Optimization of event selection in search for doubly-charged Higgs bosons at ATLAS using machine learning techniques

by

Adrienne Scott
B.Eng., McMaster University, 2023

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Physics and Astronomy

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Optimization of event selection in search for doubly-charged Higgs bosons
at ATLAS using machine learning techniques

by

Adrienne Scott
B.Eng., McMaster University, 2023

Supervisory Committee

Dr. M. Lefebvre, Supervisor
(Department of Physics and Astronomy)

Dr. H. Russell, Departmental Member
(Department of Physics and Astronomy)

Supervisory Committee

Dr. M. Lefebvre, Supervisor
(Department of Physics and Astronomy)

Dr. H. Russell, Departmental Member
(Department of Physics and Astronomy)

ABSTRACT

The analysis of proton-proton collision data recorded by ATLAS during Run 2 of the LHC identified an excess of data over the Standard Model prediction in both the $W^{\pm}Z$ and $W^{\pm}W^{\pm}$ vector boson scattering processes. These excesses could be attributed to resonances of the singly-charged ($H_5^{\pm}$) and doubly-charged ($H_5^{\pm\pm}$) Higgs bosons, which are hypothesized by the Georgi-Machacek (GM) model. To investigate this excess and assess its compatibility with the GM model, a dedicated search is being performed for the $H_5^{\pm}$ and $H_5^{\pm\pm}$ bosons where they are produced by vector boson fusion and decay to $W^{\pm}Z$ and $W^{\pm}W^{\pm}$ respectively. In this thesis, the selection of the $H_5^{\pm\pm}$ signal region is optimized by training a neural network to discriminate signal events from background events. The characteristics of the $H_5^{\pm\pm}$ events vary significantly with mass, which leads to undesired behaviour when training a single network for a large mass range. A number of strategies are devised to address this problem; the best solution is to modify the weighting of different simulated masses during training. The neural network is used to define a new $W^{\pm}W^{\pm}$ signal region which has a greater sensitivity to the GM model compared to a cuts-based approach.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Michel Lefebvre, for the innumerable lessons about physics, science, and life, and for sharing your infectious enthusiasm for particle physics.

My ATLAS collaborators, including, but not limited to, Heather Russell, John McGowan, and Kyle Lau, for providing thoughtful guidance and feedback on my work.

My partner Josh, for supporting me through the busiest days and answering my many questions about machine learning.

My friends and family, for boosting my morale during thesis writing.

My cat Zoe, who reminds me to take breaks to lie in the sun.

The Natural Sciences and Engineering Research Council of Canada, for providing financial support during my Master's degree.

Chapter 1

Introduction

The purpose of this thesis is to improve the event selection for a doubly-charged Higgs boson ($H_5^{\pm\pm}$) search at ATLAS. This search is specifically for a $H_5^{\pm\pm}$ with a mass between 200 GeV and 3 TeV that is produced by vector boson fusion and decays to $W^{\pm}W^{\pm} \rightarrow lvlv$. The doubly-charged Higgs event selection is optimized by training a neural network (NN) to differentiate $H_5^{\pm\pm}$ signal events from background events. One of the main challenges with performing event classification for this analysis is the significant variation in the characteristics of the $H_5^{\pm\pm}$ events across the mass range of interest. As a result, the neural network performs poorly at low mass when the training set includes all of the simulated mass points combined. Several techniques are explored to address this problem, including feature importance optimization, parameterized neural networks, and applying different weights to different mass points.

The Large Hadron Collider (LHC) currently provides the best environment to search for new particles at the electroweak scale since it is the highest energy particle collider in the world. Extended Higgs sectors are not prohibited by the Standard Model (SM) and could address major questions in physics [1]. Thus, searching for additional Higgs bosons is an important part of maximizing the physics potential of the ATLAS experiment.

This thesis begins with an overview of the Standard Model of particle physics in Chapter 2. This is followed by a description of the LHC and the ATLAS detector in Chapter 3. In Chapter 4, the doubly-charged Higgs search and the existing signal region event selections are introduced, motivating the optimization of the signal region using machine learning. Chapter 5 begins with a "naive" approach to neural network classification, and subsequently explores modifications to improve the classification at low mass. Finally, in Chapter 6, the performance of the optimized network is presented and the neural network signal region is compared to a cuts-based signal region.

Chapter 2

Theory

2.1 The Standard Model

The Standard Model (SM) of particle physics is a framework for understanding the fundamental particles in our universe and the way they interact through the fundamental forces. The term "fundamental" (or "elementary") in particle physics means that there is no way to further decompose the particle or force into smaller constituents.

There are four known fundamental forces: the electromagnetic (EM) force, the strong force, the weak force, and the gravitational force. The SM successfully explains the first three forces, but does not yet explain gravity, which is one of the present mysteries in physics. In the SM, forces are described as the exchange of virtual force carrying particles which transfer momentum from one particle to another. For instance, the repulsion of two electrons due to electromagnetism is the result of an exchange of virtual photons, the force mediating particle of the EM force. The other force mediating particles are the gluon for the strong force, and the W and Z bosons for the weak force. A particle interacts with (or "feels") a force if it carries the corresponding charge: electric charge for the EM force, colour charge for the strong force, and weak isospin for the weak force. The strength of the fundamental forces depends on the distance and energy scale. At low energies, the strong force is the strongest, followed by the EM force, and then the weak force (very logical!).

The fundamental particles of the SM are categorized based on their spin, as shown in Figure 2.1. The force carrying particles have a spin of 1, and are thus referred to as *vector bosons*. The only spin 0 (scalar) fundamental particle is the Higgs boson, which was the last particle in the SM to be observed in 2012 [2]. The fermions have a spin of 1/2, and are grouped into three "generations" in increasing mass. The first generation includes up quarks, down quarks, and electrons, which together form the constituents of the materials that we interact with on a daily basis. All fermions have a corresponding antiparticle which has the same mass and spin but opposite charge. Each type of fermion is referred to as a *flavour*.

The fermions are subdivided into two main categories: quarks, which carry colour charge and interact with the strong force, and leptons, which do not. There are three colour charges: red (r), green (g), and blue (b). Antiparticles carry corresponding "anticolour": antired ($\bar{r}$), antigreen ($\bar{g}$), and antiblue ($\bar{b}$). Quarks prefer to exist in colour singlet states. These can be formed by a quark-

antiquark pair ($r\bar{r}, b\bar{b}$, or $g\bar{g}$), called a *meson*. Colour singlets can also be formed by three quarks or three antiquarks which all carry different colours (rgb or $\bar{r}\bar{g}\bar{b}$), called *baryons*. Mesons and baryons both belong to a group of particles called *hadrons*, which includes any particle consisting of two or more quarks in a bound state. In addition to the strong force, quarks also interact with the EM force and the weak force. In each quark generation, there are two quarks, one of which has an electric charge of $+\frac{2}{3}$ and the other which has an electric charge of $-\frac{1}{3}$.

All leptons interact with the weak force, but only electrons, muons and tau leptons interact with the EM force. In each lepton generation, there is a particle of electric charge -1 and a corresponding neutrino. Each generation has an assigned lepton number; particles have lepton number of $+1$ and antiparticles have a lepton number of -1. Lepton number must be conserved in particle physics interactions.

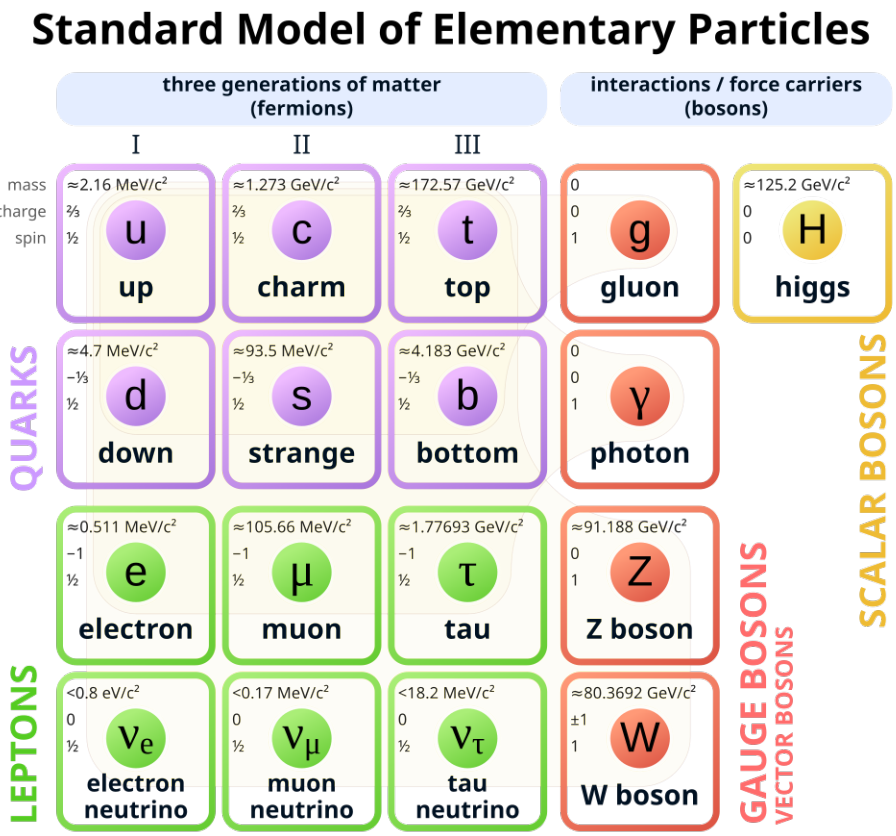


Figure 2.1: The Standard Model of particle physics, from [3]. More information on the SM can be found in [4].

2.1.1 Quantum Electrodynamics

Quantum Electrodynamics (QED) describes electromagnetic interactions, which are mediated by the massless photon (γ). From Noether's Theorem, we know that each symmetry corresponds to a conservation law. In this case, the U(1) global symmetry of QED gives rise to the conservation of

electric charge, Q. The EM interaction is uniquely predicted by a U(1) gauge symmetry and cannot change the colour, flavour, or charge of a particle. There is only one EM interaction vertex, which connects a photon and two electrically charged particles.

2.1.2 The Electroweak Sector

The weak force obeys an $\mathrm{SU(2)_L}$ gauge symmetry and is mediated by the W and Z vector bosons. Global $\mathrm{SU(2)_L}$ symmetry ensures that weak isospin, $\mathrm{I_W^{(3)}}$, is conserved. Unlike photons, the W and Z vector bosons self-interact — many examples can be found in this thesis!

The W^+ and W^- bosons mediate the weak charged current interaction, and the Z boson mediates the weak neutral current interaction. Unfortunately, weak isospin is not as straightforward as electric charge. The handedness of a particle (either right-handed or left-handed) affects its weak isospin. The weak charged current interaction couples only to particles with non-zero weak isospin, which are left-handed particles and right-handed antiparticles. The neutral current interaction, on the other hand, couples to both right- and left-handed particles, and right- and left-handed antiparticles.

The charged current interaction couples leptons within the same generation. For example, an electron neutrino can interact with a $W^\pm$ to produce an electron. The charged current interaction cannot couple leptons in different generations since lepton number is always conserved (in the limit of $m_\nu \to 0$). A lepton can also interact through a neutral current interaction, but it will not change flavour.

Similar to leptons, the charged current interaction also couples quarks within the same generation. However, it is also possible for a charged current interaction to couple quarks from different generations, although it is less likely. This is called quark mixing, and the strength of mixing between different quark flavours is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix. Quarks also interact with the Z through the neutral current interaction, but will not change flavour.

Above a certain energy scale, the electromagnetic and weak interactions can be unified under a single electroweak (EW) theory. The electroweak (EW) sector obeys $\mathrm{SU(2)_L \times U(1)_Y}$ gauge symmetry. The electroweak symmetry has a corresponding weak hypercharge, $\mathrm{Y = 2(Q - I_W^{(3)})}$, which is also a conserved quantity in the SM.

2.1.3 Quantum Chromodynamics

Quantum chromodynamics (QCD) describes the interactions of gluons, the mediating particle of the strong force, and quarks, the fermions that carry colour charge. QCD obeys the $\mathrm{SU(3)_C}$ gauge symmetry, and the global $\mathrm{SU(3)_C}$ symmetry enforces that colour charge is conserved. The strong force interaction vertex connects quarks with different colour charges. Therefore, gluons always have a colour charge which consists of a colour and anticolour. Since there are three colour charges, and the gluon cannot be a colour singlet, there are eight possible gluon states:

$$r\bar{g},\ g\bar{r},\ r\bar{b},\ b\bar{r},\ g\bar{b},\ b\bar{g},\ \frac{1}{\sqrt{2}}(r\bar{r} - g\bar{g}) \quad \text{and} \quad \frac{1}{\sqrt{6}}(r\bar{r} + g\bar{g} - 2b\bar{b}).$$

Gluons are able to self-interact at vertices consisting of 3 or 4 gluons, which leads to colour confinement.

2.2 The Higgs Sector

Without the Higgs mechanism, the massive W and Z bosons break the gauge invariance of the electroweak sector. This is a problem for the Standard Model, which needs gauge invariance to obey unitarity and describe the forces. Fortunately, the Higgs mechanism provides an elegant solution which gives mass to the W and Z through spontaneous symmetry breaking. Another consequence of this symmetry breaking is the existence of a massive scalar particle, the Higgs boson (H). The Higgs mechanism also gives mass to the fermions through their interactions with the Higgs field. The strength of the coupling of the Standard Model fermions to the Higgs boson is proportional to the fermion mass.

2.2.1 Extended Higgs Sectors

The current Higgs model, with a single Higgs boson, is the simplest possible explanation for our observations of fundamental particles [4]. There are no strict limitations, however, which prevent the Higgs model from being more complex, leading to additional scalar states [1]. Extended Higgs sectors provide avenues for addressing open problems in particle physics, such as CP violation, vacuum stability, and dark matter [1]. Some proposed theories for extended Higgs sectors include the two-Higgs-doublet model, singlet extensions, scalar dark matter models, and the Georgi-Machacek model [1]. The latter will be the focus of this thesis.

Chapter 3

The ATLAS Experiment

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a hadron accelerator and collider that occupies an underground ring with a circumference of 26.7 km [5]. There are two vacuum tubes in the ring, which accelerate bunches of protons in opposite directions before they collide at designated interaction points (IPs). There are four main experiments located around the LHC: ATLAS (IP1), CMS (IP5), LHCb (IP8), and ALICE (IP2; heavy ion collisions) [5]. The protons are accelerated in stages, shown in Figure 3.1. The first stage is the LINAC2, followed by the booster (1.4 GeV), the Proton Synchrotron (25 GeV), the Super Proton Synchrotron (450 GeV) and finally the LHC ring, where proton bunches are accelerated to 6.5 TeV (6.8 TeV for Run 3) [5]. This results in proton– proton collisions at a center-of-mass energy of $\sqrt{s} = 13$ TeV (13.6 TeV for Run 3) [6].

The proton bunches at the LHC can be approximated by two uniform Gaussian bunches colliding head on near the speed of light. In this case, the instantaneous luminosity per bunch crossing is

$$\mathcal{L}_b = \frac{N_1 N_2 f_r}{4\pi\sigma_x\sigma_y}, \tag{3.1}$$

where N_1 and N_2 are the number of protons in each bunch and σ_x and σ_y are the standard deviation of the Gaussian beams in the x and y directions, respectively [8]. The revolution frequency, f_r, is 11 246 Hz for protons at the LHC [9]. The instantaneous luminosity is then $\mathcal{L} = n_b \langle\mathcal{L}_b\rangle$, where n_b is the number of bunches and $\langle\mathcal{L}_b\rangle$ is the average luminosity per bunch crossing. The peak instantaneous luminosity in Run 2 was $2.1 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ [6].

Another important quantity is the mean number of interactions per bunch crossing, μ, which is proportional to the total inelastic cross-section, σ_{inel} [9]:

$$\mu = \frac{\sigma_{\text{inel}}\mathcal{L}_b}{f_r}. \tag{3.2}$$

Figure 3.2 shows the luminosity recorded by ATLAS during Run 2 as a function of μ. Larger values of μ correspond to increased pileup conditions in the detector.

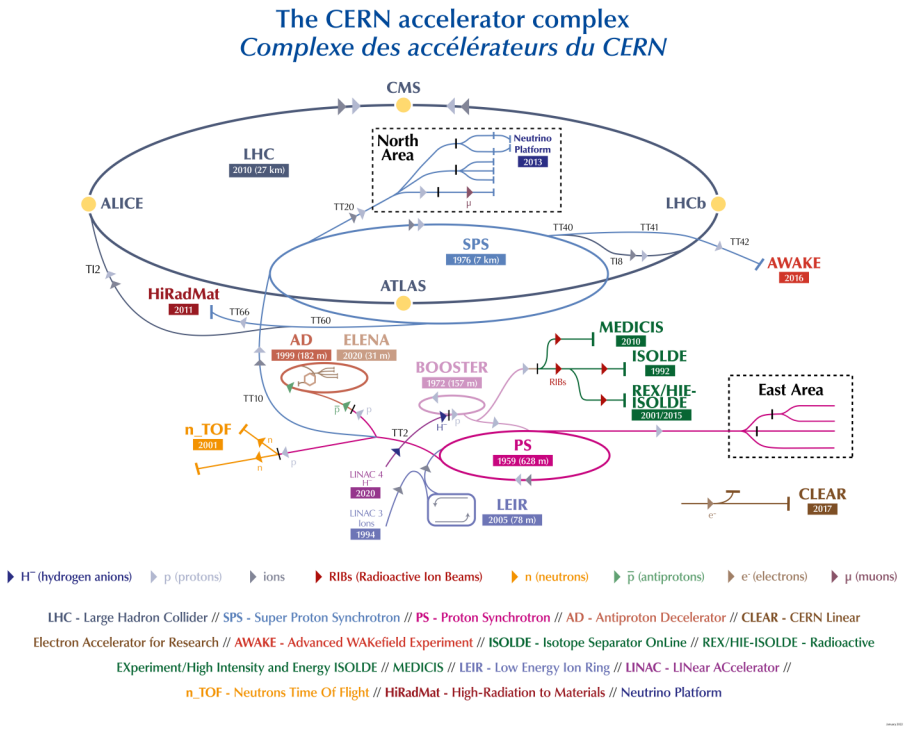


Figure 3.1: The CERN Accelerator complex (Image from [7]).

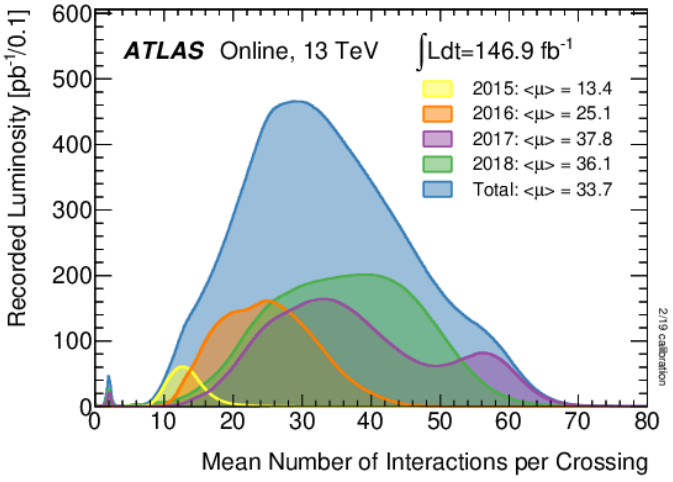


Figure 3.2: The luminosity recorded by ATLAS during each year of data taking in Run 2, from [6]. $\langle\mu\rangle$ refers to the average value of μ over some time period.

For a process with cross-section σ, the number of events that are expected in a given time period, N_{exp}, is proportional to the integrated luminosity:

$$N_{\text{exp}} = \sigma \int \mathcal{L}(t)dt. \tag{3.3}$$

Thus, performing precise measurements of the cross-section of particle physics processes requires a precise knowledge of the luminosity. Furthermore, increasing the luminosity allows for a greater number of events to be observed for a particular process. This is the motivation for the High-Luminosity LHC, which is set to begin in 2030 [10].

3.2 The ATLAS Detector

3.2.1 Detector Overview

The ATLAS detector, shown in Figure 3.3, is located ∼100 m underground at IP1 of the LHC [11]. It spans 25 m in height, 44 m in length, and weighs around 7000 tonnes [6]. It consists of five main systems: an inner detector, an electromagnetic calorimeter, a hadronic calorimeter, a muon system, and a magnet system.

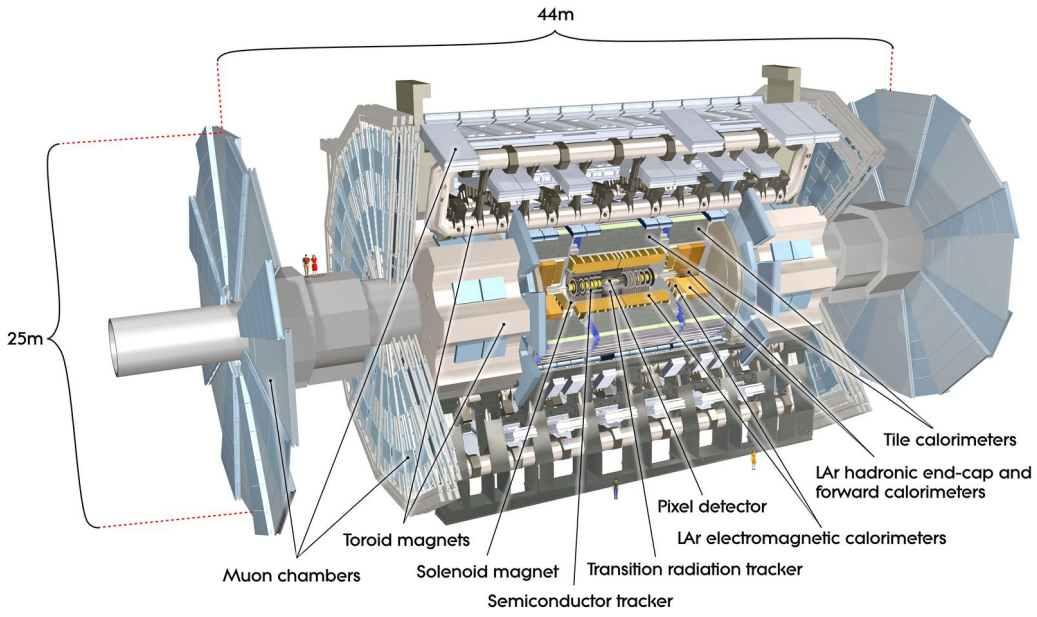


Figure 3.3: The ATLAS detector, from [11].

The inner detector (ID) performs tracking of charged particles using two main detection methods. The pixel detector and the semiconductor tracker (SCT) both use silicon sensors which provide precise spatial information about particle trajectories [6]. The transition radiation tracker (TRT) uses narrow drift tubes filled with a gas mixture that ionizes when charged particles pass through [6]. The

spatial resolution is worse than the silicon sensors, but the TRT nevertheless contributes important information to p_T reconstruction and electron identification [6].

The electromagnetic calorimeter (ECal) measures the energy of photons and electrons by fully containing their electromagnetic showers [11]. It also absorbs some energy from other particles, such as hadrons and muons. It is a sampling calorimeter which uses lead as the absorber and liquid argon as the sensitive medium [11]. The alternating layers of absorber and liquid argon are arranged in an accordion shape with a constant gap between them [11]. This ensures full coverage and uniform performance in the ϕ direction.

The hadronic calorimeter consists of a tile calorimeter, a liquid argon hadronic end-cap calorimeter (HEC) and a liquid argon forward calorimeter (FCal) [11]. Its purpose is to absorb energy from hadrons, most of which are in hadronic jets produced by quarks and gluons. The tile calorimeter is a sampling calorimeter that uses steel as the absorber and plastic scintillating tiles as the sensitive medium [11]. The emitted light is read out via wavelenght shifting fibres and photomultiplier tubes [11]. The HEC provides coverage in the end-cap region and differs from the electromagnetic calorimeter in both geometry and absorbing medium (the HEC uses copper rather than lead) [11]. The FCal completes the coverage of the hadronic calorimeter and uses both copper and tungsten as the absorber for compactness [11].

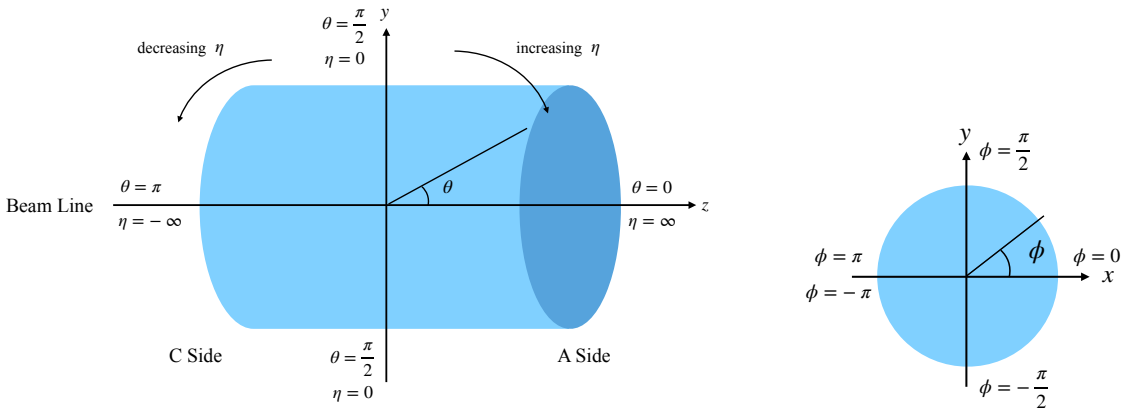
The muon system (MS) is the outermost layer of the ATLAS detector. Tracks in the MS almost always originate from a muon, since all other particles experience more energy loss in the calorimeters and are stopped before reaching the MS. The MS uses several different technologies to detect muons, including resistive plate chambers, thin gap chambers, and monitored drift tubes [6]. All of these technologies detect particles through gas ionization, and collect positive and negative charges at an anode and cathode respectively.

The magnet system produces magnetic fields which deflect charged particles due to the Lorentz force. This enables the determination of the momentum and charge of charged particles in the detector, since the direction of curvature will indicate the charge, and the radius of curvature will be proportional to the momentum. The solenoid magnet produces an axial magnetic field of 2 T in the ID so that charge and momentum can be determined using reconstructed tracks from the ID [6]. The toroid magnet produces a magnetic field which deflects muons, improving the resolution of the muon momentum reconstruction by matching tracks from the MS to tracks from the ID [6].

3.2.2 Detector Geometry

The origin of the ATLAS coordinate system is defined to be the nominal interaction point. The positive $x-$direction points towards the centre of the LHC, the positive $y-$direction points upwards, and the positive $z-$direction points towards side A of the detector, forming a right-handed coordinate system. The angle ϕ is measured in the $x-y$ plane with respect to the positive $x-$axis, and the angle θ is measured with respect to the positive $z-$axis [11].

As observers of particles colliding in the z direction, the lab frame will always be boosted in the z direction with respect to the center-of-mass frame of the parton-parton collision. Differences in ϕ will be invariant under longitudinal boost, but not differences in θ. Therefore, it is more practical



(a) (b)

Figure 3.4: The right-handed ATLAS coordinate system, shown from (a) the side view, outside the LHC ring and (b) the transverse view, from the A side of the detector.

to use the rapidity y or pseudorapidity η,

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \qquad \text{and} \tag{3.4}$$

$$\eta = \lim_{m \to 0} y = -\ln \tan \frac{\theta}{2}, \tag{3.5}$$

since differences of y and differences of η are both invariant under longitudinal boosts. The pseudorapidity can also be used to define an angular distance between two objects, ΔR, which is also invariant under longitudinal boosts:

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

The preferred four-vector representation for particles in ATLAS is $p^\mu = (p_\mathrm{T}, \eta, \phi, E)$ where p_T is the magnitude of the particle momentum in the transverse ($x - y$ plane, where y is the $y-$axis not rapidity) direction. The detector coordinate system is visualized in Figure 3.4.

3.2.3 Object Identification and Reconstruction

There are three types of physics objects that are of particular importance to this thesis: hadronic jets, electrons, and muons. Each of these objects is reconstructed by synthesizing information from multiple detector systems. A variety of tools are used to maximize the likelihood that the reconstructed physics object is correctly identified. The reconstruction procedure and identification methods are described in greater detail below.

