TÉCNICO LISBOA

Study of $t\bar{t}H$ production with $H \to b\bar{b}$ in ATLAS at the HL-LHC

António Manuel Mendes Jacques da Costa

Thesis to obtain the Master of Science Degree in

Engineering Physics

Supervisors: Dr. José Ricardo Morais Silva Gonçalo
Prof. Pedro Morais Salgueiro Teixeira de Abreu

Examination Committee

Chairperson: Prof. Jorge Manuel Rodrigues Crispim Romão
Supervisor: Dr. José Ricardo Morais Silva Gonçalo
Members of the Committee: Prof. António Joaquim Onofre Abreu Ribeiro Gonçalves
Prof. Michele Gallinaro

November 2018

Acknowledgments

I would like to warmly thank Dr. Ricardo Gonçalo for his support and invaluable advices and discussions, without which this work would not have been possible. Being supervised by Ricardo was both a pleasure and a privilege. My gratitude is extended to Prof. Patricia Muiño for the enthusiastic discussion, and to Prof. António Onofre and Prof. Pedro Abreu, for all the help provided.

I would also like to thank Ana Luísa Carvalho, Aidan Kelly, Bruno Alves and Ricardo Barrué, for a most enjoyable working atmosphere and support. I also thank Aidan for the generation of the $b\bar{b}j$ samples. Additionally, special thanks go to Prof. Liliana Apolinário, for the generation of the dijets samples and supplementary advices, and to João Martins, for the configuration of the LIP's machines. A special thanks also goes to Prof. Pedro Ferreira, for reviewing my theoretical chapter, and for the interesting literature discussions. Thanks also to Emanuel Gouveia and Duarte Azevedo, for introducing me to new software tools.

Working at LIP was very rewarding. It started in the first years of my degree and culminated in this master thesis. I got to know interesting and motivated people in these years, that helped me throughout this path, and for that I am grateful to all of them. This work was partially supported by Fundação para a Ciência e Tecnologia, FCT (Projects No. CERN/FIS-PAR/0008/2017).

I also thank the Boosted ttH(bb), ttH(bb) and Higgs Prospects conveners at CERN, for giving me the opportunity to present my work at the meetings and for the instructive discussions. I would also like to thank Clement Helsens and Michele Selvaggi for the help with the DELPHES framework.

I am also greatly thankful to Duarte Drumond, Gonçalo Castro, João Ferreira, Miguel Gonçalves, Pedro David, Ana Jorge, Beatriz Belbut and Margarida Cordeiro, for their patience and support throughout my thesis. My gratitude is extended to my cousin Nuno Camarneiro and my high school Physics teacher Carlos Portela, for raising my interest in Physics, and namely in Particle Physics.

Finally, I must express my very profound gratitude to my parents and sister, for their love and everything they have done for me. Thank you for supporting me and my choices, even if it means that I will be far away from home.

Resumo

Esta tese apresenta uma estratégia alternativa para analisar experimentalmente o processo $t\bar{t}H(H \to b\bar{b})$ no LHC e no futuro regime de alta luminosidade. Ao contrário de estratégias correntemente implementadas em colaborações experimentais, esta análise explora técnicas de substructura de jactos hadrónicos e foca-se na reconstrução de bosões de Higgs com alto momento, de forma a obter sensibilidade de sinal com uma análise baseada em cortes. O fundo de $t\bar{t}$+jactos pode ser constrangido na análise proposta através de uma região de controlo com muito pequena contaminação de sinal. Usando esta estratégia de análise, o processo $t\bar{t}H(H \to b\bar{b})$ poderá ser observado no LHC, singularmente no canal semi-leptónico, tendo uma significância associada de 5.41 ± 0.12 para uma luminosidade integrada de $300\,\text{fb}^{-1}$. Esta mesma luminosidade integrada está associada a uma significância de 6.13 ± 0.11, no HL-LHC com um detector melhorado. É esperada uma incerteza de 18% na força de sinal do processo $t\bar{t}H$ no LHC, com uma luminosidade integrada de $300\,\text{fb}^{-1}$, que se reduz para 5% no HL-LHC com uma luminosidade integrada de $3000\,\text{fb}^{-1}$. Adicionalmente, o acoplamento do bosão de Higgs ao quark top é esperado poder ser medido com uma incerteza de 35$\%$ no fim das operações do LHC, utilizando a estratégia proposta com uma luminosidade integrada de $300\,\text{fb}^{-1}$. Esta incerteza diminui para 17$\%$ no caso do HL-LHC, com uma luminosidade integrada de $3000\,\text{fb}^{-1}$. É possível ainda implementar a estratégia usando jactos re-clustered, sem perdas de eficiência.

Palavras-chave: Bosão de Higgs, Yukawa, Força de Sinal, Quarks Top, HL-LHC, Subestrutura de Jactos

Abstract

A feasibility study for an experimental analysis searching for $t\bar{t}H(H \to b\bar{b})$ production at the LHC and its high luminosity phase is presented in this thesis. Unlike search strategies currently being used in experimental collaborations, the present analysis exploits jet substructure techniques and focuses on the reconstruction of boosted Higgs bosons, to obtain sensitivity to the signal in a simple cut-based analysis. The $t\bar{t}$+jets background may be constrained in the proposed analysis through a control region with very small signal contamination. Using this analysis strategy, the $t\bar{t}H(H \to b\bar{b})$ process could be observed at the LHC, in the semi-leptonic channel alone, with a significance of 5.41 ± 0.12 for an integrated luminosity of $300\,\text{fb}^{-1}$. For the same integrated luminosity, in the High Luminosity LHC scenario with an upgraded detector, a significance of 6.13 ± 0.11 may be obtained. The top Yukawa coupling could be measured with a 35% uncertainty using an integrated luminosity of $300\,\text{fb}^{-1}$ of LHC data and of 17% at the HL-LHC scenario with an integrated luminosity of $3000\,\text{fb}^{-1}$. In the same luminosity scenarios, the signal strength is equally expected to have a 18% and 5% uncertainty, respectively. It was also found that re-clustered jets may be used without loss of efficiency.

Keywords: Higgs Boson, Yukawa, Signal Strength, Top Quarks, HL-LHC, jet substructure

Contents

List of Tables

List of Figures

Glossary

2HDM	Two Higgs-Doublets Models
ATLAS	A Toroidal LHC Apparatus
BDRS	Butterworth-Davison-Rubin-Salam
BDT	Boosted Decision Tree
BSM	Beyond Standard Model
C/A	Cambridge-Aachen
C2HDM	Complex Two Higgs-Doublets Models
CERN	European Organization for Nuclear Research
CMS	Compact Muon Solenoid
CP	Charge-Parity
DAQ	Data Acquisition
DNN	Deep Neural Network
ECAL	Electromagnetic Calorimeter
ECF	Energy Correlation Functions
EF	Event Filter
HCAL	Hadronic Calorimeter
HL-LHC	High Luminosity Large Hadron Collider
HLT	High Level Trigger
HTT	Hardware-based Tracking for the Trigger
Hz	Hertz
ID	Inner Detector
L0	Level 0
L1	Level 1
LHC	Large Hadron Collider
LO	Leading Order
LS3	Long Shutdown 3
MC	Monte Carlo
MEM	Matrix Element Method
ML	Multilepton
MVA	Multivariate Analysis

NLO	Next-to-Leading Order
NNLL	Next-to-Next-to-Leading Log
NNLO	Next-to-Next-to-Leading Order
PDF	Parton Distribution Function
PF	Particle-Flow
PV	Primary Vertex
QCD	Quantum Chromodynamics
SCT	Semiconductor Tracker
SM	Standard Model
SV	Secondary Vertex
TRT	Transition Radiation Tracker
eV	Electron Volt

Chapter 1

Introduction

Milan Kundera, in the Unbearable Lightness of Being, wrote that the most important questions are the most naive ones. In fact, one of the most innocent and relevant questions one can think of concerns the composition of our own bodies. And while the quickest answer seems trivial, resulting of our observation of the world, the complete and real answer is much more complex and hard to get. Indeed, we are not simply just made of flesh and bone but instead of smaller and simpler constituents, known as atoms. Atoms consist of nuclei with electrons orbiting around each one, and the nuclei are then composed of neutrons and protons. By studying these it was discovered that they are made of quarks and gluons, that along with the electrons, constitute some of the so-called elementary particles.

The whole collection of elementary particles then ranges from massless particles to a particularly massive one, the top quark - interesting due to the great disparity between its mass and the masses of the others massive particles. On the other hand, these particles interact with each other, and with the goal of describing Nature as precisely as possible, the Standard Model of Particle Physics was developed by theorists and experimentalists throughout the last decades. An essential piece of the Standard Model is the Higgs Mechanism, without which the origin of the mass of some particles would be left unexplained. Strongly linked to this mechanism is the Higgs boson, discovered in 2012 at CERN, the European Organization for Nuclear Research.

This was one of the major results of the Large Hadron Collider (LHC), the world's largest and most powerful particle accelerator. In it, particles, as protons, are accelerated to close to the speed of light, and collide in four points, one of which is the position for the ATLAS detector. This detector, along the CMS one, are generalist experiments, as they are used to perform searches for new particles and precision measures of the Standard Model, and therefore checking its validity.

One of these analyses is focused on the production of the Higgs boson in association with two top quarks, or $t\bar{t}H$, a particularly relevant process as it allows for a direct measurement of the coupling of the Higgs to the top. By precisely measuring this coupling it is then possible to attest the Standard Model or to find evidence for new physics, as if these two particles don't interact as predicted, other explanations are needed. The motivation for new physics theories comes from the fact that, whilst the Standard Model is able to explain most of what we see in Nature, it yet fails to provide an answer to why there

is so much more matter than anti-matter in the universe, where antimatter are particles with opposite physical charges, among other unanswered questions, such as the hierarchy of particle masses or the origin of dark matter. The solution for this could actually be in the Higgs sector, and its coupling to the top can be used as a probe of possible new physics.

Unfortunately, the $t\bar{t}H$ production is a rare process, with low statistics associated, and this contributes to the difficulty of having a precise measure of the coupling. The study of rare processes though, among other physics goals, is the reason for a future upgrade of the LHC, resulting in the High Luminosity LHC (HL-LHC). With the future collider and its new features the number of collisions per second will then increase, and more data will be collected, allowing further detailed studies.

Adding to the rarity of the process it comes the fact that, in the collisions at the LHC, not only the process of interest for an analysis is produced. In fact, more collisions happen at the same time, generating many other particles. The processes are then divided in signal, concerning the relevant process under study, and backgrounds, for all the processes that mimic the signal.

In this thesis a study for the $t\bar{t}H$ process is presented, with the Higgs decaying into two bottom quarks, one of the tops decaying into quarks, and the other decaying to a lepton, a neutrino and a bottom quark (semileptonic channel). In this case the main irreducible background is $t\bar{t}b\bar{b}$, and the uncertainties on its modelling are one of the dominant causes for not having a precise result on this channel. The goal of this work is therefore to present an alternative strategy for experimentally analysing this channel, using the High Luminosity LHC framework, in such a way that these adversities are overcome. It uses as key features characteristics of the event, as hadronic jet substructure information, and identifies as efficiently as possible the Higgs in the event, in order to reconstruct its mass and have a clear peak around its nominal mass value, against continuous distributions for the backgrounds. As a result of this work the amount of collected statistics (luminosity) needed for the observation of the $t\bar{t}H$ process in the semileptonic channel is obtained, as well as the uncertainty on the Higgs coupling to the top quark by the end of the LHC and HL-LHC.

For an adequate contextualization of the thesis subjects, the Standard Model is initially explained in the first chapter of the thesis, along with the process under study and a scenario of new physics.

The second chapter intends to describe the experimental apparatus used to perform particle physics analyses, covering the LHC and its future upgrade to the HL-LHC, and focusing on the ATLAS detector, with an explanation of each of its constituents and their respective use when detecting particles.

The third chapter covers the analysis tools required to perform this study, from the simulation frameworks used for the event generation, to the algorithms and variables that will be used throughout the work, namely to identify the Higgs boson.

In the fourth chapter the current state of the art regarding the study of $t\bar{t}H$ process is presented, supported by experimental and theoretical studies, and with different strategies being proposed. One of these served as a preliminary basis for this work, being designed to identify the tops and the Higgs in the events.

The fifth chapter is then dedicated to the implementation of this strategy, adapted to the HL-LHC, and this is followed by the optimization of the strategy, where the top identification is dropped and the Higgs

reconstruction is also improved. Furthermore, a different implementation scheme is presented, along with a proposed region to control the backgrounds. The optimized strategy is in addition implemented for the current LHC framework, in order to check for differences with respect to the High Luminosity scenario.

The sixth chapter contemplates the results of the strategy proposed in this thesis, stating the minimum integrated luminosities necessary to observe the $t\bar{t}H$ process, as well as the uncertainties on the coupling for the LHC and HL-LHC. With these results, the goals of this work are accomplished.

Finally, in the last chapter of this thesis, the conclusions regarding this work are drawn, along with some suggestions for future improvements.

Chapter 2

High Energy Physics

The Standard Model (SM) encodes our current understanding of the elementary particles present in Nature, and of their interactions. It grew and took its present form in the second half of the last century, from a growing body of theoretical and experimental results. In addition to elementary matter particles and interactions, it includes the Higgs mechanism, which occupies a central role in the SM and in the generation of the masses of some particles.

This chapter therefore starts by explaining the Standard Model in the first section, covering its particles and describing the Higgs mechanism, and introducing the coupling of the Higgs to the top quark. In the second chapter the study of the $t\bar{t}H$ process is motivated and presented for different final states, with the associated advantages and drawbacks. Finally, a brief discussion of physics beyond the Standard Model is given, along with some theories that could be probed through the $t\bar{t}H$ process.

2.1 Standard Model of Particle Physics

The Standard Model of Particle Physics (SM) is the result of tremendous theoretical and experimental efforts for several decades and it describes the strong, electromagnetic and weak forces through gauge field theories. This description uses a quantum field theory, where each type of particle is described by a different field. Fermion interactions, for instance, are in this context mediated by gauge bosons, and also by the Higgs boson (Figure 2.1). Each particle is characterised by a mass, spin and charge. Bosons (particles of integer spin) are classified as scalar if they have spin 0, and gauge/vector bosons if they have spin 1, while fermions are particles with half-integer spin.

Fermions are divided in two categories: six quarks and six leptons. There are three families or generations of quarks (composed of an ”up-type” and ”down-type” quark), the only difference between each family and the next one being the increasing mass of the particles in each family. Likewise for the leptons, each family containing a charged lepton and a neutrino, the mass of the charged leptons (not necessarily that of the neutrinos) growing from one family to the next. Moreover, while quarks are charged under the strong, electromagnetic and weak interactions, leptons don’t feel the effect of the strong force, and the neutrinos, specifically, also don’t interact electromagnetically. In addition, gauge

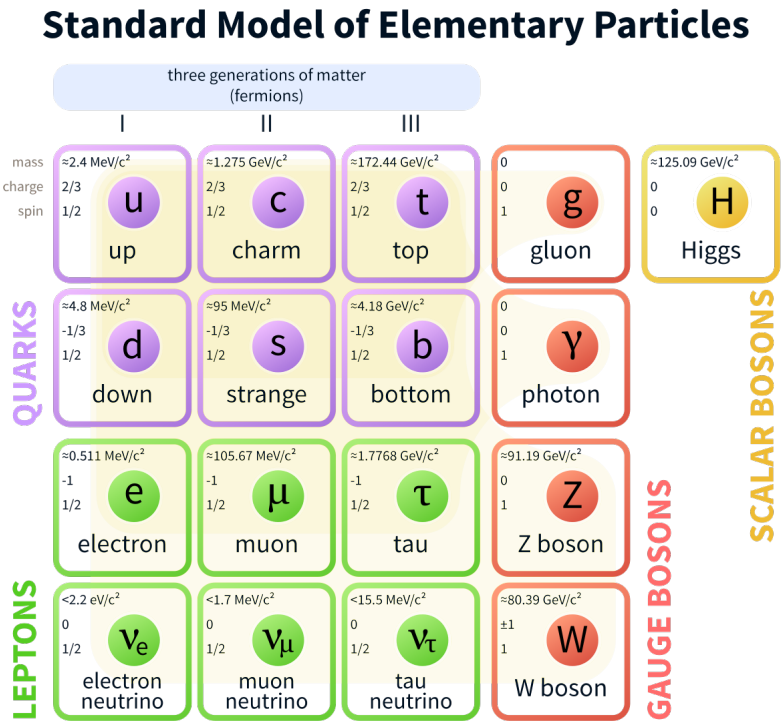

Figure 2.1: The Standard Model particles [1].

bosons are the spin 1 carriers of the interactions: gluons are responsible for the strong force, photons for the electromagnetic force and the W^+, W^- and Z bosons for the weak force [2].

In the SM, gluons and photons are massless while the Z, W^+ and W^- bosons, quarks and charged leptons have a mass. Neutrinos were experimentally demonstrated to have a non-zero mass, leading to the 2015 Nobel Prize in Physics. The origin of their mass is the object of much research at the moment. The other masses, however, are generated through the Higgs mechanism that also introduces a scalar boson, the Higgs boson [3–5]. The discovery of the Higgs boson was, therefore, of the utmost importance. In 2012, the ATLAS and CMS experiments at CERN observed a new particle consistent with the Higgs boson [6, 7].

In order to better understand this mechanism, it is useful to look to the Standard Model lagrangian,

$$\mathcal{L} = -\frac{1}{4}G^k_{\mu\nu}G^{k\mu\nu} - \frac{1}{4}W^a_{\mu\nu}W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} + i\overline{\psi}\not{D}\psi + \overline{\psi}_{L_i}y_{ij}\psi_{R_j}\phi + |D_\mu\phi|^2 - V(\phi) \qquad (2.1)$$

In this formulation, the first three terms include the $SU(3)_C$, $SU(2)_L$ and $U(1)_Y$ gauge bosons self-interactions, respectively, while the fourth encodes the interactions of gauge bosons with fermions, and the propagators of these latter particles. The fifth term accounts for the couplings of the fermions to the Brout–Englert–Higgs field (Higgs field for simplicity, from now on). The sixth term then represents the interactions of the gauge bosons to the Higgs field, as well as its propagators. Finally, the last term stands for the Higgs potential.

In this lagrangian, ψ represents the fermions spinor fields, and ϕ is a scalar Higgs doublet, while y_{ij} stands for the Yukawa couplings between fermions and the Higgs. The indices L and R stand for the

left-handed and right-handed chiralities, respectively. Moreover, $G^k_{\mu\nu}$ is the $SU(3)_C$ field-strength tensor, $W^a_{\mu\nu}$ is the $SU(2)_L$ corresponding tensor, while for the $U(1)_Y$ group the field-strength tensor is $B_{\mu\nu}$. The $SU(3)_C$, $SU(2)_L$ and $U(1)_Y$ field-strength tensors are defined, respectively, as

$$G^k_{\mu\nu} = \partial_\mu G^k_\nu - \partial_\nu G^k_\mu + g_s f^{kjl} G^j_\mu G^l_\nu \tag{2.2}$$

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g\epsilon^{abc} W^b_\mu W^c_\nu \tag{2.3}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{2.4}$$

In these equations, f^{kjl} are the group structure constants and ϵ^{abc} is the completely anti-symmetric tensor in 3 dimensions. Additionally, B_μ is the $U(1)_Y$ gauge boson, while W^a_μ are the $SU(2)_L$ gauge bosons. G^k_μ represent the eight gluon fields of the $SU(3)_C$ group. On the other hand. *g*, *g'* and g_s are the $SU(2)_L$, $U(1)_Y$ and $SU(3)_C$ coupling constants, respectively.

Furthermore, in the SM lagrangian, D_μ is the covariant derivative, defined as

$$D_\mu = \partial_\mu - ig_s t_k G^k_\mu - igT_a W^a_\mu - ig'\frac{Y}{2}B_\mu \tag{2.5}$$

and $\not{D} = \gamma^\mu D_\mu$. In Equation 2.5, Y is the hypercharge (the $U(1)_Y$ generator), and is defined as $Q = I_3 + \frac{Y}{2}$, where Q is the electric charge and I_3 is an additional quantum number, isospin. Isospin is a quantum number associated to group representations. Namely, isospin doublets have $I_3 = 1/2$, while singlets have $I_3 = 1$. Additionally, *a* runs from 1 to 3, and $T_a = \frac{1}{2}\tau_a$, where τ_a are the Pauli matrices and the generators of the $SU(2)_L$ group. On the other hand, $t_k = \frac{1}{2}\lambda_k$ are the $SU(3)_C$ generators (*k* runs from 1 to 8), where λ_k are the Gell-Mann matrices.

In fact, the SM is ruled by the $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge theory, and this imposes some constraints on the terms allowed in the lagrangian. Namely, the $SU(2)_L \times U(1)_Y$ symmetry forbids mass terms for the gauge bosons and the fermions, as they would not be invariant under the respective gauge transformations. In the boson case, due to the vector field transformations, mass terms such as $m^2 A_\mu A^\mu$, where A_μ is the photon field, are prohibited. The fermion case is related to the fact that the weak $SU(2)_L$ interaction only transforms particles with left-handed chiral component, and leaves the right-handed chiral particles unchanged, and therefore $m\overline{\psi}\psi$ terms are not allowed. In fact, this term can be decomposed in left and right chiralities,

$$m\overline{\psi}\psi = m(\overline{\psi}_L\psi_R + \overline{\psi}_R\psi_L) \tag{2.6}$$

and since ψ_L is an isospin doublet and ψ_R is an isospin singlet, the product of these two fields is not a number. Indeed, a scalar doublet is required.

Nevertheless, while the full lagrangian is invariant under the $SU(3)_C \times SU(2)_L \times U(1)_Y$, the vacuum is only $SU(3)_C \times U(1)_{EM}$ (EM stands for electromagnetic) invariant, due to the spontaneous symmetry breaking, resulting from the Higgs mechanism. This mechanism starts by firstly introducing the smallest fundamental $SU(2)_L$ scalar Higgs doublet ϕ, with hypercharge $Y = +1$, defined as

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.7}$$

Then, it is necessary to add the Higgs potential, generally expressed as

$$V(\phi) = \mu^2|\phi|^2 + \lambda|\phi|^4 \tag{2.8}$$

where μ and λ are almost free parameters. It should be noted that λ must be positive, in order to have vacuum stability, that is, that there exists an absolute minimum of the potential. On the other hand, μ^2 can be positive or negative.

If $\mu^2 > 0$, the vacuum has a single minimum, at $\phi = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and this represents a trivial case, as there is no symmetry breaking. Contrary to this scenario, with $\mu^2 < 0$, there is an infinite number of vacua that satisfy the condition

$$\frac{dV}{d\phi} = 0 \Leftrightarrow |\phi| = \sqrt{\frac{-\mu^2}{\lambda}} = v \tag{2.9}$$

and for this case, the vacuum then has a value different from zero, the so-called vacuum expected value (VEV or v).

In this case, the vacuum configuration with $\phi_1 = \phi_2 = \phi_4 = 0$ and $\phi_3 = v$ is chosen,

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{2.10}$$

introducing a symmetry breaking. This situation is represented in Figure 2.2. It is then useful to study the system under small perturbations around this minimum, and for that four shifted fields are introduced, H, θ_1, θ_2 and θ_3, resulting in

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} \theta_1 + i\theta_2 \\ v + H + i\theta_3 \end{pmatrix} \tag{2.11}$$

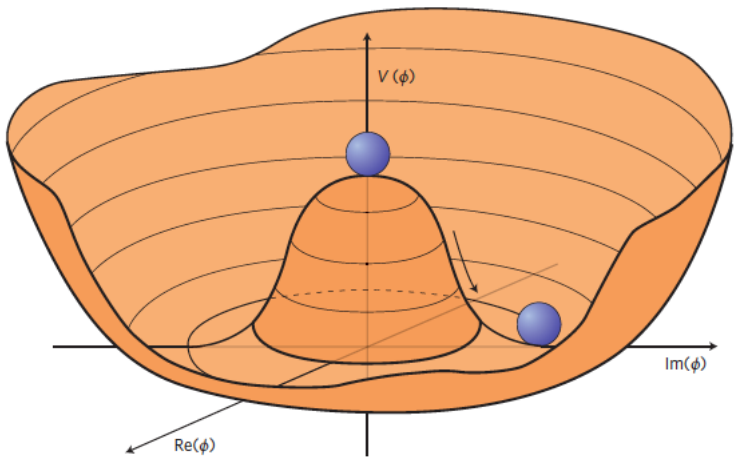

Figure 2.2: The Higgs potential [8].

Then, rewriting the lagrangian in terms of the shifted fields, and expanding the $|D_\mu\phi|^2$ term, results

in massive gauge fields, and an additional massive scalar, the Higgs boson. In particular, the W^1 and W^2 fields mix to form the W^+ and W^- bosons, as defined in Equation 2.12, and the W^3 and B field mix to form the Z boson and the photon, defined, respectively, as

$$W^{\pm} = \frac{1}{\sqrt{2}}(W_{\mu}^{1} \mp iW_{\mu}^{2}) \tag{2.12}$$

$$Z_{\mu} = \frac{gW_{\mu}^{3} - g'B_{\mu}}{\sqrt{g^2 + g'^2}} \tag{2.13}$$

$$A_{\mu} = \frac{gW_{\mu}^{3} + g'B_{\mu}}{\sqrt{g^2 + g'^2}} \tag{2.14}$$

In this process, massless particles appear, the so-called Goldstone bosons, one for each broken symmetry, but these terms disappear when changing to the unitary gauge.

It is then possible to extract the masses for the gauge bosons,

$$M_W = \frac{1}{2}vg \tag{2.15}$$

$$M_Z = \frac{1}{2}v\sqrt{g^2 + g'^2} \tag{2.16}$$

$$M_A = 0 \tag{2.17}$$

for the W bosons, Z boson and photon, respectively. Note the fact that the photon remains massless, as the vacuum is invariant under the $U(1)_{EM}$ symmetry. In fact, there exists charge conservation of the vacuum, with $Q\,\phi_0 = 0$, as $Y = +1$ and $I_3 = -\frac{1}{2}$.

In addition, the Higgs boson also has an associated mass,

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

The SM does not predict the Higgs mass, as λ is a free parameter. On the other hand, it was possible to arrive at a value of v of 246 GeV, through the study of muon decay measurements. It is possible to measure the coupling strength of the muon to the W boson in this decay, and establish a relation with the vacuum expectation value, using $\frac{G_F}{\sqrt{2}} = \frac{g^2}{8M_W^2} \Leftrightarrow v = (\sqrt{2}G_F)^{-1/2}$ [2]. Moreover, recent observations from ATLAS and CMS resulted in a observed mass of 124.97 ± 0.24 GeV for the Higgs boson [9].

On the other hand, the fermion masses come from the interactions of these particles with the Higgs field, the so-called Yukawa couplings (y_f). While the previous mass terms for the fermions were not invariant under the $SU(2)_L \times U(1)_Y$ symmetry, it is possible to construct terms combined with the Higgs boson, which become invariant under this symmetry, and therefore are allowed in the lagrangian. Moreover, the Yukawa coupling is proportional to the fermion mass, according to $y_f = \sqrt{2}\frac{M_f}{v}$.

2.2 $t\bar{t}H$ Production

The measurement of the couplings of the Higgs boson to fermions is therefore of the utmost importance, and the natural place to start are the couplings to the third generation of particles, as the coupling is proportional to the particle mass, and the processes involving these couplings have higher associated cross sections. In fact, throughout 2018, ATLAS and CMS observed the coupling of the Higgs to top [10, 11] and bottom quarks [12, 13], and tau leptons [14, 15]. This was an important step in what regards these analyses, but nevertheless there is still a long path ahead, as precise measurements of these couplings are necessary in order to confirm the validity of the Standard Model, or to find evidence of new physics.

An interesting case is the coupling of the Higgs boson to the top quark, as the top is much heavier than the other fermions, for no apparent reason, and therefore has the largest Yukawa coupling. Moreover, if new physics is accessible at the LHC, and being this the strongest coupling, deviations from the SM interaction would be more evident. This coupling can be measured in the production of the Higgs boson in association with top quarks, $t\bar{t}H$, where there is the possibility of having direct access to the $t\bar{t}H$ vertex.

The $t\bar{t}H$ process is characterised by a small rate due to the large invariant mass of the final state objects, contributing only around 1% to the total Higgs boson production cross-section (Figure 2.3). In an attempt to compensate this, searches often look for the Higgs decaying to bottom (b) quarks, as this decay is associated to the largest branching ratio of the Higgs particle (Figure 2.4). Representative leading-order Feynman diagrams for signal and main irreducible background associated with this process are shown in Figures 2.5 and 2.6, respectively.

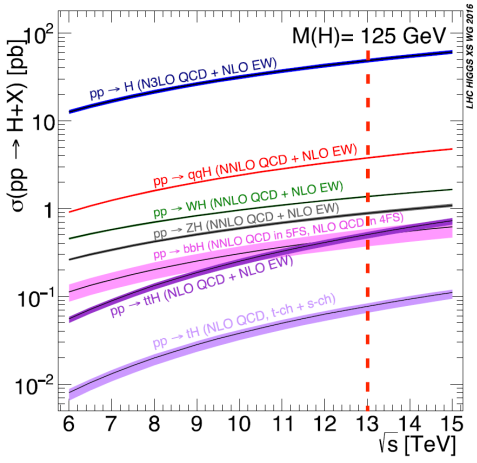

Figure 2.3: Cross sections as a function of the energy in the center of mass [16]

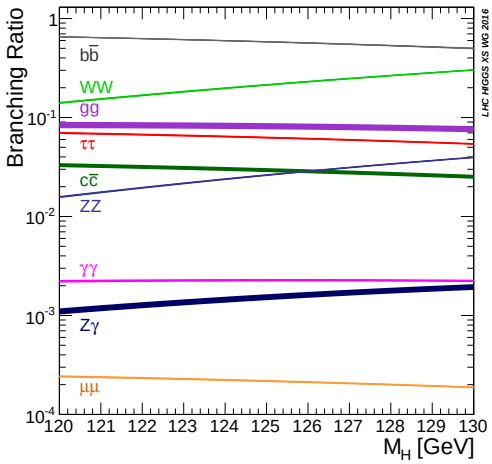

Figure 2.4: Branching ratios for Higgs decays as a function of its mass [16]

Top quarks decay $(95.7 \pm 3.4)\%$ of the time to a b quark and a W boson (experimental value stated in Reference [18]), which can then decay either hadronically, to a quark-antiquark pair, or leptonically, to a charged lepton and a neutrino. The most probable decay modes for the top quark are presented in table 2.1.

The bottom quark gives origin to a hadron characterised by having a considerable lifetime. After a *pp*

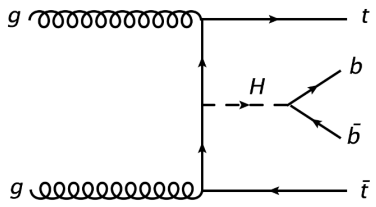

Figure 2.5: Representative Feynman diagram for signal, $H \to b\bar{b}$ channel [17].

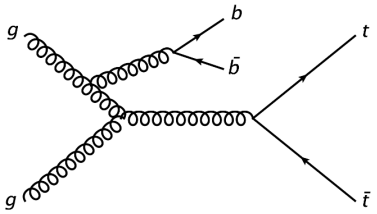

Figure 2.6: Representative Feynman diagram for main irreducible background, $t\bar{t}b\bar{b}$ channel [17].

Table 2.1: Probabilities of top decay modes [18]

Top Decay Mode	Probability (%)
$e\nu_e b$	13.3 ± 0.6
$\mu\nu_\mu b$	13.4 ± 0.6
$\tau\nu_\tau b$	7.1 ± 0.6
$q\bar{q}b$	66.5 ± 1.4

collision at 13 TeV it will travel a length of the order of a centimeter [18] from the primary vertex (PV) and decay into other sub-products, forming a secondary vertex (SV) within the jet of hadrons (Figure 2.7). This particular signature distinguishes the bottom quark decay from the rest, and requires dedicated identification algorithms.

