. 2014. DOI: 10.48550/ARXIV.1411.4085.

[32] B. Andersson et al. 'Parton fragmentation and string dynamics'. In: *Physics Reports* 97.2 (1983), pp. 31–145. ISSN: 0370-1573. DOI: https://doi.org/10.1016/0370-1573(83)90080-7.

[33] J.-C. Winter, F. Krauss and G. Soff. 'A modified cluster-hadronisation model'. In: *The European Physical Journal C* 36.3 (Aug. 2004), pp. 381–395. DOI: 10.1140/epjc/s2004-01960-8.

[34] Manuel Bähr et al. 'Herwig++ physics and manual'. In: *The European Physical Journal C* 58.4 (Nov. 2008), pp. 639–707. DOI: 10.1140/epjc/s10052-008-0798-9.

[35] Johannes Bellm et al. 'Herwig 7.0/Herwig++ 3.0 release note'. In: *The European Physical Journal C* 76.4 (Apr. 2016). DOI: 10.1140/epjc/s10052-016-4018-8.

[36] Torbjörn Sjöstrand et al. 'An introduction to PYTHIA 8.2'. In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024.

[37] Enrico Bothmann et al. 'Event generation with Sherpa 2.2'. In: *SciPost Phys.* 7 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034.

[38] Paolo Nason. 'A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms'. In: *Journal of High Energy Physics* 2004.11 (Nov. 2004), pp. 040–040. DOI: 10.1088/1126-6708/2004/11/040.

[39] Stefano Frixione, Paolo Nason and Carlo Oleari. 'Matching NLO QCD computations with parton shower simulations: the POWHEG method'. In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), pp. 070–070. DOI: 10.1088/1126-6708/2007/11/070.

[40] Simone Alioli et al. 'A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX'. In: *Journal of High Energy Physics* 2010.6 (June 2010). DOI: 10.1007/jhep06(2010)043.

[41] H. B. Hartanto et al. 'Higgs boson production in association with top quarks in the POWHEG BOX'. In: *Phys. Rev. D* 91 (9 May 2015), p. 094003. DOI: 10.1103/PhysRevD.91.094003.

[42] Stefano Frixione, Giovanni Ridolfi and Paolo Nason. 'A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction'. In: *Journal of High Energy Physics* 2007.09 (Sept. 2007), pp. 126–126. DOI: 10.1088/1126-6708/2007/09/126.

[43] Johan Alwall et al. 'MadGraph/MadEvent v4: the new web generation'. In: *Journal of High Energy Physics* 2007.09 (Sept. 2007), pp. 028–028. DOI: 10.1088/1126-6708/2007/09/028.

[44] The ATLAS collaboration. 'Jet reconstruction and performance using particle flow with the ATLAS Detector'. In: *The European Physical Journal C* 77.7 (July 2017). DOI: 10.1140/epjc/s10052-017-5031-2.

[45] Matteo Cacciari, Gavin P Salam and Gregory Soyez. 'The anti-k_t jet clustering algorithm'. In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. DOI: 10.1088/1126-6708/2008/04/063.

[46] ATLAS Collaboration. *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*. 2022. DOI: 10.48550/ARXIV.2211.16345.

[47] S. Agostinelli et al. 'Geant4—a simulation toolkit'. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8.

[48] David J. Lange. 'The EvtGen particle decay simulation package'. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 462.1 (2001). BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, pp. 152–155. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00089-4.

[49] *ATLAS Pythia 8 tunes to 7 TeV data*. Tech. rep. Geneva: CERN, 2014. URL: https://cds.cern.ch/record/1966419.

[50] S. Schumann and F. Krauss. 'A parton shower algorithm based on Catani-Seymour dipole factorisation'. In: *Journal of High Energy Physics* 2008.03 (Mar. 2008), p. 038. DOI: 10.1088/1126-6708/2008/03/038.

