Improvement of the top-quark reconstruction for the top-quark mass measurement using machine learning techniques

Bachelor thesis

UNIVERSITY OF FREIBURG

Department of Mathematics and Physics

Joschka Birk

Supervisor: Prof. Dr. Gregor Herten

Submission date: July 2, 2019

Acknowledgement

First of all I would like to thank Prof. Dr. Gregor Herten for giving me the opportunity to write my bachelor thesis in his working group.
Especially, I would like to thank Dr. Andrea Knue for the great supervision, always having an open ear and providing helpful advice whenever necessary during the whole work of this thesis. Additionally, I would like to thank Manuel Guth for also helping me in case of any problems. Finally, I thank the whole Herten group for the great time I had during the course of my thesis.

Abstract

The measurement of the top-quark mass is important for precision tests of the Standard Model. The precision of the m_{top} measurement is limited by systematic uncertainties. A measurement of m_{top} in the lepton+jets channel at $\sqrt{s} = 8\,\text{TeV}$ was performed by the ATLAS experiment [1] using a boosted decision tree (BDT) for the event selection optimisation. The BDT was used to reject wrongly reconstructed events in order to reduce the systematic uncertainty of the top-quark mass. In the thesis presented here, investigations on the event selection and reconstruction for the m_{top} measurement at $\sqrt{s} = 13\,\text{TeV}$ are performed using simulated events. These investigations are achieved using both a BDT similar to the 8 TeV analysis as well as using a deep neural network (DNN), which is a more sophisticated machine-learning technique. In comparison to the BDT selection at 8 TeV, the matching efficiency of the BDT selection at 13 TeV is slightly lower but leads to a larger purity. It is also found that variables with a small separation can be removed from the training without decreasing the BDT performance significantly. First studies with the DNN improve the purity of the event selection compared to the BDT selection and lead to a better m_{top} resolution. The DNN training therefore provides a new classifier of promising potential for the full measurement of m_{top}.

Kurzzusammenfassung

Die Messung der Top-Quark-Masse ist wichtig für Präzisionstests des Standardmodells. Die Genauigkeit der m_{top}-Messung ist durch systematische Unsicherheiten begrenzt. Eine Messung von m_{top} im lepton+jets Zerfallskanal bei $\sqrt{s} = 8\,\text{TeV}$ wurde vom ATLAS-Experiment [1] unter Verwendung eines Boosted Decision Trees (BDT) für die Optimierung der Ereignisauswahl durchgeführt. Der BDT wurde verwendet, um falsch rekonstruierte Ereignisse zu verwerfen und dadurch die systematische Unsicherheit der Top-Quark-Masse zu reduzieren. In der hier vorliegenden Arbeit werden Untersuchungen zur Ereignisauswahl und -rekonstruktion für die m_{top}-Messung bei $\sqrt{s} = 13\,\text{TeV}$ unter Verwendung simulierter Ereignisse durchgeführt. Diese Untersuchungen werden sowohl mit einem BDT ähnlich der 8 TeV-Analyse als auch mit einem tiefen neuronalen Netz (DNN) durchgeführt, welches eine anspruchsvollere Methode des maschinellen Lernens darstellt. Im Vergleich zur BDT-Selektion bei 8 TeV ist die Matching-Effizienz der BDT-Selektion bei 13 TeV etwas geringer, führt jedoch auch zu einer größeren Reinheit. Außerdem zeigt sich, dass Variablen mit kleiner Separationskraft aus dem Training entfernt werden können, ohne die BDT-Leistung signifikant zu beeinträchtigen. Erste Studien mit dem DNN verbessern die Reinheit der Ereignisauswahl im Vergleich zur BDT-Selektion und führen zu einer besseren m_{top}-Auflösung. Das DNN-Training bietet daher eine neue Klassifikation mit vielversprechendem Potenzial für die vollständige Messung von m_{top}.

Contents

1 Introduction

To understand the fundamental principles of nature, physicists are looking for theories that describe the elementary particles of our universe. At the same time, highly complex experiments are carried out to find experimental proof for these theories as well as to look for new physics which can not be explained by any theory yet.

At the Large Hadron Collider (LHC) at CERN (Conseil Européen pour la Recherche Nucléaire) protons are accelerated to velocities close to the speed of light and are then made to collide in order to investigate the behaviour of the elementary particles and the fundamental forces between them.

The Standard Model of particle physics is a quantum field theory that describes these particles and interactions. With the discovery of the Higgs boson by the two experiments ATLAS [2] and CMS [3] at the LHC in 2012, the last missing piece of the Standard Model was found. Since then, further precision measurements have been performed to test the consistency of the Standard Model.

One of these precision measurements is the measurement of the top-quark mass. Even though the top quark is an elementary particle, its mass is approximately as the mass of a tungsten atom. Having such a large mass, precision measurements of m_{top} are essential to test not just the consistency of the Standard Model but also to obtain information on fundamental interactions at the electroweak symmetry-breaking scale.

Due to its large mass, the top quark has an extremely short lifetime, which is shorter than the time necessary to form hadrons. Therefore, to measure the top-quark properties, it can simply be reconstructed using the information of the final state particles of its decay. At the LHC, top quarks are primarily produced in top-antitop-pairs ($t\bar{t}$-pairs), which can be measured in three different decay channels, that are defined by the decay modes of the W boson.
The lepton+jets channel is characterised by the presence of a charged lepton, four jets and missing transverse momentum. The reconstruction of the $t\bar{t} \rightarrow$ lepton+jets events measured in the ATLAS detector requires a correct assignment of the measured jets to the corresponding quarks on parton level. Due to the large combinatorial background, this reconstruction is very challenging, but can be solved using kinematic likelihood methods combined with machine learning techniques.

This thesis presents investigations and improvements of the event reconstruction and selection in the lepton+jets channel for the ATLAS m_{top} measurement at a centre-of-mass energy of 13 TeV. The analysis is based on the ATLAS m_{top} measurement in the lepton+jets channel at 8 TeV [1].

In chapter 2 of this thesis an overview of the Standard Model is given followed by a brief description of the ATLAS detector in chapter 3. The most important properties of the top quark, its production and decay channels as well as the ATLAS m_{top} measurement in the lepton+jets channel at 8 TeV are described in chapter 4. In chapter 5, the object definition and the preselection used for the studies in this thesis are explained. The event

reconstruction methods and the multivariate classification methods used in these studies are introduced in chapter 6. Studies of the improvement of the event reconstruction and event selection are presented in chapter 7 and 8. Finally, the results of these studies are summarised in chapter 9.

2 The Standard Model of particle physics

The Standard Model of particle physics (SM) is a quantum field theory which describes the elementary particles that all matter is made of as well as the fundamental interactions between these particles. The SM includes three of the four fundamental forces, namely the electromagnetic force, the strong force and the weak force, while the gravitational force is not included in the SM. However, due to the extremely small masses of the elementary particles, the gravitational force can be neglected when describing interactions between elementary particles.
The SM was developed in the late 1970s [4, 5, 6, 7] and has been tested in several experiments since then. In 2012, the last missing piece of the SM was found with the discovery of the Higgs Boson [2, 3]. This chapter gives a brief overview of the SM. If not stated otherwise, this chapter is based on [8, 9, 10, 11].

The elementary particles in the SM are divided into *fermions*, which have a spin of 1/2, *gauge bosons* with spin 1 and the Higgs boson, which has spin 0. The fermions and the gauge bosons are descibed in the first two sections of this chapter, followed by a brief introduction to the Higgs mechanism and the Higgs boson.

2.1 Fermions

There are two types of fermions: leptons and quarks. An overview is given in Table 2.1. The fermions are sorted into three generations with each generation containing two quarks, a charged and a neutral lepton. The particles of the second and the third generation have the same properties as the particles in the first generation except their mass, which is increasing with the number of the generation.

Table 2.1: List of the known fermions and their properties. They are grouped in three generations, each generation containing two quarks and two leptons. Values taken from [12].

	Leptons			Quarks		
Gen.	Flavour	Charge [e]	Mass [MeV]	Flavour	Charge [e]	Mass [MeV]
I	e	-1	0.511	u	$^2/_3$	$2.2^{+0.5}_{-0.4}$
	ν_e	0	$< 2 \cdot 10^{-6}$	d	$^{-1}/_3$	$4.7^{+0.5}_{-0.3}$
II	μ	-1	105.7	c	$^2/_3$	$1\,275^{+25}_{-35}$
	ν_μ	0	< 0.19	s	$^{-1}/_3$	95^{+9}_{-3}
III	τ	-1	1776.9	t	$^2/_3$	$173\,000 \pm 400$
	ν_τ	0	< 18.2	b	$^{-1}/_3$	$4\,180^{+40}_{-30}$

The leptons are described by their electric charge Q, electron number L_e, muon number L_μ and tau number L_τ [1]. While the electron (e), the muon (μ) and the tau (τ) have an

[1] L_e, L_μ and L_τ are +1 for particles of the corresponding flavour, -1 for the respective antiparticles and 0 for particles of a different flavour.

electric charge[2] of $-e$, the neutrinos have no electric charge. In addition to the six leptons shown in Table 2.1 there are also six corresponding antiparticles, each with negative lepton numbers[3] and negative electric charge compared to the respective particle, which means that there are twelve leptons in total.
Quarks do also appear in three generations and are described by their electric charge, their flavour and an additional property, their so called *colour charge*. The up (u), charm (c) and the top-quark (t) have an electric charge of $^2/_3\,e$ while the down (d), strange (s) and bottom quark (b) have an electric charge of $-^1/_3\,e$. Since there are these six different quark flavours (u, d, c, s, t, b) and three different colour charges[4] (green, blue and red), there are 36 quarks in total considering that for every possible combination of flavour and colour charge there is a corresponding antiparticle with the signs of the electric charge flipped.

2.2 Gauge bosons

Gauge bosons are the mediator particles of the interactions between the elementary particles. These gauge bosons are the photon for the electromagnetic force, the $W^{\pm}$ bosons and the Z^0 boson for the weak force and eight gluons for the strong force. The gauge bosons of the SM and their properties are listed in Table 2.2. Additionally, the graviton, the theoretical boson for the gravitational force and the corresponding strength is included in the table to provide a comparison of its strength compared to the other forces. Like stated before, the gravitational interaction is not included in the SM but can be neglected when describing the interactions between elementary particles due to its small strength.

Table 2.2: Gauge bosons of the Standard Model and their properties. The hypothetical graviton with its coupling strength is also listed in the table, but it is not included in the standard model. Values taken from [8, 12].

Gauge bosons

	Mass [GeV]	Electric charge [e]	Interaction	Strength
g (gluon)	0	0	strong	1
γ (photon)	$< 10^{-27}$	$< 10^{-35}$	electromagnetic	1/137
$W^{\pm}$	80.379 ± 0.012	± 1	weak	10^{-6}
Z^0	91.1876 ± 0.0021	0	weak	10^{-6}
graviton	not included in the SM		gravitation	10^{-31}

The SM is based on a gauge theory including several free parameters which can be constrained by measured observables like for example the cross section of a specific interaction or the masses of particles.
In the 1960s, the electromagnetic force and the weak force have been unified into the electroweak force by Glashow [4], Salam [5] and Weinberg [7] based on the gauge group

[2] In the context of electric charge, $e = 1.602\,176\,6208(98) \times 10^{-19}$ C [12] is the elementary charge of the electron.
[3] The lepton numbers are the electron number, the muon number and the tau number.
[4] Every colour has a corresponding anticolour. colour and anticolour of the same kind cancel each other out.

$SU(2)_L \times U(1)_{Y_W}$ [5]. The index L refers to the fact, that it includes only left-handed states and $Y_W = 2(Q - T_3)$ is the weak hypercharge with Q being the electric charge and T_3 the third component of the weak isospin. The resulting mediator particles are three massive particles namely the two charged W bosons and the neutral Z boson as well as the massless photon.
In contrast to all other interactions, the charged weak interaction is the only one which is able to change the flavour of quarks. This occurs when a up (down)-type quark decays via $q \to q' W$ into a down (up)-type quark and a W boson or vice versa. Flavour changing decays via a neutral current, i.e. a flavour changing decay including the radiation of a Z^0 boson, have not been observed yet. The probabilities for the different possible flavour changing decays can be written in a unitary 3×3 matrix [13, 14], the Cabbibo-Kobayashi-Maskawa matrix (CKM matrix)

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

where $|V_{ij}|^2$ represents the probability of a quark of flavour i changing into flavour j if it interacts with a W boson. However, the off-diagonal entries do not have very large values as a flavour change to the corresponding up-type or down-type quark in the same generation has the highest probability.
The strong interaction is based on a $SU(3)_C$ [6] gauge symmetry with C being the colour charge. This group has 8 generators, which is why there are 8 different types of gluons. These gluons only couple to colour charged particles and therefore only gluons and quarks interact via the strong force. However, gluons carry colour charge themself which is why gluons do not only couple to quarks but also to each other. The strength of this coupling, which is described by the coupling constant of the strong interaction α_S depends on the energy scale. This *running of* α_s is shown in Figure 2.1.

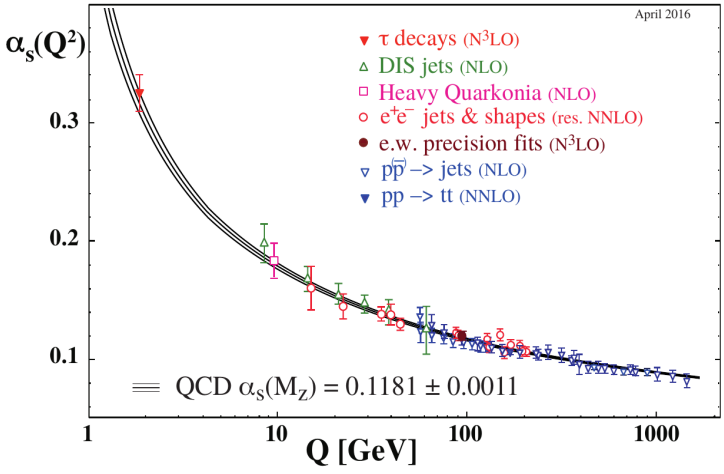


Figure 2.1: Running of the coupling constant α_s of the strong interaction as a function of the energy scale Q [12].

As can be seen in the figure, the coupling of the strong force decreases for high energies (small distances), which is called *asymptotical freedom*. Another characteristic of the

[5] SU(2) is the special unitary group of 2×2 matrices and U(1) is the unitary group with $n = 1$.
[6] SU(3) is the special unitary group of 3×3 matrices.

strong force is an effect called *confinement*, which means that coloured particles do not appear as isolated particles but form colourless bound states, called *hadrons* [15]. The process of the production of these hadrons is called *hadronisation*.
Combining the gauge groups of all three interactions included in the Standard Model, the gauge group of the Standard Model is

$$\mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_{Y_W} . \tag{2.2}$$

However, if the three interactions of the SM are introduced in this way, the gauge bosons included in the standard model would be massless. Since the Z boson and the W bosons are known to have mass from experiment, the SM had to be extended by an additional sector describing the origin of the elementary particles masses.

2.3 The Higgs mechanism

By introducing an additional scalar field Φ, the *Higgs field*, the three physicists Brout, Englert [16] and Higgs [17] solved this problem. This adds an additional term to the SM lagrangian

$$\mathcal{L}_{\text{Higgs}} = (\partial_\mu \Phi)^\dagger (\partial^\mu \Phi) - V(\Phi) \tag{2.3}$$

where the potential of the Higgs field is given by

$$V(\Phi) = \mu^2 (\Phi^\dagger \Phi) + \lambda (\Phi^\dagger \Phi)^2 , \tag{2.4}$$

For $\mu^2 < 0$ and $\lambda > 0$, this field has a non-zero vacuum expectation value v. A sketch of the higgs potential can be seen in Figure 2.2.

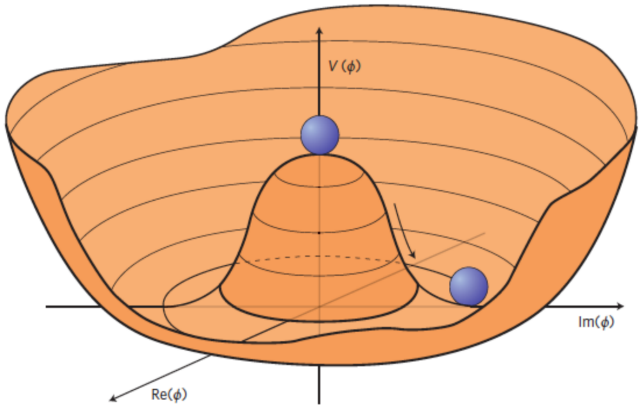


Figure 2.2: Sketch of the Higgs potential. Choosing any of the points in the circle of minima breaks the symmetry spontaneously [18].

This non-zero vacuum expectation value results in a *spontaneous symmetry breaking*, which generates masses for the $\mathrm{W}^\pm$ and the Z^0. Furthermore, due to the self interaction of the Higgs field, a new particle occurs: the Higgs boson. It has charge 0 and spin 0 and was discovered at the Large Hadron Collider at CERN by the two experiments ATLAS [2] and CMS [3] in 2012. The mass of the Higgs boson is measured to be $m_H = (125.18 \pm 0.16)\,\mathrm{GeV}$ [12].

Additionally, the Higgs boson explains the masses of the fermions as it couples to all fermions with a strength proportional to their mass. This coupling is called the *Yukawa coupling* and the mass of the fermion M_f can be written as

$$M_f = Y_f \frac{v}{\sqrt{2}}, \tag{2.5}$$

with $v \approx 246\,\text{GeV}$ being the vacuum expectation value of the Higgs field and Y_f the Yukawa coupling constant of the fermion. Since the top quark is the heaviest fermion, it is predicted to have the largest Yukawa coupling to the Higgs boson. Its value is close to unity, which is why the top-quark is expected to play a special role in the mechanism of electroweak symmetry breaking.

3 The ATLAS experiment and the Large Hadron Collider at CERN

The *Conseil Européen pour la Recherche Nucléaire* (CERN) is a european research organisation exploring the basic constituents of matter and the fundamental interactions between them. It was founded in 1954, is located at the Franco-Swiss border near Geneva and has currently 23 member states. Providing a large range of particle accelerator facilities, CERN's primary mission is fundamental research. However, CERN also does a lot of research in the fields of material science and computing.

By accelerating particles to velocities close to the speed of light and then making them collide, CERN studies the elementary particles of the SM and their interactions. This chapter gives a brief overview of the Large Hadron Collider (LHC), a proton-proton collider at CERN, and the ATLAS experiment at the LHC.

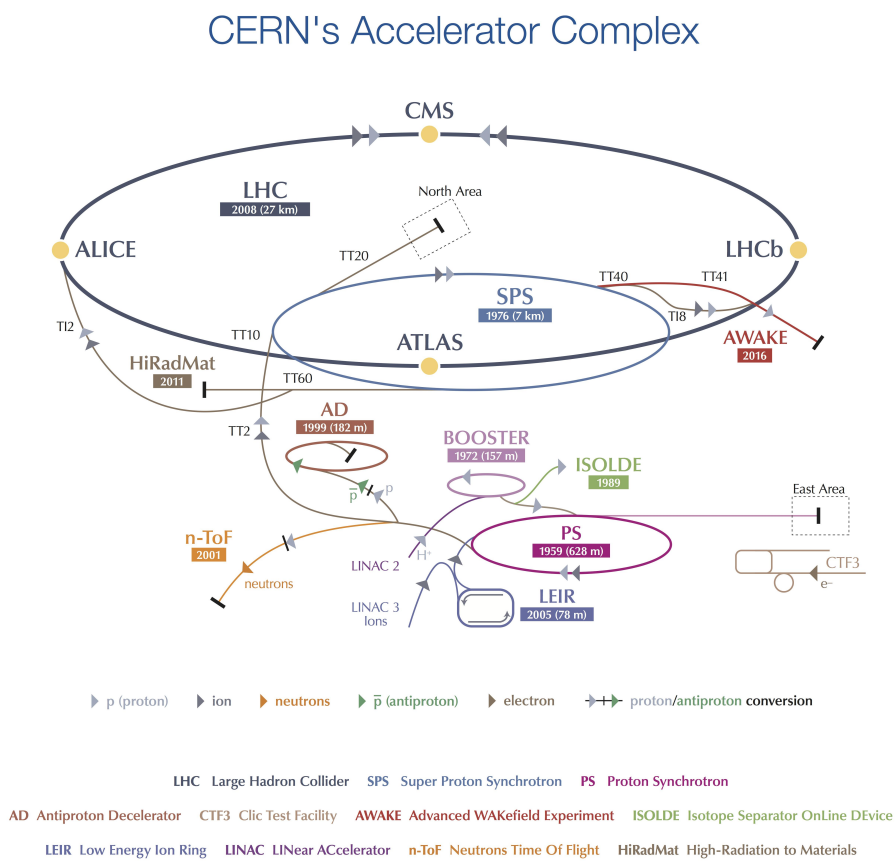


Figure 3.1: Accelerator complex at CERN [19]

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a circular particle accelerator with a circumference of 27 km located at CERN. It is designed to accelerate bunches of protons (p) to velocities near the speed of light and then collide these protons at a centre-of-mass energy of up to currently 13 TeV, making it the most powerful particle accelerator built so far. Besides from accelerating protons, the LHC is also used to accelerate and collide heavy ions (A). Before entering the LHC, the particles are accelerated in several steps using different pre-accelerators. The full accelerator complex is shown in Figure 3.1.

To produce the protons for the experiment, simple hydrogen gas is used by separating the protons from the electrons with an electric field. Afterwards, these protons are accelerated via the linear accelerator Linac to an energy of 50 MeV. The protons are then further accelerated with the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS), which are all circular accelerators. Afterwards, with an energy of 450 GeV, the protons are injected to the LHC, where they reach their final energy of 6.5 TeV. The protons are injected into the LHC packed in 2808 bunches per beam with each bunch containing $1.2 \cdot 10^{11}$ protons.

There are four collision points along the 27 km ring each corresponding to one of the four main experiments at the LHC. These four experiments and a brief description of their research areas are listed in the following .

- **ATLAS** (**A** **T**oroidal **L**HC **A**pparatu**S**) [20] and **CMS** (**C**ompact **M**uon **S**olenoid) [21] are general-purpose detectors measuring both p-p and A-A collisions. These two particle detectors are built for the search for new particles and interactions as well as for precision measurements to test the consistency of the Standard Model.
- **LHCb** (**L**arge **H**adron **C**ollider **b**eauty experiment) [22] is a detector used for measurements of rare decays of b-hadrons and precision measurements of CP violation. In contrast to the other detectors at the LHC, this detector is built in forward direction and not symmetrically around the beam axis.
- **ALICE** (**A** **L**arge **I**on **C**ollider **E**xperiment) [23] is a general-purpose detector designed for measurements at extreme values of energy density and temperature. The main goal of the ALICE detector is to investigate further the strong interaction and the quark-gluon plasma produced in heavy-ion-collisions.

3.2 ATLAS detector

The ATLAS detector is in terms of volume the largest particle detector ever built for a collider experiment. It has the shape of a cylinder with a length of 46 m and a diameter of 25 m and is located one hundred meters below ground at the LHC at CERN. The detector has a total weight of around 7000 tonnes and consists of different layers of detector-subsystems namely the inner detector (ID), the calorimeters and the muon spectrometer [20]. These detector subsystems, as well as the magnet and trigger systems of the ATLAS detector are briefly explained in the following. An overview of the detector and its constituents can be seen in Figure 3.2.

The coordinate system used in this thesis is the one used by ATLAS, which is a right-handed coordinate system with the origin at the centre of the detector and the z-axis

along the beam line. The x-axis points to the centre of the LHC and the y-axis points upwards. Points in the transverse (i.e. x-y) plane are described by cylindrical coordinates (r,ϕ), with ϕ being the azimuthal angle around the z-axis. The pseudorapidity is defined as

$$\eta = -\ln\tan(\theta/2) \tag{3.1}$$

with θ being the polar angle measured from the z-axis. The pseudorapidity is used instead of the the angle θ, as $\Delta\eta$ has the advantage of being Lorentz invariant. Angular distances are measured in units of ΔR, which is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}\,. \tag{3.2}$$

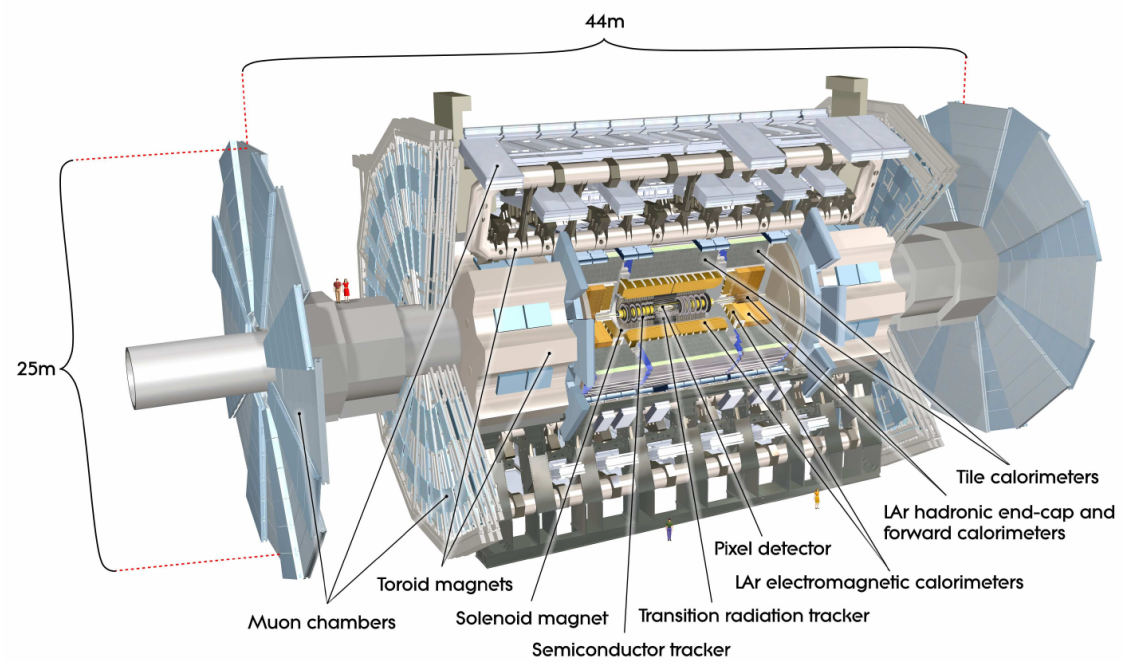


Figure 3.2: Overview of the ATLAS-Detector and its subcomponents [20].

3.2.1 Inner Detector

The ID, which can be seen in Figure 3.3, is built to measure the tracks and the momenta of all charged particles with a pseudorapidity of up to $|\eta| < 2.5$ and to reconstruct secondary vertices. It is divided into three different subsystems aligned concentrically around the beam axis as well as three equal subsystems in the end caps and is enveloped by a solenoid magnet which provides a magnetic field of strength 2 T. The silicon pixel detector is installed closest to the beamline and consists of 80 million pixels of size $50 \times 400\,\mu\text{m}^2$. Surrounding the pixel detector there is the semiconductor tracker, also made out of silicon and composed of 4088 two-sided modules with over 6 million implanted readout strips allowing the determination of the particle's track with an accuracy of up to 17 μm. Finally, the third layer of the inner detector, the transition radiation tracker has approximately 351 000 channels each assigned to a drift tube with a diameter of 4 mm [20]. The tubes are filled with a gas mixture of 70 % Xe, 27 % CO_2, and 3 % O_2 and the spaces between the tubes are filled with polymer fibres in the barrel region and with polymer foils in

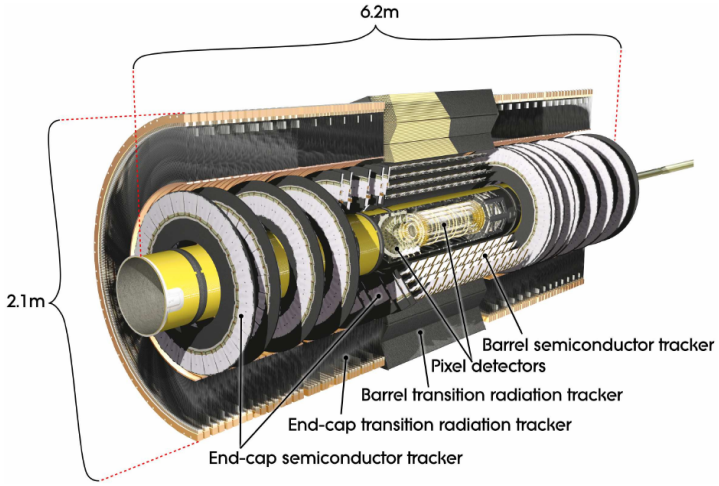


Figure 3.3: Schematic representation of the inner detector [20].

the endcaps region. When high energy particles travel through the material between the tubes, transition radiation is produced, which is then measured in the straw tubes. Therefore, this part of the ID provides additional information about the particle type as the amount of transition radiation of particles with the same energy depends on the type of the particle [24].

