Meirin in Meyrin: Machine learning in the $t\bar{t}Z$ 2ℓ channel and preparing ATLAS data for education worldwide

Meirin Oan Evans
University of Sussex, United Kingdom
Submitted for the degree of Doctor of Philosophy to the University of Sussex in September 2022.

Abstract

Run 2 of the Large Hadron Collider at CERN provided proton–proton collisions at a centre-of-mass energy of 13 TeV from 2015 to 2018. The ATLAS experiment is one of the experiments recording proton–proton collisions at the Large Hadron Collider, having recorded 139.0 $\text{fb}^{-1}$of data in Run 2. Analysis of this dataset has provided an unprecedented arena to undertake precision measurements of the top quark, the heaviest fundamental particle discovered to date. Indeed, the Large Hadron Collider is sometimes called a “top quark factory". The interaction between top quarks and Z bosons provides a precise probe of the electroweak force, by studying the associated production of a Z with a top-antitop quark pair ($t\bar{t}Z$) at the LHC.

This thesis focuses on the 2-Lepton-Opposite-Sign channel of the $t\bar{t}Z$ process, which is being analysed for the first time with the full Run 2 dataset. By studying the 2-Lepton-Opposite-Sign channel, it is possible to reconstruct properties of the Z boson directly, due to sensitive lepton identification and kinematic measurement by the ATLAS detector. Therefore, this channel provides efficient Z boson reconstruction in studying the $t\bar{t}Z$ process. Multivariate analyses have been developed for this thesis to select $t\bar{t}Z$ 2-Lepton-Opposite-Sign events within Run 2 LHC data. The multivariate analyses included variables regarding leptons, jets and missing transverse momentum in the final state. Z bosons were reconstructed in selected $t\bar{t}Z$ 2-Lepton-Opposite-Sign events, providing additional variables for multivariate analyses. Including the 2-Lepton-Opposite-Sign channel in the overall cross-section measurement of the $t\bar{t}Z$ process provided a more sensitive cross-section measurement using the full Run 2 ATLAS dataset, due to the extra statistics compared to the 3-Lepton and 4-Lepton channels alone.

In addition to the multivariate analyses developed for the ATLAS measurement of the $t\bar{t}Z$ process, similar multivariate analyses were developed using ATLAS Open Data; proton–proton collision data collected by ATLAS, processed and made available to the public. The purpose of developing multivariate analyses with ATLAS Open Data was to create, test and demonstrate educational uses of particle-physics data. Experimental and simulated data for the $t\bar{t}Z$ process, the code to perform multivariate analyses, various educational resources and extensive documentation are now publicly available for students around the world to analyse and learn for themselves.

Statement

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature:. . . . .

Acknowledgements

This thesis would not have been possible without the presence of all those who surrounded me, both professionally and personally.

I acknowledge that this work would not have been possible without everyone working on the CERN accelerator complex and the ATLAS detector. Data from the ATLAS detector constitute my sources of information. My research procedure was to analyse data from the ATLAS detector.

I am deeply grateful to my supervisor, Kate Shaw, for her endless support and confidence in my abilities in the past four years as a PhD student at the University of Sussex. I appreciate all her time and effort, particularly during the write up of this thesis.

I would like to thank all my current and previous colleagues and friends at the University of Sussex and CERN. You all made my time at Sussex, home and CERN great!

Thanks to all my colleagues in the ttZ analysis and Data & Tools group, especially Arturo, Iacopo and Tom for the great work and experience.

I owe my deepest gratitude to my family for their continuous support, and love. Words can't express my love for them, which is why I don't express such words often enough.

An additional thanks to Sir Tom Jones and Dame Shirley Bassey for their continued support.

Finally, thanks to all my friends all over the world who accompanied me in my life up to now.

Contents

List of Tables

List of Figures

Acronyms

2ℓOS 2-Lepton-Opposite-Sign. 10, 13, 14, 15, 19, 20, 21, 23, 24, 28, 45, 60, 61, 63, 66, 69, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 83, 84, 86, 91, 92, 93, 94, 96, 98, 100, 103, 104, 105, 106, 107, 108, 109, 110, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 125, 129, 133, 137, 138, 139, 140, 141, 143, 144, 147, 151, 152, 153, 156, 157, 159, 160, 162, 163, 164, 167, 168

ATLAS A Toroidal LHC ApparatuS. 9, 11, 13, 15, 18, 19, 20, 28, 34, 40, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 58, 59, 60, 61, 62, 63, 64, 67, 68, 69, 76, 77, 78, 79, 80, 81, 82, 83, 84, 86, 98, 99, 103, 104, 109, 116, 141, 143, 144, 146, 152, 156, 160, 162, 163, 164, 167, 168

AUC area under curve. 14, 22, 85, 105, 106, 109, 110, 115, 116, 119, 123, 129, 130, 131, 132, 137, 138, 139, 141, 167

BDT Boosted Decision Tree. 10, 11, 13, 14, 15, 20, 21, 23, 24, 86, 87, 88, 91, 93, 103, 104, 105, 106, 107, 108, 109, 116, 123, 139, 141, 160, 161, 162, 163, 164, 167

BSM Beyond the Standard Model. 27

CERN Conseil Europeean pour la Recherche Nucleaire. 9, 17, 43, 44, 144, 168

CKM Cabibbo–Kobayashi–Maskawa. 13, 34, 35, 36, 39

CMS Compact Muon Solenoid. 27

CP Charge-Parity. 35

csv comma separated values. 11, 15, 144, 156, 157, 160

DNN Deep Neural Network. 10, 11, 13, 14, 15, 21, 22, 87, 90, 91, 103, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 163, 164, 165, 167, 168

DT Decision Tree. 20, 86, 87

EM Electromagnetic. 18, 33, 50, 51, 52, 60, 61

FSR final-state-radiation. 17, 38

HS Hard Scatter. 19, 65

ID Inner Detector. 43, 47, 48, 49, 50, 53, 60, 61, 67

IP Impact Parameter. 62, 67

ISR initial-state-radiation. 17, 38, 59, 98, 99, 140

JER Jet Energy Resolution. 64, 98, 99

JES Jet Energy Scale. 19, 64, 65, 98, 99

JVT Jet Vertex Tagger. 19, 65

LAr Liquid Argon. 18, 51, 52

LHC Large Hadron Collider. 9, 17, 20, 27, 28, 31, 33, 36, 37, 39, 43, 44, 45, 46, 50, 57, 68, 71, 72, 80, 82, 84, 98, 99, 143, 167, 168

LO Leading Order. 18, 38, 56, 57

MC Monte Carlo. 20, 21, 22, 28, 59, 60, 62, 66, 67, 71, 85, 93, 96, 97, 98, 99, 105, 106, 107, 108, 114, 129, 133, 134, 135, 161, 168

MET Missing Transverse Momentum. 10, 19, 46, 52, 61, 68, 150

ML Machine Learning. 15, 55, 71, 84, 156, 157, 158, 160

MS Muon Spectrometer. 9, 18, 43, 52, 53, 60, 67

MVA Multi-Variate Analysis. 10, 13, 15, 20, 21, 28, 55, 62, 66, 71, 75, 79, 84, 85, 86, 87, 88, 91, 93, 94, 95, 96, 98, 100, 101, 103, 104, 108, 109, 114, 116, 120, 121, 122, 123, 139, 141, 150, 156, 157, 160, 162, 163, 164, 167, 168

NLO next-to-leading order. 18, 56, 57, 58

NN Neural Network. 87, 88, 89

NNLO next-to-next-to-leading order. 56, 58

NP nuisance parameter. 98, 100, 101, 139

OS opposite-sign. 28, 72

OSSF opposite-sign-same-flavour. 20, 21, 22, 93, 96, 97, 105, 106, 107, 108, 113, 114, 118, 120, 121, 122, 129, 130, 131, 132, 133, 134, 135

PCBT Pseudo Continuous B-Tagging. 66

PDF Parton Distribution Function. 18, 39, 56, 57, 59, 98, 100

PU pile-up. 19, 65, 98

QCD Quantum Chromodynamics. 9, 32, 33, 34, 35, 56, 58

QED Quantum Electrodynamics. 9, 32, 33, 35

QFT Quantum Field Theory. 29, 32, 33, 35

ROC receiver operation characteristic. 14, 20, 21, 22, 85, 105, 106, 107, 109, 112, 113, 115, 116, 117, 118, 119, 129, 130, 131, 132, 137, 138, 141, 167

SF Scale Factor. 66, 98, 99

SM Standard Model. 9, 17, 27, 29, 30, 31, 32, 33, 35, 37, 45, 52, 100

WP Working Point. 62, 66

1 Introduction

The party's just getting started. This is where the fun starts.

Kobe Bryant [1]

In particle physics, a theory called the "Standard Model (SM) of particle physics" attempts to describe all observed phenomena at the scale of subatomic particles [2–5]. The SM has been successful in describing all phenomena observed by the Large Hadron Collider (LHC) [6] and previous collider experiments [7, 8].

The top quark (t) [9, 10], along with its antiparticle the antitop quark ($\bar{t}$), are the heaviest known particles in the SM [11]. The production of top-quark pairs ($t\bar{t}$) has been measured with a great level of precision [12]. Properties of the top quark have been explored by the LHC in great detail, owing to the large centre-of-mass energy and luminosity at the LHC. One such property is the top quark's weak neutral-current couplings [13], the top quark's interaction with the Z boson [14, 15], a carrier of the weak force. The Z boson is the third heaviest particle in the SM [11], therefore the interaction between a top quark and Z boson is the third highest-scale direct interaction, and therefore it is of interest to probe.

The production of top-quark pairs ($t\bar{t}$) in association with Z bosons is a rare process in the SM [16]. The coupling to the Z boson is not yet well constrained and its value can vary significantly in many models including physics Beyond the Standard Model (BSM) [17–22].

The $t\bar{t}Z$ process is furthermore an irreducible background to several searches for BSM phenomena, such as supersymmetric models [23–25] or searches for vector-like quarks and four-top production [26–28]. The Compact Muon Solenoid (CMS) collaboration has carried out similar searches for such phenomena [29, 30]. Measurements of important SM processes are also affected by $t\bar{t}Z$ background. Examples include $t\bar{t}$ production in association with a Higgs boson [31, 32] or single top-quark production in association with a Z boson [33]. Precise measurements of the $t\bar{t}Z$ process are thus of particular interest.

This thesis focuses on the 2ℓOS channel of the $t\bar{t}Z$ process where the Z decays into 2 opposite-sign (OS) leptons (electrons [34] or muons [35]). This channel was studied because the Z boson can be reconstructed directly from the 2 leptons, without any ambiguity over which leptons are from the Z decay.

An ATLAS measurement of the $t\bar{t}Z$ 2ℓOS channel at a centre-of-mass energy of 13 TeV was conducted with the proton–proton collision data collected during the years 2015 and 2016, corresponding the an integrated luminosity of 36.1 fb^{-1} [36]. In Ref. [37], the ATLAS Collaboration provided the first measurements of the $t\bar{t}Z$ differential cross section using the full dataset from Run 2 of the LHC, corresponding to an integrated luminosity of 139 fb^{-1}.

This thesis presents an extended and refined measurement of the $t\bar{t}Z$ 2ℓOS channel using the full set of data collected by the ATLAS experiment during Run 2 of the LHC.

The thesis is structured as follows:

- Chapter 2 briefly introduces the theoretical background for this thesis;
- Chapter 3 summarises the different parts of the ATLAS detector used in this analysis;
- Chapter 4 describes the data and physics objects used;
- Chapter 5 introduces the strategy for the $t\bar{t}Z$ 2ℓOS analysis of this thesis;
- Chapter 6 contains the results of the signal regions;
- Chapter 7 discusses the education work that forms part of this thesis;
- A conclusion of the analysis is drawn in Chapter 8.

The author's specific contributions were:

- creating event displays for the $t\bar{t}Z$ process;
- validating input datasets over a range of variables in terms of distributions, shapes, and data vs. MC agreement;
- verifying signal vs. background separation provided by a range of variables;
- setting up, training, and evaluating various MVA models for classification of the $t\bar{t}Z$ 2ℓOS process;
- comparing the performance of the various MVA models;
- creating a data pipeline to go from data used for physics analysis to simplified data formats for Open Data;
- creating the datasets used as input for Open Data analyses;
- writing example physics analyses for use with Open Data;
- writing corresponding documentation for datasets and example analyses;
- testing datasets and example analyses for Open Data.

2 The Standard Model, the most precise theory in science

Snow White never could read particle physics for very long.

Anon [38]

This chapter briefly introduces the Standard Model (SM) of particle physics. The first part discusses the various particles and forces of the SM. The second part introduces Quantum Field Theory (QFT), the theory underpinning the SM. The third part focuses on the production and decay mechanisms of the top quark and Z boson.

2.1 The Standard Model primer

The SM formulated in the 1960s and 1970s represents humanity's best understanding of nature at scales smaller than atoms [2–4, 39–42]. Figure 2.1.1 shows the particle content of the SM. There are two fundamental classes of particles, defined by a property called spin, with fermions having a spin of 1/2 and bosons having a full integer spin. Fundamental fermions include quarks [43] and leptons [3], whilst fundamental bosons can be vector-type (spin 1) or scalar-type (spin 0). In the SM the fundamental particles interact via 3 fundamental forces: the strong force, electromagnetic force and weak force. These fundamental forces are carried by gauge bosons, (Table 2.1.1), which is how fermions interact. The coupling constants in Table 2.1.1 are terms that appear in the equations for the respective forces, to quantify the relative strength of different interactions. Since strengths are quantified in relative terms, they have no units. Gravity is not included in the SM.

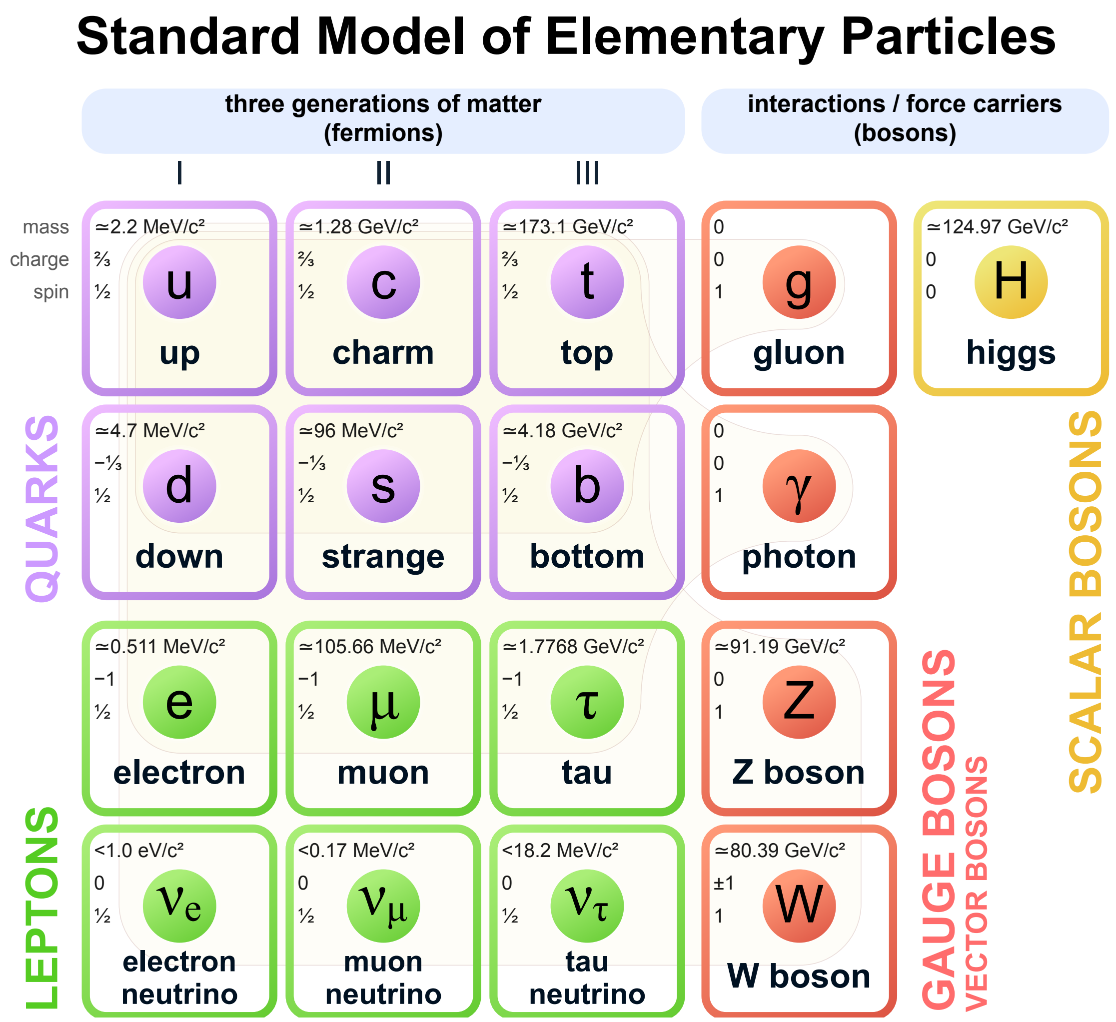


Figure 2.1.1: The particles of the SM. The three generations of matter (fermions) are shown in the left columns, 1, 2 and 3. In the right columns are the force carriers (bosons). In purple are the quarks. In green the leptons. In red are the gauge (vector) bosons. In yellow the only scalar boson of the SM. The mass, charge, symbol and name of each particle are shown in each box. Faint outlines around the symbols show which matter particles interact with which gauge bosons. Figure from Ref. [44].

Force	Carrier	Coupling constant	Strength
Strong	gluon	α_s	~ 1 [45]
Electromagnetic	photon	α_{EM}	1/137 [46]
Weak	$W^{\pm}$ and Z	α_W	1/10000 [47]

Table 2.1.1: The known force carrier bosons. The forces are listed in decreasing order of strength of their coupling constants. The electromagnetic force is 137 times weaker than the strong force. The weak force is 10000 times weaker than the strong force.

Matter generations

Fundamental particles of matter are duplicated into three generations, the only difference between the fundamental properties of the three generations of matter is their masses. Matter particles of generations 2 and 3 are unstable and tend to decay to lighter particles, to reach lower energy states. Up quarks, down quarks and electrons (Section 4.3.1) of generation 1 are abundant in atoms that make up all objects around us. Up quarks, charm quarks, and top quarks are grouped into "up-type quarks". Down quarks, strange quarks, and bottom quarks are grouped into "down-type quarks". Electrons, muons, and taus are grouped into "charged leptons".

Neutrinos (Section 4.3.4) are needed to complete generation 1 along with up quarks, down quarks and electrons. Neutrinos are not seen in the objects around us, but are involved in interactions of the weak force, such as beta decay. Neutrinos come in three types; electron neutrino, muon neutrino, and tau neutrino, corresponding to the three types of charged leptons. Matter particles of generations 2 and 3 are only abundant in high energy processes, such as at the LHC, at the beginning of the universe, and in high-energy collisions in the Earth's atmosphere.

Summary of particles

The up-type quarks (u, c, t) carry an electric charge of +2/3e, whilst the down-type quarks (d, s, b) carry charge -1/3e. Charged leptons (e, μ, τ) carry a charge of -1e. Each charged lepton has a corresponding neutrino (ν_e, ν_μ, ν_τ). All fundamental fermions have a corresponding anti-particle, with an anti-particle's charge obtained by multiplying the corresponding particle's charge by -1. A plot of the masses of fundamental particles is shown in Figure 2.1.2. The four heaviest particles in the SM are the top quark, Higgs boson, Z boson and $W^{\pm}$ respectively [11].

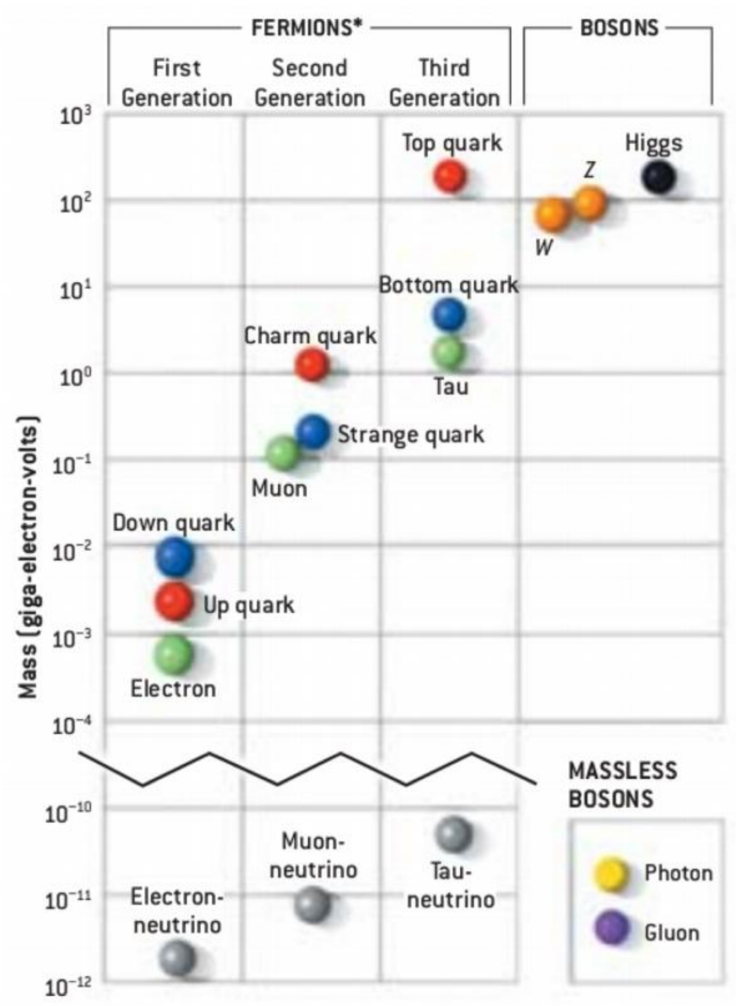


Figure 2.1.2: Masses of the particles of the SM. The three generations of matter (fermions) are shown in the left columns, first, second and third. In the right column are the force carriers (bosons). In red are the quarks with charge +2/3e. In blue the quarks with charge -1/3e. In green the charged leptons. In grey the neutral leptons. The gauge bosons for the weak nuclear force are in orange. Other bosons are in individual colours. Figure from Ref. [48].

Hadrons

Quarks bind together to form hadrons, an example of which is the proton. Hadrons can be classified as baryons containing three quarks (e.g. proton) or mesons containing one quark and one antiquark (e.g. pion). These combinations of three quarks, or one quark and one antiquark, ensure that hadrons have a total colour charge of zero.

2.2 Quantum Field Theory

The SM is underpinned by rigorous mathematics and theory, in the form of a Quantum Field Theory (QFT) [49]. All fundamental particles exist in the form of quantum fields. Three QFTs are of particular importance to this thesis: Quantum Electrodynamics, electroweak theory and Quantum Chromodynamics.

2.2.1 Quantum Electrodynamics (QED)

Quantum Electrodynamics (QED) is the QFT describing the Electromagnetic (EM) force [50]. It formulates the interaction of the photon as the carrier of the EM force. If a particle has an electric charge it can interact via the EM force through the photon. This means that QED describes the EM interaction of charged leptons and quarks, via the exchange of photons. When two charged particles interact via the EM they exchange virtual photons, from particle 1 to particle 2 or vice versa. Neutrinos do not interact via the EM force as their electric charge is zero.

2.2.2 Electroweak theory

At the energies of the LHC, the EM force unites with the weak force to become one “electroweak force" via the Glashow-Weinberg-Salam theory. The QFT groups relevant to Electroweak theory are $SU(2)_L$ and $U(1)_Y$. $SU(2)_L$ is for weak isospin and $U(1)_Y$ is for hypercharge. The S in $SU(2)_L$ stands for special, the U for unitary [51]. (2) means that the matrices of the $SU(2)_L$ group are 2×2. L indicates interaction with left-handed particles. (1) means that the matrices of the $U(1)_Y$ group are 1×2 column matrices.

$SU(2)$ matrices are of the form

$$\begin{pmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{pmatrix},$$

where α and β are complex numbers satisfying the condition $|\alpha|^2 + |\beta|^2 = 1$ [52]. The bar over α and β denotes a complex conjugate.

Electroweak interactions are gauge-invariant under $SU(2)_L \to U(1)_Y$ gauge transformations [53–55]. Following the same principle of unifying the weak force and EM force into the electroweak force at energies accessible by the LHC, Grand Unified Theories attempt to unify the electroweak force and strong force at higher energies [56].

Particles of the SM are left-handed whereas anti-particles of the SM are right-handed. It is these particles that interact with the electroweak force. Handedness appears in the SM Lagrangian [49] as terms that contain the $SU(2)_L$ symmetry of the weak force. Yukawa coupling is an interaction between particles according to the Yukawa potential

$$V(r) \propto -\frac{1}{r}e^{-r\mu} \tag{2.2.1}$$

,

where r is the distance between the two particles and μ is the mass of the particle mediating the interaction. Yukawa couplings appear in the SM Lagrangian as mass terms. Yukawa couplings are discussed later in the context of the top quark in Section 2.3.2.

2.2.3 Quantum Chromodynamics (QCD)

Quantum Chromodynamics (QCD) is the QFT describing the strong force [57], the strongest of all known fundamental forces, being 137 times stronger than the next strongest force, the EM force [46]. QCD interactions are mediated through 8 gluons, the force carriers of the strong force. The strong interactions are mediated by the gluons described by the $SU(3)_C$ group as QCD. The S in $SU(3)_C$ stands for special, the U for unitary [51]. (3) means that the matrices of the $SU(3)_C$ group are 3×3. C indicates interaction with colour charge. In analogy to electric charge being the charge of the EM force, colour charge is the charge of the strong force. There are three colour charges in nature; red, green, and blue (as well as anti-red, anti-green, and anti-blue for anticolour in antiparticles). If a particle has a colour charge it can interact via the strong force through gluons. The $SU(3)_C$ group has 8 generators for the 8 gluons. Quarks are the only fermion to have colour charge and experience the strong force. The properties of QCD mean that the strong force gets stronger at smaller scales, meaning shorter distance, a term called asymptotic freedom [58]. As the only fundamental fermions that experience the strong force, this increase leads to the fact that isolated quarks hadronise into jets in ATLAS, as briefly discussed in Section 4.3.2 (with the exception of the top quark, as discussed in Section 2.3). The fact that quarks cannot exist in isolation is termed colour confinement [59].

The structure of the proton

Protons are composite objects that comprise of three valence quarks (uud) held together by gluons. The valence quarks need to have colour adding up to being colourless (e.g. red, green, and blue). Due to the nature of QCD, protons also contain a sea of quark-antiquark pairs. Sea quark-antiquark pairs are always red+anti-red, green+anti-green, or blue+anti-blue. Together, valance quarks, sea quarks, and gluons are called "partons". A pictorial representation of a proton is shown in Figure 2.2.1.

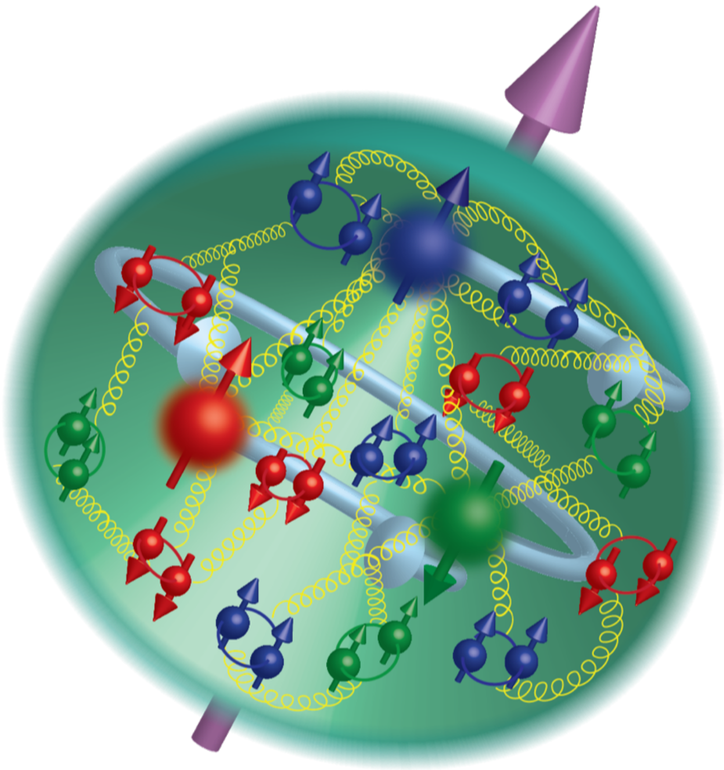


Figure 2.2.1: An artist's impression of a proton. The three "valence quarks" are the largest spheres. The other spheres are "sea quarks". Gluons are represented by yellow springs. Figure from Ref. [60].

CKM matrix

The Cabibbo–Kobayashi–Maskawa (CKM) matrix quantifies the strength of the weak-interaction coupling between different quarks. A high value in the CKM matrix means that the pair of quarks are likely to be seen together in weak interactions. The CKM matrix is shown in Table 2.2.1. The fact that pairs of quarks that are in different generations can couple together via the weak interaction is called “mixing". Like QCD discussed in 2.2.3, the weak interaction field theory formulation is also with $SU(3)$ groups. Having three dimensions to the matrices of the weak interaction is what allows mixing between the different generations of quarks, through a mixing phase δ. Mixing would not be possible in an $SU(2)$ group structure with two generations.

$\lvert V_{ud}\rvert = 0.97370 \pm 0.00014$	$\lvert V_{us}\rvert = 0.2245 \pm 0.0008$	$\lvert V_{ub}\rvert = 0.00382 \pm 0.00024$
$\lvert V_{cd}\rvert = 0.221 \pm 0.004$	$\lvert V_{cs}\rvert = 0.987 \pm 0.011$	$\lvert V_{cb}\rvert = 0.0410 \pm 0.0014$
$\lvert V_{td}\rvert = 0.0080 \pm 0.0003$	$\lvert V_{ts}\rvert = 0.0388 \pm 0.0011$	$\lvert V_{tb}\rvert = 1.013 \pm 0.030$

Table 2.2.1: Cabibbo–Kobayashi–Maskawa (CKM) matrix as given by the best experimental measurements of weak interactions [11]. The fact that $|V_{tb}|$ is close to 1 means that the top quark decays almost 100% of the time to a bottom quark (along with a W-boson).

Renormalisation and factorisation scales

Renormalisation scale is introduced to QFT to account for the involvement of high-momentum virtual particles in interactions [61], e.g. high-momentum virtual photons in QED. Factorisation scale is introduced to QFT to account for the possibility of a massless particle radiating another massless particle [62], e.g. a gluon radiating another gluon in QCD. Uncertainties on these scales give rise to systematic uncertainties based on theory. Renormalisation and factorisation scales are discussed later in the context of systematics for the main analysis of this thesis in Section 6.5.

2.3 The top quark

The top quark remains one of the most interesting fundamental particles to study at particle colliders, as is discussed during this section. It has been of interest since before its discovery and remains of interest today due to its unique properties.

2.3.1 Historical background

The top quark was predicted by Kobayashi and Maskawa in 1973 as a result of their work on kaon decays [63]. To explain Charge-Parity (CP) violation observed in kaon decays, a third generation of fermions was required. Up until then, only two generations had been found. Therefore a third generation containing an up-type quark of electric charge +2/3e and a down-type quark of electric charge of -1/3e was predicted. This third generation up-type quark is now known as the top quark, and this third generation down-type quark is now known as the bottom quark. After much experimental work, the top quark was discovered by the CDF [9] and DØ [10] Collaborations in 1995. It is in generation III of the matter fermions, together with the bottom quark.

2.3.2 Unique properties of the top quark

Some unique properties of the top quark are:

- it is the heaviest known particle in the SM, with mass $m_t = 172.76 \pm 0.30$ GeV [11];

- the large width to its mass peak, $\Gamma = 1.42^{+0.19}_{-0.15}$ GeV [11], which is inversely proportional to lifetime;
- its short lifetime, $\tau = (3.29^{+0.90}_{-0.63}) \times 10^{-25}$ s [11]. This is shorter than the timescale of hadronisation [64] ($\approx 10^{-24}$ s [65]), and therefore when produced it decays almost 100% of the time into a W-boson and b-quark (2.2.3), giving one a unique opportunity to study the bare quark directly;
- its large Yukawa coupling to the Higgs boson, $\lambda_t = 1.16^{+0.24}_{-0.35}$ [11]. The fact that the top-Higgs Yukawa coupling is predicted to be close to 1 (as discussed in Ref. [66]), leads to the high mass of the top quark. Experimentally, λ_t is found to be in agreement with 1. The proximity of the top-Higgs Yukawa coupling to 1 could be a coincidence, but it could also hint at some new physics Beyond the Standard Model. Since the top quark is the heaviest fundamental particle, its interaction with the Higgs boson is strongest amongst the fundamental particles, providing a promising avenue to measure the Higgs Yukawa coupling to fermions.

These unique properties of lifetime, decay and Yukawa coupling make the top quark an important field of study within LHC physics.

Production and decay

The main production of top quarks at the LHC is via the production of a top-antitop pair, with a measured cross-section of $830 \pm 0.4 \pm 36 \pm 14$ pb at $\sqrt{s} = 13$ TeV [12][1]. Cross-section quantifies how likely a process is to occur, a higher cross-section means higher probability of occurrence. The main production Feynman diagrams for the production of a top-antitop pair are shown in Figure 2.3.1. Gluon–gluon fusion (a) is the most common production mechanism at the LHC at 90%.

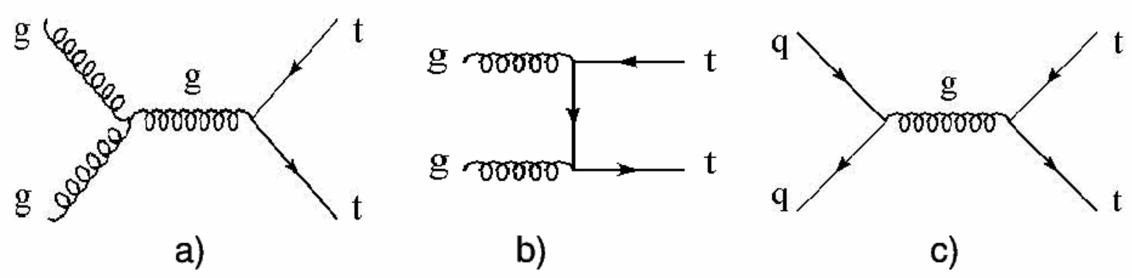


Figure 2.3.1: Feynman diagrams depicting the most common top-antiop pair production mechanisms. (a) is gluon-gluon fusion. (b) is gluon-gluon splitting. (c) is quark-antiquark annihilation.

The CKM matrix of Table 2.2.1 indicates that top quarks decay almost 100% of the time to W-bosons and bottom quarks. This most common of top-quark decays is shown in Figure 2.3.2.

[1] A "barn" (b) is 10^{-28} m^2 and a "pico" (p) is a 10^{-15} multiplier, therefore pb is 10^{-43} m^2

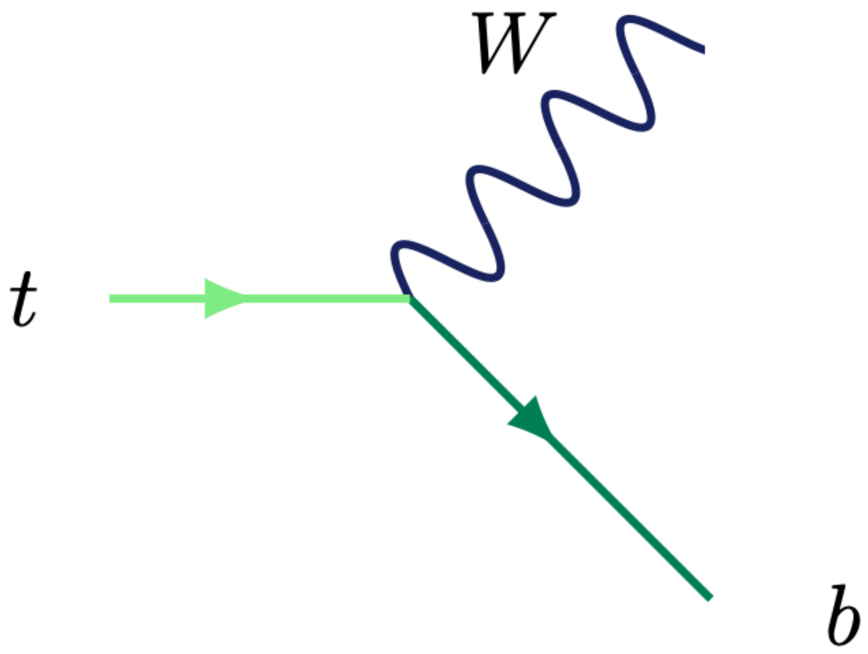


Figure 2.3.2: Feynman diagram depicting the most common top-quark decay to a b-quark and W-boson.

2.4 The Z boson

The Z boson is a force carrier of the weak force. A neutral gauge boson was predicted by theories unifying the weak and electromagnetic forces into a single electroweak force [2–4], adding to the charged gauge bosons ($W^{\pm}$) that had already been discovered. The Z boson was observed for the first time by the UA1 [14] and UA2 [15] Collaborations in 1983.

2.4.1 Unique properties of the Z boson

Some unique properties of the Z boson are:

- it is the third heaviest particle of the SM, with mass $m_Z = 91.1876 \pm 0.0021$ GeV [11];
- the large width to its mass peak, $\Gamma = 2.4952 \pm 0.0023$ GeV [11];
- its short lifetime, $\tau \approx 3 \times 10^{-25}$ s [11], similar to the top quark;

2.5 $t\bar{t}Z$ production

The associated production of a top quark, antitop quark and Z boson is one of the highest energy processes possible to probe at the LHC [67], occurring at energies $\gtrapprox 450$ GeV. Production mechanisms of the $t\bar{t}Z$ process need to be measured to verify their agreement with SM predictions and determine properties of the $t\bar{t}Z$ process. I will now introduce the production mechanisms of the $t\bar{t}Z$ process.

2.5.1 $t\bar{t}Z$ production Feynman diagrams

Leading Order (LO) Feynman diagrams for the production of $t\bar{t}Z$ are shown in Figure 2.5.1. They can be divided into diagrams in which the Z boson is produced via initial-state-radiation (ISR) or via final-state-radiation (FSR). One LO diagram exists for ISR and three for FSR. The FSR diagrams can be further divided into one quark-antiquark initiated process and two gluon-gluon initiated processes. The FSR process in which two gluons fuse into one gluon is the most common $t\bar{t}Z$ production mechanism (Figure 2.5.1 a), followed by the FSR process in which the two gluons split into $t\bar{t}$ pairs (Figure 2.5.1 b), followed by the FSR quark process (Figure 2.5.1 c), followed by the ISR quark process (Figure 2.5.1 d). The quark-antiquark processes are less common than the gluon-gluon processes because the antiquark is a sea quark within the proton. A discussion of why the presence of a sea quark makes a process less common will be given in Section 4.1.1.

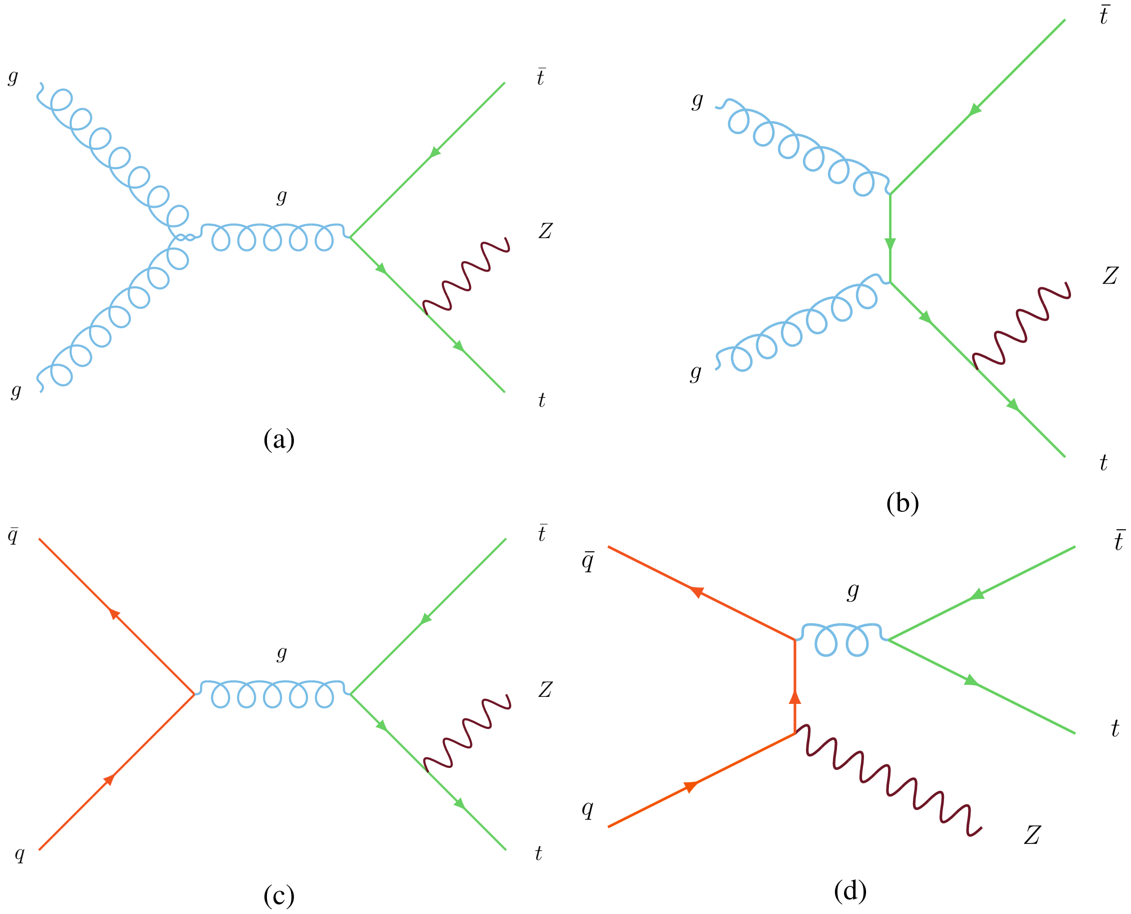


Figure 2.5.1: Feynman diagrams for the leading-order production mechanisms of the $t\bar{t}Z$ process. The different colours are simply to help distinguish between different particles.
(a) The FSR gluon-gluon-fusion process is the most common $t\bar{t}Z$ production mechanism.
(b) The FSR gluon-gluon-splitting process is the second most common $t\bar{t}Z$ production mechanism.
(c) The FSR quark-antiquark process is the third most common $t\bar{t}Z$ production mechanism.
(d) The ISR quark-antiquark process is the fourth most common $t\bar{t}Z$ production mechanism.
Diagrams produced using Ref. [68].

2.5.2 Production rates

Since $t\bar{t}Z$ production is similar to $t\bar{t}$ production, it is useful to quantify the relative production rate of $t\bar{t}$ through gluon processes and through quark processes. No experimental measurements have distinguished between production measurements of $t\bar{t}Z$, therefore numbers for $t\bar{t}$ are given as an indication. However, since $t\bar{t}Z$ is essentially a $t\bar{t}$ process with the extra radiation of a Z boson, the fraction of gluon processes vs. quark processes for $t\bar{t}Z$ are similar to the fraction for $t\bar{t}$. At a centre-of-mass energy of 13 TeV at the LHC, the production of top-antitop quark pairs ($t\bar{t}$) occurs ~90% of the time through gluon processes and ~10% through quark processes. Measurements of the $t\bar{t}Z$ process in this thesis do not distinguish between production mechanisms.

At a centre-of-mass energy of 13 TeV, the production cross-section of $t\bar{t}Z$ is predicted to be [16]

$$\sigma_{t\bar{t}Z} = 0.863^{+0.07}_{-0.09}\,(\text{scale}) \pm 0.03\,(\text{PDF} + \alpha_S)\,\text{pb}.$$

The scales that form part of this uncertainty were introduced in Section 2.2.3, α_S in Section 2.1, and pb in Section 2.3.2. PDFs will be introduced in Section 4.1.1. This cross-section is ≈3 orders of magnitude below the production cross-section for $t\bar{t}$ (Section 2.3.2).

2.6 Top quark decay

The probability of transition from a top quark to a bottom quark with the emission of a W boson is given by the CKM matrix as 99.8% [11]. Following this, the W boson decays either hadronically ($W \rightarrow q\bar{q}'$) or leptonically ($W \rightarrow \ell\nu_\ell$). In hadronic decays, q and $\bar{q}'$ could be any of the flavours u, d, c, s, b, as long as their charges add up to the same charge as the W boson. The prime superscript in $\bar{q}'$ indicates that the antiquark is of different flavour to to the quark q. Hadronic W decays occur (67.41 ± 0.27)% of the time [11]and leptonic decays, $\ell = e, \mu, \tau$, occurs (32.72 ± 0.30)% of the time [11]. For a top-antitop pair production, the final $t\bar{t}$ state depends predominantly on the decays of the two W bosons. This thesis focuses on the all-hadronic $t\bar{t}$ decay, $t \rightarrow bW^+ \rightarrow bq\bar{q}'$ and $\bar{t} \rightarrow \bar{b}W^- \rightarrow \bar{b}q\bar{q}'$. This can be shortened to $t\bar{t} \rightarrow bW^+\bar{b}W^- \rightarrow bq\bar{q}'\bar{b}q\bar{q}'$ and finally to $t\bar{t} \rightarrow bq\bar{q}'\bar{b}q\bar{q}'$. All-hadronic is also known as “alljets". This is chosen as it is the most common $t\bar{t}$ decay, with a probability of (45.44 ± 0.26)%. Top quark decay branching fractions (related to probabilities) are shown in Figure 2.6.1.

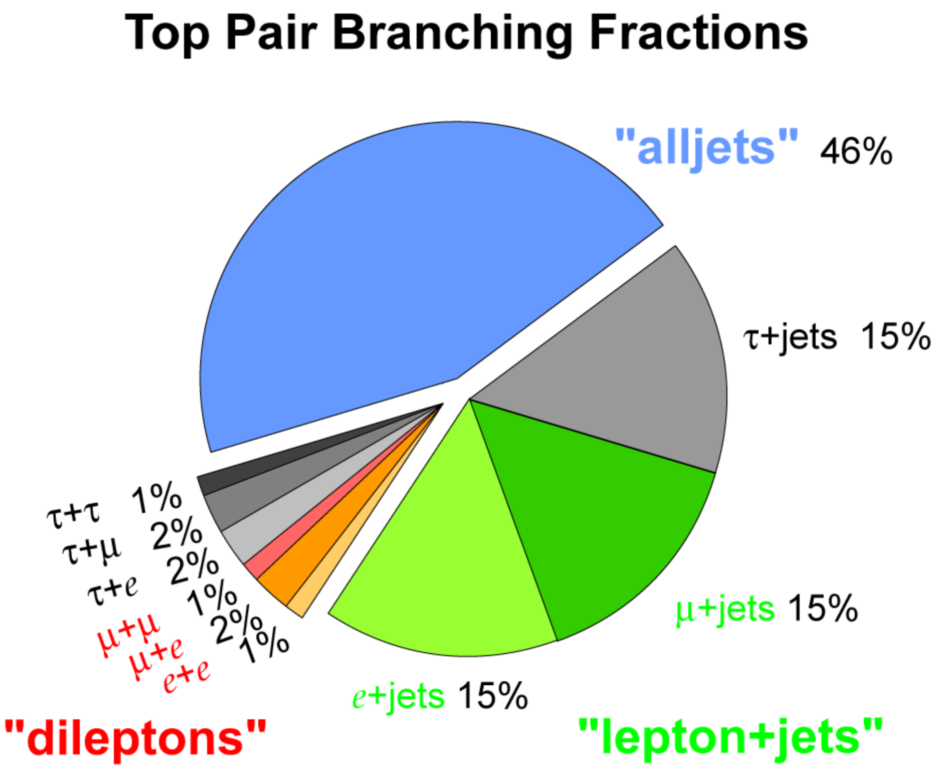


Figure 2.6.1: Pie chart showing the branching fractions of $t\bar{t}$ decay. Branching fractions are related to relative probabilities. This thesis focuses on the “all-hadronic" channel where both W bosons decay hadronically, as it provides the most statistics. Figure from Ref. [69].

2.7 Z boson decay

Z bosons decay (69.91 ± 0.06)% of the time to $q\bar{q}'$, (3.363 ± 0.004)% of the time to e^+e^- and (3.363 ± 0.007)% of the time to $\mu^+\mu^-$ [11]. Though $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$ are not the most probable Z boson decays, they are focused on in this thesis because they provide the cleanest signature in the ATLAS detector since electrons and muons can be reconstructed directly, as briefly discussed in Section 4.3.1 and Section 4.3.3. Z boson decay branching fractions are shown in Figure 2.7.1.

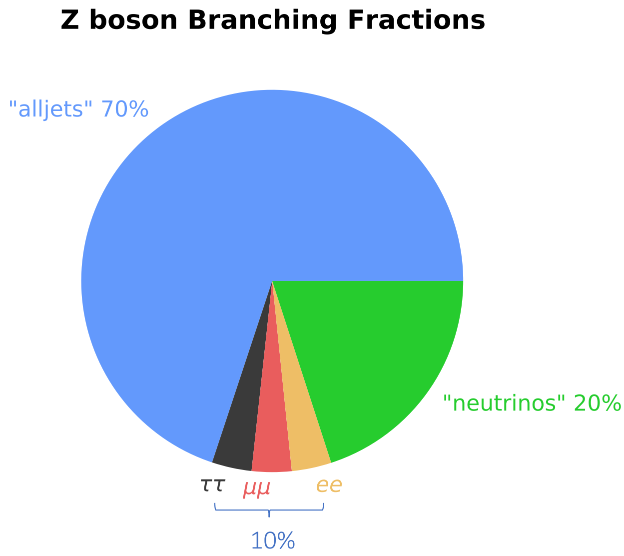


Figure 2.7.1: Pie chart showing the branching fractions of Z boson decay. Branching fractions are related to relative probabilities. This thesis focuses on the ee and $\mu\mu$ channels, since these provide the best avenues for Z boson reconstruction. Figure produced in Python [70].

3 CERN, the LHC, and ATLAS

The clever people at CERN are smashing particles together in the hope that Doctor Who will turn up and tell them to stop.

Ben Aaronovitch [71]

The LHC at CERN is currently the highest energy particle accelerator in the world. During the data taking run of this thesis, it operated at a world-record centre-of-mass energy of 13 TeV. This chapter introduces the background information about CERN, the LHC, and A Toroidal LHC ApparatuS (ATLAS) that is needed for the remainder of the thesis. This chapter introduces the following concepts:

1. CERN and the LHC;
2. the physics programme and detector requirements for ATLAS;
3. how the LHC and ATLAS work together;
4. ATLAS's Inner Detector;
5. ATLAS's Calorimeters;
6. ATLAS's Muon Spectrometer;
7. ATLAS's magnet system;
8. ATLAS's trigger system;
9. tracks in ATLAS.

3.1 CERN and the LHC

CERN is a particle physics laboratory located around the French-Swiss border near Geneva. The main CERN site is in Meyrin, Switzerland. The laboratory contains many experiments and an accelerator complex. The whole complex of particle accelerators is shown in Figure 3.1.1. The steps in the journey of a proton before colliding within ATLAS are:

1. extraction from a hydrogen gas bottle
2. kick-start by a LINAC to 50 MeV [72];
3. acceleration by the BOOSTER to 2 GeV [73];
4. acceleration by the PS to 26 GeV [74];
5. acceleration by the SPS to 450 GeV [75];
6. acceleration by the LHC to 6.5 TeV [6];
7. collisions with other protons within ATLAS at 13 TeV [76].

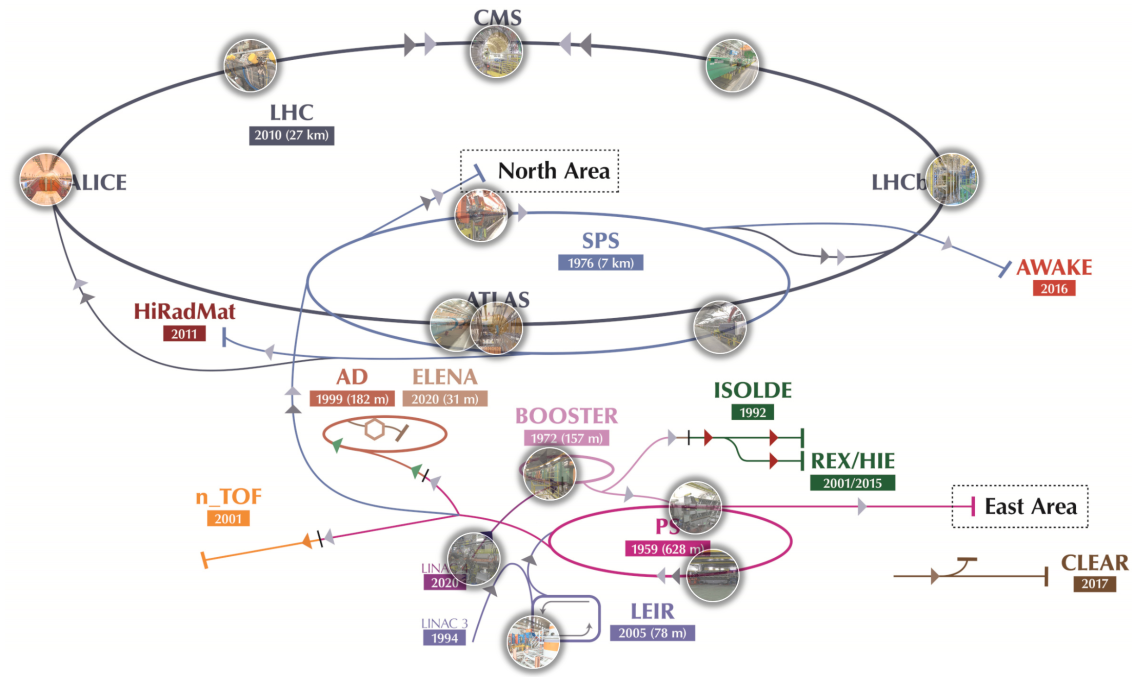


Figure 3.1.1: Schematic diagram of the CERN accelerator complex, with all steps in the protons' acceleration chain labelled. A proton's journey from start to finish will be LINAC → BOOSTER → PS → SPS → LHC → ATLAS. Figure is taken from Ref. [77].

The Large Hadron Collider (LHC) started collecting data at a centre-of-mass energy of 7 TeV in 2010, continuing at 7 TeV into 2011. In 2012, the centre-of-mass energy increased to 8 TeV. 2010-2012 constituted “Run 1" of the LHC. After “Long Shutdown 1" for maintenance and upgrades, “Run 2" of the LHC started at 13 TeV in 2015 and continued to 2018. This thesis used data collected during all of Run 2.

“Integrated luminosity", L_{int} is a measure of how much data have been collected during particle physics collisions, defined by

$$L_{int} = \int \frac{1}{\sigma} \frac{dN}{dt} dt, \tag{3.1.1}$$

where σ is the total cross-section and dN is the number of events detected over a certain time (dt). Therefore, the total cross-section defines luminosity. Run 2 produced 139.0 fb^{-1}of ATLAS data that are ready for physics analysis. The evolution of the integrated luminosity collected by ATLAS over Run 2 is shown in Figure 3.1.2. Data that can be used for physics analyses such as the $t\bar{t}Z$ 2ℓOS analysis for this thesis are described as "Good for Physics".

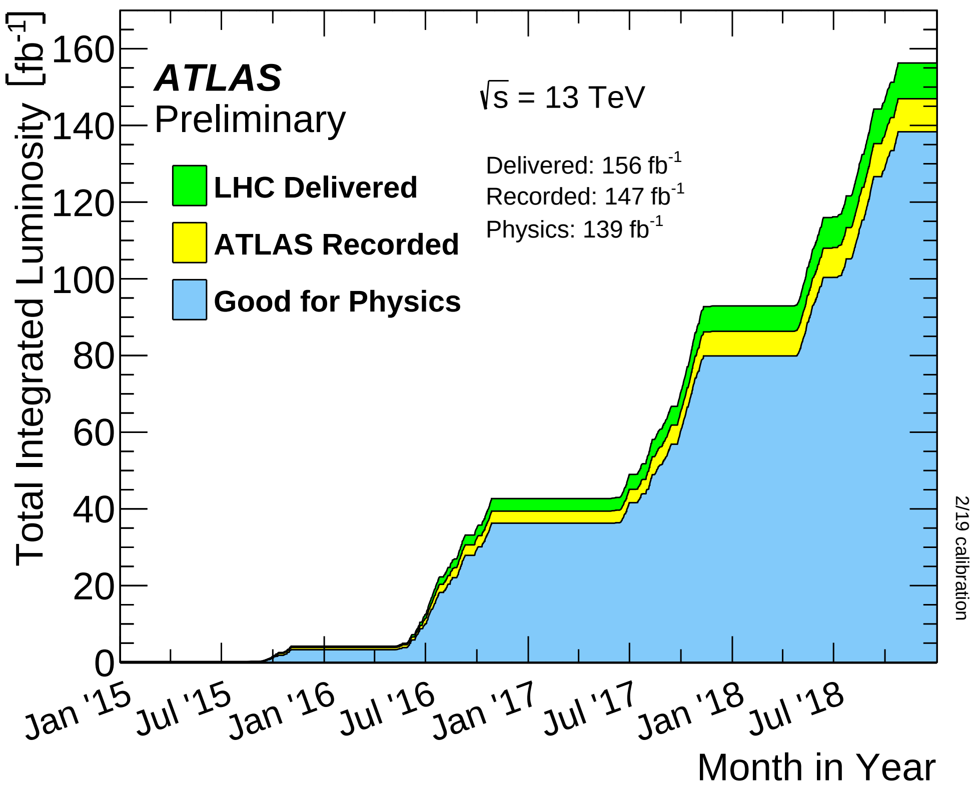


Figure 3.1.2: Cumulative integrated luminosity over time during Run 2 of the LHC. The legend label of interest for this thesis is Good for Physics.

