Separating $t\bar{t}$ and HH Final States Using Neural Networks

Master's Thesis at the Faculty of Physics of
Ludwig-Maximilians-University of Munich

submitted by

Youn Jun Cho

born in Seoul

Munich, on 14.06.2024

Separation von $t\bar{t}$ und HH Endzuständen mit Neuronalernetze

Masterarbeit der Fakultät für Physik der
Ludwig-Maximilians-Universität München

vorgelegt von

Youn Jun Cho

geboren in Seoul

München, den 14.06.2024

Abstract

Immense research has been conducted on the Higgs Boson since its discovery. However, the sheer amount of background events in high energy particle colliders complicate the study of its physical properties. Research on the Higgs self interactions is no exception. The corresponding cross section is small compared to many competing processes with similar final states. For example, the top anti-top quark pair decays and the Higgs self interactions can have equivalent final states such as $b^+b^-W^+W^-$, but their cross sections are roughly 1 μb and 30 fb respectively, such that the two mass distributions overlap. To separate these two, feed forward neural networks were applied onto the data simulated with the MCatNLO event generators Herwig and MadGraph5. However, given a reasonable computation time, these event generators could not generate enough top background events to train the neural network. Therefore, certain features of these data beyond kinematics were modified in order to generate sufficient training data. Overall, this usage of the neural network was effective in separating the two end states.

Contents

Chapter 1

Introduction

The Higgs boson was first proposed in 1964 as a scalar boson responsible for particle mass generation in the Standard Model. Its existence was required to explain the metastability of the electroweak vacuum, and its discovery in 2012 at the Large Hadron Collider (LHC) was met with relief. Furthermore, it was found that the Higgs boson has mass (125.11 ± 0.11) GeV, suggesting that the Higgs boson also couples to itself. However, this self interaction remains unobserved to this day since its occurrence is rare compared to many competing processes with similar final states. [12] The Top anti-Top quark pair decay ($t\bar{t}$) is one example, and this is the background to be separated in this thesis. To do so, $t\bar{t}$ backgrounds were first selected from the Monte Carlo simulated $t\bar{t}$ events generated by Pythia6. It was not possible to collect enough $t\bar{t}$ backgrounds in di-Higgs decays given a reasonable computation time. Therefore, the $t\bar{t}$ events were analyzed to find under what conditions they could mimic Higgs self interaction events. These conditions were imposed on all events to create a large number of $t\bar{t}$ backgrounds. These $t\bar{t}$ backgrounds were then used to train feed forward neural networks. Observables for the jet pairs, such as their masses and their angular distributions, were constructed to facilitate this process. The ultimate purpose of this thesis is to support the ATLAS collaboration and its endeavour to measure the Higgs self coupling constant λ in the Higgs potential. [4] This thesis proceeds as follows.

Chapter 2 discusses the relevant particle physics theory, including gauge theory and detector physics. Chapter 3, introduces what artificial neural networks are and suggests how they can be relevant in physics. Chapter 4 explains the procedures to find the conditions under which the $t\bar{t}$ decays are likely to mimic the Higgs self interactions. Chapter 5 discusses the procuring of the datasets fed into neural networks as well as the training of the neural network to separate $t\bar{t}$ backgrounds in HH events. Chapter 6 summarizes the thesis and offers possible outlooks. Natural units $c = \hbar = 1$ are used all throughout this thesis, with electric charges in the units of electron charge e.

Chapter 2

Particle Physics

The Standard Model (SM) is the most experimentally well-verified model of known elementary particles and their non-gravitational interactions. Elementary particles are point-like particles without any known substructure. [7] Among these are Higgs bosons, and investing their self-interacting properties is the aim of this thesis. To do so, this chapter first elaborates on the relevant theoretical aspects of the SM. This chapter then concludes with describing the experimental setup for particle physics.

2.1 Gauge Theory

Two branches of field theory have provided the theoretical foundations for the SM. One is quantum field theory (QFT), which lies at the interface between quantum mechanics and special relativity. Quantum fields are field operators on Hilbert spaces, satisfying the canonical commutation relations from quantum mechanics. Fluctuations around quantum fields represent particles. QFT is by construct Poincaré invariant like special relativity: that is, invariant under spacetime translations and Lorentz transformations.

The other branch is gauge theory (GT), where the Lagrangian is invariant under local gauge transformations. [15] As rotations on fields, gauge transformations are said to be local if their phases dependent on each point on the manifold, and global if their phases are constants. The group of gauge transformations that leave the Lagrangian invariant is called the gauge group. If the gauge group is abelian, then the GT is said to be abelian. In GT, fields that transform amongst each other under local gauge transformations form multiplets, identified with charges associated to the gauge group.

This section first presents an overview of the SM: the SM phenomenology is discussed along with the construction of the SM from QFT. This section then elaborates on each sectors of the SM from a field theory perspective. This section concludes with discussing the role of the Higgs field in the SM.

2.1.1 The Standard Model

The first objective of this subsection is to outline the SM phenomenology summarized in tables (2.1) and (2.2). The SM fermions have spin 1/2 and constitute all known matter. Each has its own anti-particle with the same mass but with opposite charges. The SM fermions with color charges red, green, or blue are called quarks, and those without are called leptons. Both quarks and leptons have six flavors as shown in table (2.1). Quarks are either up-like with electric charges $+2/3$, or down-like with $-1/3$. Up-like and down-like quarks are each arranged in ascending order of mass, each forming columns in table (2.1). Leptons are either electron-like with electric charges -1, or neutrino-like with 0. Electron-like and neutrino-like quarks are each arranged in ascending order of mass, each forming columns in table (2.1). Each row in table (2.1) then forms a generation of SM fermions. [7]

At high energies, each SM fermion is either right handed if its intrinisic and orbital angular momenta are parallel, and left handed if antiparallel. However, right-handed neutrinos have not yet been observed. The handedness of a fermion is described by a charge-like quantum number T_3, the third component of weak isospin T. Right handed fermions form weak isospin singlets $T = 0$. Left handed fermions form weak isospin doublets $T = 1/2$, with components $T_3 = +1/2$ for up-like quarks and neutrino-like leptons, while $T_3 = -1/2$ for down-like quarks and electron-like leptons. [7]

Quarks form colorless bound states called hadrons with the exception of top quarks, whose lifetimes are too short to form bound states. Hadron types are determined by their valence quarks. Hadrons with three valence quarks are composite fermions called baryons. Hadrons with one quark antiquark valence pair are composite bosons called mesons. Hadrons are assigned charge-like quantum numbers called baryon numbers B : baryons have $B = 1$, anti-baryons $B = -1$, and all other particles have $B = 0$. Leptons are given charge-like quantum numbers called lepton numbers L. Each generation of leptons is given a separate lepton number. [7]

The SM interactions are electromagnetic, strong, and weak. Each of these are mediated by spin 1 parity odd bosons gauge bosons. The SM gauge bosons are photons, gluons, and weak bosons. Photons mediate electromagnetic interactions between electrically charged particles. Photons are electrically neutral and thus do not self-interact. Gluons mediate strong interactions between color charged particles. Gluons are color charged and thus do self-interact. Weak bosons $W^{\pm}$ and Z mediate weak interactions between SM fermions and among themselves. Z bosons mediate interactions between all SM fermions. W bosons mediate interactions between electron-like and their neutrino-like leptons or between up-like and down-like quarks. W bosons interact among themselves and with Z bosons and photons. Weak interactions conserve weak isospin. [7]

	Quarks		Leptons	
Name	Up (u)	Down (d)	Electron (e)	Neutrino (ν_e)
Mass	2.16 MeV	4.67 MeV	0.511 MeV	< 0.8 eV
Charge	$+2/3$	$-1/3$	-1	0
Name	Charm (c)	Strange (s)	Muon (μ)	Neutrino (ν_μ)
Mass	1.27 GeV	93.4 MeV	106 MeV	< 0.19 eV
Charge	$+2/3$	$-1/3$	-1	0
Name	Top (t)	Bottom (b)	Muon (τ)	Neutrino (ν_τ)
Mass	173 GeV	4.18 GeV	1.78 GeV	< 18.2 MeV
Charge	$+2/3$	$-1/3$	-1	0

Table 2.1: The masses and the charges of all SM fermions. The values are obtained from the particle listings in the Particle Data Group. All SM fermions have spin 1/2. [12]

The SM bosons are either gauge bosons or Higgs bosons, as summarized in table (2.2). Higgs bosons are spin 0 parity even bosons responsible for particle mass generation, interacting only with massive particles. Higgs bosons are massive and thus self interact. Higgs bosons also interact with all fermions and weak bosons but not with photons and gluons, which are massless.

Electric charges Q, baryon numbers B, and lepton numbers L are conserved in all SM interactions. However, particle four momenta are not always conserved due to quantum fluctuations prescribed by the Heisenberg uncertainty principle. Particles that do not satisfy the relativistic dispersion relations are said to be off-shell. Particles that do are said to be on-shell. [7]

	Gauge				Scalar
Name	Photon (γ)	Gluon (g)	$W^\pm$ Boson	Z Boson	Higgs Boson (H)
Mass	0 GeV	0 GeV	80.4 GeV	91.2 GeV	125 GeV
Charge	0	0	± 1	0	0

Table 2.2: The masses and the charges of all SM bosons. The values are obtained from the particle listings in the Particle Data Group. The gauge bosons have spin 1, while the scalar bosons have spin 0. [12]

The second objective of this subsection is to outline the construction of the SM from QFT and GT. First, impose local gauge invariance on Dirac Lagrangians of free spin 1/2 spinor fields corresponding to SM fermions. This introduces massless vector fields for each gauge group generators. The massless vector fields contribute additional terms to the Lagrangian such that the new Lagrangian is gauge invariant. Such terms represent interactions mediated by the massless vector fields between charged fields. The charged fields are just the spinor fields if the gauge group is abelian. The charged fields also include the massless vector fields if the gauge group is non-abelian. The charged fields interact by coupling to the massless vector fields, and the interaction strengths are

defined by parameters called coupling constants. Certain massless vector fields become massive upon spontaneous symmetry breaking (SSB), where the vacuum state breaks the usual gauge invariance. SSB is implemented by scalar fields called Higgs fields corresponding to Higgs bosons. The vector fields correctly correspond to the SM gauge bosons upon SSB. [15]

The resulting Lagrangian is used to find scattering matrix elements, which can be used to calculate observables. Therefore, if the scattering matrix elements diverge to infinity, the Lagrangian must be regularized. This involves introducing cutoff energies, which define the energy scales at which such a Lagrangian is valid. Such is possible because the Lagrangian is not yet written in terms of physical quantities. Rewriting the Lagrangian in terms of physical quantities is called renormalization, which involves redefining the fields and modifying the coupling constants between them such that the observables remain unaffected. Upon renormalization, the coupling constant acquires dependence on energy scales: physically, this reflects the variation of measured charges of particles with distance, which occurs due to vacuum polarization. [15] Furthermore, observables such as decay widths and cross sections can be computed perturbatively in powers of the coupling constant, matching experimental results with high accuracy.

This altogether establishes the SM as a $SU(3)_C \times SU(2)_L \times U(1)_Y$ GT of quantum fields, where C denotes the color charges, L the left handed fermions, and $Y := 2(Q - T_3)$ the weak hypercharges. There are three parts to the SM Lagrangian. The first part is the Lagrangian for quantum chromodynamics (QCD), the $SU(3)_C$ GT of strong interactions. The second part is the Lagrangian for electroweak theory, the $SU(2)_L \times U(1)_Y$ GT of electroweak interactions. The third part is the Lagrangian for the Higgs sector, responsible for particle mass generation. Concepts needed for this thesis are to be further elaborated in the following subsections. [15]

2.1.2 Quantum Electrodynamics

Consider a free spin 1/2 spinor field $\psi = \psi(x)$ with mass m defined on spacetime with coordinates $x^\mu = (t, \vec{\mathbf{x}})$, then its Poincaré invariant Lagrangian is given by the Dirac Lagrangian of the form:

$$\mathcal{L}_D = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi, \tag{2.1}$$

where the partial derivatives are with respect to x^μ, and Einstein notations are assumed. $\bar{\psi} := \psi^\dagger \gamma^0$ is the anti-spinor, and the γ matrices satisfy the Clifford algebra:

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}, \tag{2.2}$$

where $\eta^{\mu\nu} = \text{diag}\,(1, -1, -1, -1)$ is the Minkowski metric. The γ matrices in the Dirac representation are:

$$\gamma^0 = \begin{pmatrix} \mathbb{1}_2 & 0 \\ 0 & \mathbb{1}_2 \end{pmatrix}, \quad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \tag{2.3}$$

where $\mathbb{1}_N$ is the $N \times N$ identity matrix and σ^i the Pauli matrices for $i = 1, 2, 3$. [13] Applying the Euler Lagrange equations gives the Dirac equation:

$$(i\gamma^\mu \partial_\mu - m)\psi = 0, \tag{2.4}$$

which has general solutions of the form:

$$\psi(x) = \int \frac{d^3\vec{\mathbf{p}}}{(2\pi)^3 2E(\vec{\mathbf{p}})} \sum_{s=1,2} u_s(\vec{\mathbf{p}}) a_s(\vec{\mathbf{p}}) e^{-ip_\mu x^\mu} + v_s(\vec{\mathbf{p}}) b_s^\dagger(\vec{\mathbf{p}}) e^{ip_\mu x^\mu}, \tag{2.5}$$

where $E(\vec{\mathbf{p}}) = \sqrt{\vec{\mathbf{p}}^2 + m^2}$, and $p_\mu = (E, \vec{\mathbf{p}})$ is the four momentum. $a_s(\vec{\mathbf{p}})$ and $b_s^\dagger(\vec{\mathbf{p}})$ are creation operators for spinors $u_s(\vec{\mathbf{p}})$ and anti-spinors $v_s(\vec{\mathbf{p}})$ respectively, each satisfying the anti-commutation relations

$$\{a_s(\vec{\mathbf{p}}), a_{s'}^\dagger(\vec{\mathbf{p}}')\} = \{b_s(\vec{\mathbf{p}}), b_{s'}^\dagger(\vec{\mathbf{p}}')\} = (2\pi)^3 2E(\vec{\mathbf{p}}) \delta_{ss'} \delta^{(3)}(\vec{\mathbf{p}}' - \vec{\mathbf{p}}). \tag{2.6}$$

Next, consider the behavior of $\mathcal{L}_D$ under $U(1)$ gauge transformations. Parametrizing $U \in U(1)$ as $U = e^{i\theta}$ with $\theta \in \mathbb{R}$ its phase shows that $\mathcal{L}_D$ is invariant under global $U(1)$ transformations on ψ :

$$\psi \to \psi' = \psi e^{i\theta} \Longrightarrow \mathcal{L}_D \to \mathcal{L}_D' = \bar{\psi}'(i\gamma^\mu \partial_\mu - m)\psi' = \mathcal{L}_D \tag{2.7}$$

but not under local $U(1)$ transformations on ψ :

$$\psi \to \psi' = \psi e^{i\theta(x)} \Longrightarrow \mathcal{L}_D \to \mathcal{L}_D' = \bar{\psi}'(i\gamma^\mu \partial_\mu - m)\psi' = (1 + i\partial_\mu \theta(x))\mathcal{L}_D. \tag{2.8}$$

However, local $U(1)$ gauge invariance can be imposed onto $\mathcal{L}_D$ by replacing the partial derivatives ∂_μ with the $U(1)$ covariant derivatives $D_\mu := \partial_\mu + ieA_\mu$, where e is called the coupling strength of ψ to A_μ, and $A_\mu = A_\mu(x)$ is a vector field that transforms under local $U(1)$ gauge transformations as

$$A_\mu \to A_\mu' = A_\mu - \frac{1}{e}\partial_\mu \theta(x). \tag{2.9}$$

A_μ with mass M then contributes additional terms given by the Proca Lagrangian:

$$\mathcal{L}_P = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{1}{2} M^2 A_\mu A^\mu, \tag{2.10}$$

where $F_{\mu\nu} := (ie)^{-1}[D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the A_μ field strength tensor. The first term is invariant under local $U(1)$ gauge transformations defined in equation (2.9). However, the second term is not. Therefore, $\mathcal{L}_P$ is gauge invariant only if $M = 0$. [13]