3.2.2 Calorimeters

The calorimeters are used to determine the energy of the particles and are designed to force the particles to deposit all of their energy within the calorimeter. Therefore, the calorimeters are built out of material with which particles interact very strongly. The interactions between high energy particles and the detector material lead to particle showers which can be measured up to a pseudorapidity of $|\eta| < 4.9$ [20]. However, neutrinos pass the calorimeters without losing energy making it impossible to measure their energies directly. All calorimeters used in the ATLAS detector are sandwich calorimeters, which means that they are composed of alternating layers of active and passive material, also referred to as absorber material. The particles interact with the passive material leading to particle showers which are measured with the active material. A schematic representation of the calorimetry system is shown in Figure 3.4.

Electromagnetic calorimeter

The electromagnetic calorimeter of the ATLAS detector is important for the identification and reconstruction of electrons and high energy photons. It consists of a barrel part covering a range of $|\eta| < 1.475$ and two end cap components over the range $1.375 < |\eta| < 3.2$. Lead plates are used as absorber medium and liquid argon (LAr), which is cooled down to -183 °C, is used as the active medium, which is ionized by the particles created in the shower. In the part of the barrel region, a fine granularity is chosen to perform precision measurements of electrons and photons while the rest of the electromagnetic calorimeter features a coarser granularity as this area is mainly used for jet reconstruction and measurements of missing transverse momentum [20].

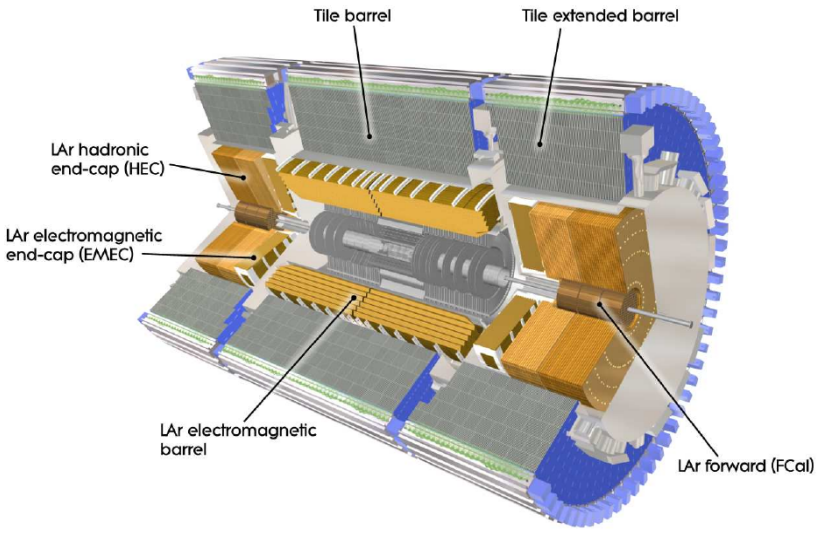


Figure 3.4: Schematic representation of the calorimetry system of the ATLAS detector [20].

Hadronic calorimeters

There are different types of hadronic calorimeters built into the ATLAS detector: the tile calorimeter, the LAr hadronic end-cap calorimeter (HEC) and the LAr forward calorimeter (FCal). These calorimeters play an important role in the reconstruction of hadronic jets and are designed to stop nearly all particles besides from muons and neutrinos from entering the muon system.
The tile calorimeter forms another layer of the onion-like structure of the detector and is located next to the electromagnetic calorimeter. This calorimeter envelops both the barrel parts of the other components mentioned so far with its own barrel part as well as their end cap parts with its extended barrel parts. This provides measurements for pseudorapidities $|\eta| < 1.7$. This calorimeter uses steel as its absorber medium and contains 500 000 plastic scintillator tiles which are read-out via photo-multiplier tubes. Like the electromagnetic calorimeter, both the HEC and the FCal use liquid argon as their active medium while copper and tungsten are used as passive medium. The HEC and the FCal are located at the end cap of the detector close to the beamline to detect particles near the beam axis providing measurements in the ranges $1.5 < |\eta| < 3.2$ (HEC) and $3.2 < |\eta| < 4.9$ (FCal) [20].

3.2.3 Muon spectrometer

As muons do not interact via the strong force, they penetrate the hadronic calorimeter without losing much energy. Furthermore, muons do interact electromagnetically but since they are much heavier than electrons, their energy loss due to bremsstrahlung and ionization is much smaller than for electrons. Hence they do also penetrate the electromagnetic calorimeter without losing much energy. Therefore, the muon's coordinates, charge and momentum are measured in the muon spectrometer located around the calorimetry system in combination with the muon tracks measured in the inner detector.
The muon system is a combination of four different detector subsystems, installed in so called muon chambers, and strong magnetic fields. An overview of the different subsystems can be seen in Figure 3.5. The Thin gap chambers (TGC) and the Resistive plate

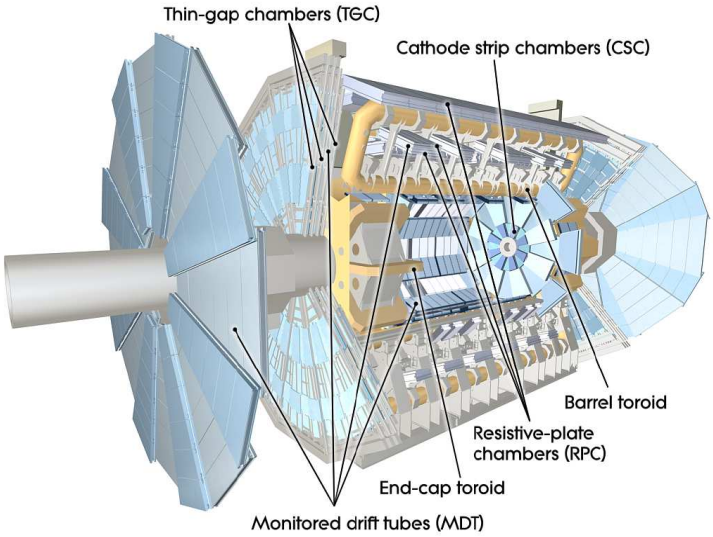


Figure 3.5: Different constituents of the ATLAS muon spectrometer.

chambers (RPC) are used for triggering and provide a second coordinate measurement. The other two subsystems, namely the Monitored Drift Tubes (MTD) and the Cathode Strip Chambers (CSC) are used for precision measurements in the principal bending direction of the magnetic field [20].

3.2.4 Magnet system

The magnet system consists of three toroid magnets and a solenoid magnet. The solenoid magnet is installed between the inner detector and the calorimetry system to provide additional information for the reconstruction of low energy particles and fast decaying particles. This solenoid magnet has an axial magnetic field of 2 T. Two of the three toroids are located at the two end caps while the third one is assembled surrounding both the calorimeters and the two end cap magnets. These toroid magnets produce a magnetic field of up to 4 T and each of them consists of 8 air-core coils.

3.2.5 Trigger system and data acquisition

The proton-proton collisions in the ATLAS detector occur with an extremely high frequency of about 1 billion collisions per second. Processing and storing every single event would result in a combined data volume of more than 60 terabytes per second. As this amount of data cannot be stored, the trigger system selects events with distinguishing characteristics which are then processed and stored for later analysis. A schematic view of the ATLAS trigger system is shown in Figure 3.6.
This selection process is divided in two stages. First, the Level-1 hardware trigger selects events based on a subset of information from the calorimeters and the muon spectrometer, which reduces the data to at most 100 000 events per second. Afterwards, the High Level Trigger (HLT), which is a software based trigger, refines the signal by applying selection algorithms using either the whole information of the event or only data of specified regions of the detector. About 1000 events pass the HLT selection and are stored for offline analysis.

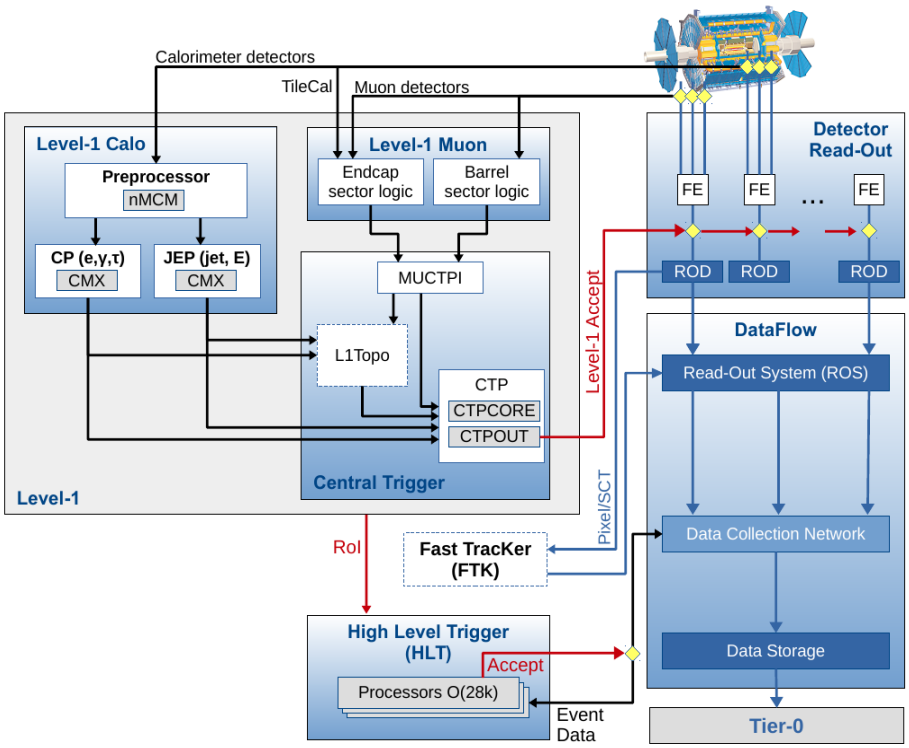


Figure 3.6: Schematic view of the ATLAS trigger system [25].

4 The top quark

Since the mass of the top quark is, like all fermion masses, a free parameter in the Standard model, it plays an important role in measurements for the consistency of the SM. Furthermore, being close to the scale of electroweak symmetry breaking, the top mass is an important quantity for calculations of the stability of the electroweak vacuum [12]. This chapter contains an overview of the the top-quark properties as well as its most important production processes and the decay channels.

4.1 Properties of the top quark

Compared to all other elementary particles of the SM, the top quark has an extremely large mass of (172.69 ± 0.49) GeV [1]. Hence it has a corresponding lifetime of $\approx 0.5 \times 10^{-24}$ s which is why it is expected to decay before top-quark flavoured hadrons can form. Therefore, it is the only "free" quark that can be observed, and it can be reconstructed from its decay particles.
As the mass of the W boson is with $(80.379 \pm 0.012)\,\text{GeV}$ [12] larger than the mass of all other quarks besides the t, the top quark is the only particle with a real W boson in its decay products. This weak decay is discussed in more detail in section 4.4. Top quarks are produced either as single-top-quarks via the weak interaction, referred to as single-top-quark production or in pairs via the strong interaction, referred to as top-quark pair production.

4.2 Top-quark pair production

The top-quark pairs produced at the LHC are produced via hard scattering processes between the constituents of the colliding protons, namely the quarks and the gluons, which are described by perturbative Quantum chromodynamics [26]. The total cross section for $t\bar{t}$-production can be described via the factorisation theorem using the parton distribution functions (PDF), $f_i(x_i, \mu^2)$, of the quarks and gluons in the proton. The PDF represents the probability density for a parton of type i to carry the fraction x_i of momentum compared to the corresponding total momentum of the proton. Using the partonic cross section $\hat{\sigma}^{ij}$, the total cross section for the $t\bar{t}$-production can be calculated as [26, 27]

$$\begin{aligned}\sigma^{t\bar{t}}(\sqrt{s}, m_t) = \sum_{i,j=q,\bar{q},g} \int dx_i\, dx_j\, f_i(x_i, \mu^2)\, f_j(x_j, \mu^2) \\ \times\, \hat{\sigma}^{ij\to t\bar{t}}(\rho, m_t^2, x_i, x_j, \alpha_s(\mu^2), \mu^2)\ ,\end{aligned} \tag{4.1}$$

where $f_i(x_i, \mu^2)$ and $f_j(x_j, \mu^2)$ are the PDFs for the parton of two protons, μ is the energy scale of the hard scattering, m_t the top mass and $\rho = 4\, m_t^2/\sqrt{\hat{s}}$. The effective centre-of-mass energy $\hat{s}$ is defined as $\hat{s} = x_i x_j s$ and the indices i, j run over all quarks, antiquarks and gluons.
The leading order processes for the top-antitop pair-production are shown in Figure 4.1. The process $q\bar{q} \to t\bar{t}$, referred to as $q\bar{q}$ annihilation, can be seen in Figure 4.1a. This process was with a fraction of approximately 85% the dominant process in the $t\bar{t}$-production at the TEVATRON proton-antiproton collider. However, at the LHC gluon fusion is the

underlying process for the majority of the produced $t\bar{t}$-pairs with around 90% and only 10% $q\bar{q}$ annihilation [12]. The Feynman diagrams for the gluon fusion processes are shown in Figure 4.1b - Figure 4.1d.
The difference between the $t\bar{t}$ production at the TEVATRON and the LHC can be explained by the parton distribution functions used in Equation 4.1 and the necessary momentum fraction x the colliding partons need to carry to produce a $t\bar{t}$ pair. Assuming that the two colliding partons carry the same fraction $x := x_i \approx x_j$ of momentum, the reduced centre-of-mass energy becomes $\hat{s} = x^2\, s$.

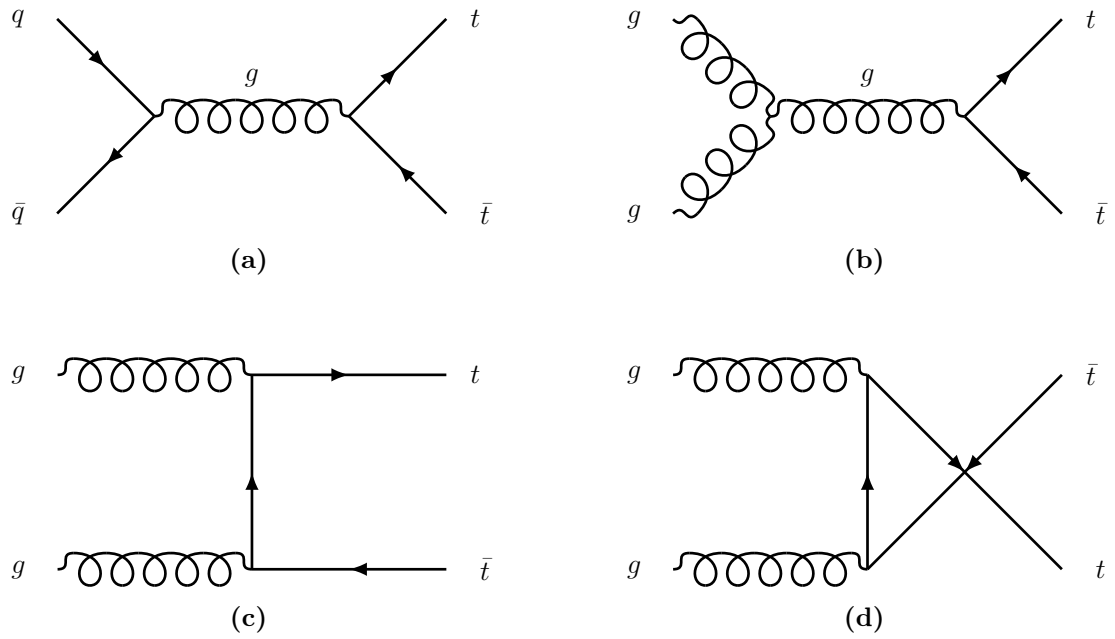


Figure 4.1: Leading order Feynman diagrams of $t\bar{t}$-production. **(a)** shows the production via $q\bar{q}$ annihilation, while **(b)** - **(d)** show the gluon fusion processes.

To produce a top-antitop pair, $\sqrt{\hat{s}}$ needs to be bigger than two times the top quarks rest mass, which leads to the condition $x \gtrsim 2\, m_t/\sqrt{s}$. Since the LHC features a centre-of-mass energy of 13 TeV, the threshold for x to produce a $t\bar{t}$ pair is approximately 0.027 compared to around 0.18 at the TEVATRON with a centre-of-mass energy of 1.96 TeV [26]. As can be seen in Figure 4.2, the PDF values multiplied with the corresponding value of x at $x \approx 0.18$ are higher for the quarks, which is why the $q\bar{q}$ annihilation dominates at the TEVATRON. However, for $x \approx 0.027$, the values of the gluon PDF are much higher than the values of the quark PDF and hence gluon fusion is the dominant process for $t\bar{t}$ production at the LHC.

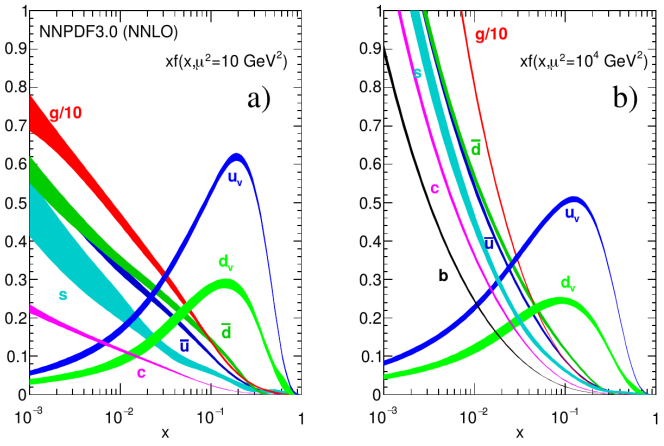


Figure 4.2: Parton distribution functions of the proton at two energy scales. The gluon distributions are scaled down by a factor of 10 [12].

4.3 Single-top-quark production

In addition to the $t\bar{t}$ pair production, top quarks are also produced as single quarks through the elektroweak interaction. The leading order Feynman diagrams are shown in Figure 4.3. The single-top-quark production can be divided into three separate processes, namely the t-channel production, the Wt production and the s-channel production.

As can be seen in Figure 4.3, all of these processes involve the Wtb vertex by either changing the flavour of a quark from b to t via the interaction with a charged W boson or a W boson decaying into a t quark and a $\bar{b}$ quark. In principle, the b quark could be replaced by a d quark or s quark in all three diagrams, but these processes are strongly suppressed due to the entry $|V_{tb}| = 1.019 \pm 0.025$ [12] in the CKM matrix. With $|V_{tb}|$ being so close to unity, the t quark interaction with a W boson includes in nearly 100 % a b quark.

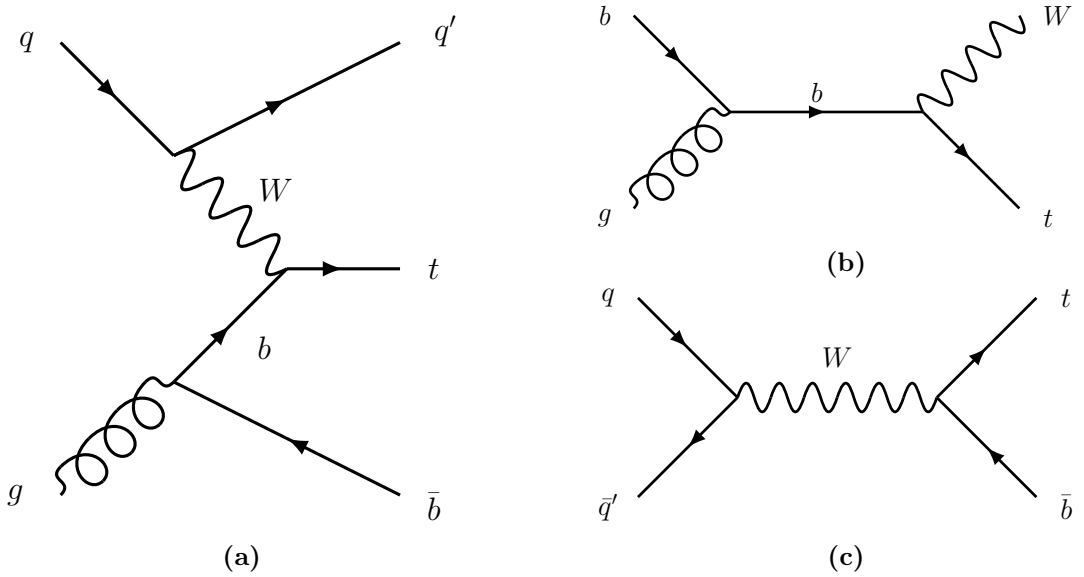


Figure 4.3: Feynman diagrams of the leading order processes for the single top quark production: **(a)** t-channel production, **(b)** Wt production, **(c)** s-channel production.

The cross section for the single-top-quark production is the composition of the cross sections of the three underlying processes. However, at the LHC the cross section of the t-channel process is expected to be more than three times as large as the cross section of the other two processes combined [12]. Hence with measured t-channel cross sections of $\sigma_t(tq) = 156 \pm 5\,(\text{stat.}) \pm 27\,(\text{syst.}) \pm 3\,(\text{lumi.})\,\text{pb}$ and $\sigma_t(\bar{t}q) = 91 \pm 4\,(\text{stat.}) \pm 18\,(\text{syst.}) \pm 2\,(\text{lumi.})\,\text{pb}$ [28] at 13 TeV compared to the measured cross section of the $t\bar{t}$ production $\sigma_{t\bar{t}} = 817 \pm 13\,(\text{stat.}) \pm 103\,(\text{syst}) \pm 88\,(\text{lumi.})\,\text{pb}$ [29], $t\bar{t}$ production dominates the top quark production at the LHC at 13 TeV.

4.4 Top-quark decay

Since the top quark itself decays in almost 100 percent into a real W boson and a b quark, its decay channels are classified based on the W boson decays. The W boson can either decay leptonically via $W \to \ell\,\nu_\ell$ or hadronically via $W \to q\bar{q}'$. The b quarks and the quarks from a hadronically decaying W boson can be seen in the detector as jets and the leptons of a leptonically decaying W boson are measured with the track and momentum of the charged lepton ℓ and the missing transverse momentum corresponding to the neutrino ν_ℓ. The three resulting decay channels of a $t\bar{t}$ pair are [12]:

- **all-jets**: $t\bar{t} \to W^+\,b\,W^-\,b \to q\,\bar{q}'\,b\,q''\,\bar{q}'''\,\bar{b}$ (45.7%)
- **lepton + jets**: $t\bar{t} \to W^+\,b\,W^-\,b \to q\,\bar{q}'\,b\,\ell^-\,\bar{\nu}_\ell\,\bar{b} + \ell^+\,\nu_\ell\,b\,q''\,\bar{q}'''\,\bar{b}$ (43.8%)
- **dilepton**: $t\bar{t} \to W^+\,b\,W^-\,b \to \ell^+\,\nu_\ell\,b\,\ell^-\,\bar{\nu}_\ell\,\bar{b}$ (10.5%)

Figure 4.4 shows a sketch of the $t\bar{t}$ decay. The three decay channels differ in their reconstruction techniques and both the all-jets and the dilepton channel are very hard to reconstruct.

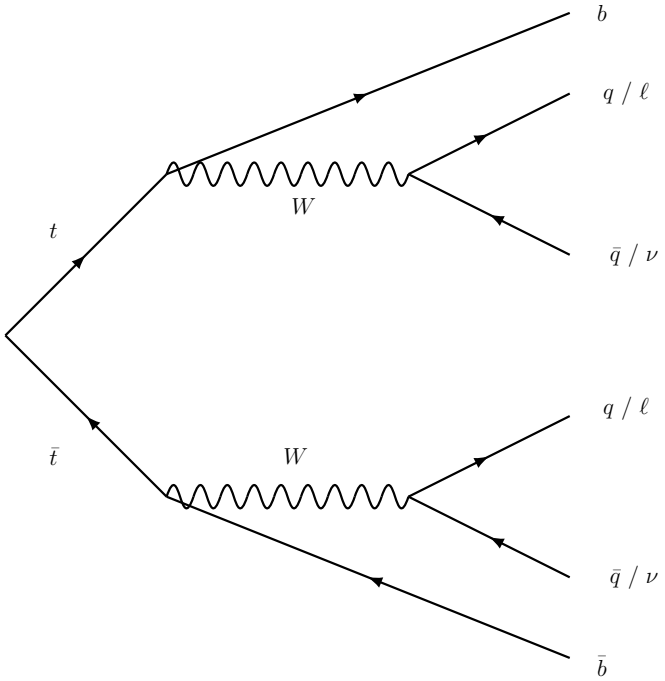


Figure 4.4: Sketch of the $t\bar{t}$ decay. The top quarks decay into bottom quarks and W Bosons, which can then either decay hadronically or leptonically.

In the all-jets channel there are at least six jets, but extra QCD radiation, i.e. quarks and gluons, from the coloured particles in the event can lead to additional jets. Therefore, the all-jets channel has a very large multi-jet background making a measurement in this channel very challenging.
The dilepton channel on the other hand includes two neutrinos, which means that the tracks and the momentum of two involved particles cannot be measured. Therefore, the reconstruction of a dilepton event is also quite challenging, as the missing transverse energy corresponds not only to one particle. Furthermore, this channel only makes up around 10 % of all $t\bar{t}$ decays, which means that the statistics in this channel are very low compared to the statistics in the other two channels.
However, the lepton + jets channel, which is used in the studies of this thesis, is easier to reconstruct. Since there is only one neutrino involved in the decay, the missing transverse momentum corresponds to this neutrino and the multi-jet background is not as large as in the all-jets channel.

4.5 $m_{\mathbf{top}}$ measurement in the lepton + jets channel with the ATLAS experiment at 8 TeV

This section gives a brief summary of the analysis performed in the ATLAS top-quark mass measurement at a centre-of-mass energy of 8 TeV in the lepton + jets channel [1] as this method will be used for the measurement at a centre-of-mass energy of 13 TeV analysis as well. In the following, the ATLAS top mass measurement at 8 TeV in the lepton + jets channel is referred to as the *8 TeV analysis* and the upcoming similar analysis using the data recorded at a centre-of-mass energy of 13 TeV is referred to as the *13 TeV analysis*.

In this analysis, a three-dimensional template fit to the measured data is performed in order to measure the top-quark mass m_{top}, a jet scale factor JSF and a relative b-to-light jets scale factor bJSF. The fit is performed as unbinned likelihood fit, where the templates utilised are functions obtained from a fit to simulated events and are parameterised as functions of one or more of the three parameters introduced above. To obtain the template functions for this fit, Monte Carlo (MC) simulated samples are used. The analysis chain to get these templates is

$$\text{MC Simulation} \rightarrow \underbrace{\text{preselection} \rightarrow \text{KLFitter} \rightarrow \text{BDT selection}}_{\text{event selection and reconstruction}} \rightarrow \text{template functions}\,.$$

Once these template functions are determined, the event reconstruction and selection procedure is also applied to the data and an unbinned likelihood fit is performed to the reconstructed observables. The different steps of the measurement and the analysis optimisation performed are briefly explained in the following.

4.5.1 Event reconstruction and event selection

Before applying the reconstruction algorithm, a first preselection is performed. This preselection intends to select potential events originating from the $t\bar{t} \rightarrow$ lepton + jets channel. The triggers applied in this analysis require the presence of a single high energy electron or muon. Further preselection requirements are[1]:

[1] More details about the object definition and selection can be found in Ref. [1].

- Events must have at least one primary vertex with at least five associated tracks, each having a minimum transverse momentum of 0.4 GeV.
- Selected events must contain exactly one reconstructed isolated charged lepton with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ that matches the charged lepton that fired the lepton trigger.
- If the charged lepton is a muon, $E_{\mathrm{T}}^{\mathrm{miss}} > 20\,\mathrm{GeV}$ and $E_{\mathrm{T}}^{\mathrm{miss}} + m_{\mathrm{T}}^{\mathrm{W}} > 60\,\mathrm{GeV}$ are required [2].
- If the charged lepton is an electron, $E_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$ and $m_{\mathrm{T}}^{\mathrm{W}} > 30\,\mathrm{GeV}$ are required.
- At least four jets with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.5$ are required to be present in the event.
- The events are required to have exactly two b-tagged jets, using the 70 % efficiency working point.

Events which pass this preselection are then reconstructed by the KLFitter framework as described in section 6.1. The algorithm assigns four of the jets in the event to the jets from the top quark decay, b_{had}, b_{lep}, q_1^W, q_2^W and therefore provides a full kinematic reconstruction of the event.

