SEARCH FOR THE $t\bar{t}HH(b\bar{b}b\bar{b})$ NONRESONANT PRODUCTION IN THE LEPTONIC FINAL STATES USING MACHINE LEARNING TECHNIQUES AT THE CMS EXPERIMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GAMZE SÖKMEN ŞAHİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
PHYSICS

JANUARY 2024

Approval of the thesis:

SEARCH FOR THE $t\bar{t}HH(b\bar{b}b\bar{b})$ NONRESONANT PRODUCTION IN THE LEPTONIC FINAL STATES USING MACHINE LEARNING TECHNIQUES AT THE CMS EXPERIMENT

submitted by **GAMZE SÖKMEN ŞAHİN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Physics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Seçkin Kürkçüoğlu
Head of Department, **Physics** _____

Prof. Dr. Mehmet Tevfik Zeyrek
Supervisor, **Physics, METU** _____

Examining Committee Members:

Prof. Dr. Meltem Serin
Physics, METU _____

Prof. Dr. Mehmet Tevfik Zeyrek
Physics, METU _____

Prof. Dr. Altuğ Özpineci
Physics, METU _____

Prof. Dr. Bora Işıldak
Physics, Yıldız Teknik University _____

Prof. Dr. Muhammed Deniz
Physics, Dokuz Eylül University _____

Date:26.01.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Gamze Sökmen Şahin

Signature :

ABSTRACT

SEARCH FOR THE $t\bar{t}HH(b\bar{b}b\bar{b})$ NONRESONANT PRODUCTION IN THE LEPTONIC FINAL STATES USING MACHINE LEARNING TECHNIQUES AT THE CMS EXPERIMENT

Sökmen Şahin, Gamze
Ph.D., Department of Physics
Supervisor: Prof. Dr. Mehmet Tevfik Zeyrek

January 2024, 182 pages

This thesis presents a search for the production of a top quark-antiquark pair associated with a pair of Higgs bosons in both semileptonic and dileptonic final states using machine learning techniques. The candidate $t\bar{t}HH$ events are selected with criteria both targeting a lepton and jets decay channels and two leptons and jets decay channels of the $t\bar{t}$ system and the decay of the double Higgs bosons into two bottom quark-antiquark pairs. The dilepton (DL) study is performed for the first time by using the proton-proton collision data collected between the years 2016 and 2018 at the CERN Large Hadron Collider (LHC) in the Compact Muon Solenoid (CMS) experiment at a center-of-mass energy of $\sqrt{13}$ TeV. The semileptonic (SL) study is also performed for the first time with the upgraded CMS detector at the CERN High-Luminosity(HL)-LHC using proton-proton collisions at a center-of-mass energy of $\sqrt{14}$ TeV by using simulated samples. In order to increase the sensitivity of both searches, selected events are fed into a multi-classifier deep neural network. For the SL channel, the discriminant outputs of the DNN are split into several b jet multiplicity categories with different expected signal and background rates.

A simultaneous maximum likelihood fit is performed to evaluate the expected sensitivity reach for each channel. For the Run 2 study in the DL channel, no deviation from the background-only hypothesis is observed. A 95% confidence level upper limit on the $\mathrm{t\bar{t}HH}$ production cross section is observed at 94.23 times the standard model (SM) prediction for an expected value of 69.25 for the collision data collected at an integrated luminosity of 41.5 fb^{-1}. The HL-LHC study in the SL channel is expected to exclude $\mathrm{t\bar{t}HH}$ production down to 3.14 times the SM cross section with 3000 fb^{-1} of data. The sensitivity for Minimal Composite Higgs Model scenarios is also presented.

Keywords: LHC, CMS, Higgs boson, top quark, leptonic

ÖZ

CMS DENEYİNDE MAKİNE ÖĞRENİMİ TEKNİKLERİ KULLANILARAK LEPTONİK SON DURUMLARDA REZONANT OLMAYAN $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$ ÜRETİMİNİN ARAŞTIRILMASI

Sökmen Şahin, Gamze
Doktora, Fizik Bölümü
Tez Yöneticisi: Prof. Dr. Mehmet Tevfik Zeyrek

Ocak 2024 , 182 sayfa

Bu tez, hem yarı leptonik hem de iki leptonlu son durumlar kullanılarak bir top kuark-antikuark çifti ile bir çift Higgs bozonunun üretiminin araştırılmasını sunmaktadır. Makine öğrenimi teknikleri kullanılarak yapılan bu araştırmada, aday $\mathrm{t\bar{t}HH}$ olayları, $\mathrm{t\bar{t}}$ sisteminin bir lepton ile çoklu jetlere bozunma kanalı ve iki lepton ile çoklu jetlere bozunma kanalını hedefleyen kriterlerle ve çift Higgs bozonunun iki b kuark-antikuark çiftine bozunması ile seçilmiştir. Çift lepton kanalındaki (DL) çalışma, CERN Büyük Hadron Çarpıştırıcısı'nda (LHC) Kompakt Muon Solenoidi (CMS) deneyinde, 2016 ve 2018 yılları arasında toplanan proton-proton çarpışma verileri kullanılarak ilk kez gerçekleştirilmiştir ve merkez-kütle enerjisi $\sqrt{13}$ TeV'dir. Yarı leptonik (SL) çalışma da, CERN Yüksek-Işıklılık HL-LHC'de geliştirilmiş CMS dedektörü ile proton-proton çarpışmaları kullanılarak $\sqrt{14}$ TeV merkez-kütle enerjisinde ve simüle edilmiş örneklerle ilk kez gerçekleştirilmiştir. Her iki aramanın hassasiyetini artırmak için, seçilen olaylar çoklu-sınıflandırıcı derin sinir ağı (DNN) ile analiz edilmiştir. SL kanalı için, DNN'nin ayırıcı çıktıları, farklı beklenen sinyal ve arka plan oranlarına

sahip birkaç b jet çoğulluk kategorisine ayrılmıştır. Her kanal için beklenen hassaslik çalışmaları için eş zamanlı bir maksimum olabilirlik uyumu gerçekleştirilmiştir. Run 2 için yapılan DL çalışması, 2017 yılında toplanan çarpışma verileri için 41.5 fb^{-1} entegre ışıklılıkta SM kesitinin 94.23 katına kadar $t\bar{t}HH$ hassiyet ölçümü yapmayı beklemektedir.HL-LHC'deki SL kanal çalışması, 3000 fb^{-1} ışıklılıktaki veri ile SM kesitinin 3.14 katına kadar $t\bar{t}HH$ hassiyet ölçümü yapmayı beklemektedir. Minimal Kompozit Higgs Modeli senaryoları için duyarlılık çalışması da sunulmuştur.

Anahtar Kelimeler: LHC, CMS, Higgs bozonu, top kuark, leptonik

To my beloved family

ACKNOWLEDGMENTS

Throughout the journey of my doctoral research at CERN, I had the distinct honor of engaging with an environment enriched by the expertise of distinguished physicists and scientists. The opportunity to conduct the majority of my research in such an esteemed setting has been a privilege. Central to this invaluable experience has been my supervisor, Prof. Dr. Mehmet Zeyrek, to whom I owe a profound debt of gratitude. For his enduring encouragement, patience, and the significant role he has played in my development as a researcher and individual, I extend my deepest and most heartfelt thanks.

The importance of Sezen Sekmen and Özgür Şahin in my PhD journey cannot be overstated. Their continuous support, provided with generosity and insight, has been a critical element for the success of my thesis. Their readiness to engage in challenging discussions, share their invaluable experiences, and offer encouragement and guidance has been instrumental in navigating the complexities of my research. Without their assistance, my academic achievements would not have been attainable. My heartfelt appreciation to them is boundless.

Special acknowledgment for my analysis team, including Aurore Savoy-Navarro, Sezen Sekmen, Özgür Şahin, Maxwell Chertok, and Wei Wei. My particular thanks to Aurore, for always being there to assist me and guiding me towards the best possible paths for my research. Working with her taught me about doing my best and has greatly influenced how I approach challenges, encouraging me to push the limits.

Many thanks to the Thesis Investigation Committee members, Prof. Dr. Mehmet Zeyrek, Prof. Dr. Altuğ Özpineci and Prof. Dr. Bora Işıldak, for their guidance throughout my research.

I gratefully acknowledge Turkish Energy, Nuclear and Mineral Research Agency (TENMAK) for their financial support, which was fundamental in making my PhD research at CERN possible.

I would like to express my deepest gratitude to my family, my beloved parents and İpek for all their motivation and love throughout this journey. I am thankful to my fluffy companion Psi for the soothing purrs, which has been a source of joy and comfort, especially during my stressful times. Most importantly, I reserve my deepest appreciation for my love and guiding star. His presence has transformed every moment of doubt into one of confidence and possibility. His endless support and unwavering belief in my abilities have been the cornerstone of my journey. Without his love, encouragement, and faith, none of this would have been conceivable. To Özgür, for being my rock and my inspiration, I am eternally grateful.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xii

LIST OF TABLES xvi

LIST OF FIGURES xix

LIST OF ABBREVIATIONS xxv

CHAPTERS

1 INTRODUCTION 1

2 THEORETICAL FRAMEWORK 5

2.1 Standard Model 5

2.1.1 Elementary Particles of the Standard Model 5

2.1.2 The Lagrange formalism and the gauge transformation 8

2.1.3 Quantum Chromodynamics 9

2.1.4 Electroweak theory 10

2.1.5 The Higgs Mechanism 11

2.1.6 Higgs Boson Discovery, Production and Decay Channels . . . 15

2.2 Beyond the Standard Model 20

3 CERN LARGE HADRON COLLIDER AND THE COMPACT MUON SOLENOID EXPERIMENT . . . 23

3.1 The Large Hadron Collider . . . 23

3.2 The Compact Muon Solenoid detector . . . 26

3.2.1 Inner tracking detector . . . 29

3.2.2 Electromagnetic Calorimeter . . . 30

3.2.3 Hadron calorimeter . . . 32

3.2.4 Muon system . . . 33

3.2.5 Triggers and data acquisition system . . . 34

3.2.6 Offline computing . . . 37

3.3 Monte Carlo event simulation in proton-proton collisions . . . 37

3.4 High Luminosity Large Hadron Collider . . . 38

4 OBJECT RECONSTRUCTION . . . 43

4.1 Particle-flow algorithm . . . 43

4.2 Primary vertex . . . 44

4.3 Muons . . . 45

4.4 Electrons . . . 46

4.5 Jets . . . 47

4.5.1 Jet Energy Correction . . . 50

4.5.2 B tagged jets and identification . . . 51

4.6 Missing Transverse Momentum . . . 53

5 MACHINE LEARNING AND ITS APPLICATIONS IN HIGH ENERGY PHYSICS . . . 55

5.1 Machine Learning Flowchart and Hyperparameters . . . 57

5.2 Neural Networks and Deep Learning 59

5.2.1 Convolutional Neural Networks 60

5.2.2 Graph Neural Networks 63

6 PHYSICS MOTIVATION FOR $t\bar{t}HH$ PRODUCTION MECHANISM . . . 65

7 FULL RUN-2 ANALYSIS DETAILS AND STRATEGY 71

7.1 Analysis Method 71

7.2 Data and Simulation Samples 72

7.3 Trigger requirements 81

7.4 Object and Event Selection 85

7.4.1 Muon reconstruction and selection 86

7.4.2 Electron reconstruction and selection 88

7.4.3 Jet reconstruction and selection 89

7.4.4 b jet identification 90

7.4.5 Event Selection 91

7.5 Event Variables for DNN 94

7.6 Graph Attention based jet assignment - GATJA 104

7.7 Data-Monte Carlo Comparison in a Control Region 110

7.8 Event Categorization 114

8 STRATEGY FOR THE $t\bar{t}HH$ SEARCH IN THE SEMILEPTONIC DECAY CHANNEL AT THE HL-LHC 119

8.1 Analysis Method 120

8.2 Simulated Samples for Phase-2 Study 121

8.3 Objects used in the Phase-2 HL-LHC analysis in the single lepton channel 122

8.3.1 Object and baseline event selection 123

8.4 Event Variables for DNN 124

8.5 Event Categorization 127

9 RESULTS AND INTERPRETATION 133

9.1 Measurement of the $t\bar{t}HH$ process in the dileptonic decay channel with the 2017 Run 2 data 133

9.2 Search for the $t\bar{t}HH$ process in the semileptonic decay channel at the HL-LHC 138

10 CONCLUSION 143

REFERENCES 147

APPENDICES

A COMPARISON OF PRE-LEGACY AND ULTRA-LEGACY DISTRIBUTIONS 163

B ELECTRON MVA WORKING POINT STUDY 169

C SIGNAL AND BACKGROUND COMPARISONS FOR INDIVIDUAL BACKGROUNDS FOR THE HL-LHC STUDY 173

D STATISTICAL METHODS 177

CURRICULUM VITAE 181

LIST OF TABLES

TABLES

Table 3.1 The essential components needed for detecting and recognizing the SM particles with their distinct signatures are the primary subsystems. . . 27

Table 5.1 Comparison and definitions of three main ML types. 56

Table 6.1 Cross sections computed at the NLO QCD for ZHH, WHH, VBF HH, $t\bar{t}HH$, and tjHH at 14 TeV center-of-mass energy. 66

Table 6.2 Cross-sections for ttH and ttHH Processes at 13 and 14 TeV. 67

Table 7.1 Integrated luminosity per year and total. 73

Table 7.2 2016preVFP datasets used in the analysis. 73

Table 7.3 2016postVFP datasets used in the analysis. 74

Table 7.4 2017 datasets used in the analysis. 74

Table 7.5 2018 datasets used in the analysis. 74

Table 7.6 Branching ratios of the H, W and Z bosons for the decay channels considered in this analysis (top) and the final state contribution of a top quark anti quark pair decaying dileptonically (bottom). 76

Table 7.7 Production campaigns given for each year. 76

Table 7.8 $t\bar{t}HH \rightarrow b\bar{b}b\bar{b}$ signal samples. 77

Table 7.9 $t\bar{t}$ background samples. 77

Table 7.10 $t\bar{t}4b$ background samples. 78

Table 7.11 $t\bar{t}b\bar{b}$ background samples. 78

Table 7.12 $t\bar{t}Htob\bar{b}$ background samples. 79

Table 7.13 $t\bar{t}Ztob\bar{b}$ background samples. 80

Table 7.14 $t\bar{t}ZZ \rightarrow b\bar{b}b\bar{b}$ and $t\bar{t}ZH \rightarrow b\bar{b}b\bar{b}$ background samples. 80

Table 7.15 List of triggers used for the 2016 data. 81

Table 7.16 List of triggers used for the 2017 data. 82

Table 7.17 List of triggers used for the 2018 data. 82

Table 7.18 Global tags used for data and simulation. 85

Table 7.19 MET filters used in the analysis 85

Table 7.20 Definition of all objects used in the DL channel 86

Table 7.21 Baseline selection applied to muon objects. 87

Table 7.22 Baseline selection applied to electron objects. 89

Table 7.23 Baseline selection and flags applied to jet objects. 90

Table 7.24 Baseline selection and flags applied to b jet objects. 91

Table 7.25 The reference b jet efficiencies for each year and working points. . . 91

Table 7.26 Baseline event selection. 93

Table 7.27 Event quantities calculated for all events passing the baseline selection, including object properties, invariant masses, angular variables. . . . 94

Table 7.28 List of the main DNN hyperparameters values for the baseline selection case requiring: ≥ 4 jets, ≥ 3 b-jets. 116

Table 8.1 List of MC simulated samples for this study, and event yields after baseline selection normalized to 3000 fb^{-1}. 122

Table 8.2 Definition of all objects used in the analysis. 123

Table 8.3 Event selection and search channels. 124

Table 8.4 Event quantities calculated for all events passing the baseline selection, including object properties, invariant masses, angular variables. . . . 125

Table 8.5 List of the main DNN hyperparameters values for the baseline selection case requiring: $\geq$ 4 jets, $\geq$ 3 b jets. 129

Table 9.1 Experimental and theoretical systematic uncertainties considered in this study. 134

Table 9.2 Cross sections at 13 TeV with the uncertanties $\pm$QCD Scale (%) and $\pm$(PDF+α_s) (%) for the signal and all the considered backgrounds in this study. 134

Table 9.3 Upper limits on the signal strength shown for the $\mathrm{t\bar{t}HH}$ signal, with considering both systematical and statistical uncertainties. 136

Table 9.4 Upper limits on the signal strength shown for the $\mathrm{t\bar{t}HH}$ signal, with considering only statistical uncertainties. 136

Table 9.5 List of systematic uncertainties applied in this analysis, their values and their effects on the SM signal and background yields. 139

Table B.1 Comparison for WP80. 170

Table B.2 Comparison for WP90. 171

Table B.3 Comparison for WP98. 172

LIST OF FIGURES

FIGURES

Figure 2.1 Elementary particles of the SM. Charge, mass, color and spin information are provided. 7

Figure 2.2 Representation of a 'Mexican Hat' potential that leads to spontaneous symmetry breaking. 13

Figure 2.3 The production cross sections of the SM Higgs boson are presented. 17

Figure 2.4 The branching ratios for a SM Higgs boson with m_H = 125GeV (left) represented by a pie chart and as a function of Higgs mass (right). 18

Figure 2.5 The Higgs boson discovery showcased in two channels with high resolution. 18

Figure 2.6 Parameters for signal strength are derived for various production modes (left) and for decay channels (right). 19

Figure 3.1 The CERN accelerator complex. 24

Figure 3.2 The 15 m long superconducting dipole magnets for the LHC at CERN. 26

Figure 3.3 Slice view of the CMS detector showing how various particles interact with its subsystems. 28

Figure 3.4 An overview of the CMS detector, with a person included for scale reference. 28

Figure 3.5 The layout of the CMS tracker system. 30

Figure 3.6 Cross section of an ECAL detector quadrant and the layout of the ECAL Endcap Calorimeter. 31

Figure 3.7 General architecture of the CMS Trigger and DAQ System (left) and simple chart of the system with the trigger rates at each step (right). 36

Figure 3.8 Projected peak and integrated luminosity in the HL-LHC for the nominal (left) and ultimate (right) detector parameters. 40

Figure 3.9 The LHC's planned roadmap for the upcoming decade and beyond. 40

Figure 4.1 Comparison of the topologies of sequential kT and anti-kT jet algorithms. 50

Figure 5.1 Comparison of a biological neuron and a perceptron. 59

Figure 5.2 Deep Neural Networks representation with many hidden layers. . 61

Figure 5.3 Convolutional Neural Network flowchart with feature extraction and classification parts. 61

Figure 5.4 Sliding of a 3×3 kernel over an input. 62

Figure 5.5 Illustrations showcasing two diverse classification scenarios using graphs in HEP. 63

Figure 6.1 Total cross sections at the NLO in QCD for the six largest HH production channels at pp colliders. 65

Figure 6.2 Total cross sections at the LO and the NLO in QCD for HH production channels, plotted as a function of the self-interaction coupling λ. 66

Figure 6.3 Representative Feynman diagrams for the $t\bar{t}HH$ process. 67

Figure 6.4 Representative diagrams for the $t\bar{t}HH$ production process at the LO within MCHM. 69

Figure 7.1 Feynman diagram of the $t\bar{t}HH$ production process in the dilepton channel. 72

Figure 7.2 Distributions of several lepton kinematical variables to observe the trigger efficiency for the baseline selection. 83

Figure 7.3 2D plots representing the p_T values of the leading and subleading electrons (top-left), muons (top-right), and overall leptons (bottom) for the 2017 $t\bar{t}HH$ signal sample. 83

Figure 7.4 2D plots representing the p_T values of the leading and subleading electrons (top-left), muons (top-right), and overall leptons (bottom) for the 2017 $t\bar{t}(SL)Htob\bar{b}$ sample. 84

Figure 7.5 2D plots representing the p_T values of the leading and subleading muons (top-left), electrons (top-right), and overall leptons (bottom) for the 2017 $t\bar{t}(DL)Htobb$ sample. 84

Figure 7.6 Distributions of the various kinematical variables comparing the $t\bar{t}HH$ signal samples with (red) and without (blue) b tag corrections applied conditions for 2017. 92

Figure 7.7 Invariant mass distributions comparing the $t\bar{t}HH$ signal (red) with $t\bar{t}ZH$ (green) and $t\bar{t}ZZ$ (yellow) backgrounds for the 2017 samples. 97

Figure 7.8 Distributions of different discriminating variables for leptons for the 2017 samples. 100

Figure 7.9 Distributions of different discriminating variables for jets and b-tagged jets for the 2017 samples. 101

Figure 7.10 Distributions of different discriminating variables for jets and b-tagged jets for the 2018 samples. 102

Figure 7.11 Distributions of different discriminating variables for jets and b-tagged jets for the 2016 samples. 103

Figure 7.12 Higgs boson mass distribution showing the b jets matched or unmatched to the Higgs boson. 105

Figure 7.13 GATJA flowchart. 106

Figure 7.14 A simplified view of the attention mechanism used for each b-tagged jet. 107

Figure 7.15 Hyperparameters and input varibles employed during the GATJA training process are listed. 108

Figure 7.16 Higgs boson node (top left), top quark node (top right) and other objects node (bottom) of the GATJA output for the 2017 case. 108

Figure 7.17 Higgs boson node of the GATJA output for the 2018 case. 108

Figure 7.18 Distributions of GATJA outputs obtained for 8 b-tagged jets. . . 109

Figure 7.19 Distributions of the Data-MC comparison for the jet and b-tagged jet related variables for the 2017 samples. 111

Figure 7.20 Distributions of the Data-MC comparison for the jet and b-tagged jet related variables for the 2017 samples. 112

Figure 7.21 Distributions of the Data-MC comparison for the lepton related variables for the 2017 samples. 113

Figure 7.22 DNN workflow in the DL channel. 114

Figure 7.23 Top 20 ranked event features contributing to the DNN training are listed. 115

Figure 7.24 Confusion matrices showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2017 results. 117

Figure 7.25 Confusion matrix showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2017 results. 117

Figure 7.26 Confusion matrices showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2018 results. 117

Figure 8.1 Feynman diagram of the $t\bar{t}HH$ production process in the single lepton channel. 120

Figure 8.2 Distributions of different discriminating variables after the baseline selection applied. 126

Figure 8.3 Probability density functions for different discriminating variables after the baseline selection applied. 127

Figure 8.4 Distributions of different discriminating variables after the baseline selection applied. 127

Figure 8.5 Schematic view showing the categorization of events using a DNN-based procedure 129

Figure 8.6 Final discriminant distributions for SM $t\bar{t}HH$ are shown for three different n_{bjet} categories; $n_{bjet} = 3$ (top) $n_{bjet} = 4$ (middle), and $n_{bjet} > 4$ (bottom). 130

Figure 8.7 Final discriminant distributions for the $t\bar{t}HH$ $MCHM_5^{C2}$ benchmark point case (left) and $t\bar{t}HH$ $MCHM_{14}^{D7}$ benchmark point case (right) are shown. 131

Figure 9.1 Final discriminant distributions for the $t\bar{t}HH$ signal node and $t\bar{t}$, $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t}ZZ$, and $t\bar{t}ZH$ background nodes. 135

Figure 9.2 The 95% upper limits on the signal strength shown for the $t\bar{t}HH$, $t\bar{t}HH+t\bar{t}ZH$ and $t\bar{t}HH+t\bar{t}ZH+t\bar{t}ZZ$ processes for different scenarios of systematic uncertainties. 137

Figure 9.3 The top 30 impacts and their effects on the signal strength in the fit to pseudo data. 137

Figure 9.4 The 95% upper limits on the signal strength shown for the SM $t\bar{t}HH$, $t\bar{t}HH + t\bar{t}ZH$ and $t\bar{t}HH + t\bar{t}ZH + t\bar{t}ZZ$ processes for different scenarios of systematic uncertainties. 140

Figure 9.5 The 95% upper limits on the signal strength for the $t\bar{t}HH$ $MCHM_5^{C2}$ and $t\bar{t}HH$ $MCHM_{14}^{D7}$ processes for different scenarios of systematic uncertainties. 141

Figure A.1 The PreLegacy and UltraLegacy $t\bar{t}HH$ samples comparison. . . . 164

Figure A.2 The PreLegacy and UltraLegacy $t\bar{t}H \to b\bar{b}$ samples comparison. 165

Figure A.3 The PreLegacy and UltraLegacy $t\bar{t}DL$ samples comparison. . . . 166

Figure A.4 The PreLegacy and UltraLegacy $t\bar{t}Z \to b\bar{b}$ samples comparison. 167

Figure C.1 Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $t\bar{t}H$ and $t\bar{t}Z$ backgrounds. 173

Figure C.2 Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $t\bar{t}4b$ background. . . 174

Figure C.3 Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $t\bar{t}ZZ$ and $t\bar{t}ZH$ backgrounds. 175

LIST OF ABBREVIATIONS

ABBREVIATIONS

CERN	European Organization for Nuclear Research
CMS	Compact Muon Solenoid
BEH	Brout–Englert–Higgs
BSM	Beyond the Standard Model
CL	Confidence Level
DL	Dileptonic
DNN	Deep Neural Network
fb	Femtobarn
GeV	Giga electron Volt
HL-LHC	High Luminosity-Large Hadron Collider
LO	Leading Order
MCHM	Minimal Composite Higgs Model
NLO	Next-to-Leading Order
SL	Semileptonic
SM	Standard Model
TeV	Tera electron Volt
VBF	Vector Boson Fusion
QCD	Quantum Chromodynamics
QED	Quantum Electrodynamics

CHAPTER 1

INTRODUCTION

High-energy physics aims to uncover the smallest building blocks of nature, known as elementary particles, and understand the fundamental forces lying behind their interactions. The standard model (SM) of particle physics stands as a main guide in understanding these particles and their behaviors. A key tool in this journey is the Large Hadron Collider (LHC) located at the European Organization for Nuclear Research (CERN) in Switzerland. The LHC, the biggest accelerator ever constructed, plays a crucial role in creating high-energy collisions between particles, with the aim of reaching the limits of the SM and uncovering potential phenomena beyond its scope, known as beyond the SM (BSM). Theoretical groundwork laid by the SM has led to remarkable experimental successes. One of the most important discoveries is undoubtfully the Higgs boson [1–4], a missing piece in the SM puzzle. Its discovery was announced in 2012 by the colloborations of the A Toroidal LHC ApparatuS (ATLAS) and Compact Muon Solenoid (CMS) detectors at the LHC [5, 6]. This groundbreaking discovery triggered extensive investigations into the Higgs sector, with a focus on tests to verify the SM and contributions from physics beyond the SM (BSM).

The interaction of the Higgs boson with other elementary particles is anticipated to be proportional with their masses. Given that the top quark is the heaviest known elementary particle, it couples to the Higgs boson with a Yukawa coupling constant close to unity [7]. Consequently, processes like $\mathrm{t\bar{t}HH}$ provide a means to access the Top-Higgs Yukawa coupling. Unlike the $\mathrm{t\bar{t}H}$ process, the $\mathrm{t\bar{t}HH}$ process not only grants access to the triple Higgs coupling but also differs from single Higgs production by excluding interference terms in the access to the triple Higgs coupling. The investigation of these crucial couplings, in conjunction with the production processes

of $\mathrm{t\bar{t}H}$ and double Higgs (both through gluon fusion and vector boson fusion (VBF)), underscores the significance of the $\mathrm{t\bar{t}HH}$ process within the SM. Consequently, one of the primary objectives of this thesis is to search for nonresonant $\mathrm{t\bar{t}HH}$ production and advance towards measuring the Higgs trilinear coupling.

In addition to the SM framework, the $\mathrm{t\bar{t}HH}$ process opens a window to the physics beyond the SM (BSM) [8], providing an additional motivation for this thesis. Deviations from the SM, evident in both signal strength and kinematic distributions, as well as the potential identification of high-mass resonances, offer diverse ways for exploring physics beyond the SM. This thesis puts particular emphasis on the $\mathrm{t\bar{t}HH}$ process within the framework of Minimal Composite Higgs Models (MCHM) [9–11]. Two MCHM scenarios, as studied in [12], are chosen as case studies, and, for the first time, an experimental analysis is conducted within the context of the CMS experiment and the High Luminosity-LHC (HL-LHC) to predict the sensitivity. This investigation provides insightful initial perspectives into the observability of MCHM-like scenarios at the HL-LHC, highlighting the potential to enhance the chances of detecting the $\mathrm{t\bar{t}HH}$ process at the HL-LHC.

The thesis consists of two complementary studies, each focusing on the production of a top quark-antiquark pair $\mathrm{t\bar{t}}$ associated with a pair of Higgs bosons (HH), while exploring different signal topologies. In the first study, the investigation centers on the production of a $\mathrm{t\bar{t}}$ pair associated with a HH pair, where the top quark pair decays dileptonically and the Higgs boson pair decays hadronically into b quark-antiquark pairs. This analysis utilizes the full Run 2 dataset, corresponding to an integrated luminosity of 137.60 fb^{-1} at a 13 TeV center-of-mass energy. The inclusive $\mathrm{t\bar{t}HH}$ cross section at next-to-leading order (NLO) in QCD is calculated as $0.775^{+1.5\%}_{-4.3\%} \pm 3.2\%$ fb, with the first uncertainty originating from the QCD scale dependence and the second from parton distribution functions [13, 14]. The second study focuses on the search for the production of a $\mathrm{t\bar{t}}$ pair decaying semileptonically associated with a Higss boson pair also decaying into b quark-antiquark pairs within the SM framework. The analysis is conducted in the context of the High-Luminosity Large Hadron Collider (HL-LHC), which is expected to operate at a center-of-mass energy of 14 TeV and aims for a total integrated luminosity of 3000 fb^{-1}. The SM production cross section for the $\mathrm{t\bar{t}HH}$ process is computed at NLO QCD to be $0.948^{+1.7\%}_{-4.5\%} \pm 3.1\%$

fb [13]. This study benefits from upgraded detector conditions, providing an increased integrated luminosity with operation at a higher center-of-mass energy and a higher production cross section. To optimize the conditions for the HL-LHC, this work is conducted with simulated samples reproducing the key features of the upgraded CMS detector designed for the HL-LHC.

In the DL channel study performed with the full Run 2 data, both W bosons from the top quarks decay into leptons, either $\mathrm{e}^+\mathrm{e}^-, \mu^-\mu^+, \mathrm{e}^-\mu^+$, or $\mathrm{e}^+\mu^-$, and their corresponding neutrinos. Additionally, two Higgs bosons decay into a bottom quark-antiquark pair individually. Hence, the final state consists of two leptons, their corresponding neutrinos, and six b-tagged jets. The branching ratio for this decay signature is the lowest one among the other decay channels with a value of 3.6%, but it gives a rather clean signature. On the other hand, the HL-LHC study in the SL channel, has a W boson decaying into a lepton (either e or μ) together with its corresponding neutrino and another W boson decaying hadronically into two quarks. Including the Higgs boson pair both decaying hadronically into bottom quarks, a final state with eight relatively high transverse momentum (p_T) jets, at least six of which are b-tagged, arises. The branching ratio for this specific decay channel has a value of 11.4%. For both studies, various potential constraints and limitations are encountered. These include a reduced expected number of jets due to detector acceptance, the potential for jet merging, and challenges related to b-tagging efficiency. Consequently, this results in a lower signal production rate when compared to the high rates of competing physics backgrounds. Therefore, a relatively loose baseline selection is applied to the events to optimize the overall sensitivity of the analysis. In particular, both analyses require a minimum of 4 jets and at least 3 b-tagged jets in the final state selection. Additional requirements or filters are applied to other objects, adjusted according to the specific requirements of each analysis.

The dominant background contribution in this final state originates from the QCD production of top pair+jets, encompassing all associated production cases involving light, c, or b quark jets. Additional backgrounds include the $\mathrm{t\bar{t}}$H, $\mathrm{t\bar{t}}$Z, $\mathrm{t\bar{t}}$ZZ, and $\mathrm{t\bar{t}}$ZH production processes, where both Z and H bosons decay into a b quark pair. Alongside the primary background arising from the production of top quark pairs with additional b quarks, this analysis is affected by a combinatorial background due

to multiple b-quark jets in the final state. This brings a necessity for the reconstruction of the invariant mass of the Higgs boson and the assignment of b-quark jets in the final state to their originating mother particles, either the Higgs boson or the top quark. To address this, a jet assignment method based on graph attention layers is specifically developed for the Run 2 analysis. After applying the baseline selections, multi-classifier deep neural networks (DNNs) categorize selected events into multiple categories, based on their consistence with the signal or various background processes. The final DNN discriminants are used for obtaining the upper limits on the signal strength.

The study performed for the HL-LHC in the SL channel is published by the CMS collaboration in a Physics Analysis Summary (PAS) [15] and also contributes to the Snowmass White Paper publication [16].

The thesis is structured in 10 chapters. Chapter 1 gives an introduction to these studies by mentioning the motivation behind and the analysis strategies followed. Chapter 2 presents a brief overview of the SM and the BSM physics, including the importance of the Higgs boson. Chapter 3 is dedicated to a review of the LHC and the CMS experiments and explains methods to produce simulated events in proton-proton collisions. It also introduces the upgraded LHC, HL-LHC, conditions and the time schedule. Chapter 4 gives the idea of the object reconstruction in the CMS detector. Chapter 5 explores machine learning techniques, offering a general understanding and their specific applications within the context of this thesis. Chapter 6 presents the physics motivation for studying the $\mathrm{t\bar{t}HH}$ production mechanism in terms of the SM and the BSM physics. Chapter 7 concentrates on the analysis methodology developed for the Run 2 study, while Chapter 8 details the analysis methodology applied in the HL-LHC study. Lastly, Chapter 9 presents the results obtained from the two studies, accompanied by insights into their interpretation and Chapter 10 concludes the thesis.

CHAPTER 2

THEORETICAL FRAMEWORK

2.1 Standard Model

The standard model (SM) of particle physics represents the current understanding of elementary particles and their interactions. Developed over the last five decades, it has been extensively validated through several experiments, effectively predicting and explaining various physics processes. Despite its remarkable success, the SM is being questioned by particle physicists due to conceptual limitations and unexplained experimental observations, leading to the proposal of new theories that extend beyond the standard model (BSM).

2.1.1 Elementary Particles of the Standard Model

The SM provides a framework to understand the nature of the elementary particles and four fundamental particle interactions—gravity, electromagnetism, weak interaction, and strong interaction—encompassing electromagnetic, weak, and strong forces within the $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge group [17–22].

The strong nuclear force acts at a distance of about one fermi (or 10^{-15} meters). In contrast, the weak force, responsible for radioactive decay, operates at a much shorter range of 10^{-17} meters, making it about 10^{-5} times weaker at low energies. The electromagnetic force, dominating everyday physics, has an infinite range and strength determined by the fine structure constant ($\alpha \approx 10^{-2}$). Gravity, the fourth force, also has an infinite range and a low energy coupling ($\tilde{1}0^{-38}$), making it too weak to be observed in typical laboratory experiments. The SM successfully explains

how forces like electromagnetism, the strong force, and the weak force interact with particles. On the other hand, fitting the gravity, which is a significant force in our daily lives into the SM is challenging due to the complexities of the mathematical frameworks for describing the micro (quantum theory) and macro (general relativity) worlds while working together. Luckily, experiments at the particle physics level are minimally affected by gravity, allowing the SM to work well without including gravity in its equations. While gravity becomes dominant at larger scales, like planets or our bodies, the SM acts efficiently in explaining the fundamental forces [23,24].

The SM includes two main particle groups called fermions and bosons depending on their 'spins'.

Fermions serve as the fundamental elements of matter, characterized by a half-integer spin (S = 1/2). They are further subcategorized as leptons and quarks. The three charged leptons are the electron (e), muon (μ), and tau (τ), accompanied by their associated electrically uncharged neutrinos known as the electron, muon, and tau neutrino (ν_e, ν_μ, ν_τ). While e, μ, and τ leptons are interacted via both electromagnetic and weak interactions, their corresponding neutrinos can interact only via weak interactions. This makes the neutrinos hard to be detected. Another member of fermions is the quark families consisting of the up (u) and down (d) quarks, the charm (c) and strange (s) quarks, and the top (t) and bottom (b) quarks. Besides the varying particle masses, these three families are essentially identical. The strong force, also known as the nuclear force, binds quarks together. Quarks possess color charges — red (R), green (G), and blue (B) — with corresponding anti-color charges. Each quark or antiquark carries one of these color charges, and they cannot exist independently, a phenomenon referred to as color confinement. Consequently, color-charged quarks form color-neutral groups, called hadrons. Baryons, composed of three quarks (e.g., proton (uud)), and mesons, formed from a quark and an antiquark (e.g., neutral pion (π ($u\bar{u}$ or $d\bar{d}$))), are examples of these composite structures.

Additionally, in accordance with the Pauli exclusion principle, fermions, with their half-integer spins, obey the rule that two or more identical particles cannot occupy the same quantum state within a quantum system simultaneously.

Bosons can be divided into two sub categories, namely gauge bosons and the Higgs

boson. Gauge bosons serve as mediators, transmitting forces between particles. These include the photon (γ, the electromagnetic force carrier), gluon (g, the strong force carrier), and W and Z bosons (weak force carrier). They all have a spin S = 1. While photons and Z bosons are electrically neutral, gluons carry a color charge allowing them to interact with other gluons. In contrast, W bosons can carry either a positive or a negative electric charge. On the other hand, the Higgs boson, one of the most important discoveries in the particle physics, possesses distinct characteristics. Unlike gauge bosons, it is an elementary boson and does not mediate any of the fundamental forces. Instead, the Higgs boson is linked to the Higgs field, known as an unseen, uniform "cloud" throughout space believed to determine the mass of all particles. With a spin S = 0, it is electrically neutral and has a significant mass — over a hundred times heavier than a proton. More details on the Higgs boson are provided in Section 2.1.6.

All the details of the elementary particles of the SM are presented in Figure 2.1, adapted from [25], and further information are provided in References [17–22] .

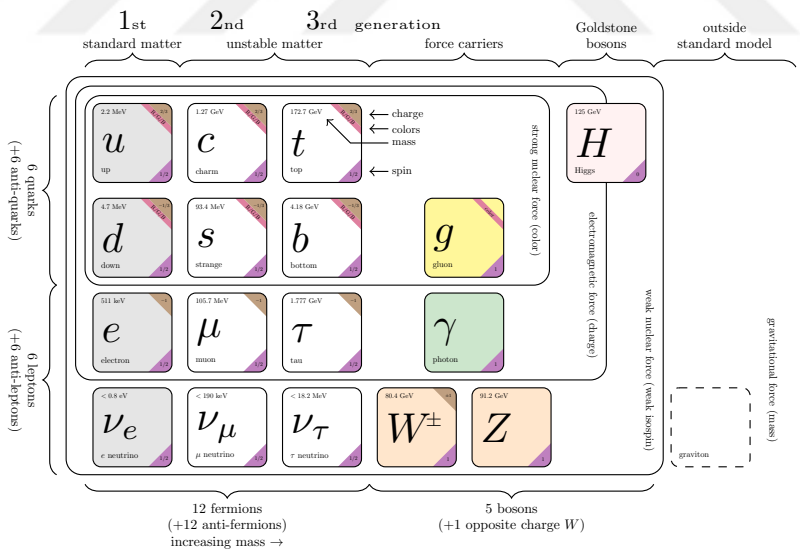

Figure 2.1: Elementary particles of the SM. Charge, mass, color and spin information are provided.

2.1.2 The Lagrange formalism and the gauge transformation

The Lagrange function is a mathematical tool utilized to find the stationary points of a constrained problem. Within the field theories context, it is employed to minimize the action, making it a useful framework for describing dynamic systems under specific constraints.

The system's evaluation is determined using the Euler-Lagrange equations, expressed as:

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i)} \right) - \frac{\partial \mathcal{L}}{\partial \phi_i} = 0 \tag{2.1}$$

Here, $\mathcal{L}$ represents the Lagrangian of the system, and ϕ_i denotes the fields. The solutions to this equation yields the equations of motion for the system of interest, and the particles of the SM emerge through the quantization of these fields.

Introducing invariance under local gauge transformations brings about a set of gauge potentials coupling to scalar and fermion matter fields. For instance, ensuring invariance for fermions involves local transformations such as:

$$\psi(x) \to e^{iq_\psi \alpha(x)} \psi(x) \tag{2.2}$$

Here, ψ is the wave function of a spin-1/2 particle, $\alpha(x)$ is a scalar phase, and q_ψ is the electric charge. This introduces a vector potential A_μ coupled through the electromagnetic field Lagrangian:

$$\mathcal{L} = i\bar{\psi}\gamma_\mu D_\mu \psi - m\bar{\psi}\psi - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} \tag{2.3}$$

The gauge covariant derivative D_μ is defined as

$$D_\mu = \partial_\mu q_\psi A_\mu(x) \tag{2.4}$$

and

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.5}$$

Simultaneously, A_μ undergoes a transformation:

$$A_\mu(x) \to A_\mu(x) - \partial_\mu \alpha(x). \tag{2.6}$$

The Lagrangian described in the Equation 2.3 characterizes the interactions of charged fermions within Quantum Electrodynamics (QED). The scalar phase transformations, associated with an integer electric charge, correspond to the unitary group $U(1)_{em}$. Solutions to the Euler-Lagrange equations for this Lagrangian provide insights into the nature of QED interactions.

2.1.3 Quantum Chromodynamics

The introduction of color quantum numbers addresses the puzzle arising from observations of hadron states like $\Delta^{++}(uuu)$. This inclusion of three color charges is crucial in preventing low-lying fermion states from violating Pauli's exclusion principle.

Three distinct color charges — as introduced in Section 2.1.1 R, B, G — form an exact $SU(3)_C$ symmetry group. This local gauge symmetry governs interactions between color-charged quarks through gluon mediation, leading to the formulation of Quantum Chromodynamics (QCD) [17, 26]. Gluons, themselves carrying color charges, enabling the self-interaction of the mediators.

A fundamental postulate of QCD, which is already introduced in Section 2.1.1, is known as the confinement phenomenon. It states that only color singlet (colorless) particles are observable in nature. Hence, particles with color charges bind together, forming colorless composite particles. Consequently, quarks and gluons are only observable as bound states. Another notable aspect of QCD emerges when quarks possess high energies or are in a closed position—they behave as if they are free particles,

a phenomenon called asymptotic freedom [27,28]. This allows for the construction of a simplistic model for QCD interactions using a linear potential V(r)∼r.

Similar to the QED, a Lagrangian is formulated for QCD, incorporating three color charges

$$\mathcal{L} = i\bar{\psi}\gamma_\mu D^\mu \psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - (q_s\bar{\psi}\gamma^\mu\lambda\psi)A_\mu, \tag{2.7}$$

where λ represents the Gell-Mann matrices, and q_s is the corresponding color charge. Notably, the QCD Lagrangian introduces an additional term due to the self-interaction of gluons comparing to the QED Lagrangian (Equation 2.3).

2.1.4 Electroweak theory

Enrico Fermi proposed a theory to explain the neutron's β decay using an effective interaction with a coupling constant G_F. However, this theory is only accurate within a specific energy range, which is the energy scale of $m_{W^\pm}$. The Glashow-Salam-Weinberg [29–31] model provides a more comprehensive solution by unifying Quantum Electrodynamics (QED) and weak interactions. This unification, combining electromagnetic and weak forces, is a significant step toward grand unified theories (GUT). Observations show that weak interactions have a unique chiral structure, unlike QED or QCD. Only left-handed particles and right-handed antiparticles interact through the weak force. Consequently, the mediators of weak interaction, Z^0 and $W^\pm$, exclusively interact with left-handed doublets. The simplest group containing these doublets is $SU(2)$. The electroweak theory (EWK) unifies $SU(2)_L$ and $U(1)_Y$, introducing hypercharge (Y) to preserve weak interaction in flavor space.

$$\frac{Y}{2} = Q - T_3, \tag{2.8}$$

where $T_3 = \sigma_3/2$ and Q stands for the electromagnetic charge operator.

Similar to the QED and QCD, the covariant derivative is expressed as:

$$D_\mu = \partial_\mu - 2ig'\frac{1}{2}YB_\mu - igTW_\mu \tag{2.9}$$

Here, g and g' are coupling constants, B_μ is the gauge field, T is a vector of Pauli matrices, and W_μ is a three-vector gauge field.

The Lagrangian for electroweak interaction is expressed as:

$$L = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}W^{\mu\nu}W_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \tag{2.10}$$

where W_n and $B_{\mu\nu}$ are field strength tensors. The following relations are solved to find the mass eigenstates of the EWK bosons

$$\begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B \\ W^3 \end{pmatrix}$$

and

$$W^\pm = W^1 \pm iW^2 \tag{2.11}$$

where θ_W is called the EWK mixing angle.

The EWK bosons' mass eigenstates arise from these fields, addressing the absence of mass terms for gauge fields in the Lagrangian. However, from the experimental evidence, it is known that the physical $W^\pm$ and Z bosons, as well as fermions are massive. This challenge is resolved through the spontaneous symmetry breaking mechanism involving the Higgs field and Higgs particle.

2.1.5 The Higgs Mechanism

The $SU(2)_L$ symmetry in the electroweak theory, providing an accurate description of weak interaction, is not experimentally observed due to measurable masses of fermions and Z and $W + \pm$ bosons. This discrepancy suggests a hidden sym-

metry, leaving only a $SU(1)$ symmetry associated with electromagnetic interactions and a massless photon in the electroweak sector. The mechanism responsible for breaking $SU(2)_L \times U(1)_Y$ symmetry and giving mass to the SM bosons is called the Brout–Englert–Higgs (BEH) mechanism [1, 2, 4].

The BEH mechanism introduces a complex scalar field ϕ, which undergoes a spontaneous symmetry breaking in the Lagrangian. According to the Goldstone theorem, such spontaneous breaking of continuous symmetry leads to the occurrence of a massless scalar particle, known as Goldstone boson. Consequently, the number of Goldstone bosons in a theory corresponds to the number of broken generators within the symmetry group. However, in the context of gauge theories, such as the SM, the full picture is more nuanced. In these theories, the massless gauge bosons in the initial state gain mass through the absorption of Goldstone bosons. Hence, the number of massive gauge bosons in a gauge theory undergoing spontaneous symmetry breaking aligns with the number of broken generators.

Prior to electroweak symmetry breaking, all four electroweak gauge bosons — W^1, W^2, W^3, and B^0 — are massless. On the other hand, experimental observations reveal one massless gauge boson (γ), and three massive gauge bosons (W^+, W^-, and Z), leading a spontaneous symmetry breaking

$$SU(2)_L \times U(1)_Y \rightarrow U(1)_{em}. \tag{2.12}$$

The scalar fields need to contain at least three degrees of freedom for the mechanism to work. The simplest way to do this is by introducing a complex, scalar $SU(2)$ doublet Φ such that

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{2.13}$$

Now, a scalar Lagrangian including both a kinematic and a potential term is defined,

$$\mathcal{L} = (D^\mu \Phi)^\dagger (D_\mu \Phi) - V(\Phi) \tag{2.14}$$

where first term is the kinetic term and the second term, i.e., the scalar potential $V(\Phi)$ is defined as

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda(\phi^\dagger \phi)^2. \tag{2.15}$$

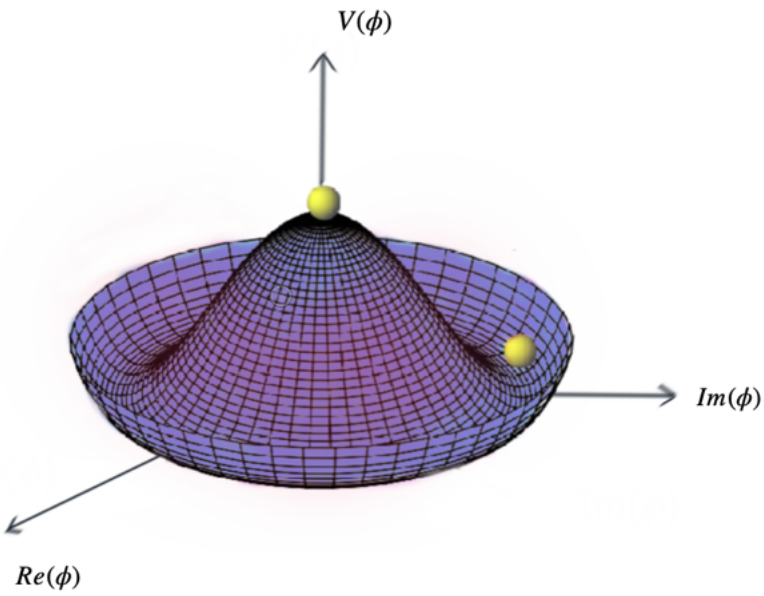

Figure 2.2: Representation of a 'Mexican Hat' potential that leads to spontaneous symmetry breaking. It illustrates the Higgs boson as a massive, spin-zero particle vibrating radially, fluctuating between the hat's center and edge.

For vacuum stability, the parameter λ must be positive. The Higgs potential function, also known as the "Mexican hat" potential, is illustrated in Figure 2.2 and the two cases for the minimum value of $V(\Phi)$ are distinguished based on the sign of μ^2 [32]:

- **If μ^2 > 0,** $V(\Phi)$ is always positive having a minimum at the origin such that

$$< 0|\Phi|0 >= \Phi_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.16}$$

  Hence, no spontaneous symmetry breaking occurs.

- **If μ^2 < 0,** unlike the previous case, there is no minimum located at the origin. Thus, the neutral component of Φ results in a non-zero vacuum expectation

value (VEV) denoted as v and a spontaneously broken $SU(2)_L \times U(1)_Y$ symmetry.

$$< 0|\Phi|0 >= \Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \mathrm{v} \end{pmatrix} \tag{2.17}$$

resulting in a non-zero vacuum expectation value (VEV) and a spontaneously broken $SU(2)_L \times U(1)_Y$ symmetry, where v of the scalar field ϕ is given by

$$\mathrm{v} = \sqrt{-\frac{\mu^2}{\lambda}} \tag{2.18}$$

Expanding Φ around its minimum value Φ_0

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \mathrm{v} + H(x) \end{pmatrix}. \tag{2.19}$$

This implies that the only remaining scalar field belongs to the Higgs boson H, which corresponds to the last particle predicted by the SM [32].

The Lagrangian given in Equation 2.14 becomes

$$L = \ldots + \frac{1}{2} 2\mu^2 H^2 + \frac{g_2^2 \mathrm{v}^2}{4} W_\mu^+ W_\mu^- + \frac{1}{2} \frac{g_1^2 + g_2^2}{4} \mathrm{v}^2 Z_\mu Z^\mu + \ldots \tag{2.20}$$

by allowing to extract mass information for the gauge bosons such that

$$m_W = \frac{g_2 \mathrm{v}}{2} \tag{2.21}$$

$$m_Z = \frac{\sqrt{g_1^2 + g_2^2}}{2} \mathrm{v} \tag{2.22}$$

$$m_H = \sqrt{2}\mu \tag{2.23}$$

The remaining part of the Lagrangian includes interaction terms between the Higgs boson (H) and the W and Z bosons, as well as self-coupling terms of the Higgs boson. Importantly, there is no term involving the photon field (A_μ). Consequently, the photon remains massless after symmetry breaking, as anticipated, and it does not engage in interactions with the Higgs field. Also, m_H requires an experimentally found value, as there is no alternative method for accessing the parameter λ.