Hadronic jets, electrons, and muons may each have associated tracks in the ID, which are reconstructed from track seeds using a combinatorial Kalman filter [12]. The hard scatter primary vertex (PV) of an event is defined to be the vertex that intersects with the largest scalar sum of track transverse momenta [13]. When reconstructing an event, only objects that originate from within

a certain distance of the PV should be included since there are many particles that are produced at the same time from other interactions. This is enforced using the transverse impact parameter of the track relative to the beam-line, d_0, and the longitudinal distance between the corresponding point and the PV, z_0 [12].

Hadronic jets are formed from hadronized quarks and gluons, or from other hardonically decaying particles. They deposit energy in both the electromagnetic and hadronic calorimeters. If a calorimeter cell records a 4σ deviation from average noise, then a topo-cluster is formed with all adjacent cells containing a greater than 2σ deviation from average noise [13]. Jets are reconstructed by grouping these topo-clusters into cones with a radius R in the $y - \phi$ plane (where y is rapidity not the $y-$axis) using the anti-k_t algorithm [14]. In this thesis, small-R jets ($R = 0.4$) will be used since they are more likely to be produced by quarks or gluons [15]. In order to suppress pileup jets, the jet-vertex-tagger (JVT) uses the tracks associated to the jet to assess whether the jet originated from the primary vertex [13]. The track information is also used by flavour tagging algorithms to identify jets which likely originate from b or c hadrons [16]. For the purposes of this thesis, the b-tagger (GN2) will be used with the 85% working point (WP) [16].

Electrons and photons both deposit energy in ECal cells, which are grouped into topo-clusters and then superclusters [17]. Electrons are differentiated from photons based on the track information. If a supercluster is matched to a track, it is identified as an electron [17]. If a supercluster is either not matched to a track or matched to a conversion vertex, where two opposite sign electron tracks intersect, it is identified as a photon [17]. To further improve the electron identification, a likelihood (LH) method assesses each electron candidate based on its track and energy deposit information. The result of the LH method is used to define three levels of electron identification: *Loose LH, Medium LH,* and *Tight LH* [17]. In general, "Tight" refers to the most stringent criteria and "Loose" refers to the least stringent criteria. This allows each analysis to place a requirement on how closely the electrons they use should match the detector signature of an electron. Recently, a deep neural network (DNN) was found to outperform the LH method in terms of background rejection, and is used to define *Loose DNN, Medium DNN,* and *Tight DNN* criteria [18]. The electron isolation in the ID and ECal can also be used to improve the electron selection, since many background sources such as light flavour decays tend to have surrounding hadronic activity. The isolation of the electrons is determined based on the sum of the transverse momentum of the tracks contained within a certain radius of the electron ($p_{\mathrm{T}}^{\mathrm{varcone}}$), and the sum of the transverse energy of the nearby energy deposits in the calorimeters ($E_{\mathrm{T}}^{\mathrm{cone}}$) [17]. The radius of the isolation cone used to construct p_T^{varcone} varies based on the transverse momentum of the electron. Thus, the electron isolation WPs constructed from $p_{\mathrm{T}}^{\mathrm{varcone}}$ and $E_{\mathrm{T}}^{\mathrm{cone}}$ are referred to as *Loose Variable Radius (Loose VarRad)* and *Tight Variable Radius (Tight VarRad).*

Muons are identified primarily based on the presence of a track in the muon spectrometer, since other particles decay or stop in the calorimeters. Muon trajectories can be computed using only the MS, or using additional information from the ID and the calorimeters [19]. The quality of a muon candidate is determined using the number of hits on the ID and MS, the compatibility of the hits in the ID and MS, and the quality of the track fit [19]. There are three muon quality WPs: *Loose, Medium,* and *Tight.* Similar to electrons, the muon selection can also be improved by adding an isolation criteria. The muon *Tight VarRad* and *Loose VarRad* isolation criteria use

the track information and the calorimeter information separately [19]. The particle flow isolation criteria combine information from both, and are referred to as *Pflow Tight VarRad* and *Pflow Loose VarRad* [19].

Evidently, it is possible that tracks and energy deposits in the detector may be attributed to many different objects. Therefore, it is important to remove overlapping objects so that particles are not counted twice. The simulation samples that are used in this thesis were produced with the following overlap removal procedure:

1. If two electrons reconstructed in the ECal share the same track, the electron with the lower p_{T} is discarded
2. Muons reconstructed in the calorimeters that share a track with an electron are discarded
3. Electrons that share a track with a muon reconstructed in the MS are discarded
4. Jets within $\Delta R < 0.2$ of an electron are discarded
5. Electrons within $\Delta R < 0.4$ of a jet are discarded
6. Jets within $\Delta R < 0.2$ of a muon are discarded
7. Muons within $\Delta R < 0.4$ of a jet are discarded

Collisions occur at ATLAS every 25 ns, but not all collisions produce interesting interactions [11]. The trigger system ensures that only events that meet certain criteria are saved to memory. Events that are used in this thesis must pass either a single-electron or single-muon high-level trigger. The transverse momentum requirement ranges from 24 GeV to 140 GeV for electrons and from 20 GeV to 50 GeV for muons depending on the lepton identification/quality and data taking conditions.

3.2.4 Missing Transverse Momentum

The incoming bunches of protons have negligible transverse momentum when they collide at the ATLAS IP. According to momentum conservation, there should therefore be no transverse momentum in the final state of each event. Thus, if the reconstructed transverse momentum vector of an event is non-zero, it indicates that some transverse momentum was "missed" by the detector in the opposite direction. This could be due to the presence of neutrinos or minimally interacting BSM particles in the final state. Therefore, it is useful to define the missing transverse momentum vector, $\mathbf{p}_{\mathrm{T}}^{\mathrm{miss}}$, as the negative of the sum of the transverse momentum vector for each reconstructed object (the hard term) and each unused charged particle track which aligns with the hard scatter vertex (the soft term) [20]:

$$\mathbf{p}_{\mathrm{T}}^{\mathrm{miss}} = -\left(\sum_{\text{electrons}} \mathbf{p}_{\mathrm{T}}^{e} + \sum_{\text{photons}} \mathbf{p}_{\mathrm{T}}^{\gamma} + \sum_{\tau-\text{leptons}} \mathbf{p}_{\mathrm{T}}^{\tau} + \sum_{\text{muons}} \mathbf{p}_{\mathrm{T}}^{\mu} + \sum_{\text{jets}} \mathbf{p}_{\mathrm{T}}^{\mathrm{jet}} + \sum_{\substack{\text{unused} \\ \text{tracks}}} \mathbf{p}_{\mathrm{T}}^{\mathrm{track}} \right). \quad (3.7)$$

The missing transverse momentum vector is a two-dimensional vector which may be expressed as an x and y component, $\mathbf{p}_{\mathrm{T}} = (p_x, p_y)$, or as a magnitude and ϕ angle. Generally, $E_{\mathrm{T}}^{\mathrm{miss}}$ is used to refer to the magnitude of the missing transverse momentum vector, $|\mathbf{p}_{\mathrm{T}}^{\mathrm{miss}}|$.

Another useful quantity is H_{T}, which is defined similarly to $\mathbf{p}_{\mathrm{T}}^{\mathrm{miss}}$ but instead takes the scalar sum of $|\mathbf{p}_{\mathrm{T}}|$ for all of the objects in the event. In this thesis, H_{T} is calculated using selected electrons, muons, and jets.

Chapter 4

Search for $H^{\pm\pm}$

4.1 Motivation

The Georgi-Machacek (GM) model expands the Higgs sector so that it includes an additional singlet, a triplet, and a quintuplet, $(H_5^0, H_5^{\pm}, H_5^{\pm\pm})$ [21]. Each of the states in the quintuplet are degenerate in mass at tree level, with mass m_{H_5} [22]. In this theory, the $H_5^{\pm}$ and $H_5^{\pm\pm}$ are fermiophobic, and couple only to the SM W and Z bosons and the other bosons in the extended Higgs sector [22]. Therefore, if we assume that the isotriplet state is heavier than the quintuplet state, the $H_5^{\pm}$ and $H_5^{\pm\pm}$ always decay to $W^{\pm}Z$ and $W^{\pm}W^{\pm}$ respectively, due to conservation of electric charge [22].

Both $H_5^{\pm}$ and $H_5^{\pm\pm}$ can be produced through vector boson fusion (VBF) [22], where two quarks radiate vector bosons that fuse to form a single boson [23]. These two quarks produce characteristic VBF jets that have a large jet pair invariant mass and a large rapidity difference. Thus, we can search for $H_5^{\pm}$ or $H_5^{\pm\pm}$ by looking for events with two VBF jets and either the leptonic or hadronic decay products of $W^{\pm}Z$ or $W^{\pm}W^{\pm}$. The advantage of using the fully leptonic final state is that the charge of the W bosons can be determined. This significantly reduces the background for the $W^{\pm}W^{\pm}$ channel since SM processes producing two same-sign leptons are rare. Additionally, in the fully hadronic final state, it is not possible to differentiate the VBF jets from the $W^{\pm}W^{\pm}$ decay products with certainty, making it difficult to reconstruct the $H_5^{\pm\pm}$ mass. Figure 4.1 shows representative Feynman diagrams for $H_5^{\pm}$ and $H_5^{\pm\pm}$ VBF production with fully leptonic decay modes. The production cross-sections for $H_5^{\pm}$ and $H_5^{\pm\pm}$ are proportional to $\sin^2(\theta_H)$, a free parameter which characterizes the vacuum expectation value of the real and complex triplet fields in the GM model [24].

The ATLAS Collaboration performed searches for both processes shown in Figure 4.1 using Run 2 ATLAS data. In both channels, an excess of data was observed, with a local (global) significance of 2.8 (1.6) at 375 GeV in the $W^{\pm}Z$ channel [25], and a local (global) significance of 3.2 (2.5) at 450 GeV in the $W^{\pm}W^{\pm}$ channel [26]. A combined statistical analysis of the two channels revealed that the greatest excess occurs at 375 GeV with a local (global) significance of 3.3 (2.5) [22]. These results are summarized in Figure 4.2, which shows the expected and observed upper limit on $\sin(\theta_H)$ as a function of m_{H_5}. The excess of data prevented the GM model from being excluded to a greater extent in the region where the observed upper limit is larger than the expected limit. Due to its

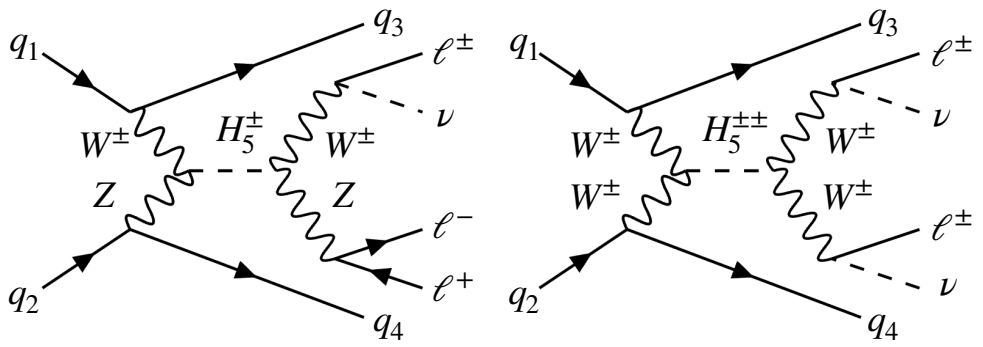
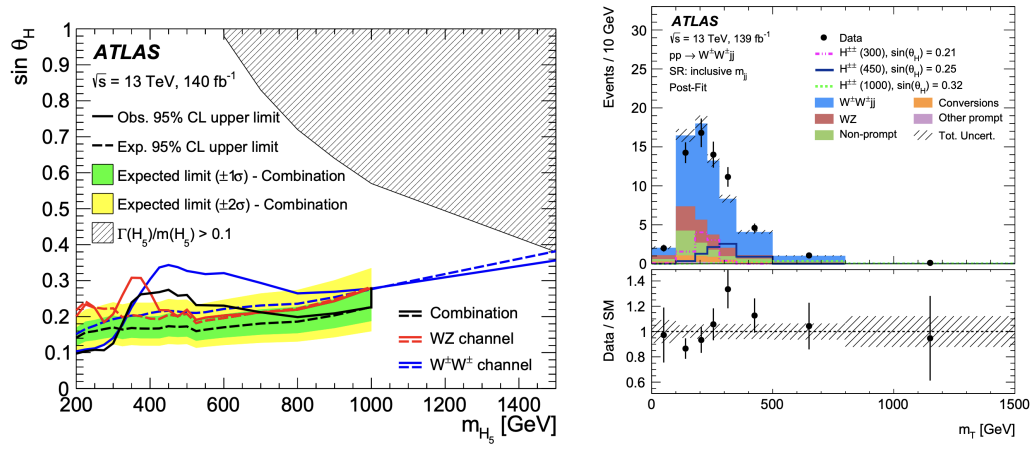


Figure 4.1: Representative Feynman diagrams for $H_5^\pm$ (left) and $H_5^{\pm\pm}$ (right) VBF production.

particular importance in this thesis, the signal region m_T distribution from the Run 2 $W^\pm W^\pm$ analysis is also shown in Figure 4.2.



(a) (b)

Figure 4.2: (a) The expected and observed limits on $\sin(\theta_H)$ from [22]. The hashed area marks the parameter space where the GM model is disfavoured due to the widths of the $H_5^\pm$ and $H_5^{\pm\pm}$ exceeding 10% of m_{H_5} [22]. (b) The post-fit m_T distribution in the signal region of the Run 2 $W^\pm W^\pm$ analysis, from [26]. The most significant excess of data over the SM expectation occurs in the 280 GeV to 350 GeV bin.

A new ATLAS analysis is being developed to investigate the $W^\pm Z$ and $W^\pm W^\pm$ excess in the context of the GM model. The objective of this analysis is to improve the limits on the GM model by optimizing the search strategy for $H_5^\pm$ and $H_5^{\pm\pm}$ and incorporating partial Run 3 (2022–23) data. In this thesis, I will present my contribution to this analysis, which is the optimization of the $W^\pm W^\pm$ signal region selection. For this reason, the $W^\pm Z$ channel will not be discussed in the remainder of my thesis. The only exception is Appendix A, which summarizes a study I performed for the $W^\pm Z$ channel of close-by-corrected isolation variables. For brevity, the $H_5^{\pm\pm}$ boson may be referred to as $H^{\pm\pm}$.

4.2 Search Strategy

To determine whether the observed data are compatible with the GM model, we perform a binned likelihood fit. This fit quantifies the significance of any $W^{\pm}W^{\pm}$ excess and sets an upper limit on $\sin(\theta_H)$. Since there are two neutrinos in the $W^{\pm}W^{\pm}$ final state, the invariant mass of the $W^{\pm}W^{\pm}$ system is difficult to estimate. Therefore, the parameter which is used to discern the $H_5^{\pm\pm}$ mass resonance is m_{T}, since it does not require the longitudinal momentum of the final state particles:

$$m_T = \sqrt{(E_T^{\ell\ell} + E_T^{\mathrm{miss}})^2 - |\vec{p}_T^{\,\ell\ell} + \vec{E}_T^{\mathrm{miss}}|^2}. \tag{4.1}$$

In the signal region, a two-dimensional maximum likelihood fit is performed with m_T and m_{jj}, the invariant mass of the leading jet pair.

The two main background sources are SM $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ events with two jets in the final state. These events can be produced through electroweak processes, such as vector boson scattering (VBS), where two vector bosons interact and produce two more vector bosons [23]. They may also be produced through QCD processes, or the interference between electroweak and QCD diagrams. Examples of the Feynman diagrams for these processes are provided in Figure 4.3. Together, the $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ events comprise ∼80% of the background events in the signal region [26]. Therefore, the accuracy of the $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ Monte Carlo (MC) simulation is very important for the observation of the doubly-charged Higgs. For this reason, they are each normalized in the fit using data in designated control regions.

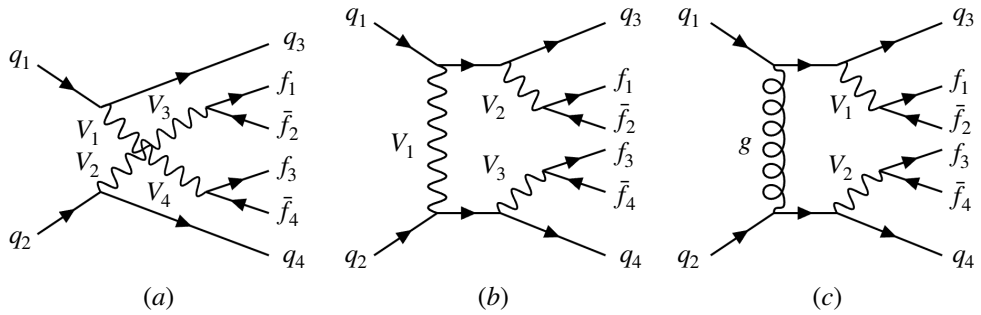


Figure 4.3: Representative Feynman diagrams for a) electroweak VV production by VBS, b) electroweak VV production without VBS, and c) QCD VV production. Here, VV could be either $W^{\pm}Z$ or $W^{\pm}W^{\pm}$.

The background contributions from $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ are estimated using MC simulations. However, some of the backgrounds are estimated using data-driven techniques since they are difficult to model accurately. The non-prompt background from V+jets and semileptonic $t\bar{t}$ events is estimated using the fake factors method. The charge mis-ID background from opposite-sign WW, Drell-Yan (Z+jets), and dileptonic $t\bar{t}$ events is estimated by weighting opposite-sign events with charge flip rates. The remaining backgrounds, which include photon conversions from $V\gamma$ events and other prompt backgrounds such as VVV or ZZ events, are estimated using MC.

4.3 Monte Carlo Simulation Samples

Monte Carlo simulation allows us to model the SM prediction for the $W^{\pm}W^{\pm}$ signal region and to compare the excess to different $H_5^{\pm\pm}$ masses. ATLAS Monte Carlo samples are generated using a multistep process. First, an MC generator is used to produce the four vectors of all of the particles in an event. Next, the "hits" produced by these particles in the ATLAS detector is simulated using GEANT4 [27]. The hits are digitized by adding real-world effects, such as detector noise and pileup, so that the output is similar to the readout of the ATLAS detector [27]. Finally, the event and its physics objects are reconstructed using the same process as real data.

In this thesis, the MC20 simulation campaign for the full Run 2 dataset will be used, which corresponds to an integrated luminosity of 140 fb^{-1}. The MC20a subcampaign corresponds to 2015 and 2016 data, MC20d corresponds to 2017 data, and MC20e corresponds to 2018 data. Each simulated sample is assigned a unique dataset identifier (DSID), which is provided here for reference.

The $H_5^{\pm\pm}$ signal samples were produced at leading order using MadGraph5aMC@NLO 3.3.1 [28] and Pythia8.307 [29]. A complete list of the simulated $H_5^{\pm\pm}$ mass points and their cross-sections is provided in Appendix B. Note that the cross-section decreases with mass — the 200 GeV mass point has a cross-section of 16.0 fb whereas the 3000 GeV mass point has a cross-section of 0.00688 fb. Background samples were generated using Sherpa 2.2 [30], with the exception of some top quark samples that were generated using Powheg [31] and Pythia8 [29]. A complete list of background samples is provided in Appendix B.

Each simulated event has a corresponding event weight which accounts for the Monte Carlo generator event weight (w_{MC}), as well as for known differences between the simulated event and the actual collision data:

$$\text{Event Weight} = w_{\text{event}} = w_{\text{MC}} \cdot w_{\text{JVT SF}} \cdot w_{\text{pileup}} \cdot w_{\text{beamspot}} \tag{4.2}$$

The jet-vertex-tagger (JVT) scale factor ($w_{\text{JVT SF}}$) accounts for differences in the JVT performance between MC and data. The pileup weight (w_{pileup}) and beam position weight (w_{beamspot}) account for differences between the simulation assumptions and the data taking conditions with regards to pileup and beam position respectively.

To compare the sum of weights of a simulated sample to the number of events observed in data, the event weights must be multiplied by the interaction cross-section and the integrated luminosity, and normalized by the sum of weights of the sample, $\Sigma_i w_i$. This new weight will be referred to as the physical event weight:

$$\text{Physical Event Weight} = w_{\text{physical event}} = w_{\text{event}} \frac{\sigma \cdot \int \mathcal{L}}{\sum_i w_i} \tag{4.3}$$

4.4 Object Selection

Object selections ensure that the electrons, muons and jets that are being used in the signal and control regions are well reconstructed by the detector. This increases the likelihood that the objects used in the analysis are correctly identified, and that their four vectors are measured accurately.

Object selections also help to reduce the presence of background in the signal region since many background events results from leptons or jets being "faked" by other particles. All objects are required to pass the overlap removal procedure described in Section 3.2.3.

There are two types of lepton selection that are used in the $W^{\pm}W^{\pm}$ channel: the "baseline" lepton selection, shown in Table 4.1, and the "signal" lepton selection, shown in Table 4.2. The baseline lepton selection is less restrictive than the signal lepton selection and is used to apply a veto on events with 3 or more baseline leptons. This reduces the background contribution of $W^{\pm}Z$ and ZZ events where the third lepton has low p_{T} or is poorly reconstructed.

Electrons	**Muons**
$p_{\mathrm{T}} > 7$ GeV	$p_{\mathrm{T}} > 3$ GeV
$\lvert\eta\rvert < 2.47$	$\lvert\eta\rvert < 2.7$
$\lvert z_0 \sin\theta\rvert < 0.5$	$\lvert z_0 \sin\theta\rvert < 1.5$
Loose DNN	Loose
Pass Overlap Removal	Pass Overlap Removal

Table 4.1: The baseline lepton selection for the $H_5^{\pm\pm}$ search.

Electrons	**Muons**
$p_{\mathrm{T}} > 27$ GeV	$p_{\mathrm{T}} > 27$ GeV
$\lvert\eta\rvert < 2.47$, excluding $1.37 < \lvert\eta\rvert < 1.52$	$\lvert\eta\rvert < 2.7$
$\lvert z_0 \sin\theta\rvert < 0.5$	$\lvert z_0 \sin\theta\rvert < 0.5$
Tight DNN	Medium
Loose VarRad	Pflow Tight VarRad
Pass Overlap Removal	Pass Overlap Removal

Table 4.2: The signal lepton selection for the $H_5^{\pm\pm}$ search.

Jets
$p_{\mathrm{T}} > 20$ GeV
$\lvert\eta\rvert < 4.5$
Baseline JVT
Non b-jet (GN2 b-tagger 85%)
Pass Overlap Removal

Table 4.3: The jet selection for the $H_5^{\pm\pm}$ search.

The signal leptons are expected to have large p_{T}, so the p_{T} requirement for signal electrons and muons is 27 GeV, which is slightly higher than the lowest p_{T} single lepton triggers. The baseline p_{T} requirements are much lower, at 7 GeV for electrons and 3 GeV for muons. This is around the lower limit that is supported for general ATLAS analyses.

All leptons must pass minimum identification or quality requirements, which helps to reduce backgrounds from jets or photons faking leptons. The electron identification requirement is Loose DNN for baseline and Tight DNN for signal. The quality requirement for muons is Loose for baseline and Medium for signal.

There is no isolation requirement for baseline leptons, but there is for signal leptons (Loose VarRad for electrons and Pflow Tight VarRad for muons). Note that these isolation variables are

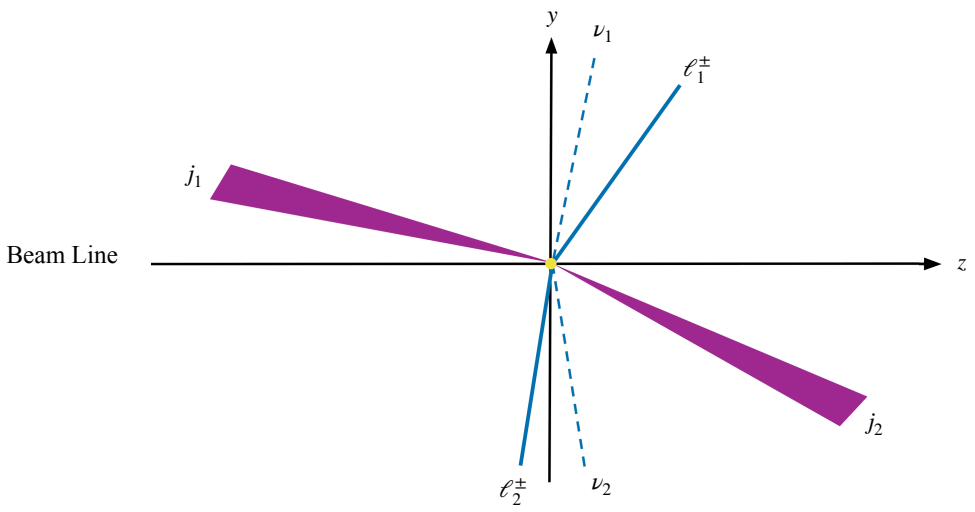


Figure 4.4: A typical event topology for a high mass $H_5^{\pm\pm}$ signal event. Each boosted W boson produces a lepton (ℓ) and a neutrino (ν) which tend to lie close together, and are located opposite to the other lepton and neutrino pair. The two jets (j) from the VBF process are in the forward regions and have a large rapidity difference.

"close-by-corrected" — meaning that a slight adjustment has been made to the normal isolation variable to account for nearby same object tracks. The close-by-correction improves the signal acceptance for the $W^{\pm}Z$ channel (see Appendix A), and so it is also used in the $W^{\pm}W^{\pm}$ channel to maintain consistency across the analysis.

All electrons and muons are required to be within the geometrical acceptance of the detector, which is $|\eta| < 2.47$ for electrons and $|\eta| <2.7$ for muons. There is a gap in the calorimeter geometry where electrons cannot be detected reliably ($1.37 < |\eta| < 1.52$), therefore this region is vetoed for signal electrons.

To ensure that the leptons originate from the primary vertex, electrons are required to have $|z_0 \sin\theta| < 0.5$, baseline muons are required to have $|z_0 \sin\theta| < 1.5$, and signal muons are required to have $|z_0 \sin\theta| < 0.5$. There is also a $|d_0/\sigma(d_0)| < 3$ requirement for muons and a $|d_0/\sigma(d_0)| < 5$ requirement for electrons, which is applied externally of this work.

The jet requirements are detailed in Table 4.3. They are required to have a minimum p_{T} of 20 GeV and fall within the detector acceptance for jets ($|\eta| < 4.5$). To reduce the background from other hadronic activity, they are required to pass the baseline JVT criteria. Events from top quarks can be suppressed by requiring that jets are not identified as b-jets using the b-tagging algorithm. The b-tagging algorithm is only available for $|\eta| < 2.5$, therefore this cut is only applied for jets in this region of the detector.

4.5 Signal Region

The signal region (SR) definition for the $W^{\pm}W^{\pm}$ channel should maximize the number of $H_5^{\pm\pm}$ events while minimizing the number of background events. To motivate the machine learning optimization

of the signal region, the present cuts-based signal region selection for this analysis is described.

Figure 4.4 visualizes a typical signal event process for a $H_5^{\pm\pm}$ produced by VBF and decaying fully leptonically via $W^{\pm}W^{\pm}$. The final state consists of two VBF jets, two leptons from the W decays, and missing transverse energy from the neutrinos. Thus, events in the signal region are required to have 2 signal leptons. It is also required that the invariant mass of the lepton pair, m_{ll}, be greater than 20 GeV due to poor modelling of low-mass processes [26]. Figure 4.5 shows that the W bosons from the $H_5^{\pm\pm}$ tend to be more back-to-back than background in the $x-y$ plane, especially at high mass. Thus, requiring that the ϕ separation of the leptons, $|\Delta\phi_{ll}|$, be greater than 1.5 improves the signal region purity. Additional requirements are placed on the ee channel due to the large background from Drell-Yan Z decays, which can contribute to the signal region if one of the electron charges is misidentified. These include the requirement that $|\eta| < 1.37$ for electrons, since the likelihood of charge flip increases in the endcap region, and the requirement that the invariant mass of the electron pair be separated from the PDG Z mass by at least 15 GeV. As discussed in the previous section, a third lepton veto is implemented to reduce the background from events with more than two leptons, such as $W^{\pm}Z$ or ZZ events.