The observables used for the fit are the reconstructed top mass $m_{\mathrm{top}}^{\mathrm{reco}}$, the reconstructed mass of the W boson m_W^{reco} and an observable called R_{bq}^{reco} which is defined as

$$R_{bq}^{\mathrm{reco}} = \frac{p_{\mathrm{T}}^{b_{\mathrm{had}}} + p_{\mathrm{T}}^{b_{\mathrm{lep}}}}{p_{\mathrm{T}}^{q_1} + p_{\mathrm{T}}^{q_2}}\,. \tag{4.2}$$

In order to simplify the template parametrisation, additional cuts are applied to the three observables used for the fit. However, only the value of $m_{\mathrm{top}}^{\mathrm{reco}}$ is obtained directly from the kinematic fit while the values of m_W^{reco} and R_{bq}^{reco} are calculated from the original four vectors as given by the jet reconstruction. The additional requirements on these three observables are:

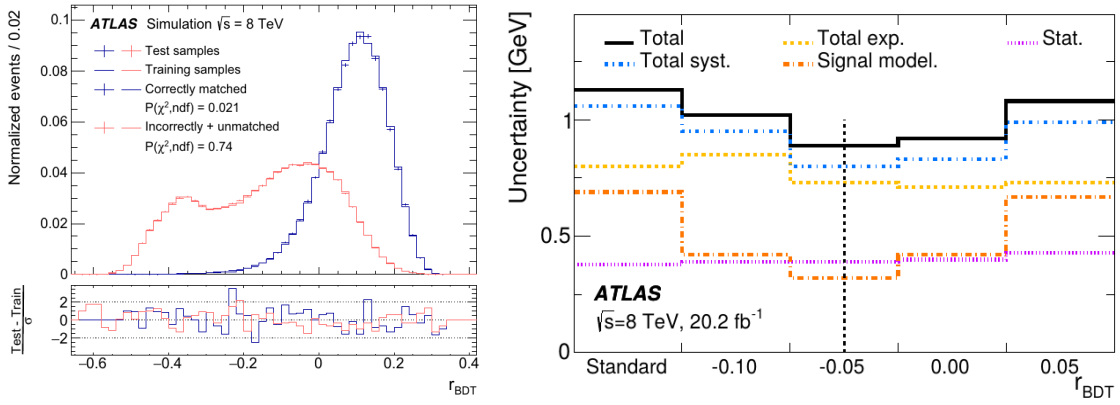
- $125\,\mathrm{GeV} < m_{\mathrm{top}}^{\mathrm{reco}} < 200\,\mathrm{GeV}$,
- $55\,\mathrm{GeV} < m_{\mathrm{W}}^{\mathrm{reco}} < 110\,\mathrm{GeV}$,
- $0.3 < R_{bq}^{\mathrm{reco}} < 3.0$.

The preselection with these additional requirements is referred to as *standard selection*.

The reconstructed events of the MC sample are matched to the closest parton level object within $\Delta R = 0.1$ for electron and muons and $\Delta R = 0.3$ for jets. After this matching, events are called *correctly matched* if they are unambiguously matched to the corresponding parton level objects, *incorrectly matched* if they are matched to a wrong parton level object (but if still all four jets from KLFitter can be matched to the four partons) and *unmatched* if one or more objects could not be matched to a parton-level object.

To reduce the fraction of incorrectly and unmatched events, these reconstructed events are further categorised using the output of a boosted decision tree (BDT). To have larger

[2] $m_{\mathrm{T}}^{\mathrm{W}}$ is the transverse mass of the W boson, defined as $\sqrt{2\,p_{\mathrm{T},\ell}\,E_{\mathrm{T}}^{\mathrm{miss}}\left(1-\cos\phi\left(\ell,\vec{E}_{\mathrm{T}}^{\mathrm{miss}}\right)\right)}$, where $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ provides an estimate of the transverse momentum of the neutrino.



(a) **(b)**

Figure 4.5: BDT selection at 8 TeV: **(a)** shows the BDT output for the training and test sample obtained by MC simulation and **(b)** shows the uncertainties of the m_{top} measurement as a function of the BDT cut compared to the standard selection [1].

statistics in the BDT training, the sample after the preselection, i.e. the sample without applying the additional cuts included in the standard selection, is used for the training. In the BDT training, correctly matched events are trained against incorrectly and unmatched events using variables from KLFitter. A list of the used variables and their description can be seen in Table 7.1. The output of the BDT is shown in Figure 4.5a. This training of the BDT intends to lower the fraction of poorly reconstructed events, assuming that those events are not well measured [1]. The trained BDT is then used to select events in the data sample.

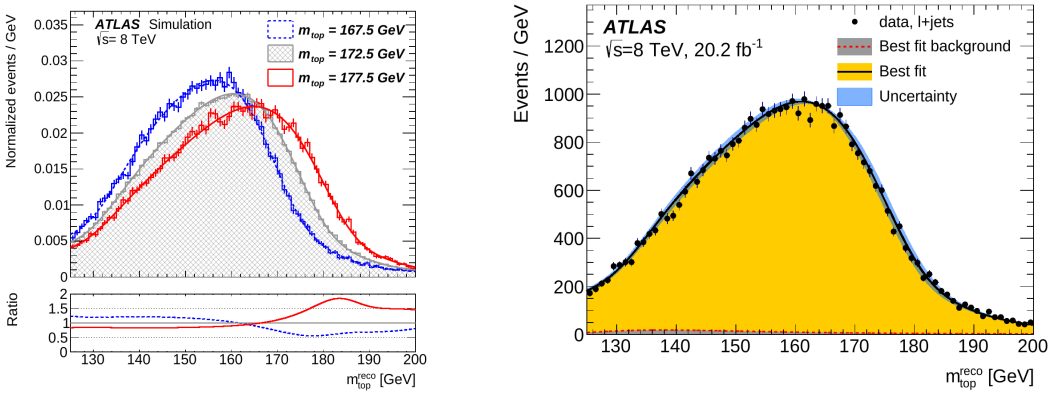
To determine the best BDT cut value[3] r_{BDT}, the full analysis is performed for different cut values. The uncertainties of the final result are shown in Figure 4.5b for both the analysis with the standard selection and the analysis with different r_{BDT} values. As can be seen in the plot, the uncertainties decrease after applying the BDT selection compared to the standard selection, but the uncertainties strongly depend on the BDT cut value. Therefore, the BDT cut value is chosen to be -0.05 since this value leads to the smallest overall uncertainty in the final result due to the reduction of the signal modelling uncertainty.

4.5.2 3D Template fit

Reconstructed events which pass the BDT selection are afterwards used to perform a 3 dimensional unbinned likelihood fit. The template parametrisation is re-done for each r_{BDT} cut value. The reconstructed mass of the W is sensitive to the jet energy scale factor JSF, R_{bq}^{reco} is sensitive to the b-jet energy scale factor bJSF and $m_{\text{top}}^{\text{reco}}$ is sensitive to both JSF and bJSF. Hence by not only fitting the top mass, but instead the two additional parameters JSF and bJSF, the systematic uncertainty due to jet energy scale (JES) and b-jet energy scale (bJES) uncertainties decreases, taking into account additional statistical components. Using the multidimensional fit reduces the JES-uncertainty-induced systematic uncertainty in m_{top} from 0.99 GeV for a 1D fit to 0.74 GeV for a 2D fit and 0.54 GeV for a 3D fit [1].

MC samples are produced for five different assumed values of m_{top} in the range from

[3]The BDT cut value is the minimum value of r_{BDT} events need to pass the BDT selection.



(a) **(b)**

Figure 4.6: (a) distribution of $m_{\text{top}}^{\text{reco}}$ for different input values of m_{top} in the MC simulation. **(b)** distribution of the data and the corresponding best fit result [1].

165.5 GeV to 177.5 GeV and the corresponding templates of $m_{\text{top}}^{\text{reco}}$ are constructed. Furthermore, for the sample at $m_{\text{top}} = 172.5\,\text{GeV}$, MC samples with different values for JSF and bJSF are produced and the corresponding templates of m_W^{reco} and R_{bq}^{reco} are constructed as well. The distributions of $m_{\text{top}}^{\text{reco}}$ together with the corresponding template is shown for three different input values of m_{top} in Figure 4.6a and the $m_{\text{top}}^{\text{reco}}$ distribution of the data and the corresponding best fit result can be seen in Figure 4.6b.

The top mass corresponding to the best fit is

$$m_{\text{top}} = (172.08 \pm 0.39\,(\text{stat.}) \pm 0.82\,(\text{syst.}))\,\text{GeV}\,, \tag{4.3}$$

with a total uncertainty of 0.91 GeV. This total uncertainty corresponds to a 19 % improvement compared to the result obtained using the standard selection. Furthermore, a combination with previous ATLAS m_{top} measurements leads to

$$m_{\text{top}} = (172.69 \pm 0.25\,(\text{stat.}) \pm 0.41\,(\text{syst.}))\,\text{GeV}\,. \tag{4.4}$$

5 Object definition and event preselection

For the studies presented in this thesis, $t\bar{t}$ events are simulated using Monte Carlo (MC) generators, followed by an event preselection and kinematic reconstruction. The event generator used for the simulation is the next-to-leading order (NLO) POWHEG-BOX (v2) [30] matrix-element (ME) event generator interfaced to PYTHIA8 (v8.23) [31] for the parton shower (PS) and hadronisation[1]. The parton distribution functions used for the simulation are NNPDF3.0NLO and NNPDF2.3LO and the PS Tune is A14 [33]. The ATLAS detector is fully simulated using GEANT 4 [34].

Only events of the lepton+jets channel and the dilepton channel are included in the Monte Carlo simulation. The fraction of all-jets events which would pass the preselection is negligibly small and which is why events of this channel are not included in the MC simulation in the first place. This chapter provides a brief overview of the object definitions used in these studies as well as the requirements of the event preselection.

5.1 Object definition

The objects used in these studies are reconstructed electrons, muons, jets and missing transverse momentum. Since the leptons in the event topology originate from the decay of a W boson, only high p_{T} leptons are accepted. The thresholds for an electron or muon to be accepted is $p_{\mathrm{T}} > 28\,\mathrm{GeV}$. Furthermore, the pseudorapidity of these objects has to be $|\eta| < 2.47$ for electrons and $|\eta| < 2.5$ for muons. If the energy cluster of an electron in the electromagnetic calorimeter is in the transition region between the barrel component and the endcap component of the calorimeter, $1.37 < |\eta| < 1.52$, the corresponding electron candidate is excluded.

Jets are reconstructed from the energy deposit in the calorimeters using the anti-k_t algorithm [35, 36] with a distance parameter of 0.4. The thresholds for accepting a jet are $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.5$. Since b quarks are essential to the reconstruction of the observed decay, b-tagging is performed using the MV2 algorithm [37] with a working point of 70 % efficiency. b-tagging is a method used to identify jets originating from a b quark by exploiting the characteristics of b-jets. These are for example the long lifetime of b-flavoured hadrons, which leads to a secondary vertex, and the high transverse momentum with respect to the jet axis of the b-hadron's decay products resulting in a wide opening angle of the jet.

5.2 Event preselection

These objects are then preselected with requirements similar to the requirements used for the analysis at 8 TeV:

- Events must have at least one primary vertex with at least two associated tracks, each having a minimum transverse momentum p_{T} of 0.4 GeV.
- Exactly one isolated lepton (electron or muon) with p_{T}>28 GeV

[1] Further settings can be found in Ref. [32], Table 2

- The charged lepton has to match the charged lepton that fired the single lepton trigger.
- The missing transverse momentum has to have $E_{\mathrm{T}}^{\mathrm{miss}} > 20\,\mathrm{GeV}$.
- At least four jets with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.5$ are required to be present in the event.
- The events are required to have exactly two b-tagged jets, using a 70 % working point.

Objects that pass this preselection are then used for the full kinematic reconstruction, which is explained in the following chapter.

6 Event reconstruction and multivariate classification methods

In this chapter, methods of event reconstruction and event classification that are used in this thesis are described. The KLFitter framework [38], explained in the first section, is used to fully reconstruct the events passing the preselection introduced in chapter 5 by performing a kinematic likelihood fit. The subsequent sections provide a brief introduction into the machine learning techniques used for the signal and background classification in this thesis, namely *boosted decision trees* (BDTs) and *artificial neural networks* (ANNs).

6.1 KLFitter

KLFitter (**K**inematic **L**ikelihood **Fitter**) is a likelihood-based reconstruction algorithm for top-related event topologies. This section explains the framework as it was applied in the top-quark mass measurement at $\sqrt{s} = 8\,\text{TeV}$ in [1] as well as in the studies presented in this thesis. If not stated differently, this section is based on [38] .
The study of top-quark properties often requires a full reconstruction of the top-quark final state using lepton and jet properties measured in the detector. In the lepton+jets channel, the decay signature is characterised by two b-jets, two jets originating from light quarks, an isolated high-p_T lepton and missing transverse momentum corresponding to the momentum of the neutrino. However, to analyse such an event, the measured jets have to be assigned to the quark they originated from. This leads to a large combinatorial background since there are a large number of possible ways to assign the measured jets to the quarks in the final state.
The algorithm requires as input the four-momenta of at least four jets j_i, the lepton four momentum and E_T^miss. The input four-momenta are then assigned to the quarks in the $t\bar{t}$ final state, namely the two b quarks, referred to as b_lep and b_had and the two light quarks, referred to as q_1 and q_2. This leads to in total 4!=24 possible assignments (permutations) like

$$\begin{aligned}(j_1,\, j_2,\, j_3,\, j_4) &\rightarrow (q_1,\, q_2,\, b_\text{had},\, b_\text{lep})\,,\\ (j_1,\, j_2,\, j_3,\, j_4) &\rightarrow (q_1,\, b_\text{lep},\, q_2,\, b_\text{had})\,,\\ &\vdots\end{aligned}$$

A large fraction of events has more than four jets measured in the detector, which is why this analysis uses up to six jets (the ones with the highest p_T) as input for KLFitter. This results in a larger CPU time per event reconstruction but leads to better efficiencies for correct matches. For six input jets there would be $6! = 720$ possible permutations, which is why some restrictions are made to these permutations. First of all, this analysis is not sensitive to a commutation of the jets assigned to q_1 and q_2, which reduces the number of possible permutations by a factor of two. Furthermore, permutations assigning b-tagged jets to non b quarks as well as permutations assigning non b-tagged jets to quarks $q_{1/2}$ are rejected.
Due to kinematics, the invariant mass of the b_lep-jet and the leptons as well as the invariant mass of the b_had-jet and the two light jets each have to be equal to the mass of the

top quark. Additionally, the invariant mass of the lepton and the missing transverse momentum as well as the invariant mass of the two light jets each correspond to the mass of a W boson. This is schematically shown in Figure 6.1. Thus for each permutation, a likelihood L is maximised with the mass of the top quark and the energies of the leptons and jets being free parameters while the mass of the W boson and the corresponding width are fixed to the measured values from the Particle Data Group [12]. This likelihood is defined as

$$\begin{aligned} L = & \mathcal{B}\left[m(q_1\, q_2)\,|\, m_W^{\text{reco}}, \Gamma_W\right] \cdot \mathcal{B}\left[m(\ell\,\nu)\,|\, m_W^{\text{reco}}, \Gamma_W\right] \cdot \\ & \mathcal{B}\left[m(q_1\, q_2\, b_{\text{had}})\,|\, m_{\text{top}}^{\text{reco}}, \Gamma_{\text{top}}\right] \cdot \mathcal{B}\left[m(\ell\,\nu\, b_{\text{lep}})\,|\, m_{\text{top}}^{\text{reco}}, \Gamma_{\text{top}}\right] \cdot \\ & \prod_i \mathcal{T}\left(E_i|\hat{E}_i\right) \ , \end{aligned} \tag{6.1}$$

with Breit-Wigner distributions $\mathcal{B}$ for the masses of the top quark and the W boson and transfer functions $\mathcal{T}$ for the measured jet and lepton energies. The transfer functions are the probability density distributions for the difference $E_{\text{meas}} - E_{true}$ between measured and true energy and are obtained from simulation. The longitudinal component of the neutrino $p_{\nu,\,z}$ is calculated from the $W \to \ell\,\nu$ mass constraint. Maximising the likelihood L leads to the maximum likelihood estimator for the parameters used in Equation 6.1 as explained in [39]. The permutation with the highest value for the maximised likelihood and the maximum likelihood estimators of $m_{\text{top}}^{\text{reco}}$ are used for the analysis. In the following, $m_{\text{top}}^{\text{reco}}$ always refers to the maximum likelihood estimator for the best permutation for a given event.

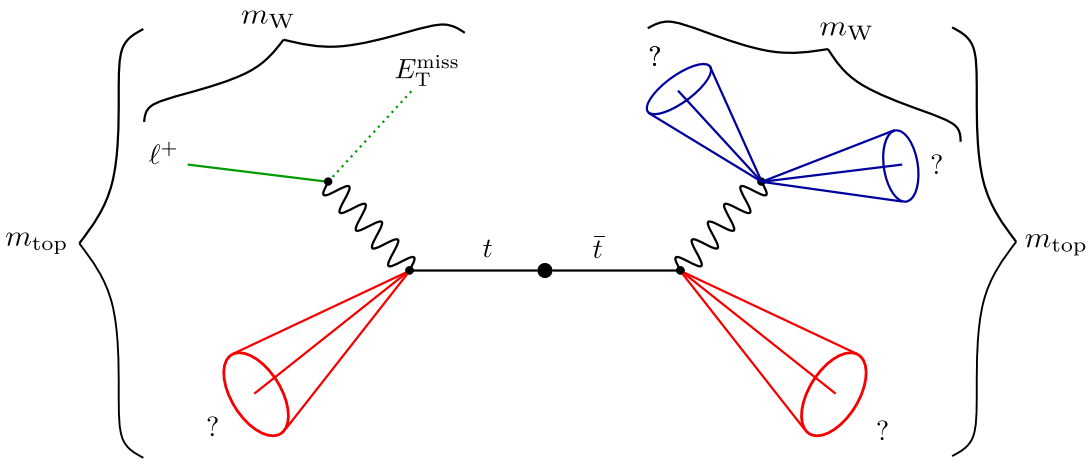


Figure 6.1: Sketch of the event reconstruction using the KLFitter framework. Both the left and the right side of the sketch corresponds to the invariant mass of the top quark. The two b-jets are drawn in red and the two light jets are drawn in blue. The lepton and the neutrino are drawn in green.

6.2 Boosted decision trees

A boosted decision tree (BDT) is a machine-learning method used for classification problems. In this particular case simulated samples are used to train a BDT which is then

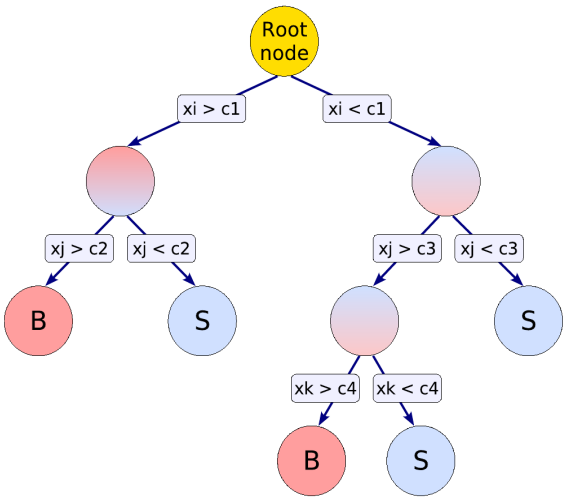


Figure 6.2: Schematic representation of a decision tree. At each node, the sample is split with respect to a specific variable. The final branches of the tree are labeled as either signal (blue) or background (red) depending on whether there are more signal or background events in the final subsample [40].

able to classify events in the data sample as either signal or background.
A single decision tree, as illustrated in Figure 6.2, is built by starting at a root node and then applying binary cuts on pre-defined variables x_i of the sample until a specified criteria is fulfilled. Which variable to split on and the corresponding cut value are determined by minimising the *loss function*

$$\mathcal{L}_{\text{split}} = \frac{n_l}{n_l + n_r} Q_l + \frac{n_r}{n_l + n_r} Q_r \,, \tag{6.2}$$

where n_l (n_r) is the number of events in the left (right) subsample after the splitting and Q_l (Q_r) is the node impurity measure, which is usually the *Gini index* [40, 41]. The Gini index is defined as

$$Q^{\text{Gini}} = p \cdot (1 - p). \tag{6.3}$$

where p is the fraction of signal events in the given subsample. Each branch is then labeled as either signal or background depending on whether more signal or background events are left in the subsample of this branch, illustrated by either the red B or blue S circles in the sketch. After building such a decision tree it can be applied to classifying data events depending on whether they correspond to a branch labeled as signal (1) or background (-1). However, single decision trees are very strongly affected by statistical fluctuations in the training sample. Therefore, the outputs $h_i = \pm 1$ of many different trees are combined to a *boosted decision tree* with a non-binary output. There are several boosting algorithms available, but the one used in this thesis is *AdaBoost* implemented in the TMVA framework [40]. The output of a BDT trained with the AdaBoost algorithm is

$$y_{\text{Boost}}(\mathbf{x}) = \frac{1}{N_{\text{trees}}} \sum_{i}^{N_{\text{trees}}} \ln\left(\frac{1 - \text{err}_i}{\text{err}_i}\right) \cdot h_i(\mathbf{x}) \,, \tag{6.4}$$

where err_i is the misclassification rate of the i^{th} tree [40]. This BDT is built by itera-

tively training single decision trees, evaluating the training sample with the current BDT, and reweighting wrongly classified events for the training of the next tree. According to Equation 6.4, events which are classified as signal by most of the decision trees result in positive values while primarily as background classified events result in negative values. The goal is to get output distributions of signal and background with only very little overlap.

The classification is then performed by defining a cut value of the BDT output which leads to either a signal classification if the BDT output of a event is higher than this value or a background classification if the BDT output is smaller than the cut value. An additional technique used for the training of BDTs is *bagging*. Bagging means that the trees are trained with randomly chosen subsamples of the full training sample, which is an additional way to stabilise the response of a classifier according to statistical fluctuations.

6.3 Artificial neural networks

Like boosted decision trees, artificial neural networks (ANNs) are algorithms which can be used to classify data after training it with a training sample. However compared to a BDT, neural networks are more complex systems and work in a quite different way. In principle, a ANN takes several variables as an input and combines them to an output of a specified dimension. This is schematically shown for an one dimensional output in Figure 6.3.

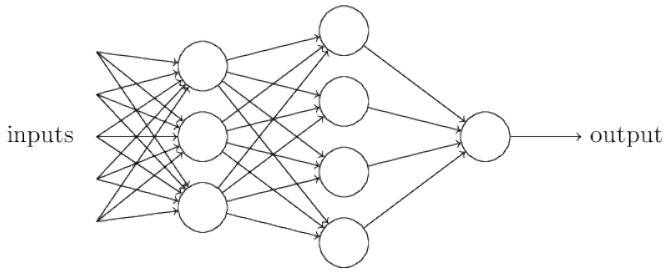


Figure 6.3: Schematic representation of a neural network [42].

In contrast to the architecture of a decision tree, which is determined through the training algorithm, the architecture of a neural network has to be specified before starting the training process. This architecture is defined by the dimension of the input layer, which is the number of input variables, the so-called *hidden layers*, which can have arbitrary dimension, and an output layer with a dimension suitable for the desired classification. Neural networks with many hidden layers are called *deep neural networks* (DNN). Figure 6.4 shows the architecture of a neural network with five input variables, three hidden layers of dimension nine, six and four, and an one-dimensional output layer.
Inspired by the human brain, the layers consist of artificial neurons, which do carry a certain activation. The neurons activation in the next layer is then calculated by taking into account all the neurons and the corresponding activation of the previous layer. Each connection is quantified by a weight w, which determines how strongly the activation of a neuron is affected by the corresponding neuron in the previous layer. The activation of the $(i+1)^{\text{th}}$ layer is then calculated by matrix multiplication

$$\vec{a}_{i+1} = f\left(\mathbf{W}_{i+1}\,\vec{a}_i + \vec{b}_{i+1}\right)\,, \tag{6.5}$$

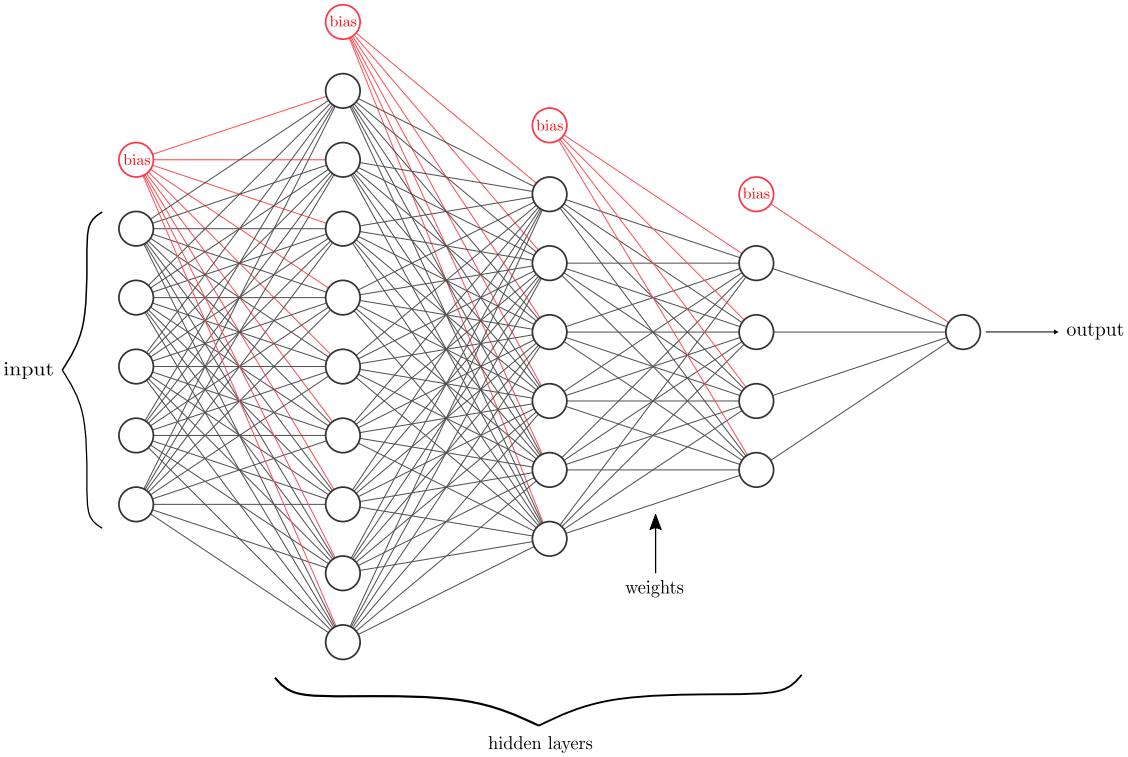


Figure 6.4: Architecture of a neural network with 5 input variables, three hidden layers of dimension 9, 6 and 4, and an one-dimensional output layer. The strength of the connection between two nodes is quantified by the weights. The activations of a layer is calculated via matrix multiplication using the activations of the previous layer, the corresponding bias and the activation function, which is not shown in the sketch.

where $\vec{a}_i$ is the vector of activations in the i^{th} layer, $\mathbf{W}_{i+1}$ is a matrix containing all the weights of the connections between the i^{th} and the $(i+1)^{\text{th}}$ layer, $\vec{b}_{i+1}$ is a bias vector, and f is an activation function which is applied to each component of its argument vector. The training of a neural network is then performed by minimising a loss function via changing the weights and biases of the ANN appropriately [42]. The loss function used in these studies is the *mean squared error*, which is defined as

$$\text{mse} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} (\hat{y}_i - y_i)^2 \ , \tag{6.6}$$

where $\hat{y}_i$ is the ANN-prediction[1] of the i^{th} event in the training sample and y_i is its true value (in this case 0 for a background event and 1 for a signal event). This minimisation can be achieved by an algorithm called *gradient descent*. This algorithm calculates the gradient of the loss function with respect to all the weights and biases of the ANN and then adjusts these parameters via a step into the negative direction of the gradient. However, in most cases, taking the full training sample into account for the gradient calculation requires too much computing time. Therefore, the training sample is split into so-called *batches* containing a given number of randomly chosen events from the original training sample. Instead of calculating the gradient step corresponding to the full training sample, the gradient is then calculated taking only the events of a single batch into account. These

[1] The prediction of the ANN is its output.

gradient steps do not provide the most efficient steps towards the minimum of the loss function but reduces the computation time considerably. Performing these calculations for all batches is called an *epoch* of the DNN training. After an epoch of the training is completed, the training sample is again randomly split into batches to start over with a new training epoch until a specified number of training epochs is completed. This version of gradient descent is called *stochastic gradient descent*. Parameters like the number of training epochs, the architecture of the ANN, the activation functions or the size of the batches are called *hyperparameters*, which are not adjusted by the training algorithm but can also be optimised with respect to a given training sample.

6.4 Overtraining

Overtraining (sometimes also called overfitting) is the effect when a trained classifier is very good at classifying the training data but performs very badly when it is applied to data it has not seen before. This means that the classifier was accidentally trained to memorise the training data instead of identifying patterns in the training sample. Hence the classifier is not robust against statistical fluctuations, which is why it performs very poorly on test data. Overtraining is often caused by either using a too small training sample, or using too many trainable parameters. A too large number of parameters gives the model the ability to adjust to the statistical fluctuations in the training sample. However, providing the model too few degrees of freedom results in a bad accuracy as the few parameters of the classifier are not sufficient to identify the patterns that are relevant for the classification. Therefore, the number of trainable parameters has to be chosen carefully. In these studies, to check for overtraining, the classifier is trained with half the sample (training sample) and then applied to the other half (test sample). Afterwards, the output of both the training and the test sample is compared to see if the classifier performs similar when applied to the test sample as when applied to the training sample.