Replacing the partial derivatives ∂_μ with the covariant derivatives $D_\mu = \partial_\mu + ieA_\mu$ and introducing a massless vector field A_μ invariant under local $U(1)$ gauge transformations therefore give the Lagrangian for quantum electrodynamics (QED):

$$\mathcal{L}_{QED} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi + \mathcal{L}_P = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi + e\bar{\psi}\gamma^\mu A_\mu \psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \quad (2.11)$$

The first term still represents the free propagation of spin $1/2$ fermions ψ with mass m. The physical meaning of e and A_μ can be understood by using the Noether theorem and referring back to the Lagrangian for classical electrodynamics. By Noether's theorem, $U(1)$ transformations on ψ yield the conserved current $j^\mu = e\bar{\psi}\gamma^\mu\psi$ with the conserved charge $Q := \int j^0 \, d^3\vec{\mathbf{x}} = e$, which can be computed using equations (2.5) and (2.6). Since the last two terms in $\mathcal{L}_{QED}$ exactly match the Lagrangian for classical electrodynamics, e can be interpreted as the electric charge of ψ and A_μ the fields for photons. Therefore, the second term represents the tree level interactions between ψ with electric charge e and A_μ. The third term represents the free propagation of photons. Therefore, it is justified to call equation (2.11) the Lagrangian for QED and to denote its gauge group as $U(1)_Q$. Note that A_μ does not self interact since $U(1)_Q$ is abelian. Observables in QED processes are computed perturbatively in powers of the QED coupling constant $\alpha_Q \propto e^2$. Since $\alpha_Q << 1$, QED processes can be predicted with high accuracy. [13]

2.1.3 Quantum Chromodynamics

QCD is the $SU(3)_C$ GT of strong interactions, where C denotes color charges R, G, B. There are two kinds of fields with color charges. One consists of spin $1/2$ spinor fields ψ_i^j with masses m_i corresponding to quarks for each flavors $j = 1, \cdots, 6$ and colors $j = R, G, B$. These form $SU(3)_C$ triplets $\psi_i = (\psi_i^R, \psi_i^G, \psi_i^B)$. The other kind consists of massless vector fields G_μ^k corresponding to gluons for each $SU(3)_C$ generator labelled by $k = 1, \cdots, 8$ and given by the Gell-Mann matrices λ_k. G_μ^k's mediate interactions between color charged fields, and the color charged fields interact with strong coupling strengths $g_S \in \mathbb{R}$. The QCD Lagrangian $\mathcal{L}_{QCD}$ is given by:

$$\mathcal{L}_{QCD} = \sum_{i=1}^{6} \bar{\psi}_i(i\gamma^\mu D_\mu - m_i)\psi_i - \frac{1}{4}\sum_{k=1}^{8} G_{\mu\nu}^k G^{k\mu\nu}, \quad (2.12)$$

where $D_\mu := \mathbb{1}_3\partial_\mu + ig_S T_k G_\mu^k$ is the $SU(3)_C$ covariant derivative with $T_k := \lambda_k/2$, and $G_{\mu\nu}^k := (ig_S T_k)^{-1}[D_\mu, D_\nu] = \partial_\mu G_\nu^k - \partial_\nu G_\mu^k - g_S f_{ij}^k G_\mu^i G_\nu^j$ is the G_μ^k field strength tensor with $f_{ij}^k = -2i\,\mathrm{Tr}(T^k[T_i, T_j])$ the $SU(3)$ structure constants. The first term in equation

(2.12) represents the free propagation of quarks. The second term represents the free propagation of gluons. The G^k_μ's transform under local $SU(3)_C$ gauge transformations such that $\mathcal{L}_{QCD}$ is left invariant. [13]

Only some observables in QCD processes can be computed perturbatively in powers of the strong coupling constant $\alpha_S \propto g_S^2$. This is because of asymptotic freedom, where coupling constants decrease asymptotically with distance: at low energies, $\alpha_S >> 1$ and thus the perturbative approach fails. This results in the confinement of quarks and gluons into colorless bound states called hadrons, and the formation of hadrons from quarks and gluons is called hadronization. Therefore, experimental approaches such as the parton model are required. Partons are point-like constituents of hadrons. Since partons are not detectable, they are described with the parton density function (PDF), defined as the probability density of finding a parton with a specific fraction of the parent hadron's longitudinal momentum at a certain resolution scale. PDFs are found by fitting into experimental data from high energy hadron collisions. [12]

In high energy hadron collisions, partons quickly hadronize into stable particles, which further decay into other stable particles. Such stable particles are detectable, leaving narrow cones of particle tracks in the detectors. These narrow cones of stable particles with partons as their vertices are called jets. [12]

2.1.4 The Electroweak Theory

Electromagnetic and weak interactions are unified into electroweak interactions at high energies. The electroweak theory is the $SU(2)_L \times U(1)_Y$ GT of electroweak interactions, where L and Y denote fields with weak isospins and weak hypercharges respectively. There are two kinds of fields with weak isospins. One kind consists of left handed spin 1/2 spinor fields corresponding to the left handed SM fermions. These form $SU(2)_L$ doublets L^i_L for leptons and Q^i_L for quarks with weak isospin $T = 1/2$, defined as:

$$L^i_L = \left\{ \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix}, \begin{pmatrix} \nu_{\mu L} \\ \mu_L \end{pmatrix}, \begin{pmatrix} \nu_{\tau L} \\ \tau_L \end{pmatrix} \right\}, \quad Q^i_L = \left\{ \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \begin{pmatrix} c_L \\ s_L \end{pmatrix}, \begin{pmatrix} t_L \\ b_L \end{pmatrix} \right\} \tag{2.13}$$

for generation indices $i = 1, 2, 3$. [15] The right handed spin 1/2 spinor fields corresponding to the right handed SM fermions form $SU(2)_L$ singlets defined as:

$$e^i_R = \{e_R, \mu_R, \tau_R\}, \quad u^i_R = \{u_R, c_R, t_R\}, \quad d^i_R = \{d_R, s_R, b_R\}. \tag{2.14}$$

There are no fields corresponding to right handed neutrinos, which have not yet been observed. The other kind of fields with weak isospins consists of massless vector fields W^k_μ corresponding to electroweak bosons for each $SU(2)_L$ generators labelled $k = 1, 2, 3$ and given by the Pauli matrices τ_k. W^k_μ's mediate electroweak interactions

between fields with weak isosopins, and the fields with weak isosopins couple to the W_μ^k's with coupling strengths $g \in \mathbb{R}$. [15]

On the other hand, all fields in equations (2.13) and (2.14) have weak charges Y. They couple to a massless vector field B_μ corresponding to another electroweak boson with coupling strengths $g' \in \mathbb{R}$, and B_μ mediates electroweak interactions between them by implementing local $U(1)_Y$ gauge transformations. Since $U(1)_Y$ is abelian, B_μ does not self interact and corresponds to another electroweak boson.

Then the W_μ^k's and B_μ contributes two Lagrangians. The first is the kinetic Lagrangian $\mathcal{L}_K$ given by:

$$\mathcal{L}_K = -\frac{1}{4}\sum_{k=1}^{3} W_\mu^k W^{k\mu} - \frac{1}{4} B_\mu B^\mu, \tag{2.15}$$

representing the free propagation of electroweak bosons. The second is the interaction Lagrangian $\mathcal{L}_I$ given by:

$$\mathcal{L}_I = -\sum_{i=1}^{3} i\Big(\bar{L}_L^i \gamma^\mu D_\mu L_L^i + \bar{Q}_L^i \gamma^\mu D_\mu Q_L^i + \bar{e}_R^i \gamma^\mu D_\mu e_R^i + \bar{u}_R^i \gamma^\mu D_\mu u_R^i + \bar{d}_R^i \gamma^\mu D_\mu d_R^i\Big), \tag{2.16}$$

where $D_\mu = \partial_\mu - ig\tau_k W_\mu^k + ig' Y B_\mu$ is the $SU(2)_L$ covariant derivative with $\tau_i = \sigma_i/2$. This represents the electroweak interactions between electroweak bosons and all SM fermions. W_μ^k's and B_μ transform under local $SU(2)_L \times U(1)_Y$ transformations such that $\mathcal{L}_K$ and $\mathcal{L}_I$ are left invariant. [15]

However, electroweak bosons are different from weak bosons and photons: electroweak bosons are massless but weak bosons are not. This necessitates the introduction of Higgs fields $H := H(x)$ defined as the $SU(2)_L$ doublet of two complex scalar fields:

$$H = \begin{pmatrix} H^\alpha \\ H^\beta \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} H_1^\alpha + iH_2^\alpha \\ H_1^\beta + iH_2^\beta \end{pmatrix} \tag{2.17}$$

responsible for the weak boson masses. Higgs fields contribute two more Lagrangians. One is from SSB, where the vacuum state breaks the usual gauge invariance. Such is given by:

$$\mathcal{L}_{SSB} = (D_\mu H)^\dagger (D^\mu H) - V(H), \tag{2.18}$$

where D_μ is the $SU(2)_L$ covariant derivative. The first term represents the kinetic term for H, and the second represents the potential term. The simplest potential in which SSB occurs takes the form of the Mexican hat shown in figure (2.1):

$$V(H) = -\mu^2 |H|^2 + \lambda |H|^4, \tag{2.19}$$

where $\mu^2 > 0$ relates to the Higgs mass, and $\lambda > 0$ to the Higgs self interactions.

Minimizing $V(H)$ with respect to H gives the vacuum manifold:

$$\mathcal{M} = \{H \in \mathbb{C}^2 : 2\langle 0|H^\dagger H|0\rangle = v^2\}/(SU(2)_L \times U(1)_Y), \tag{2.20}$$

where $|0\rangle$ is the vacuum state, and $0 \neq v := \mu/\sqrt{\lambda}$ is the vacuum expectation value. [15]

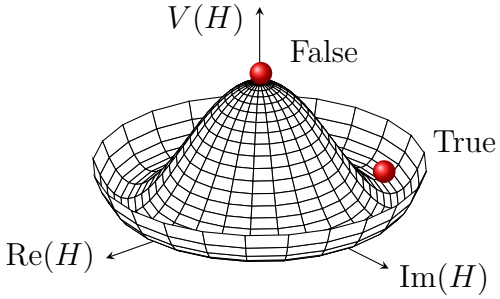


Figure 2.1: The simplest Higgs potential. The false vacuum is located at the local minimum, while the true vacuum is located at the global minimum. H transitions from the false vacuum to the true vacuum upon SSB. This demonstrates the metastability of the electroweak vacuum.

The ground state Higgs field can then be written in the unitary gauge, defined as:

$$\langle 0|H|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix}, \tag{2.21}$$

where the $SU(2)_L$ phase factor $e^{i\tau_k\theta^k}$ is scaled away using the $SU(2)_L$ gauge invariance of H and θ^k's correspond to Goldstone bosons for $k = 1, 2, 3$. There are precisely three θ^k's due to the Nambu-Goldstone theorem, which says that there is a massless spin 0 boson called the Goldstone boson for every spontaneously broken global continuous symmetry. Specifically, there is no $\tau \in SU(2)_L$ such that $\tau\langle 0|H|0\rangle = 0$ since $v \neq 0$, and this breaks the $SU(2)_L$ gauge invariance. Therefore, only one of the four $SU(2)_L \times U(1)_Y$ generators remains unbroken. The unbroken generator Q satisfies $Q\langle 0|H|0\rangle = 0$ and can be written in the unitary gauge as:

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.22}$$

This shows that $v \neq 0$ results in the SSB of $SU(2)_L \times U(1)_Y$ into $U(1)_Q$. Furthermore, the vector fields upon SSB can be written as linearly independent superpositions:

$$W_\mu^\pm = \frac{W_\mu^1 \mp iW_\mu^2}{\sqrt{2}}, \quad Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2+g'^2}}, \quad A_\mu = \frac{gB_\mu + g'W_\mu^3}{\sqrt{g^2+g'^2}}. \tag{2.23}$$

$W_\mu^\pm$ and Z_μ corresponding to weak bosons absorb θ_k's as longitudinal polarizations and thus become massive through a process called the Higgs mechanism. On the other hand, the unbroken gauge invariance yields a massless vector field A_μ corresponding to photons and thus Q represents the electric charge in the unitary gauge.

Since the three $SU(2)_L$ generators are broken, the Higgs field in equation (2.17) now has only one physical degree of freedom, which can be represented as perturbations $h(x)$ around the vacuum:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{2.24}$$

Physically, $h(x)$ represents the Higgs boson. Substituting equation (2.24) into equation (2.19) gives the Higgs potential:

$$V(h) = \mu^2 h^2 + \sqrt{\lambda}\mu h^3 + \frac{\lambda}{4} h^4, \tag{2.25}$$

and reading off the h^2 term gives the mass of the Higgs boson $M_H = \sqrt{2\mu^2} = \sqrt{2\lambda}v$. Higher order terms represent Higgs self interactions shown in figure (2.2).

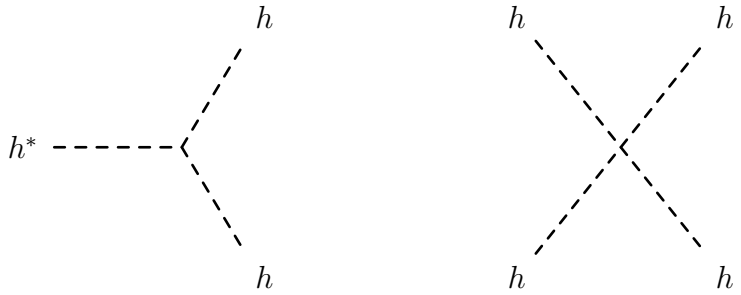


Figure 2.2: Feynman diagrams for Higgs self interactions. On the left is the trilinear coupling, and on the right is the quadrilinear coupling.

Furthermore, substituting equation (2.23) into equation (2.18) and collecting terms quadratic in the vector fields give:

$$\mathcal{L}_{SSB}^{(2)} = \frac{1}{2} g^2 v^2 W_\mu^+ W^{-\mu} + \frac{1}{8}(g^2 + g'^2) v^2 Z_\mu Z^\mu, \tag{2.26}$$

from which the mass terms can be read off as:

$$M_{W^\pm} = \frac{1}{2} g v, \quad M_{Z^0} = \frac{1}{2}\sqrt{g^2 + g'^2}\, v, \quad M_A = 0. \tag{2.27}$$

This matches the observed masses for $W_\mu^\pm$ and A_μ. There are more terms in $\mathcal{L}_{SSB}$ other than $\mathcal{L}_{SSB}^{(2)}$, representing electroweak interactions upon SSB.

However, the Higgs potential in equation (2.19) is only the simplest among many theoretically consistent ones. Since the vacuum expectation value is already measured to be $v \approx 246$ GeV, corrections to equation (2.19) or extensions to the SM would be

necessary if the measurements for λ do not satisfy $m_H = \sqrt{2\lambda}v$ from equation (2.25).