[51] Stefan Höche et al. 'A critical appraisal of NLO+PS matching methods'. In: *Journal of High Energy Physics* 2012.9 (Sept. 2012). DOI: 10.1007/jhep09(2012)049.

[52] Stefan Höche et al. 'QCD matrix elements + parton showers. The NLO case'. In: *Journal of High Energy Physics* 2013.4 (Apr. 2013). DOI: 10.1007/jhep04(2013)027.

[53] Stefano Catani et al. 'QCD Matrix Elements + Parton Showers'. In: *Journal of High Energy Physics* 2001.11 (Nov. 2001), pp. 063–063. DOI: 10.1088/1126-6708/2001/11/063.

[54] Stefan Höche et al. 'QCD matrix elements and truncated showers'. In: *Journal of High Energy Physics* 2009.05 (May 2009), pp. 053–053. DOI: 10.1088/1126-6708/2009/05/053.

[55] Lyndon Evans and Philip Bryant. 'LHC Machine'. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[56] O. Brüning, H. Burkhardt and S. Myers. 'The large hadron collider'. In: *Progress in Particle and Nuclear Physics* 67.3 (2012), pp. 705–734. ISSN: 0146-6410. DOI: https://doi.org/10.1016/j.ppnp.2012.03.001.

[57] Oliver Sim Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1.

[58] Esma Mobs. 'The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018'. In: (2018). General Photo. URL: https://cds.cern.ch/record/2636343.

[59] The LHCb Collaboration. 'The LHCb Detector at the LHC'. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08005–S08005. DOI: 10.1088/1748-0221/3/08/s08005.

[60] The ALICE Collaboration. 'The ALICE experiment at the CERN LHC'. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08002–S08002. DOI: 10.1088/1748-0221/3/08/s08002.

[61] The ATLAS Collaboration. 'The ATLAS Experiment at the CERN Large Hadron Collider'. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. DOI: 10.1088/1748-0221/3/08/s08003.

[62] The CMS Collaboration. 'The CMS experiment at the CERN LHC'. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004.

[63] *Public ATLAS Luminosity Results for Run-2 of the LHC*. Accessed: 2022/10/11. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

[64] *The Collaboration*. Accessed: 2022-11-28. URL: https://atlas.cern/Discover/Collaboration.

[65] Joao Pequenao. 'Computer generated image of the whole ATLAS detector'. 2008. URL: https://cds.cern.ch/record/1095924.

[66] Izaak Neutelings. *The ATLAS coordinate system including the LHC*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. 2021. URL: https://tikz.net/axis3d_cms/.

[67] Izaak Neutelings. *Pseudorapidity*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. 2021. URL: https://tikz.net/axis2d_pseudorapidity/.

[68] The ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331063.

[69] The ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 2*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331064.

[70] The ATLAS Collaboration. 'The upgraded Pixel detector and the commissioning of the Inner Detector tracking of the ATLAS experiment for Run-2 at the Large Hadron Collider'. In: (2016). DOI: 10.48550/ARXIV.1608.07850.

[71] The ATLAS Collaboration. 'The ATLAS Insertable B-Layer: from construction to operation'. In: (2016). DOI: 10.48550/ARXIV.1610.01994.

[72] Huaqiao Zhang (On behalf of the ATLAS Liquid Argon Calorimeter Group). 'The ATLAS Liquid Argon Calorimeter: Overview and Performance'. In: *Journal of Physics: Conference Series* 293 (Apr. 2011), p. 012044. DOI: 10.1088/1742-6596/293/1/012044.

[73] The ATLAS Collaboration. *ATLAS muon spectrometer: Technical Design Report*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331068.

[74] Joao Pequenao. 'Computer generated image of the ATLAS Muons subsystem'. 2008. URL: https://cds.cern.ch/record/1095929.

[75] Ana Maria Rodriguez Vera and Joao Antunes Pequenao. 'ATLAS Detector Magnet System'. General Photo. 2021. URL: https://cds.cern.ch/record/2770604.