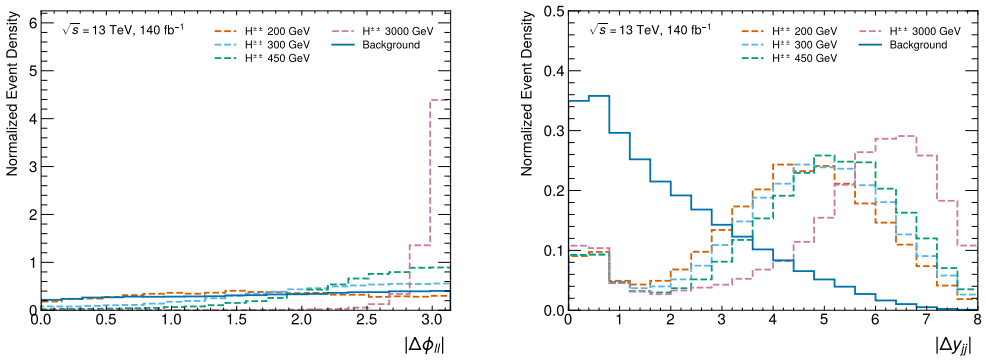


Figure 4.5: The $|\Delta\phi_{ll}|$ (left) and $|\Delta y_{jj}|$ (right) distributions for 200 GeV, 300 GeV, 450 GeV and 3000 GeV $H_5^{\pm\pm}$ signals and background (EW, QCD and interference $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples only).

We also expect there to be two VBF jets that are highly energetic. Thus, events are required to have at least two jets, with the leading jet $p_{\mathrm{T}} > 50$ GeV and the subleading jet $p_{\mathrm{T}} > 30$ GeV. To suppress the background from top processes, events that contain b-jets are rejected. Additional requirements can be made based on the jet topology of the VBF process, which is influenced by the colour connections between the initial quarks. Since there is no colour flow between the quarks that emit the vector bosons, it is unlikely for there to be hadronic activity in the central region of the detector [23]. Thus, the two VBF jets tend to lie close to the beam pipe (forward) and have a large rapidity gap (point in opposite directions), shown in Figure 4.5 [25]. This can also be understood from a more mathematical approach — the matrix element for VBF processes is maximized when the invariant mass of the jets is large and the jet scattering angle is small [32]. Therefore, for the cuts-based signal region, it is required that m_{jj} be greater than 500 GeV and $|\Delta y_{jj}|$ be greater than 2, where m_{jj} and $|\Delta y_{jj}|$ are calculated using the leading and subleading jets.

Due to the presence of two neutrinos in the final state, there should be some missing transverse

momentum in signal-like events. Therefore, it is required that $E_{\mathrm{T}}^{\mathrm{miss}}$ be greater than or equal to 25 GeV. A complete list of event selections for the cuts-based SR is provided in Table 4.4.

Cuts-Based SR Selection
Exactly two same-sign signal leptons with $p_{\mathrm{T}} > 27$ GeV ($\|\eta\| < 1.37$ in the ee channel)
$m_{ll} \geq 20$ GeV
$\|m_{ee}$ - $m_Z\| > 15$ GeV in the ee-channel
≥ 2 jets with leading and subleading jets satisfying $p_{\mathrm{T}} >$ 50 GeV and $p_{\mathrm{T}} > 30$ GeV respectively
Less than 3 baseline leptons
$E_{\mathrm{T}}^{\mathrm{miss}} \geq 25$ GeV
$m_{jj} > 500$ GeV
$\|\Delta y_{jj}\| > 2$
$\|\Delta\phi_{ll}\| > 1.5$
b-jet veto

Table 4.4: The event selection for the $H_5^{\pm\pm}$ cuts-based signal region.

The m_{T} distribution in the cuts-based $W^{\pm}W^{\pm}$ signal region for a 375 GeV $H_5^{\pm\pm}$ sample is provided in Figure 4.6. The signal sample cross-section corresponds to $\sin(\theta_H) = 0.25$, which is similar to the limit from Run 2 for this mass point [26]. As we can see, there are a significant number of background events expected in the m_{T} region near the mass peak of the 375 GeV $H_5^{\pm\pm}$ sample. The two largest background sources are electroweak $W^{\pm}W^{\pm}$ events and QCD $W^{\pm}Z$ events.

In order to improve our sensitivity to $H_5^{\pm\pm}$, we need to find a way of reducing the number of background events in the signal region. The cuts-based VBF selections ($m_{jj}, |\Delta y_{jj}|$) are well motivated by our understanding of VBF kinematics. However, they do not reject $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ background events that are produced by vector boson scattering (VBS) since they have a similar jet topology to VBF. As well, some of the cuts-based requirements may be more advantageous for high mass $H_5^{\pm\pm}$ than for low mass $H_5^{\pm\pm}$. Figure 4.5 shows that the $|\Delta\phi_{ll}|$ distribution is nearly flat for the 200 GeV signal sample. Finally, the relationships between kinematic variables may differ for signal and background, but this cannot be utilized in a regular cuts-based method. Neural networks are effective at solving complex multidimensional problems when provided with large quantities of labelled data [33]. I will show in the following chapters that it is possible to make greater use of the event kinematics and improve the background rejection rate using neural networks.

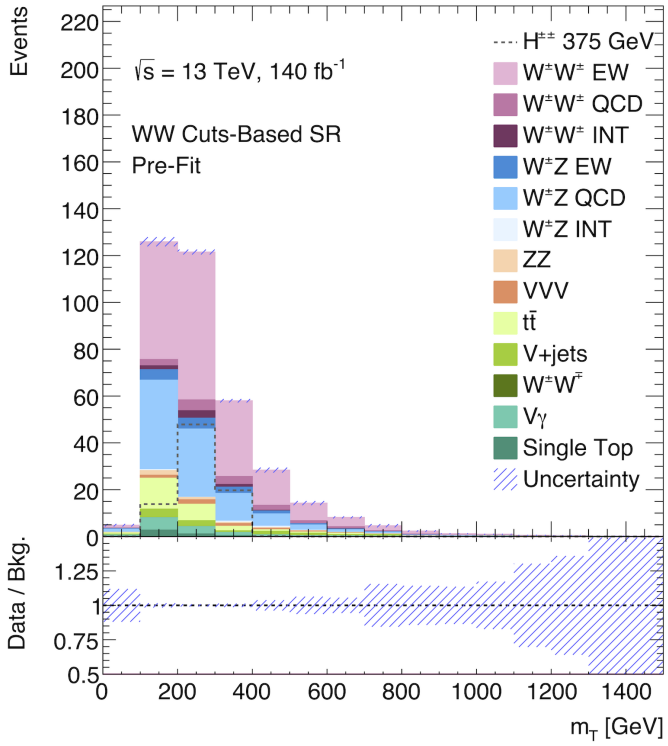


Figure 4.6: The m_T distribution in the cuts-based signal region. The signal sample is a 375 GeV $H_5^{\pm\pm}$ sample with $\sin(\theta_H) = 0.25$.

Chapter 5

Signal Region Optimization

In order to improve the signal region (SR) selection, a neural network (NN) is trained to classify $H_5^{\pm\pm}$ signal events (1) and background events (0) using simulation. The NN is then used to evaluate events on a scale of 0 to 1, with events closer to 1 being more signal-like and events closer to 0 being more background-like. A new signal region is defined to include all events which receive a NN score above a certain threshold. This threshold is referred to as the *working point* (WP).

The objective of this chapter is to develop a neural network that performs the best possible classification of $H_5^{\pm\pm}$ signal events and background events. This will maximize the purity of the SR as well as our sensitivity to the GM model. In addition to performing excellent classification, this network should also perform consistently across the $H_5^{\pm\pm}$ mass range of 200 GeV to 3000 GeV and demonstrate the ability to classify on mass points that it is has not seen before during training.

In Section 5.1, a "naive" approach to NN classification will be presented. In the following sections of Chapter 5 (5.2, 5.3, 5.4, and 5.5), different improvements will be explored.

5.1 Machine Learning Method

The type of NN that will be used is a multi-layer perceptron due to its simplicity and widespread use. All of the neural network training and evaluation is performed in Python using Tensorflow, an open-source machine learning software library created by Google [34].

5.1.1 Multi-Layer Perceptron

A multi-layer perceptron (MLP) uses layers of nodes to map a set of input values, x, to an output, $\hat{y}$, which is an estimate of the real value y. Each node in the MLP recieves N inputs from the previous network layer, $\mathbf{u} = (u_1, u_2, \ldots, u_N)$. The node then produces an output, z, by taking a weighted sum of the inputs with weights $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, adding a bias, b, and then applying an activation function, ϕ [35]:

$$z = \phi\left[\sum_i^N (w_i u_i) + b\right]. \tag{5.1}$$

The activation function used in this thesis is the Rectified Linear Unit (or ReLU), defined as $\phi(x) = \max(0, x)$ [36]. The weights and biases of the nodes in the NN are referred to as the *model parameters*, θ. The NN output is a function of the model parameters and the input values, i.e. $\hat{y} = f_\theta(x)$.

The layers between the input layer and the output are referred to as the *hidden layers*. The number of nodes in each hidden layer is the network *width* and the number of hidden layers is the network *depth*. The MLP can be used to perform binary classification by encoding each class (category) as a 0 or 1, and adding a sigmoid function after the hidden layers to bound the NN output between 0 and 1. A generic structure for a MLP classifier is provided in Figure 5.1.

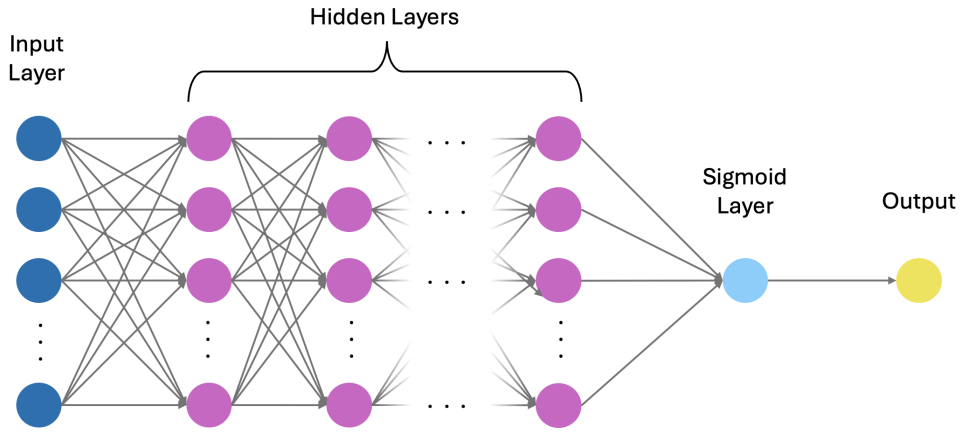


Figure 5.1: The general structure of an MLP NN used for binary classification.

Training refers to the process where the weights and biases of the nodes are tuned to maximize the agreement between the NN output, $\hat{y}$, and the true labels, y, for a set of training data. This is accomplished by optimizing the model parameters using an iterative gradient descent method. Each step, the model parameters are updated by subtracting the gradient of the loss function, $\mathcal{L}$, with respect to the model parameters [35]:

$$\theta_t = \theta_{t-1} - \lambda \nabla_\theta \mathcal{L}(x, \theta, y). \tag{5.2}$$

The loss function quantifies the magnitude of the disagreement between $\hat{y}$ and y. Provided that the learning rate, λ, is small, the loss should decrease during training, improving the NN classification until the global minimum (smallest possible loss) is found.

In practice, training is done by separating the set of training events into smaller *batches*. The model parameters are updated after each batch, and once all the batches have been seen by the model we have completed an *epoch*. The batch size and number of epochs can vary significantly depending on the machine learning context. A set of events which are not seen during training, called the validation set, are reserved to ensure that the NN model is *generalizable*. This means it is able to classify events that were not seen during training. If the training loss diverges from the

validation loss, then we say that the network is *overtrained* or *overfit*. It is best to stop training right before this point [37]. In order to ensure that this happens, we use *early stopping* to halt training when the validation loss does not decrease by a minimum amount after a certain number of epochs, called the *patience*.

A loss function that is commonly used for classification is binary cross-entropy (BCE). For a training batch containing N events with corresponding event probabilities q, the BCE loss is [35]:

$$\mathcal{L}_{\text{BCE}}(x, \theta, y) = -\sum_{i}^{N} q_i \left[(1 - y_i) \log(1 - \hat{y}_i) + y_i \log(\hat{y}_i) \right]. \tag{5.3}$$

In general, $q = \frac{1}{N}$ is used so that all events are equally weighted. However, if event weights are used, then $q_i = \frac{w_i}{\sum_j w_j}$. As the difference between the NN prediction and the correct classification increases, BCE loss penalizes the network in a logarithmically increasing manner.

For the Run 2 $W^{\pm}Z$ analysis, a variant of gradient descent called stochastic gradient descent (SGD) was used to optimize the NN [25]. This method uses a random sample of events in the training batch to calculate the gradient. This reduces the computational cost of the gradient descent and increases the likelihood of finding a global minimum compared to classical gradient descent [35]. However, SGD can be slow to converge due to its constant learning rate.

Another popular choice of optimization algorithm is Adam, which is a modified version of SGD [38]. Adam adaptively manages the learning rate, using past gradients to add momentum and using past gradients squared to slow down the descent in the steepest directions [38]. This ensures a quick and stable convergence to the global minimum, which makes it a better choice for this study.

A brief comparison was made of the two optimization algorithms to justify this change in approach. The loss curves in Figure 5.2 show that using Adam results in training that is more stable and converges on a lower minimum. Another benefit of using Adam is that the optimization is less sensitive to the initial choice of learning rate since it is managed adaptively. It was also found that a learning rate of 0.001 delivers the best results.

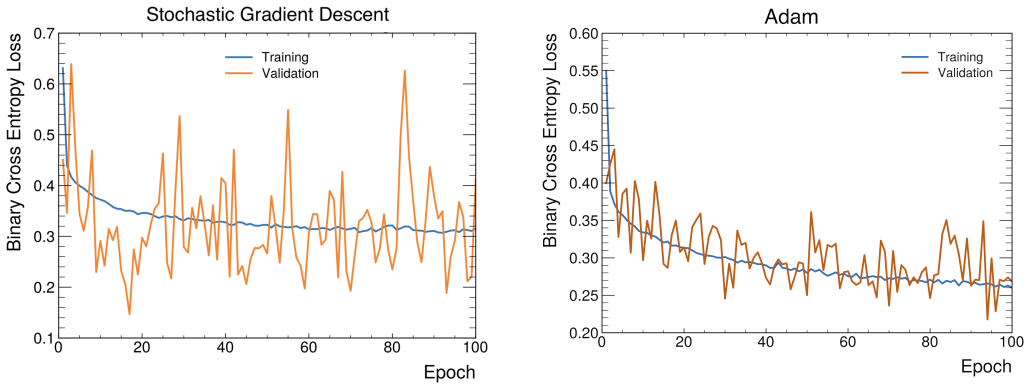


Figure 5.2: The loss curves for a network optimized with stochastic gradient descent (left) and Adam (right). The corresponding AUCs were 0.944 and 0.958 respectively (unweighted).

5.1.2 Training Region

The neural network training region (NN TR) maintains the basic requirements of the final state signature, such as having two signal leptons, two high p_{T} jets, and large missing transverse momentum. However, it is not as restrictive as the cuts-based signal region since the cuts on Δy_{jj}, $\Delta\phi_{ll}$, and m_{jj} from Table 4.4 are removed. This allows the network to develop its own "signal-like" criteria. The complete list of NN training region selections are detailed in Table 5.1.

NN Training Region Selection
Exactly two same-sign signal leptons with $p_{\mathrm{T}} > 27$ GeV ($\lvert\eta\rvert < 1.37$ in the ee channel)
$m_{ll} \geq 20$ GeV
$\lvert m_{ee}$ - $m_Z\rvert > 15$ GeV in the ee-channel
≥ 2 signal jets with leading and subleading jets satisfying $p_{\mathrm{T}} > 65$ GeV and $p_{\mathrm{T}} > 35$ GeV respectively
Less than 3 baseline leptons
$E_T^{\mathrm{miss}} \geq 30$ GeV

Table 5.1: The event selection for the NN training region. Note that the b-jet requirement is removed and the the $E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}(j_1)$ and $p_{\mathrm{T}}(j_2)$ cuts are slightly different from Table 4.4. The signal region definitions for this analysis were evolving externally from analysis.

5.1.3 MC Samples

In order to train the NN, MC simulated events are required, since they can be labelled as signal and background. The signal and background samples that are used here correspond to the Run 2 integrated luminosity of 140 fb^{-1}. This means they include the ATLAS MC subcampaigns of MC20a, MC20d, and MC20e. Eventually, a separate neural network will need to be trained for Run 3 since the centre-of-mass energy is different from Run 2.

The signal samples that are used in training are $H_5^{\pm\pm}$ samples at the following mass points (in GeV): 200, 250, 300, 350, 400, 450, 500, 550, 700, 900, 1500, and 3000. Half of the $H_5^{\pm\pm}$ samples are reserved to study the interpolation ability of the neural network (see Sections 5.4 and 6.1.2). These mass points are (in GeV): 225, 275, 325, 375, 425, 475, 525, 600, 800, 1000, and 2000. Ultimately, the final neural network will be trained on all mass points to maximize the performance and interpolation ability of the network. However, this thesis does not extend to the completion of the analysis and so this will not be studied here.

The background samples that are used in training are $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ electroweak, QCD, and interference samples with two jets in the final state. As mentioned earlier, these two backgrounds comprise roughly ∼80% of the background events in the signal region. Since the non-prompt and charge-flip backgrounds are not well-modelled, it was assumed that there would not be much benefit in including them during training. The backgrounds used in training were also chosen to ensure compatibility with previous neural network approaches (Run 2 $W^{\pm}Z$) and work being done by other analyzers. A summary of the signal and background samples used for training is provided in Table 5.2.

The signal and background events are divided into 3 sets based on the event number. If the

event number modulus 5 is equal to 0, the event is put in the test set. If the event number modulus 5 is equal to 1, the event is put into the validation set. All other events are placed in the training set. In this way, the training set consists of roughly 60% of the events, and the validation set and test set each consist of roughly 20% of the events. During hyperparameter tuning, the training and validation set performance are compared to determine the optimal hyperparameters. The test set is reserved for performance evaluation after hyperparameter tuning is complete, including the evaluation of the network performance as a function of doubly-charged Higgs mass.

Sample	$\sum w_i$ **in NN TR**	**# of Events in NN TR**
$H_5^{\pm\pm}$ 200	135.5	12738
$H_5^{\pm\pm}$ 250	124.6	15696
$H_5^{\pm\pm}$ 300	111.2	18252
$H_5^{\pm\pm}$ 350	98.71	20723
$H_5^{\pm\pm}$ 400	85.14	22430
$H_5^{\pm\pm}$ 450	71.32	23247
$H_5^{\pm\pm}$ 500	62.11	24924
$H_5^{\pm\pm}$ 550	53.35	23537
$H_5^{\pm\pm}$ 700	34.00	28235
$H_5^{\pm\pm}$ 900	19.11	29912
$H_5^{\pm\pm}$ 1500	3.942	30707
$H_5^{\pm\pm}$ 3000	0.1244	26992
$W^{\pm}W^{\pm}jj$ EW	476.3	150600
$W^{\pm}W^{\pm}jj$ QCD	216.9	413519
$W^{\pm}W^{\pm}jj$ INT	44.83	7065
$W^{\pm}Zjj$ EW	60.30	12940
$W^{\pm}Zjj$ QCD	1079.8	294439
$W^{\pm}Zjj$ INT	23.04	557

Table 5.2: The signal ($H_5^{\pm\pm}$) and background samples used for NN training, which correspond to the Run 2 integrated luminosity of 140 fb^{-1}. The signal sample cross-sections correspond to $\sin(\theta_H) = 0.25$. The middle column indicates the sum of weights in the NN training region and the right column indicates the number of events in the NN training region.

5.1.4 Metrics

In order to compare the performance of different NNs, it is necessary to define classification performance metrics. The true positive rate (TPR, also known as signal acceptance rate, recall, or sensitivity) is the percentage of positives (signal events) that are correctly classified. The false positive rate (FPR) is the percentage of negatives (background events) that are incorrectly classified. The TPR and FPR are calculated using the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN):

$$\text{TPR} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \qquad \text{FPR} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \qquad \text{(unweighted)}$$

If the events are weighted, then the TPR and FPR are calculated by summing the event weights for the TP, FN, FP and TN events:

$$\text{TPR} = \frac{\sum_{\text{TP}} w_i}{\sum_{\text{TP}} w_i + \sum_{\text{FN}} w_i} \qquad \text{FPR} = \frac{\sum_{\text{FP}} w_i}{\sum_{\text{FP}} w_i + \sum_{\text{TN}} w_i} \qquad \text{(weighted)}$$

One of the performance metrics that will frequently be used in this thesis is the area under the ROC curve (AUC). A ROC curve is a plot of the TPR against the FPR for a dense set of NN output working points. The integral of the ROC curve, the AUC, provides a metric of the classification quality across the entire range of working points. Larger values of AUC correspond to better classification — if the network performs perfect classification, then the AUC will be 1.

The AUC does not indicate which NN will provide the best possible classification at a single WP, which is ultimately how the NN will be used. For this reason, another metric that will be used is the expected signal significance (Z). Assuming no systematic error, the statistical significance of an excess over the background only hypothesis is given by [39]

$$Z = \sqrt{2\left[n \ln \frac{n}{b} - (n-b)\right]}, \tag{5.4}$$

where n is the number of measured events and b is the expected number of background events. Here, $n = s + b$, where s is the sum of weights of the signal MC samples and b is the sum of weights of the background MC samples. The significance can be used to evaluate the performance of the NN by constructing a signal region that includes all events that score above a certain NN WP. When calculating Z with a mass point that was used in training the NN, the significance needs to be assessed using only events that were not in the training set — this includes the validation and test sets which consist of 40% of events. Therefore, to estimate the significance corresponding to the entire Run 2 integrated luminosity, Equation 5.4 is used with $s \to s/0.4$ and $b \to b/0.4$.

Two additional metrics will be used in Chapter 6 which are useful to define here. The first is the classification accuracy:

$$\text{Accuracy} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}} + N_{\text{FN}}} \qquad \text{(unweighted)}$$

$$\text{Accuracy} = \frac{\sum_{\text{TP}} w_i + \sum_{\text{TN}} w_i}{\sum_{\text{TP}} w_i + \sum_{\text{FP}} w_i + \sum_{\text{TN}} w_i + \sum_{\text{FN}} w_i} \qquad \text{(weighted)}$$

The second is the background rejection rate (also called specificity or true negative rate)

$$\text{Background Rejection Rate} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FN}}} \qquad \text{(unweighted)}$$

$$\text{Background Rejection Rate} = \frac{\sum_{\text{TN}} w_i}{\sum_{\text{TN}} w_i + \sum_{\text{FN}} w_i} \qquad \text{(weighted)}$$

In general, all metrics presented in this thesis will be weighted using physical event weights (see Equation 4.3) and exceptions will be explicitly stated.

5.1.5 Class Weights

In this analysis, the total number of events in the background samples (∼880,000) is larger than the total number of events in the signal samples (∼280,000). As a result, a neural network trained on these samples will be more heavily penalized for misclassifying background than signal. Without any correction, this results in a classification distribution which peaks sharply at 0 for background events, but is quite uniform for signal. This is not desirable when using neural networks for signal region definition since for some choice of WP, the background rejection may be high, but the signal acceptance will be low.

In order to correct this bias, the class weights method is used. Two scaling factors are calculated, f_s for signal and f_b for background, such that the product of the class weight and the number of events in the class is equal for signal and background:

$$f_s = \frac{N_s + N_b}{2N_s}, \quad \text{and} \quad f_b = \frac{N_s + N_b}{2N_b}. \tag{5.5}$$

This is implemented in the loss function in Equation 5.3 by setting $q_i = \frac{f_s}{N}$ for signal and $q_i = \frac{f_b}{N}$ for background. In this way, the neural network will be equally penalized for misclassifying signal and background, regardless of the relative number of signal events to background events.

5.1.6 Input Features

There are many kinematic variables which could be used as input features for this particular event signature. It is important not to limit the choice of input features based on the cuts-based selection, since the neural network may be able to capture multidimensional relationships between features that are not obvious when comparing a one-dimensional distribution. Figures 5.3, 5.4, and 5.5 compare the signal and background distributions for each of the candidate input features. The signal distribution includes all of the $H_5^{\pm\pm}$ mass points in Table 5.2, and the background distribution includes all of the $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples in Table 5.2. Additional plots of the input features with the signal distribution split into several mass points are provided in Appendix C.

For each of the leptons and jets, there are four components of p^{μ}: p_{T}, η, ϕ, and E. The individual ϕ distributions for all objects, $\phi(j_1), \phi(j_2), \phi(l_1)$, and $\phi(l_2)$, are expected to be uniform since the proton-proton beam is not polarized. Figures 5.3 and 5.4 confirm that this is the case, and so the ϕ variables are not used for the naive approach. However, in order to fully rule out the possibility these ϕ variables are useful to the neural network, they will be included in a feature optimization procedure in Section 5.2. For jets and especially leptons, p_{T} and E are highly correlated, so only p_{T} is used. The invariant mass of the leading jet pair, m_{jj} is also important due to the fact that we expect m_{jj} to be large for VBF processes.

There are also many ways to capture the spatial relationships between objects as input features. For jets, $|\Delta y_{jj}|$ is expected to be large for signal due to the characteristic VBF jet signature, making it a good choice as an input feature. We can also consider the angular separation of the jets in the ϕ direction, $\Delta\phi_{jj}$. Figure 5.3 shows that $\Delta\phi_{jj}$ tends to be larger for background than for signal. For leptons, there are three angular variables that can be constructed: $|\Delta y_{ll}|$, $\Delta\phi_{ll}$ and ΔR_{ll}. Figure 5.4 shows that there are some differences between signal and background in these

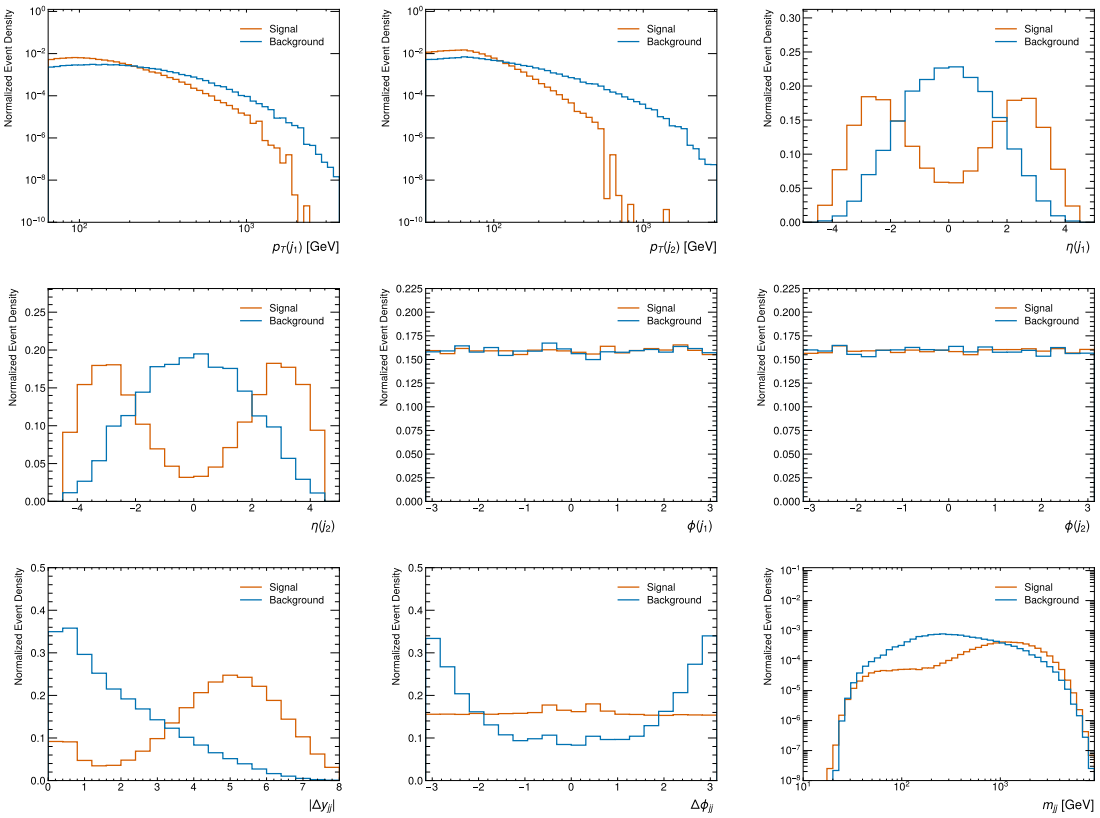


Figure 5.3: Jet input feature distributions for signal (all $H_5^{\pm\pm}$ samples in Table 5.2) and background (all $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples in Table 5.2). Samples are combined using the physical event weight from Equation 4.3. Note that individual ϕ variables are not used in the naive approach.

