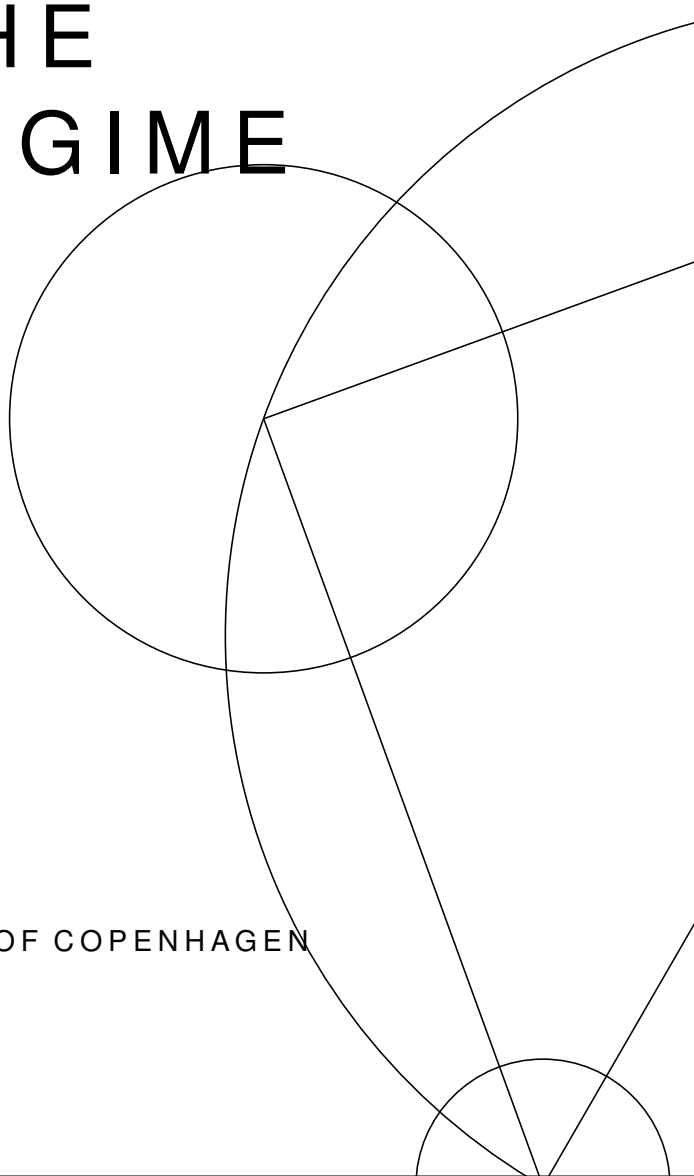
DANIEL STEFANIAK NIELSEN

AN ALTERNATIVE ANALYSIS OF THE SEMI-LEPTONIC DIBOSON FINAL STATES IN THE BOOSTED REGIME



MSC THESIS
NIELS BOHR INSTITUTE, UNIVERSITY OF COPENHAGEN

AN ALTERNATIVE ANALYSIS OF THE SEMI-LEPTONIC DIBOSON FINAL STATES IN THE BOOSTED REGIME

DANIEL STEFANIAK NIELSEN

NIELS BOHR INSTITUTE,
UNIVERSITY OF COPENHAGEN

Supervisor	Associate Professor Troels C. Petersen
Censor	Esben B. Klinkby (Technical University of Denmark)

© Daniel Stefaniak Nielsen 2016
An alternative analysis of semi-leptonic diboson final states in the boosted regime
MSc Thesis, University of Copenhagen
viii + 88 pages; illustrated, with bibliographic references

Set in 10/14 pt Palatino Linotype using pdfLaTeX with the Tufte-LaTeX package
Cover art: University of Copenhagen logo grid

Thesis submitted on the 14th of October 2016 and defended on the 28th of October 2016 for the completion of the degree of Master of Science (MSc) in Physics at the Niels Bohr Institute, University of Copenhagen.

First printing, October 2016

Contents

Abstract v
Disclaimer v
Acknowledgement vii
Introduction viii

1 Theory 1

1.1 The Standard Model 2
1.1.1 Units 3
1.1.2 Interactions and decay 3
1.1.3 Quantum Field Theory 4
1.1.4 The Higgs mechanism 8

1.2 Collider Physics 9
1.2.1 Coordinate system 9
1.2.2 Charged particle and photon interactions 9
1.2.3 Parton Distribution Functions 11
1.2.4 Factorization theorem 12
1.2.5 Luminosity 13
1.2.6 Pileup and soft processes 13

1.3 The ATLAS detector 14
1.3.1 Overview of the ATLAS detector 15
1.3.2 Inner detector 17
1.3.3 The calorimeter 19
1.3.4 Muon spectrometer 20
1.3.5 ATLAS trigger system 21
1.3.6 A note on resolutions 22

1.4 Reconstruction and software 23
1.4.1 Track reconstruction 23
1.4.2 Cluster reconstruction 24
1.4.3 Electron reconstruction and identification 25
1.4.4 Muon reconstruction and identification 25
1.4.5 Jet reconstruction 26
1.4.6 Missing transverse energy 28
1.4.7 Isolation and overlap-removal 29
1.4.8 ATLAS software 30

1.5 Boosted Decision Trees in TMVA 32

2 Analysis 35

2.1 Improving the jet charge 36
2.1.1 Datasets 36
2.1.2 Selection and cuts 37
2.1.3 Methods 37
2.1.4 Results 39
2.1.5 Charge as a separating variable 43
2.1.6 Final notes 44

2.2 Object definition 45
2.2.1 Datasets 45
2.2.2 Selection and cuts 46
2.2.3 Saved values 47
2.2.4 Monte-Carlo and data agreement 48
2.2.5 Final selection and signal region 51

2.3 Tagging W bosons 53
2.3.1 Input variables 53
2.3.2 Optimizing the BDT 56
2.3.3 Results 57

2.4 Full event tagger 60

2.5 The likelihood fit 63
2.5.1 Correlations 63
2.5.2 4 cuts 67
2.5.3 3 cuts 70
2.5.4 2 cuts 71
2.5.5 0 cuts 72
2.5.6 Final notes 74

2.6 Concluding remarks 76
2.6.1 Summary 76
2.6.2 Outlook 76

Appendices 79

A.1 Datasets 80
A.1.1 For charge calculation 80
A.1.2 For diboson studies 80

References 85

Abstract

The Large Hadron Collider (LHC) has increased the center-of-mass energy to 13 TeV for Run 2, and it will reach 14 TeV during 2017. The higher energies compared to Run 1 has allowed analyses to search for new and ever-heavier particles. This thesis will present a series of attempts at improving the diboson analyses in the boosted regime.

The charge of jets is badly reconstructed. This is due to the nature of QCD. An improvement in the calculation of jet charges as well as their subjet charge differences will open a window into W, Z, and background separation. Adding ΔR between the ghost-matched tracks and the jet to the charge calculation improved the separation of W^+ and W^-. However, the final plots of subjet charge sums and difference show very little separation power.

In ATLAS, a tagger will provide a value for a candidate particle, and cut levels (e.g. loose, medium) are prepared with given signal efficiencies. In this study, boson taggers using BDTs and the Fisher discriminant were constructed. The score from the selected tagger will be used directly in a likelihood fit without cutting. An additional tagger that uses the whole event information for the $HVT \to WW \to lvqq$ process has also been constructed, where HVT is the Heavy Vector Triplet. The simultaneous fit of the scores from the two taggers as well as m_W and m_{lvqq} provides a greater signal yield compared to simple cuts on the variables. Issues with actual W particles from background being assigned to the signal PDF has been partly addressed but needs further study. Correlations between the four variables used in the simultaneous fit also need further study.

Disclaimer

Some of the results presented in this thesis, be they plots or figures, are based on official ATLAS simulated data from the mc15c simulation project. These are the product of the authors' own work and have *not* been approved by the ATLAS Collaboration, and therefore they are labeled as 'Work in progress'. Any replication of these results should clearly reflect this fact.

Acknowledgement

I would like to thank my supervisor, Troels C. Pedersen. I was looking for a data analysis project in the particle physics field, and he was ready and full of ideas right of the bat. And he always has advice for how to proceed when I get stuck.

Another thanks goes to Lars Egholm Pedersen with whom I shared an office with for a good part of the past year. He was always ready to help with ROOT, code, and physics issues when I asked. A special thanks for angle computation code that I have used in my work.

For his help with code as well but especially his framework, I would like to give thanks to Geert-Jan Besjes. Keeping track of all the packages myself turned out to be too much work.

And last, but not least, thanks to my two master thesis office mates, Freja Thoresen and Mike Lauge, for the awesome atmosphere that made the not-so-few days of never-ending bugs seem much less bleak.

Introduction

The aim of this thesis was to look into how to improve the diboson final states analyses in the boosted regime. The boosted regime begins at approx. 500 GeV.

The thesis work turned into two studies: a minor study into improving the charge calculation of jets and a study into using scores from machine learning algorithms in a simultaneous likelihood fit with masses.

In Section 1.1, the theory of particle physics is introduced. Section 1.2 introduces the phenomenological theory of particle physics. The section serves to link the theoretical calculations of processes and the quantities that are actually measured in an experiment. Section 1.3 gives an overview of the ATLAS experiment. In Section 1.4, the technical definitions and methods which are used in this thesis are introduced. Section 1.5 shortly introduces the machine learning algorithm Boosted Decision Trees.

To improve separation of W^+, W^-, and Z particles from each other as well as background, a side study into improving the jet charge calculation was performed in Section 2.1. Hadronically decayed W and Z particles are difficult to separate only in mass because of the large smearing, so an improved charge calculation using also the charge difference between the two subjets will possibly improve the separation. The study used large-R jets, specifically $R = 1.0$. The large-R jets, also known as fatjets, will contain both decay products of hadronically decaying boosted W particles. The substructure of the fatjets were exploited by adding a ΔR dependence to the charge calculation. The jet charge equation parameters were also studied for their p_T dependence.

For the main study, the backgrounds for a diboson final states analysis are presented in Section 2.2. The backgrounds in Monte-Carlo (MC) simulations are compared to data.

Machine learning algorithms, called taggers, were constructed in Sections 2.3 and 2.4. In the former section, a W tagger was constructed from a Fisher discriminant-based algorithm and Boosted Decision Trees (BDTs). The performance of the taggers were evaluated. In the latter section, a tagger using information from the full event was constructed in the same way.

In Section 2.5, the scores from the taggers as well as the hadronically decaying m_W and the resonance m_{lvJ}, where J is the fatjet containing the $q\bar{q}$ from the hadronically decayed W particles, were used in a simultaneous fit. Before the fit, a series of cuts were applied to estimate the gain in efficiency of the uncut, simultaneous fit.

I Theory

1.1 The Standard Model

The Standard Model (SM) is the current theory of particle physics able to explain the nature of particle physics to the greatest extend. It consists of the following Quantum Field Theories (QFTs): Quantum Chromodynamics (QCD) and the electroweak interaction (EWK). QCD is the theory of the strong interaction. Electroweak theory is the unification of electromagnetism and the weak interaction[1].

[1] For historical reasons and convenience, electromagnetism and the weak interaction are often considered separately, even if they are now known to be parts of the same interaction.

Together, they explain three of the four known fundamental forces: The strong interaction, electromagnetic interaction, weak interaction, but not gravity.

A single theory unifying all forces but gravity is known as a Grand Unified Theory (GUT). A theory unifying all four forces is known as a Theory of Everything (TOE). It has so far not been possible to unify QCD with the electroweak interaction, and hence to create these theories.

Figure 1.1 shows a table of the elementary particles of the Standard Model.

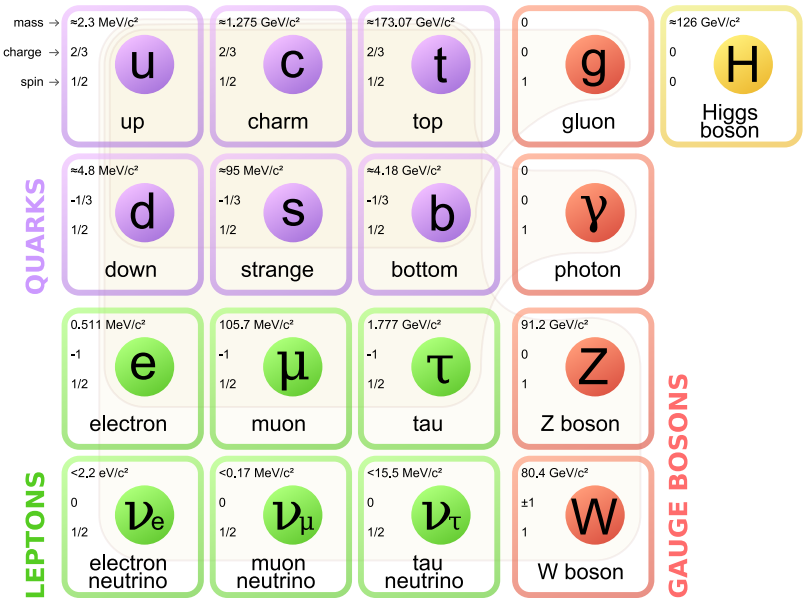


Figure 1.1: The elementary particles of the Standard Model. From [1].

Each theory of interaction is associated with one or more force carrying particle(s) (known as *gauge bosons* in the theoretical framework of QFT) that couple(s) to a quantum number.

The gluon, g, is the force carrier of the strong interaction, and it couples to the quantum number color charge. All particles with a color charge interact with the gluon. Due to the gluon itself having a color charge, it also undergoes self-interaction.

The photon, γ, is the force carrier of the electromagnetic interaction, and the associated quantum number is the (electric) charge. The photon is not itself electrically charged, and hence does not self-interact.

In the weak interaction, the force carriers are the W^+, W^-, and

Z particles, and they couple to the weak isospin, T_3.[2] The W and Z particles carry the quantum number themselves and therefore also interact with each other and themselves.

The quarks in the purple boxes of Figure 1.1 carry color charge, electric charge, and weak isospin, so they interact with all the forces of the Standard Model.

The leptons in the green boxes all carry weak isospin and therefore interact with the weak force. Moreover, the electrons, muons, and taus are also electrically charged and thus interact with the photon. No lepton has color charge, so no lepton interacts with the gluon.

The interactions are summarized in Figure 1.2.

Since neutrinos only interact through the weak force (and gravity outside of the Standard Model), they interact extremely feebly with matter. Neutrino particles moving through the Earth rarely collide with any atom at all. This poses a problem for the detection of neutrinos in particle detectors (which are known to be smaller than the Earth).

However, the Standard Model is not the final answer; many known phenomena are not explained and many parameters must be found through experiments.

Firstly, SM assumes that neutrinos are massless, which recently has been shown not to be true. Neutrino masses, gravity, dark matter and dark energy, as well as the baryon asymmetry are all not explained by the Standard Model.

Secondly, the many free parameters (mainly masses, mixing angles, and gauge couplings) of the Standard Model pose a problem for it in the sense that the underlying reasons for their values are not explained but can only be experimentally determined.

[2] Technically, T_3 is the third component of the weak isospin, but the term "weak isospin" most often refers to the third component only.

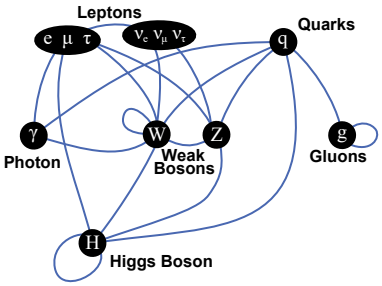


Figure 1.2: Interactions between the particles of the Standard Model. Lines looping back indicate self-interaction. From [2].

1.1.1 Units

The unit system in particle physics is called *natural units*. Instead of [kg, m, and s], [$\hbar$, c, and eV] is used. In order not to write $\hbar$ and c as often, we define $\hbar = c = 1$. In this way the units for energy, momentum and mass become electronvolts (eV), and Einstein's energy-momentum relation becomes $E^2 = m^2 + p^2$.

1.1.2 Interactions and decay

In modern particle physics, discoveries are made through violent processes where energetic particles are forced to collide to create new particles. Amongst the many created particles, an undiscovered particle may be created. This particle may be a composite particle, made up of quarks with a configuration of quantum numbers not yet measured. It may also be a new fundamental particle.

What they have in common is that almost all are unstable. They will decay to lighter particles, which in turn might decay to even lighter particles.

What drives the processes of creating new particles and their decay

is Fermi's golden rule: [3]

$$\Gamma_{fi} = 2\pi|\mathcal{M}_{fi}|^2\rho. \tag{1.1}$$

Fermi's golden rule gives us the transition probability given the (square of the) matrix element and the density of states, respectively.

In quantum mechanics it tells us the transition rate from one eigenstate (e.g. energy state) to another. This is what drives an electron to a lower energy state.

The interaction and decay probabilities in particle physics are given as *cross sections*. The name comes from the imaginary cross-sectional area of a particle where the interaction probability/rate of hitting this area is proportional to its size. Of course, in quantum mechanics, particles are not rigid balls, but the name has stuck because of the intuitive picture.

The rate of interaction per solid angle Ω of a beam of particles hitting a target is given as: [4]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{1}{I_b \rho \mathrm{d}x}\frac{\mathrm{d}N_s}{\mathrm{d}\Omega}, \tag{1.2}$$

where I_b is the current of beam particles per second, ρ is the density of target particles, $\mathrm{d}x$ is the target thickness, and $\mathrm{d}N_s$ is the rate of scatted particles.

$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}$ is known as the *differential cross section*. The *total cross section* σ is in units of area, matching its name. The unit used for cross section is barn which is $10^{-24}\,\mathrm{cm}^2$.

The definition of the differential cross section given in Equation 1.2 gives a measurement of the differential cross section that is independent of the experiment.

1.1.3 Quantum Field Theory

In a Quantum Field Theory (QFT), particles are considered excitations of their respective fields, e.g. electrons being excitations of the electron field spanning the whole Universe.

In QFT, all interactions are described by a *Lagrangian density*, $\mathcal{L}$.[3] The Euler-Lagrange equation will give, just as in classical field theory, the equations of motion. Noether's theorem states that a symmetry in the Lagrangian will give a corresponding conservation law[3].

[3] It is usually referred to as just the Lagrangian.

Local gauge invariance of a Lagrangian requires the Lagrangian to be invariant under a transformation of its fields. Making a Lagrangian gauge invariant will lead to the introduction of a gauge field with an associated gauge boson. Thus, requiring gauge invariance of the Lagrangians lead to the forces of the Standard Model.

Feynman diagrams

Calculating the matrix elements for all interactions is a long and tedious process. The *Feynman diagrams* are graphical representations of the terms in an equation calculating the matrix element. An

example can be seen in Figure 1.3. Rules for connecting lines and vertices come from the underlying conservation laws.

With Fermi's golden rule and a definition of cross section, we can combine them to obtain:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \propto |\mathcal{M}_{fi}|^2. \tag{1.3}$$

Using perturbation theory, we can calculate the matrix element for a given process. For the coulomb repulsion between two electrons of the same charge, we can draw the process like in Figure 1.3.

The process is the visualization of [3]

$$-i\mathcal{M} = \underbrace{\bar{u}(p_4)\mathrm{Vertex}_{\mathrm{QCD}}u(p_2)}_{\mathrm{I}}\underbrace{\frac{-g}{q^2}}_{\mathrm{II}}\underbrace{\bar{u}(p_3)\mathrm{Vertex}_{\mathrm{QCD}}u(p_1)}_{\mathrm{III}}. \tag{1.4}$$

The Feynman rules follow a simple recipe of combining terms when reading the diagram from right to left, top to bottom. For Equation 1.4, the following rules are applied:

Part (I): The final-state fermion, $\bar{u}$, points out of the upper vertex, connects to the QCD vertex that connects to the initial-state fermion, u, pointing into the vertex.

Part (II): The photon propagator, $\frac{-g}{q^2}$.

Part (III): Same as part (I).

p_i are the four-momenta of the particles, $g = \sqrt{\alpha_{\mathrm{EM}}}$ is the coupling constant of the interaction, q^2 is the momentum transfer, and u is the Dirac spinor, which is the magnitude of the wavefunction.

The *Mandelstam variables* are defined as:

$$\begin{aligned} s &= (p_1 + p_2)^2 = (p_3 + p_4)^2, \\ t &= (p_1 - p_3)^2 = (p_2 - p_4)^2. \end{aligned} \tag{1.5}$$

s is the square of the center-of-mass energy, $\sqrt{s} = E_{\mathrm{CoM}} = M$, which equals the invariant mass of the system.

Figure 1.3 is a so-called t-channel process, and Figure 1.4 is an s-channel process.

Figure 1.3 is only at *leading order* (also called *tree level*). Leading order diagrams describe the process, while next-to-leading order (NLO) and higher (NNLO, etc.) describe corrections. An example of an NLO process would be one of the initial-state particles radiating a particle before interacting, e.g. a gluon radiating a gluon as in Figure 1.4.

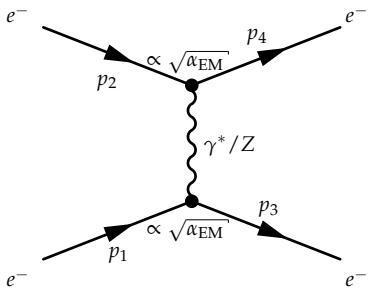


Figure 1.3: Coulomb scattering between two electrons through the exchange of a virtual photon. The strength of the interaction is related to the product of the vertices: α_{EM}.

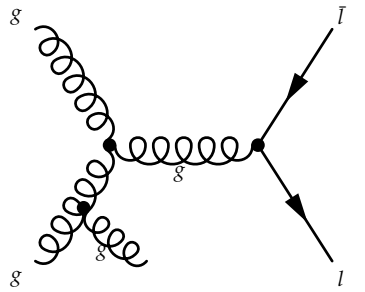


Figure 1.4: Example of an NLO process. The number of vertices dictate the level.

Resonances

Unstable particles can decay through more than one *channel*. They can decay to a variety of different final state particles through different interactions. Each individual decay channel has its own rate, Γ_i, also called *partial width*, related to the matrix element through,

$$\Gamma_i \propto |\mathcal{M}|^2. \tag{1.6}$$

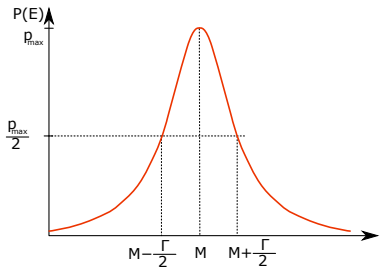


Figure 1.5: The Breit-Wigner distribution. From [5].

The decay rate, or *total width*, is the sum of the individual partial widths. The total width is related to the lifetime of the particle through $\tau\Gamma = 1$.

The fraction of decays through a single channel is given by the *branching ratio*, $\text{BR} = \Gamma_i/\Gamma$.

The photon propagator from before was proportional to $\frac{1}{q^2}$, where q^2 is the momentum transfer. Virtual exchange of massive force carriers, e.g. W and Z, must include the mass in the propagator:

$$\mathcal{M} \propto \frac{1}{q^2 - m_Z^2}. \tag{1.7}$$

Equation 1.7 is problematic. The equation will go to infinity in the limit $q^2 = m_Z^2$.

For unstable particles, their decay rate must be included as well. Doing the substitution $m \to m - i\Gamma/2$ into Equation 1.7 gives us the square root of the Breit-Wigner resonance (see also Figure 1.5): [3]

$$\mathcal{M} \propto \frac{1}{q^2 - m_Z^2 + im_Z\Gamma_Z}, \tag{1.8}$$

where Γ has become the Full Width at Half Maximum (FWHM) of the distribution, hence its name.

It is important to note that Γ is not a measurement uncertainty. The width comes from the Heisenberg uncertainty principle, $\Delta E\, \Delta(1/\Gamma) \geq 1/2$.

QCD

The local gauge symmetry group of QCD is the SU(3).[3]

The SU(3) group has 8 generators corresponding to the 8 gluons of the strong interaction. The SU(3) group introduces color charge which can have the value of R (red), G (green), or B (blue). While the quark will be in a color state of R, G, or B, gluons will be a mix of a color and anti-color.

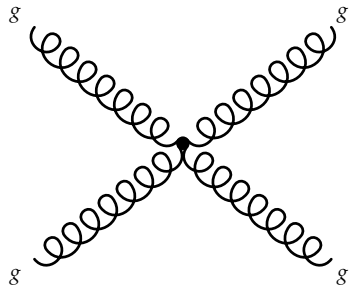


Figure 1.6: Gluon self-interaction.

Because of the gluons' non-zero color charge, they will self-interact. One example can be seen in Figure 1.6.

In QCD, the concept of *color confinement* has been introduced to explain the lack of detection of free quarks. Color confinement tells us that particles with non-zero color can not exist freely. The self-interaction of gluons is believed to be the cause of this phenomenon. Quarks interact with each other through the exchange of virtual gluons. The gluons in this exchange will also interact with each other. The greater the separation, the greater the number of self-interactions. This leads to a potential of the form

$$V(r) \propto r. \tag{1.9}$$

Trying to separate quarks will only add energy to the system. At one point, the energy added to the system allows for the creation of a new $q\bar{q}$ pair.

Because of color confinement, any quark combination must be colorless. The three common quark combinations into hadrons are

the mesons ($q\bar{q}$), baryons (qqq), and anti-baryons ($\bar{q}\bar{q}\bar{q}$). A meson is made of a quark of one color and an anti-quark of the corresponding anti-color. Mesons therefore exist in the following (normalized) superposition: [3]

$$\psi_c = \frac{1}{\sqrt{3}}(r\bar{r} + g\bar{g} + b\bar{b}), \tag{1.10}$$

where the subscript on the wavefunction denotes the color component of the full wavefunction for a meson.

The coupling strength of an interaction is, at first approximation, given by the coupling constant at the interaction vertex. However, higher-order corrections to the propagator will lead to an *effective coupling* $\alpha_S(q^2)$. For QCD, the effective coupling decreases at higher momentum transfers.

This leads to *asymptotic freedom*, in which the interaction between quarks almost vanishes are very high energies or very short distances.

Electroweak theory

In the electroweak theory, the gauge group is $U_Y(1) \otimes SU_L(2)$, where the generator of $U(1)$ is the B boson which carries *weak hypercharge* Y, and the generators of $SU_L(2)$ are W_1, W_2, and W_3 which carry the weak isospin.

The subscript L refers to the fact that the group only interacts with (sometimes explained as *feel*) the left-handed fermions.

The Higgs mechanism (explained later) will spontaneously break this symmetry into $U_{EM}(1)$ (electromagnetism) and $SU(2)$ (weak interaction).

Two of the Ws, e.g. W_2, W_3, form W^+ and W^-, while the last, W_1, will combine with B to form Z and γ through the following mixing:

$$\begin{pmatrix} Z \\ \gamma \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} = \begin{pmatrix} W_1 \\ B \end{pmatrix}. \tag{1.11}$$

θ_W is known as the Weinberg angle. [3]

From the above equation, Z and γ may appear to be very similar even if they were associated with two different interactions before the unification. Indeed, many processes can involve either one, and the propagator in the corresponding Feynman diagrams (see Figure 1.7) are therefore often denoted γ^*/Z, where the asterisk means that the photon is off-shell because it has a non-zero invariant mass. This is of course only possible because it is virtual.

The weak interaction is separated into the *charged-current* interaction (mediated by the W bosons) and the *neutral-current* interaction (mediated by the Z boson). The charged-current interaction couples to leptons within same-flavor doublets and quarks differing by one charge[4]. The neutral Z couples to two identical fermions, and will decay to a fermion and its anti-fermion.

The weak interaction couples differently to the different quarks.

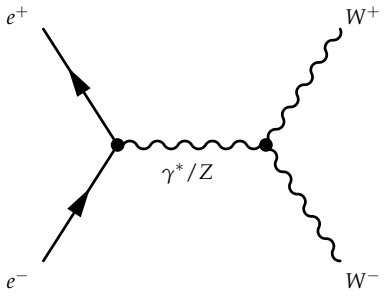


Figure 1.7: An example of an electroweak interaction.

[4] The W does, however, couple to up- and down-type quarks that are not within the same doublet.

This behavior is explained by the CKM matrix:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{1.12}$$

The *mass eigenstates* in the left vector are linear combinations of the *flavor eigenstates* of the right vector. This means that the W will actually couple to $\begin{pmatrix} u \\ d' \end{pmatrix}$, where d' is a mix given by the CKM matrix.

For example, because of V_{tb}, which has a value of above 0.99, the top quark will decay almost exclusively to Wb, because of the relation $\Gamma(t \to Wb) \propto |V_{tb}|^2$.

The branching ratios for the weak bosons are:

$$\begin{aligned} BR(W \to q\bar{q}) &\approx 2/3, \\ BR(W \to l\bar{v}) &\approx 1/3, \\ BR(Z \to q\bar{q}) &\approx 70\%, \\ BR(Z \to l\bar{l}) &\approx 10\%, \\ BR(Z \to v\bar{v}) &\approx 20\%. \end{aligned} \tag{1.13}$$

The bosons will decay approx. evenly to the three leptons.

1.1.4 The Higgs mechanism

If we try to add a mass term for the photon in the QCD Lagrangian, the Lagrangian is no longer gauge invariant.[3]

Local gauge symmetry therefore requires that the photon be massless, which is in agreement with observation. However, this also goes for the other Lagrangians. Since W and Z are massive, something is missing in the Lagrangian of the weak interaction.

The answer lies in the Higgs mechanism's *spontaneous symmetry breaking* of the electroweak gauge symmetry.

The Higgs mechanism includes two complex scalar fields in a weak isospin doublet,

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \tag{1.14}$$

in a potential of the form

$$V = \mu^2 \phi^\dagger \phi + \lambda \left(\phi^\dagger \phi\right)^2. \tag{1.15}$$

If $\mu^2 < 0$ (and $\lambda > 0$), the minimum of the potential will be a ring at a distance $v^2 = -\frac{\mu^2}{2\lambda}$ away from the Origin, as illustrated for a simpler field in Figure 1.8, where v is the *vacuum expectation value*.

The four ϕ's correspond to the four degrees of freedom needed to give mass to W^+, W^-, and Z. The last degree of freedom will give rise to the Higgs boson.

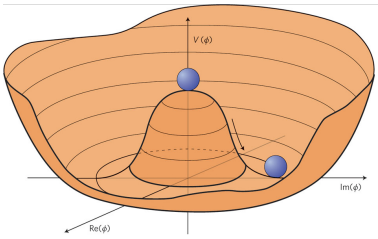


Figure 1.8: The The scalar potential for $\mu^2 < 0$ for a field $\phi = \frac{1}{\sqrt{2}}(\phi_1 + i\phi_2)$ in the potential given in the text. From [6].

1.2 Collider Physics

To study particle physics, an experimentalist needs two machines: One that creates the particles, and one that detects and measures them. For some experiments, like some neutrino experiments or cloud chambers, the Universe is the particle-creating machine.