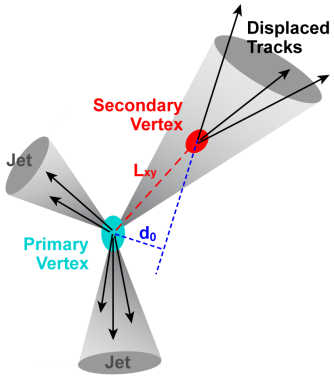

Figure 2.7: Scheme of bottom decay [19]

Jets are the result of the hadronisation of particles [20] and are deeply related to the confined nature of Quantum Chromodynamics (QCD), the theory that describes the strong force. As quarks can only appear in colourless combinations, when these start to separate at high energies, colour-anticolour pairs of quarks are produced out of the vacuum combining with the original quarks and giving origin to jets. As gluons are coloured the same happens, so in reality it is not possible to study free partons but only jets. In addition, quarks and gluons can radiate gluons, contributing to jets [21]. For this reason it is important to have an infrared and collinear safe jet clustering algorithm, for the clustering to be protected from soft (low-energy) and low-angle (with respect to the original quark or gluon) emissions, respectively, as, in QCD, such type of emissions result in divergences in higher-order perturbative calculations.

These divergences also introduce a cutoff, μ_F, called a factorisation scale, in the parton distribution

functions (PDFs), where emissions with energy below the scale are absorbed in the PDF. The scale μ_F is commonly taken as the scale of the process, generally denoted as Q, and its uncertainties are estimated by varying by a factor of two to either side of the central value [20]. Beyond tree level calculations also introduce another quantity, μ_R, called renormalisation scale, in order to handle the ultraviolet divergences. This scale also influences the value of the QCD coupling, whose dependence can be expressed in terms of a renormalisation group equation. The uncertainties for μ_R, as for μ_F, are usually estimated by choosing $\mu_R^2 = (x_\mu \mu_F)^2$ with $x_\mu = \frac{1}{2}, 1, 2$.

Furthermore, besides the low statistics, the $t\bar{t}H$ process introduces other adversities due to the strong resemblance between the main irreducible background and the signal. Also, backgrounds with jets coming from light (up, down, strange) and charm quarks, as $t\bar{t}jj$, are experimentally challenging, as these jets can be faked as bottom quarks. Moreover, in addition to the hard-scattering event of interest, events can also be categorised as pile-up. Pile-up events refer to additional proton-proton (pp) collisions occurring in the same bunch-crossing as the collision of interest, or in bunch-crossings just before and after the collision of interest. They are modelled using simulated (or real) minimum bias events, events that pass minimum trigger requirements. For each hard-scattering event, one also needs to consider what is called the 'underlying event'. This is formed by the sum of the softer radiation in the event, such as particles originating from multiple-parton interactions together with initial- or final-state radiation and the remnants of the beams.

Due to the difficulty in distinguishing $t\bar{t}H(b\bar{b})$ from the hadronic background, it is also interesting to study the $t\bar{t}H$ production with the Higgs boson decaying to WW^*/ZZ^* or $\tau\tau$. The probability of the Higgs boson decaying to a pair of W/Z bosons or a pair of tau leptons is smaller ($22\%/3\%$ and 6%, respectively [22]), but the background in these decays is smaller and easier to estimate. The representative LO Feynman diagrams for signal and main background associated with this channel are shown in Figures 2.8 and 2.9, respectively.

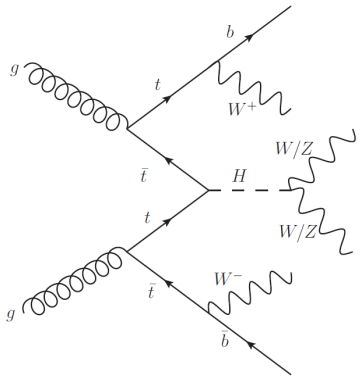

Figure 2.8: Representative Feynman diagram for signal, $H \to VV$ ($V = Z, W$ bosons) decay mode [23].

The τ lepton can decay leptonically or hadronically. The most probable decay modes are presented in table 2.2. Due to the hadronic decay modes the τ_{had} can be mistaken for a jet coming from the background and specific techniques are used in order to distinguish both.

Another interesting Higgs decay mode is $H \to \gamma\gamma$, as events can be selected with high purity. On the

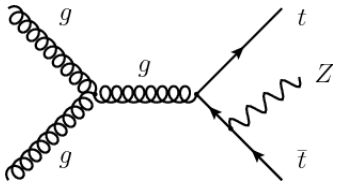

Figure 2.9: Representative Feynman diagram for background, $t\bar{t}Z$ [23].

Table 2.2: Probabilities of τ decay modes [18]

τ Decay Mode	Probability (%)
Leptonic decays	
$\mu^-\overline{\nu_\mu}\nu_\tau$	17.39 ± 0.04
$e^-\overline{\nu_e}\nu_\tau$	17.82 ± 0.04
Hadronic decays	
$\pi^-\pi^0\nu_\tau$	25.49 ± 0.09
$\pi^-\nu_\tau$	10.82 ± 0.05
$\pi^-\pi^0\pi^0\nu_\tau$	9.26 ± 0.10
$\pi^-\pi^+\pi^-\nu_\tau$	9.31 ± 0.05

other hand this decay mode has a small signal yield, so in these analyses top quarks decaying either leptonically or hadronically are targeted.

2.3 Beyond Standard Model

It is known that the SM is an incomplete theory. For example, it doesn't include the gravitational interaction and it doesn't propose a candidate for dark matter. In addition, neutrinos oscillate between flavours, implying a non-zero mass which, however, may not originate from the Higgs mechanism. Furthermore, the SM shows an hierarchy problem, as the Higgs mass is not protected from large radiative corrections and a fine-tuning is necessary to maintain the Higgs mass at the electroweak scale. Namely, when introducing one-loop mass corrections to the Higgs mass, it is possible to see that these diverge quadratically, and, taking into account the large value for the Planck Scale, it would be expected for the Higgs boson to acquire a large mass. However, the Higgs has a mass of only 124.97 ± 0.24 GeV, therefore implying that there exists an unnatural precise cancellation (fine-tuning) between the bare (non-renormalized) Higgs mass and the radiative corrections. In addition, the SM also doesn't explain the baryon asymmetry in the Universe, which may require additional sources of Charge-Parity (CP) violation. For these reasons the focus on physics Beyond the Standard Model (BSM) searches has increased, in an attempt to solve these problems.

The Higgs sector offers this possibility, as some extensions of the SM with multiple Higgs doublets account for new sources of CP violation, the so-called two Higgs-doublets models (2HDM) [24–27]. In these models the potential is still invariant under the same symmetries as in the SM but it is built with two complex Higgs doublets. As a result, new Higgs bosons are introduced, namely two neutral scalars (h and H, with h being equal to the SM Higgs), two charged Higgs scalars (H^+ and H^-) and a neutral pseudo-scalar A.

A particular model is the Complex 2HDM (C2HDM), where there are the h_1, h_2 and h_3 Higgs bosons instead of h, H, A, and these eigenstates are a CP-odd and CP-even mixture. This allows for CP-violation in the potential, providing an extra source of CP-violation to the theory. On the other hand the properties of the observed Higgs boson have been tested with improving accuracy and are therefore providing ways to test the predictions of the SM or the presence of new physics.

As mentioned before, the coupling of the Higgs boson to the top quark, when measured precisely, can be used as a probe for new physics, which can be measured in the production of the Higgs boson in association with top quarks, $t\bar{t}H$, obtaining sensitivity to the CP nature of the couplings.

The most general Lagrangian term that accounts for contributions from CP-odd and CP-even components of the couplings is defined as

$$\mathcal{L} = \kappa y_t \bar{t}(cos\,\alpha + i\gamma_5 sin\,\alpha)tH \tag{2.19}$$

where y_t is the Yukawa coupling of the Higgs boson to the top quark, and α is a CP phase ($cos\,\alpha = 1$ recovers the SM interaction while $cos\,\alpha = 0$ corresponds to the pure pseudo-scalar case). κ is a real number. Note to the fact that the pure CP-odd case was already excluded at 99.98% confidence level [28, 29], but a mixing between CP-even and CP-odd components is still allowed by experimental data.

Chapter 3

Experimental Apparatus

Dedicated experiments are necessary in order to attest the Standard Model or find evidence for new physics. In that sense, proton-proton and lead-lead collision data has been collected in the Large Hadron Collider (LHC), a particle accelerator built by CERN, the European Organization for Nuclear Research, of which ATLAS is one of the experiments.

This chapter describes the experimental conditions for which searches for the $t\bar{t}H$ process were conducted, as well as the ones for which this study is intended. It starts by describing the LHC in the first section, along with a description of the current ATLAS detector in its subsection. In the following section and subsection the future upgrade of the LHC, the so-called HL-LHC, is covered, as well as the expected upgrades to be taken in ATLAS.

3.1 The Large Hadron Collider at CERN

Located in the France–Switzerland border near Geneva, the LHC has a ring of 27 km circumference, where bunches of protons or heavy ions travel in opposite directions at close to the speed of light until they collide. Along the ring, there are several superconducting dipole magnets, of 8.3 T, to bend the beams, and quadrupole and other multipoles magnets, to focus the beams. The acceleration of the particles is provided through radiofrequency cavities. Around the LHC are located four crossing points. The ATLAS and CMS detectors are positioned in two of them, and both are multi-purpose experiments, with equivalent designs, but using different technologies, providing complementary results.

In the center of these detectors the two beams collide with a very small crossing angle, introduced in order to avoid parasitic long-range beam-beam interactions. The bunches are separated by 25 ns and collide at a center-of-mass energy of 13 TeV, producing a pile-up of around 40 additional collisions besides the hard-scattering event. In order to control the pile-up, a technique called ”luminosity leveling” has been implemented in the present Run 2. This technique reduces the luminosity and maintains it at an acceptable level, until the point where the intensity of the beams has sufficiently decreased.

Instantaneous luminosity is defined in Equation 3.1, where f_{coll} is the bunch crossing frequency, n_1 and n_2 the number of particles in each bunch and σ_x and σ_y the transverse profiles of beams, assuming

they are Gaussian. The LHC currently operates at a maximum instantaneous luminosity of around 2×10^{34} cm^{-2}s^{-1}, which is twice its nominal value. The integrated luminosity is the integral of luminosity over a period of time, as defined in Equation 3.2, and is usually expressed in inverse femtobarn (1 fb$^{-1} = 10^{39}$ cm^{-2}).

$$\mathcal{L} = f_{coll} \frac{n_1 n_2}{4\pi\sigma_x\sigma_y} \tag{3.1}$$

$$\mathcal{L}_{int} = \int \mathcal{L}\, dt \tag{3.2}$$

Luminosity plays an important role in collider physics, alongside with energy. While the latter has a direct influence in the processes' cross sections, for instance, luminosity is related to the rate, for instance, of potential occurrences of those processes in collisions. By the time of the next LHC upgrade, to the High-Luminosity LHC, in 2024, the LHC is expected to have collected around 300 fb^{-1}.

3.1.1 The ATLAS detector

The reference system used in ATLAS is a right-handed coordinate system with its origin at the nominal interaction point in the centre of the detector and the z-axis pointing in the direction of the beam pipe, the y-axis pointing upwards and the x-axis pointing towards the centre of the LHC. The polar (ϕ) and azimuthal (θ) angles are as usually defined. Pseudo-rapidity is given by Equation 3.3.

$$\eta = -ln\left[tan\left(\frac{\theta}{2}\right)\right] \tag{3.3}$$

The ATLAS detector [30], represented in Figure 3.1, has at its center the inner detector (ID), surrounded by a superconducting solenoid magnet creating a 2 T magnetic field. The ID consists of a high-granularity silicon pixel detector, placed close to the interaction region to allow for the measurement of the impact parameter of charged-particle tracks and the position of secondary vertices, and a semiconductor tracker (SCT) that also contributes to precision tracking. Furthermore it has a transition radiation tracker (TRT) that provides tracking and electron identification. The charge and momentum of charged particles is measured from the track curvature produced by the magnetic field. These components are surrounded by a lead/liquid-argon electromagnetic calorimeter (ECAL), used to detect photons and electrons, and this one by hadronic calorimeters, using liquid argon or scintillating tiles as active materials, and iron, copper or tungsten as absorbers. Hadronic jets develop in both the electromagnetic and hadronic calorimeters. Surrounding them there is the muon spectrometer that measures the deflection of muons in a magnetic field, to determine their momentum.

Due to technical aspects, such as storage capacity and readout bandwith, it is not possible to store all the collision events, hence the need for an online trigger. The ATLAS trigger is divided in two layers: Level 1 (L1) and High Level Trigger (HLT). L1 is purely hardware-based, and is responsible for reducing the event rate from 40 MHz to 100 kHz. It is composed by a muon and calorimeter sub-trigers, which information is used to find regions of interest, and a Central Trigger. On the other hand, the HLT uses

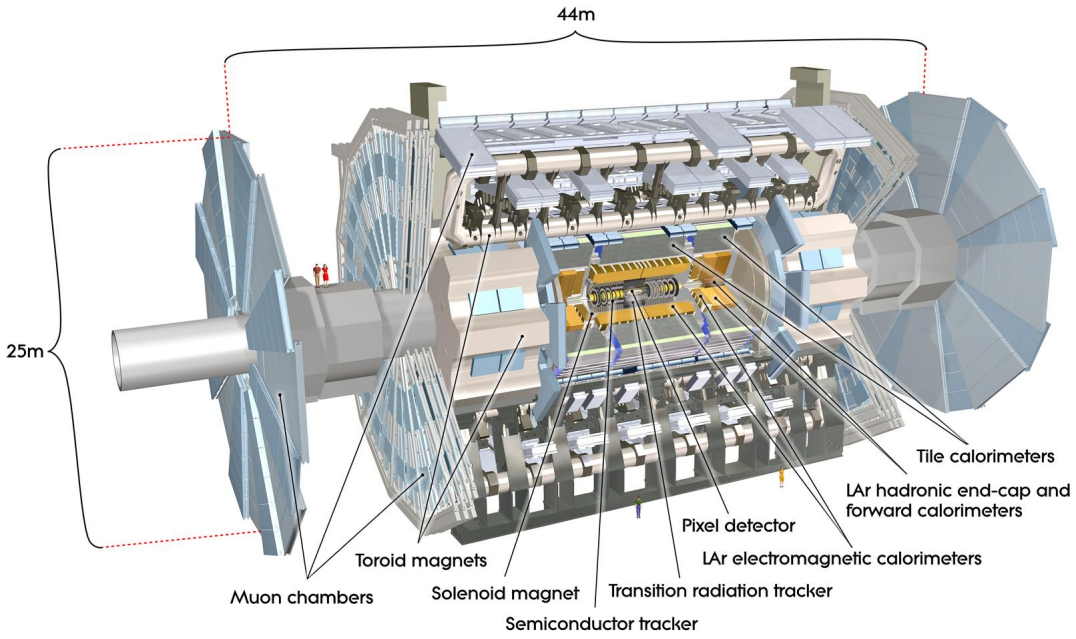

Figure 3.1: The ATLAS detector [30].

offline-like algorithms in the regions of interest to reconstruct the events and select the ones containing interesting features, further lowering the rate to 1 kHz.

Inner Detector

The ID is the first major system of the ATLAS detector and is responsible for the reconstruction of the trajectories, vertices and momenta of charged particles. For that purpose it uses the pixel detector, the SCT and the TRT for with high-precision measurements. These three parts of the ID are divided into barrel and two end-caps regions. A scheme of the ID is shown in Figure 3.2. In it, particles are detected through the ionization of the detector materials, or the creation of electron-hole pairs in semiconductors, that are separated using an electric field, and the generated charge is read.

The first component of the ID, the pixel detector, was initially composed of three layers in the barrel region and two end-cap regions, each with three disks, and it covers the pseudorapidity range $|\eta| < 2.5$. The barrel and the end-caps contain 1456 and 288 sensor modules, respectively, where each module has 46080 readout pixels, resulting in around 80 million channels. During the first LHC long shut-down an additional layer was installed, known as Insertable B-layer [31], and it contributes with roughly more 38 million pixel cells to the precise reconstruction of vertices and tracks.

The SCT consists of silicon microstrips detectors, placed in four layers in the barrel region, and nine layers in each end-cap. It counts with around 6.3 million channels to provide accurate positions for the charge particles, in the range $|\eta| < 2.5$.

Finally, the TRT is made of straw detectors, with 50 000 longitudinal straws in the barrel, and 320 radial straws in the end-caps. The barrel straws are divided in two at the center and have readout channels at both ends, while the end-cap straws have the readout at the outer radious. The TRT therefore counts

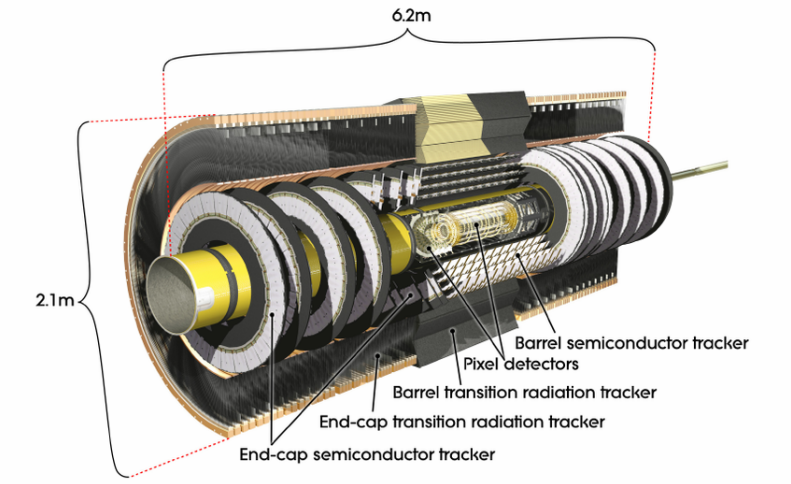

Figure 3.2: The ATLAS Inner Detector [30].

with 420 000 channels to cover the pseudorapidity range $|\eta| < 2$. Its information is particularly important for track momentum measurements, and it contributes to the identification of electron and positrons, by detecting X-ray photons emitted by these particles as they cross the detector.

Calorimeters

The calorimetry system of the ATLAS detector is composed of an electromagnetic and hadronic part, being responsible for the measurement of the energy deposited by charged and neutral particles, as well as their directions. Additionally, it is also possible to determine the missing transverse energy in the event. These calorimeters are sampling detectors, consisting of layers of absorber material alternated with active materials. When crossing the detector the particles interact with the absorbers, losing energy and forming showers. A layout of the calorimeters is presented in Figure 3.3.

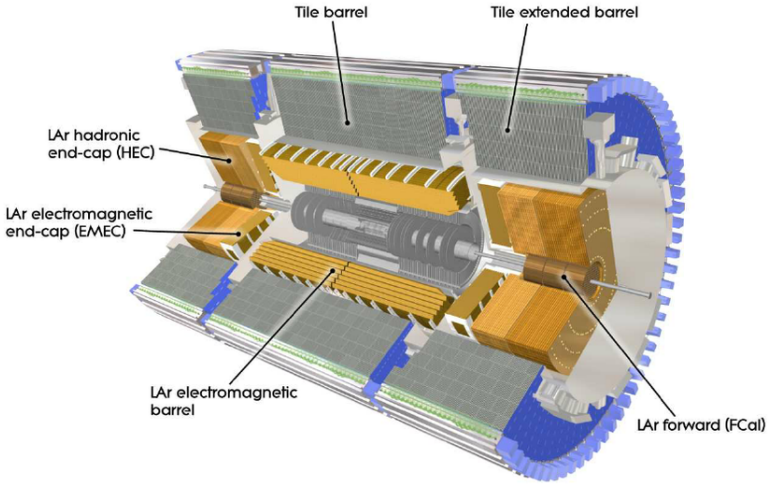

Figure 3.3: The ATLAS Calorimeters [30].

The ECAL consists of accordion-shaped lead absorber plates and electrode plates interleaved with liquid argon. The showers formed in the absorber ionize the liquid argon, and the free electrons are

collected by the electrodes. The ECAL is divided in a barrel, covering the region $|\eta| < 1.475$, and two end-caps, covering $1.375 < |\eta| < 3.200$. Furthermore, it has a total thickness of more that 22 and 24 radiation lengths in the barrel and end-caps, respectively, with the goal of containing the electromagnetic shower.

Around the ECAL there is a hadronic calorimeter (HCAL), consisting of a barrel, also known as TileCal, and two end-caps and two forward calorimeters. The TileCal uses scintillating tiles as active material and steel plates as absorber, and covers the region $|\eta| < 1.7$. In this case, the showers cause the tiles to emit light that is collected by wavelength-shifting optical fibres. The fibres then convert this light, that is in the ultraviolet region, in visible light, and send it to photo-multiplier tubes. In addition, the fibres are aluminized in the top opposite to the photo-multiplier, with the goal of maximising the efficiency of the light collected. On the other hand, the hadronic end-caps and forward calorimeters use liquid-argon as active material, and copper and tungsten, respectively. The detection of particles here goes as for the ECAL. Moreover, the end-caps cover the pseudo-rapidity region $1.5 < |\eta| < 3.2$, while the forward calorimeter covers $3.1 < |\eta| < 4.9$.

Muon Spectrometer

The muon spectrometer is the outermost component of the ATLAS detector, and intends to measure the momentum of muons deflected by large superconducting air-core toroid magnets. It is instrumented with separate trigger chambers, consisting of resistive plate chambers (RPC) and thin gap chambers (TGC), covering the range $|\eta| < 2.4$, and high-precision tracking chambers, composed of monitored drift tubes (MDT) and cathode strip chambers (CSC), covering $|\eta| < 2.7$. The magnetic bending is provided by a large barrel toroid in the range $|\eta| < 1.4$, while between $1.6 < |\eta| < 2.7$ muons are deflected by two end-cap magnets. In the transition region the bending of muon tracks is achieved through a combination of barrel and end-cap fields. Furthermore, the barrel region consists of three cylindrical layers of chambers, whilst the transition and end-cap regions have the chambers installed perpendicularly to the beam, also in three layers. A scheme of the muon system is shown in Figure 3.4.

The MDTs consist of cathode tube filled with argon and an anode wire made of tungsten-rhenium readout, and a particle is detected by the ionization of the gas. On the other hand, the CSCs are made of stripped copper cathodes and anode wires. Moreover, the RPCs are gaseous electrode plate detectors, and the TGCs are multi-wire proportional chambers.

3.2 The High-Luminosity Large Hadron Collider

In the next years the LHC will undergo a series of upgrades that will lead to the High-Luminosity Large Hadron Collider (HL-LHC) [32–34] that is expected to start operating in 2026 and to collect $3-4$ ab^{-1} of data, after ten years of operation. The upgrade's schedule is presented in Figure 3.5.

This upgrade will be undertaken between 2024 and 2026 in the so-called LS3 (Long Shutdown 3), and after it collisions will be held at a center-of-mass energy of $\sqrt{s} = 14$ TeV, in a pile-up environment

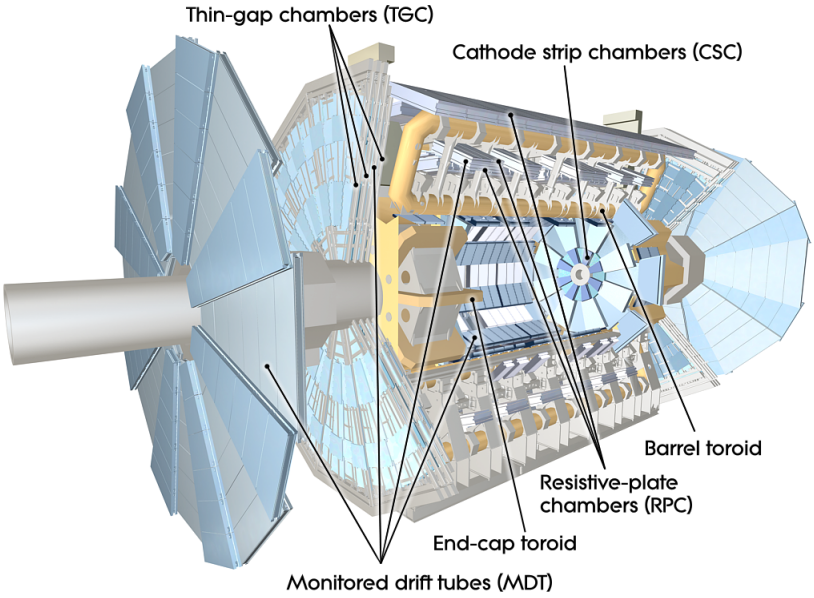

Figure 3.4: The ATLAS muon system [30].

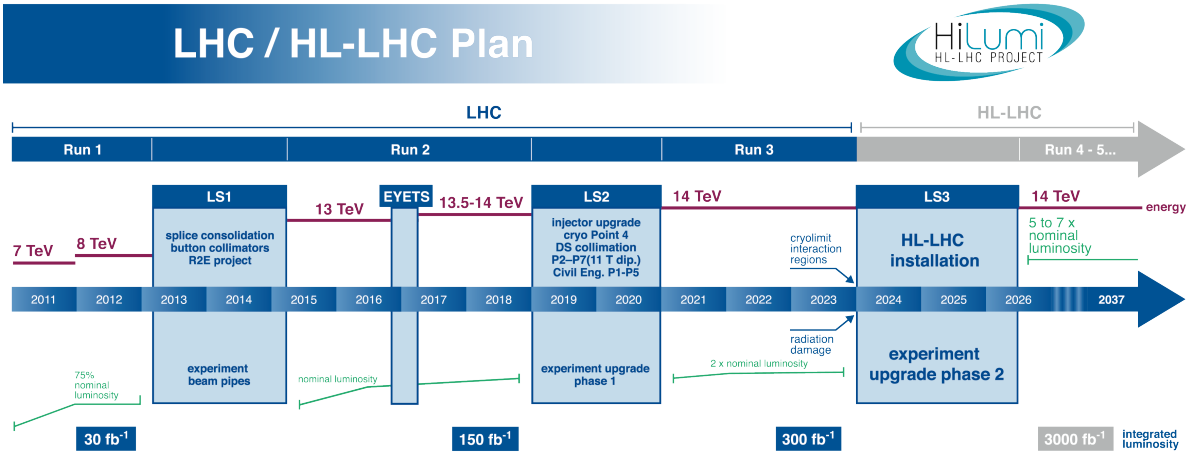

Figure 3.5: The LHC upgrade plans [33].

of 140-200 simultaneous events. It is intended to operate at an instantaneous luminosity of 7.5×10^{34} cm^{-2}s^{-1} with 25 ns bunch spacing.

An increase in luminosity translates into more collisions per bunch crossing and therefore more data collected. The increase in statistics will allow for the detailed study of processes with low cross sections, such as $t\bar{t}H$ production.

This increase in luminosity will mainly be achieved by replacing the current LHC superconducting quadrupole niobium-titanium magnets by new and more powerful magnets made of niobium-tin [32], that will allow for a better focusing of the beams and a decrease of their transverse profiles in the crossing points. Moreover, the introduction of crab cavities, that tilt the proton bunches by giving them a transverse momentum, will maximize the overlap area of the two bunches and will therefore increase the probability of collisions. In addition, two of the dipole magnets responsible for the bending of the protons

in the LHC ring will also be replaced by new ones of superconducting niobium-tin compound, capable of providing a magnetic field of 11 T.

For the so-called Phase-II upgrade the ATLAS and CMS detectors will have upgrades in each major system and will see their performances improved.

3.2.1 The ATLAS Phase-II detector

In the High-Luminosity scenario [33][1] the ATLAS trigger [35] will be divided in three major systems: Level-0 trigger, Data Acquisition (DAQ), and Event Filter (EF).

The L0 trigger will be composed of calorimeter and muons sub-triggers, receiving information at 40 MHz, followed by a Global Trigger, to refine the sub-triggers information and perform offline-like algorithms, and a Central Trigger Processor, to make a final hardware trigger decision. The rate is reduced to 1 MHz by the end of this process. Nevertheless, the hardware architecture can be split into two levels, L0 and L1, in case the pile-up conditions as so require. In that scenario, the L0 trigger will have an output rate capability of up to 2-4 MHz, while the L1 trigger will sustain rates of 600-800 kHz.

The L0 trigger output and the inner tracker information are then sent to the DAQ system, that will therefore operate at 1 MHz. It will contain a Readout and Dataflow sub-systems, to manage the data before sending it to the EF.

For the Phase-II, the EF will consist of a CPU-based processing farm, complemented by Hardware-based Tracking for the Trigger (HTT) co-processors, and further reduces the event rate to 10 kHz. The HTT is planned to receive information from the outermost ID layers, and to quickly provide track candidates for the EF.

Furthermore, the pixel tracker will be upgraded, with its pseudo-rapidity range being increased to $|\eta| < 4$. On the other hand the SCT and the TRT will be replaced by an all-silicon tracker, the ITk, that is also expected to cover the region $|\eta| < 4$. This new component will have a lower mass, which will allow to reduce the effect of photon conversions, hadronic interactions and multiple scattering, and its performance is expected to be as good as the current ID. The forward calorimeter will also suffer an improvement, resulting in higher transverse granularity, so that the large expected pile-up can be handled. Another approach to address this problem, in the forward region, will be the implementation of a high granularity timing detector [36]. This new detector will be placed in front of the end-cap calorimeters and will cover the pseudo-rapidity range $2.4 < |\eta| < 4.0$, with a timing resolution of 30 ps. The use of timing information will help reduce the effects of pile-up in this region, as it will improve the assignment of tracks to the primary vertices. Consequently, the efficiency of the b-tagging algorithms, as well as the identification of leptons, in the forward region, will be improved. Additionally, the high voltage distribution and cooling of the calorimeters will have to be improved, in order to cope with the increase in luminosity, that will result in larger energy deposits.

The muon system will also be enhanced, mainly with the goal of improving the performance of the muon related part of the trigger. This will be done by replacing the muon chambers in the region $2.0 < |\eta| < 2.4$, adding new ones in $|\eta| < 1$, and by upgrading the electronics. In addition, two new

[1] In this work the Reference scenario will be taken into account.

components will be implemented, the New Small Wheel (NSW), that will replace the first layer of the muon end-cap, and a very forward muon tagger, covering the region $2.6 < |\eta| < 4$.

The ATLAS detector performance in the LHC, along with the expected improvements in the HL-LHC scenario, is presented in Table 3.1.

Table 3.1: General performance of the ATLAS detector in the LHC and HL-LHC scenarios. The units for E and p_T are in GeV. Based in References [30] and [37].

Detector component	Resolution	η coverage		ATLAS HL-LHC Upgrade	
		Measurement	Trigger	η coverage	Improvement
Tracking	$\sigma_{p_T}/p_T = 0.05\% p_T \oplus 1\%$	± 2.5		± 4.0	Factor $\sim$ 2 better resolution
EM Calorimetry	$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$	± 3.2	± 2.5	Same	Same granularity
Hadronic calorimetry					
barrel and end-cap	$\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$	± 3.2	± 3.2	Same	Same granularity
forward	$\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$	$3.1 < \|\eta\| < 4.9$	$3.1 < \|\eta\| < 4.9$	Same	Higher transverse granularity
Muon spectrometer	$\sigma_{p_T}/p_T = 10\%$ at $p_T = 1$ TeV	± 2.7	± 2.4		NSW to reject 90% of fake muon triggers
New timing detector	-	-	-	$2.4 < \|\eta\| < 4.0$	30 ps/track

Chapter 4

Analysis Tools

Implementing an analysis requires much more than a simple idea. It frequently needs dedicated frameworks and algorithms. Namely, Monte Carlo simulations are always necessary in the scope of an analysis, either to compare to data, to estimate systematic effects, or to realize preliminary tests, as for example, looking for alternative strategies. Moreover, it is also often necessary to implement specific algorithms or variables, to identify the objects of interest in the events.

This chapter therefore starts by giving, in the first section, an overview of the methods necessary to generate event samples. In the next section, the event generation performed in the course of this thesis is then described in detail. The third section is dedicated to the explanation of the tagging of bottom quarks, and is followed in the fourth section, by the definition of variables that exploit jet substructure information to search for boosted heavy particles. Finally, the fifth and sixth sections cover two algorithms designed to tag boosted hadronic top quarks and Higgs bosons, respectively.