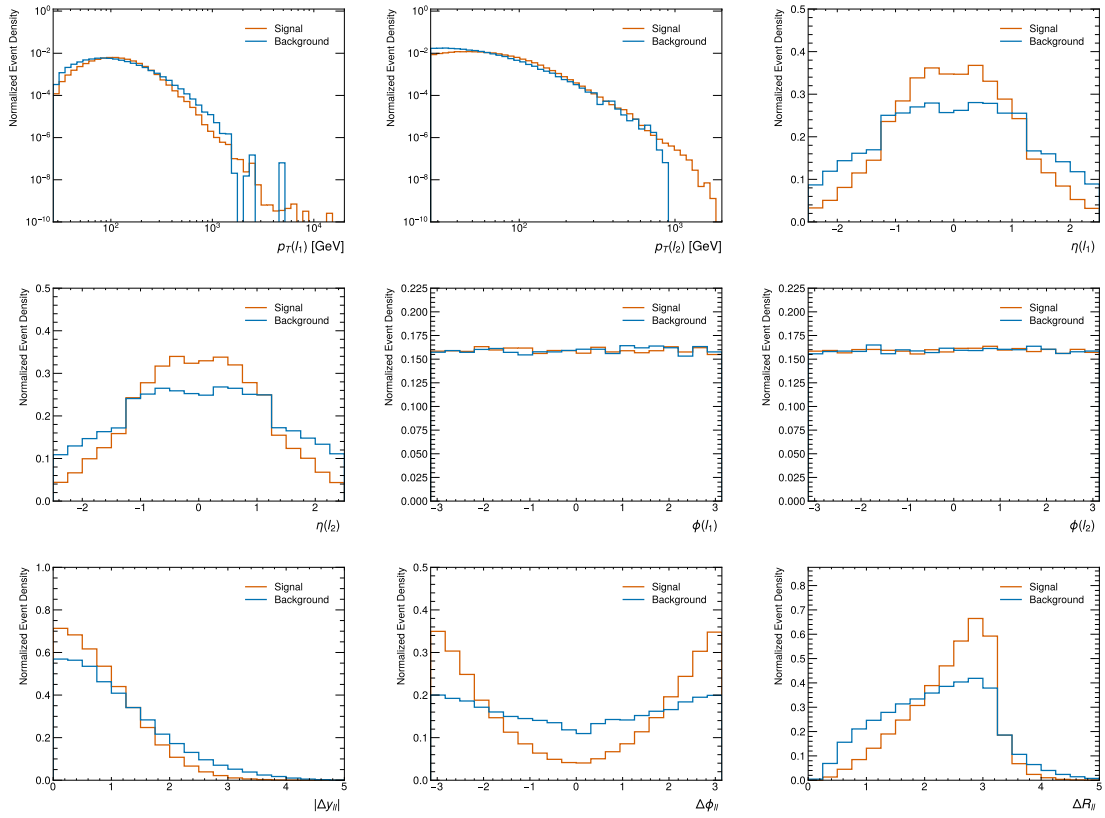


Figure 5.4: Lepton input feature distributions for signal (all $H_5^{\pm\pm}$ samples in Table 5.2) and background (all $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples in Table 5.2). Samples are combined using the physical event weight from Equation 4.3. Note that individual ϕ variables are not used in the naive approach.

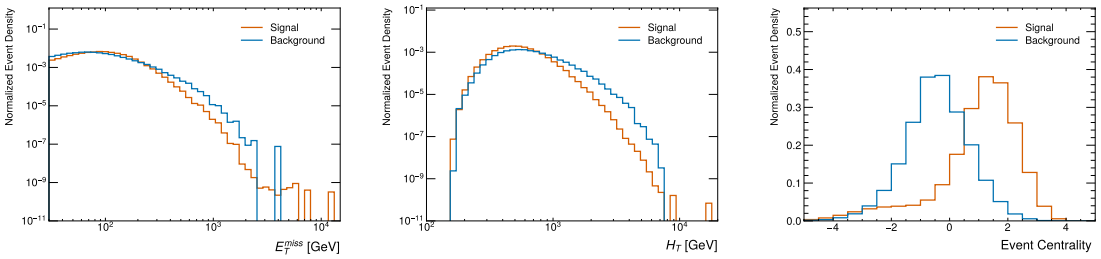


Figure 5.5: Event-level input feature distributions for signal (all $H_5^{\pm\pm}$ samples in Table 5.2) and background (all $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples in Table 5.2). Samples are combined using the physical event weight from Equation 4.3.

distributions, especially for $\Delta\phi_{ll}$ and ΔR_{ll}.

Event centrality (ξ) is a variable which is specifically useful for quantifying whether the relative positions of leptons and jets is consistent with the VBF topology. The following definition of ξ has been modified from the Run 2 $W^{\pm}Z$ analysis for the $W^{\pm}W^{\pm}$ channel. First, two terms are calculated: the difference in η between the most $-z$ pointing lepton and the most $-z$ pointing jet, and the difference in η between the most $+z$ pointing jet and the most $+z$ pointing lepton. Then ξ is defined to be the minimum of these two terms,

$$\xi = \min[\min(\eta_{l_1}, \eta_{l_2}) - \min(\eta_{j_1}, \eta_{j_2}), \max(\eta_{j_1}, \eta_{j_2}) - \max(\eta_{l_1}, \eta_{l_2})]. \tag{5.6}$$

If the jets are much more forward than the leptons (i.e. the leptons are more transverse) then both terms will be positive, yielding a positive value of ξ. If the leptons are much more forward than the jets, then both terms will be negative, yielding a negative value of ξ. Figure 5.5 shows that due to the VBF topology discussed in Section 4.5, the ξ variable tends to be larger for signal than for background.

Finally, there are two event-level kinematics which are included. The missing transverse momentum should be significant for the $W^{\pm}W^{\pm}$ process due to the presence of two neutrinos in the final state. We can also use H_{T}, which characterizes how hard the interaction was (using a classical analogy, a large H_{T} corresponds to a small impact parameter). A complete list of candidate input variables and their definitions is provided in Table 5.3.

Input	**Definition**
$p_{\mathrm{T}}(j_1), \eta(j_1), \phi(j_1)$*	$p_{\mathrm{T}}, \eta, \phi$ of leading (highest p_{T}) jet
$p_{\mathrm{T}}(j_2), \eta(j_2), \phi(j_2)$*	$p_{\mathrm{T}}, \eta, \phi$ of subleading (second highest p_{T}) jet
$\lvert\Delta y_{jj}\rvert$	$\lvert y(j_1) - y(j_2)\rvert$
$\Delta\phi_{jj}$	ϕ angle between $\phi(j_1)$ and $\phi(j_2)$ such that $\phi\epsilon[-\pi, \pi]$
m_{jj}	$\sqrt{(E(j_1) + E(j_2))^2 - \lvert\vec{p}(j_1) + \vec{p}(j_2)\rvert^2}$
$p_{\mathrm{T}}(l_1), \eta(l_1), \phi(l_1)$*	$p_{\mathrm{T}}, \eta, \phi$ of leading (highest p_{T}) lepton
$p_{\mathrm{T}}(l_2), \eta(l_2), \phi(l_2)$*	$p_{\mathrm{T}}, \eta, \phi$ of subleading (second highest p_{T}) lepton
$\lvert\Delta y_{ll}\rvert$	$\lvert y(l_1) - y(l_2)\rvert$
$\Delta\phi_{ll}$	ϕ angle between $\phi(l_1)$ and $\phi(l_2)$ such that $\phi\epsilon[-\pi, \pi]$
ΔR_{ll}	$\sqrt{(\eta(l_1) - \eta(l_2))^2 + (\Delta\phi_{ll})^2}$
E_T^{miss}	See Equation 3.7
H_{T}	Scalar sum of p_{T} of all event objects
ξ	See Equation 5.6

Table 5.3: The definitions of NN input features. *Note that ϕ was not used as an input for the naive approach.

The Pearson correlation coefficient is a normalized covariance of two random variables X and Y [40]:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{5.7}$$

If the two variables are perfectly linearly correlated, then $\rho_{x,y}$ will be 1, if they are perfectly linearly anti-correlated, then $\rho_{x,y}$ will be -1, and if they are not linearly correlated then $\rho_{x,y}$ will be 0.

Figure 5.6 shows the matrix of Pearson correlation coefficients for each of the input features

and the fit variable m_{T}. Many of the features are highly correlated and thus may be redundant. In Section 5.2, a feature optimization procedure will be performed to determine if some input features could be removed while maintaining classification performance. Many of the features are also correlated with m_T, which could be a problem if the NN introduces a bias in the fitting variable. This is explored further in Section 6.1.3.

Figure 5.7 shows the difference in Pearson correlation coefficients between signal events and background events. This reveals that there are many input variables which have different correlations for signal and background, such as $E_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}(l_1)$, or H_{T} and $p_{\mathrm{T}}(j_1)$. Although the Pearson correlation coefficients only capture linear relationships, these differences help to illustrate the complexity that machine learning could utilize to improve the event selection.

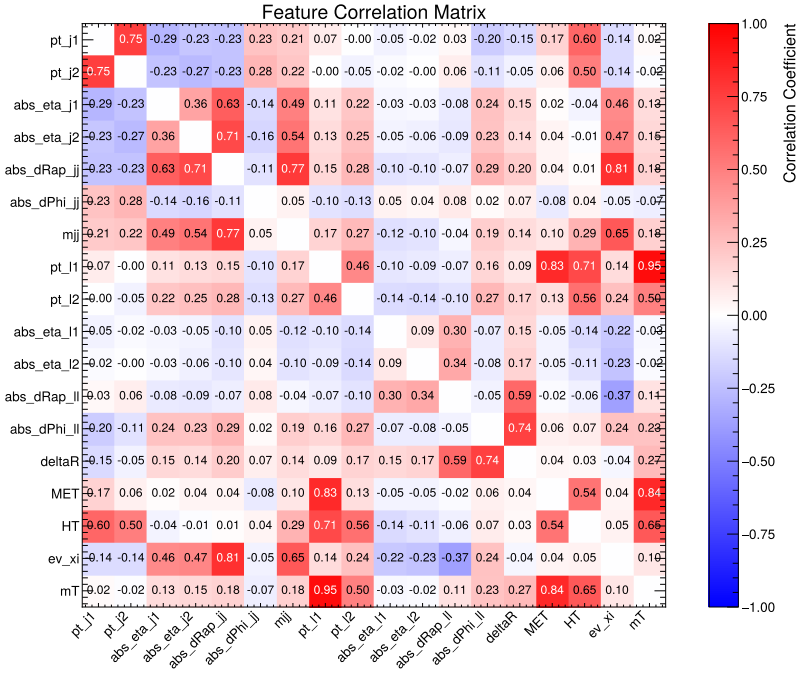


Figure 5.6: The Pearson correlation coefficients for the 17 input features used in the naive approach and m_{T}. They are calculated using all signal and background samples in Table 5.2.

5.1.7 Input Scaling

The input features must be scaled in some way prior to training. Large inputs (> 1) can cause the gradient descent to diverge, and features which have large differences in scale (e.g. m_{jj} ~100 GeV, ξ ~1) will not be weighted equally.

One method that can be used to address this is "min-max scaling". Prior to training or evaluation, the minimum value of each input is mapped to 0, and the maximum value of each input is mapped to 1. This was used to make some preliminary results, such as Figure 5.2, and Table 5.4. One of the challenges with min-max scaling is that when there are large outliers, especially for variables with large tails, the majority of the values in the distribution will be compressed into a small range.

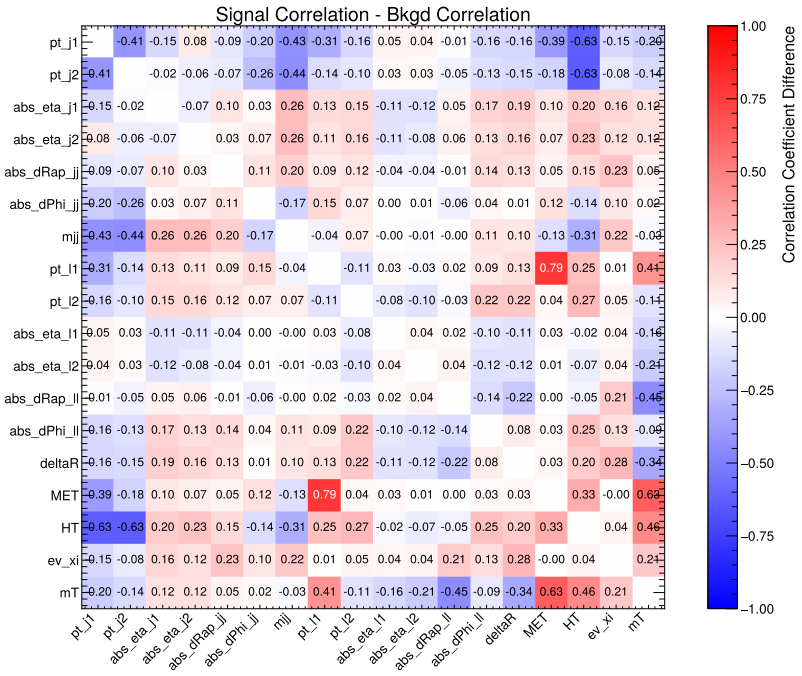


Figure 5.7: The difference in Pearson correlation coefficients between signal and background.

A more sophisticated approach is to use batch normalization layers, which are inserted between the NN inputs and the first network layer, as well as between subsequent network layers. For each training batch, the input to the batch normalization layer is scaled such that its mean is 0 and its standard deviation is 1. Next, a linear transformation is applied using learned parameters. This not only ensures that the NN inputs are scaled but also accelerates training by ensuring that the distribution of features passed between network layers does not shift during training (called internal covariate shift) [41]. Batch normalization also prevents overfitting since the mean, standard deviation, and learned parameters of the linear transformation will vary between batches [41]. During evaluation, the mean and variance from the entire training set are used, as well as the final learned parameters for the linear transformation. Batch normalization is used for all of the neural networks presented in the remainder of the thesis. Figure 5.8 shows that batch normalization dramatically speeds up training while also finding a lower minimum.

5.1.8 Neural Network Architecture

The choice of width and depth for an MLP classifier, shown in Figure 5.1, can impact several aspects of NN performance. Larger networks contain more parameters and can thus represent more complex relationships, however they are more prone to overfitting. The NN structure optimized for the Run 2 $W^{\pm}Z$ analysis was chosen as a starting point for this work [25]. It is a simple NN consisting of 2 hidden layers of 45 nodes, and a final layer of one node with a sigmoid activation function.

A brief study is performed using only MC20a samples to determine whether it would be beneficial to use a network with greater depth or width. The results in Table 5.4 show that increasing the

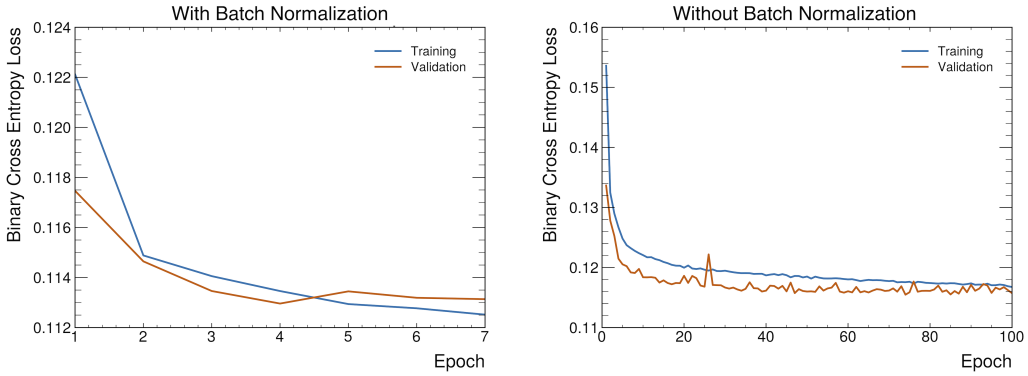


Figure 5.8: A comparison of the loss function with (left) and without (right) batch normalization. The use of batch normalization accelerates training — overtraining begins at around ∼100 epochs without batch normalization, whereas overtraining begins around ∼5 epochs with batch normalization.

Width	**Depth 2**	**Depth 3**	**Depth 4**
45	0.944	0.949	0.949
50	0.945	0.947	0.947
60	0.947	0.950	0.952

Table 5.4: The AUC for different neural network dimensions. All features from Table 5.3 are used as inputs except jet and lepton ϕ values.

network width and depth marginally improves the network performance. However, over repeated trials the improvement is not consistent. The ranking is highly dependent on the choice of patience for early stopping. Furthermore, when the highest ranking network (Width 60 Depth 4) is compared to the lowest ranking network (Width 45 Depth 2) using all campaigns in MC20, the results are reversed. The larger network has an AUC of 0.947 while the smaller network has an AUC of 0.949. This could be due to the fact that there are only 17 input features per event, which is a fairly "simple" input compared to images or audio which benefit from deep neural networks. Thus, a small network may be sufficient to capture the relationships between input features for signal and background. Since it is not clear that increasing the network size will improve the results, a width of 45 and depth of 2 is chosen for this analysis.

5.1.9 Naive Method Results

A summary of the NN training methods and hyperparameters is provided in Table 5.5. The batch size, learning rate, maximum epochs, early stopping, minimum Δ, and patience were each optimized independently through an iterative approach. The validation set AUC was used to choose the optimal hyperparameter values. The large batch size is not surprising considering that the training and validation sets are also both quite large.

Prior to training a neural network for the entire mass range, it is useful to understand the classification performance that can be achieved at each mass point using a dedicated NN. The green curve in Figure 5.9 shows the validation set AUC for 12 different networks trained at the 12 different

Batch Size	1000
Loss Function	Binary Cross Entropy
Optimizer	Adam
Learning Rate	0.001
Max. Epochs	50
Early Stopping Min. Δ	0.0002
Early Stopping Patience	3

Table 5.5: A summary of the neural network methods and hyperparameters. Note that early stopping is only used with sample weights, and not with class weights.

$H_5^{\pm\pm}$ masses in Table 5.2. In general, the NN classification improves with mass due to the fact that at larger mass points, the event kinematics are more distinctive from background: the VBF signature is enhanced, there is a large amount of transverse momentum, and the angular separation of the leptons is greater. There is, however, a slight decrease in the NN performance between 200 GeV and 300 GeV. For some kinematic distributions, the 250 GeV and 300 GeV mass points are more similar to background than the 200 GeV mass point, which slightly reverses this trend. Figure 5.10 demonstrates this effect for ΔR_{ll}, H_{T}, and the fitting variable m_{T}.

In practice, using 12 different NNs to perform 12 different hypothesis tests is impractical for our analysis. This would require each of the 12 NN SRs to be studied independently. Therefore, a network is trained on an ensemble of these 12 mass points with the objective of creating a general NN that can perform signal and background discrimination across the entire mass region. The performance of this NN is shown in the blue curve in Figure 5.9. The uncertainties for these points are estimated using the method described in Appendix E. These uncertainties were not applied to the dedicated NN points since the training conditions are quite different when using only a single mass point. At all mass points, the AUC is smaller using the general NN. This is expected due to the fact that the general network is required to accommodate many $H_5^{\pm\pm}$ masses which have varying characteristics.

Figure 5.9 also shows that the discrepancy between the dedicated NN performance and the general NN performance is much larger at low mass points than at high mass points. This can be attributed to the fact that the input distributions are quite different for mass points above and below 300 GeV relative to background. Figure 5.10 shows that the ΔR_{ll} and H_{T} distributions for the 200 GeV sample peak lower than background. At 300 GeV, the distribution of ΔR_{ll} peaks around the same value as background but peaks lower than background in the H_{T} distribution. Finally, for mass points above 300 GeV, the ΔR_{ll} and H_{T} distributions peak larger than background. During training, the NN is mostly shown signal events that are above 300 GeV. Thus, events below 300 GeV appear very different from the "average" signal event shown to the network, and are poorly classified.

The dedicated NN curve shows that it is possible to perform better classification below 350 GeV, but that these mass points are not being prioritized during training. Since the excess from Run 2 is around $m_{H_5} = 375$ GeV, and this low mass region is especially sensitive to the $\sin(\theta_H)$ parameter, we would like to improve the performance of the network at low mass.

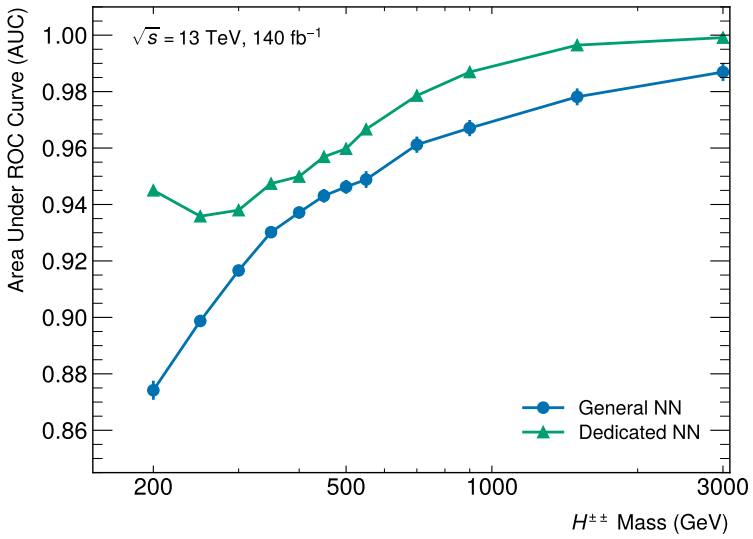


Figure 5.9: A comparison of the AUC for dedicated NNs trained for each mass point to a general NN trained on all mass points plotted as a function of the $H_5^{\pm\pm}$ mass. The AUC uncertainties for the general NN are estimated in Appendix E.

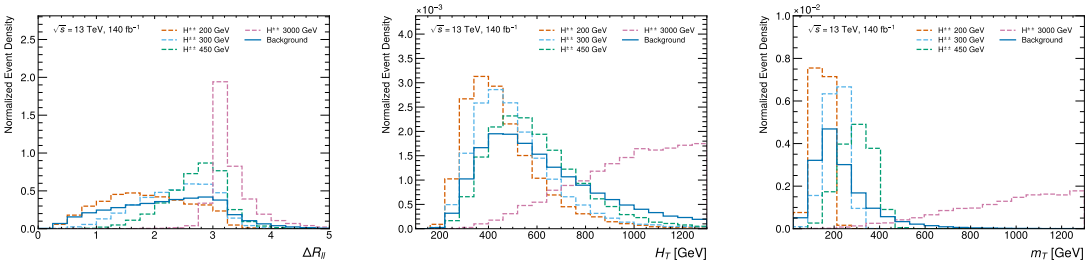


Figure 5.10: The distribution of $\Delta R_{ll}, H_{\mathrm{T}}$ and m_T for background and the 200 GeV, 300 GeV, 450 GeV, and 3000 GeV signal mass points. This shows that the behaviour of signal compared to background differs for doubly-charged Higgs masses below and above 300 GeV.

5.2 Feature Optimization

The general NN presented in the previous section uses 17 input features that were selected manually based on the fact that the signal and background distributions were different. It would be useful to know whether this set of input features could be reduced further without compromising performance, as many of the input features are correlated. It would also be useful to know if the individual ϕ variables, which were previously discarded, are in fact useful to the network.

One way to perform feature selection is by ranking features based on how much removing (or scrambling) that input feature impacts the ML performance [42]. This feature importance (FI) can be defined using the AUC metric as:

$$\mathrm{FI} = \frac{\text{AUC with all features}}{\text{AUC with feature scrambled}}. \tag{5.8}$$

The problem with selecting features based on their FI ranking is that two features that are highly correlated tend to both score poorly. When one feature is scrambled, the NN can use information from the other feature.

Therefore, it is better to perform an iterative procedure which removes the lowest ranking feature one at a time and then recalculates the FI ranking. This method is defined below:

1. Train the NN with all input features.
2. Train the NN with each input feature scrambled one at a time.
3. Rank features based on feature importance - the ratio of the original AUC to the AUC with that feature scrambled.
4. Remove the lowest ranking feature.
5. Repeat.

This feature optimization procedure could also be useful in addressing the low mass problem. If we determine which features are the most important for classification at low mass, then using these features when training the general NN may improve performance at low mass.

This feature selection procedure is performed twice: once with all of the mass points, and once with only low mass points (200, 225, 250, 275, and 300 GeV). All of the features in Table 5.3 are used, giving an initial set of 21 input features. Figure 5.11 shows that in both cases the AUC slowly begins to decrease as the number of input features is reduced, and then falls off more sharply below around 14 features. For this reason, 14 features seems to be a somewhat "natural" choice since the AUC is not significantly lower than the initial AUC.

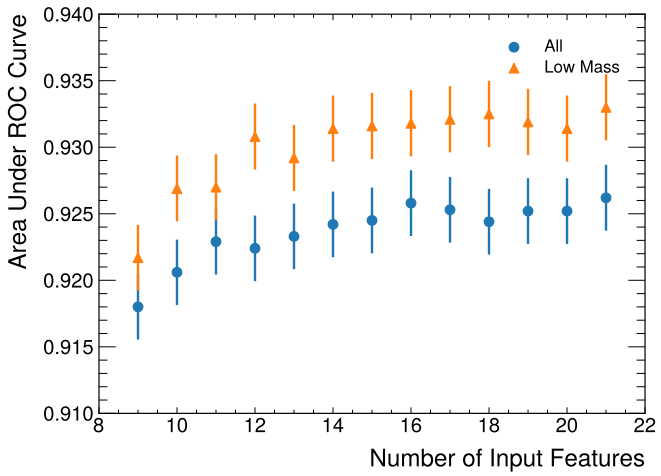


Figure 5.11: The AUC for the network training on all mass points and the network training on only low mass points throughout the feature optimization procedure. The performance generally declines as the number of input features decreases. The percentage uncertainties are attributed based on the results of Appendix E.

Table 5.6 shows the features which remain after 7 iterations (14 remaining features) for the networks trained on all mass points and only low mass points compared to the original choice of features. The main commonality between the feature sets is that they do not include the ϕ of the leptons. As well, there seems to be a complementary choice of features for the jets in either case. With all mass points, the derived jet variables are preferred over the raw jet variables, and vice versa for the low mass case. Neither optimized feature set is identical to the original 17 features. This is likely due to the fact that the margins between the feature importance values are quite small. Thus, the choice of rejected feature is inherently somewhat random.

Feature	**All**	**Low Mass**	**Original**
$p_{\mathrm{T}}(j_1)$	•	•	•
$p_{\mathrm{T}}(j_2)$		•	•
$\eta(j_1)$		•	•
$\eta(j_2)$	•	•	•
$\phi(j_1)$		•	
$\phi(j_2)$	•	•	
$\lvert\Delta y_{jj}\rvert$	•		•
$\Delta\phi_{jj}$	•	•	•
m_{jj}	•		•
$p_{\mathrm{T}}(l_1)$	•	•	•
$p_{\mathrm{T}}(l_2)$	•	•	•
$\eta(l_1)$	•	•	•
$\eta(l_2)$		•	•
$\phi(l_1)$			
$\phi(l_2)$			
$\lvert\Delta y_{ll}\rvert$	•		•
$\Delta\phi_{ll}$		•	•
ΔR_{ll}	•		•
E_T^{miss}	•	•	•
H_{T}	•	•	•
ξ	•		•

Table 5.6: Features remaining after 7 iterations of feature removal using all mass points and only low mass points compared to the original choice of features.