However, this setup is very limiting, and we therefore need a machine for which we can control more of the conditions for particle creation. In particle physics, this is for example the circular collider.

The circular collider accelerates bunches of particles (e.g. electrons, protons, or heavy ions) to great energies in both directions through the ring and brings them to collision inside an experiment. For proton-proton collisions, the center-of-mass energy in the collision is denoted as $\sqrt{s}$, where s is the Mandelstam variable $s = (p_1 + p_2)^2$. p_1 and p_2 are the four-momenta of the two beams.

1.2.1 Coordinate system

The coordinate system used for a particle at analysis level is (p_T, η, ϕ, m), where p_T (or $p_\perp$) is the transverse (or perpendicular) component of the particle's momentum, η is the pseudorapidity, ϕ is the azimuthal angle, and m is the mass.

The coordinate system is drawn in Figure 1.9 in the zy-plane.

η is related to the angle θ through

$$\eta = -\ln[\tan(\theta/2)]. \tag{1.16}$$

For a light particle, or a particle moving near the speed of light, the pseudorapidity is approximately the rapidity. In other words, for $E \gg m$,

$$\eta \approx y = \frac{1}{2}\ln\left[\frac{E + p_z}{E - p_z}\right], \tag{1.17}$$

where E is the energy and p_z the momentum along the beam axis.

Rapidity is used in experimental particle physics, because differences in rapidity are Lorentz-invariant under boost in z, rapidity is of the same order as the angle ϕ, and the number of particles per y is approximately constant. Even though pseudorapidity is not Lorentz-invariant under boost, it is used instead of rapidity[5] because of its simple relation to the angle θ.

From the two spatial coordinates, η and ϕ, the distance between two particles in an event is defined as:

$$(\Delta R)^2 = (\Delta\eta)^2 + (\Delta\phi)^2. \tag{1.18}$$

1.2.2 Charged particle and photon interactions

Charged particles transversing a medium will be subject to several phenomena. The two main sources of energy loss we will consider are ionization and bremsstrahlung.

At very low energies ($\lesssim 7\,\text{MeV}$, see Figure 1.10), an electron will ionize the medium which greatly depletes the kinetic energy of the

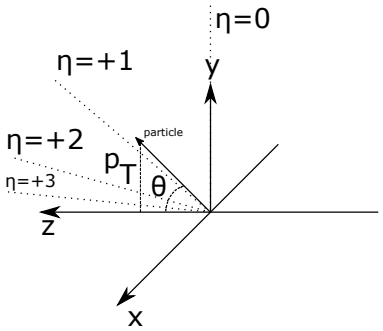


Figure 1.9: The coordinate system of ATLAS. The angle θ is related to the pseudorapidity through Equation 1.16. The x-axis points towards the center of LHC, the y-axis points upwards, and the z-axis is pointing along the beam axis. The angle ϕ (not drawn) is defined from the positive x-axis in the xy-plane.

[5] The rapidity must be used for massive objects,

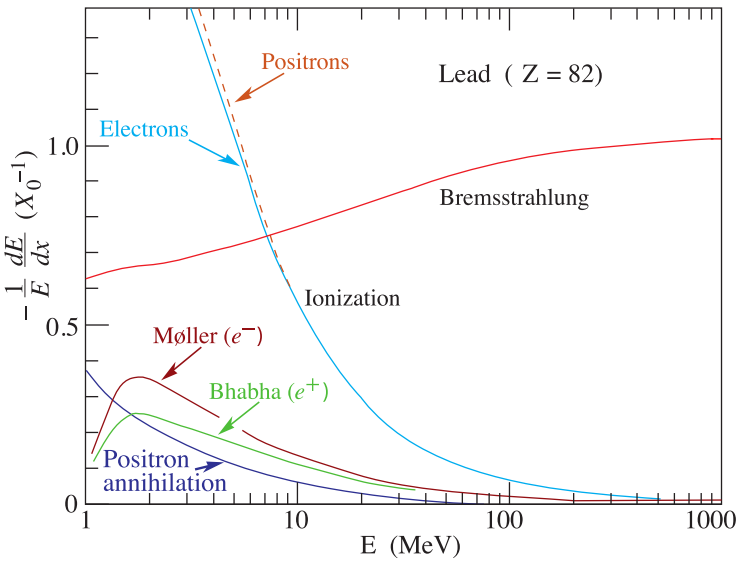


Figure 1.10: Fractional energy loss of an electron transversing lead. Adapted from [7].

electron. At higher energies, the electron will be deflected by the medium and release braking radiation, or *bremsstrahlung*.

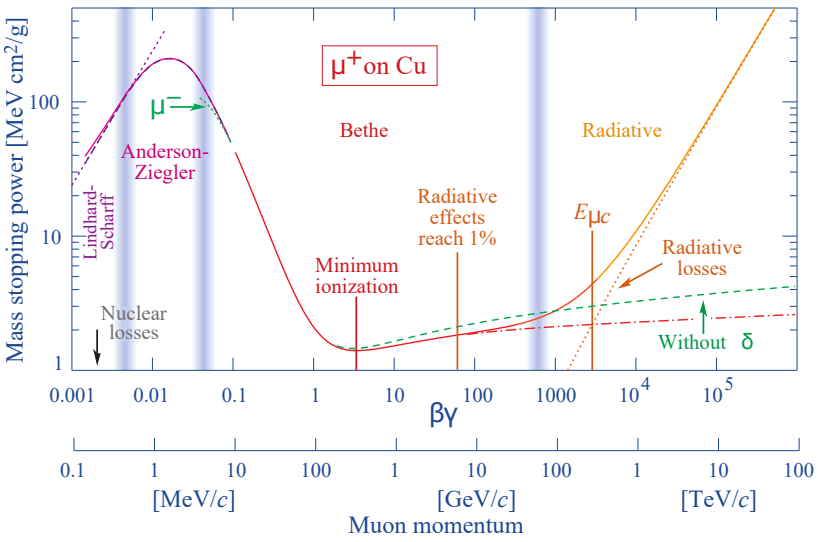


Figure 1.11: The Bethe-Bloch formula for a muon which shows the average energy loss at given energies. From [7].

Bremsstrahlung is suppressed by a factor of $\left(\frac{m_\mu}{m_e}\right)^2$ for muons, making the muons immune to bremsstrahlung for energies $\lesssim$ 1 TeV. This is seen in Figure 1.11.

Figure 1.11 depicts the Bethe-Bloch formula which gives the *average* energy loss of a particle (in this case the muon) at given energies. We see that muons with momenta in the range of a few tens of MeV to almost one TeV interact the least. Particles in this range are called *minimum-ionizing particles*, as they will transverse media (detectors) with little energy loss. This makes it near impossible to actually capture a muon in this energy range.

X_0 in Figure 1.10 is *radiation length*. The average fractional energy

loss of a particle is integrated to give: [4]

$$\langle E(x) \rangle = \frac{E_0}{X_0} \exp(-x/X_0). \tag{1.19}$$

The depths of calorimeters (detectors that measure the energies of particles) are measured in units of radiation lengths.

At energies above $2m_e$, photons will fluctuate into electron-positron pairs that annihilate back to photons. The presence of atoms' electric fields will allow the virtual process to become real, and the photons will therefore become real electron-positron pairs[6].

These electron-positron pairs will radiate bremsstrahlung photons which in turn will convert to pairs themselves.

Thus, when entering a dense medium, photons and electrons will create an *electromagnetic shower* of photons and electrons.

[6] This can happen at any time after the creation of the photon. Furthermore, because of the photons' pair production, electrons and photons will be almost indistinguishable in a detector that does not distinguish between charged and neutral particles. This can create an ambiguity between electrons and photons if the photon converted before entering the detector.

1.2.3 Parton Distribution Functions

Scattering particles off of an atom will reveal the internal structure of the atom, if the energy of the incoming particles is sufficient. This is what led Rutherford to discover the nucleus of the atom.

If we increase the energy of our probing particles, we will be able to probe the internal structure of the nucleus to find protons and neutrons. In deep scattering, particles at even greater energies probe the internal structure of the proton. This is how quarks were discovered.

At the current energy scale at which we are able to probe the quarks they seem not to reveal any substructure. They are point-like particles exactly like the electron and all the other fundamental particles.

However, as seen in the bottom section of Figure 1.12, the simple model of the proton with two up-quarks and one down-quark is insufficient. At higher energy scales, the proton will contain many virtual quarks and gluons. The three *valence quarks* of the proton will radiate gluons which split into quark pairs that we call *sea quarks*.

In collider experiments, when two protons collide, it is actually a collision between their *partons*[7]. Each parton will draw a fraction of its parent proton's momentum. This gives us a four-momentum of:

$$p_i = x_i p_{\text{proton}}(1, 0, 0, \pm 1), \tag{1.20}$$

where the sign of the last component is the direction of the beam.

If the center-of-mass energy is $\sqrt{s}$, each *proton* will carry half. We can therefore write the total energy Q (or invariant mass M) of the *parton-parton* system in terms of x.

$$Q^2 = M^2 = (p_1 + p_2)^2 = s x_1 x_2. \tag{1.21}$$

Q^2 is what was previously referred to as the "energy scale".

We can now introduce the parton distribution function (PDF) which gives the probability for extracting a given parton with a fraction x of the proton's momentum. This is what is illustrated in Figure 1.13.

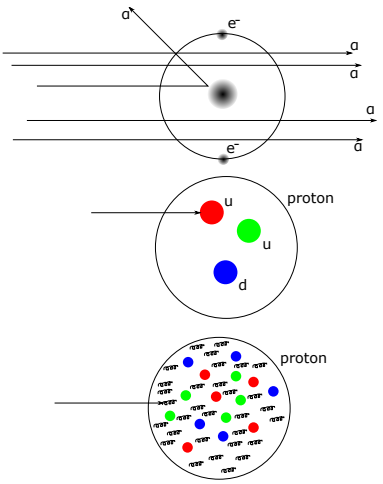


Figure 1.12: Top section: Alpha particles scattering off of the nucleus of an atom. Middle section: A particle probing the internal structure of a proton and coming in contact with one of the proton's valence quarks. Bottom section: A more detailed view of the internal structure of a proton with virtual gluons and sea quarks. This picture is especially important when the proton is accelerated to high energies as explained in the text.

[7] A parton is a quark or gluon. The name is coined by Feynman and is made from "part of hadron".

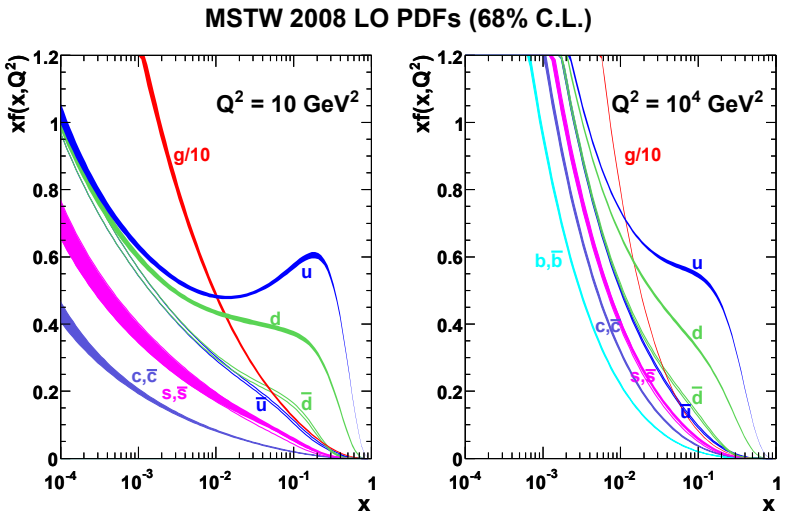


Figure 1.13: The parton distribution functions at two energy scales. From [8].

The figure shows a set of PDFs for each type of quark as well as the gluon at two energy scales. We note a few things:

- The up-quark is twice as probable as the down-quark as is expected for a proton.
- The gluon is divided by 10; in other words, the gluon is massively abundant in the proton, and the extracted parton in a collision is therefore most often a gluon.
- The second axis has the PDF multiplied by the momentum fraction ($xf(x, Q^2)$) to keep the figures in the plot. Extracting a low-momentum parton is several magnitudes times more probable.
- At higher energy scales, the probability to extract a sea quark (strange-, charm-, or bottom-quark) is greatly increased to the point where they are almost as likely as the valence quarks to be extracted at lower x.
- Related to the above, extracting anti-quarks become almost as probable as a quark. For the LHC, colliding two protons will yield a substantial amount of anti-quarks. Colliding protons and anti-protons are therefore not needed.

1.2.4 Factorization theorem

Imagine the case where two colliding protons create a particle X. We can calculate the creation of X from two quarks q_i and q_j, and we know from the PDFs how to extract the quarks q_i and q_j from the protons. The cross-section for the process $\sigma_{q_i q_j \to X}$ is independent of the PDFs. This is called the factorization theorem.

The cross-section for creating a particle X in a collision is:

$$\sigma_X = \underbrace{\sum_{q_i,q_j} \int_0^1 dx_1 dx_2 f_{q_i}(x_1, Q^2) f_{q_j}(x_2, Q^2)}_{\text{Extraction}} \cdot \underbrace{\sigma_{ab\to X}(Q^2)}_{\text{Interaction}}. \tag{1.22}$$

We have factorized the production of X in a collision into the two extractions of partons q_i and q_j via their PDFs and the production of X from q_i and q_j.

1.2.5 Luminosity

In the previous section, the differential cross section was introduced in the form of particles hitting a stationary target. In circular colliders, it is two beams colliding with each other, and we have to tweak the definition a little:

$$\sigma = \frac{N}{L\epsilon}, \tag{1.23}$$

where N is the number of observed events, L is the *integrated luminosity*, and ϵ is the total product of all efficiencies and acceptances.

For the number of observed particles N of X given the amount of data L, the production probability of X is σ.

The (instantaneous) luminosity, $L(t)$, gives the event rate.

The recorded amount of luminosity in 2015 for ATLAS can be seen in Figure 1.14.

1.2.6 Pileup and soft processes

A proton-proton collider is actually a parton-parton collider. The collisions happen between the quarks and gluons of the protons. A hard process is a collision between two partons that yields a high transverse momentum output of particles.

Most collisions will be "soft", that is low in p_T. Those include beam remnants, the decaying products of protons from which the hard partons came, and initial- and final-state radiations from the hard process.

In a collision, multiple protons will collide at a time. Most of them will be soft processes. The multiple interactions per bunch crossing (see Figure 1.15) are collectively called the *pileup* of the event.

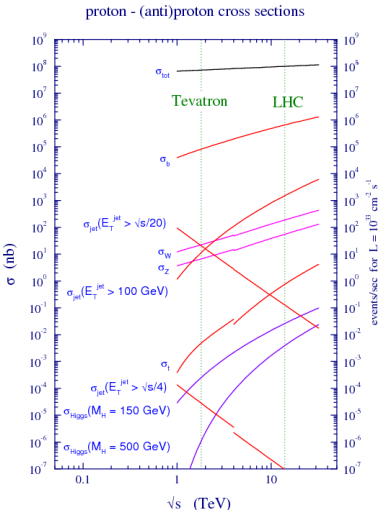


Figure 1.14: Cross sections as a function of center-of-mass energy. The total cross section of LHC is approx. 100 mb, which is one billion times the cross section of the Higgs particle! From [9].

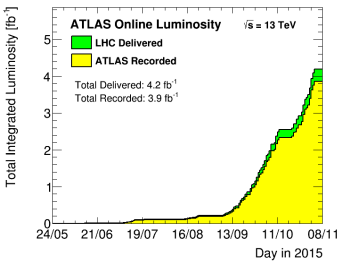


Figure 1.15: The integrated luminosity for 2015. By ATLAS Data Preparation Group.

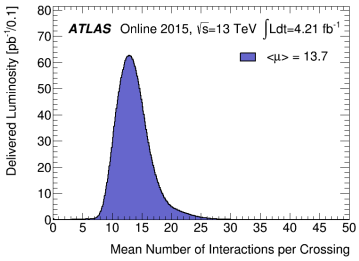


Figure 1.16: The number of interaction per crossing for 2015. By ATLAS Data Preparation Group.

1.3 The ATLAS detector

The ATLAS detector experiment is one of the four large experiments that reside on the accelerator known as the Large Hadron Collider, or the LHC. The LHC is the latest and largest accelerator installation at CERN to date. It accelerates both protons and heavy ions in both directions through the ring to great energies, and brings them to collision at one of the four experiments.

The three other large detector experiments at LHC are ALICE, CMS, and LHCb. ATLAS and CMS are general detectors. ALICE is specialized in the study of heavy ion collisions. LHCb focuses on b-physics. ATLAS also studies proton-lead and lead-lead collisions.

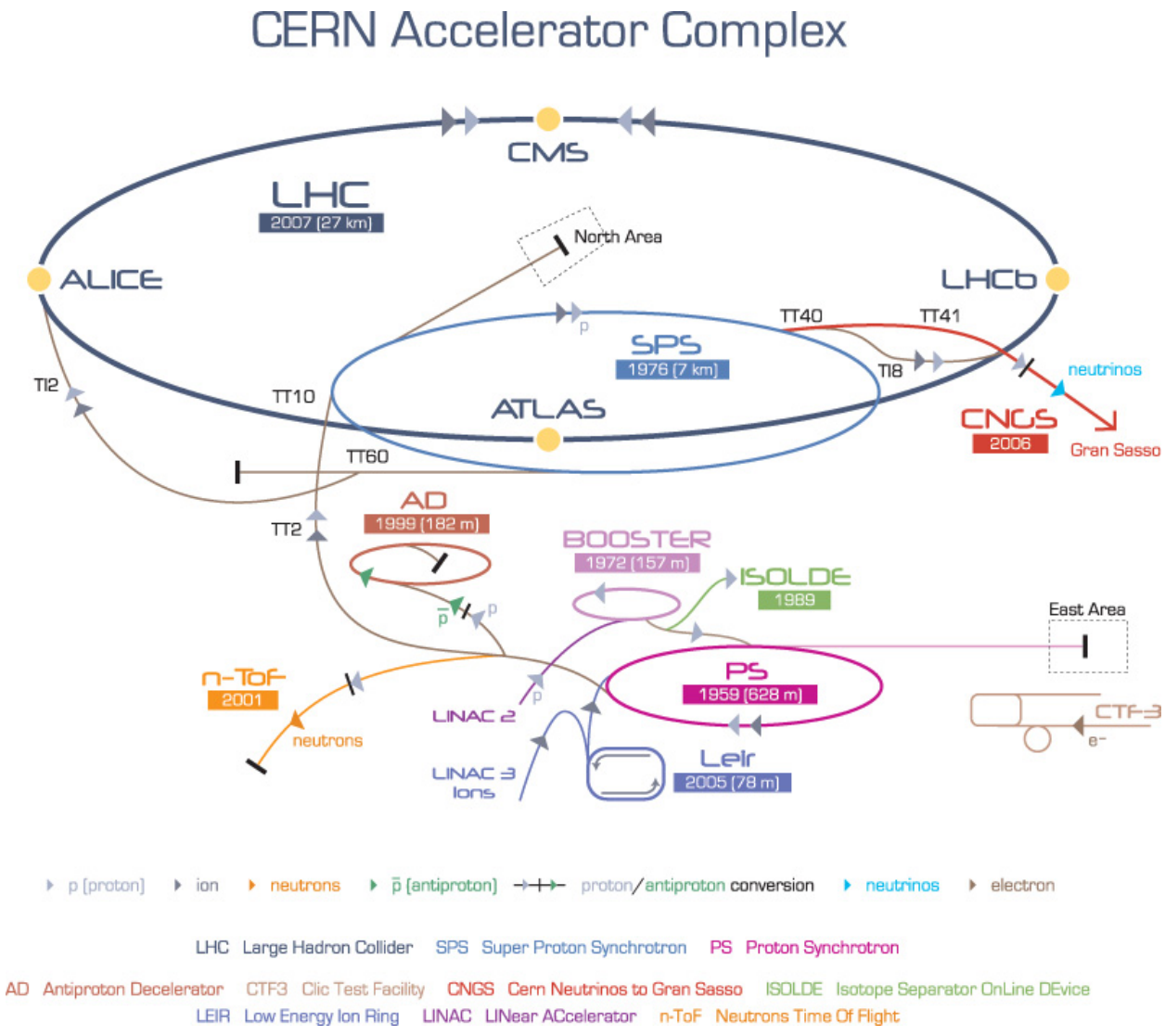


Figure 1.17: The CERN Accelerator Complex as of 2008. From [10].

The complete figure of the accelerators at CERN as of 2008 is shown in Figure 1.17. Only the large of the currently running experiments are shown in the figure, e.g. the experiments served with antiprotons by the Antiproton Decelerator (AD) are not shown.

The particles that eventually collide in ATLAS begin their journey in LINAC 2 (for protons) or LINAC 3 (for heavy ions). The LINAC accelerators are linear accelerators. The source of the protons is a hydrogen bottle in LINAC 2 from which the electrons are stripped. The protons are accelerated to 50 MeV into the Proton Synchrotron Booster. The booster further accelerates the protons to 1.4 GeV and delivers them to the PS (The Proton Synchrotron). This goes on for the PS (26 GeV) and SPS (The Super Proton Synchrotron) as well (450 GeV), until the protons have reached energies that the LHC is able to handle. The LHC will then finally accelerate the protons up to 6.5 TeV each, resulting in a center-of-mass energy of 13 TeV when they collide.[11]

The collisions happen every 25 ns. In each collision a bunch of $\mathcal{O}(11)$ protons are crossing from each direction. When the LHC is running, it has 2808 number of bunches in its ring. This brings the luminosity of the LHC to the order of: [4]

$$10^{34}\,\mathrm{cm}^2\,\mathrm{s}^{-1} = L \propto n_b N_b^2 f_{\mathrm{rev}}, \tag{1.24}$$

where n_b is the number of bunches, N_b is the number of protons per bunch, and f_{rev} is the revolution frequency of $1/(25\,\mathrm{ns}) = 40\,\mathrm{MHz}$.

The remainder of this section is based on [12].

1.3.1 Overview of the ATLAS detector

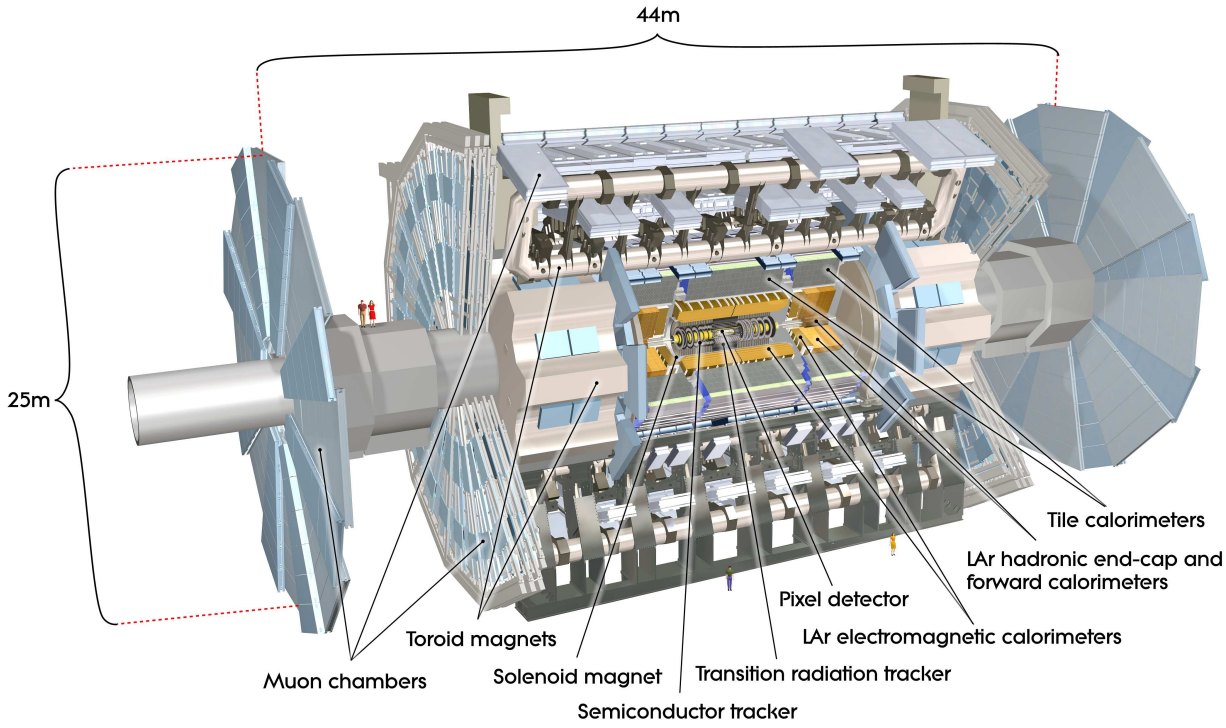


Figure 1.18: The ATLAS detector. The two people in the left side are drawn to scale. From [12].

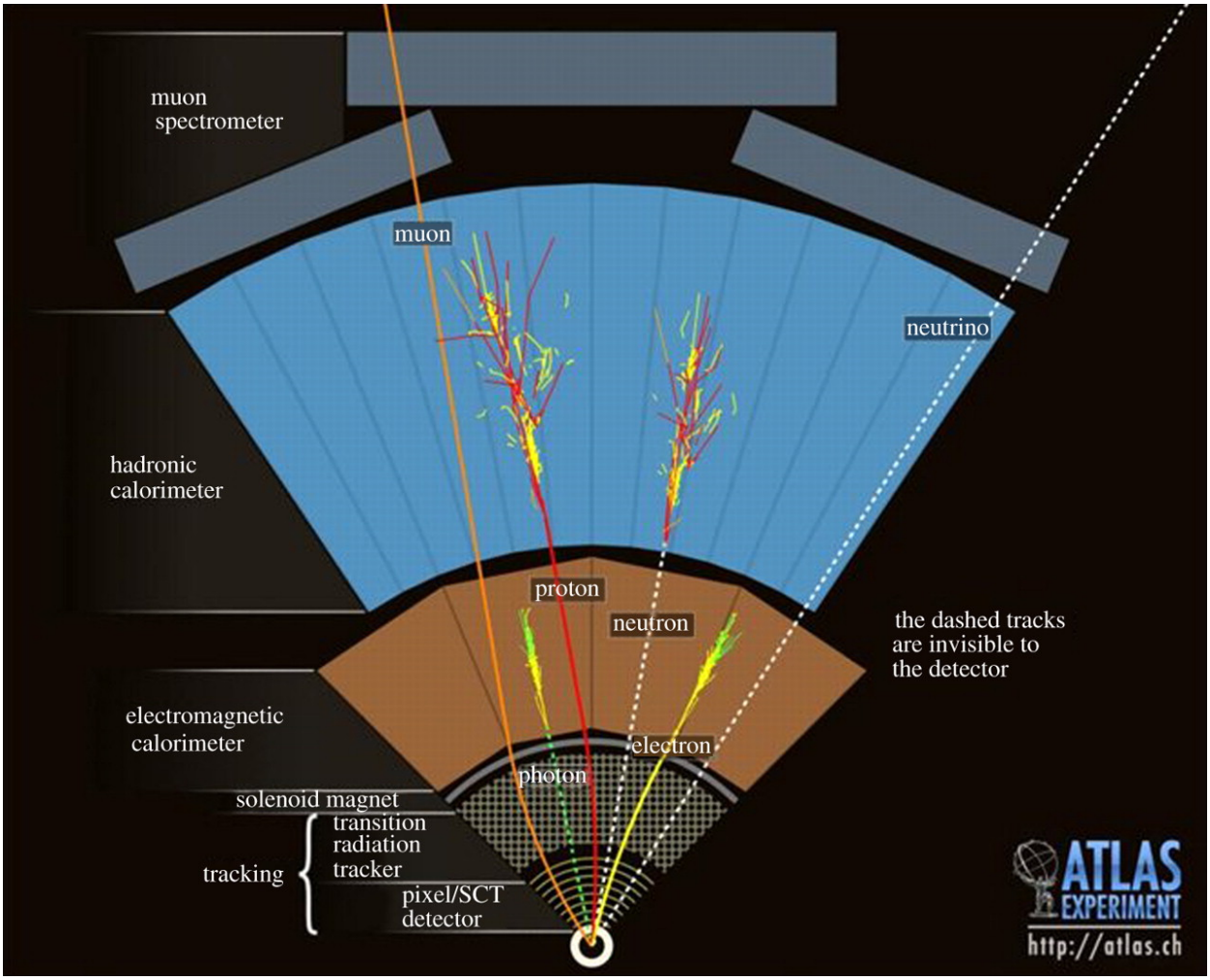


Figure 1.19: The ATLAS detector in a schematic 2D-view with the interactions of common particles. The proton in the image appears (incorrectly) not to deposit energy in the electromagnetic calorimeter.

Many of the particles, be they fundamental or composite, formed after a collision decay before they even reach the inner detector.

What is mainly detected are the following (see also Figure 1.19):

- electrons and photons in the electromagnetic calorimeter with an associated (or lack thereof) track in the tracking detectors discriminating the two,
- charged (mostly pions, though also kaons and protons) and neutral (neutrons, kaons) hadrons in both the electromagnetic and hadronic calorimeter with an associated (or lack thereof) track in the tracking detectors discriminating the two,
- muons in the tracking detectors and muon spectrometer and weakly in the calorimeters.

1.3.2 Inner detector

The inner detector (ID) is actually a set of several sub-detectors surrounded by a solenoid magnet generating a 2 T magnetic field. The inner detector will mainly be used for tracking charged particles but does also offer particle identification of some charged particles.

Due to the high number of particles created in each collision, a high granularity is needed to separate all the tracks. A strong magnetic field is needed to bend the high-p_T charged particles that can be created because of the high center-of-mass energy.

The ID covers $|\eta| < 2.5$ for the pixel and SCT, while the TRT only covers $|\eta| < 2.0$.

All the sub-detectors are installed in the barrel and end-caps.

The Pixel detector

The pixel detector consists of a barrel part that are three concentric cylinders around the beam pipe and three concentric disks in both end-caps.

Since the pixel is the module closest to the beam pipe, its resolution demand is also the greatest. Therefore, it consists of approx 80.4 million silicon pixels offering an accuracy of $10 \times 115\,\mu m^2$ in $R - \phi \times z$ ($R - \phi \times R$) in the barrel (end-caps).

For Run 2, the Insertable B-Layer (IBL) has been installed as a pixel layer in front of the three already installed pixel layers. It has been installed in anticipation of the increased luminosity, and therefore number of interactions, to deal with the increased number of tracks. It will also improve b-tagging.

The SCT

The Silicon Microstrip Tracker (SCT) consists of 4 concentric cylinders around the pixel detector in the barrel as well as 9 concentric disks in each end-cap.

In each layer of the SCT, pairs of silicon strips angled at 40 mrad with each other provide the detection of charged particles.