The final stage involves determining the masses of fermions. Since explicit fermion mass terms violate the gauge symmetries, to give masses to fermions without breaking the gauge symmetries, the SM introduces Yukawa coupling terms. These terms couple the Higgs field to both the left-handed and right-handed components of fermions. The Yukawa interactions are gauge invariant, preserving the SU(2)L×U(1) symmetry and Yukawa Lagrangian for electrons can be defined as

$$L_e^{Yukawa} = -\frac{Y_e}{\sqrt{2}}\mathrm{v}(\bar{e}_L e_R + \bar{e}_R e_L) - \frac{Y_e}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L)H \tag{2.24}$$

Thus, the Yukawa term Y_e leads to a mass term for the electron such that

$$m_e = \frac{Y_e \mathrm{v}}{\sqrt{2}}, \tag{2.25}$$

and Equation 2.24 becomes,

$$L_e^{Yukawa} = -m_e \bar{e}e - \frac{m_e}{\mathrm{v}}\bar{e}eH. \tag{2.26}$$

Masses of other fermions can be driven following the same formulasim derived for electrons.

2.1.6 Higgs Boson Discovery, Production and Decay Channels

The Brout–Englert–Higgs (BEH) mechanism was initially proposed in 1964 through the collaborative efforts of Brout and Englert [4], Higgs [1,2], and Guralnik, Hagen, and Kibble [33]. Additional insights into the mechanism were contributed in 1966

by Higgs [3] and in 1967 by Kibble [34]. In that same year, Weinberg [31] and Salam [30], building upon Glashow's 1961 work [29], extended the BEH mechanism to formulate a theory unifying EM and weak interactions. Following the discovery of the W and Z bosons by the UA1 and UA2 experiments at CERN in 1983 [35–37], the search for the Higgs boson gained prominence in particle physics.

Although the mass m_H of the SM Higgs boson is not predicted by theory, it is anticipated to be below approximately 1 TeV by general constraints. Precision EW measurements provide further constraints, suggesting that m_H is less than 152 GeV at a 95% confidence level (CL) [38]. In the past two decades, exhaustive searches for the Higgs boson were conducted at the LEP collider, establishing a lower limit of m_H >114.4 GeV at 95% CL [39]. Moreover, investigations at the Tevatron proton-antiproton collider excluded the mass range of 162–166 GeV at 95% CL [40] and reported an excess of events within the range of 120–135 GeV, as detailed in some studies [41–43].

Following several studies and proposals on the search for the Higgs boson, the LHC became a central focus and a crucial motivation for the construction of the LHC [44, 45], as well as for experiments such as ATLAS and CMS. Initial direct investigations at the LHC performed with the data collected from proton-proton collisions, encompassing an integrated luminosity of 5 fb^{-1} at a center-of-mass energy of $\sqrt{s}$ = 7 TeV. The CMS experiment, with a confidence level of 95%, ruled out a mass range spanning from 127 to 600 GeV [5], while the ATLAS experiment, at the same confidence level, excluded ranges of 111.4–116.6, 119.4–122.1, and 129.2–541 GeV [6]. In the remaining permissible mass region, both experiments reported an excess of events near 125 GeV. Despite the findings, the evidences were not enough for a certain declaration of discovery. In 2012, the proton-proton center-of-mass energy was raised to 8 TeV, and by the end of June, each of these experiments had accumulated an additional integrated luminosity of more than 5 fb^{-1}, thereby considerably boosting the search sensitivity for the Higgs boson discovery.

The identification of the Higgs boson at the LHC does not occur directly; instead, it is inferred through the analyzing of its decay products. Due to its interaction with all massive elementary particles, various production modes are observed as shown in

Figure 2.3 [46], and it offers a remarkable opportunity to investigate its couplings to various SM particles. Moreover, the decay of the Higgs boson into any pair of massive SM particles, as well as into massless particles through virtual loops, is possible. The predominant decay modes include $\mathrm{H} \to \mathrm{b\bar{b}}$, $\mathrm{H} \to \mathrm{WW}^*$, $\mathrm{H} \to \mathrm{gg}$, $\mathrm{H} \to \tau^+\tau^-$, $\mathrm{H} \to \mathrm{c\bar{c}}$, $\mathrm{H} \to \mathrm{ZZ}^*$. Decays into $\mathrm{H} \to \gamma\gamma$, $\mathrm{H} \to \gamma\mathrm{Z}$, $\mathrm{H} \to \mu^+\mu^-$ have much smaller rates. Since the decays into gluons, diphotons, and $\mathrm{Z}\gamma$ are loop-induced, they provide indirect insights into the Higgs couplings to WW, ZZ, and $\mathrm{t\bar{t}}$ in different combinations [47]. The decay branching ratios of the Higgs boson [13,48] are listed as a pie chart in Figure 2.4, adapted from [49,50]. Since the mass of the particles defines the couplings, daughter particles with higher masses have the higher decaying rate. The most predominant decay channel involves the Higgs boson decaying into a pair of bottom quarks, constituting approximately 58% of the decaying ratio. The $\mathrm{H} \to \mathrm{b\bar{b}}$ decay is in the center of this thesis and it was observed by the CMS [51] and ATLAS [52] collaborations in 2018. A comprehensive array of other Higgs boson decay modes, encompassing WW, ZZ, $\gamma\gamma$, and $\tau\tau$ [53] have been successfully identified.

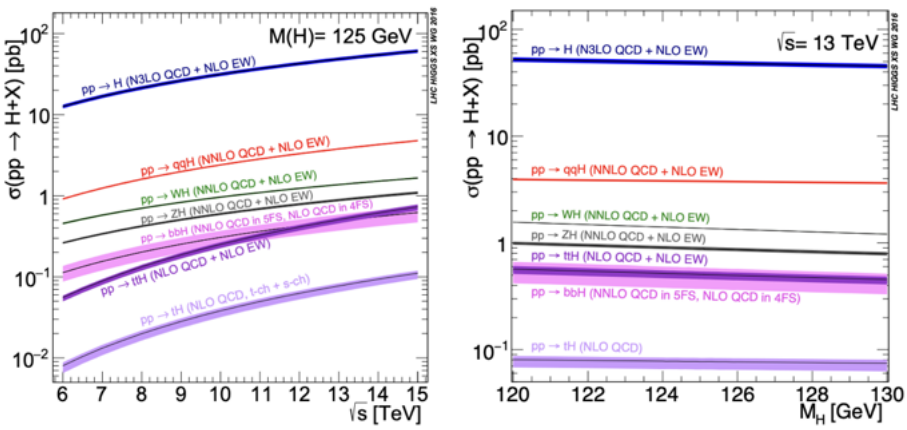

Figure 2.3: The production cross sections of the SM Higgs boson are presented. The left figure is given as a function of the center-of-mass energy, $\sqrt{s}$ and the right figure is given as a function of the Higgs mass, for proton-proton collisions.

The Higgs boson discovery is announced by the CMS and ATLAS collaborations with a combined data from all decay channels and incorporating the 7 TeV and 8 TeV datasets with integrated luminosities of 5.1 fb^{-1} and 5.3 fb^{-1} respectively. The five decay modes considered are $\gamma\gamma$, ZZ, $\mathrm{W}^+\mathrm{W}^-$, $\tau^+\tau^-$, and $\mathrm{b\bar{b}}$. An excess of events is observed above the expected background, reaching a local significance of 5.0 standard deviations at a mass approximately 125 GeV, indicative of a new particle's production.

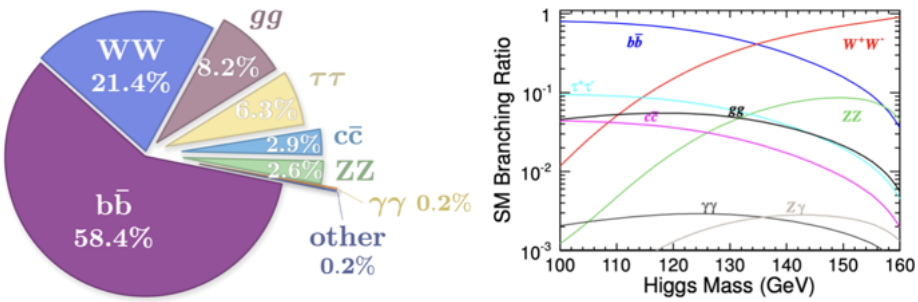

Figure 2.4: The branching ratios for a SM Higgs boson with m_H = 125GeV (left) represented by a pie chart and as a function of Higgs mass (right).

The projected significance for a SM Higgs boson of that mass is 5.8 standard deviations. In Figure 2.5, the CMS experiment presents the results from the most notable excess occurring in the two decay modes with the best mass resolution; $\mathrm{H} \to \gamma\gamma$, and $\mathrm{H} \to \mathrm{ZZ}^* \to 4\mathrm{l}$. The decay into two photons suggests that the newfound particle is a boson with spin distinct from one and a fit to these signals yields a mass of

$$m_H = 125.3 \pm 0.4(\text{stat}) \pm 0.5\,(\text{syst})\,\text{GeV}. \tag{2.27}$$

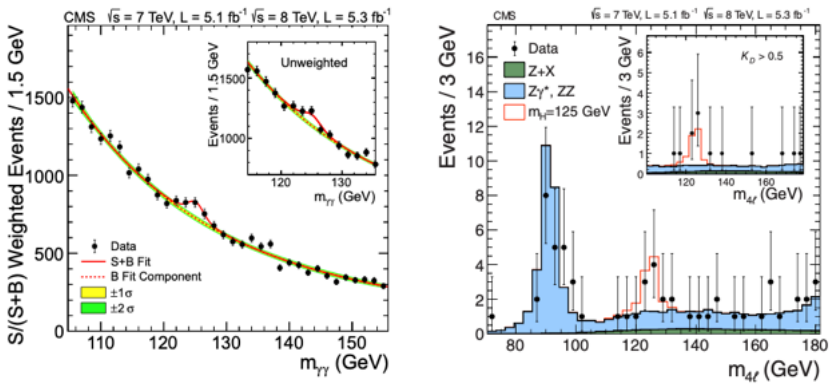

Figure 2.5: The Higgs boson discovery showcased in two channels with high resolution. Left: Illustration of the diphoton invariant mass distribution, where each event is weighted by the S/(S+B) value of its selection category. Right: Depiction of the four-lepton invariant mass distribution for the $ZZ^* \to 4l$ analysis.

In addition to its discovery during the Run 1 data collection period (2010-2013), the investigation into the properties of the Higgs boson took place during the Run 2 phase (2015-2018) at a center-of-mass energy of $\sqrt{13}$ TeV. As of the latest developments,

all expected production modes, including VH [54] and ttH [55] during Run 2, have been conclusively observed. Within the Run 2 period, the most accurate value of the Higgs boson mass is provided as $m_H = 125.38 \pm 0.11(\text{stat}) \pm 0.08\,(\text{syst})$ GeV [56].

The agreement with the SM predictions for production modes (left) and decay channels (right) with associated signal strength parameters is depicted in Fig 2.6 [57]. With a given initial (i) and final (f) states, $\mathrm{i} \rightarrow \mathrm{H} \rightarrow \mathrm{f}$, the signal strength, denoted by μ, is directly proportional to σ_i and B^f, where σ_i represents the production cross-section, and B^f is the decay branching fraction. The fits are conducted based on distinct assumptions: signal strengths per production channel ($\mu_i = \sigma_i/\sigma_i^{SM}$ with $B^f = B^f_{SM}$) and signal strengths per decay mode ($\mu_f = B^f = B^f_{SM}$withσ_i/σ_i^{SM}). These plots, illustrating the SM results, are derived from the combination of findings from various Run 2 analyses. At the time of discovery, the common μ was found to be 0.87 ± 0.23. The new combination of all Run 2 data yields $\mu = 1.002 \pm 0.057$, in excellent agreement with the SM expectation. Perfect agreement with SM expectations would yield all μ equal to one.

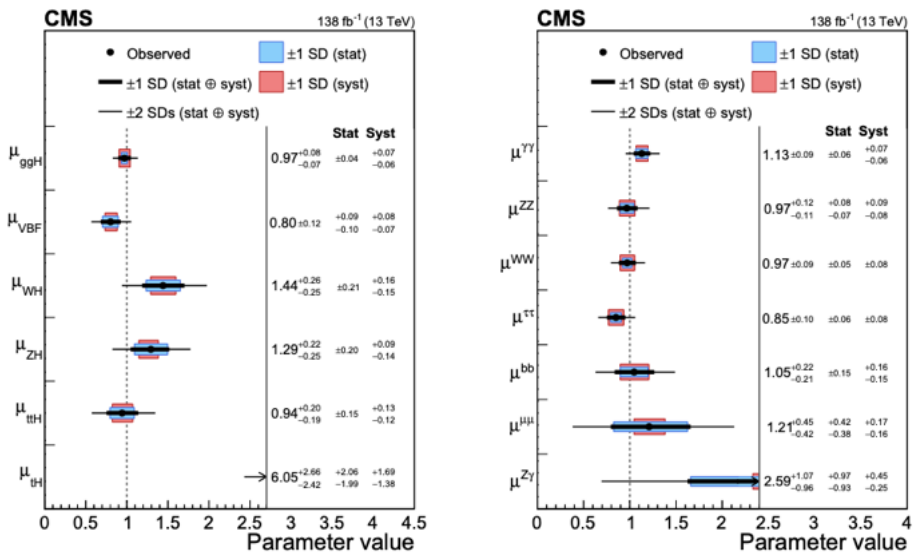

Figure 2.6: Parameters for signal strength are derived for various production modes (left) and for decay channels (right). The 1 (2) standard deviation confidence intervals are represented by thick (thin) black lines, with red and blue bands indicating the systematic and statistical components of the 1 standard deviation interval, respectively. The vertical dashed line at unity represents the values of μ_i and μ_f in the SM.

2.2 Beyond the Standard Model

As discussed in the previous subsections in details, the SM has proven itself as a reliable framework, successfully portraying various physical phenomena and accurately predicting the properties of elementary particles. Recent measurements at the LHC shows a convincing alignment with SM expectations, showcasing its effectiveness. Also, the discovery of the scalar Higgs boson with a mass $\approx$ 125 GeV at the LHC in 2012 and the on-going measurements of its properties thus far confirm this SM picture — within the current experimental precision. However, despite its successes, there still remain unanswered questions that prompt the consideration of BSM physics such that

- The SM explains mass through the Higgs boson, but questions remain regarding the nature of electroweak symmetry breaking.
- It also includes parameters determined by experimental measurements, and questions persist about the mathematical consistency of certain processes.
- Observations from cosmology, such as the dominance of dark matter (DM) and dark energy, highlight gaps in the SM's ability to account for the universe's total energy-matter density.
- The SM does not include gravity and fails to explain phenomena like neutrino oscillations, suggesting the need for a more comprehensive framework.

Addition to the points listed above, from the perspective of the Higgs boson, particularly its scalar potential, there exists a significant domain for anticipating effects of new physics in the BSM. Within this context, the Higgs field may establish connections with the DM sector through a mechanism known as a Higgs portal [58]. In addressing challenges such as the hierarchy problem, where the quadratic sensitivity of the Higgs mass parameter to the UV cutoff scale requires an "unnatural" fine-tuning of the bare mass parameter, BSM theories often involve modifications or extensions to the Higgs sector. An example of such a theory is Supersymmetry (SUSY), where supersymmetric versions of the SM incorporate at least two Higgs doublets.

In a more general sense, the introduction of new physics within the Higgs boson perspective can result in three observable effects:

- Alterations to the properties of the 125 GeV Higgs boson, encompassing changes in couplings, decay rates, and CP-properties,
- The potential existence of additional electrically neutral or charged scalar bosons,
- The initiation of interactions involving the Higgs boson (alongside other scalar bosons) with other novel particles introduced in the BSM theory, such as supersymmetric particles.

The further exploration of BSM physics, particularly at the LHC and HL-LHC (Section 3.4), offers a promising play ground to address the open questions and anomalies in the current understanding of particle physics. Whether through the detailed study of the Higgs boson, the discovery of new particles, or insights from neutrino experiments, the pursuit of BSM physics promises for a more comprehensive framework to address the observed limitations and provide a more complete understanding of the fundamental forces in the universe.

CHAPTER 3

CERN LARGE HADRON COLLIDER AND THE COMPACT MUON SOLENOID EXPERIMENT

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [44, 45] is a circular proton-proton collider that is 27 km in length and is hosted by the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. It is located at the French and Swiss borders and is situated at a depth of 45 m to 175 m below the surface. Compared to its predecessor, the Tevatron, the LHC provides much higher center-of-mass-energy and event rate. As a result, with the ability to operate at a maximum center-of-mass-energy of 14 TeV, the LHC can improve the precision of the SM measurements and investigate fundamental questions by revealing the nature of new BSM physics. The LHC uses a series of superconducting electromagnets to accelerate protons or heavy ions as two separate beams traveling in opposite directions at close to the speed of light around the LHC ring and colliding them at four different interaction points.

The final energy achieved by the LHC is the result of a series of machines with progressively increasing energies. At each step in the process, a beam of particles is accelerated to a specific energy and then injected into the next machine, where the beam gains more energy. The machines in the chain, through which the beam travels are Linac 2, Proton Synchrotron (PS) Booster, PS, and the Super Proton Synchrotron (SPS). The corresponding energies provided to the beam at each machine are 50 MeV, 1.4 GeV, 25 GeV, and 450 GeV, respectively.

After being injected into the LHC beam pipes, the two beams circulate and collide at four distinct interaction points, each equipped with a unique detector. These detec-

tors are known as CMS [59], ATLAS [60], LHCb [61], and ALICE [62]. CMS and ATLAS are general-purpose detectors designed to study a diverse range of physics phenomena, including the SM measurements, Higgs physics, and searches for new physics. Conversely, the LHCb detector is specialized in the study of antimatter, while the ALICE experiment focuses on heavy-ion collisions.

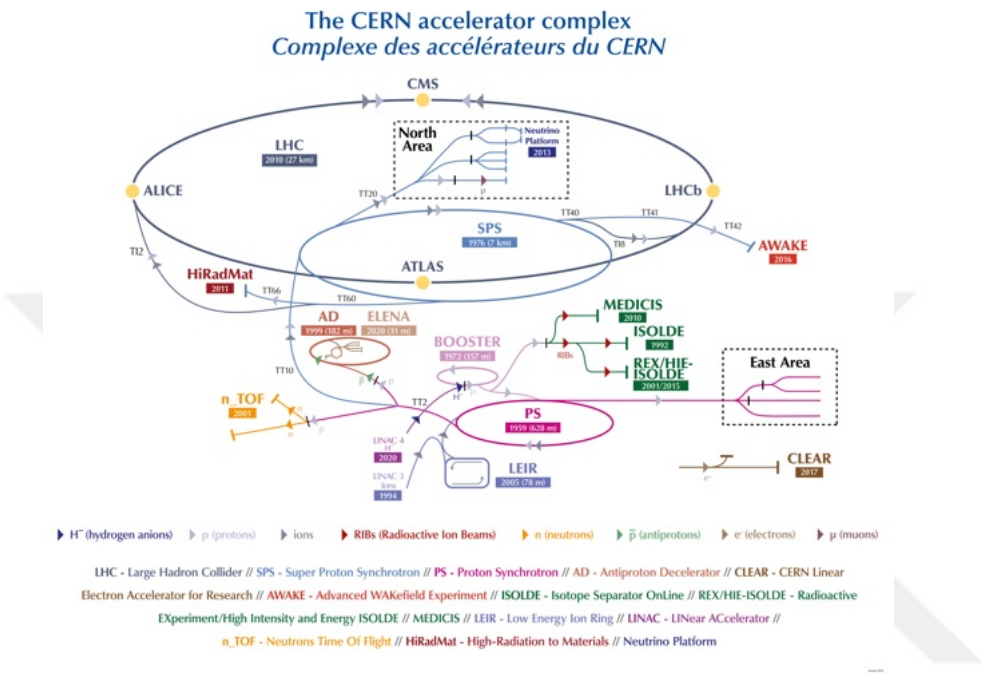

Figure 3.1: The CERN accelerator complex.

Number of events per second obtained from the LHC collisions is given by

$$N_{event} = L\sigma_{event} \tag{3.1}$$

where σ_{event} is the cross section for the event considered and *luminosity (L)* gives the quantitative definition of the particle density delivered by the LHC. Instantaneous luminosity ($\mathcal{L}$) only changes with the beam parameters

$$\mathcal{L} = \frac{{N_b}^2 n_b f_{rev} \gamma_r}{4\pi\epsilon_n \beta^*} F \tag{3.2}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam,

f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* is the value of the beta function at the collision point, and F stands for the reduction factor due to the beam crossing angle. Then, the integrated luminosity can be obtained by

$$L = \int \mathcal{L} dt. \tag{3.3}$$

The LHC is designed to collide protons with a peak luminosity of 10^{34} cm^{-2}s^{-1}, which is the highest luminosity achieved so far in particle accelerators. The LHC can also collide heavy ions such as lead (Pb) with a maximum luminosity of 10^{27} cm^{-2}s^{-1}, which corresponds to a collision energy of 2.8 TeV per nucleon. When colliding heavy ions such as lead, the collision energy per nucleon is lower than in proton-proton collisions, but the total collision energy is higher due to the large number of nucleons in the lead ions. This allows for the study of high-density nuclear matter and the formation of a quark-gluon plasma, a state of matter that is thought to have existed in the early universe.

The LHC is constructed using cutting-edge electronics, state-of-the-art design, and advanced materials. Over 1500 magnets with various functions are positioned in the former LEP tunnel to be able to meet the required specifications [63]. Niobium-titanium (Nb-Ti) superconductors are used in the accelerator magnets to bend 7 TeV proton beams, which are designed to operate at a field strength of 8.3 T (Figure 3.2). These magnets achieve the desired magnetic field by using superfluid helium cooling down to 1.9 K. While Nb-Ti superconductor cannot go beyond 9 T, many of these magnets have been tested at fields up to that ultimate level. The bending is primarily accomplished through 1232 dipole magnets, while over 300 quadruple magnets are employed for beam focusing. Compact two-in-one magnets are used in order to reduce cost and improve performance within the limited space of the accelerator ring [64].

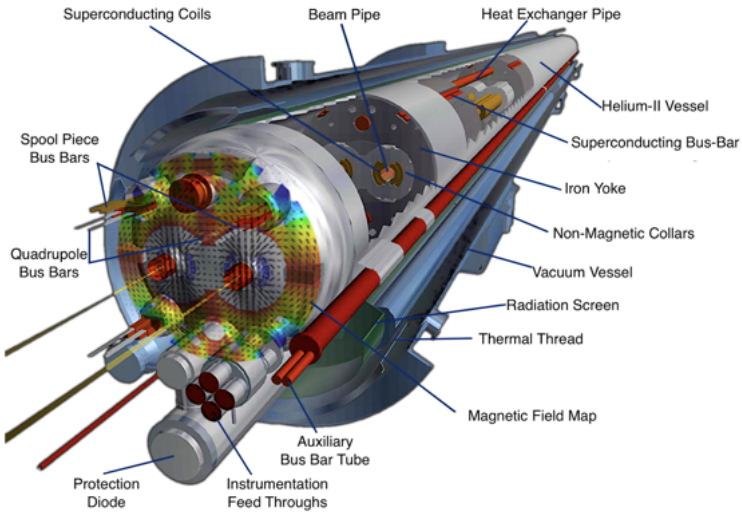

Figure 3.2: The 15 m long superconducting dipole magnets for the LHC at CERN.

3.2 The Compact Muon Solenoid detector

The CMS detector [59, 65, 66] is situated 100 m underground in an experimental cavern at Point 5 of the LHC close to the French village Cessy, and is one of two multi-purpose detectors at the LHC. The LHC's event rate necessitates low-latency measurements and high-granularity readout, which CMS achieves using quick electronics and a multi-layered, multi-channel detector structure. The CMS detector is designed as a cylinder and is composed of a central barrel region and two endcaps situated at each end of the barrel. The subdetector systems are arranged in layers around the collision point. The primary component of the CMS detector is the superconducting solenoid, which generates a high magnetic field to ensure accurate momentum and charge measurements for the tracked particles. At the interaction point (IP) of CMS, the beams are squeezed 16.7 µm and collide at a crossing angle of 285 µm. The resulting charged particles are bent using a homogeneous magnetic field of 3.8 T produced by this superconducting solenoid coil, which encloses only one fifth of the CMS detector's volume. The remaining volume is filled with a large muon system mounted in a surrounding iron return yoke, and outside of the coil, the magnetic field points in the opposite direction with a magnitude of 2 T.

The CMS physics program has challenging goals, including precise measurements

for SM, exploring the BSM physics, investigating TeV scale physics, understanding the origin of electroweak symmetry breaking, searching for dark matter, and detecting high-energy ion collisions. To achieve these objectives, CMS needs a capable detector that can measure particle interactions with high precision and speed. These requirements can be broken down into several specific areas, including efficient online triggering, good momentum resolution and identification of charged particles, good muon identification and $\mathrm{p_T}$ resolution in a wide spatial range, high electromagnetic energy resolution, good photon and electron identification, wide HCAL coverage, and hadronic shower reconstruction. The CMS detector fulfills these requirements through its multi-layered structure, which comprises different detector subsystems. The type, size, and location of these subsystems are determined by the physics requirements and radiation tolerance. Table 3.1 summarizes the main parts of the detector that are necessary to detect and identify particles in the SM. In addition, the tracking paths of these particles are shown in detail in Figure 3.3 [67], while an overview of the CMS detector [68] can be seen in Figure 3.4.

Table 3.1: The essential components needed for detecting and recognizing the SM particles with their distinct signatures are the primary subsystems. While these subsystems offer a reasonably detailed description of the physics entities individually, the CMS collaboration employs a unified approach described in Chapter 4 for reconstructing these particles.

Main subsystem	Particle	Signature
ECAL, HCAL, and tracking system	quarks	jets
ECAL and tracking system	e, γ	electromagnetic shower
ECAL and HCAL	neutrinos	missing transverse energy
Muon absorber and detectors, and tracking system	μ	ionization
vertex and tracker	c, b	secondary decay vertices

Due to the near-symmetry of the CMS detector around the beam axis, the CMS coordinate system is defined in cylindrical coordinates with the z-axis aligned along the counter-clockwise direction of the beam. The polar angle θ is measured from the z-axis, while the azimuthal angle ϕ is measured from the x-axis in the (x,y) plane. The pseudorapidity (η) coordinate is preferred over the longitudinal angle θ because differences in η are Lorentz-invariant. This property is particularly useful in hadron colliders where the boost in the z direction is unknown and varies for each collision.

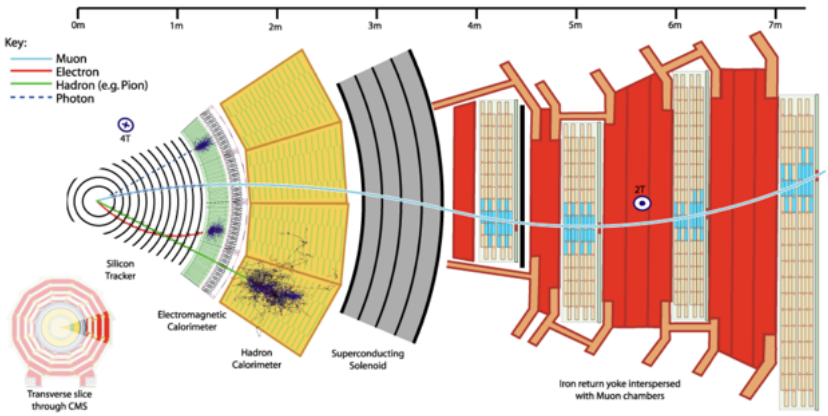

Figure 3.3: Slice view of the CMS detector showing how various particles interact with its subsystems. Muons are detected in both the tracker and muon stations before leaving the detector, while electrons leave hits only in the tracker and deposit their energy in the ECAL. Photons are identified as energy deposits in the ECAL without a corresponding track, while both charged and neutral hadrons deposit their energy in the HCAL, with matching tracker hits for charged hadrons only.

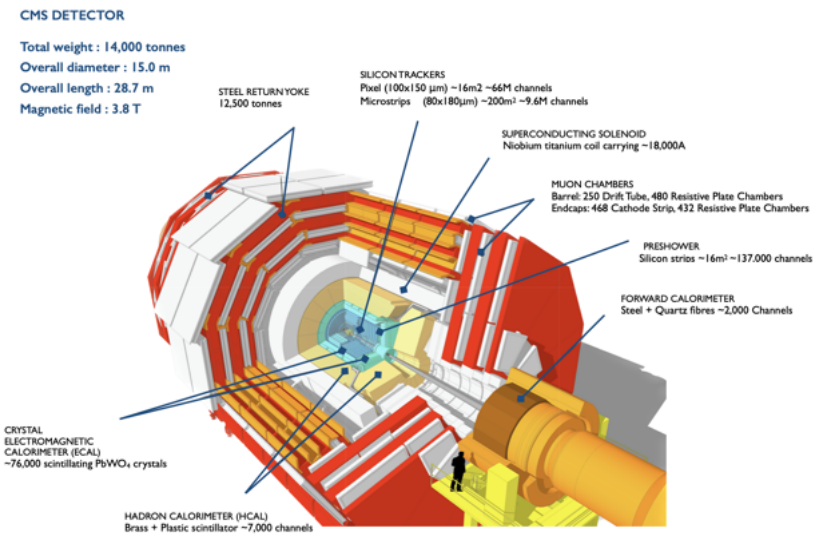

Figure 3.4: An overview of the CMS detector, with a person included for scale reference. The different subsystems of the detector are also labeled in the figure.

By using η instead of θ, it is possible to define a Lorentz-invariant angular separation between two particles, which is helpful for analyzing collisions in these types of colliders. η is defined as

$$\eta = -ln[tan\frac{\theta}{2}] \tag{3.4}$$

Also, a Lorentz-invariant angular separation between two particles is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.5}$$

Pseudorapidity is a specific form of rapidity, which is a measure of relativistic velocity that is Lorentz-invariant under longitudinal boosts. Rapidity can be considered as a hyperbolic transformation of velocity. When particle masses are negligible, rapidity reduces to η [44].

3.2.1 Inner tracking detector

The CMS tracking detector is a massive silicon tracking system that includes both a pixel detector and a silicon strip detector, each of which is enclosed by its own endcaps. Covering a range of $|\eta| < 2.5$, this system is highly advanced, offering precise measurements of charged particle trajectories and enabling the reconstruction of secondary vertices.

The pixel tracking detector is situated closest to the interaction point and is tasked with measuring the origins of tracks with exceptional precision. This is particularly important for tagged jets that originate from b quarks, distinguishing between prompt electrons and converted photons, and handling pileup events (defined in Section 4.2) by reconstructing distinct vertices. Because the sensors and front-end readout electronics are exposed to high radiation levels and a substantial flux of charged particles around the beam axis, radiation hardness is a crucial consideration. The pixel detector comprises three concentric cylindrical layers in the barrel part and two endcap disks on each side, with barrel cylinders that are 53 cm long and have radii of 4.4 cm, 7.3 cm, and 10.2 cm. The endcaps are rings with an inner radius of 6 cm and an outer radius of 15 cm, positioned at $\pm$34.5 cm and $\pm$46.5 cm along the beam axis relative to the center of the coordinate axis. The barrel consists of 768 pixel modules, while the endcaps contain 672 modules; each pixel has a surface area of 100×150 m^2. To

avoid any uncovered regions, the blades carrying the pixel modules in the endcaps are arranged in a windmill-like structure, as shown in Figure 3.5 [69]. As a result of its proximity to the interaction region, the pixel tracker is exposed to the highest particle flux in the detector, reaching up to 107 Sv at r = 10 cm.

The silicon strip detector, which encloses the pixel detector, is composed of several subsystems, including two cylindric barrels: the tracker inner barrel (TIB) and the tracker outer barrel (TOB); as well as two endcaps on each side: the tracker inner disks (TID) and the tracker endcap (TEC). The TIB is 130 cm long, while the TOB is 220 cm long. In the transverse axis, the first TEC disks are positioned at a distance of z = ±120 cm from the detector center, with the outer-most disks at z = ±280 cm. The TID is located in the gap between the TIB and the TEC. In total, the silicon strip tracker comprises 15,400 modules and records tracks at an operating temperature of -20°C. The TIB has four layers of silicon sensors, with the inner two layers made up of so-called stereo modules that provide two-dimensional measurements in both ϕ and z. The first two layers of the six TOB layers are also made of stereo modules, similar to the TIB. Each endcap on either side consists of nine disks (TEC) as well as three smaller disks called TID. Stereo modules are used in the first two rings of the TID and the first, second, and fifth ring of the TEC [59].

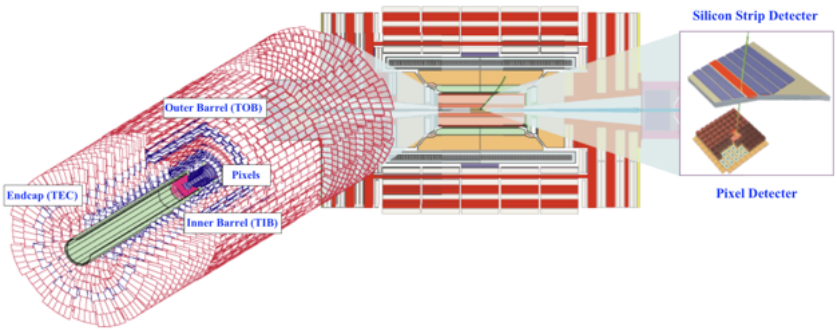

Figure 3.5: The layout of the CMS tracker system.

3.2.2 Electromagnetic Calorimeter

The CMS electromagnetic calorimeter [70] is a high-performance detector that lays between the tracker and the HCAL. It is made up of a central barrel (EB) and two endcaps (EE). By providing fine granularity and radiation resistance, it uses around

75,000 lead tungstate crystals ($PbWO_4$), arranged in a way to prevent cracks. The coverage zone of the EB is $|\eta| < 1.479$, whereas the EE covers the range of $1.479 < |\eta| < 3.0$.

A preshower detector (ES) has been installed inside the forward ECAL to be able to enhance the performance of ECAL. This is to address an issue caused by neutral pions, which can produce photons so close together that the ECAL may mistakenly label them as a single high-energy photon. Since the search for the Higgs boson was a key objective of the CMS design, and $H \rightarrow \gamma\gamma$ was a possible discovery channel, the ES was added to reduce the number of π^0 particles misidentified as high-energy photons. The preshower detector is designed as a type of sampling calorimeter. It is made up of two planes of lead radiators, which trigger showers when an electron or photon travels through them. The sensors located behind each lead plane measure both the energy deposited and the transverse shower profiles. This granularity enables the identification of closely spaced photons from π^0 and other di-photon resonances.

The lead tungstate crystals located in the ECAL scintillate when electrons or photons pass through them. The scintillation light is detected by avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps, which convert it into an electrical signal. The decay time of the scintillation light is short, with approximately 80% of the light emitted within 25 ns, making it suitable for the LHC bunch spacing.

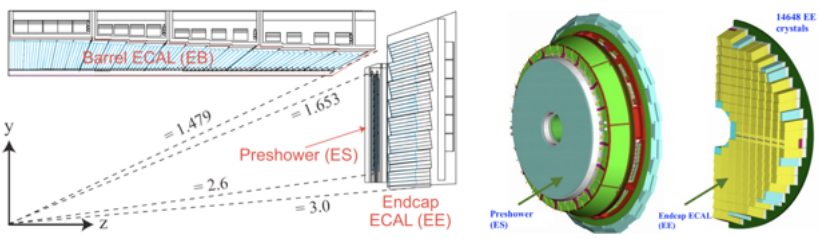

Figure 3.6: Cross section of an ECAL detector quadrant and the layout of the ECAL Endcap Calorimeter. The blue rectangles (seperated by the thicker black lines) show the PbWO4 crystals (supercrystals).

3.2.3 Hadron calorimeter

The hadron calorimeter (HCAL) [71] in the CMS detector is positioned radially between the ECAL and the magnet, covering an azimuthal range of $|\eta| < 5.2$. It consists of four parts; the barrel (HB), endcap (HE), outer (HO), and forward calorimeter (HF). Its main purpose is to measure hadron jets and estimate missing transverse energy (discussed in detail in Chapter 4) resulting from neutrinos or other new physics processes. The HO helps ensure hadrons that partially escape detection in the HB and HE are still measured, while the HF measures the forward region of CMS at 3.0 $< |\eta| < 5.0$.

The barrel and endcap subsystems of the HCAL use a sampling calorimetry approach, consisting of dense absorber material (typically brass or steel) alternating with fluorescent plastic scintillator tiles. When a hadronic particle collides with an absorber plate, it generates secondary particles that cascade through successive layers of absorber material, leading to a shower of particles. As the shower develops, the particles travel through layers of scintillators, causing them to emit blue-violet light. This light is captured by optical fibers that shift the light into the green region and transport it to readout boxes via clear optical cables. The collected light intensity provides a measure of the energy of the passing particle. Hybrid photodiodes then convert the optical signals into rapid electronic signals, which are transmitted to the data acquisition system.

In order to fully absorb a shower of particles, it typically requires approximately one meter of absorber material. However, due to the compact design of the CMS detector, it was not feasible to accommodate the entire barrel of the HCAL within the magnet coil. As a solution, the outer barrel is positioned just behind the magnet coil. This placement helps to maximize the absorption of punch-through particles, rather than generating spurious hits in the muon system.

The CMS detector includes two forward calorimeters, known as the HF, which are located at either end of the detector and extend the coverage of the HCAL up to $|\eta| =$ 5.0. The HF receives the majority of the particle energy from the collision and thus must be highly resistant to radiation. In order to withstand the high levels of radiation

in the forward region, the HF utilizes a technology based on Cherenkov radiation. To convert the light collected by the fibers into electrical signals, photomultiplier tubes (PMT) are employed.

3.2.4 Muon system

One of the primary objectives of the Compact Muon Solenoid is to detect muons, which are the more massive counterparts of electrons. Due to their weight being 200 times greater than that of electrons, muons interact with matter less, allowing them to penetrate several meters of iron and traverse the calorimeter systems without being absorbed. As a result, the muon chambers are positioned behind all other subdetectors, where they can register a signal as they are the only particles capable of doing so. Furthermore, this unique property of muons makes it difficult to misidentify them as any other particle, thus providing an exceptionally clear signature. The CMS detector owes its name to the muon system [72], which constitutes 80% of its volume and is positioned as the outermost detector. The muon system is enclosed by an iron yoke that serves the purpose of returning the magnetic flux from the solenoid, thereby producing a magnetic field of 2 T. The muon system consists of four subsystems.

Drift tube system has gas-filled containers that have a wire running through them. When charged particles pass through the gas, they produce free electrons that move towards the wire and get recorded. The CMS DT chamber is made up of twelve aluminum layers divided into three sets of four layers, called superlayers. Each superlayer contains up to sixty DTs, each of which is 4 centimeters wide. These DTs measure approximately 2 meters by 2.5 meters and are used to detect particles.

Cathode strip chamber system is utilized in the endcap disks of the CMS detector, where the magnetic field is uneven, and particle rates are high, owing to its higher radiation resistance compared to DTs. CSCs, or multiwire proportional chambers, are composed of positively charged anode wires and negatively charged copper cathode strips present in a gaseous space. Each of the 468 CSCs is trapezoidally shaped and works by ionizing the gas whenever muons pass through them. This ionization generates electrons that move towards the anode wires and positive ions that travel towards the copper cathode. The CSCs are quick, and since the wires are placed very

close together, they can be employed as input to trigger decisions, apart from their primary function of accurately measuring muons.

Resistive plate chamber system measures muons with a good spatial and timing resolution within a short period (less than 25 ns) for $|\eta| < 1.6$. RPCs employ a parallel-plate gaseous detector design that enables them to measure muons quickly, making them suitable for use in a dedicated muon trigger.

Optical alignment system

The alignment of the muon chambers and the central tracker in the CMS is crucial for accurate muon measurements, but it can be affected by various factors such as constructional tolerances, magnetic field distortions and time dependent deformations. To ensure an alignment precision of $\sim$100 μm, an optical alignment system is used to position each subsystem and calibrate the measurements based on the collected data.

3.2.5 Triggers and data acquisition system

The data output rate of the CMS detector is 40.078 MHz per bunch crossing (24.95 ns) [44, 45, 59], which exceeds the storage capacity of existing technologies. To reduce the data to a manageable size, a two-step triggering system is implemented. The first step is the Level$-$1 trigger (L1) and the second step is the High Level Trigger (HLT). L1 trigger is designed to reduce the rate of events accepted for further processing to less than 100 kHz, whereas HLT is designed to reduce this maximum L1 accept rate of 100 kHz to a final output rate of 100 Hz [73].

Level-1 trigger and readout electronics

The CMS L1 trigger system aims to quickly reduce data and transfer it to higher trigger systems with high reliability using mostly subsystem-specific systems. It operates at a remarkably quick and entirely automatic pace, searching for basic indicators of intriguing physics, such as particles carrying a significant amount of energy or appearing in unique combinations. This is accomplished by using fast electronics that can make individual decisions as low as sub-nanoseconds. The L1 trigger system employs two kinds of electronics technologies, ASICs (Application Specific Integrated

Circuits) and FPGAs (Field Programmable Gate Arrays). ASICs are radiation-hard and designed for specific tasks, making them more reliable and faster compared to other solutions, but they are expensive to replace, upgrade, or debug. FPGAs are more flexible, but are prone to radiation damage due to their logic block density and the necessity of internal storage units. Currently, both FPGA and ASIC systems are used together in the CMS trigger and data acquisition systems, with ASICs usually installed behind detectors and FPGAs placed in control rooms close to detectors. The L1 trigger system specializes in making local decision instances for each sub-detector, which are then combined to form regional trigger primitives. The Calorimeter Trigger uses information from the ECAL and HCAL to determine transverse energy, missing transverse energy, jets, jet multiplicities, and the timing of events, while the Muon Trigger matches and connects different segments of the muon system and calculates the tracks of muons, providing good momentum and timing resolution of particles. The Global Trigger has a five-level structure (input, logic, decision, distribution, and read-out) and uses information from all sub-detector trigger systems to decide whether an event will be accepted or not. After the decision, the L1 accept (L1A) signal is distributed, and accepted events are sent to the HLT with an output rate of approximately 40 kHz (with a maximum of 100 kHz) [73,74].

High-level trigger and data acquisition system

The main purpose of the High-Level Trigger (HLT) is to further reduce data by filtering events. This process has two main stages: to reconstruct physics objects and to mark events with interesting features. The HLT requires reading events at the pace of the L1 trigger output and, therefore, requires massive parallelism with a huge amount of computer power. The entire HLT process is maintained by a computer farm with more than 9,000 processor cores working at high frequencies, where the current output of the HLT is $\approx$1 kHz for Run 2.

The Data Acquisition (DAQ) system, in addition to retrieving information from the CMS detector, serves as the primary inspection point for all data from the physics collisions. It is at this juncture that intricate decisions about the neccesity of each event approved by the L1 trigger are made. This system also offers the unique opportunity to monitor the complete response of the detector to the collisions, providing

immediate feedbacks. The DAQ system is thus critical in CMS, as it executes two essential functions that define the scope of the physics program: selecting events, and overseeing and managing the CMS detector components. More specifically, when the L1 trigger sends an accept signal (L1A), approximately 700 detector front-end drivers save the data from the detector's front-end electronics. This data is then read by the readout system and held until it is forwarded to the HLT for additional processing. The builder network connects the readout and HLT systems and is made up of 500 builder units, each assembling the data for one event. Once assembled, the event is transferred to the HLT processor associated with the specific builder unit, and the HLT decision is then sent back to the builder unit. The builder unit either discards the event to free up memory or sends it to the storage manager to be written to disk [73, 74].

Figure 3.7, adapted from [75] gives the overall architecture of the CMS Trigger and DAQ System with the trigger rates at each steps. More details on the CMS trigger system can be found in [74].

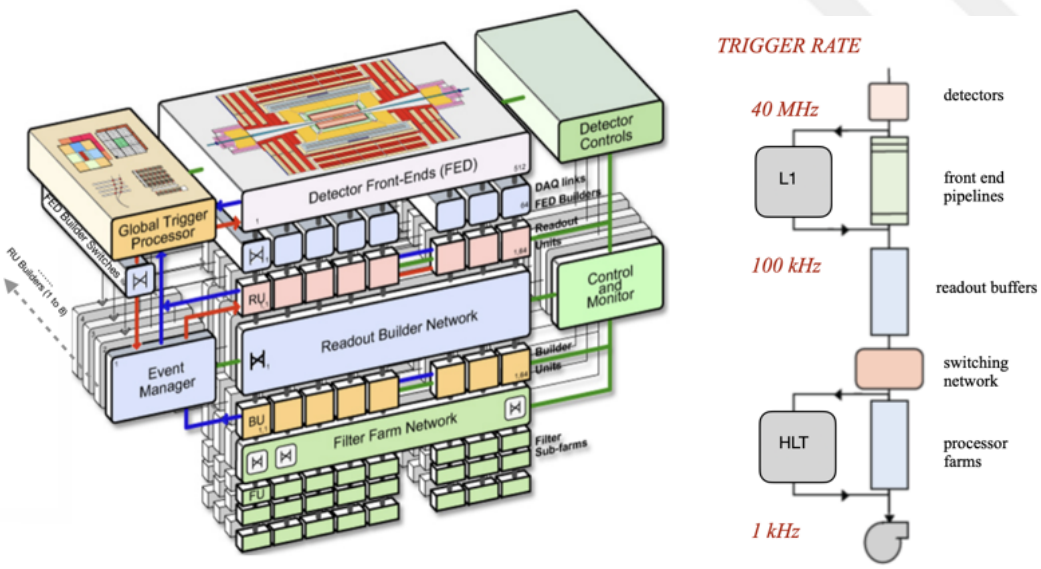

Figure 3.7: General architecture of the CMS Trigger and DAQ System (left) and simple chart of the system with the trigger rates at each step (right).

3.2.6 Offline computing

The CMS Computing Model is structured similarly to the detector and consists of multiple layers [76]. The model employs various data formats, including the DAQ RAW, RECO, AOD, and GEN. The DAQ RAW contains information collected from ASICs and FPGAs, including L1 trigger results. It is made up of reduced data from computer farms and includes L1 and HLT selection information, which is reconstructed by the event reconstruction algorithms. The RECO, or Reconstructed Data, contains information on the reconstructed objects, hits, and clusters. The AOD, or Analysis Object Data, is a reduced version of the RECO data and only includes the information required for physics analysis. Finally, the GEN data are generated Monte Carlo simulated events (explained in the next section) without detector simulation.

The CMS Computing Model has four hierarchical tier levels, with Tier-0 located at the CERN site, where raw data is stored and distributed. Tier-0 also applies the first reconstruction steps. The Tier-1 level performs reconstruction, skimming, and calibration steps, while also providing a secure mirror of the raw data. Tier-2 and Tier-3 offer local services and global grid distribution of the reconstructed data, in addition to managing overall Monte Carlo sample generation for the experiment.

3.3 Monte Carlo event simulation in proton-proton collisions

Monte Carlo (MC) simulation is an effective way to both model and understand complicated systems by using random numbers, finding broad application across diverse fields and disciplines. The working principle involves executing a substantial number of experiments or simulations with random inputs to analyze the resulting output. Further statistical analysis of these outcomes yields valuable insights into the system's behavior and makes the estimation of unknown quantities possible [77]. In particle physics, this method is widely used to model and study various physical systems. For instance, in order to perform precise measurement studies of the SM or data analysis by comparing observed particles at the LHC collisions with their corresponding theoretical predictions, MC simulated events act as the key elements. In addition to these, the accurate simulation of events for various physics processes is vital for tasks such

as calibrating the detector, fine-tuning the trigger selection, optimizing the physics analysis, modeling potential BSM interactions.

MC simulations are also often used in conjunction with Feynman diagrams, graphical representation of particle interactions, and cross section calculations at different orders of accuracy, which are described through leading order (LO) and next-to-leading order (NLO) contributions. LO in a calculation is the term with the fewest number of vertices, or interaction points, in the Feynman diagram and it is the simplest contribution to a particular process. The NLO term, on the other hand, includes additional diagrams with an additional vertex to the LO term. Simulated events based on the LO matrix elements provide a probabilistic description of the possible outcomes of a given process, whereas at NLO these events are simulated by taking into account not only the LO processes but also higher-order corrections such as loop diagrams [78, 79]. Event simulations also depends on flavour schemes (FS) chosen, which refer to the number of quark flavours considered in theoretical calculations. The choice of flavour scheme significantly influences calculations and predictions. The samples considered in this thesis are generated using both the 4-Flavour Scheme (4FS) and the 5-Flavour Scheme (5FS). The 4FS includes calculations involving up, down, strange, and charm quarks. In contrast, the 5FS extends these calculations by incorporating the bottom quark in addition to those included in the 4FS. The 5FS offers a more comprehensive approach and is utilized in scenarios where the effects of the bottom quark are significant and cannot be ignored [80].

Tools such as PYTHIA8.2 [81], DELPHES [82], GEANT4 [83, 84], POWHEG [85] and the MADGRAPH5_AMC@NLO [79] generators are used to obtain these simulated events,. Each tool applies either the LO, the NLO or both corrections to the simulated events.

3.4 High Luminosity Large Hadron Collider

During the initial operational phase of the LHC in 2011 and 2012, called Run 1, the LHC achieved a peak luminosity of 7.7×10^{33} cm^{-2}s^{-1}, surpassing 75% of its projected luminosity. It also provided approximately 25 fb^{-1} of integrated luminosity

to its two main components, ATLAS and CMS experiments. This period was marked by numerous physics results and the most significant outcome from this data was discovering the Higgs boson in 2012. Between 2015 and 2018, in the next phase called Run 2, the LHC operated at a 13 TeV center-of-mass energy. Working more efficiently, bunch spaces was halved to approximately 25 ns from the previous value 50 ns, and a peak luminosity of 2×10^{34} cm^{-2}s^{-1} was recorded in 2018. This value is known as the maximum achievable peak luminosity of the Run 2 period due to the limitations on the detector capabilities. The ongoing Run 3 phase, from 2022 to 2024, aims to further elevate the total integrated luminosity, targeting 350 fb^{-1} in the end, thereby exceeding the initial goal of approximately 300 fb^{-1} [86].

Since the LHC started, there are still many new theories in physics to explore. These include precision measurements of rare decays accurately predicted by the SM, as well as the explorations into the BSM physics, and supersymmetry. To maximize the physics potential of the LHC, CERN has initiated the High-Luminosity LHC (HL-LHC) project, with the following objectives:

- Achieving a peak luminosity of 5×10^{34} cm^{-2}s^{-1} through levelling operations
- Achieving an integrated luminosity of 250 fb^{-1} per year, with an aim of reaching 3000 fb^{-1} over approximately 12 years after the upgrade.

The performance that the HL-LHC aims is almost 10 times greater than what was initially expected for the LHC in its original design. If the HL-LHC can exceed its planned performance levels, shown in Figure 3.8 and if the updated detectors can handle a higher number of PU, averaging up to 200, then it might be possible to reach a peak luminosity of 7.5×10^{34}cm^{-2}s^{-1}. This would be about four times the highest luminosity achieved during Run 2 [86].

Significant updates to the LHC and its experiments require access to the accelerator tunnels and experimental areas, which can only be done effectively during extended shutdowns. Therefore, in addition to the data collection periods mentioned earlier, the LHC schedule includes periods of long shutdowns. These are labeled as LS1, LS2, and LS3, as shown in Figure 3.9.

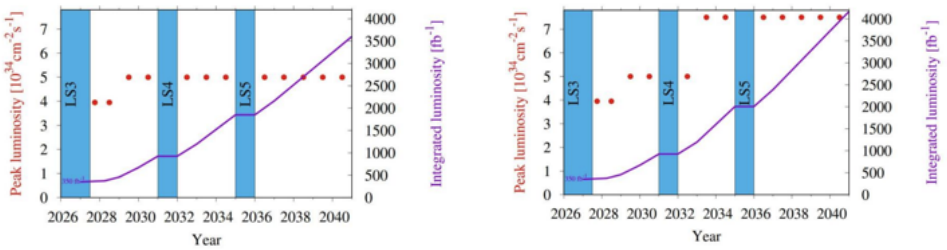

Figure 3.8: Projected peak and integrated luminosity in the HL-LHC for the nominal (left) and ultimate (right) detector parameters. Red dots represent peak luminosity predictions, while the violet line indicates total integrated luminosity over time. Forecast for peak luminosity (red dots) and integrated luminosity (violet line) in the HL-LHC era with nominal (left) and ultimate (right) HL-LHC parameters.