3.2 Physics programme and detector requirements for ATLAS

The physics programme for the ATLAS detector includes:

- measurements of processes involving the top quark;
- measurements of processes involving the Higgs boson;
- measurements of heavy-ion collisions;
- measurements of processes involving B-mesons;
- measurements of other SM processes;
- supersymmetry searches;
- di-Higgs and other diboson searches;
- searches for other exotic processes.

This broad programme necessitates a “General-Purpose Detector". This means the capability to precisely measure the energies and momenta of electrons, positrons, photons, hadrons (jets), and muons. The capability to measure Missing Transverse Momentum is also needed, to infer the presence of particles that pass through ATLAS undetected. These requirements and the parts of the ATLAS detector that achieve these requirements are discussed throughout the rest of this Chapter.

3.3 From the LHC to ATLAS

Once protons have been accelerated by the LHC up to 6.5 TeV, they collide together in the centre of the ATLAS detector, which is shown in Figure 3.3.1. The ATLAS detector is built like a cylindrical onion with many layers, in the sense that each sub-part envelopes the sub-part within it, down to the collision point.

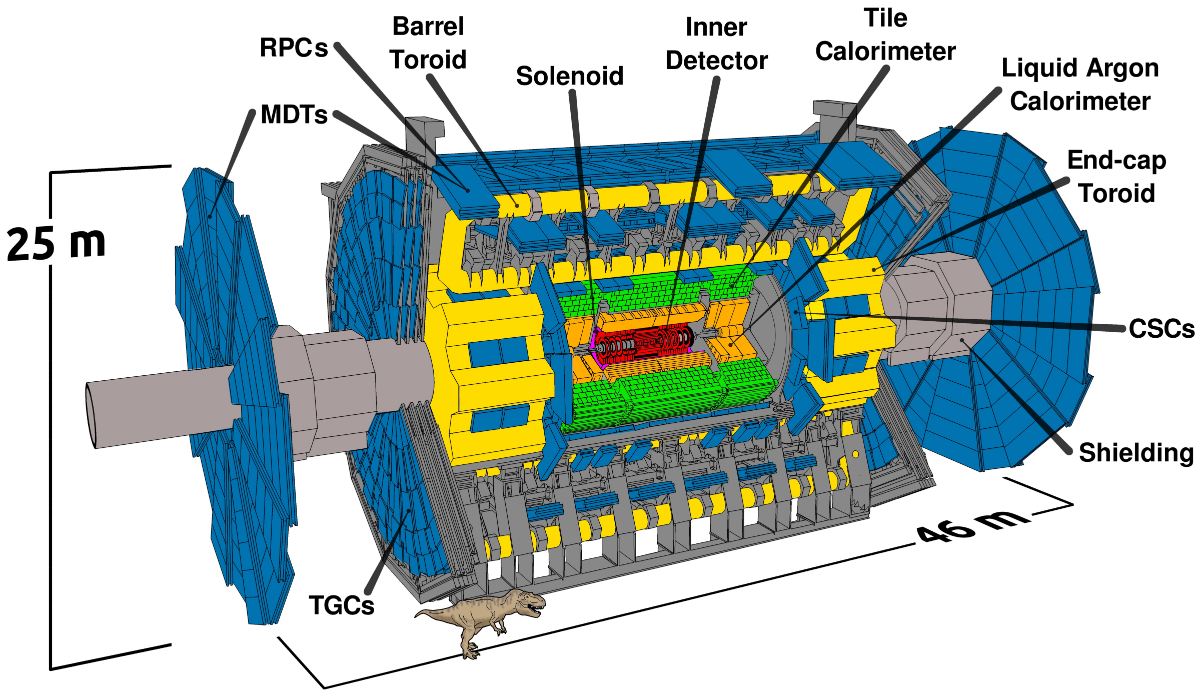


Figure 3.3.1: A cross-sectional diagram of the ATLAS detector. Labelled are all sub-parts to the ATLAS detector. A mini-T-rex is shown for scale. ATLAS is the same length as three school buses (46 m) and the same height as five giraffes (25 m). Figure adapted from Ref. [78].

A specific right-handed Cartesian coordinate system is used by the ATLAS detector, as shown in Figure 3.3.2. The positive x-axis points towards the centre of the LHC. The positive y-axis points upwards. The positive z-axis points along the counter-clockwise LHC beam direction. ϕ is the azimuthal angle in the xy plane, starting from 0 at the x-axis and increasing in value towards the positive y-axis. θ is the polar angle in the xz plane, starting from 0 at the z-axis and increasing in value towards the positive x-axis.

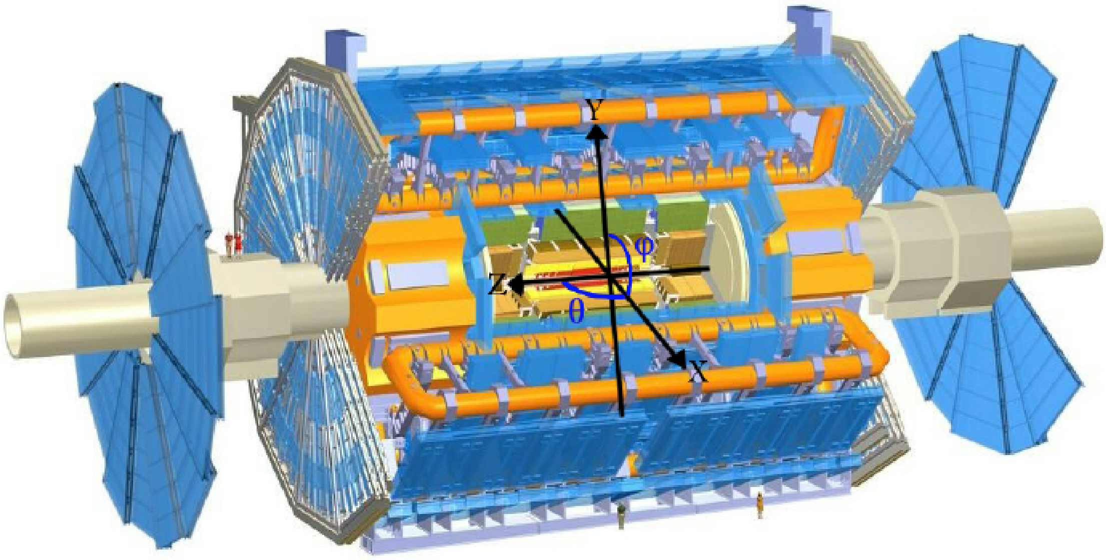


Figure 3.3.2: A cross-sectional diagram of the ATLAS detector, with the Cartesian coordinate system used by ATLAS superimposed. Figure adapted from Ref. [79].

Pseudorapidity, η is defined by

$$\eta = -\ln[\tan\frac{\theta}{2}]. \tag{3.3.1}$$

Another useful quantity to define is the angular separation between two particles, ΔR, given by

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{3.3.2}$$

r is the radial distance from the collision point,

$$r = \sqrt{x^2 + y^2 + z^2}. \tag{3.3.3}$$

The $r\phi$ resolutions and η coverage of the detector sub-systems to be discussed during this Chapter are summarised in Table 3.3.1. The sensitive coverage of many sub-systems goes up to an η of 2.5, because this is where the ID ends.

Sub-system	$r\phi$ resolution (μm)	sensitive η coverage up to
IBL	10	2.5
Pixels	10	2.5
SCT	17	2.5
TRT	130	2.0
ECAL	300-400	2.5 (excluding 1.37 - 1.52)
MS	80	2.7

Table 3.3.1: $r\phi$ resolution and η coverage of different ATLAS detector sub-systems. Sub-systems are written by increasing distance from the proton–proton collision point.

3.4 The Inner Detector

Starting from the part of the detector nearest to where collisions take place, we have the Inner Detector (ID), shown in Figure 3.4.1, taken from Ref. [80]. From inside to out, the sub-parts to the inner detector are the Insertable B-layer (IBL), Pixels, Semiconductor Tracker (SCT) and Transition Radiation Tracker (TRT). The physics objects measured by the ID are tracks, to be discussed in Section 4.3.

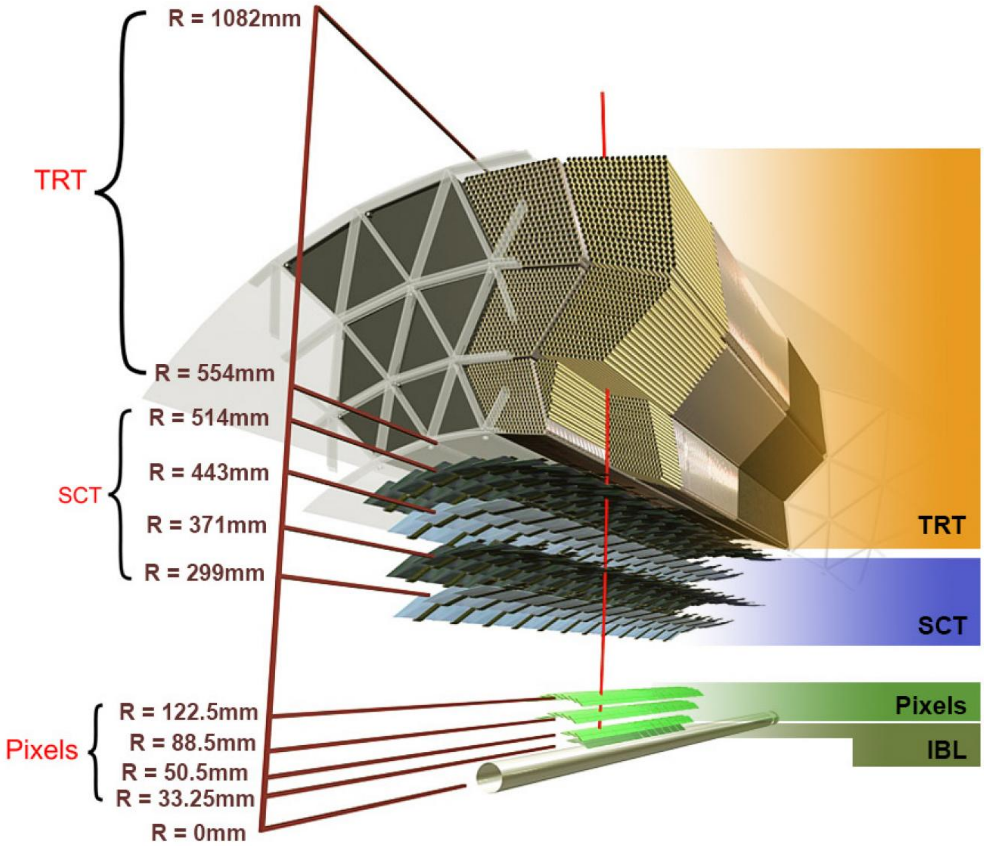


Figure 3.4.1: A cross-sectional diagram of the innermost part of the ATLAS detector, the inner detector. All sub-parts are shown and labelled in different colours, including the Insertable B-layer (IBL), Pixels, Semiconductor Tracker (SCT) and Transition Radiation Tracker (TRT), from inside to out. Figure taken from Ref. [80].

3.4.1 Insertable B-layer (IBL)

The Insertable B-layer (IBL) is the innermost layer of the ID, up to a radius of 33.25 mm from the collision point. It was inserted into the ATLAS detector as a new component in between Run 1 and Run 2 of data taking. It helps with the identification of secondary vertices associated with b-jets (Section 4.3.2), hence its name. Being the detector layer closest to the collision point, the IBL is most susceptible to radiation damage. The IBL works by measuring ionisation energy deposited by charged particles passing through it. Further details on the IBL can be found in the ATLAS paper discussing the production of the IBL [81].

3.4.2 Pixels

The Pixels of the ID are made of three separate layers. The B-layer (innermost) reaches a radius of 50.5 mm from the collision point. Next, Layer 1 reaches a radius of 88.5 mm from the collision point. Last, Layer 2 reaches 122.5 mm from the collision point. Like the IBL, the “B" in the naming of the B-layer of the Pixels also refers to its role in identifying b-jets. Like the IBL, the Pixels work by measuring ionisation energy deposited by charged particles passing through. Further details on the Pixels of the ID can be found in the Pixels Technical Design Report [82].

3.4.3 Semiconductor Tracker (SCT)

The Semiconductor Tracker (SCT) relies on semiconductor technology to track particles through the ID. The SCT consists of four layers, reaching 299 mm, 371 mm, 443 mm and 514 mm from the collision point, respectively. Like the IBL and Pixels, the SCT relies on the silicon semiconductor technology of measuring ionisation energy deposits of charged particles. Further details on the SCT can be found in the ATLAS paper discussing the performance of the SCT in Run 1 of the LHC [83].

3.4.4 Transition Radiation Tracker (TRT)

Differently to the silicon detectors closer to the beam, the Transition Radiation Tracker (TRT) is made of straw drift tubes, with a cathode wire in the centre of each. Drift tubes also rely on ionisation. This technology does not provide resolution as high as silicon. To compensate for this, 73 layers of drift tubes are layered on top of each other to provide almost continuous tracking of charged particles through the ID. Almost continuous tracking means that curvature and thus momentum can be measured. The TRT extends from 554 mm to 1082 mm in radius from the collision point. The measurement of transition radiation within the drift tubes helps distinguish between electrons and charged hadrons (mainly pions). Further details on the TRT can be found in the ATLAS paper discussing the performance of the TRT in Run 1 of the LHC [84].

3.5 The Calorimeters

Moving outwards from the inner detector, the next part of the ATLAS detector is the calorimeter. A cross-sectional diagram of the ATLAS calorimeter is shown in Figure 3.5.1, taken from Ref. [85]. The calorimeter includes the Liquid Argon (LAr) and Tile components. Different parts of the calorimeter are used for absorbing the energy deposited by electromagnetically and hadronically decay particles. Energy deposited by electromagnetically decaying particles in the electromagnetic calorimeter can be associated to photons or electrons (and positrons). Only electrons (and positrons) are used in this thesis, to be discussed further in Section 4.3.1. Energy deposited by hadronically decaying particles in the hadronic calorimeter can be associated to jets, to be discussed further in Section 4.3.2. Electromagnetic showers triggered by electrons and positrons are usually contained with the LAr calorimeter, as all their energy is absorbed within it. Further details on the LAr calorimeter can be found in the LAr calorimeter Technical Design Report [86]. Hadronic showers in the form of jets reach further than the Liquid Argon (LAr) calorimeter into the outer Tile calorimeter, where their energy is then fully absorbed. Further details on the Tile calorimeter can be found in the Tile calorimeter Technical Design Report [87]. Both the EM and hadronic calorimeters are sampling calorimeters composed of alternating layers of absorbing material and active material.

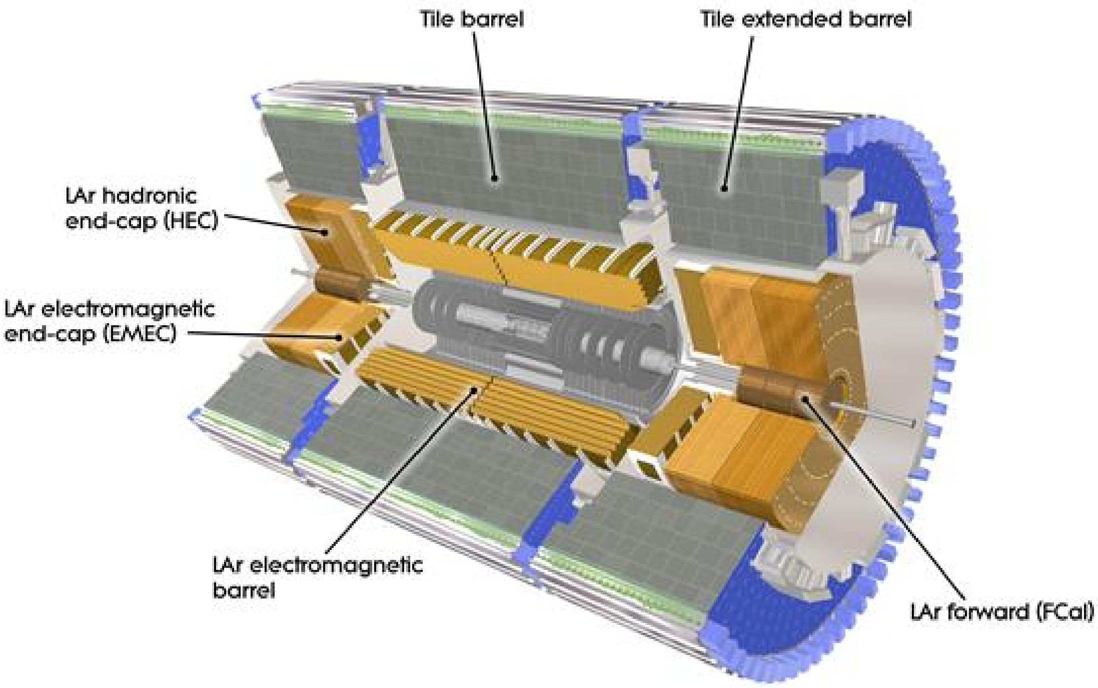


Figure 3.5.1: A cross-sectional diagram of the ATLAS calorimeter. Sub-parts are shown and labelled in different colours. The Tile calorimeter surrounds the Liquid Argon (LAr) calorimeter. Figure taken from Ref. [85].

Having shown a cross-sectional diagram of the ATLAS calorimeter in Figure 3.5.1, a longitudinal diagram of the same section of the ATLAS detector is shown in Figure 3.5.2. Figure 3.5.2 shows the accordion-like structure of an individual module of the Electromagnetic (EM) Liquid Argon (LAr) calorimeter.

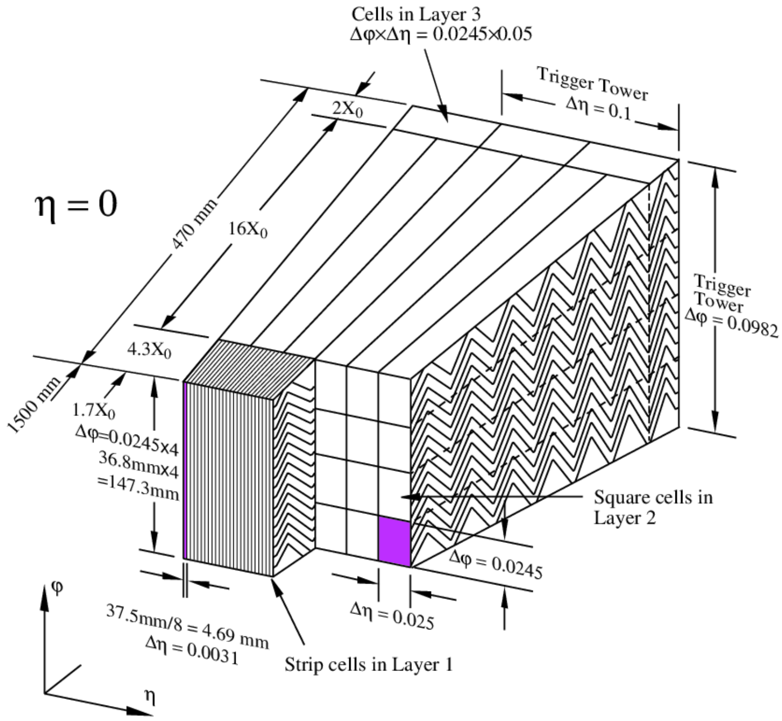


Figure 3.5.2: A longitudinal diagram of the Electromagnetic (EM) Liquid Argon (LAr) calorimeter. Cell size is labelled. Figure taken from Ref. [76].

3.6 The Muon Spectrometer

On the outer part of ATLAS outside of the calorimeters is the Muon Spectrometer (MS), shown in Figure 3.6.1. The role of the muon spectrometer is to measure the energies and momenta of muons, to be discussed further in Section 4.3.3. The toroid magnets, shown in orange in Figure 3.6.1, curve the trajectories of muons through ATLAS. It is by measuring the curvature of muons through the MS that the momenta of muons can be calculated. Further details on the MS can be found in the MS Technical Design Report [88]. All other SM particles should be stopped before the MS, other than neutrinos. Neutrinos pass through the MS, and their existence is inferred by the presence of MET, to be discussed further in Section 4.3.4.

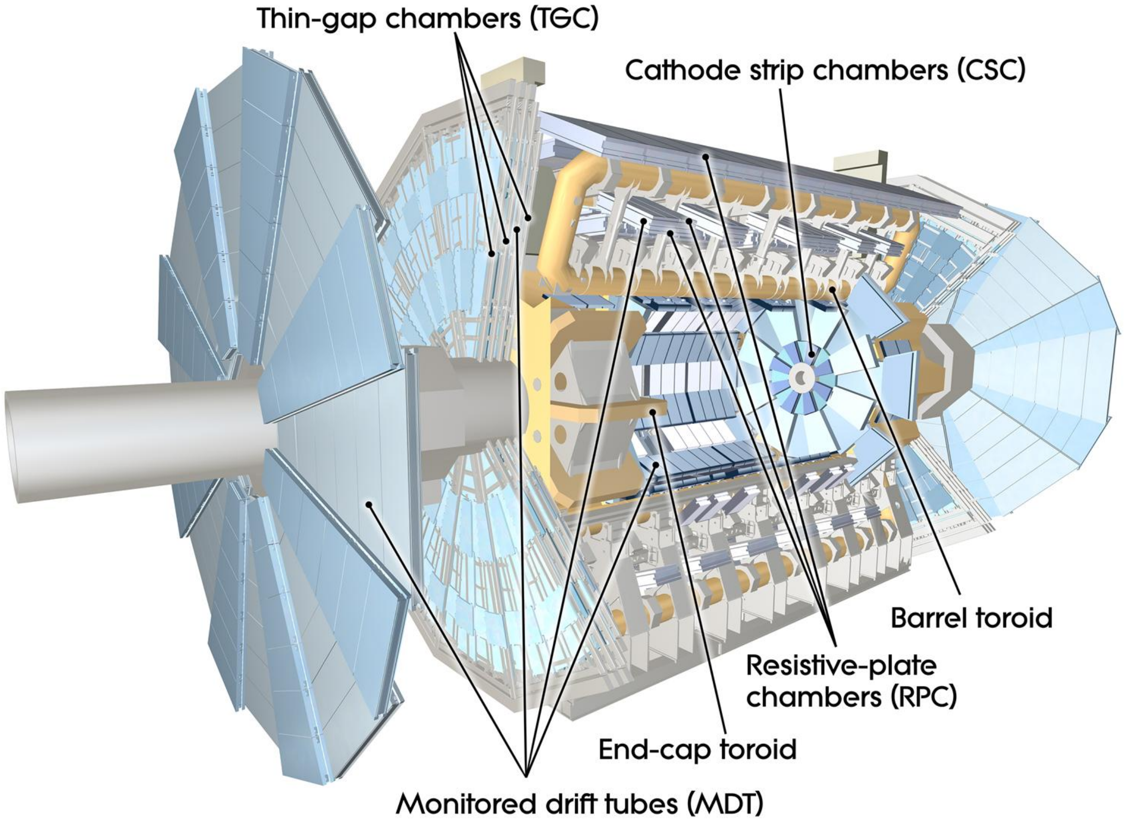


Figure 3.6.1: Schematic diagram of the ATLAS Muon Spectrometer (MS). Different sub-parts of the MS are labelled, as well as the toroid magnets. Figure taken from Ref. [76].

3.7 Magnet System

The ID is surrounded by a superconducting solenoid magnet, engulfing the sub-detector in a magnetic field of strength 2 T in the positive z-direction. The transverse-plane curvature of charged-particle tracks caused by the magnetic force of the solenoid is used to measure charged particle momentum.

Since the solenoid magnet only curves charged-particle tracks in the transverse plane, toroid magnets curve charged-particle tracks in the z-direction. Six toroid magnets are placed outside of the calorimeters for this purpose. The z-direction curvature helps with the measurement of z-momentum. The toroid magnets are labelled in orange in Figure 3.6.1.

3.8 The Trigger

Key to all of the ATLAS detector is the trigger system. The purpose of the trigger system is to reduce the event rate and data rate from the huge influx initially produced by collisions, to manageable rates that can be further processed for analysis. This decrease in event and data rates is shown in Figure 3.8.1.

Custom hardware (called the “Level 1 Trigger" or “L1 Trigger") reduces the event rate from 40 MHz to 100 kHz in less than 2.5 μs. This reduction in event rate reduces the data rate from 1.6 MB/25 ns. Custom software (called the “High-Level Trigger" or “HLT") further reduces the event rate from 100 kHz to 1 kHz and data rate from approximately 160 GB/s to approximately 1.6 GB/s in approximately 250 ms.

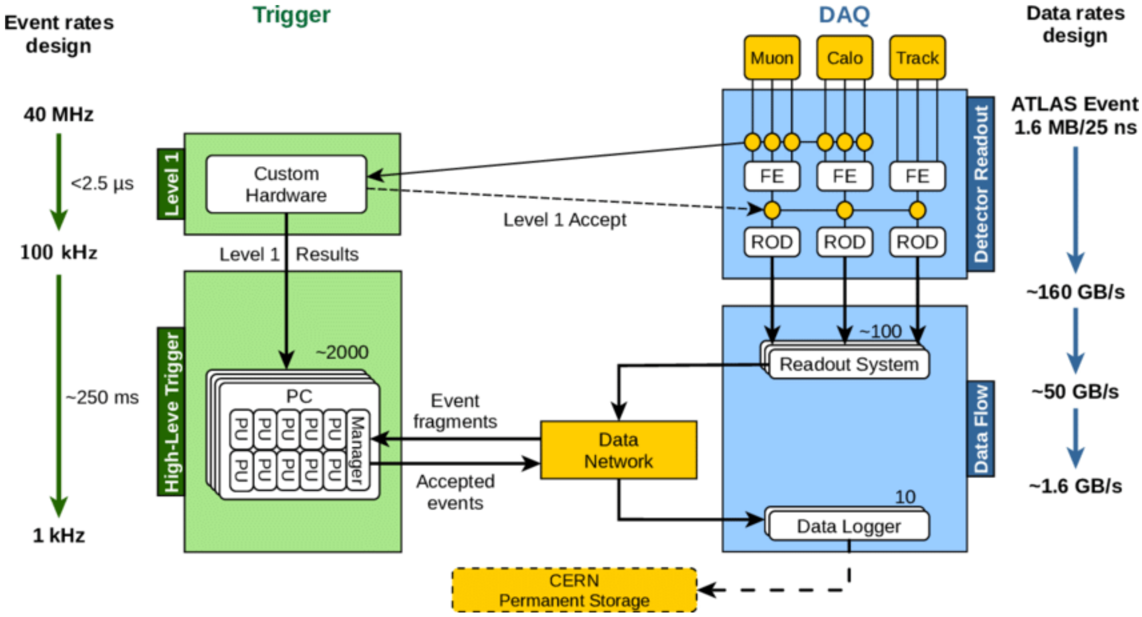


Figure 3.8.1: Flowchart for the ATLAS trigger system. Event rate is shown decreasing along the left and data rate decreasing along the right. Figure taken from Ref. [89].

4 Data and physics objects

The electron: may it never be of any use to anybody!

J.J. Thomson [90]

In order to study the $t\bar{t}Z$ process, proton–proton collision events must be simulated in order to compare measurement with theory. Information from the detector or the event simulation are then used to reconstruct physics objects such as electrons, jets, and muons in order to perform analysis on them. Finally, the Open Data from Run 2 that was processed and released to the public will be described in Section 4.4.

4.1 Event simulation

To measure the $t\bar{t}Z$ process, it is important to simulate signal and background processes, because analysis sensitivity is optimised using simulated samples. Additionally, in the context of Machine Learning [91], simulated samples permit supervised learning [92], since the simulated samples are known to be from signal or from background. This allows the Multi-Variate Analysis (MVA) algorithm to learn on simulated data then to be applied to measured experimental data without known labels because our simulations should be representative of nature. This section describes briefly the simulated samples used by the analysis. They contain the relevant object information and are produced for signal and all relevant background samples.

4.1.1 General points on simulating particle physics processes

The different parts of simulating particle physics processes are described in great detail in Refs. [11, 93–96]. Further references explain the sub-processes [97–102].

Hard-scatter

In cases of interest to this thesis, partons [97] from the incoming protons collide to give a “hard-scatter interaction" [98] that may lead to $t\bar{t}Z$ production. A hard-scatter interaction is when the outgoing particles are different to the incoming partons. Since we are colliding protons, QCD interactions will be abundant. The properties of QCD mean that quarks are observed in ATLAS as jets in our final states, to be briefly discussed in Section 4.3.2.

A Leading Order (LO) process is one in which the hard-scatter produces particles that are different to the incoming partons. A NLO process is a LO process with the emission of an additional boson. A NNLO process is a NLO process with the emission of an additional boson.

PDFs

Parton distribution functions (PDF) play an important role in simulating particle physics processes [103]. PDFs quantify the relative probability that a particular proton constituent is involved in an interaction, according to the fraction of proton momentum that constituent carries. The PDF for a widely used generator that is used in many samples of this analysis (the NNPDF3.0 next-to-next-to-leading order (NNLO) generator [104]) is shown in Figure 4.1.1. x represents the fraction of the proton momentum that a particular proton constituent carries [105]. The y-axis represents the relative probability of having a particular proton constituent with that fraction of proton momentum. The v subscript in u_v and d_v represents valence quark. Quarks without subscripts are sea quarks. The gluon line is reduced by an order of magnitude to a similar level to the quark lines. As a result of the dominance of the strong force (Section 2.1), the gluon PDF is high for a range of x values. This leads to the fact that gluon-gluon production of $t\bar{t}Z$ is the most common production mechanism.

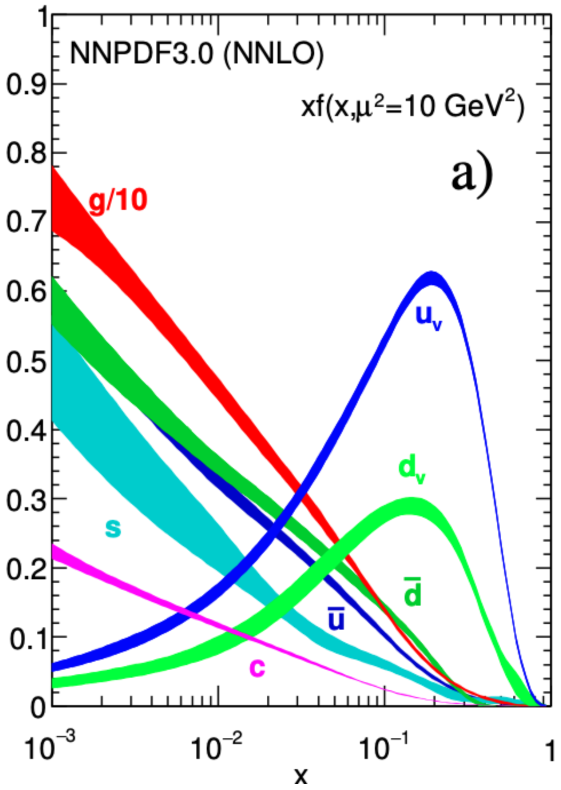
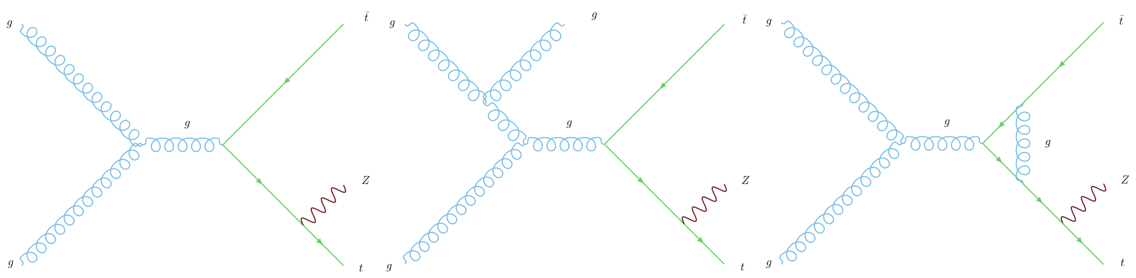


Figure 4.1.1: Graphs of Parton Distribution Function (PDF). Figure from Ref. [11]. x represents the fraction of the proton momentum that a particular proton constituent carries [105]. The y-axis represents the relative probability of having a particular proton constituent with that fraction of proton momentum.

Initial-state radiation and Final-state radiation

Not only is it vital that simulations of particle physics processes are able to simulate the Leading Order (LO) Feynman diagrams, but also that they take into account real and virtual emissions of gluons [106] (next-to-leading order (NLO)). Figure 4.1.2 shows the Feynman diagram for the most common $t\bar{t}Z$ production mechanism at the LHC, along with Feynman diagrams for the first real emission and virtual correction to this LO process.



(a) (b) (c)

Figure 4.1.2:
(a) Feynman diagram for the most common LO $t\bar{t}Z$ production mechanism.
(b) Feynman diagram for the most common LO $t\bar{t}Z$ production mechanism, with the emission of a real gluon from the initial state, making it an NLO process.
(c) Feynman diagram for the most common LO $t\bar{t}Z$ production mechanism, with the emission and absorption of a virtual gluon in the final state, making it an NLO process.
Diagrams produced using Ref. [68].

Hadronisation

QCD interactions are so strong at small distances between quarks that the process of hadronisation [107] occurs and quarks are observed in ATLAS as jets, to be briefly discussed in Section 4.3.2. Hadronisation is the process by which individual quarks and gluons (which cannot be observed directly) are dressed with other quarks and gluons. This dressing occurs because quark-antiquark pairs are produced during hadronisation from gluons. Only collections of quarks and gluons (called hadrons) can be observed directly. The formation process of colour confinement leading to hadronisation into jets is shown in Figure 4.1.3.

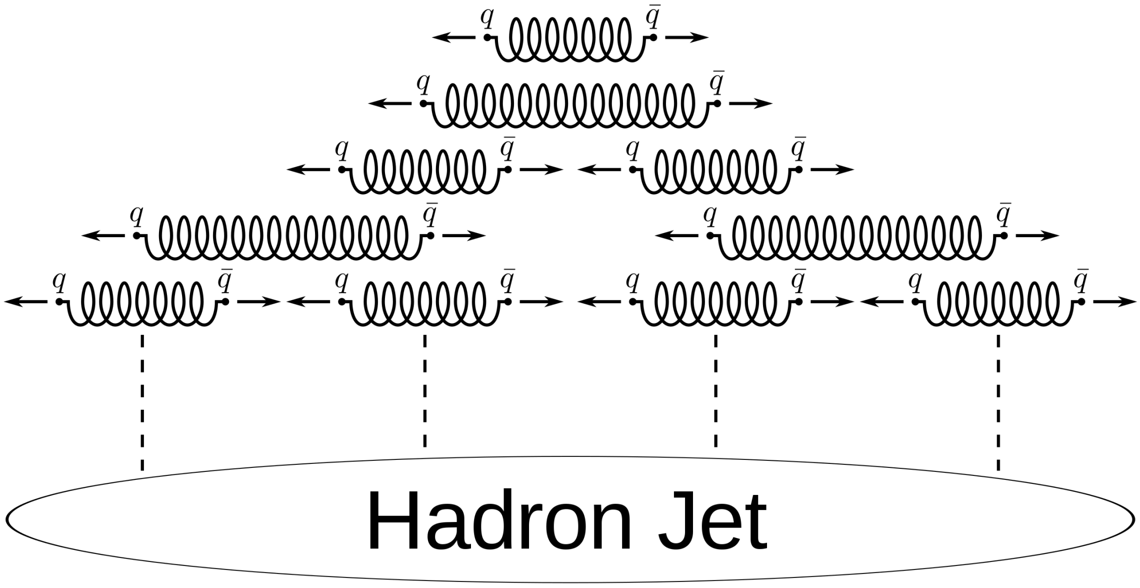


Figure 4.1.3: A pictorial diagram of the process of colour confinement leading to hadronisation into jets. Time flows forward down the diagram. Figure from Ref. [108].

Hadronisation is an important part of particle physics simulations. A diagram depicting the steps involved in simulating hadronisation is shown in Figure 4.1.4. This figure is taken from Ref. [109]. Further references are provided within the caption for Figure 4.1.4 to explain the sub-processes within it. As part of hadronisation occurs a process called "parton shower evolution". In this process, quarks emit gluons as they travel and gluons split into quark-antiquark pairs, with gluon emission then splitting repeating in a shower-like process. NLO and NNLO processes from the hard scatter (Section **??**) can also produce extra quarks and gluons. Therefore, it is necessary to avoid double counting quarks/gluons from the hard-scatter and quarks/gluons from parton shower evolution. Within the simulation of parton shower evolution there are matching procedures to avoid double counting of quarks and gluons between a) higher-order radiation from the hard-scatter (Section **??**) itself, and b) parton shower evolution. Hadronisation was also briefly introduced in Section 2.3.2.

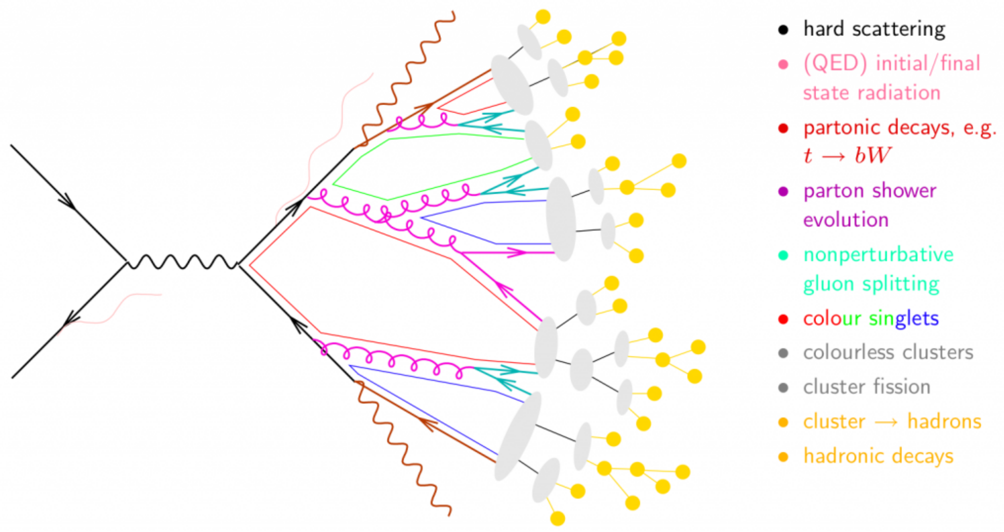


Figure 4.1.4: Diagram of the hadronisation process occurring in particle physics processes. The important steps regarding hadronisation are the ones from parton shower evolution [110], through nonperturbative gluon splitting [111] to cluster $\to$ hadrons. Hadronisation means that any quarks produced in particle physics collisions are observed in ATLAS as jets of hadrons, to be briefly discussed in Section 4.3.2. This figure is taken from Ref. [109].

4.1.2 Simulating signal and background processes

After introducing some general aspects of event simulation, it is necessary to state some specifics about the MC samples used for this thesis.

Simulating the nominal signal process

The production of $t\bar{t}\ell^+\ell^-$ is modelled using MadGraph5_aMC@NLO 2.8.1 [112] with the `NNPDF3.0NLO` [113] PDF set. Top quarks are decayed using MadSpin [114, 115]. Events are interfaced with Pythia 8.210 [116] for hadronisation, using the `A14` set of parameters [117] and the `NNPDF2.3LO` [113] PDF set. The decays of bottom and charm hadrons are simulated using EvtGen 1.2.0 [118].

Cross sections are reported in Ref. [119]. The $t\bar{t}Z$ cross section is supplemented with a correction from Ref. [120].

Quantifying uncertainties

To evaluate theoretical uncertainties of the signal prediction, alternative $t\bar{t}Z$ MC samples are considered. An alternative $t\bar{t}Z$ sample is generated with the same MadGraph5_aMC@NLO version as the nominal sample, but interfaced to Herwig 7 [121, 122] instead of Pythia 8 [116], to quantify an uncertainty associated with parton showering. Furthermore, additional samples with the same settings as the nominal $t\bar{t}Z$ sample, but with an up and down variation of the `Var3c` parameter (which is part of the `A14` [117]), are used to evaluate uncertainties associated to initial-state-radiation (ISR), following a similar approach to Ref. [120].

Simulating background processes

The MC generators used for the main backgrounds to the $t\bar{t}Z$ 2ℓOS process are Powheg-Box [2] [123] for $t\bar{t}$ and Sherpa [2.2.1] [124] for Z+jets.

4.2 Object reconstruction

This section presents the definitions of the physics objects used in the analysis, namely electrons, muons, jets and missing transverse momentum. Each of these objects leave a specific signature in the ATLAS. These signatures are used to identify individual objects. Taus are not considered in this thesis because the efficiency is low compared to electrons/muons and the events are not as “clean". The presence of neutrinos can be inferred through missing transverse momentum.

4.3 Tracks

Lepton and jet trajectories through the ATLAS detector are measured by the tracks they leave. Figure 4.3.1 shows the individual parameters that form the measurement of a track. Tracks are measured in the ID, which was briefly introduced in Section 3.4. Whilst travelling through the ID, charged particles leave hits in different layers of the ID. Tracking algorithms with efficiency ≈94% are used to reconstruct tracks from hits, i.e. tracks are reconstructed from the pattern of hits in the ID [125]. This efficiency is for the phase space $p_{\mathrm{T}} > 0.5$ GeV and $|\eta| < 2.5$.

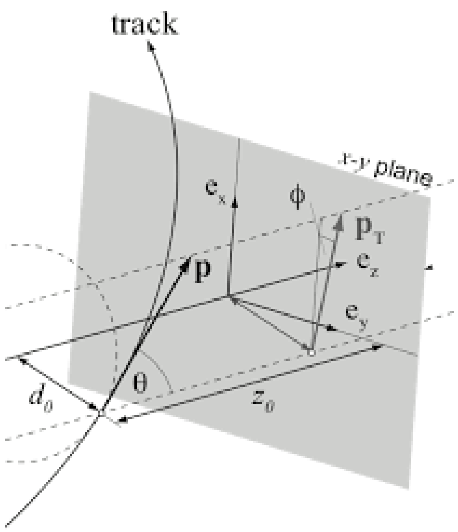


Figure 4.3.1: Diagram showing all the parameters associated with a particle track in the ATLAS detector. The tangential momentum to the particle track is the momentum vector **p**. Transverse momentum, $\boldsymbol{p_T}$, is the component of the total momentum in the $x-y$ plane. ϕ is the azimuthal angle of the track from the x-axis. θ is the polar angle of the track from the z-axis. e_x, e_y and e_z are the unit vectors in the x, y, z directions. z_0 is the z-distance of closest approach of the track to the interaction point. d_0 is the distance of closest approach of the track to the z-axis. Figure from Ref. [126].

By measuring the resultant collection of charge on silicon sensors, a position coordinate or space point is made. As a charged particle travels through the many silicon sensors, these space points can be used to reconstruct the track. Different types of particle leave different types of track in the ATLAS detector. Electrons (and positrons) leave a track in the ID and shower in the EM Calorimeter. Muons (and anti-muons) leave a track in ID and MS. Photons leave no track in the ID but shower in the EM Calorimeter. Charged hadrons leave tracks in the ID and shower in the Hadronic Calorimeter. Neutral hadrons leave no track in the ID but shower in the Hadronic Calorimeter. Neutrinos leave no track in any part of the ATLAS Detector. The different tracks are shown in Figure 4.3.2.

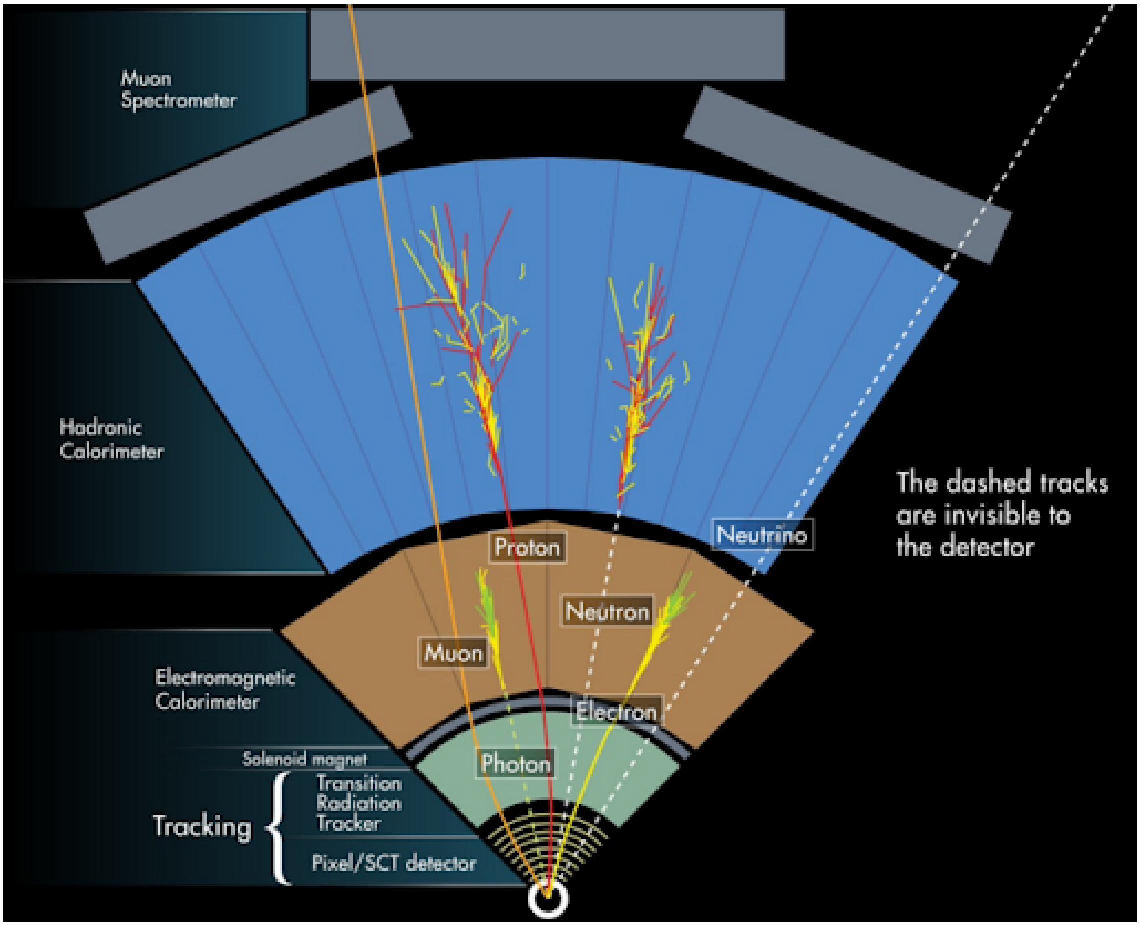


Figure 4.3.2: Cut-away view of how different particles interact with and are seen in the ATLAS detector. Electrons (Section 4.3.1) leave a curved track through all layers of the inner detector (Pixel, SCT, Transition Radiation Tracker) and deposit their energy in a shower inside the Electromagnetic Calorimeter (Section 3.5. Muons (Section 4.3.3) leave a curved track throughout all layers of the detector (Tracking, Electromagnetic Calorimeter, Hadronic Calorimeter, Muon Spectrometer). Hadrons such as protons and neutrons are observed as jets (Section 4.3.2) by showers inside the Hadronic Calorimeter. Neutrinos leave no trace in any part of the detector, but their presence is inferred as Missing Transverse Momentum (MET) (Section 4.3.4). Figure taken from Ref. [127].

4.3.1 Electrons

Electrons (and their antimatter equivalent positrons) are important to this thesis because one of the two Z-boson decay modes that constitutes the $t\bar{t}Z$ 2ℓOS process is $Z \to e^+e^-$.

Requirements

Electrons are identified in ATLAS by the presence of a track in the ID (Section 3.4 and energy deposit in the EM Calorimter (Section 3.5). The requirements for selecting electrons used in this thesis are summarised in Table 4.3.1. Likelihood-based electron identification (ID) [128, 129] is used, as it provides better background rejection compared to cut-based electron identification. The variables used in the Multi-Variate Analysis (MVA) of likelihood-based ID methods are low-level detector information like e.g.:

1. the presence or otherwise of a secondary track close to the electron candidate;
2. the ratio E_T/p_T;
3. calorimeter quantities.

Multiple Working Point (WP) (`VeryLooseLH`, `LooseLH`, `MediumLH` and `TightLH`) are supported for electrons [130]. The "tighter" the WP an electron passes, the more efficiently the electron has been reconstructed. `TightLH` electrons are used in this thesis. Furthermore, a requirement on electron isolation, corresponding to the `PLVLoose` isolation WP [34] is applied. Isolated electrons are desirable for this thesis because electrons coming from secondary processes other than the decay of a Z-boson are likely to be non-isolated. Electrons in the LAr crack-region ($1.37 < |\eta| < 1.52$) are rejected to reduce the contribution from fake or non-prompt electrons. This region in η marks the transition between different sub-systems of the ATLAS detector, and therefore the resolution is lower in this region. Table 3.3.1 of Section 3.3 shows the η coverage of different sub-systems of the ATLAS detector.

Requirements on the transverse (d_0) and longitudinal (z_0) Impact Parameter (IP) are applied to reduce contribution from charge-misidentified electrons, fake leptons, non-prompt leptons and pile-up [131]. The associated scale factors (SFs) for electron reconstruction, identification, and isolation are applied in MC, to correct for the efficiency differences between data and simulation if applying these requirements [128].

Type of requirement	Requirement
Acceptance	$p_{\mathrm{T}} > 7$ GeV
	$\lvert\eta\rvert < 2.47$
	except $1.37 < \lvert\eta\rvert < 1.52$
Impact parameter	$\lvert d_0/\sigma(d_0)\rvert < 5.0$
	$\lvert z_0 \cdot \sin(\theta)\rvert < 0.5$ mm
Quality	`TightLH`
Isolation	`PLVLoose`

Table 4.3.1: Summary of the electron object definitions.

Efficiency

An electron is identified in ATLAS with >60% efficiency in the phase space $E_T > 5$ GeV and $|\eta| < 2.5$ In addition, efficiency changes according to electron energy. Figure 4.3.3 shows how electron identification depends on electron transverse energy. Efficiency is lowest for `Tight` electrons, since the requirements to be accepted as a `Tight` electron are most stringent.

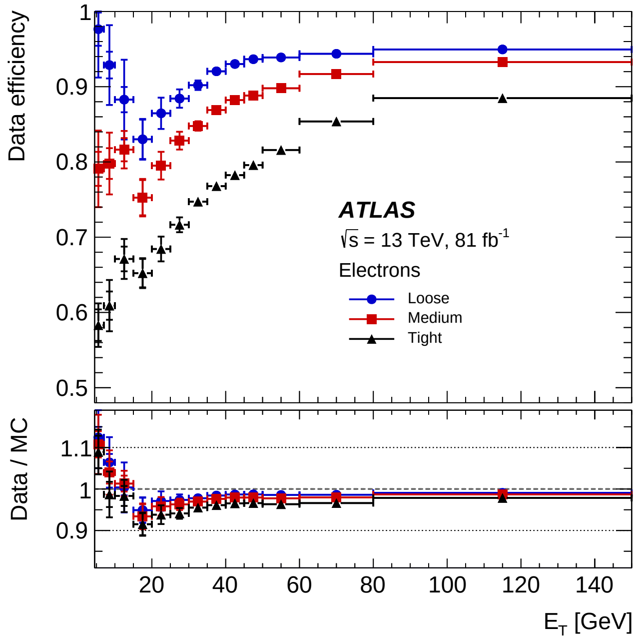


Figure 4.3.3: Graph of electron identification efficiency in data (data efficiency) on the y-axis as a function of transverse energy (E_T) on the x-axis. The lower plot is a ratio of data to simulation (Data/MC). The different markers and colours indicate different requirements on the likelihood of being an electron. `Tight` electrons in the black triangles place the most stringent requirements on electron reconstruction. `Medium` electrons in the red squares place looser requirements on electron reconstruction. `Loose` electrons in the blue circles place requirements on electron reconstruction that are looser again. This graph uses data from 2015-2017, corresponding to an integrated luminosity of 81 fb^{-1}at a centre-of-mass energy of 13 TeV. Figure taken from Ref. [128].

4.3.2 Jets

As introduced in Section 4.1.1, quarks and gluons are observed in ATLAS as jets. Jets are important to this thesis because the $t\bar{t}$ decay being measured in the $t\bar{t}Z$ 2ℓOS process is $t\bar{t} \rightarrow$ jets.

Requirements

ATLAS identifies jets by the presence of energy deposits in the Hadronic Calorimeter, which was briefly introduced in Section 3.5. The jet selection is summarised in Table 4.3.2. Jets are reconstructed using the anti-k_t jet algorithm [132] as implemented in the FastJet package [133] with topological clusters [134] as input. A pictorial representation of jets clustered with the anti-k_t jet algorithm is shown in Figure 4.3.4. Jets are calibrated with the `EMPFlow` (Electro-Magnetic Particle Flow) [135] scheme applying the jet area pile-up corrections [136]. A cluster of radius 0.4 is used for jets in this thesis, where radius $\Delta R = \sqrt{\Delta y^2 + \Delta\phi^2}$. JVT is described in a further paragraph. Jets are described in detail in Refs. [137, 138].

Type of requirement	Requirement
Collection	`AntiKt4EMPFlow`
Acceptance	$p_T > 25$ GeV $\lvert\eta\rvert < 2.5$
Jet Vertex Tagger	reject jets with $p_T < 120$ GeV and $\lvert\eta\rvert < 2.5$ and JVT < 0.59

Table 4.3.2: Summary of the jet selection criteria.

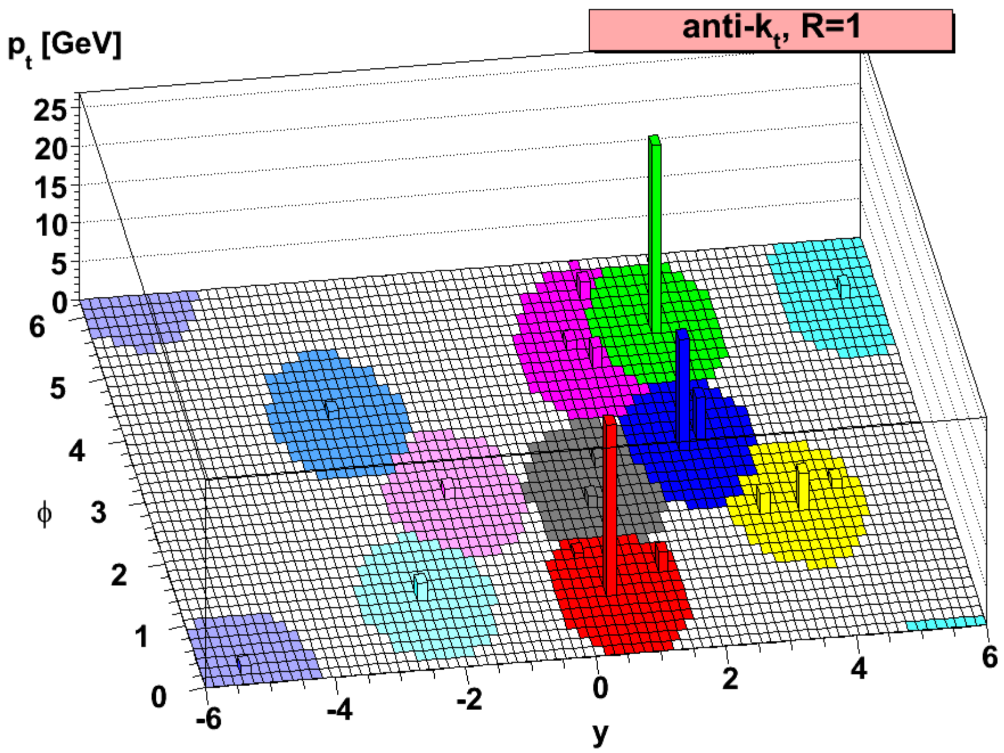


Figure 4.3.4: Jets used in this thesis are clustered together from hadrons using the anti-k_t algorithm. High p_T jets that are used in this thesis are represented by coloured circles with high bars. Ref. [132] describes the algorithm in detail. Figure taken from Ref. [132].