6.5 The ROC curve

The *Receiver operating characteristic* curve (ROC curve), is a tool to describe the performance of a classifier for binary (positive/negative) classification problems. It is a graph which presents the relation between the *false positive rate* (FPR), i.e. the fraction of wrongly classified true negative objects, and the *true positive rate* (TPR), i.e. the fraction of the correctly classified true positive objects. In the case of classifying events measured in a detector as either signal (positive) or background (negative), a signal-classified event is also referred to as a selected event. The two common variables which form the ROC curve are then the *signal efficiency*, defined as

$$\text{signal eff.} = \text{TPR} = \frac{n_{\text{signal}}\,(\text{classified as signal})}{n_{\text{signal, total}}} \tag{6.7}$$

and the *background rejection*, defined as

$$\text{background rej.} = 1 - \text{FPR} = 1 - \frac{n_{\text{background}}\,(\text{classified as signal})}{n_{\text{background, total}}}\,. \tag{6.8}$$

The background rejection decreases monotonically with the signal efficiency. An example of several ROC curves corresponding to different classifiers is shown in Figure 6.5. Since the background rejection is monotonically decreasing with the signal efficiency, the area

under the ROC curve is a way to measure the performance of a classifier. A perfect classifier would have a background rejection of 1 and a signal efficiency of 1, since a perfect classifier would separate signal and background completely. The *area under the curve* (AUC) of a perfect classifier is equal to 1 which means that the closer the AUC of a classifier gets to 1, the better it performs.

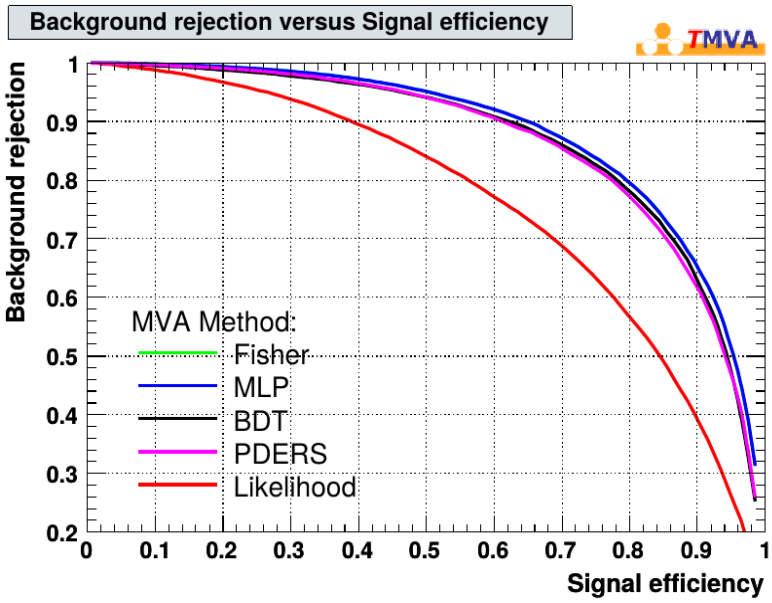


Figure 6.5: ROC curves of several classifiers [40].

7 Investigations and improvements of the BDT selection at 13 TeV

The BDT which is used in the 8 TeV analysis is trained with the sample containing all events passing the preselection. Before applying this BDT to the sample, the standard selection is performed, which means that additional cuts are applied on the observables used in the fit. This technique is used in order to have large statistics for the training since the additional cuts reduce the number of events in the sample significantly. However, the training sample which is available for the 13 TeV analysis is much larger than the training sample used in the 8 TeV analysis. Thereby, even the sample after the standard selection might still have sufficient statistics to be used for the BDT training. The training sample used for these studies contains half of the available events. In order to investigate the difference between using either the preselection sample or the standard selected sample for the training, both possibilities are tested and evaluated.

The first section of this chapter gives a short motivation for the BDT selection. Afterwards the variables which are used for the BDT training are investigated. The goal of these studies is to see if some variables used in the 8 TeV analysis could be removed to keep the training simpler. These studies are carried out using the preselection sample like in the 8 TeV analysis. In the third section, similar studies are performed with the sample after applying the standard selection. Finally, a comparison of the BDT performances at 8 TeV and 13 TeV is presented.

In all results presented here, a Monte Carlo sample with an input top mass of $m_{\text{top}} = 172.5\,\text{GeV}$ containing 7.1 million preselected $t\bar{t}$ events is used as described in chapter 5. The cuts used for the standard selection are explained in section 7.3. These additional cuts are referred to as *window cuts*.

7.1 Top-quark mass distribution after the preselection

As described in chapter 5, events passing the preselection are reconstructed with the KLFitter framework and afterwards matched to their closest parton level objects. After this matching process, events are divided into three classes, namely *correctly matched* events, *incorrectly matched* events and *unmatched* events. In the following, if not differently stated, correctly matched events are also referred to as signal events, while both incorrectly matched and unmatched events are also referred to as background events.

Figure 7.1 shows a zoomed in view of the distributions of the reconstructed top mass for both signal and background events. These distributions are again shown in Figure 10.1 with an extended $m_{\text{top}}^{\text{reco}}$ range. While the correctly matched events are distributed in a narrow, symmetric peak with its centre close to the input top mass, the incorrectly and unmatched events have a very broad distribution with large tails to higher values of m_{top}. The window cuts, which are studied in more detail in section 7.3, do already remove a large fraction of these tails. However, with a cut of $130\,\text{GeV} < m_{\text{top}}^{\text{reco}} < 200\,\text{GeV}$ on the top mass, there is still a large fraction of background events left. Hence the goal of the BDT selection is to reduce the fraction of incorrectly and unmatched events in the sample

which will lead to a more narrow $m_{\text{top}}^{\text{reco}}$ distribution.

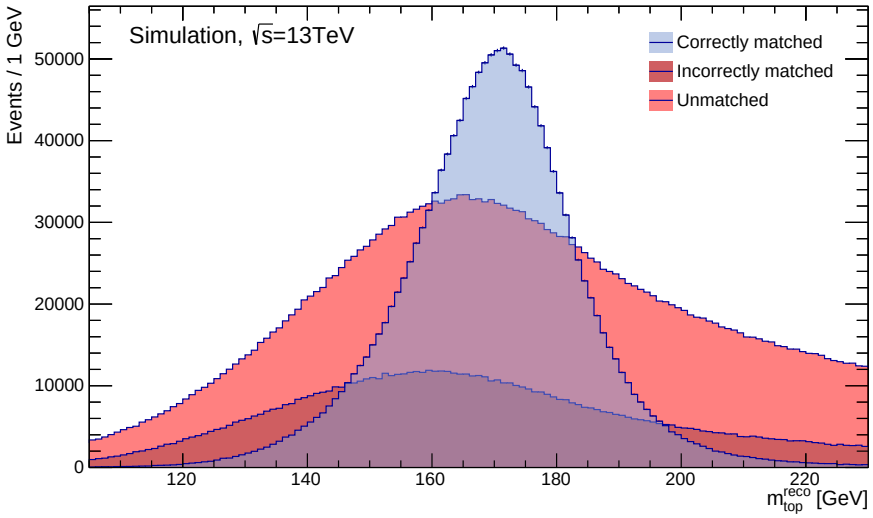


Figure 7.1: Distribution of the reconstructed top-quark mass for the three classes of events. The correctly matched events are distributed in a narrow, symmetric peak with its centre close to the input top mass while the incorrectly and unmatched events have a very broad distribution with large tails to higher values of m_{top}. The input value of the MC simulation is $m_{\text{top}} = 172.5\,\text{GeV}$.

7.2 BDT selection at 13 TeV

7.2.1 Studies on number of training variables used in BDT

The BDT selection at 13 TeV is based on the variables which have also been used in the analysis at 8 TeV. They are listed with their corresponding definitions in Table 7.1. However, many of these variables did not have much separation power, which means using them might have made the analysis more complex without increasing the performance of the BDT selection.

To investigate the effect of using these variables and to potentially simplify the analysis at 13 TeV, the performance of the BDT selection is studied for different sets of variables. The investigations are performed with all variables used in the 8 TeV analysis except the transverse momentum of the lepton $p_{\text{T}}(\ell)$. This variable is not considered in any studies, since the separation of this variable is only 0.1% in the 8 TeV analysis. In the following, the set with all 8 TeV variables besides $p_{\text{T}}(\ell)$ is referred to as the set with all variables. To quantify the separation power of a variable, the separation between the background and signal distribution[1] is defined as

$$\text{separation} = \frac{1}{2} \cdot \sum_{i}^{N_{\text{bins}}} \frac{(y_{\text{sig},i} - y_{\text{bkg},i})^2}{y_{\text{sig},i} + y_{\text{bkg},i}}\,, \tag{7.1}$$

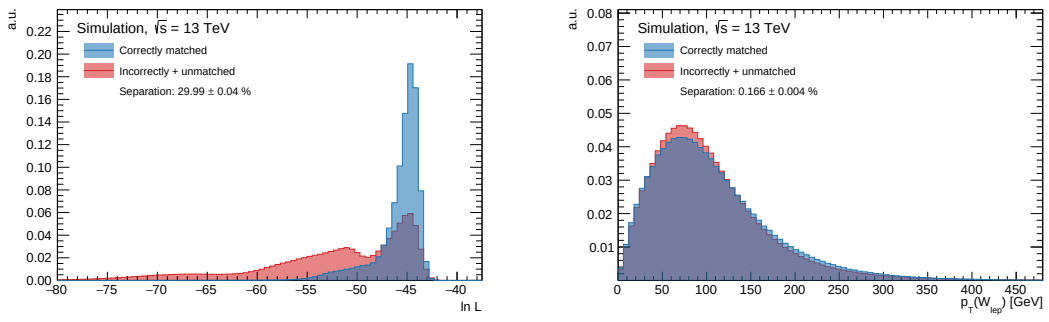
where $y_{\text{sig},i}$ and $y_{\text{bkg},i}$ are the values of the i^{th} bins of the normalised signal and back-

[1] It is assumed that the distributions are histograms with N_{bins} bins which are normalised to unity.

ground distributions of the given variable and the statistical uncertainty[2] is determined by Gaussian error propagation. The separation as defined in Equation 7.1 is equal to 1 if the two distributions do not overlap at all and becomes 0 if the normalised distributions have exactly the same values in each bin. The value of the separation is calculated for all variables and the resulting values are compared to the corresponding value calculated at 8 TeV in Table 7.2. The values of the separation at 13 TeV change a bit compared the ones at 8 TeV, which is likely caused by the difference in kinematics due to the larger centre-of-mass energy. In Figure 7.2 the distributions of both signal and background are shown for four both the variable with the largest separation, $\ln L$, and the variable with the smallest separation, $p_{\mathrm{T}}(W_{\mathrm{lep}})$. The distributions of the other variables can be found in the appendix in Figures 10.2-10.5.

Table 7.1: List of variables used for BDT training in the 8 TeV analysis [1].

Variable	Description
$\ln L$	logarithm of the event likelihood of the best permutation
$\Delta R(q,q)$	ΔR of the two untagged jets q_1 and q_2 from the hadronically decaying W boson
$p_{\mathrm{T}}(t_{\mathrm{had}})$	transverse momentum of the hadronically decaying top quark
P_{event}	relative event probability of the best permutation
$p_{\mathrm{T}}(W_{\mathrm{had}})$	transverse momentum of the hadronically decaying W boson
$p_{\mathrm{T}}(t_{\mathrm{lep}})$	transverse momentum of the leptonically decaying top quark
$m_{\mathrm{T}}(W)$	transverse mass of the leptonically decaying W boson
$p_{\mathrm{T}}(t\bar{t})$	transverse momentum of the reconstructed $t\bar{t}$ system
$\Delta R(b,b)$	ΔR of the two reconstructed b-tagged jets
N_{jets}	number of jets in the event
$E_{\mathrm{T}}^{\mathrm{miss}}$	missing transverse momentum
$p_{\mathrm{T}}(W_{\mathrm{lep}})$	transverse momentum of the leptonically decaying W boson
$p_{\mathrm{T}}(\ell)$	transverse momentum of the lepton (not used in 13 TeV study)



(a) **(b)**

Figure 7.2: Normalised distributions of **(a)** the variable with the largest separation,$\ln L$, and **(b)** the variable with the smallest separation, $p_{\mathrm{T}}(W_{\mathrm{lep}})$

[2] In this thesis, only the first significant digit of the statistical uncertainty is stated. If the first significant digit is equal to 1, the second digit is stated as well.

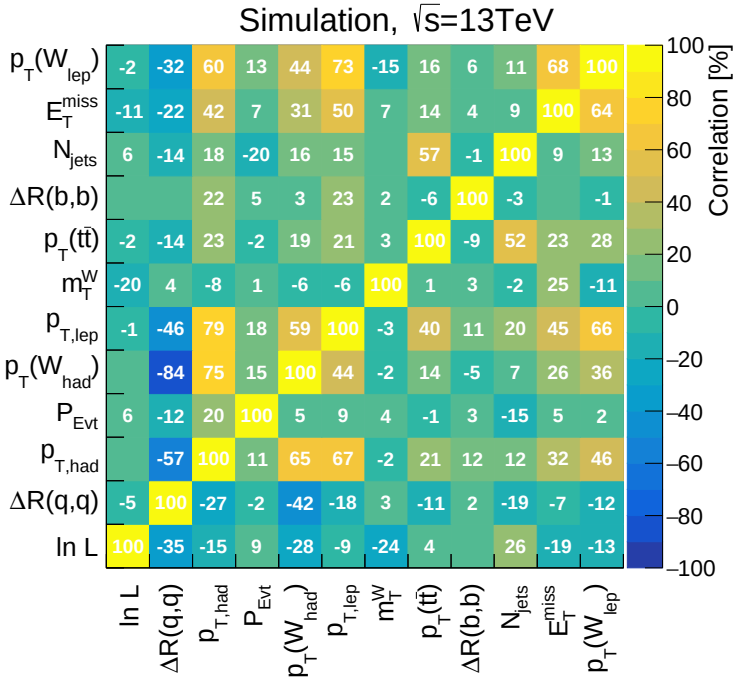


Figure 7.3: Correlation matrix of the variables used in the BDT training. The entries above the yellow diagonal show the correlations of signal events while the entries below the diagonal show the correlations of background events. The entries which do not show any number at all correspond to a correlation of 0.

The correlation matrix of the variables used at 13 TeV is shown in Figure 7.3 for both the signal events (above the diagonal) and the background events (below the diagonal). For most variables, the correlation is very small while some variables are correlated quite strongly. Those correlations are discussed in the following.
The largest correlation can be seen in the signal sample for $p_{\mathrm{T}}(W_{\mathrm{had}})$ and $\Delta R(q, q)$, which are strongly anti-correlated ($\approx -84\,\%$). However, this large correlation is expected due to momentum conservation, since the angular distance ΔR of the two light jets is decreasing for larger transverse momentum of the W boson.
In the background sample, not all jets are matched to the correct parton level object and consequently the two jets reconstructed as light jets are not necessarily the ones originating from the W decay, which is why this effect is not seen that strongly in the background sample ($\approx -42\,\%$). Furthermore, the four variables $p_{\mathrm{T}}(t_{\mathrm{had}})$, $p_{\mathrm{T}}(W_{\mathrm{had}})$, $p_{\mathrm{T}}(t_{\mathrm{lep}})$ and $p_{\mathrm{T}}(W_{\mathrm{lep}})$ are all positively correlated with each other. This can also be explained with conservation of momentum. Like before, these correlations are smaller for the incorrectly and unmatched events, which is expected. In general, introducing correlated variables to classification problems may not increase the performance of the classification method, since the new variables do not carry much new information the algorithm can learn from. In this case, this effect is not yet problematic as the correlations are, especially for background events, not too large and the separation power of these variables is still sufficient to improve the performance of the BDT. Moreover, different correlations in the signal and the background can lead to a gain in the separation power of the final classifier.

The BDT training is performed for eight different sets of variables using the TMVA framework [40]. In each training, 850 trees are built with the boosting algorithm *AdaBoost* [40]. The Gini index introduced in section 6.2 is used as the impurity measure and for each

variable the loss function is calculated for 20 values to decide which variable to split on and where. Additionally, bagging is applied with the fraction of the bagged sample being 0.5. The sample is split into a training and a test sample of same size by selecting events randomly. Starting from the variable with the smallest separation, the set is reduced by removing one additional variable at a time according to the ranking of separation in Table 7.2 each training.

Table 7.2: List of variables used for the BDT training in the 8 TeV analysis [1] with their value of separation compared to the values at 13 TeV.

Variable	Separation in % 8 TeV	Separation in % 13 TeV
$\ln L$	31	29.99 ± 0.04
$\Delta R(q,q)$	13	9.83 ± 0.02
$p_{\mathrm{T}}(t_{\mathrm{had}})$	4.3	4.590 ± 0.018
P_{event}	4.2	3.921 ± 0.017
$p_{\mathrm{T}}(W_{\mathrm{had}})$	5.0	3.222 ± 0.016
$p_{\mathrm{T}}(t_{\mathrm{lep}})$	1.7	2.322 ± 0.013
$m_{\mathrm{T}}(W)$	1.2	1.317 ± 0.009
$p_{\mathrm{T}}(t\bar{t})$	2.0	1.285 ± 0.010
$\Delta R(b,b)$	0.2	1.186 ± 0.009
N_{jets}	0.3	0.434 ± 0.006
$E_{\mathrm{T}}^{\mathrm{miss}}$	0.2	0.185 ± 0.004
$p_{\mathrm{T}}(W_{\mathrm{lep}})$	0.3	0.166 ± 0.004

For all sets of training variables, the BDT output is plotted for both the training and the test sample and the value of the separation is calculated. To check for consistency and especially for overtraining, the ratio between test and training output is calculated as well as the P-value of a χ^2 test for both signal and background. For both cases, the χ^2 test is performed with the two weighted histograms as described in [43].

Since the separation power of $p_{\mathrm{T}}(W_{\mathrm{lep}})$ and $E_{\mathrm{T}}^{\mathrm{miss}}$ is approximately the same according to their separation value, a training is done with dropping each of them separately. In contrast to the expected decrease in the separation for each dropped variable, the separation of the BDT output slightly increases when the missing transverse momentum is not used in the training. This can be seen in Table 7.3 for both the training with only $E_{\mathrm{T}}^{\mathrm{miss}}$ dropped when compared to the training with all variables as well as for the training with $E_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}(W_{\mathrm{lep}})$ dropped when compared to the training with only $p_{\mathrm{T}}(W_{\mathrm{lep}})$ left out. However, this increase is not significant but demonstrates that using more variables in a BDT training does not necessarily induce an increase in the separation. This might be caused by the fact, that decision trees are built iteratively based only on the value of the loss function after the next split instead of calculating the value of a loss function for the whole tree and then choosing the best configuration. Hence it is possible that, considering all possible decision trees, the BDT which is built with the algorithm is not the one with the best separation. Nevertheless, besides the small increase when removing

$E_{\mathrm{T}}^{\mathrm{miss}}$, the separation decreases with the number of dropped variables and the set with three dropped variables is selected for further studies. This choice is justified by accepting a small decrease of less than one percent in the separation while dropping all three variables with an individual separation less than one percent. In the following, this set of variables is referred to as the *reduced set* of variables.

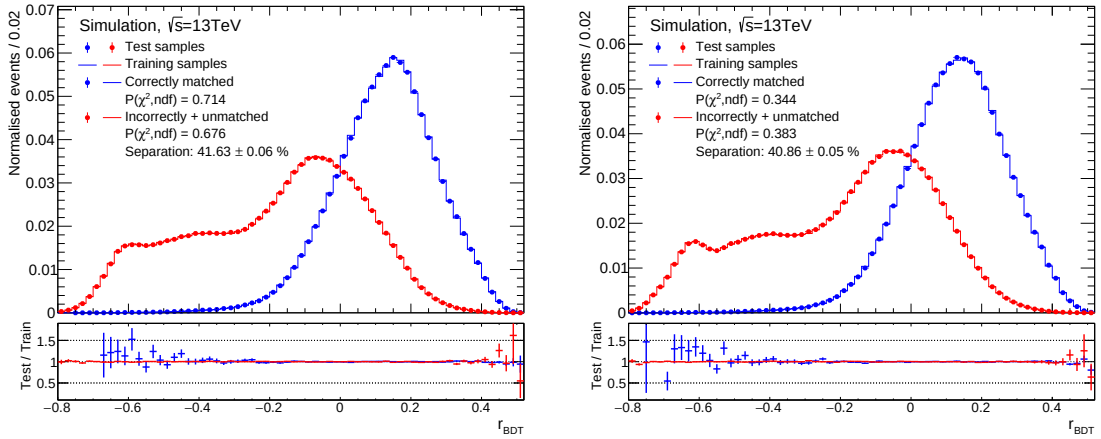
The values for the separation and the according list of dropped variables are compared in Table 7.3 and the distributions of the BDT output are shown for the training with all variables and for the training with 3 dropped variables in Figure 7.4. The separation is calculated using the BDT output of the test sample while taking into account MC weights as well as lepton and jets scale factors, such as b-tagging scale factors. As can be seen in Figure 7.4, the ratio of the BDT output of the test sample and the one of the training sample is close to unity for most bins. For bins with low statistics the ratio increases, but these larger values are acceptable since the difference is covered by statistical uncertainties. The P-values listed in the legend of the BDT output plots are quite large which also confirms the consistency between the classification of the training and the test samples.

Table 7.3: List of variables which are not used for the training compared to the set of variables used at 8 TeV with the resulting BDT separation. The set with three dropped variables, i.e. the ones with a separation lower than 1% is chosen for further studies.

Variables *not* used in the BDT training	BDT separation in %
- (all 12 variables used)	$\mathbf{41.63 \pm 0.06}$
$E_{\mathrm{T}}^{\mathrm{miss}}$	41.69 ± 0.06
$p_{\mathrm{T}}(W_{\mathrm{lep}})$	41.44 ± 0.06
$p_{\mathrm{T}}(W_{\mathrm{lep}})$, $E_{\mathrm{T}}^{\mathrm{miss}}$	41.49 ± 0.06
$\mathbf{p_{T}(W_{lep})}$, $\mathbf{E_{T}^{miss}}$, $\mathbf{N_{jets}}$	$\mathbf{40.86 \pm 0.05}$
$p_{\mathrm{T}}(W_{\mathrm{lep}})$, $E_{\mathrm{T}}^{\mathrm{miss}}$, N_{jets} , $\Delta R(b,b)$	40.75 ± 0.05
$p_{\mathrm{T}}(W_{\mathrm{lep}})$, $E_{\mathrm{T}}^{\mathrm{miss}}$, N_{jets} , $\Delta R(b,b)$, $p_{\mathrm{T}}(t\bar{t})$	40.06 ± 0.05
$p_{\mathrm{T}}(W_{\mathrm{lep}})$, $E_{\mathrm{T}}^{\mathrm{miss}}$, N_{jets} , $\Delta R(b,b)$, $p_{\mathrm{T}}(t\bar{t})$, $m_{\mathrm{T}}(W)$	39.86 ± 0.05

7.2.2 Performance of the BDT selection

The ROC curves of both the BDT with the full set of variables as well as the reduced set of variables are shown in Figure 7.5. The BDT training with all variables results in a AUC of 0.870, while the training with the reduced set of variables leads to a only a little smaller AUC of 0.867. This indicates, that removing the three variables with a separation $< 1\,\%$ does not have a large effect on the performance of the BDT.
Further investigations on the performance of the BDT selection are done by calculating the matching efficiency $\epsilon_{\mathrm{match}}$ and the selection efficiency ϵ_{sel} of the BDT selection as well as the purity π of the sample after the BDT selection. The selection efficiency is defined as the number n_{c} of correctly matched events passing the BDT selection divided by the



(a) **(b)**

Figure 7.4: Output of the BDT Training with **(a)** all variables used for the training and **(b)** the reduced set of variables (only those with a separation $> 1\,\%$). The solid lines are the output distributions of the training sample and the points are the BDT output distributions of the test sample. The background samples (training and test) are drawn in red and the signal samples are drawn in blue. The statistical uncertainty is drawn for the test sample although it is too small to be visible. For the ratio plot, statistical uncertainties for both distributions are taken into account.

sum $n_\mathrm{c} + n_\mathrm{i}$ of both correctly and incorrectly matched events passing the selection

$$\epsilon_\mathrm{match} = \frac{n_\mathrm{c}}{n_\mathrm{c} + n_\mathrm{i}} \,. \tag{7.2}$$

The total selection efficiency is the number n_sel of events passing the BDT selection divided by the total number n_total of events in the sample, which returns the fraction of events passing the BDT selection

$$\epsilon_\mathrm{sel} = \frac{n_\mathrm{sel}}{n_\mathrm{total}} \,. \tag{7.3}$$

Finally, the purity of the selected sample is defined as the fraction of correctly matched events among all the events in the sample after the selection

$$\pi = \frac{n_\mathrm{c}}{n_\mathrm{sel}} = \frac{n_\mathrm{c}}{n_\mathrm{c} + n_\mathrm{i} + n_\mathrm{u}} \,, \tag{7.4}$$

where n_u is the number of unmatched events passing the selection. To analyse the BDT performance as a function of the r_BDT cut value, the three values ϵ_match, ϵ_sel and π are calculated for different BDT cuts. Figure 7.6 shows the corresponding curves for the full set of variables. The corresponding plot for the reduced set of variables can be seen in the appendix in Figure 10.6. The statistical uncertainties are calculated via Gaussian error propagation although they are too small to be visible in the plot due to the large statistics. As expected, by increasing the cut value and thereby decreasing the selection efficiency, the matching efficiency and the purity of the sample increases monotonically with the cut value.

For a fair comparison between the BDT performance of the BDT trained with the full set of variables and the performance of the BDT trained with the reduced set of variables,

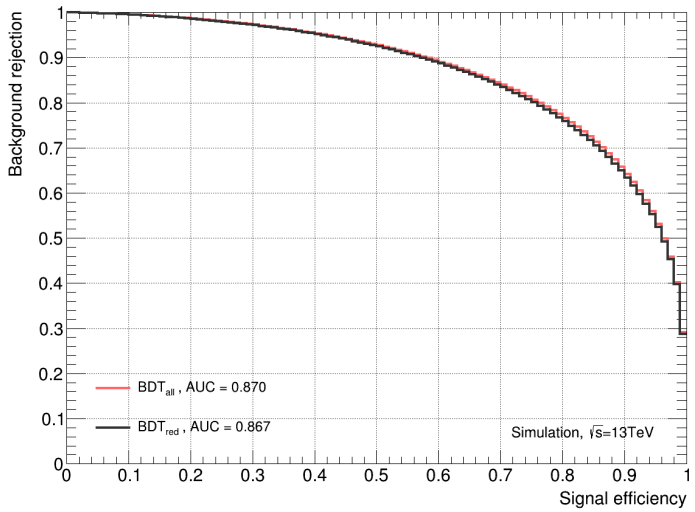


Figure 7.5: ROC curves of the BDT training using all variables (pink curve) or the reduced set of variables (grey curve).

ϵ_{match} and π are compared at approximately the same selection efficiency. In principle, this comparison can be made for any selection efficiency. However, in order to also compare these values with the BDT-performance at 8 TeV later in section 7.3, the same selection efficiency as used in the 8 TeV analysis is already chosen for this comparison.
At 8 TeV, the selection efficiency is 42.5 % [1], which is why the BDT cuts used for this comparison are chosen to correspond to the same ϵ_{sel} resulting in r_{BDT} cuts of -0.02 for the full set of variables and -0.017 for the reduced set of variables. The exact values of this comparison are listed in Table 7.4. With the same selection efficiency, the BDT with the reduced set of variables has only slightly lower values in both the matching efficiency and the purity. This shows again, that removing the variables with a separation smaller than one percent does not result in a significant drop in the performance of the BDT.

Table 7.4: Comparison of the matching efficiency and purity after the BDT selection for the full set of variables and for the reduced set at 13 TeV.

selection	ϵ_{sel} [%]	ϵ_{match} [%]	π [%]
BDT_{all} (≥ -0.02)	42.54 ± 0.03	77.35 ± 0.09	50.34 ± 0.05
BDT_{red}(≥ -0.017)	42.56 ± 0.03	76.73 ± 0.08	50.00 ± 0.05

7.2.3 Composition of the sample after the BDT selection

Since the aim of the BDT selection is to increase the fraction of correctly matched events and thus reducing the fraction of both incorrectly and unmatched events in the sample, these fractions are plotted as a function of the r_{BDT} cut in Figure 7.7 for the full set of variables. The corresponding plot for the reduced set of variables is shown in the appendix in Figure 10.7. As described in subsection 7.2.2, the fraction of correctly matched events, which is the purity of the sample, increases monotonically as a function of the cut value. However, while the fraction of unmatched events in the sample decreases monotonically, the fraction of incorrectly matched events does even increase a little for cut values below

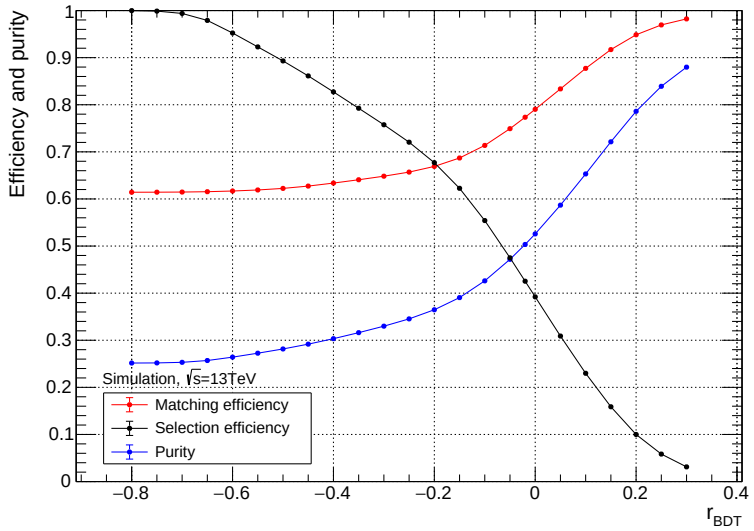


Figure 7.6: Matching efficiency, selection efficiency and purity as a function of the r_{BDT} cut value. The BDT training is performed with the full set of variables. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

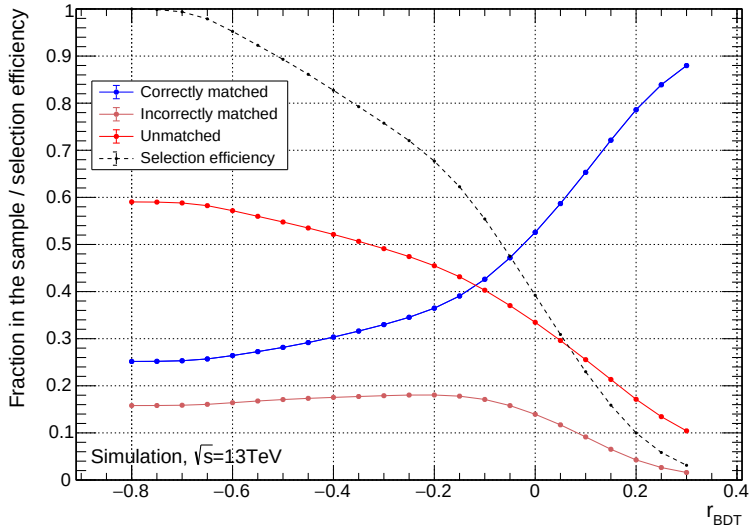


Figure 7.7: Fraction of correctly (blue), incorrectly (dark red) and unmatched (light red) events in the sample after the BDT selection as a function of the r_{BDT} cut. The training is performed with the full set of variables. The fraction of correctly matched events is equal to the purity of the sample. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

-0.2, stays relatively constant over a large range of r_{BDT} and decreases only for higher cut values. Thereby, reducing the fraction of unmatched events in the sample can only be achieved for very high cut values which means for low selection efficiencies.