4.1 Event Generation

The MADGRAPH5_AMC@NLO [38] framework is used to compute the tree-level (LO) and next-to-leading order (NLO) matrix elements of each physics process. It includes the MADSPIN [39] method that preserves the spin correlations in the decays. Alternative generators are POWHEG-BOX V2 NLO [40, 41] and SHERPA [42].

Moreover, the parton showering and hadronisation of the events is done with PYTHIA8.2 [43]. In this framework, each parton can radiate a quark or a gluon according to the probabilities obtained from the DGLAP equations [20], the QCD evolution equations for parton densities, resulting in parton showers.

The jet clustering is done via the FASTJET3 package [44] and two algorithms are particularly relevant: anti$-k_t$ [45] and Cambridge/Aachen [46, 47]. Both are infrared and collinear safe and the definitions of the distance measures have a similar structure for both. These are shown in Equations 4.1 and 4.2, where $\Delta R_{ij} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}$ is the distance between particles i and j in (y, ϕ) space (where y is rapidity and ϕ is the azimuthal angle), k_{ti}^{2p} is the transverse momentum of the $i-$th particle with respect to the beam axis, to the power of $2p$, where p is a parameter dependant of the chosen clustering

algorithm, d_{iB} is the distance between the particle i and the beam, and R is a radius parameter. For the Cambridge-Aachen (C/A) algorithm $p = 0$ and for the anti-k_t algorithm $p = -1$.

$$d_{ij} = \text{min}(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \tag{4.1}$$

$$d_{iB} = k_{ti}^{2p} \tag{4.2}$$

The anti-k_t algorithm is the most commonly used for general studies while the C/A algorithm is more suitable for the boosted regime (particles with very high transverse momentum) searches as it allows for the study of jet sub-structure, and therefore of boosted resonances decaying within a single jet.

For fast simulation of the detectors it was used the DELPHES3 [48] framework, where the simulation considers each particle crossing the detector individually. For the full simulation of the detectors the GEANT4 toolkit [49] is available.

4.2 Monte Carlo Samples

For this thesis, events were generated at a center-of-mass energy of $\sqrt{s} = 14$ TeV and at Leading Order (LO). Moreover, the semi-leptonic channel was considered, and therefore, in the samples with a $t\bar{t}$ pair, half of the events have a top decaying semi-leptonically, and the other half has an anti-top decaying semi-leptonically. For the samples with a W boson, this decays into a charged lepton and anti-neutrino or charged anti-lepton and neutrino, depending if it is a W^- or W^+, respectively. An alternative $t\bar{t}A$ signal sample was generated with the HC_UFO_V4.1 model [50], with A being a pure pseudo-scalar boson instead of the SM scalar Higgs, but also decaying into two bottom quarks.

The main generator used was MG5_AMC@NLO, with the LO NN23LO1 PDF, for consistency, except for the dijets sample, that was generated with PYTHIA8.2 and the LO CTEQ 5L PDF. Decays are made through MADSPIN.

A complete list of the generated processes, along with the values for the cross section times the branching ratio (BR) of the decays, and the number of events and generator used, for the LHC and HL-LHC scenarios, is shown in Table 4.1. The list of processes covers, besides the signal and main irreducible background, all the backgrounds that were found to be relevant for this work, as well as the $t\bar{t}A$ production. In the $t\bar{t}j$ and $b\bar{b}j$ processes, *j* stands for gluons, light quarks (up, down, strange and respective anti-particles), or charm/anti-charm quarks. On the other hand, in the dijets sample, jets can be gluons or all quarks up to the top, exclusively.

Some cuts are applied during event generation, with the goal of avoiding phase space regions where QCD completely dominates and there is poor understanding of the background description, and also to take into account experimental details. For instance, a minimum p_T of 10 GeV on leptons is required at generator level, as there is a lot of noise below that value, and the reconstruction purity is very low. The same reason is valid for the p_T cut of 10 GeV on the quarks on most samples. The remaining cases will be later explained in the text. Finally, a minimum ΔR of 0.1 is required between pairs of jets, b quarks, and between jets and leptons, to avoid QCD dominated regions, and to take into account the detector

Table 4.1: Event Generation for LHC and HL-LHC scenarios. In this table 'k' stands for thousand, 'ev' for events and 'BR' for Branching Ratios'.

Process	Cross Section $\times$ BR (pb)	# ev LHC (k)	# ev HL-LHC (k)	Generator
$t\bar{t}H(\rightarrow b\bar{b})$	0.068	500	500	MG5_aMC@NLO
$t\bar{t}b\bar{b}$	1.223	1 000	2 000	MG5_aMC@NLO
$t\bar{t}Z(\rightarrow b\bar{b})$	0.016	500	500	MG5_aMC@NLO
$t\bar{t}j$	20.300	900	2500	MG5_aMC@NLO
$W^+b\bar{b}$	17.140	400	800	MG5_aMC@NLO
$W^-b\bar{b}$	11.300	400	800	MG5_aMC@NLO
$b\bar{b}j$	175 000.000	800	1400	MG5_aMC@NLO
dijets	10 300.000	300	300	PYTHIA8.2
$t\bar{t}A(\rightarrow b\bar{b})$	0.032	-	400	MG5_aMC@NLO

resolution. All other parameters are set to their default values. A full list of the imposed cuts is presented in Table 4.2.

Table 4.2: Generator cuts. In this table *lep* and *l* stand for leptons, *g* is a gluon, *light* are light quarks, and *c* is a charm quark. In addition *b* represents a b quark, and *j* a jet.

Process	$p_{T_{lep}}$ (GeV)	$p_{T_{g,light,c}}$ (GeV)	p_{T_b} (GeV)	Min $\Delta R_{jj,bb,jl}$
$t\bar{t}H(\rightarrow b\bar{b})$		10	10	
$t\bar{t}b\bar{b}$		10	10	
$t\bar{t}Z(\rightarrow b\bar{b})$		10	10	
$t\bar{t}j$	10	100	10	0.1
$W^+b\bar{b}$		10	10	
$W^-b\bar{b}$		10	10	
$b\bar{b}j$		50	20	
dijets		300	300	
$t\bar{t}A(\rightarrow b\bar{b})$	10	10	10	0.1

The $t\bar{t}j$ process was initially generated with a minimum $p_T > 10$ GeV on the jet, that was later increased to 100 GeV. In fact, only around 250 in 200 thousand events would pass the analysis, and, from these, only 10 would have jets with a transverse momentum below 100 GeV. The difference in increasing the momentum cut is therefore associated to an error of around 4%. Nonetheless, this was the only option that could be taken, as generating the amount of required events would not be feasible, and, moreover, this background is not expected to have a mass peak. On the other hand, the 4% error will mainly be associated to a scale factor, that is nevertheless smoothed by the sidebands.

A similar situation happened with the dijets sample. The p_T cut was initially placed at 180 GeV (the initial cut was placed at high p_T as the analysis works in the boosted regime), but it was later increased to 300 GeV. Only 13 in 100 thousand events would pass the analysis, and from these only 3 would have $p_T < 300$ GeV. The associated error in this case is larger, rounding 20%. Nevertheless, the same reasons apply to this case, and, furthermore, this is a very suppressed background by the analysis, and it would even more difficult to have a proper amount of statistics for this process without the higher p_T cut.

For the $b\bar{b}j$ sample, jets coming from gluons, light or charm quarks are required to have a minimum transverse momentum of 50 GeV, and bottom quarks must have a p_T of at least 20 GeV.

Events were then processed using PYTHIA8.2, to simulate the parton showers and hadronization process. The fast simulation of the ATLAS detector is further done with DELPHES, using the ATLAS default parameter card for the LHC scenario, and with the HL-LHC default parameter card for the Phase-II scenario. These cards contain information about the detectors and their performances. An important fact worth mentioning is that the HL-LHC card considers a mixture of the ATLAS and CMS experiments, using, among other things, a magnetic field of 3 T, contrary to the 2 T used by ATLAS. Nevertheless, this is the offical card to be used for HL-LHC simulations with DELPHES.

In DELPHES a few cuts are implemented, for the same reasons as before. Namely, leptons are required to have $p_T > 10$ GeV, and have an isolation criteria, demanding that the isolation variable I is below 0.1 within $\Delta R < 0.3$, meaning that the p_T of a $R = 0.3$ jet around the lepton must be less than 10% of the lepton p_T, in order to consider it an isolated lepton. All other parameters are set to default.

4.3 Tagging of bottom quarks

The bottom quark decay process is responsible for a significant number of hadrons coming from the secondary vertex (SV) and not from the primary vertex (PV) (Figure 2.7) and this clear displaced emission in relation to the PV is an important characteristic for identifying b-jets.

The SV can be reconstructed from the tracks in the jet and has an uncertainty in position σ (σ_{d_0} and σ_{z_0} for the transverse and longitudinal projections, respectively). Then, by extrapolating all the tracks in the jet back towards the PV, the distance from the extrapolated tracks to the collision point can be measured, obtaining from the projections the transverse (d_0) and longitudinal (z_0) impact parameters. For the transverse case, if the ratio between d_0 and σ_{d_0}, also called the transverse impact parameter significance, is close to 1 then the likelihood that we are in the presence of a b-quark is small; however if this value is much larger than 1, then we can do the identification with a good degree of confidence. The transverse and longitudinal impact parameter significances, along with secondary vertex properties (invariant mass, number of tracks, distance to PV, for instance), and jet kinematic variables (p_T, η), are usually used as input to artificial neural networks, that then provide further discriminating power between jets originating from bottom quarks and those coming from charm or light quarks.

For each b-tagging efficiency there is a associated c-jet tagging efficiency, for b-tagged jets actually containing a charm quark, and a light quark mistag probability, for incorrect b-tagged jets coming from light quarks. The chosen working points for ATLAS, for the HL-LHC [51] and for the current run of the LHC [*Expected flavour tagging performance in release 21* - Protected data], are presented in Table 4.3. These points result of a balance between b-tagging efficiency and background rejection, as a higher b-tagging efficiency is associated with higher background contamination, and vice-versa.

Table 4.3: b-tagging working point

Scenario	b-tagging efficiency	c-tagging efficiency	light quark mistag probability
LHC	61%	4.5%	0.08%
HL-LHC	65%	3%	0.07%

However, when using a fast detector simulator such as Delphes, it is not possible to use vertex information, as in a real analysis. The reason for this lies with the fact that in such a simulation tracks are not valid, as in the framework there are no reconstructed tracks from points in the ID, but instead parametrized tracks.

The b-tagging in this work is therefore based on a ΔR criterion, with $\Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$. The process starts by collecting in each event all the b and c quarks with a PYTHIA particle status of 23, which corresponds to particles belonging to the hard scattering process being simulated. Moreover, when applying b-tagging on jets it is mandatory that these have a $\eta < 2.5$. The ID's range will be increased to $\eta < 4.0$, but improvements are only expected at the moment for $\eta < 2.5$, with comparison to the current ATLAS scenario, and pseudo-rapidity values above that are still under optimization in terms of b-tagging. Therefore, this work uses a b-tagging range of $\eta < 2.5$.

The minimum ΔR between each jet axis and any b quark in the collection is then computed. For a jet, if this distance is below or equal to 0.3 a random number is generated between 0 and 1. In order to have a b-tag on this jet this number must not be greater than 0.65, simulating a 65% b-tagging efficiency working point.

On the other hand, the ΔR between the jet and the collection of c quarks is computed if the ΔR between the jet and the b quark is greater than 0.3. If the value for this ΔR is below or equal to 0.3 another random number is generated between 0 and 1. It must not be greater than 0.03 in order to b-tag the jet, for a 3% c-tagging efficiency, simulating the misidentification of a c-quark initiated jet as a b jet.

Finally, if the ΔR between the jet and the c quark is greater than 0.3 it is generated a random number between 0 and 1. This number has to be smaller or equal to 0.0007, for a 0.07% light quark mistag probability, so that we have a b-tag. The cuts on ΔR between the jets and the generator level b and c quarks were based on the distributions of the minimum ΔR between these objects. One example is presented in Figure 4.1.

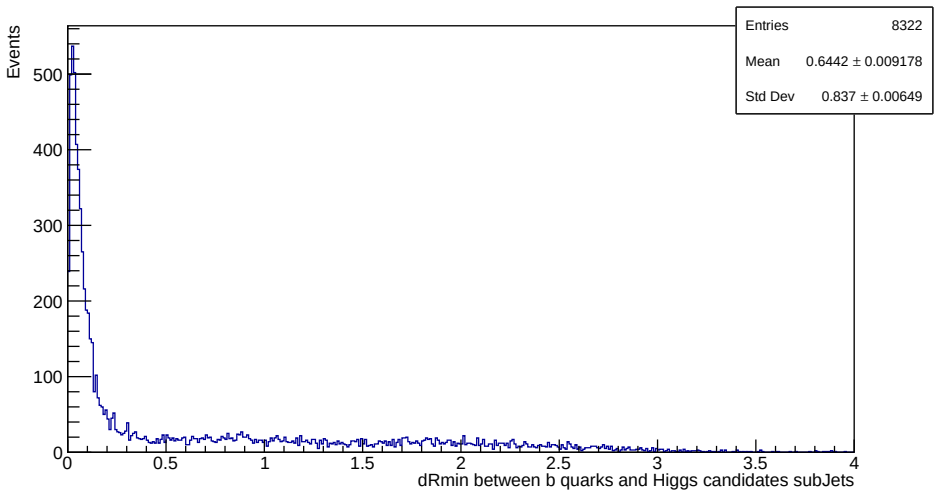

Figure 4.1: ΔR_{min} between generator level b quarks and Higgs subJets.

For each event the random seed of the generator is differently initialised so that each b-tag runs

independently of the others. As a confirmation that there is no bias in the random generator, several tests were performed. An example, the number of times each random number is generated, for a sample of 500 thousand events of $t\bar{t}H$ production, is presented in Figure 4.2.

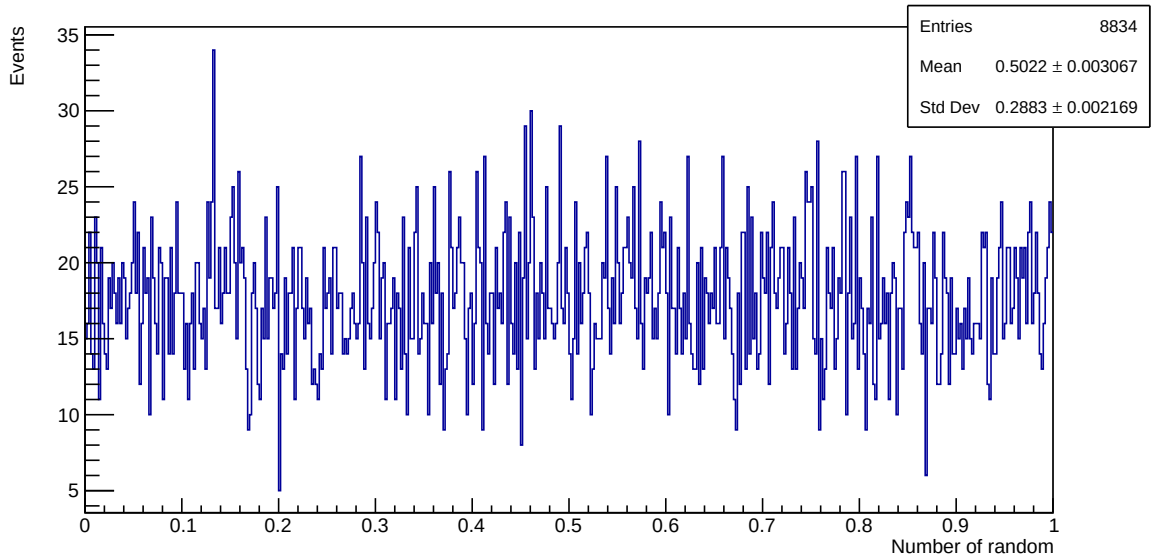

Figure 4.2: b-tagging random numbers.

4.4 Jet Substructure

The composition of a pure QCD jet is energetically different from a jet coming from boosted objects with an equivalent invariant mass. While a QCD jet results mostly from soft and collinear radiation of a light quark or a gluon, a jet originating from a hadronically-decaying top quark, on the other hand, is composed of three hard substructures, associated to the three quarks.

The jet substructure information can therefore be used to identify particles such as the Higgs boson and the top quark, through the use of specific variables, such as N-subjettiness, or dedicated functions, as for instance the Energy Correlation Functions.

4.4.1 N-subjettiness

N-subjettiness (τ_N) [52] is an event-shaped variable that looks for the energy flow inside jets. It therefore reflects the likelihood of having N hard substructures inside a jet, which can be particularly useful when searching for boosted heavy particles. In the boosted regime, the decay products of an object, such as the Higgs, will be collimated, and will be inside a single large jet. This variable can then be used to discriminate between a jet of interest and a QCD jet.

N-subjettiness is defined in Equation 4.3. In this equation, the k index stands for a constituent particle of the input jet, and $p_{T,k}$ is the transverse momentum of that particle. $\Delta R_{S,k}$ is again the distance in the

rapidity-azimuth plane, in this case between a constituent particle k and a candidate subjet S. Moreover, the factor d_0 is equal to $\Sigma p_{T,k} R_0$, where R_0 is radius of the input jet.

$$\tau_N = \frac{1}{d_0} \sum p_{T,k} min(\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}) \tag{4.3}$$

Through this definition it is possible to see that a jet with N hard substructures will have $\tau_N \approx 0$, with the jet radiation spatially aligned with its subjets. On the other hand, a QCD jet will have its radiation more dispersed in space, and therefore $\tau_N >> 0$, meaning that jet has in principle at least $N+1$ subjets.

For the computation of this variable and the determination of the subjets of a large jet, the exclusive k_t algorithm [53] is used. This algorithm runs initially in the same way as the regular (inclusive) k_t algorithm, looking for the minimum between the distances d_{iB} and d_{ij}, where *i,j* are particles and B is the beam. If d_{iB} is the smallest distance, then particle i is removed from the list of particles, while the usual algorithm keeps it.

However, τ_N is not always sufficient in order to discriminate between an object of interest and a QCD jet. In the case of a Higgs decaying into two b quarks, for instance, the variable of interest would naively be τ_2. It happens nevertheless that there are also QCD jets with small values of τ_2. On the other hand, it is expected that a Higgs candidate has larger values of τ_1, but again, the same can happen for QCD jets. QCD jets with large τ_1 usually also have large values of τ_2, so, in reality, for this case, a better discriminating variable is the ratio $\frac{\tau_2}{\tau_1}$. In the same line of thought, the best variable to look for a hadronically-decaying top is the ratio $\frac{\tau_3}{\tau_2}$.

For example, the $\frac{\tau_2}{\tau_1}$ values for the Higgs candidates versus its masses are presented in Figures 4.3 and 4.4, for the $t\bar{t}H$ and $t\bar{t}b\bar{b}$ processes, respectively. As it is possible to see, $t\bar{t}H$ shows a peak at low values of $\frac{\tau_2}{\tau_1}$ in the Higgs mass region, while for $t\bar{t}b\bar{b}$ the values are more dispersed in mass and towards large values of the ratio.

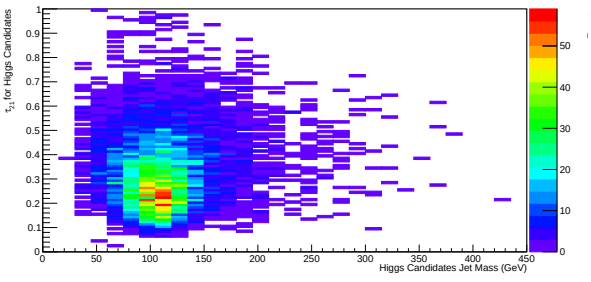

Figure 4.3: τ_{21} for Higgs candidates in $t\bar{t}H$ simulated events.

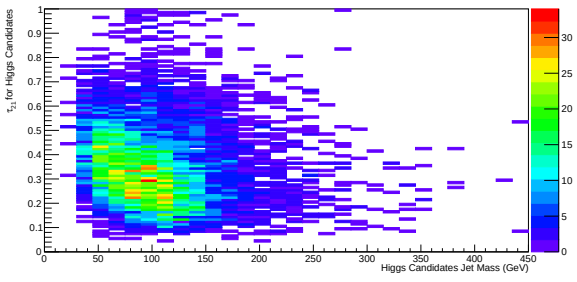

Figure 4.4: τ_{21} for Higgs candidates in $t\bar{t}b\bar{b}$ simulated events.

4.4.2 Energy Correlation Functions

In order to exploit jet substructure information, another option is to implement energy correlation functions [54, 55]. In this case, the sensitivity to the N-prong substructure is achieved through $(N+1)$-point correlation functions, that use energy and angle information of the jet. Moreover, contrary to N-subjettiness, these functions don't require subjet finding methods. In addition, these functions account

for an angular exponent β, that needs to be bigger than 0 in order to be infrared and collinear safe.

The Energy Correlation Functions (ECF) are defined in Equation 4.4, where the index i_k stands for a particle and p_{T,i_k} is its transverse momentum, J is the input jet, and $\Delta R_{f,k}$ is the distance between particles *f* and *k* in the rapidity-azimuth plane.

$$ECF(N,\beta) = \sum_{i_1<i_2<\cdots<i_N \in J} \left(\prod_{a=1}^{N} p_{T,i_a} \right) \left(\prod_{b=1}^{N-1} \prod_{c=b+1}^{N} \Delta R_{i_b,i_c} \right)^{\beta} \tag{4.4}$$

From these functions it is possible to define the ratio presented in Equation 4.5, that exploits the fact that, for a jet with N subjets, $ECF(N+1,\beta) \approx 0$, being much smaller than $ECF(N,\beta)$. This ratio is expected to behave like τ_N.

$$r_N^{\beta} = \frac{ECF(N+1,\beta)}{ECF(N,\beta)} \tag{4.5}$$

Nevertheless, the most interesting variables are the double ratios. In this thesis two of these double ratios were used, C_2 and D_2, in order to discriminate between Higgs candidates and background objects. These two variables are dimensionless and are defined in Equations 4.6 and 4.7.

$$C_2 = \frac{r_2^{\beta}}{r_1^{\beta}} = \frac{ECF(3,\beta)\,ECF(1,\beta)}{ECF(2,\beta)^2} \tag{4.6}$$

$$D_2 = \frac{r_2^{\beta}\, r_0^{\beta}}{(r_1^{\beta})^2} = \frac{ECF(3,\beta)\,ECF(1,\beta)^3}{ECF(2,\beta)^3} \tag{4.7}$$

As for the N-subjettiness ratio $\frac{\tau_2}{\tau_1}$, the smaller the values for C_2 and D_2 are, the most likely is for a jet to have 2 hard substructures. The optimal values for the β parameter then depend on the resonance mass, and usually two regimes are tested, with $\beta = 0.5$, for masses around or above 200 GeV, and $\beta = 2.0$, for masses around or bellow 100 GeV. On the other hand, the p_T of the objects also influences the discriminating power, with higher values of β being preferred in cases of higher boosts.

As example, the distributions for D_2 for the Higgs candidates versus their masses, for $t\bar{t}H$ and $t\bar{t}b\bar{b}$, are presented in Figures 4.5 and 4.6, respectively, using $\beta = 2.0$. Again it is possible to see that the background is dispersed toward larger values of the variable, compared to signal.

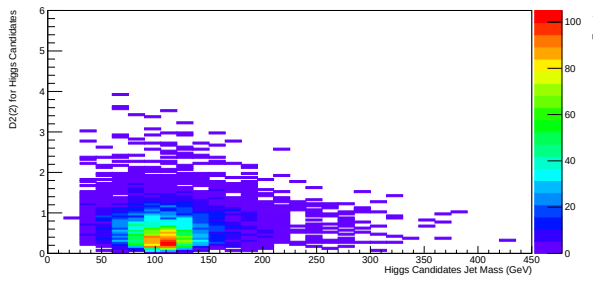

Figure 4.5: D_2 for Higgs candidates for $t\bar{t}H$, with $\beta = 2.0$.

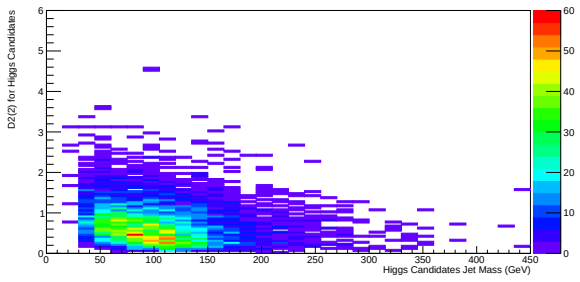

Figure 4.6: D_2 for Higgs candidates for $t\bar{t}b\bar{b}$, with $\beta = 2.0$.

4.5 Tagging of boosted top quarks

In order to search for a boosted hadronic top quark, one possibility is to use a dedicated algorithm, such as HEPTOPTAGGER2 [56, 57]. The main idea behind this algorithm is to receive as input a fat jet and then look for three hard substructures inside it, corresponding to the b quark and to the two quarks resulting of the decay of the W boson.

HEPTOPTAGGER2 is designed to take a C/A jet (jet *i*), of radius R=1.8, and to undo the last step of the clustering, such that it has two subjets, jets i_1 and i_2, with $m_{i_1} > m_{i_2}$. The so-called mass drop condition is then applied, which consists in first checking the relation between the mass of the hardest subjet and the initial jet. If $m_{i_1} < f_{drop}\, m_i$, with the mass drop threshold $f_{drop} = 0.8$, both subjets are kept, while otherwise only jet i_1 is kept, with i_2 being considered a jet coming from pile-up or an underlying event. Then, requiring a minimum mass for the subjets of 30 GeV, the remaining subjets are further decomposed or added to a collection of relevant substructures.

From this collection the three hardest subjets are selected and filtered with the C/A algorithm, with the filtering removing the pile-up and underlying event contamination. In the filtering process the algorithm keeps up to five hard substructures, in order to have into consideration gluon radiation of two quarks. These five subjets are then reclustered into three, that should correspond to the top decay products, and the invariant mass of the triplet is computed. The candidate only proceeds in the algorithm if this mass falls in the [150,200] GeV mass window.

Then, assuming the particles resulting of the top decay are massless, that is $p_i^2 \approx 0$, the expression in Equation 4.8 holds, where m_t is the top mass, m_{123} is the triplet mass and m_{ij} is the invariant mass of particles i and j.

$$m_t^2 \equiv m_{123}^2 = (p_1 + p_2 + p_3)^2 = (p_1 + p_2)^2 + (p_1 + p_3)^2 + (p_2 + p_3)^2 = m_{12}^2 + m_{13}^2 + m_{23}^2 \tag{4.8}$$

Considering that one of the m_{ij} should be equal to the W boson mass, there are then still two degrees of freedom left to fully describe the kinematics, which is solved by introducing two variables, $\frac{m_{23}}{m_{123}}$ and $arctan\left(\frac{m_{13}}{m_{12}}\right)$. The top candidate is then required to satisfy specific conditions on the mass plane of these two variables, such that $t\bar{t}$ processes are enhanced against pure QCD and W+jets backgrounds.

Finally, the top candidate is required to have a minimum p_T of 200 GeV for consistency, and in the case of more than one top candidate, the algorithm chooses the one with its mass closest to the top mass.

4.6 Tagging of boosted Higgs

As for the top quark case, it is possible to exploit substructure information of jets in order to find a boosted Higgs decaying into two bottom quarks. One of the possibilites is to use the BDRS Higgs Tagger [58], that tries to find two hard substructures inside a initial jet of radius R. The complete procedure is shown

in Figure 4.7.

Similarly to HEPTOPTAGGER2, the algorithm undoes the last step of the clustering, breaking the initial jet into two subjets, jets i_1 and i_2, with $m_{i_1} > m_{i_2}$. The mass drop condition is then required, with a mass drop threshold $f_{drop} = 0.9$ in this case. Moreover, in order to keep the two subjects, the decay must not be too asymmetric, which is achieved by demanding that $\frac{min(p_{T_1}^2, p_{T_2}^2)}{m_i^2} \Delta R_{i_1,i_2}^2 > y_{cut}$, with $y_{cut} = 0.09$. If any of these criteria fails then only jet i_1 is kept, and the iterative procedure starts again.

Taking the BDRS Higgs candidate, the general procedure is to then ask for a b-tag in each of its subjets. In addition, the two subjets must have each a $p_T > 30$ GeV and $|\eta| < 2.5$. Moreover, the candidate is then filtered, again to remove PU and UE contamination, keeping up to three hard substructures, to account for gluon radiation of one of the b quarks.

Finally, a minimum p_T of 200 GeV (or equivalent) for the Higgs candidate is required.

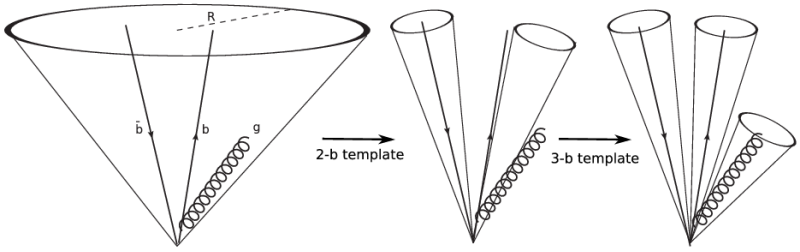

Figure 4.7: Higgs tagging procedure [58].

Chapter 5

State of the Art

Several searches for $t\bar{t}H$ production have been conducted in the last years, with a (> 5 sigma) observation being made in 2018 by ATLAS and CMS. This was achieved by attacking the process on many fronts, with different strategies associated to each case. Nevertheless, there is still work to do, as the analysis sensitivity is generally limited by large background modelling uncertainties or lack of statistics, and a precise measurement of this process is required in order to understand if the coupling of the Higgs boson to the top quark is indeed as predicted in the SM or not.

In this chapter an overview of the experimental analyses performed by ATLAS and CMS is given in the first section, with different Higgs decay modes being explored. In the second section, theoretical studies proposing different strategies for current or future analyses are presented.

5.1 Experimental Status

Throughout the year of 2018, the ATLAS and CMS experiments announced the observation of the production of the Higgs boson in association with two top quarks [10, 11]. This observation, along with the observation of the Higgs decaying into bottom quarks and taus [14, 15], proves the coupling of the Higgs boson to the third generation of fermions, results of the utmost importance in the process of attesting the SM.

The observation of the $t\bar{t}H$ process by ATLAS was obtained through the combination of several analyses, targeting different Higgs boson decay modes: $H \to b\bar{b}$ [17], $H \to WW/ZZ$ and $H \to \tau\tau$ (multilepton or ML channel) [23], $H \to \gamma\gamma$ [10, 59] and $H \to ZZ^* \to 4l$ [10, 60]. The analyses for the $H \to b\bar{b}$ and ML decay modes used 36.1 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV collected in 2015 and 2016, while the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4l$ analyses were conducted with an integrated luminosity of 79.8 fb^{-1}, and includes the data collected in 2017.

Each of these different decay modes has its own advantages and particular challenges, complementing each other to some extent. Namely, while the $H \to b\bar{b}$ decay is the one associated with larger signal statistics but lower purity, the ML channel has a lower associated background, and, on the other hand, the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4l$ decays have events with higher signal purity but lower signal rates.

For the $H \to b\bar{b}$ decay mode [17], the signal is modelled using MADGRAPH5_AMC@NLO at NLO and using a set of NLO parton density functions (PDF), for consistency. Moreover, top quarks are decayed using MADSPIN. The $t\bar{t}$ background is generated using POWHEG-BOX V2 NLO at next-to-next-to-leading order (NNLO) in QCD including resummation of next-to-next-to-leading logarithmic (NNLL) soft gluon terms and it is categorised according to the number and flavour of additional jets in the event. The $t\bar{t}V$ ($V = W, Z$ bosons) backgrounds are also generated at NLO using MADGRAPH5_AMC@NLO. The parton shower and hadronisation are modelled by PYTHIA8.2 for all samples.

Events are divided in single-lepton and dilepton channels, depending on the number of light charged leptons ($l = e, \mu$) in the final state as result of the decays of the top quarks. In these events leptons and jets are reconstructed and have to pass specific requirements in order to suppress background. Furthermore jets are clustered using the anti-k_t algorithm with a radius parameter of 0.4 and a minimum p_T of 25 GeV, and are b-tagged via an algorithm using multivariate techniques and according to different working points depending on the tightness of the criteria they pass (*loose*, *medium*, *tight*, *very-tight*).