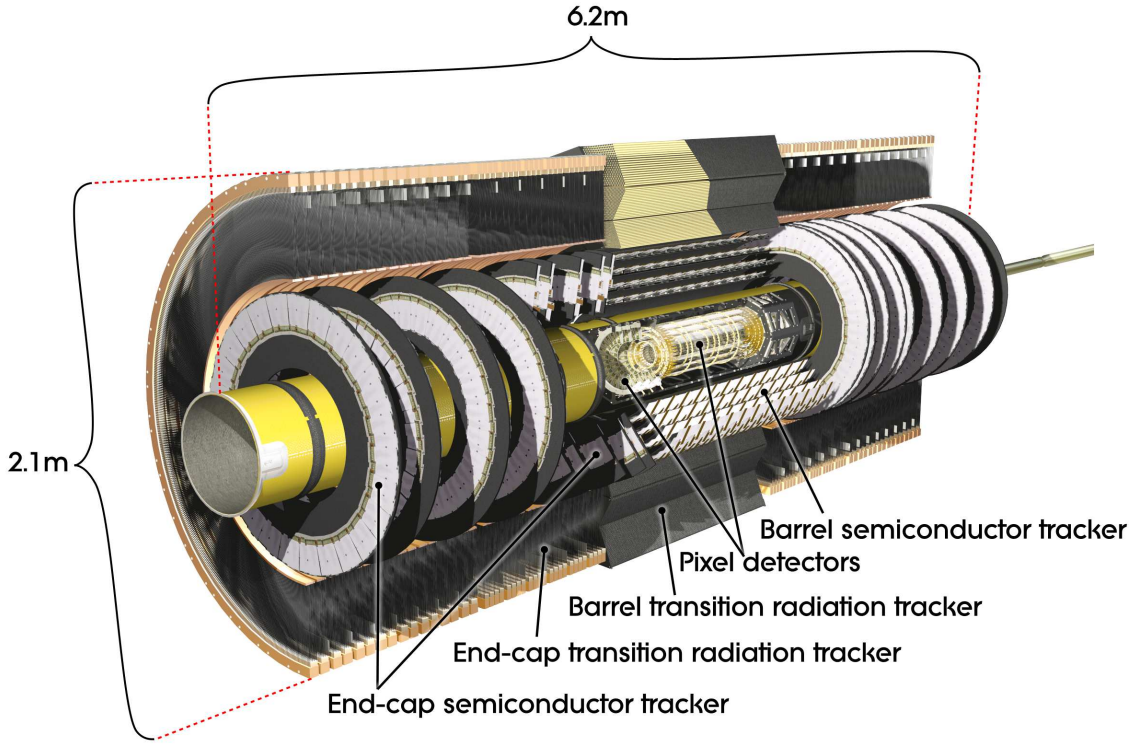


Figure 1.20: The Inner Detector of ATLAS. From [12].

The accuracy of the SCT is $17 \times 580\,\mu m^2$ in $R - \phi \times z$ ($R - \phi \times R$) in the barrel (end-caps).

The TRT

The Transition Radiation Tracker (TRT) consists of 73 layers of tubes in the barrel and 160 in each end cap. The TRT straws provide no information in η as the straws are parallel to the beam in the barrel and aligned radially from the beam pipe in the end-caps.

Each tube consists of various gases and a straw in the middle. When a charged particle traverses the tube, it will ionize the gases. The *drift electrons* will drift towards the straw that is kept at a potential of 1530 V.

The TRT tubes initially contained 70% xenon which would give transition radiation from electrons, thus providing particle identification information. Due to leaks, parts of the TRT has had the xenon replaced with argon as xenon is too expensive to refill. The tubes also contain approx. 28% CO_2 and approx. 2% O_2 to aid the process.

The TRT will only be able to give a distance between the straw and the incident particle, which will create so-called *drift circles* for which there is an inherent left-right symmetry.

However, the many layers of straws and greater distance to the beam pipe will result in a low momentum uncertainty. On average, a track will leave 36 hits in the TRT which will also resolve any left-right ambiguity of the drift circles.

1.3.3 The calorimeter

The calorimeter also consists of several parts. In the barrel, the electromagnetic calorimeter (ECAL) surrounds the solenoid magnet followed by the hadronic calorimeter (HCAL).

In the end-caps, there is also an ECAL and HCAL as well as a forward calorimeter (FCAL).

The ECAL and HCAL span $|\eta| < 3.2$, while the FCAL spans $3.1 < |\eta| < 4.9$.

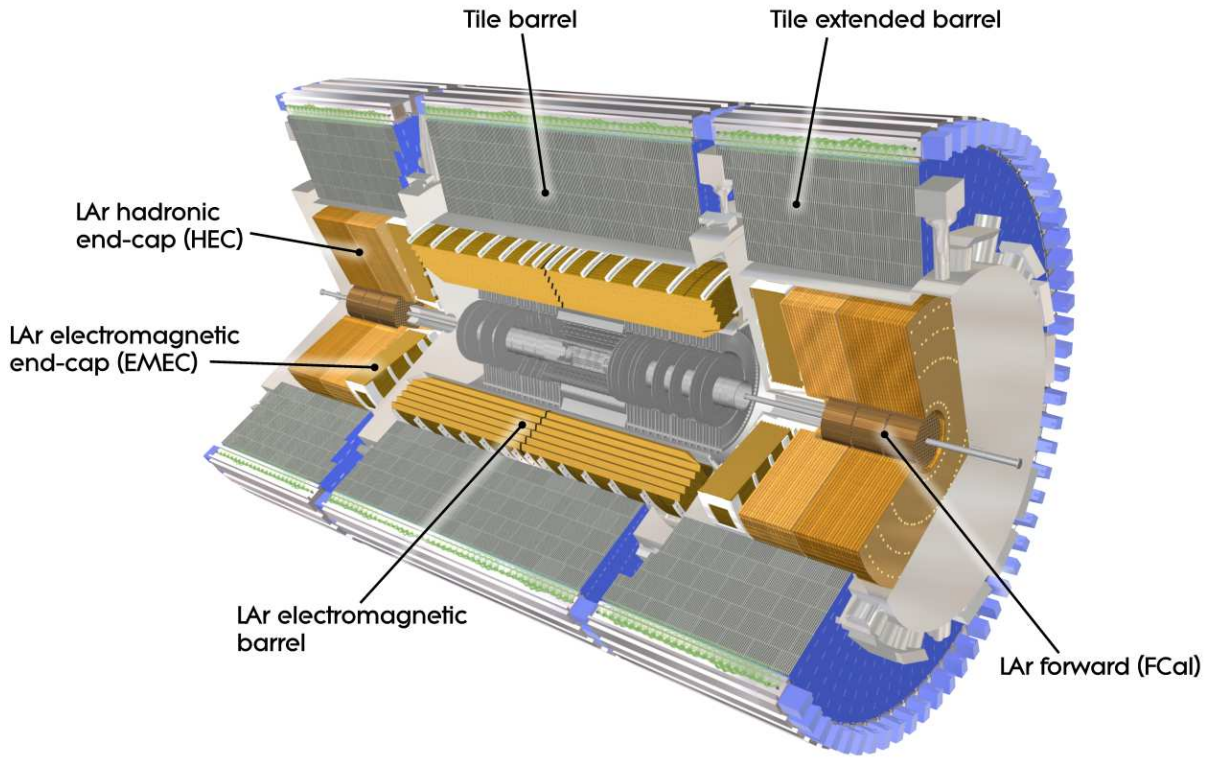


Figure 1.21: The calorimeters of ATLAS. From [12].

The electromagnetic calorimeter

The ECAL is a sampling calorimeter with lead as the passive material and liquid argon (LAr) as the sampler. As seen in Figure 1.22, the ECAL is accordion-shaped, so any transversing particle would not slip through the otherwise existing cracks of the calorimeter.

The ECAL consists of three sections in the barrel. The first section consists of strips that only have a resolution in η, though at $1/8$ of the second section. The second section has a resolution of maximum 0.025×0.025 in $\Delta\eta \times \Delta\phi$.

The resolutions for the end-caps in $|\eta| < 2.5$ are similar to the second section of the barrel ECAL.

The ECAL is more than 22 radiation lengths deep in the barrel and more than 24 in the end-caps.

The ECAL also has a presampler in $|\eta| < 1.8$ that measures energy

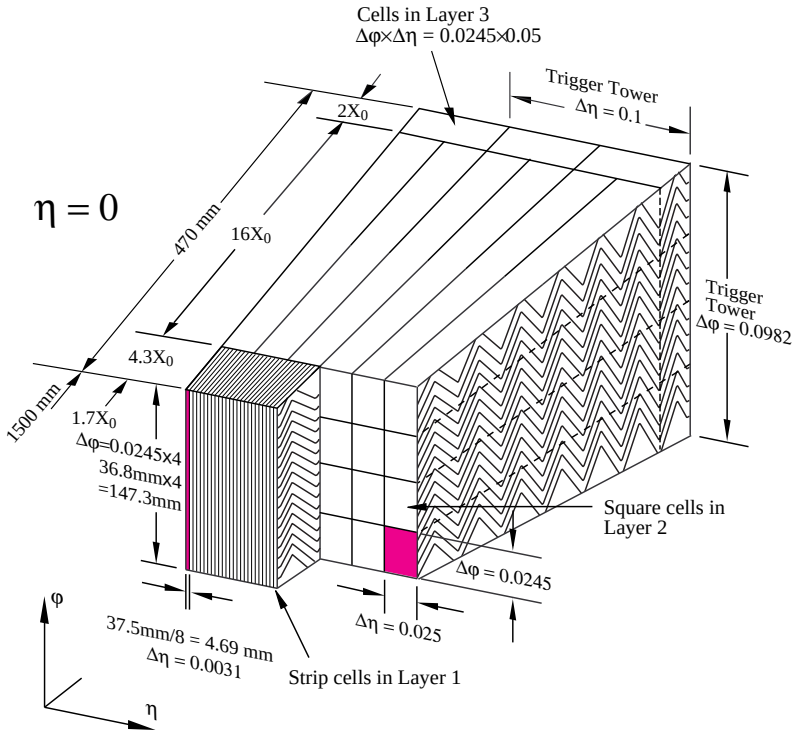


Figure 1.22: A section of the barrel ECAL. From [12].

lost in the material (magnet, cooling system, scaffolding) right in front of the ECAL.

The hadronic calorimeter

The barrel HCAL is a scintillator tile calorimeter which is also a type of sampling calorimeter like the ECAL. The passive material is steel while sampling material is scintillating tiles that are read out by wavelength shifters connected to PMTs.

The HCAL's end-caps consists of two wheels on either side segmented into two sections, effectively making four layers on each side. The end-caps are copper-LAr sampling calorimeters.

The forward calorimeter

The FCAL has a granularity of up to 5.4×4.7 and is only approx. 10 radiation lengths deep. It is also a sampling calorimeter with LAr as the active material.

Each end-cap consists of three modules. The first module has copper as the passive material and is used for measuring electromagnetic showers, while the two others use tungsten as the passive material and are used for the hadronic measurements.

1.3.4 Muon spectrometer

The muon spectrometer (MS) is the outermost detector of ATLAS. It covers $|\eta| < 2.7$ for measurements and $|\eta| < 2.4$ for triggering.

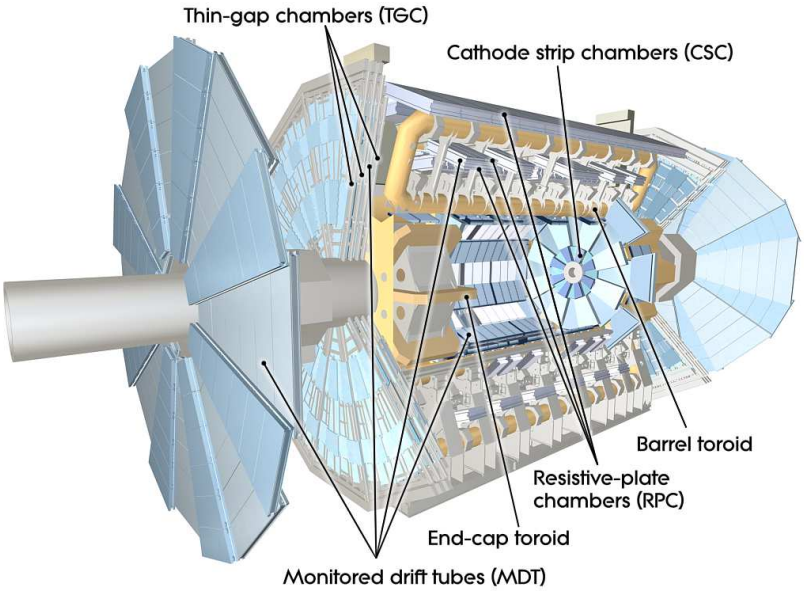


Figure 1.23: The muon spectrometer of ATLAS. From [12].

In the barrel, it consists of several sub-detectors surrounded by a magnet providing a 0.5 T toroidal magnetic field. In the end-cap the magnets provide a 0.5 T toroidal magnetic field.

The precision measurements are performed by the Monitored Drift Tubes (MDTs) aided by the Cathode Strip Chambers (CSCs) in the end-caps. The Resistive Plate Chambers (RPCs) in the barrel and Thin Gap Chambers (TGCs) in the end-caps provide coarse measurements as well as triggering information.

The MDT chambers cover $|\eta| < 2.7$, except for the first layer only covering $|\eta| < 2.0$, and consists of up to 8 layers of drift tubes that provide an average resolution of 35 µm per chamber in z and no measurement in ϕ.

The CSCs ($2.0 < |\eta| < 2.7$) are mutliwire proportional chambers. They provide a resolution of 40 µm $\times$ 5 mm in $R \times \phi$ in the end-cap.

The RPC ($|\eta| < 1.05$) and TGC ($< 1.05 < |\eta| < 2.7$(2.4 for triggering)) also provide the ϕ coordinate that the MDT is missing.

1.3.5 ATLAS trigger system

Whether to keep an event is determined by the *trigger*. It consists roughly of two systems, L1 and HLT, that perform the decision in two steps.

The level 1 trigger (L1) is a hardware (analog) trigger with limited amount of detector information available. The L1 trigger will bring the event rate down to approx. 100 kHz (75 kHz in Run 1).

At L1, only muon and coarse calorimeter information is available. The trigger will look for high p_T particles, jets, and missing energy.

The High Level Trigger (HLT) is a software (digital) trigger that has the full event information available from which it can reconstruct tracks as well as use the additional information to make its final

decision. The HLT brings the event rate down to approx. 400 Hz (200 Hz in Run 1).

1.3.6 A note on resolutions

The SCT offers an accuracy similar to the pixel detector, but the more layers compared to the pixel detector and the longer distance from the interaction point greatly reduces the relative uncertainty of its momentum measurement[8]. This makes the SCT the most important sub-detector for charged particle tracking. This also goes for the TRT, which can offer an approximate resolution due to the on average 36 hits and much longer distance to the interaction point, even if the single hit accuracy is much lower.

[8] The relative uncertainty goes like $\sigma_{p_T}/p_T \propto p_T/(BL^2)$, where B is the magnitude of the magnetic field, and L is the distance.[4]

As the inner detector determines the momentum of a track from the bending of the particle, naturally the relative uncertainty grows with increasing p_T. The calorimeters, on the other hand, will see a decrease in their relative uncertainty with increasing E_T that goes like $1/\sqrt{E}$.

The point at which the calorimeter uncertainty falls below the ID's is at approx. 40 GeV.

The ID, however, has a much greater spatial resolution.

For these reasons, particles in ATLAS are usually reconstructed with η and ϕ from the ID. For muons, p_T is reconstructed from the ID and MS. For hadrons, the p_T (E_T) is reconstructed from the ID (calorimeters for $E_T \gtrsim 40\,\text{GeV}$). For electrons and photons, E_T is mainly reconstructed from the calorimeters.

The resolutions are shown in Table 1.1.

Table 1.1: Relative momentum resolutions for the main sub-detectors.

Sub-detector	Resolution
ID	$\frac{\sigma_{p_T}}{p_T} = 0.05\% \cdot p_T + 1\%$
ECAL	$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} + 0.7\%$
HCAL	$\frac{\sigma_E}{E} = \frac{50\%}{\sqrt{E}} + 3\%$
FCAL	$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}} + 10\%$
MS	$\frac{\sigma_{p_T}}{p_T} = 10\%$ at $p_T = 1\,\text{TeV}$

1.4 Reconstruction and software

1.4.1 Track reconstruction

Tracks are extrapolated lines between hits in the three inner detectors as well as the muon spectrometer. Only charged particles leave tracks as only they ionize the silicon trackers and TRT gas, etc.

A p_T cut is applied to reduce reconstruction time to remove what is mostly pile-up and has a high uncertainty. Therefore, any track with $p_T < 0.5\,\text{GeV}$ is discarded. Charged particles with $p_T < 450\,\text{MeV}$ will spiral around in the detector and not leave the TRT.

The track reconstruction begins with a seed. The seed is the first set of points that mark the beginning of the reconstruction. Because of the higher single hit precision of the pixel and SCT detectors, hits in them are used as seeds for the tracks that will later be extrapolated into the TRT.[4]

The reconstruction procedure used in ATLAS is the Kalman filter. The Kalman filter is an iterative method that propagates the track from layer to layer through the SCT, recalculating the position and direction as well as the covariance matrix at every layer. At every layer, the calculated track is compared to the nearest hit, and a decision is made whether to keep the point.

The extrapolation continues into the TRT.

When the track is reconstructed, the algorithm is then rerun in the opposite direction. Now starting in the TRT, the track is refitted given the full covariance matrix.

When all tracks have been identified this way, the algorithm starts in the TRT and backtracks to find particles not found in the silicon. This could be long-lived particles that may have decayed inside the inner detector or photons that have converted inside the ID.

Track coordinate system

The coordinate system for describing tracks is not the same as used for analysis. The tracks parameter are not measured in relation to the origin. The origin is defined as center of the detector in which the collisions happen.

The tracks are defined at their *perigee* which is the point at which the track that is closest to the beam line.

The perigee parameters are: [4]

- $\boldsymbol{d_0}$: The shortest distance from the track to the z-axis. This is known as the transverse impact parameter.
- $\boldsymbol{z_0}$: The z-coordinate. It is known as the longitudinal impact parameter.
- $\boldsymbol{\phi_0}$: The usual azimuthal angle in the xy-plane.
- $\boldsymbol{\theta}$: The usual polar angle in the zy-plane.
- $\boldsymbol{q/p}$: The charge over the magnitude of the momentum.

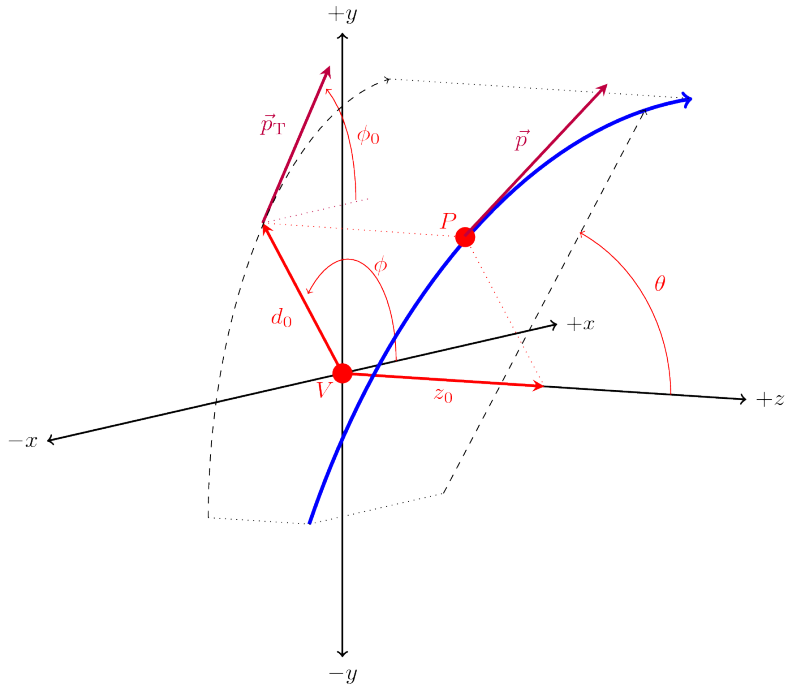


Figure 1.24: The perigee parameters. From [13].

The charge of particle is then the sign of q/p. This of course assumes that the charged particle is only either $+1$ or -1. Amongst the charged particles in the Standard Model, only the quarks are not charged either $+1$ or -1, but they will form jets of particles that do have integer charges. Any composite particle (e.g. Δ^{++}) will quickly decay to particles with charges $+1$, 0, or -1.

Track-vertex association

The reconstructed tracks are extrapolated to the beam line. Any crossing tracks will form a vertex. The average number of interactions in a single bunch-crossing were approx. 14 in 2015 (see Figure 1.16). Those are 14 individual initial vertices to which we have extrapolated our reconstructed tracks.

The primary vertex is the one with the highest $\sum_{i \in \text{tracks}} p_{\text{T}i}^2$. It is the only vertex on the beam line in the event that most analyses will use.

The other vertices are pile-up, but must be just as well reconstructed, so their tracks can be removed from the event during analysis.

1.4.2 Cluster reconstruction

No physics object reconstructed within the calorimeter uses the individual cells directly. Instead, EM clusters (in case of electrons in the ECAL) or topoclusters (in case of jets) are created from the cells. This reduces contributions from noise and greatly reduces reconstruction time for jet algorithms.

Clusters of energy deposits are created from 3×7^9 cells in $\Delta\eta \times \Delta\phi$ from the middle layer of the barrel ECAL[14]. All cells in the longitudinal direction of the cluster are summed to create towers. The EM clusters are then seeded by towers with $E_T > 2.5\,\text{TeV}$[15].

The jets used in this study are calorimeter jets constructed from topo-clusters. The topo-cluster is generated from a seed cell that has a signal 4 standard deviations above noise levels. The neighboring cells must have a signal 2 sigma above noise to be included in the seed. Topo-clusters are then formed from those cells[16].

[9] In Run 1, different sizes were used for electrons and photons.

1.4.3 Electron reconstruction and identification

Electron *candidates* are identified by a shower in the ECAL following a track in the ID[17]. This also means that *central electrons* are only reconstructed within the ID's $|\eta|$ range of 2.5. Electrons outside of this range are reconstructed as *forward electrons*, but these will not be used in this study.

EM clusters are initially formed in the ECAL, and tracks from the ID are extrapolated into the ECAL. If a track is within the cluster, the match is considered an electron candidate.

EM clusters without tracks are considered photons.

Care is taken to distinguish between prompt electrons, converted photons, electron bremsstrahlung, and other effects.

With the electron container filled with electron candidates, *electron identification* is performed. A series of discriminating variables (leakage into the HCAL, shower shapes, number of hits in the three IDs, etc.) are calculated, and three categories are created for use in analyses.

The three cuts (loose, medium, and tight) increase background rejection at every step, and have signal efficiency and background rejection rate pre-computed for easy comparison between analyses.

1.4.4 Muon reconstruction and identification

Muons candidates are identified in various ways: [18]

- **Combined muons:** Tracks are reconstructed in the MS and extrapolated into the ID to match its reconstructed tracks.
- **Stand-alone (or extrapolated) muons:** Tracks reconstructed in the MS that do not match an ID track, however the extrapolated track must loosely match in interaction point.
- **Segment-tagged:** Tracks reconstructed in the ID are extrapolated into the MS and match only one layer in the MS.
- **Calorimeter-tagged muons:** A track in the ID that matches a deposit in the calorimeter compatible with a minimum-ionizing particle like the muon is identified. This is mainly for muons in $|\eta| < 0.1$ which is the MS crack.

The muons also have identification cuts (loose, medium, tight). The tight cut level (which will be used in this study) requires that the

muon candidate is identified as a combined muon.

1.4.5 Jet reconstruction

Because of the strong interaction's color confinement, quarks created in the collision as well as quarks from decayed particles will form jets which are collections of hadronized quarks.

The energetic quarks will initially radiate gluons that split into quarks, and all these quarks will combine to form hadrons. The jet will have neutral and charged hadron components, and they will also leave energy deposits in the ECAL before showering in the HCAL.

The three popular jet algorithms in ATLAS are all based on the following formula:

$$d_{ij} = \min\left(p_{Ti}^{2p}, p_{Tj}^{2p}\right)\frac{\Delta R_{ij}}{R_0}, \tag{1.25}$$

where i and j are topo-clusters, ΔR is the distance between them, and R_0 is approx. the size of the jet.

The parameter p determines the algorithm. k_t-algorithm uses $p = 1$, the Cambridge-Aachen algorithm uses $p = 0$, and the anti-k_t-algorithm uses $p = -1$.

The algorithms will determine d_{ij} and $d_{kB} = p_{Ti}^{2p}$ (k being any topo-cluster, and B being the beam). They will combine the [four-momenta of the] topo-clusters i and j into a single topo-cluster continuously, until $d_{kB} > d_{ij}$ is satisfied for a given topo-cluster. The topo-cluster k is then called a jet.

This is repeated until all topo-clusters have been combined into jets. The reconstruction is performed by the FastJet algorithm[19].

All the topo-clusters that make up each jet are saved as a collection of *jet constituents*.

The anti-k_t-algorithm clusters jets around hard (high-p_T) particles and performs best of the three against pile-up effects.

The usual R_0 sizes for regular jets are 0.4 and 0.6. Larger values of R_0 (e.g. 1.0) are used for so-called *fatjets*. They will be used in this study to capture both jets formed in the $W \to qq$ decays, as the decay products of W particles with high p_T will merge.

Fatjets will still contain a substructure resembling two subjets, even at high transverse momenta. This will be exploited later when separating W jets from background jets.

The fatjets will capture a substantial amount of pileup because of their large size and often small subjets. The fatjets therefore need *grooming*. The grooming method used in ATLAS and this study is called *trimming*.

The trimming algorithm takes the jet constituents and reclusters them into smaller jets. The most used configuration (within ATLAS and used in this study) is to recluster inclusively all the constituents of the jet into subjets with $R = 0.2$ using the k_t algorithm. Any subjet for which $p_T^i / p_T^J < f_{\text{cut}}$, with $f_{\text{cut}} = 5\%$ in ATLAS and this study, are removed from the jet. The process can be seen in Figure 1.25.

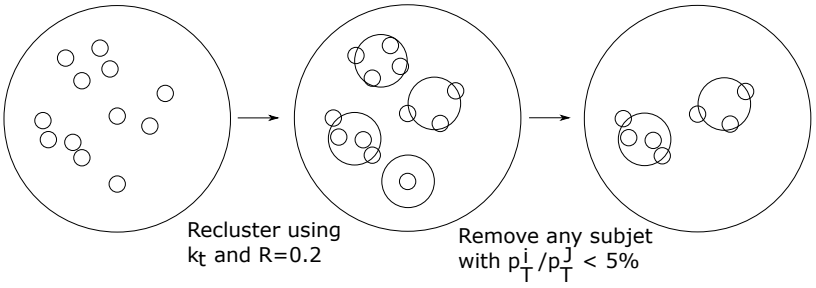


Figure 1.25: Jet trimming.

Tracks from the inner detector that are within ($\Delta R < 1.0$) of the area of the fatjets will be *ghost-matched* to the jet. The ID tracks have, unlike the calorimeter topo-clusters, charge information. This will be used for the jet charge study.

Substructure of fatjets

This study will use fatjets to find the hadronically decayed W particles. Fatjets are used because the two jets initiated by the quarks from the W will merge in the very boosted regime, $p_T^J > 1\,\text{TeV}$. Already at 200 GeV, a fatjet with $R = 1.0$ is used to capture both quark jets.

Therefore, a set of substructure variables for the fatjets have been devised which will separate hadronically decayed W particles from background jets.

The different substructure variables used in likelihood study are: D_2, $\sqrt{d_{12}}$ (also called split12 or kT splitting scale), width, τ_{12}^{wta}, aplanarity, FoxWolfram$_{20}$, μ_{12} (also called mass-drop), planar flow, sphericity, major and minor axes of thrust, and $\sqrt{z_{12}}$ (also called zcut12). Their definitions can be found in Refs [20, 21].

The most separating variables are defined in the following way:

- $\boldsymbol{D_2}$: It is defined as

$$D2 = ECF_3 \frac{ECF_1^3}{ECF_2^3}.$$

  The N-point Energy Correlation Functions (ECF_N) are p_T-weighted sums of the dR of single, pairs of or triplets of jet constituents. The equations are therefore:

$$\begin{aligned}
\text{ECF}_1 &= \sum_{i \in J} p_{T_i}, \\
\text{ECF}_2 &= \sum_{i<j \in J} p_{T_i} p_{T_j} \Delta R_{ij}, \\
\text{ECF}_3 &= \sum_{i<j<k \in J} p_{T_i} p_{T_j} p_{T_k} \Delta R_{ij} \Delta R_{ik} \Delta R_{jk}.
\end{aligned}$$

  $D2$ is sensitive to jets that show two-prong decay.

- $\boldsymbol{\tau_{12}^{\text{wta}}}$: It is defined as:

$$\begin{aligned}
\tau_{21}^{\text{wta}} &= \frac{\tau_2^{\text{wta}}}{\tau_1^{\text{wta}}}, \\
\tau_1^{\text{wta}} &= \frac{1}{\tau_0^{\text{wta}}} \sum_{i \in J} p_{T_i} \Delta R_{1,i}, \\
\tau_2^{\text{wta}} &= \frac{1}{\tau_0^{\text{wta}}} \sum_{i \in J} p_{T_i} \min(\Delta R_{1,i}, \Delta R_{2,i}), \\
\tau_0^{\text{wta}} &= \sum_{i \in J} p_{T_i} \Delta R.
\end{aligned} \tag{1.26}$$

  where ΔR is the jet radius, $\Delta R_{N,i}$ is the distance between the hardest constituent of subjet N and jet constituent i.

  For τ_N^{wta}, the jet is reclustered into exactly N subjets using the k_t algorithm.

  τ_N^{wta} is known as "N-subjettiness", i.e. τ_{21}^{wta} will give a small number for jets that show a substructure of exactly two subjets.

  The regular (non-wta) version uses the centers of the subjets instead of the hardest constituents. The wta-version is often used because of its increased discriminatory power.

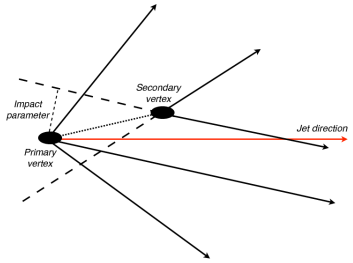


Figure 1.26: A second identified vertex. From [22].

b-tagging

Quark flavor tagging is a crucial part of many analyses, e.g. tagging of bottom quarks is used in suppressing $t\bar{t}$ backgrounds as they will decay to Wb with a BR of over 99%.

Tagging of quarks is based on several discriminating variables. Especially jets formed by the hadronization of bottom quarks can be identified by a secondary vertex coming from the decay of the b-hadron in the jet, as shown in Figure 1.26. Invariant mass, energy ratio between vertices, and number of vertices have shown to separate b-quarks from lighter quarks[12].