Higgs fields also generate spinor field masses through Yukawa interactions, contributing another Lagrangian:

$$\mathcal{L}_Y = -Y^e_{ij}\bar{E}^i_L H e^j_R - Y^u_{ij}\bar{Q}^i_L \tilde{H} u^j_R - Y^d_{ij}\bar{Q}^i_L H d^j_R + \text{h.c.}, \tag{2.28}$$

where tildes denote charge conjugation $\tilde{H} = i\sigma_2 H^*$, and +h.c. consists of Hermitian conjugate terms. Y^e, Y^u, Y^d are Yukawa matrices, which have entries given by coupling strengths of electron-like, up-like, and down-like spinor fields respectively to Higgs fields. Yukawa matrices for neutrinos are excluded since right handed neutrinos have not been observed yet. Yukawa matrix entries relate to fermion masses, which become more conspicuous upon SSB. [15]

Upon SSB, equation (2.28) becomes:

$$\mathcal{L}_Y = -\frac{v}{\sqrt{2}} \sum_f \bar{f}_L Y^f f_R + \text{h.c.}, \tag{2.29}$$

where $f = e, u, d$ denotes electron-like, up-like, down-like spinor fields respectively. Since $Y^f Y^{f\dagger}$ is Hermitian, there exist diagonal matrices M^f and unitary matrices U^f such that $Y^f Y^{f\dagger} = U^f (M^f)^2 U^{f\dagger}$ by the spectral decomposition theorem. This implies that Y^f can be written as $Y^f = U^f M^f K^{f\dagger}$ for any unitary matrix K^f. After performing the change of basis $f_L \to U^f f_L$ and $f_R \to K^f f_R$, equation (2.29) can be written as:

$$\mathcal{L}_Y = -\sum_f m^f \bar{f}_L f_R + \text{h.c.}, \tag{2.30}$$

where $m^{f,j} = \frac{v}{\sqrt{2}} M^{f,jj}$ is the mass of the fermion of type $f = e, u, d$ in generation j. This demonstrates how Higgs fields generate particle masses. Altogether, the electroweak Lagrangian $\mathcal{L}_{EW}$ can be written as:

$$\mathcal{L}_{EW} = \mathcal{L}_K + \mathcal{L}_I + \mathcal{L}_{SSB} + \mathcal{L}_Y, \tag{2.31}$$

where the exact forms of each terms change upon SSB. The SM Lagrangian can then be written as:

$$\mathcal{L}_{SM} = \mathcal{L}_{EW} + \mathcal{L}_{QCD} + \sum_{i=1}^{6} m_i \bar{\psi}_i \psi_i. \tag{2.32}$$

The last term accounts for the quark mass terms in $\mathcal{L}_{QCD}$, which have already been included in $\mathcal{L}_Y$. [15]

2.2 High Energy Collisions

Many aspects of the Standard Model such as the existence of a Higgs boson has been precisely tested at the Large Hadron Collider (LHC), a hadron to hadron synchrotron-type collider operated by the European Organization for Nuclear Research (CERN) at energy scales of TeV. With 27 kilometers in radius, the LHC goes through multiple detectors, including A Toroidal LHC ApparatuS (ATLAS). The ATLAS detector is cylindrical in shape, with multiple layers that each measure different observables needed to analyze high energy collisions. [4]

This section starts with explaining observables to be measured from particle colliders. General components of particle detectors as well as their roles in measuring necessary observables are then presented, with more focus on those of the ATLAS detector. This section concludes with describing how such measurements are used for analysis.

2.2.1 Collider Observables

High energy collisions are subject to special relativity. In special relativity, Lorentz transformations Λ^ν_μ are antisymmetric tensors that map four vectors v^μ in one reference frame to four vectors $v'^\nu = \Lambda^\nu_\mu v^\mu$ in another reference frame such that their Minkowski norms $||v|| := \eta_{\mu\nu} v^\mu v^\nu = v_\mu v^\mu$ are left invariant, with $\eta_{\mu\nu}$ the Minkowski metric. This implies that $\eta_{\mu'\nu'} = (\Lambda^T)^\mu_{\mu'} \eta_{\mu\nu} \Lambda^\nu_{\nu'}$. Lorentz transformations include rotations in three spatial dimensions and the rotation-free ones called boosts.

Furthermore, $\Lambda^\mu_\nu = \Lambda^\mu_\nu(\vec{\boldsymbol{\beta}})$ can be parametrized by boost vectors $\vec{\boldsymbol{\beta}}$, defined as:

$$\vec{\boldsymbol{\beta}} := \frac{\vec{\mathbf{p}}}{E}, \tag{2.33}$$

where $\vec{\mathbf{p}}$ is the momentum and E is the energy of the particle. Closely related to $\vec{\boldsymbol{\beta}}$ are boost factors γ, defined as:

$$\gamma := \frac{1}{\sqrt{1 - \vec{\boldsymbol{\beta}}^2}} = \frac{E}{M}, \tag{2.34}$$

where $M := ||p|| = \sqrt{E^2 - \vec{\mathbf{p}}^2}$ is the mass and $p^\mu = (E, \vec{\mathbf{p}})$ is the four momentum of the particle. [7] Lorentz transformations on $p^\mu = (E, \vec{\mathbf{p}})$ in one reference frame yield $p'^\mu = (E, \vec{\mathbf{p}}')$ in another given by:

$$\vec{\mathbf{p}}' = \vec{\mathbf{p}} + \gamma\Big(\frac{\gamma\vec{\boldsymbol{\beta}} \cdot \vec{\mathbf{p}}}{1+\gamma} + E\Big)\vec{\boldsymbol{\beta}}, \quad E' = \gamma E + \gamma\vec{\boldsymbol{\beta}} \cdot \vec{\mathbf{p}}, \tag{2.35}$$

demonstrating that γ parametrizes the mixing between the space-like and the time-like components of four vectors upon Lorentz transformations. [12]

γ can be written in terms of rapidity w, defined as:

$$w := \cosh^{-1}\gamma = \tanh^{-1}|\vec{\boldsymbol{\beta}}| = \frac{1}{2}\ln\frac{E+|\vec{\mathbf{p}}|}{E-|\vec{\mathbf{p}}|}. \tag{2.36}$$

The rapidity difference Δw between two particles is invariant under longitudinal boosts, along the axis $\hat{z}$ of particle beams. Similarly, the beam rapidity w_z along $\hat{z}$, defined as:

$$w_z := \tanh^{-1}\beta_z = \frac{1}{2}\ln\frac{E+p_z}{E-p_z}, \tag{2.37}$$

is also invariant under longitudinal boosts. [12] The transverse momentum p_T perpendicular to $\hat{z}$ and the polar angle ϕ in the transverse plane are both invariant under longitudinal boosts since p_T and ϕ are defined on the plane perpendicular to $\hat{z}$. The components of p^μ can then be written as:

$$p_x = p_T\cos\phi, \quad p_y = p_T\sin\phi, \quad p_z = M_T\sinh w_z, \quad E = M_T\cosh w_z, \tag{2.38}$$

where $M_T = \sqrt{M^2+p_T^2}$ is the transverse mass. [12] However, this is not what particle detectors can measure due to the relativistic speeds involved with high energy collisions. Instead of w_z, the detectors measure the pseudorapidity η, defined as:

$$\eta := -\ln\tan\frac{\theta}{2}, \tag{2.39}$$

where θ is the angle between p_z and $\hat{z}$. Taking the relativistic limit $p >> M$ in equation (2.37) and defining $\cos\theta := p_z/|\vec{\mathbf{p}}|$ yield η in equation (2.39). η is invariant under longitudinal boosts since w_z is too. Applying $p >> M$ onto equation (2.38) also yields:

$$p_x = p_T\cos\phi, \quad p_y = p_T\sin\phi, \quad p_z \approx M_T\sinh\eta, \quad E \approx E. \tag{2.40}$$

The detectors measure p_T, ϕ, η, E for all detectable particles. These kinematic variables are then used to calculate the p^μ's to analyze high energy collisions. Such analysis is facilitated by introducing the angular distance ΔR, defined as:

$$\Delta R := \sqrt{(\Delta\phi)^2+(\Delta\eta)^2}, \tag{2.41}$$

which is also invariant under longitudinal boosts. Specifically, this is useful in particular for jet reconstruction at hadron colliders, where the parent jet is deduced from its daughters by four momenta conservation. [12]

Other than these kinematic variables, particle detectors also measure decay widths and cross sections. The decay width Γ of a particle is the probability per unit time that the particle decays and has the units of energy in natural units. Γ defines the particle

lifetime $\tau := 1/\Gamma$, defined as the time taken for the number of such particles to decay by a factor of e. Γ also defines the branching ratio $B_i := \Gamma_i/(\sum_k \Gamma_k)$ of a particular decay mode i of the particle. B_i can be seen as the likelihood that the particle takes the decay mode i. On the other hand, the cross section σ in a collision is the area of impact needed for a process to occur and has the units of barns $\mathrm{b} = 10^{-28}\,\mathrm{m}^2$ in SI units. σ can be seen as a measure of the probability that a particular process occurs upon collisions between two particles. Both Γ and σ are dependent on energy scales.

2.2.2 Particle Detection

The first objective of this subsection is to outline the observables measured at each layers of the ATLAS detector. Particles produced in high energy collisions at the LHC pipes first reach the inner detector, which measures the positions and the momenta of electrically charged particles. The inner detector has three main species of layers. The innermost layers constitute the pixel detector, which records the initial positions of the charged particles. The surrounding layers constitute the semiconductor tracker, which records the trajectories of the charged particles. Semiconductors are used to minimize the energy loss and to maximize the detector sensitivity. The outermost layers constitute the transition radiation tracker, which records the outgoing particle types. The toroidal magnet surrounding the inner detector helps the tracking of electrically charged particles. [4]

Particles then reach the calorimeter, which measure their energies. The calorimeter has two main species of layers. The inner layers constitute the liquid Argon calorimeter, which measures the energies of electrons, photons, and hadrons. Sheets of heavy metal are inserted within these layers to absorb incoming particles, creating new particles with lower energies. These particles then ionize the liquid argon between these layers. The resultant electric currents are then used to determine the energies of the original particles. The outer layers constitute the tile calorimeter, which measures the energies of remaining hadronic particles. Within the tile calorimeter are layers of steel that produces new particles, and layers of scintillators that produce photons with intensities proportional to the energies of the original particles. [4]

The calorimeter can stop almost all known particles except muons and neutrinos, which then reach the muon spectrometer. The muon spectrometer has five kinds of detectors. One consists of monitored drift tubes, made of aluminum and filled with gas. Muons passing through the tubes displace electrons from the gas, producing electric signals. These signals are then used to record the muon trajectories. The other four kinds are collectively referred to as fast response detectors and consist of resistive plate chambers, thin gap chambers, small strip thin gap chambers, and micromegas detectors. The first two provide muon triggers. The last two are used in high intensity LHC collisions to

quickly detect muons with high precision. The fast response detectors altogether give an estimate of the muon momenta. The overall structure of the ATLAS detector is summarized in figure (2.3). [4]

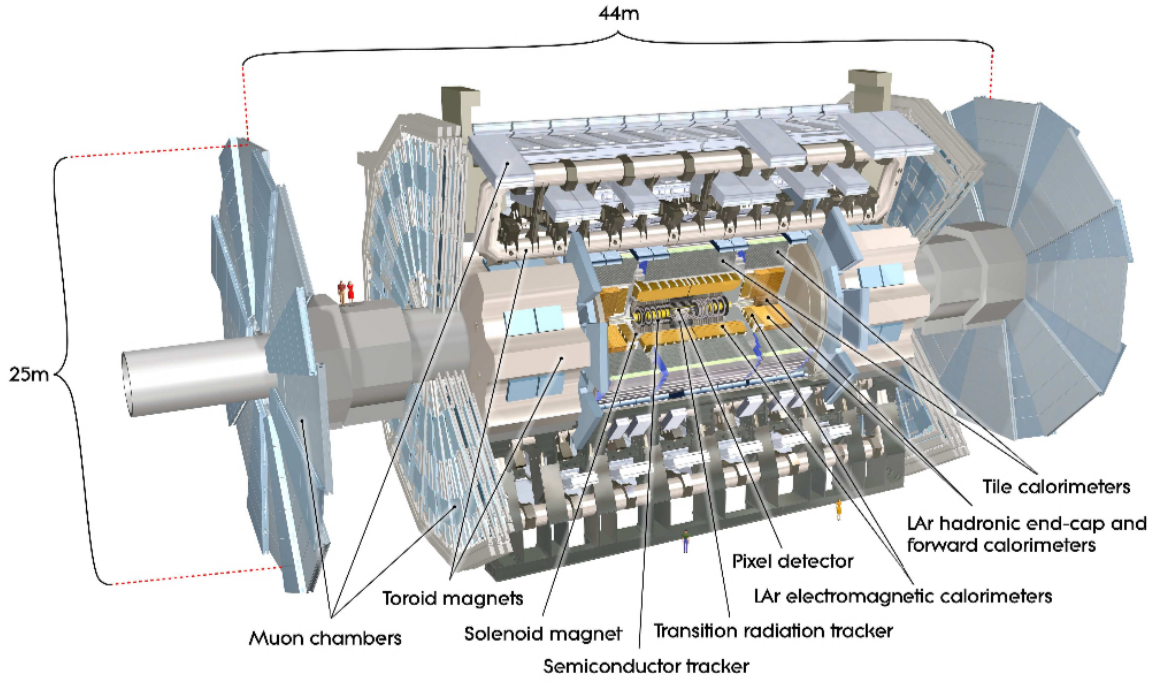


Figure 2.3: A schematic cross section through the ATLAS detector, highlighting its main components.

There are particles that cannot be detected. For example, neutrinos cannot be detected, and their kinematics must be inferred from transverse momenta and energy missing from the total. Quarks and gluons cannot be detected either because they hadronize before leaving any tracks in detectors. Their behavior can only be modeled using PDFs. There are also particles like tauons, whose lifetimes are too short to be detected. Such particles must be reconstructed from their daughters. [12]

2.2.3 Breit Wigner Distributions

Reconstructing an unstable particle from its daughters in high energy collisions yields a mass distribution that follows the Breit Wigner distribution (BWD) with a probability density function of the form:

$$f_{BW}(E) = \frac{K}{(E^2 - M^2)^2 + M^2\Gamma^2}, \tag{2.42}$$

where E is the energy scale, M is the unstable particle mass, and Γ is its decay width. K is a proportionality constant of the form:

$$K \propto \frac{\mu M \Gamma}{\sqrt{\mu + M^2}},$$

where $\mu = \sqrt{M^2(M^2 + \Gamma^2)}$. [5] That the mass distributions of unstable particles follow the BWD can be derived from their S matrix elements, which take the form:

$$\mathcal{M} \propto \frac{\sqrt{K}}{(E^2 - M^2) + iM\Gamma}.$$

Equation (2.42) reflects the idea that an unstable particle with mass M can be seen as a resonance peaking at $E = M$ with a natural decay width Γ. However, in high energy collision experiments, particle beams produce such resonances with uncertainties around the peaks $E = M$. Such uncertainties follow Gaussian distributions and can be accounted for in equation (2.42) using Gaussian BWDs (GBWDs), with probability density functions of the form [5]:

$$f_{GBW}(E) = \int_{-\infty}^{\infty} \frac{K}{(E'^2 - M^2)^2 + M^2\Gamma^2} \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{(E' - E)^2}{-2\sigma^2} dE'. \tag{2.43}$$

Chapter 3

Neural Networks

In biological nervous systems, electric signals enter the dendrites through synapses and travel to the neuron cell body, which then processes them. If the output signals exceed a certain threshold, the neuron cell body sends them to the axons, which connect to other neurons through synapses. Nervous systems learn by adjusting the coupling strengths between neurons. Neural networks (NNs) are idealizations of these biological observations. For example, biological neuron cells and coupling strengths between them correspond to activation functions and weights for artificial neurons respectively. [11]

NN based machine learning, or deep learning, has been increasingly successful in both identifying patterns in given data and generalizing them to unseen data. There are many kinds of NNs, each specialized for different tasks. For example, feed forward NNs (FFNNs) are useful for classifying input data. There are also convolutional NNs for image recognition, recurrent NNs for language processing, and generative adversarial NNs for generating output data. All of these are made possible by enhanced graphical processing units, which can simultaneously perform multiple tensor computations in high dimensional phase spaces. Increased availability of data also played a part. [11]

This chapter elaborates on the general structure of NNs and their learning algorithms, with more focus on FFNNs. Common problems in their learning processes such as vanishing gradients and overfitting are also outlined, along with possible solutions such as regularization and hyperparameter tuning.

