Machine Learning Techniques for Density Estimation of Top-Quark Production in the ATLAS Detector

Luka Vomberg

Masterarbeit in Physik
angefertigt im Physikalischen Institut

vorgelegt der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität
Bonn

Jun 2022

I hereby declare that this thesis was formulated by myself and that no sources or tools other than those cited were used.

Bonn, 02.06.22 . . . .

Date Signature

1. Gutachter:	Prof. Dr. Ian Brock
2. Gutachterin:	Prof. Dr. Klaus Desch

Contents

CHAPTER 1

Introduction

Scientists have searched for the most fundamental building blocks of matter – and of nature itself – for a long time. Even predating the concept of quantitative science, natural philosophers thought about the complex world we live in in terms of simple indivisible elements. The Greek word *atomon*, meaning "indivisible" is still used today, though its use to describe the chemical elements turned out to be premature. Today we know that atoms are not indivisible at all, but that they are made up of protons, neutrons, and electrons. Even more, two of these turned out to still be further divisible.

The advent of particle accelerators in the first half of the 20th century allowed for many important discoveries in the smallest realm of nature. Their technical improvements allowed for more precise measurements, and their rising capabilities regarding the energy of accelerated particles allowed for the discovery of many different particles. The emerging "particle zoo" of hundreds of new particles seemed at the time to be chaotic and with little internal order. Out of the desire to find fundamental rules that govern all these particles, the standard model was created.

Some decades later, we are now able to produce and detect all the particles the standard model predicts. This was only made possible by advances not only in accelerator design, but also by advances in computing. Today it is almost unthinkable to perform any high-energy analysis without the use of computers. Therefore, any advance in computational science, be it in hardware or software, is also a potential advance in physics. In turn, progress in physics has caused progress in many other fields, such as medicine [1], or famously by playing a large role in the creation of the internet as we know it today.

One modern development which is helping particle physics to progress is machine learning. New machine learning techniques are being created constantly, and their applicability to particle physics needs to be investigated to not miss out on an analysis method that could prove to be valuable. Because of this, the goal of this thesis is not only to gain information about a specific physics problem in t-quark production but also to test the method itself. As such, *anomaly detection with density estimation* [2] is not only used as the analysis technique of choice but is also a subject of study in its own right.

The first chapter of this thesis gives an overview of the most important aspects of the standard model. After that, the experimental setup used as a basis for the simulations used is explained. Then, an interference effect in t-quark production is explained, including two proposals for methods to treat this interference for further analysis. Finally, a machine learning technique for anomaly detection is employed in an attempt to study the differences between the two proposed methods.

CHAPTER 2

Standard Model of Particle Physics

In the following, a brief overview of the SM is given. For a more rigorous approach see [3], [4], or any other recent textbook on particle physics.

2.1 Particles

The elementary particles of the standard model are displayed in Fig. 2.1. They are principally divided into fermions, which make up matter, and bosons, which are responsible for the fundamental interactions. The defining property of a fermion is that its spin is a half-integer, while bosons have an integer spin. All fundamental fermions have spin $1/2$, while all bosons have spin 1, except for the Higgs boson, which has spin 0.

The 12 elemental fermions are further divided into quarks and leptons. There are 3 charged leptons, the electron (e), the muon (μ), and the tau (τ), each of which has the same electric charge. Their masses increase in the order listed. Additionally, there are 3 uncharged leptons, the neutrinos ν_l, which form a family with their respective charged lepton. Quarks exist in 6 distinct flavours and have a fractional electric charge. They can be grouped into two categories based on their electric charge. Three of them, the up (u), charm (c) and the top (t), are positively charged at $+2/3$ of the electron charge. The quarks in this group are referred to as up-type-q_u. In addition, there are also three down-type-quarks q_d with an electric charge of $-1/3$. These are the down (d), strange (s) and bottom (b) quarks. All fermions f have an antimatter partner $\bar{f}$ with equal mass but opposite charge.

2.2 Interactions

The SM encompasses three interactions or forces. Each of these interactions is based on a group of local transformations under which the Lagrangian is invariant. The Higgs-Brout-Englert mechanism is sometimes also counted among the fundamental forces. It is a necessary part of the standard model because, without it, the masses of the fermions and bosons would break the gauge symmetry. The Higgs boson belonging to this field was the last particle of the SM to be discovered in 2012.

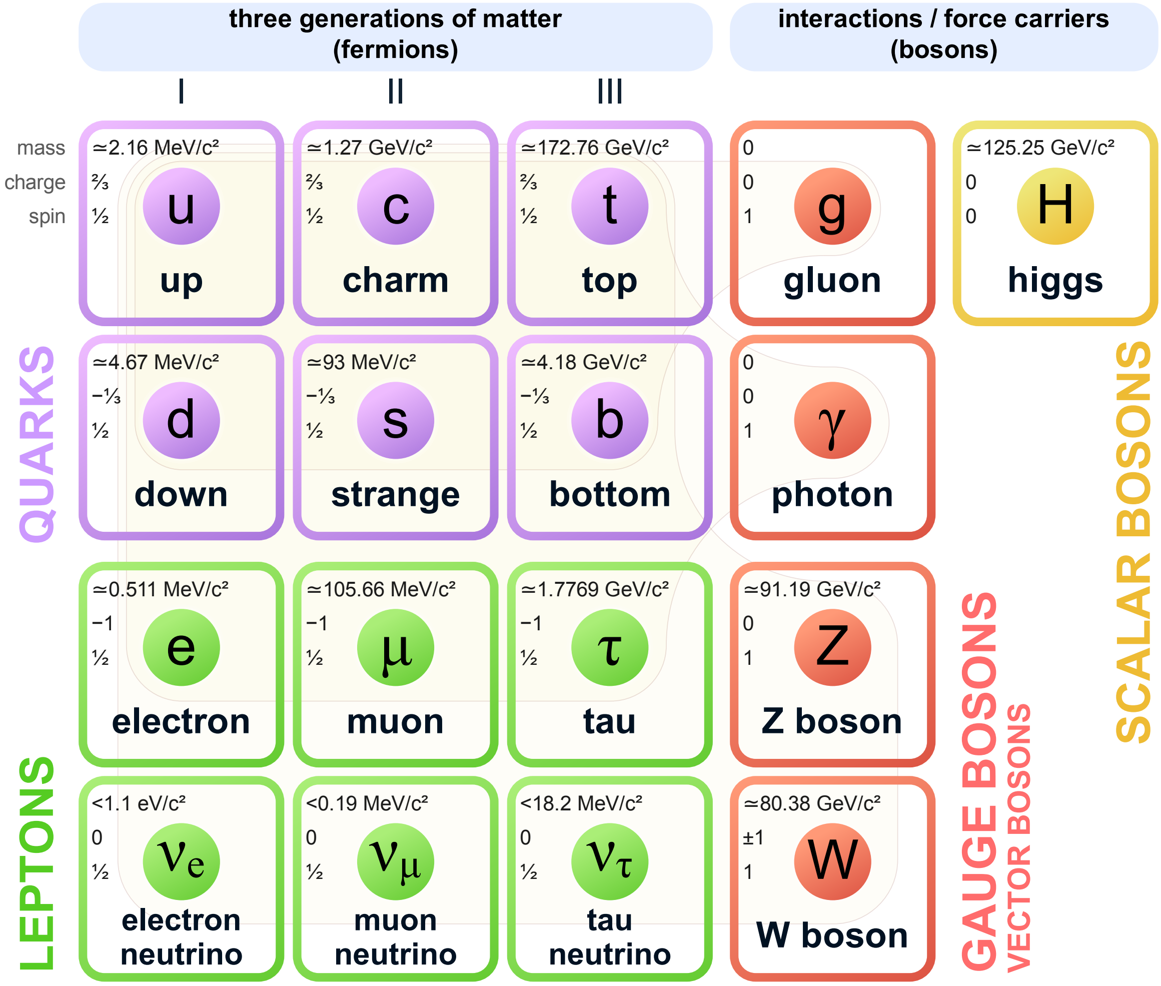


Figure 2.1: The standard model of elementary particles, displayed in a style similar to the periodic table of elements. The original figure [5] was updated to show the most recent values from the particle data group [6]. These do not include a recent measurement of the $W^{\pm}$ mass by the CMS group, which is incompatible with other current measurements. [7]

2.2.1 Quantum Electrodynamics (QED)

QED governs the interactions of electrically charged particles. Its symmetry is described by the unitary group $\mathrm{U}(1)$. The interaction is mediated by its gauge boson, the photon, which is chargeless. It is responsible for the bonding of electrons to the nuclei, for the formation of molecules, and what we experience as “touch” in the macroscopic world, among many other phenomena.

2.2.2 Quantum Chromodynamics (QCD)

The strong force is described by Quantum Chromodynamics (QCD), which only affects particles with a colour charge. It is described by the special unitary group $\mathrm{SU}(3)_\mathrm{C}$. The particles with colour charge are the quarks and gluons. Colour charge can take the values red, green, blue, or their respective anticolour. The term colour and the names of its possible values are not to be taken literally, it was chosen as a loose analogy to real colours. Hadrons are only stable if they are colour neutral (white). A particle made up of three quarks, which possess all three different (anti)colour charges, is called a baryon, while a hadron consisting of 2 quarks with the same colour-anticolour charge is called a meson.

The force carrier of QCD is the gluon. The gluon itself also carries a superposition of colour and anticolour charge, which makes it possible for gluons to directly interact with each other, and also to interact with themselves. This causes the field strength between two quarks to stay equal irrespective of the distance between them. The energy in the gluon field between the quarks rises with further separation, and at some point, it becomes energetically favourable to create a $q\bar{q}$ pair, with which the initial quarks can recombine. This happens until the colour charge of the final particles created is “white”. This process is called hadronisation. From this, it follows that no quark is stable by itself.

Among the hadrons are the protons and neutrons that make up most of the common matter in the shape of chemical elements. The proton is the lightest baryon, all others are more massive and can thus decay. If the initial baryon is massive enough, it might go through multiple decays. The neutron is only stable when bound in an atomic nucleus consisting of multiple protons and neutrons. The strong force is also responsible for the stability of these nuclei. They are bound by a residual force, similar to how molecular bonds can be created by residual electromagnetic fields, even though the atoms are electrically neutral.

2.2.3 (Electro)Weak Interaction

The weak interaction is transmitted by the $W^\pm$ and Z bosons and acts on all fermions, as well as on each other. In contrast to QED and QCD, its gauge bosons are massive. Due to the mass of these bosons, they decay quickly, which causes their range to be limited. Therefore it only affects subatomic processes, such as the radioactive β-decay. The weak force is also the only force not to conserve parity. It follows that weak force couples differently to fermions and antifermions, depending on their chirality, which is the projection of the particles' spin on its momentum vector. It is also the only interaction able to change the flavour of quarks. Therefore it is mainly responsible for the decay of heavy quarks into lighter ones. The weak force does not couple to the mass eigenstates of the quarks like the other interactions do, but to separate weak eigenstates. They can be related to each other by the unitary Cabbibo-Kobayashi-Maskawa (CKM) matrix V_CKM [6]:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_\mathrm{CKM} \times \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \text{ with} \tag{2.1}$$

$$V_{\mathrm{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} = \begin{pmatrix} 0.97401 & 0.22650 & 0.00361 \\ 0.22636 & 0.97320 & 0.04053 \\ 0.00854 & 0.03978 & 0.99917 \end{pmatrix}. \tag{2.2}$$

The element V_{ij} describes a part of the coupling strength between the quarks i and j. The probability of one being transmuted to the other is proportional to $|V_{ij}|^2$. These transmutations only happen by interaction with a charged $W^{\pm}$ and not with a Z. Due to charge conservation, which is respected by all interactions, an up-type quark can only become a down-type quark, as a charged W is emitted, and vice versa. The electromagnetic and weak interactions can be unified in the Glashow-Salam-Weinberg Theory [8–10]. The gauge group of this part of the SM is $\mathrm{SU}(2) \times \mathrm{U}(1)$. The total gauge group of the standard model is then:

$$\mathrm{SU}(3)_{\mathrm{C}} \times \mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)_{Y} \tag{2.3}$$

2.3 Limits of the Standard Model

Despite the great success of the SM to explain and predict the behaviour of fundamental particles, it has serious flaws. The most obvious one is that the fourth force, gravity, is not accounted for at all. It is explained with great success in general relativity. Both theories have been tested and found to be exceedingly good at explaining many natural phenomena, as well as making testable predictions, however, the theories are not compatible. Another unexplained phenomenon are neutrino oscillations, which require neutrinos to have a non-zero mass [11]. The SM however predicts zero mass neutrinos. The predictions of the SM regarding the total mass that exists in the universe also differ by about a factor of 5 from cosmological observations [12]. According to the SM, the baryon number is a conserved quantity, meaning that the number of baryons minus the number of anti-baryons is constant. The observable universe is almost exclusively made up of baryons, with only trace amounts of anti-baryons. This asymmetry cannot be explained within the SM. There are many ongoing searches for physics beyond the SM, but no experimental evidence could be found so far.

CHAPTER 3

The Large Hadron Collider (LHC) and the ATLAS Detector

3.1 Large Hadron Collider

The LHC is a collider type accelerator, located at the European Organisation for Nuclear Research (CERN), which was established in 1954 near Geneva. It is the largest accelerator in the world with a circumference of 26.7 km. The LHC is the successor of the Large Electron-Positron Collider (LEP) and is housed in the same tunnel about 100 m underground. The operations of the LHC began in 2008. Since then two runs of physics data taking were conducted, the first from 2009 to 2013. During this time, a beam energy of 4 TeV was reached. After some maintenance and upgrades, the LHC was operational again from 2015 to 2018. It was possible to substantially increase the beam energy to 6.5 TeV for this period. All data used in this thesis is from the second run, and all simulated events are simulated according to run 2 specifics. As of the writing of this thesis, a third run has started, but has not yet yielded usable data.

The LHC can be operated with protons or lead ions. The protons are created by ionising hydrogen. The lead is produced from a solid block, which is partially vaporised and then ionised as well. Subsequently, the particles are accelerated through a series of pre-accelerators before they finally arrive in the LHC itself. This is necessary because a single machine cannot accommodate particles at energies ranging from nearly zero to 6.5 TeV. Most of the pre-accelerators were already built for other experiments at CERN, preceding the LHC. Therefore it was practical to use the already well established infrastructure instead of building a second one. The smaller accelerators at CERN are also used for experiments requiring less energetic particles. The whole accelerator chain and some of the major experiments can be seen in Fig. 3.1. The ions are passing through a linear accelerator first. Next, the particles are sent through a series of synchrotrons, which is a type of circular accelerator that utilises an alternating electric field, through which the charged particles pass multiple times, gaining energy on each revolution. The Proton Synchrotron Booster, the Proton Synchrotron and the Super Proton Synchrotron accelerate the protons to 1.4 GeV, 25 GeV and 450 GeV respectively.[1]

Finally, the particles reach the LHC, which is also a synchrotron. They are injected in opposing directions into two separate beam pipes. Here they are accelerated up to an energy of 6.5 TeV (6.8 TeV in run 3). The beam can be stored in this state for many hours with only small losses. To keep it on

[1] These are the energies for run 2. For run 3 these are 160 MeV, 2 GeV, 26 GeV and 450 GeV.

its circular path, 1232 supraconducting dipole magnets with field strengths of 8.3 T are employed. In addition to this, 474 quadrupole magnets are utilised to focus the beam, as well as some sextupole, octupole and decapole magnets, which correct small imperfections in the magnetic fields of the dipoles.

There are four interaction points where the opposing beams are steered to collide. At each such point, there is one detector. The two general-purpose detectors ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) do not specialise in detecting a certain type of event or particle but are designed to cover a wide range of measurements. They both have cylindrical symmetry around the beam pipes and cover nearly the entire solid angle. The other detectors are LHCb (LHC beauty) and ALICE (A Large Ion Collider Experiment), which specialise in the detection of B mesons and heavy-ion collisions respectively.

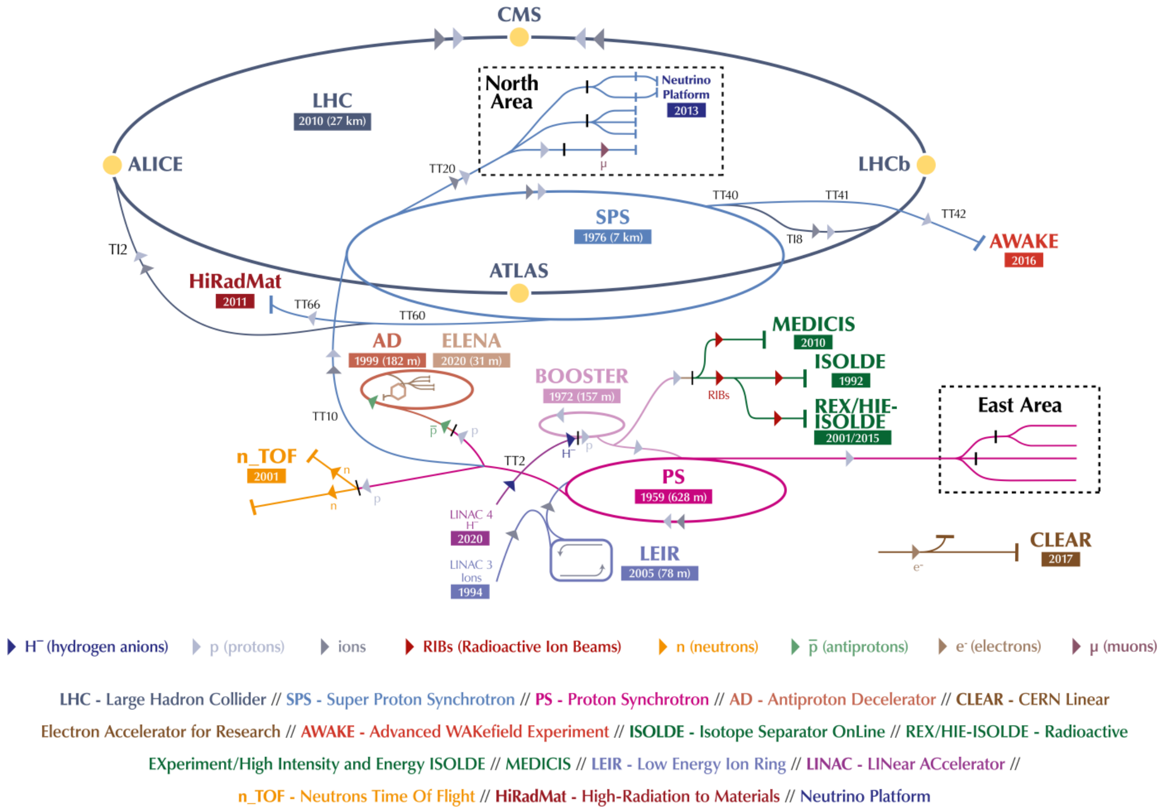

CMS
LHC
2010 (27 km)
North Area
Neutrino Platform
2013
ALICE
TT20
LHCb
TT40
TT41
TT42
AWAKE
2016
SPS
1976 (7 km)
TI8
TI2
HiRadMat
2011
TT66
ATLAS
TT60
AD
1999 (182 m)
ELENA
2020 (31 m)
MEDICIS
2010
ISOLDE
1992
BOOSTER
1972 (157 m)
RIBs
REX/HIE-ISOLDE
2001/2015
East Area
TT10
n_TOF
2001
TT2
PS
1959 (628 m)
LINAC 4
H⁻
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
CLEAR
2017
e⁻
μ
p
p̄
n
H⁻
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
μ (muons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE-ISOLDE - Radioactive EXperiment/High Intensity and Energy ISOLDE // MEDICIS // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // Neutrino Platform

Figure 3.1: The accelerator complex at CERN [13]. Schematic is not shown to scale. The major accelerating structures and experiments are shown with the year of their installation. For LHC operation the particles are passing through pre-accelerators before injection into the LHC storage ring.