Figure 3.9: The LHC's planned roadmap for the upcoming decade and beyond. It illustrates both the collision energy (depicted by the upper line) and luminosity (represented by the lower line). During LS2, there are key developments including the consolidation of the LHC, excavation for the High-Luminosity LHC (HL-LHC), as well as enhancements to the LHC injectors and the Phase 1 upgrade of the LHC detectors. Following the completion of LS3, the LHC will transition into its high-luminosity configuration.

The CMS detector will undergo significant upgrades to exploit the enhanced physics potential due to increased luminosity and to handle the challenging operational conditions of the HL-LHC [87–91] in 2025. In particular, to be able to maintain the increased PU rate and associated increase in flux of particles, these upgrades will include improvements in granularity to manage channel occupancy, increased data bandwidth, and enhanced trigger capabilities to maintain manageable trigger rates without compromising physics potential. Enhanced radiation resistance is also a key

focus of these upgrades.

The L1 trigger will be upgraded to support higher rates and latency, up to approximately 750 kHz and 12.5 μs, respectively. The upgraded L1 will incorporate input from the silicon strip tracker, enabling real-time track fitting and particle-flow (PF) reconstruction [92] at the trigger level. The HLT aims to reduce the event rate by a factor of about 100 to 7.5 kHz.

Both pixel and strip tracker detectors will be completely replaced. These upgrades aim to enhance granularity, lower the budget by minimizing the material inside the tracking volume, improve radiation resistance, extend geometrical coverage, and ensure efficient tracking up to $|\eta| = 4$. The tracker will also provide information on high-momentum tracks to the L1 trigger, a function currently reserved for the HLT, and will track low-momentum particles, i.e., around 2 GeV.

The muon system will receive upgrades to the electronics of the CSC, RPC and DT. New muon detectors based on improved RPC and gas electron multiplier (GEM) technologies will add redundancy, extend coverage to $|\eta| = 2.8$ and enhance trigger and reconstruction performance in the forward region.

The barrel ECAL will operate at lower temperatures to reduce noise from radiation damage. Its front-end electronics will be upgraded to handle the demands of the L1 trigger, including increased sampling rates and precision timing capabilities. The barrel and endcap HCAL will feature silicon photomultipliers (SiPMs) for readout.

A new combined high granularity calorimeter (HGCal) will replace the existing endcap electromagnetic and hadron calorimeters, extending coverage from $|\eta| = 1.5$ to $|\eta| = 3$. The redesigned calorimeter will feature a construction combining lead tungsten and stainless steel absorbers, with silicon sensors positioned in the front section to serve as the primary active material. Toward the back section, at greater distances from the beam, the calorimeter will incorporate plastic scintillator tiles, which will be read out using SiPMs. This design will enable the calorimeter to deliver detailed spatial data in both the transverse and longitudinal planes. Additionally, it will be equipped with a 320 MHz sampling rate, enhancing its timing accuracy for photons. This high-precision timing, along with its structural design, will significantly enhance

the calorimeter's ability to distinguish between different types of particles, like electrons, photons, tau leptons, and jets, and to more effectively manage and reject PU.

Furthermore, a new precision timing detector for minimum ionizing particles (MTD) is planned for both barrel and endcap regions. This detector aims to achieve timing resolutions between 30 to 40 ps will extend coverage to $|\eta| = 3$. This addition is expected to significantly enhance the CMS's ability to reconstruct interaction vertices in four dimensions, countering the performance challenges posed by high PU rates.

A comprehensive summary of the CMS detector upgrade and the expected performance improvements of the physics objects (electrons, photons, taus, jets, and missing energy) is provided in [87–91, 93–96].

CHAPTER 4

OBJECT RECONSTRUCTION

The CMS detector functions as a particle flow detector, merging data from several subdetectors to figure out particle kinematics and types through a method called reconstruction. Muons leave traces in the tracker, small energy accumulations in the calorimeters, and exit with traces in the outer muon systems. Electrons and photons deposit energy in the ECAL, but the trajectory of electrons in the tracker provides additional information with the path of electrons in the tracker giving more data. Jets leave signatures in all subdetectors, while particles interacting weakly, such as neutrinos, are not directly seen by the CMS detector. However, their presence can be inferred from the observed momentum imbalance in the transverse direction.

4.1 Particle-flow algorithm

The CMS detector is adept at employing the particle-flow (PF) algorithm [92, 97] due to its detailed tracking system and high granular ECAL. This algorithm seeks to enhance object reconstruction by making use of the information gathered from different components of the detector. The CMS PF algorithm consists of three main elements:

- Tracks reconstructed from hits in the tracker,
- Tracks reconstructed from traces in the muon system,
- Energy deposits in the calorimeters.

Elements possibly linked to a single object are connected to create a basic module.

The algorithms for making these connections vary depending on the element type and the object in question. For instance, tracks in the muon system are adjusted to align with the internal tracks if the adjustment results in an acceptable χ^2 p-value. For electrons, tracks are linked with nearby ECAL clusters because of bremsstrahlung.

Rebuilding entities follows an order from the most likely to the least likely to be reconstructed, gradually removing unrelated elements from the less reliable reconstructions. Due to the clear conditions in the muon system, muons are the first to be reconstructed, followed by electrons, taus, photons, and hadrons.

4.2 Primary vertex

At a projected luminosity of $10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$, which is anticipated for the LHC, the CMS tracker is predicted to encounter roughly 1000 charged particles every time proton bunches cross at 25 ns intervals. These particles result from the overlapped particle interactions, called "pileup (PU)", in the detector. It is necessary to separate these overlapping interactions from the actual pp collision, which is aimed to analyze. Furthermore, due to the limited time resolution of the detector, PU can also arise from preceding or subsequent bunch crossings. The task of accurately reconstructing tracks in this cluttered setting is considerably difficult. There's a balancing act between maintaining a high success rate in identifying tracks and minimizing the occurrence of incorrect tracks. These incorrect or "fake" tracks could arise from unrelated hit combinations or from an inaccurately reconstructed real particle pathway that includes unrelated hits [98].

The Primary Vertex (PV) reconstruction [98] focuses on accurately pinpointing the location of the key collision event. To increase collision occurrences, hadron beams are tightly focused in the transverse plane at the CMS interaction point. Even with this increased density, the number of collisions follows a Poisson distribution, and only a small number of extra collisions occur alongside the main one. However, reducing the impact of these extra PU collisions can be tough due to constraints in spatial precision.

The PV reconstruction process starts with selecting a group of tracks. Various quality

conditions for tracks are applied, such as the transverse impact parameter, the number of tracker hits, and the normalized-track χ^2. Chosen tracks are sorted based on how close they are to the beam axis in the transverse direction. The PV is reconstructed using these track clusters with an adaptive vertex fitter. Each track gets a weight based on its closeness to the vertex candidate [99].

In line with other adaptive techniques, fluctuating tracks are not discarded, but are assigned lower weights. The algorithm keeps iterating until the position of the vertex candidate stabilizes at a predefined level.

4.3 Muons

Local reconstruction of the muons involves utilizing information from muon chambers such as Resistive Plate Chamber (RPC), Drift Tube (DT), and Cathode Strip Chamber (CSC) to map the path of a muon. Hits made by ionization in these chambers are used to determine the precise locations of hits. Different algorithms are used depending on the chamber technology. For instance, in a DT chamber, hit reconstruction relies on the transverse distance between the wire and the intersection of the muon trajectory. In a CSC layer, the muon's location is found by combining data from cathode strips and anode wires. Hit reconstruction in an RPC chamber involves clustering of hit strips. These reconstructed hits from the muon chambers are then used in combination with the inner tracker to reconstruct muon tracks in the CMS experiment. The process of reconstructing tracks in CMS for proton-proton collisions involves different techniques for different types of tracks [100]. **Standalone-muon tracks** use a Kalman-filter technique to gather information from all muon subdetectors, beginning with DT or CSC segments. **Tracker muon tracks** are created by propagating tracker tracks to the muon system and matching them to DT or CSC segments, as well as extrapolating tracker tracks to the muon system. **Global muon tracks** are formed by matching standalone-muon tracks with tracker tracks. The combined fitting of the Kalman filter is used, incorporating information from both the tracker track and standalone-muon track to refine the reconstruction.

Almost all muons produced within the muon system's acceptance are reconstructed

either as a global muon track or as a tracker muon track, often as both. Global muons and tracker muons that share the same tracker path are joined into a single candidate. The reconstructed muons are fed into the CMS PF algorithm [100].

The details of the muons used in this thesis are given in Sections 7.4.1 and 8.3.1.

4.4 Electrons

In the CMS experiment, the reconstruction of electrons [101] is done with high precision and efficiency. Electrons, which interact electromagnetically, leave a distinct signature in the ECAL, marked by isolated energy collections that are also linked with tracks in the silicon tracker. To reconstruct electron signals in the ECAL crystals, a fitting process is employed, using multiple template functions to remove the contribution from unwanted PU signals that occur at different times. As electrons move through materials before getting to the ECAL, they can undergo interactions, like emitting bremsstrahlung photons or turning photons into electron-positron pairs. Consequently, when an electron reaches to the ECAL, it may no longer be a single particle but instead a shower of multiple electrons.

To regain the energy of the main electron, a special algorithm combines clusters from the individual particles into one unit. Additionally, an electron's trajectory, changed by the emission of bremsstrahlung photons and resulting in changes in curvature, is precisely estimated using a special tracking method called the Gaussian sum filter (GSF). In the last step of electron reconstruction, the true amount of electron energy is determined by calculating the difference between the momentum of the outermost path and the innermost path segment. This data, along with the information collected through previous steps, is integrated into the overall reconstruction process of GSF electrons.

The electron reconstruction process within the CMS experiment is fully integrated into the PF framework and follows the same fundamental principles as the reconstruction of other particles.

The details of the electrons used in this thesis are given in Sections 7.4.2 and 8.3.1.

4.5 Jets

Jets are the experimental evidence of quarks and gluons produced in high-energy processes such as head-on proton-proton collisions. As quarks and gluons possess a net color charge and cannot exist in isolation due to a phenomenon called color confinement, first introduced in Section 2.1, they are not directly observable in nature. Instead, they combine and undergo a process called hadronization, resulting in the formation of color-neutral hadrons. The bunch of hadrons moves together like a spray and is referred to as a jet. In the CMS detector, these jets leave marks in various parts like the tracker and calorimeters. These traces are utilized in conjunction with jet algorithms to construct a reconstructed jet [102]. Jet reconstruction algorithms can be categorized and discussed under two main headings:

Cone algorithm

One can picture a jet as a cone-shaped spray of particles. The cone algorithm aims to reconstruct this spray by assuming a cone shape. The central line of this cone is taken as the direction of the jet, and the energy within this cone is considered as the jet's energy. There are two main types of cone algorithms. In the subset of cone algorithms, one particle is chosen as the midpoint, and cones with a predefined radius are constructed around it. However, these algorithms are not considered infrared and collinear safe (IRC safe) without the inclusion of certain thresholds. On the other hand, iterative cone algorithms begin with the hardest object as a seed and construct clusters within a predetermined cone. In the calculation of softer objects, these clusters are eliminated. Compared to the midpoint approach, iterative algorithms are more resilient against soft objects but are still not collinear safe.

To ensure IRC safety, cone algorithms can be implemented without the concept of a seed. One such practical implementation is the Seedless Infrared-safe Cone (SIScone) algorithm. It iteratively considers a subset S of all particles, allowing the construction of a cone that encapsulates only the subset S.

Sequential algorithms

Sequential clustering algorithms operate on the assumption that particles within jets

exhibit small variations in transverse momenta. As a result, these algorithms group particles based on their momentum characteristics, leading to jets with varying areas in the $(\eta - \phi)$ space. Similar to cone algorithm, sequential clustering algorithms are also designed to be IRC safe.

In sequential algorithms, there are two measures of distance that are important. The first one is the distance between two particles, and the second is the distance between the particle and the beam of protons. The algorithm combines particles into jets by considering these distances. If the distance between two particles is small, they are combined. If the distance between a particle and the beam is small, that particle is considered a jet.

The initial distance variable in sequential clustering algorithms refers to the measure of separation between two particles

$$d_{ij} = min(p_{ti}^a, p_{tj}^a \times \frac{R_{ij}^2}{R}) \tag{4.1}$$

where a is an exponent corresponding to a particular clustering algorithm, $\mathrm{R}_{\mathrm{ij}}^2 = (\eta_{\mathrm{i}} - \eta_{\mathrm{j}})^2 + (\phi_{\mathrm{i}} + \phi_{\mathrm{j}})^2$ is the $(\eta - \phi)$ space between two particles whereas R (radius parameter), within the range of 0.4 to 0.7, determines the final size of the jet.

The second distance variable, known as the momentum space distance between the beam axis and the particle detected is given by

$$d_{iB} = p_{ti}^a \tag{4.2}$$

Sequential clustering algorithms follow a step-by-step procedure to construct jets. The algorithm begins by identifying the minimum value from the set of distances $\{d_{ij}, d_{iB}\}$. If d_{ij} is the minimum, particles i and j are merged into a single particle (ij) by summing their four-vectors. Subsequently, particles i and j are removed from the list of particles. On the other hand, if d_{iB} is the minimum, particle i is labeled as a final jet and removed from the particle list. The algorithm repeats this process until one of two things happen:

- In inclusive clustering, the algorithm continues until all particles are incorporated into a jet, with the distance between the jet axes R_{ij} exceeding the pre-set radius parameter R.
- In exclusive clustering, the algorithm stops once the desired number of jets has been achieved. By employing this sequential clustering approach, jets are progressively formed, either by merging particles or designating individual particles as jets, until the desired criteria are met.

Despite minor differences, all sequential clustering algorithms (k_T, anti k_T, and the Cambridge/Aachen (C/A) algorithm) follow a similar approach [103].

For the k_T, the anti k_T, and the Cambridge/Aachen (C/A) algorithms, a parameter takes the value of 2, -2, and 0, respectively.

By substituting the value of 2 for the parameter a in Equation 5.1 and Equation 5.2, these equations can be modified as follows:

$$\begin{aligned} d_{ij} &= min(p_{ti}^2, p_{tj}^2) \times \frac{R_{ij}^2}{R} \\ d_{iB} &= p_{ti}^2 \end{aligned} \tag{4.3}$$

Following the equations above, the low $\mathrm{p_T}$ is dominant. As a consequence, the k_T algorithm exhibits a preference for clustering soft particles initially.

The a value corresponding to the anti-k_T algorithm, -2, resulting in the following equations

$$\begin{aligned} d_{ij} &= min\left(\frac{1}{p_{ti}^2}, \frac{1}{p_{tj}^2}\right) \times \frac{R_{ij}^2}{R} \\ d_{iB} &= p_{ti}^2 \end{aligned} \tag{4.4}$$

For the anti-k_T algorithm, high p_T values are dominant and hard particles are preferred to be clustered first. Figure 4.1 compares the topologies of the k_T and the anti-k_T algorithms [103].

Finally, for the C/A algorithm, if the a value is inserted, the above equations become

$$
\begin{aligned}
d_{ij} &= \frac{R_{ij}^2}{R} \\
d_{iB} &= 1
\end{aligned}
\tag{4.5}
$$

In this case, both of the distance variables are not depend on the momentum.

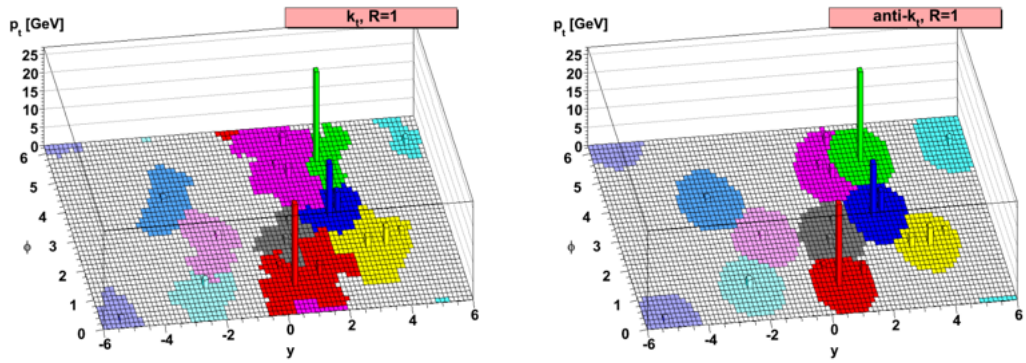

Figure 4.1: Comparison of the topologies of sequential kT and anti-kT jet algorithms. The anti-kT algorithm is characterized by cones with smoother boundaries compared to the kT algorithm. Moreover, the cones in the anti-kT algorithm are centered around harder objects, resulting in a distinct clustering pattern.

There are three different types of jets that a detector can identify, and these are classified based on how data from different parts of the detector are used together [104]. These three types are Calorimeter (CALO) jets, Jet-Plus-Track (JPT) jets, and PF jets. CALO jets are reconstructed by using the energy deposits detected in the calorimeter towers, while JPT jets are formed by incorporating tracking information to enhance the energy response and resolution of the calorimeter jets, following the Jet-Plus-Track algorithm. Finally, the reconstruction of the PF jets involves clustering the four-momentum vectors of PF candidates. As discussed in Section 3, the PF algorithm combines data from different parts of the CMS detector to find and reconstruct all visible particles in a given event. These particles could be muons, electrons, photons, or hadrons.

4.5.1 Jet Energy Correction

The measurement of jet energies, affected by both the detector response and its efficiency, along with specific details of the jet clustering algorithm employed, leads

to discrepancies in the recorded jet energies compared to those at the particle level. The main objective of jet energy calibration (JEC) [104] is to establish a connection between the average measured energy of the detector jet and the energy of the actual particle jet it corresponds to. The actual particle jet is formed by clustering all the stable particles originating from the fragmenting parton, as well as the particles resulting from the underlying event (UE) activity, using the same clustering algorithm applied to the detector jets. To achieve this, a correction factor C is applied to each component of the raw jet four-momentum vector p_{μ}^{raw} (with components represented by μ in the following)

$$p_{\mu}^{cor} = C.p_{\mu}^{raw} \tag{4.6}$$

The correction factor C consists of various components, namely the offset correction C_{offset}, the MC calibration factor C_{MC}, and the residual calibrations C_{rel} and C_{abs}, which handle the relative and absolute energy scales, respectively. The offset correction eliminates the excess energy caused by noise and PU, while the MC correction addresses the major irregularities in η and non-linearities in p_T. Lastly, the residual corrections take into account the minor deviations between the data and the simulation. The different components are implemented consecutively according to the equation provided:

$$C = C_{offset}(p_{\mu}^{raw}).C_{MC}(p_T', \eta).C_{rel}(\eta).C_{abs}(p_T'') \tag{4.7}$$

where p_T' represents the $\mathrm{p_T}$ of the jet after applying the offset correction, and p_T'' denotes the $\mathrm{p_T}$ of the jet after all preceding corrections have been applied.

4.5.2 B tagged jets and identification

In high-energy physics (HEP) detectors, the identification of particles carrying color charge becomes highly challenging due to the process of hadronization. However, this complexity offers a unique opportunity to categorize numerous rare SM and new physics processes. Among the various partons, only jets originating from b quarks,

and to a lesser extent c quarks, can be distinguished and effectively isolated. This distinction is possible due to the longer lifetime and greater mass of b and c quarks. The occurrence of b quarks (and consequently b jets) holds significance in a wide range of intriguing new physics phenomena and top quark and Higgs boson decay channels.

The CMS detector is well-equipped for identifying b jets (b tagging), due to its precise tracking capabilities and robust lepton identification systems among other notable features. Various reconstructed objects, such as tracks, vertices, and identified leptons, can be utilized to construct observables that differentiate between b jets and jets originating from light particles.

Identification algorithms [105] employ feature sets associated with either the impact parameter or the secondary vertex, and in some cases, a combination of both to enhance the discrimination power. The impact parameter represents the vertical distance between the decay point (secondary vertex) and the interaction point (primary vertex). The algorithm's effectiveness is influenced by the proximity of the tracking detector hits to the interaction point, requiring at least two out of the eight hits to be detected in the pixel detector. On the other hand, the secondary vertex is an additional vertex observed following the primary collision vertex. b quarks, unlike other partons, possess longer decay lengths from the interaction point and decay into harder colorless particles, thereby producing a secondary vertex.

Several machine learning techniques, described in detail in the next chapter, have recently been utilized to improve b tagging performances. For example, the DeepCSV algorithm uses artificial neural networks (ANNs) to combine secondary vertex and track-based lifetime information. A more recent algorithm developed for Run 2, DeepFlavour, uses a deep neural network with a larger set of features, including properties of charged particles, neutral constituents, and secondary vertices [106]. Notably, the DeepFlavour tagger shows a 4% absolute improvement in b tagging efficiency compared to DeepCSV, with a mistag rate for light jets at 0.1%.

Each algorithm produces a discriminator value for each jet. By setting minimum thresholds on these discriminators, loose ("L"), medium ("M"), and tight ("T") operating points are defined. These operating points correspond to misidentification

probabilities of approximately 10%, 1%, and 0.1%, respectively, for jets originating from light particles, at an average jet $\mathrm{p_T}$ of around 80 GeV/c. This is critical for reducing background events and enhancing the accuracy of measurements and searches for new physics [105].

All selections applied to the jets and b jets in this thesis are given in Sections 7.4.3 and 7.4.4.

4.6 Missing Transverse Momentum

Neutral particles, such as neutrinos, usually go through collider detectors without making any direct impact. To figure out if such particles are present, one can look at the total momentum in the event and see if there is a discrepancy, indicating something went undetected. This discrepancy in momentum, particularly in the plane perpendicular to the direction of the particle beams, is termed "missing transverse momentum", initially defined in Section 3 and in the introduction part of this chapter, symbolized as $\tilde{\mathrm{p_T}}^{\mathrm{miss}}$. The amount of this missing momentum is denoted as $\mathrm{p_T^{miss}}$.

$\mathrm{p_T^{miss}}$ is essential for distinguishing specific particle events, such as the leptonic decays of W bosons and top quarks, from background noise that doesn't contain neutrinos, such as multijet and Drell-Yan events. It also plays a vital role in the search for new particles that interact weakly and have long lifetimes. Some theories that extend BSM, like supersymmetry, predict events with a large amount of $\mathrm{p_T^{miss}}$.

However, calculating or "reconstructing" $\vec{p_T}^{miss}$ is a challenging process. There are several factors that can lead to inaccurate p_T^{miss} measurements, such as errors in measuring particle momentum, misidentifying particles, detector malfunctions, particles hitting areas of the detector that are not well-instrumented, cosmic rays, and beam-related particles.

The CMS detector has come up with three different algorithms for reconstructing $\vec{p_T}^{miss}$:

- PF p_T^{miss}: This method utilizes the PF technique, making use of information from various sub-detectors to reconstruct each individual particle.

- Calo p_T^{miss}: This approach relies on the energies measured in the calorimeter and the geometrical arrangement of the calorimeter towers.

- TC p_T^{miss}: This method starts with the Calo p_T^{miss} and then makes adjustments by incorporating information from tracks that are reconstructed in the inner tracker, after accounting for the energy these tracks are expected to deposit in the calorimeter.

As mentioned above, estimated magnitude of $\vec{p_T}^{miss}$ can sometimes be lower than the actual value due to various factors. The jet energy corrections, which are detailed in Section 4.5.1, are extended to the calculations of $\vec{p_T}^{miss}$, and are known as type-I corrections to the p_T^{miss}. Ideally, due to rotational symmetry, p_T^{miss} should not depend on the ϕ angle of the detector axis. However, in practice, p_T^{miss} is influenced by the ϕ angle due to factors such as calorimeter efficiencies, displacement of the beamspot, and non-uniform detector responses. A correction for this effect is applied by shifting the origin of the coordinate system in the transverse plane [107].

CHAPTER 5

MACHINE LEARNING AND ITS APPLICATIONS IN HIGH ENERGY PHYSICS

Particle physics, especially in the post-Higgs boson discovery era, aims to fully exploit the potential of the LHC and its upgraded version, the HL-LHC, alongside ongoing and future neutrino experiments. The HL-LHC, with a ten times improvement in the integrated luminosity compared to the current LHC, presents new challenges in terms of event size, data volume, and complexity. Meeting the physics objectives of precision in probing the SM and exploring the BSM physics requires solid solutions to these challenges. At this point, Machine Learning (ML) offers a promising approach to enhance both algorithmic performance and computational efficiency in experimental high energy physics (HEP).

In a general sense, ML implies that a computer program's performance improves over increased experience for a specific class of tasks and performance metrics [108]. The primary goal is to automate the process of analytical model building for tasks like object detection or natural language translation. This is accomplished through the application of algorithms that iteratively learn from problem-specific training data, enabling computers to uncover hidden insights and complex patterns without explicit programming [109]. ML, especially in tasks involving high-dimensional data such as classification, regression, and clustering, offers notable applicability. By assimilating knowledge from past computations and identifying regularities within extensive databases, ML contributes to making reliable and repeatable decisions. Table 5.1 shows the three categories of ML depending on the data considered and the problem in question [110].

Table 5.1: Comparison and definitions of three main ML types.

Supervised Learning	
Data Type	Labeled data
Process Description	Uses labeled data (training set) to train a model to learn patterns. Minimizes difference between predicted outputs and true labels. Once trained, the model predicts or classifies features in new, unseen data (test set). Used for regression and classification.
Applications	Medical diagnosis, image classification, natural language processing
Unsupervised Learning	
Data Type	Unlabeled data
Process Description	No labels are provided. The model discovers patterns and relationships in the data by itself. Includes dimensionality reduction and does not require data labeling. Used for clustering, association, and dimensionality reduction [109].
Applications	Scientific discovery, anomaly detection, image analysis
Reinforcement Learning	
Data Type	Feedback-based data
Process Description	Self-teaching through trial and error to maximize rewards. Algorithms learn from the outcomes of past actions and receive feedback after each action to assess its effectiveness.
Applications	Automated systems, decision-making processes

5.1 Machine Learning Flowchart and Hyperparameters

Solving a problem by using ML techniques involves a sequence of steps such as collecting and preparing data, creating a model, training and evaluating the model, tuning hyperparameters to optimize results, and making predictions. In this section, several terms and hyperparameters frequently used in ML are defined, underpinning the concepts introduced in the following sections.

Data collection and preprocessing

Collecting a sufficient amount of data is a crucial prerequisite for achieving desired outcomes. This data needs to be preprocessed, which removing any noisy or missing elements, scaling input features mostly by normalizing them between 0 and 1 to ensure stable training. Additionally, categorical data and text need to be converted into real-valued numbers.

Model building

Next step is building a model according to needs and requirements of the problem. A model is a set of one or more layers with learnable parameters that maps an input to an output. Layers are made up of nodes and each node is allocated a weight, providing a significant contribution to the overall output of the model. The weighted inputs are summed up and then a bias, denoting the presence of systematic errors that can lead the model to consistently make incorrect predictions, is added. The weights and biases are the trainable parameters and updated on the basis of the error at the output.

Backpropagation, loss and activation function

The above process is known as *back-propagation*. The backpropagation algorithm operates by calculating the gradient of the *loss function* concerning each weight. This computation involves utilizing the chain rule to compute gradients layer by layer, with the process beginning from the last layer and moving backward. This approach minimizes the redundancy in the calculation of intermediate terms in the chain rule and continues until the desired output is reached. The loss function serves as a mathematical tool for assessing how well the algorithm represents the dataset. It quantifies the discrepancy between predicted and actual values within the model. Once the gradients are determined through backpropagation, an optimization algorithm comes into

play, adjusting the model's parameters in a manner that minimizes the loss. Also, the amount that the weights are updated during training is referred to as the *learning rate*. Then, a special function called *activation function* [111] is applied to make both inputs and outputs non-linear and the backpropogation possible.

Datasets and other hyperparameters

Some of other parameters in an ML training includes different sample sets such as validation, training, and test, along with key metrics like the number of epochs, batch size, dropout and an evaluation factor called *Area Under the Curve* of the *Receiver Operating Characteristic* (AUC-ROC) curve [112, 113]. The training dataset represents the input data used for model fitting, while the validation dataset offers an unbiased assessment of model performance on the training data, particularly during hyperparameter tuning. Finally, test dataset is the sample of input data used to provide an unbiased evaluation of a final model fit on the training dataset. The test dataset serves as an unbiased evaluation of the final model fitted on the training data. The batch size, a hyperparameter, determines the number of samples (i.e., a single row of input data) processed before updating internal model parameters. Additionally, the number of epochs, another hyperparameter, specifies how many times the learning algorithm iterates through the entire training dataset. Additionally, when dealing with a limited amount of training data, complex relationships within an ML model often arise from sampling noise. Consequently, these relationships may be present in the training set but may not reflect the true patterns in real test data, even if both are drawn from the same distribution. This phenomenon leads to overfitting, and various methods have been developed to address it. One such method is known as dropout [114], which accomplishes this by randomly disabling neurons during training. The AUC-ROC curve, on the other hand, is a metric to evaluate classification performances in ML algorithms. Optimization of the parameters mentioned above is crucial to increase the model performance and obtain a better classification. Details on the optimization techniques can be found in Ref. [115, 116].

5.2 Neural Networks and Deep Learning

A neural network (NN) is an ML algorithm inspired by the structure of a human neuron. The brain is a complex network of billions of neurons that interact through electrical and chemical signals at synapses. Within this intricate communication system, information originating from one neuron is transmitted to others. This transmission is facilitated by an electrical impulse known as 'action potential.' For effective transmission to occur, the input signals, or impulses, must possess sufficient strength to surpass a predefined threshold. Only when this threshold is exceeded, a neuron can activate the signal and allow it to be transmitted as output to other neurons. Perceptron, first introduced by Frank Rosenblatt in 1957 [117], is a type of artificial neuron in a neural network and used as a building block. In the context of perceptrons, the electrical signals in actual neurons are analogously represented as numerical values. The perceptron models this representation by assigning a weight to each input value. This process mirrors the biological scenario by calculating the weighted sum of inputs, similar to the total strength of input signals, and then applying a step function to determine the perceptron's output. This output, resembling biological neural networks, is subsequently transmitted to other perceptrons. Perceptrons are binary classifiers, where the output is categorized as one of two final possibilities.

A figurative comparison of a biological neuron and a perceptron is presented in Figure 5.1.

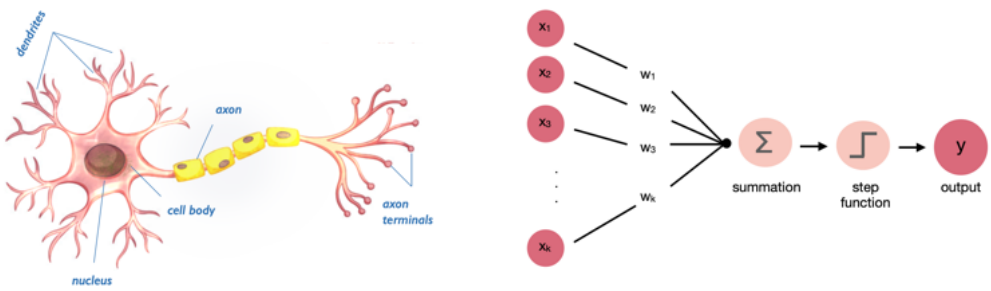

Figure 5.1: Comparison of a biological neuron and a perceptron.

Neural Networks (NNs), on the other hand, can be considered as a more complex and sophisticated extension of perceptrons. They are used to uncover complex connections between inputs and outputs in various applications, such as image recognition

and medical diagnosis. Three key layers are present in an NNs. The input layer, where information like text or numbers is received, is followed by hidden layers in the middle. These hidden layers, which can be a single layer, as in a perceptron, or multiple layers, perform computations on input data to identify patterns. The result is obtained in the output layer after rigorous computations by the hidden layers. The performance of NNs is determined by factors described in the previous section such as weights, biases, learning rate, activation function and batch size.

Deep neural networks (DNNs) are like advanced versions of regular NNs. They have many hidden layers arranged in a complex way. Unlike simple networks, they use advanced operations and more complicated neurons, allowing them to understand raw data and figure the important aspects of a specific task out. This ability is known as deep learning (DL). In a fully connected DNN, there exists an input layer followed by hidden layers connected sequentially. Each neuron in these layers receives input from the neurons in the previous layer or directly from the input layer. The output of one neuron serves as the input for other neurons in the subsequent layer, and this sequential process continues until the final layer generates the network's output. Applying the steps described in Section 5.1, the layers of the neural network undergo a series of nonlinear transformations, enabling the network to acquire intricate representations of the input data. DL can be used for three ML categories mentioned earlier, specifically supervised learning, unsupervised learning, and reinforcement learning. Over time, various architectures in DL have been introduced [118–120]. While almost any architecture can be applied to different tasks, certain architectures are better suited for specific types of data. These architectural variations, one is presented in Figure 5.2 [121], are primarily defined by the types of layers, neural units, and connections they incorporate.

5.2.1 Convolutional Neural Networks

In this thesis, convolutional neural networks (CNNs) are used as a DL architecture. CNN stands out from its predecessors as it can figure out important features all by itself, without needing human guidance [122,123]. These networks are used in many areas such as computer vision [124], speech processing [125], and Face Recogni-

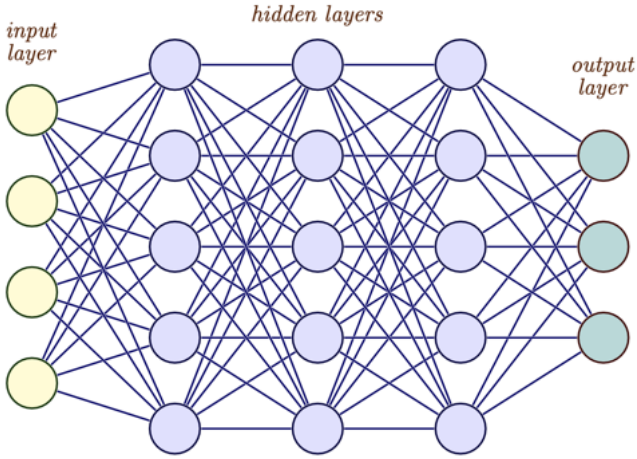

Figure 5.2: Deep Neural Networks representation with many hidden layers.

tion [126]. Like regular NNs, the design of CNNs is also inspired by how neurons work in human and animal brains. In a typical CNN, there are four main parts, namely, input layer, convolutional layers, pooling layer, and fully connected layers as represented in Figure 5.3.

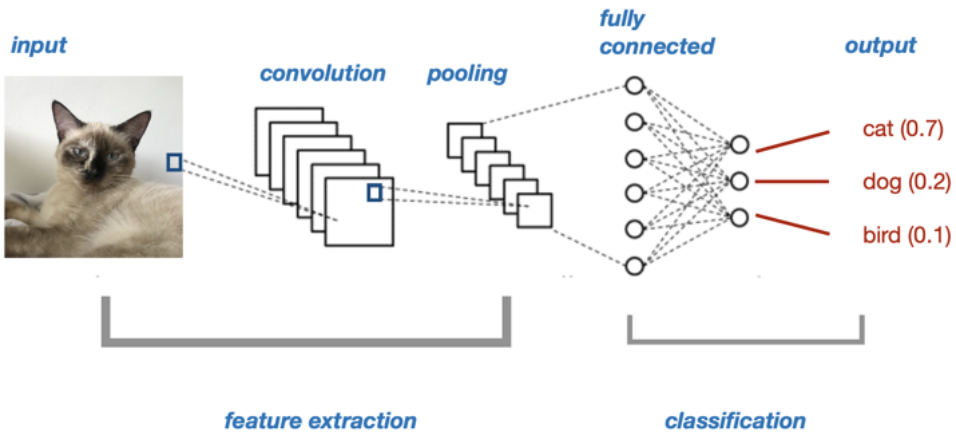

Figure 5.3: Convolutional Neural Network flowchart with feature extraction and classification parts.

Input layer is the place where input is provided to the model. The input generally is a three dimensional image or a sequence of images with a width, a height, and a depth. **Convolution layers** play a crucial role in extracting features from the input dataset. As illustrated in Figure 5.4, it utilizes a collection of learnable filters called *kernels*, applying them to the input images. These filters are essentially smaller matrices, with dimensions like m $\times$ m (where m can be 2, 3, and so on). They traverse the whole

input image data step by step (called *stride*) by calculating the dot product between the kernel weight and the corresponding patch of the input image. The result of applying a filter to the entire input is called a feature map, providing a map of where the feature is present or absent in the input. Convolutional layers can be expressed as a matrix, enabling the creation of transposed convolutional layers to conduct the reverse operation, often referred to as deconvolution. These layers can learn to expand a higher-dimensional representation from a feature map. Convolutional layers apply an *activation function* to the output of the preceding layer allowing to add non-linearity to the model without changing the output's dimension. The most commonly used activation function in CNNs is known as rectified linear units (ReLU) [127,128]. This function aids the network in grasping non-linear connections among input features, thereby enhancing the network's robustness in recognizing diverse patterns. **Pooling layer** is used to systematically decrease the the representation's dimension, thereby reducing the number of parameters and the computational complexity of the mode. In the majority of CNNs, this adjustment is carried out by max-pooling layers, utilizing 2 $\times$ 2 kernels and a stride of 2 along the spatial dimensions of the input. Consequently, this process scales down the activation map to one-quarter of its original size while retaining the depth volume as the same. In the final stage, **fully connected layers** consist of neurons directly linked to the neurons in the two adjacent layers, without establishing connections to any intermediary layers. They produce class scores from the activation functions, to be used for classification. Through the iterative application of convolutions on input data, CNNs acquire the ability to discern high-level features and understand the relationships among them. This capability proves especially advantageous in the field of HEP.

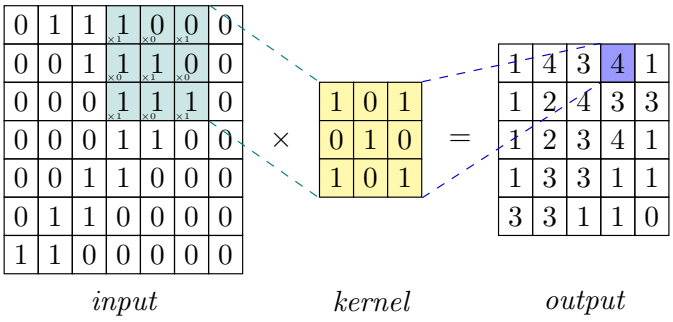

Figure 5.4: Sliding of a 3 $\times$ 3 kernel over an input.

5.2.2 Graph Neural Networks

LHC comprises various detectors and sub-detectors, yielding highly intricate and heterogeneous data. Certain data in HEP can be partially interpreted as images, leading to the CNN applications with enhanced performance. However, image representations may encounter challenges with irregular detector geometry or sparse projections, and any possible information loss in image representations may limit the extraction of data-related information. On the other hand, this data can be handled as a collection of items and can be transformed into a graph representation by incorporating an adjacency matrix. Graphs are a natural and powerful way of representing many complex systems employing mathematical formulas to represent objects (nodes) and their relationships (edges) in a unified structure. This graph-based representation offers a flexible environment to apply more advanced DL techniques by reducing the limitations effective for other techniques mentioned above such as fully connected and convolutional layers. Graph neural networks (GNNs), developed by Scarselli et al. [129] in 2008 are a type of DL architecture that focuses on learning functions for graph-structured data by incorporating strong relational inductive biases. In GNNs, there are building blocks, like fully connected layers, that handle the computations needed for messages and their propagation within the graph enabling the computation of sophisticated edge, node, and graph-level outputs [130,131]. In Figure 5.5 [131], two examples of how the HEP data is formulated as graphs are shown.

Figure 5.5: Illustrations showcasing two diverse classification scenarios using graphs in HEP. The left one involves categorizing events with various types of physics objects, while the right figure focuses on classifying jets based on their associated particles.

Graph Attention Networks (GATs) [132], a specialized form of GNNs, employ an 'attention mechanism' to determine the significance of features in graph structures.

This mechanism, widely used in DL, serves as a method to emphasize certain elements over others within a given context. By assigning weights to each input attribute based on its relevance to the output, GATs can dynamically focus on the most crucial information. This selective engagement enhances the model's ability to identify and interpret critical dependencies and relationships within the data, thereby improving its capacity to discern complex patterns.

GATs operate on a sophisticated mechanism that begins with input encoding, where the input data sequence is transformed into a set of representations suitable for the attention mechanism. A query vector, which is a numerical representation that encodes the current state or focus of the model, is then generated based on the model's current context. This is followed by the creation of key-value pairs from the input representations, where keys determine the relevance of the data, and the values contain the actual information. The model computes the similarity between the query vector and each key to assess their relevance. These similarity scores are then transformed into attention weights using a Softmax function, highlighting the importance of each key-value pair. The attention weights are applied to the values, resulting in a weighted sum that aggregates the most relevant information from the input. This weighted sum forms a context vector, encapsulating the focused information from the input, which is then integrated with the model's current state or hidden representation. This integration provides additional context for subsequent steps or layers of the model. The entire process, from query generation to the integration of the context vector, is dynamically repeated at each step or iteration of the model, allowing it to adaptively focus on different parts of the input data, thereby enhancing its ability to process and interpret complex information structures efficiently.

In this thesis, GATs are used to assign objects to their originating particles by graphically analyzing the connections between objects and their adjacent entities (neighbours). The attention mechanism enables the model to achieve a 'higher representation' of the objects (particles). This means it can understand and represent the particles not just in isolation but also in terms of their contextual relationships and interactions with their neighbours within the graph. This approach allows for a more nuanced and accurate representation and understanding of particle dynamics and interactions. Section 7.6 explains how GATs are used in this study.

CHAPTER 6

PHYSICS MOTIVATION FOR $t\bar{t}HH$ PRODUCTION MECHANISM

Higgs production in association with a top anti-top quark pair, referred to as $t\bar{t}HH$ production mechanism plays an important role in the Higgs Physics both in the SM and the BSM. It is in interplay with both HH (gluon fusion and VBF), and $t\bar{t}H$ productions. Figures 6.1 and 6.2 [133] show the series of processes that occur in pp collisions involving the production of double Higgs with a large cross section range from $\sqrt{s}$ = 8 TeV to $\sqrt{s}$ = 100 TeV and at $\sqrt{s}$ = 14 TeV, respectively. These processes enable access to a new region in the exploration of the Higgs sector. Table 6.1 lists the cross sections computed at the NLO QCD [13] for $t\bar{t}HH$ in comparison to those for VBF (HH) and tjHH at $\sqrt{s}$ = 14 TeV.

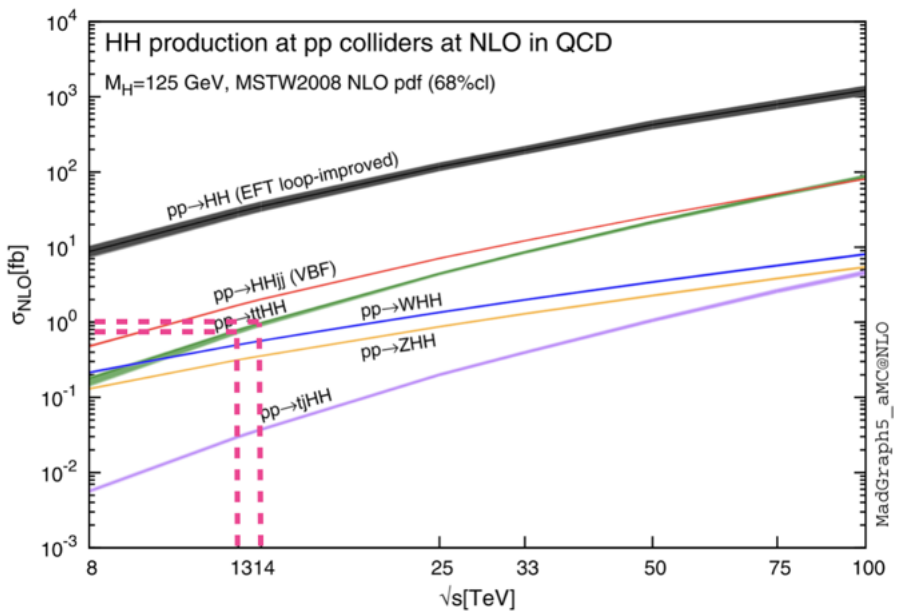

Figure 6.1: Total cross sections at the NLO in QCD for the six largest HH production channels at pp colliders. The thickness of the lines corresponds to the scale and PDF uncertainties added linearly.

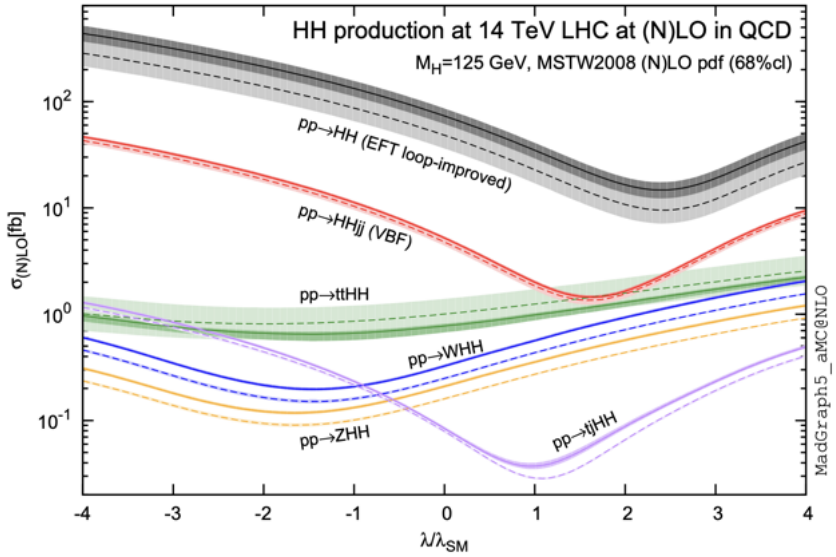

Figure 6.2: Total cross sections at the LO and the NLO in QCD for HH production channels, plotted as a function of the self-interaction coupling λ. The dashed (solid) lines and light-(dark) colored bands represent the LO (NLO) results, with the uncertainties due to PDF scales added linearly. The SM values of the cross-section are obtained when $\lambda/\lambda_{SM} = 1$.

Table 6.1: Cross sections computed at the NLO QCD for ZHH, WHH, VBF HH, $\mathrm{t\bar{t}HH}$, and tjHH at 14 TeV center-of-mass energy.

$\sqrt{s}$(TeV)	ZHH	WHH	VBF HH	ttHH	tjHH
14	$0.359^{+1.9\%}_{-1.3\%} \pm 1.7\%$	$0.573^{+2.0\%}_{-1.4\%} \pm 1.9\%$	$1.95^{+1.1\%}_{-1.5\%} \pm 2.0\%$	$0.948^{+3.9\%}_{-13.5\%} \pm 3.2\%$	$0.0383^{+5.2\%}_{-3.3\%} \pm 4.7\%$

The information provided in Table 6.2 can be supplemented by comparing the production cross sections for $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}HH}$ processes at 13 and 14 TeV [13].

The study of double Higgs production, either through direct processes or via vector boson fusion (VBF), is currently ongoing in both ATLAS and CMS experiments [134, 135], alongside the already observed $\mathrm{t\bar{t}H}$ production process [55]. The $\mathrm{t\bar{t}HH}$ production process is in particular linked with both the $\mathrm{t\bar{t}H}$ production process and the standard double Higgs production, offering unique and intriguing additional features.

The aim of this research is to improve upon current studies on $\mathrm{t\bar{t}HH}$ production at the LHC by developing potential novel methodologies. Another objective of the study is to explore this production at the HL-LHC, which offers increased cross-section and luminosity compared to the LHC.

Table 6.2: Cross-sections for ttH and ttHH Processes at 13 and 14 TeV.

Process	$\sqrt{s}$	**Cross-section (fb)**	**QCD Scale Uncertainty Uncertainty**	**PDF + α_s Uncertainty**
$t\bar{t}H$	13 TeV	507.1	$+5.8\% - 9.2\%$	$\pm 3.6\%$
$t\bar{t}H$	14 TeV	613.7	$+6.0\% - 9.2\%$	$\pm 3.5\%$
$t\bar{t}HH$	13 TeV	0.775	$+1.5\% - 4.3\%$	$\pm 3.2\%$
$t\bar{t}HH$	14 TeV	0.949	$+1.7\% - 4.5\%$	$\pm 3.1\%$

The top quark Yukawa coupling refers to the interaction between the Higgs boson and the heaviest particle in the standard model, the top quark. This interaction is explored through the processes of Higgs boson production and decay. Due to the substantial mass of the top quarks, the Higgs boson cannot directly decay into them; instead, the coupling is indirectly probed via quantum loops in its decay processes. Direct examination of this coupling is feasible only during production scenarios, specifically when the Higgs boson is produced alongside a pair of top quarks, providing straightforward access to the magnitude of the top quark Yukawa coupling. In the SM, as illustrated in Figure 6.3, the LO mechanisms for $t\bar{t}HH$ production include the Yukawa vertex, accounting for 80% of the total cross-section, and the trilinear Higgs coupling ($_{HHH}$), contributing to the remaining 20% of the total cross-section.

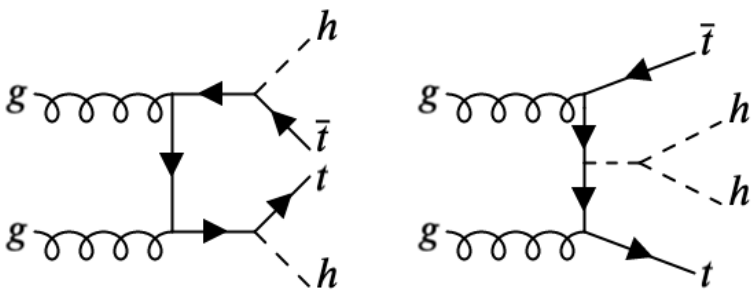

Figure 6.3: Representative Feynman diagrams for the $t\bar{t}HH$ process. These illustrate the two distinct physical subprocesses: the Yukawa vertex (left) and the Higgs trilinear self-coupling (right) as expected in the SM.

The Yukawa vertex component offers an additional opportunity to measure the $t\bar{t}HH$ coupling and assess its CP structure. In contrast to the $t\bar{t}H$ process, the $t\bar{t}HH$ production process provides access to the triple Higgs coupling. Differing from double Higgs production, $t\bar{t}HH$ does not involve interference terms that would allow for

access to the triple Higgs coupling. These characteristics highlight the significant relevance of the $\mathrm{t\bar{t}HH}$ process within the Standard Model, beyond its intrinsic interest.

The exploration of the Higgs boson will remain at the center of the HL-LHC program. It will include precise measurements of the Higgs boson couplings, probing of its tensor structure and the search for rare SM and BSM decays. The enormous dataset will give access to all the p-p production processes and decays of the Higgs boson. Hence, a key motivation of this thesis is the potential detection of $\mathrm{t\bar{t}HH}$ processes in the BSM physics, especially within the context of Minimal Composite Higgs Models (MCHM). Unlike the $\mathrm{t\bar{t}H}$, MCHM features an additional radiated Higgs boson that creates more complex interactions allowing possible new physics effects.

There are two key types of contributions in MCHM:

Resonant Processes: These are highly relevant in MCHM, as the models predict the existence of vector-like top partners with a 2/3 charge. These partners can decay into a tH channel, leading to a $\mathrm{t\bar{t}HH}$ final state as described below in details. Their production is mainly through QCD pair processes. Previous searches for these resonances in this decay channel and combined searches in the bW, tZ, and tH channels have set constraints on such vector-like resonances [136, 137].