[76] Peter Jenni et al. *ATLAS high-level trigger, data-acquisition and controls: Technical Design Report*. Technical design report. ATLAS. Geneva: CERN, 2003. URL: https://cds.cern.ch/record/616089.

[77] Aranzazu Ruiz-Martinez and ATLAS Collaboration. *The Run-2 ATLAS Trigger System*. Tech. rep. Geneva: CERN, 2016. DOI: 10.1088/1742-6596/762/1/012003.

[78] The ATLAS collaboration. 'Operation of the ATLAS trigger system in Run 2'. In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10004–P10004. DOI: 10.1088/1748-0221/15/10/p10004.

[79] Beijiang Liu et al. 'Applications of Machine Learning at BESIII'. In: *EPJ Web of Conferences* 214 (2019). Ed. by A. Forti et al., p. 06033. DOI: 10.1051/epjconf/201921406033.

[80] 'The Machine Learning landscape of top taggers'. In: *SciPost Physics* 7.1 (July 2019). DOI: 10.21468/scipostphys.7.1.014.

[81] ATLAS Collaboration. 'Performance of *b*-jet identification in the ATLAS experiment'. In: *Journal of Instrumentation* 11.04 (Apr. 2016), P04008. DOI: 10.1088/1748-0221/11/04/P04008.

[82] P. Baldi, P. Sadowski and D. Whiteson. 'Searching for exotic particles in high-energy physics with deep learning'. In: *Nature Communications* 5.1 (July 2014). DOI: 10.1038/ncomms5308.

[83] R. Santos et al. 'Machine learning techniques in searches for $t\bar{t}H$ in the $H \to b\bar{b}$ decay channel'. In: *Journal of Instrumentation* 12.04 (Apr. 2017), P04014. DOI: 10.1088/1748-0221/12/04/P04014.

[84] Anna Hallin et al. *Resonant anomaly detection without background sculpting*. 2022. DOI: 10.48550/ARXIV.2210.14924.

[85] Giuseppe Carleo et al. 'Machine learning and the physical sciences'. In: *Reviews of Modern Physics* 91.4 (Dec. 2019). DOI: 10.1103/revmodphys.91.045002.

[86] Alexander Radovic et al. 'Machine learning at the energy and intensity frontiers of particle physics'. In: *Nature* 560.7716 (Aug. 2018), pp. 41–48. DOI: 10.1038/s41586-018-0361-2.

[87] Dan Guest, Kyle Cranmer and Daniel Whiteson. 'Deep Learning and Its Application to LHC Physics'. In: *Annual Review of Nuclear and Particle Science* 68.1 (Oct. 2018), pp. 161–181. DOI: 10.1146/annurev-nucl-101917-021019.

[88] Li Deng. 'The MNIST database of handwritten digit images for machine learning research'. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[89] David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams. 'Learning representations by back-propagating errors'. In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. DOI: 10.1038/323533a0.

[90] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980.

[91] Jing An, Lexing Ying and Yuhua Zhu. 'Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients'. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=iQQK02mxVIT.

[92] Izaak Neutelings. *Neural Networks*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. 2021. URL: https://tikz.net/neural_networks/.

[93] Vinod Nair and Geoffrey E. Hinton. 'Rectified Linear Units Improve Restricted Boltzmann Machines'. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.

[94] Nitish Srivastava et al. 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[95] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. DOI: 10.48550/ARXIV.1502.03167.

[96] Manzil Zaheer et al. *Deep Sets*. 2017. DOI: 10.48550/ARXIV.1703.06114.

[97] Izaak Neutelings. *B tagging jets*. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/. 2021. URL: https://tikz.net/jet_btag/.

[98] *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2273281.

[99] *Measuring the b-jet identification efficiency for high p_T jets using multijet events in proton−proton collisions at $\sqrt{s} = 13$ TeV recorded with the ATLAS detector*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2804062.

[100] *Secondary vertex finding for jet flavour identification with the ATLAS detector*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2270366.

[101] *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*. Tech. rep. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2645405.