1.4.6 Missing transverse energy

Energy and momentum conservation laws dictate that the transverse energy after a collision should be zero within the uncertainties of the detectors and barring processes that would fake missing transverse energy.

The E_T^{miss}, $\not{E}_T$, missing E_T, or simply MET, is a very important quantity. Regular Standard Model neutrinos, but also many hypothetical particles amongst dark matter, SUSY, and more, all move through ATLAS undetected.

Since the MET is calculated when everything has been reconstructed using the whole detector, it is important that the uncertainty of each and every sub-detector is well-modeled, noise is suppressed, everything is calibrated, and every low-p_T track is accounted for, as the many small contributions will smear the MET. Neutral particles with low energies are badly or not at all measured by the calorimeter and will therefore add to the MET.

To first order, the MET is calculated as: [23]

$$E_{x(y)}^{\text{miss}} = -E_{x(y)}^{\text{miss,e}} - E_{x(y)}^{\text{miss,fl}} - E_{x(y)}^{\text{miss,ø}} - E_{x(y)}^{\text{miss,jets}} - E_{x(y)}^{\text{miss,}^-} + E_{x(y)}^{\text{miss,soft}}, \tag{1.27}$$

which are the missing energies (component-wise) of the calibrated electrons, photons, taus, jets, muon, and soft terms, respectively. The soft term is from energy deposits not associated with any object. From the Cartesian components, the transverse and azimuthal components of the MET can be calculated. However, the polar angle can not be reconstructed from the transverse (missing) energies.

Neutrino reconstruction

Neutrinos will be reconstructed from the missing energy. MET, however, only has the transverse components, and can therefore only give p_T and ϕ. To reconstruct the full four-momentum, η will have to be inferred somehow. The solution is to *assume* that the missing energy is the neutrino, and then pair the neutrino with a selected lepton of an event to form the W: $W^\mu = l^\mu + v^\mu$. The assumed neutrino is assumed to come from a W. The mass of the W is known, and from the equation, η_v can be reconstructed [24]:

$$\eta_v = \eta_l \pm \operatorname{arccosh}\left(\frac{M_W^2}{2p_{T_l}p_{T_v}} + \cos\Delta\phi\right). \tag{1.28}$$

If the argument of arccosh is smaller than one, η of the neutrino is set to the pseudorapidity of the electron.

If the argument is larger, the equation has two solutions. The solutions that produces the smallest $|\eta_v|$ is chosen.

1.4.7 Isolation and overlap-removal

In ATLAS, physics objects are created from the whole event and not removed iteratively until all energy in the event has been categorized. This leads to scenarios where one object overlapping with another will contribute its energy to the energy measurement of the other object. It may even be that both objects are created from the same particle.

The solution is to apply isolation criteria and to remove overlaps as well.

Isolation in ATLAS is based on topoetcone and ptvarcone. topoetcone (topo-E_T-cone) is the sum of E_T of the calorimeter topo-clusters in a cone of a given size around the object excluding the object itself. ptvarcone (p_T-var-cone) is the sum of p_T of tracks in the ID in cone of variable size excluding the object itself. ptvarcone has a fixed cone size until a lower threshold where it will shrink until it reaches its minimum size.

For electrons, topoetcone20 and ptvarcone20 are used, where 20 is $\Delta R < 0.2$. ptvarcone20 has a cone size of 0.2 for electrons with $p_T > 50\,\text{GeV}$.

For muons, topoetcone20 and ptvarcone30 are used.

The cuts on the variables depend on the isolation level chosen by the analyses.

The overlap removal procedure has been somewhat harmonized in Run 2, and all analyses are recommended to perform the same overlap removal [25]. The overlap removal will be applied to calibrated objects that have passed initial pre-selection cuts of identification, isolation and various kinematics[10], but before the final objects have been selected. This ensures that a badly identified object does not remove a good object.

[10] In the case of jets, events are removed if they have jets constructed from noisy calorimeters. Furthermore, the jets are removed if they fail a jet-vertex-association (JVT) cut as those jets will most likely be from pileup. In case of fat-jets, they must be groomed beforehand.

Overlap-removal procedures for the objects relevant to the likelihood study are given in Table 1.2.

Object	Procedure
Electron	The electron is removed if it shares inner detector tracks with a muon, or there is a jet within $0.2 < \Delta R < 0.4$. However, the jet is removed instead, if the jet is within $\Delta R < 0.2$ of the electron.
Muon	The muon is removed if a jet with more than two tracks is within $\Delta R < 0.4$ of the muon. However, if jet has fewer tracks, it is removed instead.
Fatjets	The fatjet is removed if it is within $\Delta R < 1.0$ of an electron.

Table 1.2: Overlap-removal for some physics objects.

1.4.8 ATLAS software

The framework used in ATLAS for reconstruction of events is Athena. Athena can also be used for data analysis. However, EventLoop is becoming the standard framework for ntuple-creation[11]. Due to the large number of packages needed to correct various physics objects, frameworks are usually built on top of EventLoop. Popular ones are CxAOD (used by the Higgs and Exotics groups), SUSYTools (used by the SUSY groups), QuickAna (generic framework developed by the top group), and various others.

[11] ntuples are the final containers of physics objects that will be used for data analysis.

For the likelihood study, a homemade framework adapted from a work progress framework using SUSYTools made by a colleague[12] has been used. The framework will apply all calibrations, pre-selections, overlap removals, etc. in accordance with the harmonization group's recommendations.

[12] Geert-Jan Besjes

Data quality

Looking back at Figure 1.14, where the total integrated luminosity for 2015 was presented, the luminosity delivered by LHC was $4.2\,\text{fb}^{-1}$, and the luminosity recorded by ATLAS was $3.9\,\text{fb}^{-1}$. However, the actual amount of data available for most analyses is only $3.2\,\text{fb}^{-1}$. This is determined by the Good Runs List (GRL) which is updated by the ATLAS Data Preparation Group.

Every run is divided into 1 minute blocks called lumiblocks. For each lumiblock, the quality of the data is evaluated. If a part of a detector was misaligned, a module malfunctioned, parts of the

detector went out of sync, or for any other reason the integrity of the data was in question, the lumiblock would be marked as bad.

Pileup-reweighting

It is not known before-hand exactly what the average number of interactions per crossing, $\langle\mu\rangle$, will be. Therefore, MC is generated with a wider range of pileup. The mc15a simulation project in ATLAS came with an anticipated profile that needed to be reweighed to match data. mc15c has the final profile of 2015 and includes an anticipated profile for 2016.

The reweighting works by creating a histogram of the average number of interactions per crossing from a data file as well as a configuration file for each MC sample. The PileupReweightingTool (PRW) will then select events at random so the final profile of the MC will match data.

Trigger and trigger matching

ATLAS has triggers for all physics objects and for some combinations. The naming scheme of the triggers are very strict. For example:

$$\underbrace{\text{HLT}}_{\text{Trigger}}_\underbrace{\text{mu}}_{\text{Object}}\underbrace{\text{20}}_{p_T-\text{threshold}}_\underbrace{\text{iloose}}_{\text{Optional identification cut}}_\underbrace{\text{L1MU15}}_{\text{Level 1 trigger information}} \tag{1.29}$$

An analysis will require that a trigger at the HLT level fired for a given object (in this case muons). The minimum p_T level at which the trigger reaches its maximum efficiency follows. A trigger will fire for objects above this threshold. Some triggers include an optional identification cut. The final piece is the level 1 trigger level. Due to the rate at which the L1 has to operate, the requirements are lowered slightly to keep the efficiency high.

The selected objects of an event are then matched to the objects that triggered the event.

1.5 Boosted Decision Trees in TMVA

The tool used for machine learning is Toolkit for Multivariate Analysis (TMVA) [26], which comes bundled with ROOT 6.

TMVA has several machine learning algorithms: Boosted Decision Trees (BDTs), Neural Networks (NNs), etc. The machine learning algorithm considered in the likelihood study is the BDT.

A (non-boosted) decision tree will cut on the variable that give the greatest separation between signal and background. A branch will form from this node for both the case of an event succeeding or failing the cut. In both new nodes, the greatest separating variable after the first cut will then be cut on. It is important to note that the separation power of variables change after cutting on a variable. The tree will continue this until reaching a defined depth. A classifier from this tree will then give a 1 or -1 answer as to whether the tested event was deemed a signal or background event.

BDTs are collections, or *forests*, of trees. After the first tree has been created, a boosting algorithm will assign higher weights for misclassified events such that the next tree will cut on a different set of variables. After continuing this process for some defined number of iterations, each tree will be able to (to some degree) recognize and separate their own piece of the N-dimensional variable space. The classifier will run the variables through the forest and give the average of the tree outputs.

The adjustable parameters for the BDTs in TMVA that can lead to greater performance are:

- NTrees. It is the number of trees grown in a forest.
- MaxDepth. It is the maximum depth of a tree.
- nCuts. At each node, every variable is cut into nCuts pieces.
- MinNodeSize. Each node is required to have a minimum percentage of training events available.
- AdaBoostBeta. The parameter for the boosting procedure.

Overtraining in the machine learning sense is a (large) discrepancy between training and testing performance. In physics, one could also consider an algorithm overtrained if it becomes sensitive to quirks or events not found in data but specific to some Monte Carlo generators.

This level of overtraining is combated by considering two important aspects:

- **Input variable coherence:** Variables must match between data and simulation to a great extent. A simple cut will separate a variable, but machine learning algorithms will cut on the same value many times.
- **Using another generator for testing:** Generators use different models in their algorithms, e.g. for hadronization. This will create nuances in the MC data. Generators are continuously

tuned to match actual data, but many generators are optimized for different processes, some at NLO or greater, so their correspondence to data or each other is not guaranteed. Training against a sample from one generator and testing against a sample from another will give indications as to when the algorithm has trained beyond the separability in data.

In any case, if MC data and correlations between variables are well-modeled by the simulation(s), overtraining a machine learning algorithm will only lead to worse performance on data. This can for example be seen as a less pronounce peak in the mass histogram.

II Analysis

2.1 Improving the jet charge

The motivation for this section is to improve the charge calculation of jets in order to improve the separations of actual W and Z bosons from background jets as well as to improve the separations of W^+, W^-, and Z from each other.

The charges calculated from jets should match the charges of the particles that initiated them. The classification can possibly be further improved by also considering the charge *difference* of the subjets.

W bosons, to conserve charge, decay to an up-type quark and a down-type antiquark (or vice-versa), when decaying hadronically, e.g. they decay to $u\bar{d}$ which has the charge $\frac{2}{3}e + \frac{1}{3}e = 1\,e$.

This leads to table 2.1.

Table 2.1: Charge sums and differences of the two quarks that the bosons decay into. $Q_1 \geq Q_2$ are the charges (in units of e) of the subjets identified in the fatjet. Background jets are also included.

Particle	$3(Q_1 - Q_2)$	$Q_1 + Q_2$
W	1	1
Z	2 or 4	0
quark [initiated bkg. jet]	ill-defined	$\pm 1/3$ or $\pm 2/3$
gluon [initiated bkg. jet]	ill-defined	0

Background jets have on average no substructure, so the charge difference of the two constructed subjets is ill-defined. The charge sum is assumed to be the total sum of the jet which should match the charge of the parton that initiated the jet.

However, simply summing the charges of all the tracks in a jet will give wildly fluctuating values when tracks are missed by the detector. Weighting has to be introduced to give stable charge calculations and the best performance. The exact equations will be discussed in Section 2.1.3.

For the analysis we will only consider the separation of positive and negative W bosons. At the end of this section we will return to Z and background jets with our improved calculations of charge to examine the improved separation power of jet charge.

2.1.1 Datasets

The datasets used are listed in Tables A.1 (backgrounds) and A.2 (signals) in Appendix A.1.

The signal samples are generated from the hypothetical W' particle decaying fully-hadronically through the channel $W' \to WZ \to qqqq$. A fair selection of resonance masses for the hypothetical W' particle has been chosen to span the p_T range from 200 GeV to 2000 GeV.

The background samples are generated from the same setup as the signals but consist of dijets initiated by gluons and light quarks. The samples are cut on the p_T of the p_T-leading jet and split into the datasets listed in the table. The background covers the same p_T range as the signal samples.

All datasets in this section are based on the mc15a simulation

project within ATLAS. mc15a datasets were the first datasets for Run 2 which includes the bunch-spacing of 50 ns and an anticipated pileup profile for 2015.

2.1.2 Selection and cuts

From each event, the trimmed and calibrated anti-k_T $R = 1.0$ jet made from calibrated topo-clusters with the highest p_T is selected.

The event is afterwards discarded if the jet fails any of the following cuts:

- $p_T > 200\,\text{GeV}$
- $|\eta| < 2.0$
- $1\,\text{GeV} < m_J < 200\,\text{GeV}$

For signal, the jet will then be truth-matched. The truth W and Z particles must have decayed hadronically. If the truth W or Z is within $\Delta R < 0.6$ of the reconstructed jet, the reconstructed jet is truth-matched. If zero or more than one truth W or Z particles match, the event is discarded.

The truth particle's type (W or Z) as well as charge is saved.

The ghost-matched tracks of the jets are saved. The values saved are: ϕ, η, p_T, charge, and ΔR between track and fatjet as well as hits in the SCT and pixel detectors and several track quality variables (e.g. the perigee parameters d_0, z_0, etc.). The hits and track quality variables were used as part of the analysis but they showed no improvement in separation.

2.1.3 Methods

Currently, ATLAS calculates the charge of a jet using the following equation: [27]

$$Q_J = \left(\frac{1}{p_T^J}\right)^{\kappa} \sum_{i \in \text{tracks}} q_i \cdot (p_T^i)^{\kappa}, \quad \kappa = [0.3, 0.6, 1.0]\,. \tag{2.1}$$

The three κ values are the values used in the reference. Later, internal studies focus on values 0.3, 0.5, and 0.7. Optimal κ values have been shown to lie in the range 0.3 to 1.0, in any case.

This can, however, be improved. Two cases will be investigated:

- **Fatjet method:** Using equation 2.1 with $\kappa(p_T^J)$ as a function of the jet transverse momentum, p_T^J, gives us the equation:

$$\begin{aligned} Q_J &= \frac{1}{\sum_{i \in \text{tracks}} W_{pt}^i} \sum_{i \in \text{tracks}} q_i W_{pt}^i, \\ W_{pt}^i &= (p_T^i)^{\kappa(p_T^J)}. \end{aligned} \tag{2.2}$$

  Here, the weight has been defined as W_{pt}^i. The term in the sum is therefore the same as in equation 2.1 except for κ being a

function of the jet transverse momentum. If κ does depend on p_T^J, and a simple function can be fitted against it, it will give us a set of hyper-parameters (the fit parameters for κ) that may improve separation.[1]

[1] In the case of a p_T^J dependence, if we cannot fit against a simple function, or we choose not to fit, the result is a κ value that jumps discontinuously from p_T^J bin to bin. That would have to be smoothed, otherwise this discontinuousness would propagate through the analysis. The solution is to determine κ in each bin and find a function that fits the the κ evolution as p_T^J increases, so we can perform our final charge calculation without bins (and hence without discontinuous jumps).

The normalization is slightly different between equations 2.1 and 2.2. In the updated equation, the sum of the weights is divided out to normalize the weight. This means that we use only the momentum from the tracks (and hence only from the inner detector) directly in the equation, and the momentum from the jet (the calorimeter) indirectly in the p_T^J binning and in κ. The new normalization should avoid the risk of asymmetry.

- **Subjet method:** The subjet structure of the fatjets can hopefully be exploited by calculating each of the two subjet charges. The equation for this method will then be:

$$\begin{aligned} Q_J &= Q_1 + Q_2, \\ Q_n &= \frac{1}{\sum_{i\in\text{tracks}_n} W_{pt}^i} \sum_{i\in\text{tracks}_n} q_i W_{pt}^i W_{dR}^i, \\ W_{pt}^i &= (p_T^i)^{\kappa(p_T^J)}, \\ W_{dR}^i &= \exp\left[-\left(\frac{dR_i}{R(p_T^J)}\right)^{\beta(p_T^J)}\right]. \end{aligned} \tag{2.3}$$

The set tracks_n is the set of the ghost-matched tracks which were closest to subjet n. The distance is calculated as the usual distance in $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ between the subjet center and the i'th track. If the closest distance is larger than 0.2, the track is discarded:

$$\text{if } \min(dR_{\text{subjet 1,track } i}, dR_{\text{subjet 2,track } i}) > 0.2 \text{ then remove track } i.$$

A new weight, W_{dR}^i, has been introduced which weighs the tracks by their distance to the closest subjet. Since the hardest tracks are expected to be close to the center of the subjets, and any track farther away than $\Delta R = 0.2$ is discarded, it can be argued that this weight will serve the same purpose as the regular W_{pt}^i weight. Never-the-less, the weight is introduced and comes with its own two parameters, R and β. This expression has been chosen because its behavior very much depends on the value of β. As can be seen in Figure 2.1, the function bends the other way at $\beta \approx 1$. This generic expression should be able to match any simple ΔR dependence, should there be one.

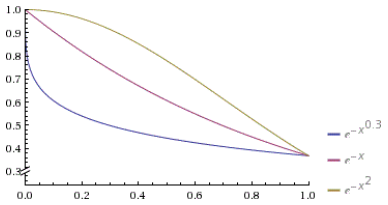


Figure 2.1: $W_{dR}^i(R = 1)$ for different values of β, showing the versatility of the expression. The yellow line ($\beta = 2$) is the normal distribution, and the blue line ($\beta = 0.3$) resembles exponential decay.

With the equations defined, we set to maximize the performance of our equations. The standard way is to define the positively charged particles as signal and negatively charged as background (or vice-versa) and then calculate a ROC curve from which one can quantize the performance by listing the signal efficiency and background rejection at the cut $Q = 0$.

Since we want to use a minimizer to find the optimal values for the parameters, we have to be wary of the above method. The number of

particle that have passed a cut is an integer. Taking the ratio of passed particles to the total number of particles will result in a constant value around the minimum. This will trip ROOT's Minuit minimizer.

Instead, we shall use a similar method for performance validation. When the charges of all jets have been calculated, the means and uncertainties of the W^+ and W^- charge distributions are each calculated. The method for measuring the efficiency is then calculating the z-score between W^+ and W^-:

$$z = \frac{|\mu_{W^+} - \mu_{W^-}|}{\sqrt{\sigma^2_{W^+} + \sigma^2_{W^-}}}. \tag{2.4}$$

Larger values of z mean greater separation between the positive and negative bosons. Minimizing $-z$ is a possibility, as the smallest adjustments to the parameters will always lead to slightly different calculations of the charges and hence not trip the minimizer.

The z-score, however, is only well-defined for Gaussian-like distributions. Asymmetric distributions may suffer from a lower z-score, even if they separate positive and negative jets better. The integral of the ROC curve does not suffer the same limitations. It has, however, not been possible to compute the ROC curve consistently. Initial studies have shown the difference in the obtained parameters to be small when using the ROC curve integral compared to the z-score. Therefore, the z-score was used for the separation measure for the jet charge study.

2.1.4 Results

The data sample has been binned in p_T^J, and all minimizations and calculations have been done in each bin. The bins are

$$[200, 280, 410, 510, 710, 920]\,\text{GeV}.$$

The width of the bins have been chosen such that there are 5000 W^+ and W^- particles each in every bin until approx. 950 GeV, where the jets become boosted to such an extend that any substructure disappears in the detector. At this point, the subjet method will break down, and jets with transverse momentum beyond 920 GeV have therefore not been included in this part of the study.

Firstly, we maximize the separation (using equation 2.4) by running the minimizer on equation 2.3 (the subjet method) in each p_T^J bin. We compare the calculated charges against the originally calculated charges using the standard ATLAS method.

The result can be seen in Figure 2.2. The subjet method (named 'new' in the figure) is drawn with the full lines, while the charges calculated using the current ATLAS method (named 'old' in the figure) is drawn with dashed lines.

Since the minimizer minimizes $-z$ and not a χ^2, the reported uncertainties on the parameters are incorrect. This is fixed by using the bootstrap method which weights the calculated charges, during

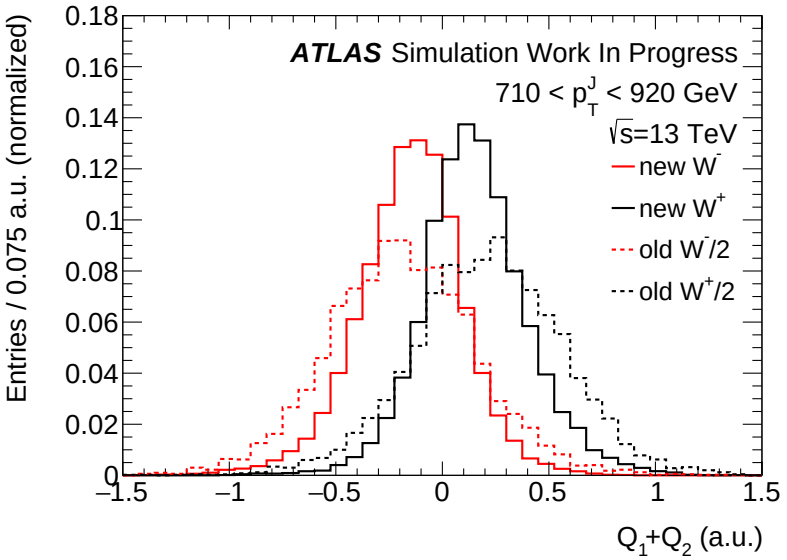


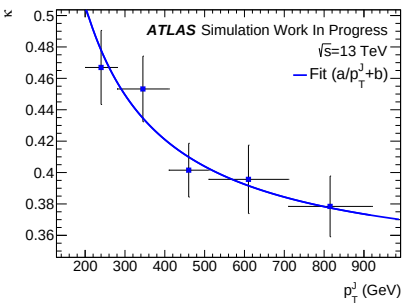
Figure 2.2: An example of the distribution of W^+ and W^- charges after maximizing the separation using equations 2.3 and 2.4. 'new' (full lines) in the legend refers to the subjet method (after maximized separation), while 'old' (dashed lines) refers to the current ATLAS method (untouched). The separation can be seen in Figure 2.4 by comparing the last red and black points to see the difference between the new and old methods, respectively.

the minimization, with a number drawn from a Poisson distribution with $\lambda = 1$.

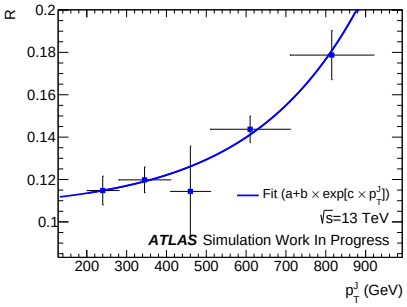
The bootstrap method has been executed 100 times within each p_T^J bin to calculate the actual standard deviations of the parameters κ, R, and β. From the 100 values of each parameter, the standard deviation is calculated in the usual way:

$$\sigma_x = \sqrt{\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

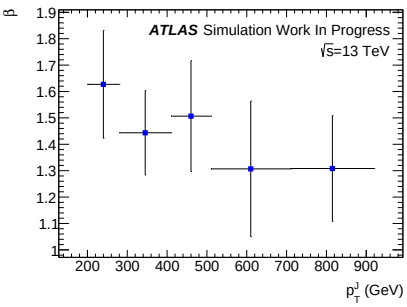
The results can be seen in Figure 2.3.



(a) κ as function of p_T^J.



(b) R as function of p_T^J.



(c) β as function of p_T^J.

Figure 2.3: The most likely values obtained for the weight parameters κ, R, and β after maximizing the separation of W^+ and W^- using equations 2.3 and 2.4. The error bars are the standard deviations calculated from the bootstrapping method.

It can immediately be seen that it cannot be ruled out that β is constant in p_T^J. κ and R, however, seem to express clear trends. The two graphs have been fitted with functions with few parameters in order to keep the number of degrees of freedom high. The fit probabilities, p, are calculated from the χ^2s of the fits and listed as well.

The downwards going trend of κ has been fitted against a hyper-

bola,

$$\kappa = \frac{a}{p_T^J} + b,$$
$$p = \text{Prob}(\chi^2 = 0.78, NDF = 5 - 2) = 0.86.$$

The upwards going trend of R has been fitted against an exponential,

$$R = a + b \cdot \exp\left[c \cdot p_T^J\right],$$
$$p = \text{Prob}(0.34, 2) = 0.84.$$

A first or second order polynomial will fit the graphs just as well. However, we do not expect κ to go below zero like a first-order polynomial would otherwise suggest, or R to increase for p_T^J going in the negative direction as a second-order polynomial would do.

The functions (with the hyper-parameters obtained from the fits) are then used to set the parameter values when again calculating the charges. We are not minimizing this time; we have found our final equation and its (hyper-)parameters, and the result of this is named 'final' in Figure 2.4.

The final expressions for the parameters are:

$$\begin{aligned} \kappa &= \frac{34(12)\ \text{GeV}}{p_T^J} + 0.34(3), \\ R &= 0.11(3) + 0.004(13) \cdot \exp\left(0.004(4)\frac{1}{\text{GeV}} \cdot p_T^J\right), \\ \beta &= 1.45(9), \end{aligned} \tag{2.5}$$

with p_T^J in units of GeV.

So far, we have only used the subjet method, and we have seen a slight improvement. However, the complexity has increased from the current ATLAS implementation to the subjet method. To verify that each addition to the current implementation further improves separation, we make intermediate equations.

We fill in the steps between the two equations by

- using the fatjet method, which is the same as the ATLAS implementation apart from κ being p_T^J dependent,
- and using the subjet method with just the W_{pt}^i weight.

Finally, the bootstrap method has also been applied to the fatjet method.

The uncertainties obtained from bootstrapping the fatjet and subjet methods are added to or subtracted from the central parameter values, and the separations are recalculated, until the worst separation is found. This is done to obtain a quantitative proxy to a systematic uncertainty on the separations. However, this method is known to over-estimate the uncertainty. The systematic uncertainties are then defined as:

$$\sigma_{\text{syst}} = 1/2(z_{\text{optimal}} - z_{\text{worst}}).$$

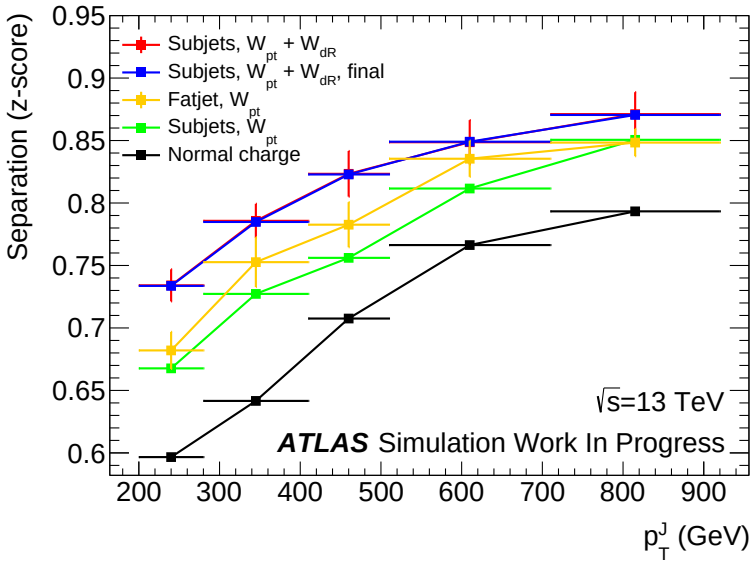


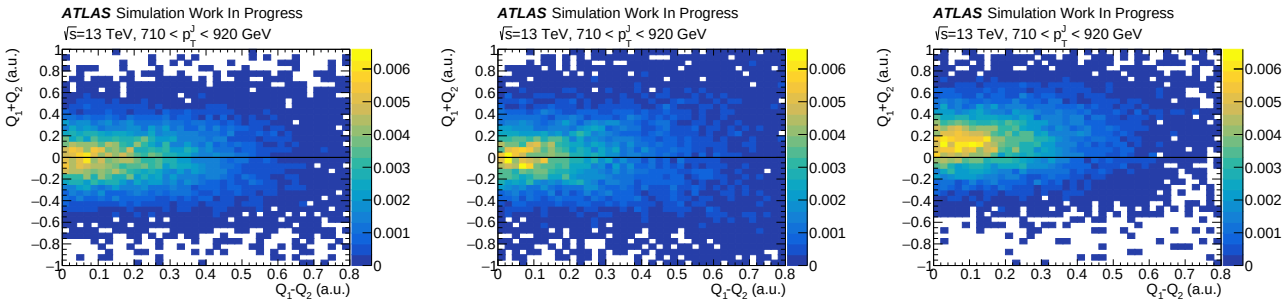
Figure 2.4: Except for the 'final' version, for each method in each p_T^J bin, the separation has been maximized by varying its parameters. From bottom to top: The black line is the current ATLAS implementation, the green line is the subjets method without the W_{dR}^i weight, the orange line is the fatjet method, the red and blue lines are the full subjet methods. The blue ('final') line is not the result of minimization, and is instead the actual application of the subjet method with all its (hyper)parameters defined. The uncertainties on the red and orange lines are purely systematic and determined with help from the bootstrapping method and is explained in the text.

In Figure 2.4 the stepwise improvement of the new charge calculation can be seen. The blue ('final') line performs just as well as the subjet method that was subjected to minimization which suggests that the parameters' p_T^J dependence is well-modeled by their expressions in Equation 2.5.

It is interesting that the subjet method with just W_{pt}^i actually performs worse than just the fatjet method. Further studies are required to find the reason for this.

Separation is greatly increased by exploiting the p_T^J dependence of κ, as can be seen by comparing the black and orange line. However, except for the first bin, the orange and red lines lie within approx. two standard deviations, and the full subjet method does therefore not necessarily improve separation. The uncertainties are the previously mentioned systematic uncertainties which are over-estimated, so a more detailed study may find the separation between the two lines to be greater than obtained here.

2.1.5 Charge as a separating variable



(a) Z (b) background jets (c) W^+

Figure 2.5: Subjet charge differences and sums for (a) the Z bosons, (b) background jets (quark and gluon initiated), and (c) the W^+ bosons for $710 < p_T^J < 920$ GeV. A black line has been drawn through $Q_1 + Q_2 = 0$ to show the (a)symmetry of the charge sums.

With the improved jet charge calculation, we seek to use this to separate bosons from background and bosons from each other. During the minimization, the subjet charges have been saved individually with $Q_1 \geq Q_2$. In Figure 2.5, the subjet charges have been added and subtracted for Z jets, gluon plus quark initiated background jets, and W^+ jets. The Z and background jets look almost indistinguishable.

It seems from Figure 2.5 that it is not possible to separate Z bosons from background using both the charge sum and the charge difference of the two subjets. However, as expected, the positive W bosons are mostly on the positive side of the subjet charge sums. W^- mirrors W^+ and is therefore not shown.