Non-Resonant Processes: The $\mathrm{t\bar{t}HH}$ process plays a significant role in the non-resonant production, particularly in situations involving heavy resonances. This process largely determines the cross section in these scenarios. At the LO, it is governed by diagrams specific to MCHM, as illustrated in the upper section of Fig. 6.4. There are three key physical subprocesses involved here: the Yukawa vertex and the Higgs trilinear self-coupling, both of which are also present in the SM, and an additional 'double Higgs' Yukawa vertex, which is unique to composite Higgs models. However, it's important to note that this 'double Higgs' interaction only contributes a small percentage to the total cross section. The non-resonant $\mathrm{t\bar{t}HH}$ process thus bears a close resemblance to the SM $\mathrm{t\bar{t}HH}$ process [12].

Changes in the $\mathrm{t\bar{t}H}$ production within MCHM are primarily due to the effects of Higgs compositeness and interactions with vector-like quarks (VLQs). These changes are all channeled through modifications in the top Yukawa coupling, denoted as y_t. This

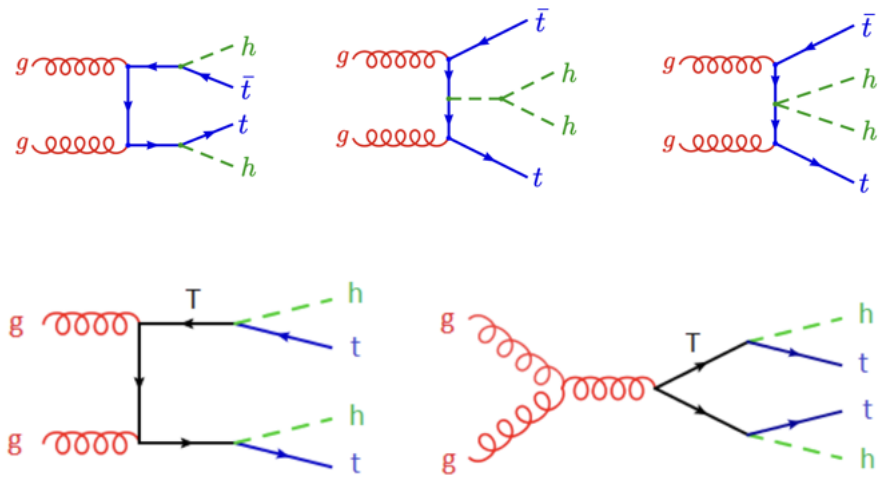

Figure 6.4: Representative diagrams for the $t\bar{t}HH$ production process at the LO within MCHM. Top: the non-resonant part of the $t\bar{t}HH$ production process, illustrating the three distinct physical subprocesses: the Yukawa vertex, the Higgs trilinear self-coupling and the "double Higgs" Yukawa vertex arising in composite Higgs scenarios. Bottom: the resonant part illustrating the QCD pair production of the top heavy partners, with their top-Higgs decay.

relationship can be mathematically represented as follows:

$$\sigma_{\mathrm{MCHM}}(ttH) = \left(\frac{y_t}{y_t^{\mathrm{SM}}}\right)^2, \sigma_{\mathrm{SM}}(ttH). \tag{6.1}$$

This equation shows that while the kinematic distributions of $t\bar{t}H$ in MCHM remain the same as in the SM, the total production rates differ.

On the other hand, these modifications have implications for the non-resonant contributions to the $t\bar{t}HH$ process, resulting in even more significant changes in signal strength compared to the SM. In this thesis, the MCHM_5 and MCHM_{14} models are specifically explored, which offer unique insights and predictions about the $t\bar{t}HH$ process, diverging from the predictions of the SM [12,138] as detailed in the CERN Yellow report on the perspectives for the HL/HE-LHC [138]. In both MCHM scenarios, the presence of resonances leads to increased $t\bar{t}HH$ cross sections. Specifically, in MCHM_5, there are three heavy top partners with a 2/3 charge (T^1, T^2, T^3), and in MCHM_{14}, there are six (T^1 to T^6). These top partners predominantly decay into the

tH channel, contributing to a $\mathrm{t\bar{t}HH}$ final state and altering the kinematic distributions of these events in comparison to the SM scenario.

For the $\mathrm{MCHM_5}$ and $\mathrm{MCHM_{14}}$ scenarios, two benchmark points C2 and D7 are selected respectively to showcase the aforementioned features as detailed in a phenomenological reference study [12]. These are chosen as relevant examples of "low scale" Minimal Composite scenarios, i.e. possibly accessible at the HL-LHC. These benchmark points are labelled as $\mathrm{MCHM_5^{C2}}$ and $\mathrm{MCHM_{14}^{D7}}$ following the numbering in Ref. [12]. The phenomenology study found that, for both cases, the $\mathrm{t\bar{t}H}$ signal strength with respect to the SM expectation is close to 1, amounting to 0.94 and 0.93 for $\mathrm{MCHM_5^{C2}}$ and $\mathrm{MCHM_{14}^{D7}}$, respectively. Hence, it would be difficult for measurements in $\mathrm{t\bar{t}H}$ to conclusively state a possible deviation from the SM expectation, even at the HL-LHC. TThe signal strength of $\mathrm{t\bar{t}HH}$ in comparison to the SM expectation is notably elevated, measuring at 1.47 for $\mathrm{MCHM_5^{C2}}$ and 2.15 for $\mathrm{MCHM_{14}^{D7}}$. The higher potential sensitivity in the $\mathrm{t\bar{t}HH}$ measurements holds the potential to detect scenarios like MCHM at the HL-LHC.

In summary, through deviations from the SM both in the signal strength and in the kinematic distributions, as well as through the possibility of identification of high mass resonances, the $\mathrm{t\bar{t}HH}$ process provides a variety of ways to explore the BSM physics.

CHAPTER 7

FULL RUN-2 ANALYSIS DETAILS AND STRATEGY

With the motivation described in Chapter 6, the study focuses on the production of a top quark-antiquark pair associated with a pair of Higgs bosons, where the top quark pair decays dileptonically and the Higgs boson pair decays hadronically into b-quark-antiquark pairs. The full Run 2 dataset of 137.60 fb^{1} of integrated luminosity, at a center-of-mass energy of $\sqrt{13}$ TeV is used. The inclusive $\mathrm{t\bar{t}HH}$ cross section computed at NLO QCD is $0.775^{+1.5\%}_{-4.3\%} + 3.2\%$ fb where the first uncertainty originates from the QCD scale dependence and the second from parton distribution functions [13,14].

7.1 Analysis Method

The signal topology, as illustrated in the Figure 7.1, involves multiple jets, multiple b jets, and moderate missing transverse energy arising from neutrinos released during W boson decays. Additionally, the presence of precisely two leptons is required. These leptons can be either two electrons, two muons, or one electron and one muon.

Muon and tau leptons, unlike electrons, have the ability to undergo decay through weak interactions. However, due to their different masses, they exhibit distinct decay behaviors. Muons, with a mass lower than the lightest meson (the pion, π), cannot decay into final states that involve hadrons. Instead, muon decay predominantly results in lighter particles, such as electrons, neutrinos, and other leptons. On the other hand, tau leptons have a mass more than twelve times greater than that of π object. This substantial mass allows the τ to undergo more diverse decay processes. In τ decay, the W boson involved in the decay can disintegrate into quark-antiquark pairs

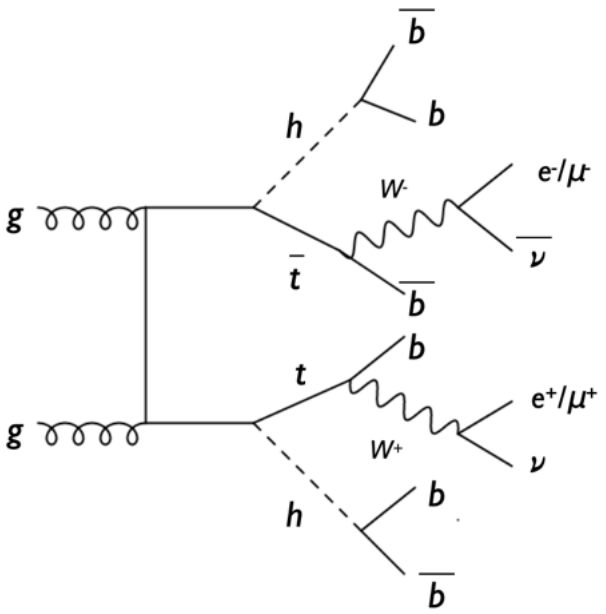

Figure 7.1: Feynman diagram of the $t\bar{t}HH$ production process in the dilepton channel.

(in about two thirds of the cases), in addition to producing leptons. Consequently, τ decays involve a combination of hadronic and leptonic final states [139]. Hence, τ objects are not considered in this analysis and the focus lies on studying the production of top quark-antiquark pairs associated with Higgs bosons using electrons and muons as the primary final-state leptons, given the simplicity of their decay modes compared to tau leptons.

Following the object selection and the application of baseline criteria, multi-classifier Deep Neural Networks (DNNs) are employed to categorize the selected events into multiple categories, distinguishing between signal and various background processes. These DNN discriminants play a crucial role in signal extraction and are critical for determining the upper limits on signal strength.

7.2 Data and Simulation Samples

Data samples

The primary datasets considered in the analysis of the data collected in 2016, 2017,

and 2018 run periods are listed in Tables 7.2 to 7.5, respectively [140]. Although the study first started with the pre-legacy samples, shortly after it moved to samples produced with the ultimate Run 2 reconstruction (ultra-legacy (UL)). Detailed explanation of these terms and a comparison study is presented in Appendix A. The total integrated luminosity corresponds to 137.62 fb^{-1}, with per-year luminosities as listed in Table 7.1 [141].

Table 7.1: Integrated luminosity per year and total.

Year	Integrated Luminosity
2016	36.31 fb^{-1}
2017	41.48 fb^{-1}
2018	59.83 fb^{-1}
Total	137.62 fb^{-1}

Table 7.2: 2016preVFP datasets used in the analysis.

Sample	Run Range
/MuonEG/Run2016B-ver1_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	272760-273017
/MuonEG/Run2016B-ver2_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	273150-275376
/MuonEG/Run2016C-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	275656-276283
/MuonEG/Run2016D-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276315-276811
/MuonEG/Run2016E-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276831-277420
/MuonEG/Run2016F-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	277932-278807
/DoubleEG/Run2016B-ver1_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	272760–273017
/DoubleEG/Run2016B-ver2_HIPM_UL2016_MiniAODv2_NanoAODv9-v3/NANOAOD	273150–275376
/DoubleEG/Run2016C-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	275656-276283
/DoubleEG/Run2016D-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276315-276811
/DoubleEG/Run2016E-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276831–277420
/DoubleEG/Run2016F-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	277932-278807
/DoubleMuon/Run2016B-ver1_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	272760–273017
/DoubleMuon/Run2016B-ver2_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	273150–275376
/DoubleMuon/Run2016C-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	275656-276283
/DoubleMuon/Run2016D-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276315-276811
/DoubleMuon/Run2016E-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	276831–277420
/DoubleMuon/Run2016F-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	277932-278807

Table 7.3: 2016postVFP datasets used in the analysis.

Sample	Run Range
/MuonEG/Run2016F-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	278769-278808
/MuonEG/Run2016G-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	278820-280385
/MuonEG/Run2016H-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	281613-284044
/DoubleEG/Run2016F-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	278769-278808
/DoubleEG/Run2016G-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	278820-280385
/DoubleEG/Run2016H-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	281613-284044
/DoubleMuon/Run2016F-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	278769-278808
/DoubleMuon/Run2016G-UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD	278820-280385
/DoubleMuon/Run2016H-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD	281613-284044

Table 7.4: 2017 datasets used in the analysis.

Sample name	Run Range
/MuonEG/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	297047-299329
/MuonEG/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	299368-302029
/MuonEG/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	302031-302663
/MuonEG/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	303824-304797
/MuonEG/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	305040-306460
/DoubleEG/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	297047-299329
/DoubleEG/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	299368-302029
/DoubleEG/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	302030-302663
/DoubleEG/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	303818-304797
/DoubleEG/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	305040-306460
/DoubleMuon/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	297047-299329
/DoubleMuon/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	299368-302029
/DoubleMuon/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	302031-302663
/DoubleMuon/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	303824-304797
/DoubleMuon/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	305040-306462
/DoubleMuon/Run2017G-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	306526-306826
/DoubleMuon/Run2017H-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	306896-307082

Table 7.5: 2018 datasets used in the analysis.

Sample	Run Range
/MuonEG/Run2018A-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	315257-316995
/MuonEG/Run2018B-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	317080-319310
/MuonEG/Run2018C-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	319337-320065
/MuonEG/Run2018D-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	320500-325175
/DoubleMuon/Run2018A-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	315257-316995
/DoubleMuon/Run2018B-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	317080-319310
/DoubleMuon/Run2018C-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	319337-320065
/DoubleMuon/Run2018D-UL2018_MiniAODv2_NanoAODv9-v2/NANOAOD	320500-325175
/EGamma/Run2018A-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	315257-316995
/EGamma/Run2018B-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	317080-319310
/EGamma/Run2018C-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD	319337-320065
/EGamma/Run2018D-UL2018_MiniAODv2_NanoAODv9-v3/NANOAOD	320500-325175

Simulated samples

The signal and background events, on the other hand, are modelled using the MC simulated event samples listed in Tables 7.8 to 7.14. The events are generated at the NLO with POWHEG [85] or MADGRAPH5_AMC@NLO [79], or at leading order (LO) with PYTHIA [81], depending on the process. Parton showering and hadronization are simulated with PYTHIA and the parameters for the underlying event description correspond to the CP5 tune [142], where CP stands for "CMS PYTHIA8", for all signal and background processes.

In this study, $\mathrm{t\bar{t}HH} \rightarrow \mathrm{b\bar{b}b\bar{b}}$ is considered as the signal sample. It is generated with the MADGRAPH5_AMC@NLO generator, where $\mathrm{t\bar{t}}$ decay is inclusive. The inclusive $\mathrm{t\bar{t}HH}$ cross section is normalised to $0.775^{+1.5\%}_{-4.3\%} + 3.2\%$ fb taken from calculations at NLO QCD where the first uncertainty originates from the QCD scale dependence and the second from parton distribution functions [13, 14]. The effective $\mathrm{t\bar{t}HH}$ cross section is obtained as $\sigma \times (\mathcal{B}_{\mathrm{H}\rightarrow\mathrm{b\bar{b}}})^2 = 0.775\mathrm{fb} \times (0.5824)^2 = 0.263$ fb. Signal samples used for three years are given in Table 7.8.

Background samples and their corresponding cross section values [13, 22, 143] are listed in the Tables 7.9 to 7.14 for all three years. Details for the production campaigns mentioned in the sample names for each year are provided in Table 7.7. The MADGRAPH5_AMC@NLO generator is also used to produce the $\mathrm{t\bar{t}} + 4\mathrm{b}$, $\mathrm{t\bar{t}Z} \rightarrow \mathrm{b\bar{b}}$, $\mathrm{t\bar{t}ZH} \rightarrow \mathrm{b\bar{b}b\bar{b}}$ and $\mathrm{t\bar{t}ZZ} \rightarrow \mathrm{b\bar{b}b\bar{b}}$ background processes. Additionally, events for the background processes $\mathrm{t\bar{t}}$, $\mathrm{t\bar{t}} \rightarrow \mathrm{b\bar{b}}$ and $\mathrm{t\bar{t}H} \rightarrow \mathrm{b\bar{b}}$ ($\mathrm{t\bar{t}}$ with inclusive, SL, and DL) are generated with POWHEG at the NLO in the 5 flavour scheme (5FS), whereas $\mathrm{t\bar{t}} \rightarrow \mathrm{b\bar{b}}$ is in the 4 FS.

Furthermore, the sample tables include integrated cross-section calculations that incorporate relevant branching ratios for the Higgs, W, and Z boson decays, seperately provided in Table 7.6 [22, 144].

Last but not least, v9 of NANOAOD [145], which is a flat ntuple format readable with plain ROOT is used for all the samples.

In this study, a dedicated analyzer is used to get ntuples and histograms. At the analyzer stage, event counts are adjusted by a factor termed "weight", which is computed

using the formula

$$w = \frac{\sigma \times \mathcal{B} \times \mathcal{L}}{number\ of\ events} \tag{7.1}$$

where σ is the cross section of the process, $\mathcal{B}$ is the branching ratio of the considered boson and $\mathcal{L}$ is the integrated luminosity belonging to the considered year.

Table 7.6: Branching ratios of the H, W and Z bosons for the decay channels considered in this analysis (top) and the final state contribution of a top quark anti quark pair decaying dileptonically (bottom).

Decay mode	Branching ratio [fb^{-1}]
$\mathrm{H} \to \mathrm{b\bar{b}}$	0.5824
$\mathrm{W} \to \mathrm{had}$	0.6741
$\mathrm{W} \to \ell\nu$	0.3259
$\mathrm{Z} \to \mathrm{b\bar{b}}$	0.15
$\mathrm{t\bar{t}} \to \mathrm{W^+bW^-\bar{b}} \to \ell^+\nu_\ell \mathrm{b}\ell'^-\bar{\nu}_{\ell'}\mathrm{\bar{b}} = 10.5\%$	

Table 7.7: Production campaigns given for each year. These are abbreviated in the sample names with the symbol of each year itself.

Year	Production campaign
[2016]	RunIISummer20UL16NanoAODv9
[2016APV]	RunIISummer20UL16NanoAODAPVv9
[2017]	RunIISummer20UL17NanoAODv9
[2018]	RunIISummer20UL18NanoAODv9

Table 7.8: $t\bar{t}HH \to b\bar{b}b\bar{b}$ signal samples. Effective cross section is provided in the bottom part.

Channel	Year	Dataset
$HH \to b\bar{b}b\bar{b}$	2016	/TTHHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/TTHHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTHHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/TTHHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2

$\sigma \times (\mathcal{B}_{H\to b\bar{b}})^2 = 0.775\text{fb} \times (0.5824)^2 = 0.263$ fb

Table 7.9: $t\bar{t}$ background samples. Effective cross sections are provided in the bottom part for two channels, respectively.

Channel	Year	Dataset
SL	2016	/TTToSemiLeptonic_TuneCP5_13TeV−powheg−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v1
		/TTToSemiLeptonic_TuneCP5_13TeV−powheg−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTToSemiLeptonic_TuneCP5_13TeV−powheg−pythia8/ [2017] − 106X_mc2017_realistic_v9−v1
	2018	/TTToSemiLeptonic_TuneCP5_13TeV−powheg−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v1
DL	2016	/TTToTo2L2Nu_TuneCP5_13TeV−powheg−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v1
		/TTToTo2L2Nu_TuneCP5_13TeV−powheg−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTToTo2L2Nu_TuneCP5_13TeV−powheg−pythia8/ [2017] − 106X_mc2017_realistic_v9−v1
	2018	/TTToTo2L2Nu_TuneCP5_13TeV−powheg−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v1

$\sigma \times 2.(\mathcal{B}_{W\to had} \times \mathcal{B}_{W\to l\nu}) = 831760$ fb $\times 0.6741 \times 0.3259 = 212842.37$ fb
$\sigma \times (\mathcal{B}_{W\to l\nu})^2 = 831760$ fb $\times (0.3259)^2 = 88341.9$ fb

Table 7.10: $t\bar{t}4b$ background samples. Effective cross section is provided in the bottom part.

Channel	Year	Dataset
$t\bar{t}+4b$	2016	/TT4b_TuneCP5_13TeV−madgraph−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/TT4b_TuneCP5_13TeV−madgraph−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TT4b_TuneCP5_13TeV−madgraph−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/TT4b_TuneCP5_13TeV−madgraph−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2

Table 7.11: $t\bar{t}b\bar{b}$ background samples. Effective cross sections are provided in the bottom part for two channels, respectively.

Channel	Year	Dataset
SL	2016	/TTbb_4f_TTToSemiLeptonic_TuneCP5−powheg−Openloops−Pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v1
		/TTbb_4f_TTToSemiLeptonic_TuneCP5−powheg−Openloops−Pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTbb_4f_TTToSemiLeptonic_TuneCP5−powheg−Openloops−Pythia8/ [2017] − 106X_mc2017_realistic_v9−v1
	2018	/TTbb_4f_TTToSemiLeptonic_TuneCP5−powheg−Openloops−Pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v1
DL	2016	/TTbb_4f_TTTo2L2Nu_TuneCP5−powheg−Openloops−Pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v1
		/TTbb_4f_TTTo2L2Nu_TuneCP5−powheg−Openloops−Pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTbb_4f_TTTo2L2Nu_TuneCP5−powheg−Openloops−Pythia8/ [2017] − 106X_mc2017_realistic_v9−v1
	2018	/TTbb_4f_TTTo2L2Nu_TuneCP5−powheg−Openloops−Pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v1

$\sigma \times 2.(\mathcal{B}_{W\to had} \times \mathcal{B}_{W\to l\nu}) = 1452\text{ fb} \times 0.6741 \times 0.3259 = 318.988\text{ fb}$
$\sigma \times (\mathcal{B}_{W\to l\nu})^2 = 1452\text{ fb} \times (0.3259)^2 = 154.218\text{ fb}$

Table 7.12: $t\bar{t}Htob\bar{b}$ background samples. Effective cross sections are provided in the bottom part, for three different samples.

Channel	Year	Dataset
$H \to b\bar{b}$	2016	/ttHTobb_M125_TuneCP5_13TeV−powheg−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/ttHTobb_M125_TuneCP5_13TeV−powheg−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/ttHTobb_M125_TuneCP5_13TeV−powheg−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/ttHTobb_M125_TuneCP5_13TeV−powheg−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2
$H \to b\bar{b}$, $t\bar{t}SL$	2016	/ttHTobb_ttToSemiLeptonic_M125_TuneCP5_13TeV−powheg−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/ttHTobb_ttToSemiLeptonic_M125_TuneCP5_13TeV−powheg−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/ttHTobb_ttToSemiLeptonic_M125_TuneCP5_13TeV−powheg−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/ttHTobb_ttToSemiLeptonic_M125_TuneCP5_13TeV−powheg−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2
$H \to b\bar{b}$, $t\bar{t}DL$	2016	/ttHTobb_ttTo2L2Nu_M125_TuneCP5_13TeV−powheg−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/ttHTobb_ttTo2L2Nu_M125_TuneCP5_13TeV−powheg−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/ttHTobb_ttTo2L2Nu_M125_TuneCP5_13TeV−powheg−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/ttHTobb_ttTo2L2Nu_M125_TuneCP5_13TeV−powheg−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2

$\sigma \times \mathcal{B}_{H\to b\bar{b}} = 507.1\text{ fb} \times 0.5824 = 295.34\text{ fb}$
$\sigma \times \mathcal{B}_{H\to b\bar{b}} \times 2.(\mathcal{B}_{W\to had} \times \mathcal{B}_{W\to l\nu}) = 507.1\text{ fb} \times 0.5824 \times 2.(0.6741 \times 0.3259) = 129.76\text{ fb}$
$\sigma \times \mathcal{B}_{H\to b\bar{b}} \times (\mathcal{B}_{W\to l\nu})^2 = 507.1\text{ fb} \times 0.5824 \times 0.3259 \times 0.3259 = 31.368\text{ fb}$

Table 7.13: $t\bar{t}Ztob\bar{b}$ background samples. Effective cross section is provided in the bottom part.

Channel	Year	Dataset
$Z \to b\bar{b}$	2016	/TTZTobb_TuneCP5_13TeV−amcatnlo−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/TTZTobb_TuneCP5_13TeV−amcatnlo−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTZTobb_TuneCP5_13TeV−amcatnlo−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/TTZTobb_TuneCP5_13TeV−amcatnlo−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2

$\sigma \times \mathcal{B}_{Z\to b\bar{b}} = 841 \text{ fb} \times 0.15 = 126.15 \text{ fb}$

Table 7.14: $t\bar{t}ZZ \to b\bar{b}b\bar{b}$ and $t\bar{t}ZH \to b\bar{b}b\bar{b}$ background samples. Effective cross sections are provided in the bottom part, for two different samples.

Channel	Year	Dataset
$ZZ \to b\bar{b}b\bar{b}$	2016	/TTZZTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/TTZZTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTZZTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/TTZZTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2
$ZH \to b\bar{b}b\bar{b}$	2016	/TTZHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016] − 106X_mcRun2_asymptotic_v17−v2
		/TTZHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2016APV] − 106X_mcRun2_asymptotic_preVFP_v11−v1
	2017	/TTZHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2017] − 106X_mc2017_realistic_v9−v2
	2018	/TTZHTo4b_TuneCP5_13TeV−madgraph−pythia8/ [2018] − 106X_upgrade2018_realistic_v16_L1v1−v2

$\sigma \times (\mathcal{B}_{Z\to b\bar{b}})^2 = 1.98 \text{ fb} \times (0.15)^2 = 0.045 \text{ fb}$
$\sigma \times \mathcal{B}_{Z\to b\bar{b}} \times \mathcal{B}_{H\to b\bar{b}} = 1.535 \text{ fb} \times 0.15 \times 0.58 = 0.134 \text{ fb}$

7.3 Trigger requirements

The events in the dilepton channel are chosen online using dilepton triggers for various lepton combinations ($\mathrm{ee}, \mathrm{e}\mu, \mu\mu$) and single-lepton triggers for electrons and muons (e, μ). The trigger paths employed to select events in different years of Run 2 can be found in Tables 7.15 to 7.17. Some of the triggers mentioned in the tables are exclusively used for specific run periods where they were unprescaled, i.e., capturing specific events or processes of interest without any reduction in the event rate. To select events, a logical OR operation is performed between the various triggers applicable to each channel, ensuring that events are not counted twice. For each dilepton stream ($\mathrm{ee}, \mathrm{e}\mu, \mu\mu$), only triggers corresponding to the same lepton stream are considered, and the remaining trigger paths are vetoed to prevent double counting. The inclusion of single lepton triggers alongside dilepton triggers is a strategic choice aimed at maximizing the efficiency of event selection.

Table 7.15: List of triggers used for the 2016 data.

Stream	Trigger Paths dilepton channel	Run Era
e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	B-H
	HLT_Ele27_WPTight_Gsf	B-H
$e^\pm\mu^\pm$	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	B-H
	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	B-H
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL	B-H
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	B-H
	HLT_Ele27_WPTight_Gsf	B-H
	HLT_IsoMu24	B-H
	HLT_IsoTkMu24	B-H
$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL	B-G
	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ	H
	HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL	B-G
	HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ	H

Table 7.16: List of triggers used for the 2017 data.

Stream	Trigger Paths dilepton channel	Run Era
e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL	B-F
	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	B-F
	HLT_Ele32_WPTight_Gsf	B-F
$e^\pm\mu^\pm$	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	B-F
	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	B-F
	HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	B-F
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	B-F
	HLT_Ele32_WPTight_Gsf	B-F
	HLT_IsoMu24_eta2p1	B-D
	HLT_IsoMu27	B-F
$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ	B-F
	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	B-F
	HLT_IsoMu24_eta2p1	B-D
	HLT_IsoMu27	B-F

Table 7.17: List of triggers used for the 2018 data.

Stream	Trigger Paths dilepton channel	Run Era
e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL	A-D
	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	A-D
	HLT_Ele32_WPTight_Gsf	A-D
$e^\pm\mu^\pm$	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	A-D
	HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	A-D
	HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	A-D
	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	A-D
	HLT_Ele32_WPTight_Gsf	A-D
	HLT_IsoMu24	A-D
$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8	A-D
	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	A-D
	HLT_IsoMu24	A-D

Trigger efficiency study is performed by checking the ratio of events with the trigger applied to those without the trigger applied. Only a set of the 2017 samples, namely $\mathrm{t\bar{t}HH} \to \mathrm{b\bar{b}b\bar{b}}$, $\mathrm{t\bar{t}(SL)H} \to \mathrm{b\bar{b}}$, and $\mathrm{t\bar{t}(DL)H} \to \mathrm{b\bar{b}}$ are considered. Figure 7.2 shows the comparison of the several kinematic distributions with and without trigger applied conditions for $\mathrm{t\bar{t}HH} \to \mathrm{b\bar{b}b\bar{b}}$ sample, whereas Figures 7.3 to 7.5 present the 2D comparisons of $\mathrm{t\bar{t}(SL)H} \to \mathrm{b\bar{b}}$ and $\mathrm{t\bar{t}(DL)H} \to \mathrm{b\bar{b}}$ samples, respectively. Overall, the efficiency is found quite high.

The trigger performance in simulation is corrected to account for any differences to data. Trigger scale factors are extracted from Ref...., where the values are derived from pre-legacy samples. As this analysis employs the ultra-legacy samples, trigger efficiency discrepancy between these two sample sets is calculated and found to be 1%. This difference is added as a systematic uncertainty.

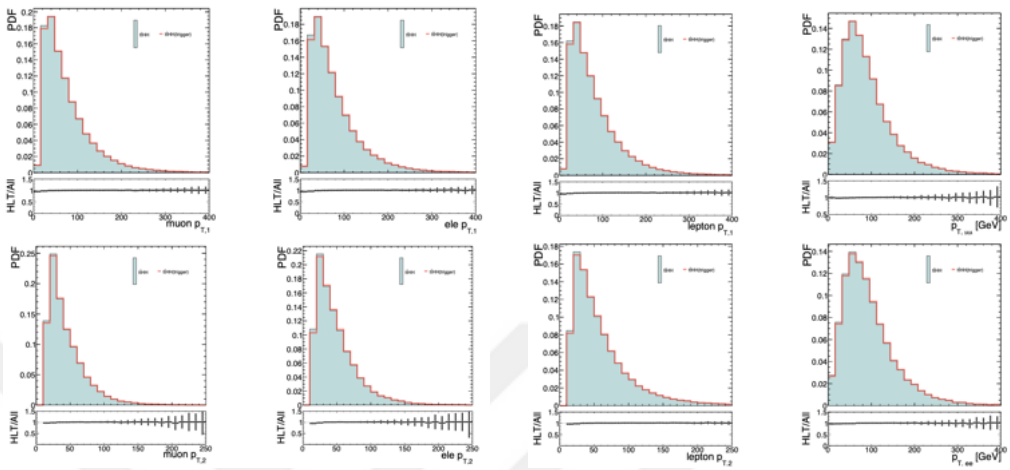

Figure 7.2: Distributions of several lepton kinematical variables to observe the trigger efficiency for the baseline selection. Results are given for the $\mathrm{t\bar{t}HH}$ signal, where red and blue colors show the with and without trigger cases, respectively.

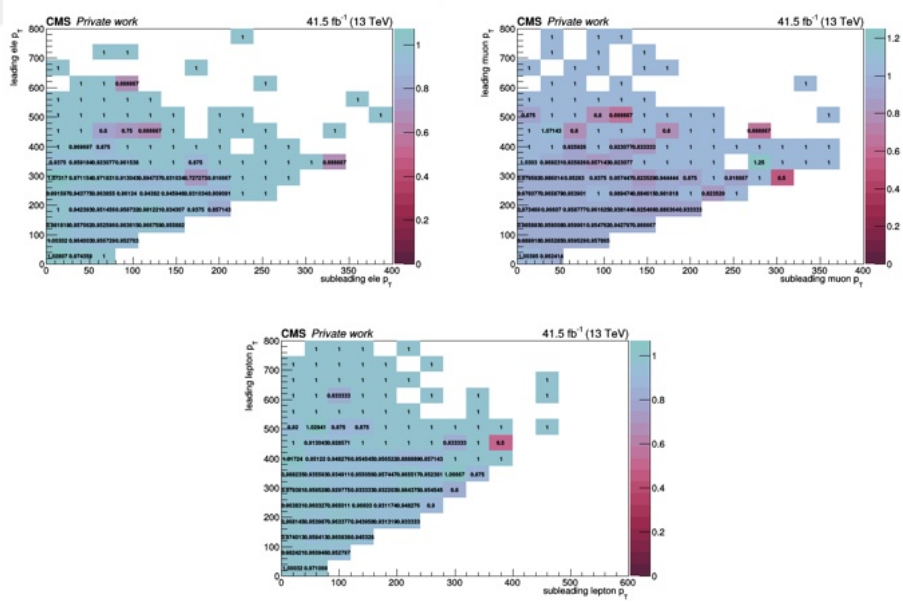

Figure 7.3: 2D plots representing the p_T values of the leading and sub-leading electrons (top-left), muons (top-right), and overall leptons (bottom) for the 2017 $\mathrm{t\bar{t}HH}$ signal sample.

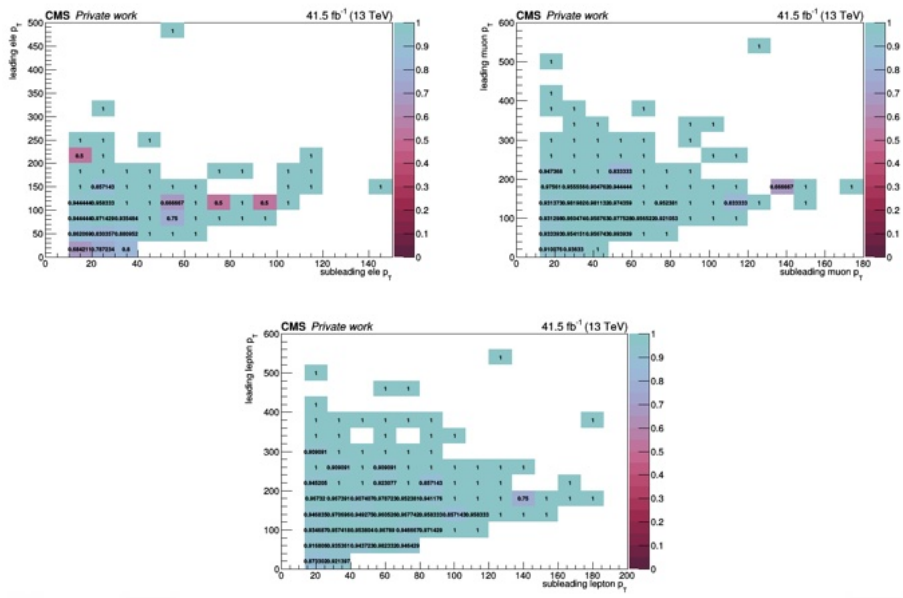

Figure 7.4: 2D plots representing the p_T values of the leading and sub-leading electrons (top-left), muons (top-right), and overall leptons (bottom) for the 2017 $\mathrm{t\bar{t}(SL)Htob\bar{b}}$ sample.

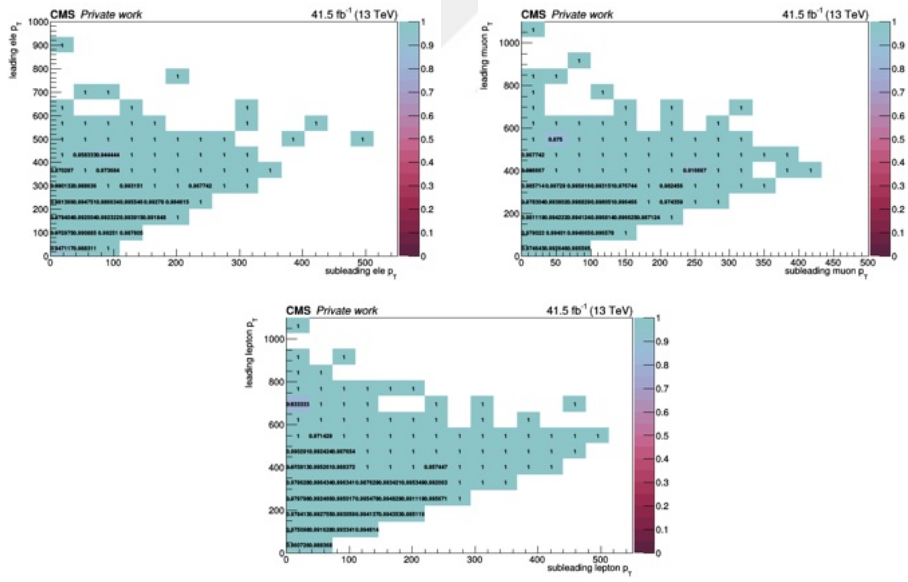

Figure 7.5: 2D plots representing the p_T values of the leading and sub-leading muons (top-left), electrons (top-right), and overall leptons (bottom) for the 2017 $\mathrm{t\bar{t}(DL)Htob\bar{b}}$ sample.

7.4 Object and Event Selection

Chapter 4 provides an overview of the fundamental principles behind CMS particle reconstruction and the definitions of the objects considered in this study. However, every analysis customizes its own identification, isolation and selection algorithms of the objects to optimize the sensitivity.

For this specific study, reconstruction of the full Run 2 analysis is performed with the CMSSW_10_6_28 version. The global tags [146–148] listed in Table 7.18 are used to get the detector and calibration conditions.

Table 7.18: Global tags used for data and simulation.

Year	Data	MC
2016	106X_dataRun2_v27	106X_mcRun2_asymptotic_v17
2017	106X_dataRun2_v20	106X_mc2017_realistic_v9
2018	106X_dataRun2_v24	106X_upgrade2018_realistic_v16

Finally, to carry out an event cleaning for the detector noise, beam halo effects, and other factors, MET filters are applied for both data and MC. These filters, listed in Table 7.19, are implemented in accordance with the recommendations given by the Jet/MET Physics Object Group (JME POG) [149].

Table 7.19: MET filters used in the analysis

Event Filter	**2016**	**2017**	**2018**
Flag_goodVertices	✓	✓	✓
Flag_globalSuperTightHalo2016Filter	✓	✓	✓
Flag_HBHENoiseFilter	✓	✓	✓
Flag_HBHENoiseIsoFilter	✓	✓	✓
Flag_EcalDeadCellTriggerPrimitiveFilter	✓	✓	✓
Flag_BadPFMuonFilter	✓	✓	✓
Flag_BadPFMuonDzFilter	✓	✓	✓
Flag_eeBadScFilter	✓	✓	✓
Flag_ecalBadCalibFilter	-	✓	✓

Objects, object IDs applied and their corresponding p_T and η cuts are listed in Table 7.20.

Table 7.20: Definition of all objects used in the DL channel

Object	ID (object property)	p_T(GeV)	$\|\eta\|$
Leading Electron	medium (mvaFall17V2Iso_WP90)	> 25	< 2.4
Sub-leading Electron	medium (mvaFall17V2Iso_WP90)	> 15	< 2.4
Leading Muon	tight (tightId)	> 25	< 2.4
Sub-leading Muon	tight (tightId)	> 15	< 2.4
Jets	tight (including TightLepVeto)	> 30	< 2.4
b jets (medium)	medium	> 30	< 2.4
b jets (loose)	loose and not medium	> 30	< 2.4
light jets	below loose WP	> 30	< 2.4

7.4.1 Muon reconstruction and selection

In this analysis two types of muons are considered as the leading and the subleading muon. Muon objects are required to fulfill the "tightId" quality criteria as well as an isolation criteria based on the $\delta\beta$-corrected PF-relative isolation with an isolation cone of 0.4 named as pfRelIso04_all. Loose isolation criteria is selected with $\%98$ efficiency by requesting pfRelIso04_all < 0.25. The objective of a "tight muon identification" is to filter out muons that come from in-flight decay and hadronic punch-through. A muon is classified as a "tight muon" if it satisfies several specific conditions [100]:

- Firstly, the muon has to qualify as a "loose muon" and its tracker track should comprise hits from a minimum of six layers in the inner tracker. At least one of these hits must be a pixel hit.
- Secondly, the muon has to be identified as both a tracker muon and a global muon.
- Thirdly, the tracker muon must show segment matching in a minimum of two muon stations.

- Fourthly, the global muon fit should yield a chi-squared per degree of freedom (χ^2/dof) value that is less than 10 and should include a minimum of one hit originating from the muon system.

- Lastly, the tight muon needs to be consistent with the primary vertex, meeting the following sub-criteria: the transverse impact parameter ($|dXY|$) must be less than 0.2 cm and the longitudinal impact parameter ($|dz|$) must be less than 0.5 cm.

These rules together formulate the guidelines for categorizing a muon as a tight muon.

Additionally, there are kinematic requirements based on the muon's $\mathrm{p_T}$and η. For the leading muons, the $\mathrm{p_T}$ threshold is 25 GeV, while it is 15 GeV for subleading muons. On the other hand, the absolute value of η ($|\eta|$) is required to be less than 2.4 for both leading and subleading muons.

The baseline selection applied to muons is summarised in Table 7.21 and is the same for three run years.

Table 7.21: Baseline selection applied to muon objects.

	muon
muon ID	tightId
min. p_T [GeV]	25 (leading), 15 (sub-leading)
max. $\|\eta\|$	2.4
isolation ID	pfRelIso04_all
isolation cut	< 0.25

Dedicated $\mathrm{p_T}$ and η dependent scale factors are applied to the MC events in order to improve the agreement of the reconstruction, ID, and isolation efficiencies with the efficiencies in data, following the recommendations of the Muon Physics Object Group [150].

7.4.2 Electron reconstruction and selection

The categorization of electrons follows a similar approach as muons, distinguishing between leading and subleading electrons. For electron selection, a Multivariate Analysis (MVA) based identification method is used. This approach is especially effective at achieving optimal differentiation at low $\mathrm{p_T}$values, demonstrating strong performance and facilitating high-efficiency multi-lepton analyses. The specific electron ID version utilized in this thesis is mvaFall17V2Iso_WP90. It incorporates three PF isolation components as input variables, ensuring a signal efficiency of 90%. This identification method is trained on DY+Jets MC samples where prompt electrons are treated as signal and unmatched plus non-prompt electrons are regarded as background [151]. The isolation criteria is selected as pfRelIso03_all (< 0.06), again based on the $\delta\beta$-corrected relative PF-relative isolation with an isolation cone of 0.3.

Beyond the WP90, there exist two additional working points denoted as WP98 and WP80, corresponding to loose and tight criteria, respectively. Prior to the selection of the current working point, a comprehensive and dedicated study was conducted to evaluate signal and background efficiencies by comparing event yields for all three working points. Ultimately, WP90 was chosen as the preferred option since it offered the most optimal signal-to-background ratio. WP80 was found to reduce signal efficiency, while WP98 increased both signal and background efficiency. Detailed comparative tables can be found in the Appendix B for further reference.

As in the case of muons, kinematic requirements on $\mathrm{p_T}$are decided as 25 GeV and 15 GeV for the leading and the subleading electrons, respectively. The $|\eta|$ condition they need to fulfill is selected as 2.4. Electrons are also discarded if they are tracked in the η values $|\eta_{Supercluster}| < 1.4442$ and $|\eta_{Supercluster}| > 1.5560$, which correspond to the barrel and the endcap regions, respectively. The baseline selection applied to electrons is summarised in Table 7.22.

Dedicated $\mathrm{p_T}$ and η dependent scale factors are applied to the MC events in order to improve the agreement of the reconstruction, and ID efficiencies with the efficiencies in data, following the recommendations of the Electron and Photon Physics Object Group [152].

Table 7.22: Baseline selection applied to electron objects.

	electron
electron ID	mvaFall17V2Iso_WP90
min. p_T [GeV]	25 (leading) , 15 (subleading)
max. $\|\eta\|$	2.4
$\|\eta_{\text{Supercluster}}\| <$	1.4442
$\|\eta_{\text{Supercluster}}\| >$	1.5560
isolation ID	pfRelIso03_all
isolation cut	< 0.06

7.4.3 Jet reconstruction and selection

As a default implementation of NANOAOD, this study employs slimmed jets, specifically AK4 PFJets CHS with applied JECs, following the initial selection criteria ($p_T > 15$ GeV) [153]. The CHS algorithm [97] uses reconstructed vertex information obtained from charged-particle tracks. For the selection of the primary pp interaction vertex, the algorithm considers track jets formed through the anti-kT clustering algorithm [103], with tracks associated with the vertex as inputs. Additionally, the associated $\vec{p}_{T,tracks}^{miss}$, which represents the negative vector sum of $\mathrm{p_T}$for these jets, is taken into account. The primary pp interaction vertex, also known as the "leading vertex" (LV), is identified as the reconstructed vertex with the highest summed physics-object p_T^2. Other reconstructed collision vertices are referred to as PU vertices. The CHS algorithm employs tracking information to identify particles originating from PU events after PF candidates have been reconstructed but prior to any jet clustering. This procedure involves removing charged particle candidates associated with a reconstructed PU vertex. Specifically, a charged particle is considered associated with a PU vertex if it has been utilized in the fit for that particular PU vertex. Charged particles not linked to any PU vertex, along with all neutral particles, are retained [154]. To ensure the quality of the jets, multiple criteria are applied following the recommendations provided by the JME POG. For all three years, jets are required to satisfy the "jetId==6" criteria, indicating successful passage of both the tight and the TightLepVeto ID. Moreover, in order to mitigate the impact of PU events, jets with pT $<$ 50 GeV must meet the criteria of the loose working point for the "pileup

jet ID". PUID BDT is exclusively trained on jets with $p_T < 50$ GeV, which represents the region with the highest concentration of PU jets and where the PUID is most necessary. As a result, this criterion is not extended to jets with $p_T > 50$ GeV [155]. In the 2017 and 2018 samples, the flag "puId==4" is utilized. However, for the 2016 sample, this same flag indicates the meeting of the loose ID criteria while failing to meet the medium and tight ones. In place of "puId==4," the 2016 sample exclusively uses the flag "puID==1" [153]. The baseline selection applied to jets is summarised in Table 7.23.

Table 7.23: Baseline selection and flags applied to jet objects.

	2016	2017	2018
jet ID	jetId==6	jetId==6	jetId==6
pileup jet ID	puId==1	puId==4	puId==4
min. p_T [GeV]	30	30	30
max. $\|\eta\|$	2.4	2.4	2.4

7.4.4 b jet identification

In this study, the btagDeepFlavB flag is used as a b tagging discriminant to identify jets originating from b-quark decays. Both the medium and the loose b jets are taken into account for calculating different kinematical variables as described in Section 7.5. Medium b jets need to pass medium working point, whereas loose b jets need to pass the loose working point and are explicitly required to fail the medium b tagging criteria in order to avoid overlaps. A jet is called light jet if it fails the loose b tag working point. As mentioned in Chapter 4, these operating points correspond to misidentification probabilities of approximately 10%, 1%, and 0.1%, respectively, for jets originating from light particles, at an average $\mathrm{p_T}$ of around 80 GeV/c.

Details of the working points for each year are given in Table 7.24. The reference b jet efficiencies for each year and working points are calculated before data/MC SFs on a QCD sample with jet $\mathrm{p_T} > 30$ GeV and provided by CMS B Tagging and Vertexing Physics Object Group (BTV POG) [156] (Table 7.25).

Typically, there are variations between the efficiency of identifying b jets and the

Table 7.24: Baseline selection and flags applied to b jet objects.

	2016 preVFP	2016 postVFP	2017	2018
btagDeepFlavB (medium)	> 0.2598	> 0.3657	> 0.3040	> 0.2783
btagDeepFlavB (loose, not medium)	> 0.0508 < 0.2598	> 0.0816 < 0.3657	> 0.0532 < 0.3040	> 0.0490 < 0.2783
btagDeepFlavB (light jet)	≤ 0.0508	≤ 0.0816	≤ 0.0532	≤ 0.0490
min. p_T [GeV]	30	30	30	30
max. $\|\eta\|$	2.4	2.4	2.4	2.4

Table 7.25: The reference b jet efficiencies for each year and working points.

Working point	2016 preVFP	2016 postVFP	2017	2018
loose	87.3%	86.3%	91%	91.5%
medium	73.3%	71.4%	79.1%	80.7%
tight	57.5%	54.7%	61.6%	65.1%

likelihood of incorrectly identifying non-b jets as b jets in both data and simulation. As the b jet related variables plays an important role in the DNN training, taking these discrepancies into account by applying correction factors, or scale factors, to both the simulated samples and the data is necessary. The b tagging and light flavor jets efficiency scale factors are provided by the BTV POG [156]. The corrections are determined based on the b tagging discriminant value, as well as the jet's $\mathrm{p_T}$ and $|\eta|$, and it is calculated independently for both light-flavour and b jets. Effect of the b-tagging SFs to the event variables effectively used in the DNN training is also studied. The comparison distributions of some of these variables for the $\mathrm{t\bar{t}HH}$ signal with and without b-tag corrections cases are given in Figure 7.6.

7.4.5 Event Selection

Once the object properties are established, the baseline event selection criteria outlined in Table 7.26 are applied. Besides the primary hard interaction, various other processes contribute to the final state. These include multi-parton interactions (MPI),

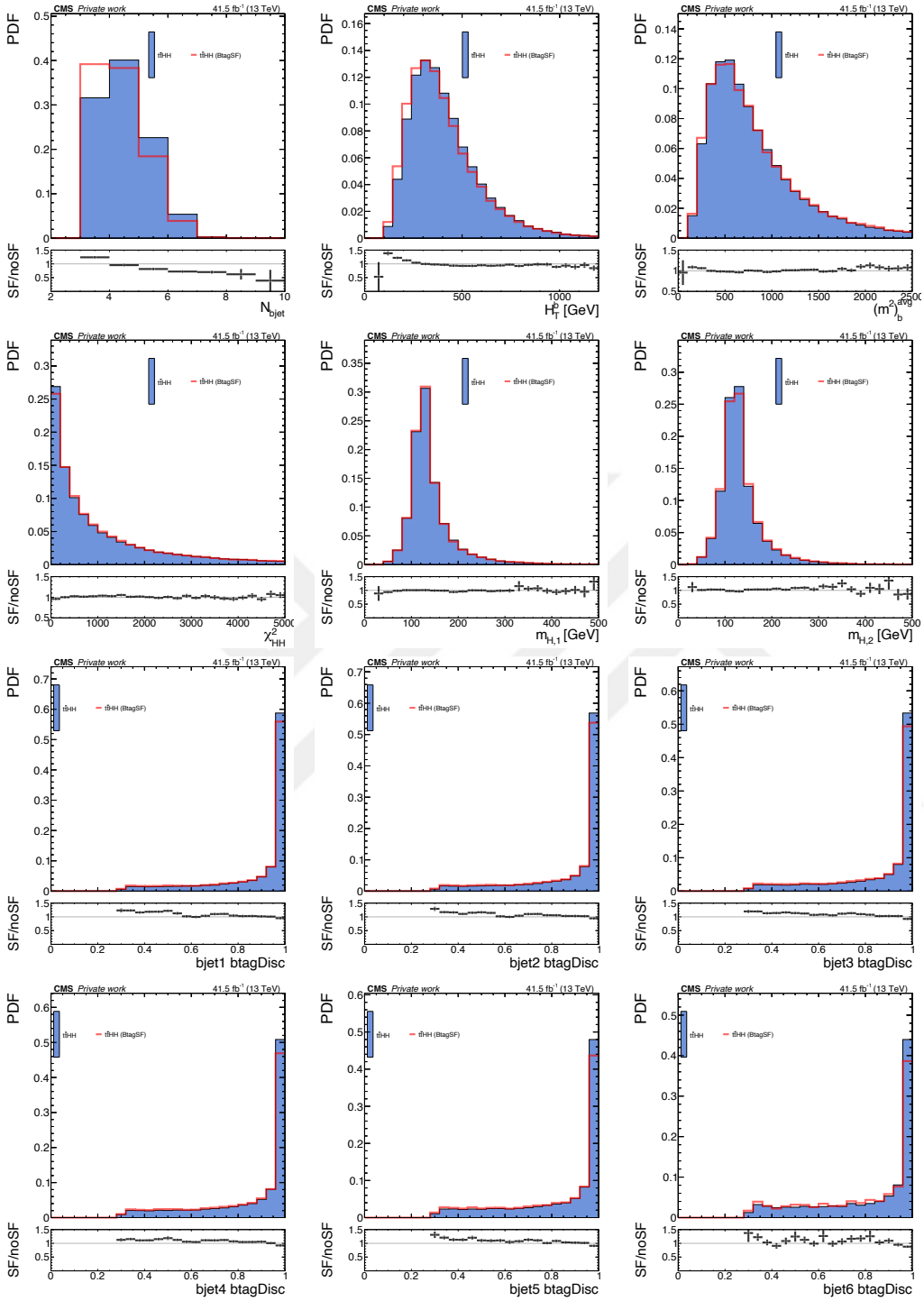

Figure 7.6: Distributions of the various kinematical variables comparing the $t\bar{t}HH$ signal samples with (red) and without (blue) b tag corrections applied conditions for 2017.

extra proton-proton interactions (pileup), initial-state radiation (ISR), and final-state radiation (FSR). These processes mainly add light-flavor jets, potentially resulting in more than 4 jets in the final state. Therefore, the event selection in this study requires a minimum of 4 jets, accommodating these extra jets. However, some jets may not be fully reconstructed due to the limitations of the detector's coverage. This is particularly true for the b-tagging algorithm, which may struggle to precisely identify b jets. To mitigate this, events are selected based on having at least 4 jets, with the additional condition that 3 of these jets must be effectively b-tagged. Moreover, events are chosen based on the presence of precisely two oppositely charged leptons, such that e^+e^-, or $\mu^\pm e^\mp$, $\mu^+\mu^-$. To be selected, events must pass a trigger relying on online leptons with a flavor content consistent with the two leptons chosen offline (e.g., $\mu^+\mu^-$ events must satisfy a requirement of events to pass a dimuon or single muon trigger), as already described in Section 7.3. Additionally, the invariant mass of the selected lepton pair (m_{ll}) must exceed 20 GeV to suppress events from heavy-flavour resonance decays and low mass Drell–Yan processes. In the case of same-flavor channels, events are rejected if the m_{ll} lies between 76 GeV and 106 GeV, effectively reducing the contribution from Z+jets events. Furthermore, events must meet the condition $p_T^{miss} > 40$ GeV to suppress background contributions, for instance, those arising from QCD-multijet production, and to account for the neutrinos coming from the W-boson decay.