[102] *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2255226.

[103] *Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2718948.

[104] Patrick T. Komiske, Eric M. Metodiev and Jesse Thaler. 'Energy flow networks: deep sets for particle jets'. In: *Journal of High Energy Physics* 2019.1 (Jan. 2019). DOI: 10.1007/jhep01(2019)121.

[105] The ATLAS collaboration. 'A neural network clustering algorithm for the ATLAS silicon pixel detector'. In: *Journal of Instrumentation* 9.09 (Sept. 2014), P09009. DOI: 10.1088/1748-0221/9/09/P09009.

[106] *The Optimization of ATLAS Track Reconstruction in Dense Environments*. Tech. rep. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2002609.

[107] *Track Reconstruction Performance of the ATLAS Inner Detector at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2015. URL: http://cds.cern.ch/record/2037683.

[108] The ATLAS collaboration. 'Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector'. In: *Phys. Rev. D* 97 (7 Apr. 2018), p. 072016. DOI: 10.1103/PhysRevD.97.072016. URL: https://link.aps.org/doi/10.1103/PhysRevD.97.072016.

[109] *DeXTer: Deep Sets based Neural Networks for Low-p_T $X \to b\bar{b}$ Identification in ATLAS*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2825434.

[110] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[111] Francois Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[112] Joschka Birk et al. *umami-hep/puma: v0.1.9*. Version v0.1.9. Nov. 2022. DOI: 10.5281/zenodo.7382284.

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

C. List of Acronyms

SM Standard Model

QED Quantum electrodynamics

QCD Quantum chromodynamics

EW electroweak

CKM Cabibbo–Kobayashi–Maskawa

LEP Large Electron-Positron Collider

LHC Large Hadron Collider

PDF parton distribution function

ISR initial state radiation

FSR final state radiation

PS parton shower

ME matrix element

LO leading order

NLO next-to-leading order

NNLO next-to-next-to-leading order

PFlow Particle Flow

CERN European Organization for Nuclear Research

ID inner detector

SCT semiconductor tracker

TRT transition radiation tracker

IBL insertable b-layer

LAr liquid argon

HEC hadronic end-cap calorimeter

FCal hadronic forward calorimeter

EMEC electromagnetic end-cap calorimeter

MDT monitored drift tube

CSC cathode-strip chamber

RPC resistive plate chamber

TGC thin gap chamber

HLT high-level trigger

MC Monte Carlo

ML machine learning

HEP high energy physics

LR learning rate

NN neural network

AI artificial intelligence

MSE mean squared error

GD gradient descent

SGD stochastic gradient descent

ReLU rectified linear unit

PV primary vertex

SV secondary vertex

IP impact parameter

DIPS Deep Impact Parameter Sets

RNN Recurrent Neural Network

HF heavy flavour

List of Figures

List of Tables

Erklärung gem. der geltenden Prüfungsordnung:

Hiermit versichere ich, dass

1. ich die eingereichte Masterarbeit bzw. bei einer Gruppenarbeit meinen Anteil entsprechend gekennzeichneten Anteil der Arbeit selbständig verfasst habe,
2. ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Inhalte als solche kenntlich gemacht habe und
3. die eingereichte Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.

_____ _____

Ort, Datum Unterschrift

Auszug Prüfungsordnung M.Sc.
Vom 19. August 2005 (Amtliche Bekanntmachungen Jg. 36, Nr. 46, S. 269–293) in der Fassung vom 25. September 2020 (Amtliche Bekanntmachungen Jg. 51, Nr. 66, S. 328–337)

§20 (8) Bei der Einreichung hat der/die Studierende schriftlich zu versichern, dass

1. er/sie die eingereichte Masterarbeit beziehungsweise bei einer Gruppenarbeit seinen/ihren entsprechend gekennzeichneten Anteil der Arbeit selbständig verfasst hat,
2. er/sie keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Inhalte als solche kenntlich gemacht hat und
3. die eingereichte Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.