3.1.1 Collider Kinematics

The energies listed above are the energies of single particles in the beam. For physical purposes it is useful to define the center-of-mass energy $\sqrt{s}$. It is an important characteristic value for any accelerator, because it is the energy available to produce new particles in a collision. It is defined as:

$$\sqrt{s} = \sqrt{(E_1 + E_2)^2 - (\boldsymbol{p}_1 + \boldsymbol{p}_2)^2}, \tag{3.1}$$

where E_i and p_i are the energy and 3-momentum of the beams. For the LHC with its 2 opposing beams with $\boldsymbol{p}_1 = -\boldsymbol{p}_2$ this simplifies to:

$$\sqrt{s} = 2E = 2 \cdot 6.5\,\text{TeV} = 13\,\text{TeV}. \tag{3.2}$$

Another key performance characteristic of an accelerator is its luminosity L. It is a measure of the number of events produced per area and time. A slightly simplified formula for the case of colliding beams can be given as [14]:

$$L = \frac{f N_1 N_2 n_\text{b}}{4\pi\sigma_x\sigma_y}. \tag{3.3}$$

At the LHC there are $N_1 = N_2 = 1.15 \times 10^{11}$ particles in each bunch, of which there are $n_\text{b} = 2556$ with a revolution frequency $f = 11.145 \times 10^3\,\text{s}^{-1}$. The beams are assumed to be Gaussian in the transverse plane with standard deviation $\sigma_i = 2.5\,\mu\text{m}$. Eq. (3.3) also assumes that the collision is exactly head on and that the bunches are uniformly distributed along the longitudinal plane. The LHC reached a luminosity of $L = 1.9 \times 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ in proton-proton collisions [15]. The Luminosity also relates directly to the number of expected events N:

$$N = \sigma\mathcal{L}_\text{int} = \sigma \int L\,\mathrm{d}t\,, \tag{3.4}$$

where $\mathcal{L}$ is the integrated luminosity and σ the cross-section of one specific scattering process. To gain as many measurements as possible it is important to optimise the luminosity by means of the beam properties, as well as use as much of the beamtime for actual data taking.

3.1.2 Proton Proton Scattering

As explained in Chapter 2, the proton is not a fundamental particle, but a composite structure. In a simplified picture, it is made up of 2 u-quarks and 1 d-quark, which are referred to as the valence quarks. The masses of these make up only about 1 % of the total mass of a proton, 938.3 MeV. The rest of the mass is created as an emergent property of the interactions of the quarks with each other [16]. The quarks are constantly exchanging gluons, which in turn can create $q\bar{q}$ pairs. These pairs do not need to be $u\bar{u}$ or $d\bar{d}$, but can also be a heavier quark flavour. These additional quarks are called sea quarks. The contents of a proton are often referred to as partons, a term which can refer to any of the valence quarks, gluons or sea quarks or combinations thereof.

If a proton is involved in a collision, it can either scatter elastically, which means that the proton leaves the process as a whole, or inelastically. In inelastic scattering, only a single parton is scattered individually, thus breaking the proton apart, or exciting it into a higher energy state. In such a scattering, the relevant energy of the collision is not the energy of the proton as a whole, but that of the parton which gets scattered. Each parton carries a fraction of the total proton momentum. The fraction of this parton momentum over the energy of the boson mediating the interaction is called the Bjorken x. This variable defines the probability of scattering off any particular type of parton in the proton. Roughly

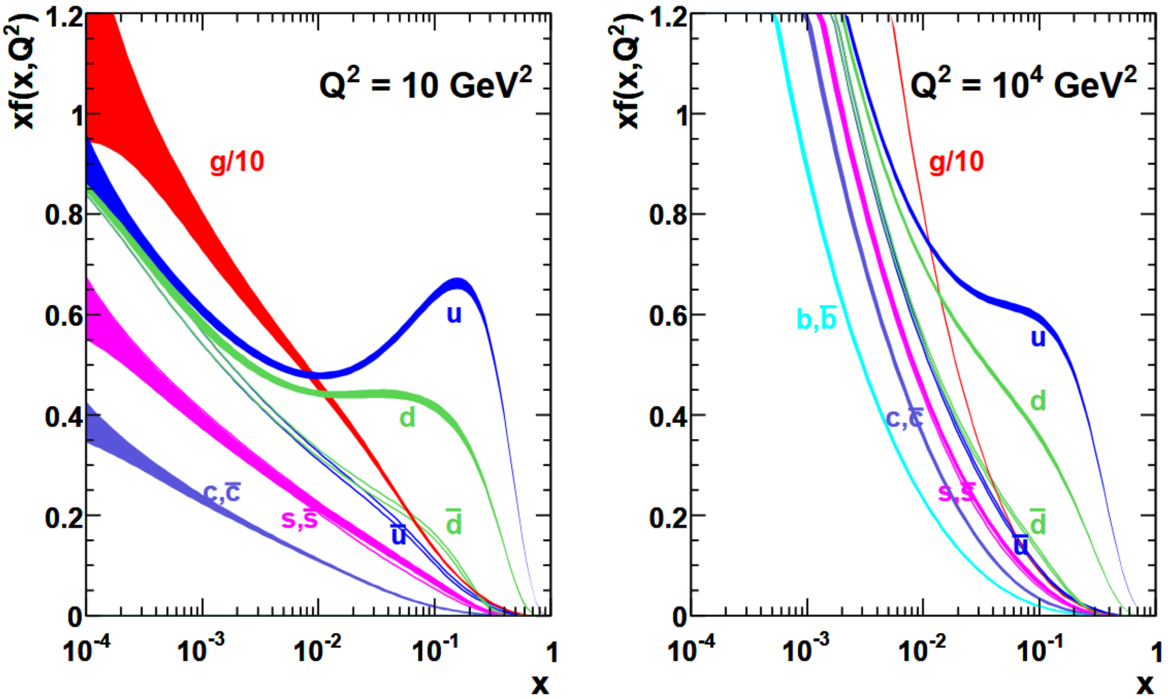


Figure 3.2: The parton distribution functions as a function of the Bjorken x at two different momentum transfers Q. The line width corresponds to uncertainty estimates. At higher momentum transfer Q^2 the scattering of sea quarks and gluons is more likely than at lower momentum transfers. [17]

speaking, the probability to scatter of a gluon or sea quark rises as x becomes smaller. In other words, the higher the energy of the collision, the higher the probability to scatter of a parton that is not one of the sea quarks. This behaviour is presented in Fig. 3.2. The scattering of heavier quark flavours becomes more likely when the momentum transfer Q is larger. At the LHC quarks up to the b can thus be "found" without specifically producing them first.

3.2 The ATLAS Detector

ATLAS (A Toroidal LHC ApparatuS) is the largest collider detector ever built. Its geometry roughly describes a 46 m long cylinder with a 25 m diameter weighing about 7 000 t. A schematic is shown in Fig. 3.3. As a general-purpose detector, its task is to be able to measure precisely as many different final states as possible. To achieve this, different subsystems are working complementary to each other to create the most complete picture of the collision events.

3.2.1 Coordinate System

In ATLAS coordinates the z-axis is defined to follow the beam pipe. The positive x direction is towards the centre of the LHC and the y-axis points upwards. The polar angle θ is defined as the angle between the x-y plane and z-axis and the azimuthal angle ϕ around the z-axis. In practice it is often more

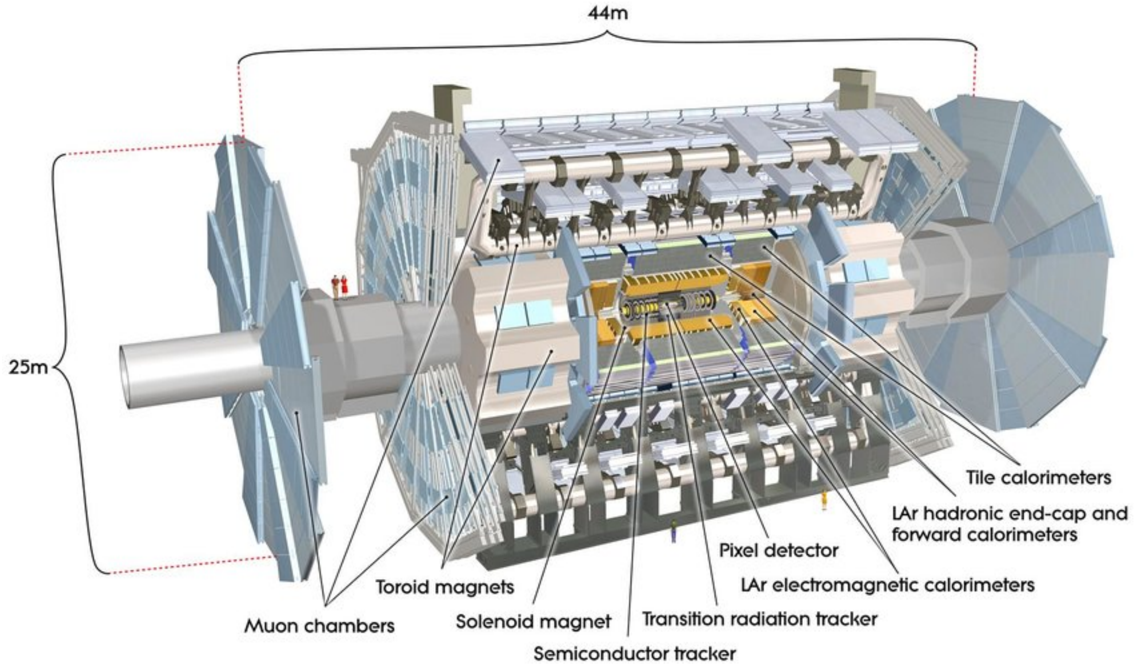

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 3.3: Schematic of the ATLAS detector. The subsystems are layered around the central interaction point. True to Scale. [18]

convenient to use the pseudorapidity η instead of θ:

$$\eta = \frac{1}{2}\log\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{3.5}$$

In contrast to the polar angle, differences between pseudorapidities are invariant under boosts along the z-axis. The angular difference between two objects can then be given as:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{3.6}$$

3.2.2 Detector Systems

The ATLAS Detector is a combination of multiple detector systems. Each of these parts specialises in a specific type of measurement. Combining the information gathered by each of these is necessary to gain insight into the collision events. The subsystems all tend to be cylindrically symmetrical around the z-axis, with a barrel portion surrounding the central beam pipe and an endcap per side on the xy-plane to cover high η events. In Fig. 3.4 the different layers of the barrel section of the ATLAS detector are shown schematically in a cross-sectional view.

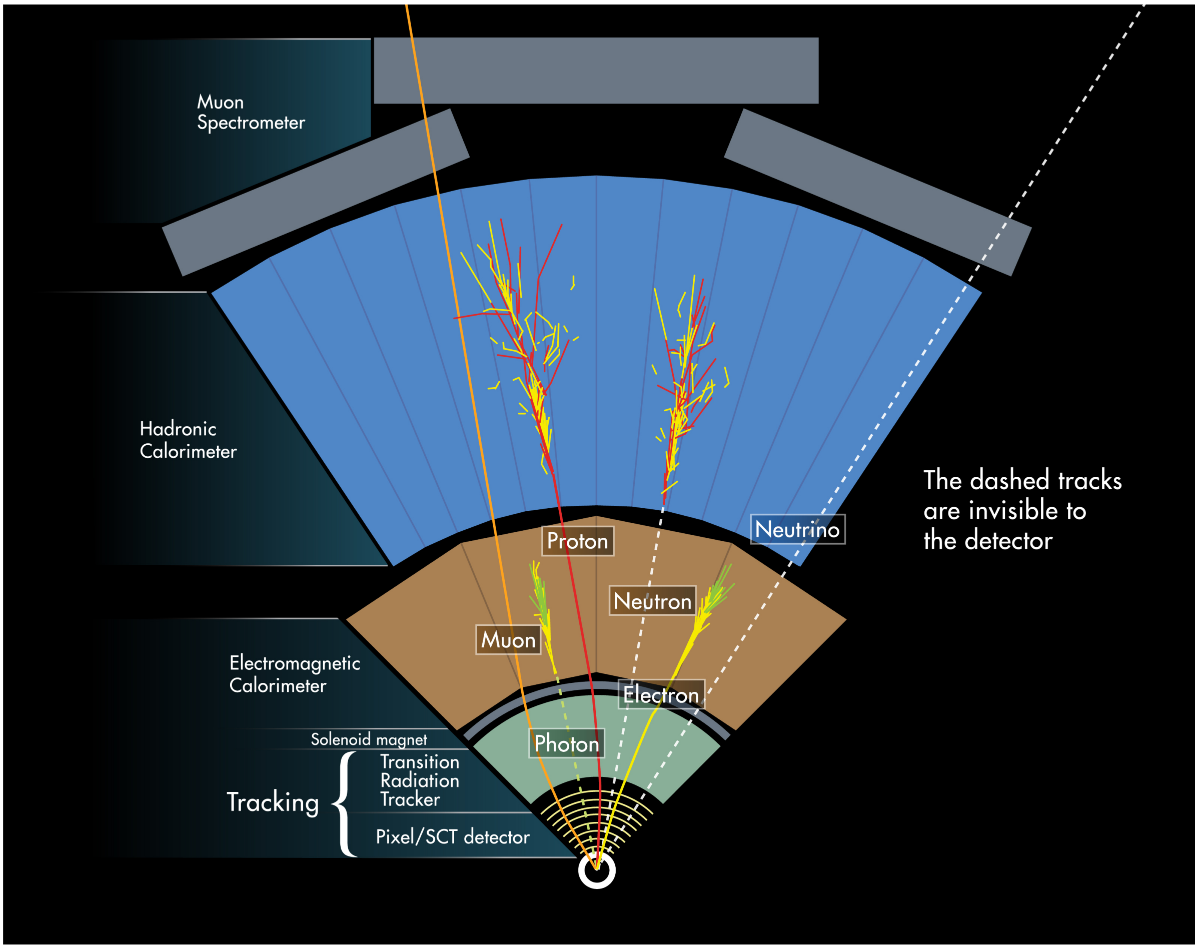


Figure 3.4: Schematic of the ATLAS Detectors subsystems. The innermost layers track electrically charged particles. In the ECal mainly electrons and photons are stopped and their energy measured. Hadrons are producing showers mainly in the HCal and muons are tracked in the ID and MS before leaving the detector volume. Neutrinos are not detected at all. [19]

The Inner Detector (ID)

The innermost part of the ATLAS detector is the ID [20]. A 2 T magnetic field produced by a supraconducting solenoid spans its entire volume. Due to this charged particles passing through are following curved paths. From their trajectory, the particle's charge and momentum can be inferred. The ID is comprised of 3 subsystems that are mainly aimed at measuring these trajectories with good spatial resolution. The ID covers the volume within $|\eta| < 2.5$.

Pixel Detector The pixel detector is the detector system closest to the beampipe at a distance of only 3.3 cm from the interaction point. It is composed of four layers of semiconductor chips. In this type of material, a signal is measured when a charged particle passes through. By separating the chips into many pixels, the precise location of the particle can be measured at multiple points. The first layer is the Insertable B-Layer (IBL), which was inserted before run 2. The IBL delivers a spatial resolution of $8 \times 40\mu\text{m}^2$, while the other 3 layers are slightly less precise at $10 \times 115\mu\text{m}^2$. [21]

Silicon Microstrip Tracker (SCT) The SCT is a semiconductor detector as well. It is made up of four layers of chips in the barrel section and 9 layers in the endcap. Their number and arrangements were chosen to make sure nearly all particles pass through at least 4 layers. The spatial resolution of the SCT is $17 \times 580\mu\text{m}^2$. [20]

Transition Radiation Tracker (TRT) The TRT is a gaseous detector with approximately $3{\times}10^5$ straw tubes with a diameter of 4 mm. Each of these functions as a proportional-mode drift tube. The tubes contain a wire and a Xe gas mixture. Between the wire and the walls of the tube, a potential difference of 1.5 kV is created. Charged particles passing through the gas release electrons from it, which are accelerated towards the wire. During their drift, they can hit other electrons and thus cause a cascade that can be measured. The amount of deposited charges is directly proportional to the deposited energy of the initial collision with the wire. The spatial resolution is approximately 120 µm. This is lower than that of the semiconductor detectors but works complementary to them because the number of hits detected in the TRT is much higher. The volume between the tubes contains polymer fibres which cause transition radiation from highly relativistic particles. This means it affects electrons the most and can serve in their identification. [22]

Calorimeters

The task of the calorimeters is to measure the energy of particles passing through. This is achieved by stopping the particles completely and measuring the energy deposited in the detector material. Since different types of particles behave differently when passing through matter, different types of calorimeters are used in conjunction. They also make some measurements on the trajectory of particles, but the spatial resolution is poor compared to the ones achievable in the ID. The entire calorimeter setup of ATLAS is shown in Fig. 3.5. The Electromagnetic Calorimeter (ECal) is designed to primarily measure photons and electrons, while the Hadronic Calorimeter (HCal) measures the energies of both neutral and electrically charged hadrons.

Electromagnetic Calorimeter For high-energy photons or electrons, the main interactions with matter are pair production and bremsstrahlung respectively. In both of these radiative processes, the initial state

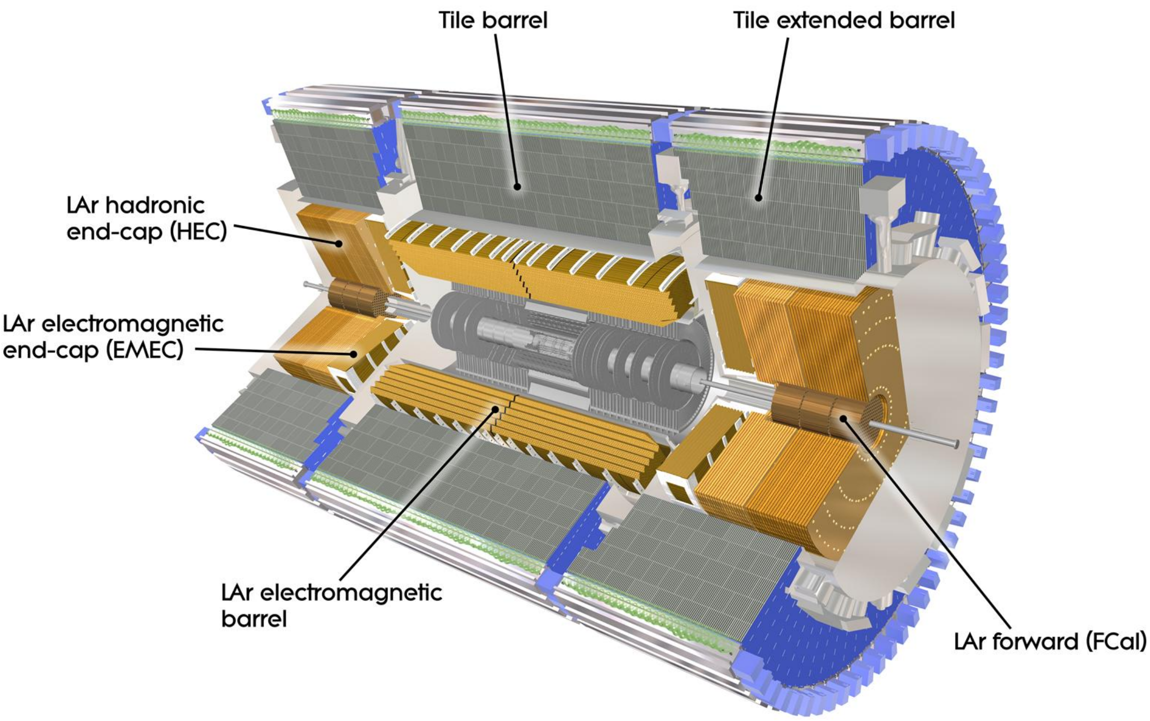

Tile barrel
Tile extended barrel
LAr hadronic end-cap (HEC)
LAr electromagnetic end-cap (EMEC)
LAr electromagnetic barrel
LAr forward (FCal)

Figure 3.5: The ATLAS Calorimeter System. [23]

has one particle, while the final state has two. Due to this, the energy per particle is reduced after each scattering. The produced particles can scatter again, thus creating an electromagnetic shower of photons, electrons and positrons. At low energies, ionisation becomes the main scattering process. The charges freed this way induce a measurable signal from which the deposited energy is inferred.