Table 7.26: Baseline event selection.

	Baseline Selection
Number of jets	$>= 4$
Number of medium b-tagged jets	$>= 3$
Number of leptons	$= 2$
MET (GeV)	> 40
Sign and flavour of leptons	e^+e^-, $\mu^\pm e^\mp$, $\mu^+\mu^-$
Min. $m_{e^+e^-/\mu^+\mu^-}$ [GeV]	20
$m_{e^+e^-/\mu^+\mu^-}$ [GeV]	<76, >106

7.5 Event Variables for DNN

The primary objective of an experimental physicist is to comprehend and derive meaningful insights from the data they gather. Consequently, it is crucial to extract features, known as variables in experimental physics, guided by the principles of physics theories or hypotheses. To achieve this goal, this section focuses on discussing the variables employed in the current analysis.

In all events that meet the baseline selection criteria, a collection of event variables is derived to describe the event's topology and kinematics. These variables, which will be elaborated upon in Section 7.8 in the context of DNN training, exhibit distinct characteristics for both signals and backgrounds. The details of these variables, as presented in Table 7.27, cover a range of features that can effectively discriminate between signal and background events in this signal-background separation analysis.

Table 7.27: Event quantities calculated for all events passing the baseline selection, including object properties, invariant masses, angular variables. Jets and b jets are ordered according to decreasing p_T. See text for further details.

Group	Variables
Object multiplicities	N_{jets}, N_{bjets}, N_{light}
Object 4-momenta (objects are ordered by decreasing p_T)	$\mathrm{p_T}$ & $\lvert\eta\rvert$ jet 1, 2, 3, 4, 5, 6 $\mathrm{p_T}$ & $\lvert\eta\rvert$ bjet 1, 2, 3, 4, 5, 6 $\mathrm{p_T}$ & $\lvert\eta\rvert$ lepton 1, 2 $\mathrm{p_T}$ & $\lvert\eta\rvert$ dimuon and dielectron $\mathrm{p_T}$ & $\lvert\eta\rvert$ muon and electron 1, 2
Hadronic transverse momenta	H_T, H_T^b, H_T^{light}, H_T^{lepton}, S_T
Mass variables	m_j^{avg}, m_b^{avg}, m_{light}^{avg}, $(m^2)_b^{avg}$ $m_{jjj}^{maxp_T}$, $m_{jbb}^{maxp_T}$, $m_{\mu\mu}$, m_{ee}
Angular separation variables	$\Delta\eta_{jj}^{avg}$, $\Delta\eta_{bb}^{avg}$, $\Delta\eta_{bb}^{max}$ ΔR_{jj}^{avg}, ΔR_{bb}^{avg}, ΔR_{jj}^{min}, ΔR_{bb}^{min} $\Delta R_{jj,mass}^{min}$, $\Delta R_{jj,p_T}^{min}$, $\Delta R_{bb,mass}^{min}$, $\Delta R_{bb,p_T}^{min}$
Optimized χ^2 values	χ^2_{HH}, χ^2_{ZZ}, χ^2_{ZH}
Invariant masses	$m_{H,1}$, $m_{H,2}$, $m_{Z,1}$, $m_{Z,2}$, $m_{ZH,Z}$, $m_{ZH,H}$
Reconstructed Higgs momenta	$p_{T(H,1)}$, $p_{T(H,2)}$
Event shape variables	aplanarity, centrality sphericity, trans sphericity, C value, D value (for jets and b jets)

The fundamental variables encompass information related to object multiplicities, including jets and b jets, as well as the transverse momenta and pseudorapidities of the initial six jets and b jets. Since the Run 2 study explores the dilepton channel, the analysis extends to include $\mathrm{p_T}$and η values not only for the leading and subleading leptons but also for the e and μ objects in the dielectron, dimuon, and mixed electron-muon event categories.

Meanwhile, more complex variables involve computations like the scalar sum of transverse momenta (H_T), the calculation of average mass (m^{avg}), or the determination of average mass-squared ($(m^2)_b^{avg}$) for b jets. Additionally, various angular quantities are established based on differences in pseudorapidity ($\Delta\eta$) and azimuthal separation ($\Delta\phi$) between pairs of jets or b jets. Moreover, average or maximum values for $\Delta\eta$ between two jets or two b jets are considered. Another noteworthy variable is ΔR, which is defined in Section 3.2 in details. Average and minimum ΔR values between two jets or two b jets are served as inputs to DNN. Other complex variables with high discrimination power are explained in detail below.

Higgs and Z bosons χ^2

Additional strongly discriminative variables can be generated by forming the diboson systems HH, ZZ, and ZH, with each boson assumed to undergo decay into b quark pairs. To reconstruct these systems, the most probable decay products for the bosons are determined using a minimization process involving three distinct χ^2 variables:

$$\begin{aligned} \chi^2_{XY} &= \frac{(m_{j_1j_2} - m_X)^2}{\sigma^2_{j_1j_2}} + \frac{(m_{j_3j_4} - m_Y)^2}{\sigma_{j_3j_4{}^2}}, \\ [X, Y] &= [H, H],\ [H, Z],\ [Z, Z] \end{aligned} \tag{7.2}$$

The symbols m_{j_kjl} represent the invariant mass, specifically the sum of the four-vectors identified in the subscripts. In this context, m_Z corresponds to the mass of the Z boson, which is 91.2 GeV$/c^2$, while m_H signifies the mass of the Higgs boson, which is 125 GeV$/c^2$. The quantities σj_kj_l refer to the mass resolution, and these are determined by taking into account the propagation of the p_T-dependent energy resolutions associated with the input jets, denoted as j_k and j_l. the energy resolution of individual input jets is taken as a function of JER. In accordance with the

details provided in Section 7.4.4, the analysis considers both medium and light b jets, each designated for specific b jet multiplicity scenarios. When an event involves four medium b jets, all four are included in the analysis. In cases where there are more than four medium b jets, all b jets are factored into the computation of the χ^2, and the combination of four b jets yielding the smallest χ^2 value is chosen. In scenarios with three medium b jets and at least one loose b jet, all medium b jets are obligatory components in the χ^2 calculation, with the fourth jet selected from the loose b jets pool. For events characterized by three medium b jets and no loose b jets, all medium b jets are mandated to participate in the χ^2 computation, with the fourth jet chosen from among the light jets. In all these instances, an iterative process is employed to assess all potential permutations of candidate b jet and jet configurations, ultimately selecting those configurations that yield the lowest χ^2 values [15].

The reconstructed Higgs and/or Z invariant masses giving the minimum χ^2 values are also used as discriminating variables. For χ^2_{HH}, $m_{H,1}$ and $m_{H,2}$ are the $m_{j_k j_l}$ values that are closest and second closest to the Higgs boson mass. For χ^2_{ZZ}, $m_{Z,1}$ and $m_{Z,2}$ are the $m_{j_k j_l}$ values that are closest and second closest to the Z mass. For χ^2_{ZH}, $m_{ZH,H}$ and $m_{ZH,Z}$ are the $m_{j_k j_l}$ values that are closest to the Higgs boson and Z boson masses, respectively. These mass variables also have discriminative properties and the distributions shown in Figure 7.7 give a comparison among the signal sample $\mathrm{t\bar{t}HH}$, $\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}ZZ}$ backgrounds.

Event shape variables

Event shapes typically serve as parameters that describe the geometric characteristics of the energy-momentum distribution within an event. These parameters directly reflect how an event appears when observed in the detector. In the context of hadron collisions, where the center-of-mass frame of the interaction is frequently boosted along the beam axis, event shape observables are commonly defined using transverse momenta. This choice is made because transverse momenta remain Lorentz invariant even when subjected to such boosts, making them a suitable basis for characterizing event shapes in this scenario [157–159].

Event shape variables are calculated using the eigenvalues, namely $(\lambda_1, \lambda_2,$ and $\lambda_3)$, of the full momentum tensor of the event, denoted as M_{xyz}

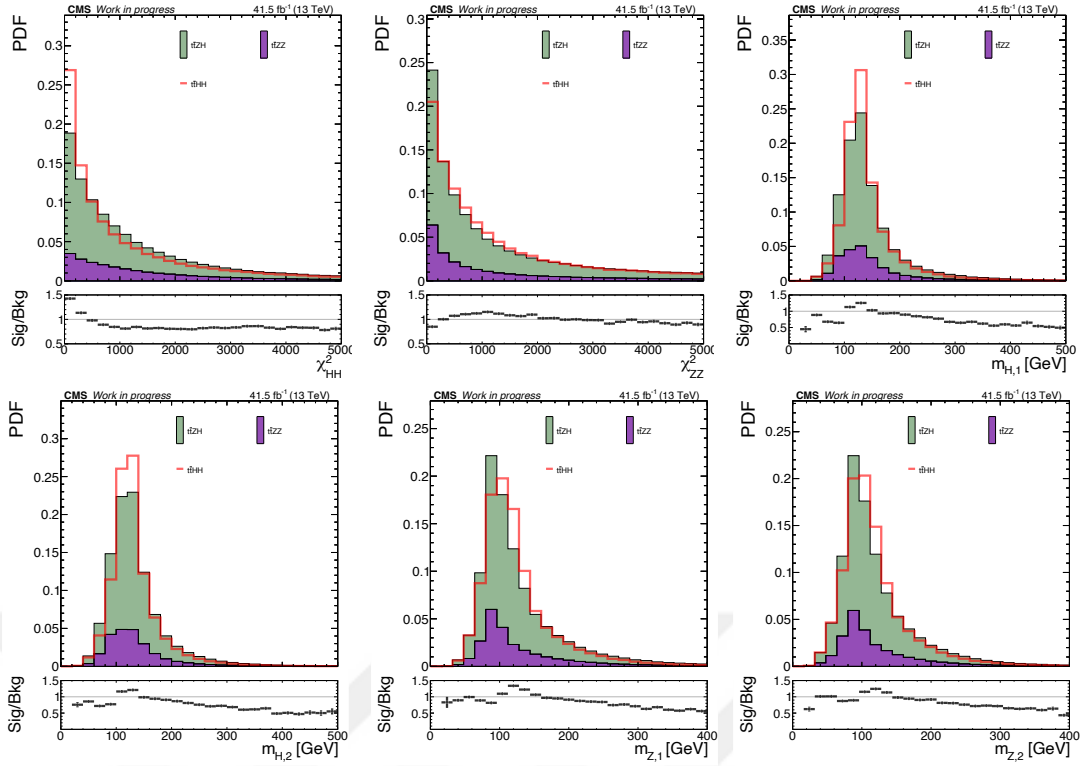

Figure 7.7: Invariant mass distributions comparing the $t\bar{t}HH$ signal (red) with $t\bar{t}ZH$ (green) and $t\bar{t}ZZ$ (yellow) backgrounds for the 2017 samples.

$$M_{xyz} = \sum_i \begin{pmatrix} p_{xi}^2 & p_{xi}p_{yi} & p_{xi}p_{zi} \\ p_{yi}p_{xi} & p_{yi}^2 & p_{yi}p_{zi} \\ p_{zi}p_{xi} & p_{zi}p_{yi} & p_{zi}^2 \end{pmatrix} \tag{7.3}$$

The sum is performed over all jets employed in the measurement. The individual eigenvalues are then normalized and arranged such that $(\lambda_1 > \lambda_2 > \lambda_3)$. These eigenvalues are required to sum to 1, as stated by $\sum_i \lambda_i = 1$ by definition. These eigenvalues serve as the basis for defining three observables known as sphericity, transverse sphericity, and aplanarity.

Sphericity and transverse sphericity quantify the total transverse momentum relative to the sphericity axis established by the four-momenta employed in the event shape measurement, particularly the first eigenvector. The range of allowed values for Sphericity is $0 \leq S < 1$, with a balanced dijet event corresponding to S=0 and an isotropic event yielding S=1. On the other hand, transverse sphericity is constructed using the two largest eigenvalues, and its typical range aligns with the allowed range,

$0 \leq S_\perp < 1$. These observables are computed as follows:

$$S = \frac{3}{2}(\lambda_2 + \lambda_3), \qquad S_\perp = \frac{2\lambda_2}{(\lambda_1 + \lambda_2)} \tag{7.4}$$

Aplanarity quantifies the degree of transverse momentum within or perpendicular to the plane defined by the two leading jets, solely considering the smallest eigenvalue of M_{xyz}, denoted as λ_3. Its valid range is $0 \leq A < 1/2$, and typically, measured values fall within $0 \leq A < 0.3$, with values close to zero indicating events that are relatively planar. It is expressed as:

$$A = \frac{3}{2}\lambda_3 \tag{7.5}$$

It is also possible to define event shape variables without explicitly referring to a specific axis. The most well-known examples of these are the C and D parameters:

$$C = 3(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1), \qquad D = 27\lambda_1\lambda_2\lambda_3 \tag{7.6}$$

Centrality

This topological variable utilizes the angular dependence of event yields from various processes along the beam-line axis. Signal events typically display localization on the transverse plane, while background processes tend to have a more uniform distribution across the azimuthal angle. The centrality of an event is quantified as follows:

$$Centrality = \frac{\sum\limits_{jets,l} p_T}{\sum\limits_{jets,l} p} \tag{7.7}$$

Fox-Wolfram Moments

The Fox-Wolfram moments [160], a distinct category within event shape variables, offer a unique perspective on event topologies. Unlike other variables that primarily

consider angular changes between jets and fixed axes, Fox-Wolfram moments take into account both the momenta of the jets and the total angular separation between pairs of jets within an event. This dual-stage dependency sets them apart, making them a valuable tool for characterizing the final state topology in e^+e^- annihilation processes [161].

The Fox-Wolfram moments are defined by

$$H_l = \frac{4\pi}{2l+1} \sum_{m=l}^{-l} \left| \sum_i Y_l^m(\theta_i, \phi_i) \frac{|\vec{p_i}|}{\sqrt{s}} \right|^2 \tag{7.8}$$

where $\mathrm{Y_l^m}$ represents the spherical harmonics and θ and ϕ are the spherical coordinates. By expending the square the following form can be obtained

$$H_l = \frac{4\pi}{2l+1} \sum_{m=l}^{-l} \left(\sum_i Y_{lm}(\theta_i, \phi_i) \frac{|\vec{p_i}|}{\sqrt{s}} \right) \left(\sum_i Y_{lm}^*(\theta_j, \phi_j) \frac{|\vec{p_j}|}{\sqrt{s}} \right) \tag{7.9}$$

$$= \sum_{ij} \frac{\vec{p_i}\vec{p_j}}{s} \frac{4\pi}{2l+1} \sum_{m=l}^{-l} Y_{lm}(\theta_i, \phi_i) Y_{lm}^*(\theta_j, \phi_j) \tag{7.10}$$

$$= \sum_{ij} \frac{\vec{p_i}\vec{p_j}}{s} P_l(\cos \Omega_{ij}) \tag{7.11}$$

Mathematically, the Fox-Wolfram moments are defined using spherical harmonics and spherical coordinates, resulting in a formulation that relates them to Legendre polynomials (P_l) and the total angle between jets (Ω). They are computed by scanning through and summing over all possible combinations of jets, including identical jet combinations. In the analysis, Fox-Wolfram moments $\mathrm{H_0}$, $\mathrm{H_1}$, , $\mathrm{H_2}$, $\mathrm{H_3}$, and $\mathrm{H_4}$ are used as inputs to a Deep Neural Network (DNN), both for jets and b jets. Additionally, the ratios $\mathrm{R_i} = \frac{\mathrm{H_i}}{\mathrm{H_0}}$ provide another set of DNN inputs, further emphasizing the importance of Fox-Wolfram moments in characterizing event topologies.

Figures 7.8 and 7.9 show the distributions of certain variables, demonstrating significant discrimination power between the signal and backgrounds for the 2017 samples. Similarly, Fig. 7.10 illustrates some of these distributions for the 2018 samples,

whereas Fig 7.11 shows these for the 2016 samples. All distributions are normalized to luminosity, 49.5 fb^{-1}, 59.7 fb^{-1} and 36.3 fb^{-1}, respectively. The ratio plots in the lower panels are generated by normalizing signal and total background yields to 1, facilitating a direct comparison of the shapes between the signal and total background. Similar distributions comparing signal with individual background categories are shown in Appendix C.

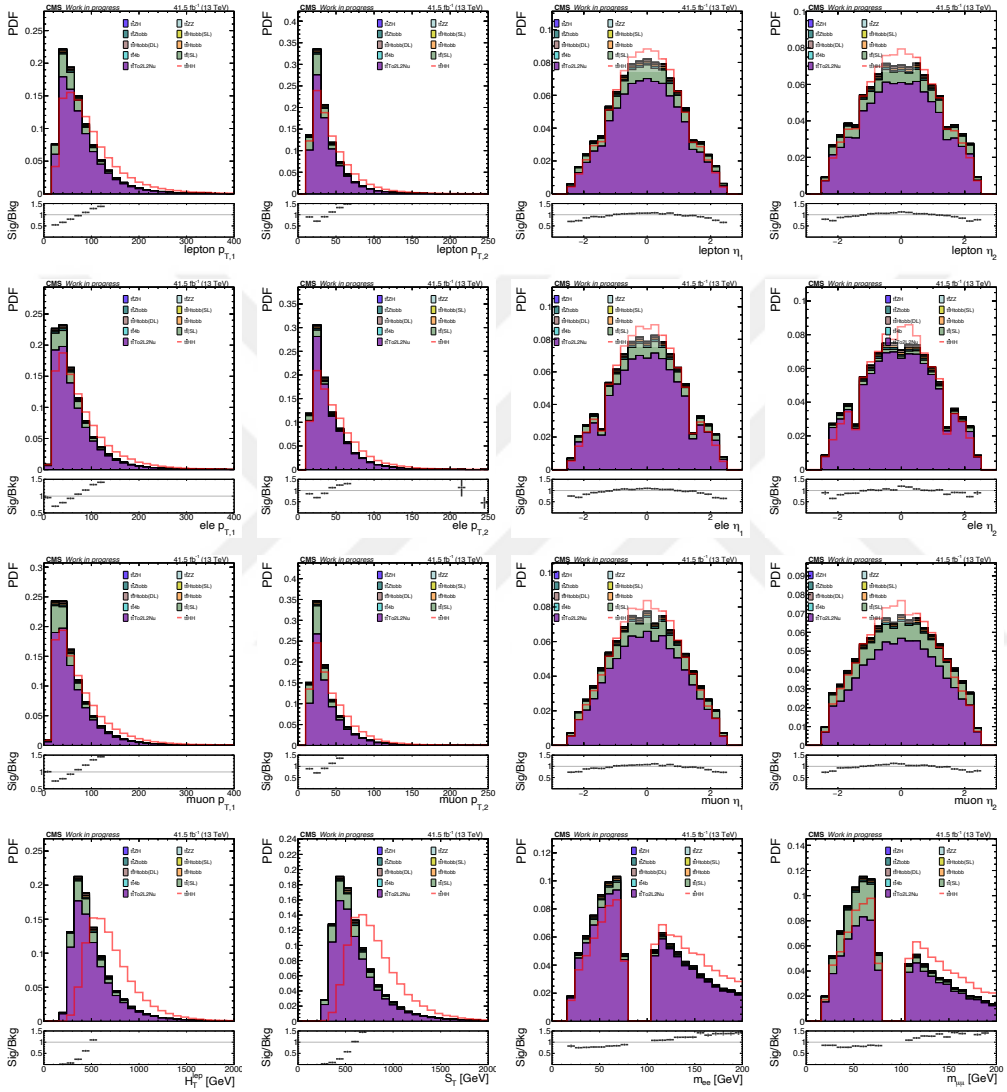

Figure 7.8: Distributions of different discriminating variables for leptons for the 2017 samples. The baseline selection is applied to the $t\bar{t}HH$ signal and the backgrounds, normalized to 41.5 fb^{-1} luminosity.

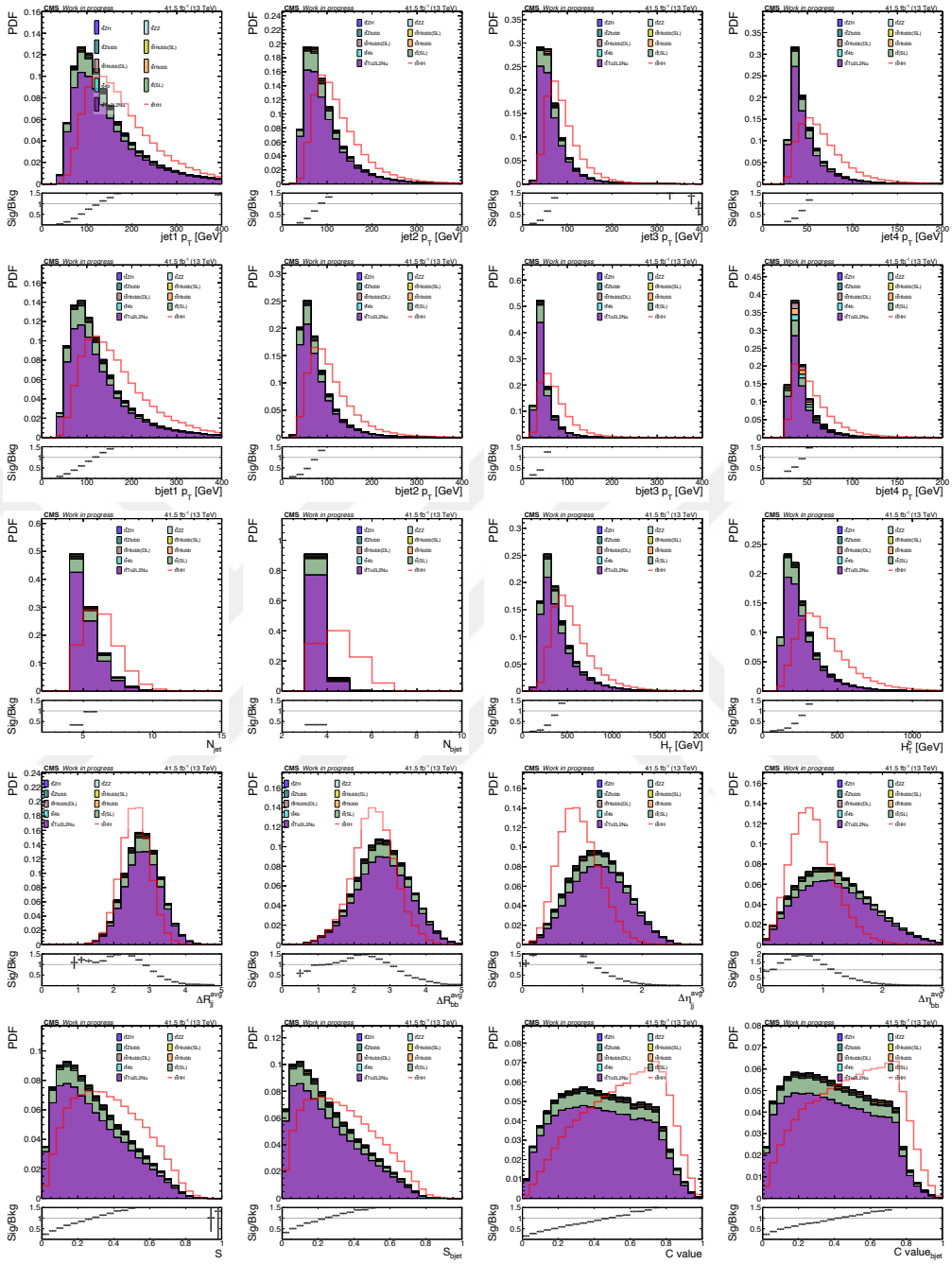

Figure 7.9: Distributions of different discriminating variables for jets and b-tagged jets for the 2017 samples. The baseline selection is applied to the $t\bar{t}HH$ signal and the backgrounds, normalized to 41.5 fb^{-1} luminosity.

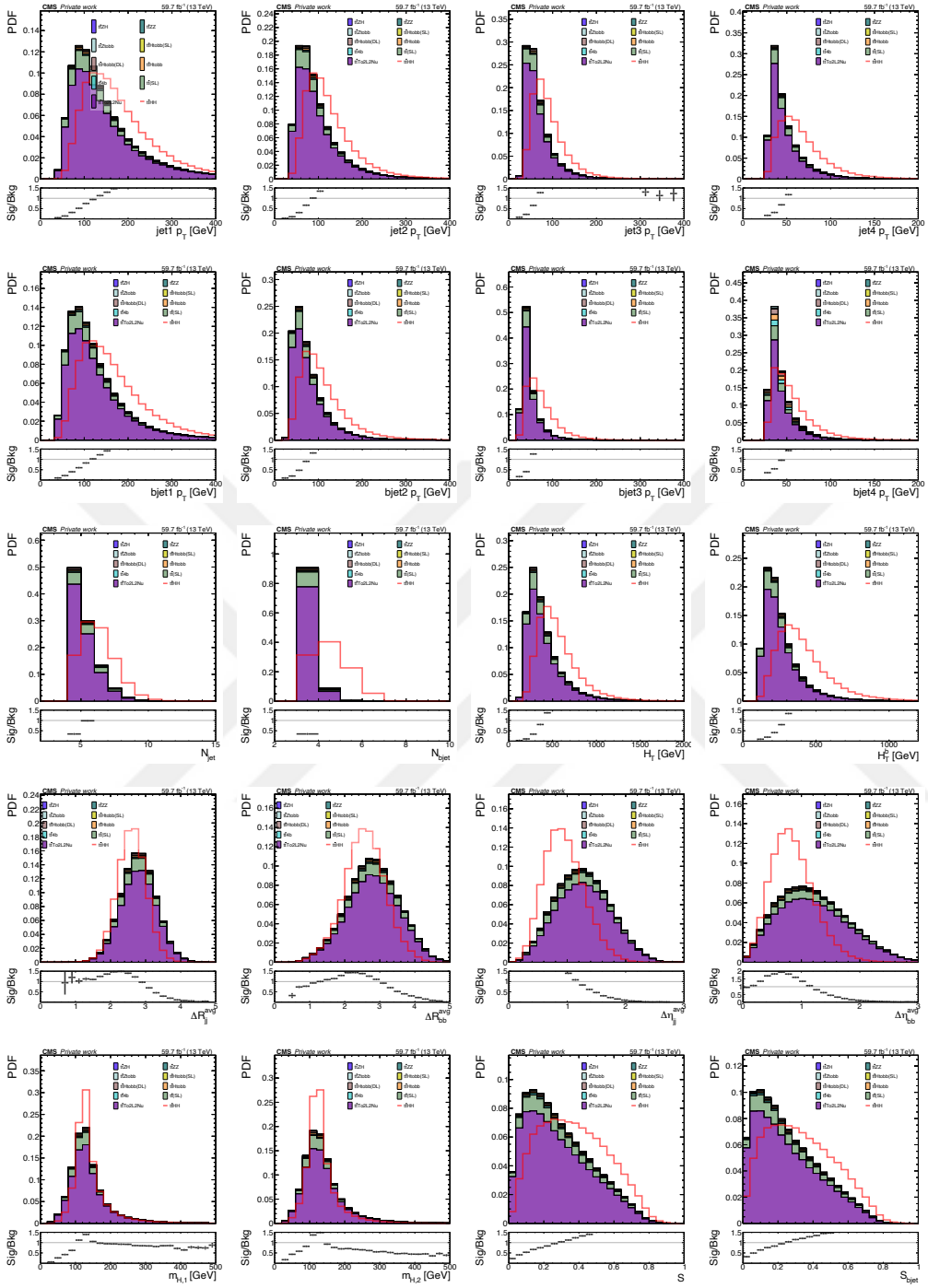

Figure 7.10: Distributions of different discriminating variables for jets and b-tagged jets for the 2018 samples. The baseline selection is applied to the $\mathrm{t\bar{t}HH}$ signal and the backgrounds, normalized to 59.7 fb^{-1} luminosity.

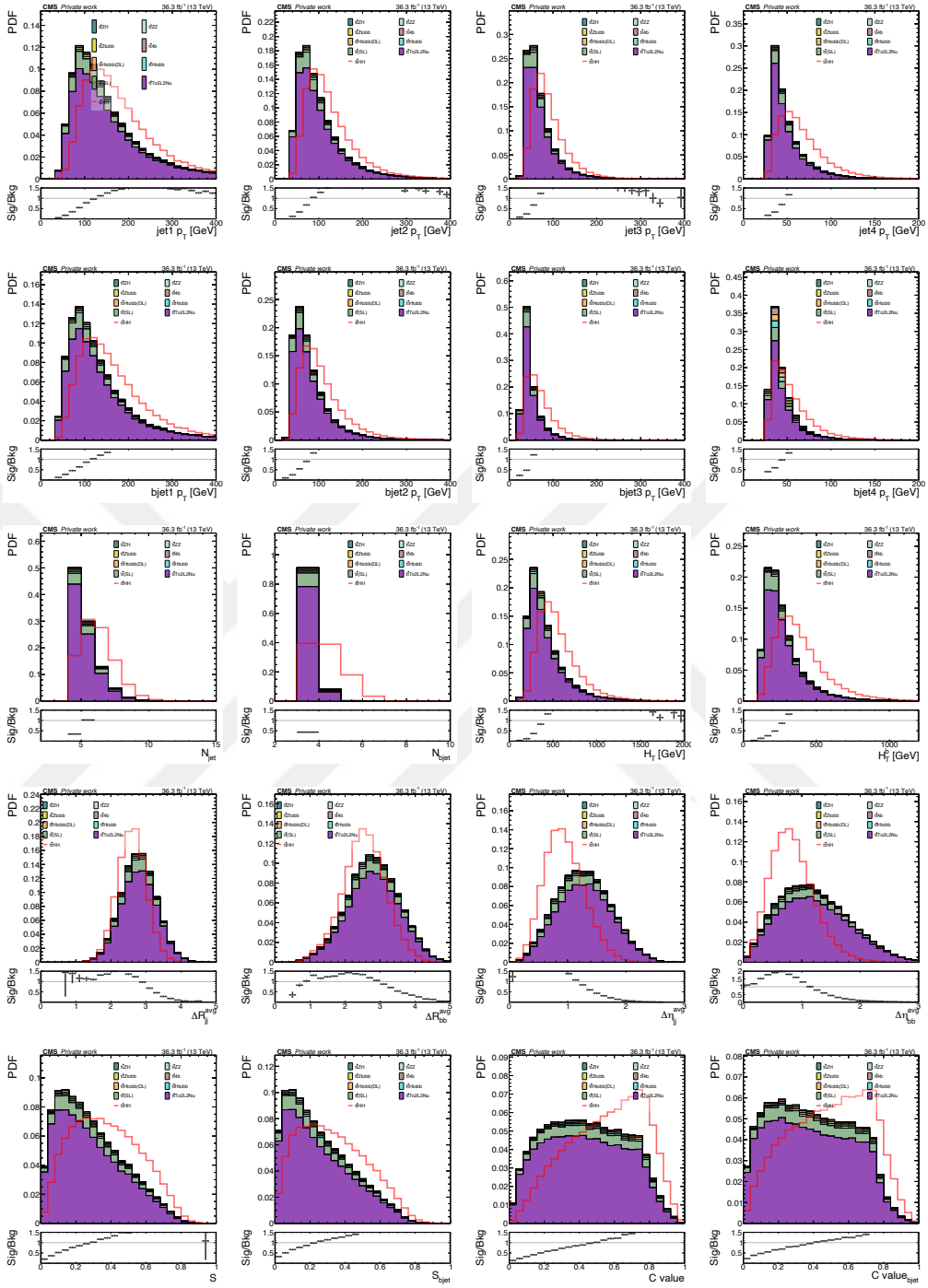

Figure 7.11: Distributions of different discriminating variables for jets and b-tagged jets for the 2016 samples. The baseline selection is applied to the $t\bar{t}HH$ signal and the backgrounds, normalized to 36.3 fb^{-1} luminosity.

7.6 Graph Attention based jet assignment - GATJA

The total cross-section of the $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$ signal is substantially reduced after multiplying it by the decaying branching ratios of the Higgs boson into b jets and the W boson into dilepton. Additionally, performance of the detector efficiency and acceptance significantly affects the observed event yield. Moreover, the predicted number of background events overhelmingly surpasses that of $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$ events. One of the predominant sources of the background is $\mathrm{t\bar{t}b\bar{b}}$ system, generated in conjunction with two additional b jets, and the other is the $\mathrm{t\bar{t}jj}$ process, wherein at least one of the two jets produced alongside the $\mathrm{t\bar{t}}$ system is incorrectly identified as a b jet. Hence these two processes have a comparable final state signature to the signal process $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$. An additional complex background challenge is posed by the $\mathrm{t\bar{t}ZZ(b\bar{b}b\bar{b})}$ and $\mathrm{t\bar{t}ZH(b\bar{b}b\bar{b})}$ samples. Analogous to the HH system in the signal sample, ZZ and ZH diboson systems decay into 4 b jets, yielding a nearly indistinguishable final state. This similarity results in the observation of almost mimical kinematic distributions across these three samples. Thus, the precise assignment of b jets in the final states to their original mother particle becomes critical for extracting clear signal information. As introduced in Section 7.5, the distinguishing factor between the $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$ process and the irreducible background processes lies in the spectrum of the invariant mass of the b jets (m_{bb}). It has a peak shape in the case of $\mathrm{t\bar{t}HH(b\bar{b}b\bar{b})}$, corresponding to the mass spectrum of the Higgs boson at the generator level. Consequently, this distribution serves as a powerful discriminator. However, accurately identifying the two b jet pairs originating from the Higgs bosons decay, and thereby constructing the m_{bb} variable, is very challenging due to the large number of jets present in the event. A minimum of six jets is expected, consisting of two from the $\mathrm{t\bar{t}}$ system and four from the decay of the Higgs boson.

A novel object-based assignment algorithm, namely Graph Attention Based Jet Assignment (GATJA) is introduced for the purpose of assigning b-tagged jets to specific particles, including Higgs bosons, top quarks, and other relevant quantities that are present in the related decay chain. This allocation or reconstruction task cannot be accomplished by considering single objects in isolation, particularly b-tagged jets within the context of this analysis. Instead, one needs to consider two or more

objects along with a connecting variable among them. As such, this algorithm efficiently groups physics objects, guided by a physics-motivated distance parameter, which provides the necessary information for the reconstruction of heavier particles.

The algorithm's approach begins with a selected b-tagged jet and its neighboring objects, determined by the chosen distance metric, such as the inverse χ^2 of the Higgs, which effectively reduce combinatorial background. In the analysis, all b jets within the event are strategically employed, as each b jet is paired with another in a manner that minimizes the χ^2. This process is performed for every b-tagged jet, regardless of its origin from the Higgs boson or not.

Figure 7.12 shows the invariant mass distribution of the Higgs boson reconstructed by using a b-tagged jet pair. A pair matching to a Higgs boson (green) and another one not matching (blue) indicate that partial mass reconstruction is still possible. Importantly, this observation challenges the idea that the second matched b-tagged jet should necessarily be the closest neighbor to the first. In fact, it could also be the second closest neighbor. Even though the system is still functional with just two b-tagged jets, using three ensures to account for these complexities and to offer a more comprehensive analysis. This not only aligns with the minimum number of b-tagged jets required for the baseline selection defined in Section 7.4 but also serves as a fail-safe mechanism.

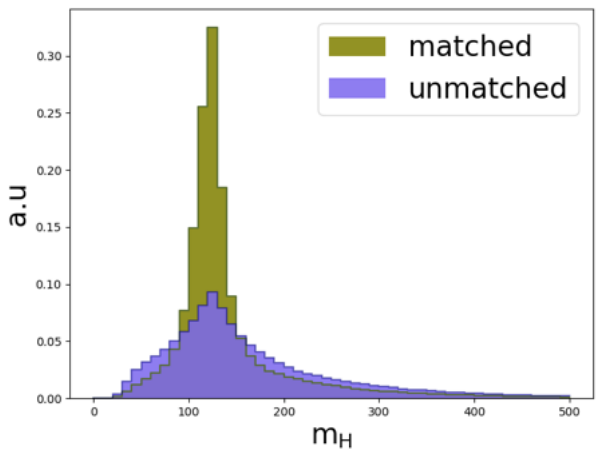

Figure 7.12: Higgs boson mass distribution showing the b jets matched or unmatched to the Higgs boson.

Figure 7.13 illustrates the overall working mechanism of GATJA. The four momentum variables and b tagging discriminant output associated with the b-tagged jet in question and its neighboring b-tagged jets serve as inputs to the self-attention layer. The outputs of this mechanism, namely the encoded parameters, are fed into a simple DNN with skip connections. Another set of inputs is depicted in Figure 7.13 as 'event variables,' encompassing general variables such as the four momentum of leptons, which is crucial for reconstructing the top quark, MET, b jet multiplicity, and H_T. These variables constitute additional inputs for the DNN. In the final step, the network gives three scores, which show the probability of the b-tagged jet in question originating from H boson, top quark, or any other particles.

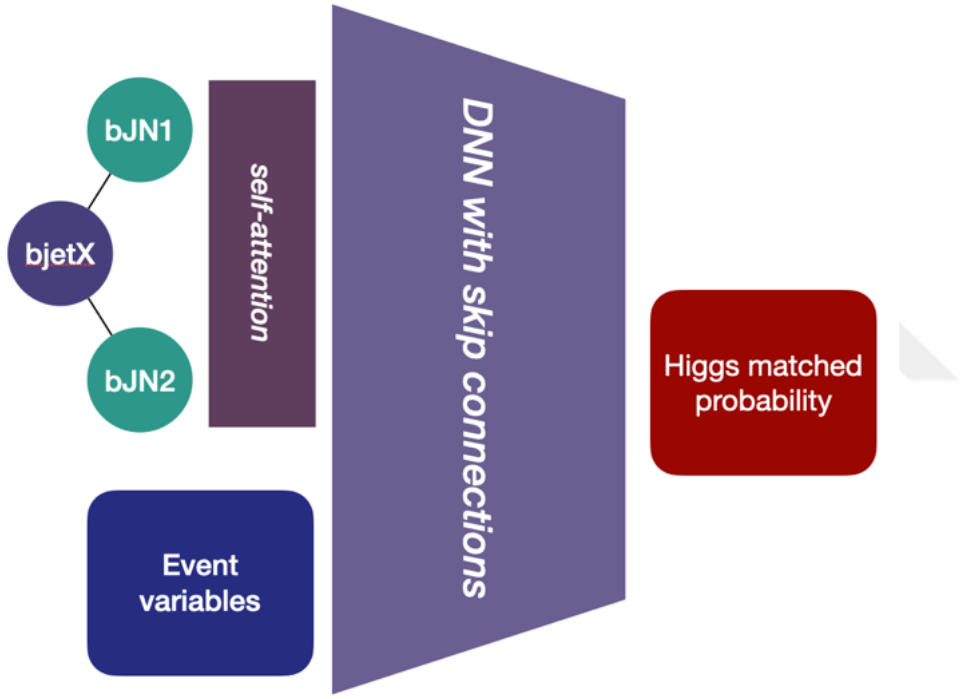

Figure 7.13: GATJA flowchart.

As shown in Figure 7.14 in detail, the attention mechanism includes some dense layers to create the encoders, from encoders, the weight matrix is created by using the softmax activation function. Then the matrix is multiplied by the first b-tagged jet that is interested in to obtain the output, i.e. the higher representation. This process can be repeated for the all available b-tagged jets but the network sees only one at each instance.

Instead of employing conventional methods like scaling with the adjacency matrix, the utilization of Graph Attention Networks by accomplishing the construction of particle interaction graphs shows a significant potential for the improvement of the current analysis.

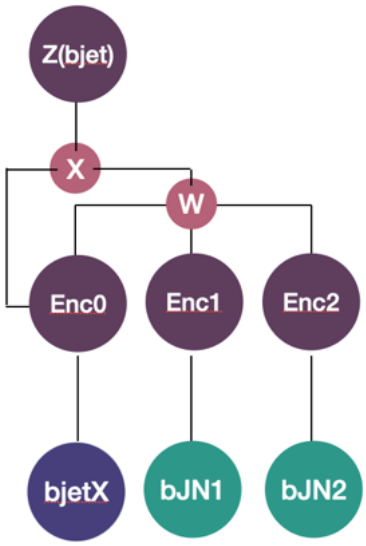

Figure 7.14: A simplified view of the attention mechanism used for each b-tagged jet.

The training of GATJA applies to any sample characterized by the $\mathrm{t} + * \rightarrow \mathrm{b}+$ or $\mathrm{h}+ \rightarrow \mathrm{b} + *$ topology. For this specific analysis, the $\mathrm{t\bar{t}HH}$ signal and all the background samples listed in Section 7.2 are utilized, excluding $\mathrm{t\bar{t}ZH}$. To handle the complexity of this diverse dataset, a multiclass classifier with three distinct outputs is employed: "Higgs boson", "top quark", and "others". These outputs provide the matched probability of a b-tagged jet to a mother particle, as indicated in Figure 7.13. Currently, the focus is on the top quark and Higgs boson, but the method's adaptability allows for potential expansion to include the Z boson as well.

Some technical details such as model hyperparameters and input variables for both the attention layer and the DNN training are provided in Figure 7.15.

Figures 7.16 and 7.17 show the GATJA scores of three nodes (Higgs, top and others) with a clear separation of matched (red) and unmatched (green) b-tagged jets to the corresponding particles for the years 2017 and 2018, respectively. Figure 7.18 shows the distributions of several GATJA variables serving as inputs to the DNN model.

Characteristic	*Description*
Outputs	Higgs matched, top matched, or others
Training	500 epochs with early stopping
Optimization	Large batch optimizer with LR decay
Model Size	5M parameters; trains in 30 mins (30 variables)

For the bjet in question and neighbor bjets	
bjet Inputs	jet 4-vector + btag discriminant output

Event Inputs	
Variables	'jetAverageMass', 'bJetAverageMass', 'bJetAverageMassSqr', 'jetHT', 'bjetHT', 'lightjetHT', 'jetNumber', 'bjetNumber', 'leptonPT1', 'leptonEta1', 'leptonPT2', 'leptonEta2'...

Figure 7.15: Hyperparameters and input varibles employed during the GATJA training process are listed.

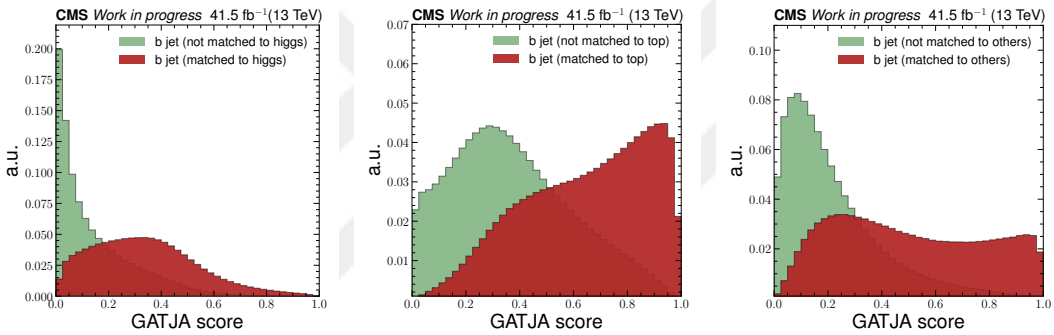

Figure 7.16: Higgs boson node (top left), top quark node (top right) and other objects node (bottom) of the GATJA output for the 2017 case. It shows the distribution to Higgs boson score given by the output of GATJA

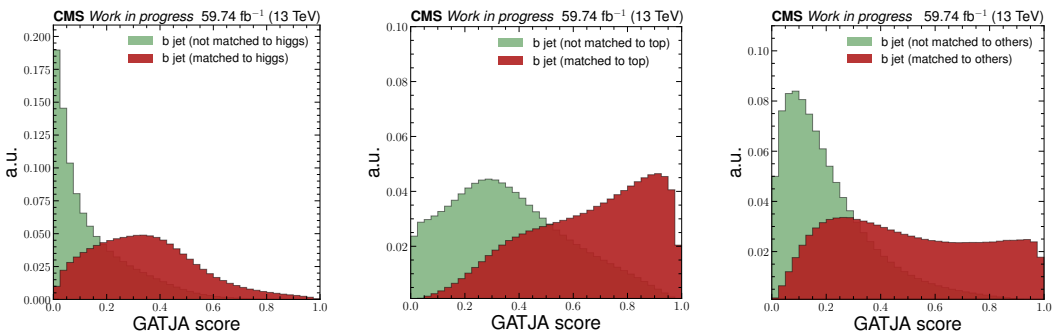

Figure 7.17: Higgs boson node of the GATJA output for the 2018 case. It shows the distribution to Higgs boson score given by the output of GATJA

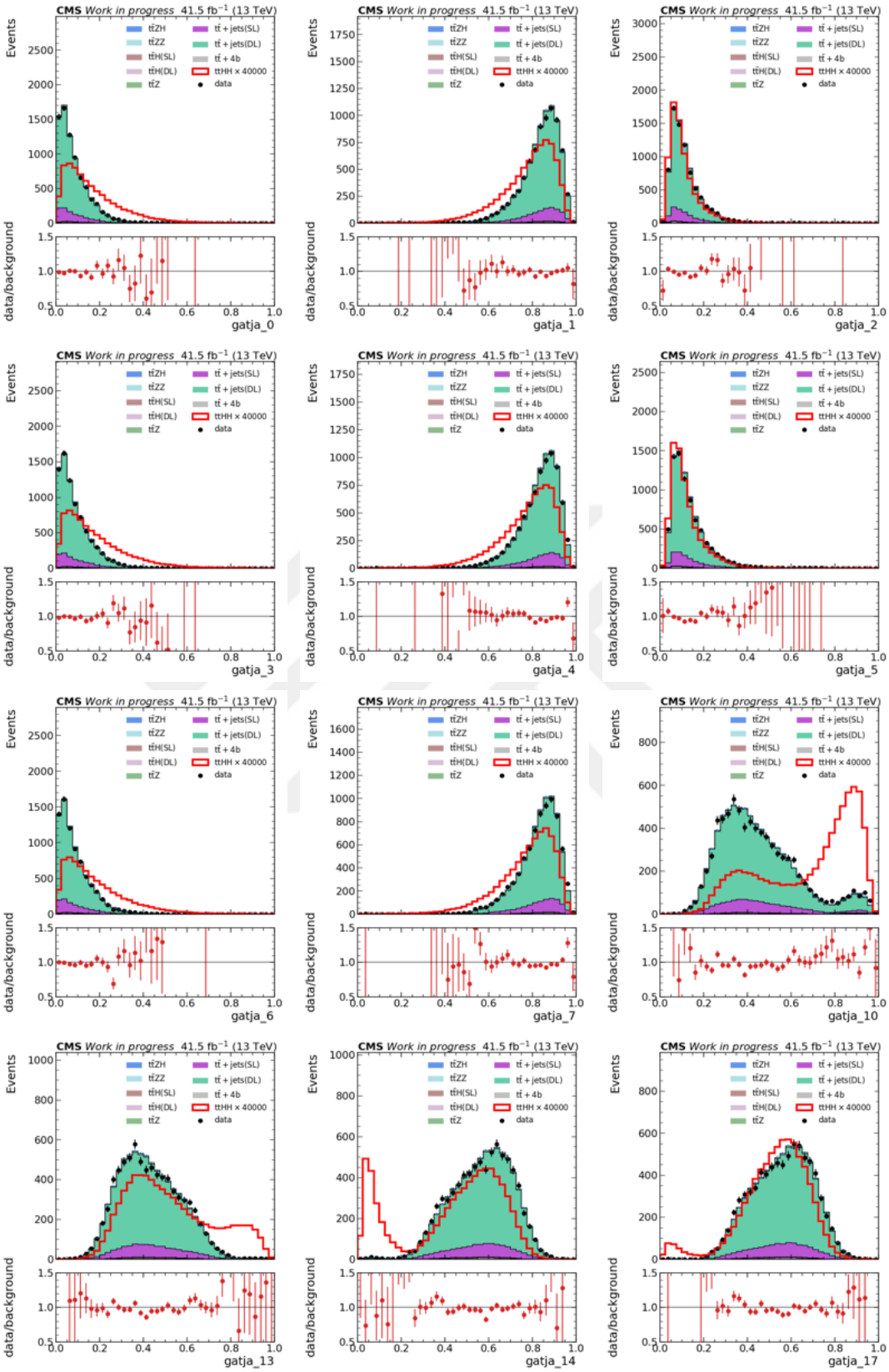

Figure 7.18: Distributions of GATJA outputs obtained for 8 b-tagged jets. They serve as inputs to the DNN model. Data comparison is also presented.

7.7 Data-Monte Carlo Comparison in a Control Region

The accurate representation of observed data by the MC simulated samples is crucial, as the final measurement considerably depends on the simulated samples mentioned in the previous subsections. To verify the effectiveness of the MC simulation, one can compare it to data within a specific control region defined by:

$$n_{jet} \geq 4 \;\&\; n_{bjet} = 2 \;\&\; n_{lep} = 2.$$

This control region is chosen as the physics involved is already well-understood. Such criteria ensure that the control region does not overlap with the signal region, which is identified by different parameters

$$n_{jet} \geq 4 \;\&\; n_{bjet} \geq 3 \;\&\; n_{lep} = 2.$$

Several event variables are compared for the data and the background MC samples used in this analysis. Overall, the sum of these MC samples are in a good agreement with the the full Run 2 dataset, corresponding to an integrated luminosity of 137.60 fb^{-1} data. Figures 7.19 to 7.21 depict these comparisons across various event variable distributions for the 2017 run period.