3.1 Structure

NNs are graphs with neurons as their nodes. Each neuron $i = 1, \cdots, M$ at a discrete time step t takes input signals $s_j(t)$ for $j = 1, \cdots, N$ with weights w_{ij} and activates only when the weighted output $\sum_{j=1}^{M} w_{ij} s_j(t)$ exceeds its bias b_j as prescribed by its

activation function f_i, which outputs:

$$s_i(t+1) = f_i\Big(\sum_{j=1}^{N} w_{ij} s_j(t) - b_i\Big). \tag{3.1}$$

These outputs are then sent to other neurons as inputs until reaching the neurons that output the final signals. Succeeding neurons intake with their own weights the outputs from their predecessors. Therefore, the weights of each neuron represent the coupling strengths between neurons. NNs learn by updating these weights, which amounts to optimization in weight spaces. The agent must take additional measures to ensure that the NNs not only learn patterns in given datasets but also generalize to unseen datasets. Such is called training. [11]

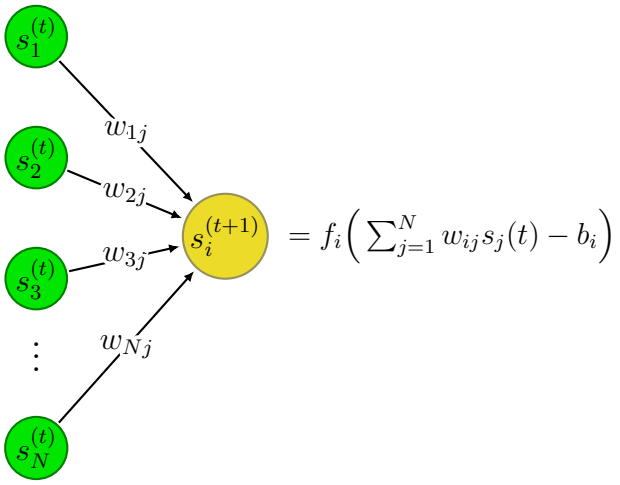


Figure 3.1: An illustration of how a neuron takes inputs $s_1(t), \cdots, s_N(t)$ at a time step t and outputs $s_i(t+1)$ at the next time step $t+1$ subject to the activation function f_i and the bias b_i.

Non-interacting neurons that share both inputs and output directions form layers, which come in three general kinds: the input layer that intakes raw data, the hidden layers that process the data, and the output layer that outputs the results. Layers closer to the output layer are said to be deeper and are schematically placed toward the right. NNs with hidden layers are called deep neural networks (DNNs). [11]

DNNs structured sequentially in terms of layers are called FFNNs. This means that all connections between neurons are one-way: neurons can only feed their outputs forward, only to the neurons in the immediately right layer. There are no connections that skip layers nor those that point backwards. If all signals $s_j(t)$ through a layer in a DNN are updated simultaneously, they can instead be denoted as s_j^l, where l is the layer number counting from the left. [11]

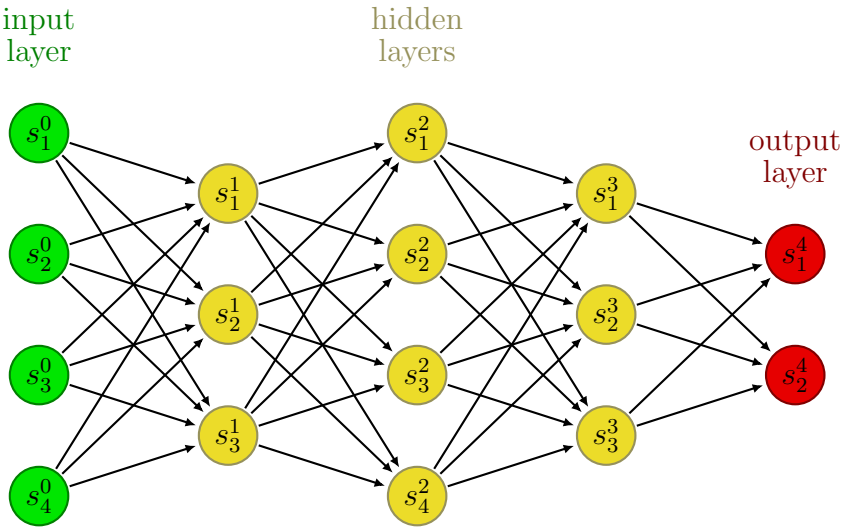


Figure 3.2: An example of a fully connected feed forward neural network. Each node is a neuron. The j-th neuron in the l-th layer is denoted by its output s_j^l.

3.2 Activation Functions

Activation functions are chosen to be non-linear to produce a more complex variety of outputs. Otherwise, the output would just be a different representation of the inputs since compositions of linear functions are still linear. For example, consider a sample with M features and a NN that takes each of the features as input signals. Running S samples through the NN with a linear function gives a system of S linear equations with M input variables. If $S > M$, then this system of linear equations is linearly dependent, implying that there are no input variables that can produce such outputs if the activation functions are linear. If the activation functions are non-linear, such outputs can be produced even when $S > M$. [11]

This subsection discusses two categories of activation functions. The first includes Rectified Linear Units (ReLUs), defined as

$$\mathrm{ReLU}(x) = \max(0, x), \tag{3.2}$$

which are commonly chosen for the neurons in DNNs. Note that ReLUs vanish for negative arguments, so not all neurons are activated simultaneously. In addition, both ReLUs and their derivatives are monotonic, helping the output converge. For these reasons, ReLUs are computationally efficient and are thus preferred in DNNs. However, their insensitivity to negative arguments could result in dead neurons, which return nothing but zero. Dead neurons are problematic for the whole NN because they no longer participate in the learning process, affecting all neurons connected to them. Such unwanted stagnation in the learning process is called the vanishing gradient problem.

This can be dealt with Leaky ReLUs defined as:

$$\text{LReLU}(x) = \max(\alpha x, x), \tag{3.3}$$

where $0 < \alpha < 1$ is the leakage constant to be initialized by the agent. Nonetheless, there are cases where even Leaky ReLUs fail to respond correctly to negative arguments. [11]

The second category of activation functions includes Sigmoid functions, defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{3.4}$$

which are commonly chosen for the output neurons in binary classification problems because their outputs can be interpreted as the probabilities of their inputs belonging to one out of the two classes, as per $\sigma : (-\infty, \infty) \to (0, 1)$. They are monotonic, with the derivatives

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)), \tag{3.5}$$

which are not monotonic and peak at $\sigma'(0) = \frac{1}{4}$. Their outputs do not converge as fast as those of ReLUs due to higher computational complexity. Furthermore, their asymptotic behaviors at $\lim_{x\to\infty} \sigma(x) = 1$ and $\lim_{x\to-\infty} \sigma(x) = 0$ result in saturated neurons, which cease to noticeably update the inputs to the subsequent neurons. This is another case of the vanishing gradient problem. [11]

As generalizations of Sigmoid functions, Softmax functions defined as

$$\sigma(x_1, \cdots, x_N)_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}} \tag{3.6}$$

for $j = 1, \cdots, K$ are preferred activation functions for output neurons in classification problems with K classes. Their outputs can also be interpreted as the probabilities of their inputs belonging to each of the multiple classes. Softmax functions may also suffer vanishing gradient problems. [11]

3.3 Weights and Biases

As mentioned in section 3.1, NNs learn by updating their weights. One way NNs use to update weights is backpropagation, which corresponds to learning by trial and error in biological nervous systems. Backpropagation can be implemented through gradient descent, while there are other ways to do so. Subsection 3.3.1 explains how weights and biases are initialized in general. Subsection 3.3.2 discusses how backpropagation is implemented particularly for the supervised learning of DNNs.

3.3.1 Initialization

Weights must be initialized first so that the NN updates them as it learns. Optimal initialization methods depend on activation functions. For example, He-initialization works best with ReLU variants: the weight distribution W for a NN with M input neurons is initialized to follow a normal distribution

$$W \sim N\Big(0, \frac{2}{M}\Big), \tag{3.7}$$

with mean 0 and standard deviation $\frac{2}{M}$. On the other hand, Xavier initialization works best with Sigmoid variants: the weight distribution W for a NN with M input neurons and N output neurons is initialized to follow a normal distribution

$$W \sim N\Big(0, \frac{2}{M+N}\Big), \tag{3.8}$$

with mean 0 and standard deviation $\frac{2}{M+N}$. In contrast, biases are initialized to zeros or small random constants in practice. [2]

3.3.2 Backpropagation

Labeled data are raw data pre-identified by the agent to help models produce correct outputs, and these correct outputs are called targets. Supervised learning (SL) is learning based on labeled data. In particular, SL for DNNs is about minimizing errors between the outputs and their targets, and measures for these errors are loss functions.

Choosing a particular loss function among the many depends on the task at hand. Commonly used is the mean squared loss (MSL), defined as

$$\mathcal{L}_{MSL} = \frac{1}{N}\sum_{i=1}^{N} |O_i - T_i|^2, \tag{3.9}$$

where O_i for $i = 1, \cdots, N$ is the output from each of the N output neurons, and T_i is the corresponding target. Since each O_i is a function of weights and biases, $\mathcal{L}$ is too.

As demonstrated, loss functions are generally non-negative functions of the outputs and are zero only if each output matches its targets. Since all outputs are functions of weights and biases, loss functions are too. Minimizing a loss function is therefore equivalent to updating weights and biases accordingly. Weights and biases are updated by backpropagation, particularly in SL: NNs test how close their outputs with current weights are to their targets using loss functions, and then update their weights and biases until the losses are minimized. backpropagation is commonly implemented by gradient descent (GD): for DNNs, weights w_{ij}^l in all layers $l = 1, \cdots, L$ each with N_l

neurons are updated for each run through all given samples, as per the update rule:

$$w_{ij}^l \to w_{ij}^l - \eta \left\langle \frac{\partial \mathcal{L}}{\partial w_{ij}^l} \right\rangle \tag{3.10}$$

for $i, j = 1, \cdots, L$, where the learning rate η is the step size in each run, and the angular brackets denote the averages over a fixed number of randomly selected samples. [11] The b_i^l's are updated similarly. Each run through all given samples is called an epoch, and a set of samples processed for updating weights is called a batch. Learning rates, numbers of epoches, and batch sizes must be properly chosen by the agent. The partial derivatives are calculated backwards using the chain rule:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^l} = \frac{\partial \mathcal{L}}{\partial s_{k_L}^L} \frac{\partial s_{k_L}^L}{\partial s_{k_{L-1}}^{L-1}} \cdots \frac{\partial s_{k_{l+1}}^{l+1}}{\partial s_{k_l}^l} \frac{\partial s_{k_l}^l}{\partial w_{ij}^l}, \tag{3.11}$$

where the repeated indices $k_m = 1, \cdots, N_m$ for $m = l, \cdots, L$ are summed over. This demonstrates that the activation functions must be at least piece-wise differentiable for GD to be applicable. The name backpropagation comes from this difference, where the outputs are updated forward, while the losses are updated backwards. GD is best implemented by ADAM optimization algorithms. [9]

Removing the angular brackets in equation (3.10) gives stochastic gradient descent (SGD), where weights and biases are updated for each sample. Since there is no longer a need to compute the average over all samples, SGD is computationally cheaper. However, this comes at the cost of convergence: outputs obtained with GD are more likely to be closer to the targets than those obtained with SGD especially in case there are outliers. For FFNNs in particular, the feed forward structure of the outputs ensures the convergence of outputs to their targets. [8]

3.4 Regularization

Datasets are samples drawn from unknown data distributions. Learning with datasets involve extracting trends while neglecting noises. Successful learning models should be able to generalize these trends to unseen datasets. In that sense, two kinds of problematic learning models exist: the underfitting ones and the overfitting ones.

Underfitting models are those that fail to learn trends in datasets. Underfitting occurs because the models are too simple to capture all samples in the datasets. Such can be solved rather easily by increasing the complexities of the learning models. For DNNs, this amounts to increasing the number of hidden layers or by increasing the number of neurons in each layers.

On the other hand, overfitting models are those that are too sensitive to noises that

they fail to generalize: they learn noises on top of trends so that generalizing to unseen datasets returns high errors. Overfitting occurs because such models are too complex compared to the number of samples in the datasets. A range of methods used to reduce overfitting in learning models is called regularization. Regularization is effectively implemented by modifying loss functions. For NNs, dropouts are as effective. This section discusses these commonly used regularization methods in detail.

3.4.1 Modifying Loss Functions

SL models update weights by backpropagation until their loss functions are minimized. Modifying these loss functions efficiently reduces overfitting by altering the update rule in equation (3.9). Such is implemented by adding a so-called regularizer $\tilde{\mathcal{L}}$ to the loss function $\mathcal{L}$:

$$\mathcal{L}' = \mathcal{L} + \lambda\tilde{\mathcal{L}}, \tag{3.12}$$

where the regularization strength $\lambda \geq 0$ is a parameter to be determined by the agent. λ must be carefully adjusted because it affects both model biases and model variances.

The model bias is the systematic error in the model's outputs compared to their targets. Said differently, low model bias implies the model's ability to extract trends. Therefore, overfitting occurs when the model bias is too low. In contrast, the model variance is the difference between the model's outputs over the whole dataset and those over its subsets. In other words, higher model variance implies the model's inability to neglect noises. Therefore, overfitting occurs when the model variance is too high. Since overfitting models react more sensitively to the changes in their loss functions, the λ's need to be small. Such is possible when the model biases are proportional to the λ's and the model variances are inversely proportional.

There are more details to equation (3.12): a common choice for $\mathcal{L}$ is the MSL defined in equation (3.9), and $\tilde{\mathcal{L}} = \tilde{\mathcal{L}}(w_{ij}^l)$ is added as a Lagrange multiplier in order to reduce weights w_{ij}^l. This is because the larger the weights are, the more sensitive the model is to the changes in its inputs, implying overfitting. Furthremore, there are three common choices for $\tilde{\mathcal{L}}$. The first is the L^2 regularizer, defined as:

$$\tilde{\mathcal{L}}_2 = \sum_{i,j,l} |w_{ij}^l|^2, \tag{3.13}$$

which ensures that the average magnitude of all weights is minimized. The second is the L^1 regularizer, defined as:

$$\tilde{\mathcal{L}}_1 = \sum_{i,j,l} |w_{ij}^l|, \tag{3.14}$$

which allows more weights to be zero, rather than minimizing the average magnitude

of all weights. The third is the cross entropy regularizer, defined as:

$$\tilde{\mathcal{L}}_{CE} = \sum_k p_k \ln p_k, \tag{3.15}$$

which is used for classification problems with K classes. $p_k = p_k(w_{ij}^l)$ is the probability that an output belongs to the class $k = 1, \cdots, K$. Furthermore, entropy is maximized in uniform distributions because all classes occur with equal probabilities. In other words, cross entropy regularization brings outputs closer to a uniform distribution, where entropy is maximized. This helps overfitting models to balance out their weights so that the outputs are more uniformly distributed. Note that cross entropy regularizers and entropies have opposite signs: this implies that cross entropy regularizers are minimized when entropies are maximized. [11]

3.4.2 Dropouts

Dropouts are designed to regularize NNs in particular. In dropouts, each neuron is dropped based on a dropout probability P defined as the likelihood of each neuron to be deactivated in an epoch. As a result, all neurons are forced to participate equally well in the training process. This helps the weights even out and thus reduces overfitting. Higher P generally implies stronger effects on overfitting, but excessively high P results in underfitting. P must thus be carefully adjusted by the agent. Dropouts are as effective as modifying loss functions when it comes to regularization because of their ability to adjust the weights.

Generalizations of dropouts such as dropconnects are also effective in regularizing NNs. In dropconnects, each connection between neuron is dropped based on a dropconnect probability, defined as the likelihood of each connection to be deactivated in an epoch. As a result, all weights are given equal chance to be updated in the training process. This again helps the weights even out and thus reduces overfitting. [11]

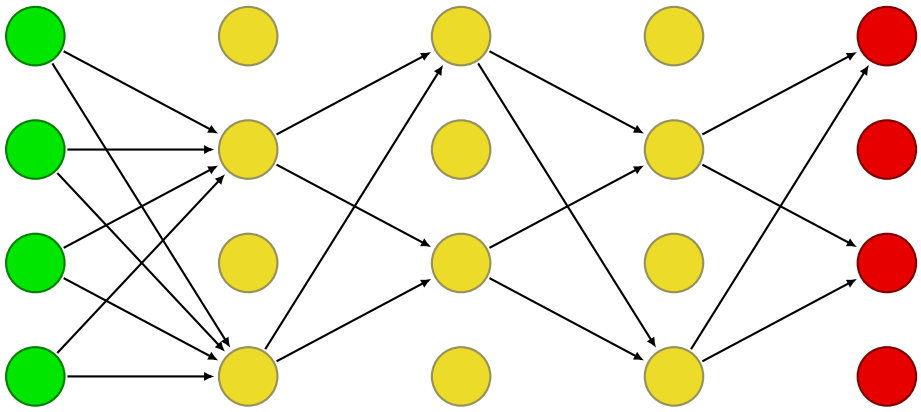


Figure 3.3: An example of a dropout regularization with a dropout probability of 40%.