In order to assess the impact of the feature optimization, a neural network is trained on each of the three feature sets in Table 5.6. The AUC for each feature set is plotted as a function of $H_5^{\pm\pm}$ in Figure 5.12. The feature set optimized for low mass points generally performs slightly worse than the original feature set. Meanwhile, the feature set optimized for all mass points performs slightly worse at low mass and slightly better at high mass, accentuating the classification disparity between low mass and high mass. That being said, the difference between each of these three feature sets is not much larger than the scale of the AUC uncertainties. Therefore, the differences may be a greater reflection of the randomness of the NN optimization rather than genuine differences between the quality of the feature sets.

This study shows that a smaller feature set could be used without significantly compromising the network performance. However, since both optimized feature sets do perform slightly worse below 400 GeV than the original feature set, it is best to continue using the original feature set.

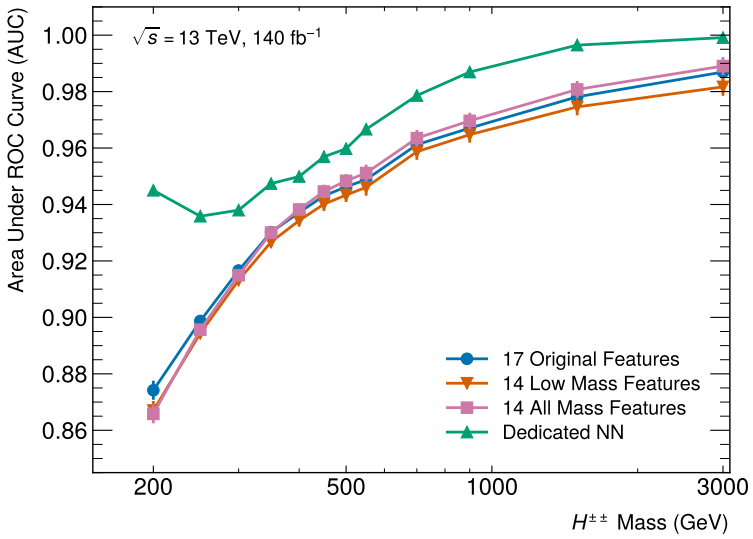


Figure 5.12: A comparison of the performance of NNs trained on the three different feature sets in Table 5.6 plotted as a function of $H_5^{\pm\pm}$ mass. The AUC uncertainties are estimated in Appendix E, with the exception of the dedicated NN.

5.3 Event Weighting

In the previous section, it was found that the choice of input features does not have a significant impact on the NN performance at low doubly-charged Higgs masses. Instead of modifying the input features, what if we modified the weighting of the signal events in the loss function? The current class weights method is very simplistic, since it treats each signal event equally. As a result, the performance of the NN across the mass range is dependent on both the set of mass points used in training as well as the number of events in each signal sample. Since the set of generated signal samples is fixed, we focus our attention on the number of events per sample. The signal samples that have a greater number of events will contribute more to the loss function of the NN and thus the NN will pay more "attention" to classifying these samples. In the $H_5^{\pm\pm}$ samples, there are arbitrarily more events in the higher mass samples than the lower mass samples — there are 7,643 events in the training set for the 200 GeV sample but 16,195 events in the training set for the 3000 GeV sample. This could enhance the performance of the NN at high mass and suppresses the performance of the NN at low mass, as was observed in Figure 5.9. In order to counteract this effect, two weighting methods are proposed: democratic class weights, and physical event weights.

5.3.1 Democratic Class Weights

The democratic class weights method is similar to regular class weights, but assigns a different factor for each of the signal samples, $f_{s,i}$. The signal sample factors each have an additional term which

forces the sum of weights to be equal for each signal sample:

$$f_{s,i} = \frac{N_s}{nN_{s,i}} \frac{N_s + N_b}{2N_s}, \quad \text{and} \quad f_b = \frac{N_s + N_b}{2N_b}$$

where n is the number of different signal mass points that are being used. The factors $f_{s,i}$ are largest for the low mass samples with less events, and smallest for the high mass samples with more events, as expected.

Figure 5.14 shows that using the democratic class weights slightly improves the performance at low mass. This is expected due to the fact that the weighting in the loss function for low mass samples is larger than when using regular class weights. Interestingly, it also slightly improves the performance at high mass, giving an overall improved performance.

5.3.2 Physical Event Weights

Another approach is to use the physical event weight from Equation 4.3 to weight each event in the loss function. In this way, the importance of the event in training is proportional to the physical probability of observing a similar event in the detector. Due to the sharp drop-off in doubly-charged Higgs cross-section with increasing mass, shown in Figure 5.15, this method should prioritize NN performance at lower mass points.

In order to use the physical event weights in training, the sum of weights for signal and background must be equalized to ensure that signal classification and background classification are equally important during training. This can be accomplished by multiplying each physical event weight for signal by the factor f_s and each physical event weight for background by the factor f_b:

$$f_s = \frac{\Sigma_s w_i + \Sigma_b w_i}{2\Sigma_s w_i}, \quad \text{and} \quad f_b = \frac{\Sigma_s w_i + \Sigma_b w_i}{2\Sigma_b w_i}. \tag{5.9}$$

The resulting weights form a distribution that has a long tail for positive values and a small number of events with a negative event weight. There are two problems with this. First, the large outlier event weights could cause problems by heavily influencing training. Second, negative event weights could be confusing for the NN during training since the network is rewarded for poorly classifying these samples. That being said, the batch size for training is 1000, which could be sufficient for the negative weights to reduce the importance of other nearby events — similar to their intended function as an MC weight. Nevertheless, it is not clear that using negative event weights will consistently give meaningful results.

Thus, to ensure that the training weights are not exceedingly large or negative, two methods are explored. The first method, min-max scaling, rescales the physical event weights such that the minimum physical event weight is 0 and the maximum physical event weight is 1. This requires all physical event weights to be non-negative and sets an upper limit on the value of the physical event weight. However, the outliers are still very large relative to the rest of the events. The second method, percentile min-max scaling, addresses this problem by rescaling the physical event weights such that the 1st percentile physical event weight is 0 and the 99th percentile physical event weight is 1. All weights below the 1st percentile are set to 0 and all weights above the 99th percentile are set

to 1. The distributions for the two scaling methods are compared in Figure 5.13. The event weight distributions with standard min-max scaling are more compressed than with percentile min-max scaling due to outliers. For the 200 GeV $H_5^{\pm\pm}$ sample, the effect of outliers when using standard min-max scaling is especially pronounced. There is a single event whose physical event weight is much smaller than any other events, producing a gap in the scaled event weight distribution between 0.1 and 0.45. In both cases, after scaling, the sum of weights of signal and background are equalized using Equation 5.9.

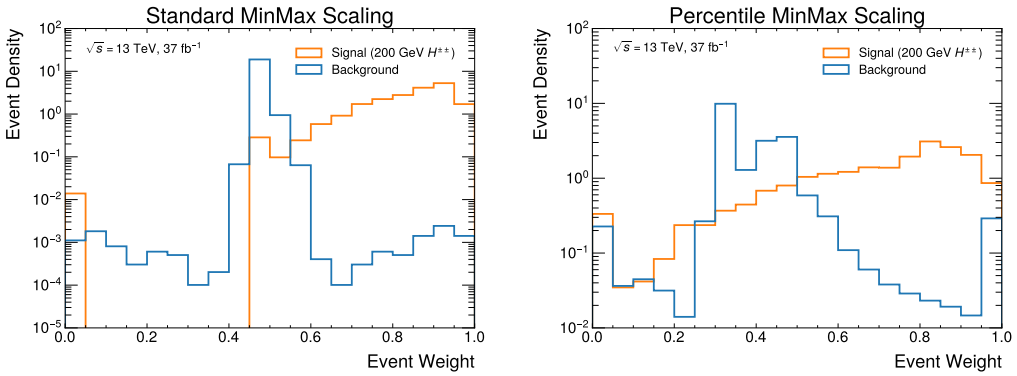


Figure 5.13: The event weight distribution for all background samples and a 200 GeV signal sample rescaled using standard min-max scaling (left) and percentile min-max scaling (right). These plots were produced using only MC20a.

Three NNs are trained with physical event weights scaled three different ways: no min-max scaling, min-max scaling, and percentile min-max scaling. The results in Table 5.7 show that the NN trained using event weights scaled with percentile min-max scaling slightly outperformed the other methods. Although the discrepancy may not be significant, the percentile min-max method is the most well-motivated and robust, and will be the default method for scaling physical event weights for the remainder of the thesis.

Method	**AUC**
No MinMax Event Weights	0.954
MinMax Event Weights	0.953
Min Max Percentile Event Weights	0.955

Table 5.7: A comparison of the NN performance for three different event weight scaling methods. The baseline significance with no NN cut is 1.18.

The performance of the NN trained with physical event weights is shown in Figure 5.14 (labelled "Physical Event Weights"). This method significantly improves the performance at low mass, due to the higher cross-section at low mass. Thus, the neural network prioritizes the correct classification of low mass events during training.

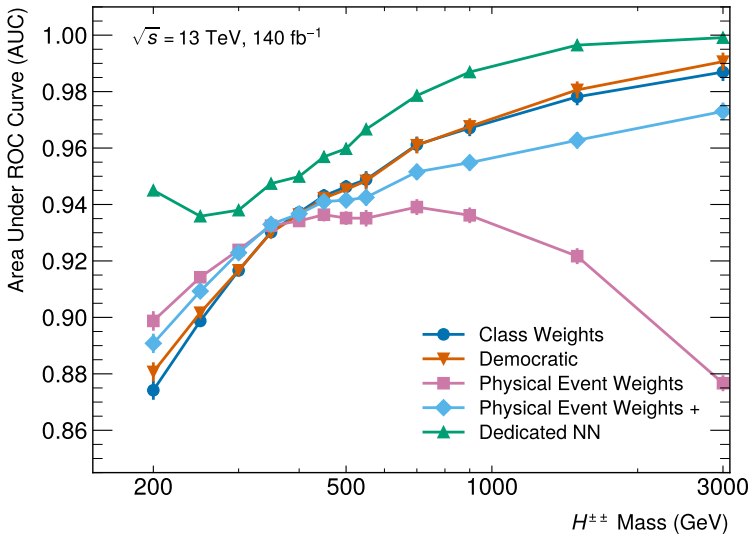


Figure 5.14: A comparison of the AUC for different sample weighting methods plotted as a function of $H_5^{\pm\pm}$ mass. The democratic weights method (Section 5.3.1) slightly improves the performance at low mass and high mass compared to the original class weights method (Section 5.1.5). The physical event weighting (Section 5.3.2) significantly improves the performance at low mass but also significantly reduces the performance at high mass. The performance at high mass was recovered by adjusting the cross-section used in training, and is therefore labelled "Physical Event Weights +" (Section 5.3.3). The AUC uncertainties are estimated in Appendix E, with the exception of the dedicated NN.

5.3.3 Physical Event Weights Power Law Modification

As evidenced by Figure 5.14, the physical event weights method also significantly reduces the performance at high mass points. This is due to the fact that the cross-section declines so quickly at high mass that the cross-section at 3000 GeV is a factor of 10^3 smaller than the cross-section at 200 GeV (see Figure 5.15). Thus, the scaled physical event weights for the high mass samples are much smaller than the low mass samples, and so the high mass events have very little influence on the training of the model parameters. As a result, the neural network more frequently misclassifies high mass samples compared to when the low mass samples and high mass samples are more evenly weighted (either with class weights or democratic weights). Evidently, suppressing the weighting at high mass is effective in boosting the performance of the network at low mass. However, it may be desirable to recover some performance at high mass if it does not invoke too much of a cost for the low mass range.

In order to slightly boost the performance at high mass, the cross-sections are modified before applying them to the physical event weights used in training. This is done by forcing the cross-section to obey a power law relation of the form $\sigma = Am_{H_5}^p$, where A and p are fitted variables. This preserves the steep drop off at high mass and is linear in the plot of $\log\sigma$ vs. $\log m_{H_5}$,

$$\log\sigma = \log A + P\log m_{H_5} = P\log m_{H_5} + k, \tag{5.10}$$

which allows for the values of k and P to be easily determined given two mass points and their corresponding cross-sections. In this case, the two lowest mass points (200 GeV and 250 GeV) are used, yielding a value of 2.88 for k and a value of -1.32 for P. This new cross-section relation is then applied to signal samples by multiplying the physical event weights by the following factor:

$$F = \frac{e^k m_{H_5}^P}{\sigma_{\text{old}}}. \tag{5.11}$$

As shown in Figure 5.15, the resulting modified cross-sections decline linearly on a log-log scale, as expected. The cross-section weight at 3000 GeV is now only a factor of 50 smaller than the cross-section weight at 200 GeV.

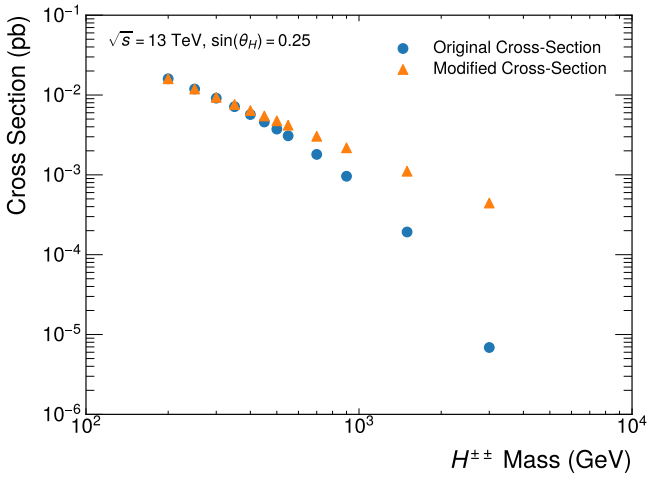


Figure 5.15: The original cross-sections used for sample weighting during NN training in Section 5.3.2 and the modified cross-sections used in Section 5.3.3. Only odd numbered DSIDs are shown, since these are the samples used in training, and also the samples whose cross-sections were directly provided by theory.

Training a neural network with these modified physical event weights, denoted "Physical Event Weights +" in Figure 5.14, significantly improves the performance at high mass, with only a small penalty at low mass. Evidently, determining which method is best depends on the objectives of the analysis. However, if the goal is to ensure consistent performance across the mass region, relative to the dedicated NN score, then this modified physical event weights method is the best option.

5.4 Parameterized Neural Network

The parameterized neural network (PNN) is a new NN architecture that has recently gained popularity in high energy physics [43]. PNNs are used when performing a classification or regression task where the samples are related by a physics parameter [43]. The idea is to include this physics parameter as an input feature during training, so that the network can benefit from correlations between the input features and the physics parameter, and in turn develop enhanced generalizabil-

ity across the range of the physics parameter. In this case, a PNN is implemented by adding the doubly-charged Higgs mass as an input parameter for signal events during training. For background events, the mass is randomly chosen from the discrete distribution of $H_5^{\pm\pm}$ mass points used in training. During evaluation, one cannot simply remove this input feature or know the signal mass a priori, so the mass input for all test set events is also randomly chosen from the training mass points.

There are two motivations for trying the PNN method. First, the PNN has been shown to improve interpolation ability between simulated samples as a function of mass [43]. This is an important requirement for this analysis since it is highly unlikely that the doubly-charged Higgs mass, if it exists, will align with one of the simulated doubly-charged Higgs mass points. Second, the performance of the general NN is much worse than the dedicated NN at low mass. Perhaps the parameterized NN will benefit from having information about the simulated mass points since it could identify relationships between input features and $H_5^{\pm\pm}$ mass.

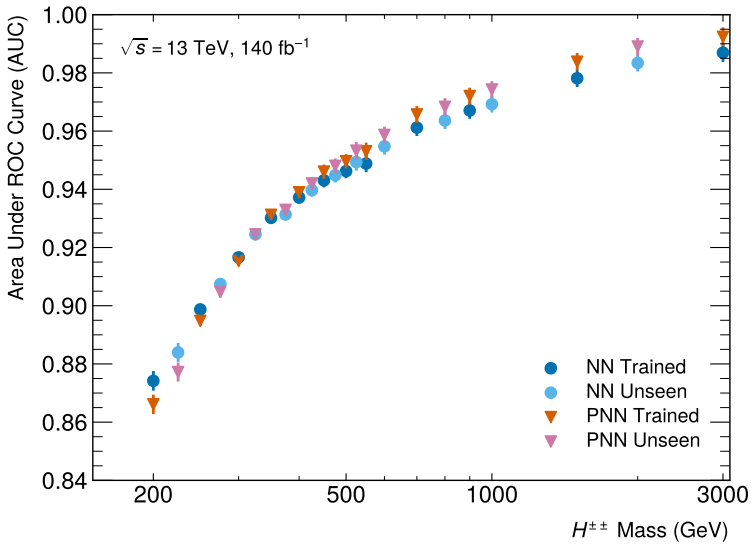


Figure 5.16: The AUC for mass points that were seen during training and unseen during training for a PNN and a regular NN plotted as a function of $H_5^{\pm\pm}$ mass. The AUC uncertainties are estimated in Appendix E.

The PNN and general NN are both trained on half of the signal mass points, and then evaluated on all of the signal mass points. Figure 5.16 shows that the PNN is not superior at interpolating between mass points nor does it improve the performance at low mass compared to a regular ensembling method. Both the PNN and the regular NN demonstrate interpolation ability since there is not a significant drop in performance between mass points which were used in training and mass points which were not used in training.

There are several factors that could explain this result. First, the set of simulated signal samples is quite large, which allows the regular ensembling method to also become generalizable without any special techniques. Second, the PNN has many inherent limitations. Even though the simulated

mass points are known for signal during training, the masses for background are randomly assigned, which may confuse the network. As well, when the PNN performance is evaluated using the test set, all of the events, including signal, are assigned a random mass point. The input node for the mass point cannot simply be removed for evaluation, but the mass of a potential signal event in data will evidently not be known in advance. It would be interesting to know whether the discrepancy between the simulated mass point and the randomly assigned mass point of a signal event has an impact on classification. The NN may be more confused by seeing a low mass $H_5^{\pm\pm}$ event with a high mass NN input than if the mass point was omitted completely.

5.5 Separate Signal Regions

The clear separation of $H_5^{\pm\pm}$ event characteristics below and above ∼350 GeV raises the question: should the signal region be split in two, one for low mass and one for high mass? To answer this question, three different signal region "split points" are tested: 300 GeV, 350 GeV, and 400 GeV. In each case, one network is trained on mass points up to and including the split point, and another network is trained on mass points including and above the split point. The signal and background events are balanced using the democratic weighting method described in Section 5.3.1, so that each mass point used in training has equal weight. Figure 5.17 shows the resulting AUC across the mass range. The 350 GeV split appears to be the best choice out of the three options, since the performance at the split point is the same for both the low mass and high mass network. Overall, there is a significant improvement in classification at both high mass and low mass. However, there is always a loss in performance near the split point compared to the current best network, Physical Event Weights +. This is due to the fact that the network tends to perform better in the middle of the mass range which it was trained for than at the edges. This is not ideal since the excess region of 375 GeV is important in the analysis.

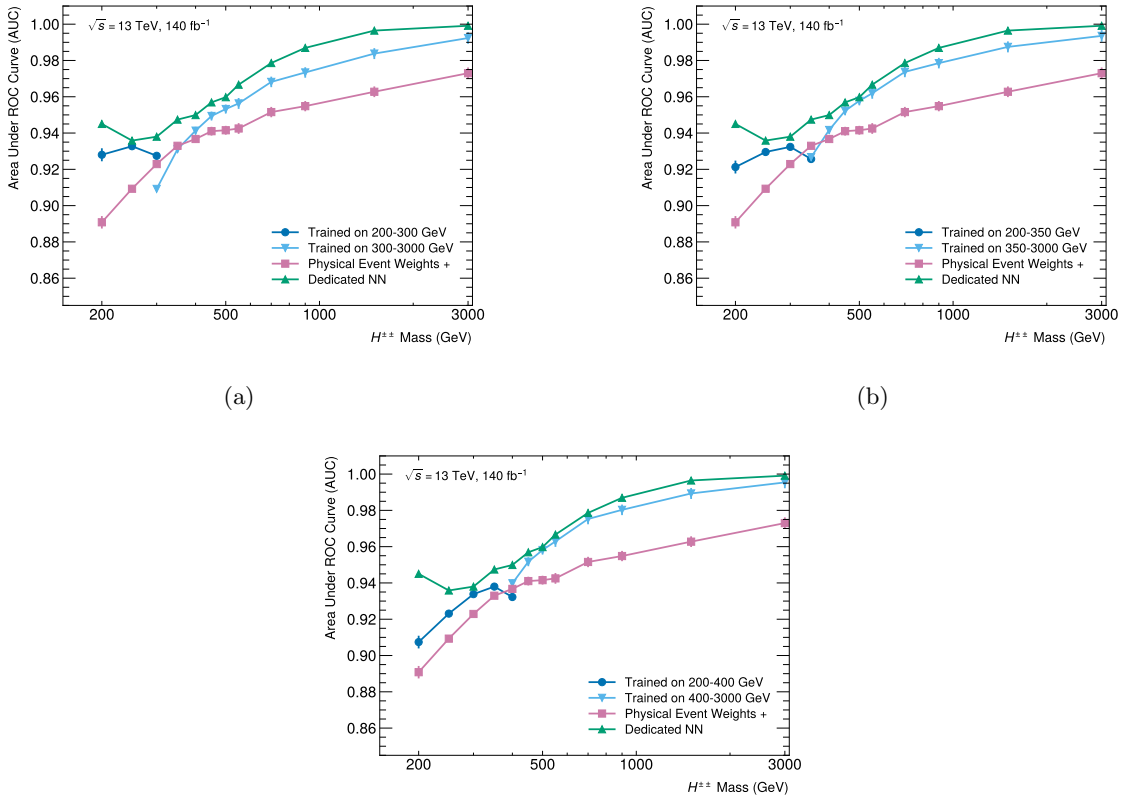


Figure 5.17: A comparison of the AUC for a SR split point at (a) 300 GeV, (b) 350 GeV, and (c) 400 GeV plotted as a function of $H_5^{\pm\pm}$ mass. Two NNs are trained, one for the region below the split point, and one for the region above the split point. The AUC uncertainties are estimated in Appendix E, with the exception of the dedicated NN.

Chapter 6

Results

Based on the results from Chapter 5, the neural network trained with the modified physical event weights ("Physical Event Weights +") is the optimal neural network. This method boosts performance at low mass, without heavily sacrificing performance at high mass. There are several other key aspects of the optimized NN performance that will be studied in Section 6.1. This includes verifying that the network is not overtrained, assessing the interpolation ability of the network, and determining whether there are any undesirable correlations between the fitting variable, m_{T}, and the NN score.

After completing these checks, the next task is to determine whether a signal region defined based on the NN score will improve the analysis compared to the cuts-based approach. The main metric that will be used to quantify this improvement is the expected upper limit on the $\sin(\theta_H)$ parameter of the GM model. These results will be presented in Section 6.2.

It is important to note that the training region for this chapter is slightly different from Chapter 5. Table 6.1 shows the updated NN training region selection, with all changes indicated in bold font. The m_{jj} cut was added since there is a generator-level cut on m_{jj} at 100 GeV for signal samples. Therefore, there are very few signal events for which $m_{jj} < 200$ GeV. The other changes were made to harmonize with the cuts-based signal region in Table 4.4 which was being developed concurrently with Chapter 5.

The changes to the NN training region will impact the NN metrics, since the number and composition of signal and background events in the training region will be different. However, these changes should not significantly impact the optimal choice of network architecture, hyperparameters or training method. The sum of weights and number of events in the new training region (Table 6.2), is similar to that of the previous training region (Table 5.2). There are less background events in the new training region, likely due to the addition of the m_{jj} requirement. There are also more signal events, due to the looser jet p_{T} and $E_{\mathrm{T}}^{\mathrm{miss}}$ requirements. Thus, the new NN training region has a higher signal purity overall.

New NN Training Region Selection
Exactly two same-sign signal leptons with $p_{\mathrm{T}} > 27$ GeV ($\lvert\eta\rvert < 1.37$ in the ee channel)
$m_{ll} \geq 20$ GeV
$\lvert m_{ee}$ - $m_Z\rvert > 15$ GeV in the ee-channel
≥ 2 signal jets with leading and subleading jets satisfying $p_{\mathrm{T}} >$ **50** GeV and $p_{\mathrm{T}} >$ **30** GeV respectively
Less than 3 baseline leptons
$E_{\mathrm{T}}^{\mathrm{miss}} \geq$ **25** GeV
$m_{jj} >$ 200 GeV
b-jet veto

Table 6.1: The event selection for the NN training region in Chapter 6.

Sample	$\sum w_i$ **in NN TR**	**# of Events in NN TR**
$H_5^{\pm\pm}$ 200	155.7	14596
$H_5^{\pm\pm}$ 250	143.2	18041
$H_5^{\pm\pm}$ 300	127.8	20932
$H_5^{\pm\pm}$ 350	114.3	23928
$H_5^{\pm\pm}$ 400	97.9	25795
$H_5^{\pm\pm}$ 450	82.6	26861
$H_5^{\pm\pm}$ 500	71.3	28559
$H_5^{\pm\pm}$ 550	61.0	26874
$H_5^{\pm\pm}$ 700	39.1	32425
$H_5^{\pm\pm}$ 900	22.0	34377
$H_5^{\pm\pm}$ 1500	4.63	36007
$H_5^{\pm\pm}$ 3000	0.152	33002
$W^{\pm}W^{\pm}jj$ EW	414.8	131067
$W^{\pm}W^{\pm}jj$ QCD	162.9	310165
$W^{\pm}W^{\pm}jj$ INT	41.0	5909
$W^{\pm}Zjj$ EW	47.0	10061
$W^{\pm}Zjj$ QCD	693.6	183890
$W^{\pm}Zjj$ INT	18.4	427

Table 6.2: The signal ($H_5^{\pm\pm}$) and background samples used for NN training in Chapter 6, which correspond to the Run 2 integrated luminosity of 140 fb^{-1}. The signal sample cross sections correspond to $\sin(\theta_H) = 0.25$. The middle column indicates the sum of weights in the updated NN training region and the right column indicates the number of events in the updated NN training region.

6.1 Optimized Neural Network Results

6.1.1 Optimized NN Training

In order to determine whether the optimized NN is overtrained, we can look at the loss curves and the NN score distributions for both the training set and the validation set, shown in Figure 6.1. The loss curves show that the network converges quickly due to the large number of events and the use of batch normalization. Early stopping halted training at seven epochs, and the weights were restored to the optimal values at epoch four. The validation set loss is very close to the training set loss at this point, which shows that there is likely no overtraining. This is confirmed by the NN score distribution, which shows that the training set and validation set classification distributions are nearly indistinguishable for both signal and background. The network is able to perform equally well on unseen events, due to the large quantity of events in the training set and the use of early stopping. Thus, we can expect that the network will also perform well for other simulated events that are not included in the training set or for real data.

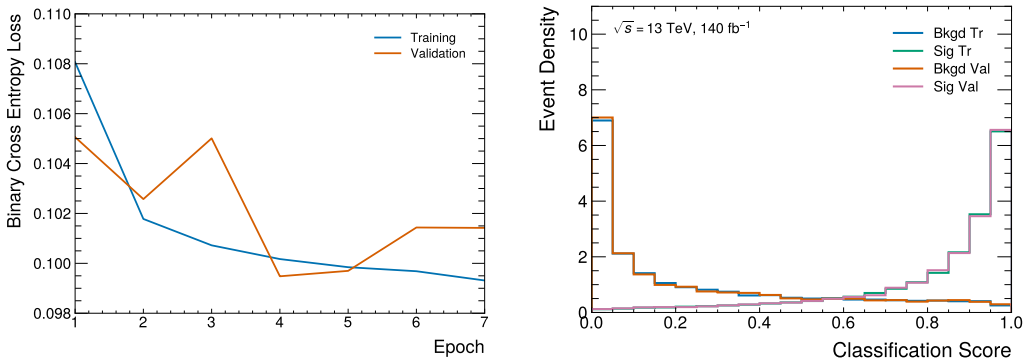


Figure 6.1: The loss curve (left) and NN score distribution (right) for the optimized NN. Both plots demonstrate that the network is not overtrained — the loss curve for the training set and the validation set is similar and the NN output distribution is nearly identical.