The cross-section of the radiative processes depends strongly on the atomic number Z of the detector material. Therefore, lead radiators are used in the Ecal to facilitate particle showers. It is then necessary to measure the number of charges which are finally created by these showers. For this purpose Liquid Ar (LAr) is better suited than solid heavy metals with large Z. A large volume of LAr would be necessary to make sure the electromagnetic shower is entirely contained, making it unpractical to use only LAr. For this reason, a series of alternating layers of radiators and LAr is used to exploit the favourable properties of both. The trade-off for this approach is that some of the energy is deposited in the metal and not measured. As a consequence, the energy measurement is only a sample of the total deposited energy. By calibrating the LAr with incoming particles of known energy, it is still possible to infer the total energy deposited in real measurements.

The Ecal consists of the LAr electromagnetic barrel and the LAr electromagnetic end-cap (EMEC). It is completely hermetic in ϕ. This is achieved by folding the layers into an accordion-shape. The LAr barrel covers the region up to $|\eta| < 3.2$. [24]

Hadron Calorimeter Due to the large number of hadrons that can potentially be created or decay within the detector, hadronic particle showers can manifest in more varieties than electromagnetic showers. Since some parts of the showers can traverse a longer distance, the HCal has a larger volume than the Ecal. The barrel portion of the HCal is made up of the Tile Barrel (TileBar) and the Tile Extended barrel (TileExt). Together they cover the region $|\eta| < 1.7$. The same trade-off between effective radiators and active detector material as in the Ecal needs to be made. The tiles in the barrel are made from a combination of steel plates with a high cross-section for many hadronic processes and plastic scintillators connected to photo-multiplier tubes.

The HCal End-Caps (HEC) are constructed in a similar fashion to the Ecal, with the difference, that the radiator material is copper instead of lead. The HEC extends to $|\eta| < 3.2$. In addition, the Forward Calorimeter (FCal) employs a combination of copper and tungsten. The FCal covers the region $|\eta| < 4.9$, which is closest to the beam pipe. [24]

Muon Spectrometer (MS)

Most particles are stopped somewhere in the ID or calorimeter systems. Muons however interact relatively little with matter, meaning they can penetrate those layers. To still gain information from them, the Muon Spectrometer is used. It is unfeasible to completely stop the muons though. Like the ID, the MS employs a magnetic field to bend the trajectories, yielding information on the muons' momenta. Since the muon mass is known, the corresponding energy can be calculated.

Four different detector technologies are employed in the MS. The Resistive Plate Chambers (RPCs) and Thin Gap Chambers (TGCs) are used as triggers, quickly responding with a signal when there is a signal consistent with a muon. The Monitored Drift Tube chambers (MDTs) and Cathode Strip Chambers (CSCs) are used to accurately track the muons via their induced signal.

The muon chambers cover the volume $|\eta| < 2.7$. Due to necessary support structures for the detector, there are a few gaps in the coverage. [25]

3.2.3 Event Reconstruction

Many types of particles are produced and decay again before they can reach any part of the detector. Therefore only the relatively stable products of their collisions can be measured. Each particle detected thus corresponds to only one puzzle piece of the whole picture. Only by combining them is it possible to understand what happened to produce them. For this reason, all measurements made by the different detector components need to be combined. The track information from the ID and MS are used to recreate an initial vertex of the event. This is the position where the final state particles originated. If there are multiple such vertices, the one whose corresponding objects have the highest transverse momenta is called the primary vertex, while the other ones are called pileup vertices. If there are vertices originating from a track of another particle, they are called secondary vertices.

The hits in the calorimeters are clustered and their energy is assigned to a particle track if one is consistent with it. If there is none, an electrically neutral particle can be assumed to have deposited the energy. Some examples of objects that are relevant to this thesis are given here.

Jets are a spatially grouped assembly of hadronic particles. They are typically produced via hadronisation of quarks, gluons or τ. The jets' signatures in the detector are then caused by the long-lived particles in their decay chain, such as π or $K^{\pm}$. The calorimeter hits are grouped into clusters, which correspond to one hadron each if the clustering is correct. If track information from the ID is also available, it will be used together with the calorimeter clusters to reconstruct the initial particle that caused the jet. In ATLAS, mostly the anti-k_t algorithm is used for reconstruction. [26, 27]

Electrons leave tracks in the ID and cause clusters in the EM. For reconstruction, the track is interpolated to the calorimeter. An algorithm decides if a cluster matches this track. If it does, the energy and momentum of the electron can be computed. As a further security against misidentification another multivariate algorithm is used to determine a likelihood value. Based on this value the electron is considered “tight”, “medium” or “loose”. This information is saved. It is then up to the analyst to decide which threshold is appropriate for a given analysis. The reconstruction efficiency of electrons is dependent on their transverse energy E_T and the choice of tightness. It varies from around 65 % for the lowest energy electrons with a tight restriction to almost 100 % for loose electrons with $E_\mathrm{T} > 70\,\mathrm{MeV}$. The misidentification efficiency, also called fake rate, of jets as electrons is estimated around 1–7×10^{-5} %. It is lowest for tight supposed electrons at high E_T. [28]

Muons potentially leave tracks in the ID and the MS. These tracks are first reconstructed independently from each other and then matched. They are identified based on a different multivariate algorithm that depends on what information is available. The efficiency of muon reconstruction is approximately 99 %, except for the gap in the MS at $\eta = 0$. [25]

***b*-jets** are jets that originate from the decay/hadronisation of a b-quark. Since t-quarks decay almost exclusively to b-quarks, b-jets offer a useful discriminant in identifying t-quarks in an event. b-jets have some distinct features, such as the b hadron carrying about 70 % of the initial quark mass. Additionally, b-quarks have a lifetime long enough to traverse a few mm before decaying, thus causing a secondary vertex. Again, there is a multivariate algorithm assigning a score to each potential b-jet. When analysing data, a decision needs to be made on how large the purity of the samples used needs to be. When a high purity is desired, only jets with a large score are considered, and the rest is discarded. This comes at the disadvantage of also discarding more of the true b-jets. Therefore, a trade-off needs to be made and a working point needs to be chosen in accordance with the needs of the analysis. The working points recommended for ATLAS analyses are shown in Table 3.1 together with the estimated purities and rejection rates of other objects that could be misinterpreted as b-jets. [26]

Efficiency [%]	Purity [%]	c-jet rejection	τ-jet ejection	Light-jet rejection
60	99.0	34	184	1 538
70	97.5	12	55	381
77	95.2	6	22	134
85	89.7	3.1	8.2	33

Table 3.1: Efficiencies of the b-tagging working points recommended for ATLAS analyses. Efficiency is the proportion of b-jets that are still included when this working point is chosen. Rejection refers to the number of jets of which one is expected to be misidentified as a b-jet. The values are estimated using simulated samples. Numbers are taken from [26, 29].

3.3 Monte Carlo and Data Samples

It is often useful to have simulations of certain processes, for example, to estimate the values in Table 3.1 above. These simulations are Monte Carlo Methods (MC). MCs describe a wide field of numerical simulation techniques. In particle physics applications, they are used to simulate events in high-energy collisions. This takes into account the productions of new particles, their decays, hadronisation and many more physical effects. In addition, a simulation of the detector response to the simulated particles is carried out as well. All produced samples are assigned a weight during the simulation. When the events are counted, for example, to fill a histogram, each event does not increment the histogram count by 1, but by its weight [30]. Taking this into account, the MC samples, or just MC, can be compared to data to evaluate if the data and the simulations fit together. From this, it can be inferred if the model that the simulations are based on is in accordance with the measurements. The MC samples used in this thesis are created to correspond to 1 fb^{-1} in total. When they are compared to data, their weights must be multiplicated by a factor, that takes into account how much data is used for the comparison, e.g. if the full 139 GeV of available run 2 data is used, this factor would need to be 139.

CHAPTER 4

Top-Quark Production

4.1 Top-Quark Properties

The t-quark was theorised to exist as a part of the third generation of matter in 1973 [31]. 25 years later, its discovery at Fermilab was announced by the CDF collaboration [32]. The t-quark is the heaviest particle in the SM with a mass of (172.76±0.30) GeV [6]. Its decay width is $(1.42\,^{+0.19}_{-0.15})$ GeV, which corresponds to a lifetime of $\tau_t \approx 5 \times 10^{-25}$ s [6, 26]. This is smaller than the typical hadronisation time. Therefore the t-quark decays as a free particle, free of hadronisation effects that could shroud the decay [33], which is useful to measure the CKM element V_{tb}. The other quark flavours in contrast only decay after they have hadronised. An example of this advantage is that in the decay $t \longrightarrow bW$, the spin polarisation of the b is exactly that of the t [26]. This would not be the case if the t-quark was in a bound state. With its large mass, the t-quark is also close to the electroweak scale and could give insight into this area of the SM [33]. This makes the t-quark a valuable probe.

4.2 Top-Quark Production at the LHC

4.2.1 Production Channels

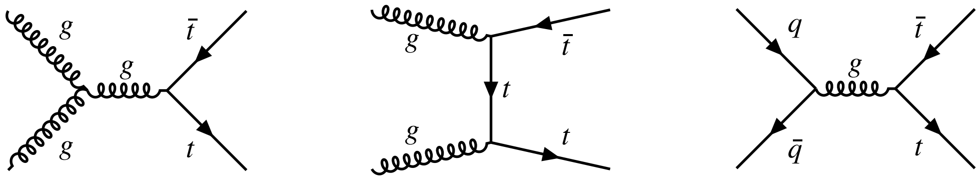
t-quarks can be produced either by the strong or the weak interaction. The strong interaction conserves quark flavour, therefore production is only possible by creating a $t\bar{t}$ pair. Some possible Leading Order (LO) Feynman diagrams of this process are displayed in Fig. 4.1. LO means that the processes displayed are the ones with the lowest order in the coupling constants and therefore the diagrams contributing most to the total cross-section. Since each vertex contributes to the order of coupling constants, these are the diagrams with the least vertices.

The most recent calculation of total cross section for $t\bar{t}$ production results in [34]:

$$\sigma_{t\bar{t}}(\sqrt{s} = 13\,\text{TeV}) = (830\ \pm 0.4\,(\text{stat.})\ \pm 36\,(\text{sys.}))\,\text{pb}. \tag{4.1}$$

This calculation takes into account all diagrams in pp collisions up to Next-to-Next-to-Leading-Order (NNLO).

The production of a single t-quark cannot be done via the strong force. However, there are three distinct single t-quark production processes via the weak interaction. These are possible in the s and



(a) Gluon fusion, s-channel **(b)** Gluon fusion, t-channel **(c)** $q\bar{q}$ annihilation

Figure 4.1: Some examples of leading order $t\bar{t}$ production via the strong interaction. At $\sqrt{s} = 14\,\text{TeV}$ gluon fusion $gg \to t\bar{t}$ is the dominant process, making up about 90 % of total $t\bar{t}$ production at ATLAS. [26]

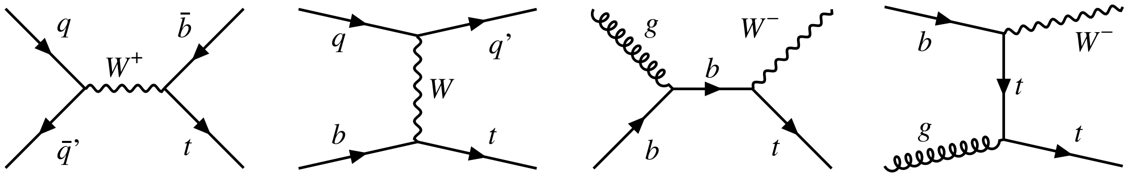
t-channel and via associated production with a W boson; the latter is process is called tW. An example for each type of signature is given in Fig. 4.2. The production cross sections are [6]:

$$\sigma_{tW\text{-channel}}(\sqrt{s} = 13\,\text{TeV}) = (71.7\ \pm 1.8\ \pm 3.4)\,\text{pb}, \tag{4.2}$$

$$\sigma_{t\text{-channel}}(\sqrt{s} = 13\,\text{TeV}) = (216.99\,^{+9.04}_{-7.71})\,\text{pb}, \tag{4.3}$$

$$\sigma_{s\text{-channel}}(\sqrt{s} = 13\,\text{TeV}) = (10.32\,^{+0.10}_{-0.36})\,\text{pb}. \tag{4.4}$$

For a discussion on the uncertainties and methods used to obtain these values, see [35].



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

Figure 4.2: **(a)** and **(b)**: Single t-quark production in s and t-channel. **(c)** and **(d)**: The tW process under the assumption of a five-flavour scheme, in which the proton parton-distribution-function includes a b-quark contribution. If the four flavour scheme is applied, the b-quark needs to be created in pair production from a gluon first. [26]

4.3 Top-Quark Decay

The t-quark can only decay via the weak interaction. The CKM-Matrix introduced in Section 2.2 strongly suppresses the possible decays into d and s-quarks. Therefore the decay $t \longrightarrow bW$ is the dominant process. From this it follows that all $t\bar{t}$-processes, as well as tW will contain two W bosons. The W boson can decay leptonically $W^{\pm} \longrightarrow \bar{l}\nu_l/l\bar{\nu}_l$, or hadronically $W \longrightarrow q_d q'_u$. The branching rates are [6][1]:

$$\mathcal{B}_{\text{lep}} = (32.72 \pm 0.30)\,\%, \tag{4.5}$$

$$\mathcal{B}_{\text{had}} = (67.41 \pm 0.27)\,\%. \tag{4.6}$$

[1] [6] gives the error for each lepton family separately. The error for $\mathcal{B}_{\text{lep}}$ is taken as the sum of squares of these errors

There are then three possible channels in which two W-bosons can decay. Their branching ratios are:

$$\mathcal{B}^{\star}_{\text{dilep}} = (10.71 \pm 0.20)\,\%, \tag{4.7}$$

$$\mathcal{B}_{\text{lep,had}} = (44.11 \pm 0.44)\,\%, \tag{4.8}$$

$$\mathcal{B}_{\text{dihad}} = (45.44 \pm 0.36)\,\%. \tag{4.9}$$

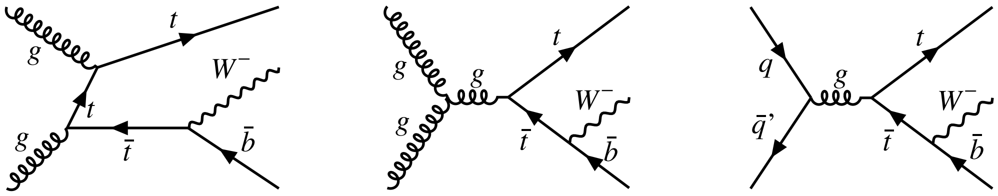
In the dilepton channel the least amounts of jets are involved. This makes the detector signature easier to identify in this channel. Therefore it is the preferred channel to study the tW process, even though it has the lowest branching ratio [26]. The dilepton channel branching ratio given in Eq. (4.7) includes decays into τ, which can also decay hadronically and create jets. Removing these events from the branching ratio for the dilepton channel as well, but keeping the ones in which the τ decays leptonically, finally yields a branching ratio of:

$$\mathcal{B}_{\text{dilep}} = (6.51 \pm 0.09)\,\% \tag{4.10}$$

for the dilepton channel. This ratio is independent of whether the production channel was tW or $t\bar{t}$.

4.4 tW and $t\bar{t}$ Interference

The diagrams describing the tW processes in Section 4.2.1 are of LO. Some example diagrams for tW at NLO are shown in Fig. 4.3. The former are referred to as singly resonant since they are resonant in the production of a real t-quark, while the latter are examples of doubly resonant diagrams, because they are potentially resonant in the production of both t and $\bar{t}$ [33]. The LO $t\bar{t}$ and NLO tW diagrams are identical to the $t\bar{t}$ LO diagrams in Fig. 4.1. This in turn means that the definitions of tW and $t\bar{t}$ are only cleanly separable in LO. In NLO, and consequently also in higher orders, tW is an ill-defined process [26, 33, 36, 37]. Due to identical diagrams being present in $t\bar{t}$ LO and tW NLO, the processes can be thought of as interfering with each other. Due to the much larger cross-section of $t\bar{t}$ in LO, the interference contribution is larger than tW at LO in total [33]. However, it is not sufficient to include only LO contributions when calculating observables [33].



(a) Gluon fusion, s-channel **(b)** Gluon fusion, t-channel **(c)** $q\bar{q}$ annihilation

Figure 4.3: Some examples of doubly resonant NLO tW production diagrams that are equivalent to $t\bar{t}$ diagrams [36].

4.4.1 Diagram Removal and Diagram Subtraction Schemes

Despite the issues mentioned, it is still seen as beneficial to define tW in a way that is usable when comparing theoretical predictions and experimental data. To achieve this, [33] suggests two different schemes, which aim at clearly separating $t\bar{t}$ and tW at NLO.

Consider the amplitudes of the singly resonant tW diagrams $\mathcal{A}^{tW}$ and the doubly resonant tW NLO diagrams $\mathcal{A}^{t\bar{t}}$. Their combined amplitude is then:

$$\mathcal{A} = \mathcal{A}^{tW} + \mathcal{A}^{t\bar{t}}. \tag{4.11}$$

It follows for the squared amplitude, which is relevant to the calculation of observables:

$$|\mathcal{A}|^2 = \left|\mathcal{A}^{tW}\right|^2 + \left|\mathcal{A}^{t\bar{t}}\right|^2 + 2\Re\left(\mathcal{A}^{tW}\mathcal{A}^{t\bar{t}\dagger}\right) \equiv \mathcal{S} + \mathcal{D} + \mathcal{I}. \tag{4.12}$$

The first scheme to define the tW process is called Diagram Removal (DR). In DR, the doubly resonant diagrams are removed. This is not a theoretically sound procedure; it violates the gauge invariance. In [33] it is shown that this is however not a problem in practice. By using multiple different choices of gauges, they calculate some observables and find that the difference between the gauges is compatible with zero. The modified Eq. (4.12) is then just:

$$\left|\mathcal{A}^{\mathrm{DR}}\right|^2 = \left|\mathcal{A}^{tW}\right|^2 = \mathcal{S}. \tag{4.13}$$

The second scheme proposed is called Diagram Subtraction (DS). In this scheme a term is subtracted at the level of the cross-section:

$$\left|\mathcal{A}^{\mathrm{DS}}\right|^2 = \mathcal{S} + \mathcal{D} - \hat{\mathcal{D}} + \mathcal{I} = \mathcal{S} + \mathcal{I} + \delta. \tag{4.14}$$

$\hat{\mathcal{D}}$ is designed to locally cancel out the contributions from the doubly resonant diagrams. "Locally" refers to the kinematic region in which the virtual t-quark is on-shell. In contrast to DR, it is possible to carry out this subtraction without violating gauge invariance, although it is also this requirement which stops δ from being exactly zero. [33]

The t production and decay are interesting for many different reasons, among them the rare possibility of studying the decay of an unbound quark. In many such analyses, the differences between the DR and DS schemes are sources of significant systematic uncertainties [38]. The goal of this thesis is to study the difference between these two schemes. A better understanding of the effects of the interference term would allow the models describing the t-quark production and decay to improve and thus improve on the uncertainties of their predictions.