Jets are kept only if they have $p_T > 25$ GeV and are inside a pseudorapidity (η) range of $|\eta| < 2.5$. The ATLAS detector has slightly different responses to jets of different energy and at different pseudo-rapidity (η) within the detector. In a similar way to the identification efficiency of muons and electrons varying with energy or momentum, the uncertainty on the measurement of jet energy varies with jet transverse momentum. Ref. [139] describes Jet Energy Scale (JES) in detail. Jet energy measurements not only depend on jet transverse momentum, but also jet pseudorapidity. Jet Energy Resolution (JER) drops rather rapidly at $|\eta| > 2.5$ for reconstructed jets of high energy, which is part of the reason for using jets with $|\eta| < 2.5$ in this thesis. Ref. [139] describes JER in detail.

The `JetVertexTagger` (JVT) [140] is employed in order to mitigate pile-up effects, and specifically reject any jets with $p_{\mathrm{T}} < 120$ GeV, $|\eta| < 2.5$ and for which JVT < 0.59. Pile-up jets are jets coming from interactions other than the specific parton–parton collision that has resulted in the process of interest, such as $t\bar{t}Z$. Figure 4.3.5 shows how selecting high JVT scores removes a large fraction of pile-up jets. Selecting JVT > 0.59 would remove most pile-up jets, but also some Hard Scatter (HS) jets that are of interest. This is why jets with $p_{\mathrm{T}} > 120$ GeVand $|\eta| < 2.5$ are kept no matter their JVT. Jets with $p_{\mathrm{T}} > 120$ GeVand $|\eta| < 2.5$ are likely to be of interest and are unlikely to arise from pile-up. 120 GeV is chosen as a threshold as any jets with $p_{\mathrm{T}} >$ are likely to originate from high-energy processes, of which $t\bar{t}Z$ is an example.

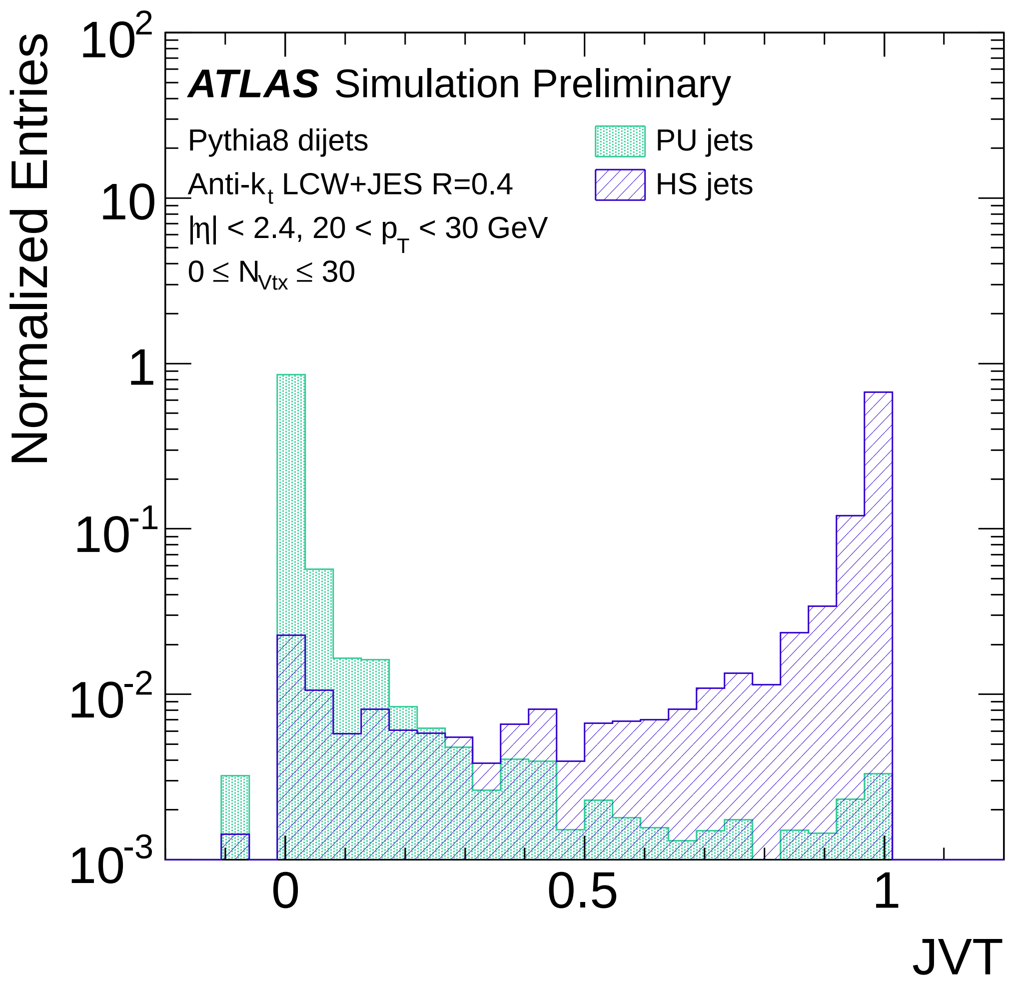


Figure 4.3.5: A Jet Vertex Tagger (JVT) is used to suppress the effect of low-p_T jets that are not part of the hard interactions that produce the processes of interest for this thesis. This graph shows how JVT score differs for Hard Scatter (HS) and pile-up (PU) jets. HS jets tend to give a high JVT score and PU jets tend to give a high JVT score. This graph uses simulated Pythia 8 [116] jets clustered with the anti-k_t algorithm [132]. Ref. [140] describes JES in detail. Figure taken from Ref. [140].

b-jets

The most important jets for this thesis are b-jets, because a top quark decays 99.8% of the time to a W boson and a b-quark [11] (See Chapter 2). Figure 4.3.6 shows a pictorial diagram of how b-jets are distinguished from other jets. A b-hadron tends to travel a few cm from the primary vertex to a secondary vertex before decaying. This means there tends a d_0 impact parameter (distance of closest approach) of a few cm between the secondary and primary vertices. This in contrast to light jets that do not have a secondary vertex.

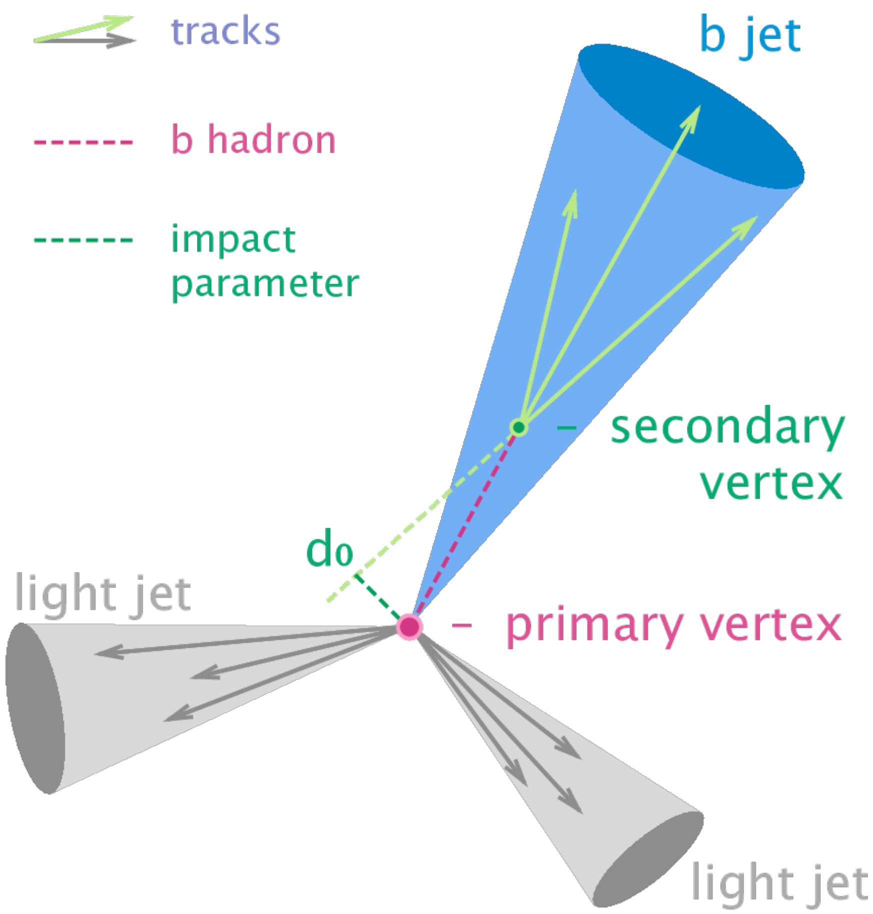


Figure 4.3.6: Pictorial diagram of how b-jets are distinguished from light jets. b-jets are of particular importance to this thesis, since top quarks almost always decay to b-quarks. Figure from Ref. [141].

b-tagging algorithms

To distinguish b-jets from other jets and implement the idea of Figure 4.3.6 algorithmically, b-tagging algorithms need to be used. In this thesis, a specific algorithm from the DL1 family is used, called DL1r. DL1r is a high level MVA tagger which uses lower level information and MVA techniques as inputs. Ref. [142] describes DL1 in detail. b-tagging algorithms such as DL1r define several Working Points (WPs), according to b-jet reconstruction quality. Which WP is passed can be used to define a variable Pseudo Continuous B-Tagging (PCBT), that can eventually be used as input for analysis.

b-tagging algorithms are not perfect in identifying b-jets. Even with the b-tagging algorithms that reject most light and c jets, corrections are needed to fully compare b-jets in data and MC. Such corrections are applied to MC in the form of SFs. Scale factors are applied to simulation, so that data and simulation agree.

4.3.3 Muons

Muons (and their antimatter equivalent anti-muons) are important to this thesis because one of the two Z-boson decay modes that constitutes the $t\bar{t}Z$ 2ℓOS process is $Z \to \mu^+\mu^-$.

Requirements

Muons are identified using both the ID (Section 3.4) and MS (Section 3.6), to associate tracks between both detector elements. Muon selection criteria for this thesis are summarised in Table 4.3.3. ATLAS supports the following working points for muon identification: `Loose`, `Medium`, `Tight`, `HighPt` and `LowPt` [35]. Muons in this thesis have to pass `Tight` quality requirements and have $p_{\mathrm{T}} > 7$ GeV, as well as $|\eta| < 2.5$.

As it is the case for electrons, requirements on transverse (d_0) and longitudinal (z_0) IP are applied to reduce contribution from fake leptons, non-prompt leptons and pile-up [35]. Muons selected for this thesis need to pass the `PLVLoose` isolation WP [143]. The associated SFs for identification and isolation are applied as multiplicative factors to the MC event weight, to correct for the efficiency differences between data and Monte Carlo [35].

Type of requirement	Requirement
Acceptance	$p_{\mathrm{T}} > 7$ GeV $\lvert\eta\rvert < 2.5$
Impact parameter	$\lvert d_0/\sigma(d_0)\rvert < 3.0$ $\lvert z_0 \cdot \sin(\theta)\rvert < 0.5$ mm
Quality	`Tight`
Isolation	`PLVLoose`

Table 4.3.3: Summary of the muon object definitions.

Efficiency

Muons are identified in ATLAS with ≈ 99% efficiency in the phase space $p_{\mathrm{T}} > 5$ GeV and $|\eta| < 2.5$ In the same way as electrons, muon identification efficiency changes according to muon energy. Figure 4.3.7 shows how muon identification depends on muon transverse momentum. Figure 4.3.7 is a graph of muon identification efficiency in data and MC on the y-axis as a function of transverse momentum (p_T) on the x-axis. The lower plot is a ratio of data to simulation (Data/MC). Areas where Data/MC is different to 1 are corrected for using scale factors. Open markers indicate MC and closed markers indicate Data. The different colours indicate different methods of measuring muon identification efficiency. This graph measures `Medium` muons using data from 2015, corresponding to an integrated luminosity of 3.2 fb^{-1}at a centre-of-mass energy of 13 TeV. Figure 4.3.7 is for `Medium` muons, but the same principle of varying efficiency applies to `Tight` muons.

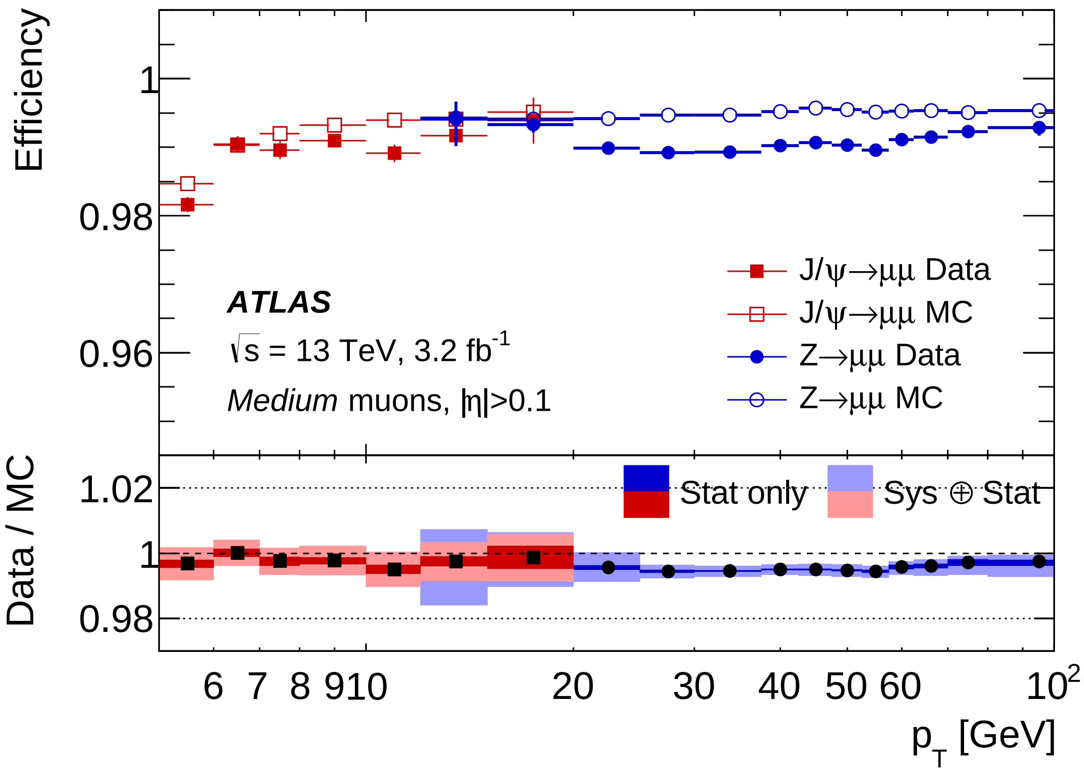


Figure 4.3.7: Muon identification efficiency as a function of transverse momentum. Figure taken from Ref. [35].

4.3.4 Missing Transverse Momentum

The total transverse momentum before a proton–proton collision in ATLAS is zero. Conservation of momentum can be applied in the transverse plane to conclude that the total transverse momentum after a collision is also zero. By measuring the transverse momentum of all detectable physics objects and tracks not associated with physics objects, the negative vector sum of the measured transverse momentum is calculated. When “Missing Transverse Momentum" (MET) is greater than approximately 30 GeV, it may indicate that a particle has passed through ATLAS undetected. The undetected particles relevant to this thesis are neutrinos. 30 GeV is chosen as a threshold because mis-measurements of the transverse energies of visible particles can lead to smaller values of MET.

4.4 Open Data

Open Data are a commitment by the LHC experiments to share their data with the public, but we go beyond this by supplying tools and examples for the public (mainly students) to use. Like all four large LHC experiments [144–146], the ATLAS experiment provides Open Data [147–151], allowing students to access real proton-proton collision data collected by ATLAS, along with tools, software and documentation all accessible on the ATLAS Open Data website [152] and CERN Open Data Portal [153]. These data can then be used for teaching, training and outreach [154] outside the ATLAS Collaboration around the world [155–162]. The intended target audience is primarily university students, and a number of universities have now incorporated lab courses into their degree programs based on the Open Data. Students can then analyse various processes [163–169], including $t\bar{t}Z$. Chapter 7 presents various analyses of the $t\bar{t}Z$ 2ℓOS process, and compares them with analyses using full ATLAS data.

10 fb^{-1}of measured data were produced with the variables of Table 4.4.1. Further descriptions of the variables can be found in Ref. [170].

Prefix	Variables
	runNumber, eventNumber, channelNumber, mcWeight, XSection, SumWeights
scaleFactor	PILEUP, ELE, MUON, PHOTON, TAU, BTAG, LepTRIGGER, PhotonTRIGGER, TauTRIGGER, DiTauTRIGGER
lep	n, truthMatched, trigMatched, pt, eta, phi, E, z0, charge, type, isTightID, ptcone30, etcone20, trackd0pvunbiased, tracksigd0pvunbiased, pt_syst
met	et, phi, et_syst
jet	n, pt, eta, phi, E, jvt, trueflav, truthMatched, MV2c10, pt_syst
photon	n, truthMatched, trigMatched, pt, eta, phi, E, isTightID, ptcone30, etcone20, convType, pt_syst
largeRjet	n, pt, eta, phi, E, m, truthMatched, D2, tau32, pt_syst
tau	n, pt, eta, phi, E, charge, isTightID, truthMatched, trigMatched, nTracks, BDTid, pt_syst
ditau	m

Table 4.4.1: Information content of the 13 TeV ATLAS Open Data. Variable names are prefixed with the left column, e.g. lep_pt

5 $t\bar{t}Z$ analysis strategy

Measurement is like laundry. It piles up the longer you wait to do it.

Amber Naslund [171]

This chapter introduces the motivation and strategy for the main analysis carried out for this thesis: a machine-learning classification of the $t\bar{t}Z$ 2ℓOS process. This analysis used the full Run 2 data of the LHC from 2015-2018, at a centre of mass energy of 13 TeV and an integrated luminosity of 139.0 fb^{-1}. The first analysis of the $t\bar{t}Z$ process in the 3ℓ and 4ℓ channels with full Run 2 data was published in EPJC [37]. This analysis including the 2ℓOS channel will be published after this thesis. In brief, the analysis strategy is to use MVA to classify $t\bar{t}Z$ 2ℓOS signal from background processes, contributing to a $t\bar{t}Z$ cross-section measurement. Multiple MVA are tested and compared for best performance in classifying between $t\bar{t}Z$ 2ℓOS signal and background processes.

In this chapter, Section 5.1 introduces the motivation to study the $t\bar{t}Z$ 2ℓOS process. After defining specific signal regions in the $t\bar{t}Z$ 2ℓOS channel (Section 5.2), event visualisations for various $t\bar{t}Z$ channels are shown in Section 5.3. ML concepts are then used in these defined 2ℓOS signal regions (Section 5.4). Using the signal regions, validation plots for the input variables used in the initial MVAs need to be made (Section 5.5). Some new MVA input variables used for this thesis also needed to be validated (Section 5.6).

The author's specific contribution to this chapter was to:

- create event displays for the $t\bar{t}Z$ process, in the 2ℓOS, 3ℓ, and 4ℓ channels;
- validate input datasets over a range of variables in terms of distributions, shapes, and data versus MC agreement. This validation was done for approximately 50 variables, some used in MVAs and some not;
- verify signal versus background separation provided by a range of variables. The same $\approx$50 variables were checked for separation power.

5.1 Motivation

Reasons for which the $t\bar{t}Z$ 2ℓOS process are of interest are introduced in this section. To do this, different possible $t\bar{t}Z$ decays are discussed.

Motivation to study the $t\bar{t}Z$ 2ℓOS channel

The $t\bar{t}Z$ process is of interest to study because it is one of the highest energy processes possible to probe at the LHC. Together, the masses of the top quark, antitop quark and Z boson amount to approximately 440 GeV. The combination of the large branching ratio (≈46%) of a $t\bar{t}$ all-hadronic decay and the clean reconstruction of $Z \to e^+e^-$ or $Z \to \mu^+\mu^-$ makes the $t\bar{t}Z$ 2-Lepton-Opposite-Sign (2ℓOS) channel attractive to study. $t\bar{t}Z$ decay branching fractions are shown in Figure 5.1.4. When $t\bar{t}Z$ is produced, $t\bar{t}Z \to bq\bar{q}'\bar{b}q\bar{q}'\ell^+\ell^-$ ($\ell = e, \mu$), occurs (3.05 ± 0.01)% of the time. This 3.05% is different to the 2ℓOS $Z \to \ell\ell$ percentage shown in Figure 5.1.4 because $t\bar{t}Z \to bq\bar{q}'\bar{b}q\bar{q}'\ell^+\ell^-$ ($\ell = e, \mu$) does not include taus. In addition, the 2ℓOS is of particular interest because it allows simpler reconstruction of the Z boson, due to the dileptons arising from the Z boson decay being of opposite-sign (OS). By combining the 2ℓOS channel with other $t\bar{t}Z$ decay channels, a combined measurement of the $t\bar{t}Z$ production cross section can be performed.

Origin of different $t\bar{t}Z$ decay channels

This thesis focuses on the 2ℓOS channel with $Z \to \ell^+\ell^-$, since it provides both a high branching fraction for $t\bar{t}$ decay and simplicity in reconstructing the Z boson. Table 5.1.1 shows how different $t\bar{t}Z$ decay channels are obtained from $t\bar{t}$ decay branching fractions and Z boson decay branching fractions. Figures 5.1.1-5.1.3 show examples of diagrams for the different channels. There are two possible combinations of $t\bar{t}$ decay and Z boson decay that lead to the 2ℓOS channel. "all-hadronic + neutrinos" in Table 5.1.1 indicates that the "allj-hadronic" ("alljets") fraction and "neutrinos" fractions from Figure 2.7.1 for Z branching ratios are added together, whereas "lepton+jets" in Table 5.1.1 indicates the "lepton+jets" fraction from Figure 2.6.1 for $t\bar{t}$ branching ratios. All calculations in this pie chart involving "leptons" include τ.

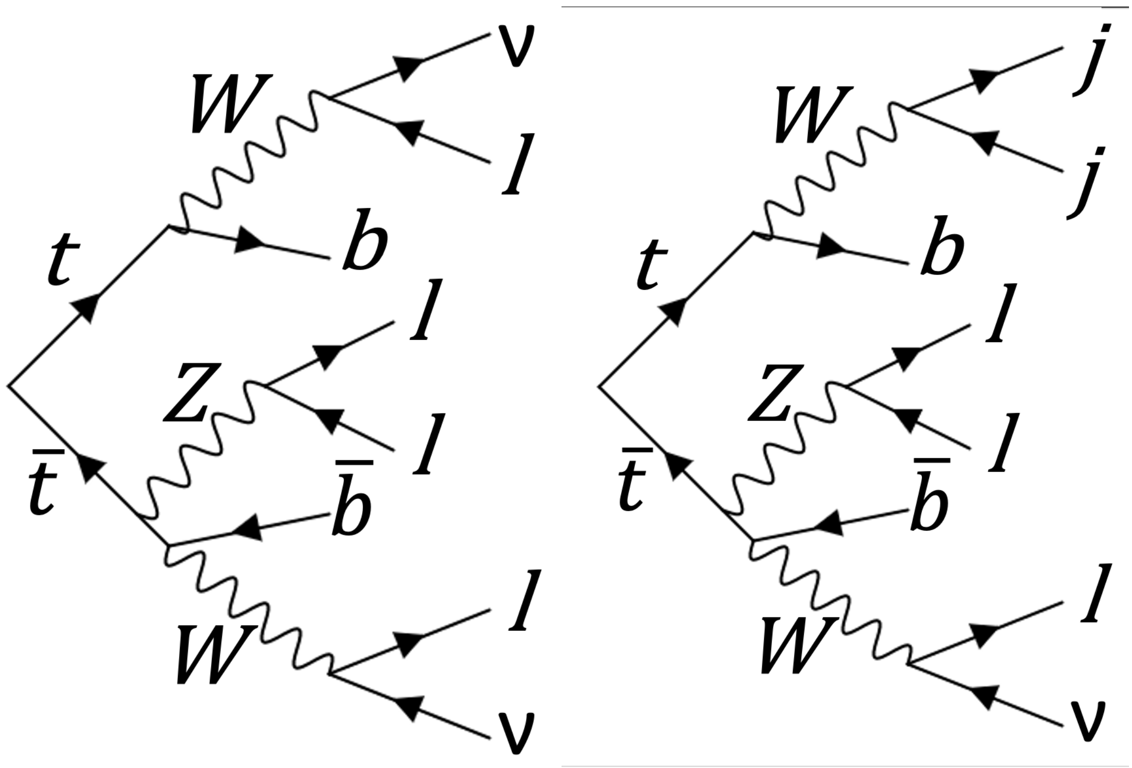


Figure 5.1.1: Example diagrams for the (left) 4ℓ channel and (right) 3ℓ channel of the $t\bar{t}Z$ process.

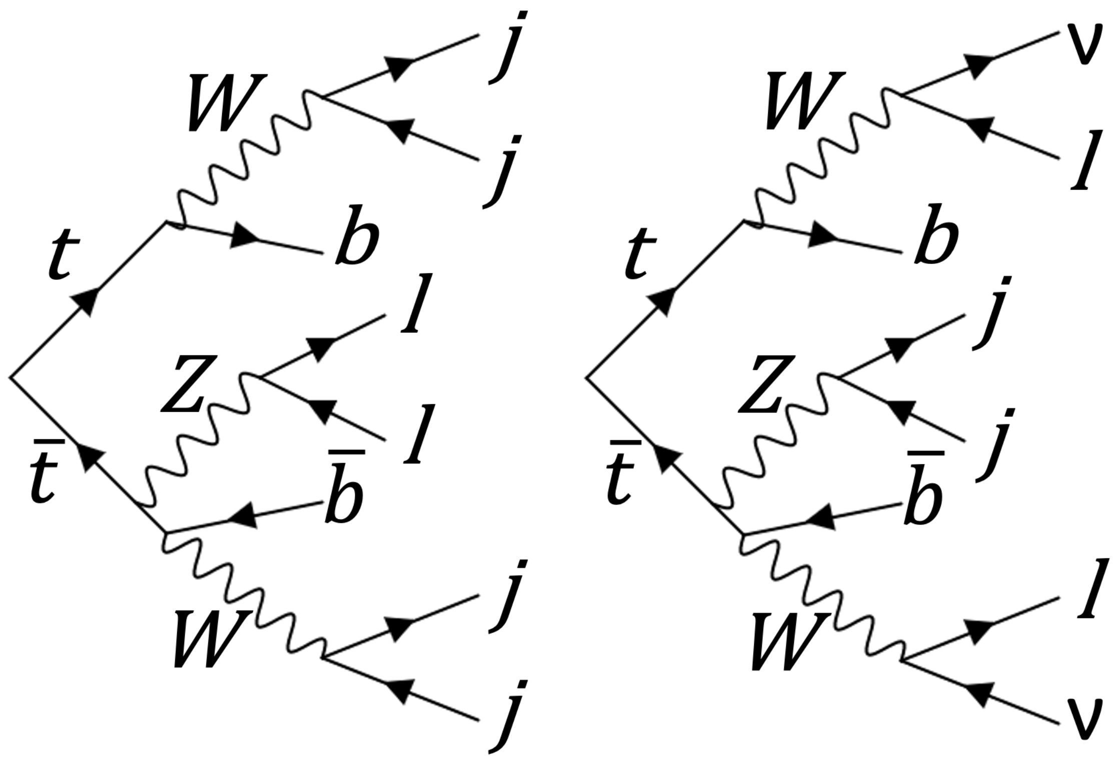


Figure 5.1.2: Example diagrams for the (left) 2ℓOS (with $Z \to \ell\ell$) channel and (right) 2ℓOS (with $Z \to jj$ or $Z \to \nu\nu$) channel of the $t\bar{t}Z$ process.

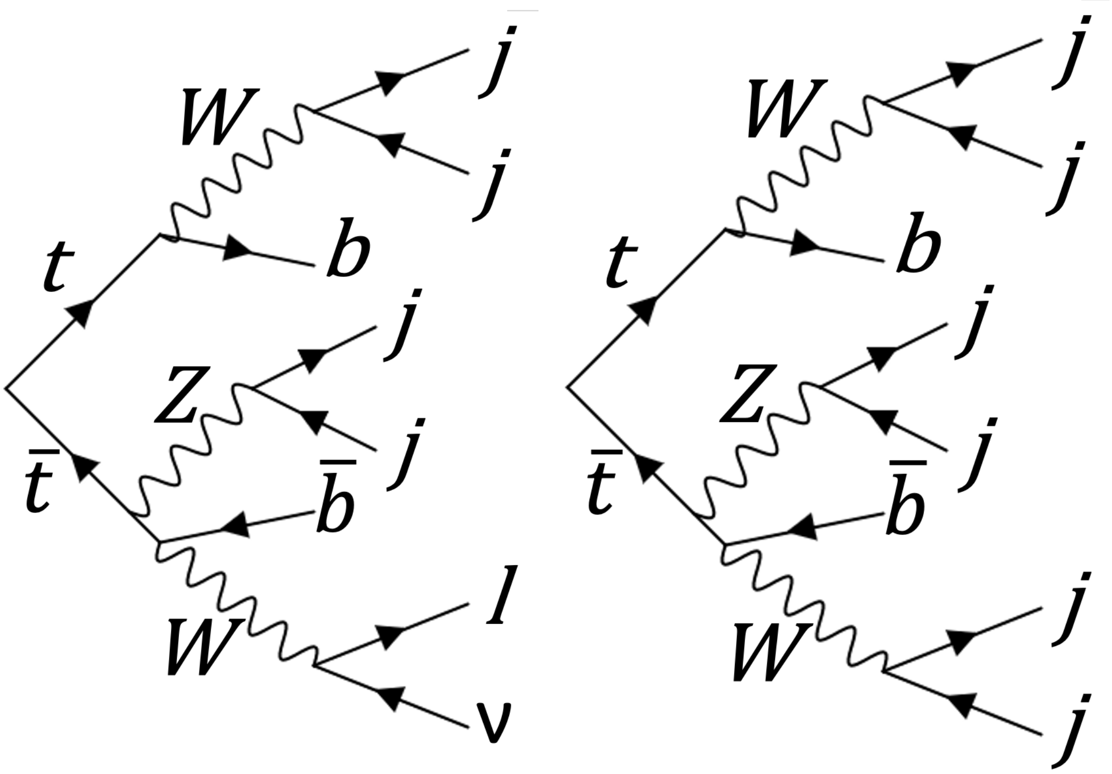


Figure 5.1.3: Example diagrams for the (left) 1ℓ channel and (right) 0ℓ channel of the $t\bar{t}Z$ process.

$t\bar{t}Z$ channel	$t\bar{t}$ fraction	Z boson fraction	branching ratio (%)
4ℓ	dileptons	dileptons	0.9
3ℓ	lepton+jets	dileptons	4.5
2ℓOS	all-hadronic	dileptons	4.6
2ℓOS	dileptons	all-hadronic + neutrinos	8.1
1ℓ	lepton+jets	all-hadronic + neutrinos	40.5
0ℓ	all-hadronic	all-hadronic + neutrinos	41.4

Table 5.1.1: Table showing how different $t\bar{t}Z$ decay channels are obtained from $t\bar{t}$ decay branching fractions and Z boson decay branching fractions.

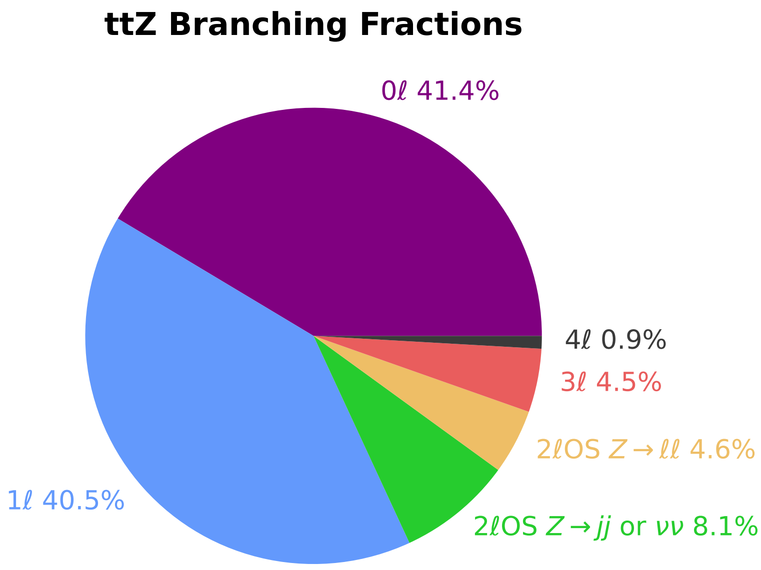


Figure 5.1.4: Pie chart showing the branching fractions of $t\bar{t}Z$ decay. This thesis focuses on the 2ℓOS channel, since it provides a compromise between high branching fraction and simplicity in reconstructing the Z boson. Figure produced in Python [70].

5.2 Signal regions

It is necessary to define specific regions with different selection criteria, to be able to optimise the MVAs in each region. Combining the separately optimised regions can lead to a better overall signal versus background separation. Separation is defined as the sum

$$\frac{1}{2}\sum_{n=1}^{Nbins}\frac{(s_i - b_i)^2}{s_i + b_i}, \tag{5.2.1}$$

where s_i is the number of signal events in bin i, b_i is the number of background events in bin i, and the sum is performed over all bins, $Nbins$. Separation is zero for identical signal and background distributions, and one for signal and background distributions without any overlap.

Signal region definition is done by imposing some initial selections common to each region, then specific selections for each region. These selections are presented later in Section 5.2.3.

5.2.1 Initial dilepton selections

The analysis in the dilepton signal region (2ℓOS) aims to select events where the top-quark pair decays hadronically (jets only) and the Z decays dileptonically into two charged leptons via $Z \to e^+e^-, \mu^+\mu^-$. Events featuring hadronically decaying tau leptons that originate directly from either the Z (via $Z \to \tau^+\tau^-$) or the W bosons from the $t\bar{t}$ system (via $W \to \tau\nu_\tau$) are removed and are not considered.

In addition to the preselection criteria of Section 4, a summary of the definitions of the dilepton signal regions is provided in Table 5.2.2. The minimum requirement on the transverse momentum for the leading and subleading lepton, in the dilepton signal regions is 30 and 15 GeV, respectively. Since a dileptonic decay of the Z boson is being searched for, an opposite-sign-same-flavour (OSSF) lepton pair is required. An invariant mass within ±10 GeV of the m_Z value as quoted in the PDG (91.12 GeV) [11] is required for this OSSF lepton pair. The sum of the two lepton charges has to be 0. Additional requirements are imposed on the total number of reconstructed jets and b-tagged jets in the event. All selected jets are required to satisfy $p_{\mathrm{T}} > 25$ GeV, as specified in the object definitions.

5.2.2 Backgrounds

The two main backgrounds to the $t\bar{t}Z$ 2ℓOS process are $t\bar{t}$ and Z+jets. The $t\bar{t}$ background originates from $t\bar{t}$ di-leptonic decay, with the radiation of two extra gluons, as shown in Figure 5.2.1(b). The Z +jets background originates from Z boson di-leptonic decay, with the radiation of three extra gluons, one of which decays to $b\bar{b}$, as shown in Figure 5.2.1(c). The inclusive cross-section for $t\bar{t}$ production at 13 TeV was measured by ATLAS in the dilepton channel to be 826.4 ± 3.6 (stat) ± 11.5 (syst) ± 15.7 (lumi) ± 1.9 (beam) pb [172]. The $Z \longrightarrow e^+e^- + \geq 5$ jets cross-section was measured by ATLAS at 13 TeV to be 0.357 ± 0.013 (stat) ± 0.069 (syst) ± 0.009 (lumi) pb [173]. The $Z \longrightarrow \mu^+\mu^-$ $+ \geq 5$ jets cross-section was measured by ATLAS at 13 TeV to be 0.354 ± 0.012 (stat) ± 0.068 (syst) ± 0.009 (lumi) pb [173]. These cross-sections are at least an order of magnitude greater than the cross-section for $t\bar{t}Z$ 2ℓOS, which is about 0.04 pb. Table 5.2.1 compares the cross-sections for $t\bar{t}Z$ 2ℓOS and its two main backgrounds. As well as diagrams for the two main backgrounds, a diagram for the $t\bar{t}Z$ 2ℓOS signal is shown in Figure 5.2.1. The selection requirements imposed to reduce background while keeping signal are introduced later in Section 5.2.3.

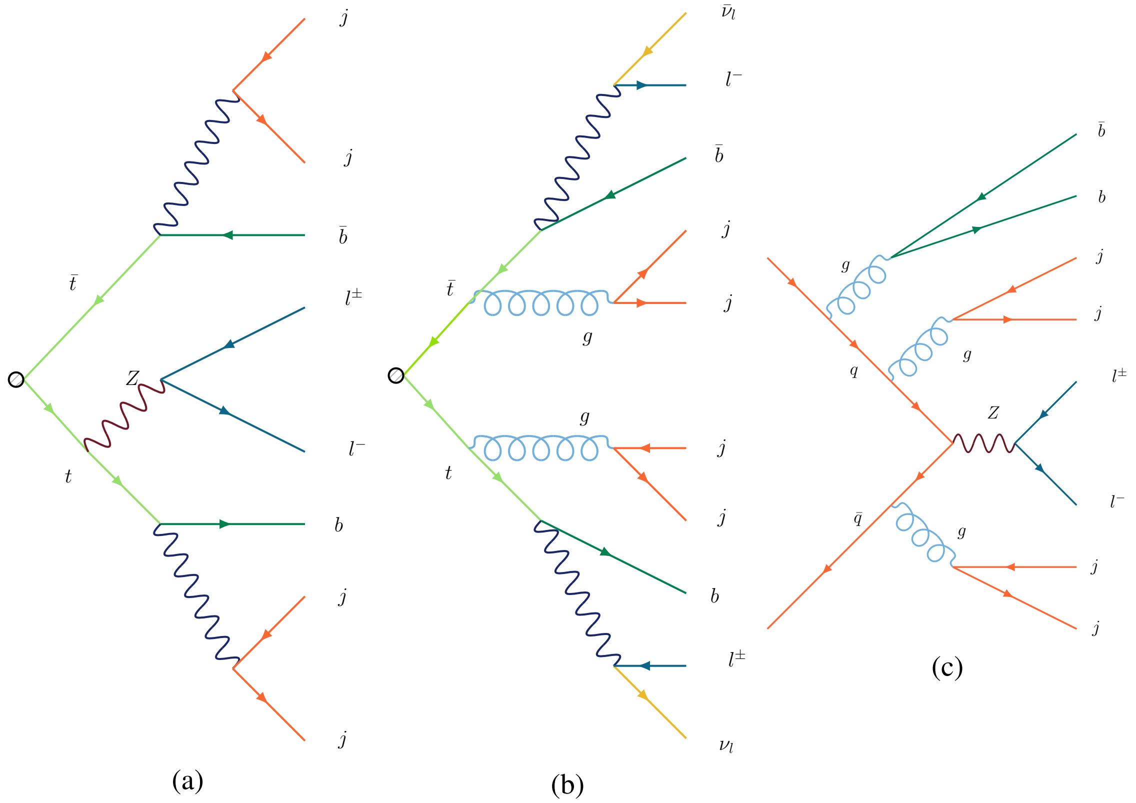


Figure 5.2.1: Feynman diagrams for the signal and main backgrounds of the $t\bar{t}Z$ 2ℓOS channel. (a) $t\bar{t}Z$ 2ℓOS decay, which is the signal process for this thesis. (b) $t\bar{t}$. (c) Z+jets. Diagrams produced using Ref. [68].

Process	Cross-section (pb)
$t\bar{t}Z$ 2ℓOS	$\approx$0.04
$t\bar{t}$ dileptonic	826.4 ± 3.6 (stat) ± 11.5 (syst) ± 15.7 (lumi) ± 1.9 (beam) [172]
$Z \rightarrow e^+e^- + \geq 5$ jets	0.357 ± 0.013 (stat) ± 0.069 (syst) ± 0.009 (lumi) [173]
$Z \rightarrow \mu^+\mu^- + \geq 5$ jets	0.354 ± 0.012 (stat) ± 0.068 (syst) ± 0.009 (lumi) [173]

Table 5.2.1: The main signal and background processes involved in the $t\bar{t}Z$ 2ℓOS analysis, along with their cross-sections.

5.2.3 Definition of dilepton signal regions

A combination of three different signal regions are used and will be referred to as 2ℓ-Z-$1b6j$, 2ℓ-Z-$2b5j$ and 2ℓ-Z-$2b6j$. Diagrams showing the $t\bar{t}$ decay within $t\bar{t}Z$ that leads to these regions are shown in Figure 5.2.2. The production of 2 b-jets along with 4 other jets (2b6j) corresponds to nominal all-hadronic $t\bar{t}$ decay. 2ℓ-Z-2b5j can occur when a single jet goes out of acceptance in the ATLAS detector. 2ℓ-Z-1b6j can occur when a b-jet is not identified.

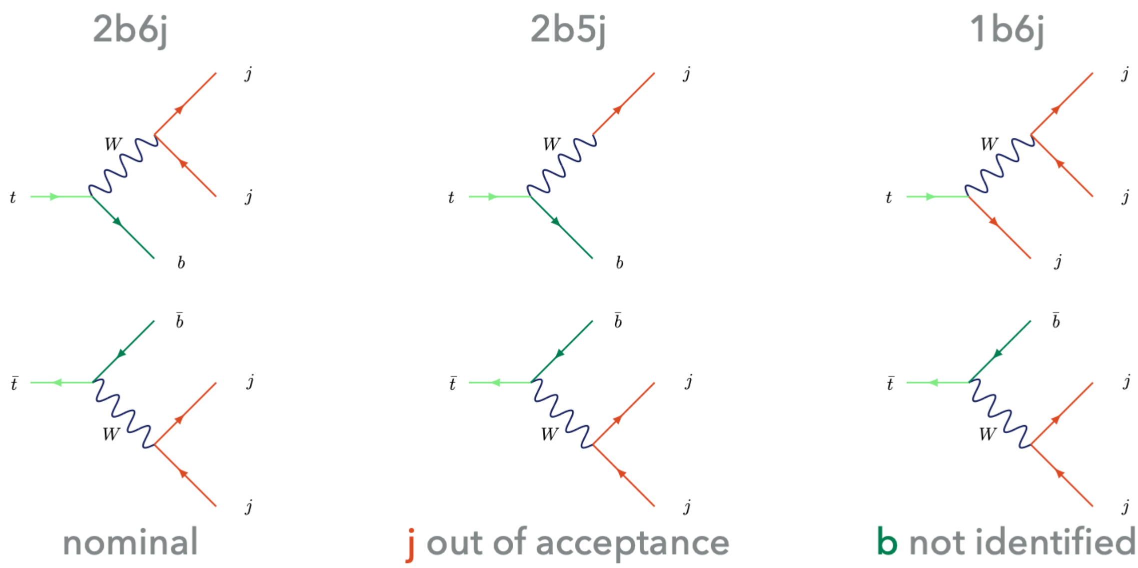


Figure 5.2.2: Feynman diagrams depicting the $t\bar{t}$ decay within $t\bar{t}Z$ that leads to the three separate regions used in this thesis. The Z boson within $t\bar{t}Z$ decays dileptonically in each case, therefore is not shown. The production of 2 b-jets along with 4 other jets (2b6j) corresponds to nominal all-hadronic $t\bar{t}$ decay. 2ℓ-Z-2b5j can occur when a single jet goes out of acceptance in the ATLAS detector. 2ℓ-Z-1b6j can occur when a b-jet is not identified.

There is a mix of Z+b and Z+c events in the 2ℓ-Z-1b6j region (as opposed to the 2ℓ-Z-2b5j and 2ℓ-Z-2b6j containing mostly Z+b) because one c-jet from Z+c can fake a b-jet. There is little $t\bar{t}$ background in the 2ℓ-Z-1b6j region (compared to 2ℓ-Z-2b5j and 2ℓ-Z-2b6j) because $t\bar{t}$ hadronic decays are very likely be identified with 2 b-jets, as introduced in Section 2.2.3. Validation plots showing the distribution of several variables in the dilepton regions can be found in Section 5.5.

Variable	2ℓ-Z-$1b6j$	2ℓ-Z-$2b5j$	2ℓ-Z-$2b6j$
N_ℓ ($\ell = e, \mu$)		$= 2$	
		1 OSSF lepton pair with $\lvert m^Z_{\ell\ell} - m_Z \rvert < 10$ GeV	
p_{T} (ℓ_1)		> 30 GeV	
p_{T} (ℓ_2)		> 15 GeV	
$N_{\mathrm{jets}}(p_{\mathrm{T}} > 25$ GeV$)$	≥ 6	$= 5$	≥ 6
$N_{b-\mathrm{jets}}$@77%	$= 1$	≥ 2	≥ 2

Table 5.2.2: The definitions of the dilepton signal regions. A combination of the regions 2ℓ-Z-$1b6j$, 2ℓ-Z-$2b5j$ and 2ℓ-Z-$2b6j$ is used.

5.3 Visualisation of $t\bar{t}Z$

Event displays are useful tools to help one visualise final state topology and with the debugging of reconstruction and simulation software and physics analysis. For the $t\bar{t}Z$ analysis in particular, event displays can be used to showcase suitable $t\bar{t}Z$ candidates, and show the difference between channels. Though this thesis focuses on the $t\bar{t}Z$ 2ℓOS channel, event displays were also produced for the $t\bar{t}Z$ 3ℓ and 4ℓ channels. The 3ℓ and 4ℓ event displays were published as part of the first analysis of the $t\bar{t}Z$ 3ℓ and 4ℓ channels with full Run 2 data [37]. The 2ℓOS event display is not yet published, because the 2ℓOS is not yet published. Event displays were all produced by the author of this thesis, with the ATLAS Virtual Point 1 software [174]. Through event displays, different physics objects can be seen to leave different signatures in the ATLAS detector.

2ℓOS event display

The selection requirements imposed to find a suitable 2ℓOS $t\bar{t}Z$ candidate event to display are shown in Table 5.3.1, along with the reasoning behind each selection requirement. These selections were imposed on events present in the final 2ℓ-Z-$2b6j$ signal region, as described in Section 5.2.3. These are more stringent requirements, to select specific candidates suitable for visualisation. Electron, jet and muon reconstruction are briefly introduced in Section 4.3.1, Section 4.3.2 and Section 4.3.3, respectively. DL1r and b-tagging algorithms more generally are briefly introduced in Section 4.3.2.

Selection requirement	Reasoning
$N_\mu == 2$	Muon reconstruction is cleaner than electrons
Muons are opposite flavour with $\lvert m_{\ell\ell}^{Z} - m_Z \rvert < 10$ GeV	Muons should come from a Z boson
$p_{\mathrm{T}}\ (\ell_1) > 30$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}\ (\ell_2) > 15$ GeV	Minimum lepton p_{T} requirements
$N_{b\mathrm{-jets}}$@77% DL1r == 2	A b-jet from the decay of each top quark
$N_{\mathrm{jets}}(p_{\mathrm{T}} > 25\,\mathrm{GeV}) == 6$	Two b-jets + four jets from two hadronic top decays
runNumber $\geq$ 315197	data17 or data18

Table 5.3.1: Selection requirements imposed to find a suitable 2ℓOS $t\bar{t}Z$ candidate event to display, along with the reasoning behind each selection requirement.

Since 2017 and 2018 data were being analysed for the first time in the $t\bar{t}Z$ 2ℓOS channel, it was desirable for the event displays to be from a 2017 or 2018 data run. After applying the selections in Table 5.3.1, the event with the highest MVA output score is shown in Figure 5.3.1. The reconstructed muons (Section 4.3.3) represented by the red lines have a transverse momentum around 290 and 90 GeV through the detector, with a dimuon mass of 87.8 GeV. The yellow bars indicate energy deposits in the calorimeter (Section 3.5). From these deposits six jets (Section 4.3.2) are identified and represented by cones. The blue jets, with transverse momentum around 250 and 155 GeV, are identified as having originated from b-quarks (Section 4.3.2). The yellow jets, with transverse momenta around 190, 80, 60 and 60 GeV respectively are not identified as having originated from b-quarks. The missing transverse momentum (Section 4.3.4) represented by the dotted white line has a magnitude around 80 GeV. The MVA used in described in detail throughout Chapter 6. This event display is to be published in an upcoming ATLAS paper on the $t\bar{t}Z$ process.

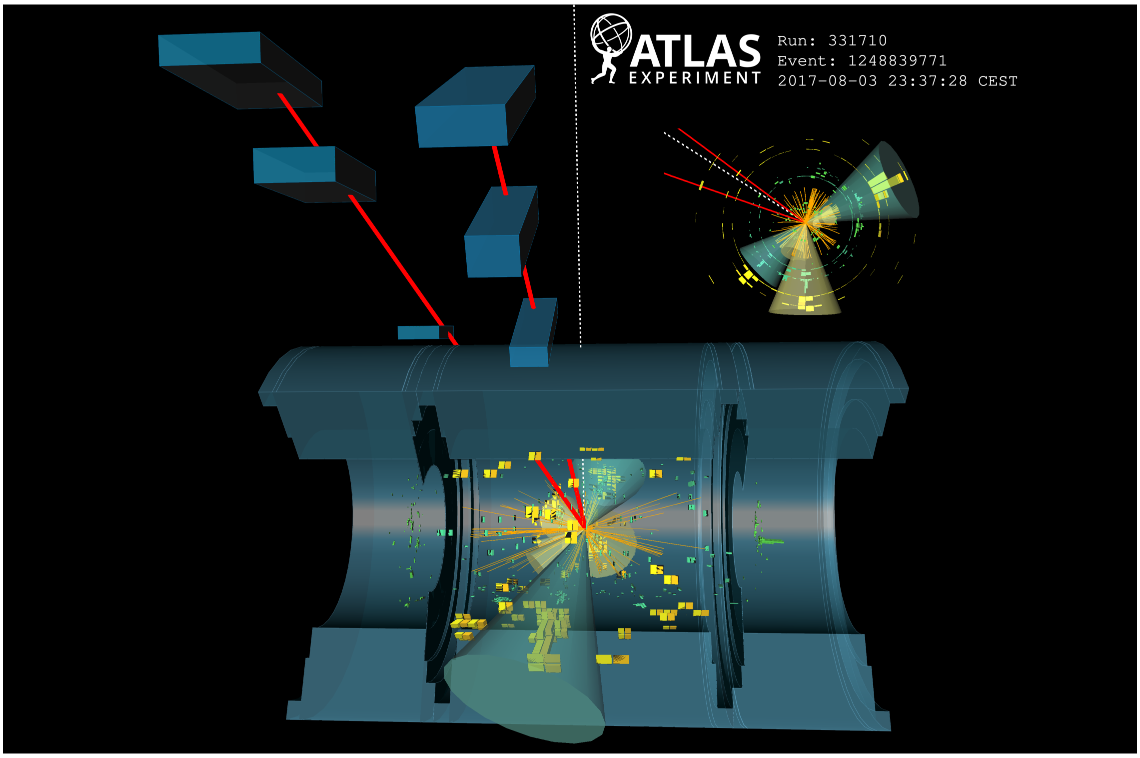


Figure 5.3.1: Display of event 1248839771 in run 331710 recorded by ATLAS in pp collisions with LHC stable beams at a centre-of-mass energy of 13 TeV on August 3rd, 2017. The topology of this candidate corresponds to $t\bar{t}Z$ production in the 2ℓOS channel.

3ℓ event display

The selection requirements imposed to find a suitable 3ℓ $t\bar{t}Z$ candidate event to display are shown in Table 5.3.2, along with the reasoning behind each selection requirement. Electron, jet and muon reconstruction are briefly introduced in Section 4.3.1, Section 4.3.2 and Section 4.3.3, respectively. MV2c10 and b-tagging algorithms more generally are briefly introduced in Section 4.3.2.

Selection requirement	Reasoning
$N_e < 3$	Only one possible lepton combination can reconstruct a Z boson
$N_\mu < 3$	Only one possible lepton combination can reconstruct a Z boson
$N_\mu == 2$	Muon reconstruction is cleaner than electrons
$N_e == 1$	Muon reconstruction is cleaner than electrons
1 OSSF lepton pair with $\lvert m^Z_{\ell\ell} - m_Z \rvert < 10$ GeV	Muons should come from a Z boson
$p_{\mathrm{T}}\ (\ell_1) > 27$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}\ (\ell_2) > 20$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}\ (\ell_3) > 20$ GeV	Minimum lepton p_{T} requirements
$N_{b\text{-jets}}$@70% MV2c10 == 2	A b-jet from the decay of each top quark
$N_{\text{jets}}(p_{\mathrm{T}} > 25\,\text{GeV}) == 4$	Two b-jets + two jets from a hadronic top decay
runNumber $\geq$ 315197	data17 or data18

Table 5.3.2: Selection requirements imposed to find a suitable 3ℓ $t\bar{t}Z$ candidate event to display, along with the reasoning behind each selection requirement.

The 70% b-tagging working point with the MV2c10 algorithm was used for the 3ℓ event display (as opposed to the 77% working point with DL1r for 2ℓOS) to reflect the different working points and algorithms in the different channels. Since 2017 and 2018 data were being used for the first time in a $t\bar{t}Z$ analysis, it was be desirable for the event displays to be from a 2017 or 2018 data run. 2017 data contained 11 events passing these selection requirements, whilst 2018 data contained 12. From these 23 events, the event with the highest leptonic top reconstruction weight (`pseudo_top_mblv`) was chosen, and the event display is shown in Figure 5.3.2. The reconstructed muons (Section 4.3.3) represented by the red lines have a transverse momentum around 250 and 65 GeV through the detector, with a dimuon mass of 90.6 GeV. The yellow bars indicate energy deposits in the calorimeter (Section 3.5). From these deposits four jets (Section 4.3.2) are identified and represented by cones. The blue jets, with transverse momentum around 250 and 60 GeV, are identified as having originated from b-quarks (Section 4.3.2). The yellow jets, with transverse momenta around 65 and 55 GeV are not identified as having originated from b-quarks. A reconstructed electron (Section 4.3.1) is represented by the green track and energy deposit. The missing transverse momentum (Section 4.3.4) represented by the dotted white line has a magnitude around 30 GeV. The difference in azimuthal angle, ϕ, between the Z boson and the lepton from the leptonic top decay, $\Delta\phi(Z, t_{\text{lep}})$, in units of rad$/\pi$ is 0.813533. This event display was published in the 2020 ATLAS paper on the $t\bar{t}Z$ process [37].

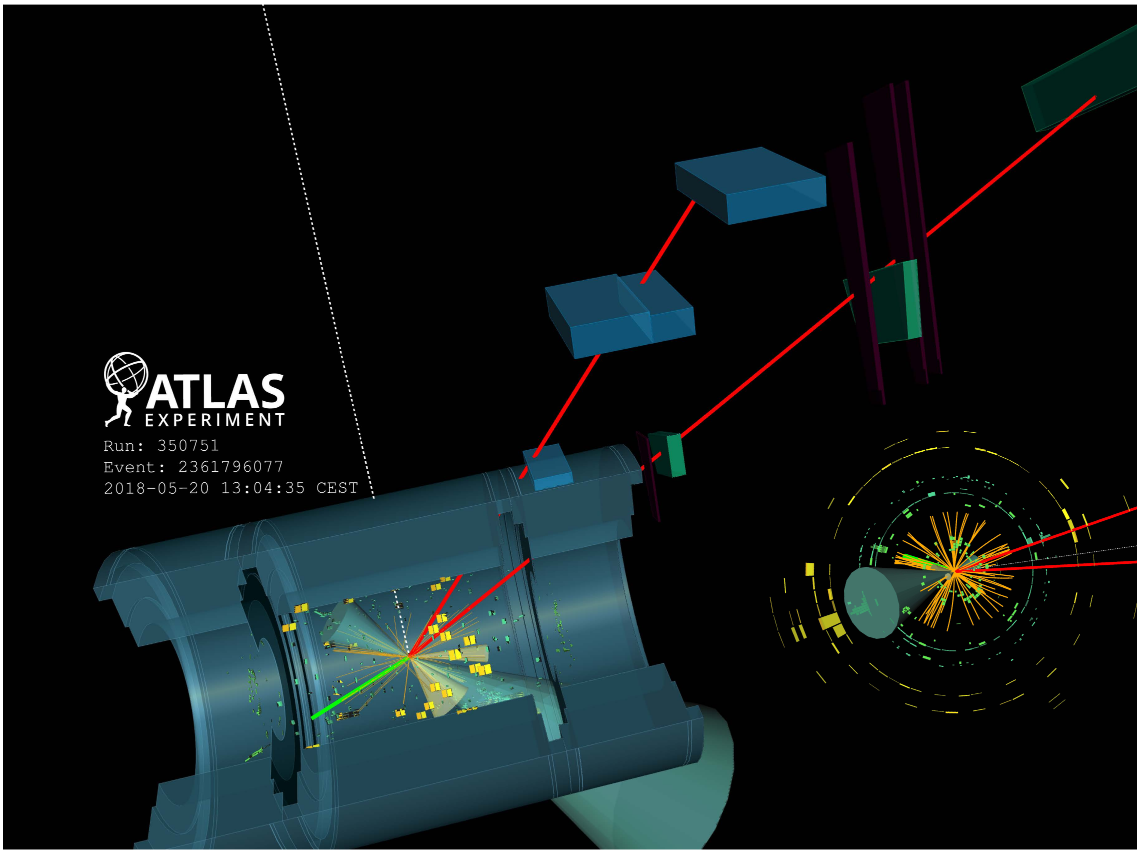


Figure 5.3.2: Display of event 2361796077 in run 350751 recorded by ATLAS in pp collisions with LHC stable beams at a centre-of-mass energy of 13 TeV on May 20th, 2018. The topology of this candidate corresponds to $t\bar{t}Z$ production in the 3ℓ channel.