3.5 Hyperparameters

As mentioned in section 3.1.1, ReLUs could result in dead neurons, possibly resulting in the vanishing gradient problem. This arises for example when the learning rate η is too high in GD: the weights turn increasingly negative, resulting in dead neurons for ReLUs. At the same time, lower η results in longer training times. Therefore, η must be properly initialized by the agent. Such overarching parameters to be determined by the agent are called hyperparameters.

η is not the only hyperparameter that define training processes: epoches and batch sizes are too. More epoches help loss functions for successful learning models reach their global minima. Similarly, smaller batch sizes facilitate the training process since the weights are updated more frequently according to equation (3.10), but this comes at the cost of longer training times.

There are also hyperparameters that define model complexities: leakage constants for leaky ReLUs from section 3.2.1, numbers of layers and numbers of neurons in each layer are some examples. Leaky ReLUs help prevent dead neurons and thus allow for more model complexities. More layers and more neurons in each layer increase model complexities and requires longer training times. Furthermore, parameters that define training strengths are also hyperparameters: regularization strength λ from section 3.4.1 and dropout probability P from section 3.4.2 are some examples.

As demonstrated, hyperparameters influence the training results significantly and thus must be carefully adjusted. The process of carefully adjusting hyperparameters to ensure the optimal training results is called hyperparameter tuning. This depends on the problem at hand and generally requires trial and error. Introducing suitable performance metrics facilitates the process, which is to be discussed more in chapter 5.

Hyperparameter tuning is often implemented within cross-validation loops to ensure enhanced performance. One instance is K-fold cross-validation. K-fold cross-validation starts with dividing the given set into K non-overlapping subsets, one of which is selected for validating and the rest for training. Repeating this training and validating procedure for each of the K subsets yields K different trained models, and the one with the least generalization error is selected. Validating a trained model means ensuring that its outputs match their targets on unseen samples in the given dataset. K-fold cross validation also helps detect overfitting: the loss functions of overfitting models evaluated on validation data diverge away from those evaluated on training data. This is because overfitting models are too distracted by the noises in the training data that they fail to generalize in the validation data. Furthermore, K-fold cross-validation itself has a regularizing effect because training and validating on each of the K partitioned datasets allows for more self-consistency by reducing model variance. [2]

Chapter 4

$t\bar{t}$ Backgrounds in HH Decays

Top pair decays $t\bar{t}$ and di-Higgs decays HH can have equivalent final states such as $b^+b^-W^+W^-$ with overlapping mass distributions. Such occurs when a $t\bar{t}$ event emits two W bosons in parallel, where one of the two W bosons has a significantly lower mass. As a result, $t\bar{t}$ events can act as backgrounds in HH events. This chapter elaborates on the methodology used to reach such a result.

4.1 $t\bar{t}$ and HH Decay Channels

Processes not explicitly forbidden by the conservation laws discussed in section 2.1.1 are allowed as per the totalitarian principle of particle physics. Therefore, particles can have multiple decay channels with different branching ratios. This section discusses $t\bar{t}$ and HH decay channels up to leading order (LO), and how the two can have equivalent final states.

4.1.1 $t\bar{t}$ Decay Channels

Top quarks are produced in pairs from hadronic collisions dominantly through quark pair annihilation $q\bar{q} \to t\bar{t}$ and gluon fusion $gg \to t\bar{t}$ at LO in QCD. Top quarks are preferably produced in pairs because single top productions yield fewer daughters and are thus harder to identify among large backgrounds. The $q\bar{q} \to t\bar{t}$ processes accounted for $\approx 85\%$ of the top quarks produced in proton anti-proton collisions $p\bar{p}$ at the Tevatron at the center of mass energy $\sqrt{s} = 1.96$ TeV. The $gg \to t\bar{t}$ processes accounts for $\approx 90\%$ of the top quarks produced in pp collisions at the LHC at $\sqrt{s} = 13$ TeV.

There are three possible categories of final states for the leading Top pair production processes, depending on how the daughter W boson pairs decay. The first is fully hadronic $t\bar{t} \to bW^+\bar{b}W^- \to bq\bar{q}'\bar{b}q''\bar{q}'''$ with a branching ratio of 45.3%. The second is semi-leptonic $t\bar{t} \to bW^+\bar{b}W^- \to bqq'\bar{b}l^-\bar{\nu}_l + bl^+\nu_l\bar{b}q''\bar{q}'''$ with 43.8%. The third is

di-leptonic $t\bar{t} \to bW^+\bar{b}W^- \to bl^+\nu_l\bar{b}l^-\bar{\nu}_l$ with 10.5%. Each lepton $l = e, \mu, \tau$ gives similar branching ratios for decays involving leptons as per the lepton universality. More quarks and gluons can be radiated from the colored particles in the $t\bar{t}$ decays and appear as jets. The $t\bar{t}$ decay modes are summarized in table (4.1), and a common $t\bar{t}$ process is depicted in figure (4.1).

Decay Channels	Branching Ratio (%)
$t\bar{t} \to bq\bar{q}'\bar{b}q''\bar{q}'''$	45.3
$t\bar{t} \to bqq'\bar{b}l^-\bar{\nu}_l + bl^+\nu_l\bar{b}q''\bar{q}'''$	43.8
$t\bar{t} \to bl^+\nu_l\bar{b}l^-\bar{\nu}_l$	10.5

Table 4.1: Top quark pair decay channels and their branching ratios, where $l = e, \mu, \tau$. Those with branching rations less than 1% are omitted. [12]

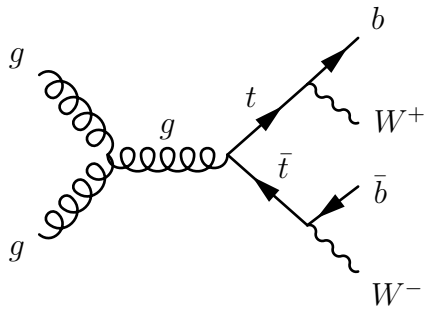


Figure 4.1: A Feynman diagram illustrating a Top pair production from gluon fusion as well as the Top pair decay into $bW^+\bar{b}W^-$.

4.1.2 HH Decay Channels

The Higgs boson dominantly decays into a bottom quark pair $H \to b\bar{b}$ with a branching ratio of 58%. The next likely fermionic final states are a tauon pair $H \to \tau^+\tau^-$ with a branching ratio of 6%, and a charm quark pair $H \to c\bar{c}$ with 3%. An excited higgs boson H^* could decay as $H^* \to W^+W^-$ or $H^* \to ZZ$, each with branching ratios 21% and 3%. The W^+ boson then decays leptonically into $W^+ \to l^+\nu_l$ with a 32% branching ratio in total. Hadronic decay channels for W^+ bosons vary, but their branching ratios sum up to 67%. The W^- decay channels are simply conjugates. The leptonic Z boson decays of the form $Z \to l^+l^-$ have a total of 10% branching ratio. Hadronic decay channels for Z bosons vary, but they have a 70% branching ratio in total.

These decays so far are all tree level, without any intermediate loops. However, higher order decays are also possible. Such could involve massless final states. For example, the Higgs boson could decay into a gluon pair $H \to gg$ with a branching ratio of 8%. Such is mostly mediated by a top quark loop, but this loop could instead be mediated

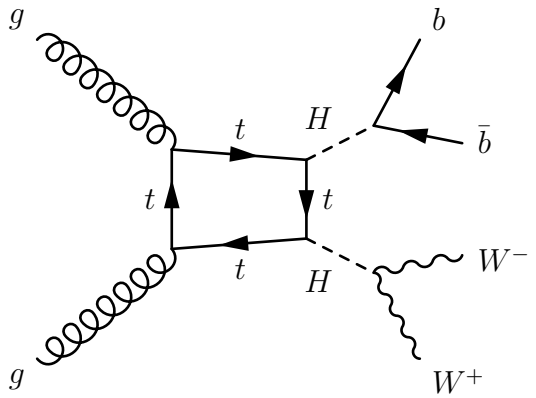


Figure 4.2: A Feynman diagram illustrating a Higgs pair production from gluon fusion through a top quark loop, as well as the di-Higgs decay into $b\bar{b}W^-W^+$.

by bottom and charm quarks 10% and 2% of the times respectively. The branching ratio increases by 70% from two loop QCD corrections at NLO. The Higgs boson could also decay into a photon pair $H \to \gamma\gamma$ with a branching ratio of 0.2%. Such is mostly mediated by W loops and less likely by t loops. The branching ratio increases by 2% from two loop QCD corrections at NLO. Though rare, $H^* \to \gamma\gamma$ along with $H \to ZZ \to l^+l^-l^+l^-$ served as the Higgs discovery channels since these were clearly identifiable among numerous backgrounds. The HH decay modes are summarized in table (4.2), and two notable HH processes are depicted in figures (4.2) and (4.3).

Comparing figures (4.2) and (4.3) shows that the final state alone is insufficient for determining whether a di-Higgs decay involves Higgs self interactions or not. Likewise, $t\bar{t}$ and HH decays can also have the similar final states such as $bW^+\bar{b}W^-$. Therefore, separating such events with similar final states requires additional measures, which are to be discussed in chapter 5. The separation can be greatly facilitated by using Monte Carlo methods such as MCatNLO and POWHEG, where fully exclusive predictions of SM processes can be made upto NLO in QCD. [12]

Decay Channel	Branching Ratio (%)
$H \to ZZ$	2.6
$H \to c\bar{c}$	2.9
$H \to \tau^+\tau^-$	6.2
$H \to gg$	8.2
$H^* \to W^+W^-$	21.5
$H \to b\bar{b}$	58.4

Table 4.2: Decay channels for Higgs bosons with $m_H = 125.1$ GeV with branching ratios greater than 1%. The rest are omitted. [12]

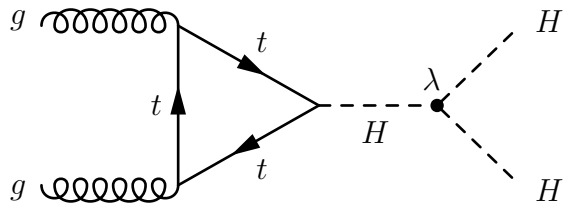


Figure 4.3: A Feynman diagram illustrating a Higgs production from gluon fusion through a top quark loop, as well as a trilinear Higgs self interaction denoted by its interaction strength λ.

4.2 Monte Carlo Event Generators

In practice, $t\bar{t}$ and HH decays are difficult to detect among large QCD backgrounds. Therefore, the use of Monte Carlo event generators (MCEGs) along with jet clustering algorithms (JCAs) can greatly facilitate the analysis on such events. Given the initial and the final particles before hadronization as inputs, MCEGs effectively simulate real high energy collisions in which the desired rare processes do occur with the irreducible backgrounds suppressed.

MCEGs do so in three steps. First, MCEGs initialize the input initial particles and their four momenta out of hard scattering processes and generate showers of particles including the input final particles. Second, MCEGs compute the four momenta and the masses of the particle showers using perturbative QCD based frameworks such as MadGraph [10] and MCatNLO [18]. This is possible because parton showers are closely packed so that strong interactions are still weak. MadGraph supports perturbations upto LO and MCatNLO upto NLO. MCEGs also compute cross sections upto this point using PDFs. Third, MCEGs hadronize quarks and gluons according to string models or cluster models.

String models are based on linear confinement: two partons with opposite color charges form the two ends of a flux tube, whose potential energy increases linearly with its length. Cluster models are based on preconfinement, where colorless combinations of partons form clusters of finite masses: the number of clusters are determined by their starting energy scales Q, while their mass distributions depend on their current energy scales Q_0 and the QCD ultraviolet cutoff Λ_{QCD}, where $Q >> Q_0 >> \Lambda_{\mathrm{QCD}}$. PYTHIA uses string models, while Herwig uses cluster models. Perturbative QCD no longer applies upon hadronization.

Throughout these three steps, MCEGs take collective QCD effects into account to better simulate physical events while suppressing irreducible backgrounds. One such example is color reconnection, where color fields in densely packed color systems recouple to nearby color fields. This effect helps model non-perturbative interactions between color fields during hadronization.

JCAs are then applied onto such events to make them more accessible for analysis. Depending on assumptions made about jets, there are largely two kinds of JCAs: cone algorithms (CAs) and sequential recombination algorithms (SRAs). CAs are built on the assumption that jets are cones with initial partons as their vertices and final particles as their bases. CAs recombine nearby particles in conic clusters, which form rigid circular boundaries in $\eta - \phi$ space. While easy to implement, CAs often result in infrared divergences. On the other hand, SRAs are difficult to implement but free of infrared divergences.

SRAs are built on the assumption that particles within jets vary in transverse momenta p_T, resulting in jets with fluctuating areas in $\eta - \phi$ space. As such, SRAs start with computing two kinds of distance measures in momentum space. The first kind d_{ij} is the distance between particles i and j, while the second kind d_{iz} is the distance between particle i and the beam axis $\hat{z}$. Specifically, these are defined as:

$$d_{ij} = \min(p_{Ti}^N, p_{Tj}^N)\frac{R_{ij}^2}{R}, \quad d_{iz} = p_{Ti}^N, \tag{4.1}$$

where R is the radius parameter that defines the final jet size, $R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ is the angular distance between the particles i and j, and N determines the SRA type; for example, the case $N = 2$ corresponds to the k_T algorithm. [3]

Upon computing all $\{d_{iz}, d_{ij}\}$, SRAs find the minimum distance $d = \min_{ij}\{d_{iz}, d_{ij}\}$ among them. If $d = d_{ij}$ for two particles i and j, then the two particles i and j are replaced with a pseudo jet, which is neither a particle nor a full jet. If $d = d_{iz}$ for some particle i, then the particle i is dropped. These pair replacement processes are repeated until certain stopping criteria are met. SRAs that stop once the minimum distance falls below a cutoff parameter d_C are said to be exclusive. SRAs that stop once no more particles remain are said to be inclusive. Inclusive SRAs however require a parameter $p_T^{\min}$ that defines the minimum transverse momenta for jets to be considered when computing $\{d_{iz}, d_{ij}\}$ in order to avoid infrared divergences in the number of jets found in an event.

JCAs assign four momenta to pseudo jets according to jet recombination schemes (RSs) for jet definition. The k_T algorithm for example is implemented in RSs such as the E_T scheme, which recombines the four momenta p_i and p_j of particles i and j into the four momenta p_r of the combined particle r such that

$$p_{Tr} = p_{Ti} + p_{Tj}, \quad \phi_r = \frac{E_{Ti}\phi_i + E_{Tj}\phi_j}{E_{Ti} + E_{Tj}}, \quad \eta_r = \frac{E_{Ti}\eta_i + E_{Tj}\eta_j}{E_{Ti} + E_{Tj}} \tag{4.2}$$

is satisfied. The subscripts T denote transverse variables perpendicular to the beam axis $\hat{z}$. k_T algorithms are best implemented by the FastJet package in the E_T scheme.

MCEGs and JCAs are often executed using ROOT, an object-oriented data analysis program developed by CERN. This yields event logs, which list information such as the identity, the four momenta, the parent, and the daughters of all particles according to the Particle Data Group (PDG) conventions. ROOT can then be used to analyze these events: ROOT offers math libraries that operate on a variety of classes such as the TLorentzVector class of four momenta, the TRandom class of random number generators, and the TH1D class of one dimesional histograms. [1] [16]

4.3 Low Mass W Bosons from $t\bar{t}$ Decays

Section 4.1 has demonstrated that $t\bar{t}$ and HH decays can have equivalent final states such as $b^+W^+\bar{b}W^-$, and section 4.3 aims to show that the two have overlapping mass distributions. This is kinematically possible when $t\bar{t}$ events each emit two W bosons in parallel, where one of the two W bosons has a significantly lower mass. This section first presents the kinematic reasoning behind this conclusion and then confirms that such is indeed the case.