7.3 BDT selection at 13 TeV with additional cuts

To investigate the effect of applying the additional window cuts to the sample before training the BDT, similar studies as presented in the previous section are performed. The cuts are applied to the three observables used in the m_{top} measurement and are a bit more stringent than in [1]. The requirements to pass this window cut selection are:

- $130\,\text{GeV} < m_{\text{top}}^{\text{reco}} < 200\,\text{GeV}$,
- $55\,\text{GeV} < m_{\text{W}}^{\text{reco}} < 110\,\text{GeV}$,
- $0.5 < R_{bq}^{\text{reco}} < 3.0$.

Introducing this additional selection is expected to remove poorly reconstructed events, i.e. unmatched and incorrectly matched events. This can be seen in Figure 7.1, as these cuts do already remove many incorrectly and unmatched events which are located in the tails of the distributions while there are only a few correctly matched events outside of the window. Applying those window cuts provides a new phase space for the BDT. In the following, studies are performed to investigate whether these cuts change the performance of the training.

Table 7.5: List of variables used for the BDT training in the studies with additional window cuts and their value of separation. For comparison, the values of the separation after the preselection are also listed.

Variable	Separation in % preselection (13 TeV)	Separation in % preselection + window cuts (13 TeV)
$p_{\text{T}}(W_{\text{had}})$	3.222 ± 0.016	9.27 ± 0.03
$\Delta R(q,q)$	9.83 ± 0.02	9.27 ± 0.03
$p_{\text{T}}(t_{\text{had}})$	4.590 ± 0.018	7.39 ± 0.03
P_{event}	3.921 ± 0.017	6.77 ± 0.03
$\ln L$	29.99 ± 0.04	5.93 ± 0.02
$p_{\text{T}}(t_{\text{lep}})$	2.322 ± 0.013	3.26 ± 0.02
$p_{\text{T}}(t\bar{t})$	1.285 ± 0.010	2.940 ± 0.019
N_{jets}	0.434 ± 0.006	2.829 ± 0.018
$p_{\text{T}}(W_{\text{lep}})$	0.166 ± 0.004	0.828 ± 0.010
$m_{\text{T}}(W)$	1.317 ± 0.009	0.522 ± 0.008
$\Delta R(b,b)$	1.186 ± 0.009	0.341 ± 0.006
$E_{\text{T}}^{\text{miss}}$	0.185 ± 0.004	0.142 ± 0.004

7.3.1 BDT variables studies after applying the window cuts

The additional selection reduces the number of the events entering the BDT selection by a little more than 50 % and the distributions of the variables used for the BDT training and the corresponding values of the separation change very strongly. The distributions are compared to the distributions without the window cuts in the appendix in Figures

10.2-10.5. The variables and their separation are listed in Table 7.5 and the correlation matrix is shown in Figure 7.8. The two variables with the largest separation are now $p_{\mathrm{T}}(W_{\mathrm{had}})$ and $\Delta R(q,q)$. Furthermore, the separation of $\Delta R(b,b)$ is now less than one percent, which leads to four variables with a separation lower than one percent.
The correlations between the BDT variables also after applying the without window cuts. In general, the difference between the correlations in the signal events and the correlations in background events are now quite small, especially the correlation between $p_{\mathrm{T}}(W_{\mathrm{had}})$ and $\Delta R(q,q)$ is now approximately the same in both signal and background. Since the window cuts are intended to remove very poorly reconstructed events from the sample, the majority of the events in the resulting background sample is expected to show more similarities with the events in the signal sample than before. Therefore, this decrease of the differences is expected.

As in the studies without the window cuts, the number of variables used in the BDT training is reduced. The training is performed with the same settings as in the studies without the window cuts. The BDT output of both the training with all variables and the training with only variables of separation larger than one percent are shown in Figure 7.9 and the list of dropped variables and the resulting BDT separation of all performed trainings can be found in Table 7.6. Again, the separation is calculated using the BDT output distribution of the test sample. The shape of the BDT output changes when applying the window cuts leading to two symmetrical distributions with a large overlap. This results in a separation of only 24 % for the training with all variables.

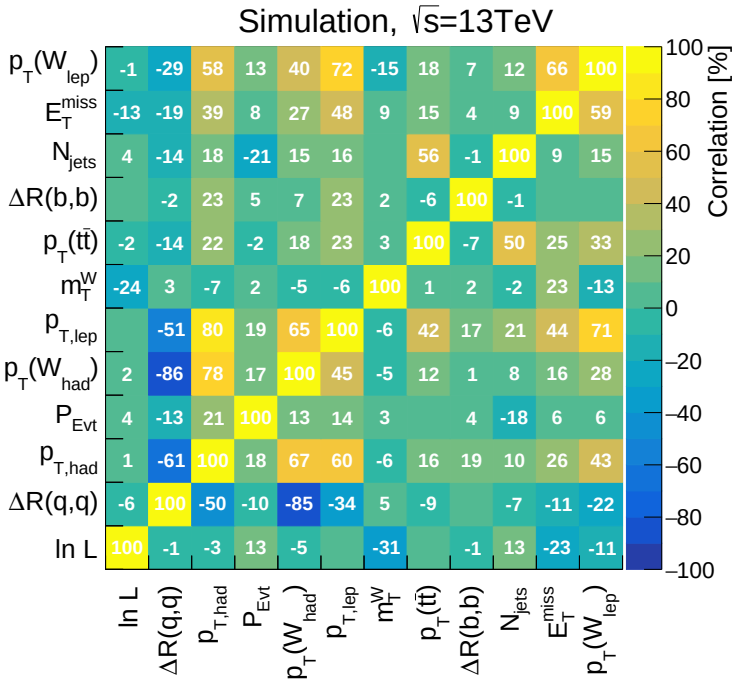
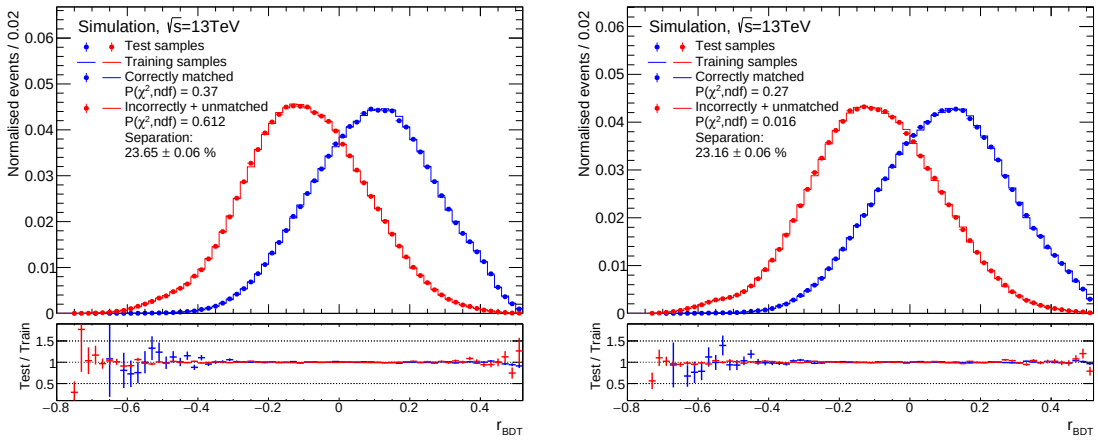


Figure 7.8: Correlation matrix of the variables used in the BDT training after applying additional window cuts. The entries above the yellow diagonal show the correlations of signal events while entries below the diagonal show the correlations of background events. In contrast to the sample without window cuts, the differences between the correlations in the signal sample and the ones in the background sample are very small. The entries which do not show any number at all correspond to a correlation of 0.

Table 7.6: List of variables which are not used for the training with additional window cuts. The set with dropping the four variables with a separation lower than 1% is chosen for further studies (reduced set).

Variables *not* used in the BDT training	BDT separation in %
- (all 12 variables used)	$\mathbf{23.65 \pm 0.06}$
$E_{\mathrm{T}}^{\mathrm{miss}}$	23.69 ± 0.06
$E_{\mathrm{T}}^{\mathrm{miss}}$, $\Delta R(b,b)$	23.50 ± 0.06
$E_{\mathrm{T}}^{\mathrm{miss}}$, $\Delta R(b,b)$, $m_{\mathrm{T}}(W)$	23.36 ± 0.06
$\mathbf{E_T^{miss}}$, $\mathbf{\Delta R(b,b)}$, $\mathbf{m_T(W)}$, $\mathbf{p_T(W_{lep})}$	$\mathbf{23.16 \pm 0.06}$
$E_{\mathrm{T}}^{\mathrm{miss}}$, $\Delta R(b,b)$, $m_{\mathrm{T}}(W)$, $p_{\mathrm{T}}(W_{\mathrm{lep}})$, N_{jets}	22.52 ± 0.06
$E_{\mathrm{T}}^{\mathrm{miss}}$, $\Delta R(b,b)$, $m_{\mathrm{T}}(W)$, $p_{\mathrm{T}}(W_{\mathrm{lep}})$, N_{jets} , $p_{\mathrm{T}}(t\bar{t})$	21.71 ± 0.06



(a) **(b)**

Figure 7.9: Output of the BDT Training with additional window cuts with **(a)** all variables used for the training and **(b)** with the reduced set of variables (only those with a separation $> 1\,\%$). The solid lines are the output distributions of the training sample and the points are the BDT output distributions of the test sample. The background samples (training and test) are drawn in red and the signal samples are drawn in blue. The statistical uncertainty is drawn for the test sample but too small to be visible. For the ratio plot, statistical uncertainties for both distributions are taken into account.

7.3.2 Performance of the BDT selection with additional window cuts

The ROC curves of the BDT trained with all variables and the BDT trained with the reduced set of variables are shown in Figure 7.10. The AUC is 0.781 for the training with all variables used and 0.778 for the reduced set of variables. Again, the total difference in the AUC between the two BDTs trained with different sets of variables is 0.003, which is quite small, and thereby the reduced set of variables leads to a similar performance as the full set of variables. The matching efficiency, selection efficiency and the purity are plotted

as a function of the r_{BDT} cut in Figure 7.11. The plot for the reduced set of variables can be seen in the appendix in Figure 10.8. As the window cut selection already removes many events before the BDT selection is applied, the graph of the selection efficiency starts at approximately 47 % for low cut values. Thereby, to obtain a selection efficiency of 42.5 % like at 8 TeV, very low cuts at -0.27 and -0.28 have to be made.

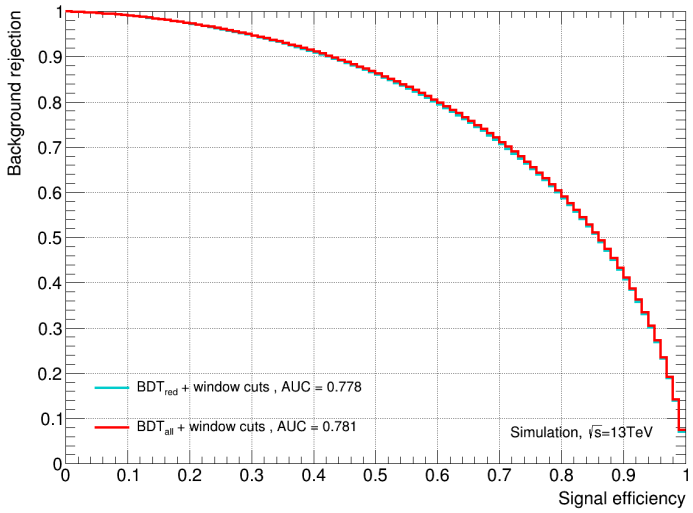


Figure 7.10: ROC curve of the BDT selection when applying the window cuts and using all variables (red curve) or the reduced set of variables (cyan curve). The cyan curve is mostly covered by the red curve and thereby only barely visible.

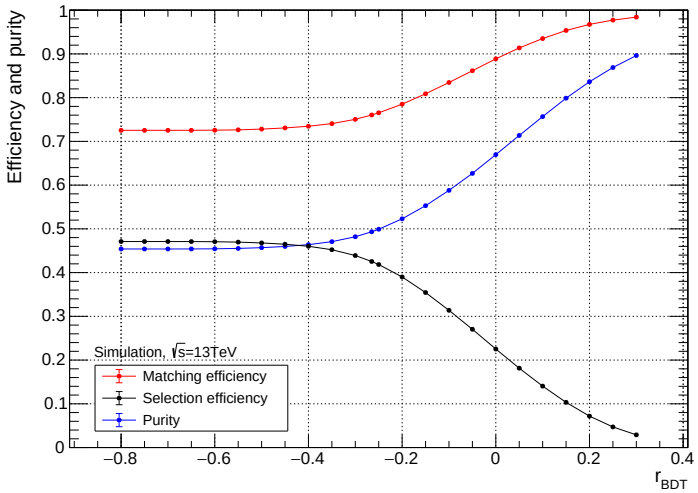


Figure 7.11: Matching efficiency, selection efficiency and purity as a function of the r_{BDT} cut. The BDT training is performed with the full set of variables and the window cuts are applied. The selection efficiency is defined with respect to the preselection, and hence it is already lower than 50 percent before applying the r_{BDT} cut. The statistical uncertainties are drawn in the graph, though they are too small to be visible.

The window cuts do not have a large effect on the composition of the sample after the BDT selection concerning the fraction of correctly, incorrectly or unmatched events. This can be seen when comparing the plot of these fractions without window cuts in Figure 7.7 and with window cuts in Figure 7.12. The corresponding plot for the training with the reduced set looks very similar and can be seen in Figure 10.9. By applying the window cuts, the fraction of unmatched events in the sample decreases from around 59 % to approximately 37 % before the BDT selection. However, for the same selection efficiency, the composition of the sample after the BDT selection is approximately the same as without window cuts.

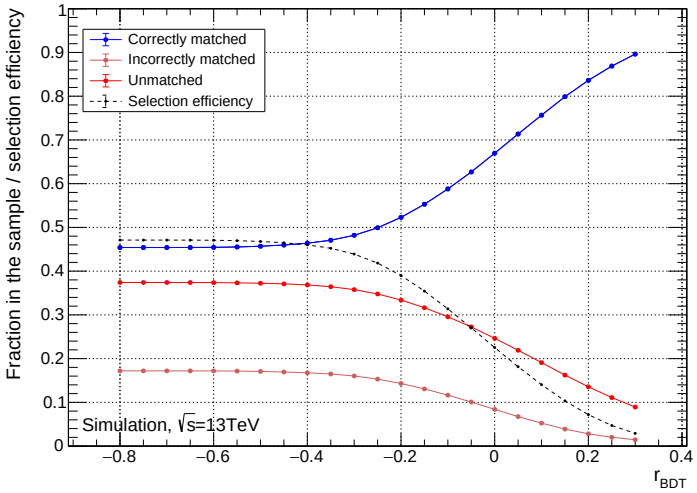


Figure 7.12: Fraction of correctly (blue), incorrectly (dark red) and unmatched (light red) events in the sample after the BDT selection as a function of the r_{BDT} cut. The fraction of correctly matched events in the sample is equal to the purity of the sample. Additional window cuts are applied before the BDT selection, which is why the selection efficiency starts at a value even below 50 %. The statistical uncertainties are drawn in the graph, though they are too small to be visible.

7.3.3 Comparison of the BDT performance at 8 TeV and 13 TeV

In Table 7.7, the performance improvement compared to the preselection is shown for both the normal BDT selection and the BDT selection with additional window cuts for a selection efficiency of approximately 42.5 %. Furthermore, the corresponding values from the 8 TeV analysis are listed in the table. As mentioned before, a selection efficiency of 42.5 % is used in the final analysis at 8 TeV, which is why this value is chosen for the comparison of the different BDTs.
Without the BDT selection, i.e. after the preselection, the matching efficiency and the purity at 13 TeV are with approximately 61 % and 25 % significantly lower than the values after the preselection at 8 TeV, which are 71% and 28 % [1]. The difference in these values might be related to the fact that the matching efficiency decreases for higher numbers of jets in the event. At 13 TeV, $\epsilon_{\mathrm{match}}$ is approximately 74 % for the subsample of events containing exactly 4 jets while the matching efficiency for six or more jets is around 43 %. As the fraction of events with six or more jets is larger at 13 TeV than the one at 8 TeV, this results in a lower overall matching efficiency although no further investigations concerning this issue are performed.

Table 7.7: Comparison of improvement in matching efficiency and purity after the BDT selection compared to the values after the preselection at 13 TeV. The values are listed for both the normal BDT selection as well as for the BDT selection with additional window cuts for the full set of variables and for the reduced set. For comparison, the corresponding values of the 8 TeV analysis [1] are also listed.

	selection	ϵ_{sel} [%]	ϵ_{match} [%]	π [%]
8 TeV	preselection	100	71	28
13 TeV	preselection	100	61.42 ± 0.06	25.16 ± 0.02
8 TeV	BDT (≥ -0.05) + window cuts	42.5	82	41
13 TeV	BDT_{all} (≥ -0.02)	42.54 ± 0.03	77.35 ± 0.09	50.34 ± 0.05
	BDT_{red} (≥ -0.017)	42.56 ± 0.03	76.73 ± 0.08	50.00 ± 0.05
	BDT_{all} (≥ -0.265) + win. cuts	42.54 ± 0.03	76.04 ± 0.08	49.33 ± 0.05
	BDT_{red} (≥ -0.276) + win. cuts	42.52 ± 0.03	76.03 ± 0.08	49.32 ± 0.05

Comparing the values corresponding to a similar selection efficiency, the BDT selection at 13 TeV appears to perform even better than the BDT selection at 8 TeV. For a selection efficiency of 42.5 %, the matching efficiency is smaller at 13 TeV but the purities after the different trainings at 13 TeV are with values between 49.3 % and 50.4 % considerably larger than the purity of 41 % at 8 TeV. This means that the unmatched events are selected with a higher efficiency at 13 TeV. It should be noted that the BDT at 8 TeV is trained with the full sample and then applied to the sample after applying the window cuts. Therefore, the comparison with the window cut studies performed here is more appropriate.

A comparison between the performance of the BDT training without applying the window cuts and with doing so shows no significant differences. The values of the training without applying the window cuts are slightly better but this can also be caused by the fact, that the training sample after applying the window cuts is only around half the size of the original sample.

In summary, the window cuts do not further increase the performance in terms of the matching efficiency, the purity or the composition of the sample. However, the performance of the BDT after applying the window cuts is not significantly worse than the performance of the BDT trained without applying the window cuts either. Therefore, applying the window cuts already before the training process is reasonable since this simplifies the analysis of the top mass measurement.

8 First studies with a DNN

In order to test if incorrectly and unmatched events can be removed more efficiently, a more sophisticated machine learning technique, a DNN, is used for the following studies. To allow the DNN to learn from the event information directly, it is fed with the four-momenta of the jets chosen by KLFitter, as well as the lepton and missing transverse momentum. Furthermore, a DNN is trained using the BDT variables to see the effects of using a more complex classification technique without changing the variables. As in the BDT studies, the DNN is trained and evaluated both before and after the window cuts are applied to the sample.

8.1 Variables used for the DNN training and the DNN setup

8.1.1 DNN variables

Using the information of the original four-momenta chosen by KLFitter provides a completely different phase space compared to the phase space of the BDT variables. Some of the variables used in the BDT training, e.g. $\Delta R(q,q)$, are observables that are calculated using the four-momenta. Being a much more complex method, the DNN is expected to be able to recognise dependencies between different quantities like $\Delta R(q,q)$ without providing the calculated value. Additionally, the DNN might be able to to recognise complex patterns between different four vectors that are not included in the variables given to the BDT, and hence could potentially further improve the separation.

The four-momenta information used for the training are the azimuthal angle ϕ, the transverse momentum p_{T}, the pseudorapidity η and the energy E. However, the azimuthal angles of the lepton and the missing transverse momentum are not used. These two variables have detector-related differences in the shape for electrons and muons. When using them in the training, it was either found that the DNN output distributions had several spikes across the output spectrum, or resulted in one single spike and were therefore not successful at all. Removing them from the training has proven to give a stable training and therefore the two variables are not used in the remainder of this thesis. This results in 4 variables for each quark, three variables of the lepton and one of the neutrino[1], which leads to 20 input variables in total. The variables and their separation are listed in Table 8.1 and the distributions of the two variables with the largest separation, the transverse momenta of the two quarks from the hadronically decaying W boson, are shown in Figure 8.1. The distributions of all variables can be seen in the appendix in Figures 10.10-10.16. Compared to the BDT variables, these new variables have very small separation values of less than 3 %. Furthermore, only three variables have a separation larger than 1 %. Note that the y-axis in the plots in Figure 8.1 have a logarithmic scale to show the small differences between the signal and background distributions.

By applying the window cuts, the distributions of the variables change quite strongly and for many variables the separation increases. The two variables with the largest separation

[1] The neutrino is only "measured" in the transverse plane which means that the only two variables of the neutrino four momentum are the azimuthal angle and the missing transverse momentum (the transverse momentum of the neutrino).

are still the transverse momenta of the two quarks originating from the W boson. The separation of the variables after applying the window cuts are also listed in Table 8.1 and the distributions of the variables after applying the window cuts can be seen in the appendix in Figures 10.10-10.16 compared to the corresponding distributions before applying the window cuts.

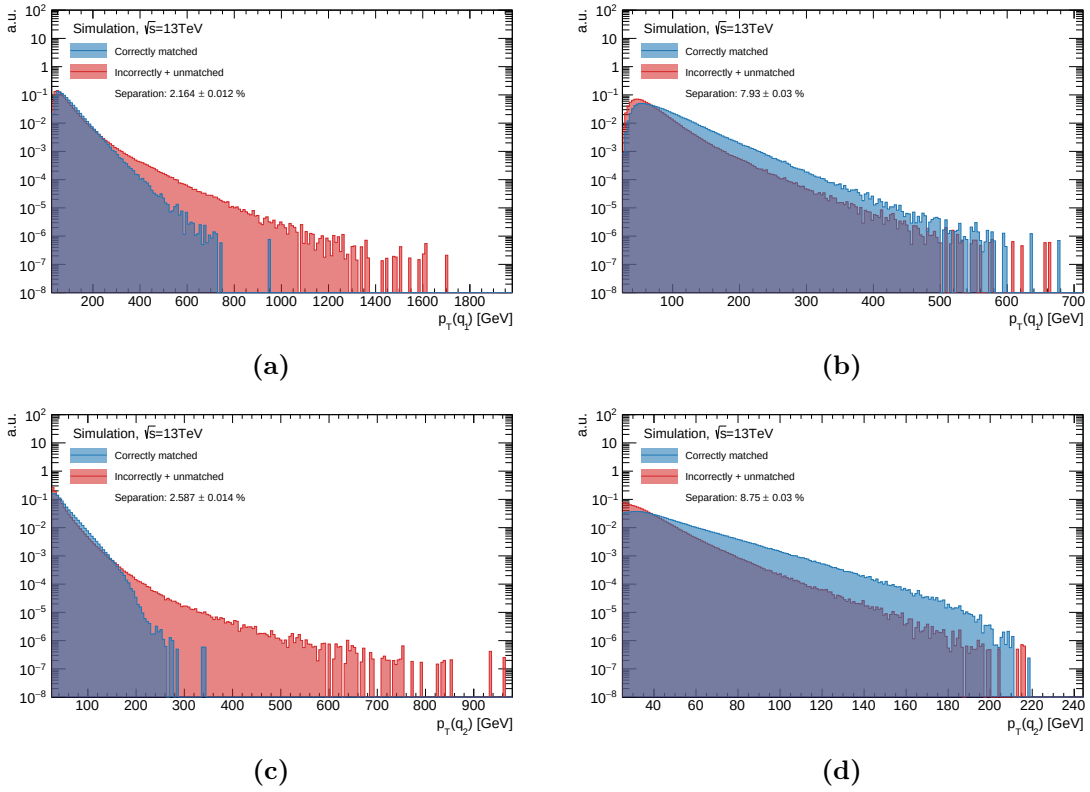


Figure 8.1: The distributions of the two DNN variables with the largest separation, for both after the preselection (left) and after the standard selection, i.e. after the additional window cuts (right): **(a)** and **(b)** show the transverse momentum of the quark q_1 from the hadronically decaying W boson as chosen by KLFitter and **(c)** show the transverse momentum of the quark q_2 from the hadronically decaying W boson as chosen by KLFitter. All plots have a logarithmic scale on the y-axis.

8.1.2 Preprocessing and architecture of the DNN

Before performing the training of the DNN, the sample is preprocessed to ensure that the values of the different variables are of the same order of magnitude. Therefore, the mean and the standard deviation of the distribution of each variable is calculated. Afterwards, the mean value is subtracted and the shifted value is then divided by the standard deviation of the original distribution. The sample is then split into a training and a test sample according to the event number of the MC generated events. Events with an odd event number are used for the test sample and events with an even event number are used for the training sample. The training of the DNN is performed with the Keras framework (release 2.2.4) [44] using the TensorFlow backend [45].

Different architectures of the DNN are tested which means that both the number of hidden layers as well as the number of nodes in these layers are modified. The performance

Table 8.1: List of the four-momenta variables used for the DNN training and their separation. The values are listed for both before and after applying the window cuts.

Variable	Separation in % (without window cuts)	Separation in % (window cuts applied)
$E\,(b_{\text{had}})$	0.115 ± 0.003	0.920 ± 0.011
$\eta\,(b_{\text{had}})$	0.0108 ± 0.0009	0.061 ± 0.003
$\phi\,(b_{\text{had}})$	0.0023 ± 0.0004	0.0049 ± 0.0008
$p_{\text{T}}\,(b_{\text{had}})$	0.164 ± 0.003	1.290 ± 0.012
$E\,(b_{\text{lep}})$	0.100 ± 0.003	1.452 ± 0.013
$\eta\,(b_{\text{lep}})$	0.075 ± 0.002	0.0106 ± 0.0012
$\phi\,(b_{\text{lep}})$	0.0025 ± 0.0004	0.0043 ± 0.0007
$p_{\text{T}}\,(b_{\text{lep}})$	0.262 ± 0.004	2.607 ± 0.018
$E\,(q_1)$	1.111 ± 0.009	3.80 ± 0.02
$\eta\,(q_1)$	0.921 ± 0.008	0.087 ± 0.003
$\phi\,(q_1)$	0.0031 ± 0.0005	0.0060 ± 0.0009
$p_{\text{T}}\,(q_1)$	2.164 ± 0.012	7.93 ± 0.03
$E\,(q_2)$	0.416 ± 0.006	2.705 ± 0.018
$\eta\,(q_2)$	0.970 ± 0.009	0.093 ± 0.003
$\phi\,(q_2)$	0.0033 ± 0.0005	0.0052 ± 0.0008
$p_{\text{T}}\,(q_2)$	2.587 ± 0.014	8.75 ± 0.03
$E\,(\ell)$	0.040 ± 0.003	0.2 ± 0.4
$\eta\,(\ell)$	0.0029 ± 0.0005	0.0061 ± 0.0009
$p_{\text{T}}\,(\ell)$	0.076 ± 0.002	0.382 ± 0.007
E_{T}^{miss}	0.197 ± 0.004	0.148 ± 0.004

of networks with different architectures and thereby different numbers of trainable parameters changes quite strongly. However, only the results of the best performing model are presented here. The best response and separation is achieved with 5 hidden layers consisting of 50, 80, 50, 4 and 2 nodes each.