In the single-lepton channel a jet re-clustering is performed using the same algorithm but with $R = 1.0$ in order to identify boosted top quark and Higgs boson candidates. In this channel events are then categorised as 'boosted' if they have at least one boosted top, with $p_T > 250$ GeV and one *loose* b-jet inside, one boosted Higgs, with $p_T > 200$ GeV and two *loose* b-tags, and an additional *loose* b-jet. Otherwise they are categorised as 'resolved' if they have at least five jets with at least two of them b-tagged with a *very-tight* working point. For both cases the reconstructed lepton must have $p_T > 27$ GeV. Events in the dilepton channel are required to have two charged leptons with opposite-sign electric charge, one with $p_T > 27$ GeV and the other with $p_T > 10$ GeV, and at least three jets, of which at least two must be b-tagged with the *medium* working point.

Events are then classified into non-overlapping regions based on the total number of jets, as well as the number of b-tagged jets and the quality of the b-jet identification. Moreover, events in the boosted single-lepton channel are not further categorised. Regions with enhanced $t\bar{t}H$ and $t\bar{t}b\bar{b}$ content are labelled *signal regions*, in which multivariate analysis (MVA) techniques are used. Other regions are labelled *control regions* and are used to constrain backgrounds and different sources of systematic uncertainties, and no multivariate techniques are applied to them.

The MVA techniques include classification boosted decision trees (BDTs), that use kinematic variables related to the b-tagged jet pair, and reconstruction BDTs that use variables related to the Higgs candidate and leptons, b-jets or tops. These are used to identify the most likely combination of leptons and jets that correspond to top quarks or Higgs boson decays, and improve the separation between signal and background.

These regions are then combined in a profile likelihood fit to test for the presence of a signal. Results for the signal strength $\mu = \sigma/\sigma_{SM}$ are presented in Figure 5.1 where all the numbers are obtained from a simultaneous fit in the two channels, but the measurements in the two channels separately are obtained keeping the signal strengths uncorrelated, while all the nuisance parameters (that encode the effects of systematic uncertainties) are kept correlated across channels. The combined signal strength is $\mu = 0.84^{+0.64}_{-0.61}$. The largest contribution to the errors comes from the systematic uncertainties, surpassing

the statistical parcel, that result of the imperfect modelling of the background, especially the $t\bar{t} \geq 1b$ production process. A signal strength larger than 2.0 is excluded at the 95% confidence level. The expected limits are calculated using the background estimate after the fit to the data. An excess of events over that expected for the "background-only hypothesis" is found with an observed (expected) significance of 1.4 (1.6) standard deviations.

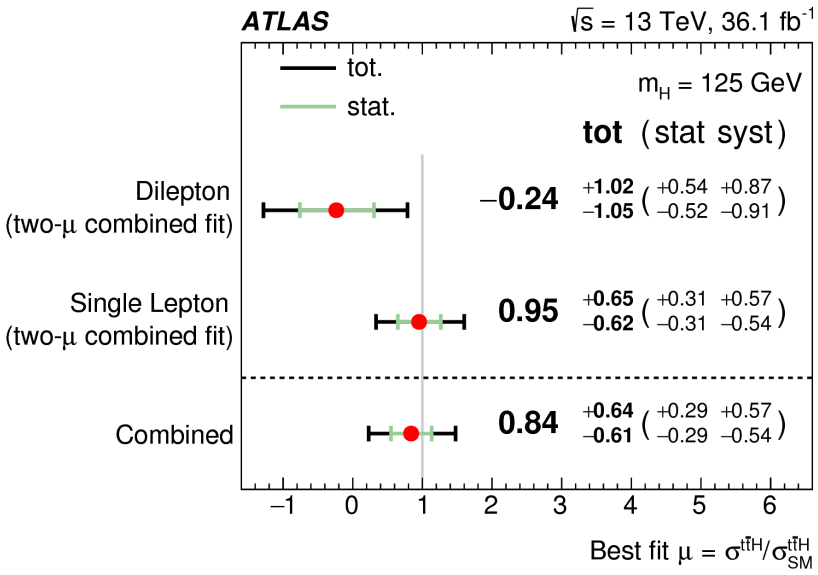

Figure 5.1: Signal strength measurements in the individual channels and for the combination, for $H \to b\bar{b}$ decay mode [17].

For the ML decay mode [23] the signal and backgrounds are generated as for the $H \to b\bar{b}$ decay mode. These decays are detected in searches for events with either a pair of same-sign charged leptons, or three or more charged leptons. Seven final states are analysed, categorised by the number and flavor of charged-lepton candidates, as presented in Figure 5.2. In these channels specific selections are applied in order to suppress the background.

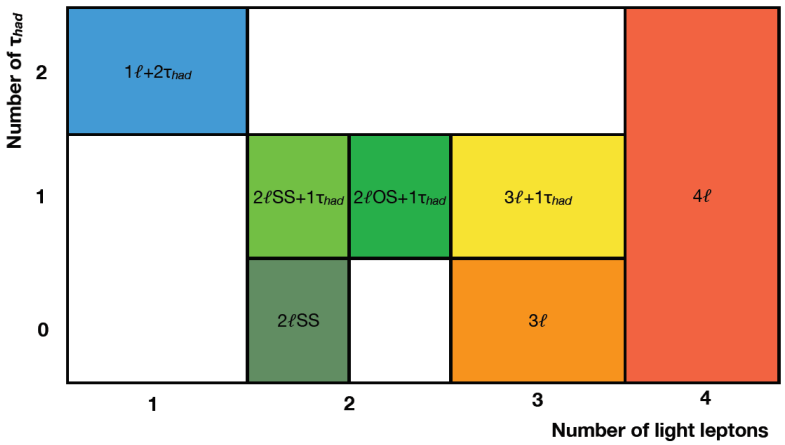

Figure 5.2: Channels used in the analysis organised according to the number of selected light leptons and τ_{had} candidates [23].

Light leptons are required to come from the interaction point in order to distinguish between prompt leptons, that come from the signal, and non-prompt leptons, resulting of hadronic jets and heavy-flavor hadron decays, and that come from the background. For the electrons it is also necessary to suppress contributions coming from photon conversions (fake electrons). For the hadronically decaying τ-lepton candidates it is necessary to reject the contributions from jet backgrounds. Furthermore, for the τ candidates three working points are defined, according to the reconstruction and identification efficiency. The loosest working point is used only for background estimates. In addition, the leading lepton must have generally $p_T > 20$ GeV and the sub-leading lepton(s) $p_T > 10$ GeV. Jets are reconstructed and are clustered using the anti-k_t algorithm with a radius parameter of 0.4 and $p_T > 25$ GeV, and are b-tagged. MVA techniques are applied in order to further reduce non-prompt leptons and to reject electrons reconstructed with an incorrect electric charge. These techniques use essentially track properties as input.

From the different final states twelve categories are defined: eight signal regions and four control regions. In the control regions comparisons between data and simulation are used to confirm the background modelling.

A maximum-likelihood fit is performed on all these twelve categories simultaneously to extract the signal strength and the results of the fit are presented in Figure 5.3. The obtained signal strength is $\mu = 1.6^{+0.5}_{-0.4}$. An excess of events over the expected background from SM processes is found, which is interpreted as an observed significance of 4.1 standard deviations for a SM Higgs boson of mass 125 GeV. The expected significance for a SM Higgs boson is 2.8 standard deviations.

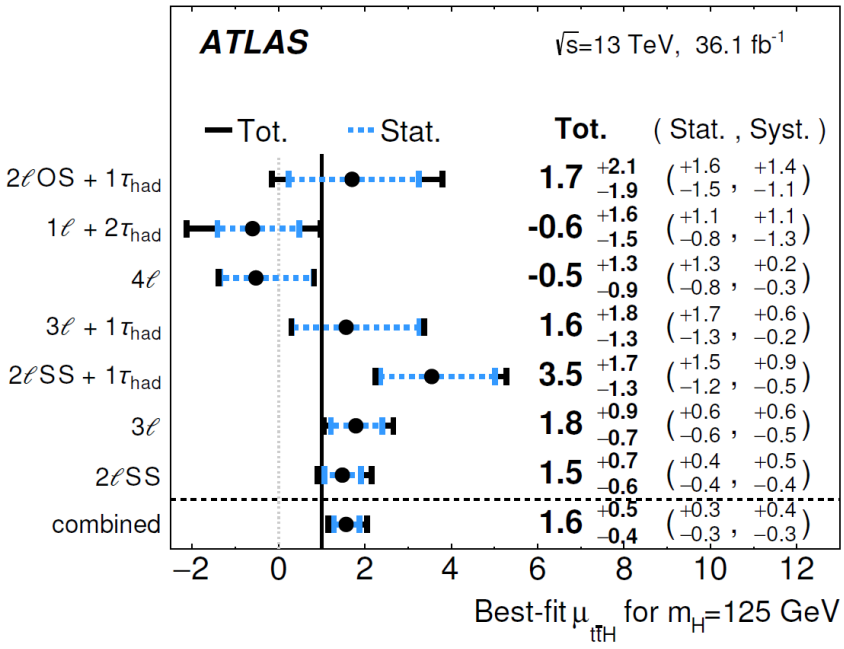

Figure 5.3: Signal strength measurements in the individual channels and for the combination, for ML decay mode [23].

For the diphoton decay mode [10, 59] signal events are generated as above, but for the continuum $\gamma\gamma$ background the SHERPA generator at LO is used. These background events are generated with

additional partons or vector bosons.

The photon candidates are reconstructed using the energy clusters in the electromagnetic calorimeter. Electrons and muons are also reconstructed. Several selection cuts are applied in order to distinguish photons from electron candidates and to reduce the background contamination, primarily associated to neutral hadrons (mainly π^0) in jets decaying into photon pairs. Two working points are defined: a *loose* criterion, primarily used for triggering and pre-selection purposes, and a *tight* criterion. Furthermore, jets are reconstructed and are clustered using the anti-k_t algorithm with a radius parameter of 0.4 and requiring a $p_T > 25$ GeV, and are b-tagged.

Events are then required to have at least two isolated photon candidates with $p_T > 35$ GeV and $p_T > 25$ GeV, that satisfy the *loose* photon identification criteria. In addition, the leading and sub-leading photon candidates must have $\frac{p_T}{m_{\gamma\gamma}} > 0.35$ and $\frac{p_T}{m_{\gamma\gamma}} > 0.25$, respectively. Moreover, the event must have an additional b-tagged jet. Two channels are further defined, hadronic and leptonic, depending if both top quarks decay hadronically or semi-leptonically. Each channel has dedicated BDTs, with the goal of further rejecting the background, that receive as input the four-momentum of photons, leptons and jets.

The Higgs boson signal is measured through a maximum-likelihood fit to the diphoton invariant mass spectrum in the range 105 GeV $< m_{\gamma\gamma} < 160$ GeV. The mass range is chosen to be large enough to allow a reliable determination of the background from collision data, using sidebands around the Higgs mass peak, and to avoid large uncertainties associated to it. The parameters of the model that define the shape of the signal distribution are determined through fits to the simulated signal samples. The background distribution comes from studies of background control samples directly obtained from data.

The observed invariant mass distribution of the selected diphoton pairs and the result of the signal-plus-background fit to this spectrum are presented in Figure 5.4. In this Figure events are weighted by $\ln(1 + S_{90}/B_{90})$, with S_{90} (B_{90}) standing for the expected $t\bar{t}H$ signal (background) in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. Furthermore, the error bars reflect 68% confidence intervals of the weighted sums. The fit assumes a signal strength $\mu = 1.4$, with an observed significance of 4.1 standard deviations, compared to an expectation of 3.7 standard deviations.

The last decay mode analysed by ATLAS is with the $H \to ZZ^* \to 4l$ [10, 60], where Z^* stands for an off-shell Z boson. For this analysis special selection cuts are applied in order to avoid overlaps between this channel and the ML one. Moreover, events must have four isolated leptons, in pairs of opposite charge but with the same flavour (four electrons, four muons, or two electrons and two muons). Jets are reconstructed and are clustered using the anti-k_t algorithm with a radius parameter of 0.4 and $p_T > 30$ GeV. Besides the four leptons the event is also required to have a b-tagged jet. Events are then divided in hadronic and leptonic regions, as for the diphoton decay mode, and dedicated BDTs are again used to distinguish between signal and background. These BDTs receive as input information regarding differences in pseudorapidity and momenta of the jets and leptons.

The mass of the four leptons must then be in the range $[115, 130]$ GeV, and a likelihood fit is performed. For this analysis no event was observed, against an expected significance of 1.2 standard deviations.

The signal strength for all mentioned Higgs decay modes in a combined fit is shown in Figure 5.5,

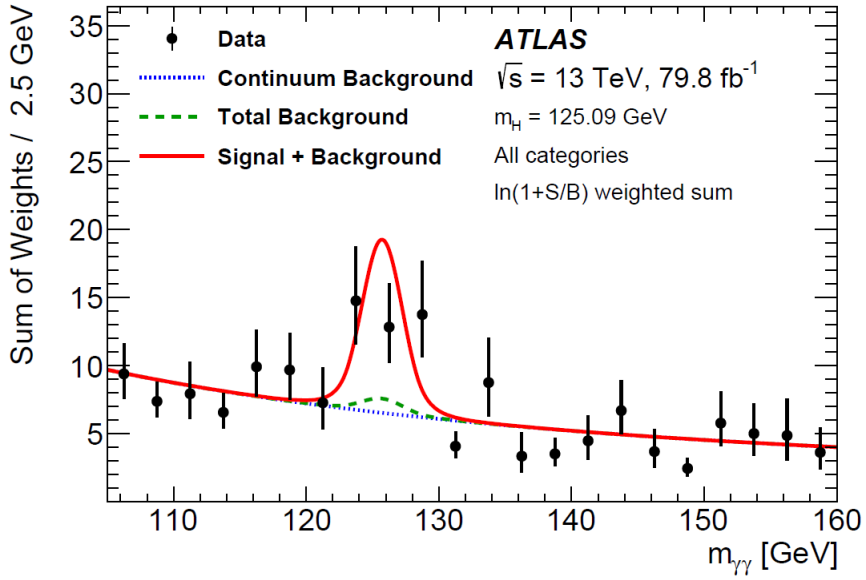

Figure 5.4: Observed invariant mass distribution of the selected diphoton pairs and signal-plus-background fit [10].

with $\mu = 1.32^{+0.28}_{-0.26}$. The combination has an observed significance of 5.8 standard deviations, compared to an expectation of 4.9 standard deviations. ATLAS combined these searches with previous analyses using 4.5 fb^{-1} at $\sqrt{s} = 7$ TeV and 20.3 fb^{-1} at $\sqrt{s} = 8$ TeV, which results in and observed (expected) significance of 6.3 (5.1) standard deviations.

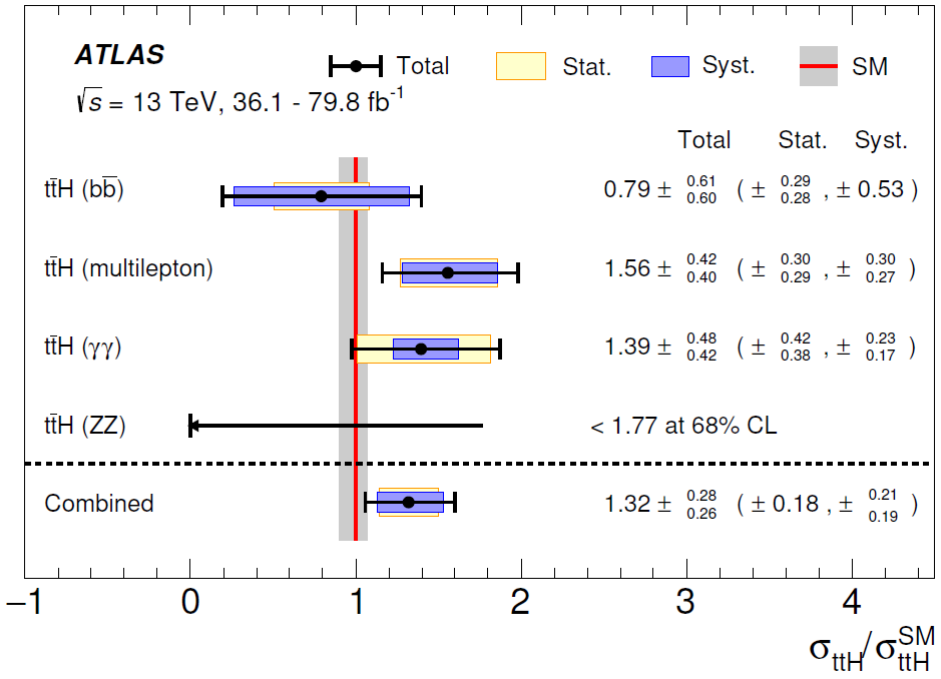

Figure 5.5: Signal strength measurements for all mentioned Higgs decay modes [10].

The cross section for this process is also measured, and its value for a center-of-mass energy of 8

and 13 TeV is presented in Figure 5.6, against the theoretical prediction of the SM. As it is possible to see, the measurements are at this point in agreement with the SM.

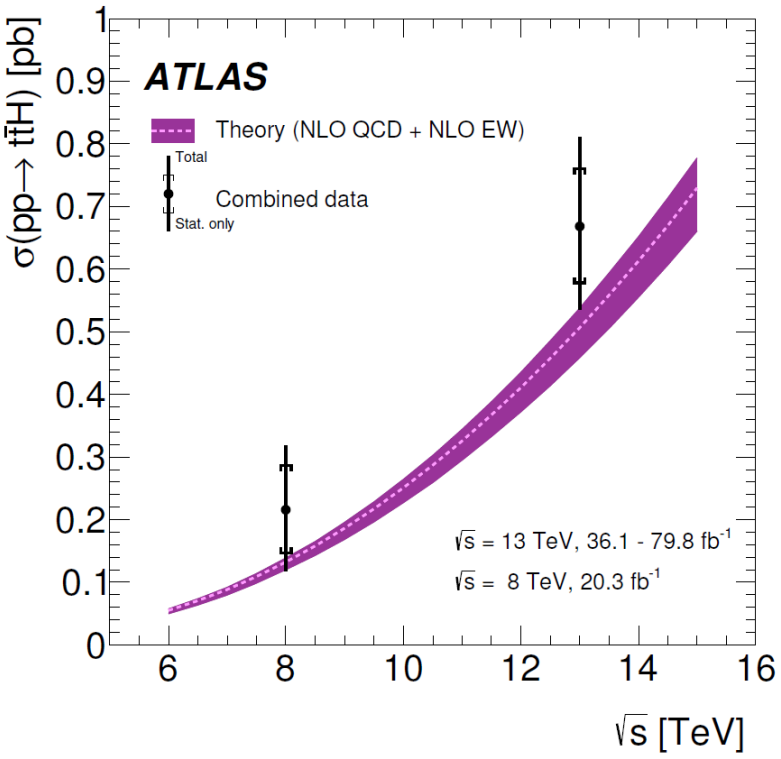

Figure 5.6: $t\bar{t}H$ cross section measurements at $\sqrt{s} = 8$ and $\sqrt{s} = 13$ TeV [10].

The $t\bar{t}H$ process, as mentioned above, was also observed by CMS in 2018 [11], using 35.9 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV collected in 2016. As in ATLAS, several Higgs decay modes are targeted: $H \to b\bar{b}$ [61, 62], $H \to WW/ZZ$ and $H \to \tau\tau$ (ML channel) [63], $H \to \gamma\gamma$ [64] and $H \to ZZ^* \to 4l$ [65].

These analyses share common points with the ones performed in ATLAS, with the obvious differences of being done in a slightly different detector. Moreover, the remaining main differences will be stated below. Another general difference in these studies, besides the detector, is the object reconstruction, that is perfomed using the Particle-Flow technique [66], that combines signals from all subdetectors. This way, the reconstruction performance is improved, as the technique identifies individual particle candidates coming from the collisions.

For the $H \to b\bar{b}$ decay mode, CMS divided the analyses in events where one or both tops decay semi-leptonically [61], and where both tops decay hadronically [62]. For the first case, apart for selection cuts equivalent to ATLAS, CMS also requires that, for the dileptonic channel, the invariant mass of the two leptons must be outside the mass window $[76, 106]$ GeV, to suppress $Z + jets$ events. Moreover, the missing transverse momentum, defined as the projection of the negative vector sum of the momenta of all reconstructed PF objects in an event on the plane perpendicular to the beams [61], must be above 20 GeV and 40 GeV for the single-lepton and dilepton channel, respectively, to take into account the neutrinos resulting from the top quark decays and further suppress backgrounds. In addition, for both channels, events must have at least four jets, with at least three of them being b-tagged.

Events are then further divided in categories and several MVA techniques are applied: BDTs, Deep Neural Networks (DNN) and Matrix Element Method (MEM) discriminant [67]. The MEM discriminant is defined as the ratio of the probability density values associated to the signal hypotheses, estimated event by event from the calculated LO signal matrix element. All these algorithms receive input variables related to the kinematics of the different particles in the event. Finally, a simultaneous likelihood fit is performed. The analysis strategy is presented in Figure 5.7.

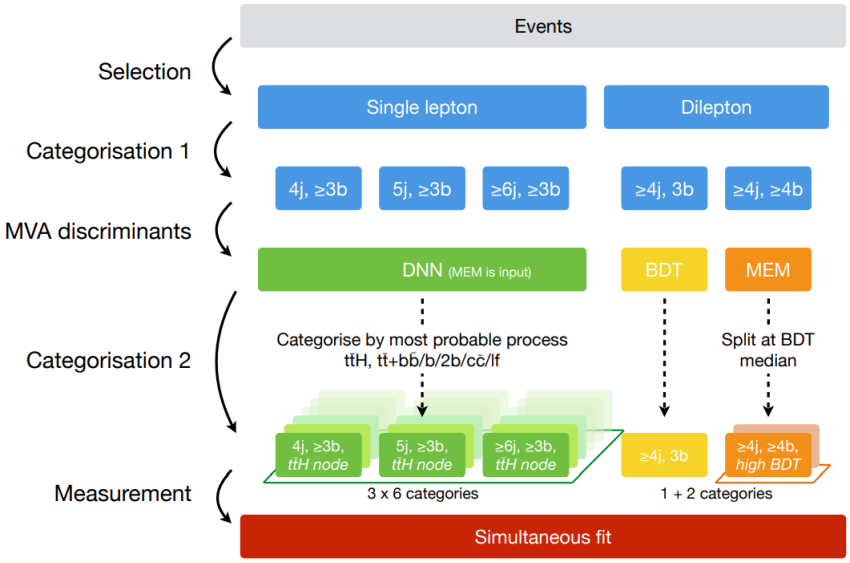

Figure 5.7: CMS Analysis strategy for $t\bar{t}H(b\bar{b})$ single-lepton and dilepton channels [61].

The results of the fit are presented in Figure 5.8. The obtained signal strength is $\mu = 0.72^{+0.45}_{-0.45}$, with an observed significance of 1.6 standard deviations, compared to an expectation of 2.2 standard deviations. Again, the main systematic uncertainties come from the modelling of the $t\bar{t}+hf$ backgrounds, where hf stands for heavy flavour quarks (charm, bottom).

The all-hadronic $H \to b\bar{b}$ analysis has a careful event selection in order to reject events that contain possible leptons coming from top quarks, in order to avoid overlap between the different regions. Moreover, events must have at least six jets with $p_T > 40$ GeV, and with at least one being b-tagged. The scalar sum of the transverse momentum of all jets in the event must be above 400 GeV. Further selections are applied in order to ensure that part of the jets come from W bosons resulting of the top quarks decays. Events are then categorised depending on the jet and b-jet multiplicity, and signal and control regions are defined.

A likelihood fit is performed on all the categories simultaneously to extract the signal strength and the results of the fit are presented in Figure 5.9. The obtained signal strength is $\mu = 0.9^{+1.5}_{-1.5}$. The observed and expected upper limits are $\mu < 3.8$ and $\mu < 3.1$, respectively, at 95% confidence levels.

For the ML decay mode [63] CMS also divides the events in several categories, depending on the lepton multiplicity and charge, and number of hadronic taus. The selection is equivalent to the one perfomed in ATLAS. The result of the likelihood fit for the signal strength is presented in Figure 5.10. The obtained signal strength is $\mu = 1.23^{+0.45}_{-0.45}$. An excess of events over the expected background from

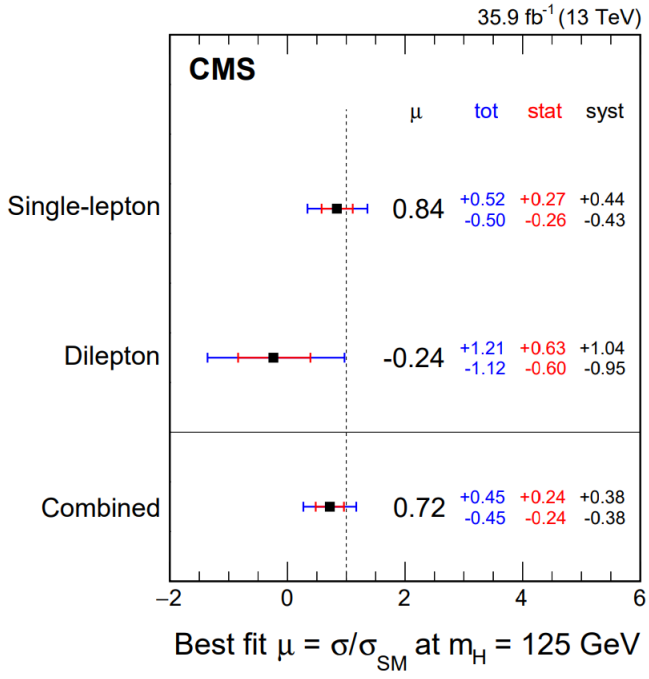

Figure 5.8: Signal strength measurements in the individual channels and for the combination, for $t\bar{t}H(b\bar{b})$ single-lepton and dilepton channels [61].

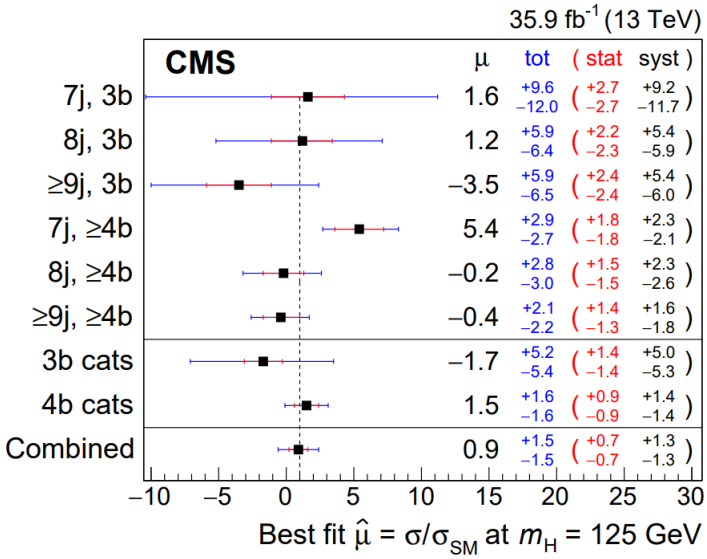

Figure 5.9: Signal strength measurements in the individual channels and for the combination, for $t\bar{t}H(b\bar{b})$ all-hadronic channel [62].

SM processes is found, which is interpreted as an observed (expected) significance of 3.2 (2.8) standard deviations.

For the $H \to \gamma\gamma$ decay mode CMS divided the search into leptonic and hadronic regions. Events must have at least one lepton or none, depending on the region. In these cases the mass windows ranges from 100 GeV to 180 GeV. Apart from that, the analysis strategy is very similar to the one implemented in ATLAS. The $m_{\gamma\gamma}$ distribution and signal-plus-background fit for both regions are presented in Figure 5.11.

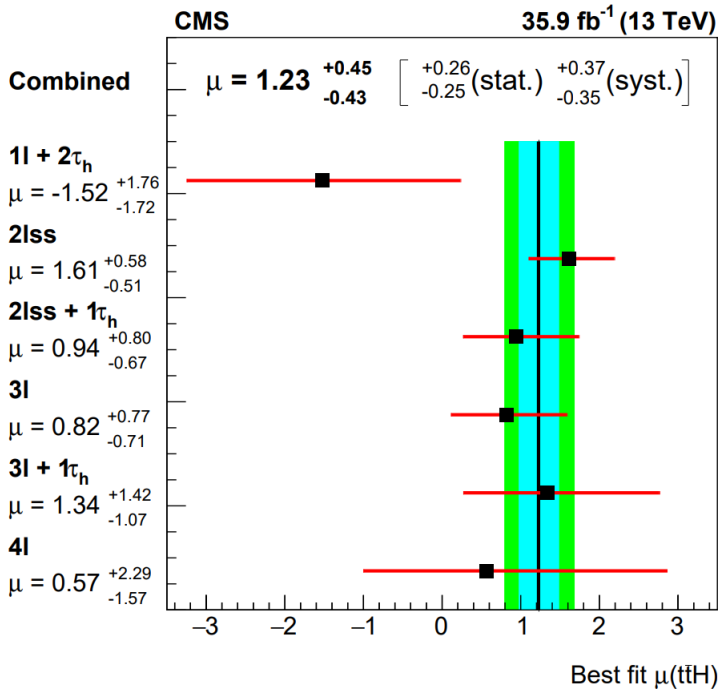

Figure 5.10: Signal strength measurements in the individual channels and for the combination, for ML channel [63].

The best-fit value is $\mu_{t\bar{t}H} = 2.2^{+0.9}_{-0.8}$ and corresponds to a 3.3 σ excess with respect to the background-only hypothesis, and is compatible with the signal strength prediction for the SM within 1.6σ .

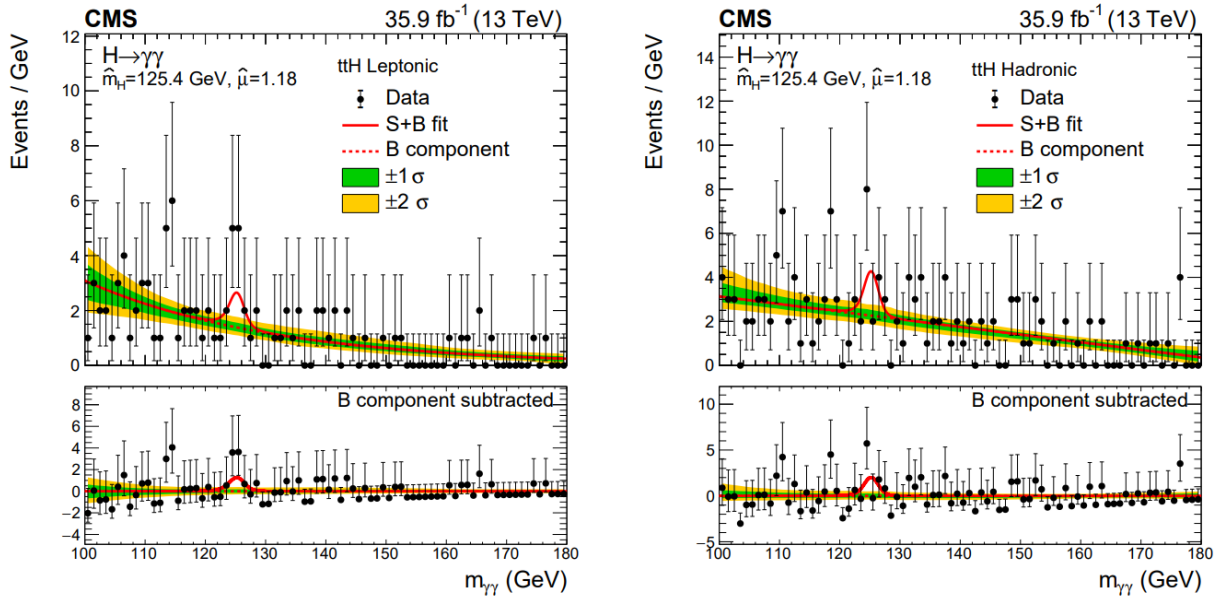

Figure 5.11: Observed invariant mass distributions of the selected diphoton pairs and signal-plus-background fit [64].