4.5 Selection Criteria

At the LHC many different particles are created in a multitude of possible processes that are not the topic of this work. To focus on the processes relevant here, selection criteria must be applied to reduce the effects of the undesired processes as much as possible, while keeping the desired signals. As mentioned in Section 4.3, the dilepton channel of the tW and $t\bar{t}$ processes involves the least jets, therefore only events with exactly 2 leptons are selected. To make sure these leptons are truly the direct product of the W boson decays they are required to each have large transverse momenta $p_{\mathrm{T},l} > 15\,\mathrm{GeV}$ and opposite charges. One of the leptons is also required to have $p_{\mathrm{T},l_1} > 28\,\mathrm{GeV}$. The decay of a Z boson could also produce two such leptons, therefore a veto is triggered and the event is discarded, if their combined invariant mass m_{ll} lies within 15 GeV of the Z mass. All desired events include two b-jets, so another requirement is that there are exactly 2 b-tagged jets, one of which with $p_{\mathrm{T},b_1} > 28\,\mathrm{GeV}$. The working point efficiency (see Table 3.1) of this selection is decided later. Finally, there should be no other jets in the event, so as not to complicate the analysis further. In summary, the selection criteria are:

- exactly 2 leptons with opposite charge and $p_{\mathrm{T},l} > 15\,\mathrm{GeV}$,
- $p_{\mathrm{T},l_1} > 28\,\mathrm{GeV}$,
- Z mass veto for $m_Z - 15\,\mathrm{GeV} < m_{ll} < m_Z + 15\,\mathrm{GeV}$,
- exactly 2 b-tagged jets,
- $p_{\mathrm{T},b_1} > 28\,\mathrm{GeV}$,
- no other jets.

CHAPTER 5

Neural Density Estimation

5.1 Density Estimation in High Dimenions

Estimating a probability density function (pdf), commonly noted as $p(\boldsymbol{x})$ from a set of available samples $\boldsymbol{x}$ is often desirable. In physics, the behaviour of particles is often probabilistic. To model these processes it is necessary to further our understanding of them. A problem in reaching this goal lies in the curse of dimensionality. As an example, imagine a hypercube of D dimensions with side length $a = 1$. Its volume is then $V = a^D = 1$. Let $p(\boldsymbol{x})$ be uniform for the volume within the hypercube. From this it follows that $p(\boldsymbol{x}) = 1$. Assume also that there is only a finite number N of samples drawn from $p(\boldsymbol{x})$ available. Now one might want to find the density around an arbitrary point $\boldsymbol{x}$. For simplicity, this can be a hypercube with side length $\epsilon \leq a$. The expectation value for the number of samples that falls into this ϵ-cube is then:

$$E_\epsilon(\boldsymbol{x}) = \frac{\epsilon^D}{V^D} N = \left(\frac{\epsilon}{V}\right)^D N. \tag{5.1}$$

Since a realistic estimation of the density can only be achieved when there are actually samples in the volume of question, $E_\epsilon(\boldsymbol{x})$ needs to be larger than 1, which is a conservative requirement. Having more samples fall into the ϵ-cube is preferable, because it typically makes the estimation easier. The exact number of samples necessary is dependent on the estimation method used. In Fig. 5.1, the number of samples necessary to fulfil the base requirement of at least one sample in the $epsilon$-cube is shown for some values of ϵ. For small values of D, it might be feasible to create more samples if one wants to estimate the density with a finer granularity. For large D the number of samples needed rises exponentially. If one wants to estimate the density for small ϵ, the required N becomes prohibitively large. In turn, if the sample size is set, this sets a soft limit to the size of ϵ.

In practice, there are many ways to estimate densities. In the following, an estimation technique is presented. This specific method attempts to circumvent the dimensionality issues presented above. Since this technique relies on Neural Networks, a short introduction to the relevant concepts is given.

5.2 Machine Learning and Neural Networks

Neural Networks (NN), as they are used in many different computational fields today, were first created as an analogy to naturally occurring neural systems. Such systems are comprised of many cells called

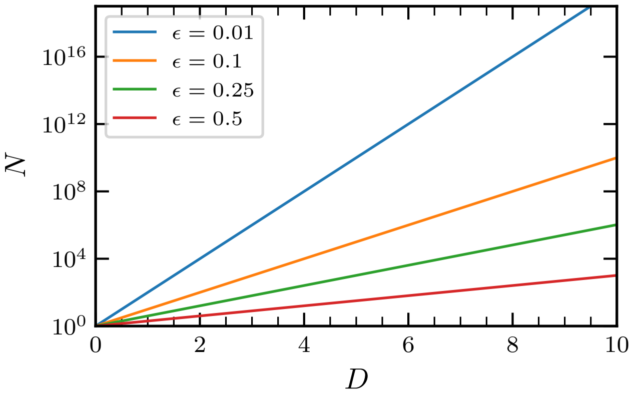


Figure 5.1: The curse of dimensionality shown for the case of subdividing a hypercube into equally large partitions. The requirement here is $E_\epsilon(\boldsymbol{x}) = 1$, so it is expected that any ϵ-cube contains one sample. The number N of samples necessary for this is shown in dependence on the dimension D for multiple values of ϵ. This example is adjusted from the one given in [39] by considering an ϵ-cube instead of an ϵ-sphere.

neurons. Each neuron is capable of receiving input from many others. If the input adds up to a threshold, the neuron activates and sends a signal. In the 1940s, it was found that networks showing this simple structure are, in principle, able to produce any logical expression satisfying certain conditions [40]. Using these findings, it was possible to build the first artificial neural network (NN) in 1958: the Perceptron [41], which was able to learn the ability to discern simple patterns in its input. Since then, many iterations of artificial neural networks have been created and by today these are usually realised as software on general-purpose computers. Due to the manifold of different techniques, network architectures, design philosophies and applications which are summarised under the umbrella term of Neural Network, only relevant and general concepts are described in the following sections.

5.2.1 Feed-Forward Networks

Any NN is comprised of neurons – called nodes – that are connected via edges. In all architectures used in this thesis, these are arranged into subsequent layers. The first layer is called the input layer. Here, the nodes' values are assigned numerical features of an observation $\boldsymbol{x}$; for example, kinematic variables of an event. The last layer is the output layer with values $\boldsymbol{z}$. The layers in between are called hidden layers. The input values are passed on to the nodes of the first hidden layer. The output value of each node in the layers is denoted n_k^l, where the superscript $l \in \{0, \dots, L\}$ is the layer and the subscript $k \in \{0, \dots, K^l\}$ the index of the node. Each node has a set of weights $\boldsymbol{w}_k^l$, one for each edge connecting from the previous layer, and a single bias b_k^l. In conjunction, these are called the parameters $\boldsymbol{\theta}$ of the network. Each input to a node is multiplied by its respecting weight and the bias is added. The result is then evaluated by a function f^l, which need not be the same for every layer in a network. This function is referred to as the activation function because, in the natural neural networks, it decides if a neuron activates and produces an output voltage or not. Collecting the node's values and the biases into column vectors for each layer, $\boldsymbol{n}^l$ and $\boldsymbol{b}^l$, as well as writing the matrix of all weights in a layer $\boldsymbol{W}^l$, whose rows correspond to $(\boldsymbol{w}^l_k)^\mathrm{T}$, it is possible to write the propagation from one layer to the next as:

$$\boldsymbol{n}^l = f^l\left(\boldsymbol{W}^l\boldsymbol{n}^{l-1} + \boldsymbol{b}^l\right), \tag{5.2}$$

where f^l is applied element-wise and $\boldsymbol{n}^{\mathbf{0}} = \boldsymbol{x}$. A NN in which the input information passes through each layer once is called a feed-forward network.

The number of layers, the number of nodes per layer, the activation functions and any other parameters, which define the structure of the NN, but not the weights and biases, are called hyperparameters (HP). They are user-determined at the creation of the NN. An example with 2 inputs, 2 hidden layers and a single output value z is shown in Fig. 5.2.

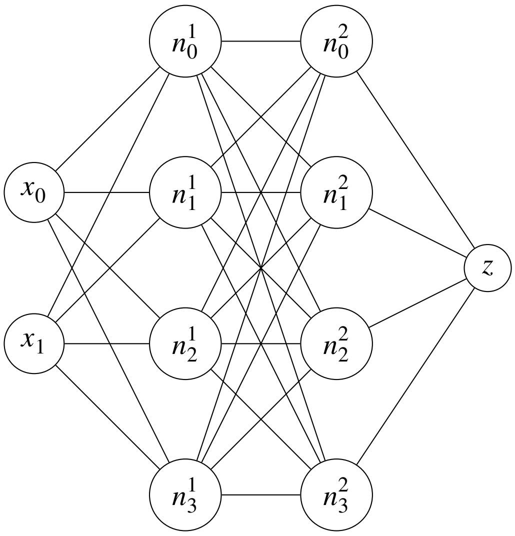


Figure 5.2: Densely connected feed-forward network with 2 input variables and a single output node. Each node calculates its output value according to equation Eq. (5.2).

5.2.2 Backpropagation and Loss Function

Typically, the parameters of a network are initialised in a random state. There exist multiple ways to initialise weights and biases that are thought to be advantageous to the network's behaviour. For this reason, different weight and bias initialisers are available in state of the art machine learning frameworks such as Keras [42] or PyTorch [43]. Even though considering the initialisers carefully is helpful, the network's output in the initialised state is most likely not useful. It is necessary to adjust the parameters systematically for the output to be meaningful. For this purpose, it is necessary to evaluate if a change leads to an improvement. This is done by calculating a loss function $\mathcal{L}(z)$, the choice of which is problem-dependent. By construction, this function is minimised if the network performs its intended task.

As an example, take a binary classifier. This network is meant to distinguish between two types of input, such as pictures of cats and dogs, or signal and background processes in a mixed sample of physics events. For such a goal, it is sensible to build a network with a single output node, whose value is the network's confidence that the input belongs to one of the categories. One could define that an output close to 1 signifies a signal event and closer to 0, a background event. A reasonable choice for a loss function would be the binary cross-entropy

$$\mathcal{L}_{\boldsymbol{\theta}}(z) = -\frac{1}{N}\sum_{j=0}^{N} z^{(j)} \log \hat{z}^{(j)} + \left(1 - z^{(j)}\right) \log\left(1 - \hat{z}^{(j)}\right), \tag{5.3}$$

where $j \in \{0, 1, ..., N\}$ is the index of the input observations, $z^{(j)} \in \{0, 1\}$ is the true label that tells to which class the input belongs, and $\hat{z}_j \in [0, 1]$ is the network's output. $\mathcal{L}_{\boldsymbol{p}}(z)$ is minimal when all observations are correctly mapped and increases when the predictions differ from the labels.

It is now possible to calculate the derivative of the loss function with respect to the model's parameters and derive a gradient $\boldsymbol{g} = \nabla\mathcal{L}_{\boldsymbol{\theta}}(z)$. The gradient points in the direction of the fastest increase in the parameter space. Therefore adjusting the parameters as

$$\boldsymbol{\theta} \to \boldsymbol{\theta} - \lambda\boldsymbol{g}, \tag{5.4}$$

the loss function decreases. Here the learning rate λ is introduced as another hyperparameter, which adjusts the magnitude of the change. This optimisation scheme is called gradient descent [44]. It is a simple and still a widely used approach. There are many other different extensions or changes to this scheme that can be used. Modern machine learning frameworks offer many optimisation schemes with different advantages and disadvantages concerning computational cost and convergence behaviour. These are called optimisers. A discussion of some widely used ones can be found in [44].

One example of a change that is often made to gradient descent is separating the available data into batches. This is useful because it reduces the memory requirements of the training process. The batches can be loaded into the memory one by one instead of all the data at once. The batchsize is also considered a hyperparameter of the NN. This procedure is called mini-batch gradient descent. The learning process is then as follows: a batch is passed through the network, the loss for this batch is evaluated, the parameters are adjusted, the next batch is passed, etc. until all events have passed once. This is called an epoch of training. Typically many such epochs are necessary until the loss converges to a minimum.

In the training scheme described above, the true labels of the input observations need to be known prior to training. This is referred to as supervised training. The network is trained on labelled observations first. If the training is successful, the network has learned to differentiate the features of the different classes. Afterwards, unlabelled observations can be passed through the network in evaluation mode, in which no further adjustment of the parameters is carried out. The network then predicts the labels for those. For example, if there is data available that is thought to come from two different processes, and there are also simulated samples available for these processes, the network can be trained on MC events and afterwards evaluate the data.

5.2.3 Autoencoders

Training a NN when there are no labels available is called unsupervised training. An example of a network architecture which can work like this is the autoencoder (AE), as shown in Fig. 5.3. The defining property of an autoencoder is that it tries to create a reconstruction $\hat{\boldsymbol{x}}$ of its input $\boldsymbol{x}$. The architecture with which this is performed is dependent on the task. A naive implementation of an AE could be built with hidden layers, that are composed of a large number of nodes. If the number of nodes in the smallest layer is larger than the input dimension, this AE is called overcomplete. Such an architecture would allow the AE to simply copy its input to its output, which is not usually desired. There are however still some use cases [45].

Due to this, a more common type of AE is the undercomplete AE, as shown in Fig. 5.3. It is comprised of an encoder and a decoder, between which there is a single bottleneck layer, which is called the code, whose number of nodes is less than the input dimension. Both encoder and decoder can have an arbitrary number of layers. For the encoder, the first of these layers typically have many nodes, while the later ones have fewer. For the decoder the inverse is true. [46]

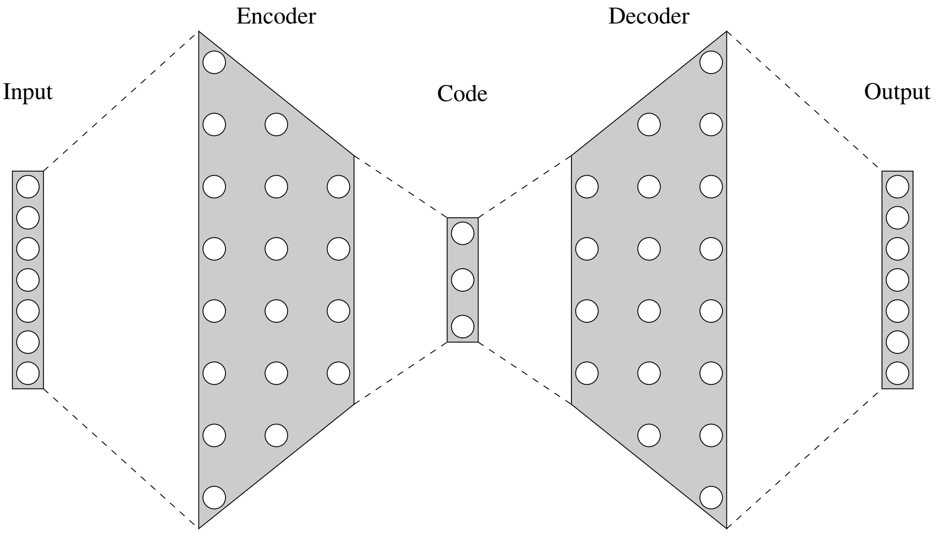


Figure 5.3: A typical autoencoder as it could be used for dimensionality reduction. Connections between nodes are not shown but implied. Input data is passed through the encoder into a bottleneck layer, called the code, and then through an encoder. After successful training, the output is supposed to be an estimation of the input. If this succeeds, the code is a low dimensional representation of the relevant input information.

Assume this undercomplete AE manages to create an output that is similar to its input. That means the network was able to recreate, from the few nodes in its code the relevant information present in its input. The information contained in the input is then represented in the code in a lower dimension and the parts of the input that are irrelevant, such as noise, are filtered out. This would not be the case if an overcomplete AE would be used. By constraining the network through the bottleneck, it is forced to learn an approximate, but more general, representation of its input. [46]

5.2.4 Generalisation of Neural Network Models

Neural networks are generally trained on samples that represent an underlying distribution. It is important to evaluate if the network is learning something that is general to the underlying distributions and not just learning specific features of the input sample. This is referred to as overtraining. Some measures need to be taken to evaluate if overtraining is an issue with a specific model. One such technique is the separation of the input samples into a training and a test subset. The model is trained only on the training subset. Then the test set is passed through the network in evaluation mode. If the results of the two sets are similar, the network has learned generalisable features of its input. If there is a large discrepancy, that means the network learned something that is specific to the training set only.

There are multiple ways to avoid overtraining. For one, it mostly happens when the amount of training samples is low compared to the complexity of the model. This means that either more input samples are needed, or the size of the network, e.g. the number of layers and/or nodes, needs to be decreased.[1]There is also the possibility of training for fewer epochs. Evaluating the test loss after each epoch can help to identify when overtraining starts and the training can be interrupted at that point. Another common method is to add a penalty term to the loss function used by the optimiser to calculate the gradient. Two common ways to calculate such a penalty are called L1 and L2 regularisation. L1 adds the absolute values of the network's weights and L2 takes the sum of the squared values. This encourages a network to prefer smaller weights when possible. All of these techniques to avoid overtraining (except for a larger input sample size) usually result in worse performance, therefore they are only employed when necessary.

5.3 Masked Autoencoder for Density Estimation (MADE)

As portrayed in Section 5.1, a major challenge for density estimation comes from the curse of dimensionality. MADE is a neural density technique that aims to resolve this issue [48]. To understand how it works, first, consider the chain rule of probability:

$$p(\boldsymbol{x}) = \prod_{d=1}^{D} p(x_d|\boldsymbol{x}_{<d}), \tag{5.5}$$

where $\boldsymbol{x}_{<d} = \left[x_1, \dots x_{d-1}\right]$. The goal of estimating the joint density $p(\boldsymbol{x})$ can thus be broken down into estimating the marginal densities on the right hand side of Eq. (5.5), and then simply multiplying them afterwards. Models attempting to do so are called autoregressive models. Now, the concept of the AE from Section 5.2.3 needs to be adjusted so that the goal of its training is no longer to output $\hat{x}_d$, but the marginal densities $p(x_d|\boldsymbol{x}_{<d})$.

For an AE to do so, its loss function needs to become minimal exactly when the likelihood $p(\boldsymbol{\theta}|\boldsymbol{x})$ becomes maximal. This is referred to as a maximum likelihood estimation [49]. For convenience, the logarithm of the likelihood is used[2]. This then leads to a natural choice for the loss function for MADE:

$$\mathcal{L}_{\boldsymbol{\theta}} = -\sum_{j=1}^{N} \log p(\boldsymbol{\theta}|\boldsymbol{x}^{(j)}) = -\sum_{j=1}^{N}\sum_{d=1}^{D} \log p(x_d^{(j)}|\boldsymbol{x}_{<d}^{(j)}). \tag{5.6}$$

When the samples used are MC events as explained in Section 3.3, their weights[3] $g^{(j)}$ can be accounted

[1] In some publications, non-generalisable learning due to a large network size is called overthinking and treated differently from overtraining [47]. This distinction is not made here.

[2] This is permissible because the logarithm is a monotonous function, meaning that $f(x)$ is maximal at the same x as $\log f(x)$. It is convenient because working with sums instead of products is often easier and the numbers appearing in models are not as large, which can easily lead to overflows in some applications – including the one presented here.