The charge difference between the two subjets has been histogrammed in Figure 2.6 which shows a slightly larger tail for the background jets. The separation power is, however, insignificant.

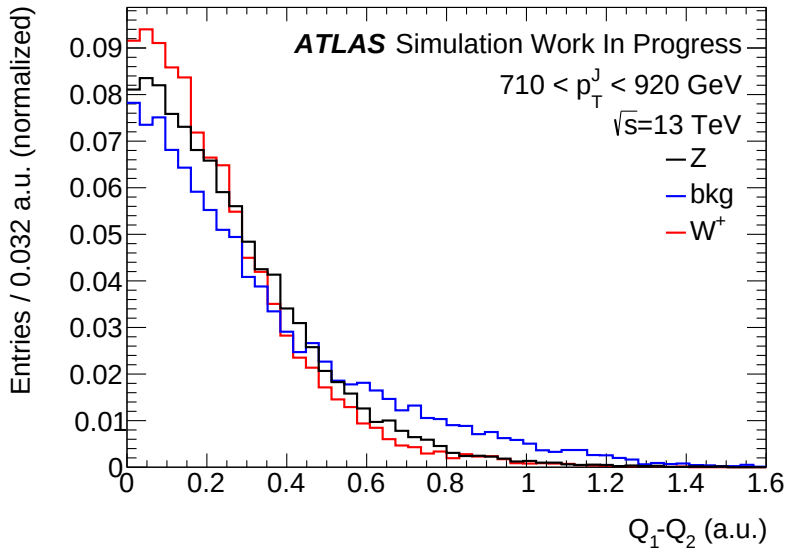


Figure 2.6: The subjet charge differences for the Z bosons, background jets, and W^+ bosons. This is the one-dimensional slice of Figure 2.5.

2.1.6 Final notes

Jet charge calculation is generator sensitive because of the difference between hadronization models (e.g. Lund string model vs. the cluster model). This study has not compared datasets from one generator to another, and the simulation has also not been compared to data, so any gain achieved in this study cannot be verified.

The analysis for this section has not included any cuts to separate signal from background. Therefore, new studies should consider including cuts on substructure variables that remove background fatjets with little two-subjet-resemblance. After these cuts, the subjet charge difference is better defined.

This study could also be combined with quark-gluon-taggers that serve to separate gluon- and quark-initiated jets. Perhaps the charge sums and difference will become more apparent after categorizing events into quark and gluon regions.

2.2 Object definition

In the previous section, it was found that the improved jet charge calculation did not improve the separation power of the jet charge. For this reason, no efforts into separating hadronically decayed W and Z bosons will be made, and the charge has not been used in this study for discrimination.

We shall now delve into the main study of the improvement of the diboson resonance search using a likelihood on machine learning outputs. In this section, the creation of the datasets will laid out, and the agreement between actual data and simulated data will be examined.

2.2.1 Datasets

The datasets used are listed in Appendix A.1.

The signal samples are generated from the hypothetical heavy vector triplet (HVT)[29]. The HVT is a hypothetical composite particle of a WV, where V is a W or Z. The HVT decay to two W bosons that further decay semi-leptonically through $HVT \to WW \to lvqq$. Resonance masses spanning from 500 GeV through 5 TeV have been analyzed and used in the study. However, only the 1.5 TeV HVT has been used in the final analysis.

The background samples are the major backgrounds in a diboson analysis: W+jets, $t\bar{t}$, and (non-resonance) dibosons. W+jets and $t\bar{t}$ are the major components, while the dibosons do contribute in the signal region (defined later).

The fourth largest background contribution comes from Z+jets, which were not included in the study. Their contribution is even smaller than dibosons, and were therefore left out for reasons that will become clear when the plots are shown below.

All datasets in this section are based on the mc15c simulation project within ATLAS. mc15c datasets contain the final pileup profile for the 2015 data of Run 2 and an anticipated pileup profile for the 2016 data.

The background samples are made from the following processes:

- **W+jets:** $q\bar{q} \to Wg \to lvq\bar{q}$ through a t-channel process, where l is an electron, muon, or tau. See Figure 2.7.
- **$t\bar{t}$:** $q\bar{q} \to g \to t\bar{t} \to WbW\bar{b} \to lvbq\bar{q}\bar{b}$, where l is an electron, muon, or tau. See Figure 2.8.
- **Dibosons:** $q\bar{q} \to VV$ through a t-channel process with a virtual quark, where V is a W or a Z decaying through any of the possible channels. See Figure 2.9.
- **Z+jets:** $q\bar{q} \to Zg \to l\bar{l}q\bar{q}$ through a t-channel process, where l is an electron, muon, or tau. See Figure 2.10. These samples were not included in the study.

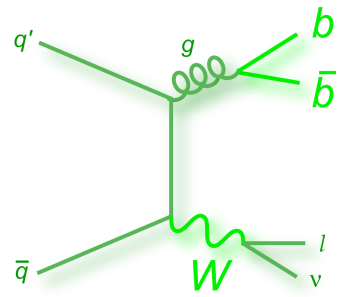


Figure 2.7: W+jets: $q\bar{q} \to Wg \to lvq\bar{q}$ through a t-channel process with a virtual quark. From [28].

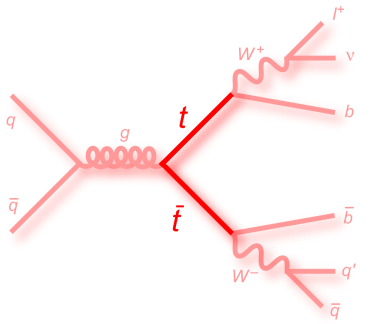


Figure 2.8: $t\bar{t}$: $q\bar{q} \to g \to t\bar{t} \to WbW\bar{b} \to lvbq\bar{q}\bar{b}$. From [28].

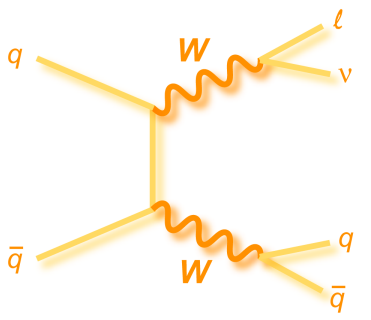


Figure 2.9: Dibosons: $q\bar{q} \to WW \to lvq\bar{q}$ through a t-channel process with a virtual quark. From [28].

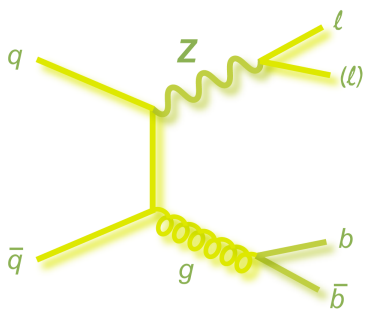


Figure 2.10: Z+jets: $q\bar{q} \to Zg \to l\bar{l}q\bar{q}$ through a t-channel process with a virtual quark. From [28].

We define the following signal region and background regions in which each background channel dominates:

- **Signal region:** $65\,\text{GeV} < m_J < 95\,\text{GeV}$ and 0 b-jets in the event.
- **Top region:** $65\,\text{GeV} < m_J < 95\,\text{GeV}$ and at least 1 b-jet in the event.
- **W+jets region:** $65\,\text{GeV} \nless m_J \nless 95\,\text{GeV}$ and 0 b-jets in the event.

These regions will only be used for validation and where explicitly mentioned. For the final analysis using the likelihood fit, a looser signal region is defined.

2.2.2 Selection and cuts

The cuts performed in this study are based on the cuts listed in the internal note of the EXOT11 group. The cuts were at one point optimized for maximum background rejection in the signal region at a signal efficiency of 50%.

The cuts could have been loosened to regain some signal efficiency (perhaps at a cost of lower significance). The machine learning methods and coming likelihood fit could have increased the final signal significance. However, this is left for a future study.

The event must pass the following cuts:

- **GRL (data) or non-zero RRN[2] (MC)**. The GRL (in case of data) tells whether the event was marked bad by the ATLAS Data Preparation Group. The PRW tool (in case of MC) gives a weight (randomly picked from a histogram) for the event to correctly weight the MC against data. If the weight is zero, there is no point in continuing.
- **Detector cleaning**. Four quality checks that are not included in the GRL. The event is rejected if the HCAL, ECAL, or SCT has an error flag set to true, or the event in the xAOD is incomplete.
- **Primary vertex check**. The primary vertex with the highest $\sum_i p_{Ti}^2$ must have at least 2 tracks to pass the event.
- **Trigger**. One of the lepton triggers defined below must have fired. The matching between the selected lepton and the lepton that fired the trigger is performed at a later stage.

[2] The Random Run Number generated by the PRW tool. If the event has a pileup outside of data, the event will be discarded.

The triggers used in this study are the lowest unprescaled[3] lepton triggers. The trigger for an event with a sole signal electron is HLT_e24_lhmedium_L1EM20VH, and for a sole signal muon it is HLT_mu20_iloose_L1MU15.

SUSYTools applies all the correction tools to all the objects of interest (electrons, muons, jets, fatjets) and applies overlap removal between them[4].

The lepton selection cuts are given by Table 2.2, and the fatjet selection cuts are given in Table 2.3.

[3] If a trigger (especially low p_T) triggers too often, its triggering rate is scaled such that the output rate from the HLT is kept within working limits. This allows for higher p_T triggers or triggers of rarer events to add more useful data to the data stream.

[4] The overlap removal methods were given in Section 1.4.7.

	Electrons		Muons	
	Veto	Signal	Veto	Signal
p_T	$> 25\,\text{GeV}$		$> 25\,\text{GeV}$	
$\lvert\eta\rvert$	$< 2.47 \notin [1.37, 1.52]$		< 2.5	
ID	LooseLH	TightLH (MediumLH for $E_T > 300\,\text{GeV}$)	Loose	Medium
Isolation	-	Tight	-	Tight
$\lvert d_0/\sigma_{d_0}\rvert$	-	< 5	-	< 3
$\lvert z_0 \sin\theta\rvert$	-	$< 0.5\,\text{mm}$	-	$< 0.5\,\text{mm}$

Table 2.2: The cuts applied to the leptons of the event. The MediumLH identification is required for $E_T > 300\,\text{GeV}$ by the Combined Performance group. The vertex cuts (two last lines) are standard cuts.

Fatjets	
p_T	$> 100\,\text{GeV}$
$\lvert\eta\rvert$	< 2.0
m	$> 35\,\text{GeV}$

Table 2.3: The cuts applied to the fatjets of the event.

After applying the cuts on the leptons and fatjets, the event must pass the following cuts after passing the previous event cuts:

- MET $> 55\,\text{GeV}$.
- At least one fatjet.
- One signal lepton and zero veto leptons.

Only one lepton (either an electron or a muon) is allowed after the lepton cuts. If an event has one or more good leptons (of either flavor), or a lepton (besides the sole signal lepton) passed the veto cuts, the event is discarded. This ensures suppression of background from two-lepton decays.

The jet collection for the fatjets is the collection of trimmed and calibrated anti-k_T $R = 1.0$ jets made from calibrated topo-clusters. The fatjet with the highest p_T will be selected.

2.2.3 Saved values

When the event has passed the cuts, the hadronically and leptonically decayed W particles are reconstructed[5]. lvJ is then reconstructed from the two W particles. The subscript J denotes the fatjet that contains both of the jets formed by the $q\bar{q}$ coming from the hadronically decayed W. Four-momenta of all particles are saved along with all the substructure variables of the had. W.

The ATLAS MV2c20 flavor tagging tool is used for tagging b-jets in the event. The highest available b-jet efficiency of 85% is chosen to eliminate the top background as much as possible. This is of course at a cost of signal efficiency, but tuning this choice is left for a future study[6]. All b-tagged jets in the event are counted. A value excluding the fatjet from the counting, if it is b-tagged, is also calculated.

Trigger matching is applied and saved to the ntuple without cutting.

The subjets of the fatjet are not reconstructed from the fatjet's cluster, because a greater precision in η and ϕ is needed for the

[5] The neutrino is reconstructed according to Section 1.4.6.

[6] In the outlook paper, an idea for an extended top tagger is presented.

subjets. Therefore, jets of radius $R = 0.2$ constructed from tracks in the ID have been ghost-matched to the fatjet. The trackjets will be used to calculate the angles of the event.

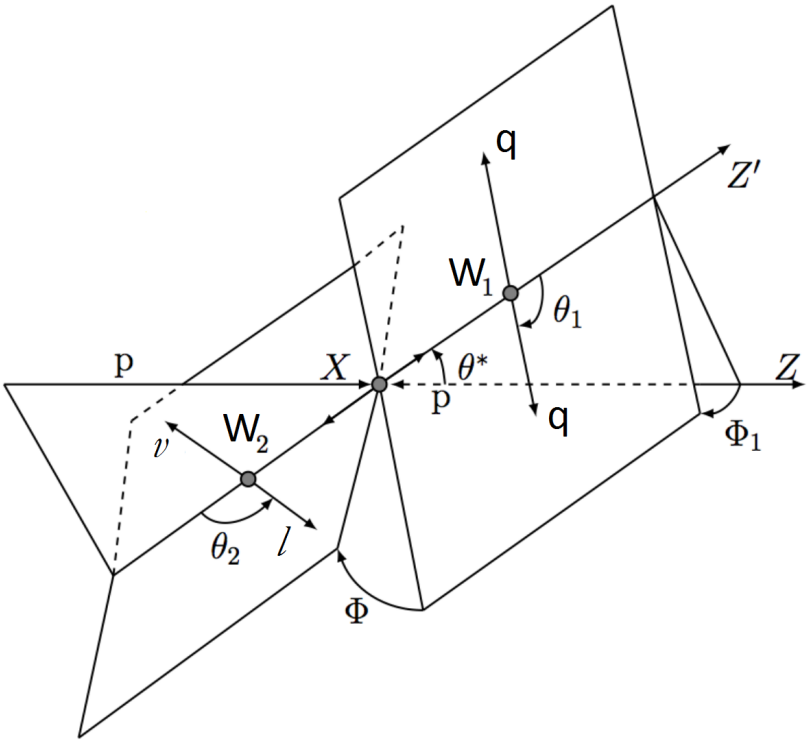


Figure 2.11: The angles of the whole event. Adapted from [30].

The angles are drawn in Figure 2.11. θ_1 is the angle between the higher-p_T trackjet and its parent W, regardless of the charge of the trackjet. θ_2 is the angle between the lepton and the direction of its parent W. Φ and Φ_1 are the angles between the planes, but they provide no separation power and are therefore not used. θ^* is the angle of W_1 to the beam line. All angles are defined in the rest frame of the $lvqq$ system, where qq are the two ghost-matched trackjets.[30]

The code for computing the angles has been lent from a colleague[7].

[7] Lars Egholm Pedersen

2.2.4 Monte-Carlo and data agreement

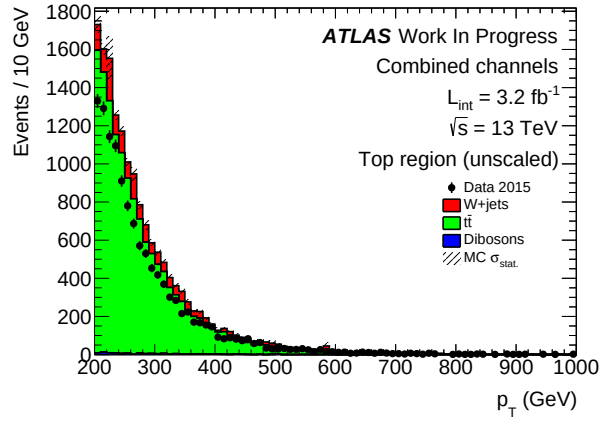
At this point, all the datasets have been analyzed by the framework and consolidated into an MC-ntuple and a data-ntuple.

We want to verify that the simulations match the data taken in the year 2015. The weights have been calculated, and the MC samples have been normalized to $3.2\,\text{fb}^{-1}$. However, there is still some discrepancy between data and MC as seen in Figures 2.12(a-b).

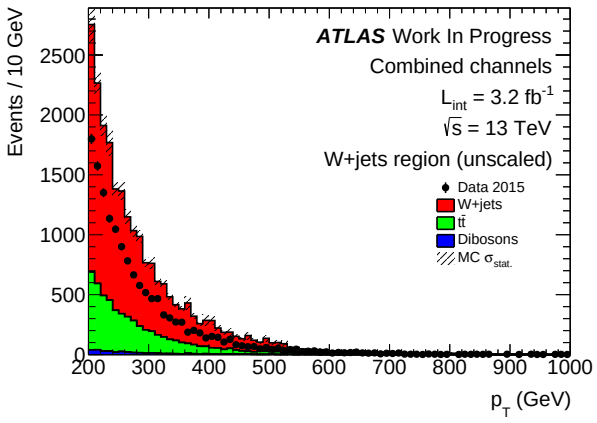
A new set of rescaling weights need to be applied on top of the previous weights.

The weights that need to be applied to the different MC samples have been calculated using Equation 2.6.

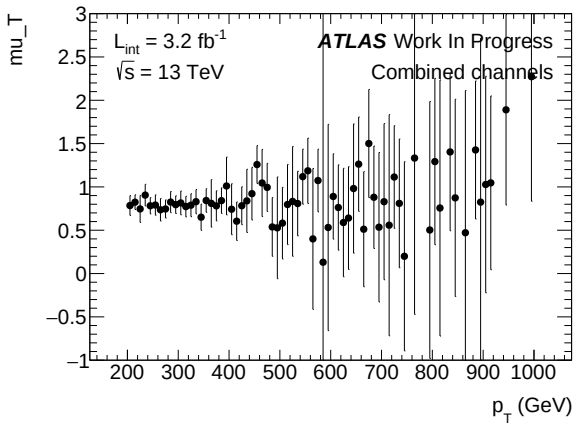
$$\begin{aligned} D_T &= \mu_T T_T + \mu_T W_T, \\ D_W &= \mu_T T_W + \mu_T W_W. \end{aligned} \tag{2.6}$$



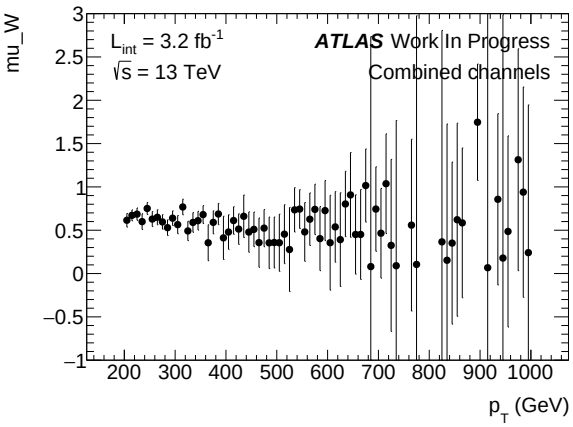
(a) p_T for had. W in the unscaled top region.



(b) p_T for had. W in the unscaled W+jets region.



(c) Weight to be applied to the top samples.



(d) Weight to be applied to the W+jet samples.

Figure 2.12: The p_T spectrum of hadronically decayed W particles in the different regions. The samples have not yet been scaled. Subfigures (c) and (d) are the calculated weights to be applied to the MC top and W+jet samples, respectively.

The subscripts on D, T, and W refer to a region (top or W+jets), while the subscripts on μ refer to a set of samples. The idea is that data (D on the LHS of the equations) must match the sum of all MC backgrounds (T for top samples and W for W+jet samples on RHS) in each region. The weights μ correct slightly for the discrepancy in the number of events. Only the two main backgrounds have been considered because it is not possible to make a diboson enriched region that can be included in the equations.

Equation 2.6 contains two equations with two unknowns for which μ_T and μ_W have been isolated:

$$\begin{aligned} \mu_T &= \frac{W_W D_T - D_W W_T}{W_W T_T - T_W W_T}, \\ \mu_W &= \frac{D_W - \mu_T T_W}{W_W}. \end{aligned} \tag{2.7}$$

The weights are calculated in each of the 80 bins, and a weighted average is computed. The average weight is applied according to Equation 2.6. The result can be seen in Figure 2.13 for which the selection for the signal region has been inverted.

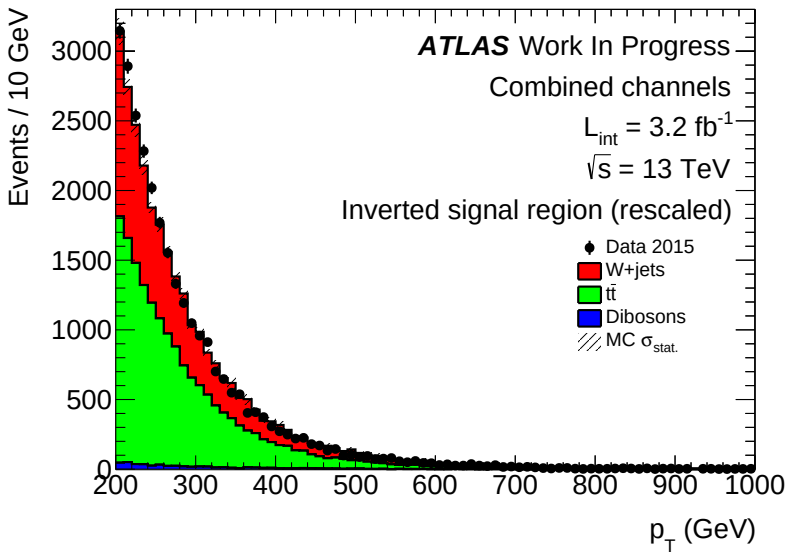


Figure 2.13: The had. W p_T after applying the rescaling weights.

The correct way of handling the rescaling would be to include the control regions in the final likelihood that would take the number of background events from the controls regions into consideration when evaluation the signal region. The smaller samples (dibosons, Z+jets) would be treated as nuisance parameters. The nuisance parameters would be Gaussians with the respective estimated background yields as mean and a fixed width.

With the background distributions rescaled to fit data, we arrive at Figure 2.14, which is the $D2$ variable for the fatjets in the signal region. The agreement is quite convincing. The lack of Z+jets as well as not including the dibosons in the rescaling might be the largest contributers for the disagreement.

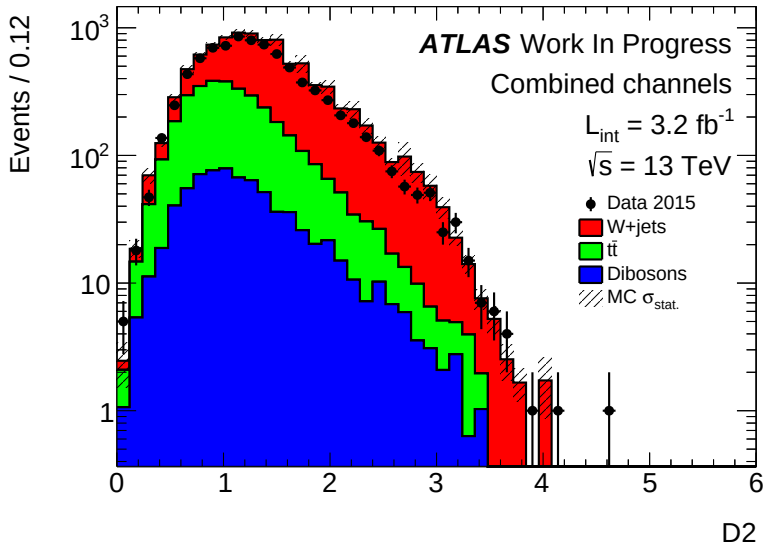


Figure 2.14: The $D2$ spectrum for the fatjets in the signal region.

The final result after selection and scale correction can be seen in Table 2.4.

	e channel	μ channel	combined
Data	18985 ± 138	23491 ± 153	42476 ± 206
Total background	18352 ± 135	22516 ± 150	40868 ± 202
W+jets	10427 ± 102	12773 ± 113	23200 ± 152
$t\bar{t}$	7627 ± 87	9377 ± 97	17003 ± 130
Dibosons	299 ± 17	366 ± 19	665 ± 26

Table 2.4: Yield table after the final corrections. The uncertainties are Poissonian.

2.2.5 Final selection and signal region

Unfortunately, the amount of statistics after applying the PRW is too low to train the BDTs properly. Therefore, for this study, the PRW is turned off, and data is therefore not included. A future study will include the 2016 data as well which will bring the available MC data to a level that is sufficient for BDT training in many variables.

The looser definition of the signal region that will be used for the likelihood fit is as follows:

- **Signal region:** $35\,\text{GeV} < m_J < 150\,\text{GeV}$ and 0 b-jets in the event.

The weight applied to the events is the total MC weight, lepton scale factors, cross sections incl. k-factors, efficiency filters, and the weight is scaled to $3.2\,\text{fb}^{-1}$. This is the same weight as for the samples with PRW but without the rescaling applied above. To have a peak to fit against, the signal is scaled by a factor of 10 throughout the rest of the study. Only the 1.5 TeV HVT decaying to WW signal has been considered for this study.

	Signal				Backgrounds			
Cut	Events	Events lost	Eff. (%)	Acc. eff. (%)	Events	Events lost	Eff. (%)	Acc. eff. (%)
Start	252	0	100	100	331314	0	100	100
$1\,\text{TeV} < m_{lvJ} < 2.5\,\text{TeV}$	250	2	99	99	11303	320011	3	3
Restart	250	0	100	100	11303	0	100	100
$400\,\text{GeV} < p_{TW} < 800\,\text{GeV}$	219	31	88	88	4261	7042	38	38
Remove buggy ThrustMin	218	1	100	87	4259	2	100	38
At least two subjets	199	19	91	80	3823	436	90	34
Trigger match	196	3	99	79	3781	43	99	33
Remove extreme weights	196	0	100	79	3614	167	96	32
$35\,\text{GeV} < m_W < 150\,\text{GeV}$	190	6	97	76	2291	1322	63	20
Zero b-jets	184	6	97	74	1591	700	69	14

Table 2.5: Cutflow table after preselection for data and MC using no PRW. The signal is only the 1.5 TeV HVT decaying to WW. The numbers are the background and scaled signal yields. The cuts are explained in the text. The line "restart" is simply put in to allow for a better view of the accumulated background efficiency.

The cutflow table for the ntuple creation without PRW can be seen in Table 2.5.

The least efficient cut on signal is the p_T range cut. This has been done to reduce any p_T dependence of the machine learning scores that will be created in the next sections.

The thrust minor substructure variable will in very few cases give an invalid value. The reason for this has not been determined, and the very few events with the invalid values have been removed without significant loss.

To recreate the angles between the subjets correctly, at least two track jets would have had to be within the fatjet area during ntuple-creation.

The background samples also contain a few events with very high weights that will distort the histograms. A proper study into this should be performed, but for this study the events with extreme weights have been discarded.

2.3 Tagging W bosons

The substructure variables used in this study were introduced in Section 1.4.5. The list is not exhaustive.

Before the full study, several more variables were tested as well, many also based on p_T and ΔR of the jet constituents. However, the variables that ultimately were not included in this study showed weak or no separation power, and removing them did not worsen the performance of the pre-study BDT.

Some variables have also not been studied. Most notably is the Shower Deconstruction (SD) algorithm. The algorithm will reconstruct the shower history and compare it against templates of known signals and backgrounds. Ongoing studies within ATLAS are showing that SD complements the other variables nicely and adds greatly to the background rejection. Unfortunately, due to complications in the configuration, the variable did not make it into the study.

All the classifiers in this section have been trained in the small mass range of $65\,\text{GeV} < m_J < 95\,\text{GeV}$ and the full range of $35\,\text{GeV} < m_J < 150\,\text{GeV}$. However, the final performance of the small range classifier was greater than the full range classifier. Therefore, the small range classifier was chosen. The rationale is that the classifier's (reduced) performance outside the signal peak will be much less significant, when the full likelihood fit is performed.

For future reference, the W classifier and the variable for the classifier score after a single evaluation will be referred to as *mva_boson*.

2.3.1 Input variables

All substructure variables have been added to a TMVA Fisher algorithm. The distributions can be seen in Figure 2.15.

The linear correlations can be seen in Figure 2.16.

The heavy correlations suggest that we can do without that many variables. To test the need for these variables, we will first calculate the separation strengths of the variables individually. The result can be seen in Figure 2.17, where the variables are sorted in descending order of separation strength.

We will now construct a Fisher that uses all of the 12 variables, and then we will subtract the least separating variable one-by-one (in the order given by Figure 2.17), re-training the Fisher each time. The separation power of the classifiers can be seen in Figure 2.18(a).

We immediately see that removing the least separating variable has a huge impact on the performance of the classifier. In Figure 2.18(b) the width variable is the fifth in the list. The performance of the classifier again drops significantly when we remove width and only keep the four first variables as listed in Figure 2.17.

We can conclude that only keeping $D2$, τ_{21}^{wta}, $\sqrt{d_{12}}$, $\sqrt{z_{12}}$ as well as width will give the strongest classifier using the least variables with only an insignificant loss in performance compared to the best case.

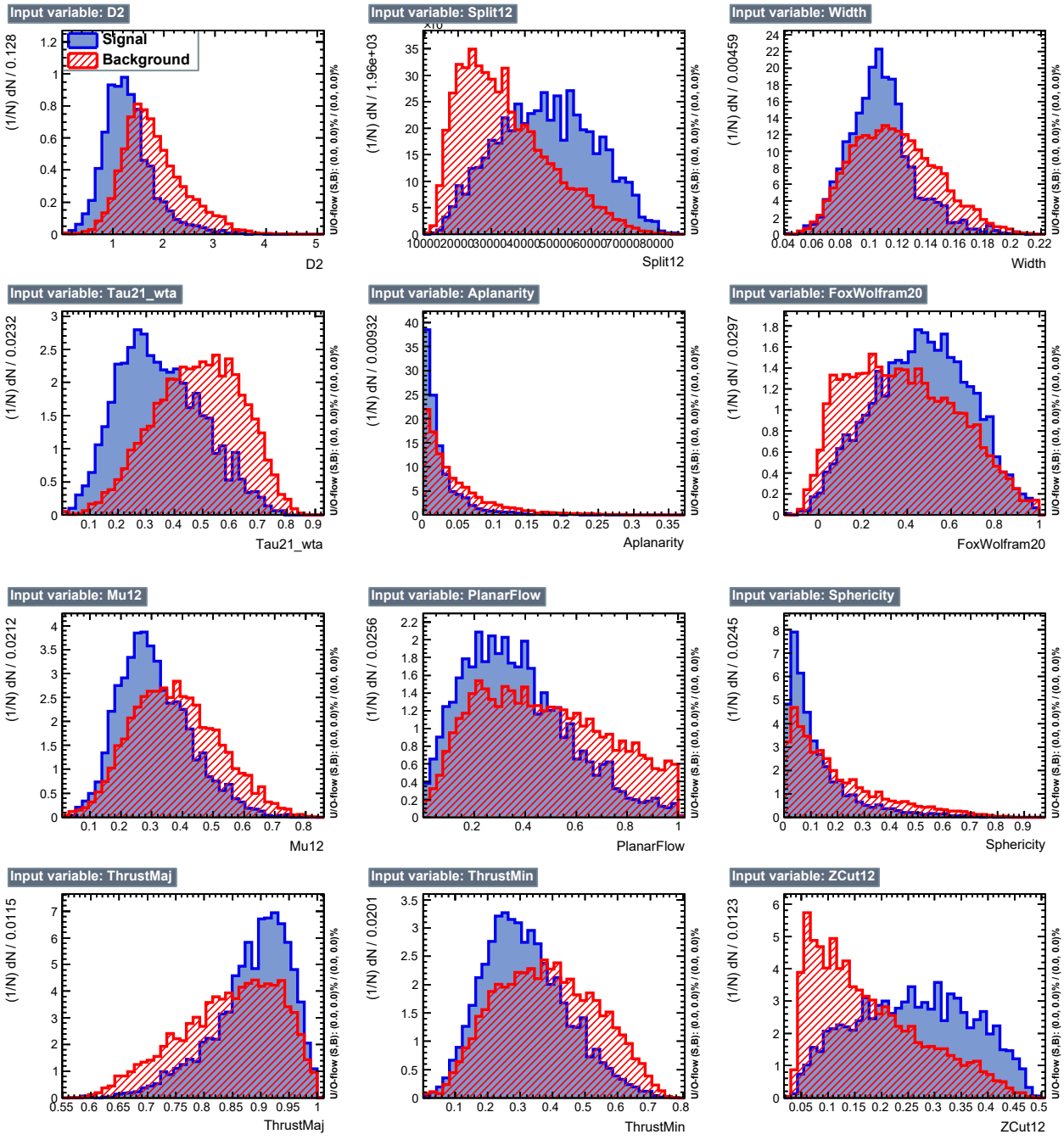


Figure 2.15: The input variables for *mva_boson*.