On the other hand, for the $H \rightarrow ZZ^* \rightarrow 4l$ decay mode [65], CMS requires, apart from the four isolated leptons, at least four jets, with at least one of them b-tagged, or at least one additional lepton. The rest of the analysis is equivalent to the ATLAS one. As in ATLAS, no event was observed, against

an expected significance of 1.0 standard deviations.

CMS then combined these analyses with previous analyses using 5.1 fb^{-1} at $\sqrt{s} = 7$ TeV and 19.7 fb^{-1} at $\sqrt{s} = 8$ TeV. The signal strength for all mentioned Higgs decay modes in a combined fit is shown in Figure 5.12, with $\mu = 1.26^{+0.31}_{-0.26}$. The combination has an observed significance of 5.2 standard deviations, compared to an expectation of 4.2 standard deviations.

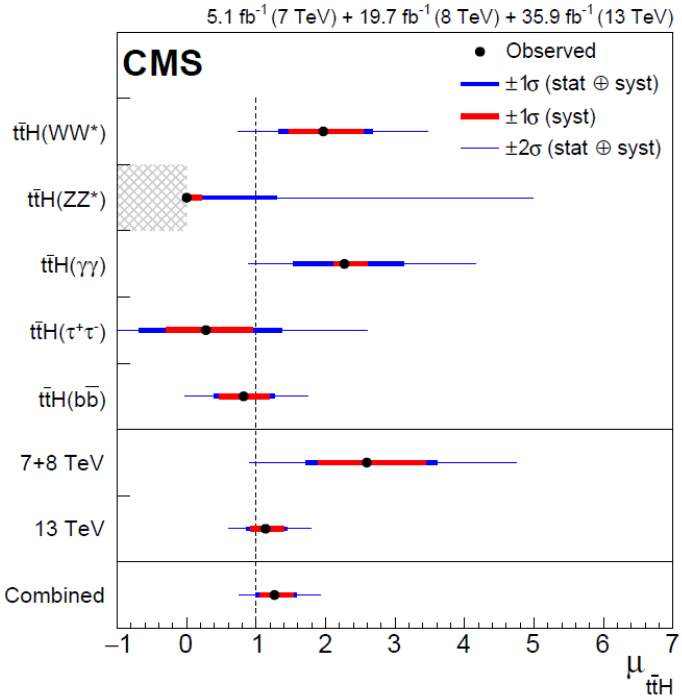

Figure 5.12: Signal strength measurements for combined Higgs decays [11].

5.2 Phenomenological Studies

In addition, others studies have been carried out based on the full reconstruction of the events by applying a kinematic fit. Furthermore angular distributions and asymmetries have been proposed in order to improve discrimination of $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal events over the dominant background, $t\bar{t}b\bar{b}$, in the semileptonic and dileptonic channels [68–70].

$$(p_{l+} + p_{\nu})^2 = m_W^2 \tag{5.1}$$

$$(p_{l-} + p_{\overline{\nu}})^2 = m_W^2 \tag{5.2}$$

$$(p_{W+} + p_b)^2 = m_t^2 \tag{5.3}$$

$$(p_{W-} + p_{\overline{b}})^2 = m_{\overline{t}}^2 \tag{5.4}$$

$$p_{\nu}^x + p_{\overline{\nu}}^x = \not{E}_x \tag{5.5}$$

$$p_{\nu}^y + p_{\overline{\nu}}^y = \not{E}_y \tag{5.6}$$

In these studies signal and background events are generated at LO and NLO with MG5_aMC@NLO and using MADSPIN for the decays, PYTHIA6/PYTHIA8 for showering and hadronisation and DELPHES3 for fast simulation of the ATLAS detector. Events are fully reconstructed. The neutrino reconstruction for both channels is based on constraints such as the ones presented in Equations 5.1-5.4. Moreover, for the dileptonic channel further constraints are needed, presented in Equations 5.5 and 5.6, where $\not{E}$ stands for missing transverse energy. In addition, p_ζ^x and p_ζ^y correspond to the momentum of particle $\zeta = (l^\pm, \nu, \overline{\nu}, W^\pm, b, \overline{b}, t, \overline{t})$ in the x and y axes.

From these events, angular distributions are constructed and asymmetries are defined as in Equation 5.7, where x_Y is a double angular product and $N(x_Y > 0)$ and $N(x_Y < 0)$ correspond to the total number of events in the corresponding angular distribution with x_Y above and below zero, respectively.

$$A_{FB} = \frac{N(x_Y > 0) - N(x_Y < 0)}{N(x_Y > 0) + N(x_Y < 0)} \tag{5.7}$$

It is shown that, even after going to NLO, event selection and full kinematic reconstruction, the shape of the new angular distributions and asymmetries is largely preserved and can be used to discriminate between the different types of signals (scalar vs. pseudo-scalar) and the dominant irreducible SM background, $t\bar{t}b\bar{b}$. One example is presented in Figure 5.13. From the $t\bar{t}$ pair rest frame one can move to the top quark rest frame, and measure the angle (θ_{l+}) between the positively charged lepton (l^+) and the axis defined by the $t\bar{t}$ pair (where the top and anti-top quarks are back-to-back). Moving then to the anti-top quark rest frame, one can correspondingly measure the angle (θ_{l-}) between the negatively charged lepton (l^-) and the axis defined by the $t\bar{t}$ pair. From these angles one arrives, for instance, at the presented distributions for different processes.

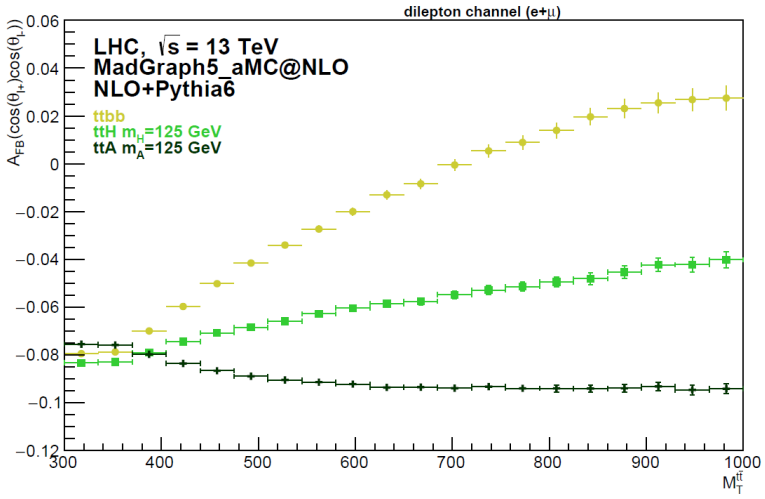

Figure 5.13: Asymmetries for $cos(\theta_{l+})cos(\theta_{l-})$ versus a lower cut on the value of $M_T^{t\bar{t}}$.

On the other hand, theorists have recently suggested a different approach in terms of analysis to the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) production process [56, 71]. In this study events are generated at LO at 100 TeV for the semileptonic channel. For the generation MADGRAPH5_aMC@NLO is used with PYTHIA8 for showering and hadronisation and DELPHES3 for fast detector simulation. An integrated luminosity of

$20\,\text{ab}^{-1}$ is assumed. The analysis strategy is based on the analysis performed for the discovery of the Higgs boson in the $H \to \gamma\gamma$ channel and consists of defining, using boosted events, side bands around the signal region in the distribution of the invariant mass of the $b\bar{b}$ pair. These side bands then control the $t\bar{t}b\bar{b}$ and $t\bar{t}$+jets backgrounds, as well as a second mass peak from the $t\bar{t}Z$ process.

The proposed event selection starts by requiring an isolated lepton with a minimum p_T of 15 GeV and $\eta < 2.5$. The event particle-flow objects (that combine information from different parts of the detector) are then clustered into 'fat' jets using the C/A algorithm with $R = 1.8$ and are required to have $p_T > 200$ GeV. If events have at least two fat jets these are passed to a top-tagging algorithm, HEPTOPTAGGER2 [56], in order to identify the tops. Cuts are then performed in order to test if the jet is a top candidate. One of these cuts is on the N-subjettiness ratio $\frac{\tau_3}{\tau_2} < 0.8$, after filtering the jet, in order to suppress QCD background. Also, the top candidate is required to be within $\eta < 4.0$.

After identifying the boosted top quark decay products, this object is removed from the event and a modified BDRS Higgs tagger [58] is applied to fat C/A jet(s) with $R = 1.2$, that are also required to have $p_T > 200$ GeV. Within the Higgs candidate jet two b-tags are required. Again, if a jet is tagged as a Higgs candidate, the associated objects are removed from the event. Moreover, the Higgs candidate is required to be within the pseudo-rapidity range $\eta < 2.5$. The remaining objects are then clustered in C/A jet(s) with $R = 0.6$, and with $p_T > 30$ GeV and $\eta < 2.5$, and one of them is required to have a b-tag. Cuts are performed to further improve these results. Cuts on the N-subjettiness ratio $\frac{\tau_2}{\tau_1} < 0.4$ are also applied in this modified algorithm, as well as a reduction of the jet radius in steps of 0.1 as long as the jet mass does not drop below $m_j < 0.8\, m_{j,orig}$. The first cuts reduce the backgrounds and also better define the mass peaks for the Higgs and Z-decays. The modified jet radius, apart from reducing underlying events and pile-up, minimises the combinatorial errors in the $m_{b\bar{b}}$ reconstructions.

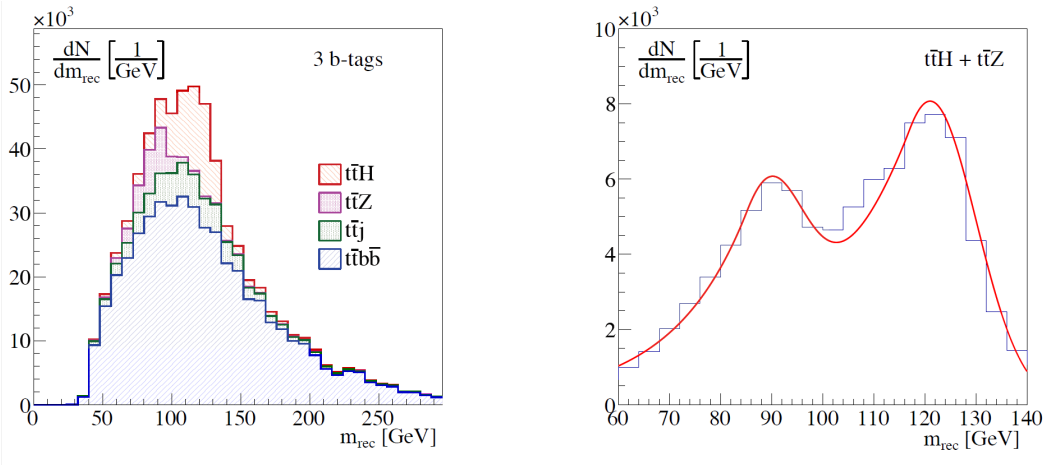

Figure 5.14: Left: Reconstructed $m_{b\bar{b}}$ for the leading jet substructures in the fat Higgs jet. Right: Double-peak fit assuming perfect continuum background subtraction. The event numbers are scaled to $\mathcal{L} = 20\,\text{ab}^{-1}$ [71].

The resulting distribution is presented in the left side of Figure 5.14. The background region between 160 and 300 GeV does not contain signal and is smooth, so it can be used to subtract the QCD background from the combined $t\bar{t}H$ and $t\bar{t}Z$ signal. On the other hand, the region between 0 and 60 GeV

needs to be checked by a full experimental analysis in order to see if it can also be used as a sideband. On the right hand side of Figure 5.14 is shown the combined fit to the Z and Higgs peaks assuming a perfect background subtraction. It is then found that using the combined fit allows to probe the top Yukawa coupling with a statistical precision of around 1%.

Chapter 6

Analysis Strategy

The analysis strategies investigated in this thesis concern the search for $t\bar{t}H$ production, in the semi-leptonic final state, and with the Higgs decaying into two bottom quarks. The goal of this work is to arrive to an efficient strategy, as simply as possible, to gain sensitivity to this channel. While the current analyses are deeply associated to machine learning methods, to identify and reconstruct the particles, this work instead uses jets with large radius to reconstruct boosted particles, such as the Higgs boson, and dedicated algorithms to identify the objects of interest.

This chapter describes the implementation of a strategy based on Reference [71] for the Future Circular Collider (FCC) at $\sqrt{s} = 100$ TeV, described in the first section. This is followed in the next section by the work done with the goal of optimising the analysis strategy for the HL-LHC resulting in a more efficient and simpler strategy.

6.1 Original Strategy

The first implementation of the analysis strategy, that will be referred to as 'original strategy' in this work, is strongly based on the approach proposed for the FCC in Reference [71], and described in detail in Section 5.2. Its key points are the use of HEPTOPTAGGER2, with $R = 1.8$ C/A jets, and a BDRS Higgs tagger, with $R = 1.2$ C/A jets, to reconstruct the boosted top quark and Higgs boson, respectively. A scheme of this strategy is shown in Figure 6.1.

In detail, this strategy starts by requiring an isolated charged lepton in the event with $p_T > 15$ GeV and $|\eta| < 2.5$. The pseudo-rapidity selection takes into account the ID acceptance, while the cut on p_T aims to suppress backgrounds with low-energy leptons, even though in the LHC triggers can only accept events with larger cuts on this variable. To model the $b\bar{b}j$ and dijets backgrounds, since generally such selection is not satisfied, the isolated charged lepton is not required, and instead one in 5000 jets is reconstructed as one, which is roughly in accordance with the fake rate of hadronic jets mimicking a charged electron (muons actually have a lower fake rate associated).

Afterwards, the calorimeter towers in the event are collected, and the ones close to isolated electrons, within a $\Delta R < 0.1$, are removed, in order to avoid overlap between objects. Muons are not considered

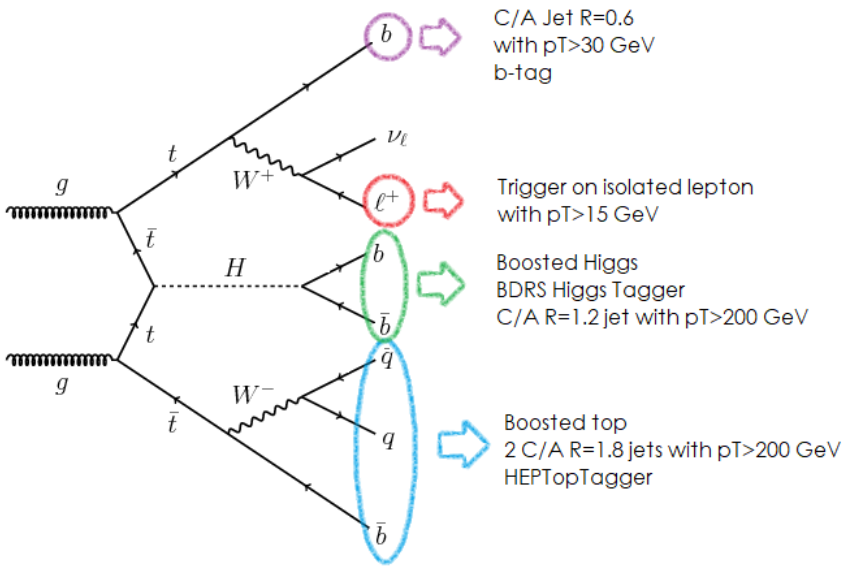

Figure 6.1: Original analysis scheme for single lepton $t\bar{t}H$.

in this overlap removal step, as their energy deposit in the calorimeters is minimal. The towers that pass this procedure form the 'tower collection', and are used as input to FASTJET, to be clustered in 'fat' Cambridge-Aachen (C/A) jets, with $R = 1.8$ and $p_T > 200$ GeV.

The event is then required to have at least two fat jets and these are passed to HEPTOPTAGGER2, to search for a boosted top. If a jet is tagged as originating from a top quark, the N-subjettiness ratio $\frac{\tau_3}{\tau_2}$ is computed and required to be $\frac{\tau_3}{\tau_2} < 0.8$. Moreover, the top candidate must be in the range $|\eta| < 4.0$. Having a top candidate satisfying these criteria, its constituents (associated towers) are removed from the tower collection.

The remaining towers are clustered with FASTJET using the C/A algorithm, with $R = 1.2$ and $p_T > 200$ GeV. The event is then required to have at least one of these jets. The chosen radius for the jets intended to contain top and Higgs candidates are backed up by generator level studies of the ΔR between the different respective decay particles, for the selected p_T.

The maximum ΔR between light or charm (c) and bottom (b) quarks, at generator level, for a $t\bar{t}H$ sample, that result of the decay of a top quark with $p_T > 200$ GeV, is shown in Figure 6.2. In it, it is possible to see that the peak is situated around a value of 1.8. On the other hand, the ΔR between generator level b quarks, coming from a Higgs boson with $p_T > 200$ GeV, is presented in Figure 6.3. In this distribution, the maximum is placed around $\Delta R \sim 1.2$, hence the use of this value for the jet radius. It should be noted that, while a larger radius for the jets would include a wider number of candidates, it would also increase the QCD contamination. QCD jets would be composed of more soft jets that, highly concentrated, could mimic hard substructures.

The $R = 1.2$ jets are passed to a BDRS Higgs Tagger with a mass drop condition of 0.9 and $y_{cut} = 0.09$. The value for the mass drop threshold was varied between 0.9 and the default value (0.667), and the difference was found to be negligible. Moreover, for the y_{cut} parameter, larger values ($y_{cut} \in [0.15, 0.30]$) and smaller values ($y_{cut} \in [0.01, 0.05]$) for y_{cut} were tested, and found to worsen the results,

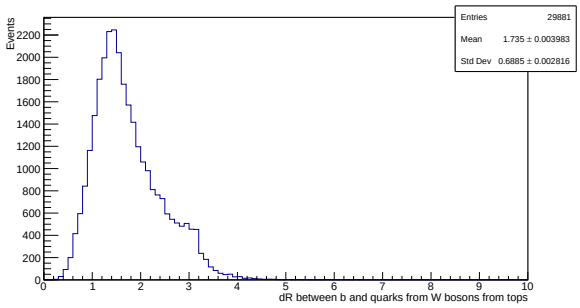

Figure 6.2: Maximum ΔR between generator level bottom and light or c quarks coming from a top quark decay with to quark momentum above 200 GeV.

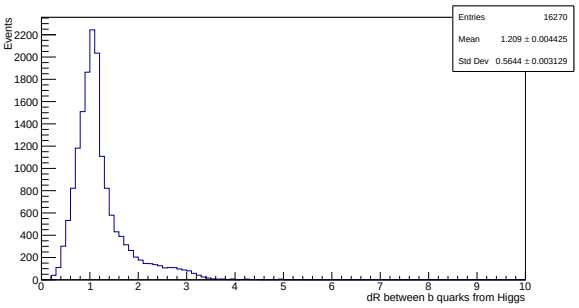

Figure 6.3: ΔR between generator level bottom quarks coming from a Higgs boson with $p_T > 200$ GeV.

so the default value was kept. In the case where BDRS tags a jet as being a Higgs candidate, its two subjets are picked and required to have each a $p_T > 30$ GeV and $|\eta| < 2.5$. While the p_T cut intends to suppress low energy quarks coming from the background, the η cut concerns the ID acceptance and so the region where b-tagging is possible. Moreover, each subjet must have a b-tag. If these requirements are satisfied, the Higgs candidate is filtered to remove pile-up and underlying event contamination, and up to three hard thinner subjets are kept in this process, as to account for a possible third subjet resulting from gluon radiation from one of the bottom quarks. Finally, the filtered Higgs-tagged jet must have $p_T > 200$ GeV and an invariant mass above 50 GeV, in order to be considered as a Higgs candidate. The motivation behind the mass cut and further optimization are described in the next section. In the case of there being more than one Higgs candidate per event, the one with the highest p_T is considered. This situation happens less than 1% of the time in $t\bar{t}H$ events. The event is required to have one Higgs candidate after this procedure and, as for the top quark case, its associated towers are removed from the tower collection.

The remaining towers in the tower collection are then sent to FASTJET, to be clustered in C/A jets, with $R = 0.6$ and $p_T > 30$ GeV. The goal of this step is to find the bottom quark coming from the leptonic top quark, and therefore the event is required to have at least one of these jets. Finally, one of these jets must have a b-tag, and this jet needs to be separated in $\Delta R > 0.4$ from other possible jets, in order for the event to be accepted as a $t\bar{t}H$ candidate.

Applying this strategy to the generated samples, for the HL-LHC scenario, mentioned in Section 4.2, we obtain a mass distribution of the Higgs candidate jets, shown in Figure 6.4. In this plot events are divided in bins of 20 GeV, and normalized to an integrated luminosity of 3000 fb^{-1}. In the figure, the colour filled distributions are stacked. In addition, the normalized mass distributions for $t\bar{t}H$ and $t\bar{t}Z$ production are presented, in order to better clarify the shape of both peaks.

The presence of signal is clear in the distribution of Figure 6.4, and the dominant backgrounds are the $t\bar{t}b\bar{b}$ and $t\bar{t}j$ processes. Moreover, the $t\bar{t}j$ distribution has significant statistical uncertainties, despite the wide bins and the large sample produced for this background. The dijets and $t\bar{t}b\bar{b}$ distributions suffers from the same problem. In fact, the peak around 130 GeV in the distribution results of the statistical fluctuations of these backgrounds. Nevertheless, the distribution for the $t\bar{t}H$ process has a peak around

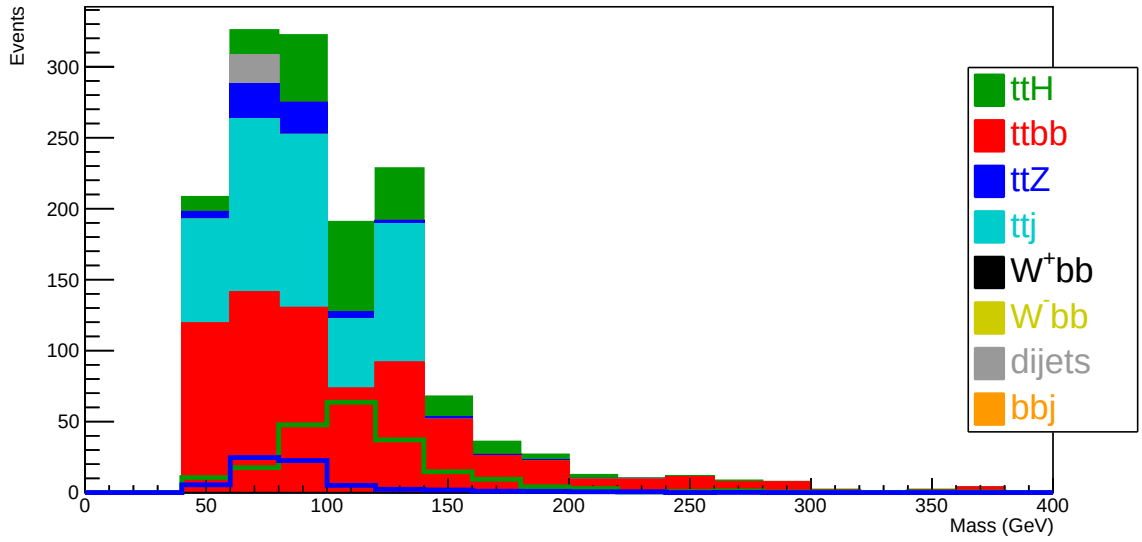

Figure 6.4: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for original analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

the Higgs boson mass, as desired, and the distribution for $t\bar{t}Z$ peaks around the Z boson mass.

A table with the number of events remaining after each selection cut (so-called cut-flow table) is shown in Table A.1. In this table, events are normalized, as before, to an integrated luminosity of 3000 fb^{-1}.

It is then possible to determine the expected significance of the signal, denoted by Z_A, using the formula in Equation 6.1. We use the significance as a figure of merit to evaluate different possible analysis strategies. The significance can be defined as the number of standard deviations necessary for a Gaussian variable to fluctuate in one direction to give a certain p-value [72]. This p-value is the probability to observe data compatible with a background-only hypothesis. Therefore, for a discovery, a minimum $Z = 5$ (sigma) is required, corresponding to a very small p-value (10^{-7}), and thus rejecting the hypothesis of not having signal (null hypothesis).

$$Z_A = \sqrt{2\left((S+B)\,ln\left(1+\frac{S}{B}\right)-S\right)} \tag{6.1}$$

However, for cases where the number of background events is much larger than the number of signal events ($S << B$), it is possible to reduce Z_A to

$$Z_A = \frac{S}{\sqrt{B}} \tag{6.2}$$

This approximation was found to have an associated error of up to 3%, for significances computed for the HL-LHC scenario, and with the strategies implemented in this thesis.

For this mass distribution of the Higgs candidates, the significance is computed in the mass region between 60 and 160 GeV, due to the wide signal spectrum, and its values for different integrated luminosities are presented in Table 6.1. The significance was computed for three different values of

integrated luminosities, $\mathcal{L} = 36, 300, 3000\,\text{fb}^{-1}$, that correspond to the collection of data used in the last $t\bar{t}H(b\bar{b})$ ATLAS analysis [17], the expected integrated luminosity by the end of Run 3, and that expected at the end of the HL-LHC, after ten years of operation, respectively. The amount of signal over the background S/B, computed in the same mass window, is also presented in this Table.

Table 6.1: Significance and S/B for different integrated luminosities. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$. Using the **original** strategy.

$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
36	0.66 ± 0.04	
300	1.92 ± 0.12	19.4 ± 1.7
3000	6.07 ± 0.38	

Analysing these numbers, it becomes clear that, even for higher luminosities, this strategy is quite inefficient, as will be seen when optimizing the analysis strategy. The main cause of this inefficiency was found to come from an inefficient top tagging by HEPTOPTAGGER2, as it was designed to search for top quarks with higher transverse momentum, of around 400 GeV or higher. Therefore, it is expected not to be as useful when requiring a top quark with a p_T of only 200 GeV. On the other hand, it is not particularly helpful in suppressing the main irreducible background $t\bar{t}b\bar{b}$, again as expected.

Moreover, removing the objects associated to the top quark candidate, frequently also results in the removal of calorimeter towers containing energy from Higgs decay products. As a consequence, the Higgs candidate is reconstructed with missing objects, ending up with a poorly reconstructed invariant mass or, in cases where the substructure of the Higgs jet was completely destroyed, the BDRS tagger fails to find a Higgs candidate, and the event is lost.

6.2 Optimization

In the optimized strategy, therefore, the use of HEPTOPTAGGER2 was dropped, as well as the clustering step in jets of $R = 1.8$. A fourth b-tag is requested, in order to compensate this fact, along with some other changes in terms of clustering technique and jet radius. Moreover, the cut on the invariant mass of the Higgs candidates is replaced by a cut on the ΔR between its two b-subjets, so as to provide a better sideband for lower Higgs candidates mass values.

6.2.1 Low mass candidates

The low-energy mass spectrum of the Higgs candidates is dominated by large contributions from different backgrounds, in a steeply falling distribution up to around 50 GeV, with around ten to twenty times more events than in the signal mass window ([60,160] GeV). For that reason, a study was conducted, firstly in order to understand the cause for this, and to find a proper way to treat these events.

The ΔR between the two b-tagged subjets of the BDRS Higgs candidates (therefore before filtering) was computed for a sample of $t\bar{t}H$ production, and is shown in Figure 6.5. Two structures can be seen

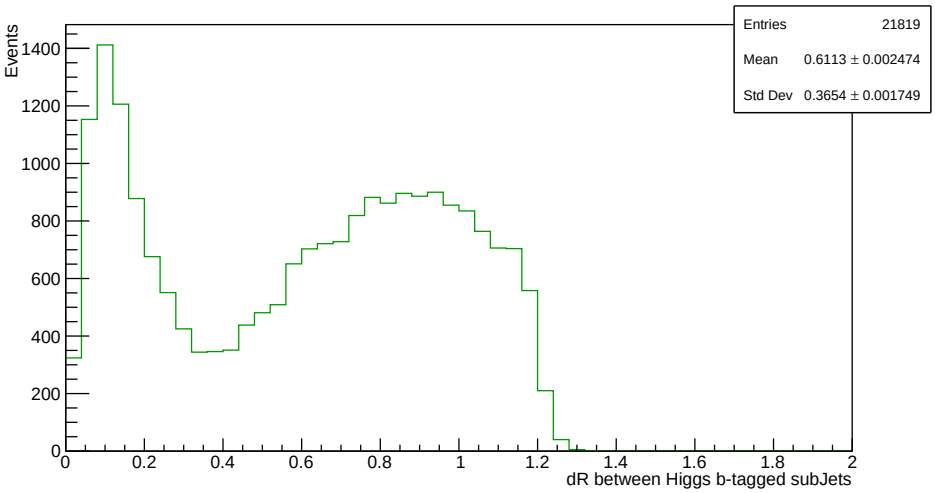

Figure 6.5: ΔR between b-tagged subjets of BDRS Higgs candidates.

in this distribution, one with small ΔR values, $\Delta R \in [0.05, 0.3]$, and another with a reasonable separation between the b-jets, $\Delta R \in [0.5, 1.2]$. It was found that Higgs candidates with values of the ΔR below 0.3 between the two b-tagged jets were also associated to lower invariant masses, ranging from 0 to around 50 GeV. Moreover, looking at the generator level distributions of the ΔR between the bottom quarks coming from a Higgs boson, it is noted that there are no events with values below 0.3.

To better understand the situation, the BDRS Higgs candidates with a $\Delta R < 0.3$ were considered 'low-mass candidates', and the ΔR between them and the generator ('gen') level Higgs boson was determined, and is shown in Figure 6.6. In this distribution it is possible to identify two clear structures, one with small ΔR values, corresponding to BDRS Higgs candidates matched to the generator level Higgs, and another with larger values, where there is no match between the two objects in the event.

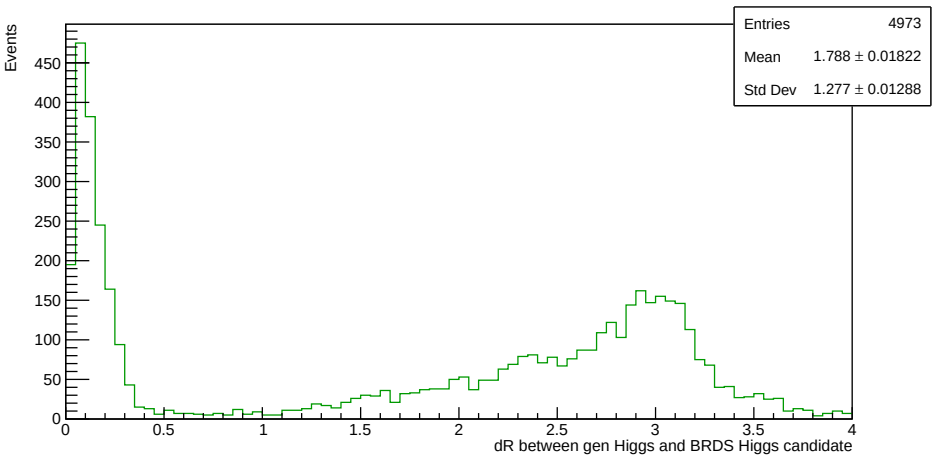

Figure 6.6: ΔR between generator level Higgs boson and BDRS Higgs candidates.

Having this in mind, the p_T of the generator level Higgs bosons was computed for both cases, considering only events with low-mass candidates, as before. The distributions for the candidates that are matched and not matched to the generator level Higgs are presented in Figures 6.7 and 6.8, respectively.