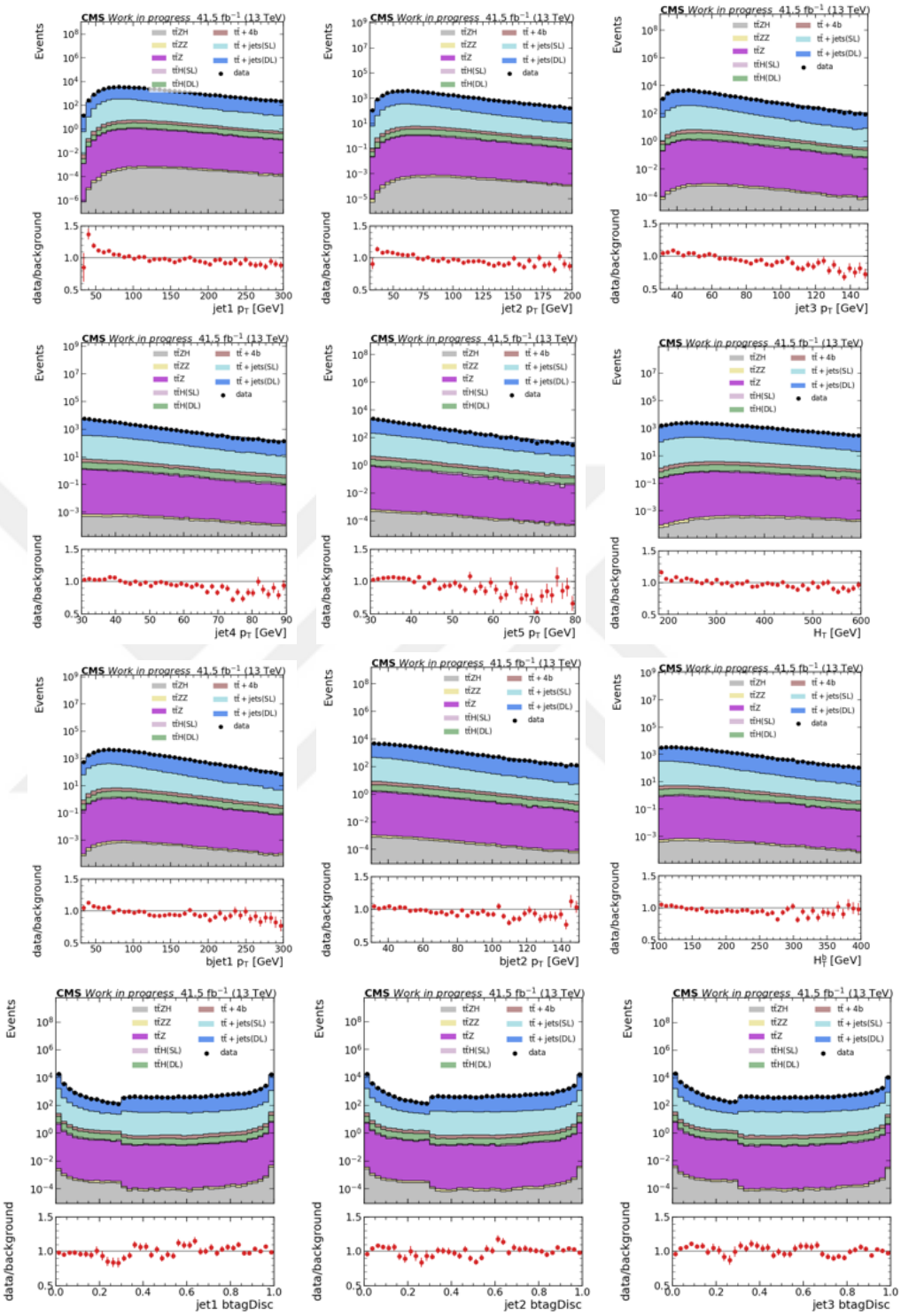

Figure 7.19: Distributions of the Data-MC comparison for the jet and b-tagged jet related variables for the 2017 samples.

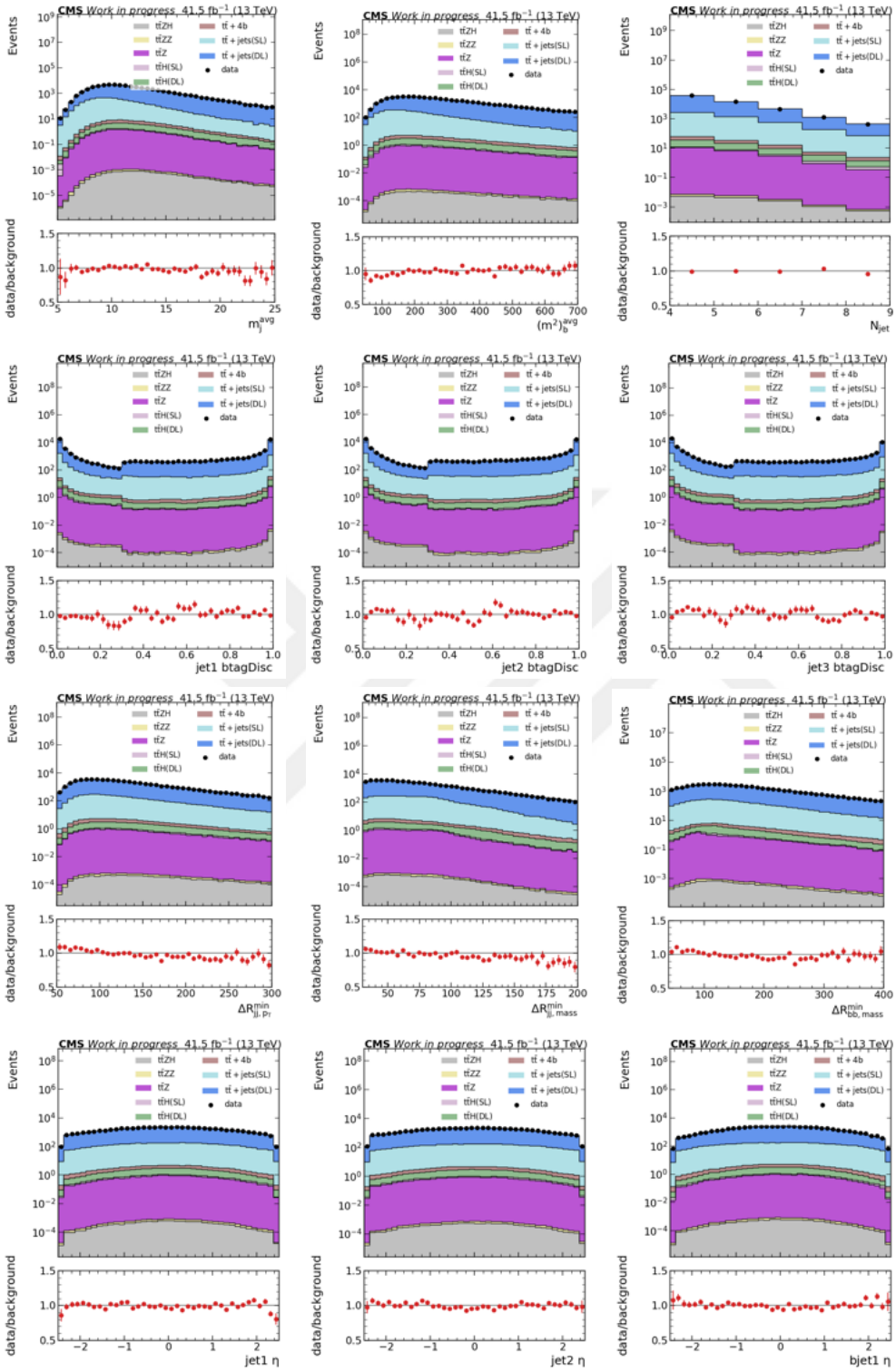

Figure 7.20: Distributions of the Data-MC comparison for the jet and b-tagged jet related variables for the 2017 samples.

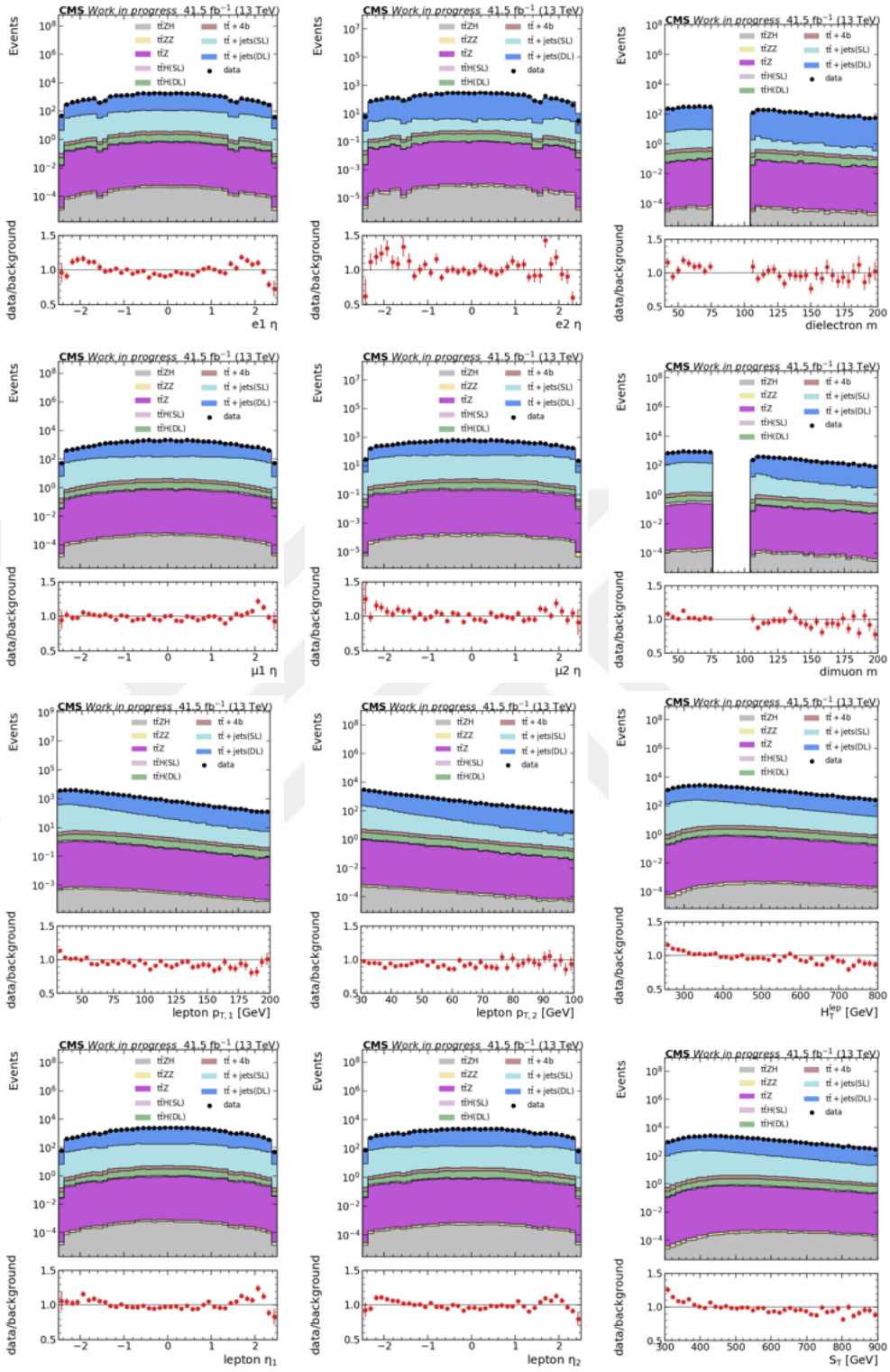

Figure 7.21: Distributions of the Data-MC comparison for the lepton related variables for the 2017 samples.

7.8 Event Categorization

The strategy followed in this analysis is based on Deep Neural Network (DNN) tools allowing an optimized event classification and analysis sensitivity. In the context of this analysis, a DNN is essentially a multi-layer perceptron (MLP) with a large number of hidden layers. DNN is employed as a classifier to distinguish between the HHto4b sample and other SM processes with similar final state signatures. The background samples used for the classification include (combined (SL) and (DL) samples), Htobb$\bar{\text{b}}$, Ztobb$\bar{\text{b}}$, ZHto4b and ZZto4b. Instead of providing a single binary-classification discriminant, the DNN model produces a separate discriminator for each process contributing to the background. Predictions of the multi-classifier model for a given event are a single value for each defined process (signal or background) ranging from 0 to 1, where 1 indicates likeliness of the event to the particular process. Sum of all these predictions are normalized to 1, therefore, process corresponding to maximum value of multi-classifier output indicates prediction of the most likely process for a given event. The overall flowchart representing the multiclass classification structure is shown in Figure 7.22. Chapter 5 provides a detailed discussion of various machine learning concepts and terms referenced in the following text.

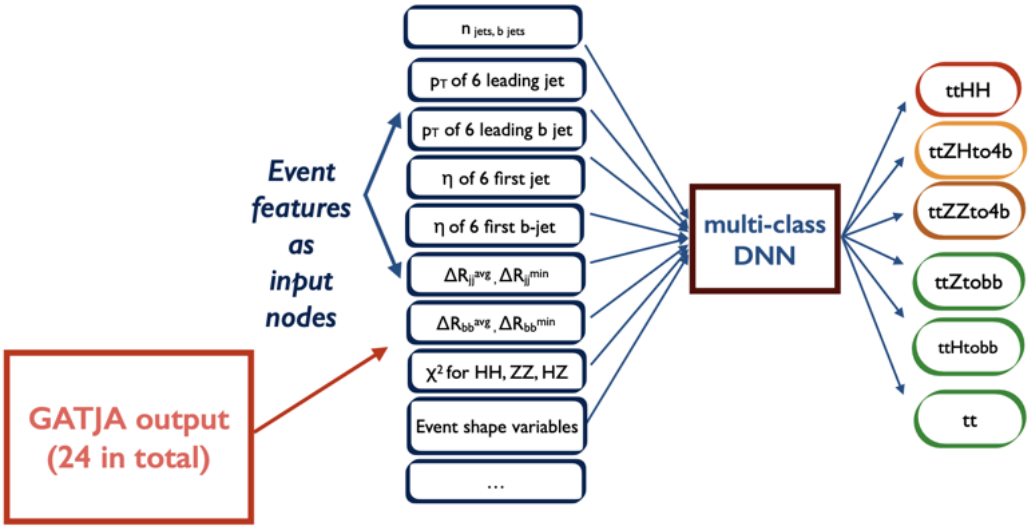

Figure 7.22: DNN workflow in the DL channel.

The multi-classifier DNN network used in the present analysis consists of fully connected neural network nodes for the consecutive layers. To prevent a bias towards

prediction performance of the training samples (over-training), independent validation and training samples are identified. Possible over-training of the network is regularized with ́dropoutĺayers, where some percentage of nodes that are selected randomly are skipped in the training pass-through. The training of the DNN is performed using all the variables listed in Table 7.27 as well as the 24 variables obtained from the GATJA outputs. Values of input nodes spanning different ranges are known to degrade the DNN classification performance. The training input features are preprocessed to be uniformly distributed over quantiles, and scaled to be ranged between 0 to 1. The obtained scale factors are applied to the evaluation samples. Feature importance study is performed to learn the variables giving higher contributions during the DNN training. Top 20 ranked event variables are listed in Figure 7.23. This list mostly include the b-tagged jet related variables as well ass GATJA outputs.

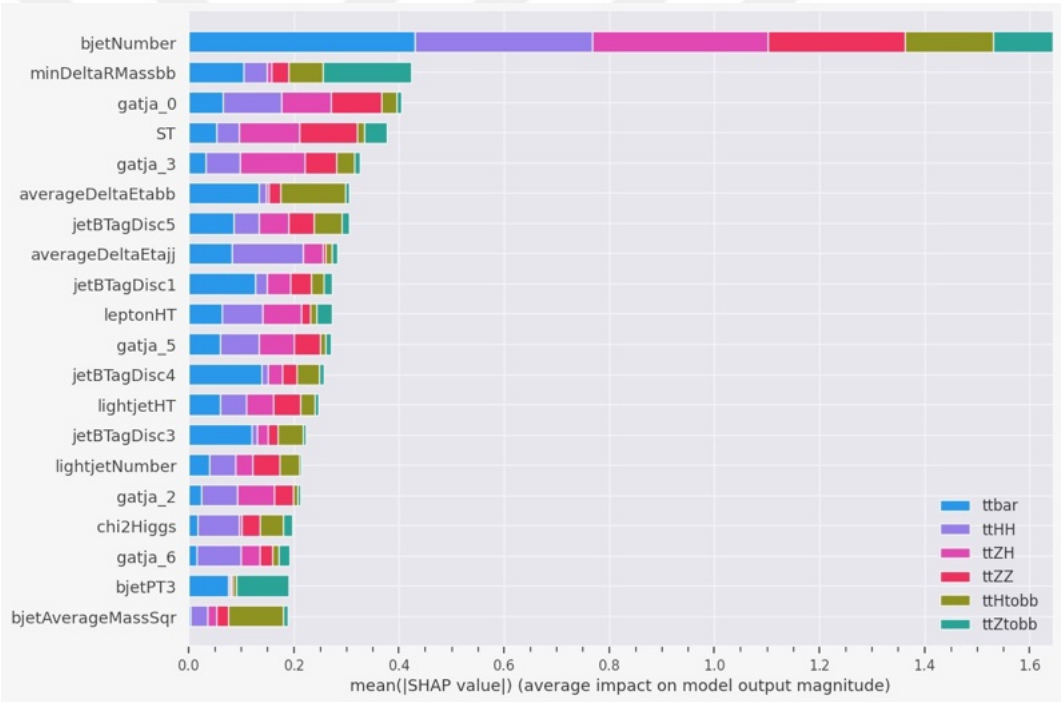

Figure 7.23: Top 20 ranked event features contributing to the DNN training are listed.

The hyperparameters of the DNN model significantly affects its performance. In the training and the validation samples various hyperparameters described in Table 8.5 are optimized. The network comprises four hidden layers with respective node configurations of 512, 256, 128, and 64. It is found that 20% dropout rate yields to a very small over-training (estimated by comparing the evaluation results from the training

and the validation samples). The batch size of 1024 is used, and the optimization is carried out using the Large Batch Optimization (LAMB) algorithm. The learning rate is dynamically adjusted using a cosine decay. The chosen activation function for the hidden layers is Rectified Linear Unit (ReLU), while the final layer employs a Softmax activation function. The categorical crossentropy loss function is employed to measure the dissimilarity between predicted and actual class distributions. Furthermore, a validation split of 20% is reserved to assess the model's generalization performance. These hyperparameter settings collectively define the configuration of the DNN model, aiming to achieve optimal predictive accuracy for the specified baseline criteria. To obtain these optimal settings a grid search on the number of nodes and hidden layers, and activation functions is performed. The setting providing the largest area under the curve (AUC) of receiver operating curve (ROC) is selected.

Table 7.28: List of the main DNN hyperparameters values for the baseline selection case requiring: ≥ 4 jets, ≥ 3 b-jets.

Hyperparameters	Selection
hidden layers	4
nodes per hidden layer	512, 256, 128, 64
dropout	0.2
batch size	1024
loss function	Categorical Crossentropy
optimizer	LAMB
learning rate	Cosine decay
activation function	ReLU
last activation	Softmax
validation split	0.2

The Figure 7.24 displays the confusion matrices for the 2017 samples, comparing scenarios with (right) and without (left) the utilization of GATJA outputs as inputs to the multiclass classifier DNN. In this case, node does not include 4b sample. On the other hand, in Figure 7.25, a confusion matrix is shown with a node including 4b sample. Figure 7.26 shows the same comparison as for Figure 7.24 with the 2018 samples.

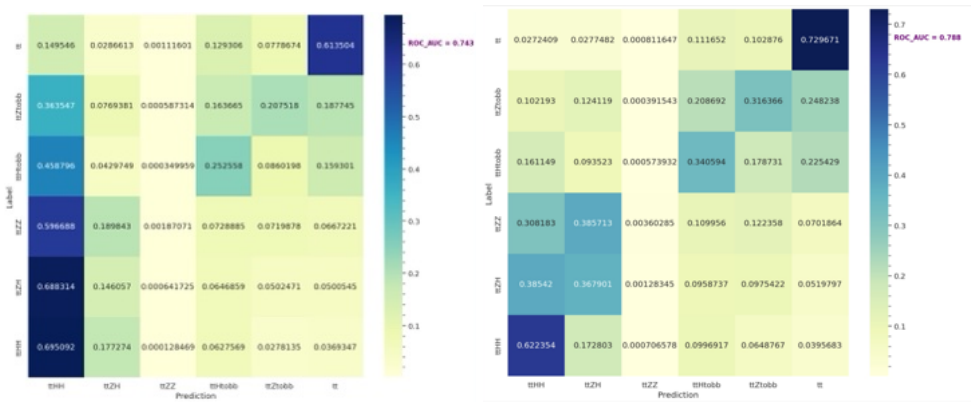

Figure 7.24: Confusion matrices showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2017 results. $t\bar{t}$ node includes $t\bar{t}SL$ and $t\bar{t}DL$ backgrounds.

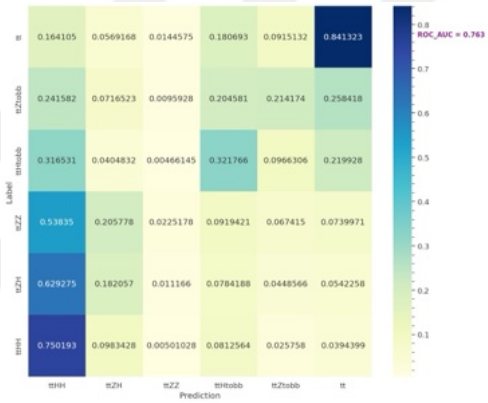

Figure 7.25: Confusion matrix showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2017 results. $t\bar{t}$ node includes $t\bar{t}SL$, $t\bar{t}DL$, and $t\bar{t}4b$ backgrounds.

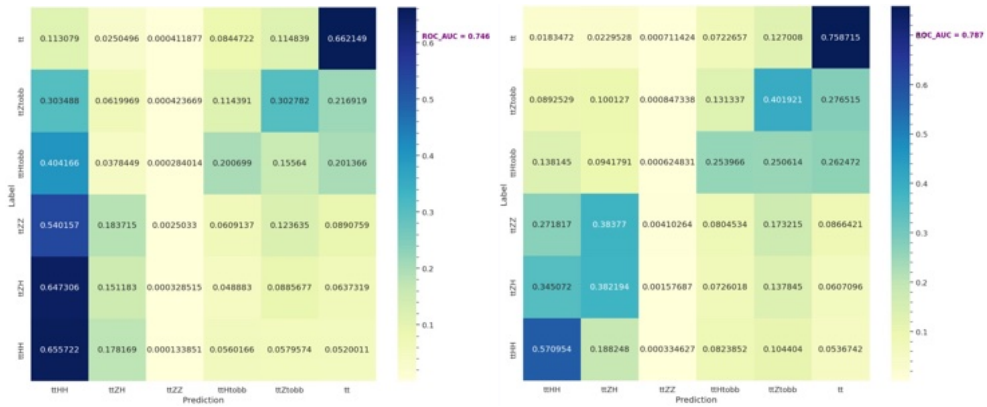

Figure 7.26: Confusion matrices showing the separation efficiency of the $t\bar{t}HH$ signal from the backgrounds in the DL channel for 2018 results.

CHAPTER 8

STRATEGY FOR THE $\mathrm{t\bar{t}HH}$ SEARCH IN THE SEMILEPTONIC DECAY CHANNEL AT THE HL-LHC

This chapter, with the support of the motivation detailed in Chapter 6, describes a search for the production of a top quark-antiquark pair associated to a pair of Higgs bosons at the HL-LHC running at a 14 TeV center-of-mass energy and with a total integrated luminosity of 3000 fb^{-1} within the framework of both the SM and the MCHM. The SM production cross section computed at NLO QCD for the $\mathrm{t\bar{t}HH}$ process is $0.948^{+1.7\%}_{-4.5\%}$ $\pm 3.1\%$ fb [13]. The cross section values for the $\mathrm{MCHM_5}$ and $\mathrm{MCHM_{14}}$ benchmark points of the $\mathrm{t\bar{t}HH}$ process are $1.47^{+1.7\%}_{-4.5\%} \pm 3.1\%$ (LO) [12] and $2.15^{+1.7\%}_{-4.5\%} \pm 3.1\%$, respectively. The production rate of HH boson events within the SM is notably low, exceeding one thousand times weaker than the rate of single H boson production. Consequently, during the forthcoming LHC operations with an $\mathcal{L}$ of a few hundred fb^1, it is anticipated that experimental outcomes will likely yield only evidence of HH production. However, the prospect of observing HH production is expected to be significantly enhanced by the enormous datasets collected with the HL-LHC. To obtain the HL-LHC conditions at best, this work is achieved within the Delphes simulation framework and reproducing the main features of the upgraded CMS detector for the HL-LHC.

This study is performed as a contribution to the Snowmass 2021 [162] carried by the the Snowmass Community. Studies of this community include a series of community-driven workshops and conferences organized by the American Physical Society (APS) and the Division of Particles and Fields (DPF) to address key questions and challenges in the field.

8.1 Analysis Method

In order to optimize both the signal extraction while retaining as much as possible of the produced events, this analysis considers the semi-leptonic (lepton + jets) decay of the top quark-antiquark pair and the decay of the double Higgs bosons into two bottom quark-antiquark pairs. In the semi-leptonic (SL) channel, one of the W boson decays to an electron or a muon and the corresponding neutrino, while the other W boson decays into two quarks. When the two Higgs bosons decay into bottom quarks, this produces a final state ideally, at leading order, with either an electron or a muon, moderate missing transverse energy and eight relatively high p_T jets, at least six of which are b quark jets as shown in Figure 8.1. However, due to several constraints related to detector acceptance, possible merging of jets, the b tagging efficiency, and the relatively low signal production rate as compared to the high rate competing physics backgrounds, a rather loose baseline selection is applied in order to optimize the overall analysis sensitivity.

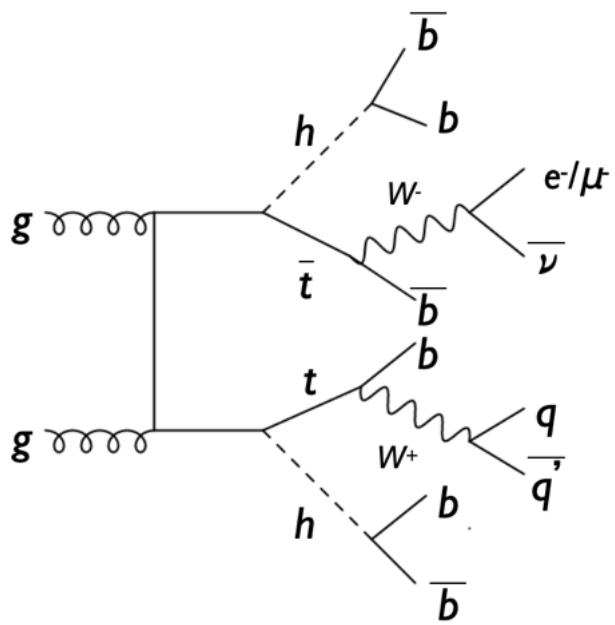

Figure 8.1: Feynman diagram of the $t\bar{t}HH$ production process in the single lepton channel.

8.2 Simulated Samples for Phase-2 Study

This research employs Monte Carlo (MC) simulation program to evaluate events. The SM $\mathrm{t\bar{t}HH}$ signal events are generated utilizing the MADGRAPH5_AMC@NLO generator at LO. Additionally, events corresponding to two MCHM benchmark points, namely MCHM_5^{C2} and MCHM_{14}^{D7}, explained in Chapter 6, are generated at LO using the same generator. The cross section calculations for these benchmarks are performed at LO, excluding uncertainties [12]. Since the MCHM does not introduce changes to the LO QCD effects, the QCD and PDF uncertainties remain consistent with the SM predictions. A significant challenge in $\mathrm{t\bar{t}H}$measurement is the production of $\mathrm{t\bar{t}}$ alongside two b jets, especially when the Higgs bosons decay into a b-quark pair. A comprehensive examination of this process at NLO+PS QCD order [163] is crucial for accurately evaluating this background in the $\mathrm{t\bar{t}}$ analysis. Furthermore, the generation of $\mathrm{t\bar{t}}$ with four b jets serves as a notable background for $\mathrm{t\bar{t}HH}$ [164]. While $\mathrm{t\bar{t}} + 2\mathrm{b}$ has an associated NLO QCD study, $\mathrm{t\bar{t}} + 4\mathrm{b}$ lacks one, with only a LO sample available. Using insights from Ref. [13] for $\mathrm{t\bar{t}} + 2\mathrm{b}$, which demonstrates that NLO QCD effects reduce scale uncertainties significantly, two distinct QCD scale uncertainties, 30% and 70%, are considered for $\mathrm{t\bar{t}} + 4\mathrm{b}$. The PDF scale uncertainty is kept consistent with $\mathrm{t\bar{t}} + 2\mathrm{b}$, specifically at +2.9% -3.5%, as mentioned in Ref. [79]. Notably, $\mathrm{t\bar{t}} + 2\mathrm{b}$ also adds to the background, and while specialized samples for this process remain unavailable, it's partially incorporated via the extensive $\mathrm{t\bar{t}}$+jets samples.

The MADGRAPH5_AMC@NLO generator is also utilized to create background processes for $\mathrm{t\bar{t}} + 4\mathrm{b}$, $\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}ZH}$, and $\mathrm{t\bar{t}ZZ}$. Moreover, background events for $\mathrm{t\bar{t}}$ and $\mathrm{t\bar{t}H}$ are produced using POWHEG at NLO. Both $\mathrm{t\bar{t}} + 4\mathrm{b}$ produced at LO and $\mathrm{t\bar{t}}$ produced at NLO use the 5FS.

Regarding the methods, all processes consider the proton's structure as described by the parton distribution functions (PDF) NNPDF3.1 [165]. The generated parton-level events undergo parton showering and hadronization via PYTHIA8.2, adhering to parameters from the CP5 tune. The simulations integrate an average pileup of 200 interactions for each event. Subsequently, the simulated signal and background events undergo processing using the DELPHES fast simulation package [82], mirroring the

anticipated response of the upgraded CMS detector in HL-LHC conditions. The efficiencies linked to object reconstruction and identification, paired with the detector's response and resolution, are parameterized grounded in the comprehensive simulation of the improved CMS detector using the GEANT4 package [83, 84].

Table 8.1 displays the simulated physics samples used in this study, with all the relevant cross sections, QCD scale, and PDF uncertainties as described above. All samples are inclusive in $\mathrm{t\bar{t}}$. The samples involving Higgs and/or Z boson production are simulated to include only the decays of Higgs or Z boson into a pair of b quarks, with the exception of the $\mathrm{t\bar{t}}Z$ sample, which is inclusive in both $\mathrm{t\bar{t}}$ and Z. Each generated event sample is normalized to the highest order cross sections available.

Table 8.1: List of MC simulated samples for this study, and event yields after baseline selection normalized to 3000 fb^{-1}.

Simulated samples	Cross sections at 14 TeV (fb) ±QCD Scale (%) ±(PDF+α_s) (%)	Event yields after baseline selection (3000 fb^{-1})
SM : $\mathrm{t\bar{t}HH} \to \mathrm{b\bar{b}b\bar{b}}$	$0.948^{+1.7\%}_{-4.5\%} \pm 3.1\%$ (NLO) [13]	308
$\mathrm{t\bar{t}}$ + jets	$984500^{+23.21\%}_{-34.69\%} \pm 4\%$ (NNLO+NNLL) [22]	21707536
$\mathrm{t\bar{t}}$ + 4b	$370^{+30.0\%}_{-30.\%} \pm 3.5\%$ (LO) [164, 166]	88977
$\mathrm{t\bar{t}H}$, $\mathrm{H} \to \mathrm{b\bar{b}}$	$612^{+6.0\%}_{-9.2\%} \pm 3.5\%$ (NLO) [13]	133960
$\mathrm{t\bar{t}Z}$	$1018^{+9.6\%}_{-11.2\%} \pm 3.5\%$ (NLO) [13]	73999
$\mathrm{t\bar{t}ZZ}$, $\mathrm{ZZ} \to \mathrm{b\bar{b}b\bar{b}}$	$2.59^{+4.3\%}_{-8.7\%} \pm 1.8\%$ (NLO) [143]	727
$\mathrm{t\bar{t}ZH}$, $\mathrm{ZH} \to \mathrm{b\bar{b}b\bar{b}}$	$1.54^{+32.2\%}_{-22.6\%} \pm 2.8\%$ (LO) [13, 166]	537
$\mathrm{MCHM_5}$: $\mathrm{t\bar{t}HH} \to \mathrm{b\bar{b}b\bar{b}}$	$1.47^{+1.7\%}_{-4.5\%} \pm 3.1\%$ (LO) [12]	377
$\mathrm{MCHM_{14}}$: $\mathrm{t\bar{t}HH} \to \mathrm{b\bar{b}b\bar{b}}$	$2.15^{+1.7\%}_{-4.5\%} \pm 3.1\%$ (LO) [12]	491

8.3 Objects used in the Phase-2 HL-LHC analysis in the single lepton channel

The objects used in this specific analysis are reconstructed starting with a dedicated PF algorithm in DELPHES. As described in Chapter 4, PF algorithm by definition correlates the basic elements from all detector layers (tracks and clusters) to identify each final-state particle, and combines the corresponding measurements to reconstruct the particle properties on the basis of this identification. Pileup subtraction is

applied at the PF candidate level via the so-called PUPPI algorithm [167], that was specifically tuned to reduce the pileup dependence on jets and missing transverse energy at 200 average pileup events. Object identification criteria are designed in the CMS GEANT4-based full simulation framework, taking into account the Phase-2 conditions. Parametrized object efficiencies and misidentification rates for these identification criteria are derived and implemented in DELPHES. Table 8.2 summarizes the objects used in this analysis.

Table 8.2: Definition of all objects used in the analysis.

Object	ID	$\mathrm{p_T}$ (GeV)	$\lvert\eta\rvert$	Isolation (I_{rel}^{PF})
Electrons (select)	medium	> 30	< 3	< 0.3
Electrons (veto)	medium	15-30	< 3	< 0.3
Muons (select)	medium	> 30	< 2.8	< 0.3
Muons (veto)	medium	15-30	< 2.8	< 0.3
Jets	loose	> 30	< 3	–
b jets (medium)	medium	> 30	< 3	–
b jets (loose)	loose and not medium	> 30	< 3	–

8.3.1 Object and baseline event selection

Electrons are identified with a boosted decision tree-based algorithm. The medium working point is selected, which has an identification efficiency of 0.8 and a misidentification rate of 0.015. Muon identification is an extension of the Run 2 identification, but takes into account the Phase-2 muon detector capabilities. The medium working point is selected, which has an identification efficiency of 0.9 and a misidentification rate of 0.01. A relative isolation variable IPF is defined for the leptons by summing the p of all PF particles within a cone of size 0.3 and dividing the sum by the lepton $\mathrm{p_T}$. Electrons and muons are required to have relative isolation less than 0.3. Pseudorapidity requirements are $|\eta| < 2.8$ (3.0) for muons (electrons). For both electrons and muons, two versions are defined in disjoint $\mathrm{p_T}$ ranges of $p_T > 30$ GeV and 15 $< p_T <$ 30 GeV, to be used for defining event selection and veto criteria, respectively. Jets are reconstructed from the PF particles using the anti-kT algorithm with a distance parameter of 0.4. A loose identification is applied by imposing criteria on variables related to energy fractions and multiplicities of various PF candidate

types clustered in a jet, in order to distinguish physical jets from those arising from calorimetry noise. To prevent overlap between selected jets and leptons, jets found within a cone of $\Delta R = 0.4$ around any of the selected leptons are vetoed. Jets are selected to have $\mathrm{p_T}>$ 30 GeV, $|\eta| < 3$. Both the $\mathrm{p_T}$ and η coverage selection applied to the leptons and jets are consistent with the realistic trigger thresholds expected at the HL-LHC [168, 169]. Jets are identified to be consistent with originating from the hadronization of b quarks (i.e. b-tagged jet) using the DeepJet [170] algorithm. The b-tagging efficiency and light flavor jet mistagging rate are parameterized based on the performance of this algorithm in full simulation and applied to reconstructed jets in DELPHES based on the jet MC truth flavor. The loose and medium b-tagging working points are used with average tagging efficiencies of ≈ 0.7 and ≈ 0.85 and gluon and light flavor jet mistagging rates of ≈ 0.01 and ≈ 0.1, respectively. Loose b jets are explicitly required to fail the medium b-tagging criterion in order to avoid overlaps. Loose and medium b jets have different roles in the event selection. The missing transverse momentum vector $\vec{p}_T^{miss}$ is computed as the negative vector $\mathrm{p_T}$ sum of all PF candidates in the event, and its magnitude is denoted as E_T^{miss}. Details are provided in Table 8.3.

Table 8.3: Event selection and search channels.

	Baseline event selection			
number of jets	≥ 4			
number of medium b jets	≥ 3			
number of selected e or μ	$= 1$			
number of veto e or μ	$= 0$			
E_T^{miss} (GeV)	> 20			
	b jet multiplicity channels			
	$= 3b$		$= 4b$	≥ 5b
	0 loose b	≥ 1 loose b		
number of medium b jets	$= 3$	$= 3$	$= 4$	≥ 5
number of loose b jets	$= 0$	≥ 1	–	–

8.4 Event Variables for DNN

In a similar manner to the full Run 2 study, event variables are computed for all events meeting the baseline selection criteria to characterize the event's topology and kinematics. Most of the variables utilized in the Run-2 study are also considered here and

can be found in Table 8.4. All the variables listed in this table have been previously described in Section 7.5. In the Phase-2 study, a newly introduced variable, the transverse momenta of reconstructed Higgs bosons ($p_{T(H,1)}$ and $p_{T(H,2)}$), is employed to distinguish the MCHM signals.

Table 8.4: Event quantities calculated for all events passing the baseline selection, including object properties, invariant masses, angular variables. Jets and b jets are ordered according to decreasing $\mathrm{p_T}$. See text for further details.

Group	Variables
Object multiplicities	N_{jets}, N_{bjets}
Object 4-momenta (objects are ordered by decreasing $\mathrm{p_T}$)	$\mathrm{p_T}$ jet 1, 2, 3, 4, 5, 6 $\lvert\eta\rvert$ jet 1, 2, 3, 4, 5, 6 $\mathrm{p_T}$ bjet 1, 2, 3, 4, 5, 6 $\lvert\eta\rvert$ bjet 1, 2, 3, 4, 5, 6
Hadronic transverse momenta	H_T, H_T^b
Mass averages	m_j^{avg}, m_b^{avg}, $(m^2)_b^{avg}$
Angular separation variables	$\Delta\eta_{jj}^{avg}$, $\Delta\eta_{bb}^{avg}$, $\Delta\eta_{bb}^{max}$ ΔR_{jj}^{avg}, ΔR_{bb}^{avg}, ΔR_{jj}^{min}, ΔR_{bb}^{min}
Optimized χ^2 values	χ^2_{HH}, χ^2_{ZZ}, χ^2_{ZH}
Invariant masses	$m_{H,1}$, $m_{H,2}$, $m_{Z,1}$, $m_{Z,2}$, $m_{ZH,Z}$, $m_{ZH,H}$
Reconstructed Higgs momenta	$p_{T(H,1)}$, $p_{T(H,2)}$ (only for MCHM DNNs)
Event shape variables	aplanarity, centrality sphericity, C value, D value
b-tag value	btagValue jet 1, 2, 3, 4, 5, 6

Figure 8.2 shows the distributions of some of these variables with particularly high discriminative power for the signal and all SM backgrounds, where all distributions are normalized to 3000 fb^{-1}. The ratio plots in the bottom panels are computed by taking signal and total background yields to be normalized to 1, allowing a direct comparison between the signal and total background shapes. Figure 8.3 shows a slightly different set of variables, where the probability density functions are compared for the $\mathrm{t\bar{t}HH}$ signal, the $\mathrm{t\bar{t}ZH}$ and the $\mathrm{t\bar{t}ZZ}$ backgrounds. Despite all 3 processes having the same final state, differences in kinematics due to different Z and H boson mass

scales are visible in the distributions, providing a discrimination with respect to $\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}ZZ}$, respectively. Finally, Fig. 8.4 shows the comparison of probability density functions of several variables between the SM $\mathrm{t\bar{t}HH}$ process with the two BSM benchmark points MCHM_5^{C2} and MCHM_{14}^{D7} introduced in Chapter 6. Here, significant differences can be observed between the SM and the MCHM processes. The MCHM signals are characterized by higher hadronic transverse activity due to the existence of heavy resonances. These heavy resonances also translate into peaks in tails of the jet and b jet $\mathrm{p_T}$ distributions, in particular for the leading ones. Similar peak structures exist in distributions of total hadronic transverse momenta calculated using all jets or only b jets. These plots demonstrate a clear discrimination between a MCHM-like resonant process from the non-resonant SM. Similar comparison plots of the SM and BSM cases for the signal with several individual background categories are shown in Appendix C.

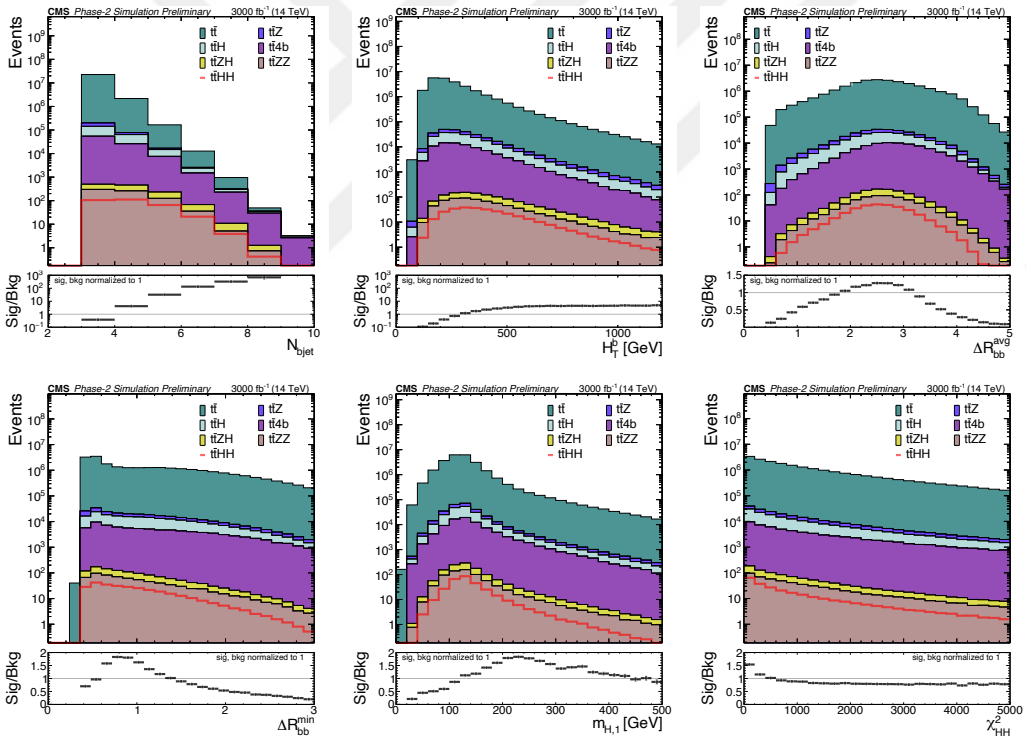

Figure 8.2: Distributions of different discriminating variables after the baseline selection applied. These compare the SM $\mathrm{t\bar{t}HH}$ signal and the SM backgrounds, normalized to 3000 fb^{-1} luminosity. Background distributions are stacked. The ratio plots in the bottom panels are computed by taking signal and total background yields to be normalized to 1.

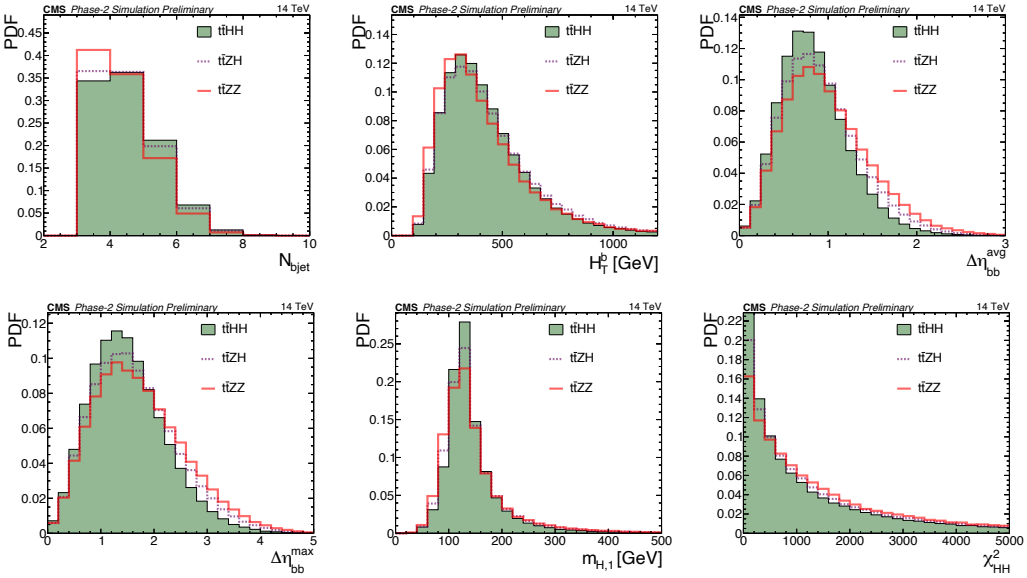

Figure 8.3: Probability density functions for different discriminating variables after the baseline selection applied. These compare the SM $\mathrm{t\bar{t}HH}$ signal shown as a filled histogram to two irreducible backgrounds : $\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}ZZ}$.

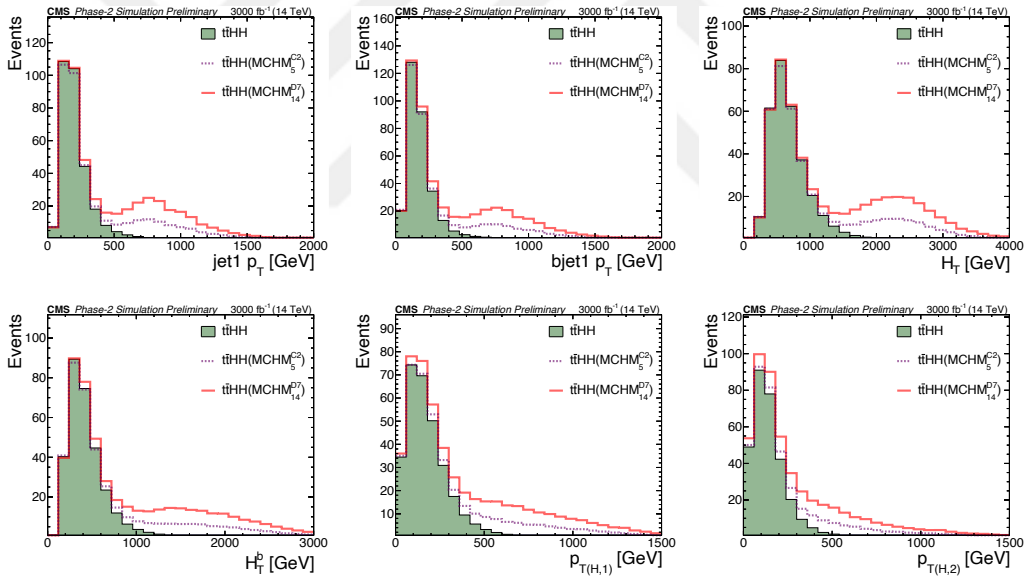

Figure 8.4: Distributions of different discriminating variables after the baseline selection applied. These compare the $\mathrm{MCHM}_5^{\mathrm{C2}}\mathrm{t\bar{t}HH}$ and the $\mathrm{MCHM}_{14}^{\mathrm{D7}}\mathrm{t\bar{t}HH}$ BSM benchmarks to the SM $\mathrm{t\bar{t}HH}$ signal, normalized to 3000 fb^{-1} luminosity.

8.5 Event Categorization

Similar to the the strategy followed in the Run 2 analysis, this analysis is also based on DNN tools allowing an optimized event classification and analysis sensitivity. In the

context of this analysis, a DNN is essentially a multi-layer perceptron (MLP) with a large number of hidden layers. DNN is employed as a classifier to distinguish between the signals and the background processes with similar final state signatures. Similar procedures are followed for the SM $\mathrm{t\bar{t}HH}$ signal and the MCHM signals. In the first step, a binary classifier is implemented using a subset of observables (40) providing highest signal over background ratio to obtain a discriminant separating the $\mathrm{t\bar{t}HH}$ or MCHM signals from $\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}ZZ}$ backgrounds. In the second step, all observables including the binary discriminant are input to a multi-classifier. In this step, instead of providing a single binary-classification discriminant, the DNN model produces a separate discriminator for each process contributing to the background. Predictions of the multi-classifier model for a given event are a single value for each defined process (signal or background) ranging from 0 to 1, where 1 indicates likeliness of the event to the particular process. Sum of all these predictions are normalized to 1, therefore, process corresponding to maximum value of multi-classifier output indicates prediction of the most likely process for a given event. The event categorization framework describing the steps explained above is shown in Figure 8.5.

The multi-classifier DNN network used in the present analysis consists of fully connected neural network nodes for the consecutive layers. To prevent a bias towards prediction performance of the training samples (over-training), independent validation and training samples are identified. Possible over-training of the network is regularized with ́dropout ́layers, where some percentage of nodes that are selected randomly are skipped in the training pass-through. It is found that 40% dropout rate yields to a very small over-training (estimated by comparing the evaluation results from the training and the validation samples).

The training of the DNN is performed using all the 58 variables listed in Table 8.4. Detailed explanation for these variables is provided in Section 7.5. Values of input nodes spanning different ranges are known to degrade the DNN classification performance. Therefore, the observables are pre-processed to be reweighted to a flat distribution and scaled to be in the range of [0, 1]. The obtained scale factors are applied to the evaluation samples. The hyperparameters of the DNN model significantly affects its performance. The hyperparameters of the DNN models described in Table 8.5 are optimized on the validation samples in order to maximize the largest

area under the curve (AUC [112]) of the receiver operating curve (ROC [112]) for the signal discriminants is selected. To obtain these optimal settings a grid search on the number of nodes and hidden layers, and activation functions is performed. Finally, samples are evaluated by the same DNN in three n_{bjet} categories as listed in Table 8.3.

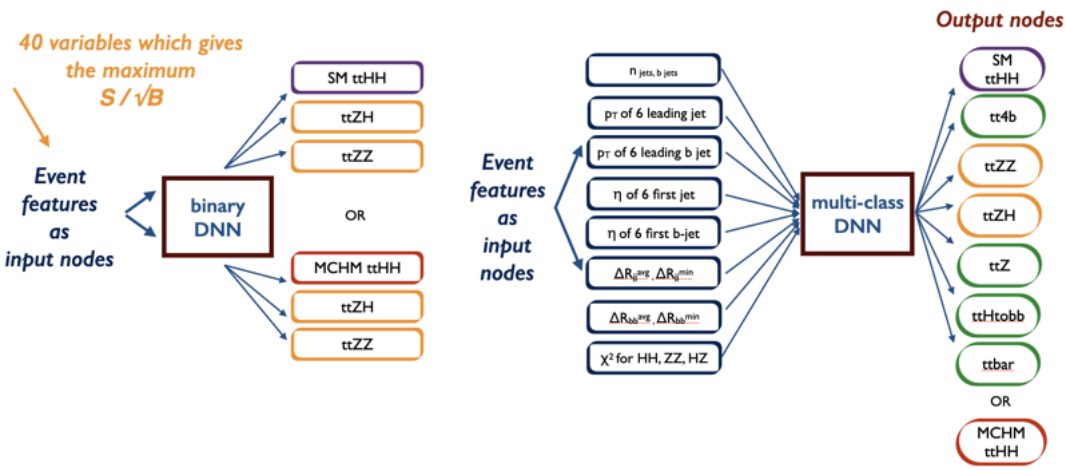

Figure 8.5: Schematic view showing the categorization of events using a DNN-based procedure

Table 8.5: List of the main DNN hyperparameters values for the baseline selection case requiring: ≥ 4 jets, ≥ 3 b jets.