4.3.1 Two Body Decays

Consider the decay of a particle with four momentum $p = (E, \vec{\mathbf{p}})$ into two particles 1 and 2 with four momenta $p_1 = (E_1, \vec{\mathbf{p}}_1)$ and $p_2 = (E_2, \vec{\mathbf{p}}_2)$ respectively. In the parent rest frame $p = (M, \vec{\mathbf{0}})$, four momenta conservation $p = p_1 + p_2$ yields the parent mass:

$$M^2 = (E_1 + E_2)^2 - (\vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2)^2 = M_1^2 + M_2^2 + 2(E_1E_2 - \vec{\mathbf{p}}_1 \cdot \vec{\mathbf{p}}_2), \tag{4.3}$$

where $M_1 = ||p_1||$ and $M_2 = ||p_2||$ are the daughter masses. Similarly, $p_1 = p - p_2$ yields the energy and the momentum of daughter 1 in the parent rest frame:

$$E_1 = \frac{M^2 - M_2^2 + M_1^2}{2M}, \quad p_1 = \frac{1}{2M}\sqrt{\lambda(M^2, M_1^2, M_2^2)}, \tag{4.4}$$

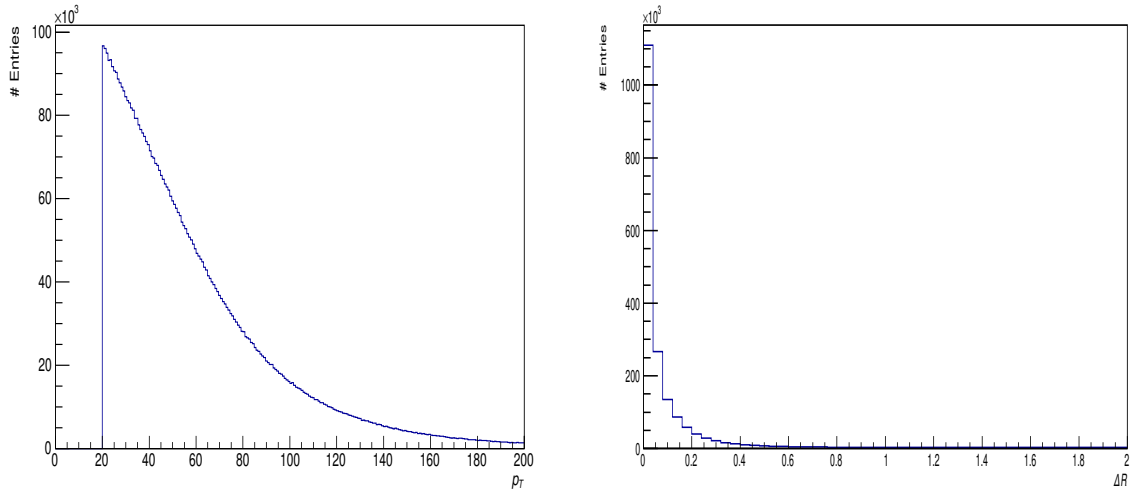
where $\lambda(\alpha, \beta, \gamma) = \alpha^2 + \beta^2 + \gamma^2 - 2\alpha\beta - 2\beta\gamma - 2\gamma\alpha$ is the Källén function. To obtain a similar equation for daughter 2, switch the subscripts 1 and 2 in equation (4.4).

Now consider the decay of an on shell Higgs boson into two W bosons. Since $M_H \approx 125$ GeV and $M_W \approx 80$ GeV, equation (4.3) can only hold if at least one of the W bosons is off shell with a lower mass. The low mass W boson has the maximum mass when the angle θ between $\vec{\mathbf{p}}_1$ and $\vec{\mathbf{p}}_2$ is zero, as the term $-\vec{\mathbf{p}}_1 \cdot \vec{\mathbf{p}}_2$ in equation (4.3) implies. This corresponds to the case where the two W bosons are emitted in parallel. Therefore, processes such as Top pair decays that emit two W bosons in the same way can act as backgrounds in Higgs decays since the mass distributions of the W boson pairs would overlap.

4.3.2 Overlapping Mass Distributions

The aim of this subsection is to show that $t\bar{t}$ and HH events do have overlapping mass distributions. To do so, PYTHIA was used to generate one million $t\bar{t}$ events with $bW^+\bar{b}W^-$ final states. MCatNLO was used to calculate the particle four momenta upto NLO along the way. FastJet was used to implement the k_T algorithm with $p_T^{\min} = 20$ GeV, $d_C = 400$ GeV^2, $R = 0.4$ in the E_T scheme. The validity of this choice is shown in figure (4.4). Figure (4.4a) shows the distribution of p_T for all jets. Since there would be too many jets satisfying $p_T \leq 20$, the choice $p_T^{\min} = 20$ GeV is justified. Figure (4.4b) shows the distribution of ΔR between all jets. Since most jet pairs satisfy $\Delta R = 0.4$, the choice $R = 0.4$ is justified. The choice $d_C = 400$ GeV$^2 = (p_T^{\min})^2$ was motivated from the definition $d_{iz} = p_{Ti}^N$ in equation (4.1).

Parent particles and jet pairs were then reconstructed using the event logs. Parents without jet daughters were reconstructed immediately from their daughters listed in the event logs. Parents with jet daughters were reconstructed by assigning to each jet daughter an inclusive jet that has the least difference in angular distance.



(a) (b)

Figure 4.4: Figure (4.4a) shows a histogram for the p_T distribution for all jets. Figure (4.4b) shows a histogram for the ΔR distribution for all jets. Both were obtained from the one million $t\bar{t}$ events.

Figure (4.5) shows a histogram of the W^+W^- and the $b\bar{b}$ pair masses from the $t\bar{t}$ events. The red data points in figure (4.5) indicate $t\bar{t}$ backgrounds in HH events, where both W^+W^- and $b\bar{b}$ pair masses are in the Higgs mass range, from 100 GeV to 150 GeV. There were only 519 such events among the million $t\bar{t}$ events. As shown in figure (4.5), $t\bar{t}$ backgrounds occur with such low probabilities because the daughter W^+W^- pair masses are unlikely to fall in the Higgs mass range, although the daughter $b\bar{b}$ pair

masses are much more likely to.

Though $t\bar{t}$ backgrounds occur with such low probabilities, they are still significant because the cross sections for $t\bar{t}$ decays are significantly larger than those for HH decays. $t\bar{t}$ decays have a cross section of $\sigma \approx 1\,\mu$b, while HH decays have $\sigma \approx 30$ fb. [12] Therefore, more events are needed to analyze $t\bar{t}$ backgrounds in HH decays. However, enough $t\bar{t}$ backgrounds cannot be generated within a reasonable computation time using MCEGs alone, since on average only ≈ 519 backgrounds appear for every million $t\bar{t}$ events. Generating enough $t\bar{t}$ backgrounds begins with finding their defining features.

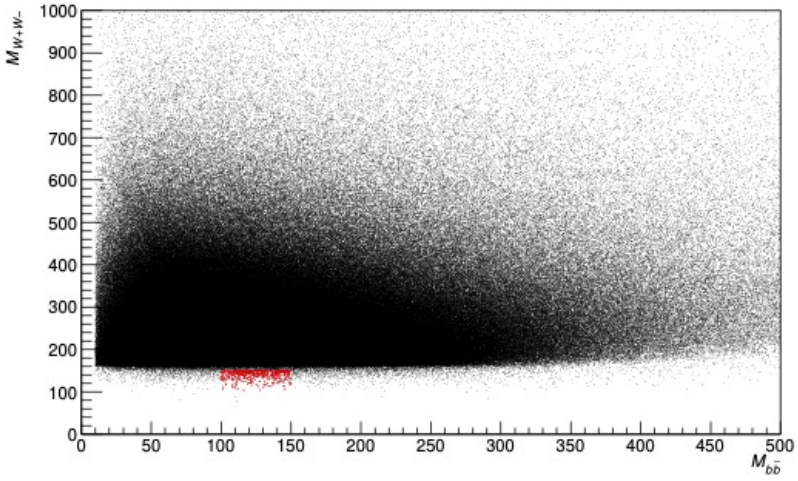
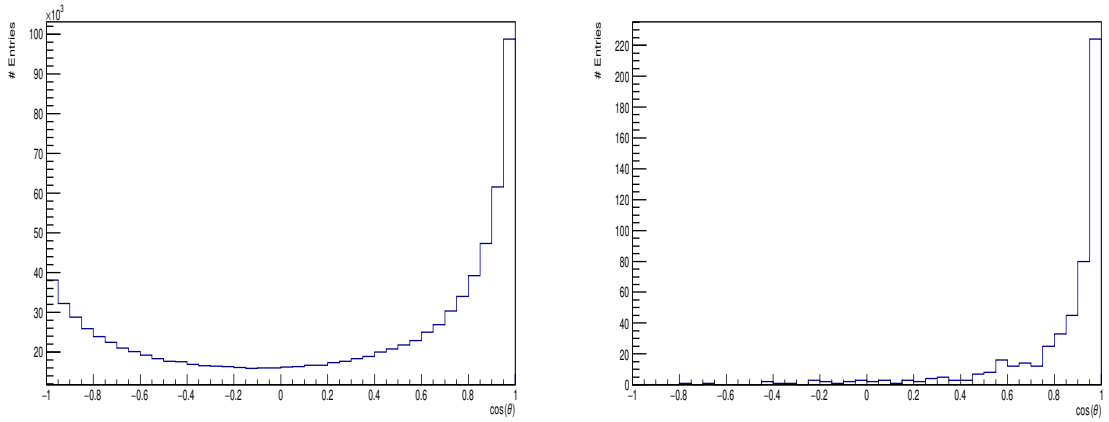


Figure 4.5: A two dimensional histogram of the W^+W^- and the $b\bar{b}$ pair masses from the one million $t\bar{t}$ events generated by PYTHIA. The data points highlighted in red are events where both $b\bar{b}$ and W^+W^- pair masses fall in the Higgs mass range, from 100 GeV to 150 GeV.

4.3.3 Low Mass W Bosons

The aim of this subsection is to discuss the defining features for $t\bar{t}$ backgrounds in HH events. To do so, the kinematic observables such as p_T, ϕ, η, E for all particles from $t\bar{t}$ events were analyzed. The kinematic observables for all particles from $t\bar{t}$ backgrounds in HH events were analyzed separately. Comparing the two analyses has shown that the angle θ between two daughter W bosons is the defining feature for $t\bar{t}$ backgrounds, other than the masses of the W bosons. This is not so obvious from figure (4.6) alone: figure (4.6a) shows the histogram of $\cos\theta$ for all $t\bar{t}$ events, and figure (4.6b) shows that of $\cos\theta$ for just the $t\bar{t}$ backgrounds. The two look similar, with most entries present in the bin with $\cos\theta = 1$. However, dividing the two histograms yields figure (4.7), which shows that $t\bar{t}$ events that emit two W bosons in parallel are overwhelmingly more likely to act as backgrounds in HH events. For example, the case $\cos\theta = 1$ occurs ≈ 10 times more likely than the case $\cos\theta = -1$ in $t\bar{t}$ backgrounds.

To summarize, three conditions must be met in order for a $t\bar{t}$ event to mimic HH events. First, one of the two daughter W bosons must be off shell, with lower mass. Second, the low mass W boson must be emitted in parallel to the other daughter W boson. Third, the daughter $b\bar{b}$ pair masses must also fall in the Higgs mass range.



(a) (b)

Figure 4.6: In figure (4.5a) is a histogram of $\cos\theta$ for all $t\bar{t}$ events, with θ the angle between the three momenta of two daughter W bosons. In figure (4.6b) is a histogram of $\cos\theta$ for just the $t\bar{t}$ backgrounds in HH events. These histograms were obtained from the one million $t\bar{t}$ events.

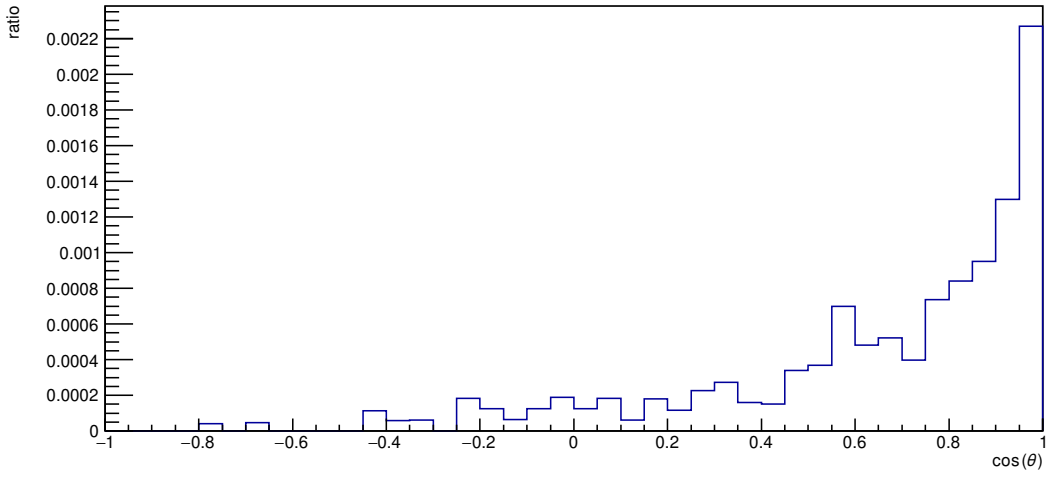


Figure 4.7: The histogram obtained by dividing the left histogram in figure (4.6) by the one on the right. This shows that $t\bar{t}$ backgrounds emit W bosons in parallel.

Chapter 5

Separating $t\,\bar{t}$ Backgrounds from HH Events

As shown in the previous chapter, $t\bar{t}$ events that emit a low mass W boson in parallel to the other daughter W boson can mimic HH events. This chapter first elaborates on how such $t\bar{t}$ backgrounds in HH events can be generated in large numbers by rotating and scaling $t\bar{t}$ events. This chapter then discusses how the HH events were procured with Gaussian smearing. This chapter finishes with how the $t\bar{t}$ backgrounds can be separated from HH events using FFNNs.

5.1 Rotating and Scaling $t\,\bar{t}$ Events

Realistic $t\bar{t}$ backgrounds in HH events can be mass produced by rotating and scaling the four momenta of low mass W bosons from $t\bar{t}$ events. This process of rotating and scaling shall be called the rotation scaling procedure (RSP). Applying the RSP to a $t\bar{t}$ event with at least one low mass W boson takes the following seven steps.

First, the whole $t\bar{t}$ event are boosted into the $t\bar{t}$ rest frame using equation (2.36). This step ensures that there are no excess transverse momenta within the $t\bar{t}$ event. Second, the t and the $\bar{t}$ events are boosted into the t and the $\bar{t}$ rest frames respectively using equation (2.36). This step boosts the t and the $\bar{t}$ events into the reference frame where the two events are back to back. Third, a daughter low mass W event is rotated so that the three momenta of the two daughter W bosons are parallel. This step ensures that the two daughter W bosons are parallel in a common reference frame. Fourth, the mass of the low mass W boson is scaled by a factor of 7/16, which was chosen so that the $b\bar{b}$ and the W^+W^- pair masses peak at the Higgs mass $M_H \approx 125$ GeV upon applying the RSP. This step breaks the four momenta conservation in low mass W events. Fifth, the low mass W event is boosted into the rest frame of the low mass W

boson using equation (2.36). This step is needed since applying equation (4.4) requires that the parent be in its rest frame. Sixth, now that the low mass W boson is in its rest frame, the four momenta of its daughters are scaled according to equation (4.4). This step restores the four momenta conservation in low mass W events. Seventh, undo all boosts in reverse order to return back to the lab frame. This step completes the RSP. The RSP simply discards the $t\bar{t}$ events without low mass W bosons.