The activation function of both the input layer and the hidden layers is the Rectified Linear Unit (ReLU) function. The output distribution of the DNN is defined in the range $[0, 1]$ where values close to 0 are background like and values close to 1 are more signal like. Therefore, the activation function of the output layer is chosen to be the sigmoid function. The loss function used in the training is the mean squared error. Each DNN training is performed over 150 epochs via gradient descent and the batch size is chosen to be 300. For all the trainings presented in the following, the loss of the DNN is plotted as a function of the training epoch.

8.2 Performance of the DNN selection without applying the window cuts

The DNN output distribution of the best training is shown in Figure 8.2 and has a separation even larger than 50 %, which is calculated using the output of the test sample. The ratio of the two histograms for the test and training sample, which is also plotted in Figure 8.2, shows very good agreement of the classification of test and training sample. The ratio plots in the DNN studies are zoomed in compared to the ratio plots in the BDT studies to make the small differences between training and test classification visible. The loss of the DNN, which can be seen in the appendix in Figure 10.17, shows no signs of overtraining as well since the difference between the training loss and the test loss is quite small. However, the corresponding P-values are very small with 0.016 for the signal and 0.012 for the background but can be explained by the small statistical uncertainties.

As done in the studies of the BDT selection, the efficiencies and purities as defined in Equations 7.2 - 7.4 are calculated as a function of the DNN cut value. The corresponding plot is shown in Figure 8.3. The matching efficiency and the purity increase and the selection efficiency decreases very fast at low cut values. This is expected, as the majority of the background events are distributed at values close to zero, which can be seen in Figure 8.2. The matching efficiency and the purity of the sample after the DNN selection are compared to the corresponding values of the BDT selection. Table 8.2 shows the matching efficiency and purity after the preselection, after the BDT selection and after the DNN selection. Once again, the comparison is done with the values corresponding to a selection efficiency of approximately 42.5 %.

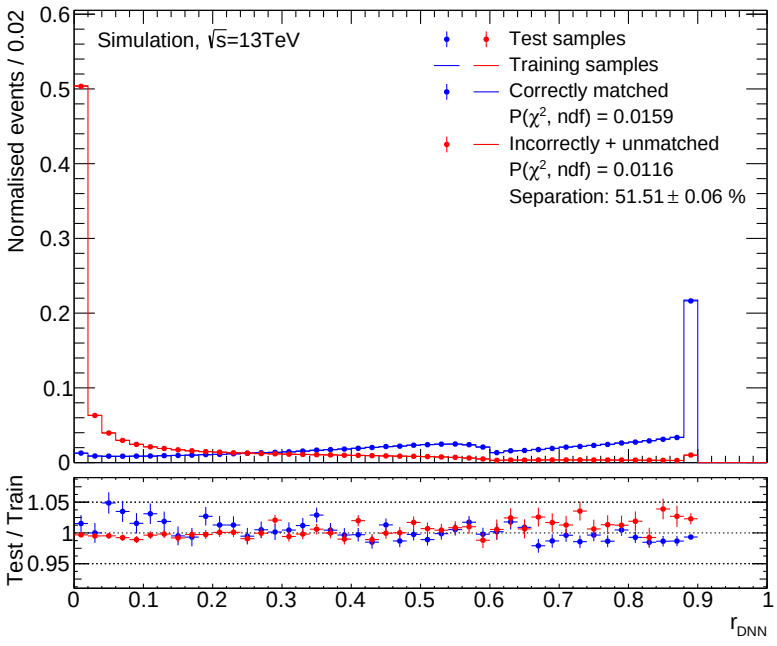


Figure 8.2: Normalised DNN output distributions of the same sample as used for the BDT studies. The solid lines are the output distributions of the training sample and the points are the BDT output distributions of the test sample. The background samples (training and test) are drawn in red and the signal samples are drawn in blue. The statistical uncertainty is shown for test sample but it is too small to be visible. For the ratio plot, statistical uncertainties for both distributions are taken into account.

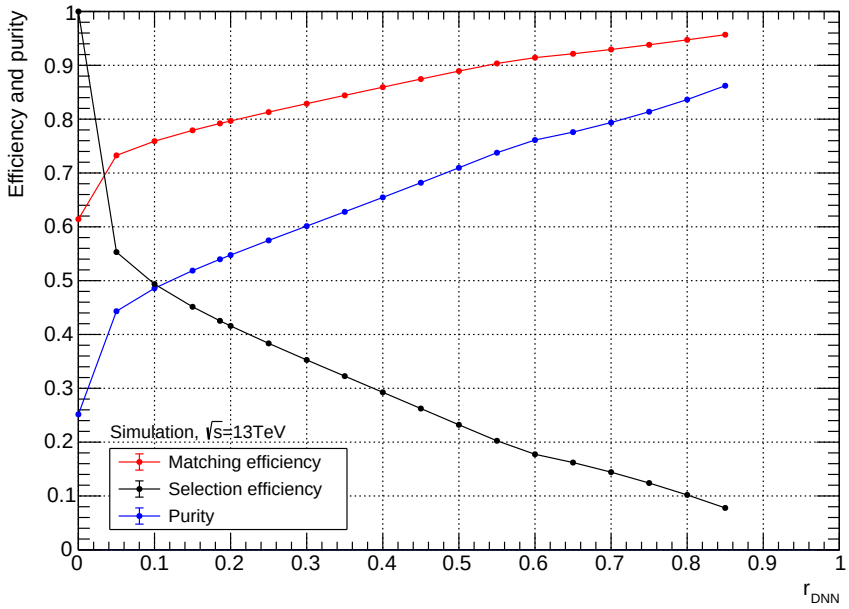


Figure 8.3: Matching efficiency, selection efficiency and purity as a function of the DNN cut value. The DNN training is performed with the information of the four-momenta of both the lepton and the jets chosen by KLFitter. The statistical uncertainties are drawn in the graph but they are too small to be visible.

Table 8.2: Comparison of the matching efficiency and purity for the BDT selection and the DNN selection.

selection	ϵ_{sel} [%]	ϵ_{match} [%]	π [%]
preselection	100	61.42 ± 0.06	25.16 ± 0.02
BDT_{all} (≥ -0.02)	42.54 ± 0.03	77.35 ± 0.09	50.34 ± 0.05
DNN (≥ 0.186)	42.53 ± 0.03	79.20 ± 0.09	53.97 ± 0.05

For a slightly higher selection efficiency, both the matching efficiency and the purity of the sample after the DNN selection are larger than the corresponding values of the sample after the BDT selection. This shows that the DNN selection performs better than the BDT selection and seems to be able to recognise characteristic patterns in the four-momenta variables. However, it should be mentioned that the training of the DNN takes, depending on the number of hidden layers and their size, several hours while the BDT training takes less than one hour[2].

The ROC curve of the DNN is shown in Figure 8.4 together with the ROC curves of the BDT training. Having an AUC of 0.905, the ROC curve of the DNN is significantly higher than the ROC curves of the BDT selection which have an AUC of 0.870 and 0.867. Furthermore, the difference in separation between the BDT output and the DNN output is with approximately 10 % quite large, which also shows that the DNN performs better than the BDT.

[2]Both trainings are performed using a CPU. Using a GPU for the DNN training is expected to be much faster.

The new approach with a DNN instead of a BDT uses both a more complex multivariate classification technique as well as different variables than the BDT. To investigate if the increase in performance is solely caused by the multivariate classification technique, the DNN is also trained and evaluated with the variables used in the BDT training. However, the architecture used for the DNN training with the BDT variables is not further optimised. This means that the same DNN architecture as used for the training with the four-momenta is chosen for the training with the BDT variables. It should be noted, that for a more precise comparison further optimisations would be necessary as the phase space provided by the BDT variables differs from the phase space provided by the four-momenta variables. Thereby, a different DNN architecture might perform better. The loss of the DNN is shown in Figure 10.18 as a function of the training epoch and the resulting DNN output of the training with the BDT variables is shown in Figure 8.5. The separation of the output distribution is 44.6 % and the outputs of the test and the train sample show very good agreement. When comparing the result of this DNN training with the output of the BDT training shown in Figure 7.4, an increase of 3 % is observed for the DNN distribution. The ROC curve of the DNN with the BDT variables is also shown in Figure 8.4. The comparison of the ROC curves also shows, that the performance of the DNN which uses the BDT variables performs slightly better than the initial BDT selection. This indicates that not only the more complex classification technique but also the usage of the four-momenta information instead of the BDT variables causes the increase in the performance.

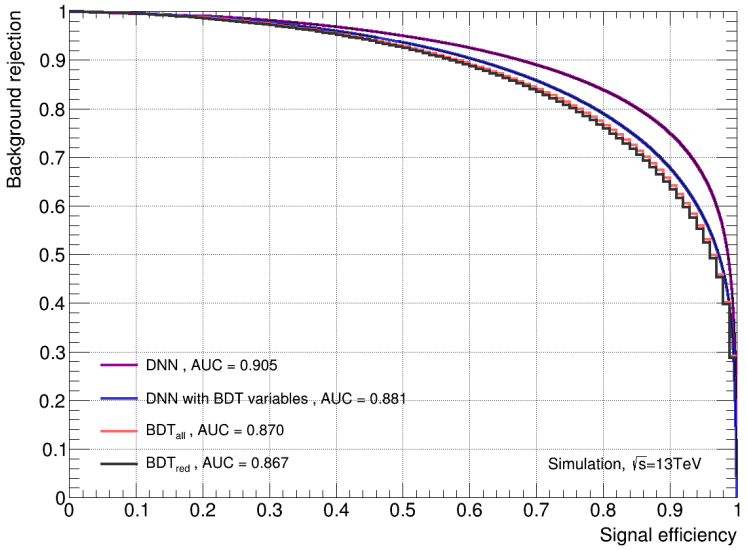


Figure 8.4: ROC curves of both the BDT trainings and the DNN training without applying the window cuts. The purple curve is the ROC curve of the DNN while the grey and pink histograms are the ROC curves of the BDT with both all variables (pink) and the reduced set of variables (grey). The ROC curve of the DNN with the BDT variables is drawn in blue.

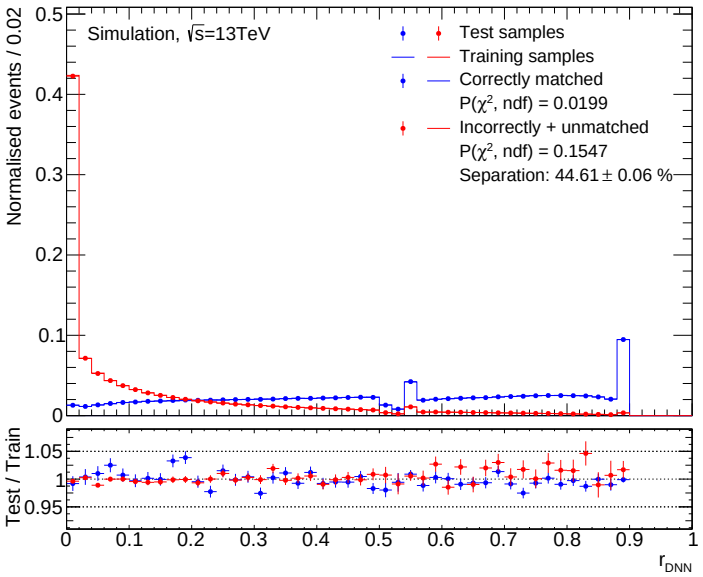


Figure 8.5: Normalised DNN output distributions when trained with the same variables as used in the BDT studies. Incorrectly and unmatched events are combined to one background. The solid lines are the output distributions of the training sample and the points are the BDT output distributions of the test sample. The background samples (training and test) are drawn in red and the signal samples are drawn in blue. The statistical uncertainty is shown for test sample but it is too small to be visibile. For the ratio plot, statistical uncertainties for both distributions are taken into account.

8.3 DNN training with additional window cuts

Since the final analysis at 13 TeV will be performed with the window cuts from section 7.3, the DNN is also trained with the sample after applying these window cuts. The architecture of the DNN is the same as the one in the studies without applying the window cuts and the preprocessing of the sample is carried out in the same way. The resulting DNN output can be seen in Figure 8.6 and the ROC curve is shown in Figure 8.7 together with the two ROC curves from the BDT training with the window cuts.

Even though the separation of most variables did increase by applying the window cuts as explained in subsection 8.1.1, the DNN does not separate the signal and the background better than in the training without applying the window cuts. Instead of showing a very clear peak of the background peak close to a DNN output of zero, the background distribution does only slowly decrease with the DNN r_{DNN}. Furthermore, both the background and the signal distribution show a peak at approximately 0.6. This indicates, that there are events in both the signal and the background sample which look very similar to the DNN leading to these small peaks at 0.6.
The separation of the DNN output is approximately 29 % and the AUC of the corresponding ROC curve is 0.809. These values are much smaller than the corresponding values of the training without applying the window cuts. However, they can not be compared directly since they are evaluated on different samples. Additionally, as explained in the BDT studies, applying the window cuts removes already a large number of background events. Therefore, the DNN trained after applying the window cuts can have a smaller separation and still perform equally well as the DNN trained without window cuts.

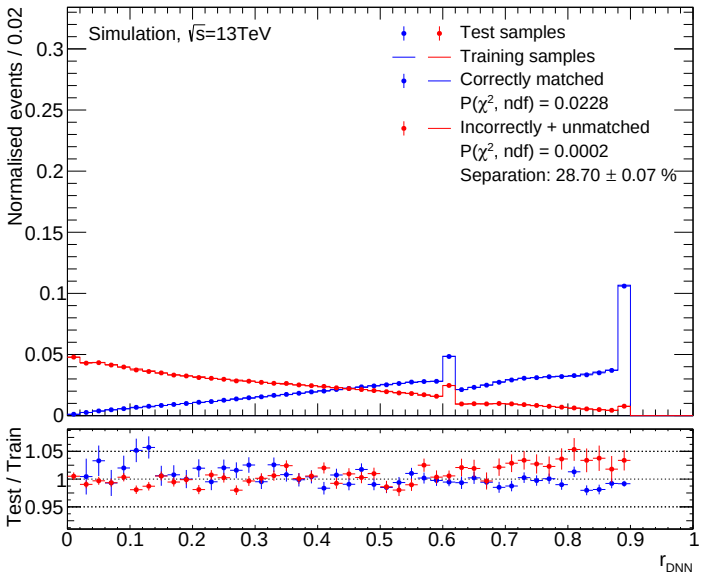


Figure 8.6: Normalised DNN output distributions when trained with the four momenta variables after applying the window cuts. Incorrectly and unmatched events are combined to one background. The solid lines are the output distributions of the training sample and the points are the BDT output distributions of the test sample. The background samples (training and test) are drawn in red and the signal samples are drawn in blue. The statistical uncertainty is shown for test sample but it is too small to be visibile. For the ratio plot, statistical uncertainties for both distributions are taken into account.

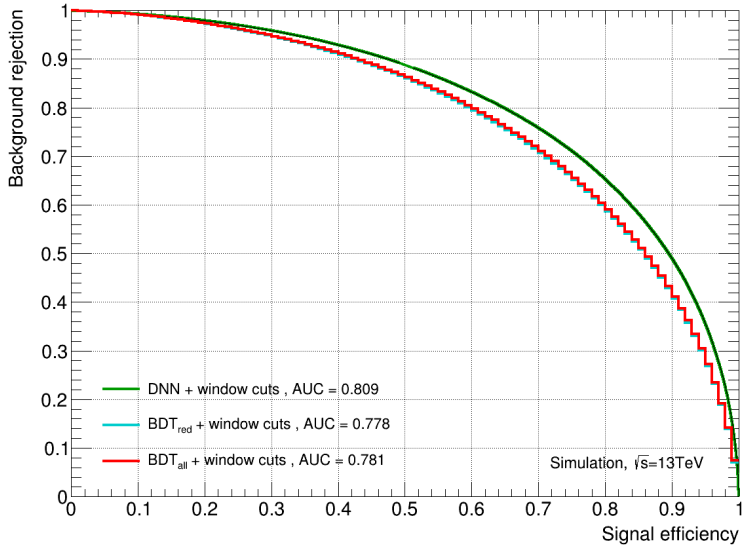


Figure 8.7: ROC curves of both the BDT trainings and the DNN training after applying the window cuts. The green curve is the ROC curve of the DNN while the cyan and red histograms are the ROC curves of the BDT with both all variables (red) and the reduced set of variables (cyan). For the ratio plot, statistical uncertainties for both distributions are taken into account.

Calculating ϵ_{match} and π at a selection efficiency of around 42.5 % leads to values of around 75 % for the matching efficiency and 49 % for the purity. The exact values are compared to the corresponding values of the BDT in Table 8.3. The efficiencies are again plotted as a function of the cut value, which can be seen in the appendix in Figure 10.20. Even though the ROC curve of the DNN has a significantly larger AUC, the performance of the DNN at a selection efficiency of approximately 42.5 % is not significantly better than the performance of the BDT. However, the selection efficiency of 42.5 % is only chosen for all the comparisons so far as it is the final selection efficiency in the 8 TeV analysis. For a comparison concerning only the performance at 13 TeV, a different selection efficiency can be chosen.
Since both the cross section of the $t\bar{t}$ production as well as the luminosity of the LHC are higher at 13 TeV, there are much more events in the 13 TeV data compared to the 8 TeV data. Therefore, lower selection efficiencies and thus higher cut values can be applied at 13 TeV while still having a smaller statistical uncertainty than in the 8 TeV measurement. Comparing the BDT and the DNN with window cuts at a selection efficiency of approximately 22.5 %, the better performance of the DNN becomes visible with an absolute increase of 2.5 % in the purity compared to the BDT selection. The exact values can also be found in Table 8.3.

Table 8.3: Comparison of the matching efficiency and purity by using the BDT selection compared to the DNN selection.

selection	ϵ_{sel} [%]	ϵ_{match} [%]	π [%]
preselection	100	61.42 ± 0.06	25.16 ± 0.02
BDT_{all} (≥ -0.02)	42.54 ± 0.03	77.35 ± 0.09	50.34 ± 0.05
DNN (≥ 0.186)	42.53 ± 0.03	79.20 ± 0.09	53.97 ± 0.05
BDT_{all} (≥ -0.265) + window cuts	42.54 ± 0.03	76.04 ± 0.08	49.33 ± 0.05
DNN (≥ 0.076) + window cuts	42.55 ± 0.03	75.78 ± 0.08	49.68 ± 0.05
BDT_{all} (≥ 0) + window cuts	22.53 ± 0.02	88.87 ± 0.12	66.97 ± 0.09
DNN (≥ 0.456) + window cuts	22.54 ± 0.02	88.45 ± 0.12	69.46 ± 0.09

Even though the DNN performs better than the BDT, the increase in the performance compared to the BDT selection is not as large as in the comparison without the window cuts (section 8.2). However, this might be further improved by optimizing the DNN architecture according to the new phase space after the window cuts. As already mentioned in the previous section, using the same DNN architecture for another phase space can lead to a worse result, as the optimised architecture of one phase space might not be ideal for the training and classification of events from another phase space.

To see the effect of both the BDT and the DNN selection on the $m_{\text{top}}^{\text{reco}}$ distribution, the resulting $m_{\text{top}}^{\text{reco}}$ distribution is shown for both the standard selection, the BDT selection and the DNN selection in Figure 8.8.

The distributions for both the BDT and DNN selection are the resulting distributions when choosing the cut values corresponding to a selection efficiency of 22.5 % as listed in Table 8.3. The distributions are normalised to show the difference between their shapes.

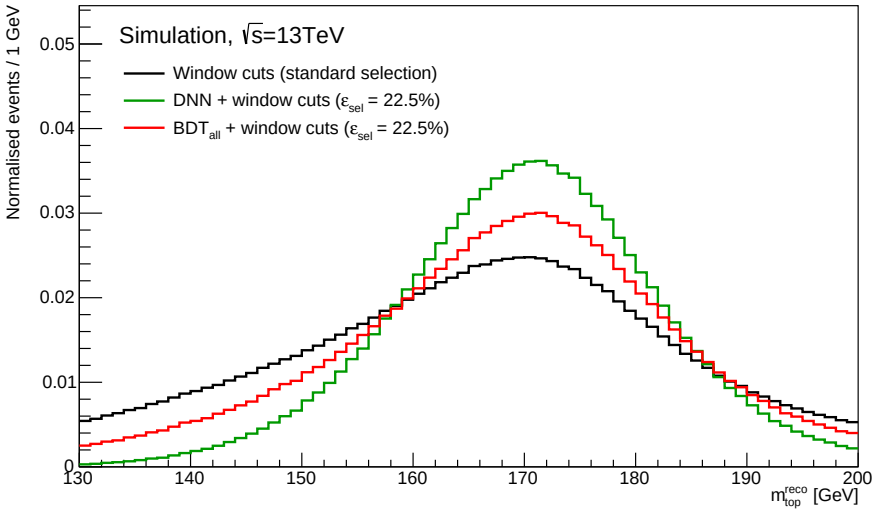


Figure 8.8: Comparison of the normalised top-mass distribution after the standard selection, the BDT selection and the DNN selection. The black curve shows the distribution as obtained after applying the window cuts (standard selection) to the preselection sample. The other two histograms show the distributions of $m_{\text{top}}^{\text{reco}}$ after applying both the window cuts and the BDT selection (red) or the DNN selection (green). The selection efficiency with respect to the preselection is 47 % for the standard selection and 22.5 % for both the BDT and the DNN selection.

However, since the distributions of the BDT selection and the DNN selection correspond to approximately the same selection efficiency, these two distributions can also be compared in terms of the absolute values of the bins as the total integral of the two original histograms is approximately the same. As can be seen in the plot, the BDT selection does already remove many events with very low values of $m_{\text{top}}^{\text{reco}}$. However, the resulting distribution when applying the DNN selection is both more narrow and especially more symmetrical. The latter fact is very useful for the parametrisation of the template function since the modelling of the unsymmetrical distribution as obtained after the standard selection is quite challenging.

In summary, it can be said that the new approach with the DNN and the new variables leads to a better performance of the selection. Further improvements of the DNN selection could be achieved by further adjusting the parameters of the DNN setup to obtain an even better separation of signal and background.

9 Conclusion

The measurement of the top-quark mass is an essential precision test of the consistency of the standard model as well as for investigations on the fundamental interactions at the scale of electroweak symmetry breaking.

Top quarks at the LHC are predominantly produced via top-quark pair production allowing the top-quark mass m_{top} to be measured with a very high statistical precision. Having an extremely short lifetime smaller than the time scale at which hadronisation processes take place, it can be measured as a "free" quark whose properties can be reconstructed using the information of its decay products.

The measurement of m_{top} is limited by systematic uncertainties. A large challenge in the m_{top} measurement in the lepton+jets channel is the reduction of the combinatorial background. Due to the large number of possible assignments of the measured jets to the corresponding particles at parton level, kinematic likelihood methods and machine learning techniques are used to find the correct assignment between these jets and the final state particles of the $t\bar{t}$ decay.

In the ATLAS m_{top} measurement at a centre-of-mass energy of 8 TeV [1], events first had to pass a basic preselection. Afterwards, the KLFitter framework was used to provide a full kinematic reconstruction of the top-quark decay. Then, to reduce the combinatorial background and simultaneously also some systematic uncertainties of the measurement, a boosted decision tree was employed in order to increase the fraction of correctly reconstructed events. This was achieved by training the BDT with MC simulated events and then applying it to the data. The reconstructed top-quark mass $m_{\text{top}}^{\text{reco}}$ was then simultaneously fitted with the two observables $m_{\text{W}}^{\text{reco}}$ and R_{bq}^{reco}.
This final fit was performed using template functions of the corresponding distributions, which were obtained from the MC simulated events. However, only a finite range of values of the three fitted observables could be considered for the fit. Therefore, additional cuts were applied to the sample, which is called standard selection. Due to the limited size of the available MC sample, the BDT used in the 8 TeV analysis was trained without applying the additional cuts of the standard selection.

The focus of this thesis was the investigation of the event reconstruction and selection for the ATLAS m_{top} measurement at a centre-of-mass energy of 13 TeV. In this context, two sets of studies were done for the selection optimisation: one based on a BDT starting from the one used at 8 TeV and the second based on a DNN, using a different set of variables. The MC sample available for the 13 TeV analysis is much larger than the one at 8 TeV. Hence the smaller sample after applying the additional cuts of the standard selection might still have large enough statistics to be used for the training of the multivariate classifier. Therefore, studies on the performance of the BDT and the DNN were done for both the sample after the preselection and the sample after the standard selection.

The BDT studies showed that for same selection efficiency the BDT selection at 13 TeV leads to a slightly smaller matching efficiency while resulting in a larger purity than the BDT selection at 8 TeV [1]. Furthermore, investigations on the BDT variables indi-

cate that variables with a very small separation can be removed from the set of training variables without a significant decrease in the BDT performance. Moreover, the comparison of the BDT performance using the sample after the standard selection leads to the same performance as the training for events passing the preselection. This means that the BDT selection could be simplified by both using only the events passing the standard selection for the BDT training and by removing variables with very low separation power.

In contrast to the boosted decision tree, the DNN was fed only with the information of the four-momenta chosen by KLFitter. These variables do provide more fundamental information about the event than the variables used for the BDT training. Being a more sophisticated classifier than a BDT, the DNN was expected to recognise relations between the four-momenta that are not included in the BDT variables. Performance studies show that for the same selection efficiency, the DNN selection leads to an even better purity than the BDT selection. In addition, utilising the DNN results in a more symmetrical distribution of the reconstructed top mass $m_{\text{top}}^{\text{reco}}$ making it easier to parametrise the corresponding template function for the fit. The performance of the DNN selection might be further improved by optimising both the architecture of the DNN as well as other hyperparameters to the given phase space.

To further investigate the performance of both the BDT and the DNN, the resulting distributions of the other two observables that are used for the fit, $m_{\text{W}}^{\text{reco}}$ and R_{bq}^{reco} should be analysed as well. In the next step of the m_{top} measurement, the systematic uncertainties will be evaluated for different cut values on r_{DNN} and r_{BDT} and compared to the corresponding result for the standard selection. Only then, the optimal classifier and cut value can be chosen. However, this task is not within the scope of this thesis.

10 Additional material

10.1 BDT studies

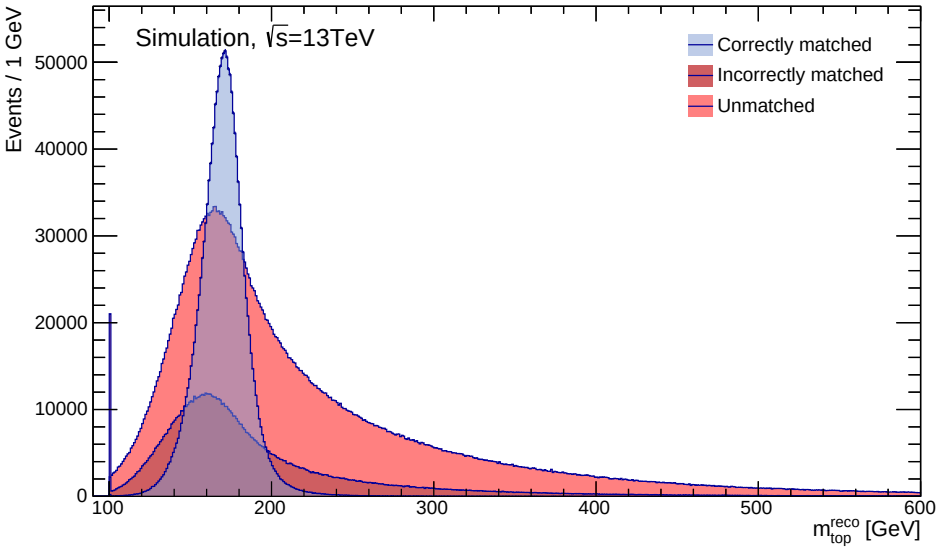


Figure 10.1: Distribution of the reconstructed top mass for the three classes of events. The correctly matched events are distributed in a narrow, symmetric peak with its center close to the input top mass while the incorrectly and unmatched events have a very broad distribution with large tails to higher values of m_{top}. The input value of the MC simulation is $m_{\text{top}} = 172.5\,\text{GeV}$. The distributions do all peak close to a value of $m_{\text{top}}^{\text{reco}} = 100\,\text{GeV}$, which seems to be a lower limit for the reconstructed top mass. However, no further investigations concerning this peak were carried out.