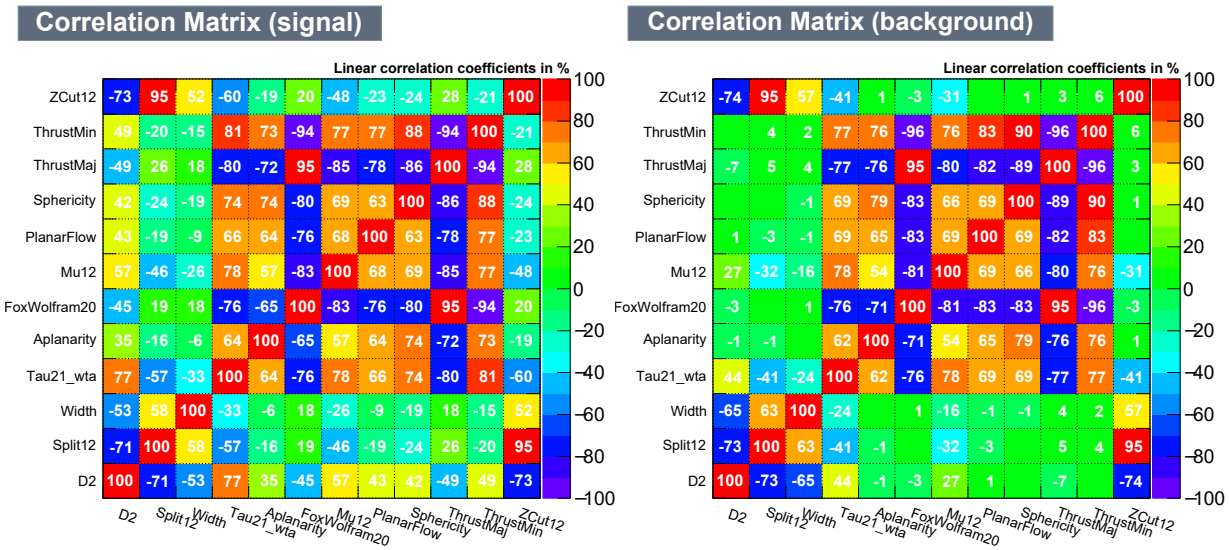


Figure 2.16: The linear correlations for *mva_boson*.

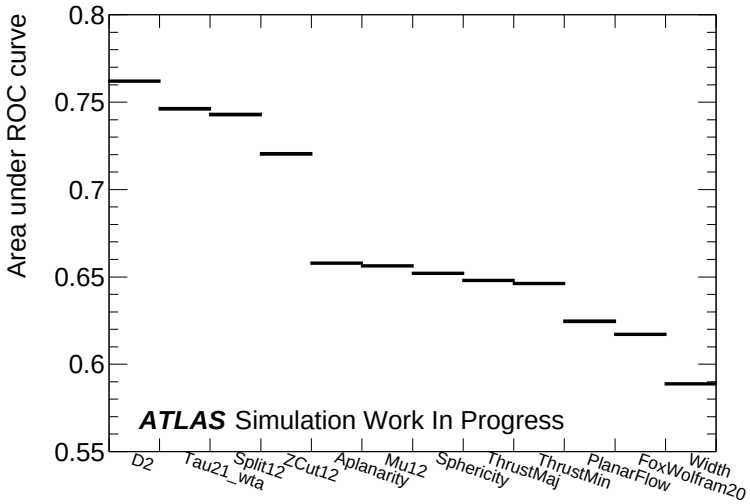
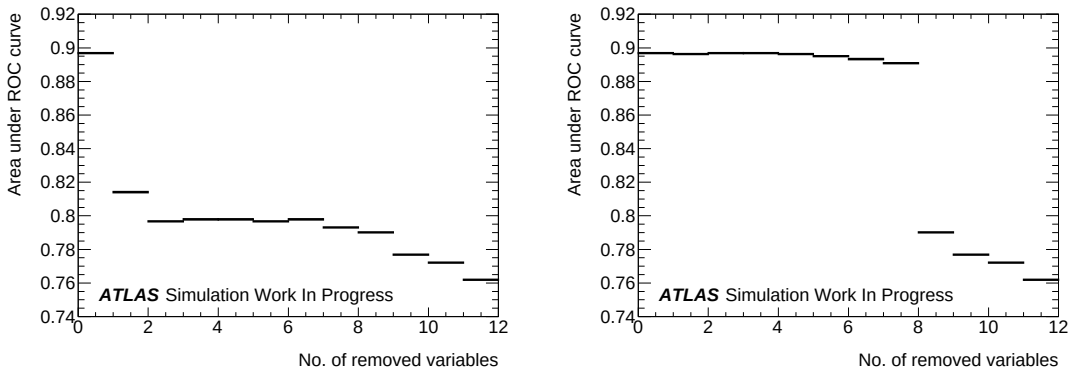


Figure 2.17: The area under the ROC curve for the variables for *mva_boson*.



(a) Width variable last.

(b) Width variable fifth.

Figure 2.18: The area under the ROC curve for a Fisher trained on all variables where the least separating variables are removed one-by-one according to Figure 2.17 in (a) and with width as the fifth variable in (b) for *mva_boson*.

For the study, however, all 12 variables have been used.

2.3.2 Optimizing the BDT

Most of the correlations between the input variables are approximately linear, so we only expect the BDTs to perform marginally better. Before comparing the BDTs to the Fisher method, we want to find the best hyper-parameters. Below is a list of the hyper-parameter values which will be tested:

$$\begin{aligned} \text{NTrees} &= 100 - 1600 \text{ in steps of } 100, \\ \text{nCuts} &= [5, 10, 20, 50, 100, 250, 500, 1000], \\ \text{MaxDepth} &= [3, 5, 7, 15], \\ \text{MinNodeSize} &= [0.001, 0.05, 0.1, 1, 5, 15, 25], \\ \text{AdaBoostBeta} &= [0.001, 0.05, 0.1, 0.5, 1], \end{aligned}$$

in p_T bins of $[200, 300, 400, 500, 600, 700, 800, 900, 1100]$. For this to work we have undone the p_T range cut that was introduced in Section 2.2.5.

A popular method is the *grid search* which runs the algorithm with every combination of the above hyper-parameters. The performance measure is usually performed through the *n-fold cross-validation*[8]. However, ROOT does not support cross-validation yet.

Instead, we run the BDT with every combination for 16 different values for *SplitSeed*[9]. This gives us 16 different classifiers. This method does not split the data samples into 16 independent training samples. This means that the classifiers will be correlated, as some training data points in one classifier will also be in another.

We will measure performance by the area under the ROC curve as well as the ratio of the training efficiency to the test efficiency at a background efficiency of 10%. The efficiencies are calculated from the ROC curves curves computed from the training and testing data samples, respectively. ROOT evaluates the classifiers using the whole data sample when calculating the integral of the ROC curve.

Without showing the results from every classifier, the two best performing classifiers are presented:

Type 1 (default settings):

$$\begin{aligned} &\text{NTrees} = 800,\ \text{MaxDepth} = 3,\ \text{nCuts} = 20, \\ &\text{MinNodeSize} = 5\%,\ \text{AdaBoostBeta} = 0.5 \end{aligned} \tag{2.8}$$

Type 2 (best training performance to test ratio):

$$\begin{aligned} &\text{NTrees} = 800,\ \text{MaxDepth} = 3,\ \text{nCuts} = 100, \\ &\text{MinNodeSize} = 1\%,\ \text{AdaBoostBeta} = 0.1 \end{aligned} \tag{2.9}$$

Type 3 (greatest ROC curve integral):

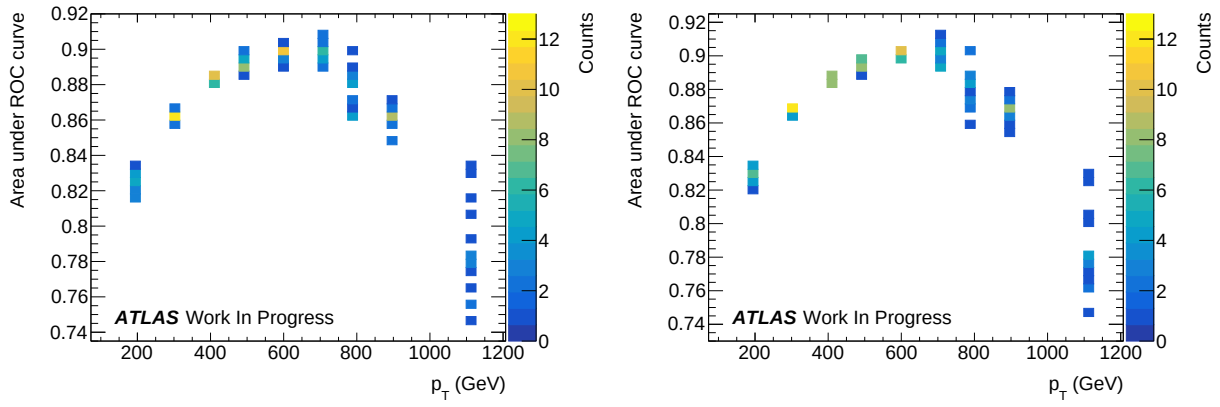
$$\begin{aligned} &\text{NTrees} = 300,\ \text{MaxDepth} = 5,\ \text{nCuts} = 100, \\ &\text{MinNodeSize} = 1\%,\ \text{AdaBoostBeta} = 0.1 \end{aligned} \tag{2.10}$$

[8] The data sample is divided into n even folds. The algorithm is run on the first $1/n$'th of the data sample, then on the second, etc. The other $n-1$ folds are left as test data in every run.

[9] ROOT splits the data sample in half, one for training and one for test. The data points are randomly placed in the training and test samples given the seed.

The results show that the value of nCuts has little to no effect on the performance. MinNodeSize and AdaBoostBeta consistently improve performance at their respective values of 1% and 0.1.

The ROC curve integrals for the two types are shown in Figure 2.19.

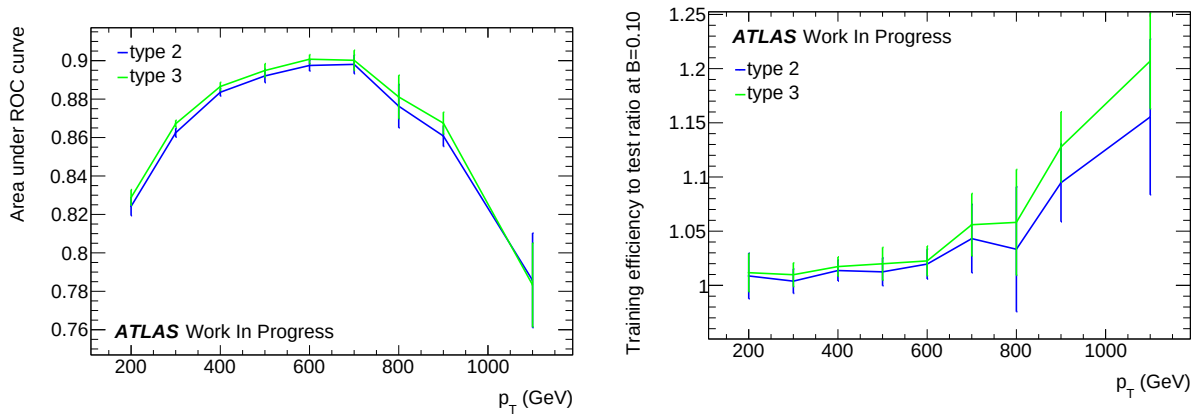


(a) Type 2.

(b) Type 3.

Figure 2.19: The area under the ROC curve for the 16 iterations for type 2 (a) and type 3 (b) for *mva_boson*.

The spread in performance for the different values of SplitSeed makes type 2 and 3 indistinguishable. This effect is better seen in Figure 2.20(a).



(a) Area under ROC curve as function of p_T.

(b) Training performance to test performance ratio at background efficiency of 0.10 as function of p_T.

Figure 2.20: Area under ROC curve (a) and training/test efficiency ratio (b) for the two types for *mva_boson*. In the left plot, a higher value is better (greater separation in the training sample). In the right plot, a lower value is better (better agreement between training and test results).

The two types have been compared in Figure 2.20. The points at each p_T are the average and RMS of the 16 classifiers. As mentioned earlier, the classifiers are correlated, so the RMS is underestimated. The figure shows the two types to be within the (underestimated) uncertainties. Figure 2.20(b) shows the ratio of the training to test efficiency at a background efficiency of 0.10. Again, the two types overlap.

2.3.3 Results

Even if the improvement gained by adjusting the BDT hyper-parameters is dubious, we will still consider the two BDTs as well as the BDT with the default configuration.

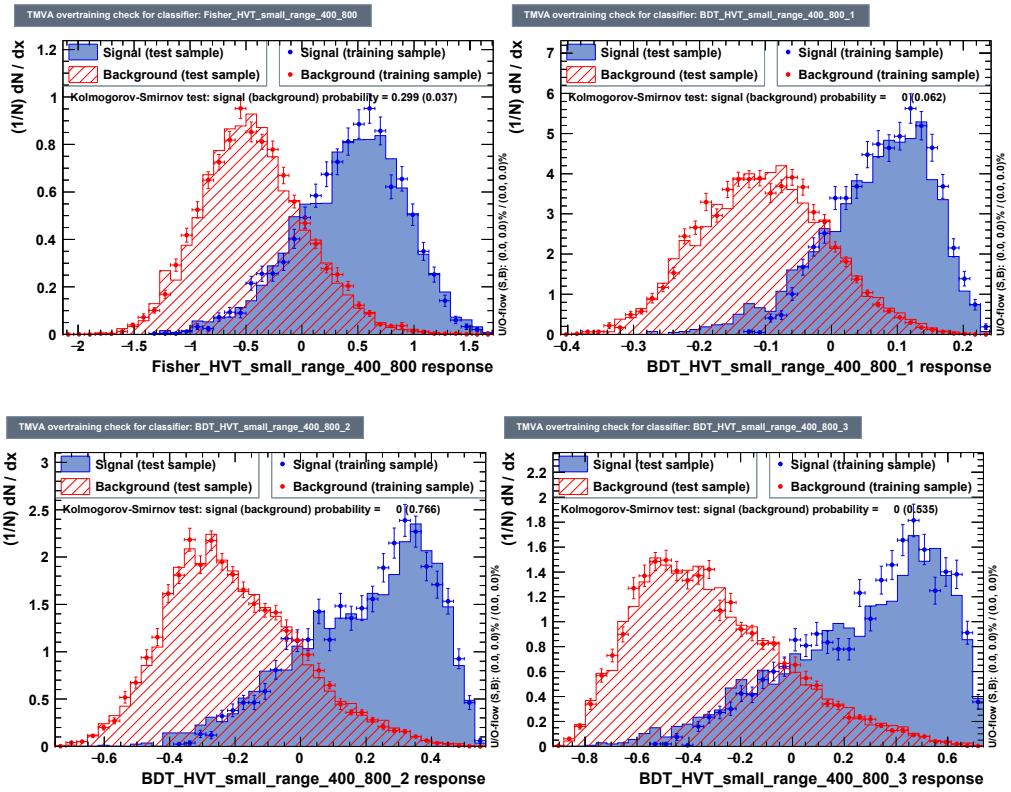


Figure 2.21: Scores from the trained classifiers for *mva_boson*.

Figure 2.21 shows the classifier scores after training. The Fisher discriminant shows little over-training while the default BDT (named BDT_HVT_small_range_400_800_1) is badly over-trained. The BDT types 2 and 3 show over-training to a lesser degree.

The ROC curves for the four methods are presented in Figure 2.22, where it is obvious that BDTs are not able to better separate signal from background.

Therefore, because of its better resilience against over-training, the Fisher-based discriminant will be used in the final analysis.

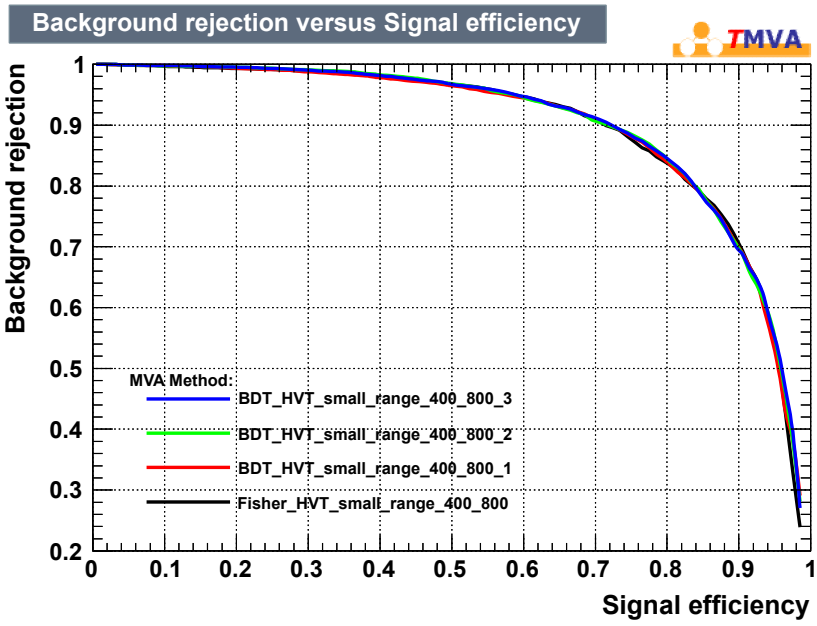


Figure 2.22: The ROC curves for the four classifiers for *mva_boson*.

2.4 Full event tagger

mva_boson will only be able to discriminate hadronically decayed W particles from background. The study will be focused on the full decay chain $X \to WW \to lvJ$. The leptonically decayed W particles are easily found[10], and a classifier for the hadronically decayed W particles has now been constructed.

However, there is still information left in the event. The cosines of the angles defined in Section 2.2.3 show separation power and are therefore included in the coming event classifier. The decaying parents, X and the two W bosons, will also distribute their energy approximately evenly among their daughters. The p_T-*balances* are defined as the p_T of the two W particles, the lepton, and the neutrino of the event each divided by the reconstructed m_{lvJ}.

For future reference, the full event classifier and the variable for the classifier score after a single evaluation will be referred to as *mva_event*.

The variables can be seen in Figure 2.23.

[10] Given the lepton and MET cuts, the lep. W is actually *assumed* to be the parent of the lepton and reconstructed neutrino. A cut on the transverse mass of the lep. W candidates to actually identify them shows little significance gain.

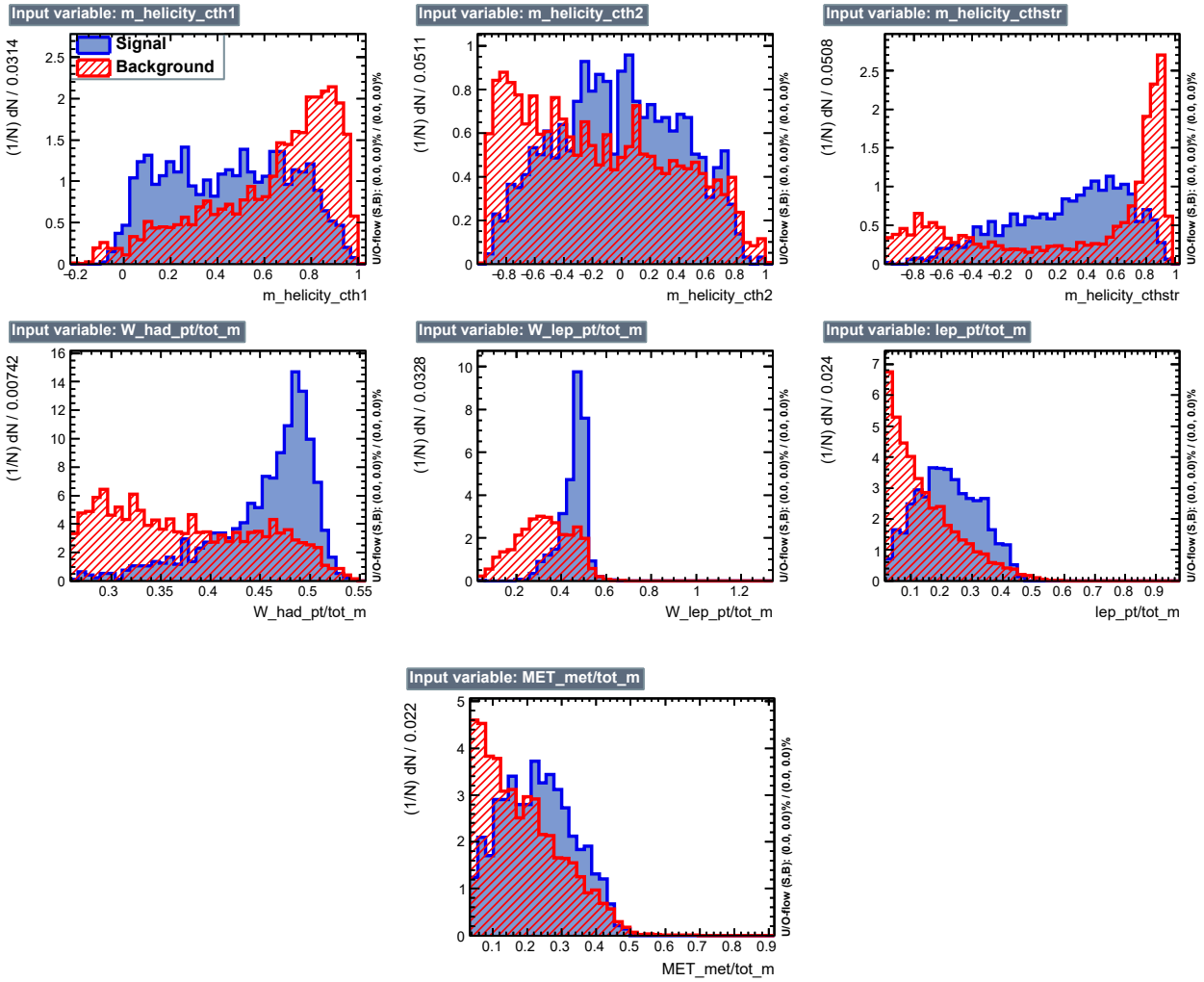


Figure 2.23: The input variables for *mva_event*.

The linear correlations are given in Figure 2.24. A Fisher and three BDTs are trained with the same configurations as for *mva_boson*. The

result can be seen in Figure 2.25.

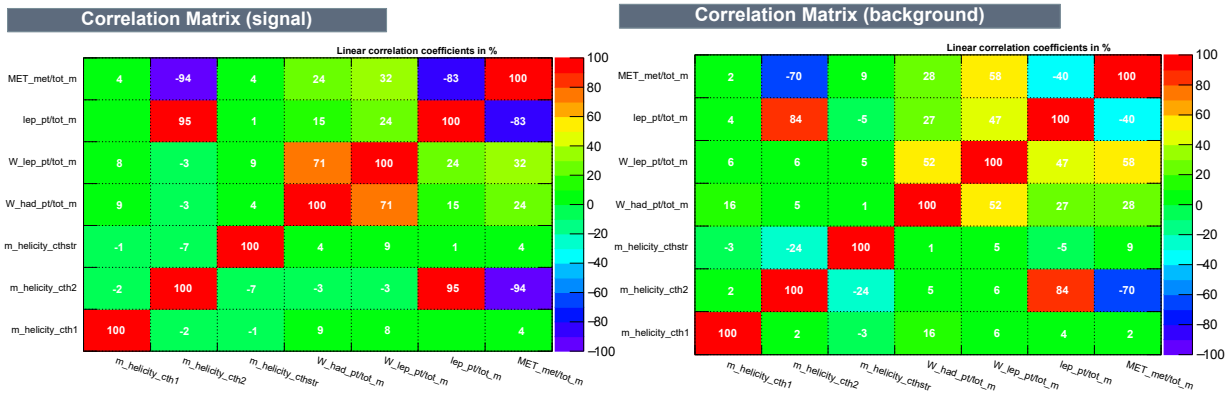


Figure 2.24: The linear correlations for *mva_event*.

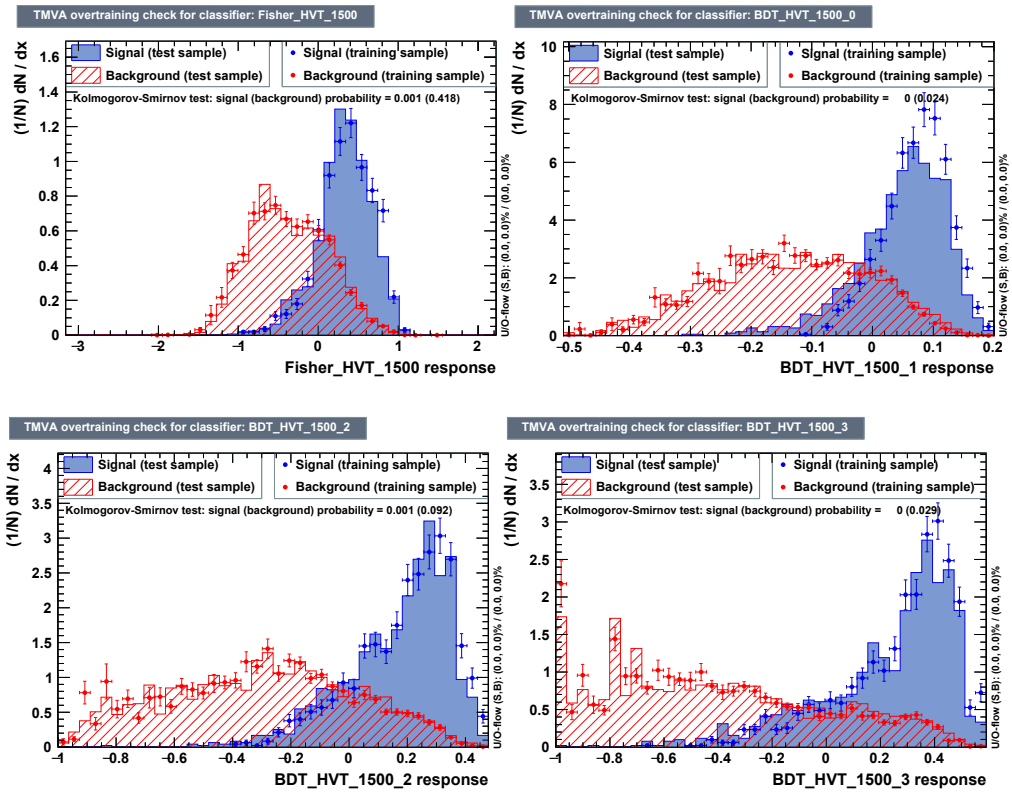


Figure 2.25: Scores from the trained classifiers for *mva_event*. Only the HVT hypothesis with $m = 1.5\,\text{TeV}$ is shown.

From the ROC curve in Figure 2.26 below, we see that the Fisher classifier performs slightly worse than the BDTs. However, the BDT with the default configuration is badly over-trained again. The two other BDTs also show worse consistency between training and test data for low scores compared to the Fisher.

The Fisher classifier is also chosen for *mva_event* for its slightly lesser over-training and its shape. Ultimately, we will want to fit the output, and two Gaussians will fit the Fisher quite well.

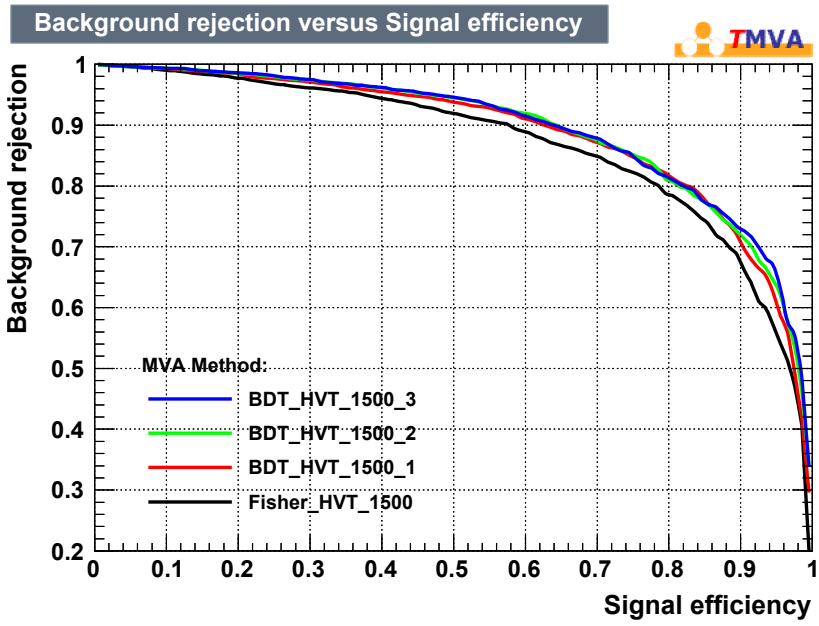


Figure 2.26: The ROC curves for the four classifiers for *mva_event*.

2.5 The likelihood fit

With the MC ntuple and the two taggers ready, we can now perform the analysis. This section will in steps move towards a simultaneous likelihood fit in our two tagger scores as well as m_W and m_{lvJ}. We will perform a series of cuts to validate our data and bring confidence in the integrity of our full fit in the end.