[3] These are entirely unrelated to neural network weights.

for by weighting the contributions to the loss function accordingly:

$$\mathcal{L}_{\boldsymbol{\theta}} = -\sum_{j=1}^{N} g^{(j)} \log p(\boldsymbol{\theta}|\boldsymbol{x}^{(j)}) = -\sum_{j=1}^{N} g^{(j)} \sum_{d=1}^{D} \log p(x_d^{(j)}|\boldsymbol{x}_{<d}^{(j)}). \tag{5.7}$$

In principle, taking the negative is not necessary but most optimisers obey the convention to minimise their argument. Each output dimension is only dependent on the input dimensions that come before it. This is referred to as the autoregressive property. [48]

To make sure the AE fulfils this property, it is necessary to remove any unwanted computational paths. This is achieved by multiplying the weight matrices $\boldsymbol{W}^l$ elementwise with masks $\boldsymbol{M}^l$, so that Eq. (5.2) reads:

$$\boldsymbol{n}^l = f^l\left(\left(\boldsymbol{M}^l \odot \boldsymbol{W}^l\right)\boldsymbol{n}^{l-1} + \boldsymbol{b}^l\right). \tag{5.8}$$

The elements of the masks $\boldsymbol{M}^l$ are binary. They can cancel out the undesirable paths by effectively setting the weights of these connections to zero while leaving the others unaffected. To find appropriate masks, each node is assigned an integer m_k^l at random, which expresses the maximum number of input connections allowed to that node. For the input and output layer $m_k \in \{1, \ldots, D\}$ and for all hidden layers $m_k \in \{1, \ldots, D-1\}$. For all layers, except the output, the masks are then constructed as:

$$\boldsymbol{M}^l_{k',k} = \begin{cases} 1 & \text{if } m^l_{k'} \geq m^{l-1}_k \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

For the output layer, the only difference is that the inequality becomes open bounded instead of close bounded. An example of a network masked in this way is shown in Fig. 5.4. Masking also takes care of the overcompleteness issue mentioned in Section 5.2.3 without having to implement a bottleneck layer at all. In the example in Fig. 5.4, it can be seen that the inputs cannot be copied to the output, since there is no computational path available to do so. An issue with a naive implementation of the masking algorithm can be seen in the output layer. For the given example no path at all is left for the output node meant to learn $p(x_2)$. This results in limited ability to learn this element of the output. To remedy this, new masks can be created for every batch of training. The authors of [48] refer to this as connectivity agnostic training. The output of MADE is also dependent on the order of the inputs, which is chosen at random. For this reason, it is also suggested to permute their ordering after each batch and then average the different results. This is referred to as order-agnostic training.

As an extension to MADE, it is possible to include further variables $\boldsymbol{m}$ in the input of the network whose density is not attempted to be learned. By doing this, it is possible to increase the information available to the network without increasing the complexity of the density to be learned. This can be achieved by connecting all nodes of the first hidden layer to additional inputs, which are not masked.

5.4 Normalising Flows

A second type of network that attempts to learn densities from samples are normalising flows (NF) [50]. The idea is based on transforming the samples with unknown density to a simple known target density via a change of variables. Usually, a standard normal distribution is chosen as the target, but other

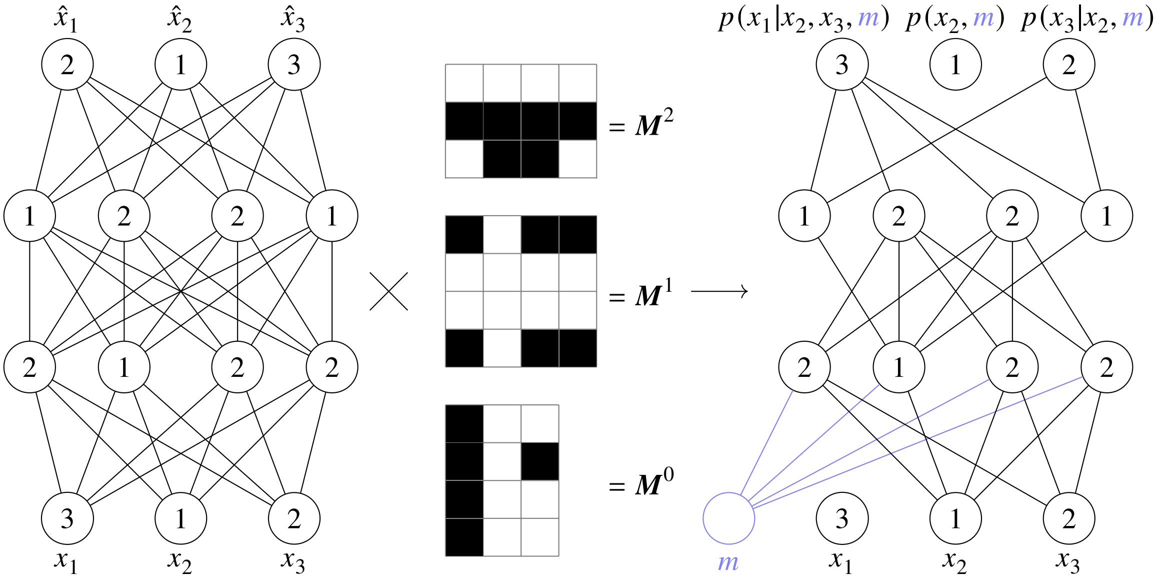


Figure 5.4: Masking scheme to transform an AE to MADE. All nodes of an AE are assigned an integer which is a count of the number of inputs the nodes are allowed to connect to. The Masks $\boldsymbol{M}^l$ are derived from this and multiplied elementwise to the weights of the layers $\boldsymbol{W}^l$. Black indicates a masked connection. On the right, potential conditional inputs, m are also indicated. Figure adapted from [48].

distributions are equally valid. A transformation of densities is described by:

$$p_X(\boldsymbol{x}) = p_Z(z) \left|\det \boldsymbol{J}\right|^{-1}, \tag{5.10}$$

where z and $\boldsymbol{x}$ are realisations of the random variables Z and X, describing the target density and the sought after density of the input respectively. $\boldsymbol{J}$ is the Jacobian matrix of the transformation $Z = f(X)$, defined as:

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_D}{\partial x_1} & \cdots & \frac{\partial z_D}{\partial x_D} \end{bmatrix}. \tag{5.11}$$

Eq. (5.10) can be used to determine the sought after density of the samples. This leads to the requirement that $Y = f(X)$ needs to be an invertible function. In practice, it is difficult to find a function that can fullfill Eq. (5.10), but multiple simple transformations can be combined, each of them incrementally transforming the distribution. This is easily possible because the composition of invertible functions is also invertible. Furthermore, the Jacobian determinant can simply be multiplied, or when using logarithmic probabilities, summed. This chain is called a normalising flow:

$$\log p_X(\boldsymbol{x}) = \log p_Z(z) \sum_{t=1}^{T} \left|\det \boldsymbol{J}^{(t)}\right|^{-1}, \tag{5.12}$$

where $t \in \{1, \ldots, T\}$ is the index of the transformation. Under the assumption that the map transforms to the target density accurately, it is now possible to find the probability of an observation $\boldsymbol{x}$. The probability of its corresponding z is evaluated on the target density, which can then be translated to $p(\boldsymbol{x})$ via Eq. (5.12). In Fig. 5.5 the principle is shown graphically.

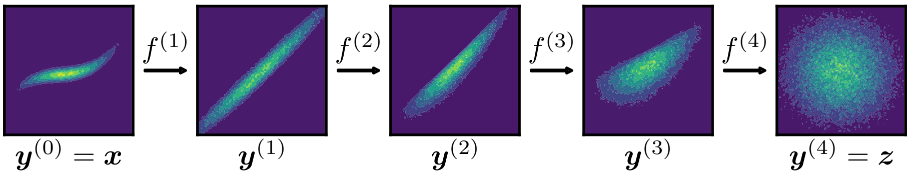


Figure 5.5: A series of transformations f_t is applied to some input data (left). The transformations are chosen, so that the output (right) is as close to a standard normal distribution as possible.

5.5 Masked Autoregressive Flows (MAF)

The major challenge in implementing a normalising flow lies in the optimisation of the transformations so that the result is a normal distribution. Another practical issue is the Jacobi determinant in Eq. (5.12) which is computationally expensive for large D.[4] Masked autoregressive flows [53] aim to solve both issues. The approach is to substantially reduce the computational cost by utilising variable transformations with a triangular Jacobian matrix, so that:

$$\log(\det(\boldsymbol{J})) = \operatorname{tr}(\boldsymbol{J}) \tag{5.13}$$

This condition leads to the requirement, that the transformations must have the shape:

$$y_d^{(t)} = f^{(t)}\left(\boldsymbol{x}_{<d}^{(t-1)}\right). \tag{5.14}$$

Here, $y_d^{(t)}$ is the d-th output of the transformation $f^{(t)}$ with index t, and $\boldsymbol{y}_{<d}^{(t-1)}$ is the subset of input variables with dimension index $<d$. This is similar to the approach of MADE, and can be achieved by applying the same masking algorithm as in Section 5.3 to a fully connected overcomplete autoencoder. The difference to pure MADE is that MAF does not attempt to map $x_d \rightarrow p(x_d|\boldsymbol{x}_{<d})$, but $x_d = y^{(0)} \rightarrow y_d^{(1)} \rightarrow \ldots \rightarrow y^{(T)} = z_d$, by using T MADE blocks in succession. Each single block transforms its inputs by applying a scale and a shift to it:

$$y_d^{(t)} = y_d^{(t-1)} \exp \alpha_d + \mu_d, \text{ with } \alpha_d = \alpha_d\left(\boldsymbol{y}_{<d}^{(t-1)}\right) \text{ and } \mu_d = \mu_d\left(\boldsymbol{y}_{<d}^{(t-1)}\right). \tag{5.15}$$

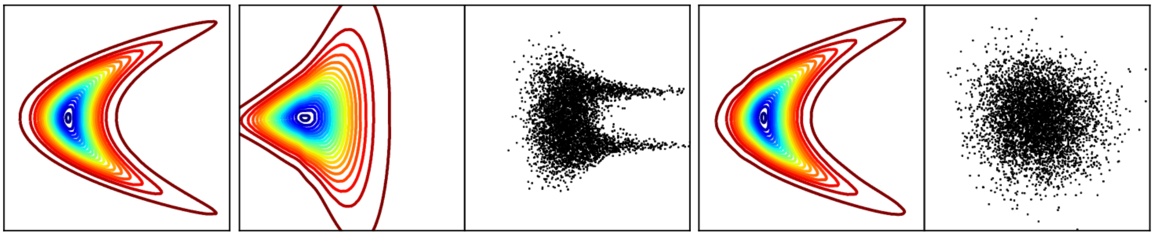
[4] Calculating a determinant scales the same as matrix multiplication [51], for which advanced algorithms can achieve a scaling of $D^{2.373}$, while most commonly used implementations scale as D^3 [52].

The network learns the functions α_d and μ_d. To achieve this, the output nodes are split into two nodes, one for each function. For the entire transformation chain, this results in:

$$\log p(\boldsymbol{x}) = -\log p(z) \sum_{t=1}^{T} \sum_{d=0}^{D} \alpha_d^{(t)}, \tag{5.16}$$

which is computationally cheap. The optimisation of the α and μ parameters is achieved with a maximum likelihood method as in MADE.

If an additional input m is included that is not included in the estimation of the density, as is suggested in Fig. 5.4, the probabilities will become conditional on these, e.g. $p(\boldsymbol{x}) \longrightarrow p(\boldsymbol{x}|m)$.



(a) Target density **(b)** Standalone MADE **(c)** MAF with $T = 5$

Figure 5.6: (**a**) The sought after density is defined as $p(x_1, x_2) = x_2 x_1$ with $x_2 = \mathcal{N}(0, 4)$ and $x_1 = \mathcal{N}(\frac{1}{4}x_2^2, 1)$. (**b**) The density learned by a single MADE Block with order (x_1, x_2), as well as the scatter plot of the supposedly normal distribution the samples are mapped to. The transformation does not succeed, therefore MADE fails. (**c**) A MAF with $T = 5$ can normalise the same input. [53]

An advantage of MAF – compared to MADE – is that it can estimate densities that have a natural ordering, such as the one displayed in Fig. 5.6. MADE has the aforementioned issue that its results are dependent on the order of its inputs. A MAF can circumvent this issue by chaining multiple transformations, each with a different input ordering. Therefore its results are order-independent, if the number T of transformations is sufficient, with no averaging over the different orders necessary. MAFs have been proven to be universal estimators, meaning that this class of transformations can learn any continuous target density to an arbitrary precision [54]. In practice, this is limited by the number of samples available for training. An example comparing MAF and MADE is shown in Fig. 5.6. A 2-dimensional target density is shown. This density is constructed so that one dimension is distributed as a standard normal distribution and the other is a normal distribution with parameters that depend on the first dimension. If the natural order of this density is not the order with which the input of MADE is supplied, MADE fails to learn the target density. However, MAF with 5 chained transformations can estimate the target density well.

CHAPTER 6

Anomaly Detection with Density Estimation (ANODE)

A common problem in many fields of quantitative analysis is that of identifying anomalies in a set of data. In this section *Anomaly Detection with Density Estimation* (ANODE) [2] is presented and applied to investigate the difference between the DR and DS schemes presented in Chapter 4.

6.1 The Method

The goal of ANODE is to create a metric which can be used to identify anomalies. This is realised as a likelihood ratio between two conditional probability density functions. The available samples are separated into two regions, called the Signal Region (SR) and Sideband (SB). The cut performing this separation is done on a variable m, which is known to have a localised anomaly. The SR is then chosen to contain most, or all, of the anomaly, while the SB should contain as little of the anomaly as possible.

To find these pdfs, the masked autoregressive flows introduced in Section 5.5 are used, although, in principle, ANODE could work with any other density estimation technique as well. The SR and SB densities $p^{\mathrm{SR}}(\boldsymbol{x}|m)$ and $p^{\mathrm{SB}}(\boldsymbol{x}|m)$ are learned by separate neural networks, which implement the flows including a conditional input m. This is achieved by training them with the samples $\boldsymbol{x}, m \in \mathrm{SR}$ and $\boldsymbol{x}, m \in \mathrm{SB}$ respectively. The variables $\boldsymbol{x}$, called the discriminating variables, are chosen to be sensitive to the anomaly. It is not necessary to know how they depend on the anomaly, just that they do, which makes ANODE a mostly model-agnostic method. Assume now that the density is estimated well, and that the choice of SR is reasonable, meaning that the anomaly is restricted to the SR, with only small contamination into the SB. The density in the SR is then the sum of the background present in the SR and that of the anomaly (ignoring normalisation):

$$p^{\mathrm{SR}}(\boldsymbol{x}|m \in \mathrm{SR}) = p^{\mathrm{SR}}_{\mathrm{ano}}(\boldsymbol{x}|m \in \mathrm{SR}) + p^{\mathrm{SR}}_{\mathrm{bkg}}(\boldsymbol{x}|m \in \mathrm{SR}). \tag{6.1}$$

The SB on the other hand should not contain any anomaly:

$$p^{\mathrm{SB}}(\boldsymbol{x}|m \in \mathrm{SB}) = p^{\mathrm{SB}}_{\mathrm{bkg}}(\boldsymbol{x}|m \in \mathrm{SB}). \tag{6.2}$$

Next, it is necessary to interpolate the SB into the SR. At this point, two MAFs are available that are capable of estimating the densities. To interpolate the SB into the SR, it is sufficient to use the SR events as the input to the model trained on the SR events: $p^{\mathrm{SB}}(\boldsymbol{x}|m \in \mathrm{SR})$. This is an advantage of density estimation utilising MAFs and would be a more complicated step with other methods. The interpolation into the SR will not sculpt any feature that is present in the SR but not in the SB. Therefore the two densities will be different in presence of an anomaly. To quantify this difference, the ratio of the two is taken:

$$R(\boldsymbol{x}|m) = \frac{p^{\mathrm{SR}}(\boldsymbol{x}|m)}{p^{\mathrm{SB}}(\boldsymbol{x}|m)} = \frac{p^{\mathrm{SR}}_{\mathrm{ano}}(\boldsymbol{x}|m) + p^{\mathrm{SR}}_{\mathrm{bkg}}(\boldsymbol{x}|m)}{p^{\mathrm{SB}}_{\mathrm{bkg}}(\boldsymbol{x}|m)}. \tag{6.3}$$

$R(\boldsymbol{x}|m)$ can be interpreted as an anomaly score, which is ≈ 1 when an event in the SR behaves as implied by the SB, and is > 1 if there is an overdensity in the SR. A cut like $R > R_{\mathrm{c}}$ can then increase the signal to background ratio, which makes it easier to further investigate the anomaly. In [2] it is shown that such a cut, if it is made at the right value of R_c, can increase the signal to background ratio from 0.6 % to around 200 %.

6.2 Variable and Region Selection

ANODE is now applied to MC samples based on the schemes DR and DS for the definition of the tW process introduced in Section 4.4.1. In this context, the differences between the two schemes are considered the anomaly. These are the interference and the error term $\mathcal{I} + \delta$ from Eq. (4.14).

6.2.1 Conditional Variable Selection

The variable m, on which the SR is defined and which will be used as the conditional input in the MAFs, needs to be chosen first. It needs to be locally sensitive to the interference. Such a variable is already known from previous analyses [26, 33, 55]:

$$m_{bl}^{\mathrm{minimax}} = \min\left[\max\left(m_{b_1 l_1}, m_{b_2 l_2}\right), \max\left(m_{b_1 l_2}, m_{b_2 l_1}\right)\right]. \tag{6.4}$$

$m_{b_i l_j}$ is the invariant mass of the system of the i-th b-jet and the j-th lepton. The indices indicate the order of the energy of the objects. The higher energy object is labelled with 1 and the lower energy object with 2.

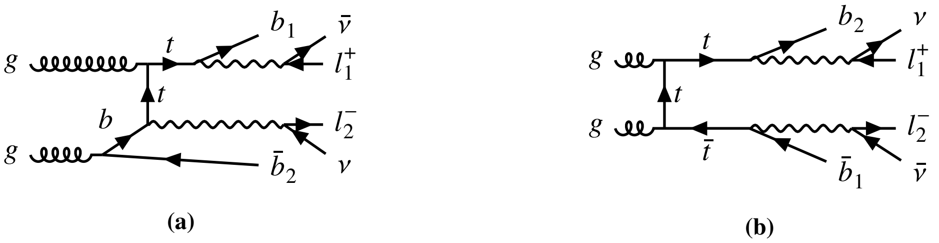


Figure 6.1: **(a)** Singly resonant t-quark production. The indices are correctly paired to correspond to originate from the same t-quark decay. **(b)** Doubly resonant t-quark production and subsequent decays. The indices are not correctly paired.

$m_{bl}^{\mathrm{minimax}}$ has been specifically constructed to be sensitive to the interference between singly and doubly resonant amplitudes. Consider the diagrams in Fig. 6.1. Assume that the internal t-quarks are on-shell, meaning they fulfil the energy-momentum relation. If the indexing of the b-quarks and the leptons is correct, their combined invariant masses $m_{bl\nu}$ would correspond to that of the t-quarks. Since the neutrino cannot be measured, $m_{b,l}$ is used as the next best alternative, and has the upper bound $m_{bl} < m_t$. This means that any $m_{bl} > m_t$ are the result of off-shell t-quarks. Since the cross-section for off-shell production, events in that region are rare. From this, it follows that singly resonant contributions are singly suppressed in this region, while doubly resonant contributions are doubly suppressed. In turn, this means that the interference between the two is enhanced compared to the background of doubly resonant contributions.

In the above, it was assumed that the indices of the objects correspond to the origin from a common t-quark decay. The correct indexing cannot be guaranteed in practice, thus $m_{b_i l_i}$ cannot be used to create a region enriched with interference. There are two possibilities for pairing up the invariant masses. They are:

$$\mathrm{A} = \{m_{b_1 l_1}, m_{b_2 l_2}\} \text{ and} \tag{6.5}$$

$$\mathrm{B} = \{m_{b_1 l_2}, m_{b_2 l_1}\}. \tag{6.6}$$

It is guaranteed that one of these is the correct pairing. If A is the true one, then both its terms must be smaller than m_t. From this it follows that $\max(\mathrm{A}) < m_t$. If B is the correct pairing, then for $\max(\mathrm{B}) < m_t$ holds true. These are the inner terms in Eq. (6.4). Since one of the pairs is guaranteed to be below m_t, the minimum of the pairs is as well. It follows that m_{bl}^{minimax}, as defined in equation Eq. (6.4), is a variable in which doubly resonant contributions are strongly suppressed above m_t [55, 56]. It is known that the neutrino energies are missing from these considerations. Since the maximal missing invariant mass due to the neutrinos is that of the W boson (assuming it to be on-shell), the region that is considered to be especially susceptible to the interference effects is found to be [38]:

$$m_{bl}^{\text{minimax}} > \sqrt{m_t^2 - m_W^2} = 153\,\text{GeV}. \tag{6.7}$$

m_{bl}^{minimax} is now considered a candidate for the definition of the SR. It is displayed for both the DR and DS Monte Carlo samples in Fig. 6.2. The two samples are distributed similarly below $m_{bl}^{\text{minimax}} = 153\,\text{GeV}$ but differ substantially above this threshold. The DS samples, which include the interference effects, contain fewer events in the SR than the DR samples. This means that the interference is destructive in this kinematic region. Therefore DS has an underdensity in the SR when compared to DR.