The RSP was applied onto the one million $t\bar{t}$ events used in chapter 4, yielding figures from (5.1) to (5.4). Figure (5.1) shows a histogram of W^+W^- pair masses before the RSP in blue and after the RSP in red, with the vertical axis in logarithmic scale; the green histogram is obtained by multiplying a factor of 200 to the blue histogram, effectively corresponding to the W^+W^- pair mass distribution of 200 million $t\bar{t}$ events before applying the RSP. The green histogram matches the red histogram up to the pair mass of ≈ 125 GeV. This follows from the self similarity of GBWDs: the Gaussian factor in equation (2.43) implies that scaling the W boson masses by a factor less than 1 shifts the W^+W^- pair mass distributions to the left. Therefore, applying the RSP has the effect of processing 200 times as many unmodified $t\bar{t}$ events upto the pair mass of ≈ 125 GeV, when it comes to producing $t\bar{t}$ events with low mass W bosons.

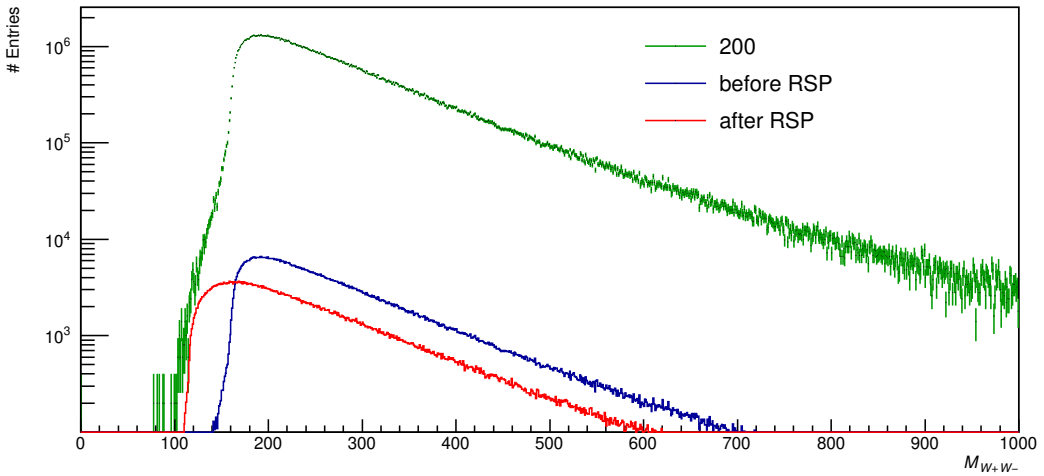


Figure 5.1: The histograms of W^+W^- pair masses before the RSP in blue and after the RSP in red. In green is the histogram obtained by multiplying a factor of 200 to the blue histogram. The vertical axis in logarithmic scale.

Next, figure (5.2) shows a histogram of $b\bar{b}$ pair masses before the RSP in blue and after the RSP in red. The red histogram shows that the number of $t\bar{t}$ events with daughter $b\bar{b}$ pair masses in the Higgs mass falls to $\approx 62\%$ after the RSP. However, comparing figures (5.1) and (5.2) shows that there are still enough of $b\bar{b}$ pairs in the Higgs mass range compared to W^+W^- pairs across all $t\bar{t}$ events. Furthermore, both red histograms in figures (5.1) and (5.2) peak at the pair masses of ≈ 125 GeV because the masses of

the low mass W bosons were scaled by a factor of 7/16 in the fourth step of the RSP. The effectiveness of the RSP in generating $t\bar{t}$ backgrounds is visualized in figure (5.3).

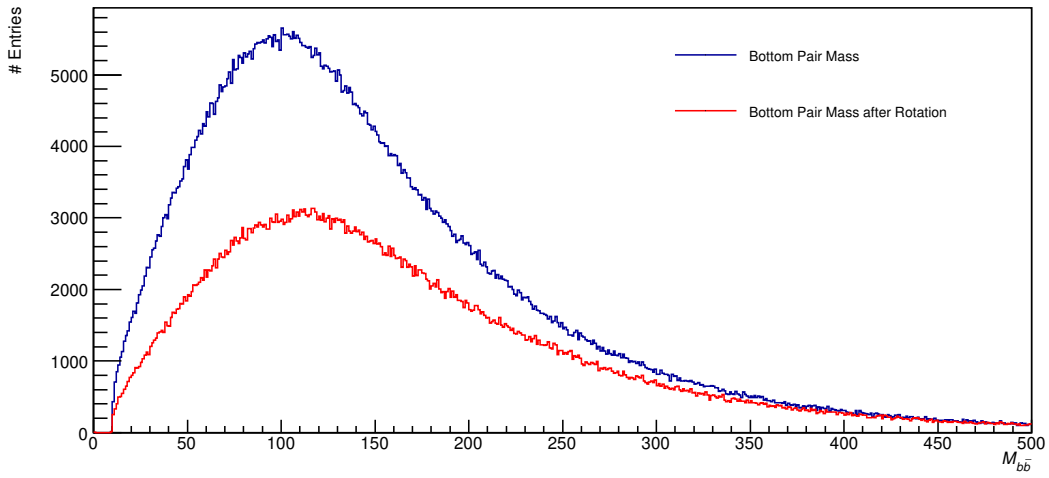


Figure 5.2: The histograms of $b\bar{b}$ pair masses before the RSP in blue and after the RSP in red.

Figure (5.3) shows a histogram of the W^+W^- and the $b\bar{b}$ pair masses upon applying the RSP. The red data points in figure (5.3) indicate $t\bar{t}$ backgrounds in HH events, where both W^+W^- and $b\bar{b}$ pair masses are in the Higgs mass range. There were 28441 such events among the million $t\bar{t}$ events, demonstrating that the RSP has increased the likelihood of producing $t\bar{t}$ backgrounds by a factor of ≈ 55, which is still less than the ≈ 200 times increase in the likelihood of finding $t\bar{t}$ events with low mass W bosons. This disparity occurs because not all $t\bar{t}$ events with low mass W bosons have both W^+W^- and $b\bar{b}$ pair masses that fall in the Higgs mass range. Nonetheless, comparing the density of the red data points in figures (4.5) and (5.3) demonstrate the effectiveness of the RSP in generating $t\bar{t}$ backgrounds.

Figure (5.4) shows the histograms of the Top quark masses reconstructed from their daughters: the blue histogram shows the Top quark mass distributions before applying the RSP, and the red histogram shows the Top quark mass distributions after applying the RSP. Applying the RSP has left the mean unchanged at 170.8 GeV and has slightly reduced the standard deviation from 8.989 to 8.982 GeV. In other words, $t\bar{t}$ events that emit low mass W bosons do not necessarily involve Top quarks with lower masses.

Next, PYTHIA was used to generate 10 million $t\bar{t}$ events. [6] Along the way, POWHEG was used to calculate particle four momenta upto NLO. FastJet was used to implement the k_T algorithm with $p_T^{\min} = 20$ GeV, $d_C = 400$ GeV^2, $R = 0.4$ in the E_T scheme. Finally, the RSP was then applied to obtain new 10 million $t\bar{t}$ events, which includes 300,053 $t\bar{t}$ backgrounds in HH events. Procuring the HH events follows next.

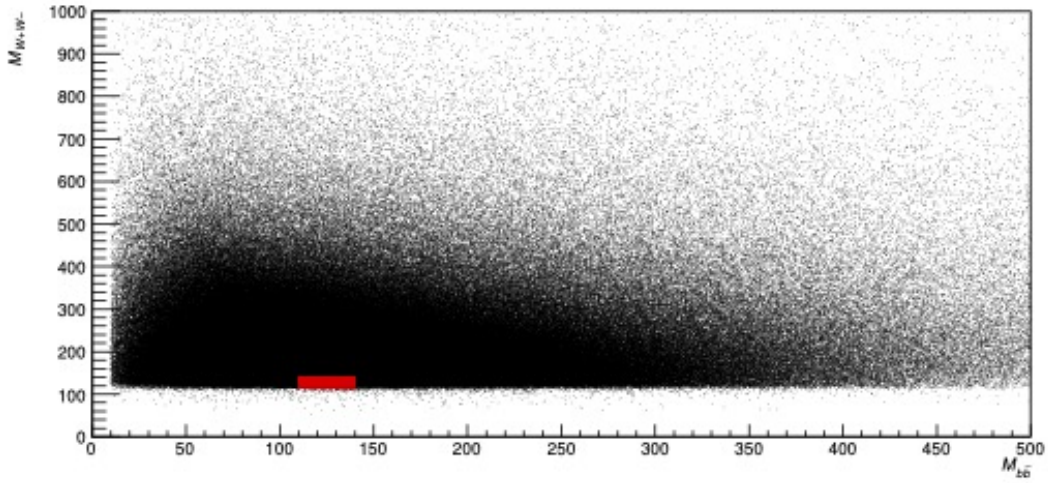


Figure 5.3: A two dimensional histogram of the W^+W^- and the $b\bar{b}$ pair masses from the one million $t\bar{t}$ events after the RSP. The data points highlighted in red are $t\bar{t}$ events where both W^+W^- and $b\bar{b}$ pair masses fall in the Higgs mass range, from 100 GeV to 150 GeV.

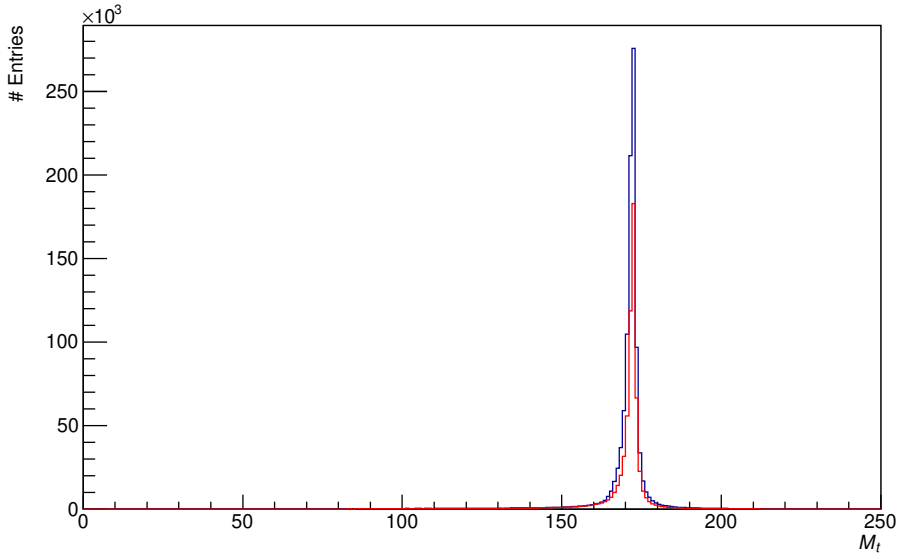


Figure 5.4: The reconstructed Top quark mass histograms from the one million $t\bar{t}$ events. In red is the histogram before applying the RSP, and in blue is the histogram after applying the RSP.

5.2 Gaussian Smearing on HH Events

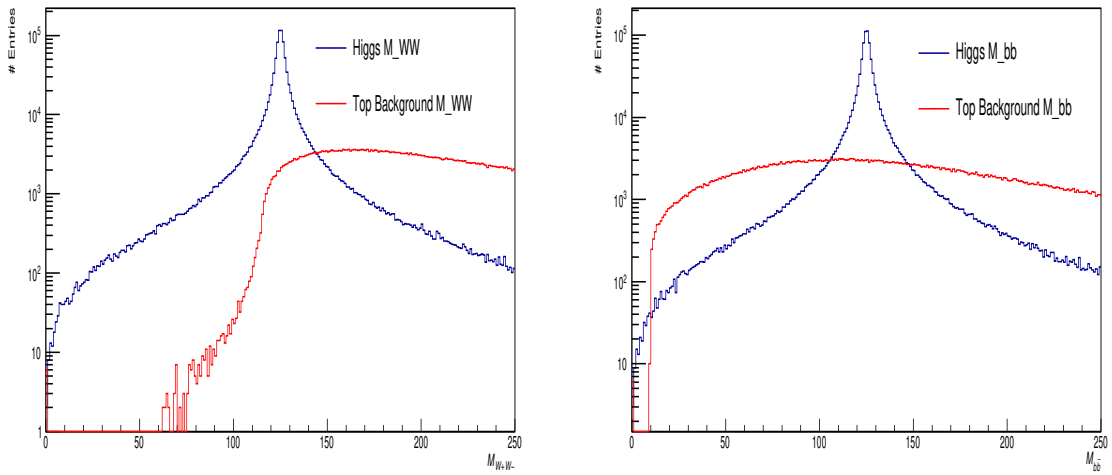
Higgs bosons do not have color charges. Therefore, the decays of Higgs bosons with quarks or gluons in the final states have excess color charged fields that recouple to other nearby color charged fields as per color reconnection discussed in subsection 4.2. This complicates the reconstruction of particles from their daughters, which was not the case for $t\bar{t}$ events because Top quarks themselves have color charges. To resolve

this complication with color reconnection, Herwig was used to generate one million HH events with $b\bar{b}W^+W^-$ final states with color reconnection settings off. [17] Madgraph was used along the way to calculate particle four momenta upto LO. FastJet was used to implement the k_T algorithm with $p_T^{\min} = 20$ GeV, $d_C = 400$ GeV^2, $R = 0.4$ in the E_T scheme: these parameter choices are again justified by the same reason illustrated in figure (4.4).

However, Herwig takes into account the natural decay width $\Gamma_H \approx 4$ MeV of Higgs bosons while generating events. [12] Such decay widths are much smaller than the energy scales relevant for this thesis. Therefore, the decay widths of the HH decay products must be broadened with Gaussian distributions to reflect their experimental resolutions. Such a process is called Gaussian smearing (GS).

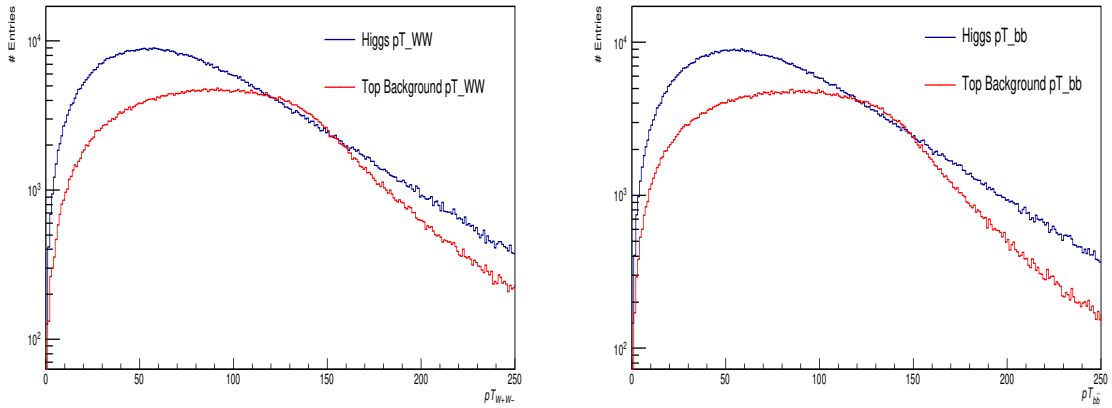
For simplicity, GS was applied only onto the direct daughters $b\bar{b}W^+W^-$ of the HH events to reflect their experimental uncertainties, realistically ranging from ≈ 10 to ≈ 20 GeV. This proceeded as follows: first, a Gaussian distribution was initialized for each component of all $b\bar{b}W^+W^-$ four momenta, and a random number was selected from each Gaussian distributions using the Gaus method of the TRandom3 class. Then, such randomly selected numbers were scaled by their GS factors. These GS factors were added to 1, and the results were finally multiplied to each component of all $b\bar{b}W^+W^-$ four momenta. The GS factors were chosen such that the reconstructed Higgs mass distributions had means at ≈ 125 GeV with standard deviations ranging from ≈ 10 to ≈ 20 GeV.