Hyperparameters	Selection
hidden layers	4
nodes per hidden layer	256, 128, 64, 32
dropout	0.4
batch size	512
loss function	Categorical Crossentropy
optimizer	Adam
learning rate	0.001
activation function	LeakyReLU
last activation	Softmax
validation split	0.3

The final signal discriminants for the SM $t\bar{t}HH$ case are shown in Fig. 8.6. Figure 8.7 shows the final discriminant distributions for the $t\bar{t}HH$ MCHM$_5$ and the $t\bar{t}HH$ MCHM$_{14}$ benchmark points. The shapes of each distribution serve as input for the final fit, based on the Combine tool (see Appendix D).

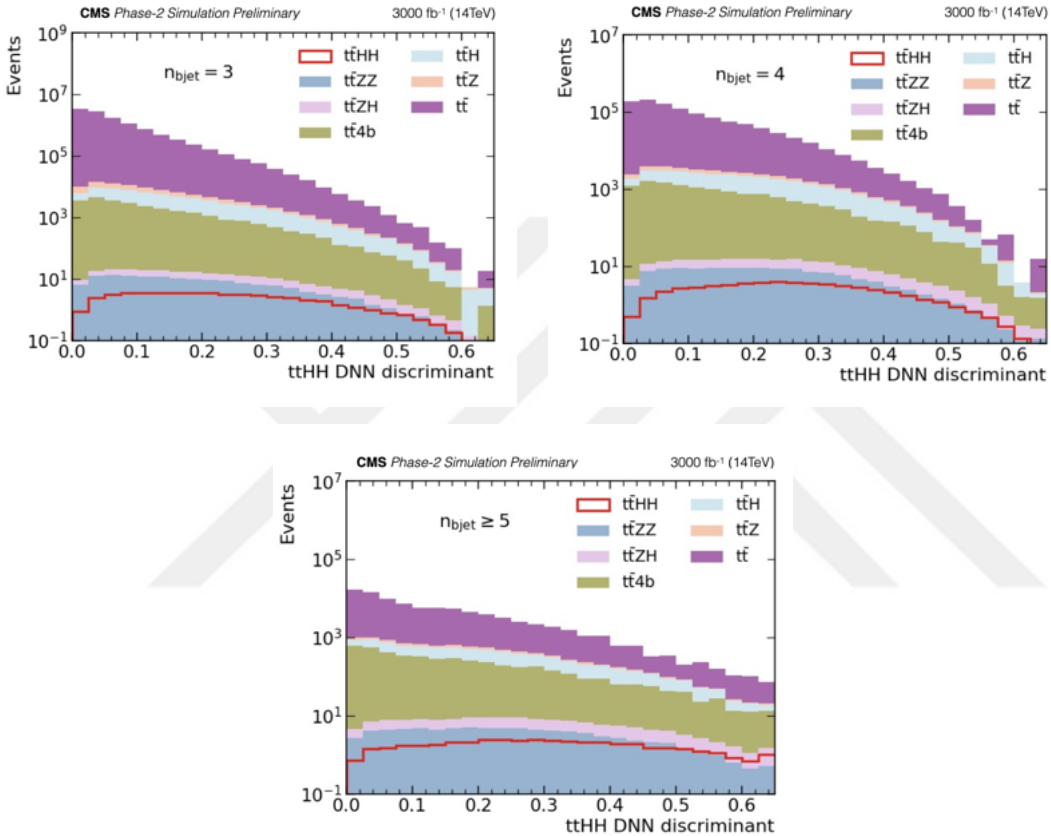

Figure 8.6: Final discriminant distributions for SM $\mathrm{t\bar{t}HH}$ are shown for three different $\mathrm{n_{bjet}}$ categories; $\mathrm{n_{bjet}} = 3$ (top) $\mathrm{n_{bjet}} = 4$ (middle), and $\mathrm{n_{bjet}} > 4$ (bottom). The plots show the expected event yields.

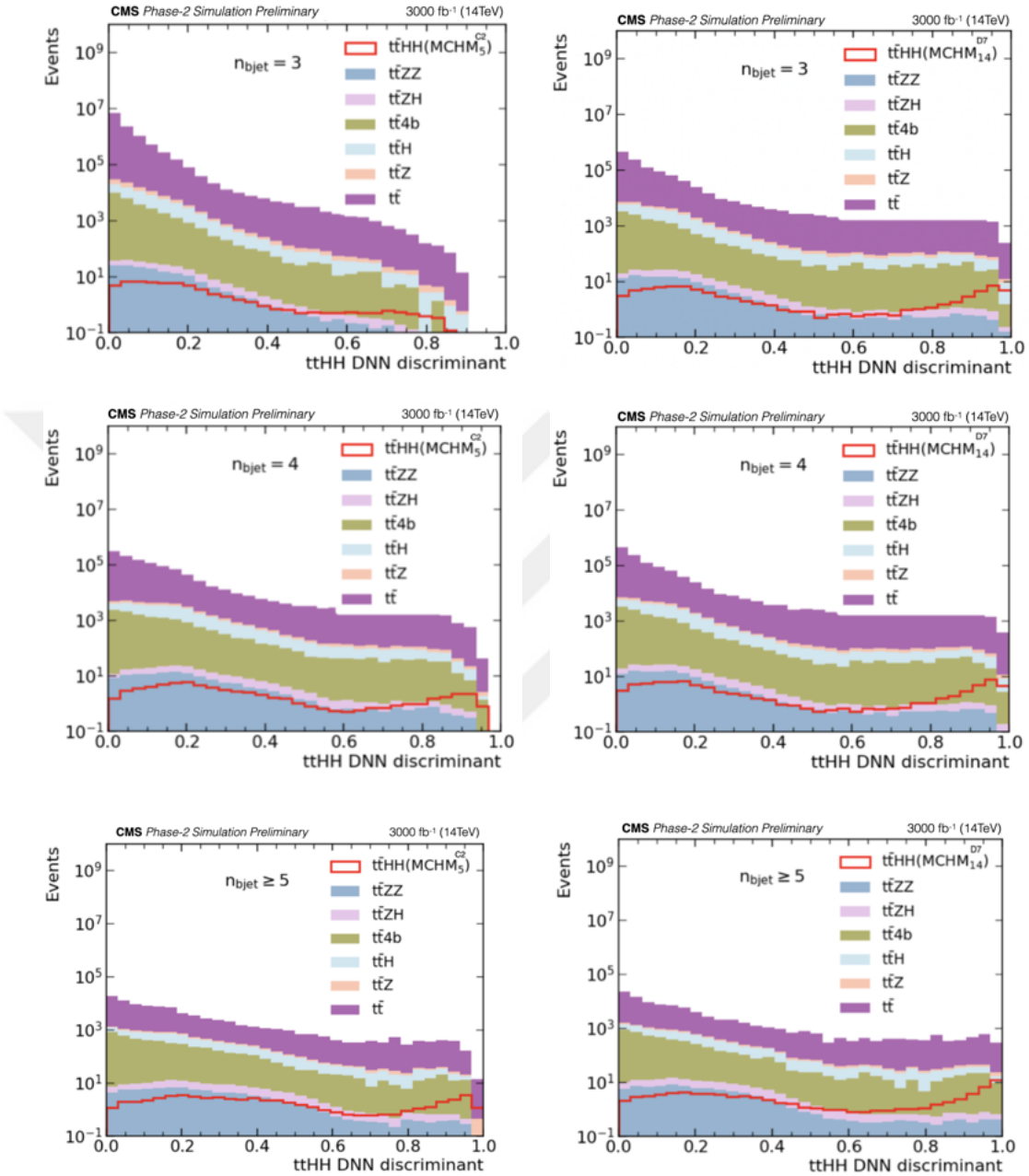

Figure 8.7: Final discriminant distributions for the $\mathrm{t\bar{t}HH}$ $\mathrm{MCHM}_5^{\mathrm{C2}}$ benchmark point case (left) and $\mathrm{t\bar{t}HH}$ $\mathrm{MCHM}_{14}^{\mathrm{D7}}$ benchmark point case (right) are shown. The plots show the expected event yields.

CHAPTER 9

RESULTS AND INTERPRETATION

In this chapter, the results of the measurement for $\mathrm{t\bar{t}HH}$ production obtained for the two complimentary study is presented. Sources of uncertainties for both the full Run 2 study in the DL channel and the HL-LHC study in the SL channel are discussed. Expected limits are presented for the $\mathrm{t\bar{t}HH}$, $\mathrm{t\bar{t}HH} + \mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}HH} + \mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}ZZ}$ processes for the 2017 run period in the DL chanel study. In the SL channel study, expected limits are obtained for the SM $\mathrm{t\bar{t}HH}$, $\mathrm{t\bar{t}HH}+\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}HH}+\mathrm{t\bar{t}ZH}+\mathrm{t\bar{t}ZZ}$ processes as well as $\mathrm{t\bar{t}HH\ MCHM_5^{C2}}$ and $\mathrm{t\bar{t}HH\ MCHM_{14}^{D7}}$ processes. The statistical methods considered here are described in Appendix D.

9.1 Measurement of the $\mathrm{t\bar{t}HH}$ process in the dileptonic decay channel with the 2017 Run 2 data

Systematic uncertanities

This study addresses two categories of systematic uncertainties: experimental and theoretical. Experimental uncertainties arise from detector effects, including particle misidentification, low efficiency, and resolution constraints within the detector sub-systems. The analysis considers experimental uncertainties related to integrated luminosity measurement, trigger efficiency, lepton scale factors, size of the MC samples, calibration of the jet energy scale and resolution, b tag scale factors applied at the event weight levels. Cross section uncertainties are considered as the theoretical ones. All systematic uncertainties with their types are listed in Table 9.1. Each rate systematic is added in the final step, while for shape systematics, varied templates with up/down variation are generated at the analyzer level. The rate inclusive cross

section uncertainties are provided in Table 9.2 for each process.

Table 9.1: Experimental and theoretical systematic uncertainties considered in this study.

Systematic uncertainties	Type
Integrated luminosity	Rate
Muon ID/Iso/Reco	Shape
Electron ID/Reco	Shape
Trigger efficiency	Shape
PDF+α_s	Rate
Pileup	Shape
Jet energy scale	Shape
Jet energy correction	Shape
b-tag uncertanities (hf/lf/hfstats1/hfstats2/lfstats1/lfstats2)	Shape

Table 9.2: Cross sections at 13 TeV with the uncertanties ±QCD Scale (%) and ±(PDF+α_s) (%) for the signal and all the considered backgrounds in this study.

Samples	σ [fb] at 13 TeV
$t\bar{t}HH$, HH $\rightarrow b\bar{b}b\bar{b}$	$0.775^{+1.5\%}_{-4.3\%} \pm 3.2\%$ [NLO]$*$
$t\bar{t}H$	$501^{+5.9\%}_{-9.3\%} \pm 3.6\%$ [NLO]$*$
$t\bar{t}$ + jets	$831760^{+19.8\%}_{-29.2\%} \pm 3.5\%$ [NNLO] $**$
$t\bar{t}$ + bb	$1452^{+37.6\%}_{-27.5\%} \pm 3.2\%$ [NLO] $***$
$t\bar{t}$ + 4b	296 [LO]
$t\bar{t}Z$	$841^{+9.6\%}_{-11.3\%} \pm 2.8\%$ [NLO]$*$
$t\bar{t}ZZ$	$1.98^{+5.2\%}_{-9.0\%} \pm 2.6\%$ [NLO]$*$
$t\bar{t}ZH$	$1.535^{+1.9\%}_{-6.8\%} \pm 3.0\%$ [NLO]$*$

Results

The methodology described in Ref [171] is used for the statistical inference of six discriminants obtained from the evaluation of the DNN algorithm. These output nodes having the optimized binning, full set of MC corrections as well as systematic uncertainties are provided for the $t\bar{t}HH$ signal node, and $t\bar{t}ZH$, $t\bar{t}ZZ$, $t\bar{t}$, $t\bar{t}Z$, $t\bar{t}H$ background nodes in the signal region and being represented in Figure 9.1. The systematical and statistical uncertainty bands on the DNN plots are also included.Simultaneous binned maximum likelihood fits are performed to these discriminants. Experimental and theoretical systematic uncertainties are considered as the nuisance parameters

of the likelihood functions, for some of which log-normal prior distributions are assumed, and the template shape variations are taken into account via continuous template morphing where shape variations are provided.

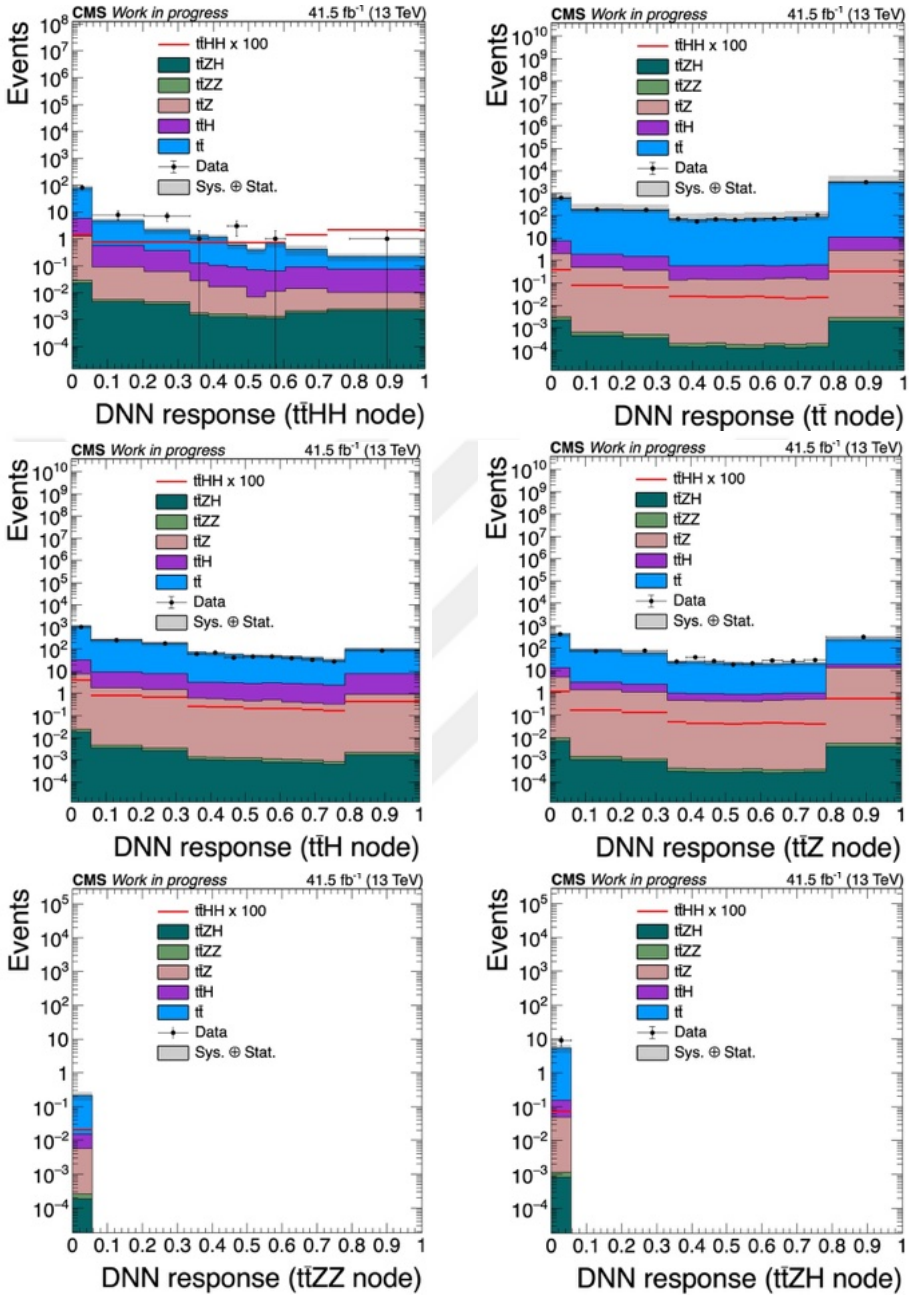

Figure 9.1: Final discriminant distributions for the $t\bar{t}HH$ signal node and $t\bar{t}$, $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t}ZZ$, and $t\bar{t}ZH$ background nodes.

The expected upper limits are computed for 41.5 fb^{-1} of data by considering both systematic and statistical uncertainties together, as well as statistical uncertainties alone. For the results with all the uncertainties,, a 95% confidence level upper limit

on the $t\bar{t}HH$ production cross section is observed at 94.23 times the SM prediction for an expected value of $69.25^{+40.02}_{-24.26}$, as presented in Table 9.3. The result for the statistical uncertainties only case is provided in Table 9.4.

Table 9.3: Upper limits on the signal strength shown for the $t\bar{t}HH$ signal, with considering both systematical and statistical uncertainties.

Syst. $\oplus$ Stat.			$1.0^{+30.4}_{-12.5}$
	95% CL upper limits on μ	Observed	94.23
		Expected (Median)	69.25
		Expected (68% CL range)	[44.99, 109.27]
		Expected (95% CL range)	[31.38, 167.05]

Table 9.4: Upper limits on the signal strength shown for the $t\bar{t}HH$ signal, with considering only statistical uncertainties.

Stat. only			$1.0^{+29.2}_{-10.8}$
	95% CL upper limits on μ	Expected (Median)	66.75
		Expected (68% CL range)	[42.86, 106.66]
		Expected (95% CL range)	[29.46, 161.59]

Similarly, the upper limit at the 95% CL on the combined $t\bar{t}ZZ + t\bar{t}ZH + t\bar{t}HH$ signal strength is expected to be $58.75^{+33.49}_{-20.29}$ when all uncertainties are considered. If $t\bar{t}ZZ$ is taken as background, the upper limit on the combined $t\bar{t}ZH + t\bar{t}HH$ signal strength is expected to be $60.50^{+34.49}_{-21.04}$. Additionally, observed upper limits are found at 87.92 and 86.78 times the SM prediction for the $t\bar{t}ZH + t\bar{t}HH$ and $t\bar{t}ZZ + t\bar{t}ZH + t\bar{t}HH$ signal scenarios, respectively. All results are illustrated in Figure 9.2.

A detailed analysis of how individual nuisance parameters influence the signal strength provides a more comprehensive insight into the statistical model. Figure 9.3 shows the 30 parameters with the highest impact and their pulls on the signal strength in the fit to pseudo data with the signal process expected from the SM.

At the time of authoring this thesis, a comprehensive analysis pipeline has been executed for the 41.5 fb^{-1} of data. In the forthcoming period, identical procedures will be applied to both the 59.7 fb^{-1} and the 36.3 fb^{-1} datasets, cumulatively contributing to the overall dataset of 137.6 fb^{-1} of data.

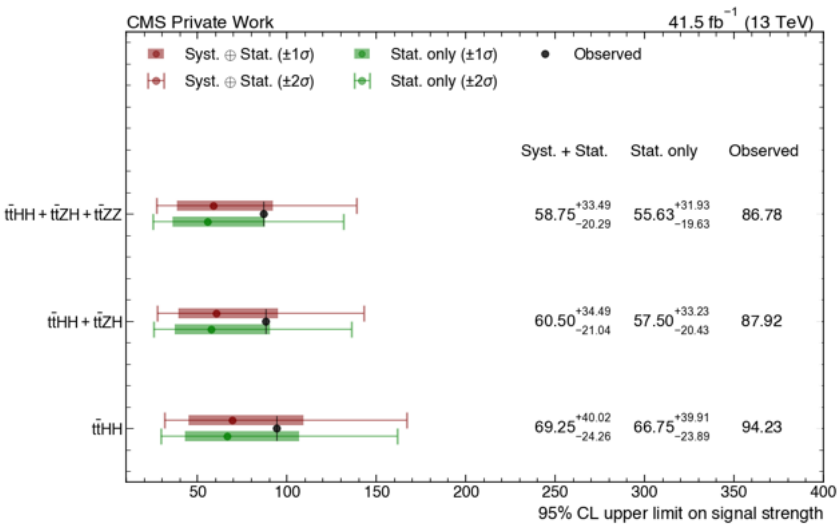

Figure 9.2: The 95% upper limits on the signal strength shown for the $t\bar{t}HH$, $t\bar{t}HH + t\bar{t}ZH$ and $t\bar{t}HH + t\bar{t}ZH + t\bar{t}ZZ$ processes for different scenarios of systematic uncertainties.

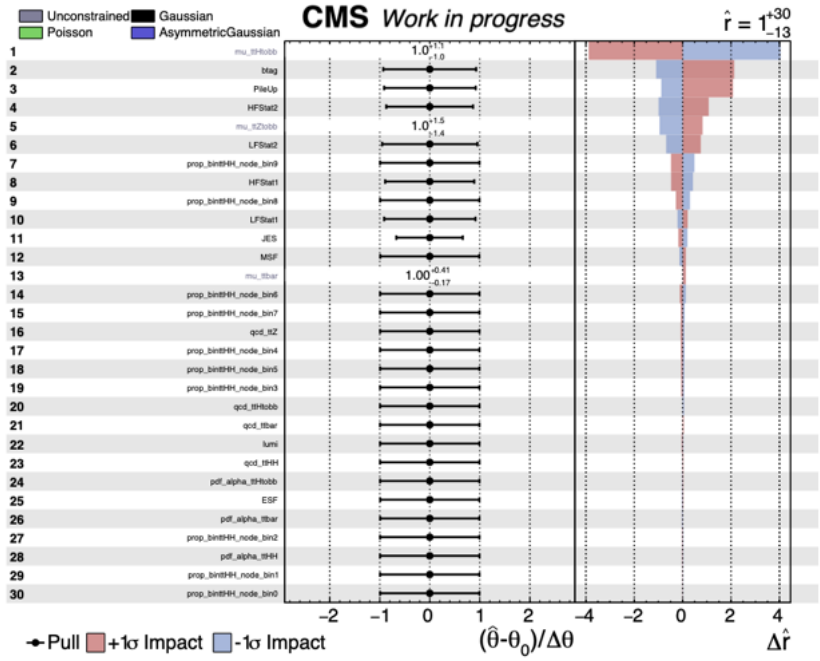

Figure 9.3: The top 30 impacts and their effects on the signal strength in the fit to pseudo data. The black lines, also known as pulls, illustrate the values of the nuisance parameters relative to their initial values before the fit and in relation to their uncertainty. The red and blue lines show the change in Δ_μ when the nuisance parameter is varied up and down by its fitted uncertainty, respectively.

9.2 Search for the $\mathrm{t\bar{t}HH}$ process in the semileptonic decay channel at the HL-LHC

Systematic uncertainties

There are several systematic uncertainties on signal and background yields due to detector effects and theory calculations which have been taken into account when predicting the sensitivity. Three scenarios for systematic uncertainties are considered in comparison to each other:

- “YR18" systematics scenario: This scenario represents the realistic systematic uncertainties assumed for the Phase-2 era, and presented in the CERN HL-LHC Yellow Report for Higgs physics [138]. A scale uncertainty of 30% is considered for the $\mathrm{t\bar{t}} + 4\mathrm{b}$ background.
- “YR18 conservative" systematics scenario: Same as above, but a scale uncertainty of 70% is considered for the $\mathrm{t\bar{t}} + 4\mathrm{b}$ background.
- Statistics-only scenario: This scenario represents the ultimate precision limit. It assumes that no systematic uncertainties are existent.

Table 9.5 lists the YR18 systematic uncertainties implemented and their effects on signal and background yields. These systematic variations either only effect the yield rate or effect the shape of the distributions, i.e. having a different percent effect on each bin. Jet energy scale, jet energy resolution and b-tagging uncertainties are applied per each object. Since jet and b jet multiplicities vary for each event, the overall effect on signal and background yields will vary accordingly. The other uncertainties are applied per event. Hence, their overall effect on the yields will be constant. Theoretical uncertainties in signal and background cross sections due to QCD scale variations and the choice of parton distribution functions including variations of α_s are also considered.

Results and Interpretation

The methodology followed for the Run 2 study is also used here for the statistical inference of the discriminant distributions obtained by the DNN training. Simulta-

Table 9.5: List of systematic uncertainties applied in this analysis, their values and their effects on the SM signal and background yields.

Uncertainty source	Uncertainty (%)	Type	Impact on signal yield (%)	Impact on BG yield (%)
Jet energy scale	0.4–3	shape	$-0.5/+0.4$	$-2.4/+2.3$
Jet energy resolution	0.4–3	shape	$-0.02/-0.6$	$-0.2/-2.2$
b tagging	1	shape	$-1.3/+1.1$	$-0.14/+0.16$
Lepton identification	0.5	rate	± 0.5	± 0.5
Luminosity	1	rate	± 1	± 1
Theory uncertainties on cross section: Values from Table 8.1.				

neous binned maximum likelihood fits are performed to the discriminants in all $\mathrm{n_{bjet}}$ categories. Experimental and theoretical systematic uncertainties are considered as the nuisance parameters of the likelihood functions, for which log-normal prior distributions are assumed, and the template shape variations are taken into account via continuous template morphing.

Limits are computed for 3000 fb^{-1} of data. As shown in Figure 8.3, it is not trivial to obtain a significant separation between the $\mathrm{t\bar{t}HH}$, $\mathrm{t\bar{t}ZH}$ and $\mathrm{t\bar{t}ZZ}$ processes. This arises from the difficulty in discriminating $\mathrm{Z \to bb}$ and $\mathrm{H \to bb}$ decays, due to their very similar kinematic characteristics and the proximity of the Z and H mass scales. We therefore quote as a main result the conservative limits obtained by treating the combination of these 3 processes as the signal during the limit calculation step. Limits are also quoted for cases where either one or both of the processes involving Z bosons are treated as backgrounds. For the YR18 systematic uncertainties scenario, the upper limit at the 95% CL on the combined $\mathrm{t\bar{t}ZZ + t\bar{t}ZH + t\bar{t}HH}$ signal strength is expected to be $0.84^{+0.34}_{-0.24}$. If $\mathrm{t\bar{t}ZZ}$ is taken as background, the upper limit on the combined $\mathrm{t\bar{t}ZH + t\bar{t}HH}$ signal strength is expected to be $1.31^{+0.53}_{-0.37}$. Furthermore, if $\mathrm{t\bar{t}ZZ}$ and $\mathrm{t\bar{t}ZH}$ are both taken as backgrounds, the upper limit on the $\mathrm{t\bar{t}HH}$ signal strength is expected to become $3.14^{+1.27}_{-0.9}$. The quoted uncertainties are obtained by the limited treatment of systematic uncertainties in this study based on current modelling of the Phase-2 detectors. A more refined study will accompany the analysis of the data that will be delivered by the HL-LHC. The uncertainties, however, give a good indication of the relative effect of different systematic uncertainty scenarios.

As described in the beginning of this section, three different theoretical scale uncertainty values have been tested for the $t\bar{t}4b$ process: 30%, 70% and leaving its normalization unconstrained in the fit procedure. Less than 2% impact is observed on the expected upper limits between the three choices. Consequently, only the two cases with 30% and 70% scale uncertainties are considered. However, analysis of the shape uncertainties due to mismodeling of the $t\bar{t}4b$ process is beyond the scope of the current study and these uncertainties are not considered in the signal strength estimation. Figure 9.4 shows a summary of the signal strength measurements for the three scenarios for systematic uncertainties.

Finally, a binned maximum likelihood fit is performed to the DNN discriminants of the $\mathrm{MCHM}_5^{\mathrm{C2}}$ and $\mathrm{MCHM}_{14}^{\mathrm{D7}}$ $t\bar{t}HH$ processes. Within the expectation of these models, the upper limits at the 95% confidence level on the signal strengths are obtained as $1.72^{+0.76}_{-0.53}$ and $1.08^{+0.43}_{-0.30}$ respectively. A summary of the signal strength measurements for the three aforementioned systematic scenarios is shown in Fig. 9.5

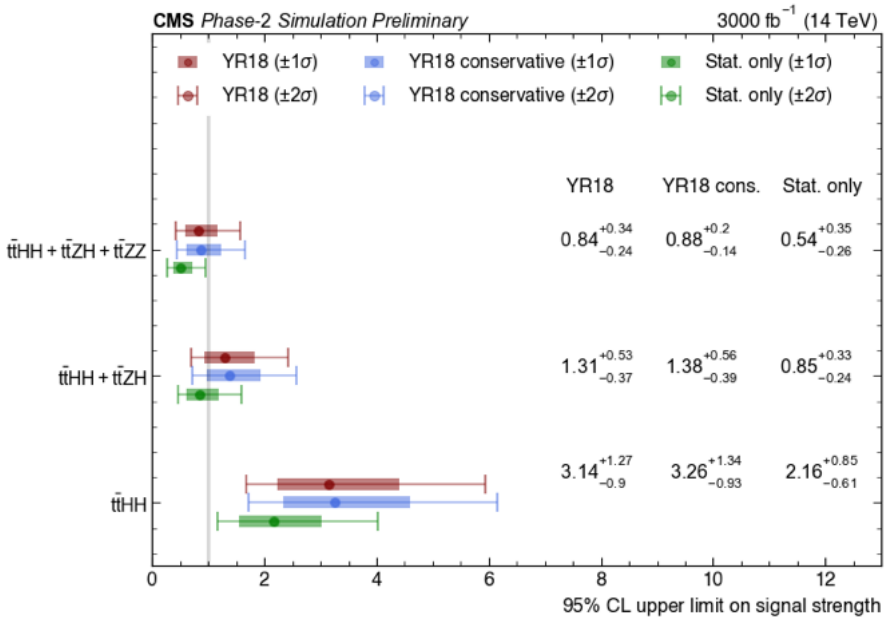

Figure 9.4: The 95% upper limits on the signal strength shown for the SM $t\bar{t}HH$, $t\bar{t}HH + t\bar{t}ZH$ and $t\bar{t}HH + t\bar{t}ZH + t\bar{t}ZZ$ processes for different scenarios of systematic uncertainties.

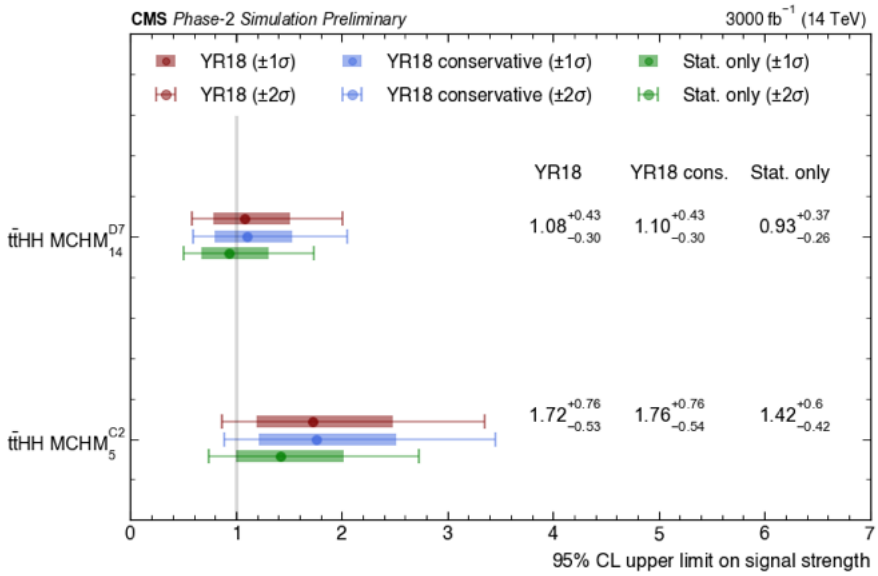

Figure 9.5: The 95% upper limits on the signal strength for the $t\bar{t}HH$ MCHM^{C2}_{5} and $t\bar{t}HH$ MCHM^{D7}_{14} processes for different scenarios of systematic uncertainties.

Prospects for BSM interpretation

This part discusses the interpretation prospects for the two showcase MCHM benchmarks considered here, based on the findings of this study. It is important to emphasize that MCHM results leading to this discussion are obtained using the same methodology developed for the SM $t\bar{t}HH$ process. The aim is to stress the various handles offered by $t\bar{t}HH$ for searching BSM.

The first handle is on the effect of MCHM on the kinematic distributions due to the charge 2/3 top partners contributions, as demonstrated by the plots and the discussion at the end of Section 8.4. The second handle is the identification of peaks in the reconstructed (top, Higgs) invariant mass distributions of the MCHM scenarios. Such peaks are expected in both MCHM benchmarks. The resonant components in MCHM^{C2}_{5} and MCHM^{D7}_{14} correspond to 41% and 63% of the total cross section, respectively. The predicted charge 2/3 top partners are in a mass range between 1.5 TeV and 2.2 TeV, and all decay with relatively high branching ratios into tH, and thus are within the reach of the HL-LHC.

The third handle comes from the signal yields and from the signal strength modifiers $\mu_{MCHM}(t\bar{t}HH)$ estimated with respect to the cross sections from the MCHM theory

prediction, as demonstrated in this analysis. The $\mathrm{MCHM}_5^{\mathrm{C2}}$ and $\mathrm{MCHM}_{14}^{\mathrm{D7}}$ signal strengths summarized in Fig. 9.5 can be interpreted as follows:

- For the $\mathrm{MCHM}_{14}^{\mathrm{D7}}$ benchmark, the 95% CL upper limit is compatible with the predicted $\mathrm{MCHM}_{14}^{\mathrm{D7}}$ cross section, with about 35% accuracy, putting this model within the reach of the HL-LHC. Discovering a deviation from the SM would constitute a preliminary step towards further BSM investigations. To establish a potential discovery, the signal strength measurements need be complemented with measurements featuring predicted resonances, such as mass spectra, branching ratios and couplings.

- For the $\mathrm{MCHM}_5^{\mathrm{C2}}$ benchmark, the upper limit expectation at 95% CL is not sufficient due to large error bars, for making a conclusive statement on the compatibility with the MCHM theory cross section. Improvements on analysis methodology and measurements on other $\mathrm{t\bar{t}HH}$ final states are needed to approach signal strength values of 1, and arrive at a more precise conclusion.

.

The above discussion featuring the MCHM scenarios as an example, highlights the overall importance of studying the $\mathrm{t\bar{t}HH}$ process for the discovery and characterization of BSM signals.

CHAPTER 10

CONCLUSION

Two complimentary studies are presented in this thesis. The first study uses full Run 2 data and describe a search for the $\mathrm{t\bar{t}HH}$ production at a center-of-mass energy of 13 TeV with a total luminosity of 137.6 fb^{-1}, where the top quark-antiquark pair decaying dileptonically, and Higgs boson pair decaying hadronically into b-quark-antiquark pairs. The second study is performed to assess the sensitivity of the HL-LHC and the Phase-2 CMS detector to standard model di-Higgs production in association with a top-antiquark pair ($\mathrm{t\bar{t}HH}$) for an integrated luminosity of 3000 $^{-1}$. The Higgs bosons are assumed to decay into b quark pairs while the $\mathrm{t\bar{t}}$ system is assumed to decay semileptonically. In addition, the study explored the HL-LHC sensitivity to beyond the standard model contributions to $\mathrm{t\bar{t}HH}$ within the context of the Minimal Composite Higgs Models (MCHM).

Both analyses investigate comparable final states characterized by a shared signature involving numerous jets, multiple b-quark jets, and moderate missing transverse energy. However, they differ in terms of lepton numbers, with the Run 2 study yielding a final state with two leptons, whereas the HL-LHC study features precisely one lepton.

After a baseline selection applied, the selected events are used for training deep neural networks that enhance signal by classifying events into signal and background categories. In the HL-LHC study, dedicated networks are trained for the SM and MCHM signals. Events are partitioned into 3 search channels having number of b-quark jets equal to 3, equal to 4 and greater then or equal to 5. A statistical analysis is performed by simultaneously fitting the multi-classifier DNN discriminants using a profile likelihood ratio method in both studies. For the HL-LHC all available categories for the three b-jet multiplicity channels are considered. Effects of various systematic uncer-

tainties for the Phase-2 conditions are taken into account.

For the Run 2 study, several experimental and theoretical uncertainties are taken into account. The upper limit at the 95% CL on the combined $\mathrm{t\bar{t}ZZ} + \mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$ signal strength is expected to be $58.75^{+33.49}_{-20.29}$ when all uncertainties are considered. If $\mathrm{t\bar{t}ZZ}$ is taken as background, the upper limit on the combined $\mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$ signal strength is expected to be $60.50^{+34.49}_{-21.04}$. The expected upper limits at the 95% CL on the signal strengths as the ratio of the $\mathrm{t\bar{t}HH}$ process to the SM expectations is obtained as $69.25^{+40.02}_{-24.26}$. For the dataset comprising 41.5 fb^{-1}, the observed upper limits for the signal scenarios $\mathrm{t\bar{t}HH}$, $\mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$, and $\mathrm{t\bar{t}ZZ} + \mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$ are respectively identified as 94.23, 87.92, and 86.78 times the SM prediction.

For the HL-LHC study, by using the DELPHES simulated samples and considering the YR18 systematic uncertainties, it is expected that the upper limit at the 95%CL on the combined $\mathrm{t\bar{t}ZZ} + \mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$ production cross section is $0.84^{+0.34}_{-0.24}$ times the SM prediction. If $\mathrm{t\bar{t}ZZ}$ is taken as background, the upper limit on the combined $\mathrm{t\bar{t}ZH} + \mathrm{t\bar{t}HH}$ production cross section is expected to be $1.31^{+0.53}_{-0.37}$ times the SM prediction. If $\mathrm{t\bar{t}ZZ} + \mathrm{t\bar{t}ZH}$ are taken as backgrounds, the upper limit on the $\mathrm{t\bar{t}HH}$ production cross section alone is expected to become $3.14^{+1.27}_{-0.9}$ times the SM prediction. For the MCHM case, the upper limits at the 95% CL $\mathrm{t\bar{t}HH}$ cross sections are obtained as $1.72^{+0.76}_{-0.53}$ times the $\mathrm{MCHM}^{\mathrm{C2}}_{5}$ prediction and $1.08^{+0.43}_{-0.30}$ times the $\mathrm{MCHM}^{\mathrm{D7}}_{14}$ prediction, respectively. Moreover, the analysis demonstrated various kinematic characteristics of the MCHM scenarios that would discriminate them from the SM $\mathrm{t\bar{t}HH}$ process.

In the conclusion of the analysis for Run 2, the most advanced methods were employed such as GATJA to obtain the results discussed. It is clear that an increase in luminosity is essential for any potential discovery. Although a discovery has not yet been made, the results achieved are valuable, providing constraints on new physics theories. Additionally, the study on the HL-LHC offers insights into the extent of improvement that high luminosity could bring. Overall, these complementary works demonstrate the importance of studying $\mathrm{t\bar{t}HH}$ as a key process in establishing the top-Higgs sector and its prospects, both on the SM and BSM fronts, in view of both Run 2 data taking period and the expected progress within the next decade, until the

completion of the HL-LHC.

Finally, the HL-LHC study is published by the CMS collaboration in a Physics Analysis Summary (PAS) [15] and also contributes to the Snowmass White Paper publication [16]. Additionally, this thesis outlines the execution of an extensive analysis pipeline on a dataset comprising 41.5 fb^{-1}. In the next phase, the same analytical methods will be extended to datasets of 59.7 fb^{-1} and 36.3 fb^{-1}, thereby contributing to the overall dataset of 137.6 fb^{-1} of data. Full results for the Run 2 study are expected to be published in the first half of 2024.

REFERENCES

[1] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Physical Review Letters*, vol. 13, no. 16, pp. 508–509, 1964.

[2] P. W. Higgs, "Broken symmetries, massless particles and gauge fields," *Physics Letters*, vol. 12, no. 2, pp. 132–133, 1964.

[3] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons," *Phys. Rev.*, vol. 145, p. 1156, 1966.

[4] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," *Physical Review Letters*, vol. 13, no. 9, pp. 321–323, 1964.

[5] CMS Collaboration, "Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV," *Phys. Lett. B*, vol. 710, p. 26, 2012.

[6] ATLAS Collaboration, "Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Phys. Rev. D*, vol. 86, p. 032003, 2012.

[7] CMS Collaboration, "Measurement of the top quark Yukawa coupling from $t\bar{t}$ kinematic distributions in the dilepton final state in proton-proton collisions at $\sqrt{s}$ = 7 TeV," *Physical Review D*, vol. 102, no. 9, 2020.

[8] C. Csáki and P. Tanedo, "Beyond the Standard Model," 2015. Proceedings of the 2013 European School of High-Energy Physics, Paradfurdo, Hungary, 5-18 June 2013, edited by M. Mulders and G. Perez, CERN-2015-004 (CERN, Geneva, 2015).

[9] K. Agashe, R. Contino, and A. Pomarol, "The Minimal Composite Higgs Model," *Nucl. Phys.*, vol. B719, pp. 165–187, 2005.

[10] G. Panico and A. Wulzer, "The Composite Nambu-Goldstone Higgs," *Lect. Notes Phys.*, vol. 913, pp. pp.1–316, 2016.

[11] M. Carena, L. Da Rold, and E. Pontón, "Minimal Composite Higgs Models at the LHC," *JHEP*, vol. 06, p. 159, 2014.

[12] C. Bautista *et al.*, "Probing the Top-Higgs sector with Composite Higgs Models at present and future hadron colliders," *JHEP*, vol. 03, p. 049, 2021.

[13] D. de Florian *et al.*, "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector," 2017.

[14] LHC Higgs Cross Section Working Group, "Current recommendations for HH cross-sections." Twiki: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWGHH?redirectedfrom=LHCPhysics.LHCHXSWGHHr87.

[15] CMS Collaboration, "Search for the nonresonant $t\bar{t}HH$ production in the semileptonic decay of the top pair and the Higgs pair decay into b quarks at the HL-LHC," CMS FTR Note CMS-PAS-FTR-21-010, CERN, 2022.

[16] ATLAS Collaboration, "Snowmass White Paper Contribution: Physics with the Phase-2 ATLAS and CMS Detectors," 2022. https://cds.cern.ch/record/2805993.

[17] D. J. Griffiths, *Introduction to Elementary Particles; 2nd rev. version*. Physics textbook, New York, NY: Wiley, 2008.

[18] M. K. Gaillard, P. D. Grannis, and F. J. Sciulli, "The Standard Model of Particle Physics," *Reviews of Modern Physics*, vol. 71, p. S96–S111, Mar. 1999.

[19] S. F. Novaes, "Standard model: An Introduction," in *10th Jorge Andre Swieca Summer School: Particle and Fields*, pp. 5–102, 1 1999.

[20] R. Mann, *An Introduction to Particle Physics and the Standard Model*. CRC Press, 1st ed., 2010. https://doi.org/10.1201/9781420083002.

[21] D. Perkins, *Introduction to High Energy Physics*. Reading, USA: Addison-Wesley, 1982.

[22] R. L. Workman *et al.*, "Review of Particle Physics," *Progress of Theoretical and Experimental Physics*, vol. 2022, p. 083C01, 2022.

[23] CMS Collaboration, "What Do We Already Know? – Forces and carrier particles." https://cmsexperiment.web.cern.ch/news/what-do-we-already-know Accessed on 6 Dec. 2023.

[24] M. K. Gaillard, P. D. Grannis, and F. J. Sciulli, "The Standard Model of Particle Physics," *Reviews of Modern Physics*, vol. 71, p. S96–S111, Mar. 1999.

[25] C. Burgard, "Standard model of physics: Tikz example." Accessed on 6 Dec. 2023 https://texample.net/tikz/examples/model-physics/.

[26] A. Pich, "Quantum chromodynamics," in *1994 European School of High-energy Physics*, pp. 157–207, 5 1995.

[27] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?," *Phys. Rev. Lett.*, vol. 30, pp. 1346–1349, June 1973.

[28] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories," *Phys. Rev. Lett.*, vol. 30, pp. 1343–1346, June 1973.

[29] S. L. Glashow, "Partial-symmetries of weak interactions," *Nucl. Phys.*, vol. 22, p. 579, 1961.

[30] A. Salam, "Weak and Electromagnetic Interactions," *Conf. Proc. C*, vol. 680519, p. 367, 1968.

[31] S. Weinberg, "A Model of Leptons," *Phys. Rev. Lett.*, vol. 19, p. 1264, 1967.

[32] J. Ellis, "Higgs Physics," in *2013 European School of High-Energy Physics*, pp. 117–168, 2015.

[33] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles," *Phys. Rev. Lett.*, vol. 13, pp. 585–587, 1964.

[34] T. W. B. Kibble, "Symmetry breaking in non-Abelian gauge theories," *Phys. Rev.*, vol. 155, p. 1554, 1967.

[35] UA1 Collaboration, "Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider," *Phys. Lett. B*, vol. 126, p. 398, 1983.

[36] UA2 Collaboration, "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider," *Phys. Lett. B*, vol. 122, p. 476, 1983.

[37] UA2 Collaboration, "Evidence for $Z^0 \to e^+e^-$ at the CERN pp Collider," *Phys. Lett. B*, vol. 129, p. 130, 1983.

[38] ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the Tevatron Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, "Precision Electroweak Measurements and Constraints on the Standard Model," Tech. Rep. PH-EP-2010-095, CERN, 2010.

[39] ALEPH, DELPHI, L3, OPAL Collaborations, and LEP Working Group for Higgs Boson Searches, "Search for the Standard Model Higgs boson at LEP,"

[40] CDF and D0 Collaborations, "Combination of Tevatron Searches for the Standard Model Higgs Boson in the W^+W^- Decay Mode," *Phys. Rev. Lett.*, vol. 104, p. 061802, 2010.

[41] CDF Collaboration, "Combined search for the standard model Higgs boson decaying to a $b\bar{b}$ pair using the full CDF data set," 2012. Submitted to Phys. Rev. Lett.

[42] CDF and D0 Collaborations, "Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in higgs boson search at the tevatron," 2012. Submitted to Phys. Rev. Lett.

[43] D0 Collaboration, "Combined search for the standard model Higgs boson decaying to $b\bar{b}$ using the D0 Run II data set," 2012. Submitted to Phys. Rev. Lett.

[44] D. Green, ed., *At the leading edge: the ATLAS and CMS LHC experiments*. Singapore: World Scientific, 2010.

[45] L. Evans and P. Bryant, "LHC Machine," *Journal of Instrumentation*, vol. 3, no. 08, p. S08001, 2008.

[46] M. Carena, C. Grojean, M. Kado, and V. Sharma, "Status of Higgs boson physics," *Review of Particle Physics*, vol. Chin. Phys. C40, p. 100001, 2016.

[47] C. Grojean, "Higgs Physics," in *8th CERN–Latin-American School of High-Energy Physics*, pp. 143–158, 2016.

[48] S. Heinemeyer *et al.*, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group*. CERN Yellow Reports: Monographs, 2013.

[49] I. Neutelings, "Piechart of SM decays." https://tikz.net/sm_decay_piechart/. Accessed on 11 Dec. 2023.

[50] D. M. Asner *et al.*, "ILC Higgs White Paper," in *Snowmass 2013: Snowmass on the Mississippi*, 2013.

[51] CMS Collaboration, "Observation of Higgs boson decay to bottom quarks," *Phys. Rev. Lett.*, vol. 121, no. 12, p. 121801, 2018.

[52] ATLAS Collaboration, "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector," *Physics Letters B*, vol. 786, p. 59–86, Nov. 2018.

[53] CMS Collaboration, "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector," *Phys. Lett. B*, vol. 779, pp. 283–316, 2018.

[54] CMS Collaboration, "Measurement of Higgs boson production in association with a W or Z boson in the $\mathrm{H} \to \mathrm{WW}$ decay channel," 2021.

[55] CMS Collaboration, "Observation of ttH production," *Phys. Rev. Lett.*, vol. 120, no. 23, p. 231801, 2018.

[56] CMS Collaboration, "A measurement of the Higgs boson mass in the diphoton decay channel," *Physics Letters B*, vol. 805, p. 135425, 2020.

[57] CMS Collaboration, "A portrait of the Higgs boson by the CMS experiment ten years after the discovery," *Nature*, vol. 607, no. 7917, pp. 60–68, 2022.

[58] G. Arcadi, A. Djouadi, and M. Raidal, "Dark Matter through the Higgs portal," *Physics Reports*, vol. 842, p. 1–180, Feb. 2020.

[59] CMS Collaboration, "The CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, p. S08004, 2008.

[60] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, no. 08, p. S08003, 2008.

[61] The LHCb Collaboration, "The LHCb Detector at the LHC," *Journal of Instrumentation*, vol. 3, no. 08, p. S08005, 2008.

[62] ALICE Collaboration, "The ALICE experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, p. S08002, 2008.

[63] M. O. Sahin, *Search for Supersymmetric Top-Quark Partners Using Support Vector Machines and Upgrade of the Hadron Calorimeter Front-End Readout Control System at CMS*. PhD thesis, Hamburg, 2017. Universität Hamburg, Diss., 2016.

[64] L. Rossi, "The LHC main dipoles and quadrupoles toward series production," *IEEE Transactions on Applied Superconductivity*, vol. 13, no. 2, pp. 1221–1228, 2003.

[65] G. L. Bayatian *et al.*, "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software," 2006.

[66] G. L. Bayatian *et al.*, "CMS Technical design report, volume II: Physics performance," *J. Phys. G*, vol. 34, no. 6, pp. 995–1579, 2007.

[67] "Overview of CMS Physics Goals and Detector. Summary of Particle Detection." Twiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperimentr41.

[68] CMS Collaboration, "SketchUp 3D model of CMS detector," 2019. https://cds.cern.ch/record/2677903.

[69] D. Barney, "Presentation for public - Introduction to CMS for CERN guides," 2013. https://cds.cern.ch/record/2629323.

[70] CMS Collaboration, "CMS: The electromagnetic calorimeter. Technical design report," Dec. 1997.

[71] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report," 1997.

[72] CMS Collaboration, "CMS, the Compact Muon Solenoid. Muon technical design report," 12 1997.

[73] S. Cittolin, A. Rácz, and P. Sphicas, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project*. Technical design report. CMS, Geneva: CERN, 2002.

[74] V. Khachatryan *et al.*, "The CMS trigger system," *JINST*, vol. 12, p. P01020, 2017.

[75] V M Ghete (on behalf the CMS Collaboration), "The CMS L1 Trigger emulation software," *Journal of Physics: Conference Series*, vol. 219, p. 032009, apr 2010.

[76] "Data Formats and Data Tiers." Twiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormatsr31.

[77] C. Z. Mooney, *Monte Carlo Simulation*. Thousand Oaks, Calif.: Sage Publications, 1997.

[78] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*. Cambridge monographs on particle physics, nuclear physics, and cosmology, Cambridge University Press, 2003.

[79] J. Alwall *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *JHEP*, vol. 07, p. 079, 2014.

[80] S. Alekhin, J. Blümlein, S. Klein, and S. Moch, "The 3-, 4-, and 5-flavor next-to-next-to-leading order parton distribution functions from deep-inelastic-scattering data and at hadron colliders," *Physical Review D*, vol. 81, Jan. 2010.