4ℓ event display

The selection requirements imposed to find a suitable 4ℓ $t\bar{t}Z$ candidate event to display are shown in Table 5.3.3, along with the reasoning behind each selection requirement. Electron, jet and muon reconstruction are briefly introduced in Section 4.3.1, Section 4.3.2 and Section 4.3.3, respectively. MV2c10 and b-tagging algorithms more generally are briefly introduced in Section 4.3.2.

Selection requirement	Reasoning
$N_\mu == 3$	Muon reconstruction is cleaner than electrons
$N_e == 1$	Muon reconstruction is cleaner than electrons
≥1 OSSF lepton pair with $\lvert m_{\ell\ell}^{Z} - m_Z \rvert < 10$ GeV	A pair of opposite-sign muons should come from a Z boson
for all OSSF combinations: $m_{\mathrm{OSSF}} > 10$ GeV	Remove low-mass resonances and photon conversions
$p_{\mathrm{T}}(\ell_1) > 27$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}(\ell_2) > 20$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}(\ell_3) > 10$ GeV	Minimum lepton p_{T} requirements
$p_{\mathrm{T}}(\ell_4) > 7$ GeV	Minimum lepton p_{T} requirements
$\ell\ell^{\mathrm{non-}Z}$ is $e^\pm \mu^\mp$	Only one final state opposite-sign-same-flavour lepton pair
$N_{b\mathrm{-jets}}$@85% MV2c10 == 2	A b-jet from the decay of each top quark
$N_{\mathrm{jets}}(p_{\mathrm{T}} > 25\,\mathrm{GeV}) == 2$	No extra jets
runNumber ≥ 315197	data17 or data18

Table 5.3.3: Selection requirements imposed to find a suitable 4ℓ $t\bar{t}Z$ candidate event to display, along with the reasoning behind each selection requirement.

The 85% b-tagging working point with the MV2c10 algorithm was used for the 4ℓ event display (as opposed to the 77% working point with DL1r for 2ℓOS) to reflect the different working points and algorithms in the different channels. Since 2017 and 2018 data were being used for the first time in a $t\bar{t}Z$ analysis, it was be desirable for the event displays to be from a 2017 or 2018 data run. 2017 and 2018 data each contained one event passing these selection requirements from the final 4ℓ-DF-2b signal region. From these two events, the event with the highest sum of jet `MV2c10` score was chosen, and the event display is shown in Figure 5.3.3. The reconstructed muons (Section 4.3.3) represented by the red lines have a transverse momentum around 140, 50 and 15 GeV through the detector, with the pair identified as having originated from a Z boson having a dimuon mass of 89.0 GeV. The yellow bars indicate energy deposits in the calorimeter (Section 3.5). From these deposits two jets (Section 4.3.2) are identified and represented by cones. The blue jets, with transverse momentum around 120 and 80 GeV, are identified as having originated from b-quarks (Section 4.3.2). A reconstructed electron (Section 4.3.1) with transverse momentum around 80 GeV is represented by the green track and energy deposit. The missing transverse momentum (Section 4.3.4) represented by the dotted white line has a magnitude around 35 GeV. The difference in azimuthal angle, ϕ, between the Z boson and $t\bar{t}$ system, $\Delta\phi(Z, t\bar{t})$, in units of rad$/\pi$ is 0.595878. This event display was published in the 2020 ATLAS paper on the $t\bar{t}Z$ process [37].

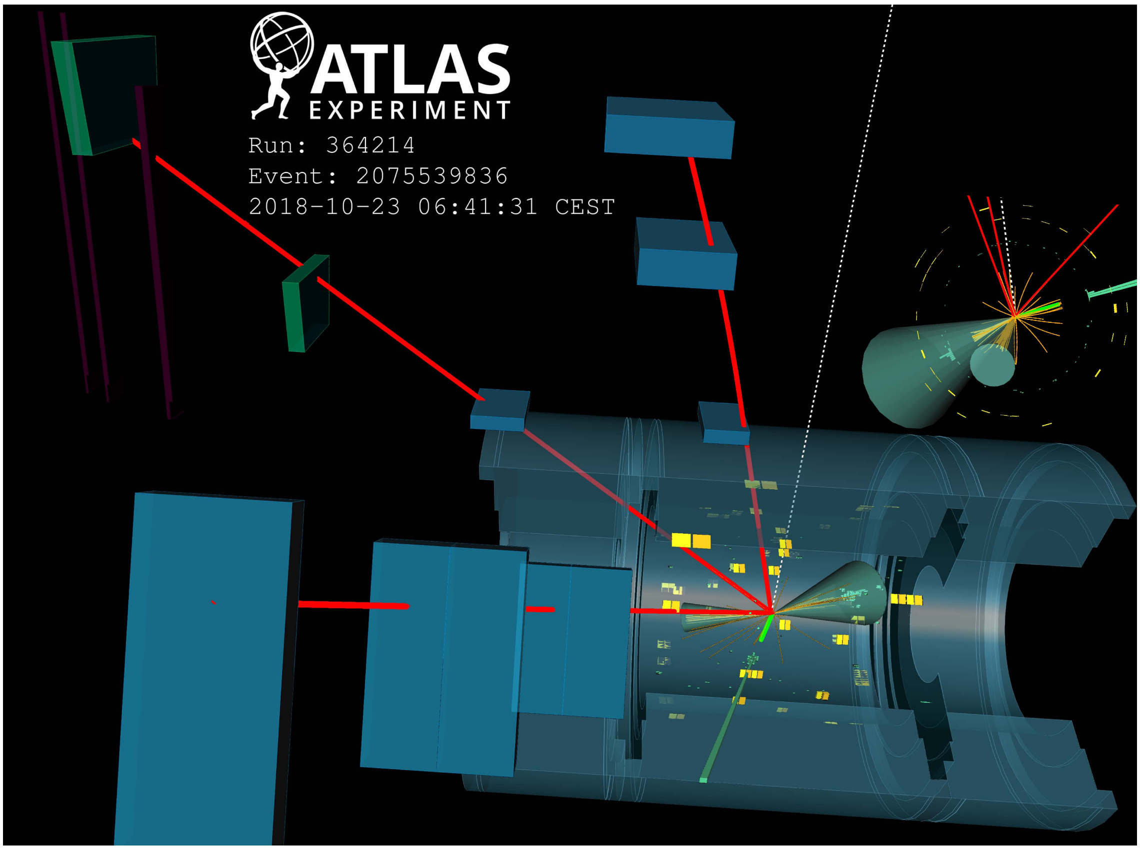


Figure 5.3.3: Display of event 2075539836 in run 364214 recorded by ATLAS in pp collisions with LHC stable beams at a centre-of-mass energy of 13 TeV on October 23rd, 2018. The topology of this candidate corresponds to $t\bar{t}Z$ production in the 4ℓ channel.

5.4 Machine Learning

Machine Learning (ML) is a key part of the $t\bar{t}Z$ 2ℓOS analysis strategy due to the fact that background processes have higher cross-sections than the signal process, therefore it is difficult to isolate signal from background. For example, as shown in Table 5.2.1 the cross-section for $t\bar{t}Z$ is $\approx$3 orders of magnitude lower than $t\bar{t}$ (one of the main backgrounds in the $t\bar{t}Z$ 2ℓOS channel).

5.4.1 Introduction to Machine Learning

An analysis in the $t\bar{t}Z$ 2ℓOS region presents two main characteristics:

- the cross-sections for background processes can be $\approx$3 orders of magnitude higher than that of the signal;
- no single variable provides high enough signal versus background separation (to be shown in Section 5.5).

High signal versus background separation is needed to be able to distinguish signal from background distributions. Separation quantifies the difference in height between signal and background distributions in each bin of a histogram. The analysis therefore uses a Multi-Variate Analysis (MVA) in the 2ℓOS channel to obtain a variable with larger separation between signal and background than an individual variable could achieve.

An MVA technique is an algorithm that takes a set of the input variables and builds an output value, called MVA output. The output has higher separation power than any input variable alone.

Training and testing

Two phases are needed when using an MVA technique: the training phase and the testing phase. Monte Carlo samples for background and signal need to be divided into two parts, usually with the same number of events. One of the sets of samples is used during the training phase and the other during the testing phase. During the training phase, input variables for signal and background events are provided to the MVA and the internal weights of the MVA are optimised to provide the best signal versus background separation.

Overtraining

After the training phase, an overtraining check is needed to verify the MVA technique's ability to identify the signal in a statistically independent sample. If the statistics in the training sample are low, the MVA technique can train on the non-physical statistical fluctuations of the training sample, which is called overtraining. In this case, the separation power on the training sample is unrealistically higher than in the testing sample. If the MVA technique is correctly trained, the separation power for the testing sample is the same as for the training sample.

Events used for MVA training can not be used in the analysis to estimate the MVA response, due to possible bias caused by overtraining. Dropping half the statistics used for MVA training would lead to increasing Monte Carlo statistical uncertainties in the analysis. To avoid this, cross training can be used.

Cross training

In cross training, the same MVA technique is used twice where the roles of the different events are switched. The first half of MC events are used to train the first MVA and test the second one, and vice versa. In the testing phase, the event is categorised by the MVA not used for training. If cross training is used, the whole MC sample can be used in the analysis. This technique is used in the results of Chapter 6.

Choosing input variables

Choosing input variables for the MVA is a balance between keeping the MVA robust against overtraining and obtaining the best separation power of the MVA output. Increasing the number of input variables leads to higher separation, but the MVA is more likely to be overtrained.

Performance metric

The dependence of rejected background events fraction on the fraction of accepted signal is quantified in a receiver operation characteristic (ROC) curve. "rejected background" is also called "true negatives" and "accepted signal" is also called "true positives". A curve further towards the top right indicates better performance. This is equivalent to a higher area under curve (AUC). It is also useful to plot the ROC curves for training and testing data sets separately, to perform another check for overtraining. If the testing line largely overlaps with the training line there is no noticeable overtraining.

The ratio of signal to background (S/B) that can be achieved using MVA output is another useful performance metric. The number of signal events divided by the square root of the number of background events is defined as statistical significance, and is another useful performance metric for MVAs. S/$\sqrt{S+B}$ is another possible metric, but does not tend to be used in cases where the background is dominant, which is the case in the $t\bar{t}Z$ 2ℓOS channel.

5.4.2 Boosted Decision Trees

Boosted Decision Trees are used as the baseline MVA algorithm for this thesis, and they were used in the previous ATLAS $t\bar{t}Z$ 2ℓOS measurement [36].

Decision Trees

Decision trees are a type of learning algorithm with a tree-like structure. A succession of questions are asked to split the input data into purer samples of separated signal and background. Each question uses a single discriminating variable to decide if an event is signal-like or background-like. In the end, an upside-down-tree like structure is formed with leaf nodes classified as signal or background. Training a Decision Tree is a process to define the cut criteria for each node. It is a binary process, starting with a cut on a variable that gives the best separation and then repeating the process for each subsample creating two new nodes at each step. The division is stopped once a certain node has reached a minimum number of events or maximum signal purity. Decision trees are amongst the most popular MVA techniques [175]. A pictorial representation of a decision tree is given in Figure 5.4.1.

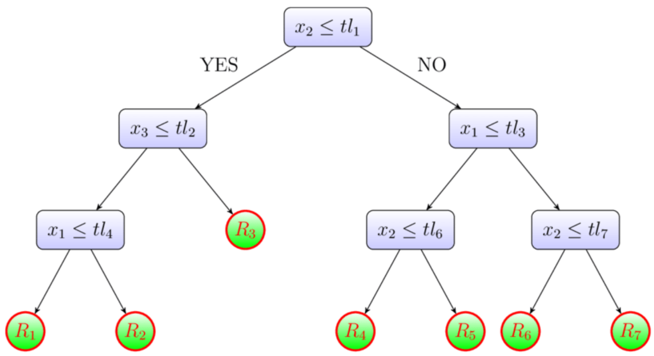


Figure 5.4.1: Diagram of a Decision Tree using three variables (x_1, x_2, x_3). These variables are compared to six numbers ($tl_1, tl_2, tl_3, tl_4, tl_6, tl_7$) whilst making decisions. Each decision aims to divide the dataset into parts that are more signal-like and others that are more background-like. Some of the leaf nodes will be rich in signal and others rich in background. Figure taken from Ref. [176].

Boosting

More information can be obtained by combing multiple decision trees, by the process of boosting, which consists of giving events in the incorrect leaf node a larger weight than events in the correct leaf node. Boosting is applied a number of times and each time a new tree is built. Using many decision trees together in the training phase is called an “ensemble" or “forest". A “likelihood” estimator is constructed from all the trees in the forest for the event being signal or background based on how often it is classified as signal. A picture of a Boosted Decision Tree (BDT) is shown in Figure 5.4.2.

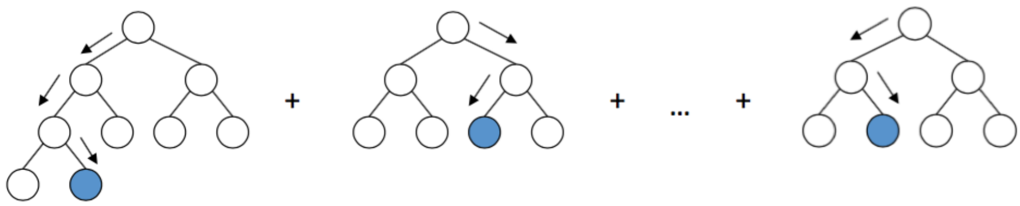


Figure 5.4.2: An ensemble of Decision Trees added together gives a Boosted Decision Tree (BDT), namely a Gradient BDT. Each Decision Tree can come to a slightly different conclusion about the leaf nodes that are rich in signal. Their results are added together to give a better total result. Figure taken from Ref. [177].

5.4.3 Neural Networks

A Neural Network (NN) is another type of MVA technique, typically providing more freedom than BDTs in terms of tuneability. Neural Network (NN)s are MVAs based on the idea of neurons in the human brain. A NN is called a Deep Neural Network (DNN) if it has more than one hidden layer of neurons, to be explained in this section. The use of Deep Neural Networks or more complex NNs is becoming more commonplace in particle physics, therefore it is crucial to test their performance for the analysis of this thesis.

Binary deep neural networks

Figure 5.4.3 shows a simple binary DNN architecture, which consists of an input layer, hidden layers and an output layer. The number of nodes in the input layer corresponds to the number of input variables or input features. Typically, the number of nodes in a hidden layer is greater than the number of nodes in the input layer. A NN can learn more information by adding more hidden layers, thus becoming a DNN. The more hidden layers, the “deeper" the network. A binary DNN contains a single node in the output layer. The number given by the single node of the output layer can be thought of as the probability that a particular event passed through the network is signal.

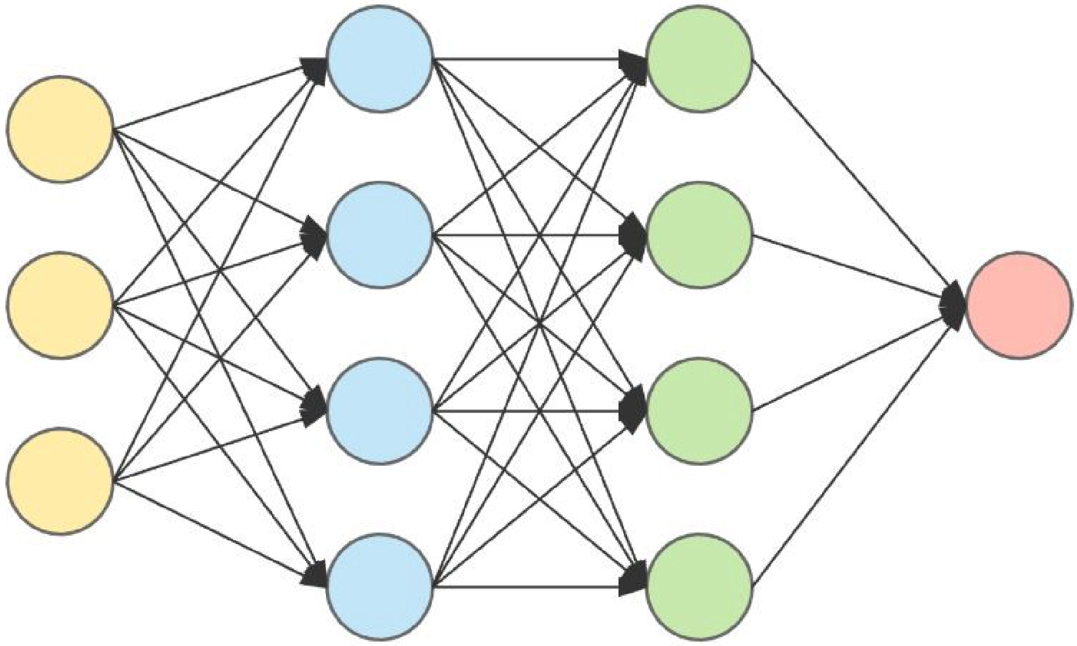


Figure 5.4.3: Diagram of a deep neural network. In this case, three input variables are used and one output. Figure from Ref. [178].

Multiclassification

The schematic diagrams of the MVA techniques shown so far have been to classify between an individual signal and the total background. If there are a number of main backgrounds that are sufficiently different from each other, it may be beneficial to classify between signal, background 1 and background 2 (or however many different backgrounds there are). This approach is called multiclassification. A multiclass NN is shown in Figure 5.4.4. Multiclass BDTs also exist.

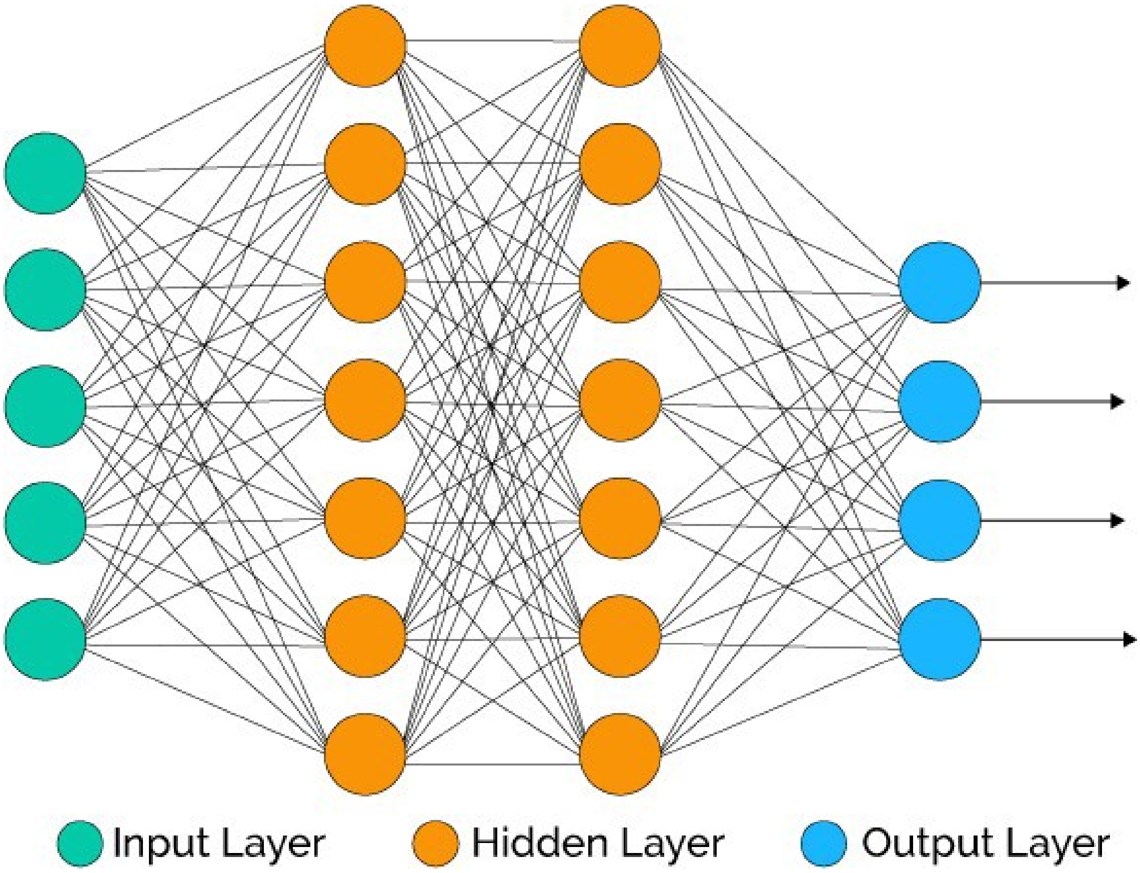


Figure 5.4.4: Diagram of a multiclass deep neural network. This is similar to the deep neural network of Figure 5.4.3, but with extra nodes in the output layer. The number from each node in the output layer can be thought as different probabilities, e.g. (1) probability of being signal, (2) probability of being $t\bar{t}$ background, (3) probability of being Z+jets background, (4) probability of being other background. Multiple nodes in the output layer makes a network "multiclass". More output layer nodes can be added to make the network classify into more classes. In this case, five input variables are used. There are two hidden layers in this network.

Loss function

NNs learn by minimising a "loss function". An epoch is a step the NN takes in the process of trying to minimise the loss function. The learning process is shown by "loss curves". If the "Train" loss curve is significantly below or significantly divergent from the "Test" loss curve, that is a sign of overtraining, i.e. the network performs much better on the training data set than the testing data set. A widely used loss function in particle physics is crossentropy. Different loss functions are used for binary classification and multiclassification, e.g. binary crossentropy and categorical crossentropy.

NN settings

Possible settings for NNs trained using tensorflow [179] through keras [180] are described in Table 5.4.1.

Option	Values	Description
InputScaling	minmax	inputs are scaled into a range between 0 and 1
Folds	2	how many folds (k-folding) should be performed during training
Nodes	50,50,50,50	Comma-separated list of neurons for each layer
Loss	categorical crossentropy	Loss function which is used in the training of a model
Epochs	100	Number of training epochs
LearningRate	0.001	Initial learning rate for the training of a model
BatchSize	32	Batch size used in training of a model
ValidationSize	0.2	Relative size of the validation set used during training of a model
Patience	30	Number of epochs with no improvement after which training will be stopped
MinDelta	0.001	Minimum change in the monitored quantity to qualify as an improvement
DropoutIndice	1,3	Layer indeces at which Dropout layers are added
DropoutProb	0.1	Probability of dropout
OutputSize	3	Number of neurons in the output layer
OutputActivation	softmax	Activation function in the output layer of a model
Metrics	Accuracy	Comma-separated list of metrics to be evaluated during training
ModelBinning	20,0,1	Custom binning using a fixed bin width in the format nbins,x_low,x_high
SmoothingAlpha	0.1	to be applied to smooth labels according using Y=Y(1-alpha)+alpha/K

Table 5.4.1: Possible settings used in DNN training using tensorflow [179] through keras [180].

5.4.4 MVA strategy

The strategy for the analysis in the $t\bar{t}Z$ 2ℓOS channel is as follows:

1. design and test BDTs for binary classification of the $t\bar{t}Z$ 2ℓOS signal against all backgrounds;
2. design and test DNNs for binary classification of the $t\bar{t}Z$ 2ℓOS signal against all backgrounds;
3. design and test DNNs for multiclassification of the $t\bar{t}Z$ 2ℓOS signal, $t\bar{t}$ background and Z boson background;
4. design and test the inclusion of new variables to use in DNNs for multiclassification of the $t\bar{t}Z$ 2ℓOS signal, $t\bar{t}$ background and Z boson background;
5. compare the performance of these different MVAs.

With the outputs of these MVAs a profile-likelihood fit to extract the $t\bar{t}Z$ 2ℓOS cross-section can be made (Section 5.7.2).

5.5 Initial MVA input variables

In the analysis, a set of 19 input variables has been considered for use in the initial MVAs. Their definitions can be found in Table 5.5.1.

Variable	Definition
ΔR_{ll}	ΔR between the two leptons
p_T^{ll}	p_T of the lepton pair
$\eta_{\ell\ell}$	η of dilepton system
p_T^{4jet}	p_T of the fourth jet
p_T^{5jet}	p_T of the fourth jet
p_T^{6jet}	p_T of the sixth jet
p_T^{1b}	p_T of the first b-jet. Jets are ordered according to p_T
H_T^{6jets}	sum of jet p_T, up to 6 jets
$N_{jet pairs}^{V mass}$	number of jet pairs with mass within a window of 30 GeV around 85 GeV
$N_{\mathrm{bjj}}^{\mathrm{top-mass}}$	number of 3 jets combinations (with exactly 1 b-tag jet) close to the top-quark mass ($\lvert M_{bjj} - M_{top}\rvert < 15$ GeV) and ($\lvert M_{jj} - M_W\rvert < 15$ GeV)
$\Delta R_{\mathrm{ave}}^{\mathrm{jj}}$	average ΔR for all jet pairs
ΔR_{bb}	cone between two jets with the highest b-tagging weight in the event
$M_{\mathrm{jj}}^{\mathrm{MindR}}$	mass of the combination between any two jets with the smallest ΔR
M_{bb}	mass of the two jets with the highest b-tag weight
$M_{\mathrm{uu}}^{\mathrm{Ptord}}$	mass of the two untagged jets with the highest p_T
M_W^{avg}	sum of the two closest 2 jet invariant masses from from jjj_1 and jjj_2 divided by 2. Not used for 2ℓ-Z-2b5j events
$\mathrm{Cent}_{\mathrm{jet}}$	scalar sum of $\mathrm{p_T}$ divided by sum of E for all jets
$H1$	First Fox-Wolfram moment, given by Equation 5.5.1 [181]
$\max M_{\mathrm{lepb}}^{\mathrm{MindR}}$	maximum mass between a lepton and the tagged jet with the smallest ΔR

Table 5.5.1: The definitions of all variables considered in the 2ℓOS channel. Jets and leptons are ordered by their p_T from the highest one. To suppress the effect of mismodelling in events with high jet multiplicity, only the first 8 jets ordered by p_T are considered when calculating these variables.

The first Fox-Wolfram momentum is defined by:

$$H1 = \sum_{i,j} \frac{\vec{p}_i \cdot \vec{p}_j}{E_{vis}^2}, \tag{5.5.1}$$

where $\vec{p}_i$ and $\vec{p}_j$ are 3-momenta of i-th and j-th object (jet or lepton) and E_{vis} is all visible energy in the event. Visible energy includes all jets and leptons (even bad jets/leptons), but not E_T^{miss}. Autocorrelations i=j are included. Correlations are only included once, i.e. the correlation between object 0 and 1 is only considered once, not 0,1 and 1,0.

Pre-fit distributions and separation of $t\bar{t}Z$ 2ℓOS signal and total background distributions for all BDT input variables were produced by the author. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. The error bars include statistical and systematic uncertainties. As an example, the input variable with highest separation is shown in Figure 5.5.1, which is ΔR_{ll}. In all 3 signal regions, there is agreement between data and MC within uncertainties.

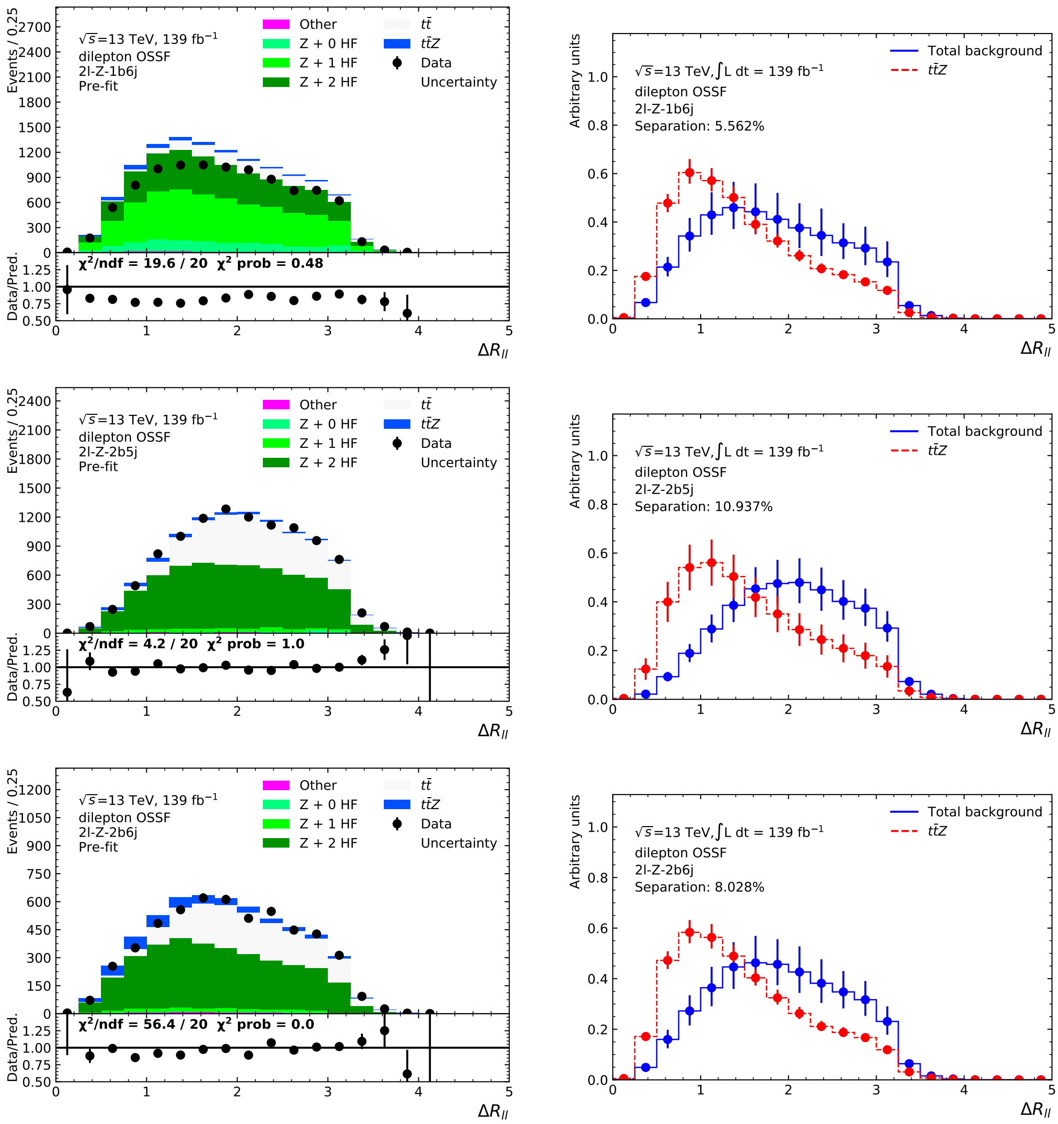


Figure 5.5.1: ΔR separation between the leptons in the regions of the dilepton OSSF channel in the regions 2ℓ-Z-2b6j (top row), 2ℓ-Z-2b5j (middle row), 2ℓ-Z-1b6j, showing the Data versus MC comparison (left column) and the separation before applying the MVA (right column).

The separations achieved by the input variables are summarised in Table 5.5.2. Table 5.5.2 orders the variables by separation in the 2ℓ-Z-2b6j region, since this is the most sensitive to $t\bar{t}Z$ 2ℓOS signal. In the input variables, no signal versus background separation >11% is shown, motivating that an MVA output variable that can provide higher signal versus background separation is desirable.

Rank	Variable	2ℓ-Z-2b6j (%)	2ℓ-Z-2b5j (%)	2ℓ-Z-1b6j (%)
1	ΔR_{ll}	8.1	10.9	5.6
2	p_T^{ll}	7.8	10.8	5.3
3	N_{jj}^{Vmass}	5.7	2.2	4.8
4	ΔR_{jj}^{ave}	5.5	8.0	4.4
5	p_T^{4jet}	4.0	6.1	4.6
6	η_{ll}	3.4	2.3	6.8
7	$Centr_{jet}$	2.8	3.4	2.1
8	M_{jj}^{mindR}	2.4	5.7	2.2
9	H_T^{6jets}	2.4	5.4	3.0

Table 5.5.2: Signal versus background separations achieved by different variables used in the initial MVAs.

5.6 New MVA input variables

In addition to the 19 MVA input variables from Table 5.5.1 that were used in the previous $t\bar{t}Z$ 2ℓOS analysis from Ref. [36], 16 new variables are chosen and added for this thesis. Most of the new variables are related to object reconstruction, allowing for better reconstruction of the $t\bar{t}Z$ 2ℓOS decay. The definitions of these new variables, along with the previous variables that are still used, can be found in Table 5.6.1.

Variable	Definition
p_T^{1jet}	p_T of the first jet
p_T^{3jet}	p_T of the third jet
p_T^{4jet}	p_T of the fourth jet
p_T^{6jet}	p_T of the sixth jet
H_T^{6jets}	sum of jet p_T, up to 6 jets
$N_{jet pairs}^{V mass}$	number of jet pairs with mass within a window of 30 GeV around 85 GeV
ΔR_{bb}	cone between two jets with the highest b-tagging weight in the event
$\mathrm{Cent_{jet}}$	scalar sum of $\mathrm{p_T}$ divided by sum of E for all jets
$H1$	First Fox-Wolfram moment, given by Equation 5.5.1
E_T^{miss}	Missing transverse momentum
ΔR_{ll}	ΔR between the two leptons
$y_{\ell\ell}$	rapidity of dilepton system
2νSM weight	weight given by the 2-neutrino-scanning-method $t\bar{t}$ reconstruction technique
1t weight	weight for reconstructing 1 hadronic top
1t1W weight	weight for reconstructing 1 hadronic top and 1 other hadronic W
2t weight	weight for reconstructing 2 hadronic tops
PCBT bin 1j	pseudo-continuous-b-tagging bin for 1st jet
PCBT bin 2j	pseudo-continuous-b-tagging bin for 2nd jet
PCBT bin 3j	pseudo-continuous-b-tagging bin for 3rd jet
PCBT bin 4j	pseudo-continuous-b-tagging bin for 4th jet
PCBT bin 5j	pseudo-continuous-b-tagging bin for 5th jet
PCBT bin 6j	pseudo-continuous-b-tagging bin for 6th jet
N_{top}^{lep}	number of leptonic top candidates
Average Min(M_{jj})	average minimum mass of jet pair per event

Table 5.6.1: The definitions of all variables considered for use in the MVA techniques. Jets and leptons are ordered by their p_T from the highest one. To suppress the effect of mis-modelling in events with high jet multiplicity, only the first 8 jets ordered by p_T are considered when calculating these variables.

The 2-neutrino-scanning-method is a reconstruction technique for dileptonic $t\bar{t}$ decay, therefore $t\bar{t}$ dileptonic decays are assigned high 2νSM weight and other processes are assigned low 2νSM weight. Separately, hadronic $t\bar{t}$ reconstruction is used, and quantified in the variables 1t weight, 1t1W weight, and 2t weight. Since the $t\bar{t}$ system in $t\bar{t}Z$ 2ℓOS signal decays hadronically, $t\bar{t}Z$ 2ℓOS signal is assigned high weight and background processes low weight. Distributions for these hadronic top reconstruction variables in the 2ℓ-Z-2b6j region are shown in Figures 5.6.1-5.6.2. They show some separation between signal and background, and are thus useful input variables for MVAs. In addition, superimposed test and train variables for signal and background are shown in Figures 5.6.1-5.6.2. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. The error bars are statistical.

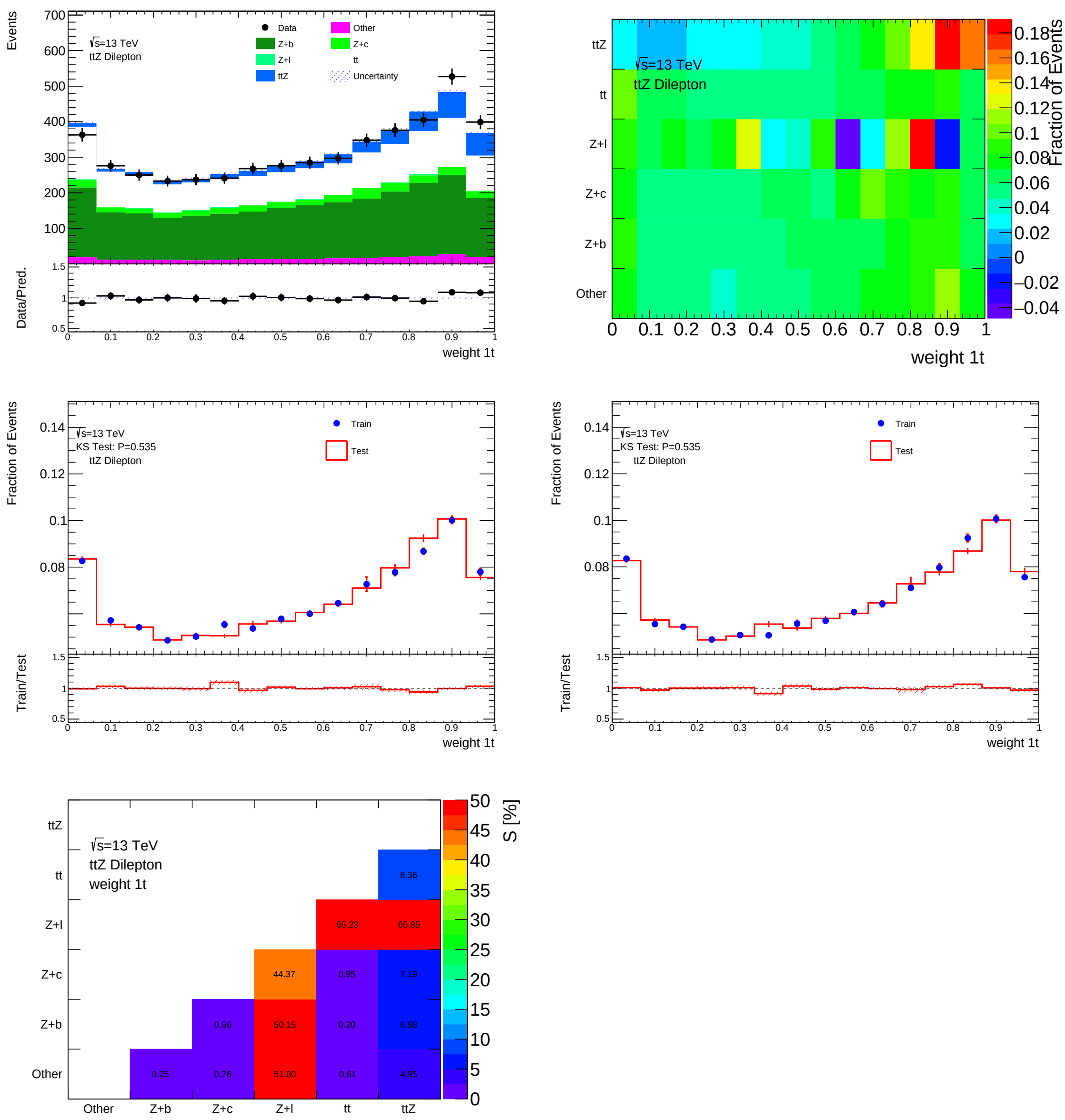


Figure 5.6.1: Weight given by the multi-hypothesis reconstruction of 1 hadronic top, in the 2ℓ-Z-2b6j region of the dilepton OSSF channel. Data versus MC comparison (top left). Separation for each individual sample (top right). Superimposed test and train variables for signal and background for 1st k-fold (middle left). Superimposed test and train variables for signal and background for 2nd k-fold (middle right). Summary of separation for each individual sample (bottom left).

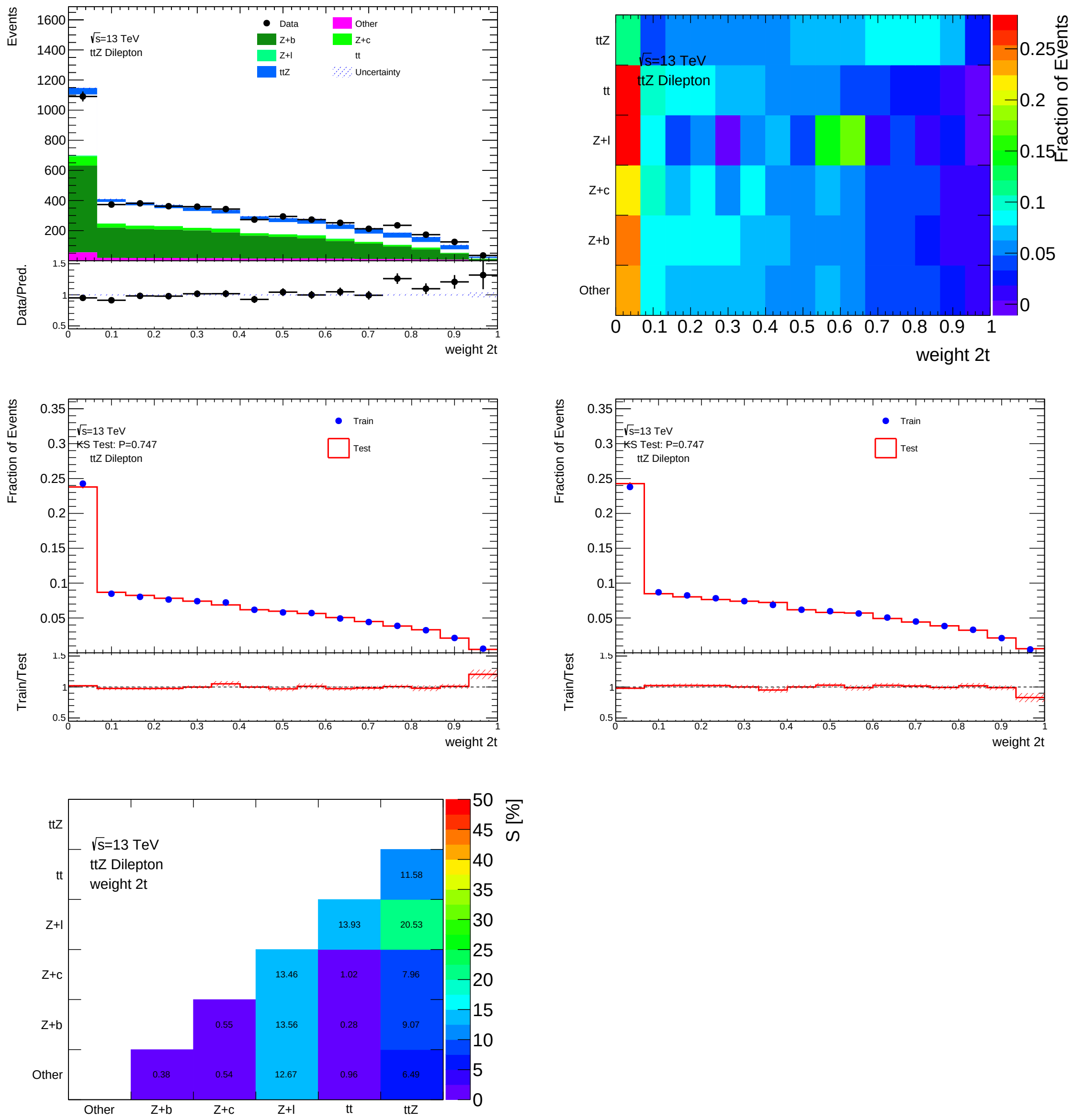


Figure 5.6.2: Weight given by the multi-hypothesis reconstruction of 2 hadronic tops, in the 2ℓ-Z-2b6j region of the dilepton OSSF channel. Data versus MC comparison (top left). Separation for each individual sample (top right). Superimposed test and train variables for signal and background for 1st k-fold (middle left). Superimposed test and train variables for signal and background for 2nd k-fold (middle right). Summary of separation for each individual sample (bottom left).

5.7 Fitting

The distributions of the MVA output from the three signal regions (2ℓ-Z-1b6j, 2ℓ-Z-2b5j, 2ℓ-Z-2b6j) are fit simultaneously together to extract a value and corresponding uncertainty for the expected $t\bar{t}Z$ "signal strength", μ. Signal strength is the ratio of measured cross-section to predicted cross-section. To avoid bias, this fit is performed with pseudo-data such that μ=1. The aim of a fit is to get uncertainties on μ=1 as low as possible. Lower errors indicate an MVA approach that better discriminates between $t\bar{t}Z$ 2ℓOS signal and background. Results of fits with different MVAs will be presented in Section 6.5.

5.7.1 Fit systematics

Included in the fit are various systematic variations to quantify systematic uncertainties on the fitted μ value. Systematic variations on the signal and background samples are included in the fit to be described in Section 5.7.2 as "nuisance parameter (NP)s". Dominant systematics are summarised in Table 5.7.1, then described throughout this subsection.

Systematic	Summary	Ref
Luminosity	Uncertainty on the luminosity estimate of 139 fb^{-1}(Section 3.1)	[182]
Pile-up reweighting	Uncertainty on the SF for PU (Section 4.3.2)	[183]
JES	Derived from test beams, LHC collision data, and simulation (Section 4.3.2)	[139]
JER	Comparison with expected fractional p_{T} resolution (Section 4.3.2)	[139]
Lepton efficiency	Associated with lepton efficiency SFs (Figure 4.3.3 & 4.3.7)	[128, 184]
$t\bar{t}Z$ showering	Alternative sample for showering (Section 4.1.2)	[121, 122]
$t\bar{t}Z$ Var3c	Variations associated to ISR (Section 4.1.2)	[120]
$\mu_R\mu_F$	Renormalisation and factorisation scales varied up and down by 2	[185]
PDF	Evaluated according to PDF4LHC prescription (Section 4.1.1)	[186]
Z+jets CKKW	Uncertainty associated with matching scale	[187]
Z+jets QSF	Uncertainty associated with resummation scale	[188]

Table 5.7.1: Dominant systematic uncertainties in the fit for $t\bar{t}Z$ signal strength.

Luminosity

The quoted integrated luminosity for the full Run 2 dataset of 139.0 fb^{-1}introduced in Section 3.1 has a relative uncertainty of 1.7% [182], following the latest recommendation of the ATLAS Luminosity Working Group. This uncertainty is applied to all samples that are derived from MC.

Pile-up reweighting

As introduced in Section 4.3.2, a Scale Factor (SF) is applied to MC to account for the difference in pile-up distributions between MC and measured data.

Jet Energy Scale

Jet Energy Scale (JES) (Section 4.3.2) and its corresponding uncertainty have been derived by combining information from test beams, measured LHC collisions, and simulation [139]. In this analysis the recommendation by the ATLAS JetEtMiss Working Group for Run 2 is used.

Jet Energy Resolution

Jet Energy Resolution (JER) (Section 4.3.2) is measured separately for MC and data, using two *in-situ* techniques described in Ref. [139]. The quadratic difference between JER for MC and data is defined as the JER systematic uncertainty.

Lepton efficiency

Systematic uncertainties for lepton efficiency are associated with SFs defined by the respective ATLAS Working Groups [128, 184]. These SFs account for differences between MC and data in the distributions of lepton identification, isolation and trigger efficiency (Figure 4.3.3 & 4.3.7).

$t\bar{t}Z$ showering

As introduced in Section 4.1.2, a systematic uncertainty for $t\bar{t}Z$ showering is obtained by comparing the nominal sample (MadGraph5_aMC@NLO +Pythia 8 [116]) to an alternative sample, MadGraph5_aMC@NLO +Herwig 7 [121, 122]. This changes the parton-shower algorithm for all jets in the $t\bar{t}Z$ decay.

ttZ Var3c

As introduced in Section 4.1.2, a systematic uncertainty quantifying different possibilities for initial-state-radiation (ISR) in the signal sample is obtained by varying the Var3c parameter in the Pythia 8 A14 tune [120]. This varies the value of α_s and therefore quantifies the effect of uncertainty on the strength of the strong nuclear force.

$\mu_R \mu_F$

To obtain a systematic uncertainty for renormalisation and factorisation scales, the scale parameters μ_R and μ_F are varied up and down by a factor of 2, and compared to the nominal prediction [185]. The variation of each scale by a factor of 2 in each direction (up and down) means there are a total of 4 systematic variations to consider:

- $0.5\times\mu_R^{nominal}$ with $2\times\mu_F^{nominal}$;
- $0.5\times\mu_R^{nominal}$ with $0.5\times\mu_F^{nominal}$;

- $2\times\mu_R^{nominal}$ with $2\times\mu_F^{nominal}$;
- $2\times\mu_R^{nominal}$ with $0.5\times\mu_F^{nominal}$.

Parton Distribution Function

Systematic uncertainties on the Parton Distribution Function (PDF) (Section 4.1.1) are evaluated according to the recommended PDF4LHC procedure [186]. This procedure includes varying the α_s parameter, as well as the choice of PDF.

Z+jets CKKW

The CKKW matching scale is part of the parton shower description in multijet events in Sherpa. A systematic uncertainty on the Z+jets CKKW matching scale is obtained by comparing the nominal scale of 20 GeV to alternative scales of 15 GeV and 30 GeV [187].

Z+jets QSF

QSF resummation scale is a cutoff point for the emission of soft gluons in simulations. A systematic uncertainty on the Z+jets QSF resummation scale is obtained by varying the QSF parameter up and down by a factor of 2, and comparing to the nominal [188].

5.7.2 Profile likelihood

Background processes are accounted for in this analysis by measuring the normalisation of some of the main SM backgrounds to the $t\bar{t}Z$ 2ℓOS process. The main backgrounds whose normalisations are measured are $Z+b$ jets and $Z+c$ jets processes. These background normalisations are measured by using parts of the output MVA distributions that are enriched in these backgrounds, whilst having low $t\bar{t}Z$ 2ℓOS signal contamination. A background-enriched region could be a region of an MVA distribution where background events tend to be observed, at the opposite end of the MVA distribution the where $t\bar{t}Z$ 2ℓOS signal events tend to be observed, e.g. the low MVA output score of a binary classifier distinguishing between $t\bar{t}Z$ 2ℓOS signal and all other processes. A background-enriched region could also be a region with additional selection requirements imposed over the initial regions (2ℓ-Z-1b6j, 2ℓ-Z-2b5j, 2ℓ-Z-2b6j), to remove as much $t\bar{t}Z$ 2ℓOS signal as possible, whilst keeping as much as possible of the background of interest.

The MVA distributions in the three signal regions (2ℓ-Z-1b6j, 2ℓ-Z-2b5j, 2ℓ-Z-2b6j), along with their systematic uncertainty variations, are used in a profile-likelihood fit to extract the $t\bar{t}Z$ signal strength by comparing measured and predicted cross-sections. A likelihood is maximised in order to find the best-fit parameters using the equation:

$$L\left(\vec{n}|\mu,\vec{\theta},\vec{k}\right) = \prod_{r\in\text{regions}}\prod_{i\in\text{bins}} \text{Pois}\left(n_{i,r}|\mu S_{i,r}(\vec{\theta}) + B_{i,r}(\vec{\theta},\vec{k})\right) \times \prod_{j\in NP} \text{Gaus}\left(\theta_j\right), \tag{5.7.1}$$

where $\vec{n}$ represents the measured data vector, with $n_{i,r}$ representing measured data yields in bin i and region r. $\vec{\theta}$ denotes the NPs that affect the number of signal events $S_{i,r}$ as well as the number of background events $B_{i,r}$ in bin i and region r. $\vec{k}$ represents free-floating normalisation parameters that also affect the number of background events. Finally, μ is the signal strength, which is the parameter of interest in the likelihood: the ratio of the measured over predicted signal cross section. The terms Pois and Gaus represent Poisson and Gaussian distributions, respectively. The full form of the likelihood can be found in Ref. [189]. In practice, the negative logarithm of the likelihood is minimised. The likelihood takes as input:

- data yields in each bin of each region;
- NPs that affect the number of signal events for each bin in each region;
- NPs that affect the number of background events for each bin in each region;
- free-floating normalisation parameters that affect the number of background events;
- the signal strength, μ.

Using multiple bins in a distribution (as is done by using the bins of an MVA distribution) provides additional information to separate signal from background, compared to using a single bin (which would be the number of events). The systematic uncertainties summarised in Table 5.7.1 are evaluated independently in each region, therefore their relative effect is different in each region.

6 $t\bar{t}Z$ Analysis results

Wisdom comes from experience. Experience is often a result of lack of wisdom.

Terry Pratchett [190]

This chapter presents results and discussion of the various MVAs that were tested for classification of the $t\bar{t}Z$ 2ℓOS process. Each MVA is shown and discussed in separate sections, increasing in complexity as the sections progress. They are compared at the end of each section. Results will also be compared to Ref. [36], the previous ATLAS published result in the $t\bar{t}Z$ 2ℓOS channel, using 36 fb^{-1}of Run 2 data. The MVAs discussed in this chapter are:

1. binary BDTs (Section 6.1);
2. binary DNNs (Section 6.2);
3. multiclass DNNs (Section 6.3);
4. multiclass DNNs with new variables (Section 6.4).

Using the final MVAs, MVA distributions are then fitted to extract the $t\bar{t}Z$ 2ℓOS signal strength. Finally, results from the 2ℓOS channel are combined with the 3ℓ and 4ℓ channels to obtain the combined $t\bar{t}Z$ signal strength. The specific contribution of the author covered throughout Section 6.1-Section 6.4 includes:

- setting up, training, and evaluating various MVA models for classification of the $t\bar{t}Z$ 2ℓOS process;
- comparing the performance of the various MVA models.

6.1 Single BDT

The first step to studying the 2ℓOS channel for this analysis was to use a single MVA per analysis region, 2ℓ-Z-1b6j, 2ℓ-Z-2b5j, 2ℓ-Z-2b6j. Each MVA would attempt to classify between $t\bar{t}Z$ 2ℓOS signal and total background.

A gradient Boosted Decision Tree (BDT) has been investigated as a multivariate technique for signal vs. background discrimination. Since a BDT was used in the previous $t\bar{t}Z$ analysis in the literature [36], a BDT was chosen as a benchmark for this analysis.

6.1.1 BDT options

Settings options of the Boosted Decision Tree can be found in Table 6.1.1. Further explanation of settings can be found in the scikit-learn [191] documentation for GradientBoostingClassifier [192]. The learning rate used here is on the same order of magnitude as the learning rate order of magnitude shown to be optimal with ATLAS Open Data.

Option	Values	Description
n_estimators	500	Number of boosting stages
max_depth	3	Maximum depth of the individual regression estimators
min_samples_leaf	$0.05N_{train}$	minimum number of samples required to be at a leaf node, where N_{train} is the number of events in the training sample
boost	Gradient	boosting type for the trees in forest
learning_rate	0.3	shrinks the contribution of each tree by ‘learning_rate‘
criterion	friedman_mse	function to measure the quality of a split
min_impurity_decrease	0.5	A node will be split if this split induces a decrease of the impurity greater than or equal to this value
SigToBkgFraction	1	Sig to Bkg ratio used in Training

Table 6.1.1: Settings used in BDT training when using single BDT per 2ℓOS region.

6.1.2 Variables used for BDT

The variables chosen for training the BDT are given in Table 6.1.2, along with the ranking of the variables used. Separate training was done for the three regions: 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j. The number of variables used for each 2ℓOS region varied: 17 for 2ℓ-Z-2b6j, 15 for 2ℓ-Z-1b6j and 14 for 2ℓ-Z-2b5j.

rank	1b6j	2b5j	2b6j
1	$\mathrm{H_T^{6jets}}$	$\mathrm{H_T^{6jets}}$	$\mathrm{H_T^{6jets}}$
2	η_{ll}	ΔR_{jj}^{ave}	ΔR_{ll}
3	$N_{jj}^{V mass}$	$N_{jj}^{V mass}$	η_{ll}
4	$\mathrm{p_T^{b1}}$	$\mathrm{M_{bb}^{pTord}}$	$N_{jj}^{V mass}$
5	$\mathrm{MaxM_{lepb}^{mindR}}$	$\mathrm{M_{jj}^{mindR}}$	$\Delta R_{\mathrm{ave}}^{\mathrm{jj}}$
6	$\mathrm{M_{jj}^{mindR}}$	ΔR_{ll}	ΔR_{bb}
7	$\mathrm{p_T^{4jet}}$	ΔR_{bb}	$\mathrm{p_T^{6jet}}$
8	ΔR_{ll}	$\mathrm{p_T^{4jet}}$	$\mathrm{MaxM_{lepb}^{mindR}}$
9	$\mathrm{p_T^{6jet}}$	η_{ll}	$\mathrm{M_W^{avg}}$
10	$\mathrm{M_W^{avg}}$	$\mathrm{p_T^{5jet}}$	$\mathrm{M_{jj}^{mindR}}$
11	$Centr_{jet}$	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{p_T^{4jet}}$
12	H1	H1	$\mathrm{M_{bb}^{pTord}}$
13	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{p_T^{ll}}$	$Centr_{jet}$
14	ΔR_{jj}^{ave}	$Centr_{jet}$	H1
15	$\mathrm{p_T^{ll}}$		$\mathrm{p_T^{b1}}$
16			$N_{bjj}^{top-mass}$
17			$\mathrm{p_T^{ll}}$

Table 6.1.2: Ranking of the variables used for BDT training, when using a single BDT per 2ℓOS region.