We will only consider *mva_boson* and *mva_event* in the range $[-1.5, 1.5]$. The signal and background yields are therefore:

$$N_s = 180 \pm 13, N_b = 1482 \pm 38, \tag{2.11}$$

where the uncertainties are assumed Poissonian.

We have a slightly smaller yield compared to Table 2.5, because of our cuts on the tagger ranges.

2.5.1 Correlations

It is assumed that the four variables used in the likelihood fit are uncorrelated.

It is, however, not entirely the case. The score from *mva_boson* correlates with the had. W mass. The effect can be seen in Figures 2.27(a-b). Each m_W bin has been averaged and plotted on top, and the profiles are fit with a linear polynomial in the range $65\,\text{GeV} < m_W < 95\,\text{GeV}$. The results for the first order polynomial fits ($y = ax + b$) are:

$$\begin{aligned} &a = 0.0122(7), \quad b = -0.74(6), \quad p = 0.013 \quad \text{(signal)}, \\ &a = 0.0137(4), \quad b = -1.26(3), \quad p = 0.002 \quad \text{(background)}, \end{aligned} \tag{2.12}$$

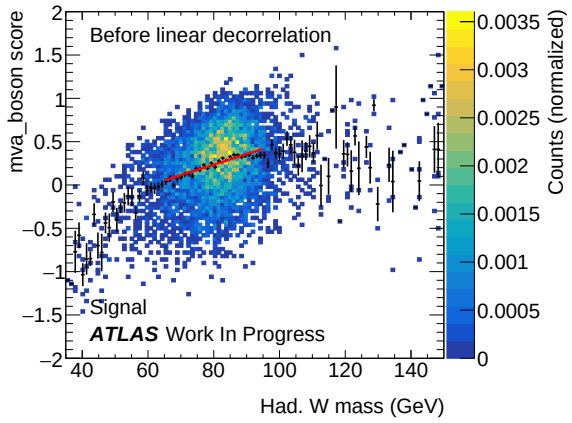
where p are the p-values obtained from the fits.

Although the slopes are not the same, the background is chosen to be entirely decorrelated (at the expense of an almost decorrelated signal). Therefore, the following transformation is performed:

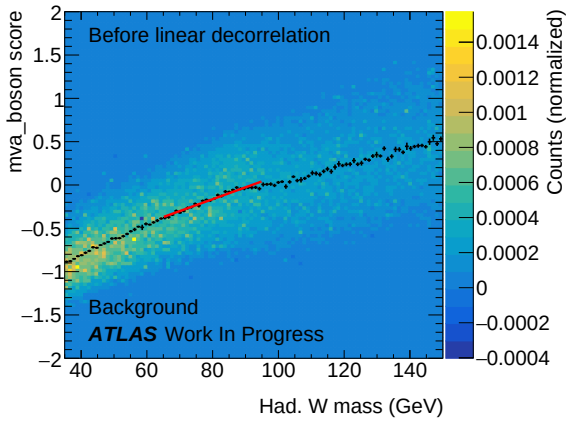
$$mva_boson \rightarrow mva_boson - (0.0137 m_W - 1.26). \tag{2.13}$$

The results of the decorrelation can be seen in Figures 2.27(c-d).

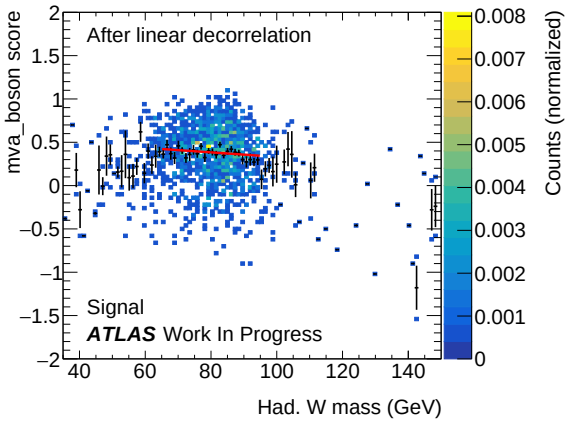
All six correlations (scatter profiles with linear correlation coefficients) can be seen in Figures 2.28 and 2.29. Please note that they only include the decorrelated *mva_boson* values.



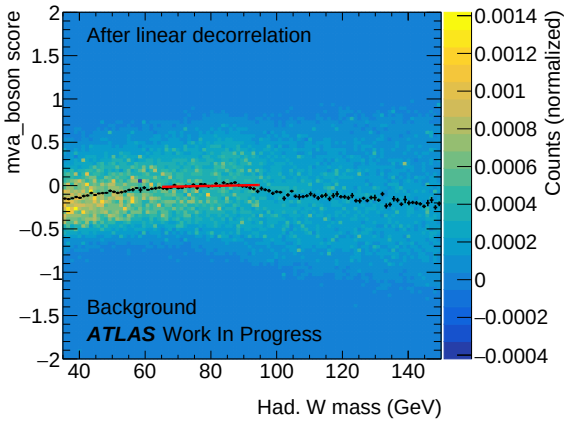
(a) Signal before decorrelation.



(b) Background before decorrelation.



(c) Signal after decorrelation.



(d) Background after decorrelation.

Figure 2.27: Scatter plots of *mva_boson* and had. W mass for signal and background before and after decorrelation using Equation 2.13. The m_W bins have been averaged and plotted on top, and the profiles are fit with a linear polynomial in the range $65\,\text{GeV} < m_W < 95\,\text{GeV}$.

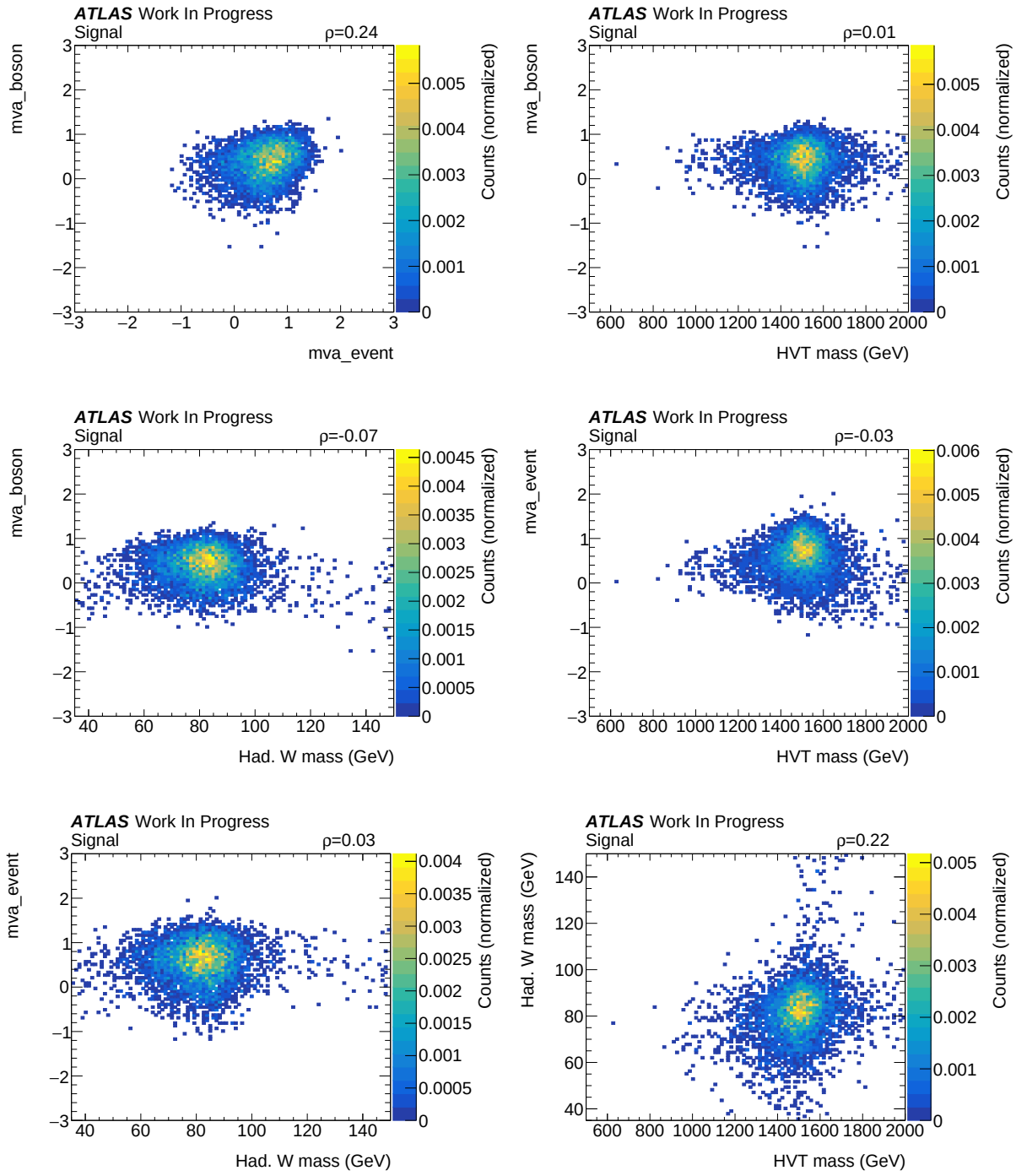


Figure 2.28: Correlations between the four variables used in the likelihood fit. Please note that they only include the decorrelated *mva_boson* values. The linear correlation factor, ρ, is also shown. The signal is the 1.5 TeV HVT particle.

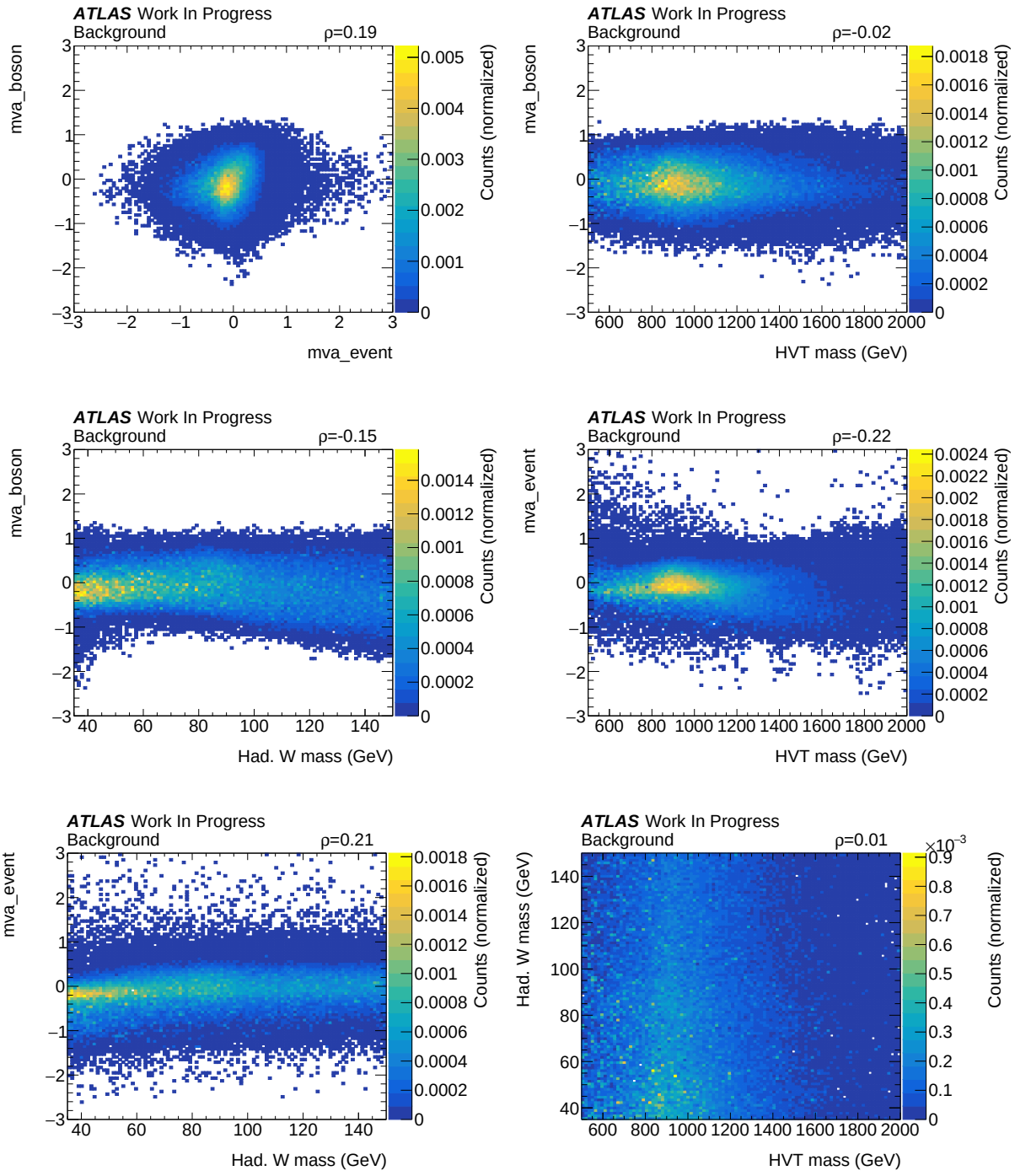


Figure 2.29: Correlations between the four variables used in the likelihood fit. Please note that they only include the decorrelated *mva_boson* values. The linear correlation factor, ρ, is also shown. The background contains all the major channels (W+jets, $t\bar{t}$, etc.).

2.5.2 4 cuts

Before creating our likelihood, we will make a benchmark. For the four variables in question, we will perform cuts on *mva_boson* and *mva_event* that maximize the significance, $\frac{S}{\sqrt{S+B}}$. We will cut the two masses, m_W and m_{lvJ}, around their peak values[11] with the ranges $\pm 15\,\text{GeV}$ and $\pm 150\,\text{GeV}$, respectively.

[11] This ought to be the nominal/theoretical mass. However, the reconstructed mass is not guaranteed to be the same.

The cuts will be performed in the following order:

- $68\,\text{GeV} < m_W < 98\,\text{GeV}$.
- $mva_boson > x$, where x is a value that maximizes $\frac{S}{\sqrt{S+B}}$.
- $mva_event > y$, where y is a value that maximizes $\frac{S}{\sqrt{S+B}}$.
- $1350\,\text{GeV} < m_{lvJ} < 1650\,\text{GeV}$.

The next cut is applied on the subset of data created by the previous cut.

All the following figures will be of signal and background of the masses drawn individually. They will show the signal and background yields as well as the signal z-scores and significances only in the above ranges. The background will be given in red, and the signal in dark blue. The figures will include the histograms before the last cut in a lighter color and with no filling. Vertical, dashed lines around the signal peaks indicate the range in which the yields, z-scores and significances have been calculated.

For the figures of the tagger scores, no previous histogram is underlaid. The peak of the significance curves are fitted with a second order polynomial to reduce sensitivity to noise. A dashed line is drawn through the figure to show the optimal cut. The figures will include their ROC curves upon which the point of the maximally significant cut is drawn.

To calculate the z-scores, the signal, and background histograms have been added and have had the background subtracted again in the given ranges. Since the uncertainties are Poissonian, the "extracted" signal uncertainty will be: $\sigma_S^2 = S + 2B$. We will therefore see an increase in the signal z-score by decreasing the background in the signal range.

The initial figures for the masses can be seen in Figures 2.30(a-b).

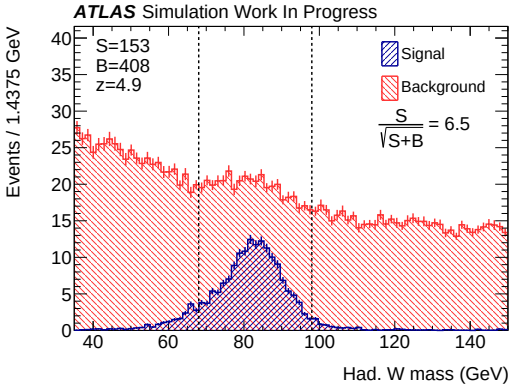
The first cut will then be applied; m_W is cut in the range $68\,\text{GeV} < m_W < 98\,\text{GeV}$. After applying the cut, we see an increase in the signal z-score and significance in m_{lvJ} in Figure 2.31(b), as would be expected.

Next, we apply the cut on *mva_boson*. The significance of a cut on it is shown against its signal and background distributions in Figure 2.32(a). Figure 2.32(b) shows the ROC curve.

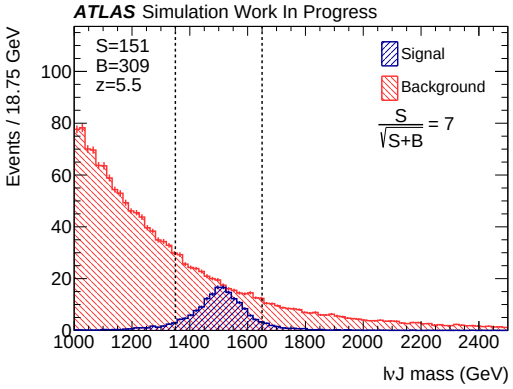
The impact of the cut on the masses can be seen in Figures 2.33(a-b).

We finally apply the cut on *mva_event*. The result is seen in Figure 2.34.

The mass distributions after the three cuts are seen in Figures 2.35(a-b).

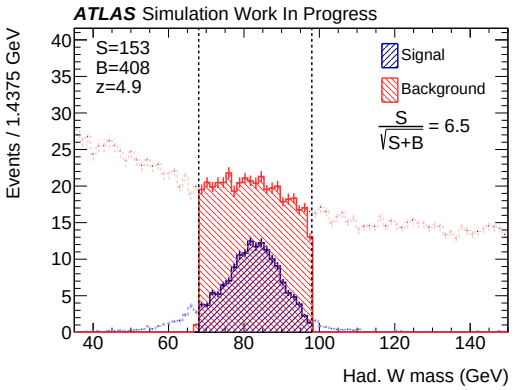


(a) m_W

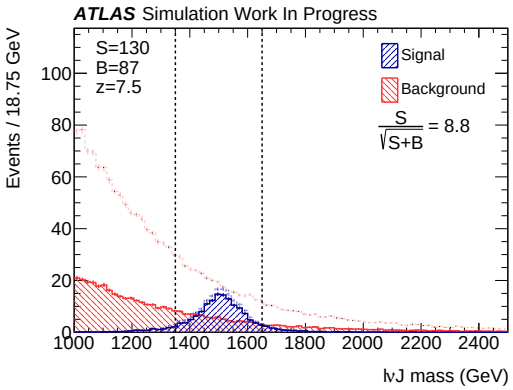


(b) m_{lvJ}

Figure 2.30: m_W and m_{lvJ} before any cutting on the tagger scores or any of the masses. See the text for detailed explanation of the figures.

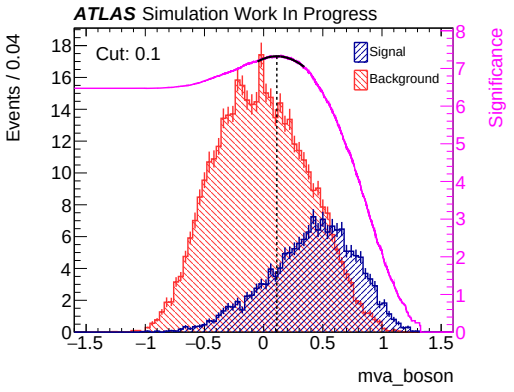


(a) m_W

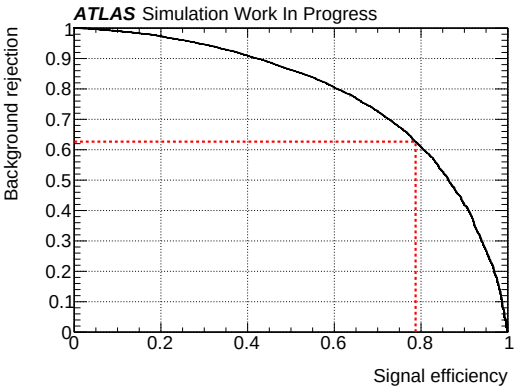


(b) m_{lvJ}

Figure 2.31: m_W and m_{lvJ} after cutting on the m_W mass. See the text for detailed explanation of the figures.

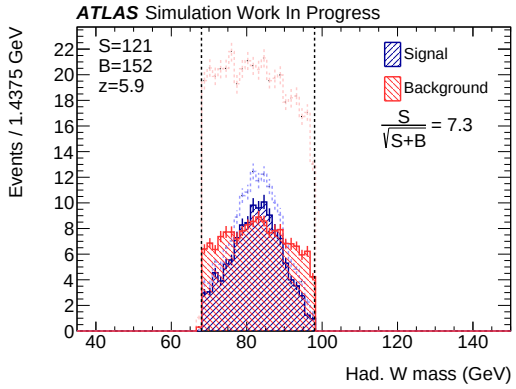


(a) *mva_boson*

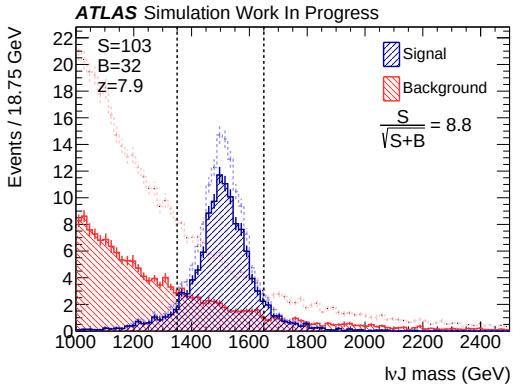


(b) ROC-curve

Figure 2.32: (a) *mva_boson* for signal and background with the cut significance overlaid, and (b) their ROC curve. See the text for detailed explanation of the figures.

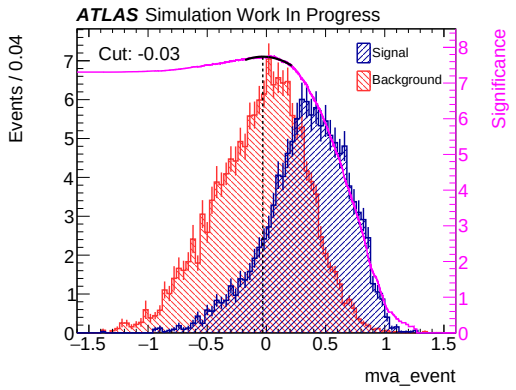


(a) m_W

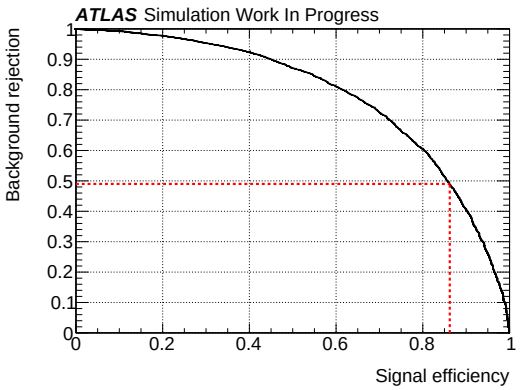


(b) $m_{l\nu J}$

Figure 2.33: m_W and $m_{l\nu J}$ after cutting on the m_W mass and *mva_boson*. See the text for detailed explanation of the figures.

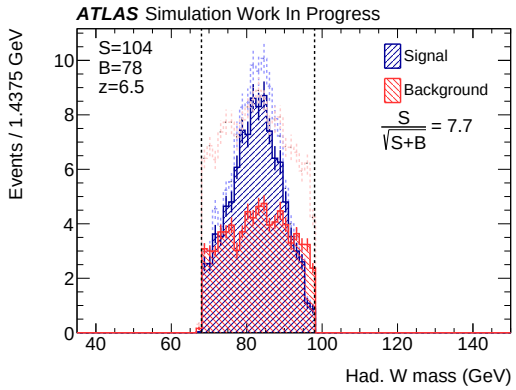


(a) *mva_event*

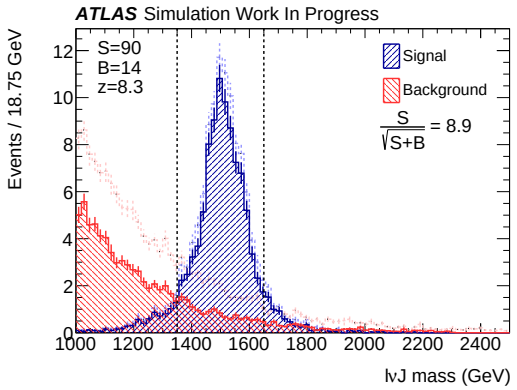


(b) ROC-curve

Figure 2.34: (a) *mva_event* for signal and background with the cut significance overlaid, and (b) their ROC curve. See the text for detailed explanation of the figures.



(a) m_W



(b) $m_{l\nu J}$

Figure 2.35: m_W and $m_{l\nu J}$ after cutting on the m_W mass and the two tagger scores. See the text for detailed explanation of the figures.

For the fourth cut, we extract the final z-score from Figure 2.35(b):

$$z_{4\text{ cuts}} = 8.3 \text{ for } 90 \text{ signal events.} \tag{2.14}$$

2.5.3 3 cuts

The result of the previous exercise was a simple cut-and-count, where we assumed pure Poisson errors. If we instead fit the final figure of m_{lvJ}, we expect to preserve its tails that a cut will inevitably remove.

The fit will be done on the sum of the signal and background distributions by RooFit [31]. The full likelihood will be a product of the background and signal PDFs (normalized to the number of background and signal events, respectively) as well as a Poisson distribution for each of them. This *extended* likelihood will, along with the Poisson distributions and when maximized, give the background and signal yields directly including errors. The fit parameters will be floating freely. The minimization of the negative log likelihood will be done with ROOT's Minuit's MIGRAD, HESSE, and MINOS routines that will find the minimum, calculate the full Hessian matrix, and calculate asymmetric errors, respectively.

The background will be fit against the following PDF:

$$p_{\text{bkg}} = \left(1 - \frac{m}{14\,\text{TeV}}\right)^{a}, \tag{2.15}$$

where a is the sole fit parameter. The equation is chosen for its simplicity and ability to model the background well.

The signal will be fit with a Gaussian with asymmetric uncertainties as well as a Breit-Wigner distribution convoluted with a Gaussian. They will all share the same mean.

The result of the fit is shown in Figure 2.36.

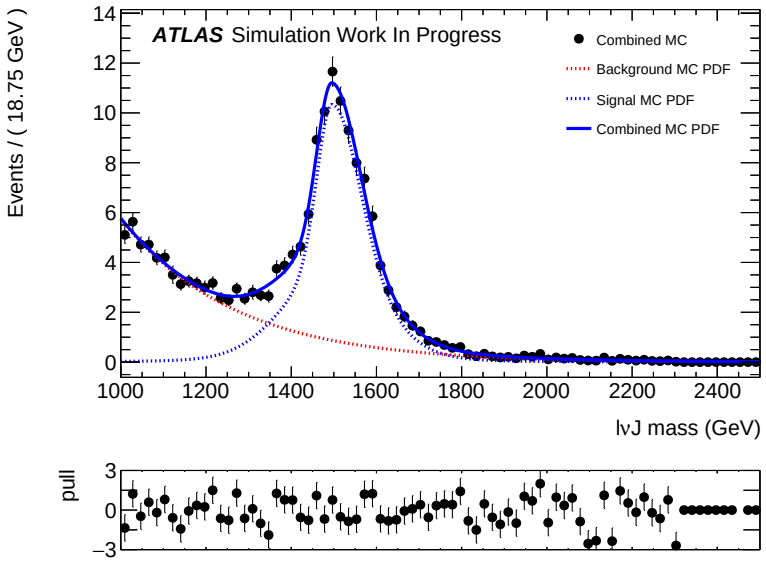


Figure 2.36: m_{lvJ} after cutting on the m_W mass and the two tagger scores. The distribution is fit with the PDFs given in the text.

The likelihood fit gives us $N_s = 10 \pm 13$, where we calculate the

z-score using the negative uncertainty:

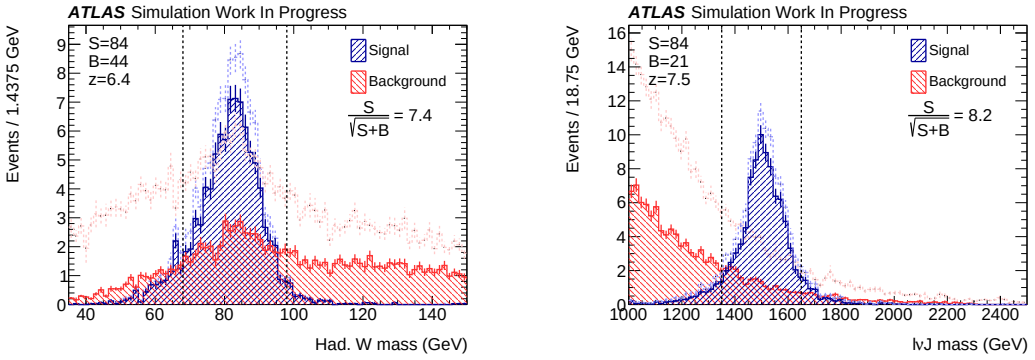
$$z_{3\text{ cuts}} = 101/13 = 7.9 \text{ for 101 signal events.} \tag{2.16}$$

We increased our signal count. However, the error reported by RooFit is significantly larger than the square root of the signal count. This could be because root propagates the errors differently.

2.5.4 2 cuts

The amount of information available to the fit could be increased. Instead of cutting on m_W, we could fit it simultaneously with m_{lvJ} to possibly increase our z-score.

We remove m_W from the cutting list, which means that we have to redo all the cutting steps to find new significance optima. However, after cutting on the two tagger scores, the background shape of m_W is changed drastically, as seen in Figure 2.37(a).



(a) m_W

(b) m_{lvJ}

Figure 2.37: m_W and m_{lvJ} after cutting on the two tagger scores. See the text for detailed explanation of the figures.

Attempts were made to fit the background with a small Gaussian on top of a polynomial. However, polynomial PDFs in RooFit are difficult to match to data, and it was therefore not possible to fit the background.

Figure 2.38 shows the background from Figure 2.37(a) split into its components drawn in front of (and not on top of) each other. The W+jets background is expected to fall off exponentially, however the *mva_boson* cut (the light red histogram of Figure 2.37(a)) changed the shape to peak around the W mass. The cut on *mva_event* further removed background at low m_W disproportionally more than at higher. This is unexpected, as the decorrelation was deliberately made such that the *mva_boson*-m_W correlation would be flat in background.

The answer might be correlations. We see from Figure 2.27(d) that the background moves slightly towards negative *mva_boson* for lower values of m_W. A cut on *mva_boson* will therefore remove relatively more background in the lower m_W region than around the W mass.

Toy data from histograms of the variables would have to be generated. *mva_boson* would then be used on the toy data. Since the toy data is uncorrelated by construction, the background should scale down independently of m_W.