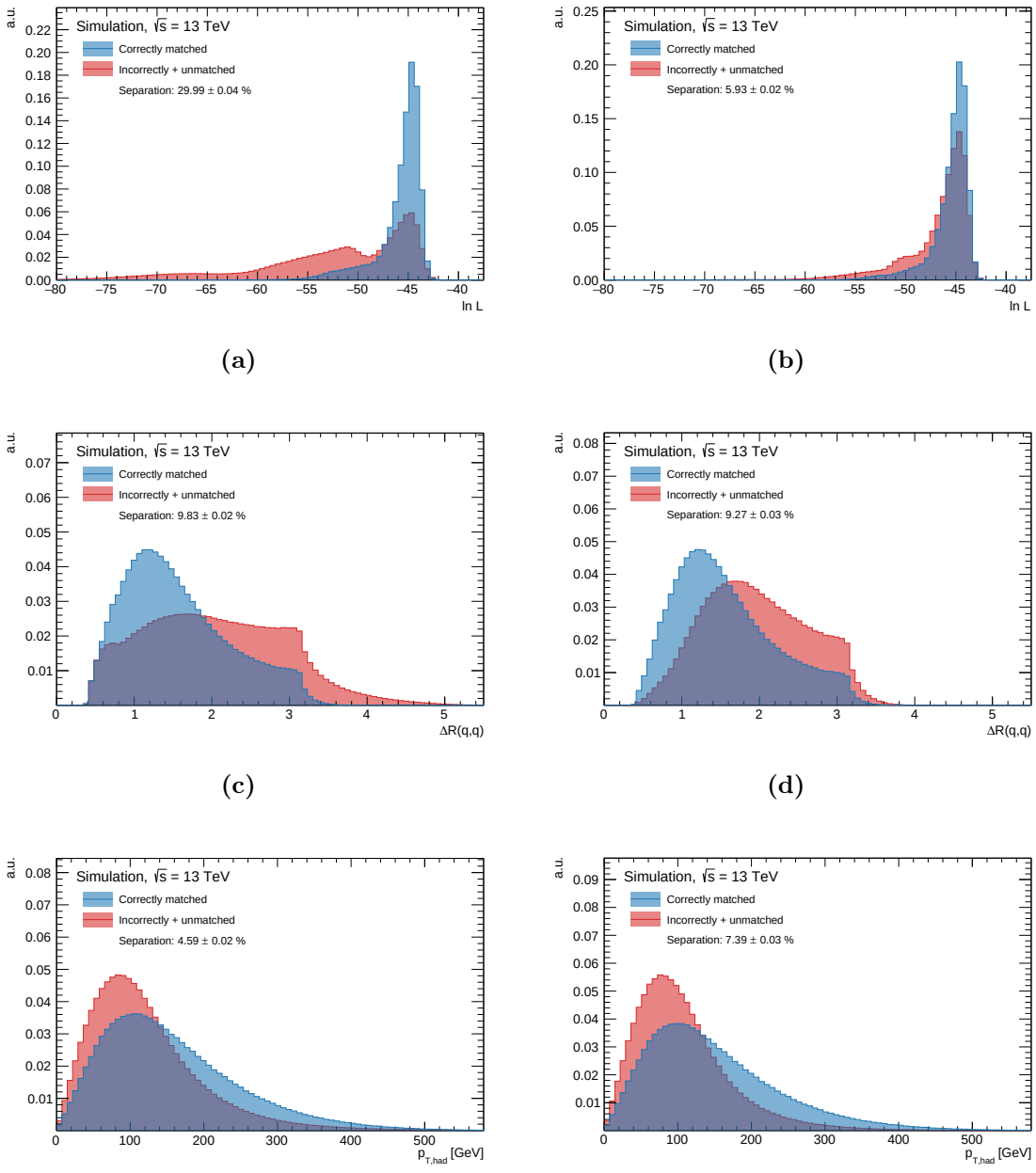


Figure 10.2: Comparison of the normalized distributions of the BDT variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the logarithm of the event likelihood, **(c)** and **(d)** show the angular distance between the two light jets and **(e)** and **(f)** show the transverse momentum of the hadronically decaying top quark.

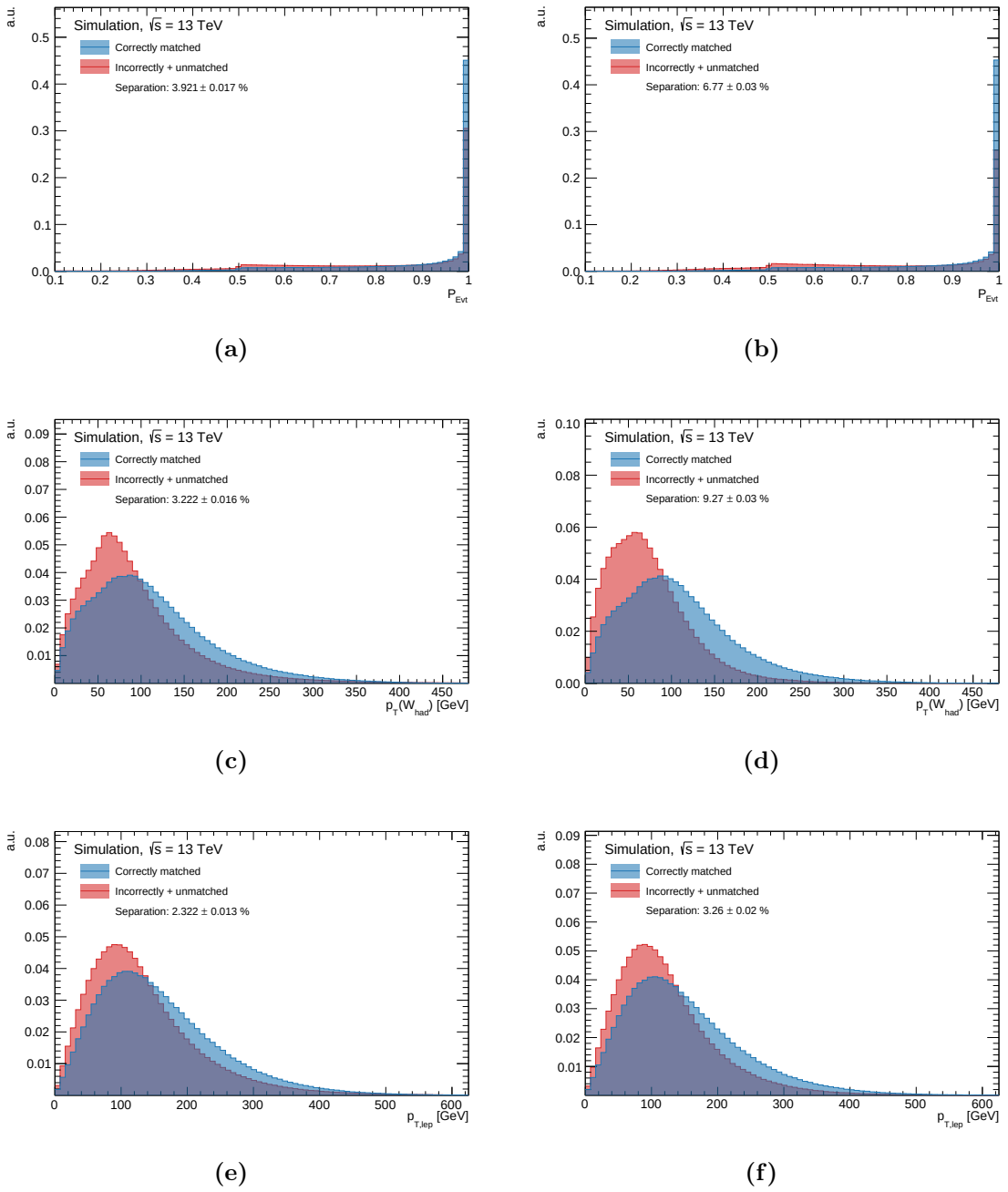


Figure 10.3: Comparison of the normalized distributions of the BDT variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the relative event probability of the best permutation, **(c)** and **(d)** show the transverse momentum of the hadronically decaying W boson and **(e)** and **(f)** show the transverse momentum of the leptonically decaying top quark.

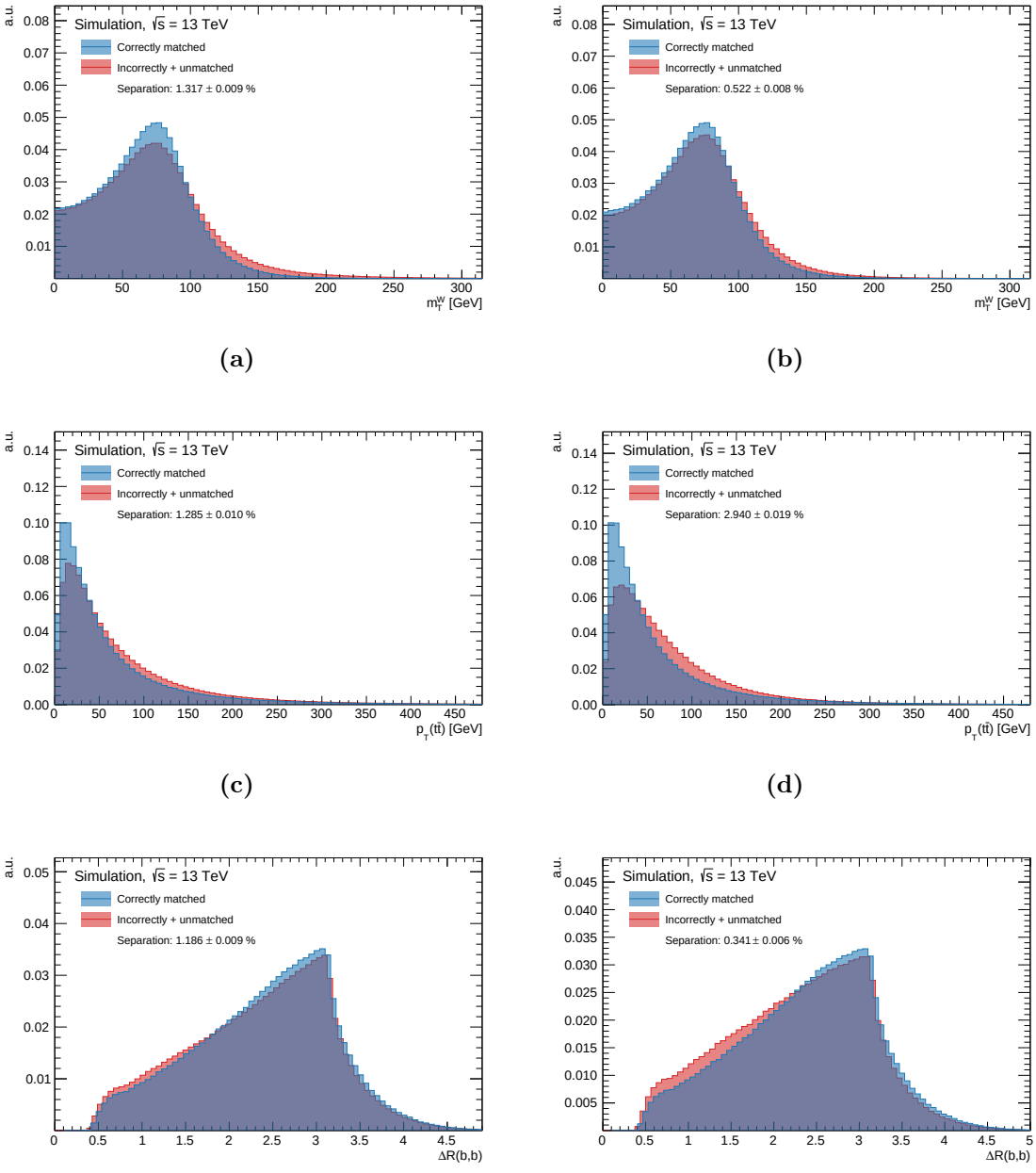
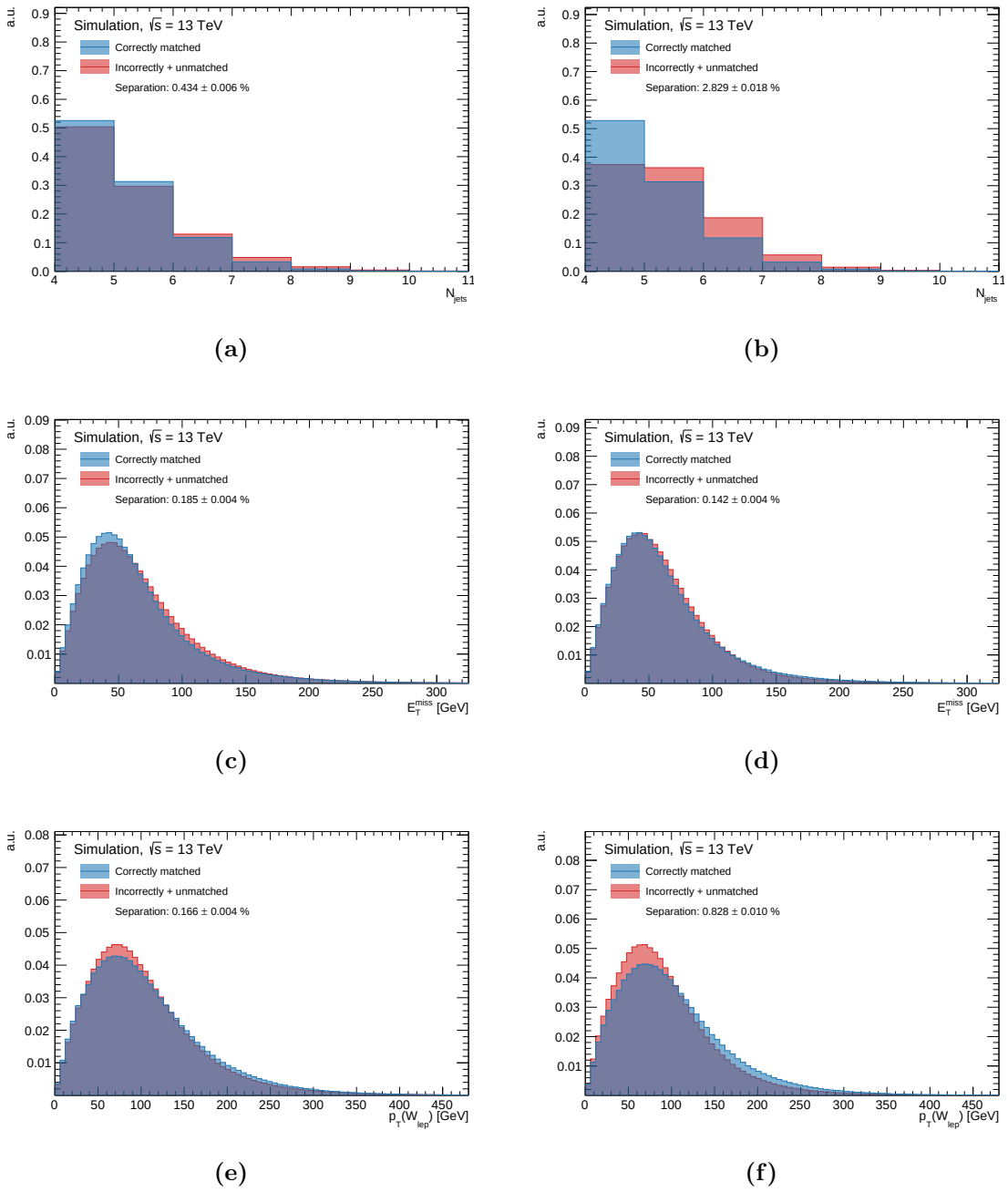


Figure 10.4: Comparison of the normalized distributions of the BDT variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the transverse mass of the leptonically decaying W boson, **(c)** and **(d)** show the transverse momentum of the reconstructed $t\bar{t}$ system and **(e)** and **(f)** show the angular distance between the two reconstructed b-tagged jets.



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 10.5: Comparison of the normalized distributions of the BDT variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the number of jets in the event, **(c)** and **(d)** show the missing transverse momentum and **(e)** and **(f)** show the transverse momentum of the leptonically decaying W boson.

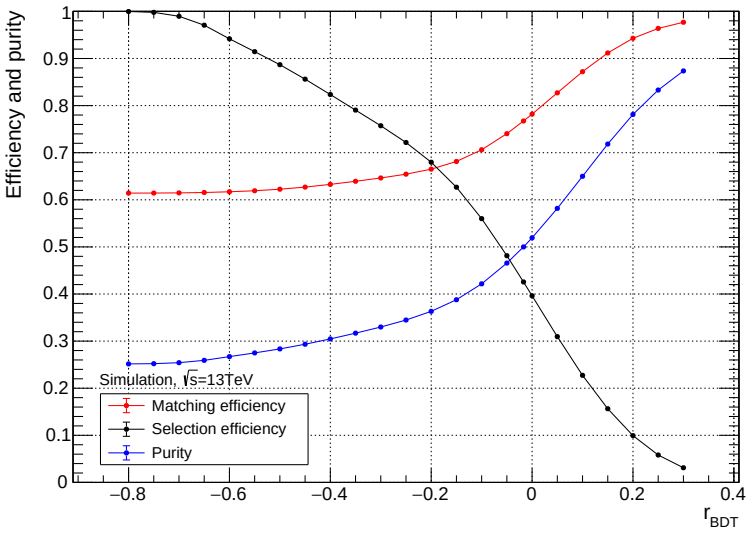


Figure 10.6: Matching efficiency, selection efficiency and purity as a function of the BDT cut value. The BDT training is performed with the reduced set of variables. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

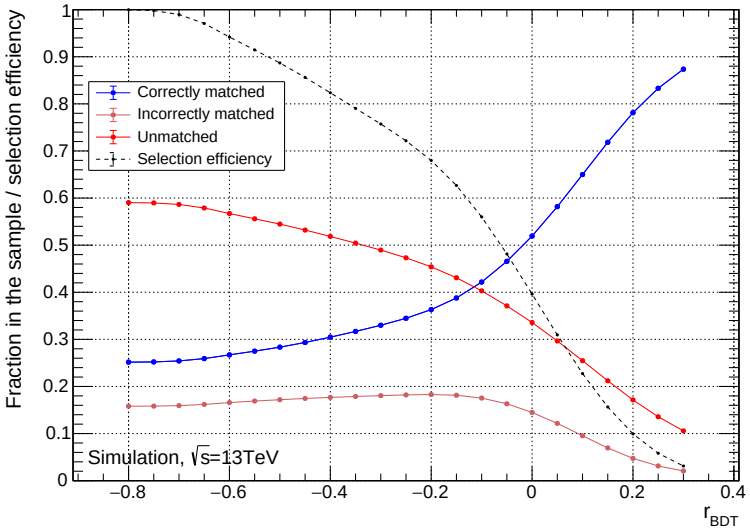


Figure 10.7: Fraction of correctly (blue), incorrectly (dark red) and unmatched (light red) events in the sample after the BDT selection as a function of the BDT cut value. The training is performed with the reduced set of variables. The fraction of correctly matched events in the sample is equal to the purity of the sample. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

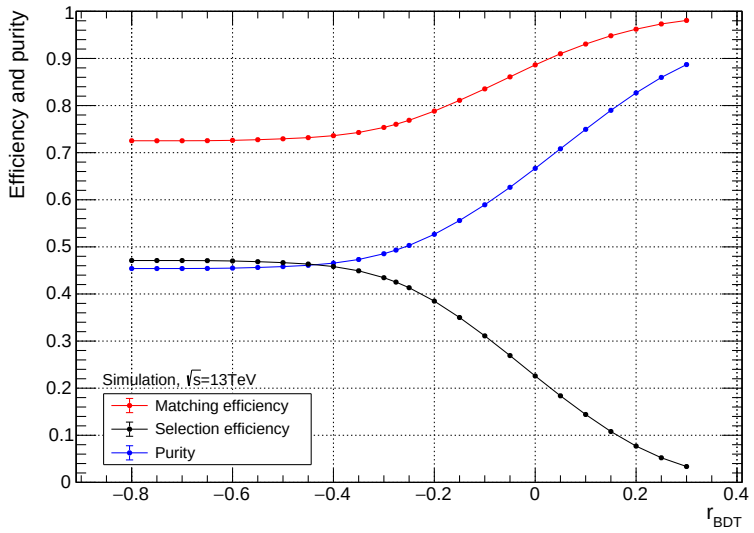


Figure 10.8: Matching efficiency, selection efficiency and purity as a function of the BDT cut value. The BDT training is performed with the reduced set of variables and the additional window cuts are applied. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

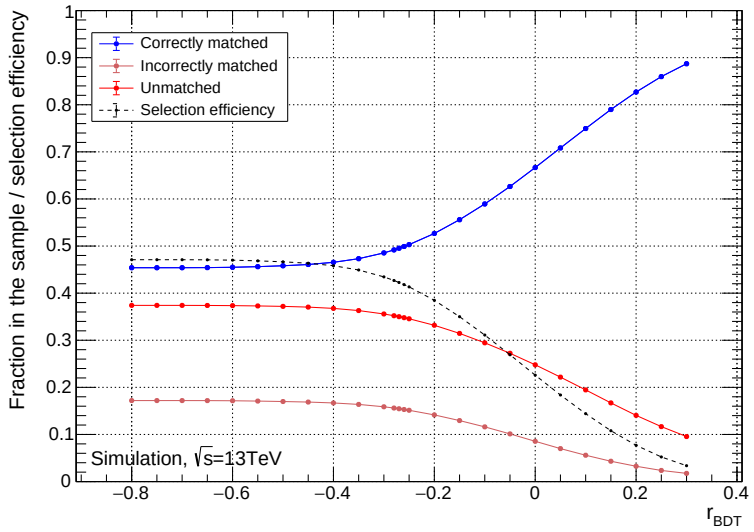


Figure 10.9: Fraction of correctly (blue), incorrectly (dark red) and unmatched (light red) events in the sample after the BDT selection as a function of the BDT cut value. The training is performed with the reduced set of variables and the additional window cuts are applied. The fraction of correctly matched events in the sample is equal to the purity of the sample. The statistical uncertainties are calculated and drawn in the graph, though they are too small to be visible.

10.2 DNN studies

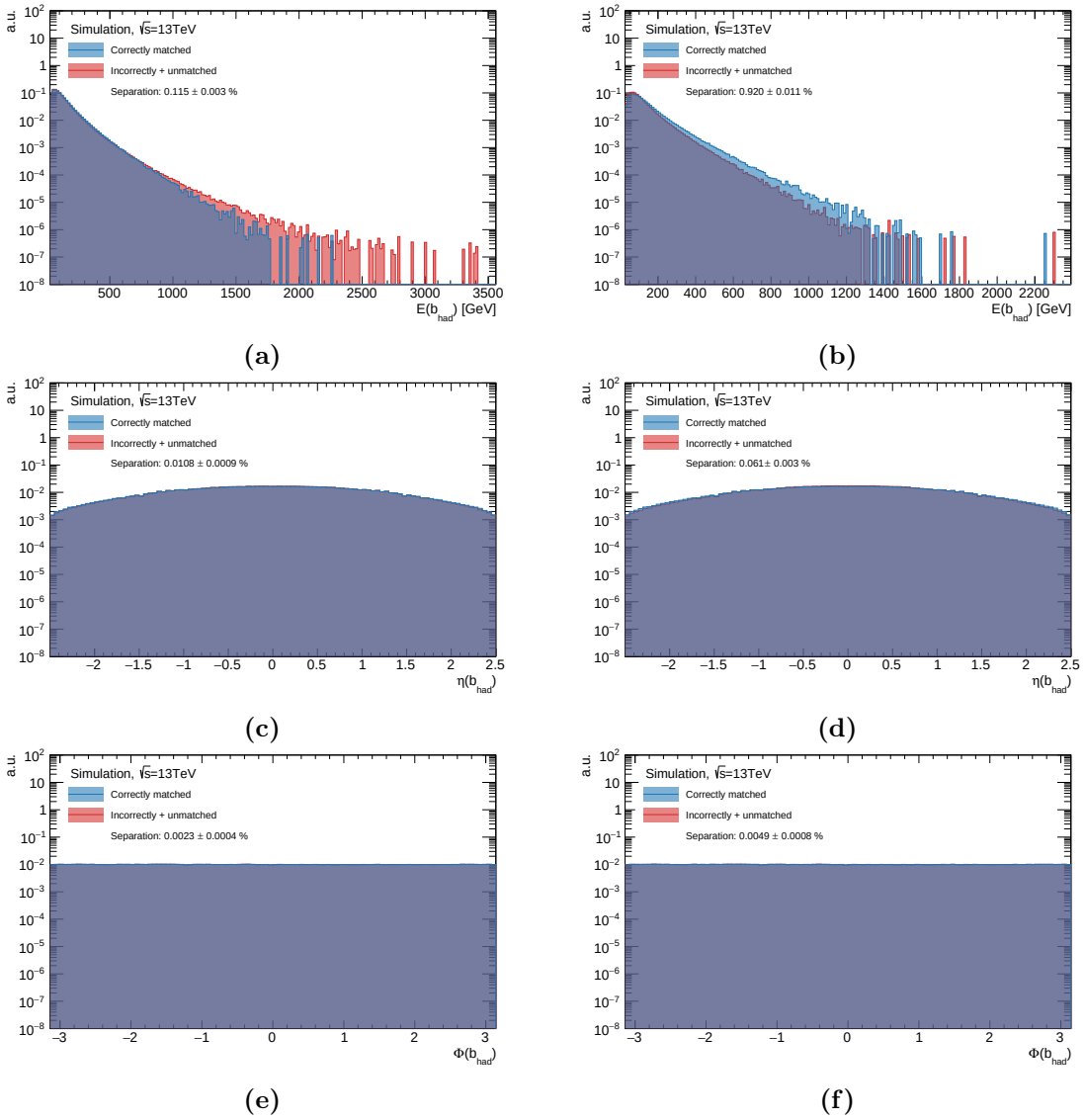


Figure 10.10: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the energy of b_{had}, **(c)** and **(d)** show the pseudorapidity of b_{had} and **(e)** and **(f)** show the azimuthal angle of b_{had}. All plots have a logarithmic scale on the y-axis.

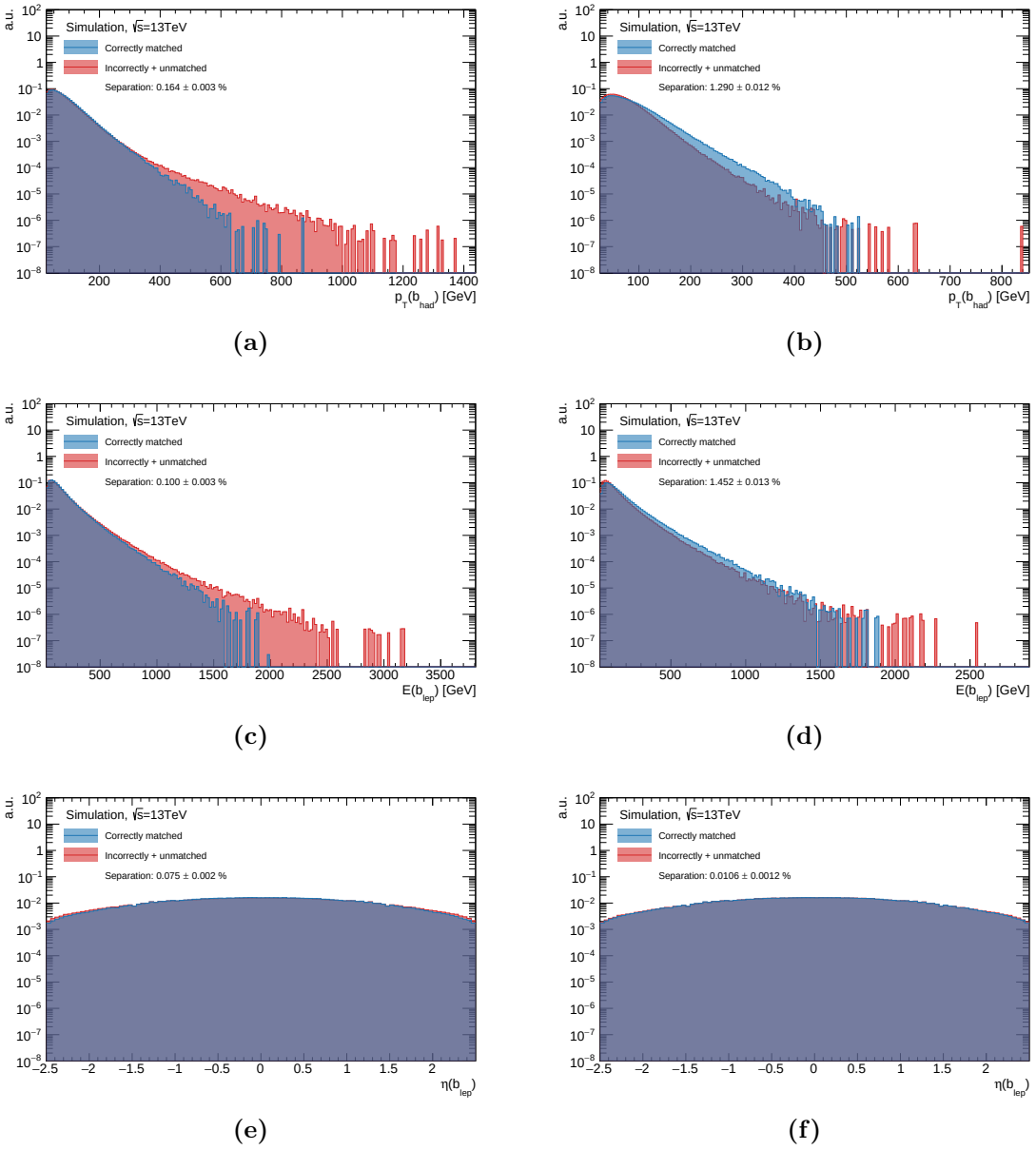


Figure 10.11: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the transverse momentum of b_{had}, **(c)** and **(d)** show the energy of b_{lep} and **(e)** and **(f)** show the pseudorapidity of b_{lep}. All plots have a logarithmic scale on the y-axis.