6.1.2 Interpolation

As discussed in Section 5.4, it is important for the NN to be able to interpolate between mass points used in training. Figure 6.2 shows that the classification performance is smooth as a function of mass, which indicates that it has good interpolation ability. There are no noticeable dips in performance for mass points that were not seen in training. This validates that a signal region constructed using the NN score should be sensitive to all $H^{\pm\pm}$ masses in the range of 200 GeV to 3000 GeV, not just at the mass points used in training.

6.1.3 Correlation with m_{T}

The distribution of m_{T} in the signal region will be used to make a fit and to estimate the mass of any $H^{\pm\pm}$ resonance. Therefore, it is useful to understand whether the NN output score is correlated

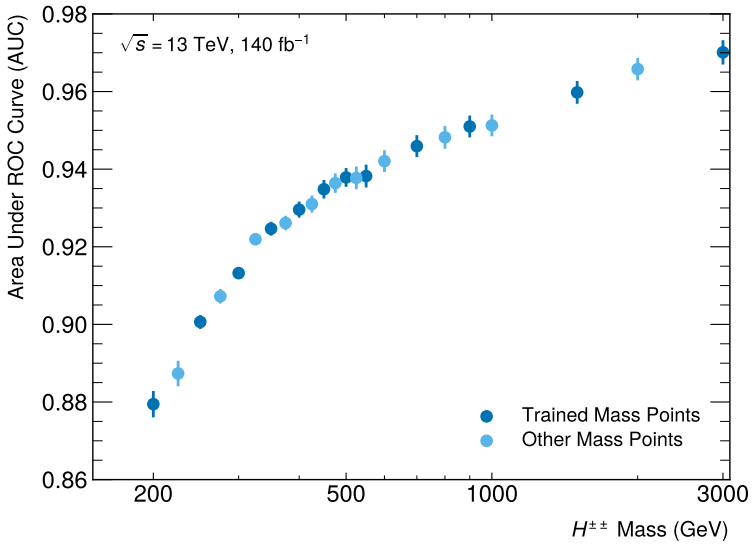


Figure 6.2: A comparison of the AUC for mass points seen in training and mass points not seen in training for the optimized NN. The AUC uncertainties are determined in Appendix E.

with m_{T}, especially since many of the input variables are highly correlated with m_{T} (see Figure 6.3).

Fortunately, the Pearson correlation coefficient for m_{T} and the NN score is 0.12 for signal, -0.06 for background, and 0.21 for signal and background combined. The correlation between m_{T} and the NN score is positive for signal due to the fact that the high mass signal samples, which tend to have a larger m_{T} value, are easier for the network to classify. Therefore, the high mass events also tend to have larger NN output values, as shown in Figure 6.4. The correlation is much larger for signal and background combined than individually since the m_{T} distribution for background peaks around 200 GeV, which is much lower than the peak of the m_{T} distribution for the majority of the $H_5^{\pm\pm}$ simulated mass points. Therefore, there should be a positive correlation between m_{T} and the NN score if the network is performing good classification. Overall, the correlation between m_{T} and the NN score is small relative to other correlations with m_{T} and the NN score, as shown in the correlation matrix in Figure 6.3. Therefore, the NN score can be used with reasonable confidence that it is not creating a bias in m_{T}.

6.2 Comparison to Cuts-Based Approach

We have found that the neural network performs excellent classification of signal and background events, and does not have any undesirable behaviour. However, we must also show that a signal region defined with the NN can improve our sensitivity to $H^{\pm\pm}$ compared to the cuts-based signal region described in Table 4.4.

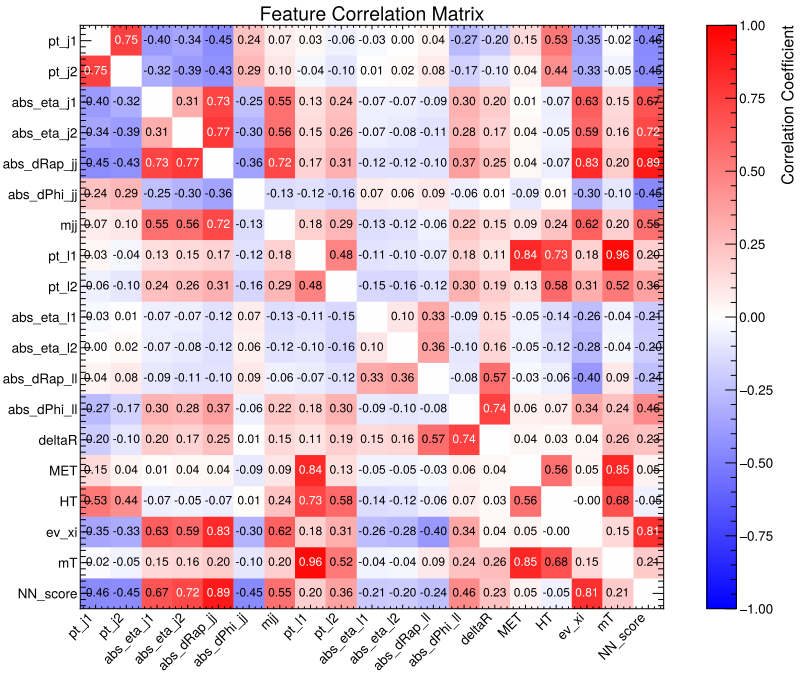


Figure 6.3: The correlation matrix for the 17 input features, m_{T}, and the NN score in the NN training region. Note that the correlations are slightly different from the correlation matrix in Figure 5.6 since the training region definitions are different.

6.2.1 Neural Network Signal Region Definition

The neural network signal region (NN SR) is defined by choosing a NN working point (WP). The NN output distribution for background and selected $H_5^{\pm\pm}$ signal masses is shown in Figure 6.4. All events which score higher than the WP will be included in the NN SR, and all events which score lower than the WP will be rejected. The choice of WP has a significant impact on the quality of the NN SR, and should be chosen intentionally.

The performance of different WPs is compared in Figure 6.5. The background rejection rate increases at larger WPs, while the signal acceptance rate decreases at larger WPs, as expected. The signal significance, which is calculated using the validation set and test set, shows that there is no single working point that provides the best signal significance for all $H_5^{\pm\pm}$ mass points. The lower WPs provide better results for the lower mass points while the higher WPs provide better results for higher mass points. This is due to the the fact that the NN output distribution peaks more sharply at one for high mass points than for low mass points, as evidenced by Figure 6.4. Therefore, the signal acceptance rate decreases with mass, especially for the higher WPs. Below 450 GeV, the benefit of the increased background rejection rate from the 0.9 WP is not worth the cost of a lower signal acceptance rate.

The WP at 0.85 provides the best significance for the $H_5^{\pm\pm}$ mass points below 350 GeV. This WP also provides comparable performance to the 0.9 WP for the 350 GeV and 400 GeV masses, which is a region of particular interest due to the resonance observed in Run 2. Since this analysis is limited by statistical uncertainty rather than systematic uncertainty, it is better to choose a slightly

looser WP if possible. Therefore, the NN SR is defined to include all events which have a NN score greater than 0.85, in addition to all selections for the NN training region found in Table 6.1.

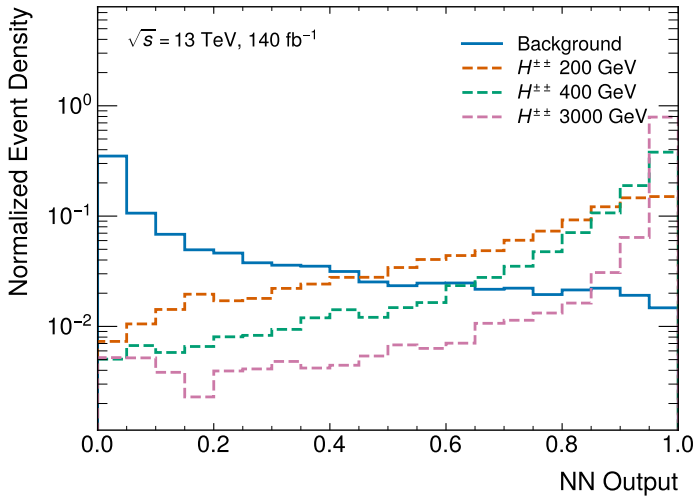


Figure 6.4: The optimized NN score distribution for background and the 200 GeV, 400 GeV, and 3000 GeV mass points.

6.2.2 Significance and Classification Metrics

The significance, accuracy, background rejection rate, and signal acceptance rate for the cuts-based signal region is compared to the selected 0.85 WP in Figure 6.5. We find that the signal significance for the 0.85 WP is nearly twice as large at every mass point than for the cuts-based signal region. Roughly four times as much data would be required to increase the significance by a comparable amount using the cuts-based approach. Although the signal acceptance rate is higher for the cuts-based selections than the 0.85 WP, the significance is larger due to the fact that background rejection rate is much higher for the 0.85 WP. Accordingly, the overall classification accuracy is also higher than the cuts-based selections. Encouragingly, the same could be said for all of the presented WPs, although they will not be used for the remainder of the thesis.

6.2.3 Asimov Fit

The expected signal significance provides a good estimate of the SR performance. However, significance does not account for the shape of the m_{T} distribution. A better way of comparing the cuts-based SR and the NN SR is to perform a binned likelihood fit with Asimov data, meaning the expected data (simulation) is used as the observed data. Even without systematic uncertainties, this allows us to assess the statistical power of the fit in terms of the constraints on background sources, the signal significance, and the expected upper limit on $\sin(\theta_H)$. In order to avoid any possible

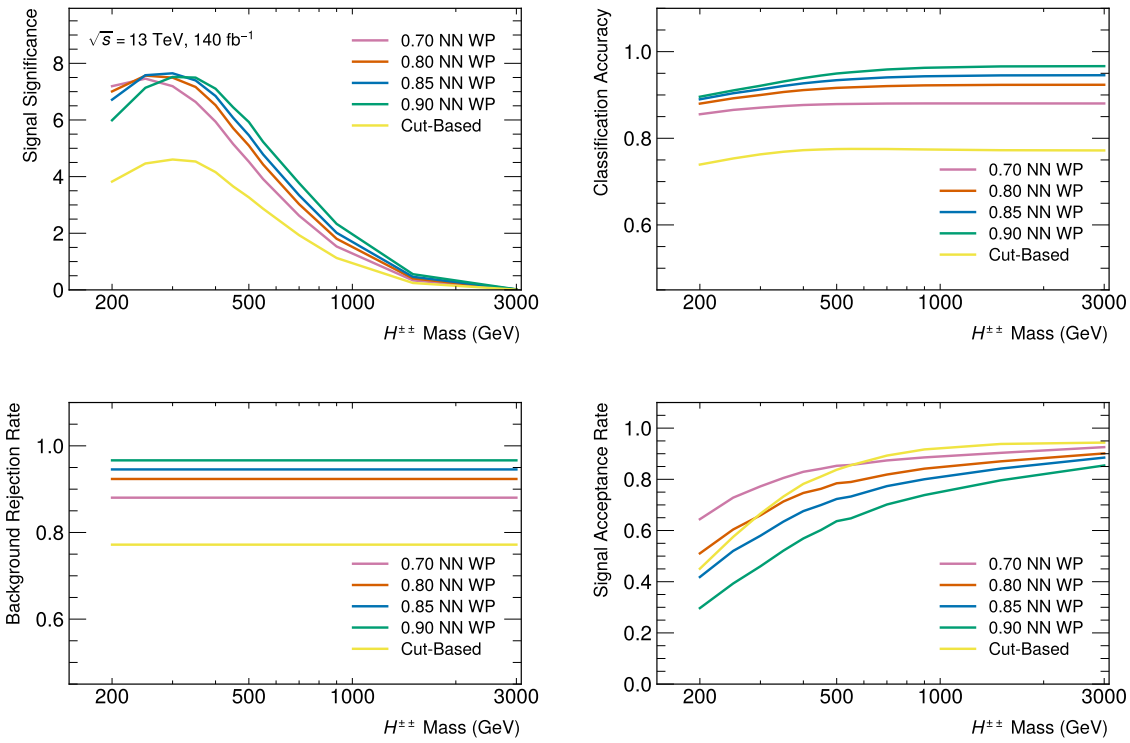


Figure 6.5: The expected signal region significance with $\sin(\theta_H) = 0.25$ (top left), classification accuracy (top right), background rejection rate (bottom left), and signal acceptance rate (bottom right) for the optimized NN plotted as a function of $H^{\pm\pm}$ mass for selected WPs. The signal significance is evaluated in the NN SR defined at the WP (or cuts-based SR) while the other metrics are evaluated in the NN training region defined in Table 6.1.

advantages from using signal events which were used to train the NN, only mass points which were not used in training are used in this section.

There are three parameters in the likelihood fit, one of which is the signal strength, $\mu(H_5^{\pm\pm})$. The signal strength can be converted to a value of $\sin(\theta_H)$ using the fact that the $H_5^{\pm\pm}$ cross-section is proportional to $\sin^2(\theta_H)$,

$$\sin(\theta_H) = \sqrt{\mu}\sin(\theta_H)_{\text{gen}}, \tag{6.1}$$

where $\sin(\theta_H)_{\text{gen}}$ is the value of $\sin(\theta_H)$ that was used to generate the signal sample. The mass points including and below 800 GeV were generated with $\sin(\theta_H) = 0.5$, but the observed upper limit on $\sin(\theta_H)$ for these mass points ranges from 0.10-0.27 [22]. Therefore, the signal strength is set to 0.25, which corresponds to $\sin(\theta_H) = 0.25$ based on Equation 6.1, to bring the cross-section of the samples closer to the current limit. The mass points above 800 GeV were generated with $\sin(\theta_H) = 0.25$ so they are not modified.

The other normalization parameters are for the two largest background sources, $W^{\pm}W^{\pm}$ and $W^{\pm}Z$. Each of these backgrounds also has a dedicated control region (CR). In this way, discrepancies between the $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ simulation and observed data in the control regions can be accounted for in the signal region by adjusting their normalization factors. The cuts-based SR and the neural

network SR share the same $W^{\pm}Z$ CR in this fit, but have different $W^{\pm}W^{\pm}$ CRs. This is done to ensure the $W^{\pm}W^{\pm}$ CRs are adjacent to and independent of each of the SRs. Since events are required to have three leptons for the $W^{\pm}Z$ control region, it is already guaranteed to be independent of both $W^{\pm}W^{\pm}$ SRs. The $W^{\pm}W^{\pm}$ CR for the cuts-based SR is defined by inverting the $\Delta\phi_{ll}$ cut, while the $W^{\pm}W^{\pm}$ CR for the neural network SR is defined by inverting the NN score cut. Additionally, a Δy_{jj} requirement is added to the $W^{\pm}W^{\pm}$ NN CR to enhance the concentration of EW $W^{\pm}W^{\pm}$ since this is the dominant $W^{\pm}W^{\pm}$ contribution in the NN SR. A full description of the regions used in the fit is provided in Table 6.3.

Selection	**CC SR**	**NN SR**	**WZ CR**	**Low $\Delta\phi_{ll}$ CR**	**Low NN CR**
Exactly two same-sign signal leptons with $p_{\mathrm{T}} > 27$ GeV ($\lvert\eta\rvert < 1.37$ in ee channel)	•	•		•	•
$m_{ll} \geq 20$ GeV	•	•		•	•
$\lvert m_{ee}$ - $m_Z\rvert > 15$ GeV in the ee-channel	•	•		•	•
≥ 2 signal jets with leading jet $p_{\mathrm{T}} > 50$ GeV and subleading jet $p_{\mathrm{T}} > 30$ GeV	•	•	•	•	•
Less than 3 baseline leptons	•	•		•	•
$E_{\mathrm{T}}^{\mathrm{miss}} \geq 25$ GeV	•	•	•	•	•
b-jet veto	•	•	•	•	•
$m_{jj} \geq 200$ GeV		•			•
$m_{jj} \geq 500$ GeV	•		•	•	
$\lvert\Delta y_{jj}\rvert > 2$	•		•	•	•
$\lvert\Delta\phi_{ll}\rvert$	>1.5			<1.5	
NN Output		>0.85			<0.85
WZ Event Selection*			•		

Table 6.3: The event selections for each of the signal regions and control regions used in the fit. *For brevity, the WZ event selection is located in Table 6.4.

Selection	**Details**
ZZ veto	Less than 4 baseline leptons
N leptons	Exactly three signal leptons
Z Candidate	Same flavour opposite-sign pair with invariant mass closest to Z boson
W Candidate	Third lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$
W, Z Selection	W candidate lepton passes Tight (LH) and Tight VarRad isolation, Z candidate leptons pass Medium (LH) and Tight VarRad isolation
Mass Window	Z candidate $\lvert m_{ll} - m_Z\rvert < 20$ GeV
Lepton Inv. Mass	$m_{lll} \geq 106$ GeV

Table 6.4: The additional requirements for the $W^{\pm}Z$ control region.

This Asimov fit uses a simplified version of the setup that was used in Run 2 [26]. The Run 2 fit was optimized for the Standard Model $W^{\pm}W^{\pm}jj$ cross-section measurement, rather than the $H_5^{\pm\pm}$ search. Therefore, I found that many aspects of the Run 2 fit yielded worse results than a simpler choice. For instance, performing a one-dimesional fit in m_{T} gave lower limits than a two-dimensional

fit in m_{T} and m_{jj}. Similarly, uniform binning gave lower limits than the variable binning used in Run 2. Eventually, this choice of binning and fit setup will need to be optimized for the $H_5^{\pm\pm}$ search, however it is not within the scope of this thesis. The complete details of the fitting configuration for each SR is provided in Table 6.5. There are 12 bins instead of 10 bins in m_{jj} for the NN $W^{\pm}W^{\pm}$ CR due to the fact that the range of m_{jj} extends to 200 GeV, and so increasing the number of bins maintains a more consistent bin size with the other CRs.

CC Regions	**Description**
CC SR	15 bins in m_{T}, 0 GeV $< m_{\mathrm{T}} <$ 1500 GeV
WZ CR	10 bins in m_{jj}, 500 GeV $< m_{jj} <$ 3000 GeV
Low $\Delta\phi_{ll}$ CR	10 bins in m_{jj}, 500 GeV $< m_{jj} <$ 3000 GeV
NN Regions	**Description**
NN SR	15 bins in m_{T}, 0 GeV $< m_{\mathrm{T}} <$ 1500 GeV
WZ CR	10 bins in m_{jj}, 500 GeV $< m_{jj} <$ 3000 GeV
Low NN Score CR	12 bins in m_{jj}, 200 GeV $< m_{jj} <$ 3000 GeV
Parameter of Interest	$\mu(H_5)$
Normalization Parameters	$\mu(W^{\pm}Z)$,$\mu(W^{\pm}W^{\pm})$
Observables	One dimensional fit in m_{T}

Table 6.5: The Asimov fit setup in the cuts-based and neural network SRs.

The pre-fit plots for the 375 GeV $H_5^{\pm\pm}$ mass point in Figure 6.6 show the background composition and signal strength for each of the SRs and CRs in Table 6.3. The $W^{\pm}Z$ CR has a very high purity of $W^{\pm}Z$ events and similar relative quantities of EW, QCD and interference events to both SRs. The low NN score CR does have a lower percentage of $W^{\pm}W^{\pm}$ events compared to the low $\Delta\phi_{ll}$ CR, mainly due to events in the lowest m_{jj} bin. In the future, it may be useful to add an $m_{jj} > 500$ GeV cut for this region. The NN SR does show significant improvement in the background rejection compared to the cuts-based SR, as expected.

The post-fit plots for each of the 375 GeV mass point are presented in Figure 6.7. The pre-fit and post-fit plots for the 225 GeV and 1000 GeV mass points can be found in Appendix D. The histograms pre-fit and post-fit are identical due to the fact that this is an Asimov fit. Therefore, the normalization parameters are unchanged from their nominal values, shown in Figure 6.8. The uncertainty on the signal strength parameter is more well-constrained with the NN SR. The background normalization parameters for the NN fit also appear to be equally or more well-constrained than the cuts-based fit.

The Asimov fit allows us to determine an expected upper limit on the signal strength, $\mu(H_5)$, for a given $H_5^{\pm\pm}$ mass point. The expected upper limit corresponds to the largest value of the signal strength for which the p-value is greater than or equal to 0.05. This means that if the observed data matches the expected data (simulation), a signal strength above the limit can be excluded with greater than 95% confidence, while a signal strength below the limit cannot be excluded. Thus, the expected upper limit is a measure of the sensitivity of the analysis to $H_5^{\pm\pm}$ production.

The expected upper limit on $\mu(H_5)$ can be converted to an expected upper limit on $\sin(\theta_H)$ with Equation 6.1. In this way, we can compare the performance of the NN SR to the cuts-based SR and the previous Run 2 limits. Figure 6.9 shows that the neural network signal region does significantly

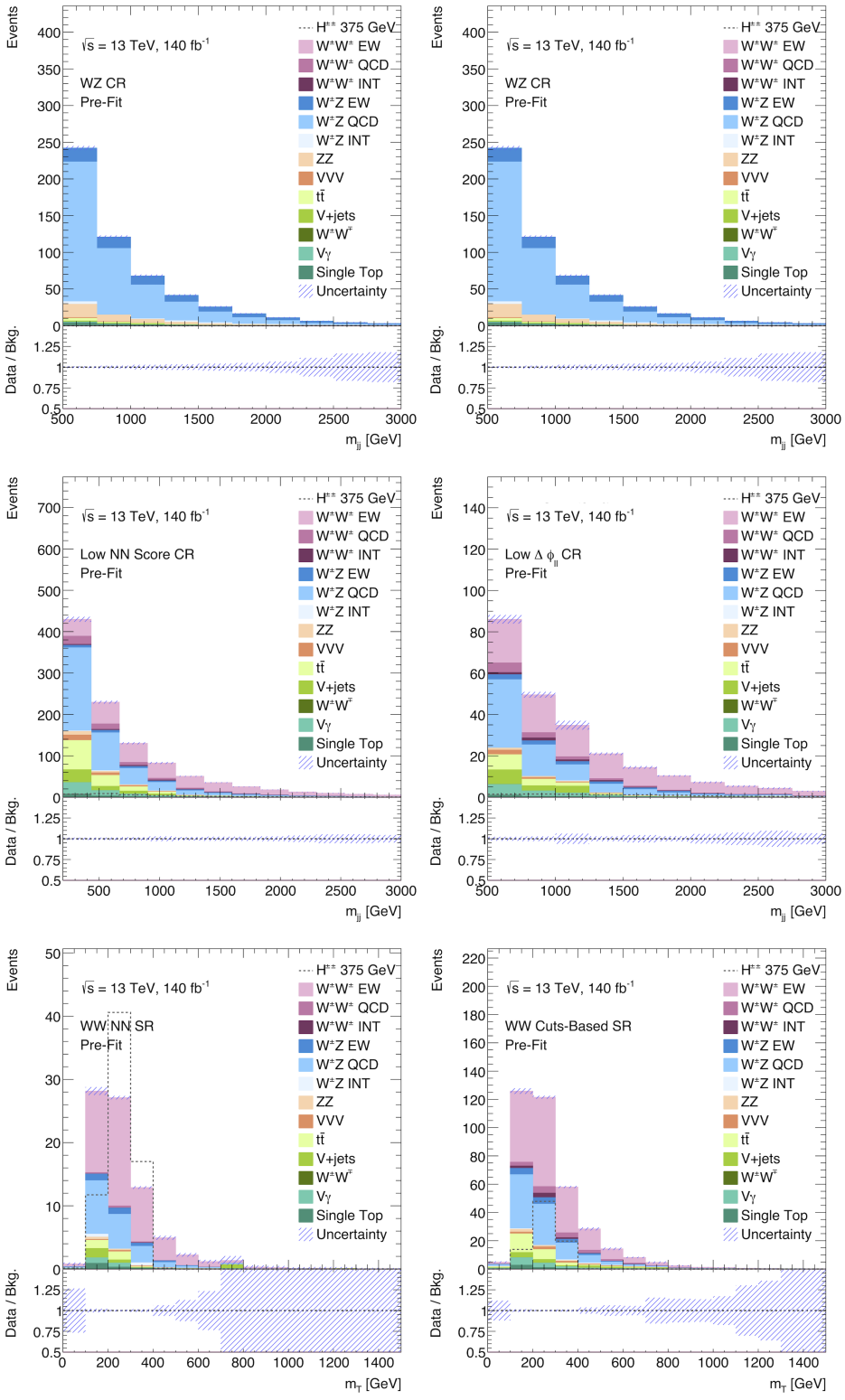


Figure 6.6: The pre-fit plots for the NN SR (left) and the cuts-based SR (right) with a 375 GeV signal sample and $\sin(\theta_H) = 0.25$.

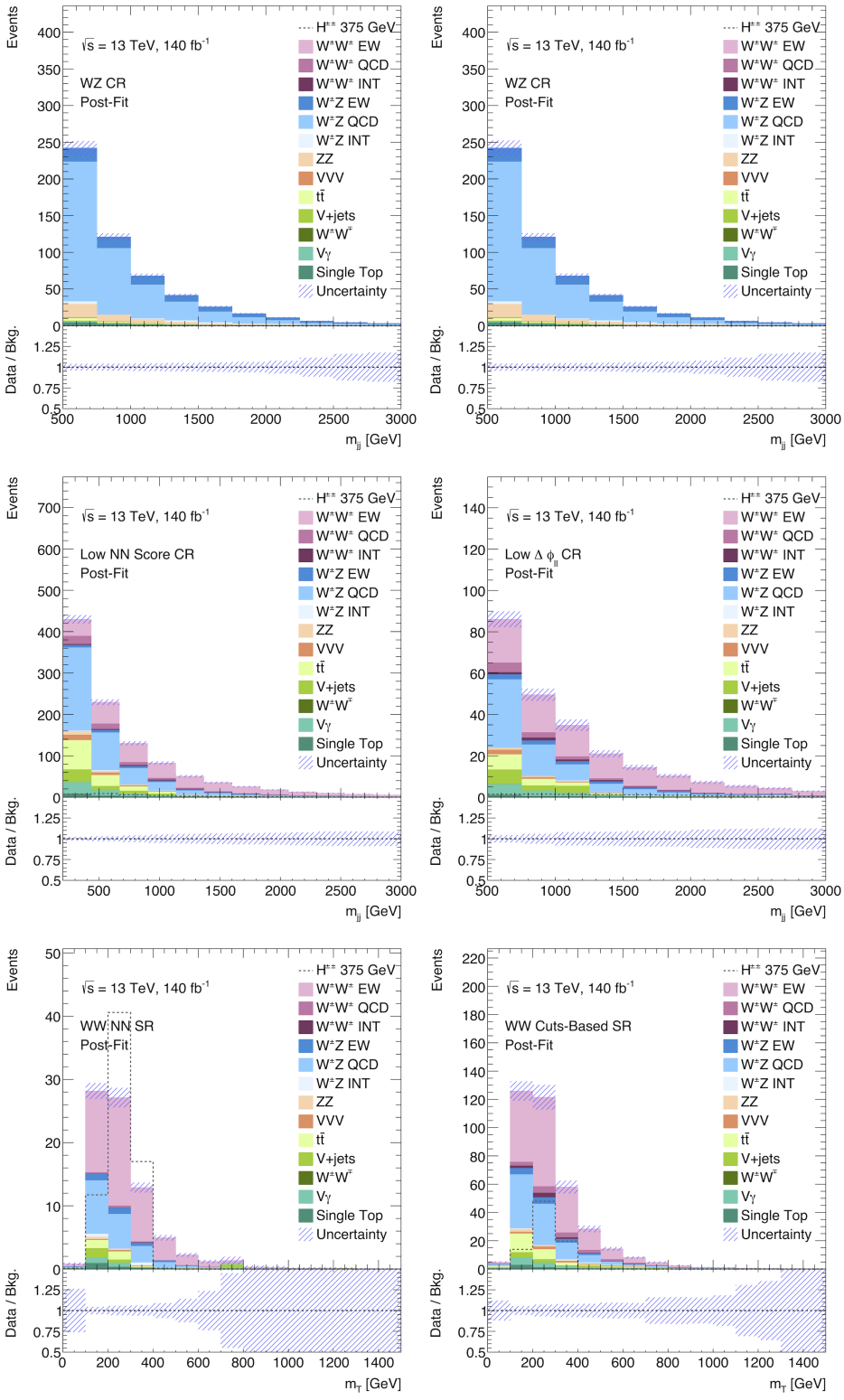


Figure 6.7: The post-fit plots for the NN SR (left) and the cuts-based SR (right) with a 375 GeV signal sample and $\sin(\theta_H) = 0.25$.