6.1.3 BDT overtraining test

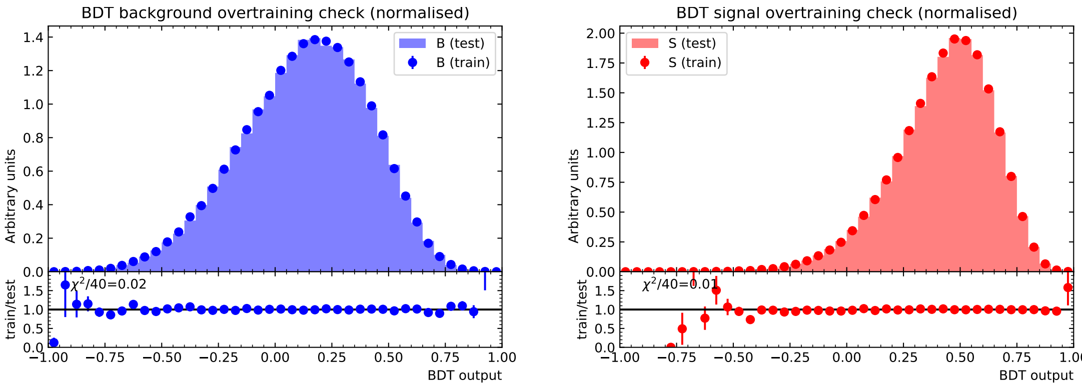
To ensure that the BDT is properly trained and its results do not depend on the event statistics, an overtraining test needs to be performed. The BDT discriminants (training and test samples) used for the 2ℓ-Z-2b6j region of the dilepton OSSF channel is shown in Figure 6.1.1, as an example. Figure 6.1.1 and other overtraining tests for the BDTs show no obvious overtraining, by the fact that the training and test samples overlap within error bars. The ROC curve for the BDT is shown in Figure 6.1.2. The fact that the red testing line largely overlaps with the blue training line means there is no noticeable overtraining.

6.1.4 BDT output distribution

Figure 6.1.3 shows the BDT output distribution in MC and data. Within uncertainties, data and MC agree, and there is >25% separation between signal and background, which is already good but can be improved upon.

6.1.5 Summary of BDT results

Table 6.1.3 summarises the main numerical metrics of this subsection using a single BDT per 2ℓOS region. Information on the number of variables is shown in Section 6.1.2. The training and test data set ROC AUCs are shown in Section 6.1.3. Separation distributions are shown in Section 6.1.4.



(Left column) Background

(Right column) Signal

Figure 6.1.1: BDT discriminants used for the 2ℓ-Z-2b6j signal region of the dilepton OSSF channel overlapping test and training samples. Background in the left column, signal in the right column. The uncertainties are MC statistics. These plots use a single BDT per 2ℓOS region.

Metric	2ℓ-Z-$1b6j$	2ℓ-Z-$2b5j$	2ℓ-Z-$2b6j$
number of variables	15	14	17
training set ROC AUC	0.7873	0.7937	0.7973
test set ROC AUC	0.7876	0.7914	0.7955
separation (%)	25.497	30.705	30.309

Table 6.1.3: Metrics that summarise the results of using a single BDT per 2ℓOS region, as described throughout this subsection. The separation is measured between the $t\bar{t}Z$ 2ℓOS signal and total background.

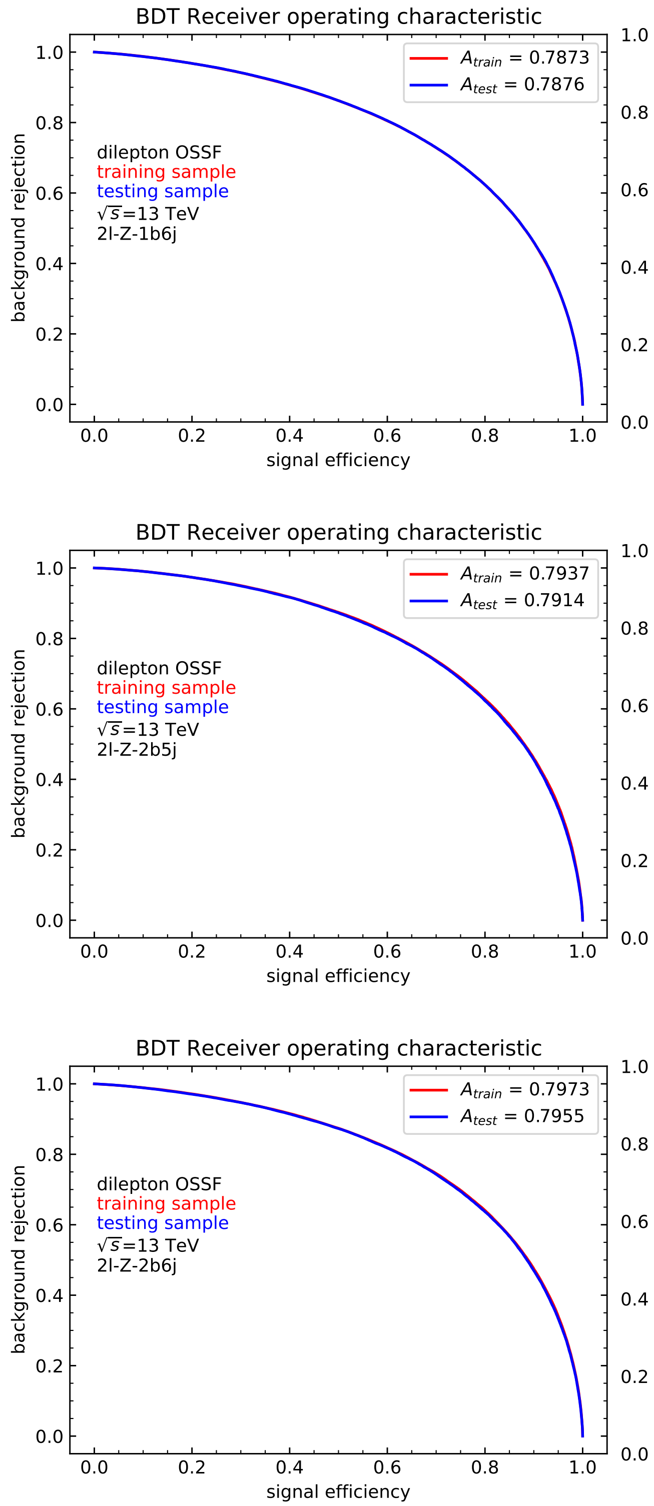
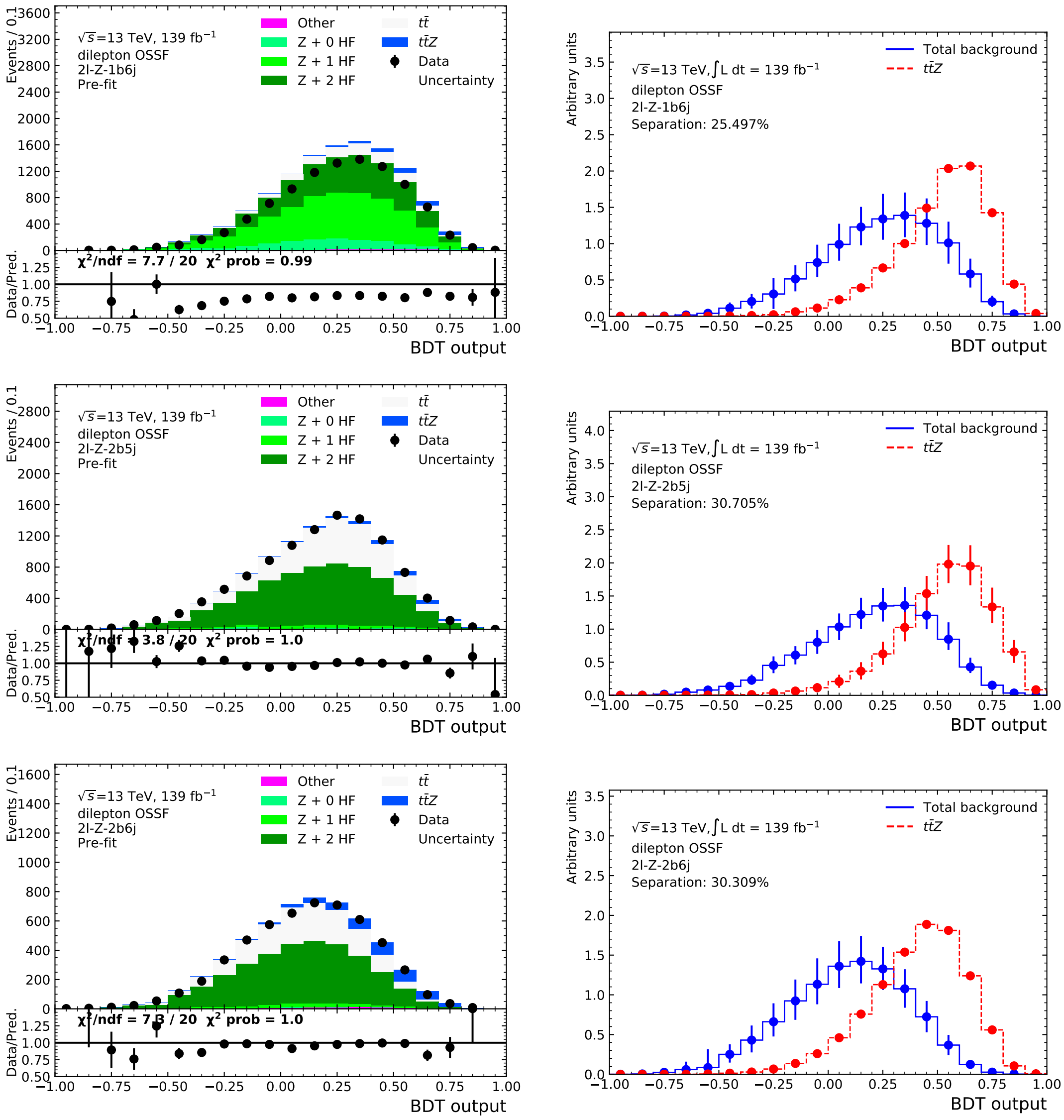


Figure 6.1.2: BDT ROC curves for the three signal regions of the dilepton OSSF channel. From top to bottom 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j regions. The errors are MC statistics. When using a single BDT per 2ℓOS region.



(Left column) Data vs. MC (Right column) Separation

Figure 6.1.3: BDT discriminants used for the three signal regions of the dilepton OSSF channel. Data vs. MC (left column) and separation of $t\bar{t}Z$ 2ℓOS signal and total background distributions (right column) in the regions of the dilepton OSSF channel. (Top row) 2ℓ-Z-1b6j. (Middle row) 2ℓ-Z-2b5j. (Bottom row) 2ℓ-Z-2b6j. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. The error bars include statistical and systematic uncertainties. When using a single BDT per 2ℓOS region.

6.1.6 Comparison with previous results

A comparison between the BDTs from the previous ATLAS publication on $t\bar{t}Z$ [36] and the BDTs developed for this thesis is shown in Table 6.1.4. All metrics other than 2ℓ-Z-1b6j Separation % improve in going from the BDTs of Ref. [36] to the BDTs developed for this thesis, which validated the MVA setup developed for this thesis. Even the 2ℓ-Z-1b6j Separation % only decreases slightly.

MVA	1b6j AUC	2b5j AUC	2b6j AUC	1b6j Separation %	2b5j Separation %	2b6j Separation %
BDT from [36]	0.618	0.585	0.615	25.6	27.6	28.6
BDT	0.788	0.791	0.796	25.5	30.7	30.3

Table 6.1.4: Comparison of metrics that summarise the results of 1) the BDTs from Ref. [36], 2) the BDTs developed for this thesis. Separation is measured between $t\bar{t}Z$ 2ℓOS signal and all backgrounds. All metrics other than 2ℓ-Z-1b6j Separation % improve in going from the BDTs of Ref. [36] to the BDTs developed for this thesis. Even the 2ℓ-Z-1b6j Separation % only decreases slightly.The AUCs quoted here are the “test set ROC AUC" values from Table 6.1.3.

Though a >16% improvement in ROC AUC is achieved compared to the BDTs from Ref. [36], a greater improvement should be possible. It is then necessary to design and test DNNs, as they could provide more flexibility in classifying the $t\bar{t}Z$ 2ℓOS process.

6.2 2ℓOS single DNN

Multiple multivariate techniques have been tested. After re-implementing the MVA strategy from Ref. [36] with BDTs as a benchmark, the next step in using the 2ℓOS channel with full run 2 data was to use a Deep Neural Network (DNN), since they offer more flexibility than BDTs. More flexibility offers the potential for better discrimination power between $t\bar{t}Z$ 2ℓOS signal and background. The single BDT implementation for this analysis can be found in Section 6.1.

6.2.1 DNN options

Settings options of the DNN can be found in Table 6.2.1. These DNNs are trained using tensorflow [179] through keras [180].

Option	Values	Description
InputScaling	minmax	inputs are scaled into a range between 0 and 1
Folds	2	how many folds (k-folding) should be performed during training
Nodes	50,50,50,50	Comma-separated list of neurons for each layer
Loss	binary crossentropy	Loss function which is used in the training of a model
Epochs	100	Number of training epochs
LearningRate	0.001	Initial learning rate for the training of a model
BatchSize	32	Batch size used in training of a model
ValidationSize	0.2	Relative size of the validation set used during training of a model
Patience	30	Number of epochs with no improvement after which training will be stopped
MinDelta	0.0001	Minimum change in the monitored quantity to qualify as an improvement
DropoutIndice	1,3	Layer indeces at which Dropout layers are added
DropoutProb	0.1	Probability of dropout
OutputSize	1	Number of neurons in the output layer
OutputActivation	sigmoid	Activation function in the output layer of a model
Metrics	Accuracy	Comma-separated list of metrics to be evaluated during training
ModelBinning	20,0,1	Custom binning using a fixed bin width in the format nbins,x_low,x_high

Table 6.2.1: Settings used in DNN training, when using a single binary DNN per 2ℓOS region.

6.2.2 Variables used for DNN

The variables chosen for training the DNN are given in Figure 6.2.1, along with the ranking of the variables used. The x-axis quantifies how much the area under curve (AUC) decreases (AUC_{nom} - AUC) when a particular variable is removed from training, compared to the AUC with all variables (AUC_{nom}). This quantity (AUC_{nom} - AUC)/AUC_{nom} is called “importance". A large AUC decrease (large horizontal bar) means the variable is important for the DNN. Figure 6.2.1 shows that H_T^{6jets} is the most important variable in all three regions 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j. Separate training was done for the three regions: 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j. The number of variables used for each 2ℓOS region varied: 15 for 2ℓ-Z-1b6j, 14 for 2ℓ-Z-2b5j and 17 for 2ℓ-Z-2b6j.

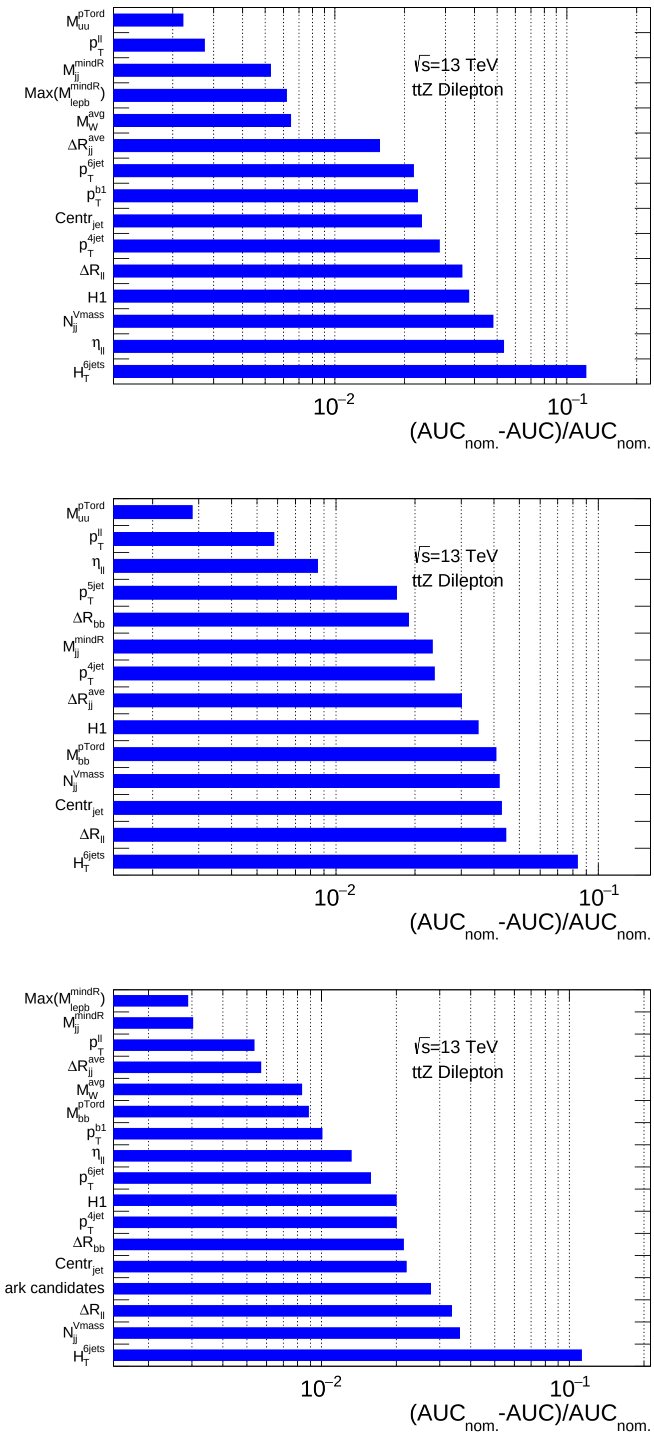


Figure 6.2.1: “Importance" of the variables used for binary DNN training. Variables with greatest values on the x-axis (e.g. $>10^{-2}$) are most “important". (Top row) 2ℓ-Z-1b6j. (Middle row) 2ℓ-Z-2b5j. (Bottom row) 2ℓ-Z-2b6j.

6.2.3 DNN visualisation

A visualisation of the DNN used in this section is shown in Figure 6.2.2 for illustration. The first layer on the left, shown as white circles, shows how many input variables are used in the DNN (17 in this example). Moving towards the right, four hidden layers of 50 neurons each are shown. The rightmost layer shows one neuron for the single output of a binary DNN.

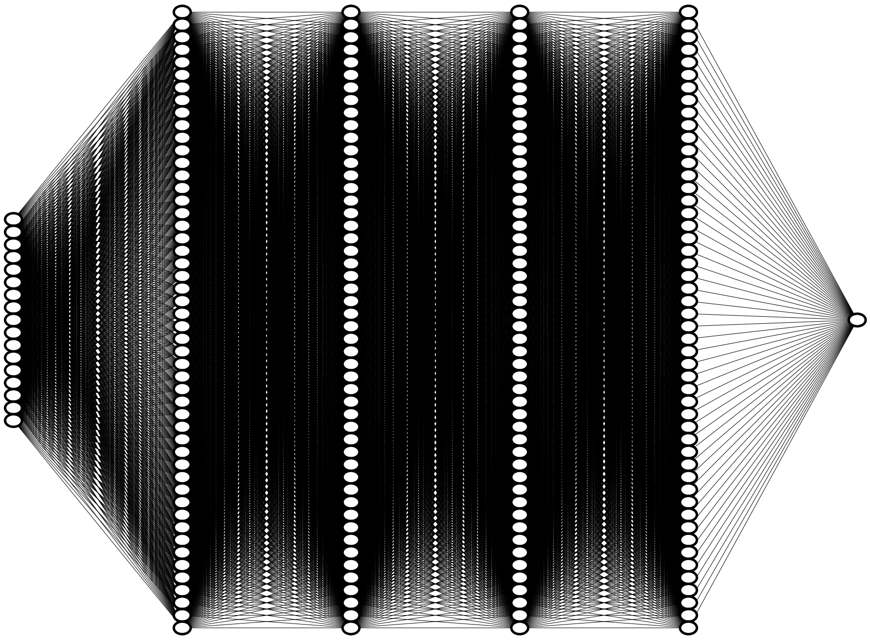
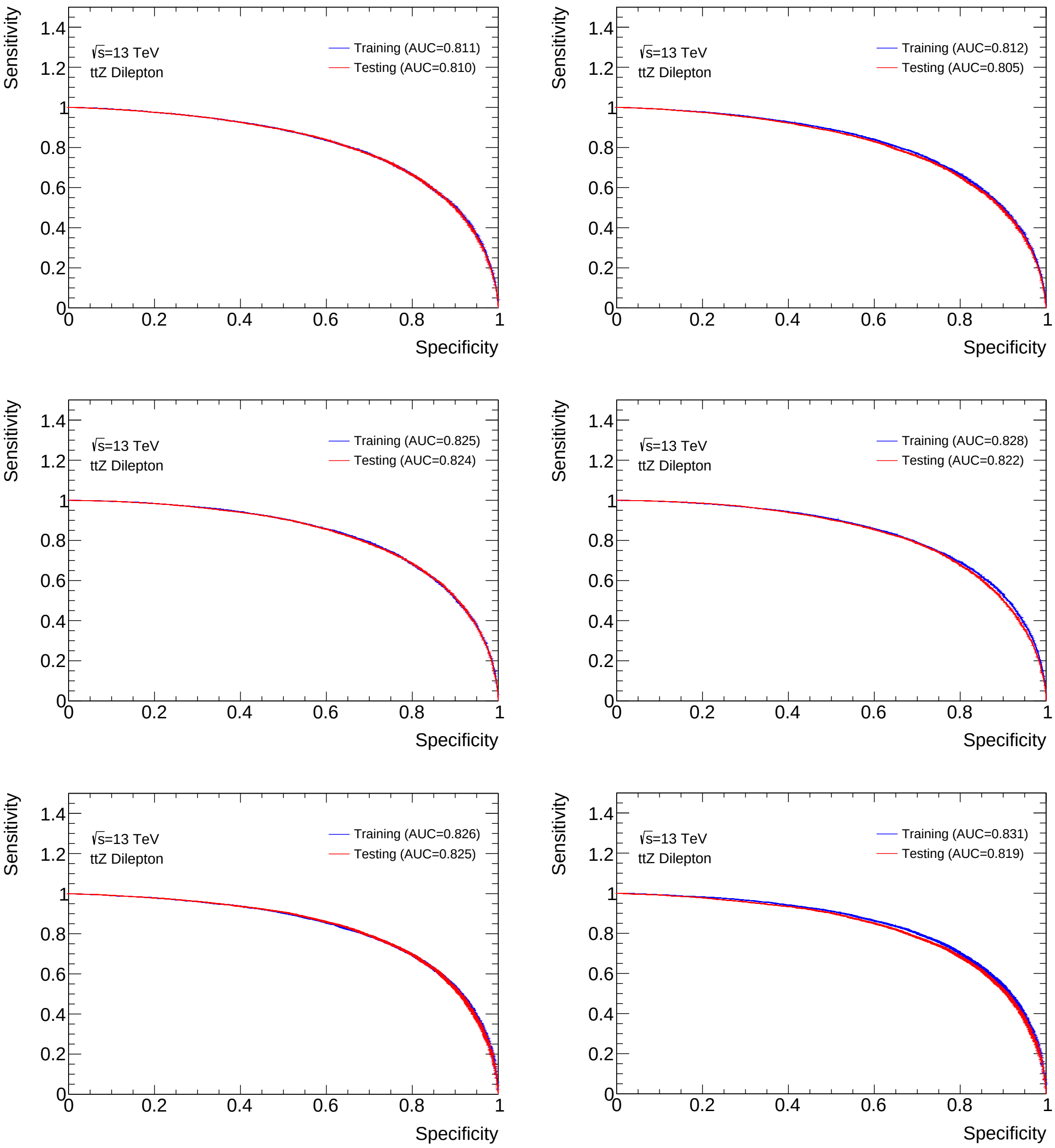


Figure 6.2.2: A visualisation of a binary Deep Neural Network used in this thesis. This network is for the 2ℓ-Z-2b6j region with 17 inputs.

6.2.4 DNN performance

The ROC curve is shown in Figure 6.2.3. The fact that the red testing line largely overlaps with the blue training line means there is no noticeable overtraining.

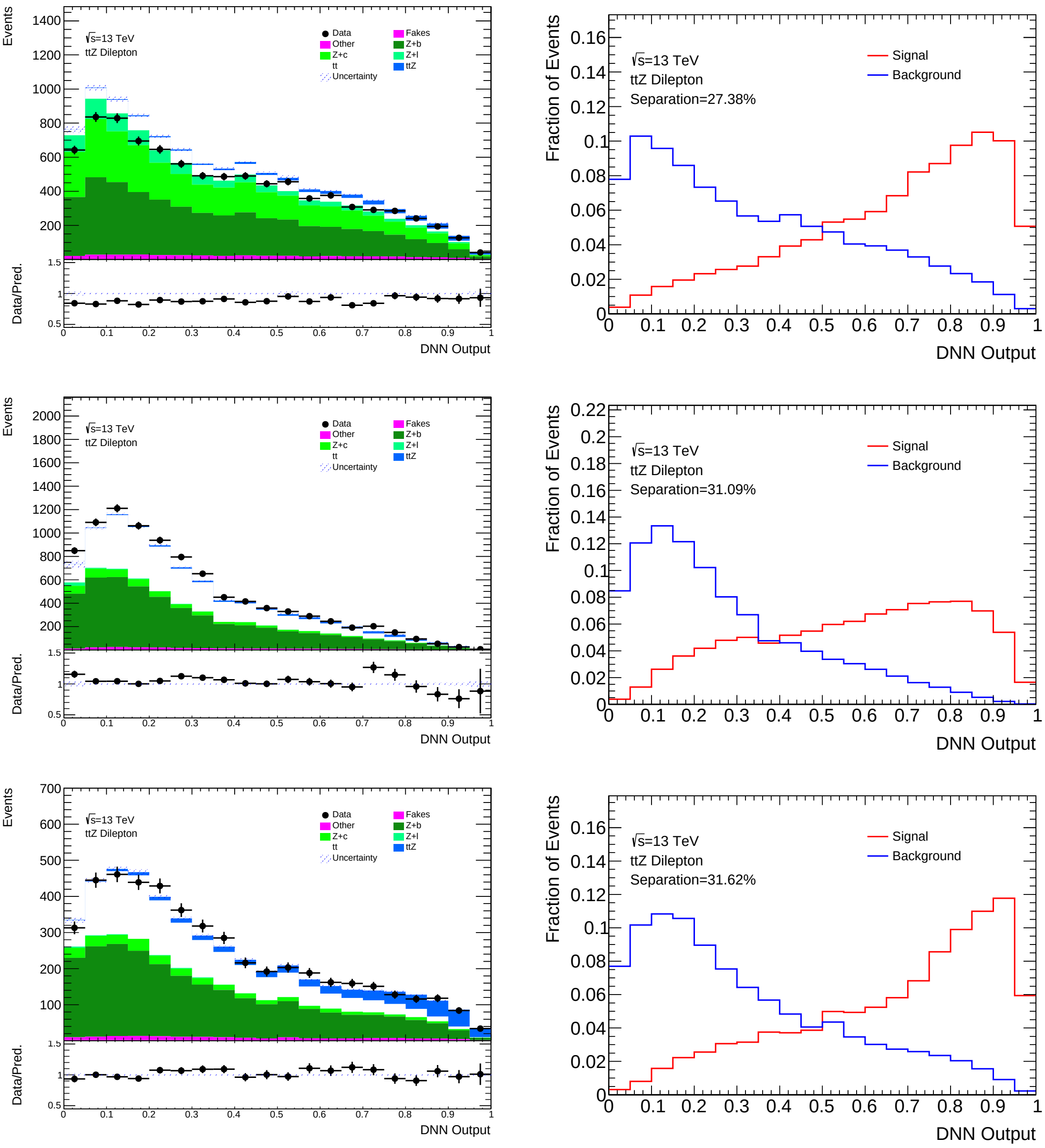


(Left column) 1st k-fold (Right column) 2nd k-fold

Figure 6.2.3: Binary DNN ROC curves for the three signal regions of the dilepton OSSF channel overlapping test and training samples. 1st k-fold in the left column, 2 k-fold in the right column. From top to bottom 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j regions.

6.2.5 DNN output distribution

Figure 6.2.4 shows the DNN output distribution in MC and data, and the separation between $t\bar{t}Z$ 2ℓOS signal and total background for the DNN output. Data and MC mostly agree, and there is >27% separation between signal and background.



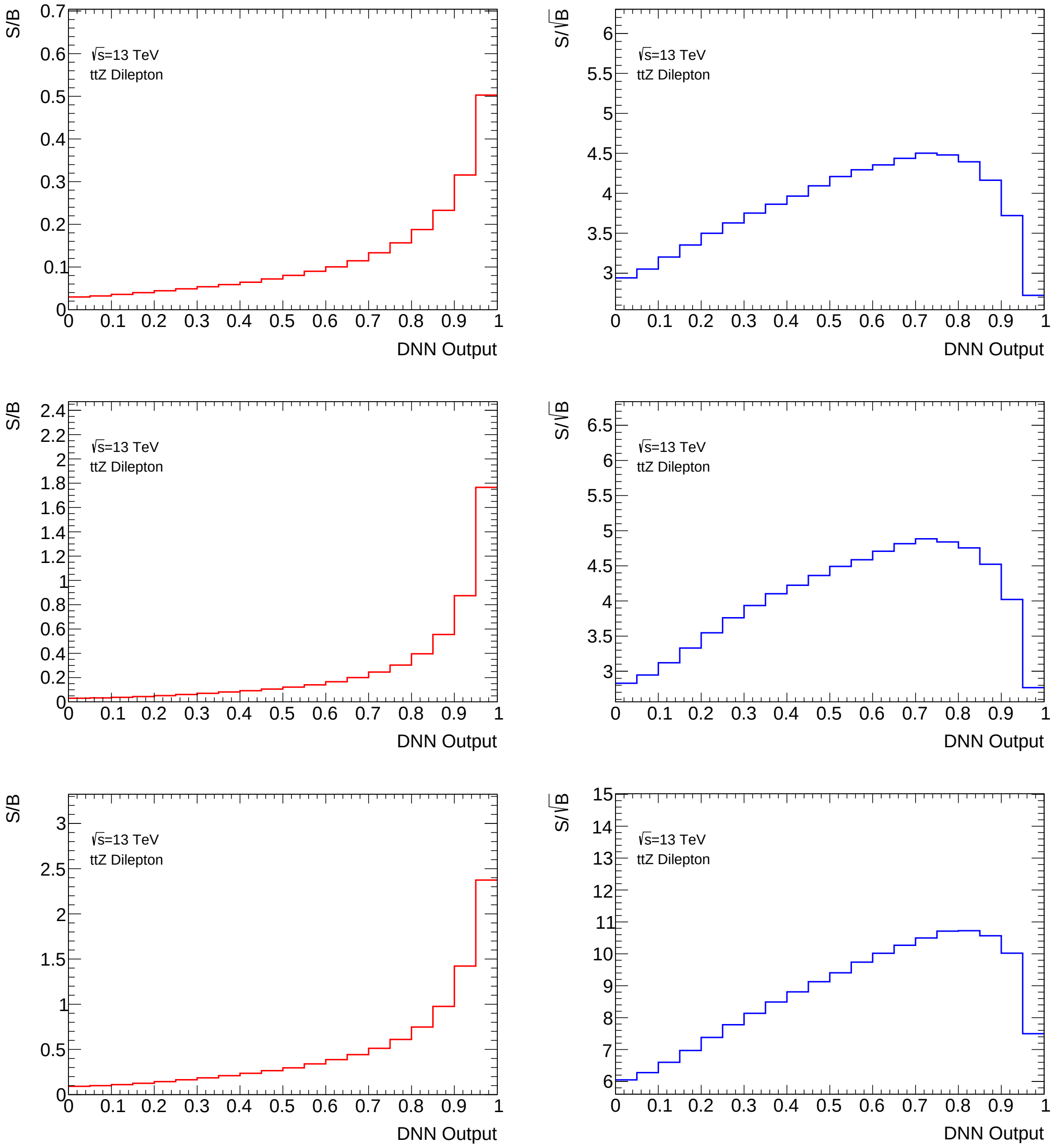
(Left column) Data vs. MC

(Right column) Separation

Figure 6.2.4: DNN discriminants used for the three signal regions of the dilepton OSSF channel, using a single binary DNN per 2ℓOS region. Data vs. MC (left column) and separation of $t\bar{t}Z$ 2ℓOS signal and total background distributions (right column), in the regions of the dilepton OSSF channel. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. The error bars are statistical. (Top row) 2ℓ-Z-1b6j. (Middle row) 2ℓ-Z-2b5j. (Bottom row) 2ℓ-Z-2b6j.

6.2.6 DNN significance measures

Signal over background (S/B) and statistical significance ($S/\sqrt{B}$) distributions are shown in Figure 6.2.5. S/B is also called purity. Figure 6.2.5 shows that statistical significances >3 ("evidence") can be achieved in the 2ℓ-Z-1b6j and 2ℓ-Z-2b5j regions alone, while statistical significance >5 ("observation") can be achieved in the 2ℓ-Z-2b6j region alone.



(Left column) S/B (Right column) S/$\sqrt{B}$

Figure 6.2.5: (Left column) Signal over background that would be achieved by selecting events above the x-axis DNN output value. (Right column) Signal over $\sqrt{\text{background}}$ that would be achieved by selecting events above the x-axis DNN output value. (Top row) 2ℓ-Z-1b6j. (Middle row) 2ℓ-Z-2b5j. (Bottom row) 2ℓ-Z-2b6j. All when using a single binary DNN per 2ℓOS region.

6.2.7 Summary of DNN results

Table 6.2.2 summarises the main numerical metrics of this subsection using a single DNN per 2ℓOS region. ROC AUCs >0.8, separations >27% and statistical significances >4 were achieved in all three signal regions. Information on the number of variables is shown in Section 6.2.2. The training and test data set ROC AUCs are shown in Section 6.2.4. Separation distributions are shown in Section 6.2.5.

Metric	2ℓ-Z-$1b6j$	2ℓ-Z-$2b5j$	2ℓ-Z-$2b6j$
number of variables	15	14	17
training set ROC AUC (fold 1)	0.811	0.825	0.826
test set ROC AUC (fold 1)	0.810	0.824	0.825
training set ROC AUC (fold 2)	0.812	0.828	0.831
test set ROC AUC (fold 2)	0.805	0.822	0.819
training set ROC AUC (average)	0.8115	0.8265	0.8285
test set ROC AUC (average)	0.8075	0.823	0.822
separation (%)	27.38	31.09	31.62
S/$\sqrt{B}$	4.5	4.9	10.8

Table 6.2.2: Metrics that summarise the results of using a single binary DNN per 2ℓOS region, as described throughout this subsection. Separation is measured between the $t\bar{t}Z$ 2ℓOS signal and total background. “training set ROC AUC (average)" is the average of “training set ROC AUC (fold 1)" and “training set ROC AUC (fold 2)". “test set ROC AUC (average)" is the average of “test set ROC AUC (fold 1)" and “test set ROC AUC (fold 2)".

6.2.8 Comparison with previous results

We now compare the DNNs to the BDT from Ref. [36], as well as the previous results for BDTs. All metrics show small improvement by ≥0.385% in going from a BDT to a DNN, which is why DNNs were used going forward.

MVA	1b6j AUC	2b5j AUC	2b6j AUC	1b6j Separation %	2b5j Separation %	2b6j Separation %
BDT from [36]	0.618	0.585	0.615	25.6	27.6	28.6
BDT	0.788	0.791	0.796	25.5	30.7	30.3
DNN	0.808	0.823	0.822	27.4	31.1	31.6

Table 6.2.3: Comparison of metrics that summarise the results of 1) the BDTs from Ref. [36], 2) the BDTs developed for this thesis, 3) the binary DNN approach developed for this thesis. Separation is measured between $t\bar{t}Z$ 2ℓOS signal and all backgrounds. All metrics improve in going from a BDT to a DNN. The AUCs quoted here are the “test set ROC AUC (average)" values from Table 6.2.2. This Table builds on Table 6.1.4 by adding the row for DNN.

Having only slightly improved upon the initial BDTs with DNNs by ≥2% in ROC AUC, it is then necessary to test whether a multiclass approach can yield a greater improvement and thus lead to a more accurate classification of the $t\bar{t}Z$ 2ℓOS process.

6.3 2ℓOS multiclass DNN with same variables as single BDT

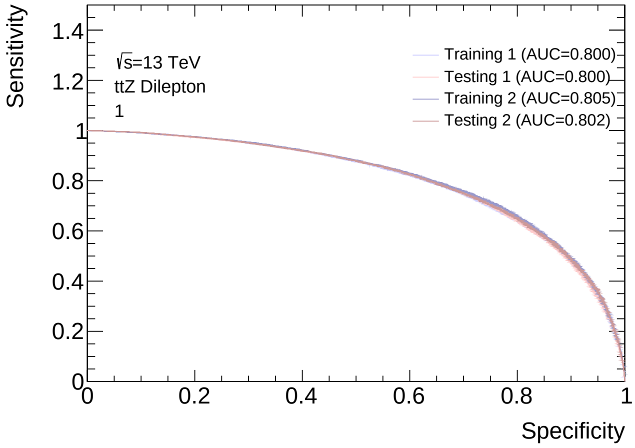
Section 6.2 shows results having trained DNNs with the same variables as the 36 fb^{-1}ATLAS $t\bar{t}Z$ paper [36], with a single output (binary classifier). This section uses the same input variables but with 3 outputs (multi-classification). The 3 outputs were 1) $t\bar{t}$ background, 2) $t\bar{t}Z$ 2ℓOS signal, 3) Z+jets background. Multi-classification allows the DNN to separately learn the characteristics of each process. In contrast, a binary classifier does not distinguish between backgrounds that are themselves physically different processes.

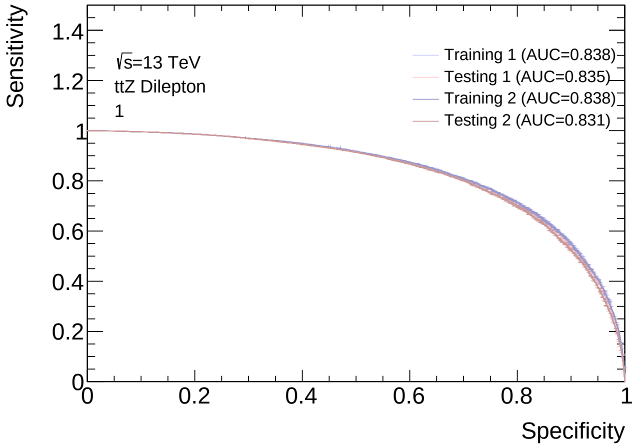
6.3.1 DNN options

These DNNs are trained using tensorflow [179] through keras [180]. The only differences compared to Table 6.2.1 of Section 6.2 is the use of categorical crossentropy rather than binary crossentropy as the loss function, and an output size of 3 rather than 1, to have an output for each of 1) $t\bar{t}Z$ 2ℓOS signal, 2) $t\bar{t}$ background, 3) Z+jets background.

6.3.2 DNN performance

The ROC curve is shown in Figure 6.3.1. The fact that the training and test curves closely overlap means there is no noticeable overtraining.





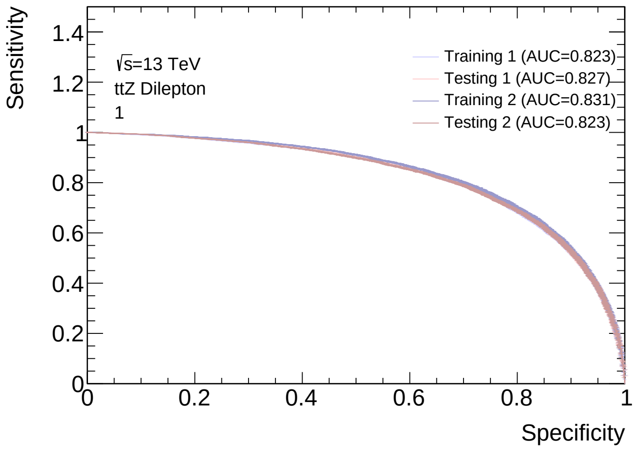


Figure 6.3.1: ROC curve for the 3 regions of the dilepton OSSF channel, when using an initial multiclass DNN per 2ℓOS region. (Top row) is for 2ℓ-Z-1b6j. (Middle row) is for 2ℓ-Z-2b5j. (Bottom row) is for 2ℓ-Z-2b6j.

6.3.3 DNN output distribution

Figures 6.3.2-6.3.4 show the separation between the classifier output being optimised in that multiclassification output node, and all other processes. e.g. separation between $t\bar{t}$ background and all other processes (including $t\bar{t}Z$ 2ℓOS signal). The fact that the separation in the bottom rows of Figures 6.3.2-6.3.4 is always <6% shows that it is difficult to separate Z+jets from other processes.

6.3.4 Summary of initial multiclass DNN results

Table 6.3.1 summarises the main numerical metrics of this subsection using a single DNN per 2ℓOS region. Information on the number of variables is shown in Section 6.2.2. The training and test data set ROC AUCs are shown in Section 6.2.4. Separation distributions are shown in Section 6.2.5. These results show that multiclass DNNs can achieve good performance (ROC AUC >0.8) and $t\bar{t}Z$ separation (>26%).

Metric	2ℓ-Z-$1b6j$	2ℓ-Z-$2b5j$	2ℓ-Z-$2b6j$
number of variables	15	14	17
training set ROC AUC (fold 1)	0.800	0.838	0.823
test set ROC AUC (fold 1)	0.800	0.835	0.827
training set ROC AUC (fold 2)	0.805	0.838	0.831
test set ROC AUC (fold 2)	0.802	0.831	0.823
training set ROC AUC (average)	0.803	0.838	0.827
test set ROC AUC (average)	0.801	0.833	0.825
$t\bar{t}$ separation (%)	16.5	11.5	14.2
$t\bar{t}Z$ separation (%)	26.3	32.1	30.2
Z+jets separation (%)	1.38	3.98	5.30

Table 6.3.1: Metrics that summarise the results of using a single multiclass DNN per 2ℓOS region, as described throughout this subsection. $t\bar{t}Z$ separation is measured between the $t\bar{t}Z$ 2ℓOS signal and total background. $t\bar{t}$ separation is measured between the $t\bar{t}$ background and all other samples (including $t\bar{t}Z$ 2ℓOS signal). Z+jets separation is measured between Z+jets background and all other samples (including $t\bar{t}Z$ 2ℓOS signal). “training set ROC AUC (average)" is the average of “training set ROC AUC (fold 1)" and “training set ROC AUC (fold 2)". “test set ROC AUC (average)" is the average of “test set ROC AUC (fold 1)" and “test set ROC AUC (fold 2)".

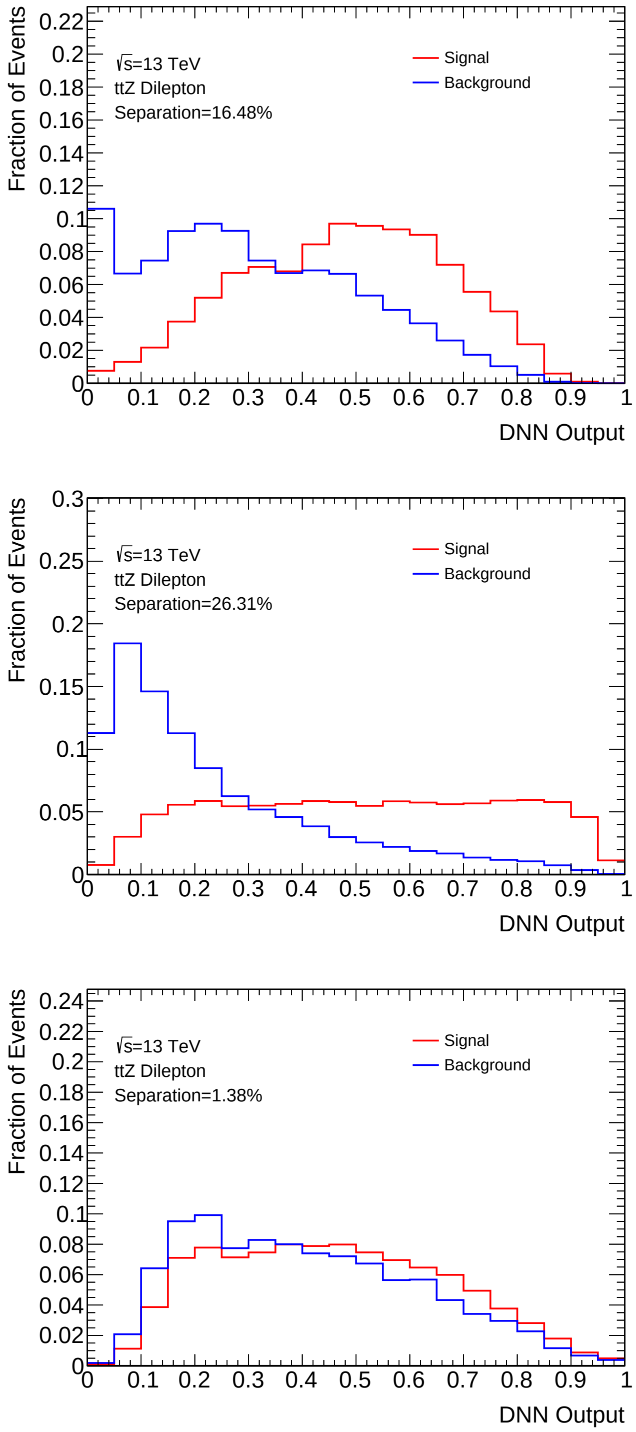


Figure 6.3.2: Separation of DNN output distributions the classifier output being optimised for, and all other processes, in the 2ℓ-Z-1b6j region of the dilepton OSSF channel, when using an initial multiclass DNN per 2ℓOS region. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. (Top row) is for the $t\bar{t}$ classifier output. (Middle row) is for the $t\bar{t}Z$ classifier output 1. (Bottom row) is for the Z+jets classifier output.

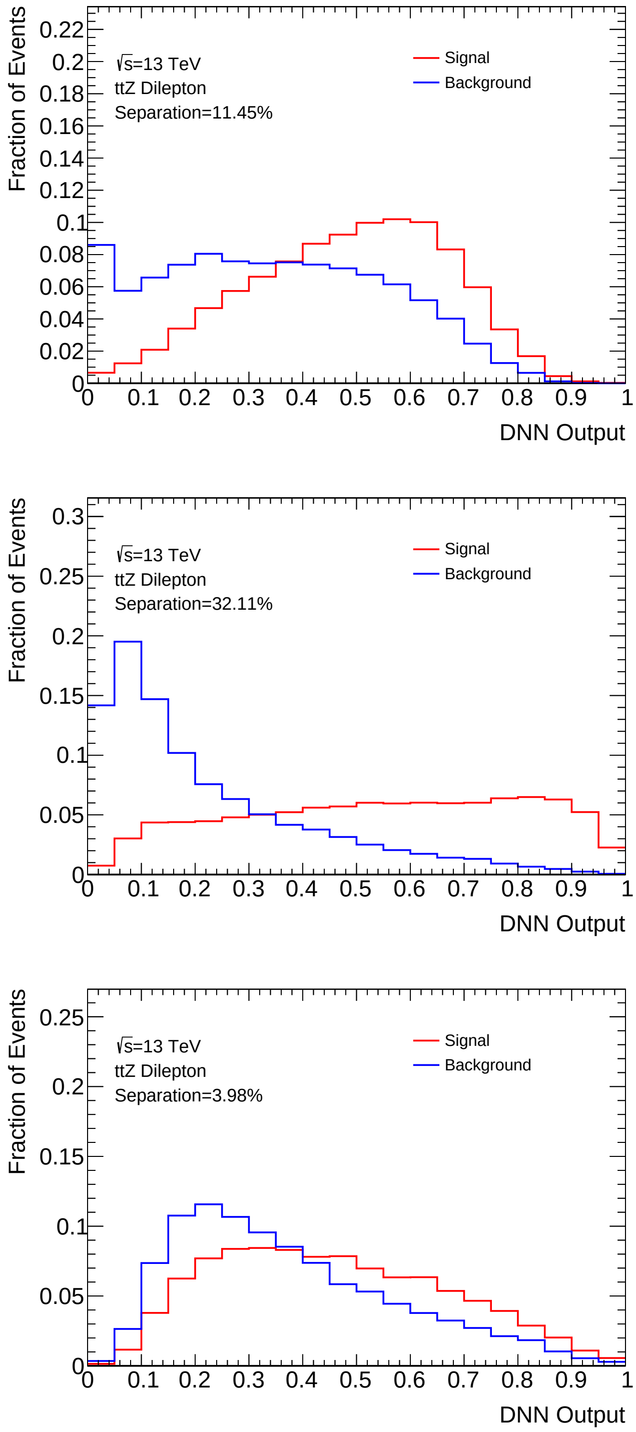


Figure 6.3.3: Separation of DNN output distributions the classifier output being optimised for, and all other processes, in the 2ℓ-Z-2b5j region of the dilepton OSSF channel, when using an initial multiclass DNN per 2ℓOS region. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. (Top row) is for the $t\bar{t}$ classifier output. (Middle row) is for the $t\bar{t}Z$ classifier output 1. (Bottom row) is for the Z+jets classifier output.

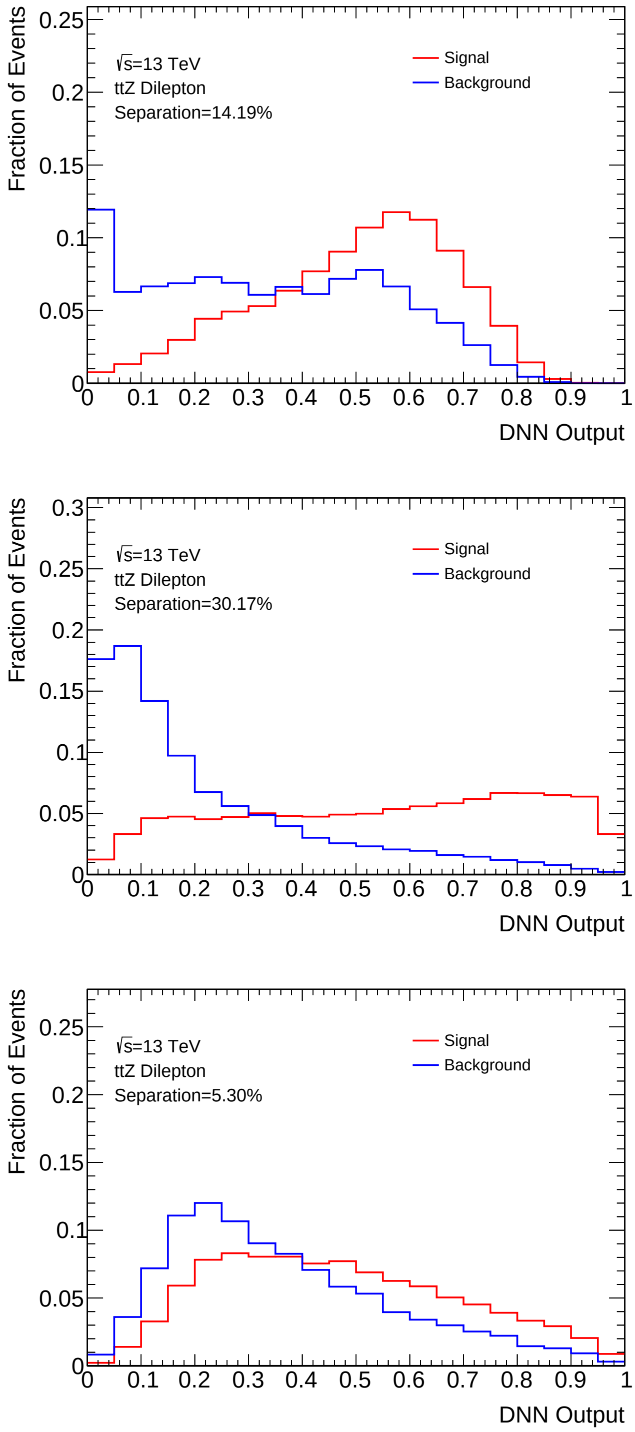


Figure 6.3.4: Separation of DNN output distributions the classifier output being optimised for, and all other processes, in the 2ℓ-Z-2b6j region of the dilepton OSSF channel, when using an initial multiclass DNN per 2ℓOS region. No cut on MVA technique output is required in separation plots, to show real separation power before cutting on the MVA output. (Top row) is for the $t\bar{t}$ classifier output. (Middle row) is for the $t\bar{t}Z$ classifier output 1. (Bottom row) is for the Z+jets classifier output.

6.3.5 Comparison with previous results

We now compare the initial multiclass DNNs to the BDTs from Ref. [36], as well as the previous results for BDTs and DNNs. Some metrics (2b5j AUC, 2b5j AUC, 2b5j separation) improve slightly (≥0.3%) in going from a binary DNN to a multiclass DNN whilst others slightly decrease. Even though only some metrics improve, the power of the multiclass approach arises is due to the extra information it gives about different background processes. Even so, all metrics for the multiclass DNN improve by a small value of ≥0.71% compared to the BDT approach from Ref. [36].

MVA	1b6j AUC	2b5j AUC	2b6j AUC	1b6j Separation %	2b5j Separation %	2b6j Separation %
BDT from [36]	0.618	0.585	0.615	25.6	27.6	28.6
BDT	0.788	0.791	0.796	25.5	30.7	30.3
DNN	0.808	0.823	0.822	27.4	31.1	31.6
multiclass DNN	0.801	0.833	0.825	26.3	32.1	30.2

Table 6.3.2: Comparison of metrics that summarise the results of 1) the BDTs from Ref. [36], 2) the BDTs developed for this thesis, 3) the binary DNN approach developed for this thesis, 4) the initial multiclass DNN approach developed for this thesis. Separation is measured between $t\bar{t}Z$ 2ℓOS signal and all backgrounds. All metrics improve in going from a BDT to a DNN.

Having only improved upon the initial BDTs with mutliclass DNNs by ≥1.3%, it is then necessary to test whether new variables can contribute to the optimisation of multiclass DNNs, for even more accurate classification of the $t\bar{t}Z$ 2ℓOS process.

6.4 2ℓOS multiclass DNN with new variables

Section 6.3 shows results having trained multiclass DNNs using the same variables as the single BDTs of Section 6.1. This section introduces some new variables and alters specific DNN settings. The aim of introducing new variables is to provide extra separation power between $t\bar{t}Z$ 2ℓOS and background.

6.4.1 Differences compared to initial multiclass DNN

Compared to the initial multiclass DNN of Section 6.3, MinDelta is increased from 0.0001 to 0.001. MinDelta is the minimum change in the monitored quantity to qualify as an improvement (accuracy in the case of this measurement). Several values of MinDelta were tested. A MinDelta of 0.0001 as in the initial multiclass DNN showed slight signs of overtraining. A MinDelta of 0.01 stopped training after around 40 epochs. To continue training past this minimum number of epochs up to the 100 epochs that is set as the maximum, a MinDelta of 0.001 was found to be a suitable balance between training for a sufficient number of epochs, whilst avoiding overtraining, after trying a few different values.

6.4.2 DNN options

Settings options of the DNN can be found in Table 6.4.1. This section changes the OutputActivation from sigmoid to softmax and introduces a SmoothingAlpha of 0.1, explained in Table 6.4.1. Softmax is used to achieve a diagonal response. The response matrices for this section can be seen later in Figure 6.4.13. SmoothingAlpha is used to try reduce overtraining. The overtraining check for this section can be seen later in Figure 6.4.5.

Option	Values	Description
InputScaling	minmax	inputs are scaled into a range between 0 and 1
Folds	2	how many folds (k-folding) should be performed during training
Nodes	50,50,50,50	Comma-separated list of neurons for each layer
Loss	categorical crossentropy	Loss function which is used in the training of a model
Epochs	100	Number of training epochs
LearningRate	0.001	Initial learning rate for the training of a model
BatchSize	32	Batch size used in training of a model
ValidationSize	0.2	Relative size of the validation set used during training of a model
Patience	30	Number of epochs with no improvement after which training will be stopped
MinDelta	0.001	Minimum change in the monitored quantity to qualify as an improvement
DropoutIndice	1,3	Layer indeces at which Dropout layers are added
DropoutProb	0.1	Probability of dropout
OutputSize	3	Number of neurons in the output layer
OutputActivation	softmax	Activation function in the output layer of a model
Metrics	Accuracy	Comma-separated list of metrics to be evaluated during training
ModelBinning	20,0,1	Custom binning using a fixed bin width in the format nbins,x_low,x_high
SmoothingAlpha	0.1	to be applied to smooth labels according using Y=Y(1-alpha)+alpha/K

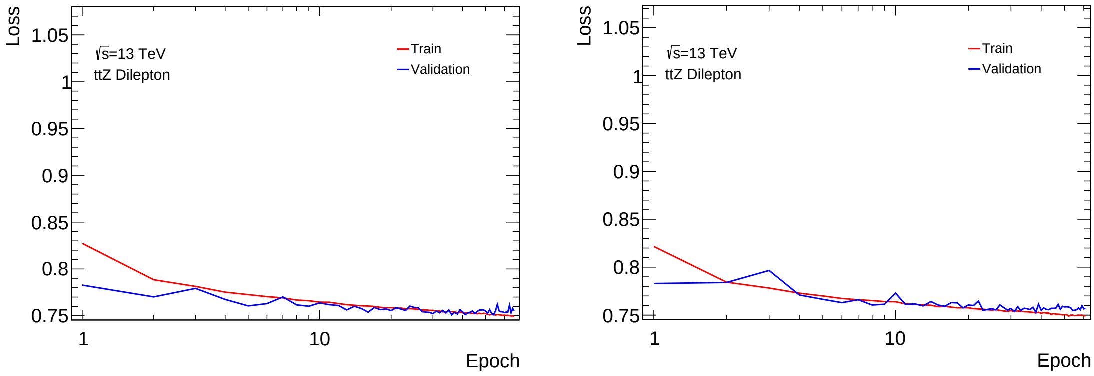
Table 6.4.1: Settings used in DNN training, when using the multiclass DNNs with new variables.