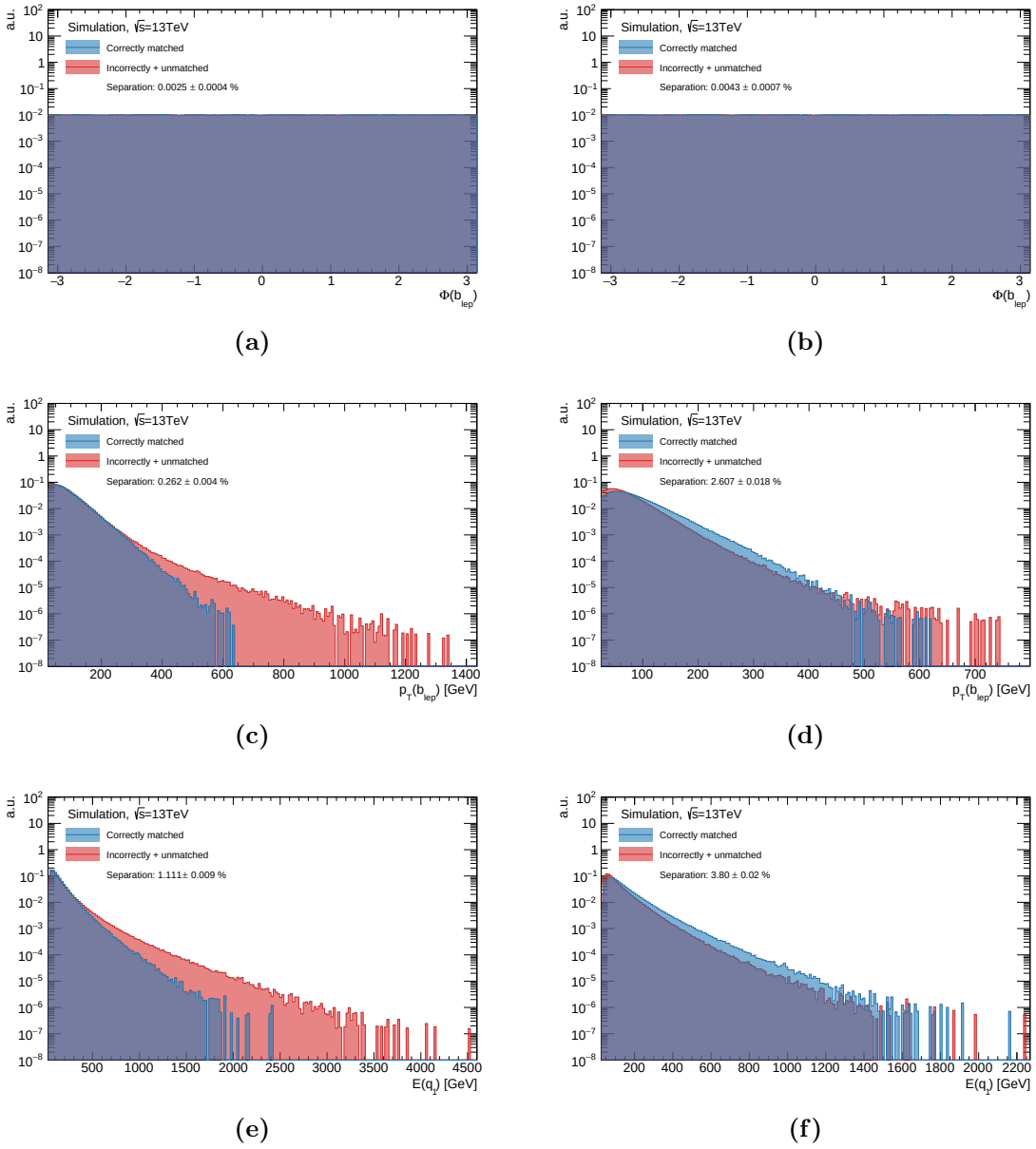


Figure 10.12: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the azimuthal angle of b_{lep}, **(c)** and **(d)** show the transverse momentum of b_{lep} and **(e)** and **(f)** show the energy of q_1. All plots have a logarithmic scale on the y-axis.

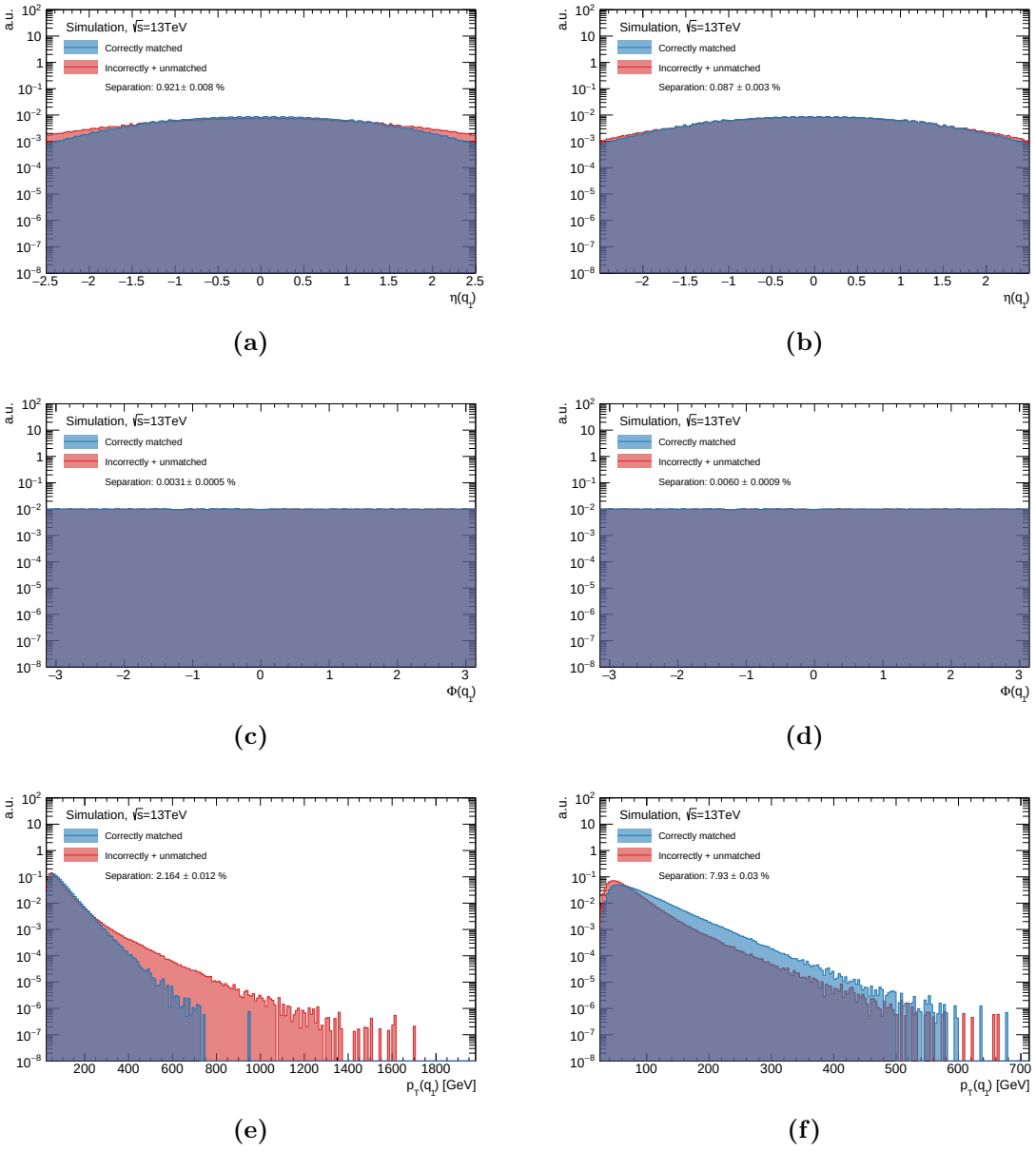


Figure 10.13: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the pseudorapidity of q_1, **(c)** and **(d)** show the azimuthal angle of q_1 and **(e)** and **(f)** show the transverse momentum of q_1. All plots have a logarithmic scale on the y-axis.

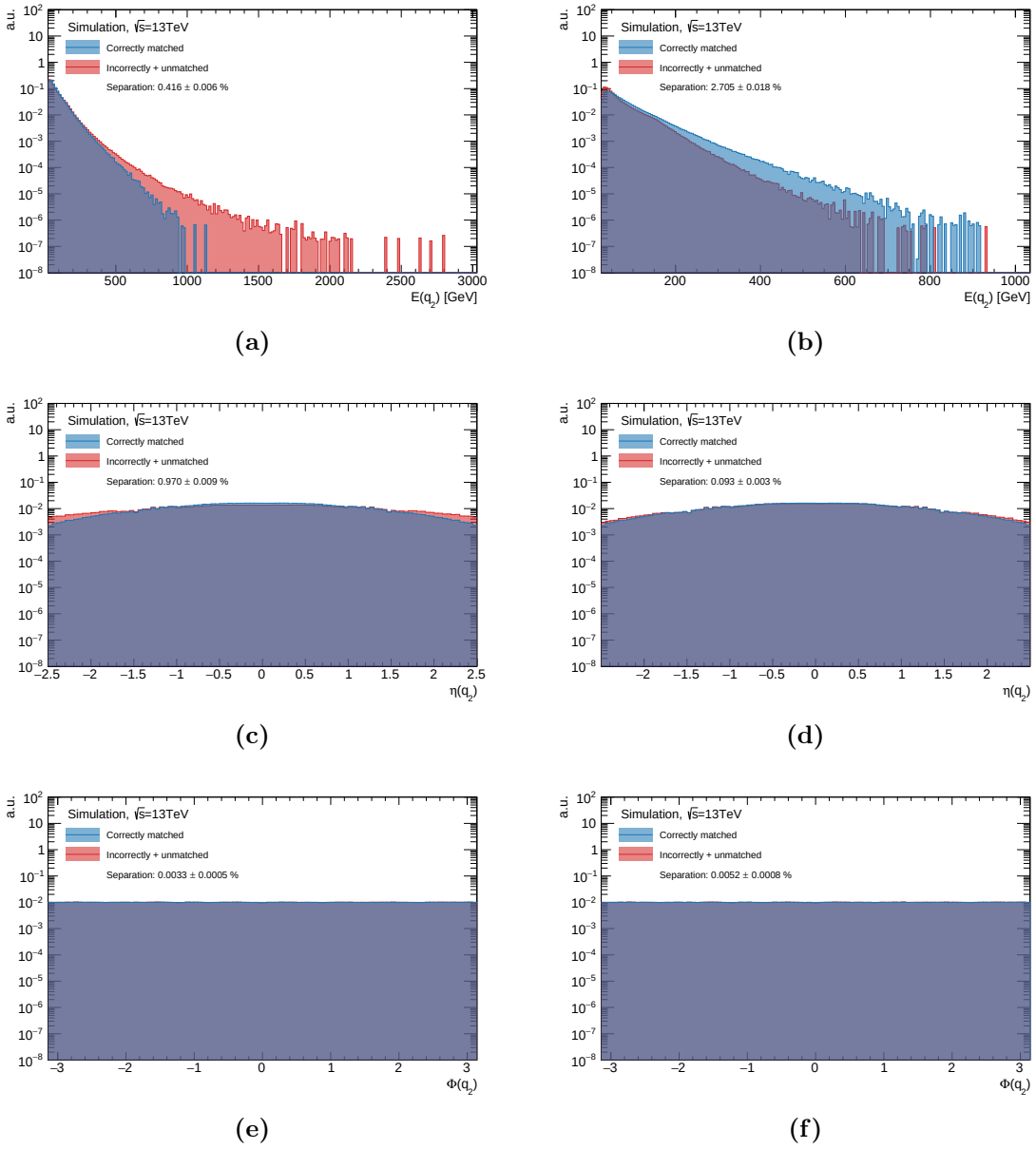


Figure 10.14: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the energy of q_2, **(c)** and **(d)** show the pseudorapidity of q_2 and **(e)** and **(f)** show the azimuthal angle of q_2. All plots have a logarithmic scale on the y-axis.

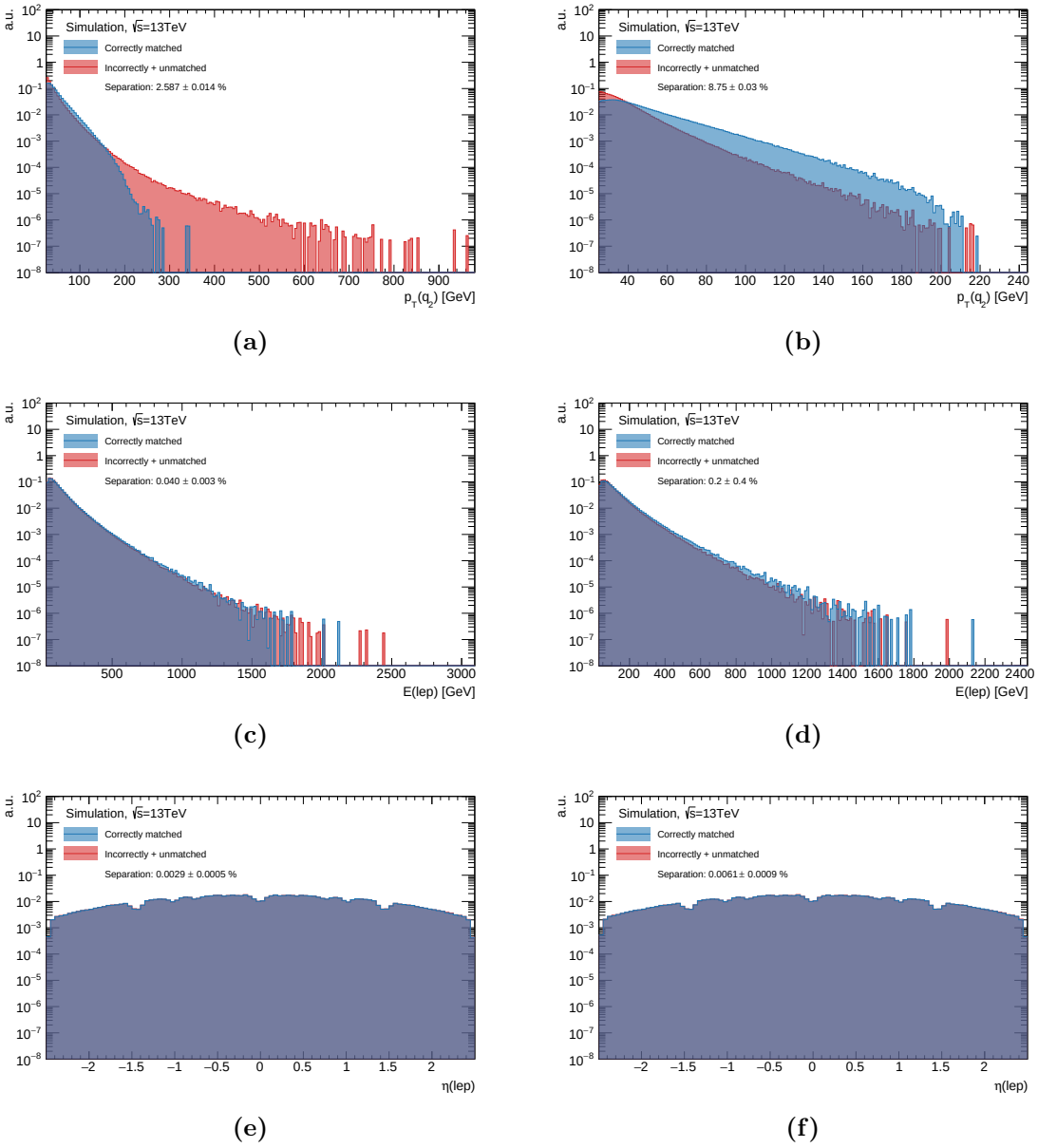


Figure 10.15: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the transverse momentum of q_2, **(c)** and **(d)** show the energy of ℓ and **(e)** and **(f)** show the pseudorapidity of ℓ. All plots have a logarithmic scale on the y-axis.

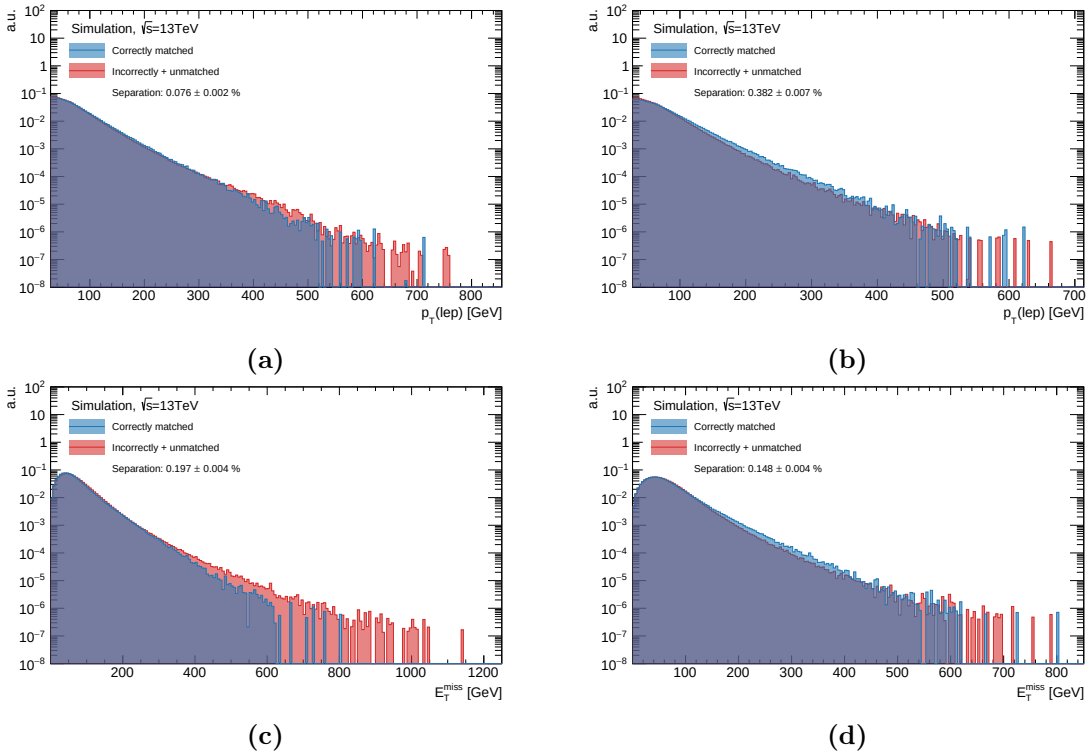


Figure 10.16: Comparison of the normalized distributions of the DNN variables before and after applying the additional window cuts as explained in section 7.3. Each row shows a variable before applying the window cuts (left) and after applying the window cuts (right). **(a)** and **(b)** show the transverse momentum of ℓ and **(c)** and **(d)** show $E_{\mathrm{T}}^{\mathrm{miss}}$. All plots have a logarithmic scale on the y-axis.

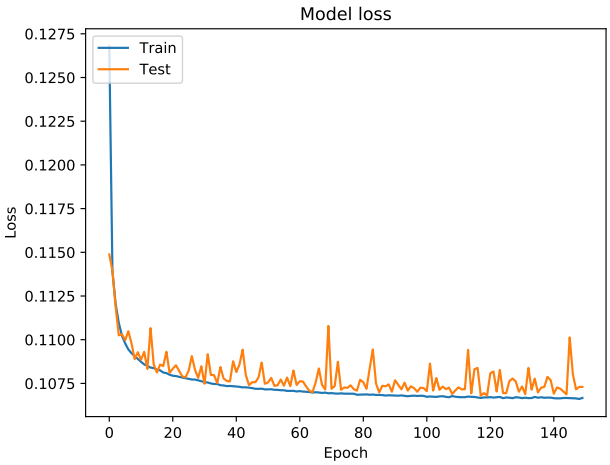


Figure 10.17: The loss of the DNN as a function of the training epoch. The training is performed with the four momenta variables but without applying the window cuts. The blue line shows the loss of the training sample while the orange line shows the loss of the test sample.

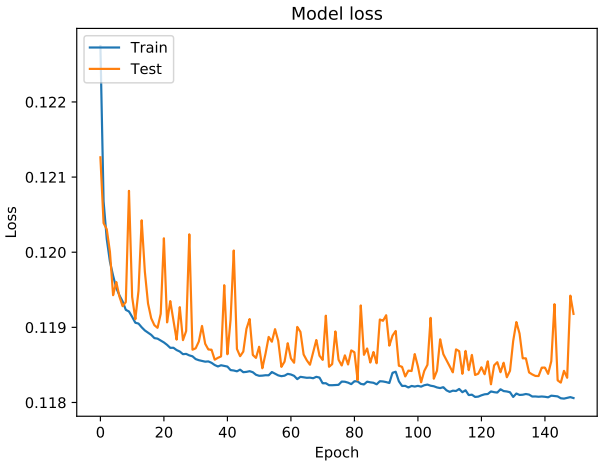


Figure 10.18: The loss of the DNN as a function of the training epoch. The training is performed with the BDT variables but without applying the window cuts. The blue line shows the loss of the training sample while the orange line shows the loss of the test sample.

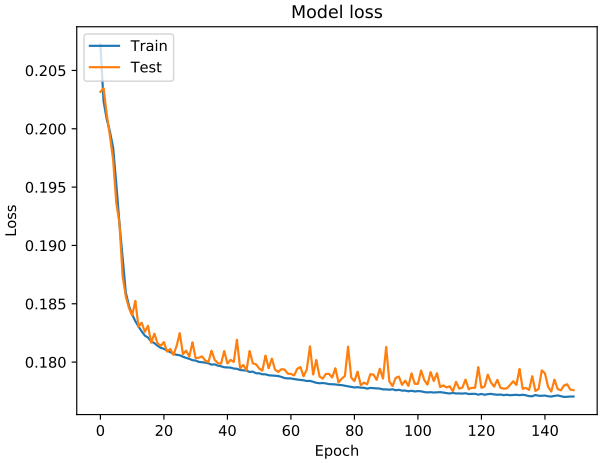


Figure 10.19: The loss of the DNN as a function of the training epoch. The training is performed the four momenta variables and with applying the window cuts. The blue line shows the loss of the training sample while the orange line shows the loss of the test sample.

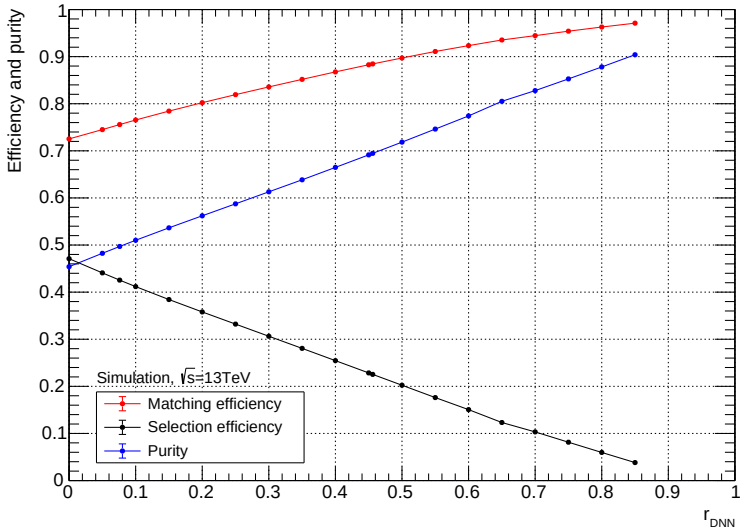


Figure 10.20: Matching efficiency, selection efficiency and purity as a function of the DNN cut value. The DNN training is performed with the information of the four momenta of both the lepton and the jets chosen by KLFitter and the window cuts are applied before performing the training. The statistical uncertainties are drawn in the graph but they are too small to be visible.

Bibliography

[1] The ATLAS Collaboration. "Measurement of the top quark mass in the $t\bar{t} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results". In: *Eur. Phys. J.* C79.4 (2019), p. 290. DOI: 10.1140/epjc/s10052-019-6757-9. arXiv: 1810.01772 [hep-ex].

[2] 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: (2012). DOI: 10.1016/j.physletb.2012.08.020.

[3] The CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: (2012). DOI: 10.1016/j.physletb.2012.08.021.

[4] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (Feb. 1961), pp. 579–588. DOI: 10.1016/0029-5582(61)90469-2.

[5] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc.* C680519 (1968), pp. 367–377.

[6] G. 't Hooft and M. Veltman. "Regularization and renormalization of gauge fields". In: *Nuclear Physics B* 44.1 (July 1972), pp. 189–213. DOI: 10.1016/0550-3213(72)90279-9.

[7] Steven Weinberg. "A Model of Leptons". In: *Physical Review Letters* 19.21 (Nov. 1967), pp. 1264–1266. DOI: 10.1103/physrevlett.19.1264.

[8] Wolfgang Demtröder. *Experimentalphysik 4*. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-01598-4.

[9] David Griffiths. *Introduction to Elementary Particles*. Wiley-VCH Verlag GmbH, Dec. 1987. DOI: 10.1002/9783527618460.

[10] Erich Lohrmann. *Hochenergiephysik*. B. G. Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden., 2005.

[11] Stefan Scherer. *Symmetrien und Gruppen in der Teilchenphysik*. Springer Berlin Heidelberg, 2016. DOI: 10.1007/978-3-662-47734-2.

[12] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001.

[13] Nicola Cabibbo. "Unitary Symmetry and Leptonic Decays". In: *Physical Review Letters* 10.12 (June 1963), pp. 531–533. DOI: 10.1103/physrevlett.10.531.

[14] 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. DOI: 10.1143/ptp.49.652.

[15] Owe Philipsen. *Quantenfeldtheorie und das Standardmodell der Teilchenphysik*. Springer Berlin Heidelberg, 2018. DOI: 10.1007/978-3-662-57820-9.

[16] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Physical Review Letters* 13.9 (Aug. 1964), pp. 321–323. DOI: 10.1103/physrevlett.13.321.

[17] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13.16 (Oct. 1964), pp. 508–509. DOI: 10.1103/physrevlett.13.508.

[18] J Ellis. *Higgs Physics*. eng. 2015. DOI: 10.5170/cern-2015-004.117.

[19] Julie Haffner. "The CERN accelerator complex. Complexe des accélérateurs du CERN". In: (Oct. 2013). General Photo. URL: https://cds.cern.ch/record/1621894.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Adrian Vogel. "ATLAS Transition Radiation Tracker (TRT): Straw tube gaseous detectors at high rates". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 732 (Dec. 2013), pp. 277–280. DOI: 10.1016/j.nima.2013.07.020.

[25] The ATLAS Collaboration. "Performance of the ATLAS trigger system in 2015". In: *The European Physical Journal C* 77.5 (May 2017). DOI: 10.1140/epjc/s10052-017-4852-3.

[26] Arnulf Quadt. *Top Quark Physics at Hadron Colliders*. Springer Berlin Heidelberg, 2007. DOI: 10.1007/978-3-540-71060-8.

[27] R. K. Ellis, W. J. Stirling, and B. R. Webber. *QCD and Collider Physics*. Cambridge University Press, 1996. DOI: 10.1017/cbo9780511628788.

[28] The ATLAS Collaboration. "Measurement of the inclusive cross-sections of single top-quark and top-antiquark t-channel production in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector". In: *Journal of High Energy Physics* 2017.4 (Apr. 2017). DOI: 10.1007/jhep04(2017)086.

[29] *Measurements of the $t\bar{t}$ production cross-section in the dilepton and lepton-plus-jets channels and of the ratio of the $t\bar{t}$ and Z boson cross-sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2015-049. Geneva: CERN, Sept. 2015. URL: http://cds.cern.ch/record/2052605.

[30] 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), p. 43. ISSN: 1029-8479. DOI: 10.1007/JHEP06(2010)043.

[31] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159 –177. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.01.024.

[32] *Improvements in $t\bar{t}$ modelling using NLO+PS Monte Carlo generators for Run2*. Tech. rep. ATL-PHYS-PUB-2018-009. Geneva: CERN, July 2018. URL: http://cds.cern.ch/record/2630327.

[33] *ATLAS Run 1 Pythia8 tunes*. Tech. rep. ATL-PHYS-PUB-2014-021. Geneva: CERN, Nov. 2014. URL: http://cds.cern.ch/record/1966419.

[34] 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 (July 2003), pp. 250–303. DOI: 10.1016/s0168-9002(03)01368-8.

[35] Matteo Cacciari and Gavin P. Salam. "Dispelling the N^3 myth for the Kt jet-finder". In: (2005). DOI: 10.1016/j.physletb.2006.08.037.

[36] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: (2008). DOI: 10.1088/1126-6708/2008/04/063.

[37] *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*. Tech. rep. ATL-PHYS-PUB-2016-012. Geneva: CERN, June 2016. URL: https://cds.cern.ch/record/2160731.

[38] J. Erdmann et al. "A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 748 (June 2014), pp. 18–25. DOI: 10.1016/j.nima.2014.02.029.

[39] G. Cowan. *Statistical data analysis*. Oxford University Press, USA, 1998.

[40] Andreas Hoecker et al. "TMVA: Toolkit for Multivariate Data Analysis". In: *PoS* ACAT (2007), p. 040. arXiv: physics/0703039.

[41] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[42] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: http://neuralnetworksanddeeplearning.com/.

[43] N. D. Gagunashvili. "Comparison of weighted and unweighted histograms". In: (2006). arXiv: physics/0605123.

[44] François Chollet et al. *Keras*. https://keras.io. 2015.

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

List of Figures

List of Tables