	DR	DS
SR	12 230	6 736
SB	20 918	24 658

Table 6.1: The number of available events in the DS and DR schemes, separated into SR and SB at $m_{bl}^{\text{minimax}} = 153\,\text{GeV}$

Some issues lie in using this variable to define the regions. While the argument given above makes it clear that the SR is indeed enriched in interference effects, it does not claim that the interference is limited to this region. If the interference is present in the SB as well, Eq. (6.2) does not hold anymore. Ideally, a further adjustment to m_{bl}^{minimax} would be made to suppress the interference in the SB, but no

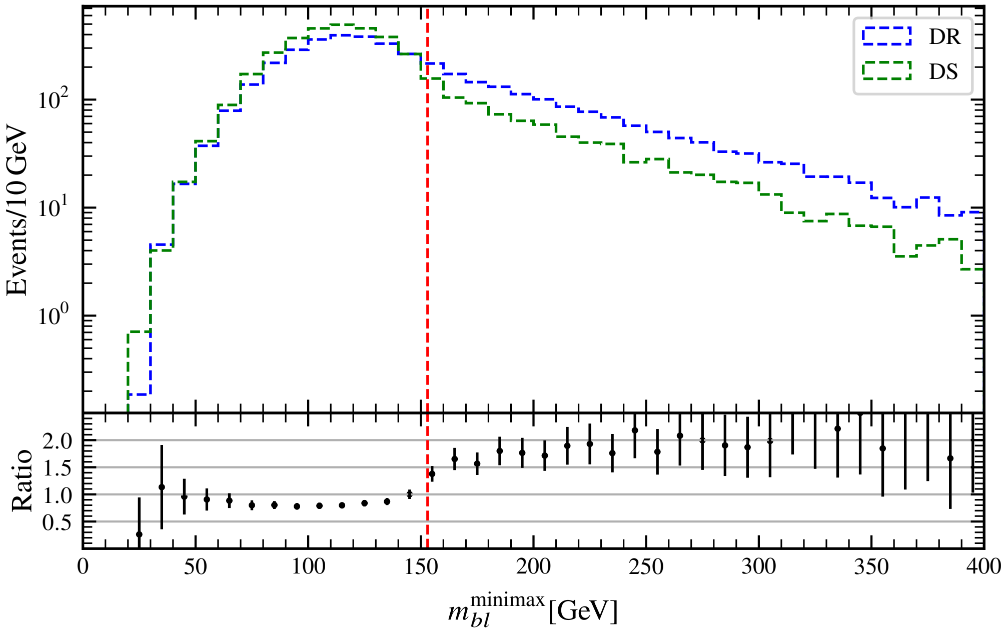


Figure 6.2: The two definition schemes DR and DS are compared in the variable m_{bl}^{minimax} as defined in Eq. (6.4). The SR candidate cutoff is indicated in red. The bottom plot shows the ratio between the bins. Events for both schemes are normalised to 1 pb^{-1}.

such adjustment could be found. Therefore m_{bl}^{minimax} is used as the region defining variable anyway.

6.2.2 Choice of Training Samples

Next, it needs to be decided which samples are used to train and evaluate the MAFs that form the basis of ANODE. Since only the DS samples contain the interference, these will be used for that purpose. The number of available events in the DR samples is however too low to train a MAF for exact density estimation. The number of available events in the two regions of the two sample sets is shown in Table 6.1. In experiments with MAFs[1], the number of events necessary to produce reasonable results was found to be on the order of 10^5, which is much more than is available for DR or DS.

For this reason, some concessions regarding the training samples need to be made. The selection described in Section 4.5 can be adjusted to include a less restrictive b-tagging working point. This was attempted, but it does not increase the available statistics sufficiently to offset the lower purity of the sample. As a solution, the DR and DS sets are combined with the $t\bar{t}$ samples. This reduces the signal efficiency in the SR substantially, however, ANODE is quoted to still work at low signal to background ratios by its authors.

While the addition of $t\bar{t}$ to the training set increases the statistics in the SB, this is not the case in the SR, because $t\bar{t}$ is doubly resonant and the SR was specifically chosen to not contain many doubly resonant events. This is remedied by choosing a less restrictive choice for the SR. ANODE requires the

[1] These experiments were carried out using the LHC Olympics 2020 Dataset [57]. This is the same set which was used by the authors of ANODE to show its capabilities in anomaly detection.

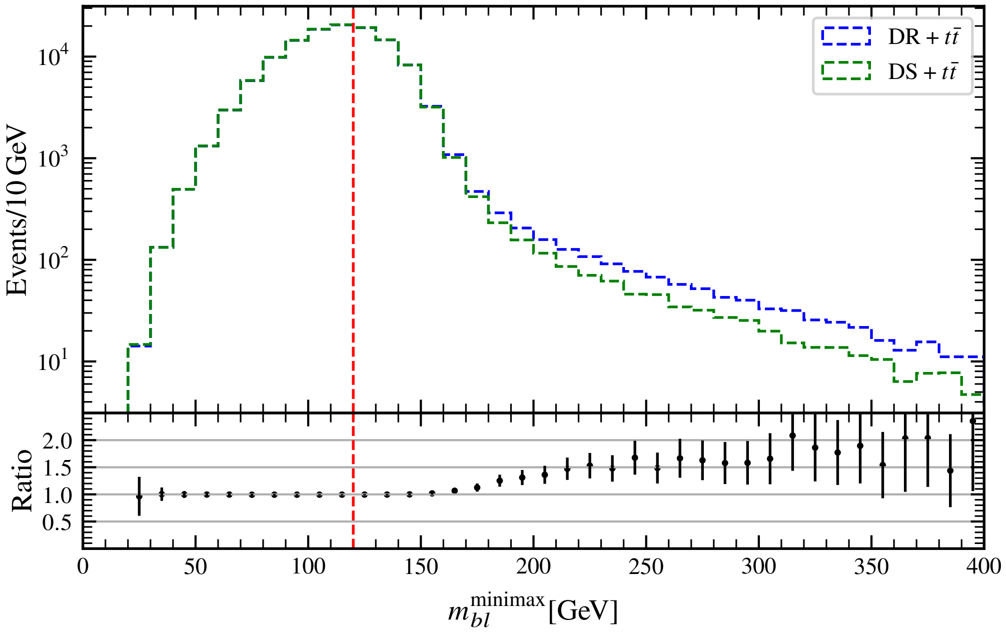


Figure 6.3: The two definition schemes DR and DS are compared in the variable $m_{bl}^{\mathrm{minimax}}$ as defined in Eq. (6.4). The SR candidate cutoff is indicated in red. The bottom plot shows the ratio between the bins. Events for all schemes are normalised to 1 pb^{-1}.

SR to include the anomaly, but it is not necessary that the SR is chosen to fit around this anomaly tightly. Because of this, and because it is unclear if the interference is constrained to the tight SR in the first place, setting the SR cutoff at a lower $m_{bl}^{\mathrm{minimax}}$ is plausible. A new and looser SR cutoff value is chosen at 120 GeV. In Fig. 6.3 the same comparison as made in Fig. 6.2 is made, but this time with the added $t\bar{t}$ events. Table 6.2 shows the number of available samples is for $t\bar{t}$ + DS/DR for the loose SR choice. The training is carried out with the DS+$t\bar{t}$ samples. It is possible to afterwards evaluate the model on the different sets individually and under the exclusion of the $t\bar{t}$ samples.

	DR	DS	DS + $t\bar{t}$
SR	1.0×10^4	8.5×10^3	6.6×10^5
SB	8.0×10^3	9.6×10^3	1.0×10^6

Table 6.2: The number of available events in the DS and DR schemes, separated into SR and SB at the cutoff $m_{bl}^{\mathrm{minimax}} = 120\,\mathrm{GeV}$

6.2.3 Discriminating Variables

Now that the SR is defined, the discriminating variables $\boldsymbol{x}$ need to be chosen. These are the variables whose density $p(\boldsymbol{x})$ is learned by the conditional MAFs. They need to be chosen in accordance with some conditions. Firstly they need to actually be discriminating, meaning that $p^{\mathrm{SR}}(\boldsymbol{x})$ must be different than $p^{\mathrm{SB}}(\boldsymbol{x})$. This is a general requirement of ANODE. In the context of studying the interference, it means that the variables $\boldsymbol{x}$ must be sensitive to it. Since only the DS samples contain the interference, this condition can be expressed as:

$$\frac{p_{\mathrm{DS}}^{\mathrm{SR}}(\boldsymbol{x})}{p_{\mathrm{DS}}^{\mathrm{SB}}(\boldsymbol{x})} \neq 1. \tag{6.8}$$

This condition is necessary but not sufficient because the differences are not guaranteed to be caused by the interference. It is possible that the pdfs are different in the SR and SB due to other effects. Therefore variables which behave differently in SR and SB due to effects other than the interference must be removed from the considerations. This is done by comparing the interference-free DR samples in the same way, but requiring them to be similar, thus introducing a second condition:

$$\frac{p_{\mathrm{DR}}^{\mathrm{SR}}(\boldsymbol{x})}{p_{\mathrm{DR}}^{\mathrm{SB}}(\boldsymbol{x})} \approx 1. \tag{6.9}$$

This condition makes sure the variables are not sensitive to any effects except the interference. Lastly, since $t\bar{t}$ samples are included in the training, the same requirement applies to them:

$$\frac{p_{t\bar{t}}^{\mathrm{SR}}(\boldsymbol{x})}{p_{t\bar{t}}^{\mathrm{SB}}(\boldsymbol{x})} \approx 1. \tag{6.10}$$

The selection of the variables must predate their actual usage in the density estimation. Therefore the densities in the conditions in Eqs. (6.8) to (6.10) are compared using normalised histograms. This is sufficient because the conditions only need to be fulfilled approximately. A two-sided two-sample Kolmogorov–Smirnov (KS) test is used to judge their similarity [58]. The KS test yields a p-value $\in [0, 1]$. For similar distributions, this value is large, while it is small for distributions which are significantly different from each other. Since cases in which the third condition is violated can still be amended by not using the $t\bar{t}$ events in the final evaluation of ANODE, it is given a lower priority than the other two when considering variables for their viability.

No variables fulfilling the second condition to a satisfactory degree were found or could be constructed. There were however some variables found in which it was approximately fulfilled in an interval of a variable. These regions were then selected by making the cuts:

$$-45\,\mathrm{GeV} < \Delta p_{\mathrm{T}}(l_1 l_2 j_1; E_{\mathrm{T}}^{\mathrm{miss}}) < 40\,\mathrm{GeV} \text{ and}$$
$$E_{\mathrm{T}}^{\mathrm{miss}} < 95\,\mathrm{GeV}.$$

The quantity $\Delta p_{\mathrm{T}}(l_1 l_2 j_1; E_{\mathrm{T}}^{\mathrm{miss}})$ is the difference between the transverse momenta between the systems separated by the semicolon. Making these cuts, the variables they are made on are satisfying the second condition by construction. They also lead to another variable fulfilling the conditions. The final choices for the discriminating variables are:

$$\begin{aligned}&E_{\mathrm{T}}^{\mathrm{miss}}\\&\Delta p_{\mathrm{T}}(l_1 l_2 j_1; E_{\mathrm{T}}^{\mathrm{miss}})\\&\left|\Delta\phi\left(l_1; l_2\right)\right|\end{aligned}$$

The variable $\left|\Delta\phi\left(l_1; l_2\right)\right|$ describes the absolute difference between the azimuthal angles of the two leptons. Initially, the difference was used without taking the absolute, but the MAFs were showing difficulties with the double peak distribution of the variable. The discriminating variables are presented in Fig. 6.4. The first condition is reasonably well satisfied. The p-values indicate that the distributions are significantly different from each other. The second condition is also reasonably well met. The third is however only somewhat fulfilled for $\Delta p_{\mathrm{T}}(l_1 l_2 j_1; E_{\mathrm{T}}^{\mathrm{miss}})$

6.3 Application to Single Top-Quark Production

6.3.1 Preprocessing

Even though the optimisation of the density estimation is carried out by the conditional MAFs, it is necessary to make some adjustments to the input variables introduced in the previous section. This preparation can be understood as a first transformation in the flow that serves to make it easier for the NN part of the flow to learn the desired probabilities.

It is known that neural networks perform better when the input they get is scaled roughly to the interval on which the employed activation functions have a changing gradient [59]. The inputs should be scaled to a similar interval for all variables, otherwise, the gradient is not guaranteed to point in the direction of the minimum in the loss function. The authors of ANODE suggest an additional procedure to transform the discriminating variables $\boldsymbol{x}$ to be distributed more gaussian-like before they are entered into the NN, while only applying a linear scale factor to the conditional input m [2].

As a first step in the preparation, each dimension of $\boldsymbol{x}$ is independently scaled to the interval $[0, 1]$:

$$x \longrightarrow \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{6.11}$$

At this point in the preprocessing, a fiducial volume is defined that spans the interval $[0.05, 0.95]$. The events are used for training whether they fall into this region or not, but [2] suggests not to use events outside of the fiducial volume in the evaluation steps of the model. The tails of the distribution are likely modelled with a large error, since density estimation in general works poorly in regions with few samples. Thus, by not considering the events outside of this volume, the accuracy of the density estimation is thought to increase. Next, a logit transformation is carried out:

$$x \longrightarrow \log\left(\frac{x}{1-x}\right). \tag{6.12}$$

The logit changes the shape of the distributions in a way which is desirable for normalising flows. Events which are initially close to 0 or 1 are stretched to a large interval, while the distribution is nearly linearly scaled around 0.5. This has multiple benefits. For one, this gets rid of hard to model discontinuous edges in the input features that appear due to the cuts made in Section 6.2.3. Secondly, the logit also reduces

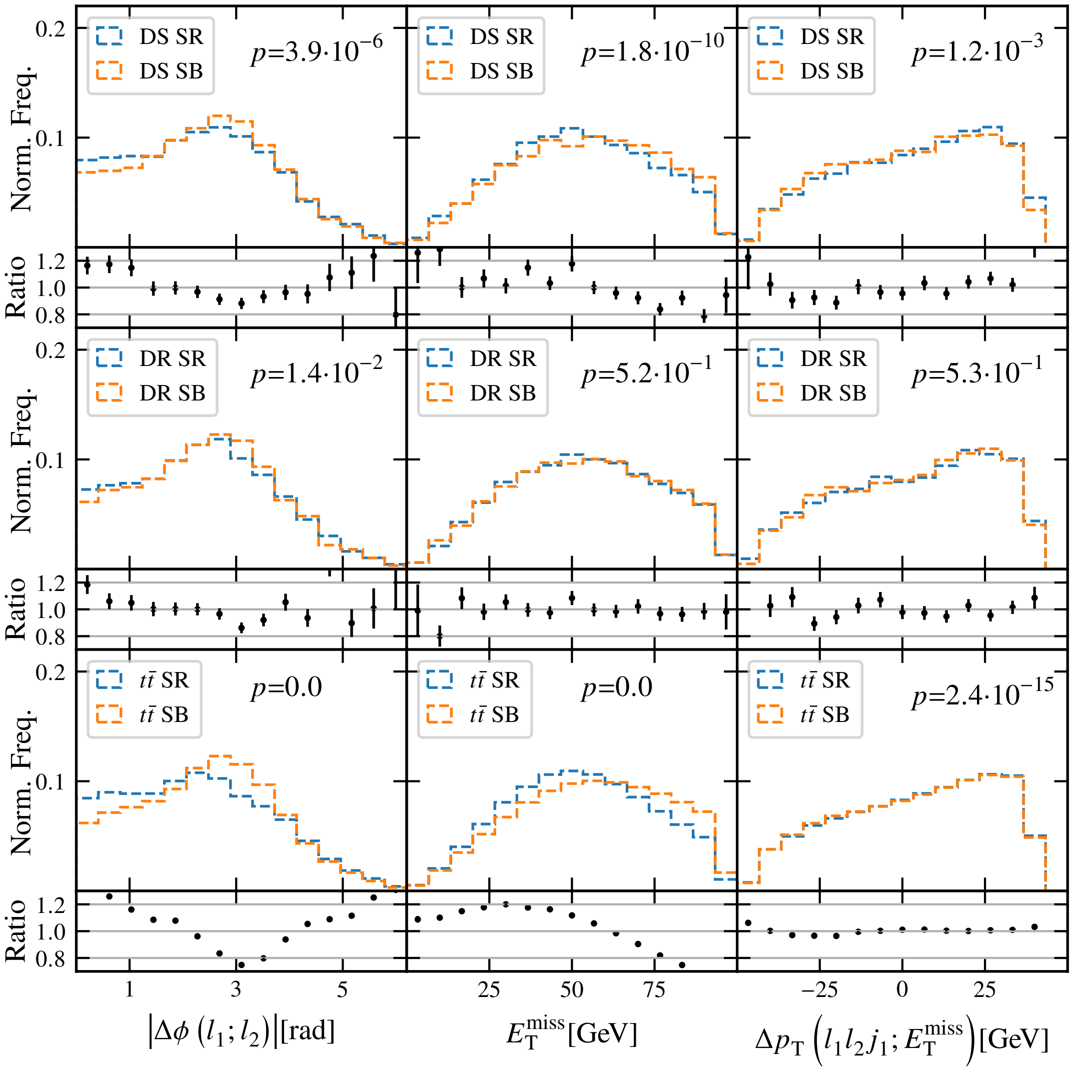


Figure 6.4: The choice for the discriminating variables is presented for the SR and SB for the DR, DS and $t\bar{t}$ samples. The sets are compared according to the conditions set on the discriminating variables in Eqs. (6.8) to (6.10). The p-values are calculated by a two-sided two-sample KS test implemented in scipy [58].

the skewness of the distributions, thus already taking a step towards a normal distribution. Finally, the last transformation shifts the distributions to 0 mean and a standard deviation of 1:

$$x \longrightarrow \frac{x - \mu_x}{\sigma_x}. \tag{6.13}$$

Due to these transformations changing the pdfs of the inputs, they need to be accounted for in the total Jacobian matrix when evaluating the flows. The transformations are performed on the SR and SB separately. The preprocessed discriminating variables of the training set are shown in Fig. 6.5.

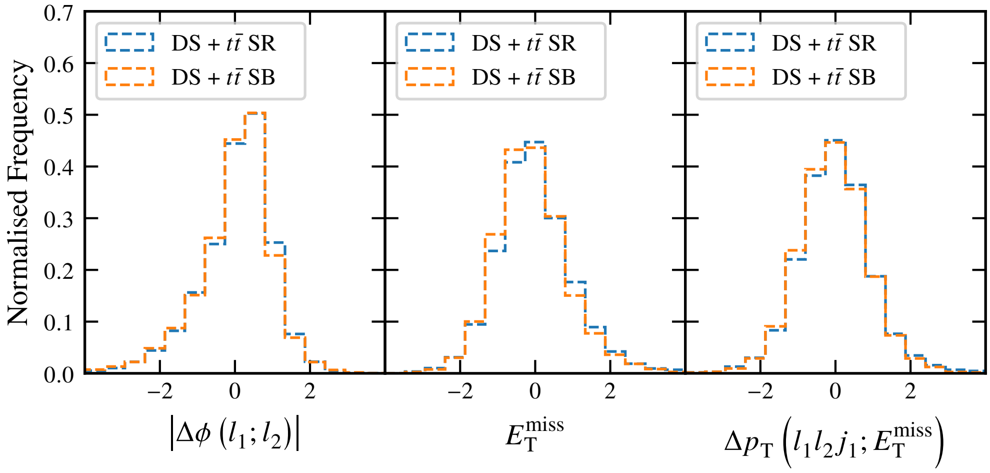


Figure 6.5: The discriminating variables after the preprocessing are shown for the training set consisting of DR and $t\bar{t}$ samples.