[81] T. Sjostrand *et al.*, "An introduction to PYTHIA 8.2," *Computer Physics Communications*, vol. 191, p. 159, 2015.

[82] J. de Favereau *et al.*, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment," *JHEP*, vol. 02, p. 057, 2014.

[83] S. Agostinelli *et al.*, "GEANT4: A Simulation toolkit," *Nucl. Instrum. Meth.*, vol. 506, no. 3, p. 250, 2003.

[84] J. Allison *et al.*, "Geant4 developments and applications," *IEEE Trans. Nucl. Sci.*, vol. 53, p. 270, 2006.

[85] S. Frixione *et al.*, "Matching NLO QCD computations with parton shower simulations: the POWHEG method," *JHEP*, vol. 11, p. 070, 2007.

[86] O. Aberle *et al.*, *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs, Geneva: CERN, 2020.

[87] CMS Collaboration, "Technical Proposal for the Phase-II Upgrade of the CMS Detector," CMS Technical Proposal CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, 2015.

[88] CMS Collaboration, "The Phase-2 Upgrade of the CMS Tracker," CMS Technical Design Report CERN-LHCC-2017-009. CMS-TDR-014, 2017.

[89] CMS Collaboration, "The Phase-2 Upgrade of the CMS Barrel Calorimeters Technical Design Report," CMS Technical Design Report CERN-LHCC-2017-011. CMS-TDR-015, 2017.

[90] CMS Collaboration, "The Phase-2 Upgrade of the CMS Endcap Calorimeter," CMS Technical Design Report CERN-LHCC-2017-023. CMS-TDR-019, 2017.

[91] CMS Collaboration, "The Phase-2 Upgrade of the CMS Muon Detectors," CMS Technical Design Report CERN-LHCC-2017-012. CMS-TDR-016, 2017.

[92] CMS Collaboration, "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector," CMS FTR Note CMS-PAS-PFT-10-001, 2010.

[93] CMS Collaboration, "A MIP Timing Detector for the CMS Phase-2 Upgrade," tech. rep., CERN, Geneva, Mar 2019.

[94] CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger," tech. rep., CERN, Geneva, Apr 2020. Final version.

[95] CMS Collaboration, "The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger," tech. rep., CERN, Geneva, Mar 2021. This is a temporary submission, just to get the LHCC reference number, to be used in this and other CMS documents.

[96] CMS Collaboration, "Expected performance of the physics objects with the upgraded CMS detector at the HL-LHC," Tech. Rep. CMS-NOTE-2018-006. CERN-CMS-NOTE-2018-006, 2018.

[97] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector," *JINST*, vol. 12, no. 10, p. P10003, 2017.

[98] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *Journal of Instrumentation*, vol. 9, pp. P10009–P10009, oct 2014.

[99] R. Frühwirth, W. Waltenberger, and P. Vanlaer, "Adaptive Vertex Fitting," tech. rep., CERN, Geneva, 2007.

[100] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$ = 13 TeV," *JINST*, vol. 13, p. P06015, 2018.

[101] CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 16, p. P05014, may 2021.

[102] R. Atkin, "Review of jet reconstruction algorithms," *J. Phys. Conf. Ser.*, vol. 645, no. 1, p. 012008, 2015.

[103] M. Cacciari, G. P. Salam, and G. Soyez, "The $\mathrm{anti-k_T}$ jet clustering algorithm," *JHEP*, vol. 4, p. 63, 2008.

[104] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS," *JINST*, vol. 6, pp. P11002–P11002, 2011.

[105] CMS Collaboration, "Identification of b-quark jets with the CMS experiment," *JINST*, vol. 8, pp. P04013–P04013, apr 2013.

[106] B. Chazin Quero, "Machine learning techniques for heavy flavour identification," *PoS*, vol. LHCP2018, p. 066, 2018. On behalf of the CMS Collaboration.

[107] CMS Collaboration, "Missing transverse energy performance of the CMS detector," *JINST*, vol. 6, p. P09001, sep 2011.

[108] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[109] C. M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer-Verlag New York, Inc., 2006.

[110] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, p. 685–695, Apr. 2021.

[111] C. Nwankpa *et al.*, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," 2018.

[112] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[113] A. M. Carrington *et al.*, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, 2022. arXiv:2103.11357.

[114] N. Srivastava *et al.*, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[115] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, p. 295–316, Nov. 2020.

[116] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," 2020.

[117] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[118] S. Leijnen and F. van Veen, "The Neural Network Zoo," *Proceedings*, vol. 47, no. 1, p. 9, 2020.

[119] S. Pouyanfar *et al.*, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–36, 2019.

[120] T. Young *et al.*, "Recent trends in deep learning based natural language processing [review article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[121] I. Neutelings, "Neural networks." https://tikz.net/neural_networks/. Accessed on 1 Jan 2024.

[122] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," 2015.

[123] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[124] W. Fang *et al.*, "Computer vision for behaviour-based safety in construction: a review and future directions," *Adv Eng Inform*, vol. 43, p. 100980, 2020.

[125] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Commun*, vol. 108, pp. 15–32, 2019.

[126] H. Li, Z. Deng, and H. Chiang, "Lightweight and resource-constrained learning network for face recognition with performance optimization," *Sensors*, vol. 20, no. 21, p. 6114, 2020.

[127] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, Omnipress, 2010.

[128] A. F. Agarap, “Deep Learning using Rectified Linear Units (ReLU),” 2019.

[129] F. Scarselli *et al.*, “The Graph Neural Network Model,” *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[130] J. Duarte and J.-R. Vlimant, *Graph Neural Networks for Particle Tracking and Reconstruction*, p. 387–436. WORLD SCIENTIFIC, Feb. 2022.

[131] J. Shlomi, P. Battaglia, and J.-R. Vlimant, “Graph neural networks in particle physics,” *Machine Learning: Science and Technology*, vol. 2, p. 021001, Jan. 2021.

[132] B. Knyazev, G. W. Taylor, and M. R. Amer, “Understanding Attention and Generalization in Graph Neural Networks,” 2019.

[133] R. Frederix *et al.*, “Higgs pair production at the LHC with NLO and parton-shower effects,” *Physics Letters B*, vol. 732, p. 142–149, May 2014.

[134] G. Aad *et al.*, “Combination of searches for Higgs boson pairs in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector,”

[135] CMS Collaboration, “Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s}$ = 13 TeV,” *Phys. Rev. Lett.*, vol. 122, no. 12, p. 121803, 2019.

[136] M. Aaboud *et al.*, “Combination of the searches for pair-produced vector-like partners of the third-generation quarks at $\sqrt{s}$ = 13 TeV with the ATLAS detector,” 2018.

[137] CMS Collaboration, “Search for vector-like T and B quark pairs in final states with leptons at $\sqrt{s}$ = 13 TeV,” 2018.

[138] M. Cepeda *et al.*, “Report from Working Group 2,” *CERN Yellow Rep. Monogr.*, vol. 7, p. 221, 2019.

[139] CMS Collaboration, “Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS,” *Journal of Instrumentation*, vol. 11, pp. P01019–P01019, jan 2016.

[140] "Run-2 UltraLegacy Datasets for Analysis." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/PdmVRun2LegacyAnalysis. Accessed on 15 May, 2022.

[141] "Luminosity recommendations for Run 2 analyses." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/LumiRecommendationsRun2#Rules_for_CMS_publications. Accessed on 27 Sept, 2023.

[142] CMS Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements," *The European Physical Journal C*, vol. 80, Jan 2020.

[143] F. Maltoni, D. Pagani, and I. Tsinikos, "Associated production of a top-quark pair with vector bosons at NLO in QCD: impact on ttH searches at the LHC," *Journal of High Energy Physics*, vol. 2016, Feb 2016.

[144] LHC Higgs Cross Section Working Group, "SM Higgs Branching Ratios and Total Decay Widths." Twiki: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageBRr22.

[145] Ehatäht, Karl, "NANOAOD: a new compact event data format in CMS," *EPJ Web Conf.*, vol. 245, p. 06002, 2020.

[146] "Analysis datasets UltraLegacy 2016." Twiki: https://twiki.cern.ch/twiki/bin/viewauth/CMS/PdmVDataReprocessingUL2016. Accessed on 16 June,2023.

[147] "Analysis datasets UltraLegacy 2017." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/PdmVDataReprocessingUL2017. Accessed on 16 June,2023.

[148] "Analysis datasets UltraLegacy 2018." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/PdmVDataReprocessingUL2018. Accessed on 16 June,2023.

[149] "MET Filter Recommendations for Run 2 & Run 3." Twiki: https://twiki.cern.ch/twiki/bin/viewauth/CMS/MissingETOptionalFiltersRun2. Accessed on 15 June,2023.

[150] Muon Physics Object Group, "User recommendations." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/MuonPOG#User_Recommendationsr231.

[151] "Multivariate Electron Identification for Run2." Twiki: https://twiki.cern.ch/twiki/bin/viewauth/CMS/MultivariateElectronIdentificationRun2. Accessed on 16 June,2023.

[152] E/gamma Physics Object Group, "EgammaUL2016To2018." Twiki: https://twiki.cern.ch/twiki/bin/view/CMS/EgammaUL2016To2018r74.

[153] "The CMS NanoAOD data tier." Twiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD#Jets. Accessed on 17 July, 2023.

[154] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data," *JINST*, vol. 15, no. 09, p. P09018, 2020.

[155] "Jet identification in high pile-up environment (PileupJetID)." Twiki: https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJetID#Recommendations_for_13_TeV_Ultra. Accessed on 21 July, 2023.

[156] CMS b-tagging Vertexing POG, "BTV Scale Factors." https://btv-wiki.docs.cern.ch/ScaleFactors/. Accessed on 16 Oct, 2023.

[157] ATLAS Collaboration, "Measurement of event shapes at large momentum transfer with the ATLAS detector in pp collisions at $\sqrt{s}$ = 7 TeV," *The European Physical Journal C*, vol. 72, no. 11, 2012.

[158] ATLAS Collaboration, "Measurement of charged-particle event shape variables in inclusive $\sqrt{s}$ = 7 TeV proton-proton interactions with the ATLAS detector," *Physical Review D*, vol. 88, no. 3, 2013.

[159] M. Dasgupta and G. P. Salam, "Event shapes in e^+e^- annihilation and deep inelastic scattering," *Journal of Physics G: Nuclear and Particle Physics*, vol. 30, no. 5, pp. R143–R181, 2004.

[160] C. Bernaciak *et al.*, "Fox-Wolfram moments in Higgs physics," *Physical Review D*, vol. 87, no. 7, 2013.

[161] G. C. Fox and S. Wolfram, "Event shapes in e^+e^- annihilation," *Nuclear Physics B*, vol. 149, no. 3, pp. 413–496, 1979.

[162] J. N. Butler *et al.*, "Report of the 2021 U.S. Community Study on the Future of Particle Physics (Snowmass 2021) Summary Chapter," 2023.

[163] F. Buccioni *et al.*, "NLO QCD predictions for $t\bar{t}b\bar{b}$ production in association with a light jet at the LHC," *JHEP*, vol. 12, p. 015, 2019.

[164] C. Englert *et al.*, "Di-Higgs phenomenology in tthh: The forgotten channel," *Physics Letters B*, vol. 743, p. 93, 2015.

[165] R. D. Ball *et al.*, "Parton distributions from high-precision collider data," *Eur. Phys. J. C*, vol. 77, no. 10, p. 663, 2017.

[166] L. Li, Y.-Y. Li, and T. Liu, "Anatomy of tthh physics at the HL-LHC," *Physical Review D*, vol. 101, no. 5, 2020.

[167] D. Bertolini *et al.*, "Pileup per particle identification," *JHEP*, vol. 10, p. 059, 2014.

[168] CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger," CMS Technical Design Report CERN-LHCC-2020-004. CMS-TDR-021, 2020.

[169] CMS Collaboration, "The Phase-2 Upgrade of the CMS Data Acquisition and High-Level Trigger," CMS Technical Design Report CERN-LHCC-2021-007. CMS-TDR-022, 2021.

[170] CMS Collaboration, "B-tagging performance of the CMS Legacy dataset 2018," Mar 2021.

[171] J. S. Conway, "Incorporating Nuisance Parameters in Likelihoods for Multi-source Spectra," *arXiv: Data Analysis, Statistics and Probability*, pp. 115–120, 2011.

[172] F. Cavallari and C. Rovelli, "Calibration and Performance of the CMS Electromagnetic Calorimeter in LHC Run2," vol. 245, p. 02027, 2020.

[173] The CMS Higgs Combine Tool Developers, "CMS Higgs Combine Tool." https://cms-hcomb.gitbooks.io/combine. Accessed on 18 Jan 2024.

[174] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[175] G. Cowan *et al.*, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, Feb. 2011.

[176] ATLAS and CMS Collaborations, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011," 2011.

[177] A. L. Read, "Modified frequentist analysis of search results (The CL(s) method)," in *Workshop on Confidence Limits*, pp. 81–101, 2000.

APPENDIX A

COMPARISON OF PRE-LEGACY AND ULTRA-LEGACY DISTRIBUTIONS

Compared to legacy reconstruction (called pre-legacy (PL) in this thesis), the Run 2 ultra-legacy (UL) reprocessing includes an improved ECAL calibration for the full Run 2 dataset. This calibration provides a better energy resolution for electrons, especially in the forward region. Also, the CMS detector requires electron candidates to pass a set of dedicated high energy electron criteria (HEEP ID) and UL improves the efficiency of this HEEP ID. In addition to those, an overall better agreement between the performance in data and in simulations is observed after the UL reconstruction [172].

The plots below compare some kinematic characteristics of the UL and the PL samples for the signal and several background production processes for the events passing the baseline selection.

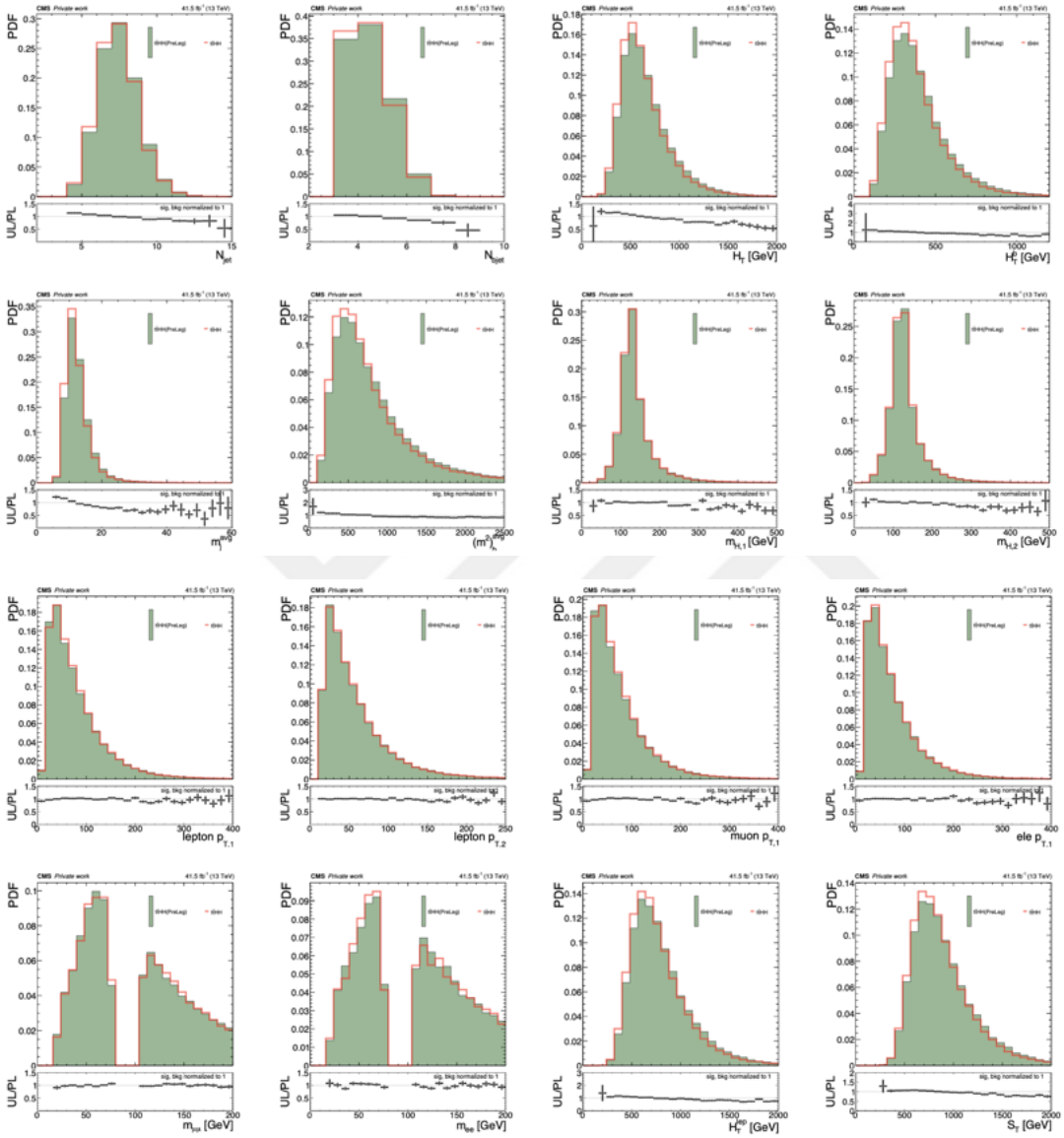

Figure A.1: The PreLegacy and UltraLegacy $t\bar{t}HH$ samples comparison. Distributions of different discrimating variables for jets and leptons after the baseline selection applied, normalized to unity.

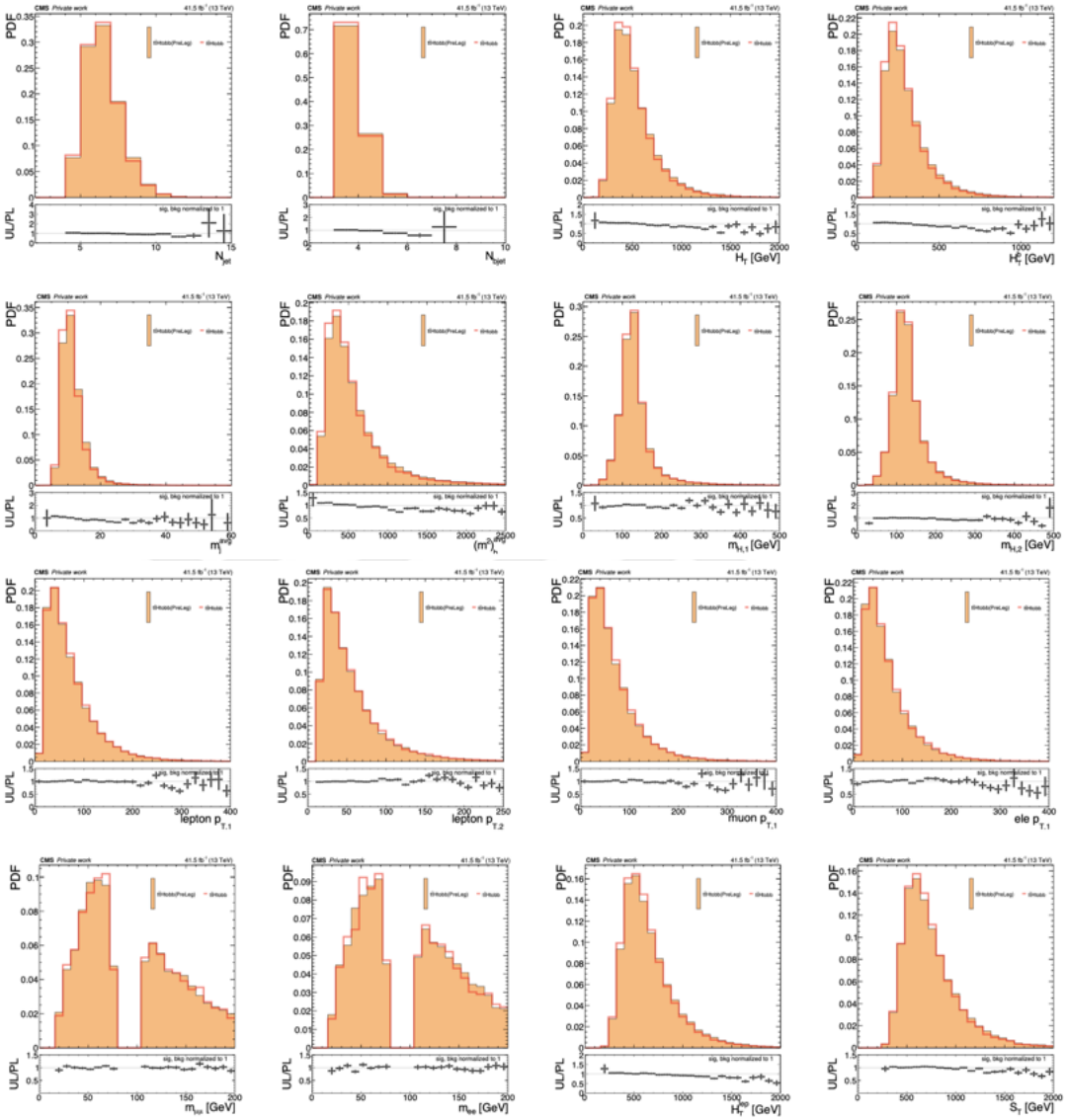

Figure A.2: The PreLegacy and UltraLegacy $\mathrm{t\bar{t}H} \rightarrow \mathrm{b\bar{b}}$ samples comparison. Distributions of different discrimating variables for jets and leptons after the baseline selection applied, normalized to unity.

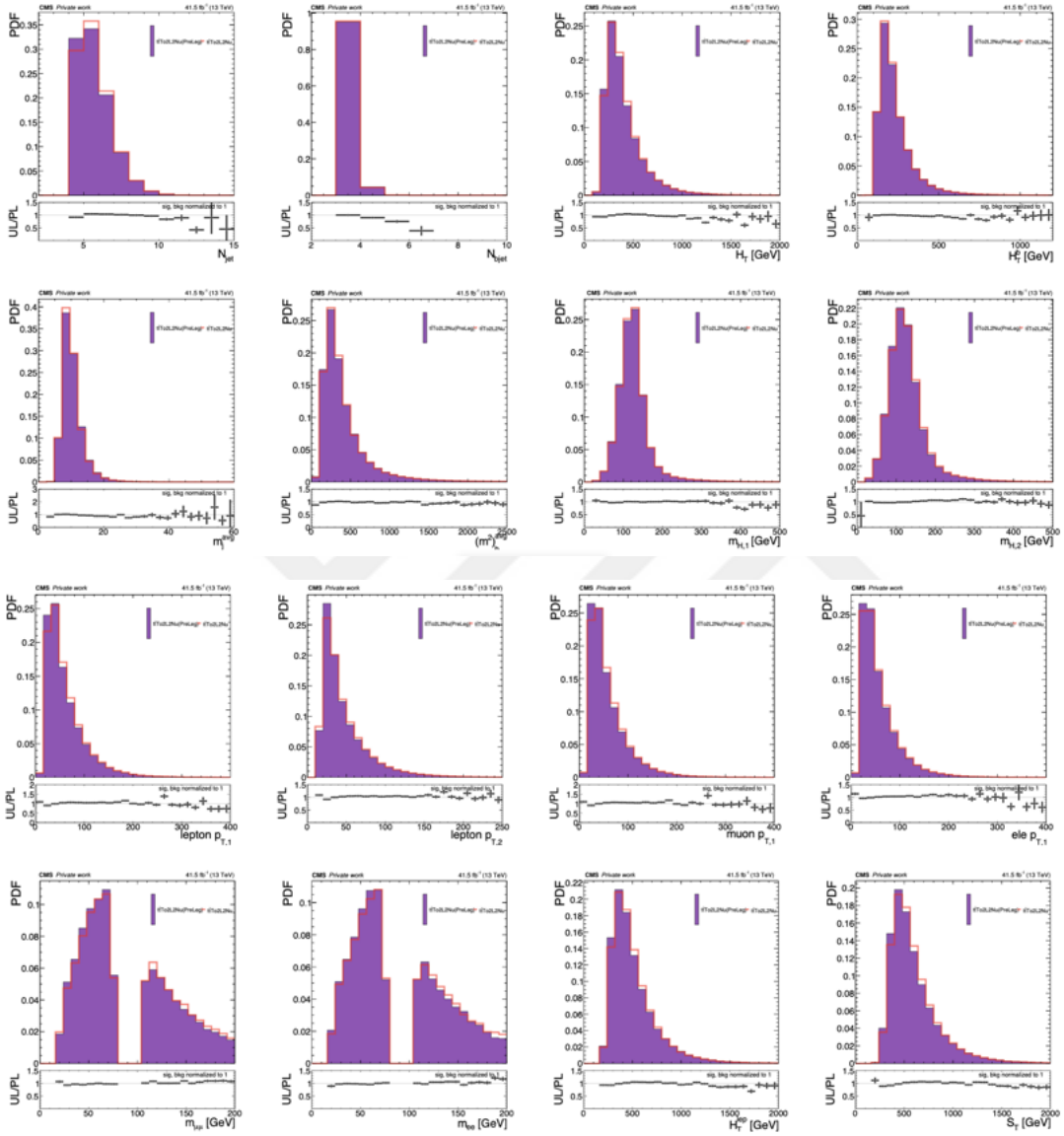

Figure A.3: The PreLegacy and UltraLegacy $t\bar{t}$DL samples comparison. Distributions of different discrimating variables for jets and leptons after the baseline selection applied, normalized to unity.

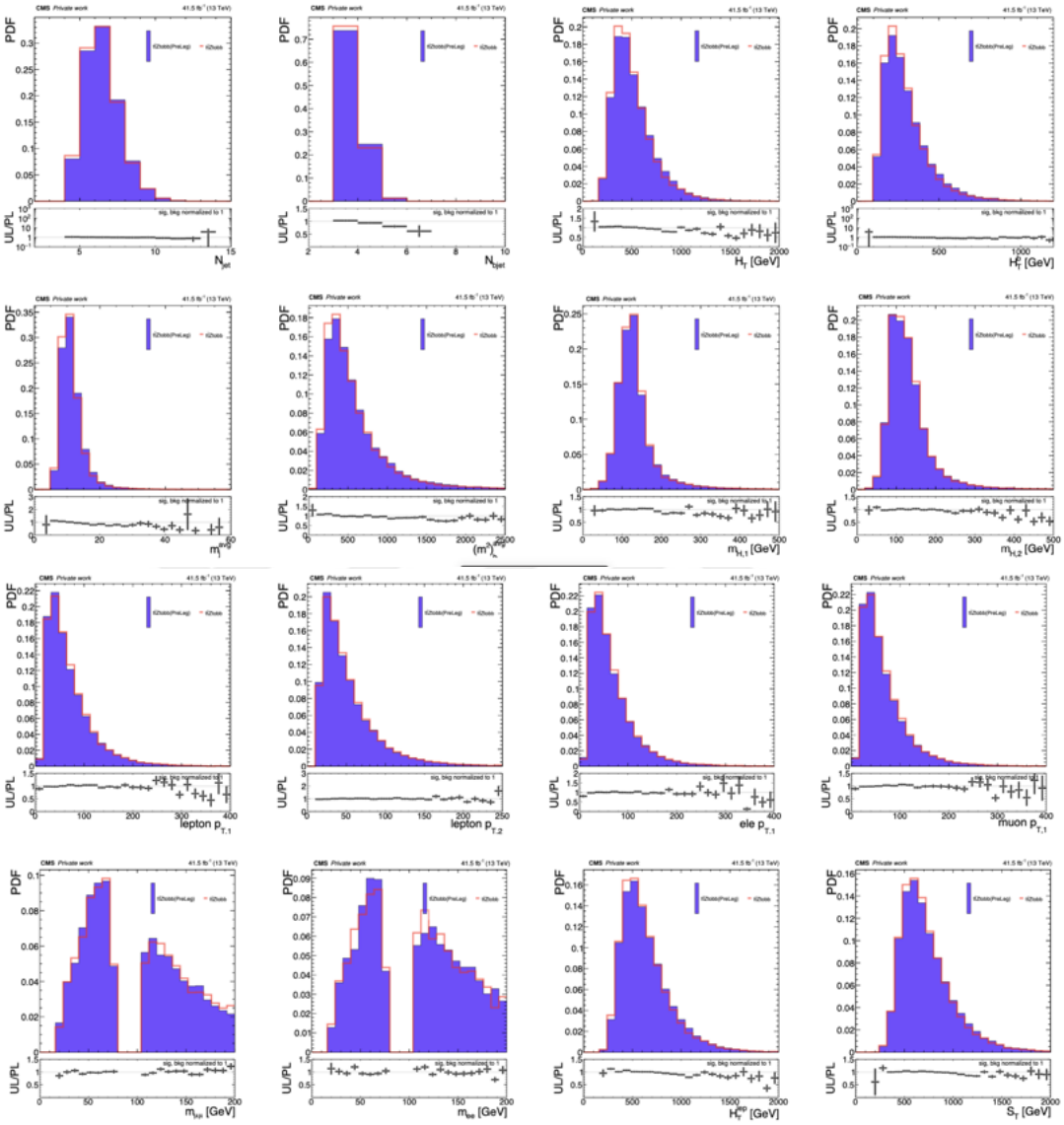

Figure A.4: The PreLegacy and UltraLegacy $t\bar{t}Z \to b\bar{b}$ samples comparison. Distributions of different discrimating variables for jets and leptons after the baseline selection applied, normalized to unity.

APPENDIX B

ELECTRON MVA WORKING POINT STUDY

In this section, the study being carried out to compare the performance of three distinct mva ID provided for electrons: mvaIDFall17V2_wp80, mvaIDFall17V2_wp90, and mvaIDFall17V2_wp98 is described. The results are presented in Tables B.1 to B.3. Each working point undergoes a baseline selection application, as outlined in Table 7.26, using dedicated test samples for both signal and background processes. After applying the baseline selection, events including exactly two leptons are counted. Subsequently, a further study is performed to identify the event yields in the remaining dataset containing at least one electron and those comprised entirely of electrons. The weighted event numbers are also provided for two cases. The weight calculation process aligns with the methodology defined in Section 7.2, however, in this context, the total luminosity of three years (137 fb^{-1}) is taken instead of individual year-specific luminosities. Moreover, the values highlighted in blue correspond to the percentage of event yields.

Table B.1: Comparison for WP80.

	$t\bar{t}HH$	$t\bar{t}SL$	$t\bar{t}DL$	$t\bar{t}Htobb$	$t\bar{t}Ztobb$	$t\bar{t}4b$	$t\bar{t}ZZ$	$t\bar{t}ZH$	total BKG	$S/\sqrt{B}$
# of events	9934000	87766000	106724000	7825000	7074000	9502000	4832000	5000000	228723000	656.85
after BS	59164	4080	197257	36764	30237	38226	28785	29696	365045	97.92
weighted #	0.2112	2326.58	22132.235	189.26	73.84	163.99	0.063	0.109	24886.077	0.0013
at least 1 ele	41869	2682	148403	26555	21774	27720	20429	20988	268551	80.79
weighted #	0.1485	1529.38	16650.82	136.71	53.17	118.92	0.026	0.08	18489.106	0.0011
event %	71	66	75	72	72	72	71	71		
dielectron	7500	214	28351	4876	3945	4988	3696	3857	49927	33.57
weighted #	0.027	122.034	3180.98	25.102	9.63	21.40	0.0046	0.014	3359.1646	0.0005
event %	13	5	14	13	13	13	13	13		

Table B.2: Comparison for WP90.

	$t\bar{t}HH$	$t\bar{t}SL$	$t\bar{t}DL$	$t\bar{t}Htobb$	$t\bar{t}Ztobb$	$t\bar{t}4b$	$t\bar{t}ZZ$	$t\bar{t}ZH$	total BKG	$S/\sqrt{B}$
# of events	9934000	87766000	106724000	7825000	7074000	9502000	4832000	5000000	228723000	656.85
after BS	68211	4939	227355	41472	34483	43591	33012	28102	412954	106.15
weighted #	0.2442	2816.42	25509.231	213.497	84.207	187.01	0.041	0.102	28810.51	0.0014
at least 1 ele	50916	3541	178501	31395	26020	33085	24656	20952	318150	90.27
weighted #	0.1815	2019.22	20027.81	161.62	63.541	141.93	0.032	0.076	22414.229	0.0012
event %	75	72	78	75	75	76	75	75		
dielectron	10346	414	38411	6492	5270	6718	5100	4401	66806	40.03
weighted #	0.0373	236.08	4309.71	33.421	12.87	28.82	0.0063	0.0165	4620.9238	0.00055
event %	13	5	14	13	13	13	13	13		

Table B.3: Comparison for WP98.

	$t\bar{t}HH$	$t\bar{t}SL$	$t\bar{t}DL$	$t\bar{t}Htobb$	$t\bar{t}Ztobb$	$t\bar{t}4b$	$t\bar{t}ZZ$	$t\bar{t}ZH$	total BKG	$S/\sqrt{B}$
# of events	9934000	87766000	106724000	7825000	7074000	9502000	4832000	5000000	228723000	656.85
after BS	82542	7721	261666	49602	40455	51088	39436	40993	490961	117.8
weighted #	0.297	4402.82	29358.93	255.35	98.79	219.167	0.049	0.15	34335.26	0.0016
at least 1 ele	65247	6323	212812	39393	31992	40582	31080	32285	394467	103.89
weighted #	0.2343	3605.63	23877.51	202.79	78.12	174.1	0.039	0.12	27938.3	0.0014
event %	80	82	81	80	80	80	79	79		
dielectron #	15501	1350	51154	9192	7332	9459	7406	7782	93675	50.65
weighted	0.0561	769.82	5739.478	47.32	17.90	40.58	0.0092	0.029	6615.14	0.0007
event %	19	17	20	19	18	19	19	19		

APPENDIX C

SIGNAL AND BACKGROUND COMPARISONS FOR INDIVIDUAL BACKGROUNDS FOR THE HL-LHC STUDY

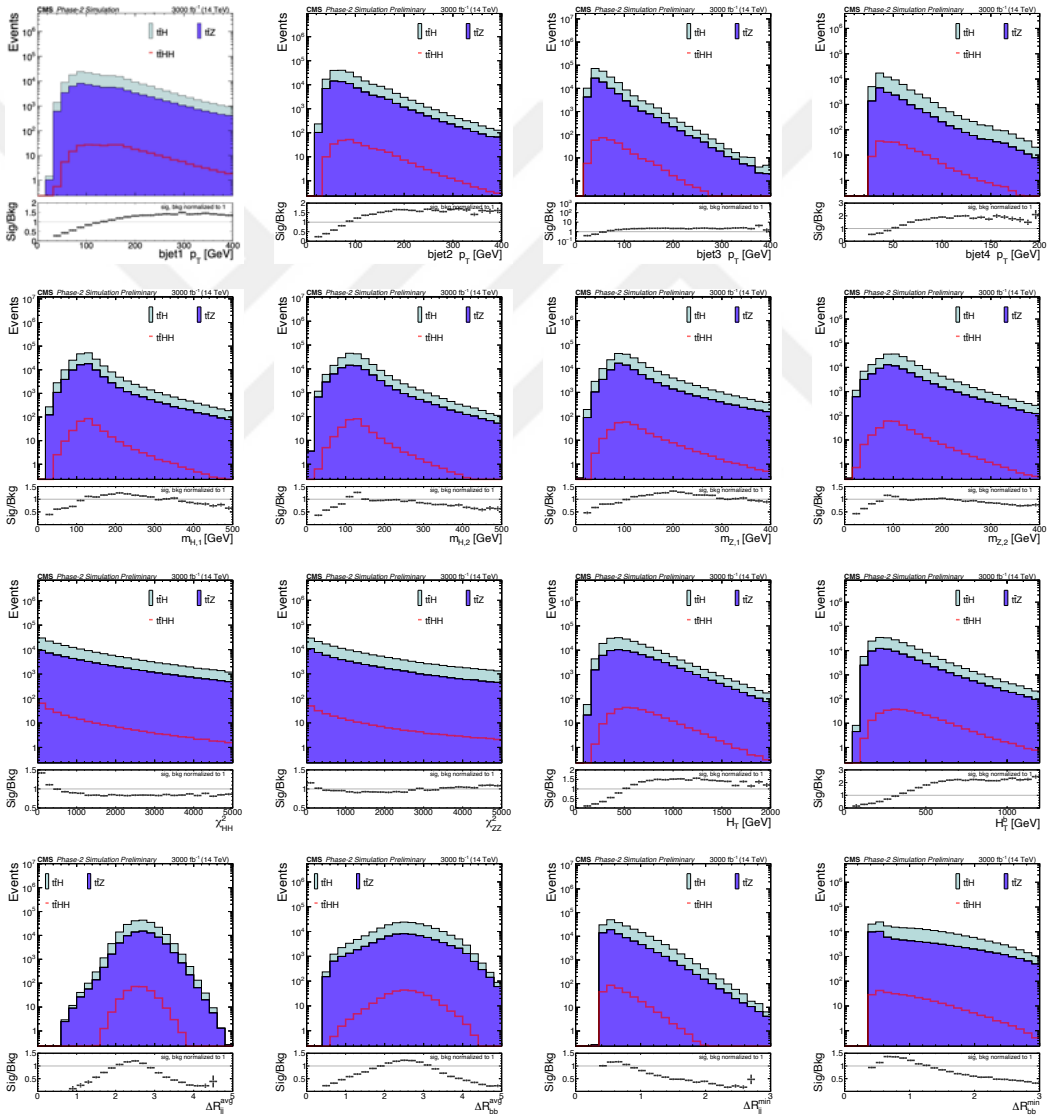

Figure C.1: Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $t\bar{t}H$ and $t\bar{t}Z$ backgrounds.

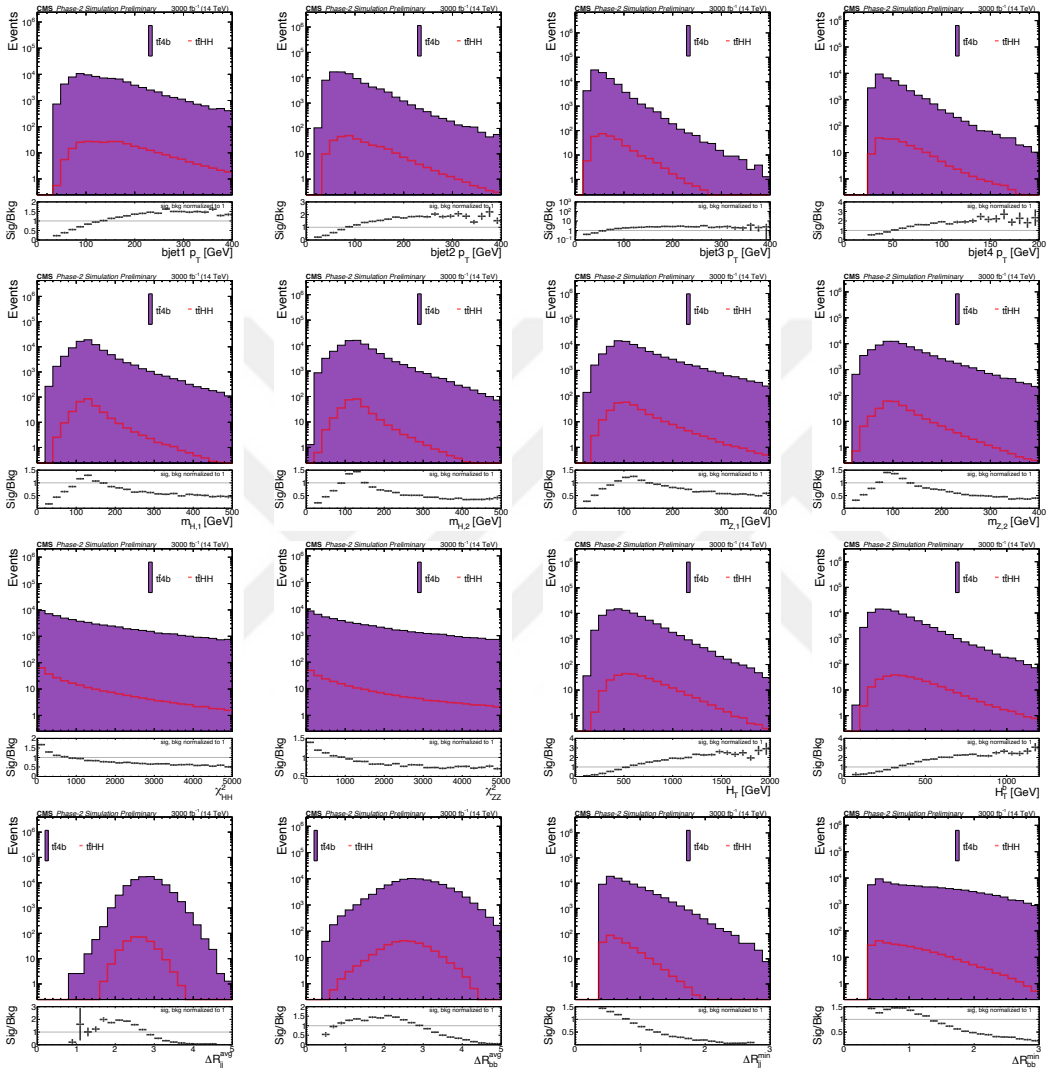

Figure C.2: Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $\mathrm{t\bar{t}4b}$ background.

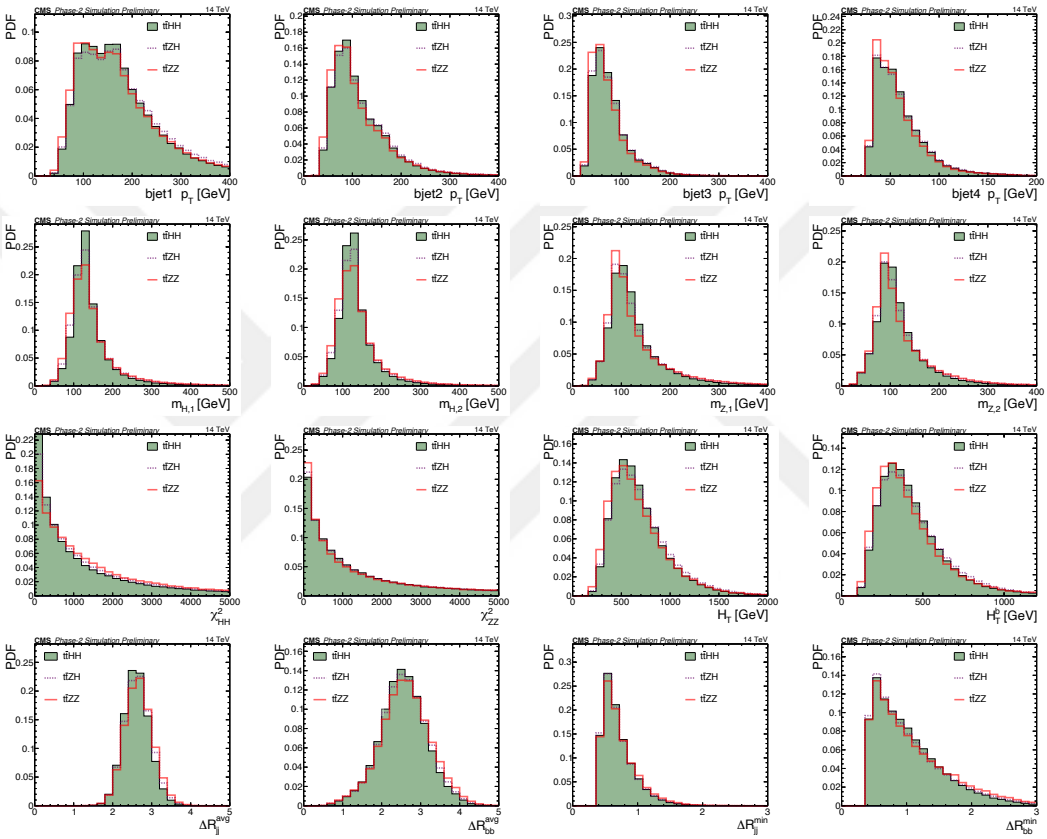

Figure C.3: Distributions of several kinematical event variables belonging to jets and b jets, comparing the signal in red and the $t\bar{t}ZZ$ and $t\bar{t}ZH$ backgrounds.

APPENDIX D

STATISTICAL METHODS

This analysis utilizes histograms with distinct content in each bin, considered as a counting experiment with an expected yield of events ν. Before analyzing the observed data in a search, the experiment forecasts yields for both the signal (S) and background (B) events, which are influenced by various experimental and theoretical uncertainties. Experimental uncertainties are figured out using specific measurements, while theoretical uncertainties are calculated by changing the model's parameters. Since these uncertainties affect the number of expected events, they are included in calculations as nuisance parameters, represented by θ. This way, both the signal and background predictions are adjusted to account for these uncertainties, expressed as S(θ) and B(θ).

In particle physics, how observables, i.e. measurable quantities, are distributed across a range of value can be determined by using a mathematical function called probability density function (*pdf*) and describing the likelihood of an observable taking on a certain value. The systematic error *pdf*, denoted as $\rho(\theta|\tilde{\theta})$, where $\tilde{\theta}$ represents the default value of the nuisance parameter, indicate the level of confidence in the true value of θ. These systematic error *pdf*s can be reinterpreted as posterior distributions derived from real or hypothetical measurements $\tilde{\theta}$, as outlined by Bayes' theorem, which is expressed as $\rho(\theta|\tilde{\theta}) \cdot \mathrm{p}(\tilde{\theta}|\theta) \cdot \pi_\theta(\theta)$. Here, $\pi_\theta(\theta)$ are hyper-prior functions for those measurements. The *pdf* $p(\tilde{\theta}|\theta)$ for the auxiliary measurement can help to constraint the likelihood of the primary measurement in a frequentist approach. The *pdf* of the observed data, as a function of a set of parameters, is the likelihood function

$$\mathcal{L}(data|\mu, \theta) = \text{Poisson}(\text{data}|\nu) \cdot p(\tilde{\theta}|\theta) \tag{D.1}$$

with the signal strength modifier μ. A more straightforward understanding of event yields for the problem of interest can be provided such that $\nu = \mu \mathrm{S}(\theta) + \mathrm{B}(\theta)$.

The final likelihood function can be adapted for scenarios with multiple independent bins, N in total. This adaptation involves taking the product of the individual likelihood functions for each bin. In each of these bins, n_i events and expect ν events are observed. Therefore, the overall likelihood function is a product of the likelihoods from each bin's counting experiment such that

$$\mathcal{L}(data|\nu) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} \tag{D.2}$$

Searching for a new signal model requires defining a null hypothesis, labeled as H_0 and representing *the known process*, together with an alternative hypothesis, denoted as H_1 and representing the model under investigation. In HEP, H_0 represents background events, whereas H_1 includes both signal and backgrounds events. To assess how well a collected data, or a hypothesis under investigation, matches H_0, one can determine a p-value. Ranging from 0 to 1, it is a statistical measure that evaluates the probability of obtaining results at least as extreme as the observed ones, under the assumption that H_0 is correct. A p-value close to 0 suggests that the difference observed is unlikely due to chance, indicating a potential discrepancy with H_0. Conversely, a p-value near 1 implies that any difference is likely due to chance, supporting H_0. A p-value is often compared to a predefined threshold, called significance level, represented as α and mostly set at 0.05 in particle physics. If the p-value is less than this level, the result is considered statistically significant, leading to the rejection of hypothesis H_0.

The μ value serves as a measure of the strength of any potential new signal. It can assume any value, with μ equaling 0 indicating the absence of new physics, while μ equaling 1 aligns with the theoretical prediction of new physics presence. The estimation of μ is crucial and is accomplished using the Maximum Likelihood Method (MLM), which assesses how well the data corresponds to the hypothesis of new physics. To test these hypotheses, a test statistic, often based on the likelihood ratio, is constructed:

$$\tilde{q_\mu} = -2\ln\frac{\mathcal{L}(data|\mu, \hat{\theta_\mu})}{\mathcal{L}(data|\hat{\mu}, \hat{\theta})}, \tag{D.3}$$

subject to the constraint $0 \leq \hat{\mu} \leq \mu$. In this equation, "data" refers to the actual experimental observation or pseudo-data (toys), and $\hat{\theta_\mu}$ represents the conditional maximum likelihood estimators of θ given the signal strength parameter μ. The pair of parameter estimators $\hat{\mu}$ and $\hat{\theta}$ maximize the likelihood. The observed value of the test statistic $\tilde{q}_\mu^{obs}$ can be calculated as a function of μ allowing for the determination of nuisance parameters $\hat{\theta}_0^{obs}$ and $\hat{\theta}_\mu^{obs}$ corresponding to the best description of the observed data for the H_0 and H_1 hypotheses, respectively.

Due to the limitations for repeating the experiment in the frequentist approach, *pdf*s are estimated through toy MC generation, denoted as $f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu, \mu^{obs})$ and $f(\tilde{q}_\mu|0, \hat{\theta}_0^{obs})$. Using these *pdf*s, the p-values p_b and p_μ for the H_0 and H_1 hypotheses are defined as follows:

$$\begin{aligned} 1 - p_b &= \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu|0, \tilde{\theta}_0^{obs}) d\tilde{q}_\mu \quad , \\ p_\mu &= \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu|\mu, \tilde{\theta}_\mu^{obs}) d\tilde{q}_\mu \quad . \end{aligned} \tag{D.4}$$

A more robust strategy for addressing signal processes that occur infrequently compared to background processes involves the CL_S limit, formulated as

$$CL_S = \frac{p_\mu}{1 - p_b}. \tag{D.5}$$

Despite being dimensionless, this metric provides an unbiased confidence level (CL) on the H_0. The decision to accept or reject a hypothesis is parallel with the α value defined earlier. If $\mathrm{CL_S} \leq \alpha$ for $\mu = 1$, the signal is excluded with a confidence level of $(1-\alpha)\mathrm{CL_S}$. In accordance with the common practice in HEP, a 95% CL is required to consider a signal model to be excluded, denoted by $\mathrm{CL_S}$= 0.05 for this thesis.

Before getting the signal strength with the real data, it is common to use a representa-

tive dataset, called Asimov data to estimate the sensitivity of an experiment. It helps in understanding the behavior of statistical methods without the random fluctuations presenting in real data.

The methods described above is performed with a tool called Combine Tool developed for the statistical analysis within the CMS experiment [173].

Additional information and in-depth explanations for each concept mentioned in this section are available in the referenced materials [174–177].

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Sökmen Şahin, Gamze
Nationality: Turkish

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	METU	2024
M.Sc.	METU	2017
B.Sc.	METU	2014
High School	Incesu Anadolu High School, Ankara	2008

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2014-Present	CERN & METU	Project researcher

FOREIGN LANGUAGES

Advanced English, Beginner French

PUBLICATIONS

- CMS Collaboration. "Search for the non-resonant $t\bar{t}HH$ in the semileptonic decay of the top pair and the Higgs pair decay into $b\bar{b}$ at the HL-LHC." CMS FTR Note, CMS-PAS-FTR-21-010. CERN (2022).
  Retrieved from http://cds.cern.ch/record/2804085?ln=en.
- ATLAS and CMS Collaborations. "Snowmass White Paper Contribution: Physics with the Phase-2 ATLAS and CMS Detectors." ATL-PHYS-PUB-2022-018, CMS-PAS-FTR-21-010. CERN, Geneva (2022).
  Retrieved from https://cds.cern.ch/record/2805993.
- Alexandru Geanta, Andrei et al. "The ECFA Early Career Researcher's Panel: composition, structure, and activities, 2021 - 2022." arXiv:2212.11238 (2022).
- Aggarwal, Anamika et al. "Results of the 2021 ECFA Early-Career Researcher Survey on Training in Instrumentation." arXiv:2107.05739 (2021).
- CMS Author in the CMS Collobaration publications.