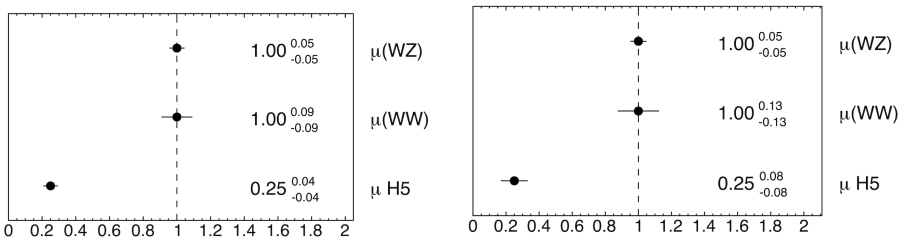


Figure 6.8: The post-fit normalization factors for the 375 GeV $H_5^{\pm\pm}$ sample with the NN SR (left) and cuts-based SR (right).

improve the expected upper limit on $\sin(\theta_H)$ at all mass points compared to the cuts-based signal region. Therefore, the NN SR is more sensitive to the GM model since it can exclude the model to a greater extent in the absence of a doubly-charged Higgs signal. The neural network signal region also improves the expected upper limit on $\sin(\theta_H)$ compared to the Run 2 analysis. However, the two limits should be compared with care as the Run 2 expected limit used real data in the control regions as well as systematic uncertainties. The latter are expected to have a small impact since this search is statistically limited.

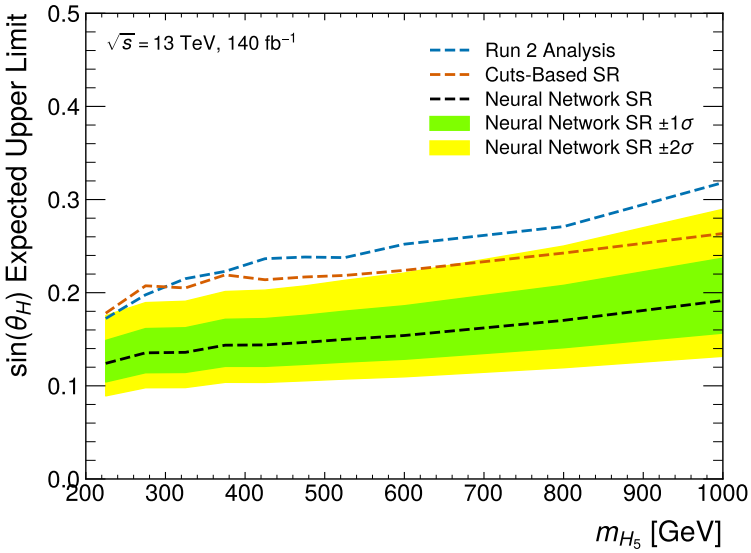


Figure 6.9: A comparison of the expected upper limit on $\sin(\theta_H)$ for the Run 2 $W^{\pm}W^{\pm}$ analysis, the cuts-based SR from the present analysis, and the NN SR developed in this thesis. The one sigma and two sigma error bands are only from statistical uncertainties.

Chapter 7

Conclusions

The search for the doubly-charged Higgs boson ($H_5^{\pm\pm}$) at ATLAS provides an exciting opportunity to use machine learning (ML) to improve our sensitivity to new physics. A signal region defined with the optimized neural network (NN) at a 0.85 working point rejected significantly more background compared to a cuts-based approach. This resulted in a greater sensitivity to the doubly-charged Higgs and the GM model at all mass points. The NN is now available to be used by the ATLAS analysis group to improve the search for the $H_5^{\pm\pm}$ and determine whether the excess observed in Run 2 leads to the discovery of a new particle.

One of the main challenges that was encountered during this study was the variation in signal characteristics across the $H_5^{\pm\pm}$ mass range, which resulted in poor classification for low mass points. In order to address this problem, several options were explored, including optimizing the feature set for low mass points, applying different weighting to different mass points during training, and using a parameterized neural network. The physical event weights method resulted in the greatest improvement, and was modified using a power law fit to counteract the decline in $H_5^{\pm\pm}$ cross-section at high mass. The event weighting methods that were developed in this thesis will hopefully provide direction and insight for researchers working on similar classification problems in particle physics.

In this work, the neural network was trained using "tabular" event data, meaning each kinematic variable is treated as an independent column in a table. When using tabular data for the $H_5^{\pm\pm}$ event selection, it was difficult to improve the classification performance beyond what can be achieved with a relatively simple NN with two hidden layers of 45 nodes. Recently, a new ML method has been proposed for high energy physics, where each particle physics event is represented by an unordered set of points which form a "point cloud" [44]. Each point in the cloud corresponds to a particle which has some associated features, such as its four vector or its charge. This preserves the essential notion of a particle physics event as a cluster of particles in space. The point cloud representation has been found to be effective when used with graph-based neural networks, and appears to benefit from deep learning in a way that is not possible with tabular data [44] [45] [46]. This could be an interesting strategy to explore for improving the $H_5^{\pm\pm}$ event selection in the future.

Appendix A

Close-by-correction

As the singly-charged Higgs ($H_5^{\pm}$) mass increases, so does the boost of the Z boson. Thus, the Z boson decay products, either a pair of electrons or a pair of muons, are likely to decay close together in the detector. The $W^{\pm}Z$ signal region selection requires that leptons must first pass a Loose Variable Radius (Loose VarRad) isolation requirement, and then the Z candidate leptons must pass a Tight Variable Radius (Tight VarRad) isolation requirement. When two leptons decay close together, they may deposit energy within each other's isolation cone, which is used to calculate the lepton isolation variables ($p_{\mathrm{T}}^{\mathrm{varcone}}$ and $E_{\mathrm{T}}^{\mathrm{cone}}$). This may cause the leptons to fail the isolation requirements and subsequently reject the signal event from the $W^{\pm}Z$ signal region.

Close-by-corrected isolation variables prevent the loss of leptons originating from a boosted Z by removing the p_{T} contribution of nearby same object tracks (for leptons that failed the isolation requirement the first time). The Loose Var Rad and Tight Var Rad isolation requirements are then constructed using the same working points as the nominal isolation variables.

To evaluate the impact of using this close-by-correction, a modified $W^{\pm}Z$ SR is defined which includes all $W^{\pm}Z$ SR requirements except the isolation requirements for the Z candidate leptons (see Table A.2). This includes both the Loose Var Rad isolation requirement for "loose" leptons and the Tight Var Rad isolation requirement for Z candidate leptons.

Two metrics are used to compare the performance of the close-by-corrected isolation requirements to the nominal isolation requirements. Lepton efficiency refers to the percentage of Z candidate leptons in the modified $W^{\pm}Z$ SR that pass the isolation requirement. Event efficiency refers to the percentage of events in the modified $W^{\pm}Z$ SR where both Z candidate leptons pass the isolation requirement.

For both the close-by-corrected and nominal case, the Tight VarRad isolation requirement is inclusive of the Loose VarRad isolation requirement. Therefore, applying the Tight VarRad isolation requirement to the Z candidate leptons is sufficient to evaluate the effect of the close-by-correction on the Z candidate acceptance.

The 3000 GeV $\mathrm{H}^{\pm}$ Monte Carlo sample is used since the largest mass point should show the most significant improvement from the close-by-correction. These results were produced using MC20a samples, but the MC20d and MC20e samples produced similar results.

For the $eZ\mu W$ and μZeW channels, the event efficiency was evaluated as a function of ΔR and

Electrons	**Muons**
$p_{\mathrm{T}} > 25$ GeV	$p_{\mathrm{T}} > 25$ GeV
$\lvert\eta\rvert < 2.47$	$\lvert\eta\rvert < 2.7$
$\lvert z_0 \sin\theta\rvert < 0.5$	$\lvert z_0 \sin\theta\rvert < 0.5$
$\lvert d_0/\sigma(d_0)\rvert < 5$	$\lvert d_0/\sigma(d_0)\rvert < 3$
Loose LH	Loose
Pass Overlap Removal	Pass Overlap Removal

Table A.1: The modified loose lepton selection for the close-by-correction study which does not include any isolation requirements.

Selection	**Details**
ZZ veto	Less than 4 soft leptons
Missing Energy	$E_{\mathrm{T}}^{\mathrm{miss}} > 25$ GeV
N leptons	Exactly three leptons
Z Candidate	Same flavour opposite-sign pair with invariant mass closest to Z boson
W Candidate	Third lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$
W, Z Selection	W candidate lepton passes Tight (LH) and Tight VarRad isolation, Z candidate leptons pass Medium (LH)
Mass Window	Z candidate $\lvert m_{ll} - m_Z\rvert < 20$ GeV

Table A.2: The modified $W^{\pm}Z$ signal region selection for the close-by-correction study which does not include any isolation requirements for the Z candidate leptons.

the lepton efficiency was evaluated as a function of ΔR, p_{T} and η. The ΔR histogram for lepton efficiency was filled twice if both Z candidate leptons passed the isolation requirement and once if only one of the Z candidate leptons passed the isolation requirement.

The event and lepton efficiency plots for ΔR in the $eZ\mu W$ channel (see top of Figure A.1) show that the nominal isolation efficiency sharply drops below a ΔR value of 0.2, which corresponds to the outermost radius of the isolation cone (for leptons with $p_{\mathrm{T}} > 50$ GeV). Meanwhile, the close-by-corrected isolation maintains high lepton and event efficiency, with only a slight reduction below a ΔR value of 0.12.

This efficiency improvement is reflected in the p_{T} and η lepton efficiency plots in the $eZ\mu W$ channel (see bottom of Figure A.1). When the Z boson is highly boosted, the leptons tend to decay close together (small ΔR) and are highly transverse (large p_{T} and small η). As a result, the lepton efficiency of the nominal isolation requirement dips down to 40% at high p_{T} and small η, while the efficiency of the close-by-corrected isolation requirement is close to 100%.

In the μZeW channel, the event efficiency and lepton efficiency for the nominal isolation requirement remains high even at small ΔR, large p_{T}, and small η (see A.2). Thus, less improvement is had by using the close-by-correction.

Figure A.3 shows that for Z decaying to electrons, there is a $\sim$40% improvement in lepton efficiency when using the close-by-corrected isolation requirements compared to the nominal isolation requirements. For the Z candidates decaying to muons, the difference is very small due to the nominal isolation variable having a high lepton efficiency.

The procedure was repeated with the addition of a close-by-corrected Loose VarRad requirement

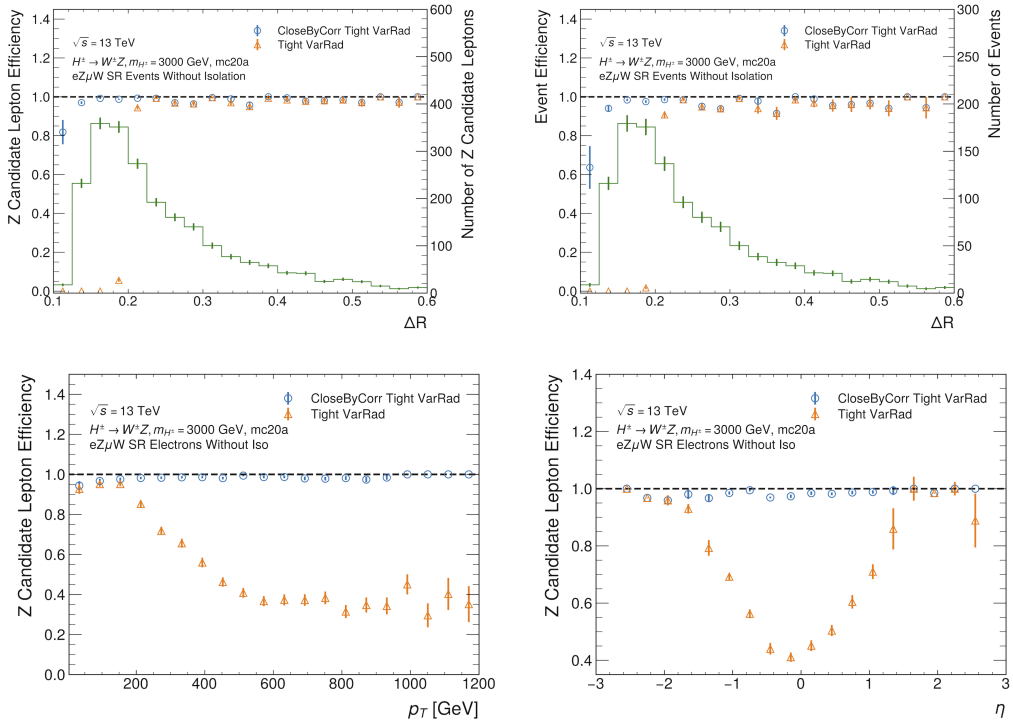


Figure A.1: The Z candidate lepton efficiency (top left) and event efficiency (top right) plotted as a function of ΔR in the $eZ\mu W$ channel. The Z candidate lepton efficiency as a function of p_{T} (bottom left) and η (bottom right).

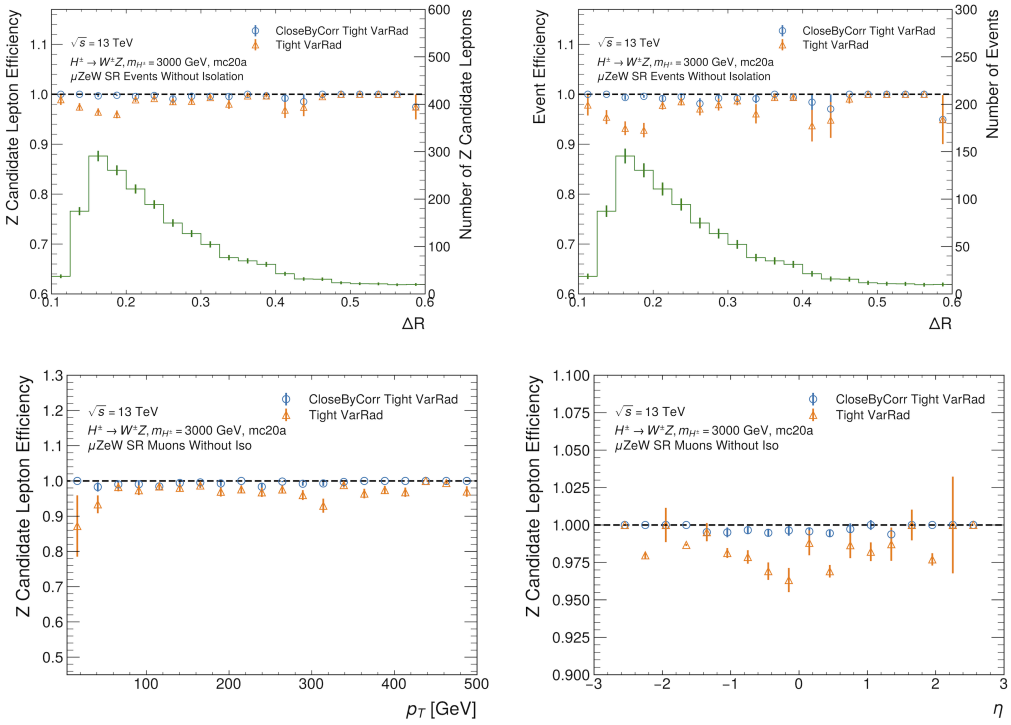


Figure A.2: The Z candidate lepton efficiency (top left) and event efficiency (top right) as a function of ΔR in the μZeW channel. The Z candidate lepton efficiency as a function of p_{T} (bottom left) and η (bottom right).

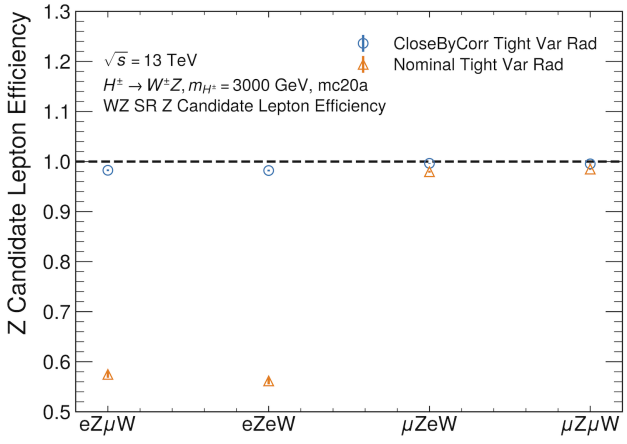


Figure A.3: The overall Z candidate lepton efficiency for different lepton channels in the modified $W^{\pm}Z$ signal region.

for loose leptons, which does slightly modify the events in the signal region due to the 4 lepton veto. However, this did not significantly affect the results. The effect of the close-by-correction for the W lepton was not evaluated. The event efficiency for the two isolation requirements was briefly studied for some background samples but was found to have a negligible effect since the Z bosons in standard model $W^{\pm}Z$ processes are not usually significantly boosted. The close-by-correction for lepton isolation variables was adopted by the analysis and is used for all subsequent isolation requirements in this thesis.

Appendix B

Monte Carlo Samples

$H_5^{\pm\pm}$ **Mass (GeV)**	**DSID**	σ **(fb)**
200	525925	16.0
225	525926	15.1
250	525927	11.9
275	525928	11.2
300	525929	9.14
325	525930	8.55
350	525931	7.15
375	525932	6.68
400	525933	5.70
425	525934	5.32
450	525935	4.59
475	525936	4.29
500	525937	3.74
525	525938	3.44
550	525939	3.09
600	525940	2.56
700	525941	1.80
800	525942	1.30
900	525943	0.961
1000	525944	0.717
1500	525945	0.192
2000	525946	0.0592
3000	525947	0.00688

Table B.1: Dataset identifiers (DSIDs) and cross-sections for Monte Carlo signal samples with $\sin(\theta_H) = 0.25$.

Sample	**Category**	**DSID**
$W^{\pm}W^{\pm}jj$ EW	$W^{\pm}W^{\pm}$	700590
$W^{\pm}W^{\pm}jj$ QCD	$W^{\pm}W^{\pm}$	700603
$W^{\pm}W^{\pm}jj$ INT	$W^{\pm}W^{\pm}$	700594
$W^{\pm}Zjj$ EW	$W^{\pm}Z$	700588
$W^{\pm}Zjj$ QCD	$W^{\pm}Z$	700601
$W^{\pm}Zjj$ INT	$W^{\pm}Z$	700592
$ZZjj$ EW	ZZ	700493, 700600, 700602, 700587
$W^{\pm}W^{\mp}$	$W^{\pm}W^{\mp}$	700589, 700593
Z+jets	V+jets	700320-5, 700792-4, 700335-7
W+jets	V+jets	700338-46
$Z\gamma$+jets	$V\gamma$	700398-401, 700710
$W\gamma$+jets	$V\gamma$	700402-4, 700709
VVV fully leptonic	VVV	364242-9
$WWW \to \ell\nu\ell\nu jj$	VVV	364336-9
$t\bar{t}$	$t\bar{t}$	410470-1
$t\bar{t}\gamma$	$t\bar{t}$	504554, 500800
$t\bar{t}Z$	$t\bar{t}$	504330, 504334, 504342
$t\bar{t}W$	$t\bar{t}$	700168, 700205
t	Single Top	410644-5, 410658-9, 601353-4
tZ	Single Top	410560

Table B.2: Dataset identifiers (DSIDs) for background Monte Carlo samples, and the background categories that are used for histograms in Chapter 6.

Appendix C

Additional NN Input Plots

In Chapter 5, the neural networks are trained using many kinematic variables which characterize each event. The input feature distributions presented in Figures 5.3, 5.4, and 5.5 combine all the different $H_5^{\pm\pm}$ signal mass points into a single distribution. However, it is also useful to observe how the kinematic distributions vary between different $H_5^{\pm\pm}$ masses. These differences help to explain why the network struggles at low mass, and to provide context for the feature optimization procedure in Section 5.2. The following plots are produced in the NN training region from Chapter 5, provided in Table 5.1. The background distribution includes the $W^{\pm}W^{\pm}$ and $W^{\pm}Z$ samples from Table 5.2, and the 200, 300, 450, and 3000 GeV $H_5^{\pm\pm}$ signal samples are all shown individually.

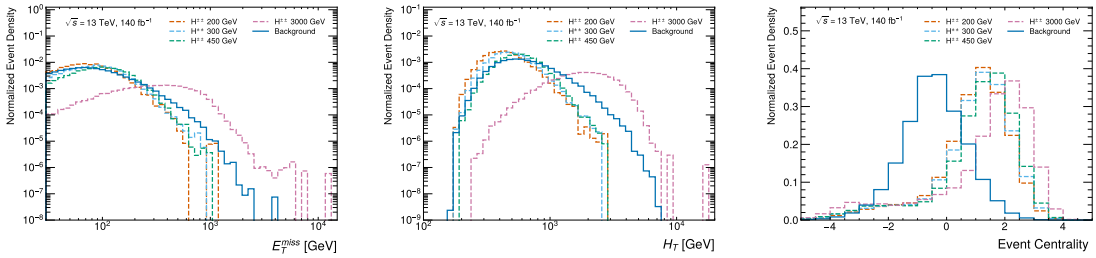


Figure C.1: Event-level input feature distributions for 200 GeV, 300 GeV, 450 GeV and 3000 GeV $H_5^{\pm\pm}$ signals and background.

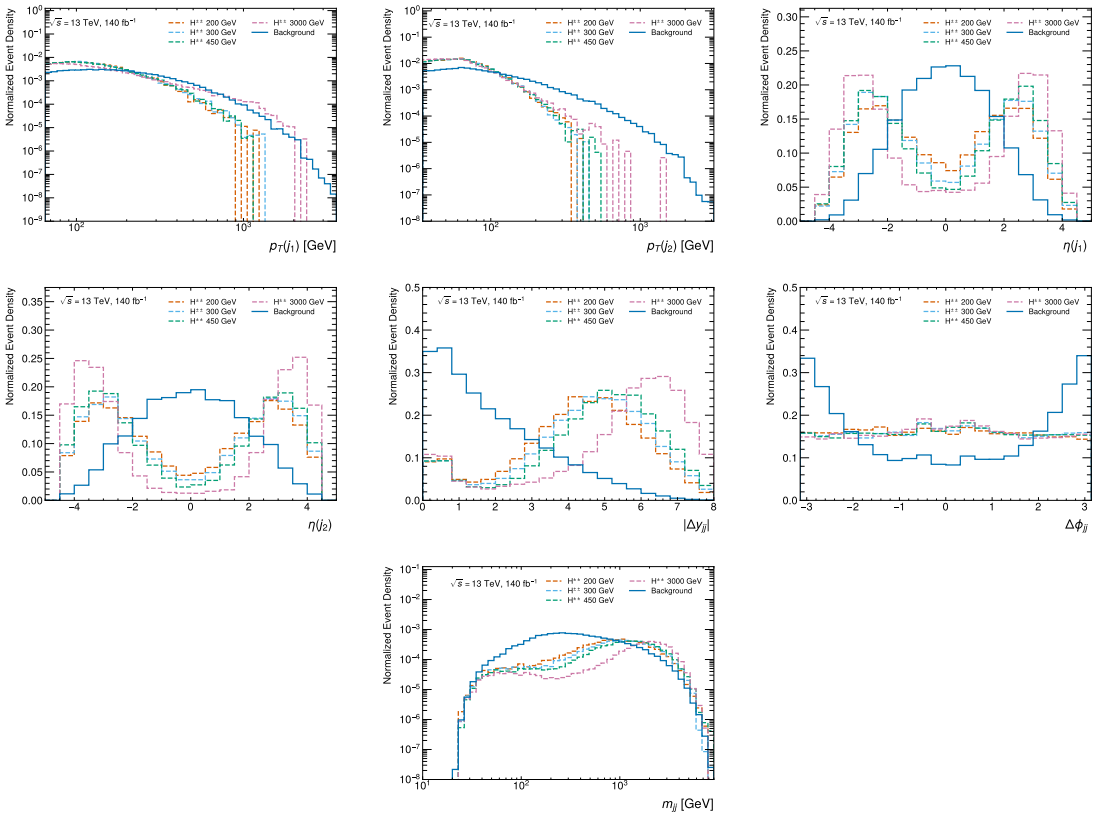


Figure C.2: Jet input feature distributions for 200 GeV, 300 GeV, 450 GeV and 3000 GeV $H_5^{\pm\pm}$ signals and background.

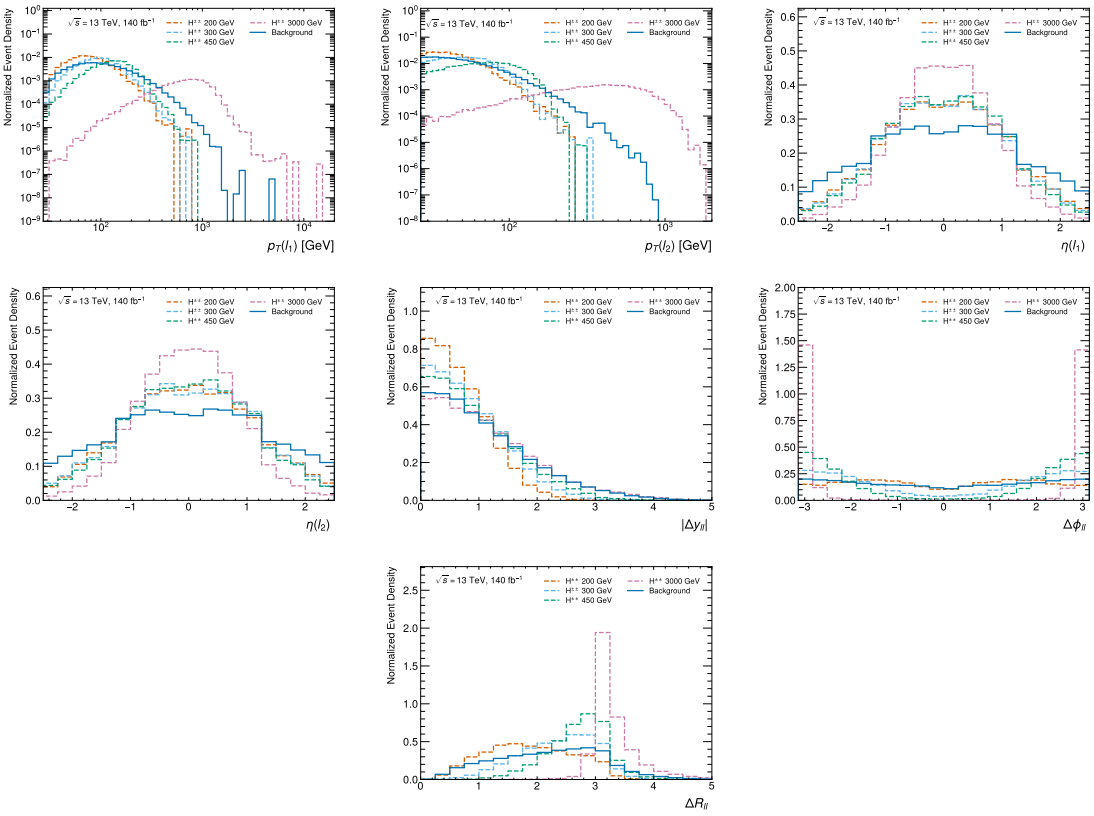


Figure C.3: Lepton input feature distributions for 200 GeV, 300 GeV, 450 GeV and 3000 GeV $H_5^{\pm\pm}$ signals and background.