6.3.2 Training the MAFs

The preprocessed discriminating variables and the conditional input m are now used as the input for the MAFs described in Section 5.5. The code executing the flows is implemented in Python and utilises the PyTorch package for machine learning [43]. It is largely based on the code used to demonstrate a proof of concept in the original ANODE publication that has been modified to fit the purpose presented here [2]. The code used in this thesis is available on Github.[2] It has been thoroughly tested to verify the integrity of the results. The results in the original publication were found to be fully reproducible with this implementation.

Network Performance Metrics

It is necessary to optimise the hyperparameters of the two MAFs. To do so it needs to be known if the created models are made "better" or "worse" by a given change in the HPs. This is only possible if there are metrics available to judge the model performance. The first choice for such a metric is the loss, however, it was found to be almost identical for a large set of different HPs, limiting its use as a

[2] https://github.com/lukavom/myANODE

performance metric. It is still useful to look at, to gauge if overtraining is a problem. Furthermore, the loss function is used to decide when to stop the training.

Another possibility is exploiting the knowledge of the desired output of a normalising flow. A normalising flow is exactly then successful when its output is a normal distribution. Initially, the marginal output distributions were compared to a univariate standard normal distribution utilizing the same 2 sided KS test that was also used in Section 6.2.3 to compare the discriminating variables. While this has some merit, the method was discarded because it is not sufficient to determine the normality of the marginal distributions. Additionally, the KS test produces a performance metric for each dimension of the input. This means that a change in an HP can cause the output in one dimension to improve while degrading in it another, making it difficult to decide if the change of the HP is an improvement.

A test for multivariate normality is necessary. The Henze-Zirkler (HZ) test is chosen for this purpose [60]. This test's available implementation in Python [61] has some drawbacks. It is computationally expensive to carry out this test. Using the entire available $\approx 10^6$ samples in the subsets was found to be unfeasible in regards to computational time and memory requirements. The test was therefore only carried out on a sample of 10^4 of the normalised events. A further problem is that the test output p-values behave unstably. Repeating the test twice, but with a different sample of the same NN output, the p-value changes by factors of up to 10. It is unclear if this is because of the small subset of events used. Since no other viable test could be found, the HZ test is used to measure the normalisation performance of the network, with the knowledge that differences in p-values need to be large to signal a significant improvement.

6.3.3 Hyperparameter Optimisation

The hyperparameter optimisation for the NNs in this thesis was carried out in the following way. Initially, a network with arbitrary HPs is trained and evaluated based on the considerations given above. A single HP is then changed by a small increment, and the new model is trained. If the results improve, the same HP is again incremented in the same direction. This is repeated until the network no longer improves. The optimal state of this parameter is then restored and the same procedure is repeated for another HP. Because the HPs are not independent of each other, this is an iterative procedure and needs to be repeated multiple times for each considered HP. This method is not guaranteed to find a global minimum in the HP space. The procedure is carried out independently for the SR and the SB, but they are found to behave similarly concerning their optimal HPs.

Instead of trying to optimise the number of epochs in the above fashion, a patience algorithm was used. A record is kept of the epoch in which the model's test loss was lowest. If the network has not improved in a number of epochs specified by the patience value, e.g. the model has “run out of patience”, the model's state is reset to the epoch of optimal loss. The final HPs that are obtained in this fashion are given in Table 6.3.

The loss curve of the training with these parameters is shown in Fig. 6.6. Since the curves for the train and test subsets are not diverging, no overtraining is implied by the loss. In Fig. 6.7 the normalised marginal distributions are shown. It can be seen that the normalisation procedure is successful. The p-values gained by performing the HZ test are given in Table 6.4. To gain some context on the meaning of these p-values, the same test was carried out with the model and data that was used in the ANODE publication, where the results derived from the normalised distributions are verifiable and found to be useful. The p-values derived from this verification are also shown in Table 6.4. In the SR the normalisation is better by multiple orders of magnitude, while in the SB the performance is comparable

	SR	SB
Transformations	23	28
Hidden Layers	5	5
Hidden Nodes	100	100
Batchsize	15 000	15 000
Learning Rate	5×10^{-3}	5×10^{-3}
L2 Regularisation	1×10^{-7}	1×10^{-7}
Epochs	39	224
Patience	120	120
Optimiser	adam	adam
Weight Init.	orthogonal	orthogonal
Bias Init.	0	0
Test Fraction	0.25	0.2

Table 6.3: The hyperparameters used to produce the results are presented. The number of hidden layers used in each transformation step is equal, and each hidden layer has the same number of nodes.

	DS + $t\bar{t}$		LHCO	
	train	test	train	test
SR	3.1×10^{-17}	6.2×10^{-5}	1.3×10^{-33}	7.8×10^{-24}
SB	4.7×10^{-1}	4.5×10^{-3}	1.4×10^{-1}	2.2×10^{-1}

Table 6.4: The p-values from a HZ-test for multivariate normality with a subsample of 10^4 events are listed. Both the p-values achieved with the MAFs optimised for this thesis and the ones which are obtained for the models used in [2] are given.

when taking into account the unstable behaviour of the HZ-test. From this comparison, the conclusion that the normalisation is precise to a useful degree can be made.

6.4 Results

After training the MAFs, two NN models are available, one of which was trained with the SR events and one with the SB events of the combined DS + $t\bar{t}$ samples. It is now possible to evaluate the density of all events with both of these models. The interpolation from the SB into the SR is achieved by evaluating the SR events via the SB model. In turn, the SR density can also be extrapolated into the SB. The anomaly score $R(\boldsymbol{x}|m)$ from Eq. (6.3) is thus calculated for all available events. This includes events from the DR set of samples, which were not used in the training. The events that are not found to be anomalous are mapped to $R \approx 1$, while events in which the SR and SB densities differ are mapped to larger or smaller values, depending on the anomaly being an overdensity or an underdensity. If the selection of discriminating variables in Section 6.2.3 was successful in eliminating all anomalies except for the interference between singly and doubly resonant contributions to the cross-section, an underdensity in DS compared to DR can be expected. This would correspond to DS having a larger portion of its events

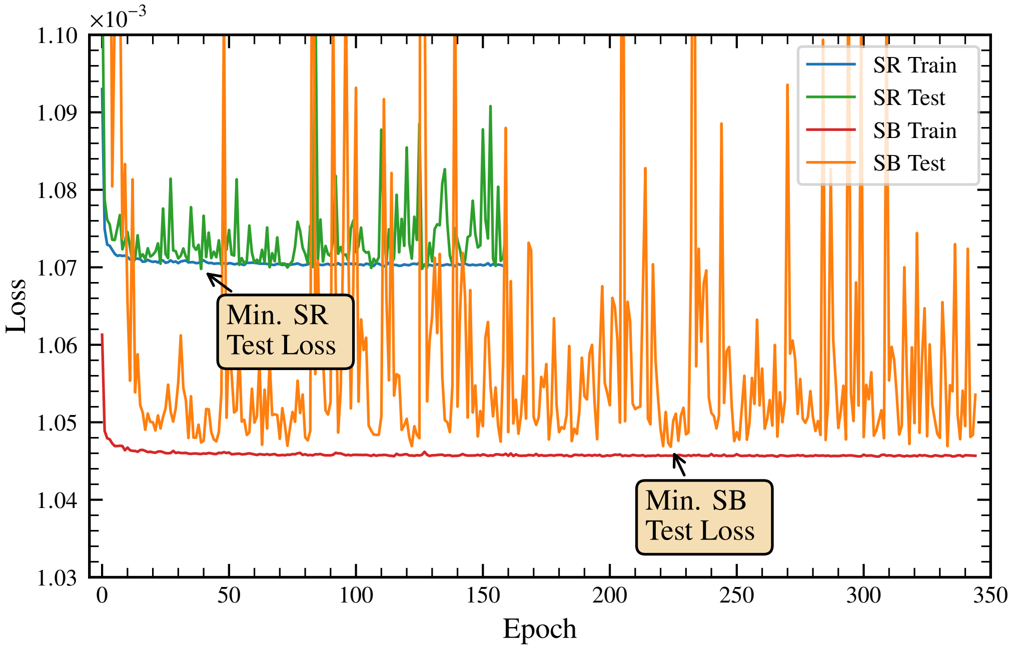


Figure 6.6: The loss curves are shown for the separate SR and SB training and both the train and test subsets respectively.

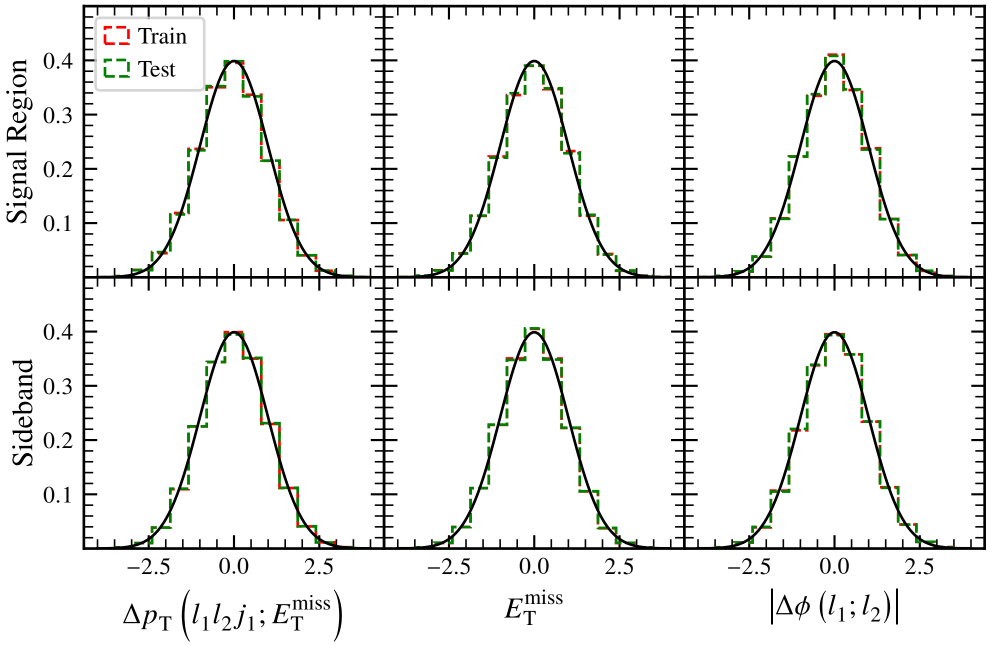


Figure 6.7: The marginal distributions after normalisation through the MAFs are compared to a standard normal distribution. The train and test subsets are displayed separately. The corresponding p-values are listed in Table 6.4.

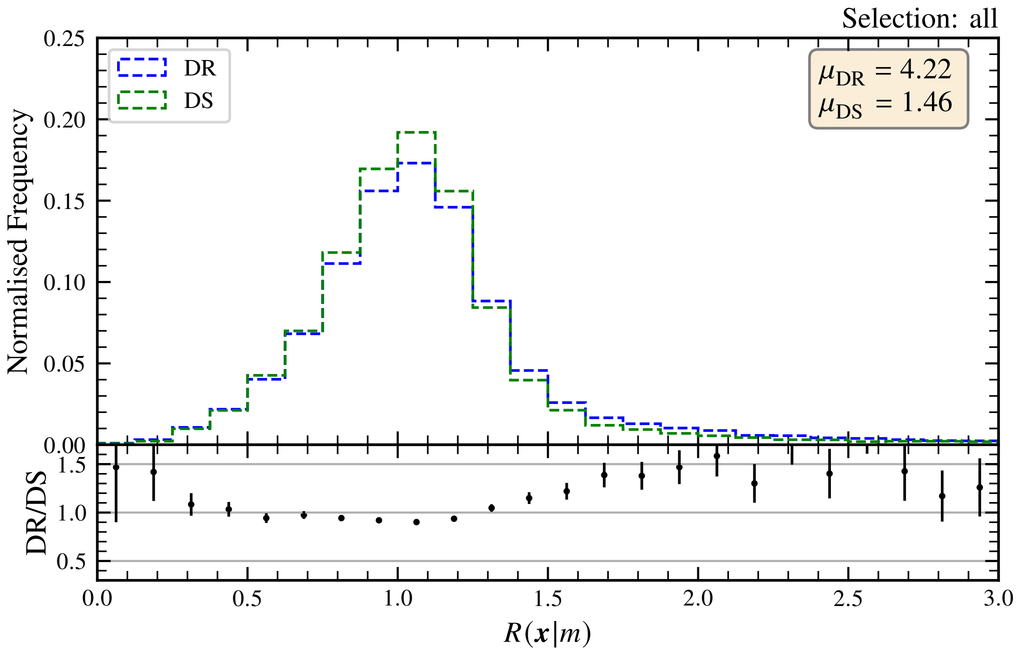


Figure 6.8: The anomaly score $R(\boldsymbol{x}|m)$ is compared for the DR and DS samples. The histogram shown includes all available events with no selection made.

at $R < 1$.

Fig. 6.8 shows a histogram of the anomaly score for all events. In the region $R < 1$ DS is more prominent than DR, indicating that the interference could be enhanced by a cut $R < R_c$. The weighted averages μ of the distribution are given as well. μ_{DS} is significantly lower than μ_{DR}, however, both values are dominated by outliers at large R, reducing their meaningfulness.

The outliers can be suppressed with the fiducial cut motivated in Section 6.3.1, which removes the events in regions in which the density is likely estimated poorly. Additionally, it is known that the interference effect is enhanced for the tight SR with $m_{bl}^{\mathrm{minimax}} > 153\,\mathrm{GeV}$, so this selection is combined with the fiducial cut. This leaves 4 405 and 2 554 samples in the DR and DS sets respectively. The result is presented in Fig. 6.9. Even though the selection should in principle have increased the ratio of DR/DS for $R < 1$ by removing more of the background, this is not the case. There is no significant difference between the two schemes visible in the R distribution. This is not the same as proof of the failure of the model to discern the interference effects, however, it seems unlikely that a cut based on these distributions is reasonable.

As a further cross-check for the normalisation procedure Fig. 6.10 shows the R distribution for the SB with $m_{bl}^{\mathrm{minimax}} < 153\,\mathrm{GeV}$ and the fiducial cut applied. Nearly all events are found around $R = 1$, with only a small tail towards larger R, as is expected when the interpolation of the SB into the SR is successful.

Another way to investigate how a potential cut on R could affect the contribution of the interference is presented in Fig. 6.11. 2D histograms of R vs. $m_{bl}^{\mathrm{minimax}}$ are shown, once without any cuts and once with events from the fiducial volume only. In these diagrams, it can be seen that the fiducial cut removes some of the fringes in the distribution. DR and DS are similarly distributed in these diagrams. To

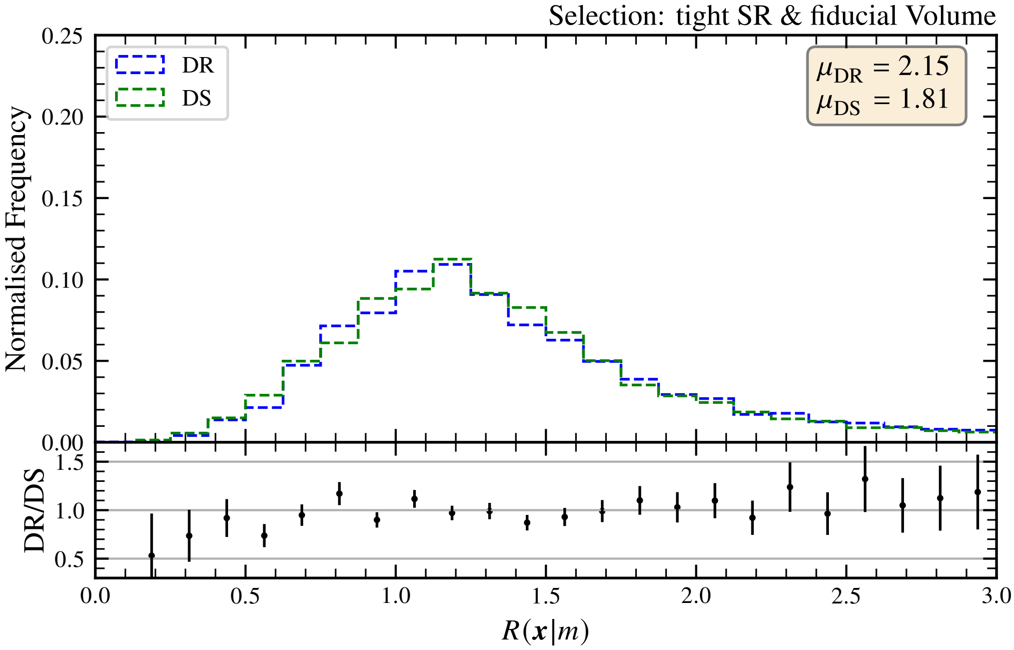


Figure 6.9: The anomaly score $R(\boldsymbol{x}|m)$ is compared for the DR and DS samples. Only events within the fiducial volume with $m_{bl}^{\mathrm{minimax}} > 153\,\mathrm{GeV}$ are selected.

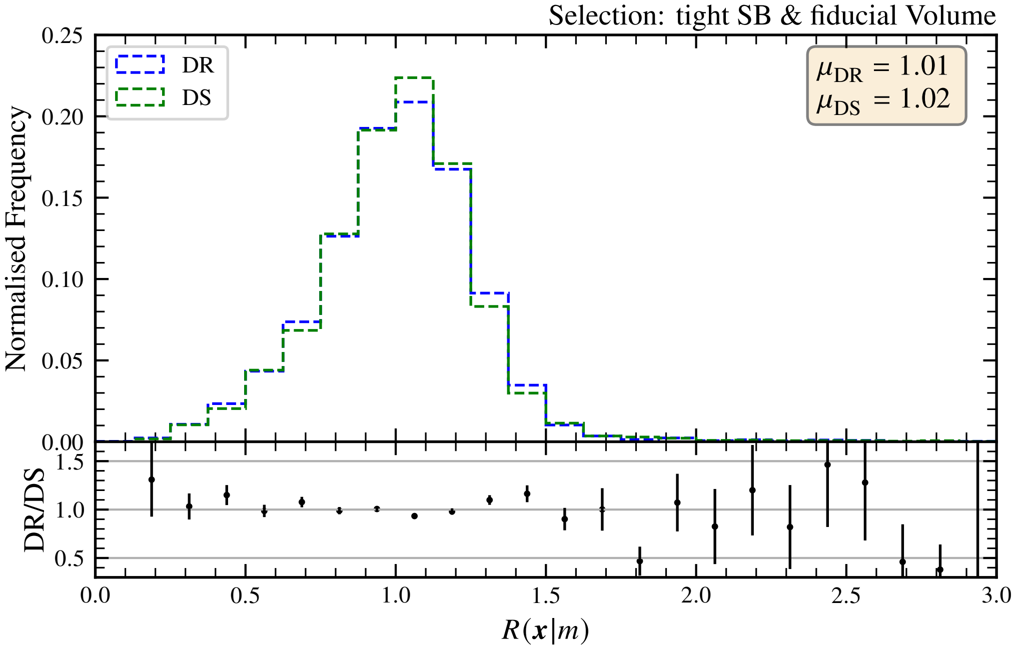


Figure 6.10: The anomaly score $R(\boldsymbol{x}|m)$ is compared for the DR and DS samples. Only events within the fiducial volume with $m_{bl}^{\mathrm{minimax}} < 153\,\mathrm{GeV}$ are selected.

better visualise the differences, the ratio between the bin counts is also shown. This ratio is smaller than 1 for large values of m_{bl}^{minimax}, and larger elsewhere, which is in agreement with Fig. 6.2. However, no obvious candidate for a cut on R, or a 2D cut can be determined from this. An interesting feature visible in Fig. 6.11 is the absence of a tail towards large R around $m_{bl}^{\text{minimax}} \approx 80 - 170\,\text{GeV}$. This is the region in which m_{bl}^{minimax} has its peak. It could be assumed that in this region of high density the density estimation was especially precise, thus reducing the spread away from $R = 1$. However, if this were the case the same would be true for low values of R and the source of this feature remains unknown.