6.4.3 DNN visualisation

The visualisations of the DNNs in this section look similar to the visualisations of the binary DNNs in Section 6.2, with the main difference being the number of nodes in the output layer. There are now 3 output nodes for multiclassification, as opposed to 1 for binary classification. The number of input nodes could also be different compared to the binary DNNs, because different numbers of input variables are optimal. The visualisations for 2ℓ-Z-2b6j, 2ℓ-Z-1b6j and 2ℓ-Z-2b5j are similar, apart from the number of nodes in the input layer.

6.4.4 Loss curves for DNN training

Loss curves visualise how neural networks evolve over the number of epochs learned for. During the learning process, the loss should decrease as the minimum of the loss function is approached. Loss curves of the DNN training for each 2ℓOS region were produced to check for overtraining. In them, the fact that the Train line does not finish at a "Loss" value significantly below the Validation line indicates that no significant overtraining is present. As an example, the loss curves for the 2ℓ-Z-2b6j region are shown in Figure 6.4.1.



(Left column) 1st k-fold

(Right column) 2nd k-fold

Figure 6.4.1: Loss curves for DNN training, using the multiclass DNNs with new variables. 1st k-fold in the left column, 2nd k-fold in the right column. 2ℓ-Z-2b6j.

6.4.5 Variables importance for DNN

The variables chosen for training the DNN are given in Figure 6.4.2-6.4.4, along with the ranking of the variables used. These ranking plots emphasise the importance of variables related to jet p_{T}, such as "Jet $p_{T,1}$" and "H_T^{6jets}". Separate training was done for the three regions: 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j. The number of variables used for each 2ℓOS region varied: 16 for 2ℓ-Z-1b6j, 15 for 2ℓ-Z-2b5j and 13 for 2ℓ-Z-2b6j.

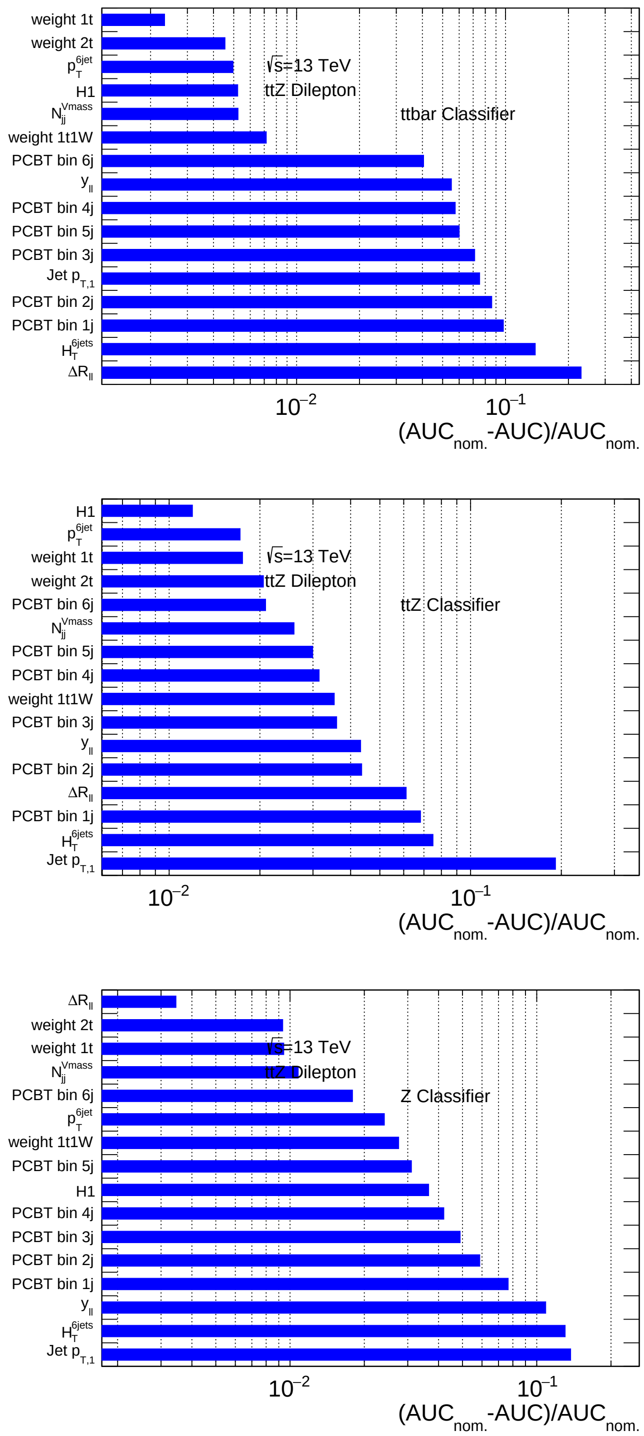


Figure 6.4.2: "Importance" of the variables used for 2ℓ-Z-1b6j DNN training, using the multiclass DNNs with new variables. Variables with greatest values on the x-axis are most "important". (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.

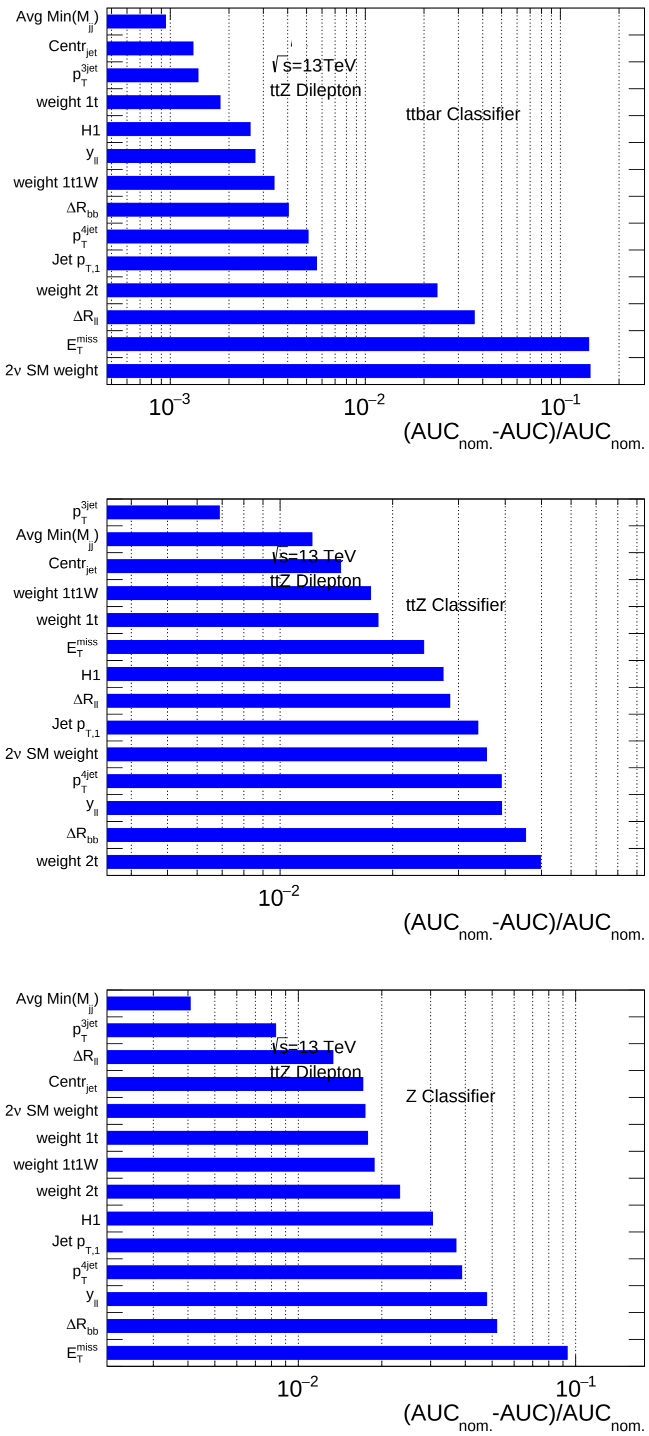


Figure 6.4.3: "Importance" of the variables used for 2ℓ-Z-2b5j DNN training, using the multiclass DNNs with new variables. Variables with greatest values on the x-axis are most "important". (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.

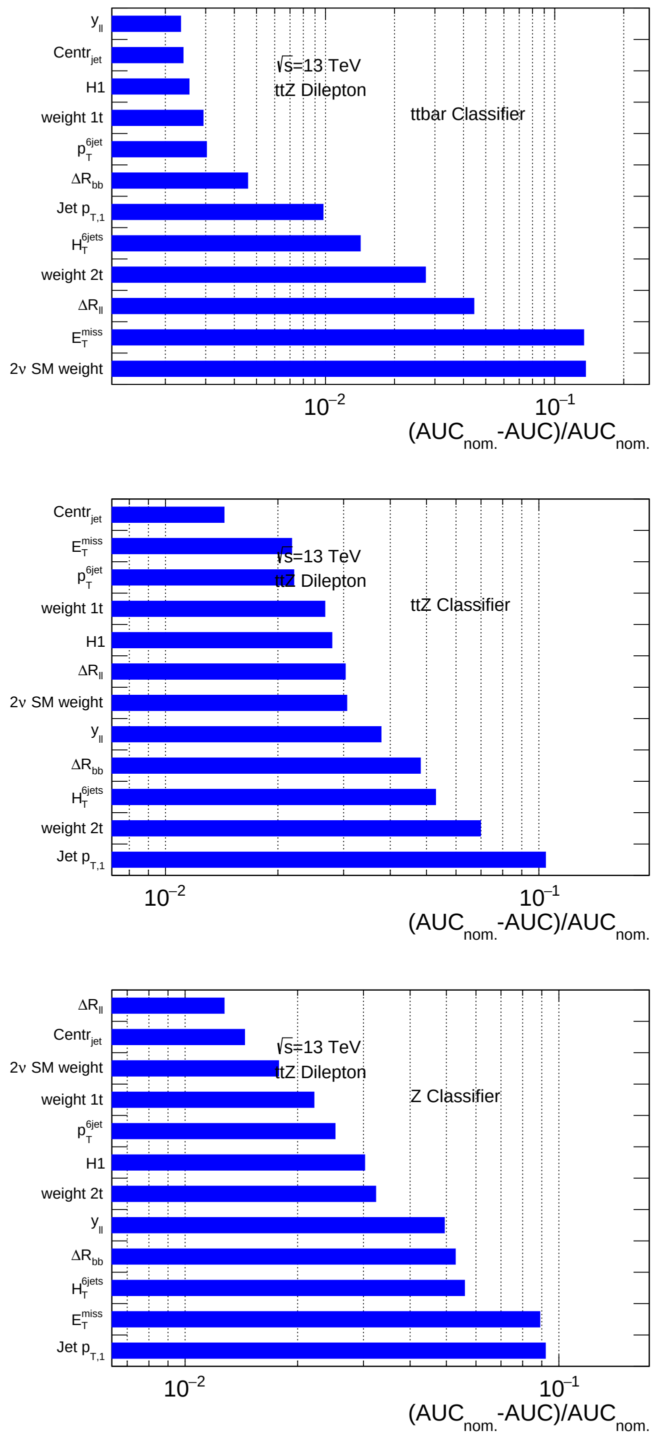
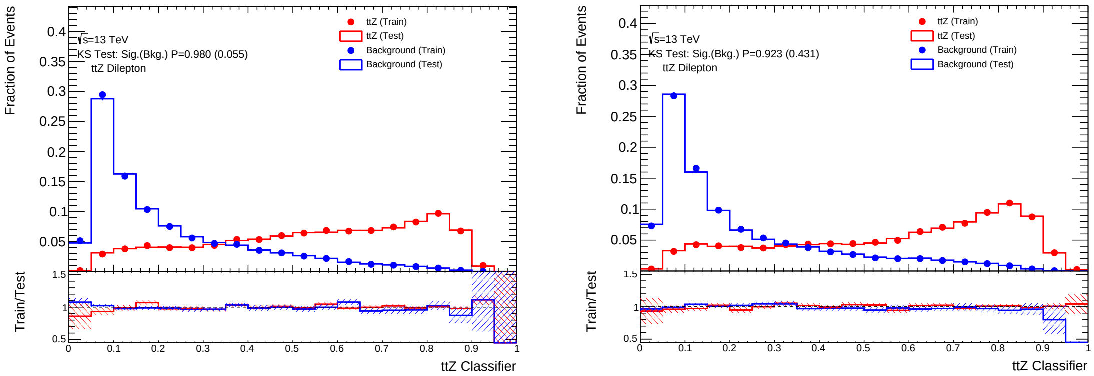


Figure 6.4.4: “Importance" of the variables used for 2ℓ-Z-2b6j DNN training, using the multiclass DNNs with new variables. Variables with greatest values on the x-axis are most “important". (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.

6.4.6 DNN overtraining test

To fully ensure that the DNN is properly trained and its results do not depend on the event statistics, another overtraining test should to be performed. In this overtraining test, the DNN discriminants (training and test samples) used for the regions of the dilepton OSSF channel are compared. An example for the 2ℓ-Z-2b6j region is shown in Figure 6.4.5. Error bars have been added to the overtraining checks. The fact that Train dots agree with Test bars within uncertainties indicates that no overtraining is present.

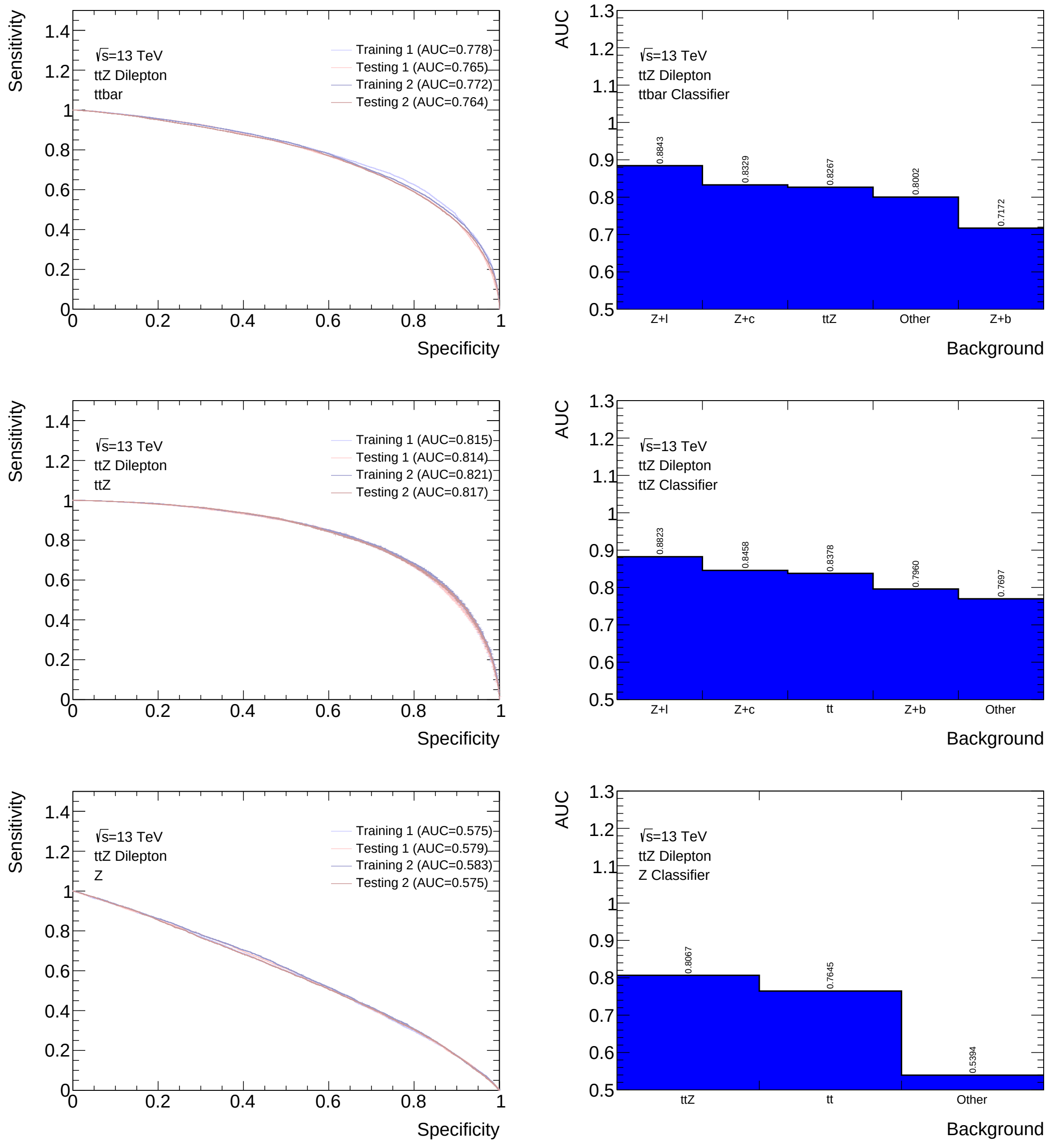


(Left column) 1st k-fold

(Right column) 2nd k-fold

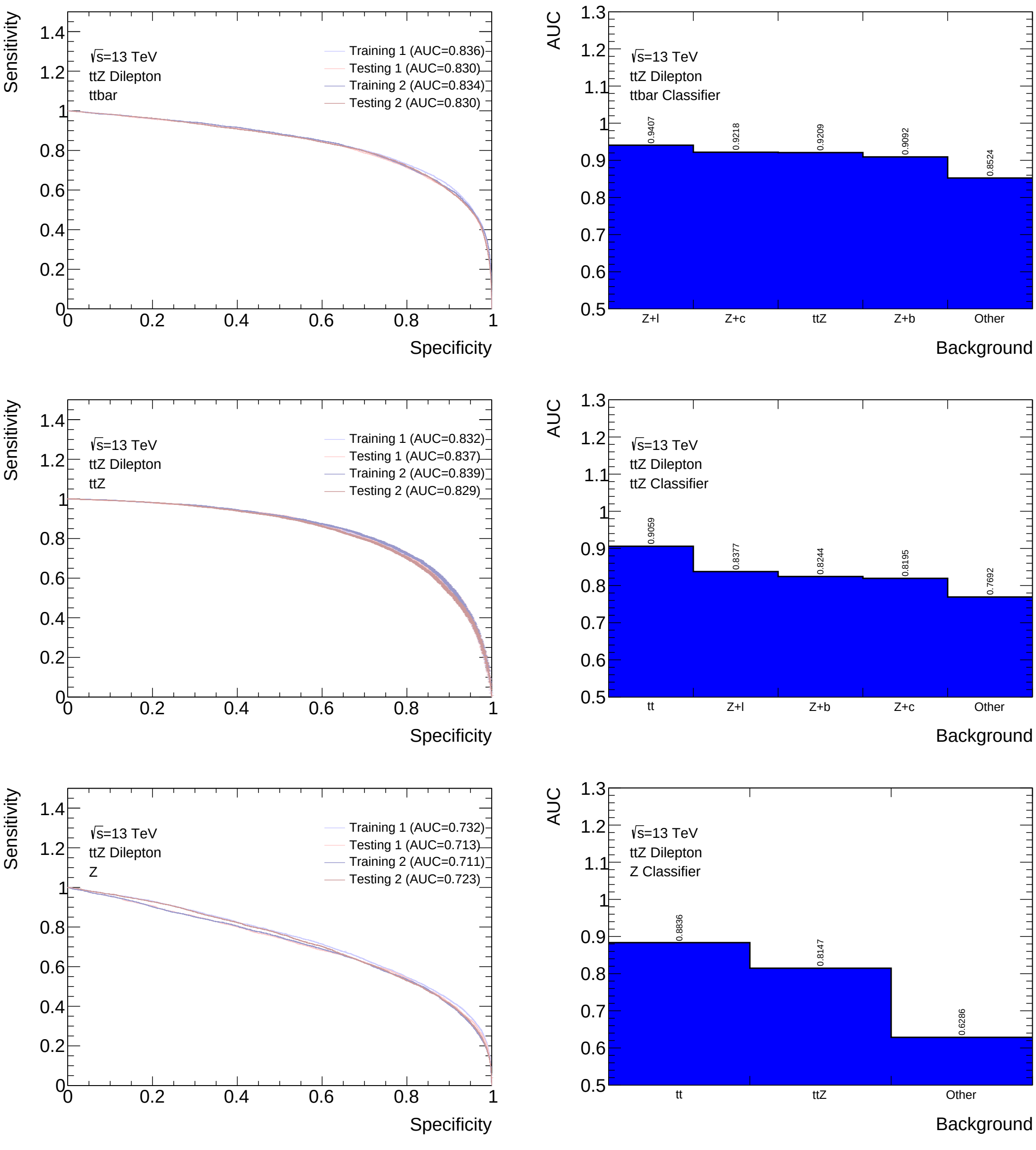
Figure 6.4.5: DNN discriminants used for the 2ℓ-Z-2b6j signal region of the dilepton OSSF channel overlapping test and training samples, using the multiclass DNNs with new variables. The errors are MC statistics.

The ROC curves for the mutliclass DNNs with new variables are shown in Figures 6.4.6-6.4.8. Note that there is no noticeable overtraining. Figures 6.4.6-6.4.8 also summarise the performance for each background. This shows how different each background is from the $t\bar{t}Z$ 2ℓOS signal. The background process most similar to $t\bar{t}Z$ 2ℓOS signal will be hardest to distinguish from $t\bar{t}Z$ 2ℓOS signal, and thus the AUC score will be lowest. The background process least similar to $t\bar{t}Z$ 2ℓOS signal will be easiest to distinguish from $t\bar{t}Z$ 2ℓOS signal, and thus the AUC score will be highest. Which process is most or least similar to $t\bar{t}Z$ 2ℓOS signal might be different for each $t\bar{t}Z$ 2ℓOS region. “Other" contains some processes with rather similar signatures to $t\bar{t}Z$ 2ℓOS signal, such as tWZ, though there are not many events that pass the $t\bar{t}Z$ 2ℓOS selection requirements of Table 5.2.2. For the other multiclass outputs ($t\bar{t}$ & Z+jets), Figures 6.4.6-6.4.8 also shows how similar other processes are to the process that the classifier output is optimising for.



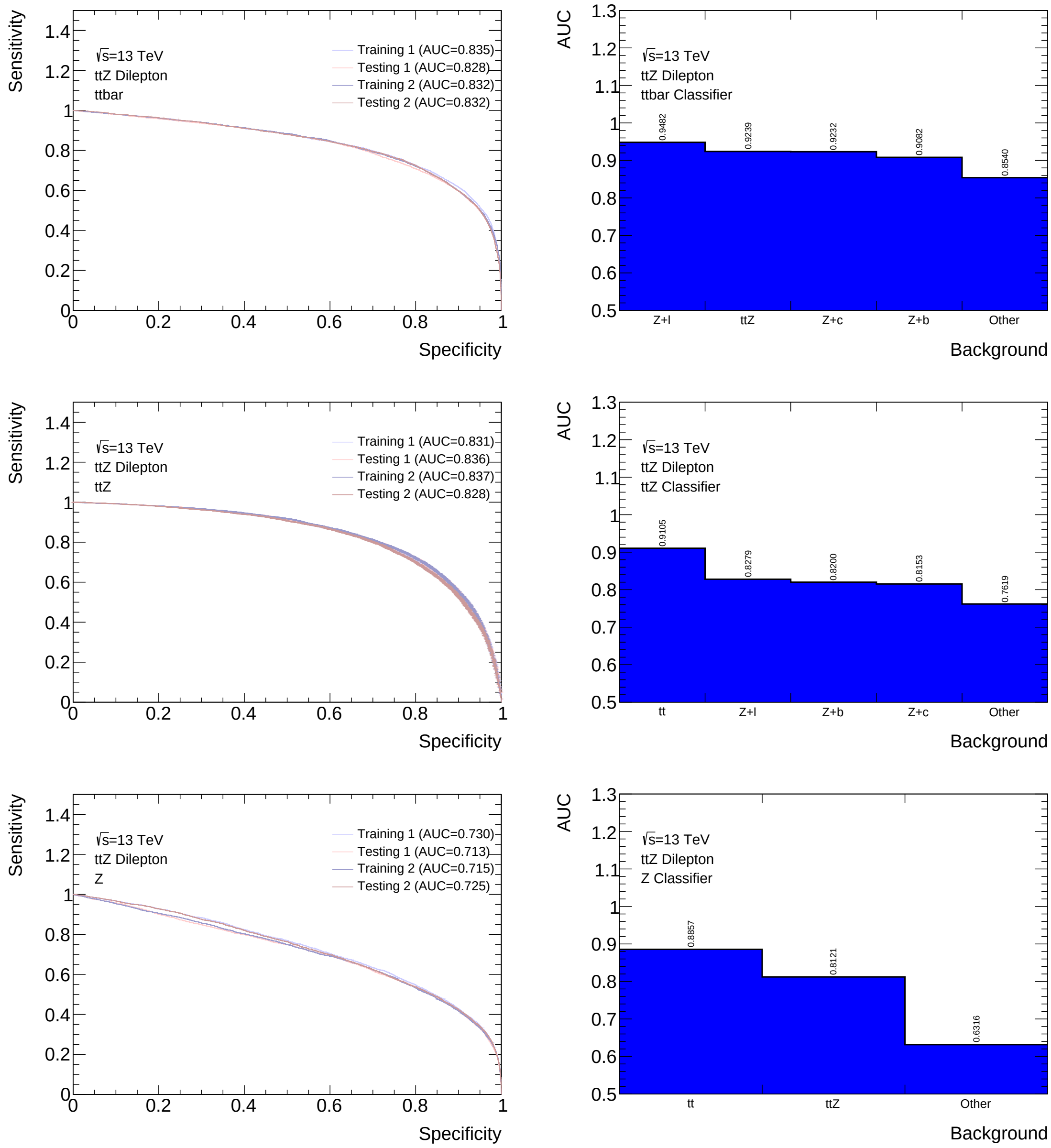
(Left column) ROC curves (Right column) AUC summary

Figure 6.4.6: ROC curves and AUC summaries for the 2ℓ-Z-1b6j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.



(Left column) ROC curves (Right column) AUC summary

Figure 6.4.7: ROC curves and AUC summaries for the 2ℓ-Z-2b5j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.

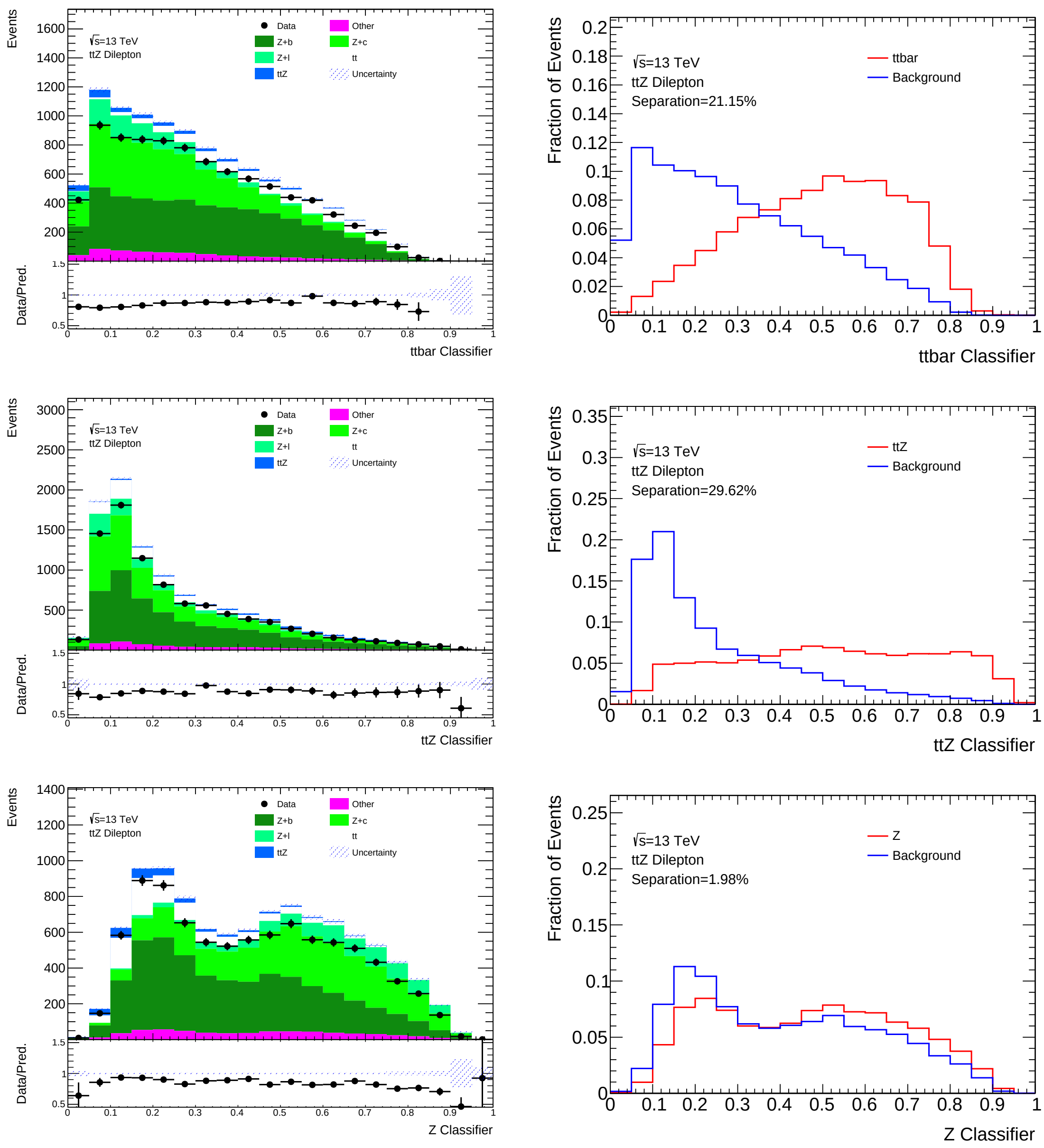


(Left column) ROC curves (Right column) AUC summary

Figure 6.4.8: ROC curves and AUC summaries for the 2ℓ-Z-2b6j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets.

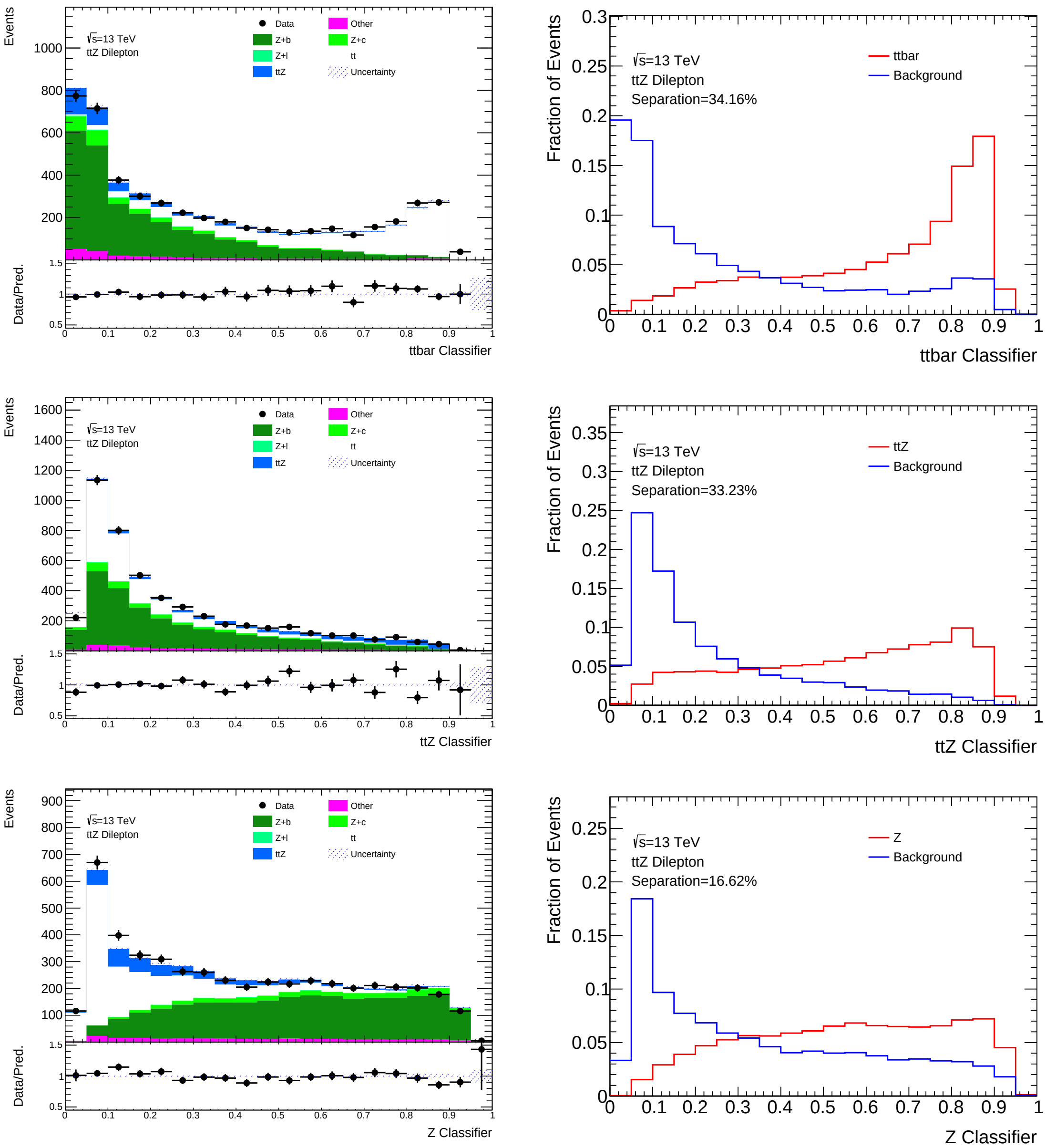
6.4.7 DNN output distribution

Figures 6.4.9-6.4.11 show the DNN output distribution in MC and data. Figures 6.4.9-6.4.11 also show the separation between $t\bar{t}Z$ 2ℓOS signal and total background for the DNN output. For all 3 classifier outputs, data and MC mostly agree, and there is >29.5% separation between signal and background. Such separation allows for better significance measures, as shown in Section 6.4.8.



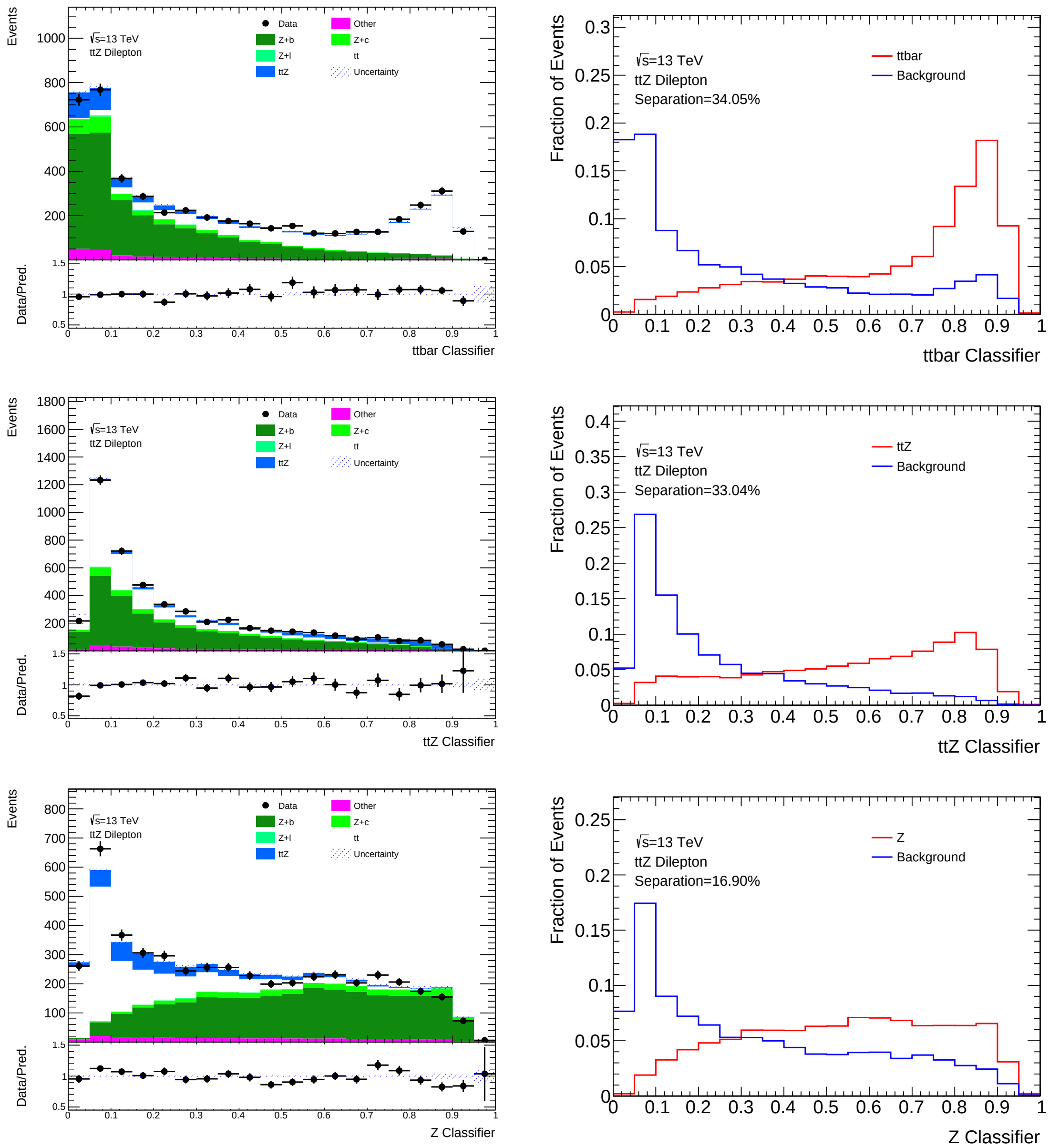
(Left column) Data vs. MC (Right column) Separation

Figure 6.4.9: DNN discriminants and separation for the 2ℓ-Z-1b6j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets. The errors are MC statistics.



(Left column) Data vs. MC (Right column) Separation

Figure 6.4.10: DNN discriminants and separation for the 2ℓ-Z-2b5j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets. The errors are MC statistics.

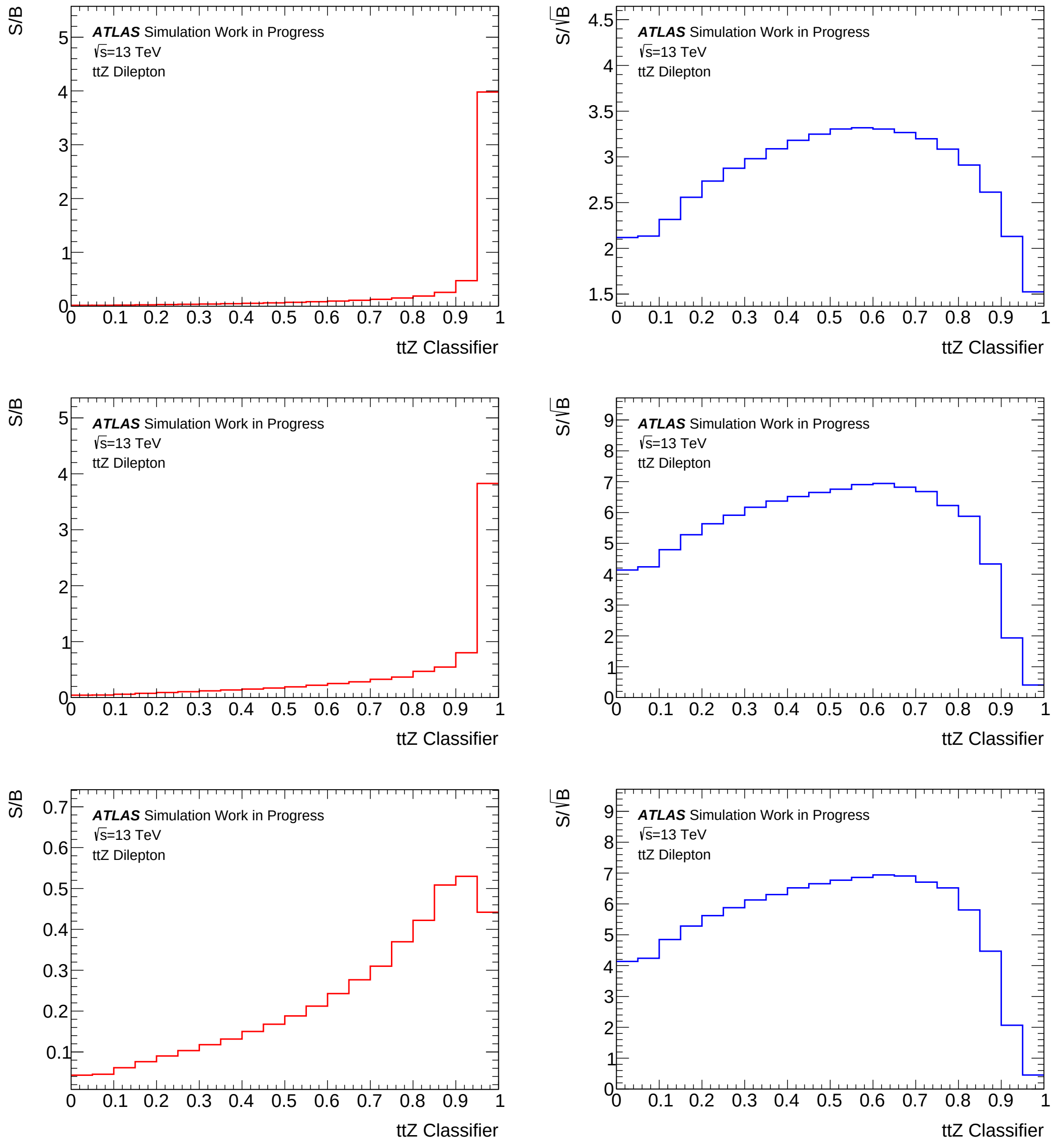


(Left column) Data vs. MC   (Right column) Separation

Figure 6.4.11: DNN discriminants and separation for the 2ℓ-Z-2b6j signal region of the dilepton OSSF channel, using the multiclass DNNs with new variables. (Top row) is for the classifier output for $t\bar{t}$. (Middle row) is for the classifier output for $t\bar{t}Z$. (Bottom row) is for the classifier output for Z+jets. The errors are MC statistics.

6.4.8 DNN significance measures

S/B and statistical significance distributions are shown in Figure 6.4.12. They show that an S/B of $\gtrsim$0.5 and statistical significance >3.3 are achievable in all 3 regions, with a statistical significance >3 being called "evidence".



(Left column) S/B   (Right column) $S/\sqrt{B}$

Figure 6.4.12: Signal over background and statistical significance that would be achieved by selecting events above the x-axis DNN output value, using the multiclass DNNs with new variables. (Top row) 2ℓ-Z-1b6j. (Middle row) 2ℓ-Z-2b5j. (Bottom row) 2ℓ-Z-2b6j.

6.4.9 DNN response

2D responses of different samples to the "ttbar Classifier" and "Z Classifier" were produced for all three signal regions. As an example, the responses for the 2ℓ-Z-2b6j region are shown in Figure 6.4.13. They show how:

- $t\bar{t}Z$ is concentrated towards low values of “ttbar Score" and low values of “Z Score", as shown in the top-left sub-figures of Figure 6.4.13;
- $t\bar{t}$ is concentrated towards high values of “ttbar Score" and low values of “Z Score", as shown in the top-right sub-figures of Figure 6.4.13;
- $Z + b$ is concentrated towards low values of “ttbar Score" and high values of “Z Score", as shown in the middle-left sub-figures of Figure 6.4.13;
- $Z + c$ is concentrated towards low values of “ttbar Score" and high values of “Z Score", as shown in the middle-right sub-figures of Figure 6.4.13;

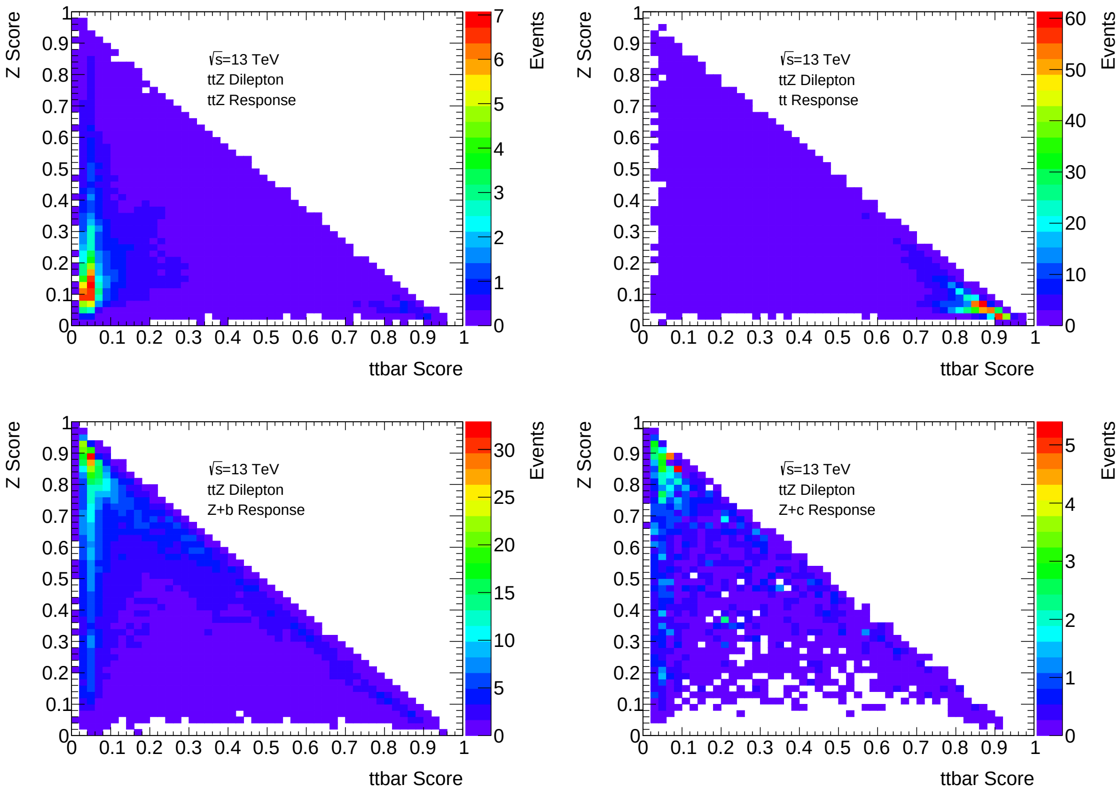


Figure 6.4.13: 2D response of different samples to the “$t\bar{t}$ Classifier" and “Z Classifier", using the multiclass DNNs with new variables in the 2ℓ-Z-2b6j region.

6.4.10 Summary of optimised multiclass DNN results

Table 6.4.2 summarises the main numerical metrics of this subsection using an optimised single DNN per 2ℓOS region. Information on the number of variables is shown in Section 5.6. The training and test data set ROC AUCs are shown in Section 6.4.6. Separation distributions are shown in Section 6.4.7.

Metric	2ℓ-Z-$1b6j$	2ℓ-Z-$2b5j$	2ℓ-Z-$2b6j$
number of variables	16	15	13
training ROC AUC (fold 1) $t\bar{t}$	0.778	0.836	0.835
test ROC AUC (fold 1) $t\bar{t}$	0.765	0.830	0.828
training ROC AUC (fold 2) $t\bar{t}$	0.772	0.834	0.832
test ROC AUC (fold 2) $t\bar{t}$	0.764	0.830	0.832
training ROC AUC (average) $t\bar{t}$	0.775	0.835	0.8335
test ROC AUC (average) $t\bar{t}$	0.7645	0.830	0.830
ROC AUC (average) $t\bar{t}$	0.76975	0.8325	0.83175
training ROC AUC (fold 1) $t\bar{t}Z$	0.815	0.832	0.831
test ROC AUC (fold 1) $t\bar{t}Z$	0.814	0.837	0.836
training ROC AUC (fold 2) $t\bar{t}Z$	0.821	0.839	0.837
test ROC AUC (fold 2) $t\bar{t}Z$	0.817	0.829	0.828
training ROC AUC (average) $t\bar{t}Z$	0.818	0.8355	0.834
test ROC AUC (average) $t\bar{t}Z$	0.8155	0.833	0.832
ROC AUC (average) $t\bar{t}Z$	0.81675	0.83425	0.833
training ROC AUC (fold 1) Z	0.575	0.732	0.730
test ROC AUC (fold 1) Z	0.579	0.713	0.713
training ROC AUC (fold 2) Z	0.583	0.711	0.715
test ROC AUC (fold 2) Z	0.575	0.723	0.725
training ROC AUC (average) Z	0.579	0.7215	0.7225
test ROC AUC (average) Z	0.578	0.718	0.719
ROC AUC (average) Z	0.5785	0.71975	0.72075
$t\bar{t}Z$ AUC in $t\bar{t}$ Classifier	0.8267	0.9209	0.9239
$Z+l$ AUC in $t\bar{t}$ Classifier	0.8843	0.9407	0.9482
$Z+c$ AUC in $t\bar{t}$ Classifier	0.8329	0.9218	0.9232
$Z+b$ AUC in $t\bar{t}$ Classifier	0.7172	0.9092	0.9082
Other AUC in $t\bar{t}$ Classifier	0.8002	0.8524	0.8540
$t\bar{t}$ AUC in $t\bar{t}Z$ Classifier	0.8378	0.9059	0.9105
$Z+l$ AUC in $t\bar{t}Z$ Classifier	0.8823	0.8377	0.8279
$Z+c$ AUC in $t\bar{t}Z$ Classifier	0.8458	0.8195	0.8153
$Z+b$ AUC in $t\bar{t}Z$ Classifier	0.7960	0.8244	0.8200
Other AUC in $t\bar{t}Z$ Classifier	0.7967	0.7692	0.7619
$t\bar{t}Z$ AUC in Z Classifier	0.8067	0.8147	0.8121
$t\bar{t}$ AUC in Z Classifier	0.7645	0.8836	0.8857
Other AUC in Z Classifier	0.5394	0.6286	0.6316
$t\bar{t}$ separation (%)	21.15	34.16	34.05
$t\bar{t}Z$ separation (%)	29.62	33.23	33.04
Z+jets separation (%)	1.96	16.62	16.90
$t\bar{t}Z$ S/$\sqrt{B}$	3.3	7.0	7.0

6.4.11 Comparison with previous results

We now compare the multiclass DNNs with new variables to the BDTs from Ref. [36], as well as the previous results for BDTs and DNNs. The optimised multiclass DNN approach provides the highest performance of all the algorithms tested, with an improvement of >4% in all metrics.

MVA	1b6j AUC	2b5j AUC	2b6j AUC	1b6j Separation %	2b5j Separation %	2b6j Separation %
BDT from [36]	0.618	0.585	0.615	25.6	27.6	28.6
BDT	0.788	0.791	0.796	25.5	30.7	30.3
DNN	0.808	0.823	0.822	27.4	31.1	31.6
multiclass DNN	0.801	0.833	0.825	26.3	32.1	30.2
optimised multiclass DNN	0.817	0.834	0.833	29.6	33.2	33.0

Table 6.4.3: Comparison of metrics that summarise the results of 1) the BDTs from Ref. [36], 2) the BDTs developed for this thesis, 3) the binary DNN approach developed for this thesis, 4) the initial multiclass DNN approach developed for this thesis, 5) the optimised multiclass DNN approach developed for this thesis. Separation is measured between $t\bar{t}Z$ 2ℓOS signal and all backgrounds. The optimised multiclass DNN approach (5) provides the best performance in all metrics.

6.5 Fitting results

The effect of different systematic uncertainties on the final result of the expected $t\bar{t}Z$ signal strength can be quantified in a so called “ranking plot". The ranking plot shows the 20 NPs (Section 5.7.1) that affect the fitted signal strength, μ, the most. The upper x-axis of a ranking plot quantifies how much a particular NP affects the fitted signal strength, shown as $\Delta\mu$. The ranking plot for the combination of multiclass DNNs with new variables in the 3 signal regions 2ℓ-Z-2b6j, 2ℓ-Z-1b6j, and 2ℓ-Z-2b5j is shown in Figure 7.3.5.

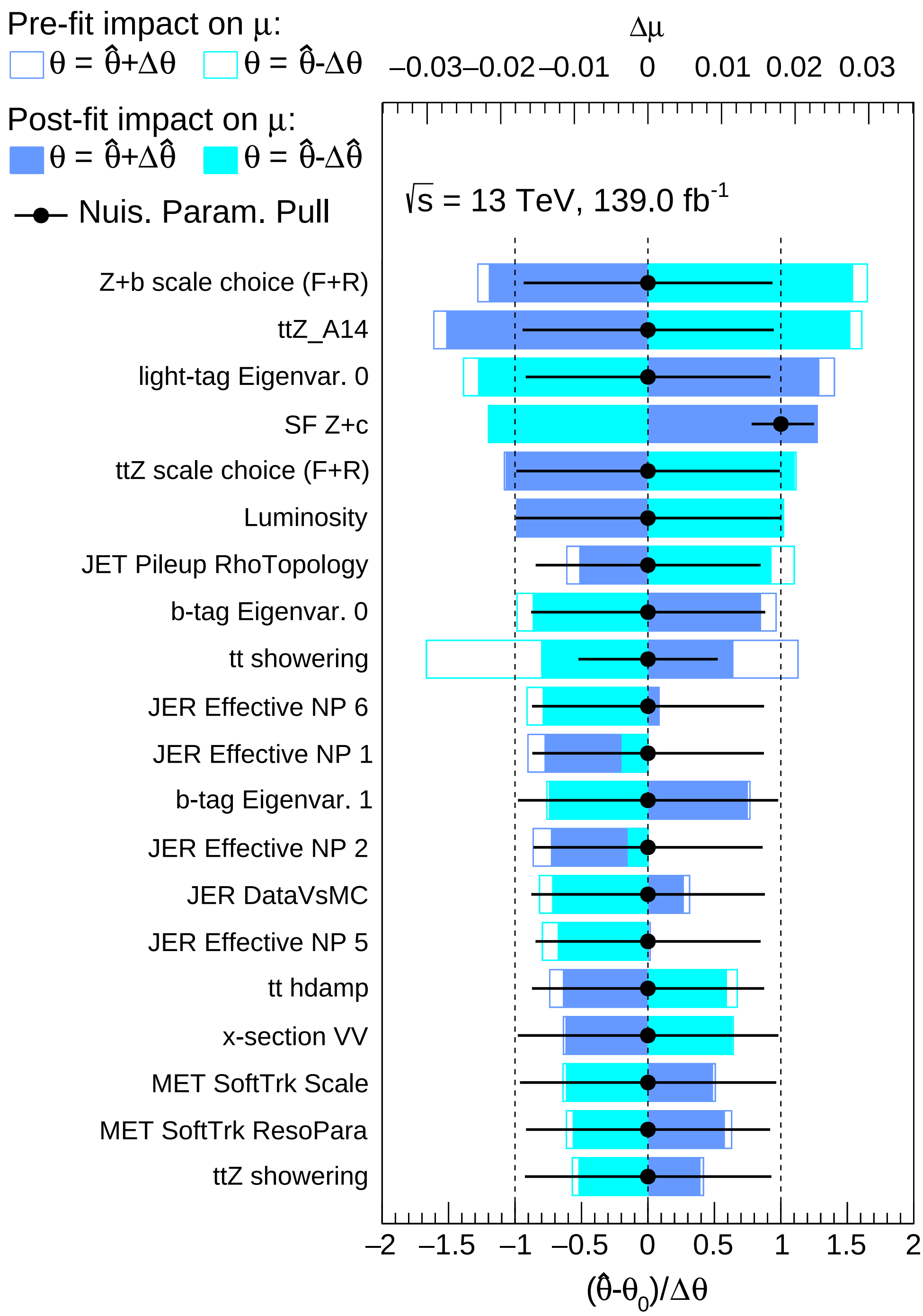


Figure 6.5.1: Ranking plot for the combination of multiclass DNNs with new variables in the 3 signal regions 2ℓ-Z-2b6j, 2ℓ-Z-1b6j, 2ℓ-Z-2b5j.

The 3 systematic uncertainties with greatest effect on the final result of the expected $t\bar{t}Z$ signal strength are:

1. combination of renormalisation and factorisation scale choice ($\mu_R\mu_F$);
2. $t\bar{t}Z$ signal A14 variation, to quantify ISR uncertainty;
3. light-tag Eigenvariation 0, related to b-tagging.

These systematic uncertainties were introduced in Section 5.7.1. The fact that the top 3 uncertainties in the ranking plot are systematic means that the $t\bar{t}Z$ 2ℓOS inclusive cross-section measurement is systematics limited.