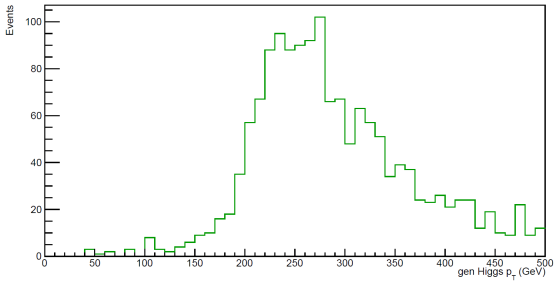

Figure 6.7: Generator-level Higgs p_T for ΔR(gen Higgs, BDRS candidate)< 0.5

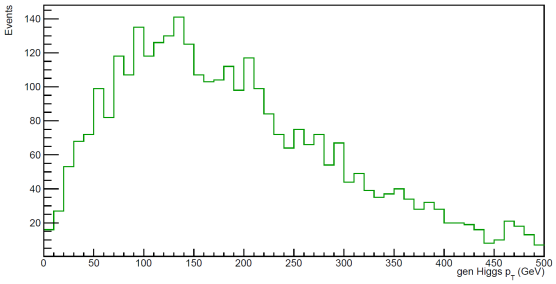

Figure 6.8: Generator-level Higgs p_T for ΔR(gen Higgs, BDRS candidate)> 0.5

It can be seen that the Higgs transverse momentum is not the main cause for the BDRS algorithm to fail reconstructing a Higgs candidate, since low-mass candidates are correctly matched to the generator level Higgs boson. In fact, it was found that these situations derive from the fact that the two candidate subjets are being b-tagged to the same bottom quark, instead of two different ones. Therefore, these subjects do not correspond to the two actual bottom quarks coming from the Higgs boson, and consequently the invariant mass of the Higgs candidate differs from values around 125 GeV. On the other hand, the low-mass candidates that are not matched in ΔR to the generator level Higgs appear to be associated to cases where the Higgs boson is not boosted enough, that is, has a p_T generally below 200 GeV. This causes the decay objects not to be collimated enough, to fit in a $R = 1.2$ jet, and naturally the BDRS Higgs tagger is then unable to properly reconstruct these candidates.

Different reconstruction procedures were conducted, in order to see if it was possible to recover these low-mass candidates. One attempt was to collect three additional b-tagged C/A $R = 0.6$ jets, with $p_T > 30$ GeV, after the removal of the top quark objects, if the ΔR between the two BDRS Higgs candidate b-tagged subjets was bellow 0.3. The invariant mass of each pairwise combination of these three b-jets would be computed, and the smallest invariant mass was taken, as it was found that this mass was generally closer to the Higgs mass. While successfully recovering poorly reconstructed Higgs candidates for the signal mass region, this procedure was also found to increase the contamination in this region with several different background sources, decreasing the significance by a factor of almost 30%, when compared to the case where the invariant mass of the Higgs candidates was required to be above 50 GeV. For this reason, this procedure was abandoned. It was also investigated if using $R = 1.4$ instead of $R = 1.2$ C/A jets would contribute to a decrease of the number of low-mass candidates, by eventually covering also candidates where the Higgs was less boosted than required, but this also proved to be ineffective.

Facing the impossibility of properly treating these cases, the Higgs candidates were requested to have an invariant mass above 50 GeV, therefore suppressing these situations. Nevertheless, it was found that requesting a $\Delta R > 0.3$ between the two b-subjets of the BDRS Higgs candidates was more useful, as it could provide a better sideband at low masses, without affecting the significance.

6.2.2 Optimized Strategy

As already mentioned, the final optimized strategy makes no use of $R = 1.8$ C/A jets and HEPTOPTAGGER2, relying instead on the identification of four jets coming from bottom quarks in the events, using also the BDRS Higgs tagger to efficiently select $t\bar{t}H$ events suppress backgrounds, and reconstruct Higgs boson jet candidates. Behind these options is the fact that removing the HEPTOPTAGGER2 algorithm from the strategy increases the analysis significance by 112%, and excluding the clustering step in $R = 1.8$ C/A jets further increases the significance by 20%. It should be noted that an alternative way to tag the top quarks, using the $R = 1.8$ jets, was investigated, through the use of the τ_{31} ratio, as it proved to be the most discriminant variable to distinguish between jets with a mass around the top quark mass and the remaining jets. However, this implementation would reduce the significance by around 30%, and was abandoned. Finally, a scheme of the optimized strategy is shown in Figure 6.9.

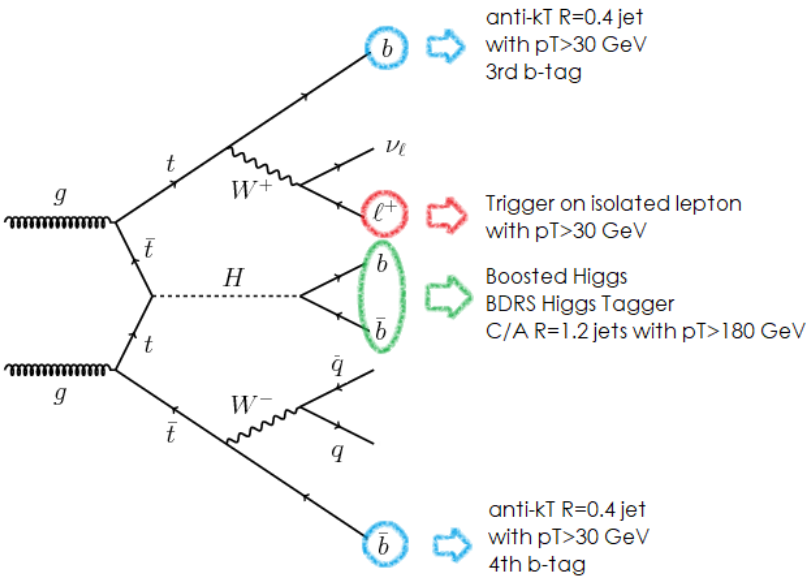

Figure 6.9: Optimized analysis scheme for single lepton $t\bar{t}H$.

In the optimized analysis strategy, the event is firstly required to have an isolated charged lepton, with $p_T > 30$ GeV and $|\eta| < 2.5$. This p_T cut reflects a more realistic minimum transverse momentum expected of lepton triggers in the High Luminosity scenario, and results in a decrease in terms of significance of around 10% with respect to a cut on 15 GeV. Again, for the $b\bar{b}j$ and dijets backgrounds, this selection is not applied, and one in 5000 jets is reconstructed as an isolated charged lepton.

The strategy then proceeds to collect the calorimeter towers in the event, and remove towers within $\Delta R < 0.1$ from isolated electrons. The resulting tower collection is used as input to FASTJET, where towers are clustered using the Cambridge-Aachen (C/A) jet algorithm with a radius $R = 1.2$, and a minimum jet transverse momentum cut of 180 GeV is applied. The event is then required to have at least one of these jets.

The BDRS Higgs tagger receives as input the $R = 1.2$ jets, and proceeds to test if there is a Higgs candidate jet, using a mass drop condition of 0.9 and $y_{cut} = 0.09$. The two subjets of the Higgs candidate

jet are then required to have each $p_T > 30$ GeV, $|\eta| < 2.5$, and to be b-tagged. The cut on the subjets' transverse momentum was tested for higher values, namely 50, 60 and 80 GeV. However, increasing the p_T value did not result in higher efficiency for signal events and further suppression of the backgrounds, and therefore the initial cut of 30 GeV was kept. The candidates that pass these selection criteria are then filtered, keeping up to three hard thinner subjets, and after this procedure are required to have $p_T > 180$ GeV. In order to remove low-mass candidates, but still keep a sideband at low values of the Higgs candidate jets mass distribution, it is demanded that the two b-tagged subjets of the Higgs candidate jet have a $\Delta R_{bb} \geq 0.3$ between them. For events with more than one Higgs candidate per event, the one with the highest p_T is considered, even though the amount of times this happens is negligible. The event is then required to have one Higgs candidate, and its associated towers are removed from the tower collection.

The minimum transverse momentum required for the $R = 1.2$ C/A and Higgs candidate jets was varied for higher and larger values, in order to optimize the significance. A value of 200 GeV was taken as reference, as it was implemented in the original strategy. The values for the difference in the significance for variable cuts with respect to the reference value, 200 GeV, are presented in Table 6.2. As can be seen, requiring a transverse momentum larger than 200 GeV results in decreasing significances, as backgrounds are not suppressed in larger proportion than the signal. On the other hand, placing the p_T threshold at values below 200 GeV was found to improve the analysis sensitivity. Nevertheless, lower p_T cuts also modify the shapes of the Higgs mass peak, enhancing values below 125 GeV, and different background could have to be taken into account. This is due to the Higgs objects being more spatially dispersed, as they have a lower boost, and therefore not fitting in a $R = 1.2$ jet. Values below 150 GeV were not explored as would be physically meaningless.

In this context, a minimum p_T of 180 GeV was found to achieve a good balance between the peak position and the significance improvement, and was therefore the implemented value. Note however that a jet radius optimization was not performed for each p_T cut, and that the significance does not vary linearly with the integrated luminosity. Therefore, different results could be obtained in different contexts.

Table 6.2: Relative significance variation for different p_T cuts on the R=1.2 C/A and Higgs candidate jets, considering the same integrated luminosity of 36 fb^{-1}. Computed from masses in range [60,160] GeV.

p_T cut (GeV)	$(Z_A - Z_A^{reference})/Z_A^{reference}$ (%)
150	+26
180	+11
200 (**reference**)	1
250	-27
300	-41
400	-66
500	-84

After removing the Higgs candidate jet associated towers, the remaining towers in the tower collection are clustered into anti$-k_t$ jets, with $R = 0.4$ and $p_T > 30$ GeV. This option was based on the fact that the change in clustering algorithm contributes to a 3% improvement in the significance, and that these are the usual type of jets used for b-tagging in a real analysis. The enhancement when using anti$-k_t$ jets with

respect to C/A jets can be seen as a consequence of the different type of clustering procedure. While the C/A algorithm clusters together towers having into consideration their distance, the anti$-k_t$ algorithm clusters sequentially towers with higher transverse momentum. This way, the anti$-k_t$ algorithm retrieves jets with hard substructure closer to the original bottom quark, contrary to the C/A jets that can have more soft jets faking jets coming from bottom quarks.

The event is required to have at least two of these jets, and two must have a b-tag. Furthermore, the two b-tagged jets must have a ΔR separation between them above 0.4, to avoid overlap of the two jets. The ΔR between the leading and sub-leading b-tagged jets, and the Higgs candidate jet is computed, and is referred to as $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$, respectively. These two variables, along with the distance between the two b-tagged subjets of the Higgs candidate jet, ΔR_{bb}, are used later in the analysis to further suppress backgrounds.

The distributions of the ΔR_{bb} values for the Higgs candidates versus its mass, before the cuts on this variable, are presented in Figures 6.10, 6.11, 6.12 and 6.13, for the signal $t\bar{t}H$, and for the main backgrounds $t\bar{t}Z$, $t\bar{t}b\bar{b}$ and $t\bar{t}j$ processes, respectively. As can be seen, the distributions for this variable differ for each process, and therefore a scan over the variable range was performed to find the optimal cut values that would result in the highest significance. For this case, a 3% improvement was found by rejecting Higgs candidate jets with $\Delta R_{bb} < 0.36$, and the same increase in significance was obtained by removing candidates with $\Delta R_{bb} > 1.28$. The same procedure was performed for $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$, and the best results in terms of significance were obtained by excluding Higgs candidate jets with $\Delta R_{b_3,H} < 0.87$ and $\Delta R_{b_4,H} < 0.88$, with each of these cuts contributing to a 2% improvement in the significance. The distributions for these variables are presented in Figures B.1 to B.8, for the signal and the main backgrounds. Therefore, these four cuts were implemented and are the last step of the analysis strategy.

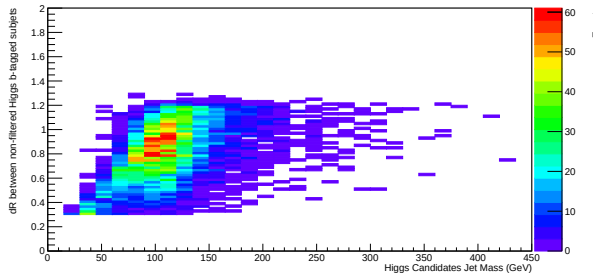

Figure 6.10: ΔR_{bb} for Higgs candidates for $t\bar{t}H$

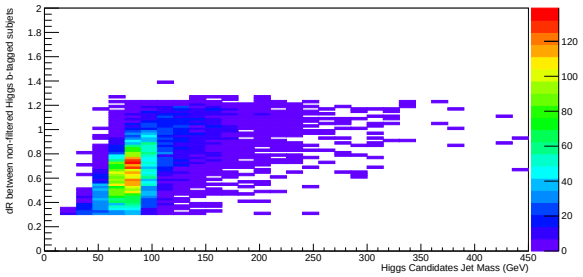

Figure 6.11: ΔR_{bb} for Higgs candidates for $t\bar{t}Z$

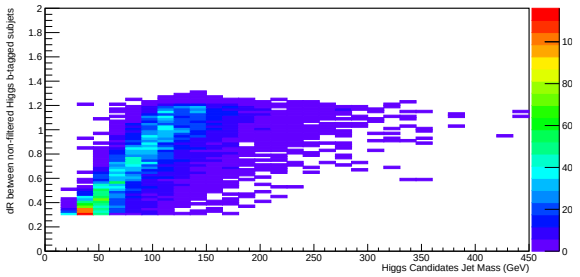

Figure 6.12: ΔR_{bb} for Higgs candidates for $t\bar{t}b\bar{b}$

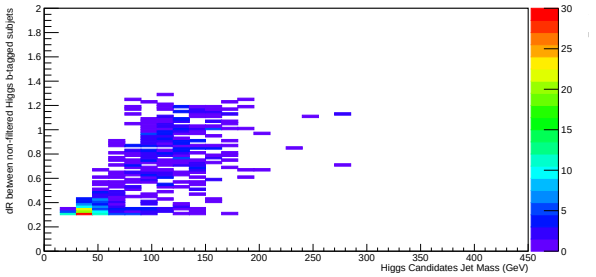

Figure 6.13: ΔR_{bb} for Higgs candidates for $t\bar{t}j$

The mass distribution of the Higgs candidate jets, resulting from the implementation of the optimized strategy, using the generated samples described in Section 4.2, for the HL-LHC scenario, is presented in Figure 6.15. In Figure 6.14, the Higgs candidate jets mass distribution for the original strategy is reproduced from Figure 6.4, as a term of comparison. Events are divided in bins of 20 GeV, and normalized to an integrated luminosity of 3000 fb^{-1}. Furthermore, the colour filled distributions are stacked. In addition, the normalized mass distributions for the $t\bar{t}H$ and $t\bar{t}Z$ are presented, in order to better clarify the shape of both peaks. The cut-flow table for the optimized strategy is presented in Table A.2.

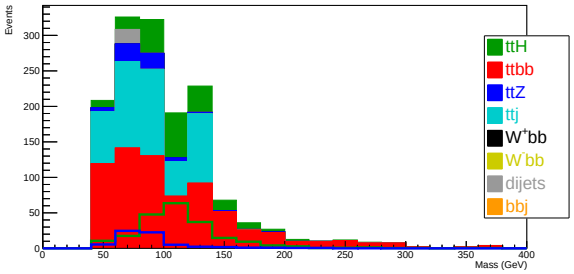

Figure 6.14: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for original analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

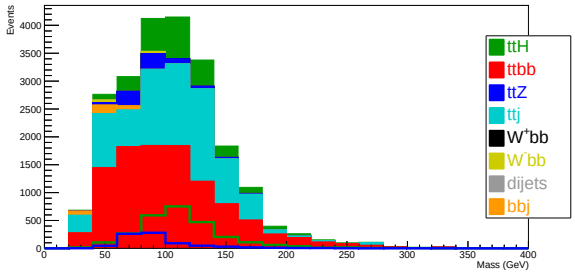

Figure 6.15: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

The peak for the $t\bar{t}H$ is well defined on top of the backgrounds, in the distribution of Figure 6.15, and the statistical fluctuations for the $t\bar{t}j$ process are much reduced. Again the dominant backgrounds are $t\bar{t}b\bar{b}$ and $t\bar{t}j$. Moreover, the contamination from $W^{\pm}b\bar{b}$ and $b\bar{b}j$ backgrounds is negligible, although these suffer from statistical fluctuations, as the strategy highly suppresses these processes.

The significance and S/B for this strategy, for different integrated luminosities, are presented in Table 6.3, along with the values for the original strategy. The significance and S/B are computed, as before, in the mass window between 60 and 160 GeV.

Table 6.3: Significance and S/B for different integrated luminosities and strategies. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Strategy	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
Original	36	0.66 ± 0.04	19.4 ± 1.7
Optimized	36	2.12 ± 0.04	15.7 ± 0.4
Original	300	1.92 ± 0.12	19.4 ± 1.7
Optimized	300	6.13 ± 0.11	15.7 ± 0.4
Original	3000	6.07 ± 0.38	19.4 ± 1.7
Optimized	3000	19.39 ± 0.33	15.7 ± 0.4

Analysing this table, it can be seen that the results improved, in terms of significance, by a factor of around 3. On the other hand, the S/B seems to slightly decrease when using the optimized strategy, but this comparison should be treated carefully, as the results for the original strategy are influenced by statistical fluctuations.

6.2.3 Discriminating Variables

While only a few variables were implemented in the last step of the optimized analysis, there were others that, despite leading to a lower improvement in significance (below 1%), could eventually result in further discrimination between the signal and the backgrounds, when used together in a multivariate method (MVA) like a boosted decision tree or a neural network. The motivation for this derives from the fact that, for these variables, linear cuts do not fully exploit their discriminating power, but non-linear selections could improve the results.

Therefore, the proposed variables to be used as input are the τ_{21} and τ_{31} ratios, and the $C2$ and $D2$ energy correlation functions, using $\beta = 2$, for the Higgs candidate jets. A β value of 2.0 was preferred to $\beta = 0.5$, as the former was associated to slightly larger significance improvements. The distributions for τ_{21}, τ_{31}, $C2$ and $D2$, for the Higgs candidates versus its masses, are presented in Figures B.9 to B.24, for the signal and the main backgrounds.

In order to obtain the best results, the cuts on ΔR_{bb}, $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$ should be removed, and these variables should also be used as input to the MVA, as the algorithm can then better exploit the different shapes of the distributions between the signal and the backgrounds.

6.3 Re-clustering Scheme

Having an optimized strategy, it was then investigated what effect of using re-clustered jets instead of large jets at the start of the analysis was. This is a realistic approach taken in experimental analyses, as every jet configuration needs a specific calibration, from the jet algorithm and radius, to its energy and mass scale. These have to be corrected in order to account for the detector response, as well as other experimental effects [73]. Therefore, large jets would require a specific jet calibration. On the other hand, using the usual small calibrated jets as input for the reconstruction of large radius jets avoids additional calibrations, and is expected to deliver the same jet performance.

The implemented strategy is the same as the optimized strategy, with the exception of the initial jets. Therefore, after requiring the isolated charged lepton, the calorimeter towers are clustered in $R = 0.4$ anti$-k_t$ jets, with $p_T > 25$ GeV and $|\eta| < 2.5$, using the FASTJET algorithm. Moreover, the object overlap removal is performed by DELPHES3 in this case. These jets are then sent as input again to FASTJET, and are re-clustered in C/A jets, with $R = 1.2$ and $p_T > 180$ GeV. The BDRS Higgs tagger then receives these large jets as input, and the strategy continues as before.

The mass distribution obtained for the Higgs candidate jets, versus the distribution using the optimized strategy are presented in the Figures 6.17 and 6.16, respectively. As before, events are divided in 20 GeV bins, and normalized to an integrated luminosity of 3000 fb^{-1}. Moreover, the colour filled distributions are stacked. Additionally, the normalized mass distributions for the $t\bar{t}H$ and $t\bar{t}Z$ are presented. The cut-flow table for this implementation is presented in Table A.3, where the events are normalized as before.

The mass distribution in Figure 6.17 is quite similar to the one obtained previously, with the clear

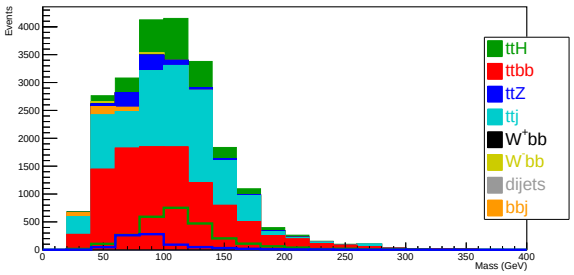

Figure 6.16: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

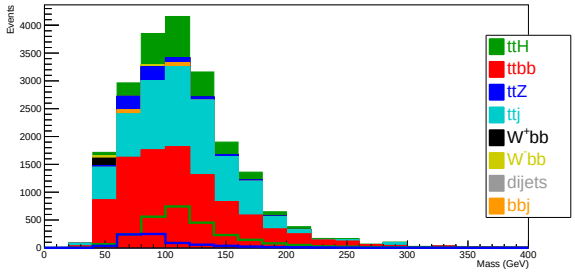

Figure 6.17: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy with re-clustering. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

presence of signal on top of the different backgrounds. Nevertheless, the $t\bar{t}j$ process still shows some statistical fluctuations, and the same happens for the other backgrounds with the exception of $t\bar{t}b\bar{b}$ and $t\bar{t}Z$, again because of the background suppressing power of the analysis.

The significance and S/B for the optimized strategy with re-clustering, for different integrated luminosities, are presented in Table 6.4, along with the values for the original and optimized strategies. Furthermore, the significance and S/B are computed in the mass window between 60 and 160 GeV.

Table 6.4: Significance and S/B for different integrated luminosities and strategies. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Strategy	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
Original	36	0.66 ± 0.04	19.4 ± 1.7
Optimized	36	2.12 ± 0.04	15.7 ± 0.4
Optimized w/ re-clustering	36	2.12 ± 0.04	15.8 ± 0.4
Original	300	1.92 ± 0.12	19.4 ± 1.7
Optimized	300	6.13 ± 0.11	15.7 ± 0.4
Optimized w/ re-clustering	300	6.12 ± 0.11	15.8 ± 0.4
Original	3000	6.07 ± 0.38	19.4 ± 1.7
Optimized	3000	19.39 ± 0.33	15.7 ± 0.4
Optimized w/ re-clustering	3000	19.34 ± 0.35	15.8 ± 0.4

When comparing the values for the significances and S/B, it is clear once again that the effect of introducing the re-clustering technique is negligible. This fact therefore strengthens the proposed optimized strategy, as it maintains the performance even when becoming closer to a real analysis implementation.

6.4 Control Region

A control region was defined in order to constrain the backgrounds, namely the $t\bar{t}j$ production, as the cross section associated to this process is around 1000 times larger than the one for signal, and jets coming from light and charm quarks can fake jets resulting of bottom quark hadronization. In this control

region, the strategy targets events with Higgs candidate jets without two b-tags.

The strategy is the same as the optimized one, but the Higgs boson reconstruction has some changes. In this case the two jets are anti-b-tagged. I.e, the probabilities associated to requiring two b-tags on the two subjets of the Higgs candidate jet, retrieved by the BDRS Higgs tagger, are complementary to the working point that is used. For instance, for the 65% b-tagging working point on the optimized strategy, the control region has, for the Higgs b-tags, a working point of 35%. The working point for the remaining b-tags is not changed.

The mass distribution of the Higgs candidate jets in the control region is shown in Figure 6.18. Again, events are divided in bins of 20 GeV, and normalized to an integrated luminosity of 3000 fb^{-1}. Furthermore, the colour filled distributions are stacked. In addition, the normalized mass distributions for the $t\bar{t}H$ and $t\bar{t}Z$ are presented. The cut-flow table for the control region is presented in Table A.4, with the events normalized to 3000 fb^{-1}.

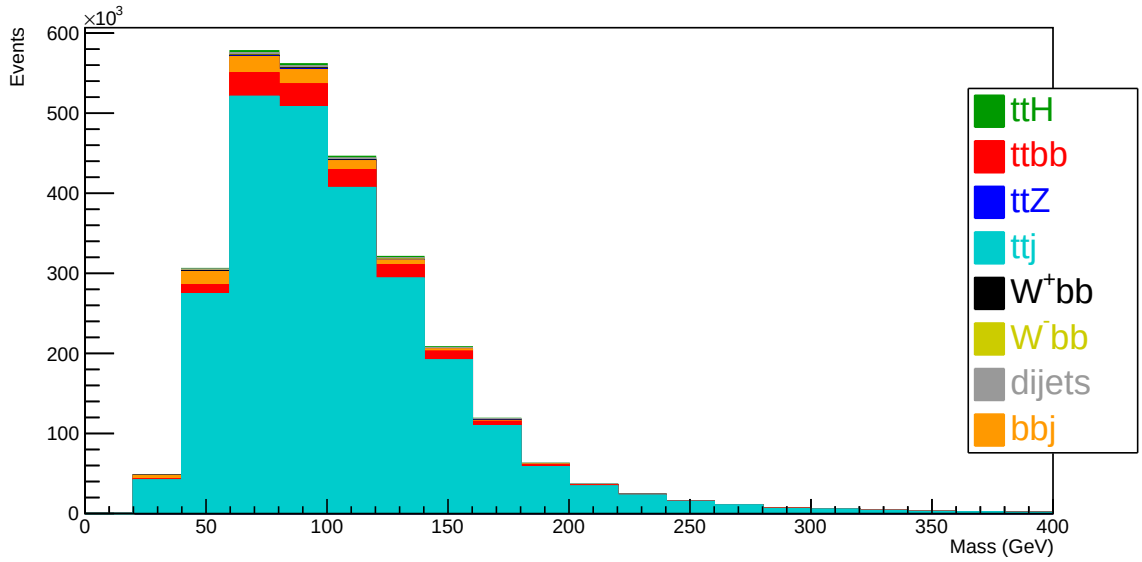

Figure 6.18: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for control region. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

As desired, this region is completely background dominated, with $t\bar{t}j$ being the predominant process. After $t\bar{t}j$, the most relevant backgrounds in this region are $t\bar{t}b\bar{b}$, $b\bar{b}j$ and dijets, in order of decreasing importance. In fact, the amount of signal in this distribution is very small in relation to the background quantity, with a S/B of only 0.5%. This region can therefore be used to control the backgrounds, in a simultaneous fit with the region covered by the optimized strategy.

6.5 Comparison with the LHC

The optimized strategy was then implemented for the LHC scenario, in order to evaluate to what extent changing from the HL-LHC card to ATLAS one, in the detector simulation, would affect the results. The ATLAS card takes into account the current ATLAS detector and, as a reminder, the samples generated

for this LHC scenario differ in size from the ones used for the HL-LHC. Moreover, the b-tagging is slightly more inefficient.

In addition, a jet energy scale formula, for the calibration of the jets momenta, is applied for the LHC scenario, according to Equation 6.3. This formula intends to rescale the jet's momenta to the value at particle level, but is nevertheless a simplification of the actual calibration performed in a real experiment, as in this case, the obtained correction considers real and simulated events, and the detector response for different objects. Moreover, this calibration is taken to be 1 for the HL-LHC, as defined in the official parameter card.

$$\text{jet energy scale formula} = \sqrt{\frac{(3 - 0.2|\eta|)^2}{p_T} + 1} \tag{6.3}$$

The mass distribution obtained for the LHC scenario is presented on Figure 6.20, while the distribution for the HL-LHC is shown in Figure 6.19. Events are divided in bins of 20 GeV, and normalized to an integrated luminosity of 300 fb^{-1}, that is the expected amount of statistics collected by the end of the LHC Run 3. Furthermore, the colour filled distributions are stacked. In addition, the normalized mass distributions for the $t\bar{t}H$ and $t\bar{t}Z$ are presented, in order to better clarify the shape of both peaks. The cut-flow tables for the LHC and HL-LHC scenarios, with the events are normalized to 300 fb^{-1}, are shown in Tables A.5 and A.6, respectively.

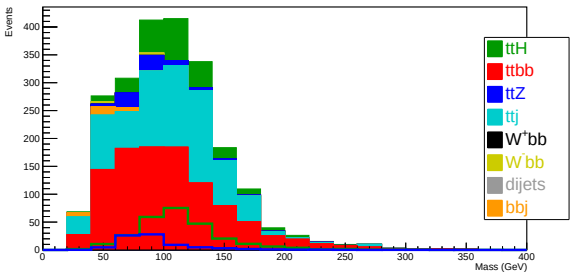

Figure 6.19: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy, and for the HL-LHC scenario. Events are normalized to $\mathcal{L} = 300\,\text{fb}^{-1}$.

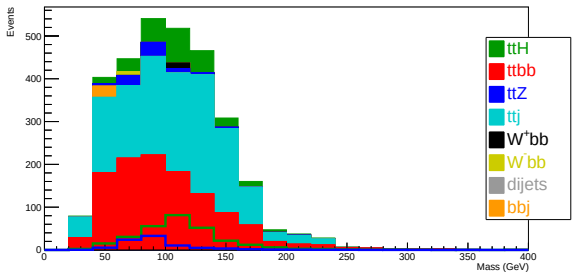

Figure 6.20: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy, and for the LHC scenario. Events are normalized to $\mathcal{L} = 300\,\text{fb}^{-1}$.

It is possible to see, comparing the two distributions, that the $t\bar{t}j$ contribution is slightly larger, deriving from the worse b-tagging. Moreover, this background and some others, as $b\bar{b}j$ and $W^{\pm}b\bar{b}$, have some statistical fluctuations due to lack of statistics. Nevertheless, the presence of signal is clear, on top of the backgrounds.

The significance and S/B of optimized strategy for the LHC scenario, for different integrated luminosities, are presented in Table 6.5, along with the values for the HL-LHC framework. The values for these variables are computed, as before, in the mass window between 60 and 160 GeV.

Analysing this table, it can be seen that the significance and S/B slightly decrease for the LHC scenario, with respect to the HL-LHC case, again because of the b-tagging being less efficient. Nonetheless, the results are nearly similar between the two scenarios, meaning that the implementation of the optimized strategy is feasible with the current ATLAS detector apparatus.

Table 6.5: Significance and S/B for different integrated luminosities and scenarios, using the optimized strategy. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Scenario	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
LHC	36	1.88 ± 0.04	11.6 ± 0.4
HL-LHC	36	2.12 ± 0.04	15.7 ± 0.4
LHC	300	5.41 ± 0.12	11.6 ± 0.4
HL-LHC	300	6.13 ± 0.11	15.7 ± 0.4
LHC	3000	17.12 ± 0.38	11.6 ± 0.4
HL-LHC	3000	19.39 ± 0.33	15.7 ± 0.4

6.6 Pure Pseudo-scalar Case

Could new physics exist, and different signals may be observed. In this work it was considered the possibility of having the production of a 125 GeV pseudo-scalar A in association with two top quarks, instead of the SM scalar Higgs boson, that also decays into two bottom quarks. As a reminder, this process is characterised by a pure CP-odd interaction with the top quarks, and has a cross section of about a half with respect to the $t\bar{t}H$ one.

The mass distribution of the Higgs candidate jets for this BSM signal sample, and the HL-LHC scenario, using the optimized strategy, is presented in Figure 6.22. On its left side, in Figure 6.21, it is shown the Higgs candidate jets mass distribution when using the SM signal sample, again with the optimized strategy. Events are divided in 20 GeV bins, and normalized to an integrated luminosity of 3000 fb^{-1}. Furthermore, the colour filled distributions are stacked. In addition, the normalized mass distributions for the $t\bar{t}A/t\bar{t}H$ and $t\bar{t}Z$ are presented, in order to better clarify the shape of both peaks. The cut-flow table for the implementation with the BSM sample is presented in Table A.7, with the events again normalized to an integrated luminosity of 3000 fb^{-1}.

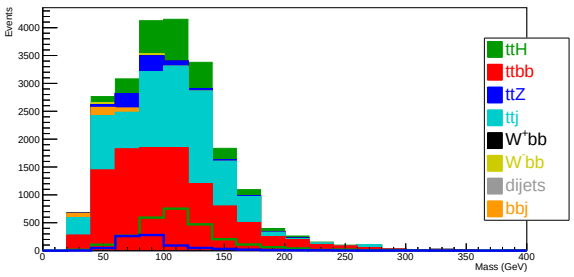

Figure 6.21: Higgs candidates mass for $t\bar{t}H$ and backgrounds, for optimized analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

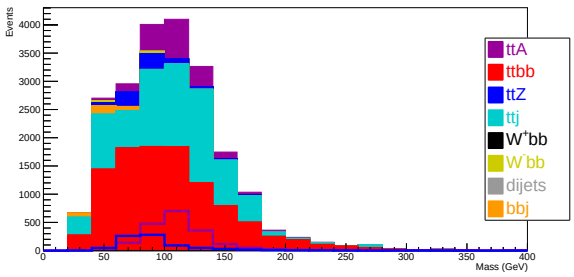

Figure 6.22: Higgs candidates mass for $t\bar{t}A$ and backgrounds, for optimized analysis strategy. Events are normalized to $\mathcal{L} = 3000\,\text{fb}^{-1}$.