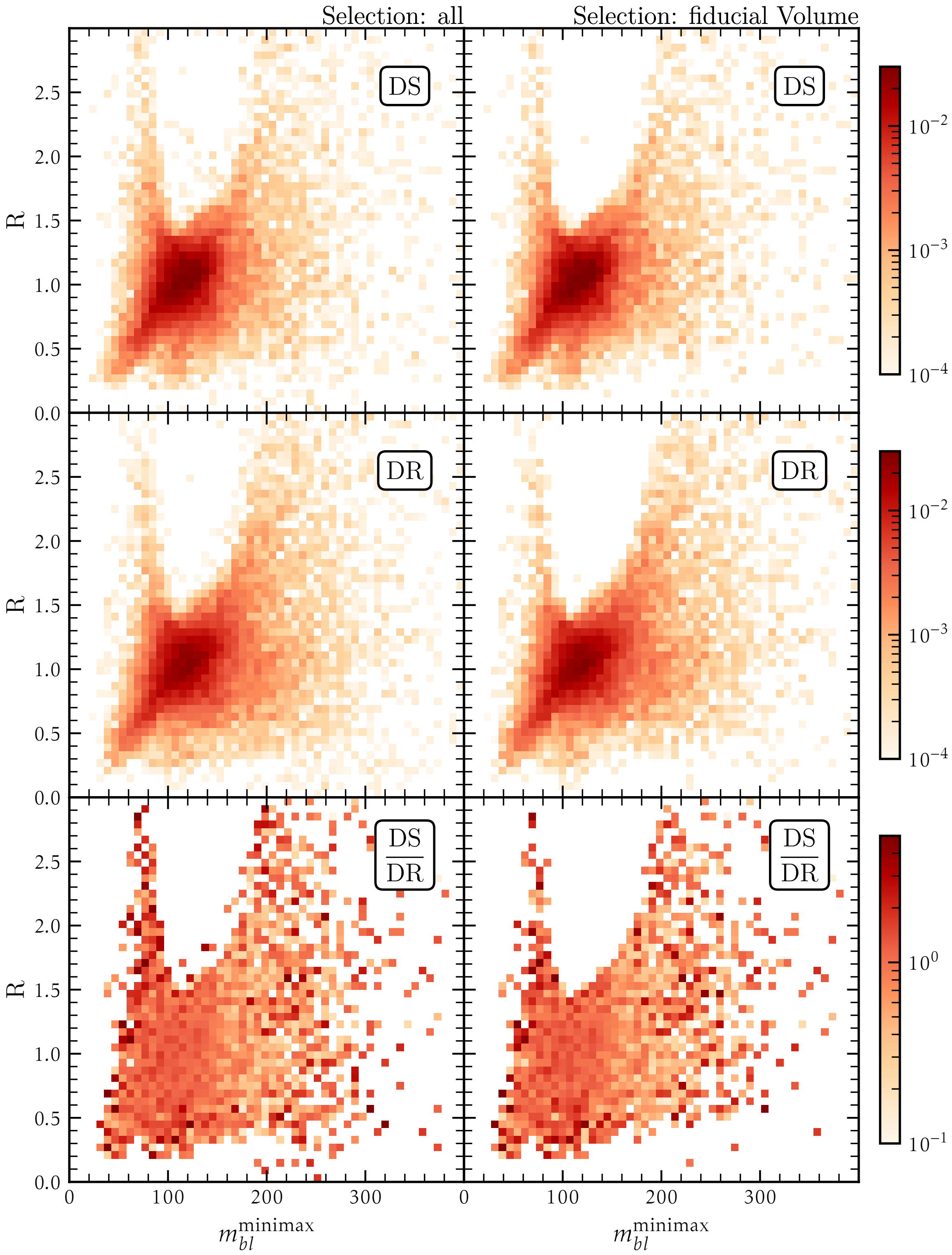


Figure 6.11: The anomaly score R vs. m_{bl}^{minimax} is shown in separate 2D Histograms for the DR and DS sample sets and their ratio. **On the left:** No selection cut is applied and all available samples are used. **On the right:** Only events within the fiducial volume are considered.

CHAPTER 7

Summary and Conclusion

Single t-quark production can serve as a valuable probe of the standard model due to its large mass. The specific interest in this thesis falls on the tW production channel. Next-to-Leading-Order corrections to this channel lead to interference effects with the $t\bar{t}$ production channel, requiring a careful redefinition of the tW process. Two proposals for such a definition are called DS and DR, with the primary difference between the two being the inclusion or exclusion of the interference effects into tW. Studying the difference between the two schemes is nontrivial because it is not possible to create MC samples which include the information on which events are affected by the interference. An anomaly detection method (ANODE), which was derived to increase the significance of a bump hunt, is introduced. An attempt is made at treating the interference effect in the available DS samples as an anomaly and use ANODE to find the kinematic regions which are especially sensitive to the interference.

The core of ANODE lies in estimating probability densities in arbitrary dimensions utilising masked autoregressive flows. These flows have been implemented and tested. The ability of the MAFs to normalise input, and thus lead to an accurate density estimate has been demonstrated. The MAFs are optimised via neural networks. The optimisation of their hyperparameters was carried out as a grid search.

To utilise ANODE, it is necessary to find a variable on which the interference is localised to define a signal region. The choice of this variable fell on m_{bl}^{minimax}, in which background contributions are suppressed in a known interval, which was then chosen as the SR. Variables with sensitivity to the interference, but with little sensitivity to other effects, were chosen for the density estimation in the signal region and the corresponding sideband. ANODE then assigns an anomaly score to each event, which is a likelihood ratio between SR and SB density estimates. The possibility of cutting this score to separate the interference effect has been investigated, however, no reasonable value for such a cut could be derived. It remains unknown if a cut on the anomaly score would yield samples enriched with the interference effect.

In conclusion, ANODE is believed to have potential in application to the issue of finding the differences between DR and DS, but more work would need to be done to achieve results that are usable for this purpose. Using MAFs for density estimation is working well, therefore the subpar results are likely caused by the choice of SR on the m_{bl}^{minimax} variable. While it can be shown that the SR $m_{bl}^{\text{minimax}} > 153\,\text{GeV}$ contains only a small contribution from doubly resonant contributions, and thus this interval is disproportionally more affected by the interference, the interference is by no means limited to this interval. It likely affects the SB regions as well, causing the SB density to include the interference as well. In turn, this means that the preconditions for successful utilisation of ANODE are not given with the choice of SR. If a variable could be constructed of which it is known that the interference effect is truly restricted to a limited interval, the same analysis could easily be repeated with the now existing tools. However, no such variable is known to me until now. A secondary issue is also the limited sample size of the MC simulations available. The required addition of $t\bar{t}$ samples for the purpose of training the MAFs might have had unforeseen effects. It could also have reduced the relative magnitude of the interference effects to below detectable levels.

Bibliography

[1] M. Dosanjh, “From Particle Physics to Medical Applications”, *From Particle Physics to Medical Applications*, 2399-2891, IOP Publishing, 2017, ISBN: 978-0-7503-1444-2, URL: https://dx.doi.org/10.1088/978-0-7503-1444-2ch1 (cit. on p. 1).

[2] B. Nachman and D. Shih, *Anomaly detection with density estimation*, Physical Review D **101** (2020), ISSN: 2470-0010 (cit. on pp. 1, 32, 33, 38, 40, 42).

[3] M. Diehl and W. Hollik, “The Standard Model: Our Picture of the Microcosm”, *Physics at the Terascale*, John Wiley Sons, Ltd, 2011 23, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527634965.ch2 (cit. on p. 2).

[4] G. Münster, *Von der Quantenfeldtheorie zum Standardmodell: Eine Einführung in die Teilchenphysik*, De Gruyter, 2019, ISBN: 9783110638547, URL: https://doi.org/10.1515/9783110638547 (cit. on p. 2).

[5] MissMJ and Cush, *Standard Model of Elementary Particles*, 2021, URL: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cit. on p. 3).

[6] P. Zyla et al., *Review of Particle Physics*, PTEP **2020** (2020) 083C01, and 2021 update (cit. on pp. 3, 4, 17, 18).

[7] CDF Collaboration, *High-precision measurement of the W boson mass with the CDF II detector*, Science **376** (2022) 170, URL: https://www.science.org/doi/abs/10.1126/science.abk1781 (cit. on p. 3).

[8] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** (1961) 579, ISSN: 0029-5582, URL: https://www.sciencedirect.com/science/article/pii/0029558261904692 (cit. on p. 5).

[9] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519** (1968) 367 (cit. on p. 5).

[10] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (21 1967) 1264, URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264 (cit. on p. 5).

[11] M. Gonzalez-Garcia and M. Maltoni, *Phenomenology with massive neutrinos*, Physics Reports **460** (2008) 1, URL: https://doi.org/10.1016%2Fj.physrep.2007.12.004 (cit. on p. 5).

[12] *Planck Publications*,
URL: https://www.cosmos.esa.int/web/planck/publications (visited on 12/05/2022) (cit. on p. 5).

[13] E. Mobs, *The CERN accelerator complex - 2019. Complexe des accélérateurs du CERN - 2019*, (2019), General Photo, URL: https://cds.cern.ch/record/2684277 (cit. on p. 7).

[14] W. Herr and B. Muratori, *Concept of luminosity*, (2006),
URL: http://cds.cern.ch/record/941318 (cit. on p. 8).

[15] ATLAS Collaboration,
Luminosity determination in pp collisions at $\sqrt{s}$ = 13 TeV using the ATLAS detector at the LHC, tech. rep., CERN, 2019, URL: http://cds.cern.ch/record/2677054 (cit. on p. 8).

[16] C. D. Roberts, *On mass and matter*, AAPPS Bulletin **31** (2021),
URL: https://doi.org/10.1007/s43673-021-00005-4 (cit. on p. 8).

[17] A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, *Parton distributions for the LHC*, The European Physical Journal C **63** (2009) 189,
URL: https://doi.org/10.1140%2Fepjc%2Fs10052-009-1072-5 (cit. on p. 9).

[18] D. Dominguez, K. P. Moles and S. Mehlhase, *ATLAS detector schematics*, (2021),
URL: https://cds.cern.ch/record/2777214 (cit. on p. 10).

[19] J. Pequenao and P. Schaffner,
"How ATLAS detects particles: diagram of particle paths in the detector", 2013,
URL: https://cds.cern.ch/record/1505342 (cit. on p. 11).

[20] The ATLAS Collaboration,
Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics, 2009,
URL: https://arxiv.org/abs/0901.0512 (cit. on p. 12).

[21] H. Pernegger, *The Pixel Detector of the ATLAS experiment for LHC Run-2*,
Journal of Instrumentation **10** (2015) C06012,
URL: https://doi.org/10.1088/1748-0221/10/06/c06012 (cit. on p. 12).

[22] A. Vogel,
ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates, tech. rep., CERN, 2013, URL: https://cds.cern.ch/record/1537991 (cit. on p. 12).

[23] J. Pequenao, "Computer Generated image of the ATLAS calorimeter", 2008,
URL: https://cds.cern.ch/record/1095927 (cit. on p. 13).

[24] ATLAS Collaboration, *AtlFast3: The Next Generation of Fast Simulation in ATLAS*, Computing and Software for Big Science **6** (2022) (cit. on pp. 13, 14).

[25] L. Pontecorvo, *The ATLAS Muon Spectrometer*, (2003),
URL: https://cds.cern.ch/record/676896 (cit. on pp. 14, 15).

[26] R. Zhang, *Inclusive and differential cross-section measurements of tW single top-quark production at s = 13 TeV with the ATLAS detector*,
PhD thesis: Rheinische Friedrich-Wilhelms-Universität Bonn, 2019,
URL: https://hdl.handle.net/20.500.11811/7942 (cit. on pp. 15, 17–19, 33).

[27] M. Cacciari, G. P. Salam and G. Soyez, *The anti-$k_tjetclusteringalgorithm$*, Journal of High Energy Physics **2008** (2008) 063, URL: `https://doi.org/10.1088%2F1126-6708%2F2008%2F04%2F063` (cit. on p. 15).

[28] *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, tech. rep., CERN, 2016, URL: `https://cds.cern.ch/record/2157687` (cit. on p. 15).

[29] *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, tech. rep., All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2016-012: CERN, 2016, URL: `https://cds.cern.ch/record/2160731` (cit. on p. 15).

[30] K. Reygers, *Statistical Methods in Particle Physics*, 2020 (cit. on p. 16).

[31] M. Kobayashi and T. Maskawa, *CP-Violation in the Renormalizable Theory of Weak Interaction*, Progress of Theoretical Physics **49** (1973) 652, ISSN: 0033-068X, eprint: `https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-652.pdf`, URL: `https://doi.org/10.1143/PTP.49.652` (cit. on p. 17).

[32] F. Abe et al., *Observation of Top Quark Production in $\overline{p}$ p*, Physical Review Letters **74** (1995) 2626, URL: `https://doi.org/10.1103%2Fphysrevlett.74.2626` (cit. on p. 17).

[33] S. Frixione, E. Laenen, P. Motylinski, C. White and B. R. Webber, *Single-top hadroproduction in association with a W boson*, Journal of High Energy Physics **2008** (2008) 029, URL: `https://doi.org/10.1088/1126-6708/2008/07/029` (cit. on pp. 17, 19, 20, 33).

[34] G. Aad et al., *Measurement of the $t\bar{t}$ production cross-section in the lepton+jets channel at $\sqrt{s} = 13$ TeV with the ATLAS experiment*, Phys. Lett. B **810** (2020) 135797, arXiv: `2006.13076 [hep-ex]` (cit. on p. 17).

[35] *NLO single-top channel cross sections*, URL: `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/SingleTopRefXsec` (cit. on p. 18).

[36] CMS Collaboration, *Observation of tW production in the single-lepton channel in pp collisions at \sqrt{s} = 13 TeV*, Journal of High Energy Physics **2021** (2021), URL: `https://doi.org/10.1007%2Fjhep11%282021%29111` (cit. on p. 19).

[37] J. Ott, *Search for Single Top tW Associated Production in the Dilepton Channel at CMS*, EPJ Web of Conferences **28** (2012) 12041, URL: `https://doi.org/10.1051%2Fepjconf%2F20122812041` (cit. on p. 19).

[38] ATLAS Collaboration, *Probing the Quantum Interference between Singly and Doubly Resonant Top-Quark Production in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector*, Phys. Rev. Lett. **121** (15 2018) 152002, URL: `https://link.aps.org/doi/10.1103/PhysRevLett.121.152002` (cit. on pp. 20, 34).

[39] G. Papamakarios, I. Murray and C. Williams,
Neural density estimation and likelihood-free inference,
Thesis (Ph.D.) University of Edinburgh, 2019, URL: https://hdl.handle.net/1842/36394
(cit. on p. 23).

[40] F. B. Fitch, *Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133.*,
Journal of Symbolic Logic **9** (1944) 49 (cit. on p. 23).

[41] F. Rosenblatt,
The perceptron: a probabilistic model for information storage and organization in the brain.,
Psychological review **65 6** (1958) 386 (cit. on p. 23).

[42] F. Chollet et al., *Keras*, https://keras.io, 2015 (cit. on p. 24).

[43] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library",
Advances in Neural Information Processing Systems 32, ed. by H. Wallach et al.,
Curran Associates, Inc., 2019 8024, URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
(cit. on pp. 24, 40).

[44] S. Ruder, *An overview of gradient descent optimization algorithms*, 2016,
URL: https://arxiv.org/abs/1609.04747 (cit. on p. 25).

[45] Y. Bengio, *Learning Deep Architectures for AI*, Found. Trends Mach. Learn. **2** (2009) 1,
ISSN: 1935-8237, URL: https://doi.org/10.1561/2200000006 (cit. on p. 26).

[46] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*,
http://www.deeplearningbook.org, MIT Press, 2016 (cit. on p. 26).

[47] W. Zhou et al., *BERT Loses Patience: Fast and Robust Inference with Early Exit*, 2020,
URL: https://arxiv.org/abs/2006.04152 (cit. on p. 27).

[48] M. Germain, K. Gregor, I. Murray and H. Larochelle,
MADE: Masked Autoencoder for Distribution Estimation, (),
URL: http://arxiv.org/pdf/1502.03509v2 (cit. on pp. 27–29).

[49] R. A. Fisher, *On the Mathematical Foundations of Theoretical Statistics*,
Philosophical Transactions of the Royal Society A **222** () 309 (cit. on p. 27).

[50] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed and B. Lakshminarayanan,
Normalizing Flows for Probabilistic Modeling and Inference,
Journal of Machine Learning Research **22** (2021),
URL: http://arxiv.org/pdf/1912.02762v2 (cit. on p. 28).

[51] A. V. Aho and J. E. Hopcroft, *The Design and Analysis of Computer Algorithms*, 1st,
USA: Addison-Wesley Longman Publishing Co., Inc., 1974 241, ISBN: 0201000296
(cit. on p. 30).

[52] J. Alman and V. V. Williams, *A Refined Laser Method and Faster Matrix Multiplication*,
CoRR (2020), URL: https://arxiv.org/abs/2010.05846 (cit. on p. 30).

[53] G. Papamakarios, T. Pavlakou and I. Murray, *Masked Autoregressive Flow for Density Estimation*,
2017, URL: https://arxiv.org/abs/1705.07057 (cit. on pp. 30, 31).

[54] C.-W. Huang, D. Krueger, A. Lacoste and A. Courville, *Neural Autoregressive Flows*, 2018, URL: https://arxiv.org/abs/1804.00779 (cit. on p. 31).

[55] ATLAS Collaboration, *Studies of $t\bar{t}/tW$ interference effects in $b\bar{b}\ell^{+}\ell^{-'}\nu\bar{\nu}'$ final states with Powheg and MG5*$_aMC@NLOsetups$, tech. rep., CERN, 2021, URL: https://cds.cern.ch/record/2792254 (cit. on pp. 33, 34).

[56] Christian Herwig, *Probing quantum interference in top production with the ATLAS detector*, ATLAS, 2018, URL: https://indico.cern.ch/event/708573/contributions/2995390/attachments/1649915/2638333/herwig_LHCTopWG.pdf (visited on 05/04/2022) (cit. on p. 34).

[57] *LHC Olympics 2020*, URL: https://lhco2020.github.io/homepage/ (cit. on p. 35).

[58] P. Virtanen et al., *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods **17** (2020) 261 (cit. on pp. 37, 39).

[59] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, 2015, URL: https://arxiv.org/abs/1502.03167 (cit. on p. 38).

[60] N. Henze and B. Zirkler, *A class of invariant consistent tests for multivariate normality*, Communications in Statistics - Theory and Methods **19** (1990) 3595, URL: https://doi.org/10.1080/03610929008830400 (cit. on p. 41).

[61] R. Vallat, *Pingouin: statistics in Python*, Journal of Open Source Software **3** (2018) 1026, URL: https://doi.org/10.21105/joss.01026 (cit. on p. 41).

List of Figures

List of Tables