As a result, the GS factor of 1% was each applied to the p_x, p_y, E components of the $b\bar{b}W^+W^-$ four momenta, and the GS factor of 2% was each applied to the p_z components in order to account for the natural difficulty of measuring variables along the beam axis $\hat{z}$. The kinematic observables associated to the Higgs bosons from their $b\bar{b}W^+W^-$ daughters after GS are shown in figures (5.5) to (5.7) in blue, along with the kinematic observables associated to the $t\bar{t}$ events after the RSP in red. Figure (5.5) shows the W^+W^- pair masses on the left and the $b\bar{b}$ pair masses on the left. Figure (5.6) shows the W^+W^- pair transverse momenta on the left and the $b\bar{b}$ pair transverse momenta on the left. Figure (5.7) shows the W^+W^- pair pseudorapidity on the left and the $b\bar{b}$ pair pseudorapidity on the left. The pair polar angles were distributed homogenously and thus were omitted. This shows that the daughter pairs from both the HH events and the $t\bar{t}$ backgrounds have the same angular distributions. Figures (5.5) to (5.7) show that there is a significant overlap between $t\bar{t}$ backgrounds and HH events in the phase space of kinematic observables. This necessitates their comparison using neural networks, which follows next.



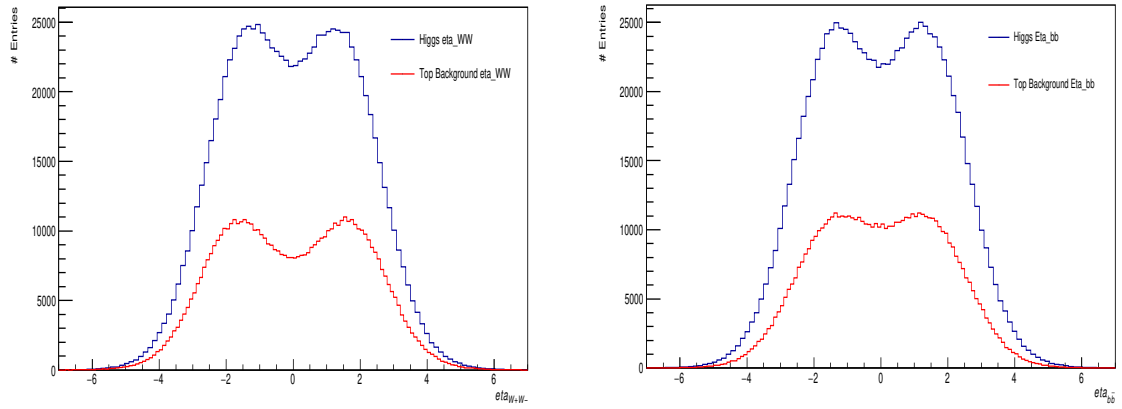
(a) (b)

Figure 5.5: Shown in figure (5.5a) in blue is the W^+W^- pair mass histogram from the one million HH events. Shown in figure (5.5a) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds. Shown in figure (5.5b) in blue is the $b\bar{b}$ pair mass histogram from the one million HH events. Shown in figure (5.5b) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds. The vertical axis of all histograms in figure (5.5) are drawn in logarithmic scale.



(a) (b)

Figure 5.6: Shown in figure (5.6a) in blue is the W^+W^- pair transverse momentum histogram from the one million HH events. Shown in figure (5.6a) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds. Shown in figure (5.6b) in blue is the $b\bar{b}$ pair transverse momentum histogram from the one million HH events. Shown in figure (5.6b) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds. The vertical axis of all histograms in figure (5.6) are drawn in logarithmic scale.



(a) (b)

Figure 5.7: Shown in figure (5.7a) in blue is the W^+W^- pair polar angle histogram from the one million HH events. Shown in figure (5.7a) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds. Shown in figure (5.7b) in blue is the $b\bar{b}$ pair polar angle histogram from the one million HH events. Shown in figure (5.7b) in red is the one from the ≈ 0.3 million $t\bar{t}$ backgrounds.

5.3 Comparison Using Neural Networks

The objective of this section is to elaborate on the separation of $t\bar{t}$ backgrounds from HH events using FFNNs. Such begins with describing the performance measures used to evaluate the NN training results. The specifics of the NN and the input datasets are then presented. The analysis on the NN's success in separating the $t\bar{t}$ backgrounds in HH events are finally elaborated.

5.3.1 Performance Measures

Given a classification problem with the target class called positive (P) and the others called negative (N), a NN can produce outputs that fall into precisely one of the four following cases. The first case is true positive (TP), where the NN correctly outputs P. The second case is true negative (TN), where the NN correctly outputs N. The third case is false positive (FP), where the NN incorrectly outputs P. The fourth case is false negative (FN), where the NN incorrectly outputs N. This categorization can be used to define performance measures that evaluate the ability of a NN to solve classification problems. One such performance measure is accuracy, defined as:

$$\text{Accuracy} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN}, \tag{5.1}$$

representing the fraction of correctly identified outputs; the # in front of each case denotes its number of occurrence. Given a properly trained NN, accuracy tends to rise

steeply at earlier epoches and stagnates at later epoches. However, accuracy is blind to overfitting because it simply stagnates or slightly oscillates as loss functions increase in value. Therefore, training methods like cross validation along with other performance measures like loss functions are still necessary.

Other performance measures include precision, recall, and specificity, each defined as:

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP} \tag{5.2}$$

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN} \tag{5.3}$$

$$\text{Specificity} = \frac{\#TN}{\#TN + \#FP}. \tag{5.4}$$

Precision represents the fraction of outputs correctly identified as P among all outputs identified as P. Recall represents the fraction of outputs correctly identified as P among all outputs that should be identified as P. Specificity represents the fraction of outputs correctly identified as N among all outputs that should be identified as N. Though not as judgemental as accuracy is, these are needed for a complete evaluation of the NN performance. For example, these three performance measures can be used for further evaluation once desired losses and accuracies are reached.

5.3.2 Set Up

There were two input datasets. One consisted of ≈ 0.3 million $t\bar{t}$ backgrounds selected from the ten million $t\bar{t}$ events with the RSP applied. The other consisted of one million HH events with the GS applied. The four kinematic observables M, p_T, η, ϕ for each W^+W^- and $b\bar{b}$ were given as input features, so the two datasets were identified with eight features in total. The ideal input features would be M, p_T, η, ϕ for all daughters $b\bar{b}W^+W^-$ such that the exact configurations of $t\bar{t}$ backgrounds and HH events can be compared. However, this was not possible because PYTHIA and Herwig initializes the particle masses in a different way so that the NN can immediately tell apart the $t\bar{t}$ backgrounds and HH events. Regardless, separating the $t\bar{t}$ backgrounds and HH events apart using NNs reduces to a binary classification task: for this task, the $t\bar{t}$ backgrounds were given labels 0, and the HH events were given labels 1. Validation splits of 0.2 were given to test different NN specifics. Testing different NN specifics has shown that the best performing NN was chosen for the classification task. Such analysis took place using a Python framework called PyTorch. [14] The specifics of the best performing NN are described as follows.

The NN used for this analysis was a FFNN with five layers. The non-output layer had eight neurons and LReLU activation functions with slope 0.1, while the output layer had two neurons and Sigmoid activation functions. He-initialization was used for

the non-output layer weights, and Xavier initialization was used for the output layer weights. Cross entropy loss functions were used. Dropout regularization with dropout ratio 5% was used. The NN was trained over 200 epoches, with batch sizes of 700 and learning rates of 5×10^{-5} in order to solve the binary classification problem of separating $t\bar{t}$ backgrounds from HH events.

5.3.3 Analyzing the Training Results

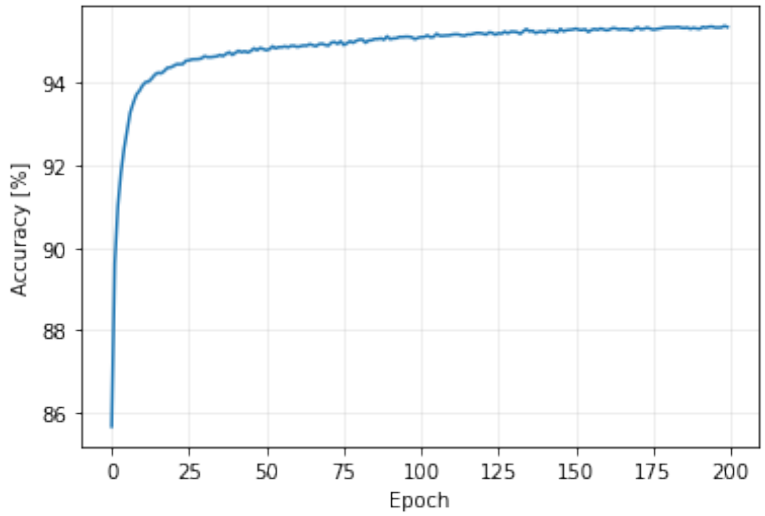


Figure 5.8: The training accuracy in separating the two datasets, as a function of 200 epoches.

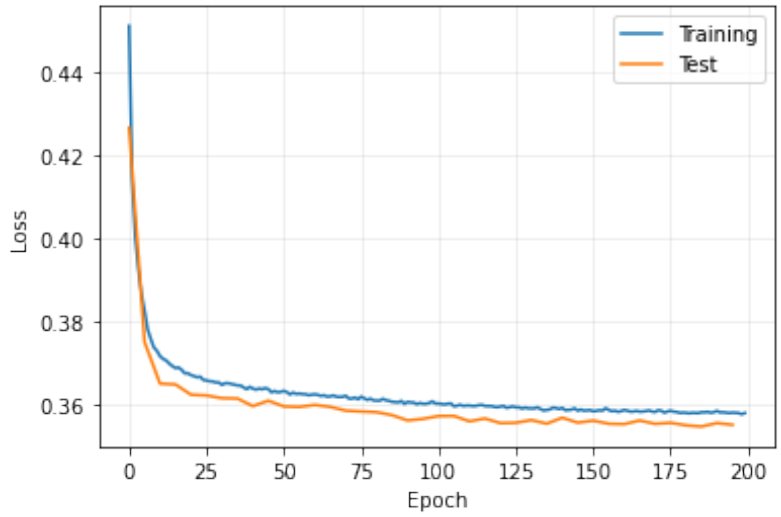


Figure 5.9: The loss functions evaluated on the training datasets in blue the loss functions evaluated on the validation datasets in yellow, as functions of epoches.

The NN was trained with the specifics detailed in the previous subsection. The results are shown in figures (5.8) and (5.9). Figures (5.8) shows the training accuracy as a function of epoches: over 200 epoches, the NN has reached a separation accuracy

of 95%, which is well above the separation accuracies of $\approx 70\%$ obtained in similar papers. The training accuracy plateaus over epoches larger than 200 and stays at $\approx 94\%$. Figure (5.9) compares the loss functions evaluated on the training datasets to the loss functions evaluated on the validation datasets: the convergence of the two leads to the conclusion that the NN was properly regularized. The plateauing of the training accuracy as well as the bottoming of the loss functions over larger epoches justify that the separation accuracy of 95% is not a mistake.

However, such a high separation accuracy is not satisfactory because the pair mass and the transverse momentum distributions of the two datasets as shown in figures (5.5) and (5.6) are easily distinguishable even to the human eye. One possible cause for this noticeable difference in such pair distributions is the application of GS onto the HH events: the RSP could have been applied instead so that the daughter $b\bar{b}W^+W^-$ masses on the two datasets match. Another possible cause for such a noticeable difference could have been reduced by carefully choosing the RSP scaling factors, the GS factors, and the Higgs mass range. This may be minor but could lead to at least a good match in the distributions from the two datasets when restricting to the Higgs mass range.

Chapter 6

Conclusion

While $t\bar{t}$ decays are rare, their large cross sections contribute significant backgrounds to the HH decays. Separating $t\bar{t}$ backgrounds is therefore necessary when analyzing Higgs self interactions among numerous backgrounds. The separation process corresponds to solving binary classification problems using neural networks. As such, this thesis has successfully separated the $t\bar{t}$ backgrounds in HH decays using feed forward neural networks with a high training accuracy of 95% in separating $t\bar{t}$ backgrounds from HH. The separation process has exploited event modification methods such as the rotation scaling procedure based on the decay kinematics. However, the training results are not completely reliable because of the different mass settings involved in generating input datasets using Monte Carlo event generators. Attempts have been made to reduce such differences in mass settings by introducing Gaussian smearing. However, this alone was not enough since the mass distributions and the transverse momentum distributions were noticeably different, even to the human eye. This issue could have been avoided completely eliminating the differences in settings by exploiting the rotation procedure. This issue could have been avoided at least in some peak around the Higgs mass by carefully choosing the parameters of the rotation scaling procedure and of Gaussian smearing. However, these approaches could not be implemented in time to obtain a more reliable separation accuracy.

Bibliography

[1] https://fastjet.fr/repo/fastjet-doc-3.4.2.pdf. [Accessed 14-06-2024].

[2] Charu C Aggarwal. *Neural networks and deep learning*. en. 2nd ed. Cham, Switzerland: Springer International Publishing, June 2023.

[3] Ryan Atkin. "Review of jet reconstruction algorithms". In: *J. Phys. Conf. Ser.* 645 (Oct. 2015), p. 012008.

[4] *ATLAS Experiment at CERN — ATLAS Experiment at CERN — atlas.cern*. https://atlas.cern. [Accessed 14-06-2024].

[5] G Breit. "Theory of resonance reactions and allied topics". In: *Nuclear Reactions II: Theory / Kernreaktionen II: Theorie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1959, pp. 1–407.

[6] *Documentation - PYTHIA 8.3 — pythia.org*. https://pythia.org/documentation/. [Accessed 14-06-2024].

[7] D. Griffiths. *Introduction to Elementary Particles*. Physics textbook. Wiley, 2008. ISBN: 9783527618477.

[8] Arnulf Jentzen and Adrian Riekert. "A proof of convergence for stochastic gradient descent in the training of artificial neural networks with ReLU activation for constant target functions". In: (Apr. 2021). arXiv: 2104.00277 [math.NA].

[9] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. eprint: arXiv:1412.6980.

[10] *MadGraph Home Page*. http://madgraph.phys.ucl.ac.be. Accessed: 2024-6-14.

[11] Bernhard Mehlig. "Artificial Neural Networks". In: *CoRR* abs/1901.05639 (2019). arXiv: 1901.05639.

[12] *Particle Data Group — pdg.lbl.gov*. https://pdg.lbl.gov/index.html. [Accessed 14-06-2024].

[13] M.E. Peskin. *An Introduction To Quantum Field Theory*. CRC Press, 2018. ISBN: 9780429983184.

[14] *PyTorch documentation &x2014; PyTorch 2.3 documentation — pytorch.org*. https://pytorch.org/docs/stable/index.html. [Accessed 14-06-2024].

[15] M.D. Schwartz. *Quantum Field Theory and the Standard Model*. Quantum Field Theory and the Standard Model. Cambridge University Press, 2014. ISBN: 9781107034730.

[16] ROOT team. *ROOT Manual — root.cern*. https://root.cern/manual/. [Accessed 14-06-2024].

[17] *The Herwig Event Generator &x2014; Herwig 7.2 documentation &x2013; Hepforge — herwig.hepforge.org*. https://herwig.hepforge.org. [Accessed 14-06-2024].

[18] Contact Us. *Theoretical High Energy Physics Group - Software - MC@NLO*. en. https://www.hep.phy.cam.ac.uk/theory/webber/MCatNLO/. Accessed: 2024-6-14.

Selbständigkeitserklärung

Ich versichere hiermit, die vorliegende Arbeit mit dem Titel

Separation von $t\bar{t}$ und HH Endzuständen mit Neuronalernetze

selbständig verfasst zu haben und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben.

Youn Jun Cho

München, den 14. Jun 2024

Declaration of Academic Integrity

I hereby declare that the following thesis with the title

Separating $t\bar{t}$ and HH end states using neural networks

is my own work, and that I have not used any sources and aids other than those stated in the thesis.

Youn Jun Cho

Munich, on 14. Jun 2024