Final values for the uncertainty on the expected $t\bar{t}Z$ signal strength, μ (which was introduced in Section 5.7), are presented in Table 6.5.1. A comparison with the previously published $t\bar{t}Z$ 2ℓOS paper (Ref. [36]) is also included. An uncertainty on μ of e.g. +0.13 means the up uncertainty on the $t\bar{t}Z$ 2ℓOS cross-section will have a relative uncertainty of 13%. A relative uncertainty of 13% is the most precise measurement of the $t\bar{t}Z$ 2ℓOS process performed by ATLAS to date. The improvement in uncertainty on μ appears greater than the improvement in ROC AUCs and separation between improvements in ROC AUC and separation can plateau, whilst still leading to large improvements in uncertainty on μ. The final result for $t\bar{t}Z$ cross-section will have a lower uncertainty still, because the 2ℓOS channel will be combined with the 3ℓ and 4ℓ channels.

MVA	μ=1
BDT from [36]	$^{+0.31}_{-0.29}$
BDT	$^{+0.17}_{-0.17}$
multiclass DNN with new variables	$^{+0.13}_{-0.12}$

Table 6.5.1: Comparison of metrics that summarise the results of 1) the BDTs from Ref. [36], 2) the BDTs developed for this thesis, 3) the multiclass DNN approach with new variables developed for this thesis. The multiclass DNN approach (3) provides the lowest uncertainties, and thus best results.

6.6 Combination results and summary

Fitted results with a precision of $\mu = 1^{+0.13}_{-0.12}$ from the 2ℓOS channel can then be combined with the 3ℓ and 4ℓ channels to obtain a more precise value of $\mu = 1^{+0.065}_{-0.063}$ [193], the most precise $t\bar{t}Z$ inclusive cross-section measurement performed by ATLAS. $\mu = 1^{+0.080}_{-0.080}$ is obtained in the 3ℓ channel and $\mu = 1^{+0.13}_{-0.12}$ is obtained in the 4ℓ channel. Therefore, the 2ℓOS channel is now as precise as the 4ℓ channel. This combined result of $\mu = 1^{+0.065}_{-0.063}$ [193] is nearly twice more precise than the previous ATLAS public result on the $t\bar{t}Z$ inclusive cross-section, $\mu = 1^{+0.115}_{-0.109}$ [37]. In other words, the 2ℓOS measurement alone is now almost as precise as the previous 3ℓ+4ℓ measurement. This combined result of $\mu = 1^{+0.065}_{-0.063}$ [193] will be published in an upcoming ATLAS paper. The development of precise MVA techniques in the 2ℓOS channel contributed to this measurement. Having measured the $t\bar{t}Z$ 2ℓOS inclusive cross-section precisely, a follow-up paper will be able to measure the differential cross-section of the $t\bar{t}Z$ 2ℓOS process. The exact variables are as yet unknown, but could include for example the transverse momentum of the Z boson and the transverse momentum of the $t\bar{t}$ system.

7 Preparing ATLAS data for education worldwide

Respect your parents. They passed school without Google.

Anon [194]

This chapter discusses the education work that forms part of this thesis - ATLAS Open Data. The ATLAS Open Data project provides open-source access to measured data, simulation, resources, and documentation for the purpose of education. ATLAS was the first LHC experiment to release real 13 TeV collision data [170, 195]. The development and testing of specific resources related to the $t\bar{t}Z$ 2ℓOS process are discussed in this chapter. It is important to point out however, that many other resources unrelated to the $t\bar{t}Z$ 2ℓOS process were also developed. All data and resources can be accessed from the ATLAS Open Data website [152]. This chapter is structured as follows:

1. discussion of the Histogram Analyser;
2. discussion of ATLAS Open Data Jupyter notebooks.

The author's specific contribution was to:

- create a data pipeline to go from 13 TeV data used for physics analysis to simplified data formats, which then allowed the creation of datasets that could be used for the $t\bar{t}Z$ 2ℓOS Histogram Analyser and Jupyter notebooks;
- create the 13 TeV datasets used as input for Open Data analyses, including those used in the $t\bar{t}Z$ 2ℓOS Histogram Analyser and Jupyter notebooks;
- write example physics analyses for use with 13 TeV ATLAS Open Data, for example the $t\bar{t}Z$ 2ℓOS Histogram Analyser and Jupyter notebooks;

- write corresponding documentation for 13 TeV datasets and example analyses, similar to the accompanying explanations given throughout this chapter;
- test 13 TeV datasets and example analyses, for example through the $t\bar{t}Z$ 2ℓOS Histogram Analyser and Jupyter notebooks.

7.1 The data

The 13 TeV ATLAS Open Data release constitutes 10 fb^{-1}of experimental data, which is approximately 1/14th of the data collected by ATLAS in Run 2. 10 fb^{-1}correspond to approximately 1000 trillion proton-proton collisions.The whole release is in .root file format, along with csv file formats for some specific processes. The variables present in the datasets were summarised in Table 4.4.1, and further information can be found in Ref. [170]. The data can be accessed through the ATLAS Open Data portal [152] or CERN Open Data portal [153]. Analysis of these data is possible through a number of tools, including the Histogram Analyser (Section 7.2) and Jupyter notebooks (Section 7.3)

7.2 Histogram Analyser

The Histogram Analyser is one of the main web-based resources that was developed for using ATLAS data for education. It allows students to apply selection requirements to histograms without the need to use computer code. It is possible to apply selection requirements on eight different variables, all of which are presented as individual histograms. This section introduces and covers the $t\bar{t}Z$ Histogram Analyser, the individual histograms that form it, and conclusions that can be drawn from three different signal regions. The $t\bar{t}Z$ Histogram Analyser is focused on because the author of this thesis was the main developer.

7.2.1 Introduction

The ATLAS Open Data Histogram Analyser [196, 197] is a web-based tool for fast, cut-based analysis of data, allowing to visualise data using online histograms with only a computer mouse. This tool shows how to differentiate between physics processes. By applying cuts to data, specific physics processes (signal) can be isolated from the background. The webpage [197] displays nine histograms of variables which can be used to isolate signal events. One can use their cursor to apply selections to a particular variable. Cutting on one histogram cuts the whole datasets, therefore changing the distributions of all 9 histograms - the effect on the other variables will be shown immediately. The Histogram Analyser helps in understanding the data and the relationship between the signal and background processes. It can simplify and speed-up the selection of cuts, before coding an analysis. The Histogram Analyser is used for an initial look at the $t\bar{t}Z$ 2ℓOS process.

7.2.2 The $t\bar{t}Z$ Histogram Analyser

The $t\bar{t}Z$ Histogram Analyser is used to help visualise rare top-quark measured data and simulations. This Histogram Analyser searches for rare top-quark processes. Data are shown by the black dots, with error bars. The error bars are statistical. The three main processes are $t\bar{t}Z$ signal, $t\bar{t}$ background and Z background. This Histogram Analyser also includes minor backgrounds, labelled as 'Other' in red. Minor backgrounds are required for data to match the total simulation. 'Other' includes single top production, WZ and ZZ diboson production and $t\bar{t}W$. Each process is represented by a different colour in the Histogram Analyser.

The Histogram Analyser displays nine histograms, shown in Figure 7.2.1 and described in the following.

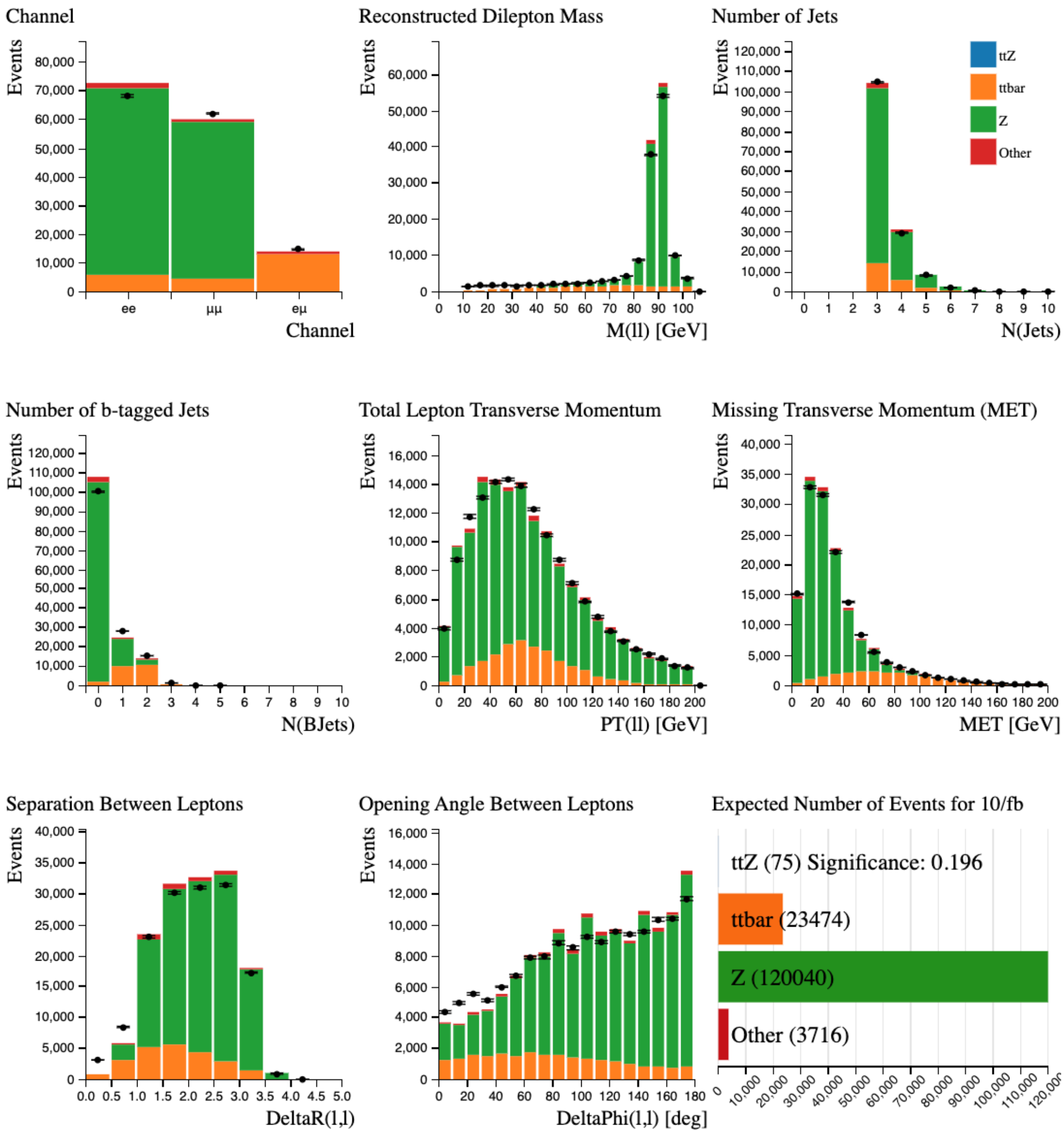


Figure 7.2.1: $t\bar{t}Z$ Histogram Analyser before any selections are applied. The 9 histograms are (top left) Channel, (top middle) Reconstructed Dilepton Mass, (top right) Number of Jets, (centre left) Number of b-tagged Jets, (centre middle) Total Lepton Transverse Momentum, (centre right) Missing Transverse Momentum, (bottom left) Separation Between Leptons, (bottom middle) Opening Angle Between Leptons, (bottom right) Expected Number of Events.

7.2.3 Expected Number of Events for 10 fb^{-1}

This histogram shows the number of events expected to be detected, reconstructed and recorded by ATLAS for 10 inverse femtobarn (10 fb^{-1}) of data, before any additional selections are made on the Histogram Analyser.

The expected number of real data events reconstructed and recorded by ATLAS is different to the number of events produced by real collisions. Some events will not be reconstructed due to the way the detector is constructed, the resolution of the sub-detectors, reconstruction efficiency and other inefficiencies.

Table 7.2.1 shows the cross-sections used by ATLAS Open Data [198], along with the expected number of events before applying additional cuts with the Histogram Analyser. With no cuts, we have 75 $t\bar{t}Z$ events, with many more background events. The majority of the background at this point is Z boson production, which can change depending on the cuts applied.

Process	Cross-section (pb)	Expected # of events
$t\bar{t}Z$	0.08258096	75
$t\bar{t}$	452.693559	23474
Z	3901.1616	120040

Table 7.2.1: Cross-sections used for the different processes of the $t\bar{t}Z$ Histogram Analyser [198], along with the expected number of events before any additional cuts are applied in the Histogram Analyser.

The **significance** of $t\bar{t}Z$ quantifies how "significant" the $t\bar{t}Z$ simulation sample is with respect to the background. It is calculated by the simplified equation:

$$\frac{\text{Number of } t\bar{t}Z \text{ events}}{\sqrt{\text{Number of background events}}}. \tag{7.2.1}$$

A larger significance value indicates better extraction of the $t\bar{t}Z$ signal amongst the backgrounds.

7.2.4 Preselections

Some pre-selections were applied to reduce the size of the datasets used as inputs to the $t\bar{t}Z$ Histogram Analyser so that the website can run quicker. These pre-selections include:

- exactly 2 leptons are required;
- decays to taus or hadrons are removed;
- events with <3 jets are removed;

7.2.5 The Histograms

Channel

The leptonic decay channels are shown in this first histogram in the top left: dielectron ee, dimuon $\mu\mu$ and electron-muon $e\mu$.

Reconstructed Dilepton Mass, M(ll)

The “Reconstructed Dilepton Mass" histogram displays the mass reconstructed from the two leptons in the final state. For $t\bar{t}Z$ 2ℓOS signal and Z background, these would originate from a Z boson. With no cuts, this peaks at 90 GeV, due to the huge Z boson contribution.

Number of Jets, NJets

The “Number of Jets" histogram displays the number of jets found in the event.

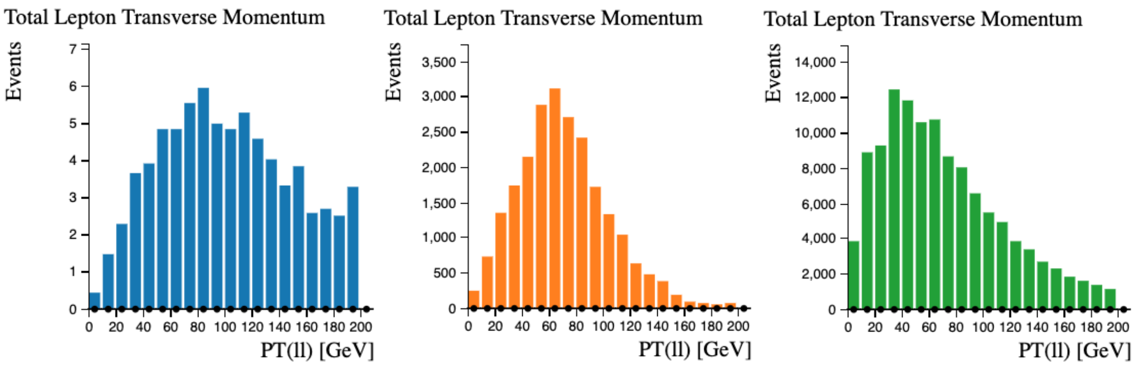
Number of b-tagged Jets, N(BJets)

Jets originating from b-quarks are identified and labelled, or **tagged**, using so-called b-tagging algorithms. b-tagged jets are expected in top quark decays, but not in leptonic W or Z boson decays.

Total Lepton Transverse Momentum, PT(l,l)

Total Lepton Transverse Momentum is the vectorial sum of the transverse momenta of the observed charged leptons.

For Z boson events, total lepton transverse momentum peaks at low values since the transverse momenta of both leptons mostly cancel each other. For the other processes this cancellation is not as pronounced, their distributions peak at between 60 and 90 GeV. This is illustrated in Figure 7.2.2.



(a) $t\bar{t}Z$ (b) $t\bar{t}$ (c) Z

Figure 7.2.2: Total Lepton Transverse Momentum (PT(ll) [GeV]) distributions for (a) $t\bar{t}Z$, (b) $t\bar{t}$, (c) Z.

Missing Transverse Momentum, MET

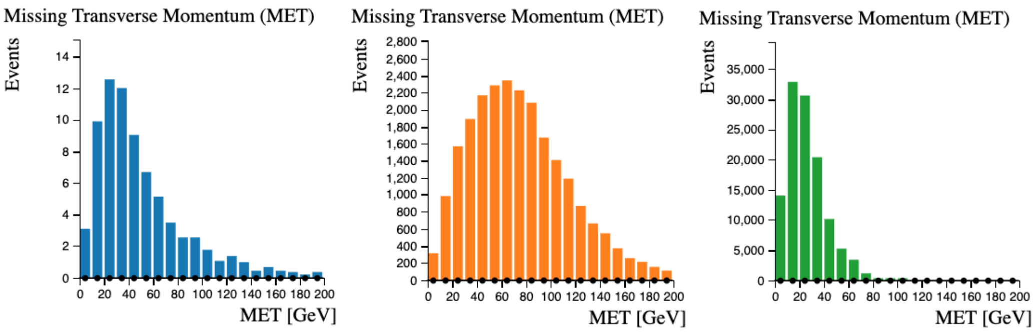
In the LHC, the initial energy of the colliding partons (quarks or gluons) along the beam axis is not known. This is due to the energy of each proton being shared and constantly exchanged between its constituents.

However, the initial momentum of particles travelling transverse to the beam axis is zero. Therefore, any net momentum in the transverse direction indicates missing transverse momentum.

Missing transverse momentum is used to infer the presence of non-detectable particles such as the neutrino. It is also expected to be a signature of many predicted physics events beyond the Standard Model, for example the lightest supersymmetric particle.

The standard abbreviation for missing transverse momentum is MET, for historical reasons.

$t\bar{t}$ decays to two leptons have two neutrinos in the final state while Z boson decays to charged leptons do not. This is illustrated in Figure 7.2.3 by the fact that the $t\bar{t}$ MET distribution peaks at higher values than the MET distributions of $t\bar{t}Z$ and Z.



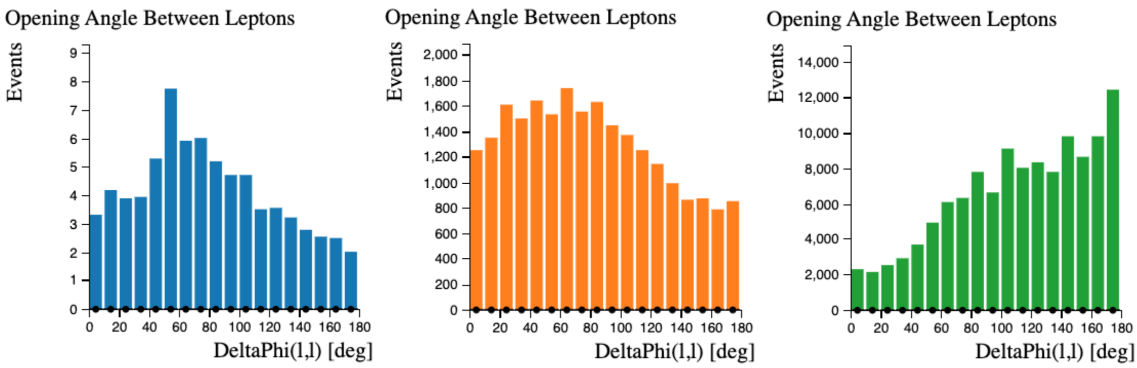
(a) $t\bar{t}Z$ (b) $t\bar{t}$ (c) Z

Figure 7.2.3: Missing Transverse Momentum (MET [GeV]) distributions for (a) $t\bar{t}Z$, (b) $t\bar{t}$, (c) Z.

Opening Angle Between Leptons, DeltaPhi(l,l)

This is the opening angle, measured in phi ϕ, between the two leptons. The azimuthal angle ϕ is measured from the x-axis, around the beam.

If the leptons are emitted back-to-back, this is displayed on the histogram as 180°. Z events show a peak at high values in contrast to all other processes, as shown in Figure 7.2.4. The reason Z events peak at higher values than other processes is because the leptons from the Z decay are emitted close to back-to-back.



(a) $t\bar{t}Z$ (b) $t\bar{t}$ (c) Z

Figure 7.2.4: DeltaPhi(l,l) distributions for (a) $t\bar{t}Z$, (b) $t\bar{t}$, (c) Z.

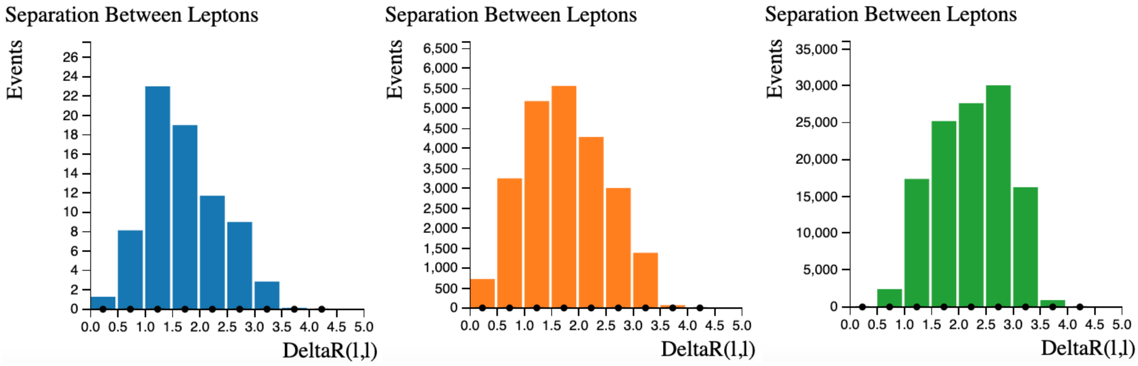
Separation Between Leptons, DeltaR(l,l)

Separation, (ΔR), is calculated using the following equation:

$$(\Delta R)^2 = (\Delta\phi)^2 + (\Delta\eta)^2, \tag{7.2.2}$$

where ϕ is the azimuthal angle between leptons and η is the pseudorapidity.

Figure 7.2.5 shows that $t\bar{t}Z$ events show a peak between 1.0 and 1.5, which is lower values than other processes, with $t\bar{t}$ peaking between 1.5 and 2.0, and Z peaking between 2.5 and 3.0.



(a) $t\bar{t}Z$ (b) $t\bar{t}$ (c) Z

Figure 7.2.5: DeltaR(l,l) distributions for (a) $t\bar{t}Z$, (b) $t\bar{t}$, (c) Z.

7.2.6 Selections for 2ℓ-Z-2b6j

Some of the variables presented in the histograms of the $t\bar{t}Z$ Histogram Analyser are shown pictorially in Figure 7.2.6.

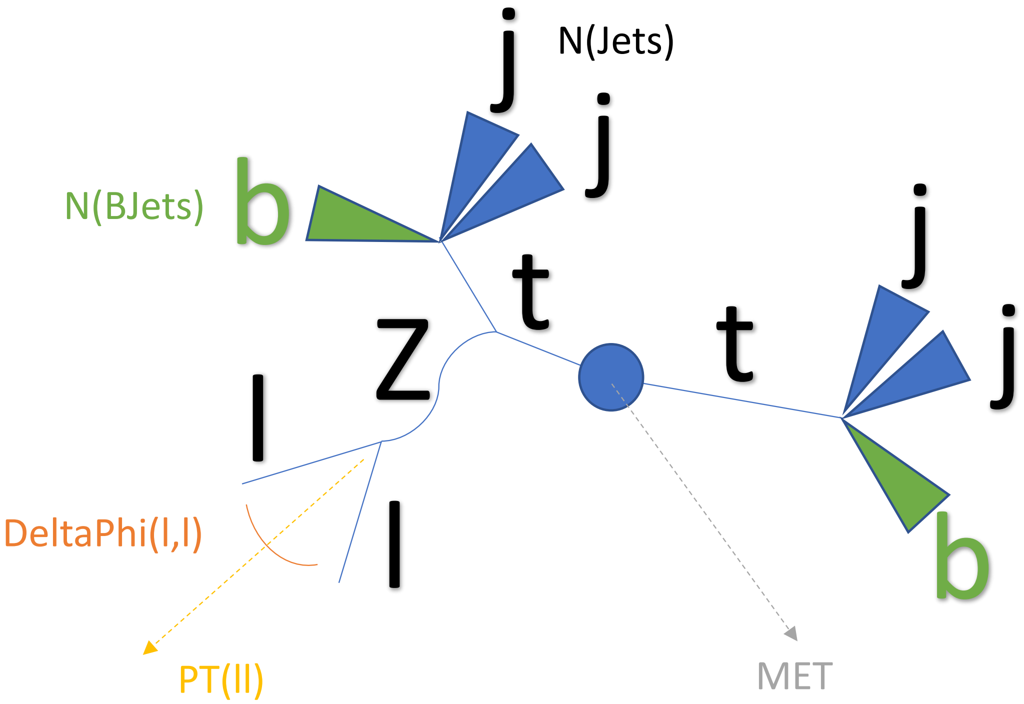


Figure 7.2.6: Schematic diagram of a $t\bar{t}Z$ decay, with some of the variables presented in the histograms of the $t\bar{t}Z$ Histogram Analyser labelled. Antiparticles are not labelled because the Z boson could be radiated from either the top or antitop.

The selections needed to define the 2ℓ-Z-2b6j region in the $t\bar{t}Z$ Histogram Analyser are:

- only the ee and $\mu\mu$ **Channels**;
- **Reconstructed Dilepton Mass** between 80 and 100 GeV;
- **Number of Jets** at least 6;
- **Number of b-tagged Jets** at least 2.

All requirements imposed so far are requirements for the 2ℓ-Z-2b6j signal region (see Table 5.2.2). The remaining variables are not used in the definitions of the final signal regions of the main analysis for this thesis (Section 6), but are used in the Multi-Variate Analysis (MVA) to described in Section 6. Therefore, exploring these variables in the Histogram Analyser can give some intuition as to what the MVA is doing to form signal-rich regions - a key learning objective of the Histogram Analyser.

These further selections are found to be optimal for increasing significance in the $t\bar{t}Z$ Histogram Analyser 2ℓ-Z-2b6j region:

- **PT(ll)** $>$ 30 GeV;
- **MET** $<$ 80 GeV;
- **DeltaPhi(l,l)** $< 140^0$;
- **Separation** $<$ 3.

The selections for the $t\bar{t}Z$ 2ℓOS channel 2ℓ-Z-2b6j region are shown in Table 7.2.2, along with the background they most help reduce. Significance achieved after making each selection sequentially is also shown in Table 7.2.2.

Variable	Selection	To reduce	Significance afterwards
Channel	e^+e^- or $\mu^+\mu^-$	$t\bar{t}$	0.197
M(ll)	80 < M(ll) < 100 GeV	$t\bar{t}$	0.179
N(Jets)	≥6	Z	0.522
N(BJets)	≥2	Z	0.885
PT(ll)	>30 GeV	Z	0.896
MET	<80 GeV	$t\bar{t}$	0.944
DeltaPhi(l,l)	$<140^0$	Z	0.968
DeltaR(l,l)	<3	Z	0.971

Table 7.2.2: Selections for the $t\bar{t}Z$ 2ℓOS Histogram Analyser 2ℓ-Z-2b6j region, along with the background process that each selection most helps reduce, and the significance achieved after making each selection. Significance quoted is by applying these selections in order.

After each selection, both the data points and the simulated Monte Carlo distributions change. The data and simulated Monte Carlo are not exactly the same, but the general agreement is very good. This shows that these processes are well understood and well modelled.

These selections are shown in Figure 7.2.7, increasing significance to 0.971.

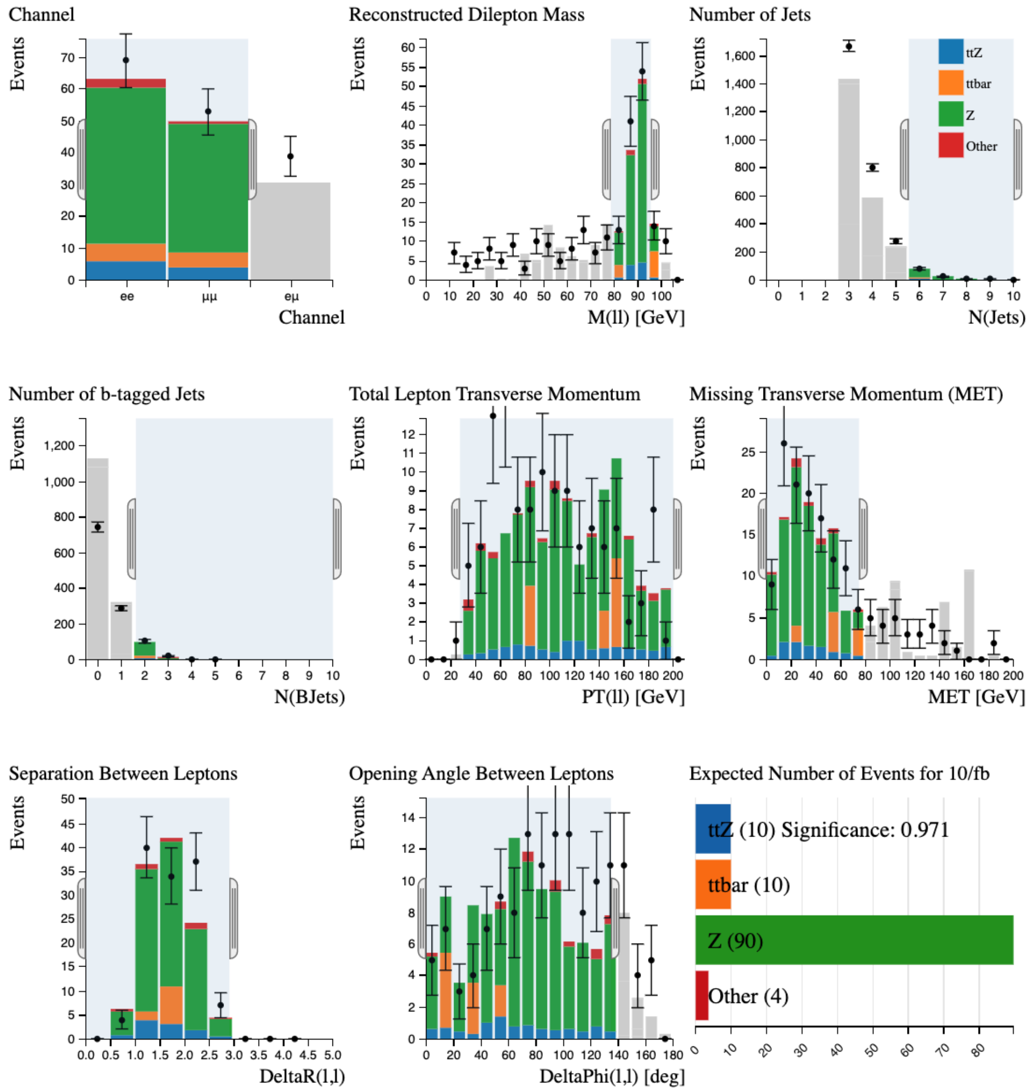


Figure 7.2.7: $t\bar{t}Z$ Histogram Analyser after applying selections for the $t\bar{t}Z$ 2ℓOS 2ℓ-Z-2b6j region. A significance of 0.971 is achieved.

No further changes in selection for any histogram increases the significance over 0.971. This indicates that the selections on Channel, M(ll), N(Jets) and N(BJets) are optimal in terms of signal region definition for 2ℓ-Z-2b6j, as is the case for $t\bar{t}Z$ 2ℓOS papers published by ATLAS [36]. The fact that the maximum significance achievable from defining a looser signal region of N(Jets)≥5 and N(BJets)≥1 indicates that the approach of defining separate signal regions can achieve higher significance than a looser signal region, e.g. with at least 5 jets rather than at least 6 jets. The significances of the separate signal regions can then be combined together to achieve a greater significance for $t\bar{t}Z$ 2ℓOS.

7.2.7 Selections for 2ℓ-Z-2b5j

To achieve a greater significance for $t\bar{t}Z$ 2ℓOS by combining signal regions, the same process can be applied to the 2ℓ-Z-2b5j signal region of Table 5.2.2 to find a significance of 0.380, shown in Figure 7.2.8. The selections for the $t\bar{t}Z$ 2ℓOS channel 2ℓ-Z-2b5j region are shown in Table 7.2.3, along with the background they most help reduce. Significance achieved after making each selection sequentially is also shown in Table 7.2.3.

Variable	Selection	To reduce	Significance afterwards
Channel	e^+e^- or $\mu^+\mu^-$	$t\bar{t}$	0.197
M(ll)	80 < M(ll) < 100 GeV	$t\bar{t}$	0.179
N(Jets)	==5	Z	0.212
N(BJets)	≥2	Z	0.329
PT(ll)	>100 GeV	Z	0.350
MET	<130 GeV	$t\bar{t}$	0.360
DeltaPhi(l,l)	$<90^0$	Z	0.380

Table 7.2.3: Selections for the $t\bar{t}Z$ 2ℓOS Histogram Analyser 2ℓ-Z-2b5j region, along with the background process that each selection most helps reduce, and the significance achieved after making each selection. Significance quoted is by applying these selections in order.

7.2.8 Selections for 2ℓ-Z-1b6j

The same process can be applied to the 2ℓ-Z-1b6j signal region of Table 5.2.2 to find a maximum significance of 0.488, shown in Figure 7.2.9. The selections for the $t\bar{t}Z$ 2ℓOS channel 2ℓ-Z-1b6j region are shown in Table 7.2.4, along with the background they most help reduce. Significance achieved after making each selection sequentially is also shown in Table 7.2.4.

Variable	Selection	To reduce	Significance afterwards
Channel	e^+e^- or $\mu^+\mu^-$	$t\bar{t}$	0.197
M(ll)	80 < M(ll) < 100 GeV	$t\bar{t}$	0.179
N(Jets)	≥6	Z	0.522
N(BJets)	==1	Z	0.472
PT(ll)	>20 GeV	Z	0.483
DeltaR(l,l)	<3	Z	0.488

Table 7.2.4: Selections for the $t\bar{t}Z$ 2ℓOS Histogram Analyser 2ℓ-Z-1b6j region, along with the background process that each selection most helps reduce, and the significance achieved after making each selection. Significance quoted is by applying these selections in order.

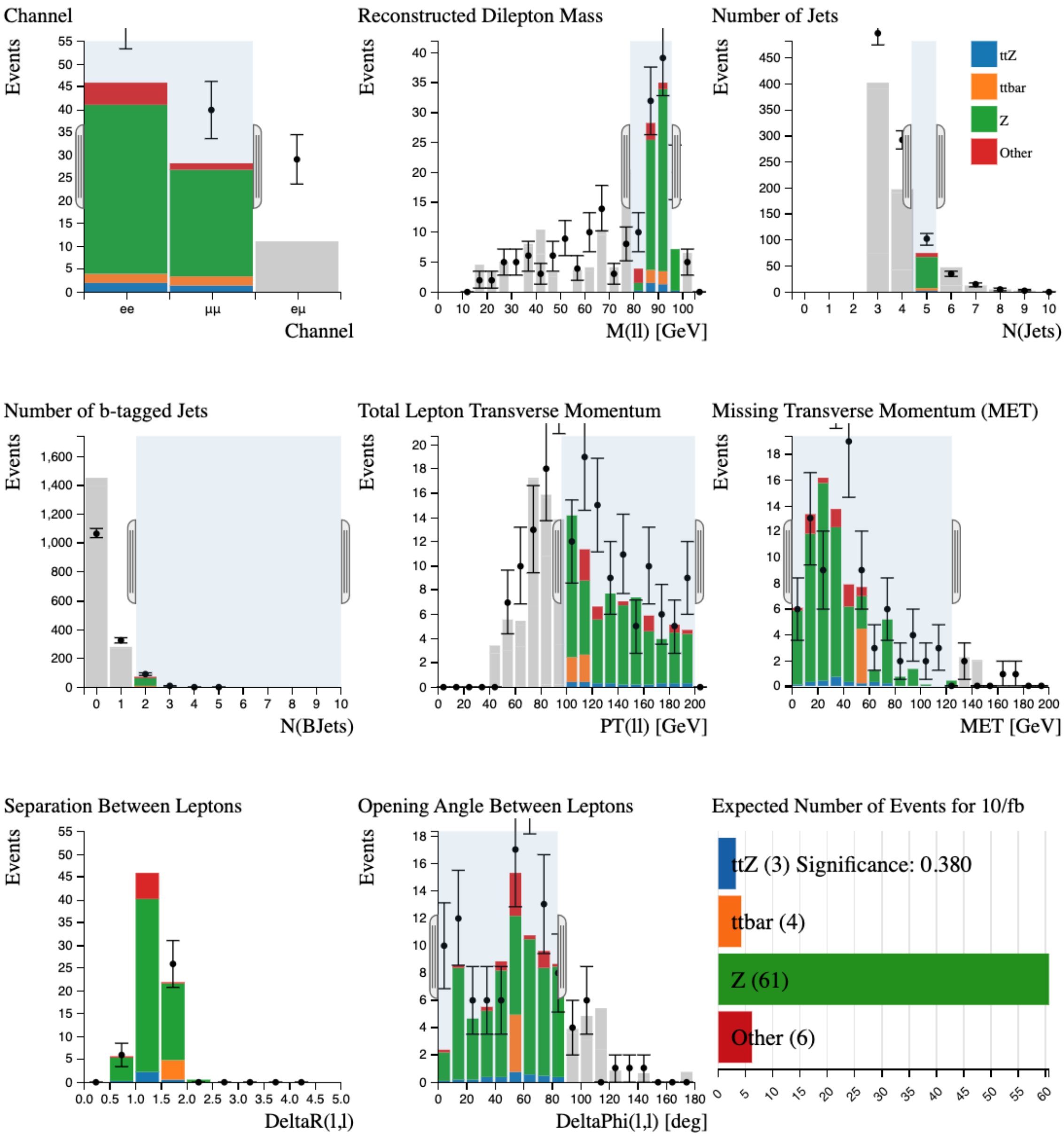


Figure 7.2.8: $t\bar{t}Z$ Histogram Analyser after applying selections for the 2ℓ-Z-2b5j signal region and optimising each variable. A significance of 0.380 is achieved.

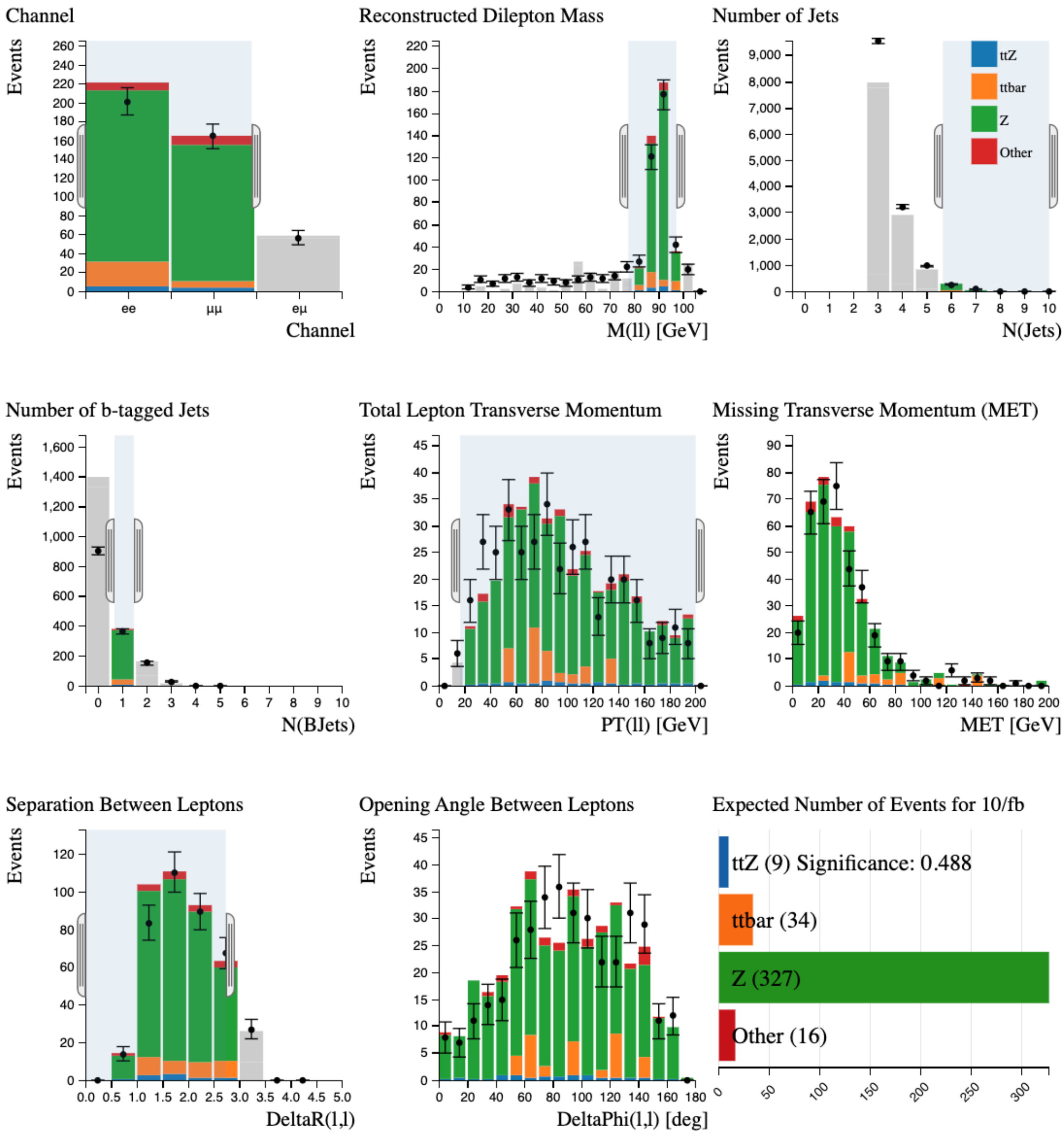


Figure 7.2.9: $t\bar{t}Z$ Histogram Analyser after applying selections for the 2ℓ-Z-1b6j signal region and optimising each variable. A significance of 0.488 is achieved.

7.2.9 Conclusion

This study indicates that an MVA will likely select:

- high PT(ll);
- low MET;
- low DeltaPhi(l,l);
- low DeltaR(l,l).

when building a signal-enriched region. No precise values can be given here because an MVA will optimise differently to the by-hand optimisation done in the Histogram Analyser. The fact that optimum selections for PT(ll), MET, DeltaR(l,l) and DeltaPhi(l,l) are different in the 3 regions illustrates why MVA training is conducted separately in different regions - because different regions will yield different optimum selections.

7.3 Jupyter notebooks

Jupyter notebooks [199] are a key online resource to introduce programming and coding, providing a very suitable arena for using ATLAS data for education. Several notebooks based on the $t\bar{t}Z$ 2ℓOS process were developed, as discussed during this section. They are presented here in sequential order of increasing difficulty.

7.3.1 Introduction

The release of the 13 TeV ATLAS Open Data was accompanied by a set of Jupyter notebooks that allow data analysis to be performed directly in a web browser [196, 200, 201]. Several notebooks with analysis examples are available, including analyses of $t\bar{t}Z$. The aim of many of these notebooks is to recreate published ATLAS results.

7.3.2 Analysis from csv

csv files are commonplace in data science outside of particle physics, therefore an analysis from csv files using ATLAS data is an opportunity to teach the transferrable skill of analysing csv files. As such, an example analysis starting from csv files and reproducing aspects of an ATLAS published result [36] is presented here.

Introduction

The csv analysis notebook [202] uses ATLAS Open Data to show the steps to implement Machine Learning in the $t\bar{t}Z$ 2ℓOS analysis, using the same input csv file as was used for the Histogram Analyser of Section 7.2. The steps taken throughout the notebook to recreate aspects of the ATLAS published result are:

1. tabulating the input data;
2. checking signal and background distributions for the variables present in the dataset;
3. checking separation between signal and background for the variables present in the dataset;

4. checking correlations between the variables present in the dataset;
5. training a MVA;
6. checking for overtraining of the MVA;
7. evaluating the performance of the MVA.

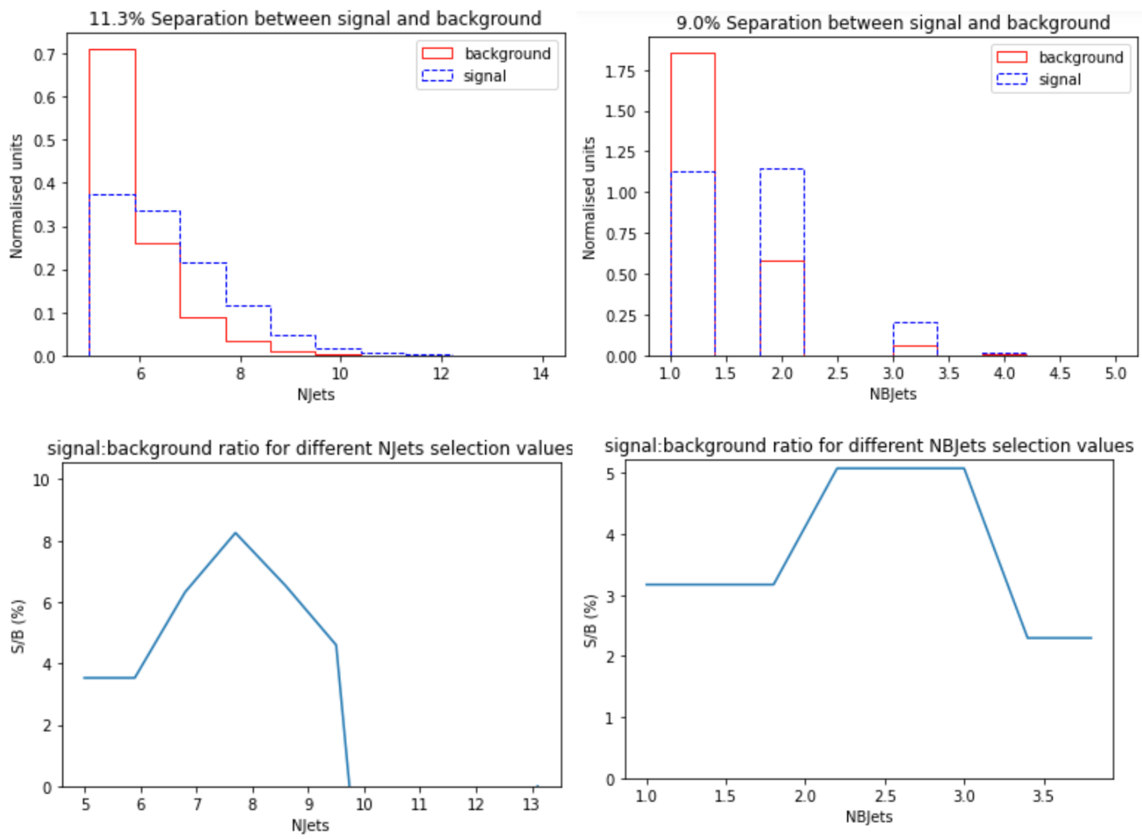
Selections

The fact that no $t\bar{t}Z$ 2ℓOS signal is visible immediately means that some selections have to be made. These selections are given in Table 7.3.1.

Reason	Code
e^+e^- or $\mu^+\mu^-$	Channel!=2
Number of jets	NJets $\geq$ 5
Number of b-jets	N(BJets) $\geq$ 1
Close to Z mass	|Mll - 91.12| < 10 GeV

Table 7.3.1: Initial selections applied to the input data in the Jupyter notebook introducing ML using $t\bar{t}Z$ 2ℓOS csv data.

After the selections of Table 7.3.1, a useful next step is to see how well signal and background are separated for each variable, and how high a signal-to-background ratio this can achieve. Such graphs are shown in Figure 7.3.1. Only 2 from 7 of the input variables are shown, for brevity.



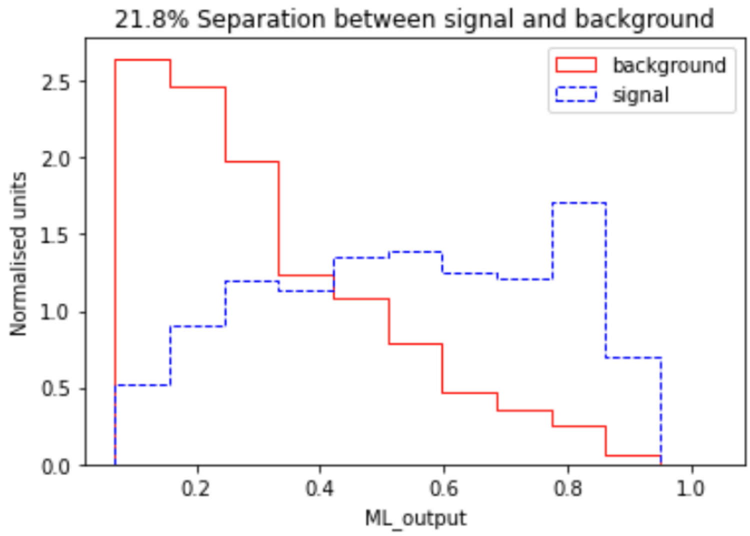
(a) Number of jets

(b) Number of b-jets

Figure 7.3.1: Separation between signal and background and signal-to-background ratio obtained by selecting above a particular value of the x-variable in question. Taking (a) NJets as an example, the starting x-value is 5. Taking the ratio of number of signal events with at least 5 jets, to the number of background events with at least 5 jets gives the S/B value at NJets=5 on the signal:background ratio plot (about 3.5%). Now imagine selecting only events with at least 7 jets. Taking the ratio of those events passing that selection gives the S/B value at NJets=7 on the signal:background ratio plot (about 6%). That is how the signal:background ratio plots are constructed.

Introducing Machine Learning

ML is introduced as a way to construct a variable that can achieve higher separation between signal and background and signal-to-background ratios. To achieve highest separation, ideally all variables would be used in the ML technique. However, for example, Mll cannot be used since values around the Z mass were selected, therefore using this sculpted distribution would lead to overtraining. To be sure all the other variables can be used, the correlations between them need to be checked. If a pair of variables is fully correlated (=1.0), using both would not add any new info. Having said this, some correlation is crucial, because this is what the ML technique exploits. No variable pair is correlated > 0.75 (absolute value), therefore each variable can be used. With a correlation check complete, the separation and signal-to-background ratio achievable using the 'ML_output' variable can be seen in Figure 7.3.2.



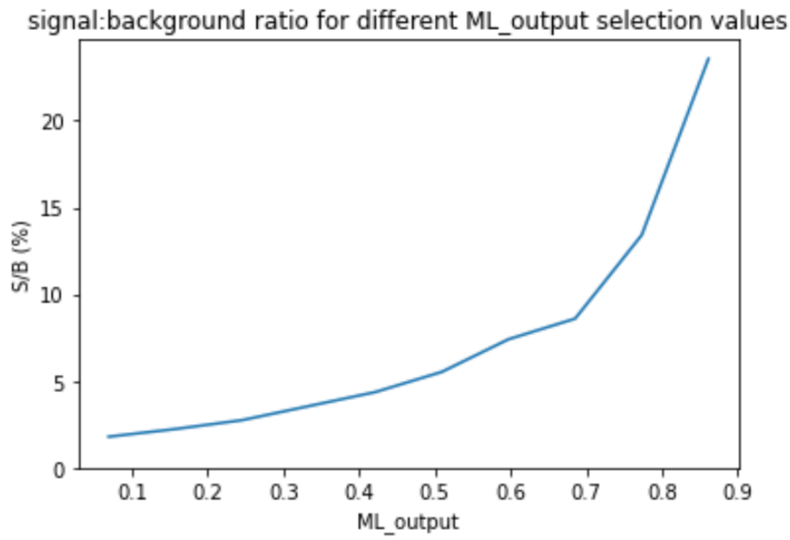


Figure 7.3.2: Separation between signal and background and signal-to-background ratio obtained by selecting above a particular value of 'ML_output'. The starting x-value is about 0.05. Taking the ratio of number of signal events with ML_output > 0.05, to the number of background events with ML_output > 0.05 gives the S/B value at ML_output = 0.05 on the signal:background ratio plot (about 2%). Now imagine selecting only events with ML_output > 0.6. Taking the ratio of those events passing that selection gives the S/B value at ML_output=0.6 on the signal:background ratio plot (about 8%). That is how the signal:background ratio plots are constructed.

ML_output compared to individual variables

The separation and S/B shown in Figure 7.3.2 is better than any of the individual variables of Figure 7.3.1 could ever have achieved. Recalling that $t\bar{t}Z$ 2ℓOS signal nominally produces at least 6 jets, including at least 2 b-jets, allows a further selection to be made, in an attempt to uncover some significant $t\bar{t}Z$ 2ℓOS signal.

Conclusion to the csv exploration notebook

After applying further selections, a significant amount of $t\bar{t}Z$ 2ℓOS signal can be seen above 0.8 in the ML_output distribution . Selecting ML_output > 0.8 would mostly eliminate background and achieve S/B 15%, as can be seen from Figure 7.3.2.

This technique of isolating signal at high ML_output allows to make precise measurements of the $t\bar{t}Z$ 2ℓOS signal process. In summary, this notebook introducing ML using $t\bar{t}Z$ shows that:

- putting data into an ML technique means only one variable has to be optimised;
- signal and background distributions are separated more when looking at ML output;
- ML achieves higher S/B than individual variables, because it finds multi-dimension correlations that give better S/B classification.

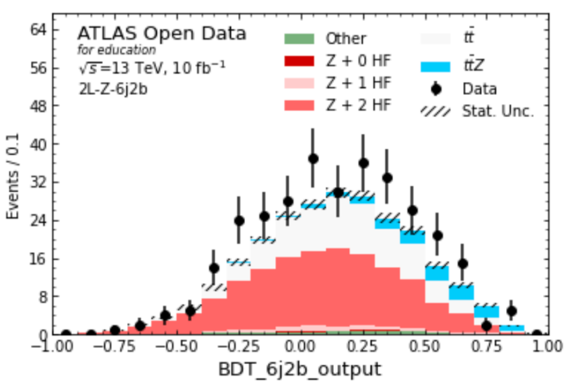
7.3.3 Full analysis

Having shown a simplified $t\bar{t}Z$ 2ℓOS analysis from csv files, similar principles can be extended to an analysis that fully reproduces a published ATLAS result [36]. The added complexity compared to the notebook of Section 7.3.2 includes:

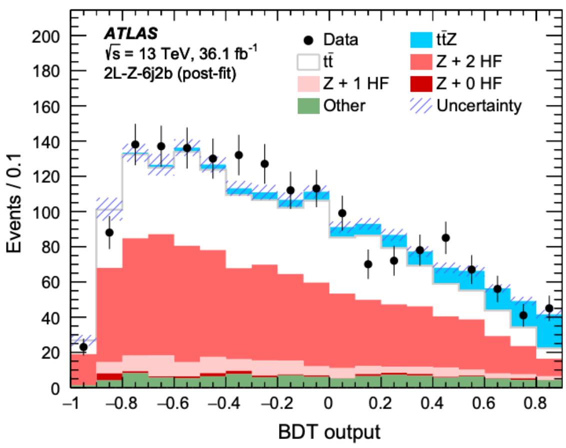
- separating the analysis into 3 different signal regions;
- defining control regions;
- creating data-driven background estimates;
- ranking MVA input variables.

Introduction

The notebook presenting a full $t\bar{t}Z$ 2ℓOS analysis [203] uses ATLAS Open Data to show the steps to implement Machine Learning in the $t\bar{t}Z$ 2ℓOS analysis, following the ATLAS published paper "Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector" [36]. In particular, this notebook aims to recreate plots from Ref. [36] using a simplified ML workflow. The first plot that can be recreated is shown in Figure 7.3.3. Similar plots to Figure 7.3.3 are recreated for the 2ℓ-Z-2b5j and 2ℓ-Z-1b6j regions.



(a) Open Data

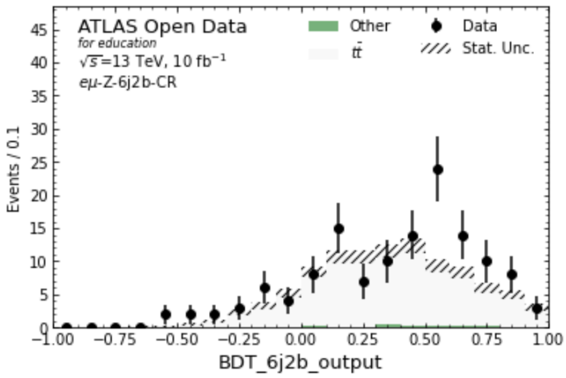


(b) Ref. [36]

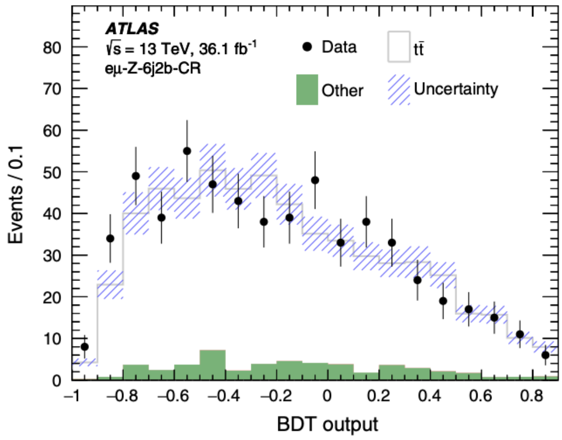
Figure 7.3.3: BDT output distributions in the signal region 2ℓ-Z-2b6j (here called 6j2b) using (a) ATLAS Open Data, (b) Ref. [36]. Considering the differences in the amount of data and the fact that not every detail from an ATLAS paper can be followed, the Open Data can reproduce this ATLAS result well. The 'Other' background contains SM processes with small cross sections producing two opposite-sign prompt leptons. The shaded band represents the total uncertainty. The last bin of each distribution contains the overflow.