As it is possible to see, the two distributions are fairly similar, and were further compared without the presence of backgrounds. In this way, the distributions shapes can be easily compared and differences induced by the different model used may be seen. The result of this comparison is shown in Figure 6.23, where events are divided in 20 GeV bins, and normalized to an integrated luminosity of 36 fb^{-1}. It can

be seen that the two distributions have similar shapes, differing on the number of events due to the lower cross section associated to the BSM production.

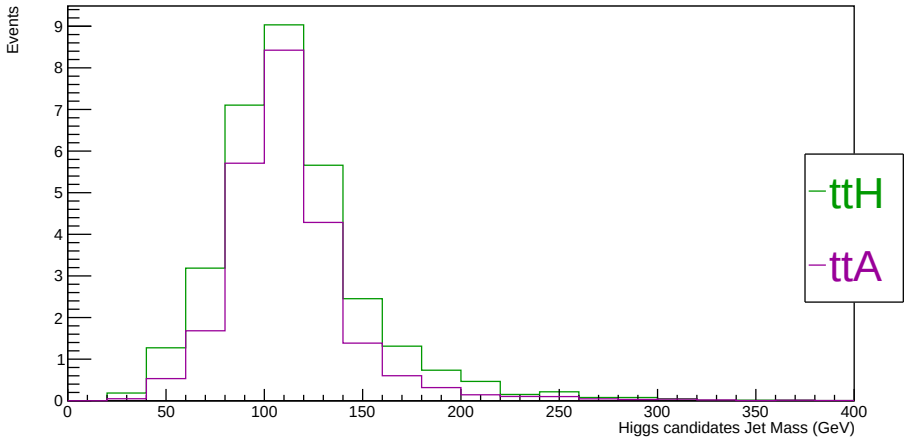

Figure 6.23: Higgs candidates mass for $t\bar{t}H$ and $t\bar{t}A$ samples, using the optimized strategy. Events are normalized to $\mathcal{L} = 36\,\text{fb}^{-1}$.

Moreover, the significance and S/B for the BSM production was computed, using the optimized strategy for the HL-LHC scenario. The results for different integrated luminosities are presented in Table 6.6, along with the values for the SM production. The values for these variables are computed, as before, in the mass window between 60 and 160 GeV.

Table 6.6: Significance and S/B for different integrated luminosities and processes, for the HL-LHC scenario. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Strategy	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
$t\bar{t}H$	36	2.12 ± 0.04	15.7 ± 0.4
$t\bar{t}A$	36	1.63 ± 0.03	12.1 ± 0.3
$t\bar{t}H$	300	6.13 ± 0.11	15.7 ± 0.4
$t\bar{t}A$	300	4.71 ± 0.10	12.1 ± 0.3
$t\bar{t}H$	3000	19.39 ± 0.33	15.7 ± 0.4
$t\bar{t}A$	3000	14.90 ± 0.32	12.1 ± 0.3

Analysing this table it is again clear the effect of the lower cross section on the $t\bar{t}A$ process, with diminished significance and S/B. In fact, the pseudo-scalar associated production would not be observed with 300 fb^{-1} in this channel, contrary to what would happen for the $t\bar{t}H$ case.

Chapter 7

Results

Having developed an optimized strategy, and tested it for different scenarios and signal processes, it is possible to arrive at the minimum integrated luminosity necessary to observe the desired processes. Moreover, a likelihood fit can be performed, in order to obtain the correspondent signal strengths.

This first section of this chapter presents the required integrated luminosity to observe the $t\bar{t}H$ process in the semileptonic channel, with the Higgs decaying into two bottom quarks. The minimum integrated luminosity necessary to observe $t\bar{t}A$ production is stated in the second section. Finally, in the last section, the uncertainties on the top Yukawa coupling for the LHC and HL-LHC scenarios are presented.

7.1 $t\bar{t}H$ Observation

The optimized strategy was implemented for the LHC and HL-LHC scenarios, and mass distributions for the Higgs candidates were obtained. The significance and S/B of optimized strategy were obtained for both scenarios and different integrated luminosities, and are presented in Table 7.1, reproduced from Table 6.5. The values for these variables were computed for Higgs candidate jets with masses ranging from 60 to 160 GeV.

Table 7.1: Significance and S/B for different integrated luminosities and scenarios, using the optimized strategy. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Scenario	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
LHC	36	1.88 ± 0.04	11.6 ± 0.4
HL-LHC	36	2.12 ± 0.04	15.7 ± 0.4
LHC	300	5.41 ± 0.12	11.6 ± 0.4
HL-LHC	300	6.13 ± 0.11	15.7 ± 0.4
LHC	3000	17.12 ± 0.38	11.6 ± 0.4
HL-LHC	3000	19.39 ± 0.33	15.7 ± 0.4

As can be seen, the $t\bar{t}H$ production could already be observed at the LHC, with an integrated luminosity of 300 fb^{-1}, using the optimized strategy. The expected significance would then be 5.41 ± 0.12.

The same integrated luminosity is associated to a significance of 6.13 ± 0.11 for the HL-LHC scenario. As mentioned before, using re-clustered jets in the analysis strategy would have a negligible effect in these results.

The signal strength is obtained by finding the minimum of $-2\,ln\,\lambda(\mu)$ [18], defined as

$$-2\,ln\,\lambda(\mu) = -2\,ln\frac{\mathcal{L}(\mu)}{\mathcal{L}(\hat{\mu})} = 2\sum_{i=1}^{N}\left[(\mu\,s_i + b_i) - n_i + n_i\,ln\left(\frac{n_i}{\mu\,s_i + b_i}\right)\right] \tag{7.1}$$

where N is the number of bins in the distribution, $\mu = (\mu_1, ..., \mu_N)$ is the signal strength in each bin, $\hat{\mu} = (\hat{\mu_1}, ..., \hat{\mu_N})$ are the corresponding best estimators, $\mathcal{L}(\mu)$ is the likelihood estimator and $\mathcal{L}(\hat{\mu})$ is the maximum likelihood estimator. Furthermore, s_i and b_i are the expected number of signal and background events, and n_i is the number of observed events.

The number of observed events, coming from a pseudo-data distribution, is randomly generated following a Poisson distribution, with mean equal to the sum of signal and background events expected in each bin for $\mu = 1$.

The distributions of $-2\,ln\,\lambda(\mu)$ for the LHC and HL-LHC scenarios, with an integrated luminosity of 300 fb^{-1} and 3000 fb^{-1}, respectively, are presented in Figure 7.1. It can be seen that the error on the signal strength decreases in the HL-LHC scenario, essentially due to the larger amount of data expected.

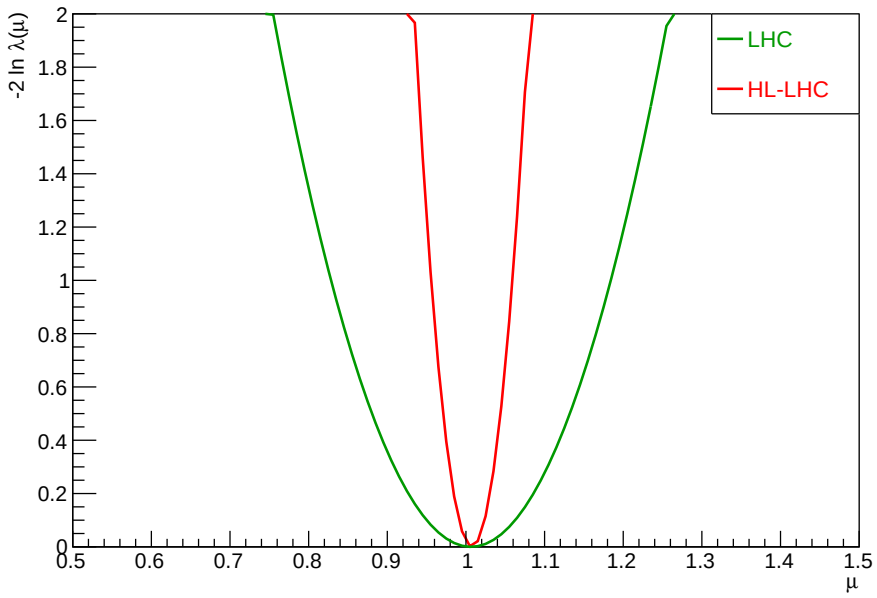

Figure 7.1: $-2\,ln\,\lambda(\mu)$ distribution for the LHC and HL-LHC scenarios, with an integrated luminosity of 300 fb^{-1} and 3000 fb^{-1}, respectively.

The values for the obtained signal strengths are shown in Table 7.2. An uncertainty on the signal strength of 18% is expected in the LHC scenario, using the optimized strategy, while this error decreases to 5% in the HL-LHC scenario.

Table 7.2: Signal strength integrated for different luminosities and scenarios, using the optimized strategy.

Scenario	$\mathcal{L}$ (fb^{-1})	Signal strength (μ)
LHC	300	0.99 ± 0.18
HL-LHC	3000	1.00 ± 0.05

7.2 Top Yukawa Coupling Measurement Uncertainty

The $t\bar{t}H$ cross section is proportional to the top Yukawa coupling squared, by

$$\sigma_{t\bar{t}H} = k\, y_t^2 \Leftrightarrow y_t = \left(\frac{\sigma_{t\bar{t}H}}{k}\right)^{\frac{1}{2}} \tag{7.2}$$

where k includes all the factors associated to a cross section computation. Considering that k has no errors associated, the uncertainty on the coupling is equal to

$$\Delta y_t = \frac{1}{2\sqrt{k\sigma_{t\bar{t}H}}} \Delta\sigma_{t\bar{t}H} \tag{7.3}$$

Having in mind that the product of the integrated luminosity $\mathcal{L}$ by a process cross section is equivalent to the number of events, $\mathcal{L}\sigma = N_{events}$, the uncertainty on the coupling reduces to

$$\Delta y_t = \frac{1}{2\sqrt{k\sigma_{t\bar{t}H}}} \frac{\Delta N_S}{\mathcal{L}} \tag{7.4}$$

where ΔN_S is the error on the number of $t\bar{t}H$ signal events, and $\mathcal{L}$ is considered not to have an associated error. The relative uncertainty on the top Yukawa coupling is then

$$\frac{\Delta y_t}{y_t} = \frac{1}{2}\frac{\Delta N_S}{\mathcal{L}\sigma_{t\bar{t}H}} = \frac{1}{2}\frac{\Delta N_S}{N_S} \tag{7.5}$$

The number of $t\bar{t}H$ events can be determined subtracting the number of background events, N_B, from the total number of events, N_T, i.e, $N_S = N_T - N_B$. The associated error is then

$$\Delta N_S = \sqrt{\Delta N_T^2 + \Delta N_B^2} \tag{7.6}$$

On the other hand, $N_B = \alpha N_{side}$, where N_{side} is the number of events in the sidebands for the Higgs candidates mass distribution, and α is a scaling factor. The number of signal events is expected to be negligible in the sidebands, but is nevertheless considered in this procedure. The error on the number of background events is then

$$\Delta N_B = \sqrt{\frac{N_B^2}{N_{side}^2}\Delta N_{side}^2} \tag{7.7}$$

For clarification, the error on each number of events is firstly computed from the non-renormalized number of events N, using $\sqrt{N}$, and is then scaled for the desired integrated luminosity.

The relative uncertainty on the top Yukawa coupling was then determined for the LHC and HL-

LHC scenarios, using the optimized analysis strategy, and are presented in Table 7.3. An integrated luminosity of 300 fb^{-1} was considered for the LHC case, as it is the expected integrated luminosity by the end of Run 3. The HL-LHC scenario considered an integrated luminosity of 3000 fb^{-1}, the expected collection of data after the HL-LHC ten years of operation. Furthermore, signal and background events were collected between 60 and 160 GeV, while the sidebands take into account events in the mass range [0,60[and]160,300] GeV.

Table 7.3: Relative uncertainty on the coupling of the Higgs boson to the top quark, using the optimized strategy in the LHC and HL-LHC scenarios. Integrated luminosities of 300 fb^{-1} and 3000 fb^{-1} are considered, respectively.

Scenario	$\mathcal{L}$ (fb^{-1})	$\Delta y_t/y_t$ (%)
LHC	300	35
HL-LHC	3000	17

A 35% uncertainty on the coupling of the Higgs boson to the top quark is expected by the end of the LHC Run 3, using the optimized strategy. This uncertainty decreases to 17% when implementing this strategy in the HL-LHC scenario, considering the whole dataset expected to be collected throughout ten years of operation.

7.3 $t\bar{t}A$ Observation

A search for the $t\bar{t}A$ production was also conducted for the HL-LHC scenario, and the significance and S/B for this case was computed, using the optimized strategy. The results for different integrated luminosities are presented in Table 7.4, reproduced from Table 6.5. The values for the $t\bar{t}H$ production are also presented, for comparison. The values for these variables were computed, as before, for Higgs candidate jets with masses ranging from 60 to 160 GeV.

Table 7.4: Significance and S/B for different integrated luminosities and processes, for the HL-LHC scenario. Computed from masses in range [60,160] GeV. Each bin in the mass distribution is considered to have a $\sqrt{N}$ error, where N is the number of events in that bin. The significance error results of the quadratic error propagation of $S/\sqrt{B}$.

Strategy	$\mathcal{L}$ (fb^{-1})	Significance ($S/\sqrt{B}$)	S/B (%)
$t\bar{t}H$	36	2.12 $\pm$ 0.04	15.7 $\pm$ 0.4
$t\bar{t}A$	36	1.63 $\pm$ 0.03	12.1 $\pm$ 0.3
$t\bar{t}H$	300	6.13 $\pm$ 0.11	15.7 $\pm$ 0.4
$t\bar{t}A$	300	4.71 $\pm$ 0.10	12.1 $\pm$ 0.3
$t\bar{t}H$	3000	19.39 $\pm$ 0.33	15.7 $\pm$ 0.4
$t\bar{t}A$	3000	14.90 $\pm$ 0.32	12.1 $\pm$ 0.3

The $t\bar{t}A$ process would not be observed with 300 fb^{-1} in the HL-LHC, contrary to what would happen for the $t\bar{t}H$ production, as the associated significance is only 4.71 ± 0.10. In fact, the observation of the BSM production would require at least 350 fb^{-1} of integrated luminosity collected at the HL-LHC, with an expected significance of 5.09 ± 0.10.

Chapter 8

Conclusions

An analysis strategy for the semileptonic $t\bar{t}H, (H \rightarrow b\bar{b})$ channel is proposed in this thesis. It starts by requiring an isolated charged lepton, and the calorimeter towers are clustered in large radius jets. The BDRS Higgs tagger is used to identify possible Higgs candidates among these jets, and two b-tags are required. Moreover, the strategy asks for two extra b-tags, using small radius jets, to account for the bottom quarks coming from the top quark decays. Cuts are also applied on variables related to jet hadronic substructure information, to further suppress backgrounds.

This optimized strategy improves the analysis significance by a factor 3 with respect to an implemented strategy based on Reference [71], in the HL-LHC scenario. In fact, is seems to be inefficient to tag the top quark unless very high p_T regions are targeted. Moreover, re-clustering jets can be used in the analysis, without affecting the results. This strategy can also be implemented in the LHC scenario, with slightly lower significances mainly due to the worse b-tagging.

The $t\bar{t}H$ process could be observed in this channel with an integrated luminosity of 300 fb^{-1} in the LHC scenario, using the optimized strategy, with a significance of 5.41 ± 0.12. The same integrated luminosity is associated to a significance of 6.13 ± 0.11 for the HL-LHC case. A pure pseudo-scalar, however, would not be observed with this integrated luminosity, having an associated significance of only 4.71 ± 0.10 for the HL-LHC scenario, using the optimized strategy.

An uncertainty on the $t\bar{t}H$ signal strength of 18% is expected in the LHC scenario, using the optimized strategy and with an integrated luminosity of 300 fb^{-1}. This error then decreases to 5% in the HL-LHC scenario, with an integrated luminosity of 3000 fb^{-1}.

A control region, to further constrain the backgrounds, is proposed, with a S/B of only 0.5%. This region targets events with Higgs candidate jets with two anti-b-tagged subjets. Apart from that, the strategy is the same as the optimized one.

The top Yukawa coupling is expected to have a 35% uncertainty by the end of the LHC Run 3, using the optimized strategy and considering an integrated luminosity of 300 fb^{-1}. This uncertainty decreases to 17% in the HL-LHC scenario with an integrated luminosity of 3000 fb^{-1}, again with the optimized strategy.

This work does not consider the effects of pile-up, neither uses a full simulation of the detector.

The analysis sensitivity is expected to decrease when introducing these realistic effects, as the object reconstruction will be more difficult to perform.

Nonetheless, some precautions were taken in order to minimize the impact of pile-up. The proposed analysis strategy requires an isolated lepton, which helps to suppress backgrounds. Moreover, the Higgs candidate jets are filtered to remove pile-up contamination, and four b-tags are required, further suppressing background processes. Working on the boosted regime also contributes to pile-up suppression.

The analysis is therefore expected to continue to be competitive in terms of significance, but this should be checked nevertheless.

A multivariate method could also be implemented, to further discriminate between the signal and the backgrounds. The different variable shapes could be exploited to larger extent using non-linear cuts than with simply linear ones.

In this sense, a set of input variables are proposed. While individually they do not contribute to a significant improvement in terms of significance, the MVA algorithm could efficiently use their combination.

The MVA should receive the τ_{21} and τ_{31} ratios for the Higgs candidate jets, along with the $C2$ [54] and $D2$ [55] energy correlation functions, using $\beta = 2$ instead of $\beta = 0.5$, as the former was found to be the best value for the parameter, slightly improving the significance. Moreover, the cuts on ΔR_{bb}, $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$ should be removed, and these variables should also be sent as input to the MVA.

Bibliography

[1] W. MissMJ. PBS NOVA, Fermilab, Office of Science, United States Department of Energy. Particle Data Group.

[2] D. H. Perkins. *Introduction to High Energy Physics*. Cambridge University Press, 4^{th} edition, 2000.

[3] P. W. Higgs. Broken symmetries, massless particles and gauge fields. *Physics Letters*, 12(2):132, 1964.

[4] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321, 1964.

[5] C. H. G.S. Guralnik and T. Kibble. Global Conservation Laws and Massless Particles. *Phys. Rev. Lett.*, 13:585, 1964.

[6] G. A. *et al.* (ATLAS Collaboration). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716(1):1–29, 2012. arXiv:1207.7214 [hep-ex].

[7] S. C. *et al.* (CMS Collaboration). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716(1):30–61, 2012. arXiv:1207.7235 [hep-ex].

[8] J. Ellis. Higgs Physics. 2013. KCL-PH-TH/2013-49, LCTS/2013-36, CERN-PH-TH/2013-315.

[9] M. A. *et al.* (ATLAS Collaboration). Measurement of the Higgs boson mass in the $H \to ZZ \to 4l$ and $H \to \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector. *Phys. Lett. B*, 784:345, 2018.

[10] M. A. *et al.* (ATLAS Collaboration). Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector. *Phys. Lett. B*, 784:173 – 191, 2018. arXiv:1806.00425v1 [hep-ex].

[11] A. S. *et al.* (CMS Collaboration). Observation of $t\bar{t}H$ Production. *Phys. Rev. Lett.*, 120:231801, 2018. arXiv:1804.02610v2 [hep-ex].

[12] M. A. *et al.* (ATLAS Collaboration). Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector. 2018. arXiv:1808.08238v1 [hep-ex], CERN-EP-2018-215.

[13] A. S. *et al.* (CMS Collaboration). Observation of Higgs boson decay to bottom quarks. 2018. arXiv:1808.08242v1 [hep-ex], CMS-PAS-HIG-18-016, CERN-EP-2018-223.

[14] M. A. *et al.* (ATLAS Collaboration). Cross-section measurements of the Higgs boson decaying to a pair of tau leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2018. ATLAS-CONF-2018-021.

[15] A. S. *et al.* (CMS Collaboration). Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett. B*, 779:283, 2018. arXiv:1708.00373v2 [hep-ex].

[16] C. Grojean. Higgs Physics. *CERN Yellow Report*, pages 143–158, 2017. CERN 2016-005, arXiv:1708.00794 [hep-ph].

[17] A. Collaboration. Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 97:072016, 2018. CERN-EP-2017-291, arXiv:1712.08895 [hep-ex].

[18] C. P. et al. (Particle Data Group). Particle Physics Booklet, 2016. Chin. Phys. C, 40, 100001.

[19] T. R. J. A. Heinson. Observation of Single Top Quark Production. *Annual Review of Nuclear and Particle Science*, 61, 2011. arXiv:1101.1275 [hep-ex].

[20] G. P. Salam. Elements of QCD for hadron colliders. *CERN Yellow Report*, pages 45–100, 2011. CERN-2010-002, arXiv:1011.5131v2 [hep-ph].

[21] R. J. Barlow. Jets in High-Energy Interactions. *Reports on Progress in Physics*, 56:1067–1143, 1993. arXiv:1602.04305 [hep-ex].

[22] M. G. B. M. Garcia, P. Musella and R. Harlander. CERN Report 4: Part I Standard Model Predictions. 2016. LHCHXSWG-DRAFT-INT-2016-008.

[23] A. Collaboration. Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector. *Phys. Rev. D*, 97:072003, 2018. arXiv:1712.08891 [hep-ex].

[24] D. F. *et al.*. Large pseudoscalar Yukawa couplings in the complex 2HDM. *JHEP*, 2015(6):60, 2015. arXiv:1502.01720 [hep-ph].

[25] G. C. B. *et al.*. Theory and phenomenology of two-Higgs-doublet models. *Phys. Rept.*, 516:1–102, 2012. arXiv:1106.0034 [hep-ph].

[26] T. D. Lee. A Theory of Spontaneous T Violation. *Phys. Lett. D*, 8:1226, 1973.

[27] S. Weinberg. Gauge Theory of CP Violation. *Phys. Rev. Lett.*, 37:657, 1976.

[28] V. K. *et al.* (CMS Collaboration). Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV. *Phys. Rev. D*, 92(1):012004, 2015. arXiv:1411.3441 [hep-ex].

[29] V. K. *et al.* (CMS Collaboration). Combined search for anomalous pseudoscalar HVV couplings in VH production and H to VV decay. *Phys. Lett. B*, 759:672, 2016. arXiv:1602.04305 [hep-ex].

[30] A. Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08003, 2008.

[31] A. Collaboration. The ATLAS Insertable B-Layer: from construction to operation. *Journal of Instrumentation*, 11(12):C12036, 2016. arXiv:1610.01994v3 [physics.ins-det].

[32] O. B. O. G. Apollinari, I. Béjar Alonso, P. Fessia, M. Lamont, L. Rossi, and L. Tavian. High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1. 2017. CERN-2017-007-M.

[33] A. Collaboration. ATLAS Phase-II Upgrade Scoping Document. 2015. CERN-LHCC-2015-020, LHCC-G-166.

[34] C. Collaboration. Technical Proposal For The Phase-II Upgrade Of The Compact Muon Solenoid. 2015. CERN-LHCC-2015-010, LHCC-P-008, CMS-TDR-15-02.

[35] M. A. *et al.* (ATLAS Collaboration). Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System. 2017. ATLAS-TDR-029, CERN-LHCC-2017-020.

[36] A. Collaboration. A High-Granularity Timing Detector (HGTD) in ATLAS: Performance at the HL-LHC. 2018. ATL-LARG-PROC-2018-003.

[37] P. Liu. Expected performance of the upgrade ATLAS experiment for HL-LHC. 2018. Talk presented at CIPANP2018, arXiv:1809.02181 [physics.ins-det].

[38] J. A. *et al.*. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 2014(7):79, 2014. arXiv: 1405.0301 [hep-ph].

[39] O. M. P. Artoisenet, R. Frederix and R. Rietkerk. Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations. *JHEP*, 2013(3):15, 2013. arXiv:1212.3460 [hep-ph].

[40] P. N. S. Frixione and C. Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 2007(11):070, 2007. arXiv:0709.2092 [hep-ph].

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

[42] T. G. *et al.*. Event generation with SHERPA 1.1. *JHEP*, 2009(02):007, 2009. arXiv:0811.4622 [hep-ph].

[43] S. M. T. Sjostrand and P. Z. Skands. A Brief Introduction to PYTHIA 8.1. *Computer Physics Communications*, 178:852, 2008. arXiv:0710.3820 [hep-ph].

[44] G. S. M. Cacciari and G. Soyez. FastJet user manual. *The European Physical Journal C*, 72(3): 1896, 2012. arXiv:1111.6097 [hep-ph].

[45] G. P. S. M. Cacciari and G. Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 2008(04):063, 2008. arXiv:0802.1189 [hep-ph].

[46] S. M. Y. L. Dokshitzer, G. D. Leder and B. R. Webber. Better Jet Clustering Algorithms. *JHEP*, 1997 (08):001, 1997. arXiv:hep-ph/9707323.

[47] M. Wobisch and T. Wengler. Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering. arXiv:hep-ph/9907280v1.

[48] J. de Favereau *et al.* (DELPHES 3 Collaboration). DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 2014(2):57, 2014. arXiv:1307.6346 [hep-ex].

[49] S. A. *et al.* (Geant4 Collaboration). GEANT4 A Simulation toolkit. *Nucl. Instrum. Meth.*, A506:250, 2003.

[50] P. A. *et al.*. A framework for Higgs characterisation. *JHEP*, 2013(11):43, 2013. arXiv:1306.6464v3 [hep-ph].

[51] M. A. *et al.* (ATLAS Collaboration). Technical Design Report for the ATLAS Inner Tracker Pixel Detector. 2017. ATLAS-TDR-030, CERN-LHCC-2017-021.

[52] K. V. T. J. Thaler. Identifying Boosted Objects with N-subjettiness. *JHEP*, 2011(3):15, 2011. arXiv:1011.2268v3 [hep-ph].

[53] M. S. S. Catani, Y..L. Dokshitzer and B. Webber. Longitudinally-invariant $k_\perp$-clustering algorithms for hadron-hadron collisions. *Nuclear Physics B*, 406(1):187, 1993.

[54] G. P. S. A. J. Larkoski and J. Thaler. Energy correlation functions for jet substructure. *JHEP*, 2013 (6):108, 2013. arXiv:1305.0007v3 [hep-ph].

[55] I. M. A. J. Larkoski and D. Nuff. Power counting to better jet observables. *JHEP*, 2014(12):9, 2014. arXiv:1409.6298v1 [hep-ph].

[56] G. P. S. T. Plehn and M. Spannowsky. Fat Jets for a Light Higgs Boson. *Phys. Rev. Lett.*, 104: 111801, 2010. arXiv:0910.5472v2 [hep-ph].

[57] T. S. G. Kasieczka, T. Plehn, T. Strebler, and G. P. Salam. Resonance Searches with an Updated Top Tagger. 2015. arXiv:1503.05921 [hep-ph].

[58] M. R. J. M. Butterworth, A. R. Davison and G. P. Salam. Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider. *Phys. Rev. Lett.*, 100:242001, 2008. arXiv:0802.2470v2 [hep-ph].

[59] M. A. *et al.* (ATLAS Collaboration). Measurements of Higgs boson properties in the diphoton decay channel with 36.1 fb^{-1} in pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2018. arXiv:1802.04146v1 [hep-ex].

[60] M. A. *et al.* (ATLAS Collaboration). Measurement of the Higgs boson coupling properties in the $H \to ZZ* \to 4l$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 2018(03):095, 2018. arXiv:1712.02304v2 [hep-ex].

[61] A. S. *et al.* (CMS Collaboration). Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton-proton collision at $\sqrt{s} = 13$ TeV. . CMS-HIG-17-026, CERN-EP-2018-065, arXiv:1804.03682v1 [hep-ex].

[62] A. S. *et al.* (CMS Collaboration). Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collision at $\sqrt{s} = 13$ TeV. . CMS-HIG-17-022, CERN-EP-2018-038, arXiv:1803.06986v2 [hep-ex].

[63] A. S. *et al.* (CMS Collaboration). Evidence for associated production of a Higgs boson with a top quark pair in final states with electrons, muons, and hadronically decaying τ leptons at $\sqrt{s} = 13$ TeV. . CMS-HIG-17-018, CERN-EP-2018-017, arXiv:1803.05485v1 [hep-ex].

[64] A. S. *et al.* (CMS Collaboration). Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. . CMS-HIG-16-040, CERN-EP-2018-060, arXiv:1804.02716v1 [hep-ex].

[65] A. S. *et al.* (CMS Collaboration). Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 2017(11):047, 2017. CMS-HIG-16-041, CERN-EP-2017-123, arXiv:1706.09936v2 [hep-ex].

[66] A. S. *et al.* (CMS Collaboration). Particle-flow reconstruction and global event description with the CMS detector. *Journal of Instrumentation*, 12(10):P10003, 2017. arXiv:1706.04965v2 [physics.ins-det].

[67] A. S. *et al.* (CMS Collaboration). Search for a standard model Higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method. *Eur. Phys. J.*, C75(6):251, 2015. arXiv:1502.02485v2 [hep-ex].

[68] S. P. A. dos Santos *et al.*. Angular distributions in $t\bar{t}H(H \to b\bar{b})$ reconstructed events at the LHC. *Phys. Rev. D*, 92:034021, 2015. arXiv:1503.07787v2 [hep-ph].

[69] S. P. A. dos Santos *et al.*. Probing the CP nature of the Higgs coupling in $t\bar{t}h$ events at the LHC. *Phys. Rev. D*, 96:013004, 2017. arXiv:1704.03565v1 [hep-ph].

[70] F. F. D. Azevedo, A. Onofre and R. Gonçalo. CP tests of Higgs couplings in $t\bar{t}h$ semileptonic events at the LHC. *Phys. Rev. D*, 98:033004, 2018. arXiv:1711.05292v2 [hep-ph].

[71] P. R. M. L. Mangano, T. Plehn, T. Schell, and H. Shao. Measuring the Top Yukawa Coupling at 100 TeV. *Journal of Physics G: Nuclear and Particle Physics*, 43(3):035001, 2016. arXiv:1507.08169v2 [hep-ph].

[72] G. Cowan. *Statistical Data Analysis*. Oxford University Press, 1998.

[73] A. S. B. Nachman, P. Nef, M. Swiatlowski, and C. Wanotayaroj. Jets from jets: re-clustering as a tool for large radius jet reconstruction and grooming at the LHC. *JHEP*, 2015(2):75, 2015. arXiv:1407.2922v2 [hep-ph].

Appendix A

Cut-flow tables

The cut-flow tables for the different strategies and scenarios presented in this thesis are presented in this section. The cut-flow table for the original strategy is presented in Table A.1, while Table A.2 shows the number of events after each step of the optimized strategy. Moreover, the cut-flow table for the optimized strategy with the re-clustering technique is presented in Table A.3. On the other hand, the cut-flow table for the control region is shown in Table A.4. Furthermore, the cut-flow tables for the the comparison with the LHC scenario are presented in Tables A.5 and A.6, respectively. Finally, the cut-flow table for the $t\bar{t}A$ search is shown in Table A.7.