Appendix D

Additional Pre-fit and Post-fit Plots

In Chapter 6, the pre-fit and post-fit plots for an Asimov fit with a 375 GeV $H_5^{\pm\pm}$ mass point were presented. An Asimov fit was also performed for other mass points: 225, 275, 325, 425, 475, 525, 600, 800, 1000, and 2000 GeV. The signal region and control region pre-fit and post-fit plots are provided here for the 225 GeV and 1000 GeV $H_5^{\pm\pm}$ mass points. These plots shows that, similar to the 375 GeV fit, there is very little signal contamination in the control regions. As well, the NN SR increases the signal region purity compared to the cuts-based SR. The normalization factors in Figures D.1 and D.4 are all fit to their nominal values —1.0 for $\mu(WW)$ and $\mu(WZ)$ and 0.25 for $\mu(H_5^{\pm\pm})$, which corresponds to $\sin(\theta_H) = 0.25$.

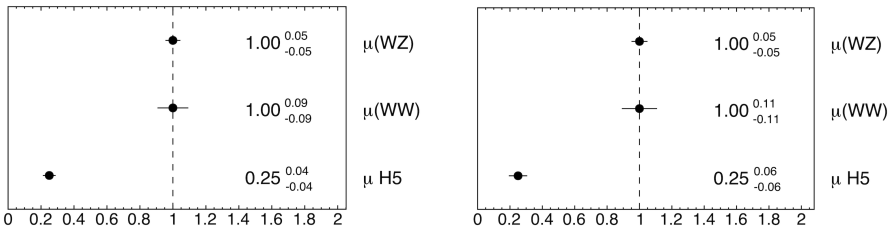


Figure D.1: The post-fit normalization factors for the 225 GeV $H_5^{\pm\pm}$ sample with the NN SR (left) and cuts-based SR (right).

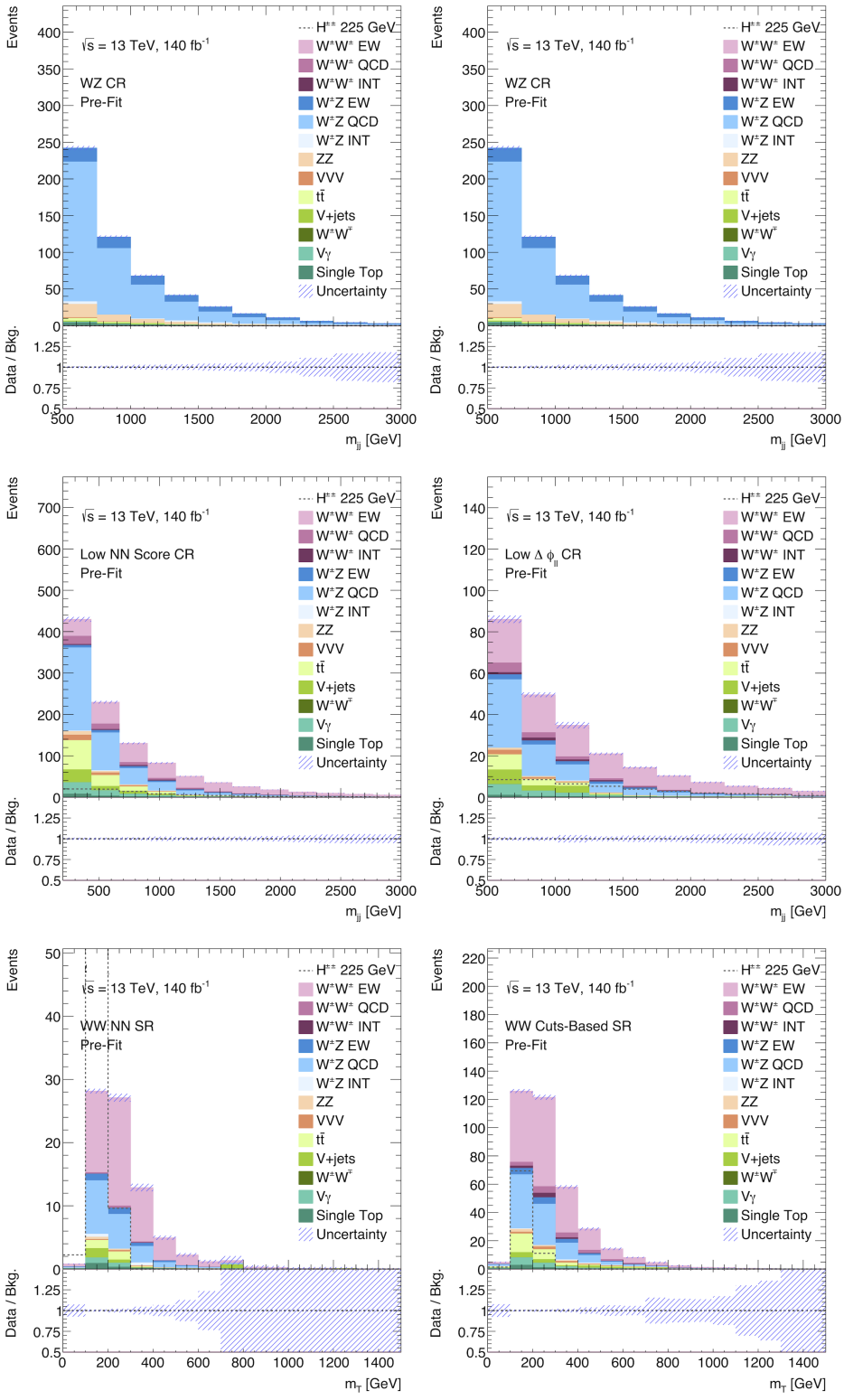


Figure D.2: The pre-fit plots for the NN SR (left) and the cuts-based SR (right) with a 225 GeV signal sample and $\sin(\theta_H) = 0.25$.

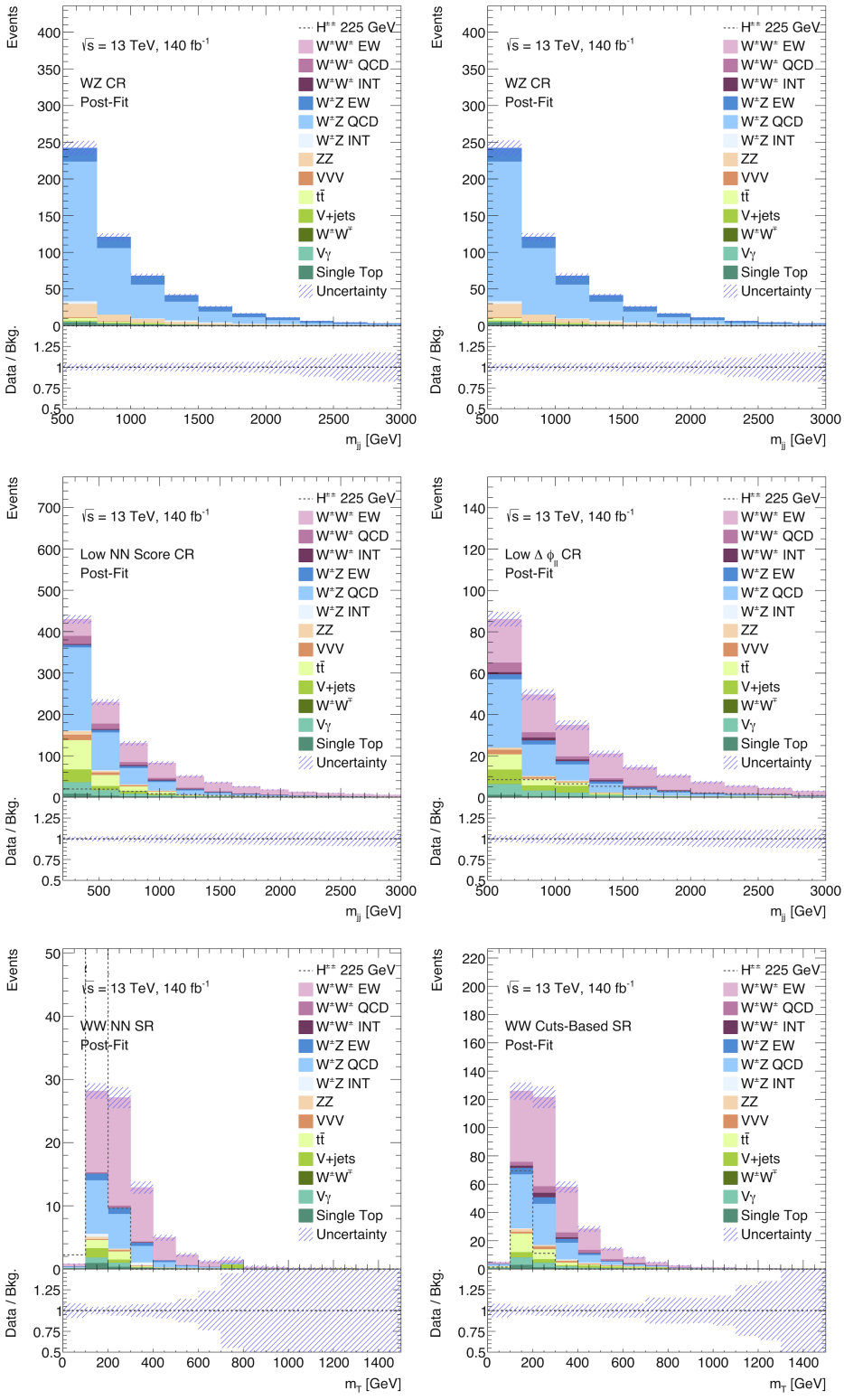


Figure D.3: The post-fit plots for the NN SR (left) and the cuts-based SR (right) with a 225 GeV signal sample and $\sin(\theta_H) = 0.25$.

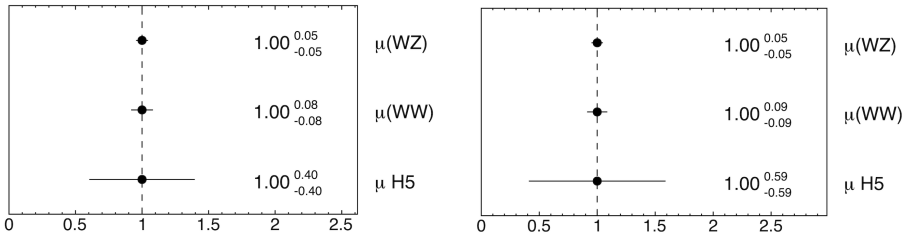


Figure D.4: The post-fit normalization factors for the 1000 GeV $H_5^{\pm\pm}$ sample with the NN SR (left) and cuts-based SR (right).

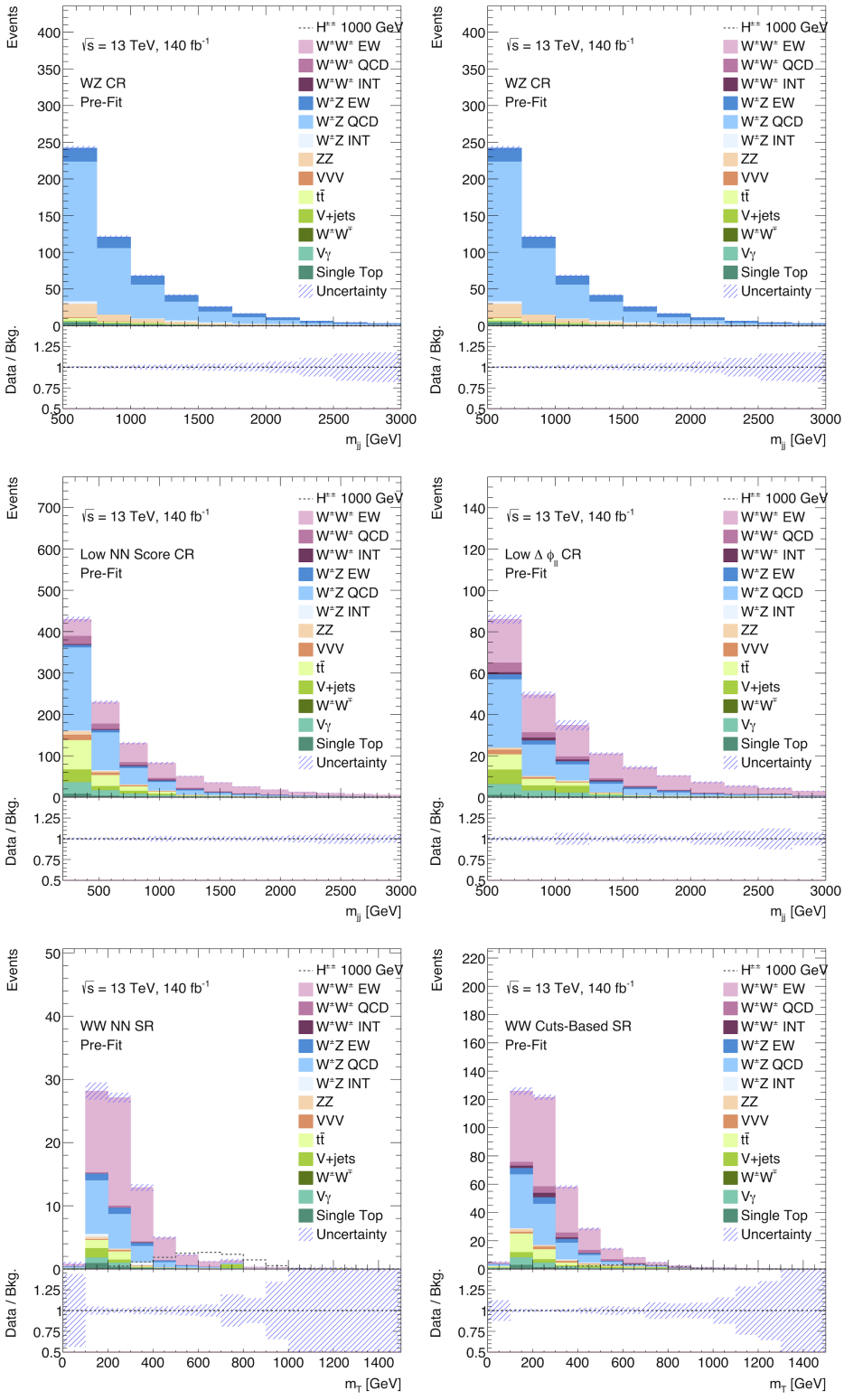


Figure D.5: The pre-fit plots for the NN SR (left) and the cuts-based SR (right) with a 1000 GeV signal sample and $\sin(\theta_H) = 0.25$.

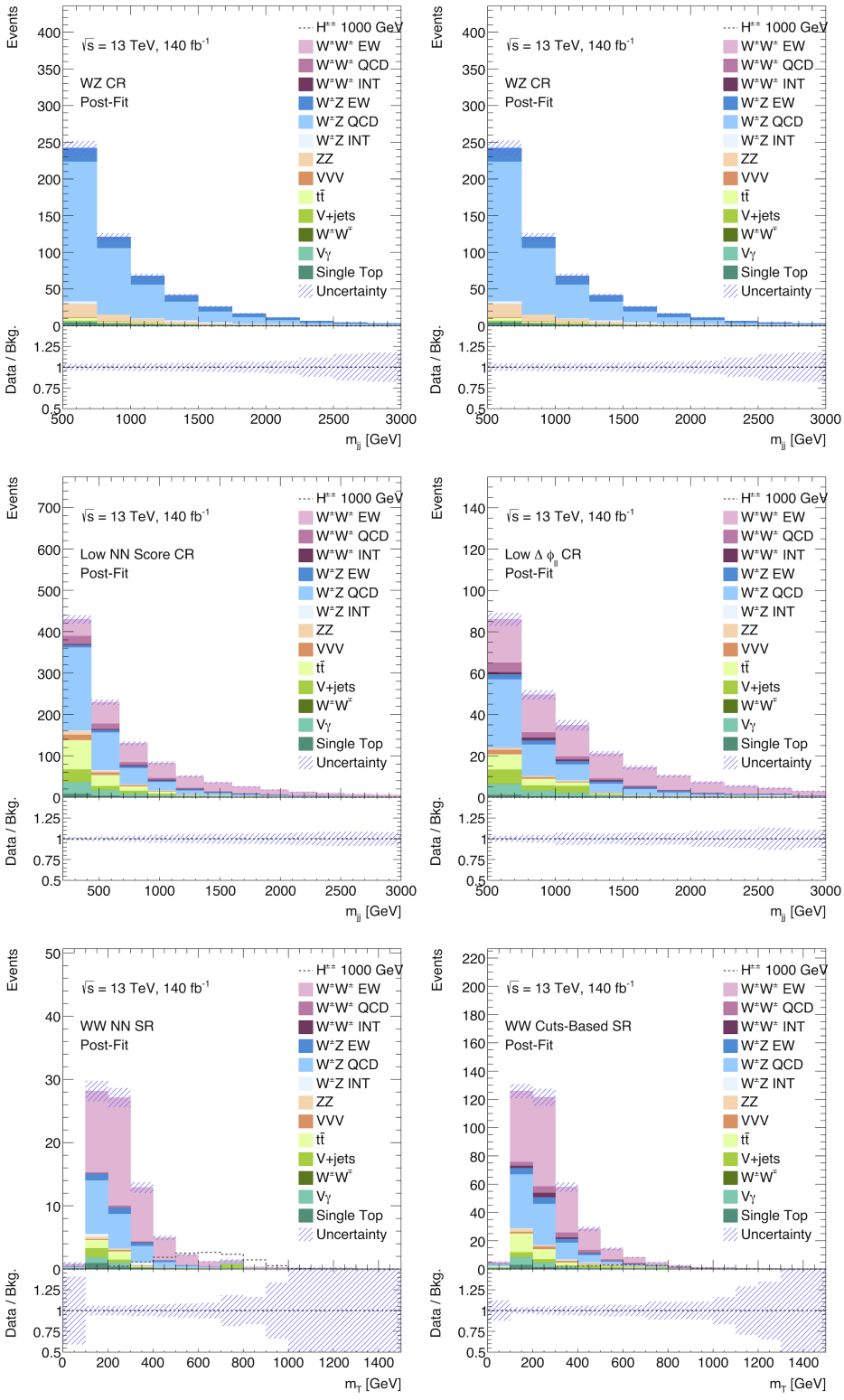


Figure D.6: The post-fit plots for the NN SR (left) and the cuts-based SR (right) with a 1000 GeV signal sample and $\sin(\theta_H) = 0.25$.

Appendix E

AUC Uncertainties

The metric that is used to compare different NN methods in this thesis is the area under the ROC curve (AUC). There are many sources of uncertainty that affect our ability to compare AUC values between different NNs and different $H_5^{\pm\pm}$ mass points. It is important to estimate these uncertainties when making claims about the NN classification performance.

Optimizing a neural network using a stochastic gradient descent (SGD) method is an inherently random process. The choice of initial model parameters, the ordering of training events, and the nondeterminism of machine learning libraries can all contribute to the variation in NN performance [47]. According to Summers and Dideen, the choice of initial model parameters is by far the largest source of instability in SGD-like optimization [47]. However, Jordan finds that models which perform differently on the test set in fact perform very similarly on the underlying test set distribution [48]. Thus, arguing that the sampling of the test set distribution is the source of the random fluctuation in NN performance.

Therefore, the uncertainty due to the model initialization and the uncertainty due to the test set sampling are both estimated here in a simple way for Figure 6.2. The training set, test set, and training configuration which are used are identical to the optimized NN in Chapter 6. Recall that for the plots of AUC as a function of $H_5^{\pm\pm}$, both the validation set and test set were used to evaluate the AUC. Therefore, "test set" here refers to both the validation set and test set.

The initial model parameters are generated with a mean of 0 and a standard deviation of 0.05 using a random seed. In order to estimate the uncertainty from the model parameter initialization, five models are trained with five different random seeds. The uncertainty in the AUC due to the choice of random seed is then the standard deviation of the AUCs for these five models.

The uncertainty in the AUC due to the statistical variation of the test set is estimated by evaluating the NN performance on four subsets of the test set. The uncertainty on the AUC due to the test set statistics is then approximated by the standard deviation of these AUCs divided by the square root of the number of subsets ($\sqrt{4} = 2$).

Table E.1 shows the uncertainty in the AUC due to the test set statistical uncertainty and the random seed as a percentage of the AUC for different $H_5^{\pm\pm}$ mass points. The uncertainty generally increases with mass from 250 GeV to 3000 GeV. This could be a reflection of the fact that the physical event weights used in training are larger at low mass, which creates greater consistency in

the NN performance. However, the mass points which have the largest uncertainty are 200 GeV and 225 GeV. This variability in classification performance is likely due to the fact that these mass points are very similar to background and also the most different from the medium to high mass samples. The two uncertainties are added in quadrature to obtain an estimate of the total percentage uncertainty, which assumes the two uncertainties are independent.

The percentage uncertainties which are determined here for the optimized NN are used to estimate the percentage uncertainties for other NNs. The uncertainty for the AUC plots as a function of mass (Figures 5.9, 5.12, 5.14, 5.16 and 5.17) are determined by multiplying each AUC by the total percentage uncertainty for that mass point, shown in the rightmost column of Table E.1. For the feature optimization AUC plot, Figure 5.11, the AUC is calculated using many mass points. Therefore, the average total percentage uncertainty in Table E.1, 0.267%, is used to estimate the uncertainty for each AUC.

Mass (GeV)	**AUC**	**Test Set Stat. Unc. (%)**	**Model Seed Unc. (%)**	**Total (%)**
200	0.879	0.16	0.35	0.39
225	0.887	0.27	0.25	0.37
250	0.901	0.14	0.14	0.20
275	0.907	0.19	0.079	0.21
300	0.913	0.12	0.075	0.15
325	0.922	0.10	0.095	0.14
350	0.925	0.14	0.13	0.194
375	0.926	0.12	0.16	0.20
400	0.930	0.12	0.19	0.22
425	0.931	0.11	0.21	0.24
450	0.935	0.15	0.21	0.26
475	0.936	0.13	0.23	0.27
500	0.938	0.11	0.24	0.26
525	0.938	0.16	0.27	0.31
550	0.938	0.17	0.26	0.31
600	0.942	0.10	0.28	0.30
700	0.946	0.071	0.29	0.30
800	0.948	0.074	0.30	0.31
900	0.951	0.061	0.29	0.30
1000	0.951	0.075	0.29	0.29
1500	0.960	0.10	0.29	0.31
2000	0.966	0.082	0.29	0.30
3000	0.970	0.081	0.31	0.32

Table E.1: The percentage AUC uncertainty for the optimized NN at each mass point. The test set and seed uncertainties are added in quadrature to obtain the total uncertainty. This data corresponds to the AUC plot in Figure 6.2

Bibliography

[1] I. Ivanov. Building and testing models with extended Higgs sectors. *Progress in Particle and Nuclear Physics*, 95:160–208, 2017.

[2] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, September 2012.

[3] Wikipedia contributors. Standard model — Wikipedia, the free encyclopedia, 2025. [Online; accessed 4-July-2025].

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

[5] E. Lyndon and P. Bryant. LHC machine. *Journal of Instrumentation*, 3:S08001, August 2008.

[6] ATLAS Collaboration. The ATLAS experiment at the CERN Large Hadron Collider: a description of the detector configuration for Run 3. *Journal of Instrumentation*, 19(05):P05063, May 2024.

[7] CERN. CERN's accelerator complex, 2025. [Accessed: 2025-07-06].

[8] W. Herr and B. Muratori. Concept of luminosity. In *CERN Accelerator School and DESY Zeuthen: Accelerator Physics*, pages 361–377, 9 2003.

[9] ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC. *The European Physical Journal C*, 83:982, 2023.

[10] CERN. New schedule for CERN's accelerators, 2024. Press Release: https://home.cern/news/news/accelerators/new-schedule-cerns-accelerators.

[11] ATLAS Collaboration. The ATLAS experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08003, August 2008.

[12] ATLAS Collaboration. Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2. *The European Physical Journal C*, 77(10):673, 2017.

[13] ATLAS Collaboration. Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector. *The European Physical Journal C*, 76(11), October 2016.

[14] M. Cacciari et al. The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[15] S. Schramm. ATLAS jet reconstruction, calibration, and tagging of Lorentz-boosted objects. Technical report, CERN, Geneva, 2017.

[16] ATLAS Collaboration. Transforming jet flavour tagging at ATLAS. 2025. arXiv preprint arXiv:2505.19689.

[17] ATLAS Collaboration. Electron and photon performance measurements with the ATLAS detector using the 2015-2017 LHC proton-proton collision data, 2019.

[18] ATLAS Collaboration. Identification of electrons using a deep neural network in the ATLAS experiment. Technical report, CERN, Geneva, 2022.

[19] ATLAS Collaboration. Muon reconstruction and identification efficiency in ATLAS using the full Run 2 *pp* collision data set at $\sqrt{s} = 13$ TeV. *The European Physical Journal C*, 81(7), July 2021.

[20] ATLAS Collaboration. The performance of missing transverse momentum reconstruction and its significance with the ATLAS detector using 140 fb^{-1} of $\sqrt{s} = 13$ TeV *pp* collisions. 2024. arXiv preprint arXiv:2402.05858.

[21] H. Georgi and M. Machacek. Doubly charged Higgs bosons. *Nuclear Physics B*, 262:463–477, 1985.

[22] ATLAS Collaboration. Combination of searches for singly and doubly charged Higgs bosons produced via vector-boson fusion in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Physics Letters B*, 860:139137, January 2025.

[23] J. Nielsen. Vector boson fusion and vector boson scattering measurements at the Large Hadron Collider. *Proceedings of Science*, LHCP2022:024, 2023.

[24] H. Logan and M. Reimer. Characterizing a benchmark scenario for heavy Higgs boson searches in the Georgi-Machacek model. *Physical Review D*, 96(9), November 2017.

[25] ATLAS Collaboration. Search for resonant WZ $\rightarrow$ *lvll* production in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *The European Physical Journal C*, 83(663), 2023.

[26] ATLAS Collaboration. Measurement and interpretation of same-sign W boson pair production in association with two jets in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2024(4), April 2024.

[27] ATLAS Collaboration. The ATLAS simulation infrastructure. *The European Physical Journal C*, 70(3):823–874, September 2010.

[28] J. Alwall et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 2014(7), July 2014.

[29] C. Bierlich et al. A comprehensive guide to the physics and usage of PYTHIA 8.3, 2022.

[30] E. Bothmann et al. Event generation with Sherpa 2.2. *SciPost Physics*, 7(3), September 2019.

[31] S. Frixione et al. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *Journal of High Energy Physics*, 2007(11):070–070, November 2007.

[32] M. Rauch. Vector-boson fusion and vector-boson scattering. 2016. arXiv preprint arXiv:1610.08420.

[33] G. Karagiorgi et al. Machine learning in the search for new fundamental physics. *Nature Reviews Physics*, 4:399–412, 2022.

[34] M. Abadi et al. TensorFlow: A system for large-scale machine learning, 2016. arXiv preprint arXiv:1605.08695.

[35] S. Navas et al. Review of particle physics. *Physical Review D*, 110(3):030001, 2024.

[36] A. F. Agarap. Deep learning using rectified linear units (ReLU). 2019. arXiv preprint arXiv:1803.08375.

[37] L. Prechelt. Early stopping - but when? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2017. arXiv preprint arXiv:1412.6980.

[39] ATLAS Collaboration. Formulae for Estimating Significance. Technical report, CERN, Geneva, 2020.

[40] Glen Cowan. *Statistical Data Analysis*. Oxford University Press, Inc., New York, 1 edition, 1998.

[41] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015. arXiv preprint arXiv:1502.03167.

[42] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[43] P. Baldi et al. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5), April 2016.

[44] H. Qu and L. Gouskos. Jet tagging via particle clouds. *Physical Review D*, 101(5), March 2020.

[45] V. Mikuni and F. Canelli. ABCNet: an attention-based method for particle tagging. *The European Physical Journal Plus*, 135(6), June 2020.

[46] V. Mikuni and F. Canelli. Point cloud transformers applied to collider physics. *Machine Learning: Science and Technology*, 2(3):035027, July 2021.

[47] C. Summers and M. Dinneen. Nondeterminism and instability in neural network optimization. 2021. arXiv preprint arXiv:2103.04514.

[48] K. Jordan. On the variance of neural network training with respect to test sets and distributions, 2024. arXiv preprint arXiv:2304.01910.