Control regions

Plots in control regions can also be recreated, shown in Figure 7.3.4 for 2ℓ-Z-2b6j as an example.Equivalent plots for the 2ℓ-Z-2b5j and 2ℓ-Z-1b6j are also recreated.



(a) Open Data



(b) Ref. [36]

Figure 7.3.4: BDT output distributions in the $t\bar{t}$ control region of 2ℓ-Z-2b6j (here called 6j2b) using (a) ATLAS Open Data, (b) Ref. [36]. Considering the differences in the amount of data and the fact that not every detail from an ATLAS paper can be followed, the Open Data can reproduce this ATLAS result well. The 'Other' background contains SM processes with small cross sections producing two opposite-sign prompt leptons, including the $t\bar{t}Z$ process, whose contribution is negligible. The shaded band represents the total uncertainty. The last bin of each distribution contains the overflow.

Data-driven $t\bar{t}$ estimates

The $t\bar{t}$ control regions exampled in Figure 7.3.4 can then be used to build data-driven estimates of the $t\bar{t}$ contribution, rather than using the MC estimates in subfigure (a) of Figure 7.3.3.

Ranking input variables

Another result from Ref. [36] that can be recreated is Table 11, showing the definitions and ranking of input variables for the BDT. This comparison is shown in Figure 7.3.5.

Definition	6j1b	5j2b	6j2b
p_T of the lepton pair	15	14	15
p_T of the 4th jet	5	1	8
p_T of the 5th jet	-	8	-
p_T of the 6th jet	2	-	2
ΔR_η between the two leptons	6	4	7
Number of jet pairs with mass within a window of 30 GeV around 85 GeV	1	2	3
Number of top-quark candidates	-	-	1
Invariant mass of the two jets with the smallest ΔR_η	13	10	17
Invariant mass of the two untagged jets with the highest p_T	9	11	-
Invariant mass of the two jets with the highest value of the b-tagging discriminant	-	5	4
Scalar sum of p_T divided by the sum of energy of all jets	14	13	16
Average ΔR_η of all jet pairs	11	3	10
Maximum invariant mass of a lepton and the b-tagged jet with the smallest ΔR_η	10	-	13
First Fox–Wolfram moment built from jets and leptons	12	12	14
Sum of jet p_T, using up to six jets	4	6	5
η of dilepton system	3	9	9
Sum of the two closest two-jet invariant masses from jjj1 and jjj2 divided by two	7	-	11
ΔR_η between two jets with the highest value of the b-tagging discriminant in the event	-	7	6
p_T of the b-tagged jet with the highest p_T	8	-	12

(a) Open Data

Definition	6j1b	5j2b	6j2b
p_T of the lepton pair	8	11	8
p_T of the 4th jet	6	12	6
p_T of the 5th jet	-	14	-
p_T of the 6th jet	9	-	11
ΔR_η between the two leptons	7	8	12
Number of jet pairs with mass within a window of 30 GeV around 85 GeV	4	6	4
Number of top-quark candidates	-	-	17
Invariant mass of the two jets with the smallest ΔR_η	13	7	14
Invariant mass of the two untagged jets with the highest p_T	15	13	-
Invariant mass of the two jets with the highest value of the b-tagging discriminant	-	10	9
Scalar sum of p_T divided by the sum of energy of all jets	2	1	2
Average ΔR_η of all jet pairs	5	4	5
Maximum invariant mass of a lepton and the b-tagged jet with the smallest ΔR_η	14	-	13
First Fox–Wolfram moment built from jets and leptons	3	2	1
Sum of jet p_T, using up to six jets	12	5	10
η of dilepton system	1	3	3
Sum of the two closest two-jet invariant masses from jjj1 and jjj2 divided by two	10	-	15
ΔR_η between two jets with the highest value of the b-tagging discriminant in the event	-	9	7
p_T of the b-tagged jet with the highest p_T	11	-	16

(b) Ref. [36]

Figure 7.3.5: The definitions and ranking of input variables for the BDT in the $t\bar{t}Z$ 2ℓOS analysis. (a) ATLAS Open Data, (b) Ref. [36]. Some similarities can be seen between (a) and (b), for example "Number of jet pairs with mass within a window of 30 GeV around 85 GeV" ranking rather highly for both. Differences between (a) and (b) can also be seen, for example "Scalar sum of p_T divided by the sum of energy of all jets" ranking highly for (b) but not so highly for (a). Jets and leptons are ordered in descending order of p_T. Only the first eight jets are considered when calculating the input variables.

Conclusion to the full analysis notebook

Using ATLAS Open Data, a full analysis of the $t\bar{t}Z$ process can be undertaken, reproducing simplified versions of the results from an ATLAS published paper [36]. Signal and control region plots can be reproduced in the same format as the ATLAS published paper [36]. The method of obtaining data-driven $t\bar{t}$ estimates used in the ATLAS published paper [36] can also be reproduced using ATLAS Open Data. The ranking of most important variables in the MVA with ATLAS Open Data in the $t\bar{t}Z$ 2ℓOS channel show similarities to the ranking of the most important variables in the MVA from the ATLAS published paper [36].

7.4 Comparisons with full ATLAS data

This section compares results from Section 7.2 and Section 7.3.3 using 10 fb^{-1}of ATLAS Open Data in simplified analyses to Section 6 using 139.0 fb^{-1}of full Run 2 ATLAS data in a full analysis. Results will be compared in terms of:

- ranking of variables by the MVAs;
- statistical significance achievable.

7.4.1 Comparison of variable ranking between Open Data and binary BDTs

Table 6.1.2 ranking input variables using BDTs with Full Run 2 data can be compared side-by-side with the information from Figure 7.3.5 ranking input variables using BDTs with ATLAS Open Data. This comparison is shown in Table 7.4.1. A number of similarities can be seen, e.g. $N_{jj}^{V mass}$ is ranked within the top 4 in each of the six BDTs, or that $\mathrm{p_T^{ll}}$ is ranked within the bottom 3 in each of the six BDTs. However, differences can be seen also, perhaps the most stark being that $N_{bjj}^{top-mass}$

is ranked 1st in the 2b6j Open Data BDT yet 16th in the 2b6j Full Run 2 BDT. This suggests that some variables are important over a range of amount of data available, whereas other variables only become more important when more data are available.

	1b6j		2b5j		2b6j	
rank	Open Data	Full Run 2	Open Data	Full Run 2	Open Data	Full Run 2
1	$N_{jj}^{V mass}$	$\mathrm{H_T^{6jets}}$	$\mathrm{p_T^{4jet}}$	$\mathrm{H_T^{6jets}}$	$N_{bjj}^{top-mass}$	$\mathrm{H_T^{6jets}}$
2	$\mathrm{p_T^{6jet}}$	η_{ll}	$N_{jj}^{V mass}$	ΔR_{jj}^{ave}	$\mathrm{p_T^{6jet}}$	ΔR_{ll}
3	η_{ll}	$N_{jj}^{V mass}$	ΔR_{jj}^{ave}	$N_{jj}^{V mass}$	$N_{jj}^{V mass}$	η_{ll}
4	$\mathrm{H_T^{6jets}}$	$\mathrm{p_T^{b1}}$	ΔR_{ll}	$\mathrm{M_{bb}^{pTord}}$	$\mathrm{M_{bb}^{pTord}}$	$N_{jj}^{V mass}$
5	$\mathrm{p_T^{4jet}}$	$\mathrm{MaxM_{lepb}^{mindR}}$	$\mathrm{M_{bb}^{pTord}}$	$\mathrm{M_{jj}^{mindR}}$	$\mathrm{H_T^{6jets}}$	$\Delta R_{\mathrm{ave}}^{\mathrm{jj}}$
6	ΔR_{ll}	$\mathrm{M_{jj}^{mindR}}$	$\mathrm{H_T^{6jets}}$	ΔR_{ll}	ΔR_{bb}	ΔR_{bb}
7	$\mathrm{M_W^{avg}}$	$\mathrm{p_T^{4jet}}$	ΔR_{bb}	ΔR_{bb}	ΔR_{ll}	$\mathrm{p_T^{6jet}}$
8	$\mathrm{p_T^{b1}}$	ΔR_{ll}	$\mathrm{p_T^{5jet}}$	$\mathrm{p_T^{4jet}}$	$\mathrm{p_T^{4jet}}$	$\mathrm{MaxM_{lepb}^{mindR}}$
9	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{p_T^{6jet}}$	η_{ll}	η_{ll}	η_{ll}	$\mathrm{M_W^{avg}}$
10	$\mathrm{MaxM_{lepb}^{mindR}}$	$\mathrm{M_W^{avg}}$	$\mathrm{M_{jj}^{mindR}}$	$\mathrm{p_T^{5jet}}$	ΔR_{jj}^{ave}	$\mathrm{M_{jj}^{mindR}}$
11	ΔR_{jj}^{ave}	$Centr_{jet}$	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{M_W^{avg}}$	$\mathrm{p_T^{4jet}}$
12	H1	H1	H1	H1	$\mathrm{p_T^{b1}}$	$\mathrm{M_{bb}^{pTord}}$
13	$\mathrm{M_{jj}^{mindR}}$	$\mathrm{M_{uu}^{pTord}}$	$Centr_{jet}$	$\mathrm{p_T^{ll}}$	$\mathrm{MaxM_{lepb}^{mindR}}$	$Centr_{jet}$
14	$Centr_{jet}$	ΔR_{jj}^{ave}	$\mathrm{p_T^{ll}}$	$Centr_{jet}$	H1	H1
15	$\mathrm{p_T^{ll}}$	$\mathrm{p_T^{ll}}$			$\mathrm{p_T^{ll}}$	$\mathrm{p_T^{b1}}$
16					$Centr_{jet}$	$N_{bjj}^{top-mass}$
17					$\mathrm{M_{jj}^{mindR}}$	$\mathrm{p_T^{ll}}$

Table 7.4.1: Comparison of ranking of the variables used for BDT training, when using a single BDT per 2ℓOS region. The comparison is performed between the BDTs using ATLAS Open Data and the BDTs using Full Run 2 data.

7.4.2 Comparison of variable ranking between Open Data and binary DNNs

Figure 6.2.1 ranking variables using DNNs with Full Run 2 data can be compared side-by-side with the information from Figure 7.3.5 ranking variables using BDTs with ATLAS Open Data. This comparison is shown in Table 7.4.2. A number of similarities can be seen, e.g. $N_{jj}^{V mass}$ is ranked within the top 3 in each of the six MVAs, or that $\mathrm{p_T^{ll}}$ is ranked within the bottom 3 in each of the six MVAs. However, differences can be seen also, perhaps the most stark being that $\mathrm{p_T^{6jet}}$ and $\mathrm{p_T^{5jet}}$ are ranked much higher in the Open Data BDTs than they are in the Full Run 2 DNNs. This again suggests that some variables are important over a range of amount of data available, whereas other variables only become more important when more data are available.

	1b6j		2b5j		2b6j	
rank	Open Data	Full Run 2	Open Data	Full Run 2	Open Data	Full Run 2
1	N_{jj}^{Vmass}	$\mathrm{H_T^{6jets}}$	$\mathrm{p_T^{4jet}}$	$\mathrm{H_T^{6jets}}$	$N_{bjj}^{top-mass}$	$\mathrm{H_T^{6jets}}$
2	$\mathrm{p_T^{6jet}}$	η_{ll}	N_{jj}^{Vmass}	ΔR_{ll}	$\mathrm{p_T^{6jet}}$	N_{jj}^{Vmass}
3	η_{ll}	N_{jj}^{Vmass}	ΔR_{jj}^{ave}	$Centr_{jet}$	N_{jj}^{Vmass}	ΔR_{ll}
4	$\mathrm{H_T^{6jets}}$	H1	ΔR_{ll}	N_{jj}^{Vmass}	$\mathrm{M_{bb}^{pTord}}$	$N_{bjj}^{top-mass}$
5	$\mathrm{p_T^{4jet}}$	ΔR_{ll}	$\mathrm{M_{bb}^{pTord}}$	$\mathrm{M_{bb}^{pTord}}$	$\mathrm{H_T^{6jets}}$	$Centr_{jet}$
6	ΔR_{ll}	$\mathrm{p_T^{4jet}}$	$\mathrm{H_T^{6jets}}$	H1	ΔR_{bb}	ΔR_{bb}
7	$\mathrm{M_W^{avg}}$	$Centr_{jet}$	ΔR_{bb}	ΔR_{jj}^{ave}	ΔR_{ll}	$\mathrm{p_T^{4jet}}$
8	$\mathrm{p_T^{b1}}$	$\mathrm{p_T^{b1}}$	$\mathrm{p_T^{5jet}}$	$\mathrm{p_T^{4jet}}$	$\mathrm{p_T^{4jet}}$	H1
9	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{p_T^{6jet}}$	η_{ll}	$\mathrm{M_{jj}^{mindR}}$	η_{ll}	$\mathrm{p_T^{6jet}}$
10	$\mathrm{MaxM_{lepb}^{mindR}}$	ΔR_{jj}^{ave}	$\mathrm{M_{jj}^{mindR}}$	ΔR_{bb}	ΔR_{jj}^{ave}	η_{ll}
11	ΔR_{jj}^{ave}	$\mathrm{M_W^{avg}}$	$\mathrm{M_{uu}^{pTord}}$	$\mathrm{p_T^{5jet}}$	$\mathrm{M_W^{avg}}$	$\mathrm{p_T^{b1}}$
12	H1	$\mathrm{MaxM_{lepb}^{mindR}}$	H1	η_{ll}	$\mathrm{p_T^{b1}}$	$\mathrm{M_{bb}^{pTord}}$
13	$\mathrm{M_{jj}^{mindR}}$	$\mathrm{M_{jj}^{mindR}}$	$Centr_{jet}$	$\mathrm{p_T^{ll}}$	$\mathrm{MaxM_{lepb}^{mindR}}$	$\mathrm{M_W^{avg}}$
14	$Centr_{jet}$	$\mathrm{p_T^{ll}}$	$\mathrm{p_T^{ll}}$	$\mathrm{M_{uu}^{pTord}}$	H1	ΔR_{jj}^{ave}
15	$\mathrm{p_T^{ll}}$	$\mathrm{M_{uu}^{pTord}}$			$\mathrm{p_T^{ll}}$	$\mathrm{p_T^{ll}}$
16					$Centr_{jet}$	$\mathrm{M_{jj}^{mindR}}$
17					$\mathrm{M_{jj}^{mindR}}$	$\mathrm{MaxM_{lepb}^{mindR}}$

Table 7.4.2: Comparison of ranking of the variables used for MVA training. The comparison is performed between the BDTs using ATLAS Open Data and the initial DNNs using Full Run 2 data.

7.4.3 Statistical significance comparison between Histogram Analyser and initial multiclass DNN

The statistical significance from Figure 6.2.5 can be compared to the significance achievable from the Histogram Analyser discussed in Section 7.2, whose final significances are shown in Figure 7.2.7, Figure 7.2.8 and Figure 7.2.9 for the 2ℓ-Z-2b6j, 2ℓ-Z-2b5j and 2ℓ-Z-1b6j channels respectively. This comparison is shown in Table 7.4.3. The Histogram Analyser only uses about 1/14th of the data used for the DNNs of Section 6.3 as this is all of the 13 TeV data currently made open by ATLAS. A more direct comparison can be made by scaling the Histogram Analyser significances by the square root of the ratio between the full Run 2 luminosity and the luminosity used in ATLAS Open Data, $\sqrt{139.0/10}$, because statistical significance scales with the square root of number of events. Even the scaled statistical significances achievable by the Histogram Analyser are about 2.5 times less than the statistical significances achievable by the DNNs. This hints at the power of DNNs in optimising for statistical significance in the $t\bar{t}Z$ 2ℓOS analysis, compared to a cut-and-count analysis.

Channel	Histogram Analyser significance	Histogram Analyser significance (scaled)	DNN significance
2b6j	0.971 (Figure 7.2.7)	3.620	10.8
2b5j	0.380 (Figure 7.2.8)	1.417	4.9
1b6j	0.488 (Figure 7.2.9)	1.819	4.5

Table 7.4.3: A comparison of the statistical significance that can be achieved using the DNNs of Section 6.3, with the Histogram Analyser of Section 7.2. It is important to remember that the Histogram Analyser uses about 1/14th of the data used for the DNNs of Section 6.3.

8 Conclusion

A conclusion is the place where you got tired of thinking.

Anon [204]

Run 2 of the Large Hadron Collider provided the highest centre-of-mass-energy collisions ever achieved in a particle physics experiment; a centre-of-mass-energy of 13 TeV. The dataset provided the largest amount of data collected by a particle physics experiment, totalling 139 fb^{-1}at 13 TeV in Run 2. Such a large dataset allowed unprecedented studies of the top quark, such as the simultaneous production of a top quark, antitop quark and Z-boson ($t\bar{t}Z$).

This thesis focuses on the first use of the 2ℓOS channel with full Run 2 data. The two main backgrounds in the 2ℓOS channel ($t\bar{t}$ and Z+jets) are much larger than the signal, and thus this channel is difficult to study.

By performing Multi-Variate Analysis in the 2ℓOS channel of the $t\bar{t}Z$ process, a new methodology was developed with multiclass DNNs to give 13% sensitivity in the $t\bar{t}Z$ 2ℓOS cross-section measurement. Numerous Multi-Variate Analysis techniques were tested and compared. Boosted Decision Trees were first tested to build on the previous ATLAS results using 36 fb^{-1}of Run 2 data [36]. Using the same variables as the initial Boosted Decision Trees, Deep Neural Networks were then built to provide more flexibility in the learning process. The Boosted Decision Trees and Deep Neural Networks built up until that point were binary classifiers. Since there are two main distinct backgrounds in the $t\bar{t}Z$ 2ℓOS channel, a multiclass approach can provide benefits over a binary approach. Having started with multiclass DNNs using the same variables as the initial BDTs, additional discriminating variables were then added. Parameters of the DNN were optimised for learning, whilst avoiding overtraining. Using the final multiclass DNNs, ROC AUCs of 0.817, 0.834 and 0.833 were found for the 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j regions respectively. Compared to the published ATLAS result in Ref. [36], this constitutes an improvement of 19.9%, 24.9% and 21.8% for the 2ℓ-Z-1b6j, 2ℓ-Z-2b5j and 2ℓ-Z-2b6j regions respectively.

With these multiclass DNNs, a precise measurement of the inclusive cross-section of the 2ℓOS channel in the $t\bar{t}Z$ process could be made, with relative uncertainties of $^{+0.13}_{-0.12}$. This is more than twice the precision of the previously published $t\bar{t}Z$ 2ℓOS measurement, where the uncertainties on the 2ℓOS channel were $^{+0.31}_{-0.29}$. In combination with the 3-lepton and 4-lepton channels, the most precise measurement of the $t\bar{t}Z$ process carried out to date could be made. The 2ℓOS inclusive cross-section measurement is systematics limited, and in particular will benefit from improvements in the MC modelling of the $t\bar{t}Z$ process and $Z+b$ process. With improvements in theory systematics like these, the next step would be to train new MVAs, potentially with new techniques.

Using ATLAS Open Data, several educational example analyses of $t\bar{t}Z$ 2ℓOS process have been presented. These analyses demonstrate 1) cut-and-count capabilities in a $t\bar{t}Z$ 2ℓOS analysis, 2) variation of learning rate, 3) an introduction to machine learning, 4) reproducing an ATLAS paper with Open Data. These analyses produce results similar to those of the main $t\bar{t}Z$ 2ℓOS analysis of this thesis (Section 6). However, the limitations of these simplified analyses also demonstrate the power of the MVAs developed for the main analysis of this thesis.

After the successful completion of Run 2 of the LHC and the associated data analysis, CERN's accelerator chain and experiments underwent major maintenance and upgrade, including ATLAS. These upgrades were in preparation for Run 3 of the LHC, in which the ATLAS experiment will be able to make even more precise measurements of rare processes such as $t\bar{t}Z$ thanks to increased statistics and improvements in detector performance. A Run 3 measurement of the $t\bar{t}Z$ process will likely include differential cross-section measurements in the 2ℓOS channel, to add to the existing differential cross-section measurements in the 3ℓ and 4ℓ channels. This will provide even deeper insights into the fundamental particles of the universe and the interactions between them.

Bibliography

[1] *Lakers need to break from their routine*, 2002, URL: https://www.latimes.com/archives/la-xpm-2002-may-09-sp-adande09-story.html (cit. on p. 27).

[2] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** (1961) 579, ISSN: 0029-5582, URL: http://www.sciencedirect.com/science/article/pii/0029558261904692 (cit. on pp. 27, 29, 37).

[3] 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 pp. 27, 29, 37).

[4] J. Goldstone, A. Salam and S. Weinberg, *Broken symmetries*, Physical Review **127** (1962) 965 (cit. on pp. 27, 29, 37).

[5] G. 't Hooft and M. Veltman, *Regularization and renormalization of gauge fields*, Nuclear Physics B **44** (1972) 189, ISSN: 0550-3213, URL: http://www.sciencedirect.com/science/article/pii/0550321372902799 (cit. on p. 27).

[6] L. Evans and P. Bryant, *LHC Machine*, Journal of Instrumentation **3** (2008) S08001, URL: https://doi.org/10.1088/1748-0221/3/08/s08001 (cit. on pp. 27, 44).

[7] T. Aaltonen et al., *Combination of the top-quark mass measurements from the Tevatron collider*, Physical Review D **86** (2012) 092003 (cit. on p. 27).

[8] S. Myers and E. Picasso, *The LEP collider*, Scientific American **263** (1990) 54 (cit. on p. 27).

[9] F. Abe et al., *Observation of Top Quark Production in $\bar{p}p$ Collisions with the Collider Detector at Fermilab*, Phys. Rev. Lett. **74** (14 1995) 2626, URL: https://link.aps.org/doi/10.1103/PhysRevLett.74.2626 (cit. on pp. 27, 35).

[10] S. Abachi et al., *Observation of the Top Quark*, Phys. Rev. Lett. **74** (14 1995) 2632, URL: https://link.aps.org/doi/10.1103/PhysRevLett.74.2632 (cit. on pp. 27, 35).

[11] Particle Data Group, *Review of Particle Physics*, PTEP **8** (2020) 083C01 (cit. on pp. 27, 31, 35–37, 39, 40, 55, 57, 65, 76).

[12] ATLAS Collaboration, *Measurement of the $t\bar{t}$ production cross-section in the lepton+jets channel at $\sqrt{s}$ = 13 TeV with the ATLAS experiment*, (2020), arXiv: 2006.13076 [hep-ex] (cit. on pp. 27, 36).

[13] A. Datta and M. Duraisamy, *Model independent predictions for rare top decays with weak coupling*, Physical Review D **81** (2010) 074008 (cit. on p. 27).

[14] G. Arnison et al., *Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider*, Physics Letters B **126** (1983) 398, ISSN: 0370-2693, URL: http://www.sciencedirect.com/science/article/pii/0370269383901880 (cit. on pp. 27, 37).

[15] P. Bagnaia et al., *Evidence for $Z^0 \rightarrow e^+e^-$ at the CERN $\bar{p}p$ collider*, Physics Letters B **129** (1983) 130, ISSN: 0370-2693, URL: http://www.sciencedirect.com/science/article/pii/037026938390744X (cit. on pp. 27, 37).

[16] A. Kulesza, L. Motyka, D. Schwartländer, T. Stebel and V. Theeuwes, *Associated production of a top quark pair with a heavy electroweak gauge boson at NLO+NNLL accuracy*, Eur. Phys. J. C **79** (2019) 249, arXiv: 1812.08622 [hep-ph] (cit. on pp. 27, 39).

[17] R. S. Chivukula, S. B. Selipsky and E. H. Simmons, *Nonoblique effects in the $Zb\bar{b}$ vertex from ETC dynamics*, Phys. Rev. Lett. **69** (1992) 575, arXiv: hep-ph/9204214 (cit. on p. 27).

[18] R. Chivukula, E. H. Simmons and J. Terning, *A Heavy top quark and the $Zb\bar{b}$ vertex in noncommuting extended technicolor*, Phys. Lett. **B 331** (1994) 383, arXiv: hep-ph/9404209 (cit. on p. 27).

[19] K. Hagiwara and N. Kitazawa, *Extended technicolor contribution to the Zbb vertex*, Phys. Rev. **D 52** (1995) 5374, arXiv: hep-ph/9504332 (cit. on p. 27).

[20] U. Mahanta, *Noncommuting ETC corrections to $Zt\bar{t}$ vertex*, Phys. Rev. **D 55** (1997) 5848, arXiv: hep-ph/9611289 (cit. on p. 27).

[21] U. Mahanta, *Probing noncommuting ETC effects by $e^+e^- \to t\bar{t}$ at NLC*, Phys. Rev. **D 56** (1997) 402 (cit. on p. 27).

[22] M. Perelstein, *Little Higgs models and their phenomenology*, Prog. Part. Nucl. Phys. **58** (2007) 247, arXiv: hep-ph/0512128 (cit. on p. 27).

[23] ATLAS Collaboration, *Search for direct top squark pair production in events with a Higgs or Z boson, and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, JHEP **08** (2017) 006, arXiv: 1706.03986 [hep-ex] (cit. on p. 27).

[24] ATLAS Collaboration, *Search for new phenomena in events with same-charge leptons and b-jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **12** (2018) 039, arXiv: 1807.11883 [hep-ex] (cit. on p. 27).

[25] ATLAS Collaboration, *Search for squarks and gluinos in final states with same-sign leptons and jets using 139 fb^{-1} of data collected with the ATLAS detector*, JHEP **06** (2020) 046, arXiv: 1909.08457 [hep-ex] (cit. on p. 27).

[26] ATLAS Collaboration, *Analysis of events with b-jets and a pair of leptons of the same charge in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **10** (2015) 150, arXiv: 1504.04605 [hep-ex] (cit. on p. 27).

[27] ATLAS Collaboration, *Search for pair production of up-type vector-like quarks and for four-top-quark events in final states with multiple b-jets with the ATLAS detector*, JHEP **07** (2018) 089, arXiv: 1803.09678 [hep-ex] (cit. on p. 27).

[28] ATLAS Collaboration, *Evidence for $t\bar{t}t\bar{t}$ production in the multilepton final state in proton–proton collisions at $\sqrt{s}$=13 TeV with the ATLAS detector*, Eur. Phys. J. C **80** (2020) 1085 (cit. on p. 27).

[29] CMS Collaboration, *Search for physics beyond the standard model in events with two leptons of same sign, missing transverse momentum, and jets in proton–proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **77** (2017) 578, arXiv: 1704.07323 [hep-ex] (cit. on p. 27).

[30] CMS Collaboration, *Search for standard model production of four top quarks with same-sign and multilepton final states in proton–proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **78** (2018) 140, arXiv: 1710.10614 [hep-ex] (cit. on p. 27).

[31] ATLAS Collaboration, *Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector*, Phys. Rev. D **97** (2018) 072003, arXiv: 1712.08891 [hep-ex] (cit. on p. 27).

[32] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, Phys. Lett. B **784** (2018) 173, arXiv: 1806.00425 [hep-ex] (cit. on p. 27).

[33] ATLAS Collaboration, *Observation of the associated production of a top quark and a Z boson in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **07** (2020) 124, arXiv: 2002.07546 [hep-ex] (cit. on p. 27).

[34] ATLAS Collaboration, *Performance of electron and photon triggers in ATLAS during LHC Run 2*, Eur. Phys. J. C **80** (2020) 47, arXiv: 1909.00761 [hep-ex] (cit. on pp. 28, 62).

[35] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **76** (2016) 292, arXiv: 1603.05598 [hep-ex] (cit. on pp. 28, 67, 68).

[36] ATLAS Collaboration, *Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ cross sections in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **99** (2019) 072009, arXiv: 1901.03584 [hep-ex] (cit. on pp. 28, 86, 94, 103, 104, 109, 116, 123, 139, 141, 152, 156, 160–162, 167).

[37] ATLAS Collaboration, *Measurements of the inclusive and differential production cross sections of a top-quark-antiquark pair in association with a Z boson at $\sqrt{s} = 13$ TeV with the ATLAS detector*, (2020), arXiv: 2103.12603 [hep-ex] (cit. on pp. 28, 71, 78, 81, 83, 141).

[38] *Snow White never could read particle physics for very long.: Mean girl quotes, that one person, funny confessions*, 2012, URL: https://www.pinterest.com/pin/107101297358534412/ (cit. on p. 29).

[39] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (9 1964) 321, URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321 (cit. on p. 29).

[40] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (16 1964) 508, URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508 (cit. on p. 29).

[41] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (20 1964) 585, URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585 (cit. on p. 29).

[42] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, Phys. Rev. **145** (4 1966) 1156, URL: https://link.aps.org/doi/10.1103/PhysRev.145.1156 (cit. on p. 29).

[43] M. Gell-Mann, *A schematic model of baryons and mesons*, Physics Letters **8** (1964) 214 (cit. on p. 29).

[44] *Standard Model of Elementary Particles*, 2022, URL: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cit. on p. 30).

[45] S. Bethke, *Determination of the QCD coupling* α_S, Journal of Physics G: Nuclear and Particle Physics **26** (2000) R27, URL: https://doi.org/10.1088/0954-3899/26/7/201 (cit. on p. 30).

[46] G. Gabrielse, D. Hanneke, T. Kinoshita, M. Nio and B. Odom, *New Determination of the Fine Structure Constant from the Electron g Value and QED*, Phys. Rev. Lett. **97** (3 2006) 030802, URL: https://link.aps.org/doi/10.1103/PhysRevLett.97.030802 (cit. on pp. 30, 33).

[47] J. W. Rohlf, *Modern Physics from A to Z0 Chapter 18*, 1994 (cit. on p. 30).

[48] *How many fundamental constants does it take to explain the universe?*, URL: https://ep-news.web.cern.ch/content/how-many-fundamental-constants-does-it-take-explain-universe (cit. on p. 32).

[49] M. Peskin, *An introduction to quantum field theory*, CRC press, 2018 (cit. on pp. 32, 33).

[50] A. I. Akhiezer, *Quantum electrodynamics*, (1965) (cit. on p. 33).

[51] F. Halzen and A. D. Martin, *Quark & Leptons: An introductory course in modern particle physics*, John Wiley & Sons, 2008 (cit. on p. 33).

[52] M. Hunacek, *Lie groups, Lie algebras, and representations: An elementary introduction, by Brian Hall. Pp. 351.£ 50. 2003. ISBN 0 387 401229 (Springer-Verlag).*, The Mathematical Gazette **89** (2005) 149 (cit. on p. 33).

[53] S. L. Glashow, *Towards a unified theory: Threads in a tapestry*, Reviews of Modern Physics **52** (1980) 539 (cit. on p. 33).

[54] S. Weinberg, *Conceptual foundations of the unified theory of weak and electromagnetic interactions*, Science **210** (1980) 1212 (cit. on p. 33).

[55] A. Salam, *Gauge unification of fundamental forces*, Reviews of Modern Physics **52** (1980) 525 (cit. on p. 33).

[56] P. Langacker, *Grand unified theories and proton decay*, Physics Reports **72** (1981) 185 (cit. on p. 33).

[57] W. Marciano and H. Pagels, *Quantum chromodynamics*, Physics Reports **36** (1978) 137 (cit. on p. 33).

[58] H. D. Politzer, *Asymptotic freedom: An approach to strong interactions*, Physics Reports **14** (1974) 129 (cit. on p. 34).

[59] A. M. Polyakov, *Quark confinement and topology of gauge theories*, Nuclear Physics B **120** (1977) 429 (cit. on p. 34).

[60] *Argonne National Laboratory*, URL: https://www.anl.gov/phy/3d-structure-of-protons-and-neutrons (cit. on p. 34).

[61] N. N. Bogoli'Ubov and D. V. Shirkov, *Introduction to the theory of quantized fields*, Introduction to the theory of quantized fields (1959) (cit. on p. 35).

[62] F. E. Close, R. Roberts and G. G. Ross, *Factorisation scale independence, the connection between alternative explanations of the EMC effect and QCD predictions for nuclear properties*, Nuclear Physics B **296** (1988) 582 (cit. on p. 35).

[63] M. Kobayashi and T. Maskawa, *CP-violation in the renormalizable theory of weak interaction*, Progress of theoretical physics **49** (1973) 652 (cit. on p. 35).

[64] L. H. Orr and J. L. Rosner, *Comparison of top quark hadronization and decay rates*, Physics Letters B **246** (1990) 221 (cit. on p. 36).

[65] T. Humanic, *Extracting the hadronization timescale in TeV proton–proton collisions from pion and kaon femtoscopy*, Journal of Physics G: Nuclear and Particle Physics **41** (2014) 075105 (cit. on p. 36).

[66] M. Carena, M. Olechowski, S. Pokorski and C. Wagner, *Electroweak symmetry breaking and bottom-top Yukawa unification*, Nuclear Physics B **426** (1994) 269 (cit. on p. 36).

[67] ATLAS Collaboration, *Top working group cross-section summary plots: Autumn 2020*, ATL-PHYS-PUB-2020-029, 2020, URL: http://cds.cern.ch/record/2746598 (cit. on p. 37).

[68] A. Aivazis, *Draw feynman diagram online*, URL: https://feynman.aivazis.com/ (cit. on pp. 38, 57, 76).

[69] C. Monini, 'Single-top s channel cross-section measurement with the ATLAS detector', PhD thesis, 2014 (cit. on p. 40).

[70] G. Van Rossum and F. L. Drake Jr, *Python reference manual*, Centrum voor Wiskunde en Informatica Amsterdam, 1995 (cit. on pp. 41, 75).

[71] B. Aaronovitch, *Moon Over Soho: Book 2 in the# 1 bestselling Rivers of London series*, Hachette UK, 2011 (cit. on p. 43).

[72] L. Arnaudon et al., *Linac4 technical design report*, tech. rep., 2006 (cit. on p. 44).

[73] K. Hanke, *Past and present operation of the CERN PS Booster*, International Journal of Modern Physics A **28** (2013) 1330019 (cit. on p. 44).

[74] S. Gilardoni et al., *Fifty years of the CERN Proton Synchrotron: Volume 2*, arXiv preprint arXiv:1309.6923 (2013) (cit. on p. 44).

[75] H. van Hees and R. Rapp, *Dilepton radiation at the CERN super-proton synchrotron*, Nuclear Physics A **806** (2008) 339 (cit. on p. 44).

[76] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003 (cit. on pp. 44, 52, 53).

[77] 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. 44).

[78] J.J. Goodson, 'Search for Supersymmetry in States with Large Missing Transverse Momentum and Three Leptons including a Z-Boson', Presented 17 Apr 2012, PhD thesis: Stony Brook University, 2012 (cit. on p. 46).

[79] J. Pequenao, 'Computer generated image of the whole ATLAS detector', 2008, URL: `http://cds.cern.ch/record/1095924` (cit. on p. 47).

[80] ATLAS Collaboration, *Alignment of the ATLAS Inner Detector in Run-2*, (2020), arXiv: `2007.07624 [hep-ex]` (cit. on pp. 48, 49).

[81] B. Abbott et al., *Production and integration of the ATLAS Insertable B-Layer*, JINST **13** (2018) T05008, arXiv: `1803.00844 [physics.ins-det]` (cit. on p. 49).

[82] ATLAS Collaboration, *ATLAS Pixel Detector: Technical Design Report*, ATLAS-TDR-11; CERN-LHCC-98-013, 1998, URL: `https://cds.cern.ch/record/381263` (cit. on p. 49).

[83] ATLAS Collaboration, *Operation and performance of the ATLAS semiconductor tracker*, JINST **9** (2014) P08009, arXiv: `1404.7473 [hep-ex]` (cit. on p. 50).

[84] ATLAS Collaboration, *Performance of the ATLAS Transition Radiation Tracker in Run 1 of the LHC: tracker properties*, JINST **12** (2017) P05002, arXiv: `1702.06473 [hep-ex]` (cit. on p. 50).

[85] ATLAS Collaboration, *Readiness of the ATLAS Tile Calorimeter for LHC collisions*, Eur. Phys. J. C **70** (2010) 1193, arXiv: `1007.5423 [hep-ex]` (cit. on pp. 50, 51).

[86] ATLAS Collaboration, *ATLAS Liquid Argon Calorimeter: Technical Design Report*, ATLAS-TDR-2; CERN-LHCC-96-041, 1996, URL: `https://cds.cern.ch/record/331061` (cit. on p. 50).

[87] ATLAS Collaboration, *ATLAS Tile Calorimeter: Technical Design Report*, ATLAS-TDR-3; CERN-LHCC-96-042, 1996, URL: `https://cds.cern.ch/record/331062` (cit. on p. 50).

[88] ATLAS Collaboration, *ATLAS Muon Spectrometer: Technical Design Report*, ATLAS-TDR-10; CERN-LHCC-97-022, CERN, 1997, URL: `https://cds.cern.ch/record/331068` (cit. on p. 52).

[89] T. Colombo, *Data-flow Performance Optimisation on Unreliable Networks: the ATLAS Data-Acquisition Case*, Journal of Physics: Conference Series **608** (2015) 012005 (cit. on p. 54).

[90] Proceedings of the Royal Institution of Great Britain **35** (1951) (cit. on p. 55).

[91] M. I. Jordan and T. M. Mitchell, *Machine learning: Trends, perspectives, and prospects*, Science **349** (2015) 255 (cit. on p. 55).

[92] R. Caruana and A. Niculescu-Mizil, 'An empirical comparison of supervised learning algorithms', *Proceedings of the 23rd international conference on Machine learning*, 2006 161 (cit. on p. 55).

[93] S. Höche, 'Introduction to parton-shower event generators', *Journeys Through the Precision Frontier: Amplitudes for Colliders: TASI 2014 Proceedings of the 2014 Theoretical Advanced Study Institute in Elementary Particle Physics*, World Scientific, 2016 235 (cit. on p. 55).

[94] J. C. Collins, D. E. Soper and G. Sterman, 'Factorization of hard processes in QCD', *Perturbative QCD*, World Scientific, 1989 1 (cit. on p. 55).

[95] R. D. Ball et al., *Parton distributions with LHC data*, Nuclear Physics B **867** (2013) 244 (cit. on p. 55).

[96] S. Carrazza, S. Forte and J. Rojo, *Parton distributions and event generators*, arXiv preprint arXiv:1311.5887 (2013) (cit. on p. 55).

[97] F. E. Close, *Quarks and partons*, tech. rep., 1976 (cit. on pp. 55, 56).

[98] T. Rogers and M. Strikman, *Multiple hard partonic collisions with correlations in proton-proton scattering*, Physical Review D **81** (2010) 016013 (cit. on pp. 55, 56).

[99] F. Bunkin and M. Fedorov, *Bremsstrahlung in a strong radiation field*, Sov. Phys. JETP **22** (1966) 844 (cit. on p. 55).

[100] T. Alexopoulos et al., *The role of double parton collisions in soft hadron interactions*, Physics Letters B **435** (1998) 453 (cit. on p. 55).

[101] R. Stock, *The parton to hadron phase transition observed in Pb+ Pb collisions at 158 GeV per nucleon*, Physics Letters B **456** (1999) 277 (cit. on p. 55).

[102] S. Dittmaier, *A general approach to photon radiation off fermions*, Nuclear Physics B **565** (2000) 69 (cit. on p. 55).

[103] S. Dulat et al., *New parton distribution functions from a global analysis of quantum chromodynamics*, Physical Review D **93** (2016) 033006 (cit. on p. 56).

[104] R. D. Ball et al., *Parton distributions for the LHC Run II*, Journal of High Energy Physics **2015** (2015) 40 (cit. on p. 56).

[105] J. D. Bjorken, *Asymptotic sum rules at infinite momentum*, Physical Review **179** (1969) 1547 (cit. on pp. 56, 57).

[106] M. Mangano, *The color structure of gluon emission*, Nuclear Physics B **309** (1988) 461 (cit. on p. 57).

[107] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross sections in deep-inelastic scattering*, arXiv preprint hep-ph/9907280 (1999) (cit. on p. 58).

[108] *File:quark confinement.svg*, URL: `https://commons.wikimedia.org/wiki/File:Quark_confinement.svg` (cit. on p. 58).

[109] A. H. Hoang, *The Top Mass: Interpretation and Theoretical Uncertainties*, 2014, arXiv: `1412.3649 [hep-ph]` (cit. on pp. 58, 59).

[110] S. Schumann and F. Krauss, *A parton shower algorithm based on Catani-Seymour dipole factorisation*, Journal of High Energy Physics **2008** (2008) 038 (cit. on p. 59).

[111] B. Andersson, G. Gustafson and L. Lönnblad, *Gluon splitting in the colour dipole cascades*, Nuclear Physics B **339** (1990) 393 (cit. on p. 59).

[112] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: `1405.0301` (cit. on p. 59).

[113] R. D. Ball et al., *Parton distributions for the LHC run II*, JHEP **04** (2015) 040, arXiv: 1410.8849 [hep-ph] (cit. on p. 59).

[114] S. Frixione, E. Laenen, P. Motylinski and B. R. Webber, *Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations*, JHEP **04** (2007) 081, arXiv: hep-ph/0702198 (cit. on p. 59).

[115] P. Artoisenet, R. Frederix, O. Mattelaer and R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, JHEP **03** (2013) 015, arXiv: 1212.3460 [hep-ph] (cit. on p. 59).

[116] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph] (cit. on pp. 59, 65, 99).

[117] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on p. 59).

[118] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. A **462** (2001) 152 (cit. on p. 59).

[119] de Florian, D. and others, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, (2016), arXiv: 1610.07922 [hep-ph] (cit. on p. 59).

[120] ATLAS Collaboration, *Modelling of the $t\bar{t}H$ and $t\bar{t}V$ $(V = W, Z)$ processes for $\sqrt{s} = 13$ TeV ATLAS analyses*, ATL-PHYS-PUB-2016-005, 2016, URL: https://cds.cern.ch/record/2120826 (cit. on pp. 59, 98, 99).

[121] M. Bähr et al., *Herwig++ physics and manual*, Eur. Phys. J. C **58** (2008) 639, arXiv: 0803.0883 [hep-ph] (cit. on pp. 59, 98, 99).

[122] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, Eur. Phys. J. C **76** (2016) 196, arXiv: 1512.01178 [hep-ph] (cit. on pp. 59, 98, 99).

[123] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv: 1002.2581 [hep-ph] (cit. on p. 60).

[124] E. Bothmann et al., *Event Generation with Sherpa 2.2*, (2019), arXiv: 1905.09127 [hep-ph] (cit. on p. 60).

[125] ATLAS Collaboration, *Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2*, Eur. Phys. J. C **77** (2017) 673, arXiv: 1704.07983 [hep-ex] (cit. on p. 60).

[126] *vertex fitting and detector alignment*, URL: https://idpasc.lip.pt/uploads/talk/file/71/Tracking-Vertexing-Slides.pdf (cit. on p. 60).

[127] C. Kourkoumelis and S. Vourakis, *HYPATIA—an online tool for ATLAS event visualization*, Physics Education **49** (2014) 21 (cit. on p. 61).

[128] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015-2017 LHC proton-proton collision data*, (2019), arXiv: 1908.00005 [hep-ex] (cit. on pp. 61–63, 98, 99).

[129] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using the 2012 LHC proton–proton collision data*, ATLAS-CONF-2014-032, 2014, URL: https://cds.cern.ch/record/1706245 (cit. on p. 61).

[130] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using 2012 LHC proton–proton collision data*, Eur. Phys. J. C **77** (2017) 195, arXiv: 1612.01456 [hep-ex] (cit. on p. 62).

[131] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 639, arXiv: 1902.04655 [hep-ex] (cit. on p. 62).

[132] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: 0802.1189 (cit. on pp. 63–65).

[133] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv: 1111.6097 [hep-ph] (cit. on p. 63).

[134] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Eur. Phys. J. C **77** (2017) 490, arXiv: 1603.02934 [hep-ex] (cit. on p. 63).

[135] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: 1703.10485 [hep-ex] (cit. on p. 63).

[136] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv: 1510.03823 [hep-ex] (cit. on p. 63).

[137] G. P. Salam, *Towards jetography*, The European Physical Journal C **67** (2010) 637 (cit. on p. 63).

[138] S. Ellis, J. Huston, K. Hatakeyama, P. Loch and M. Tönnesmann, *Jets in hadron–hadron collisions*, Progress in Particle and Nuclear Physics **60** (2008) 484 (cit. on p. 63).

[139] ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, (2020), arXiv: 2007.02645 [hep-ex] (cit. on pp. 64, 98, 99).

[140] ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, ATLAS-CONF-2014-018, 2014, URL: https://cds.cern.ch/record/1700870 (cit. on pp. 64, 65).

[141] *File:B-tagging diagram.svg*, URL: https://commons.wikimedia.org/wiki/File:B-tagging_diagram.svg (cit. on p. 66).

[142] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s}$ = 13 TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: 1907.05120 [hep-ex] (cit. on p. 66).

[143] ATLAS Collaboration, *Performance of the ATLAS muon triggers in Run 2*, (2020), arXiv: 2004.13447 [hep-ex] (cit. on p. 67).

[144] M. O. Evans, *LHC open data for the world to see*, PoS **LHCP2020** (2020) 199 (cit. on p. 68).

[145] T. McCauley, *HEP data for everyone: CERN open data and the ATLAS and CMS experiments*, PoS **ICHEP2016** (2017) 335 (cit. on p. 68).

[146] F. Socher, *ATLAS and CMS Data Release & Tools*, PoS **LHCP2016** (2016) 112 (cit. on p. 68).

[147] M. O. Evans, 'Enabling Open Science with the ATLAS Open Data project at CERN', Presented 01 Jun 2018, 2018, URL: https://cds.cern.ch/record/2630961 (cit. on p. 68).

[148] M. O. Evans, *ATLAS Open Data project: HEP for everyone*, PoS **ICHEP2018** (2019) 275 (cit. on p. 68).

[149] C. Doglioni, *The ATLAS Open Data project*, tech. rep. ATL-OREACH-PROC-2018-001, CERN, 2018, URL: https://cds.cern.ch/record/2637284 (cit. on p. 68).

[150] Y.-F. Lo, *ATLAS Open Data: software for visualization and physics analysis*, (2019), URL: https://cds.cern.ch/record/2655357 (cit. on p. 68).

[151] A. Sanchez Pineda and S. Mehlhase, *ATLAS Outreach: on the dissemination of High Energy Physics and Computer Sciences*, PoS **LHCP2019** (2019) 008 (cit. on p. 68).

[152] ATLAS Collaboration, *ATLAS Open Data*, 2021, URL: http://opendata.atlas.cern/ (visited on 23/02/2021) (cit. on pp. 68, 143, 144).

[153] CERN, *CERN Open Data*, 2021, URL: http://opendata.cern.ch/ (visited on 23/02/2021) (cit. on pp. 68, 144).

[154] L. Hagesjo, 'Making ATLAS Data from CERN Accessible to the General Public: The Development and Evaluation of a Learning Resource in Experimental Particle Physics.', Presented 12 Jun 2017, 2017, URL: https://cds.cern.ch/record/2281684 (cit. on p. 68).

[155] ATLAS Collaboration, *Review of ATLAS Open Data* 8 *TeV datasets, tools and activities*, ATL-OREACH-PUB-2018-001, 2018, URL: https://cds.cern.ch/record/2624572/files/ATL-OREACH-PUB-2018-001.pdf (cit. on p. 68).

[156] R. C. Camacho Toro, *Outreaching particle physics to Latin America: CEVALE2VE and the use of ATLAS open data*, PoS **EPS-HEP2017** (2017) 554 (cit. on p. 68).

[157] A. Adiguzel et al., *Evaluating Analysis Description Language Concept as a First Introduction to Analysis in Particle Physics*, 2020, arXiv: 2008.12034 [hep-ph] (cit. on p. 68).

[158] A. Sanchez Pineda, *The Cevale2ve case*, PoS **ICHEP2016** (2017) 322 (cit. on p. 68).

[159] C. Kourkoumelis, S. Vourakis and S. Sotiriou, *HEP interactive activities in high schools in the framework of the CREATIONS project*, PoS **EPS-HEP2017** (2017) 559 (cit. on p. 68).

[160] D. Fassouliotis, C. Kourkoumelis and S. Vourakis, 'HEP in Greek Classes', *EPJ Web of Conferences*, vol. 137, EDP Sciences, 2017 01012 (cit. on p. 68).

[161] A. Adiguzel et al., *CutLang as an analysis description language for introducing students to analyses in particle physics*, European Journal of Physics (2021) (cit. on p. 68).

[162] G. Ünel and S. Sekmen, *CutLang: A particle physics analysis description language and runtime interpreter*, Computer Physics Communications **233** (2018) 215 (cit. on p. 68).

[163] F. Ould-Saada, 'ATLAS Open Data–Development of a simple-but-real HEP data analysis framework', *EPJ Web of Conferences*, vol. 245, EDP Sciences, 2020 08023 (cit. on p. 68).

[164] J. J. C. Mckeown, 'A Contribution to ATLAS Open Data Collaboration at CERN', Presented 15 May 2019, 2019, URL: https://cds.cern.ch/record/2704159 (cit. on p. 68).

[165] S. Biryukov, 'A Contribution to the ATLAS Open Data Collaboration at CERN', Presented 2020, 2020, URL: https://cds.cern.ch/record/2724158 (cit. on p. 68).

[166] L. Thulasidharan, *J/ψ, Y and Z studies with CERN Open Data*, DESY Summer Student Report (2018) (cit. on p. 68).

[167] M. R. Di Domenico Franco, 'Reconstruction of the invariant masses of bosons of the Standard Model using public data from ATLAS Open Data', Presented 04 Oct 2017, 2017, URL: https://cds.cern.ch/record/2293251 (cit. on p. 68).

[168] A. Prieto Tirado, 'W' and Z' bosons search at $\sqrt{s} = 13$ TeV with the ATLAS detector at the LHC', Presented 11 Jun 2020, 2020, URL: https://cds.cern.ch/record/2743513 (cit. on p. 68).

[169] I. Garcia, 'Perspectivas y Evaluación de producción de Materia Oscura en asociación con un quark liviano, un quark pesado (quark b) o un boson electro débil en colisiones de particular a energías de colisión de $\sqrt(s) = 8$ TeV', Presented 14 Jul 2017, 2017, URL: https://cds.cern.ch/record/2291838 (cit. on p. 68).

[170] ATLAS Collaboration, *Review of the 13 TeV ATLAS Open Data release*, ATL-OREACH-PUB-2020-001, 2020, URL: https://cds.cern.ch/record/2707171 (cit. on pp. 69, 143, 144).

[171] *Top 6 quotes by Amber Naslund: A-Z quotes*, URL: https://www.azquotes.com/author/64233-Amber_Naslund (cit. on p. 71).

[172] ATLAS Collaboration, *Measurement of the $t\bar{t}$ production cross-section and lepton differential distributions in $e\mu$ dilepton events from pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Eur. Phys. J. C **80** (2020) 528, arXiv: 1910.08819 [hep-ex] (cit. on pp. 76, 77).

[173] ATLAS Collaboration, *Measurements of the production cross section of a Z boson in association with jets in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Eur. Phys. J. C **77** (2017) 361, arXiv: 1702.05725 [hep-ex] (cit. on pp. 76, 77).

[174] T. Kittelmann, V. Tsulaia, J. Boudreau and E. Moyse, 'The Virtual Point 1 event display for the ATLAS experiment', *Journal of Physics: Conference Series*, vol. 219, 3, IOP Publishing, 2010 032012 (cit. on p. 78).

[175] X. Wu et al., *Top 10 algorithms in data mining*, Knowledge and information systems **14** (2008) 1 (cit. on p. 86).

[176] T. Schaapherder, 'Photon Conversion Classification by Boosting Decision Trees', B.S. thesis, 2018 (cit. on p. 86).

[177] Y. Shin, *Application of Boosting Regression Trees to Preliminary Cost Estimation in Building Construction Projects*, Computational intelligence and neuroscience **2015** (2015) 149702 (cit. on p. 87).

[178] A. Dertat, *Applied deep learning - part 1: Artificial Neural Networks*, 2017, URL: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (cit. on p. 88).

[179] *Tensorflow*, URL: https://www.tensorflow.org/ (cit. on pp. 89, 90, 109, 117).

[180] K. Team, *Simple. flexible. powerful.*, URL: https://keras.io/ (cit. on pp. 89, 90, 109, 117).

[181] G. C. Fox and S. Wolfram, *Event shapes in e+ e- annihilation*, Nuclear Physics B **149** (1979) 413 (cit. on p. 92).

[182] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **76** (2016) 653, arXiv: 1608.03953 [hep-ex] (cit. on p. 98).

[183] ATLAS Collaboration, *Identification and rejection of pile-up jets at high pseudorapidity with the ATLAS detector*, Eur. Phys. J. C **77** (2017) 580, arXiv: 1705.02211 [hep-ex] (cit. on p. 98), Erratum: Eur. Phys. J. C **77** (2017) 712.

[184] G. Aad et al., *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at $\sqrt{s}$ = 13 TeV*, The European Physical Journal C **81** (2021) 1 (cit. on pp. 98, 99).

[185] S. Höche, F. Krauss, M. Schönherr and F. Siegert, *QCD matrix elements+ parton showers. The NLO case*, Journal of High Energy Physics **2013** (2013) 1 (cit. on pp. 98, 99).

[186] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, J. Phys. G **43** (2016) 023001, arXiv: 1510.03865 [hep-ph] (cit. on pp. 98, 100).

[187] N. Lavesson and L. Lönnblad, *Extending CKKW-merging to one-loop matrix elements*, Journal of High Energy Physics **2008** (2008) 070 (cit. on pp. 98, 100).

[188] S. Höche, F. Krauss, S. Schumann and F. Siegert, *QCD matrix elements and truncated showers*, Journal of High Energy Physics **2009** (2009) 053 (cit. on pp. 98, 100).

[189] K. Cranmer, G. Lewis, L. Moneta, A. Shibata and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, CERN-OPEN-2012-016, URL: https://cds.cern.ch/record/1456844 (cit. on p. 101).

[190] J. L. Collins, *The simple path to wealth: Your road map to financial independence and a rich, Free Life*, JL Collins, 2016 (cit. on p. 103).

[191] *Learn: Machine learning in python - scikit-learn 0.16.1 documentation*, URL: https://scikit-learn.org/ (cit. on p. 104).

[192] scikit-learn, *sklearn.ensemble.GradientBoostingClassifier*, 2020, URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (visited on 11/01/2021) (cit. on p. 104).

[193] O. Bessidskaia Bylund et al., *Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ production cross sections in multilepton final states using 36.1 fb^{-1} of pp collisions at 13 TeV at the LHC*, tech. rep., CERN, 2016, URL: https://cds.cern.ch/record/2235582 (cit. on p. 141).

[194] *Respect your parents. they passed school without google.: Respect your parents, funny quotes, parenting quotes*, 2014, URL: https://www.pinterest.com/pin/416934877975514282/ (cit. on p. 143).

[195] L. Serkin, 'The release of the 13 TeV ATLAS Open Data: using open education resources effectively', *EPJ Web of Conferences*, vol. 245, EDP Sciences, 2020 08026 (cit. on p. 143).

[196] M. O. Evans, *ATLAS Open Data: Data visualisation and educational physics analysis to re-discover the Higgs boson*, PoS **LHCP2020** (2020) 228 (cit. on pp. 144, 156).

[197] ATLAS Collaboration, *Get started with data visualisation*, 2020, URL: http://opendata.atlas.cern/release/2020/documentation/visualization/index.html (visited on 23/02/2021) (cit. on p. 144).

[198] Atlas-Outreach-Data-Tools, *Notebooks-collection-opendata/infofile.py at master · ATLAS-outreach-data-tools/notebooks-collection-opendata*, 2021, URL: https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/blob/master/13-TeV-examples/uproot_python/infofile.py (cit. on p. 146).

[199] T. Kluyver et al., *Jupyter Notebooks-a publishing format for reproducible computational workflows.* Vol. 2016, 2016 (cit. on p. 156).

[200] Y. Wang, *A Jupyter-based Interface Integrated with ROOT for ATLAS Open Data Analysis*, (2019), URL: https://cds.cern.ch/record/2687389 (cit. on p. 156).

[201] A. Sanchez Pineda, *Integration of ROOT NoteBook as an ATLAS analysis web-based tool in outreach and public data release projects*, PoS **ICHEP2016** (2017) 1054 (cit. on p. 156).

[202] ATLAS Collaboration, *ttZ_ML_from_csv.ipynb*, 2021, URL: https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/blob/master/13-TeV-examples/uproot_python/ttZ_ML_from_csv.ipynb (visited on 23/02/2021) (cit. on p. 156).

[203] ATLAS Collaboration, *ttZ_ML_from_root.ipynb*, 2021, URL: https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/blob/master/13-TeV-examples/uproot_python/ttZ%5C_ML%5C_from%5C_root.ipynb (visited on 23/02/2021) (cit. on p. 160).

[204] *A conclusion is the place where you got tired of thinking. at quotetab*, URL: https://www.quotetab.com/quote/by-source-unknown/a-conclusion-is-the-place-where-you-got-tired-of-thinking?source=thoughts-and-thinking (cit. on p. 167).