A.1 Original Strategy

Table A.1: Cut-flow table for the original strategy and HL-LHC scenario. Events are normalized to 3000 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	204 000 $\pm$ 288	3 669 000 $\pm$ 2 594	48 000 $\pm$ 68	60 930 000 $\pm$ 38 536	51 420 000 $\pm$ 57 489	33 900 000 $\pm$ 37 901	6 180 000 $\pm$ 11 283	105 000 000 $\pm$ 88 741
Lepton	121 905 $\pm$ 223	2 219 862 $\pm$ 2 018	28 414 $\pm$ 52	29 527 434 $\pm$ 26 826	26 667 633 $\pm$ 41 401	16 369 886 $\pm$ 26 338	-	-
2 Fat Jets	30 775 $\pm$ 112	297 165 $\pm$ 738	7 158 $\pm$ 26	8 001 961 $\pm$ 13 965	68 581 $\pm$ 2 100	31 358 $\pm$ 1 153	5 914 919 $\pm$ 11 038	419 550 $\pm$ 5 609
1 Top Tag	9 598 $\pm$ 63	75 858 $\pm$ 373	2 085 $\pm$ 14	1 908 864 $\pm$ 6 821	2 700 $\pm$ 417	932 $\pm$ 199	295 960 $\pm$ 2 469	27 300 $\pm$ 1 431
≥ 1 $R = 1.2$ jet	7 878 $\pm$ 57	61 448 $\pm$ 336	1 762 $\pm$ 13	1 630 243 $\pm$ 6 303	2 185 $\pm$ 375	890 $\pm$ 194	288 771 $\pm$ 2 439	24 150 $\pm$ 1 346
1 Higgs Tag	621 $\pm$ 16	2 258 $\pm$ 64	188 $\pm$ 4	6 337 $\pm$ 393	64 $\pm$ 64	0 $\pm$ 0	103 $\pm$ 46	525 $\pm$ 198
≥ 1 $R = 0.6$ jet	542 $\pm$ 15	1 978 $\pm$ 60	168 $\pm$ 4	5 703 $\pm$ 373	64 $\pm$ 64	0 $\pm$ 0	21 $\pm$ 21	375 $\pm$ 168
3^{rd} b-tag	212 $\pm$ 9	710 $\pm$ 36	65 $\pm$ 2	463 $\pm$ 106	0 $\pm$ 0	0 $\pm$ 0	21 $\pm$ 21	0 $\pm$ 0
ΔR cut	212 $\pm$ 9	710 $\pm$ 36	65 $\pm$ 2	463 $\pm$ 106	0 $\pm$ 0	0 $\pm$ 0	21 $\pm$ 21	0 $\pm$ 0

A.2 Optimized Strategy

Table A.2: Cut-flow table for the optimized strategy and HL-LHC scenario. Events are normalized to 3000 fb^{-1}. 'N(X)' stands for the number of events of process X. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	204 000 $\pm$ 288	3 669 000 $\pm$ 2 594	48 000 $\pm$ 68	60 930 000 $\pm$ 38 536	51 420 000 $\pm$ 57 489	33 900 000 $\pm$ 37 901	6 180 000 $\pm$ 11 283	105 000 000 $\pm$ 88 741
Lepton	99 953 $\pm$ 202	1 790 202 $\pm$ 1 812	22 908 $\pm$ 47	24 166 373 $\pm$ 24 269	18 990 177 $\pm$ 34 937	12 079 545 $\pm$ 22 625	-	-
≥ 1 $R = 1.2$ jet	52 302 $\pm$ 146	638 085 $\pm$ 1 082	11 959 $\pm$ 34	14 165 640 $\pm$ 18 581	303 378 $\pm$ 4 416	153 101 $\pm$ 2 547	6 166 837 $\pm$ 11 271	1 620 300 $\pm$ 11 024
1 Higgs Tag	5 517 $\pm$ 47	24 691 $\pm$ 213	1 629 $\pm$ 13	141 796 $\pm$ 1 859	14 398 $\pm$ 962	7 289 $\pm$ 556	1 277 $\pm$ 162	52 275 $\pm$ 1 980
≥ 1 $R = 0.4$ jet	5 204 $\pm$ 46	22 955 $\pm$ 205	1 549 $\pm$ 12	126 442 $\pm$ 1 755	2 314 $\pm$ 386	1 441 $\pm$ 247	453 $\pm$ 97	23 625 $\pm$ 1 331
3^{rd} and 4^{th} b-tags	2 856 $\pm$ 34	11 988 $\pm$ 148	869 $\pm$ 9	12 405 $\pm$ 550	129 $\pm$ 91	254 $\pm$ 104	124 $\pm$ 50	525 $\pm$ 198
$\Delta R(b_3, b_4)$ cut	2 854 $\pm$ 34	11 988 $\pm$ 148	869 $\pm$ 9	12 405 $\pm$ 550	129 $\pm$ 91	254 $\pm$ 104	124 $\pm$ 50	525 $\pm$ 198
Opt cuts	2 670 $\pm$ 33	10 508 $\pm$ 139	803 $\pm$ 9	7 872 $\pm$ 438	0 $\pm$ 0	85 $\pm$ 60	0 $\pm$ 0	300 $\pm$ 150

A.3 Optimized Strategy with re-clustering

Table A.3: Cut-flow table for the optimized strategy with re-clustering and HL-LHC scenario. Events are normalized to 3000 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	204 000 $\pm$ 288	3 669 000 $\pm$ 2 594	48 000 $\pm$ 68	60 930 000 $\pm$ 38 536	51 420 000 $\pm$ 57 489	33 900 000 $\pm$ 37 901	6 180 000 $\pm$ 11 283	105 000 000 $\pm$ 88 741
Lepton	99 953 $\pm$ 202	1 790 202 $\pm$ 1 812	22 908 $\pm$ 47	24 166 373 $\pm$ 24 269	18 990 177 $\pm$ 34 937	12 079 545 $\pm$ 22 625	-	-
≥ 1 $R = 1.2$ jet	49 590 $\pm$ 142	590 078 $\pm$ 1 040	11 277 $\pm$ 33	13 071 630 $\pm$ 17 849	277 861 $\pm$ 4 226	142 889 $\pm$ 2 461	6 137 255 $\pm$ 11 244	1 228 575 $\pm$ 9 599
1 Higgs Tag	5 146 $\pm$ 46	20 680 $\pm$ 195	1 440 $\pm$ 12	120 373 $\pm$ 1 713	10 862 $\pm$ 836	5 339 $\pm$ 476	206 $\pm$ 65	44 025 $\pm$ 1 817
≥ 1 $R = 0.4$ jet	5 041 $\pm$ 45	20 099 $\pm$ 192	1 416 $\pm$ 12	114 622 $\pm$ 1 671	2 764 $\pm$ 421	1 526 $\pm$ 254	62 $\pm$ 36	18 900 $\pm$ 1 191
3^{rd} and 4^{th} b-tags	2 710 $\pm$ 33	9 971 $\pm$ 135	775 $\pm$ 9	7 458 $\pm$ 426	129 $\pm$ 91	85 $\pm$ 60	21 $\pm$ 21	150 $\pm$ 106
$\Delta R(b_3, b_4)$ cut	2 710 $\pm$ 33	9 971 $\pm$ 135	775 $\pm$ 9	7 458 $\pm$ 426	129 $\pm$ 91	85 $\pm$ 60	21 $\pm$ 21	150 $\pm$ 106
Opt cuts	2 658 $\pm$ 33	9 837 $\pm$ 134	763 $\pm$ 9	7 214 $\pm$ 419	129 $\pm$ 91	85 $\pm$ 60	0 $\pm$ 0	150 $\pm$ 106

A.4 Control Region

Table A.4: Cut-flow table for the control region and HL-LHC scenario. Events are normalized to 3000 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	204 000 $\pm$ 288	3 669 000 $\pm$ 2 594	48 000 $\pm$ 68	60 930 000 $\pm$ 38 536	51 420 000 $\pm$ 57 489	33 900 000 $\pm$ 37 901	6 180 000 $\pm$ 11 283	105 000 000 $\pm$ 88 741
Lepton	99 953 $\pm$ 202	1 790 202 $\pm$ 1 812	22 908 $\pm$ 47	24 166 373 $\pm$ 24 269	18 990 177 $\pm$ 34 937	12 079 545 $\pm$ 22 625	-	-
≥ 1 $R = 1.2$ jet	52 302 $\pm$ 146	638 085 $\pm$ 1 082	11 959 $\pm$ 34	14 165 640 $\pm$ 18 581	303 378 $\pm$ 4 416	153 101 $\pm$ 2 547	6 166 837 $\pm$ 11 271	1 620 300 $\pm$ 11 024
1 Higgs Tag	21 406 $\pm$ 93	263 856 $\pm$ 696	4 701 $\pm$ 21	6 661 696 $\pm$ 12 742	44 478 $\pm$ 1 691	24 281 $\pm$ 1 014	3 189 601 $\pm$ 8 106	378 375 $\pm$ 5 327
≥ 1 $R = 0.4$ jet	20 278 $\pm$ 91	238 595 $\pm$ 662	4 454 $\pm$ 21	6 196 947 $\pm$ 12 290	21 596 $\pm$ 1 178	11 780 $\pm$ 707	1 926 018 $\pm$ 6 299	269 250 $\pm$ 4 494
3^{rd} and 4^{th} b-tags	14 304 $\pm$ 76	137 798 $\pm$ 503	3 152 $\pm$ 17	5 2 719 720 $\pm$ 8 142	5 978 $\pm$ 620	2 966 $\pm$ 355	17 345 $\pm$ 598	90 300 $\pm$ 2 602
$\Delta R(b_3, b_4)$ cut	14 302 $\pm$ 76	137 784 $\pm$ 503	3 151 $\pm$ 17	2 719 355 $\pm$ 8 141	5 978 $\pm$ 620	2 966 $\pm$ 355	17 345 $\pm$ 598	90 300 $\pm$ 2 602
Opt cuts	13 466 $\pm$ 74	130 873 $\pm$ 490	2 964 $\pm$ 17	2 509 219 $\pm$ 7 820	5 463 $\pm$ 593	2 754 $\pm$ 342	14 482 $\pm$ 546	81 375 $\pm$ 2 470

A.5 Comparison with the LHC

Table A.5: Cut-flow table for the optimized strategy and LHC scenario. Events are normalized to 300 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	20 400 $\pm$ 29	366 900 $\pm$ 259	4 800 $\pm$ 7	6 093 000 $\pm$ 3 854	5 142 000 $\pm$ 5 749	3 390 000 $\pm$ 3 790	618 000 $\pm$ 1 128	10 500 000 $\pm$ 8 874
Lepton	11 306 $\pm$ 21	199 345 $\pm$ 270	2 565 $\pm$ 5	3 337 305 $\pm$ 4 753	2 009 237 $\pm$ 5 082	1 279 394 $\pm$ 3 293	-	-
≥ 1 $R = 1.2$ jet	6 238 $\pm$ 16	76 435 $\pm$ 167	1 412 $\pm$ 4	2 103 189 $\pm$ 3 773	31 906 $\pm$ 640	17 136 $\pm$ 381	617 186 $\pm$ 1 128	201 390 $\pm$ 1 626
1 Higgs Tag	599 $\pm$ 5	2 801 $\pm$ 32	172 $\pm$ 1	21 779 $\pm$ 384	1 465 $\pm$ 137	797 $\pm$ 82	124 $\pm$ 16	5 276 $\pm$ 263
≥ 1 $R = 0.4$ jet	566 $\pm$ 5	2 582 $\pm$ 31	164 $\pm$ 1	19 410 $\pm$ 362	103 $\pm$ 36	42 $\pm$ 19	62 $\pm$ 11	2 783 $\pm$ 191
3^{rd} and 4^{th} b-tags	311 $\pm$ 4	1 364 $\pm$ 22	94 $\pm$ 1	2 329 $\pm$ 126	26 $\pm$ 18	8 $\pm$ 8	21 $\pm$ 7	79 $\pm$ 32
$\Delta R(b_3, b_4)$ cut	311 $\pm$ 4	1 364 $\pm$ 22	94 $\pm$ 1	2 329 $\pm$ 126	26 $\pm$ 18	8 $\pm$ 8	21 $\pm$ 7	79 $\pm$ 32
Opt cuts	285 $\pm$ 3	1 170 $\pm$ 21	85 $\pm$ 1	1 469 $\pm$ 100	13 $\pm$ 13	8 $\pm$ 8	0 $\pm$ 0	26 $\pm$ 19

Table A.6: Cut-flow table for the optimized strategy and HL-LHC scenario. Events are normalized to 300 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}H$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	20 400 $\pm$ 29	366 900 $\pm$ 259	4 800 $\pm$ 7	6 093 000 $\pm$ 3 854	5 142 000 $\pm$ 5 749	3 390 000 $\pm$ 3 790	618 000 $\pm$ 1 128	10 500 000 $\pm$ 8 874
Lepton	9 995 $\pm$ 20	179 020 $\pm$ 181	2 291 $\pm$ 5	2 416 637 $\pm$ 2 427	1 899 018 $\pm$ 3 494	1 207 954 $\pm$ 2 262	-	-
≥ 1 $R = 1.2$ jet	5 230 $\pm$ 15	63 808 $\pm$ 108	1 196 $\pm$ 3	1 416 564 $\pm$ 1 858	30 338 $\pm$ 442	15 310 $\pm$ 255	616 684 $\pm$ 1 127	162 030 $\pm$ 1 102
1 Higgs Tag	552 $\pm$ 47	2 469 $\pm$ 21	163 $\pm$ 1	14 180 $\pm$ 186	1 440 $\pm$ 96	729 $\pm$ 56	128 $\pm$ 16	5 228 $\pm$ 198
≥ 1 $R = 0.4$ jet	520 $\pm$ 5	2 296 $\pm$ 21	155 $\pm$ 1	12 644 $\pm$ 176	231 $\pm$ 39	144 $\pm$ 25	45 $\pm$ 10	2 363 $\pm$ 133
3^{rd} and 4^{th} b-tags	286 $\pm$ 3	1 199 $\pm$ 15	87 $\pm$ 1	1 241 $\pm$ 55	13 $\pm$ 9	25 $\pm$ 10	12 $\pm$ 5	53 $\pm$ 20
$\Delta R(b_3, b_4)$ cut	285 $\pm$ 3	1 199 $\pm$ 15	87 $\pm$ 1	1 241 $\pm$ 55	13 $\pm$ 9	25 $\pm$ 10	12 $\pm$ 5	53 $\pm$ 20
Opt cuts	267 $\pm$ 3	1 051 $\pm$ 14	80 $\pm$ 1	787 $\pm$ 44	0 $\pm$ 0	8 $\pm$ 6	0 $\pm$ 0	30 $\pm$ 15

A.6 Pure Pseudo-scalar Case

Table A.7: Cut-flow table for the optimized strategy and HL-LHC scenario, with BSM signal sample. Events are normalized to 3000 fb^{-1}. 'N(X)' stands for the number of events of process X. Each error is computed as $\sqrt{N(X)}$. 'Opt cuts' refer to the variable cuts implemented in the end of the analysis strategy.

Cut	N($t\bar{t}A$)	N($t\bar{t}b\bar{b}$)	N($t\bar{t}Z$)	N($t\bar{t}j$)	N($W^+b\bar{b}$)	N($W^-b\bar{b}$)	N(dijets)	N($b\bar{b}j$)
-	96 000 $\pm$ 258	3 669 000 $\pm$ 2 594	48 000 $\pm$ 68	60 930 000 $\pm$ 38 536	51 420 000 $\pm$ 57 489	33 900 000 $\pm$ 37 901	6 180 000 $\pm$ 11 283	105 000 000 $\pm$ 88 741
Lepton	46 632 $\pm$ 180	1 790 202 $\pm$ 1 812	22 908 $\pm$ 47	24 166 373 $\pm$ 24 269	18 990 177 $\pm$ 34 937	12 079 545 $\pm$ 22 625	-	-
≥ 1 $R = 1.2$ jet	26 247 $\pm$ 135	638 085 $\pm$ 1 082	11 959 $\pm$ 34	14 165 640 $\pm$ 18 581	303 378 $\pm$ 4 416	153 101 $\pm$ 2 547	6 166 837 $\pm$ 11 271	1 620 300 $\pm$ 11 024
1 Higgs Tag	3 756 $\pm$ 51	24 691 $\pm$ 213	1 629 $\pm$ 13	141 796 $\pm$ 1 859	14 398 $\pm$ 962	7 289 $\pm$ 556	1 277 $\pm$ 162	52 275 $\pm$ 1 980
≥ 1 $R = 0.4$ jet	3 616 $\pm$ 50	22 955 $\pm$ 205	1 549 $\pm$ 12	126 442 $\pm$ 1 755	2 314 $\pm$ 386	1 441 $\pm$ 247	453 $\pm$ 97	23 625 $\pm$ 1 331
3^{rd} and 4^{th} b-tags	2 066 $\pm$ 38	11 988 $\pm$ 148	869 $\pm$ 9	12 405 $\pm$ 550	129 $\pm$ 91	254 $\pm$ 104	124 $\pm$ 50	525 $\pm$ 198
$\Delta R(b_3, b_4)$ cut	2 066 $\pm$ 38	11 988 $\pm$ 148	869 $\pm$ 9	12 405 $\pm$ 550	129 $\pm$ 91	254 $\pm$ 104	124 $\pm$ 50	525 $\pm$ 198
Opt cuts	1 959 $\pm$ 37	10 508 $\pm$ 139	803 $\pm$ 9	7 872 $\pm$ 438	0 $\pm$ 0	85 $\pm$ 60	0 $\pm$ 0	300 $\pm$ 150

Appendix B

Optimization Variables

Several variables were investigated in this thesis with the goal of discriminating between signal and background. Cuts on ΔR_{bb}, $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$ proved to be associated to larger increases in significance. The first section of this appendix presents the distributions for $\Delta R_{b_3,H}$ and $\Delta R_{b_4,H}$ in Figures B.1 to B.8, respectively, for the signal and the main backgrounds. The distribution for ΔR_{bb} is shown in the thesis's body. The following section contains the distributions for the proposed variables to be used as input to a MVA algorithm, namely τ_{21}, τ_{31}, and $C2$ and $D2$ with $\beta = 2.0$, for the signal and the main backgrounds.

B.1 Implemented variables

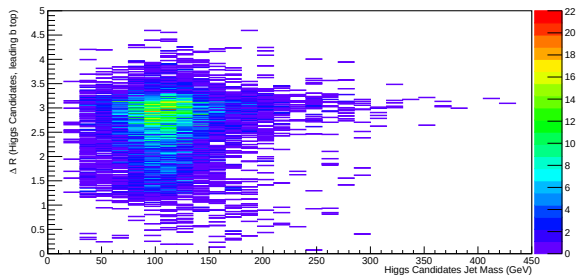

Figure B.1: $\Delta R_{b_3,H}$ for Higgs candidates for $t\bar{t}H$

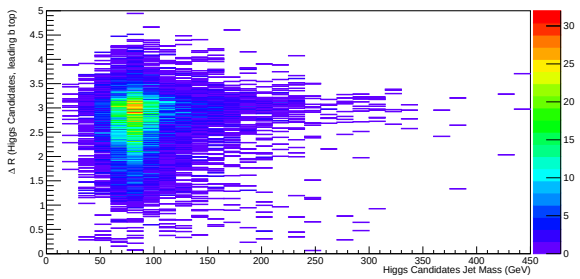

Figure B.2: $\Delta R_{b_3,H}$ for Higgs candidates for $t\bar{t}Z$

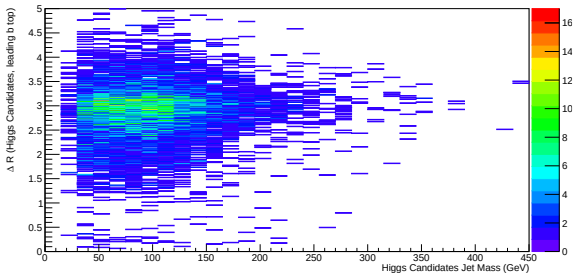

Figure B.3: $\Delta R_{b_3,H}$ for Higgs candidates for $t\bar{t}b\bar{b}$

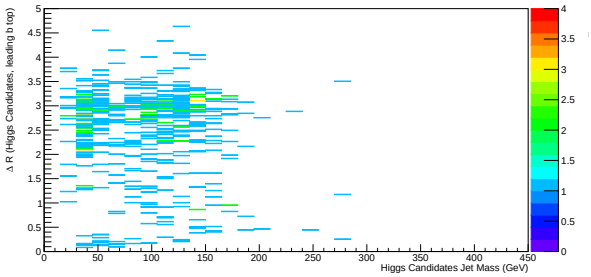

Figure B.4: $\Delta R_{b_3,H}$ for Higgs candidates for $t\bar{t}j$

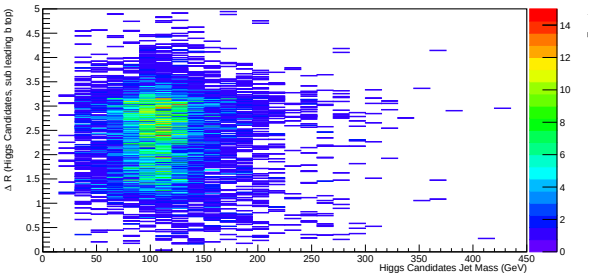

Figure B.5: $\Delta R_{b_4,H}$ for Higgs candidates for $t\bar{t}H$

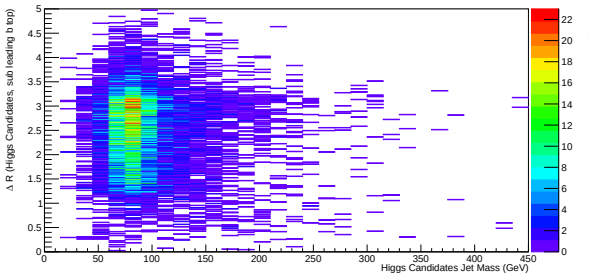

Figure B.6: $\Delta R_{b_4,H}$ for Higgs candidates for $t\bar{t}Z$

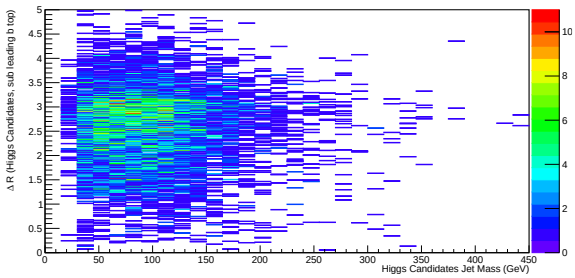

Figure B.7: $\Delta R_{b_4,H}$ for Higgs candidates for $t\bar{t}b\bar{b}$

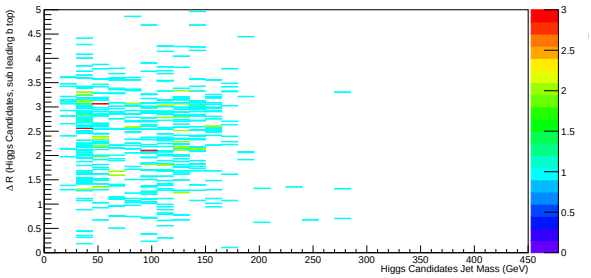

Figure B.8: $\Delta R_{b_4,H}$ for Higgs candidates for $t\bar{t}j$

B.2 Proposed MVA input variables

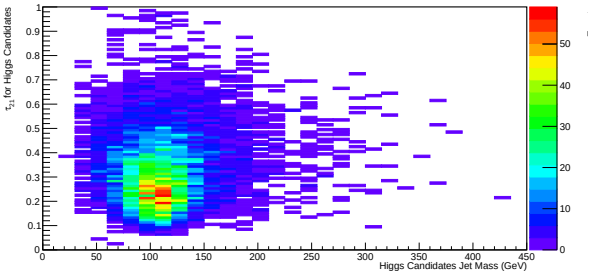

Figure B.9: τ_{21} for Higgs candidates for $t\bar{t}H$

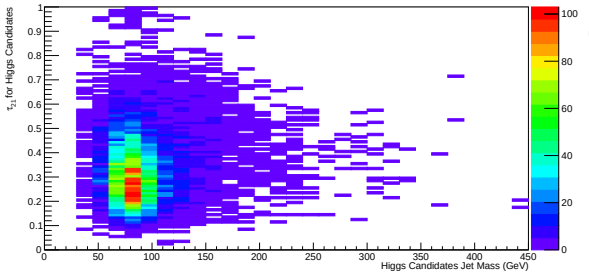

Figure B.10: τ_{21} for Higgs candidates for $t\bar{t}Z$

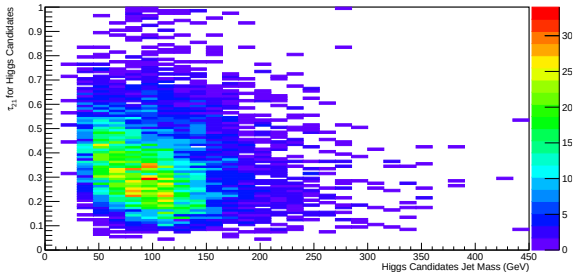

Figure B.11: τ_{21} for Higgs candidates for $t\bar{t}b\bar{b}$

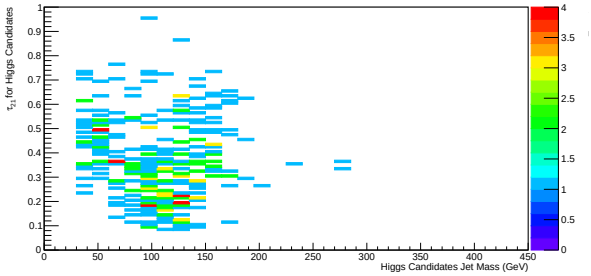

Figure B.12: τ_{21} for Higgs candidates for $t\bar{t}j$

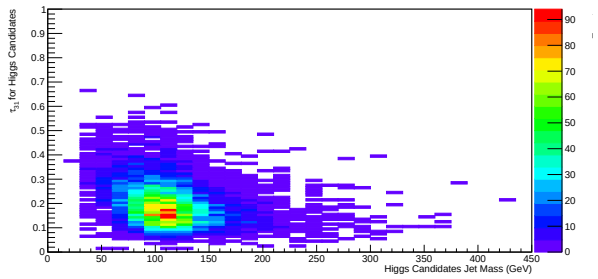

Figure B.13: τ_{31} for Higgs candidates for $t\bar{t}H$

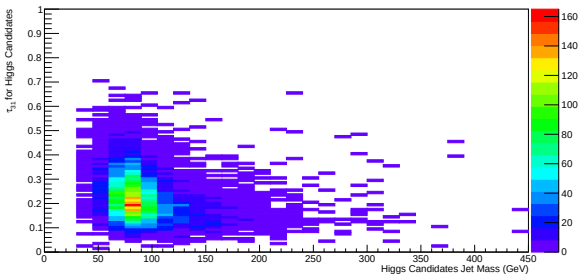

Figure B.14: τ_{31} for Higgs candidates for $t\bar{t}Z$

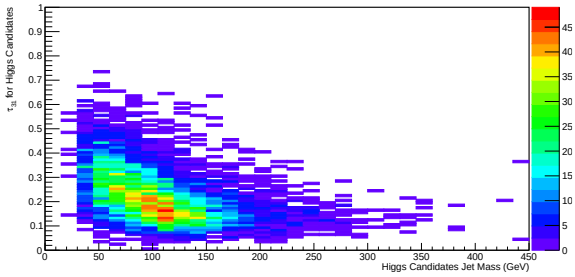

Figure B.15: τ_{31} for Higgs candidates for $t\bar{t}b\bar{b}$

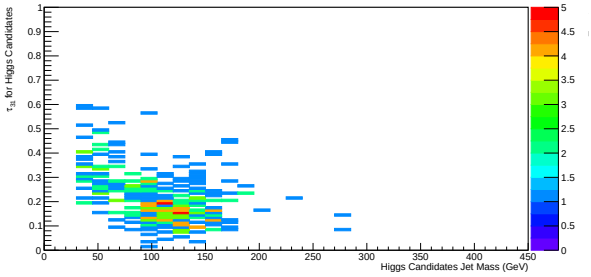

Figure B.16: τ_{31} for Higgs candidates for $t\bar{t}j$

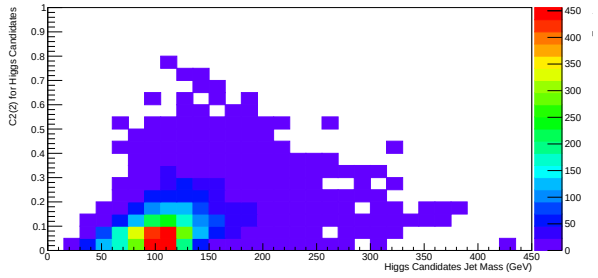

Figure B.17: C_2 for Higgs candidates for $t\bar{t}H$, with $\beta = 2.0$.

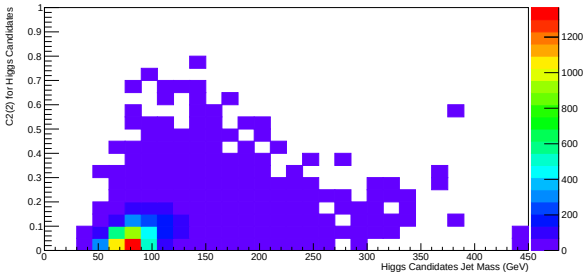

Figure B.18: C_2 for Higgs candidates for $t\bar{t}Z$, with $\beta = 2.0$

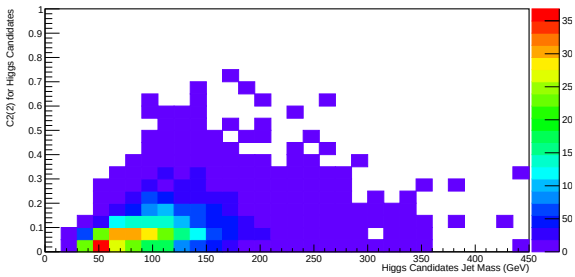

Figure B.19: C_2 for Higgs candidates for $t\bar{t}b\bar{b}$, with $\beta = 2.0$

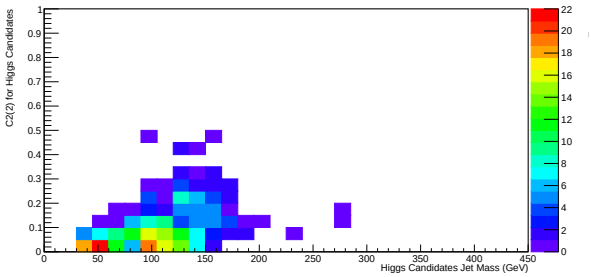

Figure B.20: C_2 for Higgs candidates for $t\bar{t}j$, with $\beta = 2.0$

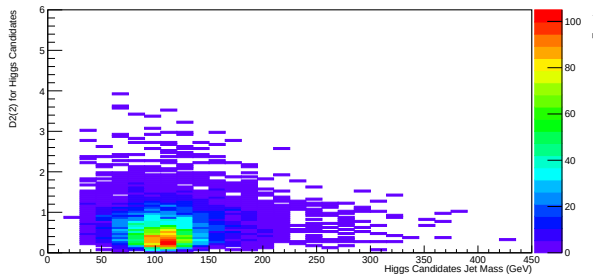

Figure B.21: D_2 for Higgs candidates for $t\bar{t}H$, with $\beta = 2.0$.

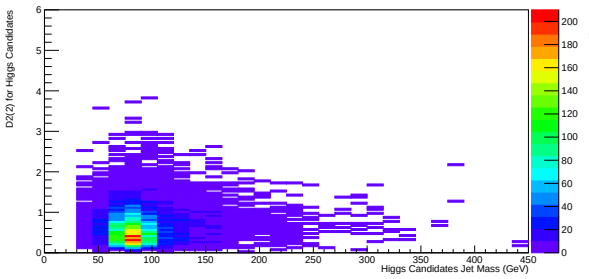

Figure B.22: D_2 for Higgs candidates for $t\bar{t}Z$, with $\beta = 2.0$

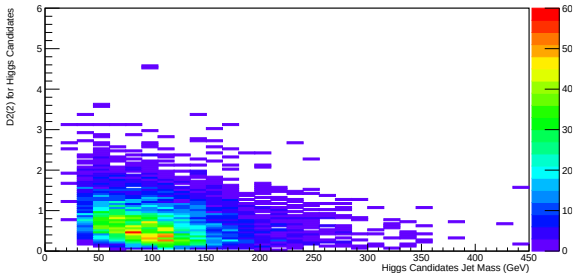

Figure B.23: D_2 for Higgs candidates for $t\bar{t}b\bar{b}$, with $\beta = 2.0$

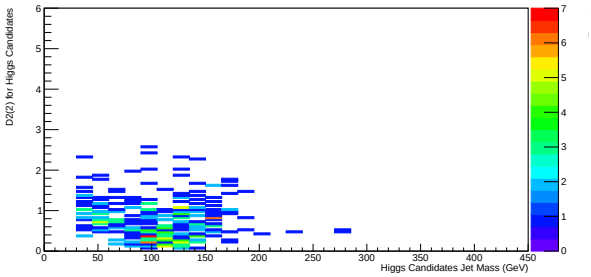

Figure B.24: D_2 for Higgs candidates for $t\bar{t}j$, with $\beta = 2.0$