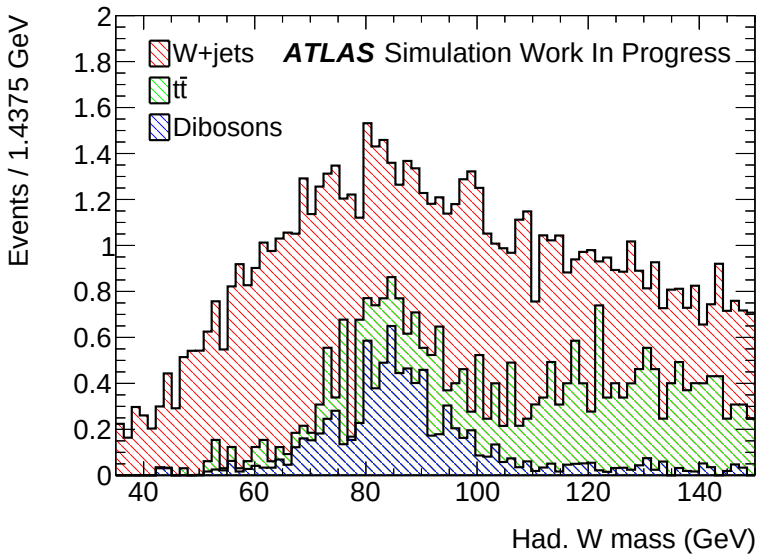


Figure 2.38: m_W after cutting on the two tagger scores. The backgrounds are drawn in front of (and not on top of) each other.

2.5.5 0 cuts

The final exercise will be attempting to perform no cutting and to fit directly on all four variables simultaneously. This is a difficult task because of the higher dimensionality. The correlations will have an ever larger impact on the shapes when performing the simultaneous fit. The rather complex shapes of the tagger score distributions will also be troublesome to fit in one dimension, and they will be even more difficult to stay fit when fitting in more dimensions.

The background PDF for m_W will be an exponential. The signal PDF will be fit with a Gaussian with asymmetric uncertainties as well as a Breit-Wigner distribution convoluted with a Gaussian. They will all share the same mean.

The background PDF for *mva_boson* is a double Gaussian with asymmetric uncertainties that do not share a common mean. The choice of a second, free Gaussian comes from the wish to capture the small bump on the left side of the distribution. Its signal PDF is a single Gaussian with asymmetric uncertainties.

The background PDF for *mva_event* is also a double Gaussian with asymmetric uncertainties that do not share a common mean. The signal PDF is a double Gaussian that share a common mean.

The PDFs for the tagger score distributions have proven difficult to stay stable when their parameters are allowed to float. Therefore, they have been fitted individually, background and signal independently, and their fit parameters have afterwards been fixed entirely.

The two mass distributions have also been fitted individually, background and signal independently. And their fit parameters will also be fixed in the simultaneous fit.

The simultaneous fit in all four variables yields $N_s = 190 \pm 16$,

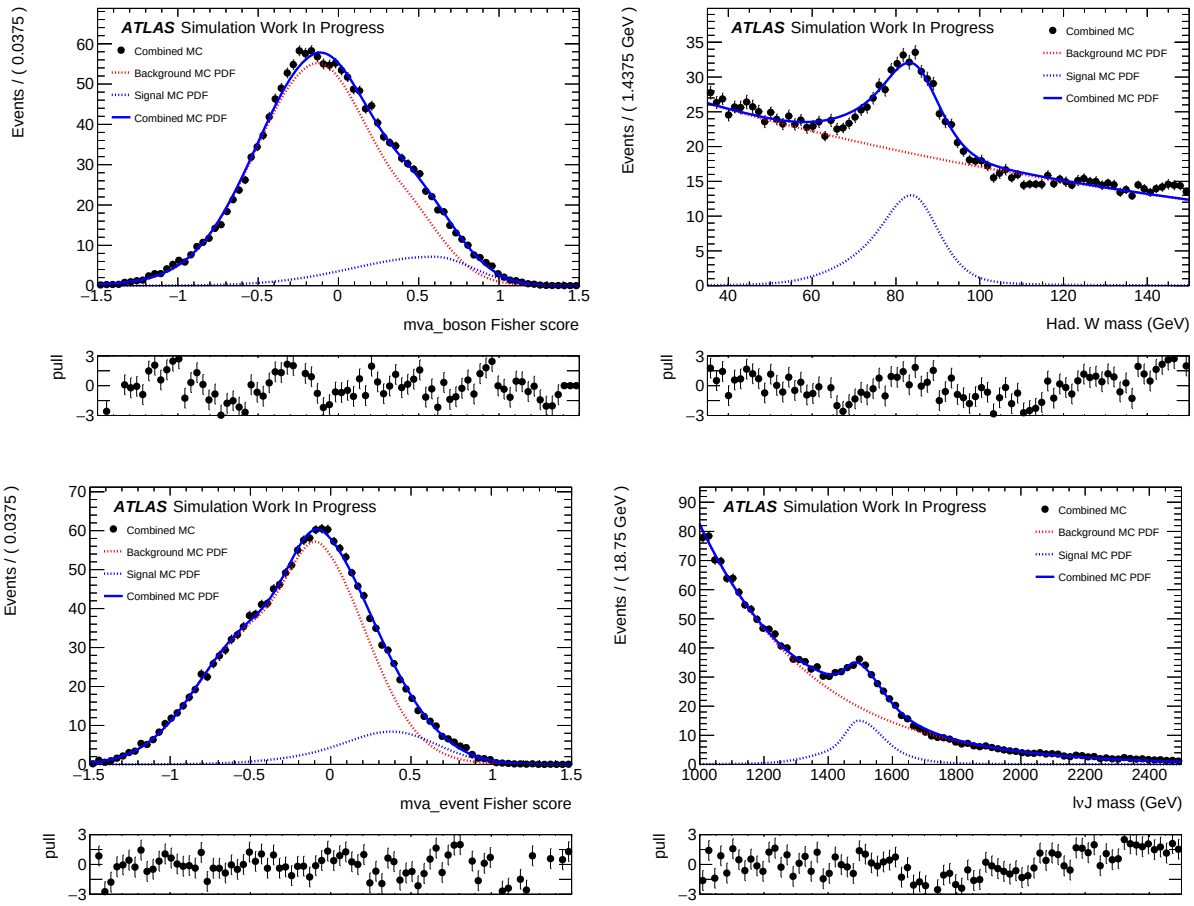


Figure 2.39: *mva_boson*, m_W, *mva_event*, and m_{lvJ} are first fit individually, signal and background individually as well. In the figure, the sum of the signals and backgrounds have been fit in each variable. Since all PDF parameters are fixed, only the yields are adjusted.

giving us a z-score of:

$$z_{\text{0 cuts, fixed mass}} = 190/16 = 12 \text{ for 190 signal events.} \tag{2.17}$$

The four-dimensional likelihood unfortunately seems to catch a signal yield slightly larger than the available amount of data. It could be that it has recognized some of the W particles from the dibosons and $t\bar{t}$ backgrounds. To test for this, the full PDF will be applied to the same dataset but without the aforementioned backgrounds.

Keeping only W+jets backgrounds

To estimate the actual efficiency of the simultaneous fit, we will now remove the diboson and $t\bar{t}$ backgrounds.

The parameters for the background PDFs will be re-determined in the usual way after which all the parameters will be fixed for the simultaneous fit. We now expect the four-dimensional fit to capture up to the amount of signal available and not beyond.

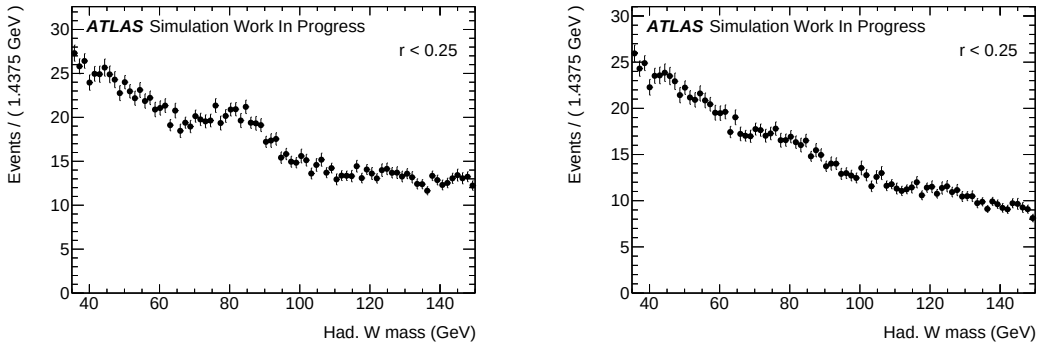
The signal yield is: $Ns = 182^{+16}_{-15}$, giving us a z-score of:

$$z_{\text{0 cuts, only W+jets}} = 182/15 = 12 \text{ for 182 signal events.} \tag{2.18}$$

The signal yield reported by the fit is very close to the amount of signal available. Either the fit is phenomenally finding all the data, or the correlations between the variables come into play. To test for this, toy data generated from histograms of the variables would have to be generated and fitted against. This is an important check for a future study.

2.5.6 Final notes

Likelihood ratios have been constructed for the final PDFs for the 0 cuts approaches with full background and W+jets background only. We expect the W+jets background to remain for low values of $r = \frac{P_{\text{sig}}}{P_{\text{sig}}+P_{\text{bkg}}}$. The result for m_W for $r < 0.25$ can be seen in Figure 2.40.



(a) m_W for 0 cuts with full background

(b) m_W for 0 cuts with only W+jets

Figure 2.40: Comparison of the two 0 cut approaches. m_W is shown for $r < 0.25$. The diboson and $t\bar{t}$ background contribution is still visible in (a).

Comparing the backgrounds in this figure to Figure 2.30(a), we see that the backgrounds in both cases are almost entirely untouched.

The small bump from the diboson and $t\bar{t}$ backgrounds is still visible in Figure 2.40(a). This is good. This means that the four-dimensional fit recognizes the backgrounds as such even if they are actual W bosons. However, this also means that further studies are needed to resolve why the approach of 0 cuts with full background yielded 190 signal events out of 180 available.

The results of the different approaches have been summed up in Table 2.6.

Method	z-score	Signal yield
4 cuts	8.3	90
3 cuts	7.9	101
0 cuts	**12**	**190**
0 cuts (only W+jets)	12	182

Table 2.6: Table of signal yields and z-scores from all the different approaches.

The final and full simultaneous fit is given in bold in the table. We see that the z-score and signal yield is much larger, however issues such as correlations and unaccounted-for-backgrounds contributing to the signal yield will have to be thoroughly examined, included in the PDFs, and quantified.

2.6 Concluding remarks

2.6.1 Summary

This thesis has presented a series of attempts at improving the diboson analyses in the boosted regime.

The jet charge study was an attempt at improving the jet charge calculation of fatjets to an extend where the fatjets charges as well as their subjet charge differences would open a window into W, Z, and background separation. Due to the nature of QCD, calculating the charge of a jet and relating it back to the quark that initiated the jet is not exactly possible. Adding the ΔR dependence to the charge calculation improved the separation of W^+ and W^-. However, the final plots of subjet charge sums and difference showed very little separation power.

The likelihood study was an attempt to make a boson tagger based on a machine learning algorithm, *mva_boson*, and use its output directly in a simultaneous likelihood fit of several variables. *mva_boson* was able to better discriminate between W and background than any of its variables were individually. It was shown that a BDT performs the same as a simple Fisher discriminant-based algorithm. A second machine learning algorithm, *mva_event*, was introduced that would exploit the whole shower history of a particle decaying to two W bosons that would further decay semi-leptonically. The BDT performed slightly better but the Fisher was used again because of its lesser over-training and more easily fitable shape. As a benchmark, the scores of *mva_boson* and *mva_event* as well as m_W and m_{lvJ} were cut in a manner similar to an ordinary search study. Fitting the m_{lvJ} yielded a higher signal, but also a higher uncertainty. The final simultaneous fit on all four variables gave the largest signal yield and z-score. However, the signal yield reported by the final PDF was higher than the actual signal yield. The issue was partly addressed as being due to real W particles from the diboson and $t\bar{t}$ backgrounds.

2.6.2 Outlook

Many issues were not entirely resolved, and several results need further examination.

The jet charge study was performed on a single dataset and also not compared to samples from another Monte-Carlo generator using another hadronization model. The study also only applied simple kinematic cuts on signal and background. A future study might therefore consider refined cutting and making use of a quark/gluon-tagger.

The likelihood study was based on a custom made framework. To save time and increase certainty in the ntuple-creations, an established framework should have been used. This would also allow for the use of 2016 data and therefore correct PRW of MC. *mva_boson* correlates with m_W. More work into the number of variables and each of their

contributions into to shape of the m_W background after cutting is needed. The final simultaneous fit in all four variables did not include the W particles from background in the background PDF. It is difficult to include this small contribution. Especially if some of the parameters must float. A study into how to build the full PDF properly is needed. This, however, is only possible when the correlations are addressed.

To further reduce the top background, the ATLAS top tagger could be used along with a cut on the invariant mass of the reconstructed top candidate. One could use a lower b-tagging efficiency cut for MV2c20. The b-tag scores of the jets with the largest b-tag scores that did not pass the b-tag cut could be used along with the aforementioned top tagger scores and invariant mass cut to make an improved top background reducer.

Appendices

A.1 Datasets

A.1.1 For charge calculation

Table A.1: Background data samples from the mc15a simulation project. The data samples contain dijets initiated by gluons and light quarks.

DSID	Name
361020	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ0W
361021	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ1W
361022	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ2W
361023	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ3W
361024	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ4W
361025	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ5W
361026	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6W
361027	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ7W
361028	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ8W
361029	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ9W
361030	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ10W
361031	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ11W
361032	Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ12W

Table A.2: Signal data samples from the mc15a simulation project. The data samples contain the channel $W' \to WZ \to qqqq$, where W' is a hypothetical particle with different mass resonances, as listed (in GeV).

DSID	Name
301254	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m400
301255	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m600
301257	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m1000
301262	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m1500
301267	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m2000
301272	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m2500
301277	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m3000
301282	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m4000
301287	Pythia8EvtGen_A14NNPDF23LO_Wprime_WZqqqq_m5000

A.1.2 For diboson studies

Table A.3: Background $t\bar{t}$ samples from the mc15c simulation project.

DSID	Name
410000	PowhegPythiaEvtGen_P2012_ttbar_hdamp172p5_nonallhad

DSID	Name
361091	Sherpa_CT10_WplvWmqq_SHv21_improved
361092	Sherpa_CT10_WpqqWmlv_SHv21_improved
361093	Sherpa_CT10_WlvZqq_SHv21_improved
361094	Sherpa_CT10_WqqZll_SHv21_improved
361095	Sherpa_CT10_WqqZvv_SHv21_improved
361096	Sherpa_CT10_ZqqZll_SHv21_improved
361097	Sherpa_CT10_ZqqZvv_SHv21_improved

Table A.4: Background diboson samples from the mc15c simulation project.

DSID	Name
363331	Sherpa_NNPDF30NNLO_Wtaunu_Pt0_70_CVetoBVeto
363332	Sherpa_NNPDF30NNLO_Wtaunu_Pt0_70_CFilterBVeto
363333	Sherpa_NNPDF30NNLO_Wtaunu_Pt0_70_BFilter
363334	Sherpa_NNPDF30NNLO_Wtaunu_Pt70_140_CVetoBVeto
363335	Sherpa_NNPDF30NNLO_Wtaunu_Pt70_140_CFilterBVeto
363336	Sherpa_NNPDF30NNLO_Wtaunu_Pt70_140_BFilter
363337	Sherpa_NNPDF30NNLO_Wtaunu_Pt140_280_CVetoBVeto
363338	Sherpa_NNPDF30NNLO_Wtaunu_Pt140_280_CFilterBVeto
363339	Sherpa_NNPDF30NNLO_Wtaunu_Pt140_280_BFilter
363340	Sherpa_NNPDF30NNLO_Wtaunu_Pt280_500_CVetoBVeto
363341	Sherpa_NNPDF30NNLO_Wtaunu_Pt280_500_CFilterBVeto
363342	Sherpa_NNPDF30NNLO_Wtaunu_Pt280_500_BFilter
363343	Sherpa_NNPDF30NNLO_Wtaunu_Pt500_700_CVetoBVeto
363344	Sherpa_NNPDF30NNLO_Wtaunu_Pt500_700_CFilterBVeto
363345	Sherpa_NNPDF30NNLO_Wtaunu_Pt500_700_BFilter
363346	Sherpa_NNPDF30NNLO_Wtaunu_Pt700_1000_CVetoBVeto
363347	Sherpa_NNPDF30NNLO_Wtaunu_Pt700_1000_CFilterBVeto
363348	Sherpa_NNPDF30NNLO_Wtaunu_Pt700_1000_BFilter
363349	Sherpa_NNPDF30NNLO_Wtaunu_Pt1000_2000_CVetoBVeto
363350	Sherpa_NNPDF30NNLO_Wtaunu_Pt1000_2000_CFilterBVeto
363351	Sherpa_NNPDF30NNLO_Wtaunu_Pt1000_2000_BFilter
363352	Sherpa_NNPDF30NNLO_Wtaunu_Pt2000_E_CMS_CVetoBVeto
363353	Sherpa_NNPDF30NNLO_Wtaunu_Pt2000_E_CMS_CFilterBVeto
363354	Sherpa_NNPDF30NNLO_Wtaunu_Pt2000_E_CMS_BFilter

Table A.5: $W \to \tau\nu$ plus jets background samples from the mc15c simulation project.

Table A.6: $W \to \mu\nu$ plus jets background samples from the mc15c simulation project.

DSID	Name
363436	Sherpa_NNPDF30NNLO_Wmunu_Pt0_70_CVetoBVeto
363437	Sherpa_NNPDF30NNLO_Wmunu_Pt0_70_CFilterBVeto
363438	Sherpa_NNPDF30NNLO_Wmunu_Pt0_70_BFilter
363439	Sherpa_NNPDF30NNLO_Wmunu_Pt70_140_CVetoBVeto
363440	Sherpa_NNPDF30NNLO_Wmunu_Pt70_140_CFilterBVeto
363441	Sherpa_NNPDF30NNLO_Wmunu_Pt70_140_BFilter
363442	Sherpa_NNPDF30NNLO_Wmunu_Pt140_280_CVetoBVeto
363443	Sherpa_NNPDF30NNLO_Wmunu_Pt140_280_CFilterBVeto
363444	Sherpa_NNPDF30NNLO_Wmunu_Pt140_280_BFilter
363445	Sherpa_NNPDF30NNLO_Wmunu_Pt280_500_CVetoBVeto
363446	Sherpa_NNPDF30NNLO_Wmunu_Pt280_500_CFilterBVeto
363447	Sherpa_NNPDF30NNLO_Wmunu_Pt280_500_BFilter
363448	Sherpa_NNPDF30NNLO_Wmunu_Pt500_700_CVetoBVeto
363449	Sherpa_NNPDF30NNLO_Wmunu_Pt500_700_CFilterBVeto
363450	Sherpa_NNPDF30NNLO_Wmunu_Pt500_700_BFilter
363451	Sherpa_NNPDF30NNLO_Wmunu_Pt700_1000_CVetoBVeto
363452	Sherpa_NNPDF30NNLO_Wmunu_Pt700_1000_CFilterBVeto
363453	Sherpa_NNPDF30NNLO_Wmunu_Pt700_1000_BFilter
363454	Sherpa_NNPDF30NNLO_Wmunu_Pt1000_2000_CVetoBVeto
363455	Sherpa_NNPDF30NNLO_Wmunu_Pt1000_2000_CFilterBVeto
363456	Sherpa_NNPDF30NNLO_Wmunu_Pt1000_2000_BFilter
363457	Sherpa_NNPDF30NNLO_Wmunu_Pt2000_E_CMS_CVetoBVeto
363458	Sherpa_NNPDF30NNLO_Wmunu_Pt2000_E_CMS_CFilterBVeto
363459	Sherpa_NNPDF30NNLO_Wmunu_Pt2000_E_CMS_BFilter

Table A.7: $W \to e\nu$ plus jets background samples from the mc15c simulation project.

DSID	Name
363460	Sherpa_NNPDF30NNLO_Wenu_Pt0_70_CVetoBVeto
363461	Sherpa_NNPDF30NNLO_Wenu_Pt0_70_CFilterBVeto
363462	Sherpa_NNPDF30NNLO_Wenu_Pt0_70_BFilter
363463	Sherpa_NNPDF30NNLO_Wenu_Pt70_140_CVetoBVeto
363464	Sherpa_NNPDF30NNLO_Wenu_Pt70_140_CFilterBVeto
363465	Sherpa_NNPDF30NNLO_Wenu_Pt70_140_BFilter
363466	Sherpa_NNPDF30NNLO_Wenu_Pt140_280_CVetoBVeto
363467	Sherpa_NNPDF30NNLO_Wenu_Pt140_280_CFilterBVeto
363468	Sherpa_NNPDF30NNLO_Wenu_Pt140_280_BFilter
363469	Sherpa_NNPDF30NNLO_Wenu_Pt280_500_CVetoBVeto
363470	Sherpa_NNPDF30NNLO_Wenu_Pt280_500_CFilterBVeto
363471	Sherpa_NNPDF30NNLO_Wenu_Pt280_500_BFilter
363472	Sherpa_NNPDF30NNLO_Wenu_Pt500_700_CVetoBVeto
363473	Sherpa_NNPDF30NNLO_Wenu_Pt500_700_CFilterBVeto
363474	Sherpa_NNPDF30NNLO_Wenu_Pt500_700_BFilter
363475	Sherpa_NNPDF30NNLO_Wenu_Pt700_1000_CVetoBVeto
363476	Sherpa_NNPDF30NNLO_Wenu_Pt700_1000_CFilterBVeto
363477	Sherpa_NNPDF30NNLO_Wenu_Pt700_1000_BFilter
363478	Sherpa_NNPDF30NNLO_Wenu_Pt1000_2000_CVetoBVeto
363479	Sherpa_NNPDF30NNLO_Wenu_Pt1000_2000_CFilterBVeto
363480	Sherpa_NNPDF30NNLO_Wenu_Pt1000_2000_BFilter
363481	Sherpa_NNPDF30NNLO_Wenu_Pt2000_E_CMS_CVetoBVeto
363482	Sherpa_NNPDF30NNLO_Wenu_Pt2000_E_CMS_CFilterBVeto
363483	Sherpa_NNPDF30NNLO_Wenu_Pt2000_E_CMS_BFilter

Table A.8: $HVT \rightarrow WW \rightarrow lvqq$ signal samples for different masses from the mc15c simulation project.

DSID	Name
302116	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m0500
302117	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m0600
302118	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m0700
302119	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m0800
302120	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m0900
302121	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1000
302122	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1100
302123	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1200
302124	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1300
302125	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1400
302126	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1500
302127	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1600
302128	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1700
302129	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1800
302130	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m1900
302131	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m2000
302132	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m2200
302133	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m2400
302134	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m2600
302135	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m2800
302136	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m3000
302137	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m3500
302138	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m4000
302139	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m4500
302140	MadGraphPythia8EvtGen_A14NNPDF23LO_HVT_Agv1_VzWW_lvqq_m5000

References

[1] Wikipedia. Standard model of elementary particles, 2006. [Online; accessed 24-August-2016].

[2] Wikipedia. Elementary particle interactions, 2007. [Online; accessed 29-September-2016].

[3] Thomson, M. *Modern Particle Physics*. Cambridge University Press, 2013.

[4] Peter Hansen. Particle detectors and accelerators. Lecture notes from course of same name, 2015.

[5] Wikipedia. Breit-wigner, 2006. [Online; accessed 06-October-2016].

[6] Wim de Boer. The Discovery of the Higgs Boson with the CMS Detector and its Implications for Supersymmetry and Cosmology. In *Time and Matter 2013 (TAM2013) Venice, Italy*, 2013.

[7] K. A. Olive et al. Review of Particle Physics. *Chin. Phys.*, C38:090001, 2014.

[8] G. Watt. MSTW 2008 LO PDFs, 2008. [Online; accessed 14-September-2016].

[9] David d'Enterria. Physics at the LHC: A Short overview. *J. Phys. Conf. Ser.*, 270:012001, 2011.

[10] CERN, Antonio Saba, and Peter Ginter. CERN Accelerator Complex, 2008. [Online; accessed 24-August-2016].

[11] LHC Machine Outreach. [Online; accessed 07-October-2016].

[12] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.

[13] Noel Dawe. tikz-track, 2016. [Online; accessed 07-October-2016].

[14] ATLAS Collaboration. Electron and photon energy calibration with the ATLAS detector using data collected in 2015 at $\sqrt{s} = 13$ TeV. Technical Report ATL-PHYS-PUB-2016-015, CERN, Geneva, Aug 2016.

[15] Georges Aad et al. Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data. *Eur. Phys. J.*, C74(10):3071, 2014.

[16] Georges Aad et al. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J.*, C73(3):2304, 2013.

[17] ATLAS Collaboration. Expected electron performance in the ATLAS experiment. Technical Report ATL-PHYS-PUB-2011-006, CERN, Geneva, Apr 2011.

[18] ATLAS Collaboration. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$=13 TeV. *Eur. Phys. J. C*, 76(arXiv:1603.05598. CERN-EP-2016-033):292. 45 p, Mar 2016.

[19] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012.

[20] Georges Aad et al. Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV. *Eur. Phys. J.*, C76(3):154, 2016.

[21] ATLAS Collaboration. Performance of Boosted W Boson Identification with the ATLAS Detector. Technical Report ATL-PHYS-PUB-2014-004, CERN, Geneva, Mar 2014.

[22] Andrea Coccaro. Track Reconstruction and b-Jet Identification for the ATLAS Trigger System. *J. Phys. Conf. Ser.*, 368:012034, 2012.

[23] ATLAS Collaboration. Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV. Technical Report ATL-PHYS-PUB-2015-023, CERN, Geneva, Jul 2015.

[24] Andreas Søgaard. Boosted bosons and wavelets. Master's thesis, Niels Bohr Institute, University of Copenhagen, 2015.

[25] Ada ms, D et al. Recommendations of the Physics Objects and Analysis Harmonisation Study Groups 2014. Technical Report ATL-COM-PHYS-2014-451, CERN, Geneva, May 2014. Internal note.

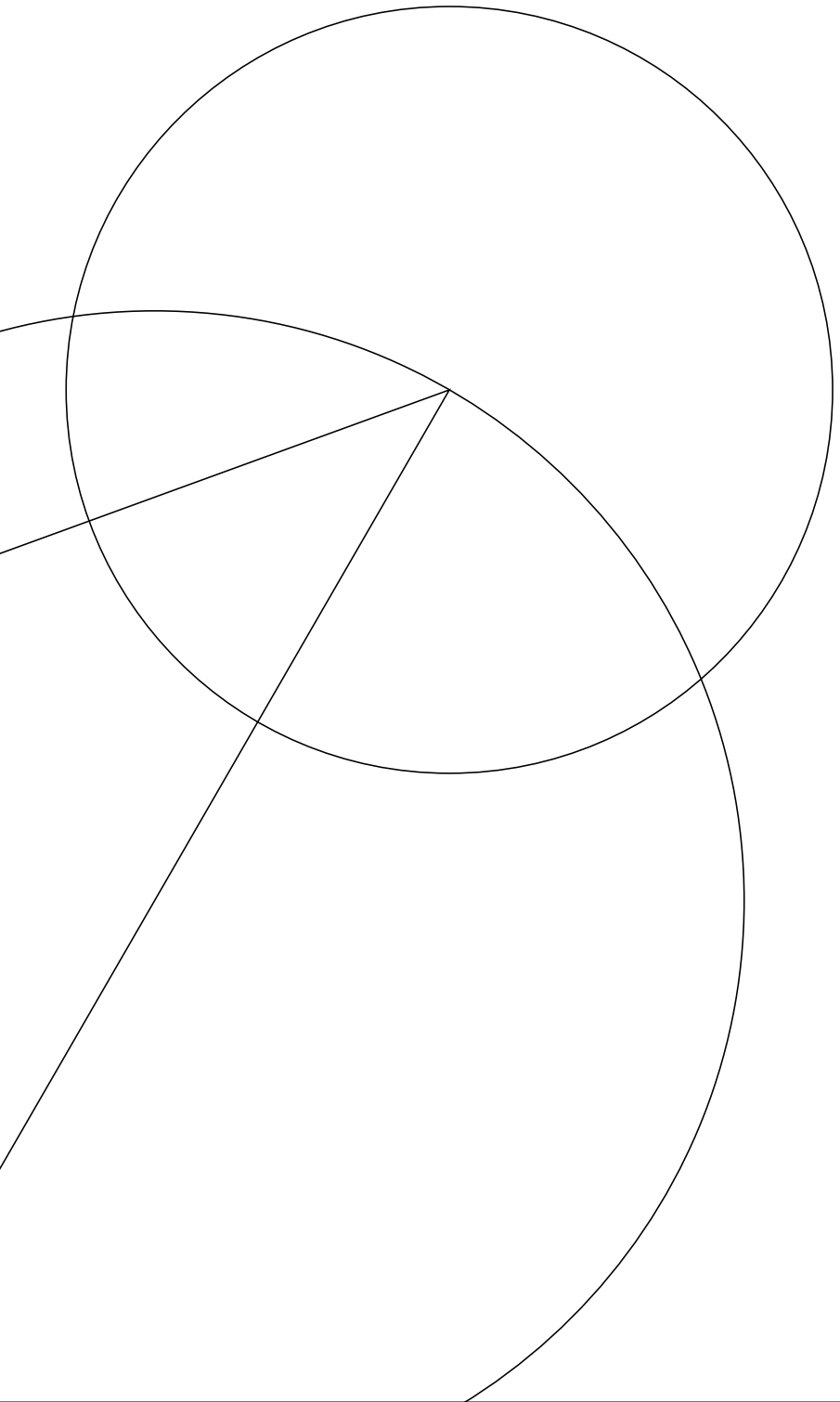
[26] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[27] ATLAS Collaboration. Jet Charge Studies with the ATLAS Detector Using $\sqrt{s} = 8$ TeV Proton-Proton Collision Data. Technical Report ATLAS-CONF-2013-086, CERN, Geneva, Aug 2013.

[28] The DZero Experiment. Useful diagrams of top signals and backgrounds, 2011. [Online; accessed 29-September-2016].

[29] Andrea Thamm, Riccardo Torre, and Andrea Wulzer. Composite Heavy Vector Triplet in the ATLAS Diboson Excess. *Phys. Rev. Lett.*, 115(22):221802, 2015.

[30] ATLAS Collaboration. Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector. *Eur. Phys. J. C*, 75(arXiv:1506.05669. CERN-PH-EP-2015-114):476. 36 p, Jun 2015.

[31] W. Verkerke and D. Kirkby. The RooFit toolkit for data modeling. *ArXiv Physics e-prints*, June 2003.

AN ALTERNATIVE ANALYSIS OF THE SEMI-LEPTONIC DIBOSON FINAL STATES IN THE BOOSTED REGIME